UNIVERSITY OF CALIFORNIA, SAN DIEGO

Statistical models for RNA biology: from single nucleotides to single cells

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Boyko Kakaradov

Committee in charge:

Professor Gene Yeo, Chair
Professor Vineet Bafna, Co-Chair
Professor Trey Ideker
Professor Wei Wang
Professor Kun Zhang

2014

The dissertation of Boyko Kakaradov is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2014

DEDICATION

To my parents, who raised me to think independently and question authority, and to my mentors who managed to teach me something despite that!

EPIGRAPH

All models are wrong,
but some are useful.
—George E. P. Box

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables x

Acknowledgements xi

Vita xiii

Abstract of the Dissertation xiv

Chapter 1 Introduction 1
1.1 RNA processing 1
1.1.1 Alternative splicing 1
1.1.2 A-to-I editing 2
1.1.3 RNP code 4
1.2 Single-cell transcriptomics 5
1.3 Statistical Modeling 7
1.4 Specific contributions 7

Chapter 2 Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data 9
2.1 Introduction 10
2.2 Methods 12
2.2.1 RNA-seq data 12
2.2.2 Native model 14
2.2.3 Gaussian model 15
2.2.4 Bootstrap technique 16
2.2.5 Robust mixture model 18
2.2.6 Practical considerations 20
2.3 Results and discussion 21
2.3.1 Accurate estimation of PSI 21
2.3.2 Consistent estimation of PSI 23
2.3.3 Runtime and efficiency 24
2.4 Conclusion 24

2.5 Acknowledgements 26
2.6 Figures 26

Chapter 3 Adenosine to Inosine RNA editing in C. elegans 31
3.1 The dsRBP and inactive editor, ADR-1, utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome 31
3.2 Introduction 32
3.3 Results 34
3.3.1 ADR1 significantly alters RNA editing of multiple mRNA targets 34
3.3.2 ADR1 binds directly to ADR2 target mRNAs in vivo 35
3.3.3 ADR1 alters RNA editing via binding to dsRNA 36
3.3.4 Binding of dsRNA by ADR1 regulates editing of transcripts 37
3.3.5 ADR1 and ADR2 co-occupy multiple transcripts in vivo 40
3.4 Discussion 41
3.5 Experimental Procedures 43
3.5.1 Maintenance of worm strains and Transgenics . 43
3.5.2 RNA Isolation and Editing Assays 43
3.5.3 Strand-specific RNA sequencing 44
3.5.4 Bioinformatics Pipeline 44
3.5.5 RNA Immunoprecipitation (RIP) Assay 44
3.5.6 Flow Cytometry 45
3.6 Acknowledgements 45
3.7 Figures 45

Chapter 4 Single-cell analysis reveals asymmetric T cell specification during adaptive immunity 51
4.1 Early specification of $CD8^+$ T lymphocyte fates during adaptive immunity revealed by single-cell gene expression analyses 52
4.2 Results 53
4.2.1 Single-cell gene expression analyses of $CD8^+$ T lymphocytes *in vivo* 53
4.2.2 Molecular heterogeneity at the single-cell level early after infection 55
4.2.3 Distinct transcriptional signatures early after infection 57
4.2.4 Predicting temporal expression of key orchestrators of $CD8^+$ T cell fates 58

4.2.5 Asymmetric partitioning of IL2Rα in is associated with distinct cellular fates 60
4.3 Discussion 62
4.4 Methods 64
4.4.1 Mice 64
4.4.2 Adoptive cell transfers and infections 64
4.4.3 Microbead-based enrichment 65
4.4.4 Lymphocyte fate tracking experiments 65
4.4.5 Antibodies and flow cytometry 65
4.4.6 Single-cell gene expression assays 66
4.4.7 Statistical analysis 66
4.4.8 T lymphocyte confocal microscopy 66
4.4.9 Data and pre-processing 67
4.4.10 Principal component analysis (PCA) 67
4.4.11 T-distributed stochastic neighborhood embedding (tSNE) 67
4.4.12 Jensen-Shannon divergence 68
4.4.13 Rationale for approach to supervised analysis of gene expression data 69
4.4.14 Robust boosting 70
4.4.15 Temporal model of $CD8^+$ T cell differentiation . 71
4.5 Acknowledgements 74
4.6 Figures 74

Chapter 5 Discussion and Future Directions 81
5.1 Biology and statistics in the era of big data 81
5.2 From read-only to read+write biology 84
5.2.1 Synthetic Genomics 85
5.2.2 Genome Editing 86
5.2.3 Open questions on RNA editing 87
5.3 Disease Diagnostics 87

Appendix A Open thoughts on science 88

Appendix B RNA editing math 89
B.1 Introduction 89
B.2 Pipeline description 89
B.3 Details of Bayesian quantification model 91

LIST OF FIGURES

Figure 1.1: RNA processing and ADAR proteins 4

Figure 2.1: Read cover of sample junction 26
Figure 2.2: Plate model for Robust Mixture 27
Figure 2.3: Comparison of PSI estimates 28
Figure 2.4: Consistency of PSI estimates 29
Figure 2.5: Consistency ratios in different tissues 30

Figure 3.1: ADR-1 alters editing at specific adenosines in multiple mRNAs. 46
Figure 3.2: ADR-1 binds ADR-2 substrates in vivo. 47
Figure 3.3: Mutation of the KKxxK Motif within the dsRBDs of ADR-1 abolishes dsRNA binding and editing regulation. 48
Figure 3.4: Impact of dsRNA binding by ADR-1 on the editing transcriptome. 49
Figure 3.5: RNA editing graphical abstract 50

Figure 4.1: Gating strategy and experimental approach for single-cell gene expression analyses of $CD8^+$ T cells 75
Figure 4.2: Effector and memory $CD8^+$ T lymphocyte subsets are molecularly distinct on a single-cell level. 76
Figure 4.3: Early heterogeneity of gene expression exhibited by individual $CD8^+$ T lymphocytes during an immune response. 77
Figure 4.4: Classifier analysis predicts eventual fates of individual $CD8^+$ T lymphocytes. 78
Figure 4.5: Temporal model predicts the differentiation paths of individual $CD8^+$ T lymphocytes. 79
Figure 4.6: Asymmetric segregation of IL-2Rα during T lymphocyte division influences the eventual fates of the daughter cells 80

Figure 5.1: Scale vs. Accuracy tradeoff in experimental design with sequencing assays 82

Figure B.1: RNA editing prior 93

LIST OF TABLES

Table 2.1: Accuracy 24
Table 2.2: Runtime 25

Table 4.1: 94 selected gene targets grouped according to their function. . 54

ACKNOWLEDGEMENTS

Looking back at my graduate school journey, I see a path through many hills and valleys (both academic and personal). Standing at the beginning of this path six years ago, I envisioned a lonely trip through uncharted territory just waiting to be explored. I am amazed at the number of people that I have met on this path and their lasting influence on me. Standing at the other end now, I realize that independent from any individual success along this path, perhaps my most significant and hard-earned achievement is learning to share my journey with others in meaningful ways.

To my mentors, I owe not only my scientific maturity but also my sense of duty to apply it well. All of you have inspired me to be ingenious, and relentless with my research, both in the scientific and engineering aspects. Trey, you were the first to show me the difference between science and engineering, and together with Yoav, you both taught me that starting with a simple idea and building on it incrementally is better suited to research than recreating all the complexity of the latest techniques at once. Pavel and Brendan, you not only taught me to formulate problems well, but also to approach them from a new perspective, leading to original and inspired solutions. You along with Alex, also guided me through the most difficult time in my career so far. Gene, you took a risk on the "black sheep" of the program and gave a new home. You challenged me to expand my interests and expertise, and most of all, you taught me to interact with a wide range of scientists for whom my work was as foreign as it was useful. Thank you all!

To my friends and colleagues, I owe not only for the general camaraderie but also for my sanity. Eric, Jesús and Steph, you are my first friends at UCSD and I am glad that we have stayed close throughout it all. The PSI group was my academic home for the middle 2 years, and I will never forget our escapades, both academic and otherwise. The genomics journal clubs at UCSD and Salk have not only been a great way to keep abreast of many developments in the field, but also a tasty and stimulating alternative to the usual Friday happy hours on campus. Too many climbing/running/swimming/biking/hiking/camping buddies to name

have kept me cheerful when I needed a respite from work. Thank you all!

To my family, I owe not only my life but also my sense of relentless curiosity. Mamo i Tatko, vashata podkrepa me pravi silen i vashite patila me uchat da ne se razseivem ot tova koeto me interesuva nai mnogo. Nadiavam se che ste gordi ne samo sas moite postizenia, no i sas vashite postizenia koito se otraziavat va mene. Vassile, tvoia hus me uchi da ne se predavam dori kogato kartite sa sreshtu men. Blagodaria vi i na trite ohliuviatki!

Finally, I want to acknowledge funding and travel support for my research from the National Science Foundation Graduate Research Fellowship Program (NSF GRFP).

Chapter 2 is adapted from **B Kakaradov**, HY Xiong, LJ Lee, N Jojic, BJ Frey. Challenges in estimating percent inclusion of alternatively spliced exons from RNA-seq data. BMC Bioinformatics. (2012). The dissertation author was the primary author of this paper, and was responsible for the research.

Chapter 3 is adapted from M Washburn*, **B Kakaradov***, B Sundararaman, E Wheeler, S Hoon, G Yeo, H Hundley. The dsRBP and inactive editor ADR-1 utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome. Cell Reports. (2014). The dissertation author was a joint first author of this paper, and was responsible for all computational research.

Chapter 4 is adapted from J Arsenio*, **B Kakaradov***, PJ Metz, SH Kim, GW Yeo, JT Chang. Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. Nature Immunology. (2014). The dissertation author was a joint first author of this paper, and was responsible for all computational research.

VITA

2007	B. S. in Mathematics Stanford University
2008	M. S. in Computer Science *with distinction in research* Stanford University
2014	Ph. D. in Bioinformatics and Systems Biology University of California, San Diego

PUBLICATIONS

H Wang, **B Kakaradov**, SR Kollins, L Karotki, D Fiedler, D Koller. (2009). A complex-based reconstruction of the S. cerevisiae interactome. Mollecular and Cellular Proteomics, doi: 10.1074/mcp.M800490-MCP200

P Medvedev, E Scott, **B Kakaradov**, P Pevzner. (2011). Error correction of high-throughput sequencing datasets with non-uniform coverage. Bioinformatics, doi: 10.1093/bioinformatics/btr208

B Kakaradov, HY Xiong, LJ Lee, N Jojic, BJ Frey. (2012). Challenges in estimating percent inclusion of alternatively spliced exons from RNA-seq data. BMC Bioinformatics, doi:10.1186/1471-2105-13-S6-S11

C Lo, **B Kakaradov**, D Lokshtanov, C Boucher. (2014) A combinatorial approach to characterizing relationships between regulatory sequences. IEEE/ACM Transactions Comp. Biology and Bioinformatics, doi:10.1109/TCBB.2014.2304294

M Washburn*, **B Kakaradov***, B Sundararaman, E Wheeler, S Hoon, G Yeo, H Hundley. (2014). The dsRBP and inactive editor ADR-1 utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome. Cell Reports, doi:10.1016/j.celrep.2014.01.011 (*equal contribution)

J Arsenio*, **B Kakaradov***, PJ Metz, SH Kim, GW Yeo, JT Chang. (2014). Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. Nature Immunology, doi:10.1038/ni.2842 (*equal contribution)

ABSTRACT OF THE DISSERTATION

Statistical models for RNA biology:
from single nucleotides to single cells

by

Boyko Kakaradov

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2014

Professor Gene Yeo, Chair
Professor Vineet Bafna, Co-Chair

With the advent of RNA sequencing and other high-throughput molecular assays, RNA biology has recently transitioned from careful curation of single-hypothesis experiments to data-driven design of multi-hypothesis investigations. Fortunately, statistical advances and increasingly powerful computers have given rise to machine learning, a computational framework which can automatically distill perpetually growing datasets into predictive models of fundamental cellular and disease processes. Finally, recent advances in microfluidics have enabled the efficient capture and interrogation of individual cells by a variety of molecular assays. My research bridges theses fields by introducing predictive statistical models

of RNA abundance and processing in single cells to uncover new insights into the regulation of RNA editing and splicing and their effects on cellular differentiation.

This dissertation collects my contributions in single-cell and statistical genomics, from low-level details of data analysis to high-level principles of cellular identity and diversity. My early contributions concentrate on building error models of RNA sequencing data in order to extract biologically-relevant signals from experimental noise and sampling biases inherent in high-throughput sequencing technologies. Specifically, I describe statistical models of RNA splicing and editing that are robust to noise from PCR duplicates or sequencing errors and to uneven sampling from incomplete reverse transcription or cDNA fragmentation biases. I then evaluate the models' self-consistency and compare their accuracy relative to a gold standard. With a solid statistical foundation for sequencing data analysis established, my latest contributions focus on developing principled methods of constructing and evaluating compelling biological hypotheses in collaboration with domain experts. Specifically, I describe a Bayesian model of A-to-I RNA editing whose high specificity helped resolve the functional difference between the catalytically-active RNA binding protein ADR-2, and its inactive homolog ADR-1. In another collaboration, I used machine learning to resolve a long-standing question in immunology regarding the asymmetric specification of T cells into two functionally distinct lineages. Here, through one of the first applications of single-cell gene expression analysis of the immune system, I demonstrate that pathogen-activated T cells undergo an early bifurcation into effector- and memory-fated populations and help identify the genes whose asymmetric expression drive this phenomenon. Together all of these contributions establish a principled statistical framework for experimental design and analysis which integrates both hypothesis- and data-driven models to validate new findings and uncover novel principles of RNA biology.

Chapter 1

Introduction

Previously dismissed as a mere intermediary of genetic information between the nucleus and cytoplasm, RNA and cellular processes dedicated to it have been subject to extensive research efforts (Blencowe 2007, Sharp 2009, Licatalosi 2010). Much more than a transient vessel of genetic information, RNA is commanding a growing role in our understanding of fundamental cellular processes. Astonishing discoveries in the mechanisms of mRNA splicing and post-transcriptional regulation, microRNA biogenesis and targeting, as well as RNA stability and localization have implicated the transcriptome as a complex and dynamic system with more variation and adaptation than the genome itself (Blencowe 2007, Sharp 2009, Licatalosi 2010). RNA expression is finely regulated for tissue- and condition-specific functions via a variety of post-transcriptional processes such as splicing and editing.

1.1 RNA processing

1.1.1 Alternative splicing

Splicing of RNA is a widespread, highly-regulated, and well-studied process. In eukaryotes, the spliceosome binds to gene transcripts (either pre-mRNA or lncRNA), removes long stretches called introns, and joins the remaining (relatively short) exons into a processed transcript. Due to variations in the sequences

of exons and splicing factor binding affinities, variations in the splicing process increases protein diversity through the stochastic exclusion of introns and inclusion of exons from a gene transcript (Yeo 2004, Pan 2008). One such process is alternative splicing, which can produce a combinatorial number of variations from an unprocessed RNA transcript. While splicing itself happens to more than 90% of multi-exon genes in mammals, this thesis addresses a common and very well-studied form called alternative splicing. Alternative splicing is especially notable in the brain, likely due to the large number and length of multi-exon genes expressed primarily in neurons and astrocytes (Pan 2008, Barash 2010). During the alternative splicing of a pre-mRNA transcript, two constitutive exons which are included in every mature transcript surround an alternative exon, which is sometimes spliced out of the pre-mRNA together with its flanking introns. This process is tightly regulated by short (2-3 nucleotides) sequence motifs called splice donor and acceptor sites at the 5' and 3' ends of each intron, by longer (6-8 nt) motifs called splicing cis-regulatory elements that serve as binding sites for the spliceosome complex itself in the exonic/intronic sequence flanking the exon (Lo 2013). Splicing is also regulated by distant trans-regulatory motifs and RNA binding proteins (RBPs) such as NOVA and RBFOX2, which serve as either enchanters or repressors of splicing activity depending on the position of their binding sites (Barash 2010, Lovci 2014).

1.1.2 A-to-I editing

A less studied mechanism for additional fine-tuning of the transcriptome is through single-nucleotide RNA editing. The editing process can insert, delete, or modify single nucleotides within an RNA transcript. While there are various forms of RNA editing, this thesis addresses the most common form: Adenosine-to-Inosine (A-to-I) editing, which is catalyzed by the ADAR (Adenosine Deaminase Acting on RNA) family of enzymes. A-to-I editing is essential to proper brain function. RNA editing of ADAR targets in the brain has been well studied, due to the high amount of editing occurring in the nervous system; however, only a handful of RNA-editing targets have been validated and accepted by the community (Li 2011,

Pickrell 2012, Lin 2012). ADAR proteins are highly expressed in brain tissue, and mis-regulation of their A-to-I editing function in the nervous system often leads to deadly phenotypes. Additionally, ADAR transcripts and proteins steadily increase during various developmental stages of the brain (Jacobs 2009). It is well known that post-transcriptional modifications are highly prevalent in the brain, resulting in an incredible amount of transcript diversity, which is required for normal brain function. Accordingly, there are likely many more ADAR targets in the brain than the few that have been revealed. While A-to-I editing is predicted to occur in the coding regions of more than 10,000 human genes (Xiao 2011), the function of RNA editing is only known in the case for a few well-studied RNA editing sites. Among them is the well-known GluR2 Q/R site, which is detailed below.

One of the most prominent examples of A-to-I editing is the Q/R site on exon 11 of the Glutamate Receptor Subunit 2 protein (GluR2). Catalyzed by the double-stranded RNA binding protein (dsRBP) ADAR2, the A-to-I conversion introduces a non-synonymous amino-acid substitution (CAG→CGG, hence Q→R) in the protein Glur2, one of the subunits for the glutamate receptor (AMPA2) which regulates calcium ion channels in neurons. This edit is conserved exclusively in brain all the way from humans to *C. elegans*. In healthy neurons, the GluR2 Q/R site is edited with 100% efficiency, while editing sites in other transcripts allow a mix of both the A and I isoform to be expressed. While inactive in other tissues, the edit renders the AMPA receptor impermeable to calcium ions in the absence of glutamate. Similar to knockout phenotype of ADR2 in worms, neurodegenerative diseases in human reduce the editing efficiency of the Q/R site, which allows Ca^{2+} ions to flood into the synaptic channels and trigger neuronal death (Maas 2006). Reduced levels of editing in the GRIA2 Q/R site are associated with defective calcium channels in patients with Amyotrophic Lateral Sclerosis (ALS), where the amount of mis-editing correlates with disease progression and neurodegenerative phenotype severity (Kawahara 2004).

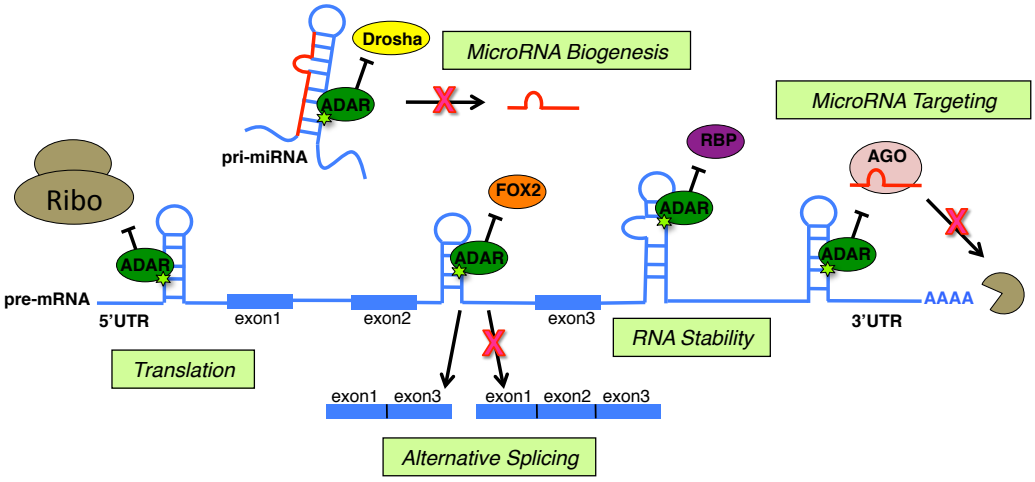


Figure 1.1: Summary of ADAR affects, both steric and catalytic.
ADAR proteins have the ability to block many cellular processes including miRNA biogenesis, miRNA targeting, RNA stability, alternative splicing, and translation by blocking other regulatory RBPs through A-to-I editing (lime star) or solely through binding.

1.1.3 RNP code

Post-transcriptional processes such as alternative splicing and A-to-I editing are regulated by a collection of interacting RNA binding proteins (RBPs) adorning various binding sites on each RNA molecule. From their transcription and splicing, through editing and silencing, to polyadenilation, export, and eventual degradation, RNAs are adorned and regulated by companion proteins which bind to specific target sites in a sequence- and structure-dependent manner. These RBPs and their targets form the mRNP code which dictates how each RNA in the nucleus will be processed, translated, or degraded (**Figure 1.1**). Disruptions in the mRNP code such as abnormal changes in alternative splicing and loss of RNA editing form the basis of many human cancers and neurological diseases such as Autism, Alzheimer's, and ALS [**?**, **?**].

ADAR proteins are essential for survival, and studying their targets and interactions with other RBPs is crucial to understanding post-transcriptional processing. In mammals, there are three ADAR proteins: ADAR1, ADAR2, and ADAR3. While they have very similar gene structure, they are expressed in differ-

ent tissues, localize to different parts of the nucleus and cytoplasm, and most importantly have sequence- and structure-specific preferences in their RNA targets. All three ADAR proteins carry multiple double-stranded RNA binding motifs: three in ADAR1, and two in both ADAR2 and ADAR3. ADAR1 also uniquely carries two Z-DNA-binding domains (Bass 2002). ADAR1 and ADAR2 are expressed in most tissues, with the exception of skeletal muscle (Heale 2009). ADAR2 is most highly expressed in the nervous system, and ADAR3 is only expressed in the brain (Chen 2000).

ADAR1 and ADAR2 bind to dsRNA as homodimers to convert an adenosine to an inosine though the selective or non-selective deamination of substrates. ADAR3 lacks this catalytic activity; however, it has been suggested that ADAR3 may compete with the catalytically active ADAR proteins for binding to substrates (Chen 2000), or dimerize with them to sequester them away from their RNA targets (Valente 2007, Cenci 2008). Most RNA editing is found to occur non-selectively and in high abundance at inverted Alu repetitive elements within introns and untranslated genic regions (Blow 2004). The hyperediting often found in these regions is thought to affect localization of the target, preventing nuclear export, and stability (Hundley 2010). RNA editing occurs in less abundance, but more specifically in coding regions, leading to synonymous and nonsynonymous codon changes (Hundley 2010). While the ADAR reaction seems simple, single base A-to-I editing can have a profound effect in several cellular processes. An inosine in an RNA transcript is recognized by all downstream events as a guanosine (G), and an A-to-I edit will be read as an A-G mismatch between the genomic DNA and complimentary DNA (cDNA). Consequently, A-to-I editing may affect the secondary structure of target mRNAs, splice site selection, miRNA targeting, and translation of amino acids crucial for protein function.

1.2 Single-cell transcriptomics

Why was single-cell sequencing named 'method of the year' in the January 2014 issue of Nature Methods? Because peering into the molecular contents of an

individual cell is not only very technically challenging, but also potentially very revealing. Measuring gene expression from individual cells as opposed to bulk samples allows us to notice stochastic differences and core similarities between cells of the same type in the context of other cell types and clarify the molecular definition of cell identity and state. Even more importantly, the natural variation present between cells of the same type serves as a statistical re-sampling which can unveil higher-order population statistics such as variance and skew in ways that simple bulk samples cannot. Any rare cells with unique molecular contents such as pluripotent or malignant cells will be lost in the pool of more ordinary cells. Moreover, the true heterogeneity of a cell type will always be underestimated by the bulk samples. There is a fundamental mathematical reason that encapsulates both of these cases and shows why n samples of individual cells will be more informative than n samples of $m >> 1$ cells each. The central limit theorem, also known as the law of large numbers, states that the mean of n independent and identically distributed random variables $x_1, \ldots, x_n$, no matter what their original distribution is (as long as it has finite mean and variance), will approximate the Gaussian distribution with mean equal to the sample mean

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} x_i = \mu$$

and variance which is n times smaller than the individual sample variance

$$\sigma_n^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n^2} = \frac{\sigma^2}{n}$$

Therefore, any biologically-relevant non-gaussian distribution which contains multiple modes (cellular subtypes) or even has wider tail (higher than normal heterogeneity) will be reflected in single-cell sampling, but will be lost in the bulk sampling! Applying high-throughput sequencing and microfuidic technologies to determine the expression and processing of RNA in single cells necessitates experimental design that is closely coupled to powerful statistical analysis, in order to harness the potential for increased precision and scale. Finally, the statistical

power of this framework for high-throughput single-cell experiments has enabled us to construct more sophisticated and precise definitions of cellular identity and function, and apply them to tracking the molecular phenotypes of T cells though immune specification and of iPS cells though neuronal differentiation.

1.3 Statistical Modeling

Many scientists rightfully mistrust statistical models – experimental biologists often because they don't fully understand the details, and computational biologists because they do. Therefore, a successful statistical model not only needs to capture as many relevant details of the known science, but also frame it's novel predictions in the context of known entities that can be interpreted by biologists and validated by statisticians. One of the dangerous pitfalls in both science and medicine is to gather data and run tests without a hypothesis because random fluctuations may be interpreted as interesting anomalies. At the same time, one of the dangerous pitfalls in statistics is to apply a model without understanding its assumptions because any prediction of the model will be biased if the real constraints on the data don't match those assumptions.

1.4 Specific contributions

In the following chapters, I capture various technical and biological signals with a variety of statistical and probabilistic models. Among them, my technical contributions consist of either designing, adapting, or simply applying various statistical models for RNA sequencing data to capture different biological signals, depending on their intended hypothesis. A majority of these models are called generative because they generate a joint probability distribution over their inputs x and any known labels t. They include: several types of error models for PCR duplication and fragmentation bias that robustly estimate the level of alternative splicing from RNA sequencing data in Chapters 2; a new Bayesian model of RNA sequencing errors that robustly estimates the level of RNA editing in Chapter

3; and adapting existing statistical tools such as Principal Components Analysis (PCA) and Hidden Markov Models (HMM) to capture single-cell gene expression through a differentiation time course in Chapter 4. My conclusions and future research directions are contained in Chapter 5.

Chapter 2

Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data

Transcript quantification is a long-standing problem in genomics and estimating the relative abundance of alternatively-spliced isoforms from the same transcript is an important special case. Both problems have recently been illuminated by high-throughput RNA sequencing experiments which are quickly generating large amounts of data. However, much of the signal present in this data is corrupted or obscured by biases resulting in non-uniform and non-proportional representation of sequences from different transcripts. Many existing analyses attempt to deal with these and other biases with various task-specific approaches, which makes direct comparison between them difficult. However, two popular tools for isoform quantification, MISO and Cufflinks, have adopted a general probabilistic framework to model and mitigate these biases in a more general fashion. These advances motivate the need to investigate the effects of RNA-seq biases on the accuracy of different approaches for isoform quantification. We conduct the investigation by building models of increasing sophistication to account for noise introduced by the biases and compare their accuracy to the established approaches.

We focus on methods that estimate the expression of alternatively-spliced isoforms with the percent-spliced-in (PSI) metric for each exon skipping event. To

improve their estimates, many methods use evidence from RNA-seq reads that align to exon bodies. However, the methods we propose focus on reads that span only exon-exon junctions. As a result, our approaches are simpler and less sensitive to exon definitions than existing methods, which enables us to distinguish their strengths and weaknesses more easily. We present several probabilistic models of of position-specific read counts with increasing complexity and compare them to each other and to the current state-of-the-art methods in isoform quantification, MISO and Cufflinks. On a validation set with RT-PCR measurements for 26 cassette events, some of our methods are more accurate and some are significantly more consistent than these two popular tools. This comparison demonstrates the challenges in estimating the percent inclusion of alternatively spliced junctions and illuminates the tradeoffs between different approaches.

2.1 Introduction

Determining the relative abundance of gene transcripts in a cell − whether in relation to each other or in relation to corresponding transcripts in other cells − is an important and long-standing problem in genomics. Since introduction of RNA-seq, a high-throughput experimental method of measuring the RNA content of a sample by reverse-transcribing it and sequencing the resultant cDNA, this problem has been illuminated by vast amounts of data and by many methods for elucidating transcript abundance (Mortazavi 2008). Current collections of RNA-seq data are rapidly growing in multiple dimensions such as species, tissues, and conditions (Wang 2009).

This data deluge necessitates more sophisticated and accurate analysis methods, which in turn create an opportunity to gain deeper insights into the role and regulation of transcript abundance in important developmental and disease processes. Undoubtedly, one important research area that can benefit from these advances is the study of RNA splicing, an essential cellular process that effectively increases the phenotypic complexity of eukaryotic organisms without necessitating an increase in their genetic complexity. Accurate measurements of

the expression levels for isoforms from a large number of genes are especially useful for research into the molecular mechanisms that regulate alternative splicing in different tissues. For example, the recent advances in the RNA splicing code that determines the relative abundance of alternatively spliced isoforms (Barash 2010) was made possible by high-throughput microarray technology. In principle, RNA-seq can lead to much richer datasets at a fraction of the cost. Thus RNA-seq technology can lead to significant new breakthroughts, as the code quality achieved by (Barash 2010) leaves a lot of room for improvement. The focus of this paper − estimation of the percent inclusion of alternatively-spliced exons from RNA-seq data − is a step toward a more accurate interpretation of the natural splicing code. This problem is complicated by several sources of bias in short read counts including those due to the cDNA fragmentation and primer amplification steps of current RNA-seq protocols (Roberts2011, Hansen 2012). These biases lead to widely varying abundances for reads from different positions in the transcript. We investigate this position-specific bias further and suggest methods to mitigate it.

Specifically, we restrict our interest only to exon-skipping events(Pan 2008, Katz 2010). The numerical quantity which captures relevant information for these events is termed percent-spliced-in (PSI). For each exon-skipping event, PSI is defined as the expression of isogorms containing the alternatively spliced exon (i.e. those containing a given cassette exon and its flanking constitutive exons) as a fraction of the total expression for both alternatively and constitutively spliced isoforms (i.e. those containing the flanking exons only) which is reported in percent. Accurate estimation of PSI is not only desirable on its own, but it can also be used to improve the resolution of differential splicing and thus improve the predictive power of the splicing code (Barash 2010).

There are several recent tools for estimating relative abundance of isoforms, which deal with position-specific biases in different ways (Katz 2010, Roberts2011, Nicolae 2011, Turro 2011). MISO can directly estimate PSI specifically for exon-skipping events (Katz 2010), while most others estimate the expression of whole isoforms from which a PSI value may be derived. This makes MISO the natural point of reference for our comparisons, but we also include Cufflinks (Roberts 2011)

in the comparisons because of its popularity and explicit modeling of fragmentation and amplification biases. However, for the task of estimating PSI, Cufflinks' focus on multi-exon isoforms appears to be detrimental, as we show in the Results section.

Our pursuit of robust estimates for PSI necessitates an appropriate measure of the uncertainty for these estimates. This additional necessity is crucial for the task of deciphering the natural RNA splicing code. Linking noisy RNA-seq read counts with the sequence determinants of RNA splicing is a hard task that requires good measurement of splicing levels even in case of transcripts with minimal coverage. For this task it is just as important to quantify the range of possible PSI values supported by the RNA-seq data, given that the position-specific bias can dramatically influence these estimates. We start by framing the classic IID sampling assumption as a Poisson model and modify it to mitigate the effect of position-specific biases. This leads to three methods of increasing complexity. We evaluate our models in terms of their accuracy and consistency. We compare our methods' accuracy to each other and to existing approaches of estimating PSI with respect to a reference set of 26 RT-PCR measurements from a human cell line. As we discussed above, we are interested in developing algorithms that provide robust estimates: A handful of highly biased positions in the transcript, from which a much larger number of reads is obtained simply due to fragmentation bias, should not unduly influence the estimate of PSI. Our results show a moderate increase in accuracy and a significant increase in consistency of our methods over the current state of the art methods for quantifying of alternative splicing events.

2.2 Methods

2.2.1 RNA-seq data

RNA-seq data was generated from a HeLa cell line by the Blencowe Lab at the University of Toronto [**?**]. The protocol consisted of polyA-selected RNA extraction, random hexamer primed reverse transcription, cDNA fragmentation (with mean insert size of 220nt), and 50nt paired-end sequencing by Illumina GA.

This dataset is publicly available on the NCBI Gene Expression Omnibus with accession number GSE26463. 305 million RNA-seq reads were sequenced and mapped to the reference human genome (NCBI build37, UCSC hg19) using TopHat, which is capable of reporting split-read alignments across splice junctions (Trapnell 2009). TopHat produced error-free alignments for 66 million reads (about 22% of the total). For each exon-exon junction, the reads that overlapped it by at least 8nt were selected and their positions were noted. Positions that contained reads mapping elsewhere were excluded. The number of 3' fragment ends (i.e. reads starts) around the junction was tabulated into a profile of read hits for each junction. This profile of read start counts is also called a read cover, in contrast to the more popular read coverage.

Figure 2.1 illustrates the actual cover profile for a representative constitutive (i.e. exclusion) junction with a relatively high total number of reads. Position-dependent biases in the read cover lead to positions with zero reads, as well as positions with many mode reads than are expected based on other positions. These two situations are sometimes treated differently, but they are essentially due to the same cause: position-dependent effects. Note that these position-dependent effects are present in the majority of junctions regardless of their underlying expression. Another source of error is mis-matched reads but, in this work, we deliberately used only error-free alignments (as opposed to the common practice of allowing a small number of mismatches) in order to differentiate the positional biases from mismatch noise. When estimating PSI, the individual read covers for each pair of alternative junctions that flank an alternative exon can be tabulated into a joint inclusion junction cover using half-counts at each position. This is common practice for analyses of alternative splicing as it is assumed that the increased sample size results in better estimates of expression. However, we note that averaging the read covers for the two alternative junctions is not appropriate when the constitutive annotation of the two flanking exons is in question, and this approach does not significantly reduce the harsh effects of positional biases.

The existing tools for isoform quantification, MISO and Cufflinks were provided with the entire alignment, not just the reads mapping to junctions. MISO

(version 0.2) and Cufflinks (version 1.2) were run with default parameters except for the paired-end read insert size and the number of samples from the appropriate posterior, which were set to 220 and 10000, respectively.

2.2.2 Native model

The first model we study makes the simplifying assumption that reads are sampled independently and identically distributed (IID) from the expressed isoforms. We refer to it as the "Native" model, because its key component, the Poisson arrival process, is a natural model for IID read coverage. This "Native" model has worked sufficiently well in the past for analysis in many respectable DNA and RNA sequencing studies (Wang 2009).

Many simple models of RNA-seq data assume, either explicitly or implicitly, that reads are sampled uniformly along the length of a transcript (Mortazavi 2008, Trapnell 2009). However, actual RNA-seq data do not follow this assumption because of multiple sequence- and position-specific biases inherent in the cDNA library preparation and sequencing (Srivastava 2010, Hansen 2010, Katz 2010, Roberts 2011). Still, we might expect this assumption to hold for sufficiently short regions on a transcript, such as the neighborhood around an exon-exon junction. In this case, the number of read starts x_p mapping to each position p near the junction should follow a Poisson distribution whose mean is estimated by $\tilde{\alpha} = \frac{1}{P}\sum_p x_p$ where the region of interest spans positions $\{1, 2, \ldots P\}$. The mean and matching variance α will estimate both the overall expression for that junction and the model's uncertainty in that expression. Unfortunately, reads are not distributed uniformly, even along short regions with sufficient coverage. As shown on Figure 2.1, the read counts covering the region within 50nt of a representative constitutive junction are highly variable and non-uniform. The corresponding cover for the two alternative junctions (not shown) contains about twice as many read counts in total, but they are split over two neighborhoods of 50nt. In general, RNA-seq data deviates from the Native Poisson model in two ways:

- the high sparsity of the data ($\sim$ 80% of positions have no reads starting at them) causes $\tilde{\alpha}$, the average cover for the region, to underestimate the

expected abundance α.

- the variance of the non-zero elements $x_p > 0$ is three times larger than that dictated by the Native model.

Note that the Poisson model describes the likelihood $P(x_p \mid \alpha)$ of observing a particular read cover profile x_p given the unknown expression α. However, we are interested in the posterior probability $P(\alpha \mid x_p)$ of the hidden expression given the observed data. This posterior can be obtained from the likelihood of the observed data and the prior over the expression through the classic Bayes' Rule:

$$P(\alpha \mid x) = \frac{P(x \mid \alpha) * P(\alpha)}{P(x)} \tag{2.1}$$

Once we have distributions over the expected expression for both the alternative (a.k.a. inclusion) and the constitutive (a.k.a exclusion) junctions, α^i and α^e respectively, we combine them to produce the posterior over the PSI estimate of this model $P(\Psi_{\text{Native}}|x_p^i, x_p^e)$ given the observed read counts over the inclusion (x_p^i) and exclusion (x_p^e) junctions, respectively. There is no closed-form expression for this distribution, but we can estimate it with the ratios of samples from the inclusion and exclusion posteriors:

$$P(\Psi_{\text{Native}}|x_p^i, x_p^e) \propto \sum_{\substack{\alpha^i, \alpha^e: \\ \frac{\alpha^i}{\alpha^i+\alpha^e}=\Psi_{\text{Native}}}} P(\alpha^i \mid x_p^i) * P(\alpha^e \mid x_p^e) \tag{2.2}$$

2.2.3 Gaussian model

In order to alleviate the shortcomings of the Native model, we propose two simple modifications which result in a new Gaussian model that is more robust to the position-specific biases present in RNA-seq data. To deal with the sparse cover and its effect on the expected expression, α, we dismiss all unmappable positions, i.e. those positions which coincide with the start of reads that map elsewhere in the reference genome or transcriptome. This leaves only the set of position indexes Q which coincide with the hits of only uniquely-mappable reads. Therefore, the normalized expression of a junction is $\gamma = \frac{1}{|Q|} \sum_{q \in Q} x_q + \frac{1}{P}$ where we have added

the pseudo-count $\frac{1}{P}$ in order to avoid dividing by zero for junctions which have no uniquely-mappable reads, e.g. those that come from homologous regions of the genome.

To deal with the high variance at positions with non-zero read count, we approximate the PSI ratio of normalized junction expressions with a Gaussian distribution. Unlike the Poisson distribution whose mean and variance are identical by definition, the link between the mean and variance of this Gaussian approximation can be relaxed in order to make the model more robust. The mean μ is estimated by the ratio of the normalized read counts for the inclusion and exclusion junctions (γ^i and γ^e, respectively). The standard deviation σ is proportional to the geometric mean of μ and its complement $1-\mu$. The variance σ^2 is normalized by the total number of uniquely mappable reads in the alternative and constitutive junction $\Gamma = \gamma^i|Q^i| + \gamma^e|Q^e|$, where $|Q^i|$ is the number of uniquely-mappable positions for the inclusion junction, and $|Q^i|$ is that for the exclusion junction. Finally, the variance is lower-bounded by an arbitrary threshold in order to avoid over-fitting the noisy RNA-seq data:

$$\tilde{\mu} = \frac{\gamma^i}{\gamma^i + \gamma^e} \qquad \tilde{\sigma}^2 = \max\left[0.01, \frac{\tilde{\mu}(1-\tilde{\mu})}{\Gamma}\right] \tag{2.3}$$

This approximation allows us to skip the Bayesian procedure and sampling approximation required by the Native model, since we can directly specify the posterior distribution of our estimate for PSI given the read counts around a junction: $P(\Psi_{\text{Gaussian}} \mid x_{p'}) \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$.

2.2.4 Bootstrap technique

To robustly estimate PSI without explicitly modeling sequence and position dependent bias, we propose a method based on randomly resampling the observed data. This method computes the degree of uncertainty in PSI by estimating the consistency within the observed dataset. It belongs to a general class of statistical methods called bootstraping that have been successfully used to model complex and unknown distributions (Davison 1997).

The bootstrap can be used to assess the uncertainty in the PSI estimates produced by any method that takes position-dependent read counts as input. Here, we use a Poisson model. We assume that there are P mappable junction positions for each exon skipping event. We observe x_p^i inclusion reads and x_p^e exclusion reads for each position $p = \{1, 2, ...P\}$. To estimate PSI from such a dataset, a simple approach assumes that for every position, x_p^i and x_p^e are generated by a Poisson distribution with real-valued underlying abundances β^i and β^j respectively. A Poisson distribution is used to model the process of how RNA-seq reads in each position arise from the true abundance of isoforms in the biological sample. Because of the IID assumption, the maxmimum likelihood (ML) estimator of β is simply the sum of the observed reads. Instead of simply using the ML estimator, we take a Bayesian approach where we assume an improper prior for $P(\beta) = 1$ for the abundances of both inclusion and exclusion variants. The posterior of β is a Gamma distribution with a shape parameter equal to 1:

$$
\begin{aligned}
P(\beta) &= 1; && (2.4)\\
P(\vec{x}|\beta) &= \prod_k P(x_p|\beta); && (2.5)\\
P(x_p|\beta) &= \text{Poisson}(x|\beta); && (2.6)\\
&= \frac{\beta^{x_p}}{x_p!}e^{-\beta}; && (2.7)\\
P(\beta|\vec{x}) &\propto P(\beta)P(\vec{x}|\beta); && (2.8)\\
&\propto \frac{\beta^{\sum_p x_p}}{(\sum_p x_p)!}e^{-\beta}; && (2.9)\\
P(\beta|\vec{x}) &= \text{Gamma}(1, 1 + \sum_p x_p), && (2.10)
\end{aligned}
$$

where $\text{Gamma}(\theta, k)$ denote the real valued Gamma distribution with scale parameter θ and shape parameter k. In this application, the shape parameter is one plus the sum of the reads across positions. The Gamma random variable in the above equation incorporates our belief of likely values of isoform abundances (β) given the observed reads, with the IID assumption for read generation across positions. However, the IID assumption described above is highly incorrect, because of position-dependent effects introduced by RNA-seq technologies. We use the

bootstrap to assess the uncertainty induced by these effects as follows. Instead of summing over the reads at all positions, we generate a sample of P positions with replacement from the observed data and then sum the reads at those positions to produce an estimate of β as described above.

The above procedure is repeated to generate a distribution of β estimates, which can be used to form a distribution of PSI. In our approach, one million β^i and β^e are generated with which one million samples of $\Psi_{\text{bootstrap}}$ are produced.

2.2.5 Robust mixture model

We propose a robust mixture model of read counts that span alternatively-spliced junctions from exon skipping events. The mixture has three components:

1. A zero-cover component to explain the empty positions arising from sparse fragmentation bias.

2. A noise component to capture the read stacks arising from the other type of positional bias.

3. A Poisson component to capture the remaining signal in the read cover.

Formulating a mixture model allows us to explicitly capture each of the two types of bias alongside the underlying signal in RNA-seq data.

For each cassette splicing event, our model links the hidden expression counts λ^i and λ^e, for the inclusion and exclusion junctions, to the unknown PSI and coverage values: $\Psi_\lambda \in \mathbb{Q}$ and $C \in \mathbb{Z}$, and to the observed read counts: $x_p^i \in \mathbb{Z}$ and $x_p^i \in \mathbb{Z}$ where $p \in \{1, 2, \ldots P\}$ are positions in the neighborhood of each junction. As before, Ψ_λ, C, and λ are linked by a deterministic relationship:

$$\Psi_\lambda = \frac{\lambda^i}{C} \quad \text{where} \quad C = \lambda^i + \lambda^e \tag{2.11}$$

Figure 2.2 shows the plate diagram for the Robust Mixture model. Its priors and factors are described in the following sections. The the priors and factors combine via Bayes' Rule (already described in Equation (2.1)) to give the posterior distribution over the hidden variables and mixture weights of this model.

Priors

- PSI: $\Psi_\lambda \sim \text{Uniform}[0,1]$
  even though the empirical distribution is closer to a convex Beta distribution with preference for extreme values of Ψ_λ, we use the least informative prior in order to gain the most information about this hidden variable of interest [**?**].

- Cover: $C \sim \text{Gamma}(\theta, k)$
  with scale parameter $\theta = 77.77$ and shape parameter $k = 0.77$ estimated from C's empirical distribution.

- Expression: A complex prior on λ^i and λ^e is induced by the priors on Ψ_λ and C through the relation in equation (2.11). We impose no further restriction on the distribution of these hidden variables.

- Mixture: The weights of the three mixture components represent the relative strengths of the signal and the two noise models. The observed sparsity of RNA-seq data (where 80% of junction-neighboring positions have no read alignments starting from them) is an upper bound on the true sparsity because we expect to see zero-cover positions in junctions with very low expression. Therefore we chose 60% sparsity as a reasonable compromise. Likewise, the observed read-stack outlier rates for the Illumina platform is a lower bound on the actual fraction of outlier reads (3% of all junction-adjacent positions have a read count that is 10× higher than the simple average).

$$p_0(m_p) = \begin{cases} 0.60 & \text{Zero Cover } (m_p = 0) \\ 0.36 & \text{Poisson Model } (m_p = 1) \\ 0.04 & \text{Read Stacks } (m_p = 2) \end{cases} \tag{2.12}$$

Factors

- Deterministic: $\lambda^i, \lambda^e \sim \delta(\lambda^i = \Psi_\lambda * C)\delta(\lambda^e = C - \lambda^i)$

- Multinomial: $m_p \sim \text{Multinomial}(c_z, c_p, c_s)$
  This factor allows our model to learn the actual mixture weights for each of the components from the observed data.

- Mixture: We use a mixture factor in order to capture each of the two biases and the actual signal in separate components. The choice for each component is motivated by the form of the signal or noise it is designed to capture.

$$x_p \mid m_p, \lambda \sim \begin{cases} \delta(x_p = 0) & \text{Sparsity } (m_p = 0) \\ \text{Poisson}(\lambda) & \text{Signal } (m_p = 1) \\ \text{Uniform}[1, L] & \text{Noise } (m_p = 2) \end{cases} \tag{2.13}$$

2.2.6 Practical considerations

Performing inference in the Native and Robust Mixture models described above is intractable due to the complex partition function that normalizes the posterior distribution $P(\Psi|x_p)$. To compute the posterior, we could use advanced approximate inference methods such as Expectation Maximization used by IsoEM (Nicolae 2011), Markov Chain Monte Carlo used by MISO (Katz 2010), and combinatorial optimization used by Cufflinks (Trapnell 2010, Roberts 2011). However, we note that discretizing the values of their parameters allows us to approximate the partition function and directly calculate the posterior distribution over the discretized PSI values: Ψ_α and Ψ_λ respectively. In contrast, the Gaussian and bootstrap models give a posterior over Ψ_γ directly, either in a closed form expression or in the form of samples from a provably exact distribution. Figure 2.3 shows that the resulting posterior distributions for all PSI estimators are well-formed, especially for junctions with sufficiently high read cover, and gives support for the viability of our discretization scheme for junctions of medium or even low read cover. Finally, performing inference with discretized parameters takes considerably less time at a minimal loss of precision. This allows our methods to analyze an entire pre-aligned RNA-seq dataset in the manner of a few minutes, while other methods take tens of hours or even days on the same task.

2.3 Results and discussion

2.3.1 Accurate estimation of PSI

In order to evaluate the accuracy of our models and compare it to that of the existing methods, we selected a validation set of 26 cassette exons with reference PSI values derived from RT-PCR experiments in HeLa cells (Saltzman 2011). The 26 events include 11 high-expression events with between 10 and 20 read starts per position, 8 medium-expression events with about 1 read start per position, and 7 low-expression events with 10 or fewer reads total across all 50 positions (≤ 0.2 read starts per position). Figure 2.3 compares the posterior distributions over PSI inferred by six different methods: our four methods described in the Methods section, and two popular tools for isoform quantification, MISO and Cufflinks. All tools shared the same input, but were able to extract varying amount of information from it. The shared TopHat alignment file included the mapping of reads to a reference set constructed only from the constitutive and alternative exons of the 26 cassette events. Our tools were able to use only the reads mapping across junctions, while MISO and Cufflinks was free to use the entire set of alignments. Furthermore, our methods did not benefit from the paired-end dependencies between the reads, while both MISO and Cufflinks were able to do so. To be fair, we note that Cufflinks is designed for whole-transcript quantification. Thus, we did not expect it to be competitive with the other methods on a highly restricted reference set consisting of only three exons per alternative splicing event

While limited, this comparison clearly shows that no particular method outperforms the others on every event. However, it does suggest that our methods are more accurate, especially when they agree with each other. We investigate the consistency of our methods in a later part of the Results section. Unfortunately there is no canonical way to measure the error between a distribution estimate and a point target. However, we modify three existing distance metrics between distributions and propose a new metric which allow us to compute the overall performance of the six methods on all 26 events. Given a PDF distribution of PSI estimates $P(x)$ and a target value ψ described by discretized Gaussian distribution

$Q_\psi(x)$ centered at the point target, ψ. We used an arbitrary standard deviation $\sigma = 0.05$ which is comparable to the accuracy needed for downstream applications of PSI estimates. The new metric directly computes the distance between a distribution and its target.

- Variation distance, which measures the total deviation between the two distributions

$$V(P, Q_\psi) = \sum_{0 \le x \le 1} \mid P(x) - Q_\psi(x) \mid \tag{2.14}$$

- Disagreement distance between CDFs, which measures the maximum deviation. In our case, the maximum is attained at the mode of either P or Q_ψ

$$S(P, Q_\psi) = \max_{0 \le y \le 1} \sum_{0 \le x \le y} P(x) - Q_\psi(x) \tag{2.15}$$

- KL divergence, which measures the asymmetric disagreement between P or Q_ψ with respect to the latter

$$D_{\mathrm{KL}}(Q_\psi \| P) = \sum_{0 \le x \le 1} Q_\psi(x) \log \frac{Q_\psi(x)}{P(x)} \tag{2.16}$$

- Novel confidence-weighted $L_{\frac{1}{2}}$ error distance, is designed to penalize distributions that distribute weight away from the target ψ

$$E_{\frac{1}{2}}(P, \psi) = \sum_{0 \le x \le 1} P(x) \| x - \psi \|_{\frac{1}{2}} \tag{2.17}$$

Table 2.1 shows the overall performance of each PSI estimation method over the 26 target events according to each of these error metrics. While our most robust methods perform well on three of these metrics, it is not surprising that MISO outperforms every other method on the remaining S-metric because it always distributes its posterior mass wider than our methods. The disagreement distance, $S(P, Q_\psi)$ rewards this extensive hedging because it is very susceptible to sampling noise which is abundant on Figure 2.3. The remaining metrics are chosen to be more robust when faced with this sampling noise.

2.3.2 Consistent estimation of PSI

In order to further investigate the consistency of PSI estimation methods, we performed a random sub-sampling procedure. This procedure chooses a random half of the positions around a junction and uses the subset of reads that start at those positions to obtain an unbiased estimate of the noise associated with the positional bias. A dataset with reduced set of positions is equivalent to a dataset with reduced signal-to-noise ratio. Comparing the PSI estimate of a method given each half of the positions can measure the consistency of that method. Figure 2.4 depicts the consistency of the most accurate methods from Table 2.1 with a non-standard 2D color visualization. We call a this visualization a constellation plot because of its superficial resemblance to images of deep-space galaxies.

We expect more consistent methods to produce consistently more similar estimates of PSI. For each method, we calculate the KL-divergence between its PSI estimate on a particular event to the PSI estimate on all other events. We compare the mean of all cross-event divergence to the divergence between PSI estimates from complementary halves of the same event. The former divergence we call the inter-exon distance, and the latter we call the intra-exon distance. Then, the ratio between the inter- and intra-exon distances is a measure of the method's consistency for that particular exon. More consistent methods will have a higher ratio over all events. Figure 2.5 compares the consistency ratios of our four methods and that of MISO using a larger dataset of over 1000 events (including the 26 validated by RT-PCR).

Consistency of the PSI estimates is especially important to the downstream uses of our methods. If only a randomly selected subset of positions are taken into account, the PSI estimate (and its uncertainty) should be very similar to the estimate that would be computed based on the complementary set of transcript positions. Thus we defined a measure of consistency of the estimator as the ratio of the average distance of the PSI distributions obtained from two different genes and the average distance from PSI distributions obtained from different position subsets of the *same* transcript. High values of this ratio indicated that using a smaller subset of the positions will not affect the estimate of PSI drastically, but

that this is not achieved in a trivial way by always estimating either a high or a very low level of exon inclusion.

2.3.3 Runtime and efficiency

While accuracy and consistency are the most important considerations for any approach of estimating PSI, runtime and efficiency are becoming increasingly relevant as the amount of RNA-seq data grows rapidly. Table 2.2 compares the runtimes of all methods on both the small validation set of 26 events and the larger set of 1051 events. To estimate the distribution over PSI values for each event, we used 10,000 samples for all methods. Sampling from the Gaussian model was direct whereas other models sampled the expression for inclusion and exclusion isoforms separately. It is not surprising that the run time of our pre-processing grows linearly with the number of RNA-seq reads, and we expect the same happens to the pre-processing subroutines of both MISO and Cufflinks. However, the estimation subroutines in the two established tools are disproportionately slower on the larger dataset than any of our simple methods, including the robust and very consistent bootstrap model.

Table 2.1: Comparison of error between different PSI estimation methods with respect to RT-PCR target. The best methods with lowest error in each row are bolded. Robust Mixture model is abbreviated to "Mixture".

Error	Native	Gaussian	Mixture	Bootstrap	MISO	Cufflinks
V	28.5	**24.1**	27.2	**24.2**	30.9	43.7
S	12.90	15.26	15.87	15.22	**9.87**	12.65
D_{KL}	264	102	**94.2**	**92.0**	220	1115
$E_{1/2}$	9.34	7.08	**6.62**	**6.65**	9.28	14.65

2.4 Conclusion

This work addressed the problem of estimating relative abundances of alternatively spliced cassette exons from the sparse and noisy evidence in RNA-seq data. First, we investigated the raw data and reviewed known fragmentation biases resulting from current RNA-seq protocols. Next, we identified position-specific

Table 2.2: Comparison of run times between different PSI estimation methods. For our methods, we report the runtime of the shared pre-processing step separately from the PSI estimation. All tests were performed on a Dell Precision T7400 workstation with 8 cores (at 3 GHz) and 32 GB of RAM. We report wall-clock times averaged over 3 re-runs then rounded to the nearest minute (or second where appropriate).

Datasets:	Validation	High-Throughput
RNA-seq reads	66 Million	145 Million
AS events	26	1051
Cufflinks	16 min	75 min
MISO	77 min	458 min
Preprocess	4 min	11 min
Gaussian	+1 sec	+2 min
Native	+2 sec	+5 min
Mixture	+6 sec	+17 min
Bootstrap	+12 sec	+29 min

anomalies affected by these biases, and proposed a modular probabilistic framework that robustly estimates the PSI and total coverage of alternatively-spliced exon junctions. Using this foundation, we framed the classic IID read sampling assumption as a Poisson model and termed the two types of position-specific deviations in the actual data as sparse cover and read stacks. Using the established framework, we proposed three novel probabilistic methods of increasing complexity, which mitigate the effects of these two biases. We compared our methods' accuracy to each other and to existing approaches of estimating PSI with respect to a reference set of 26 RT-PCR measurements from a human cell line. Our results showed a moderate increase in accuracy and a significant increase in consistency of our methods over the current state-of-the-art for quantification of alternative splicing events. While we presented and referenced several methods for quantifying alternative splicing, our goal was not to pick a single champion that is superior to all others, but to compare the strengths and weaknesses of the various approaches. We hope that these advances will enable more sensitive downstream analyses, such as better determinants of differential splicing which can eventually lead to an improved RNA splicing code.

2.5 Acknowledgements

This chapter is adapted from **B Kakaradov**, HY Xiong, LJ Lee, N Jojic, BJ Frey. Challenges in estimating percent inclusion of alternatively spliced exons from RNA-seq data. BMC Bioinformatics. (2012). The dissertation author was the primary author of this paper, and was responsible for the research.

2.6 Figures

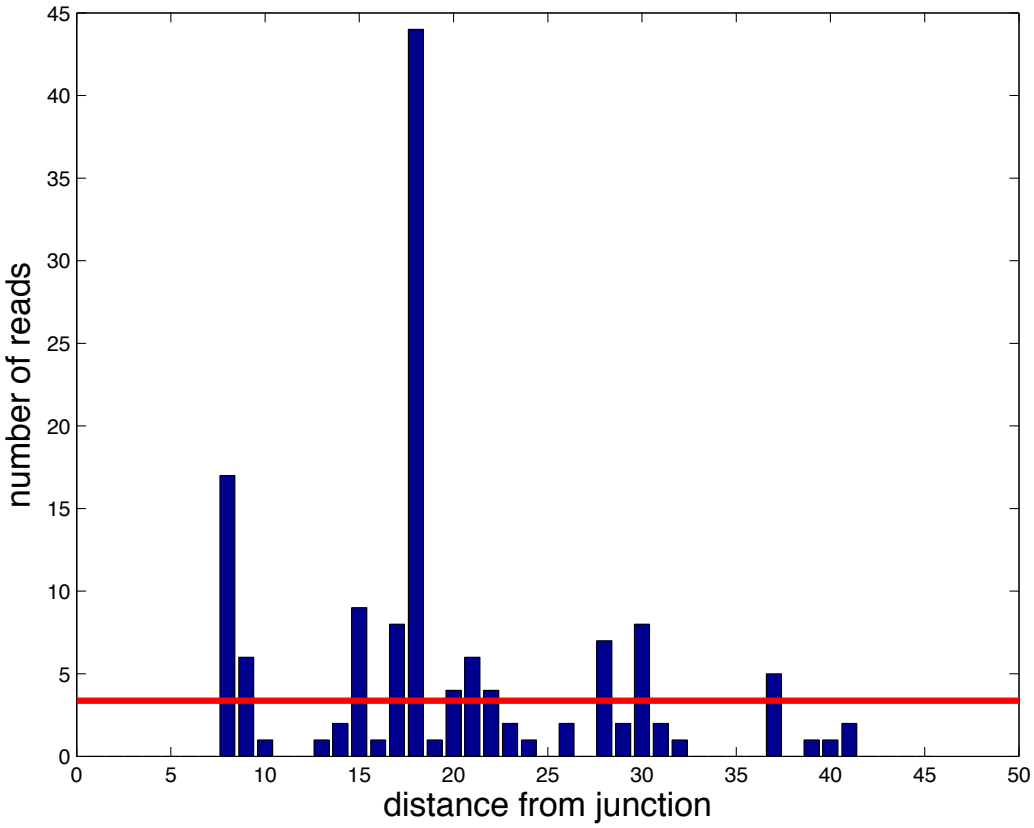


Figure 2.1: A read cover profile shows the number of read alignments (y-axis) that start at a particular distance (x-axis) from the splice junction. This histogram is a typical example of the 50nt neighborhood around a highly expressed constitutive junction. This example exhibits two types of read mapping bias: sparse coverage (empty positions) and read-stacks (tall blue bars). The horizontal line (in red) $\alpha = 3.4$ marks the average expression of the junction determined by the Native model.

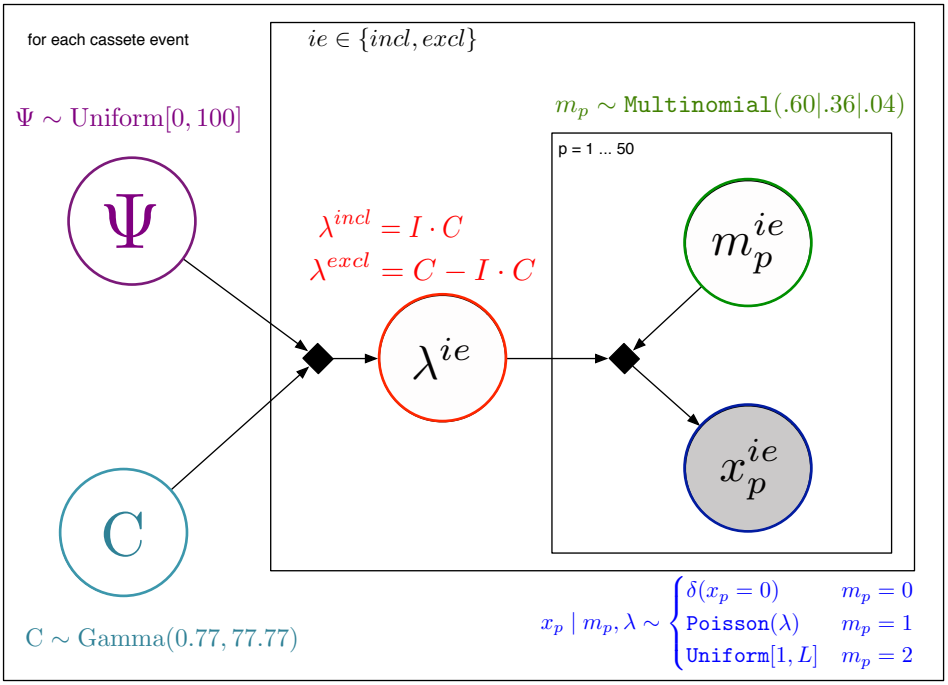
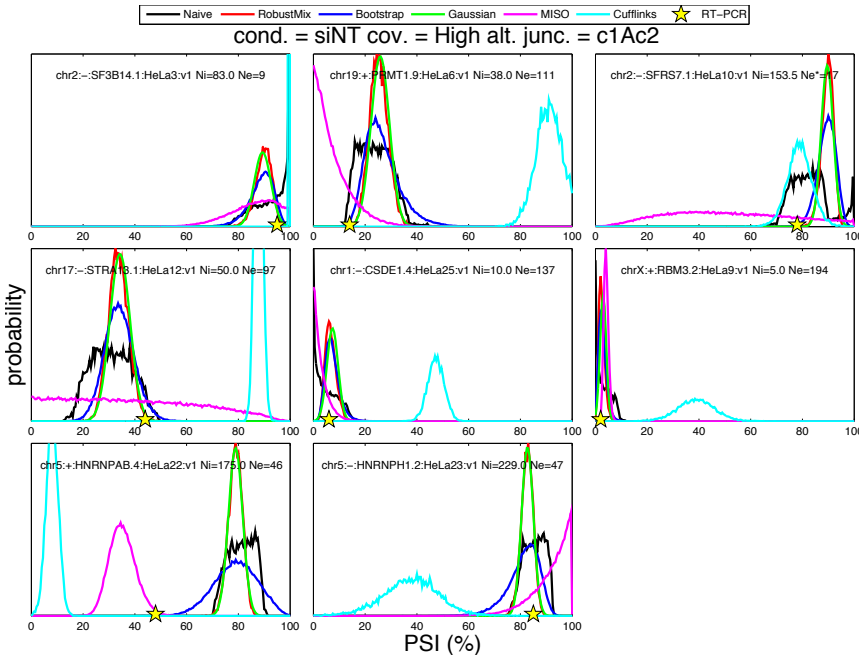
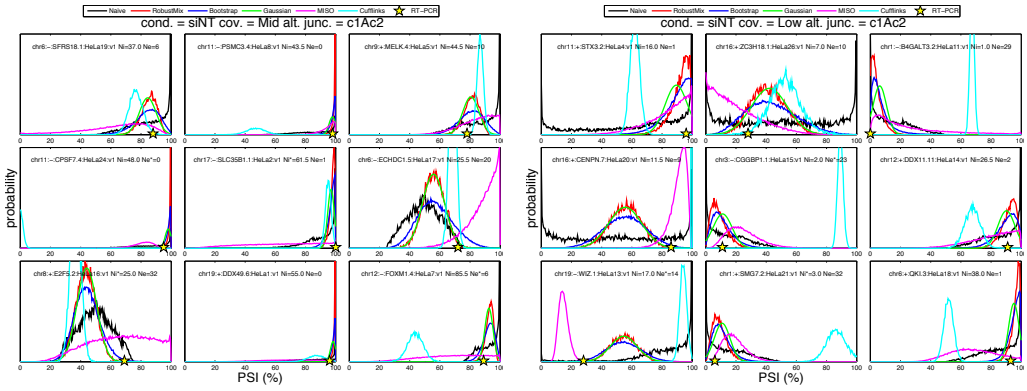


Figure 2.2: Our Mixture Model for robust estimation of PSI and coverage of cassette junctions from RNA-seq data. Only the read counts at each position (shaded x_p) are observed. The mixture components (m_p), robust expression estimates for each junction (λ^{ie}), and the overall cover (C) and percent-spliced-in (Ψ) are inferred by the model.



(A) High-cover events



(B) Medium-cover events

(C) Low-cover events

Figure 2.3: Comparison of PSI estimators of different methods for (A) high- (B)medium- and (C) low-cover junctions in a reference RT-PCR study. Each method's estimated distribution over PSI is shown in different color, and the target PSI value is shown as a yellow star on the x-axis. Methods which commit the most of their distribution mass near the star have the most accurate estimates. The text inside each plot identifies a cassette event and gives the raw number of reads mapping to the constitutive (Ne) and the average of the alternative junctions (Ni). This figure is best viewed in color.

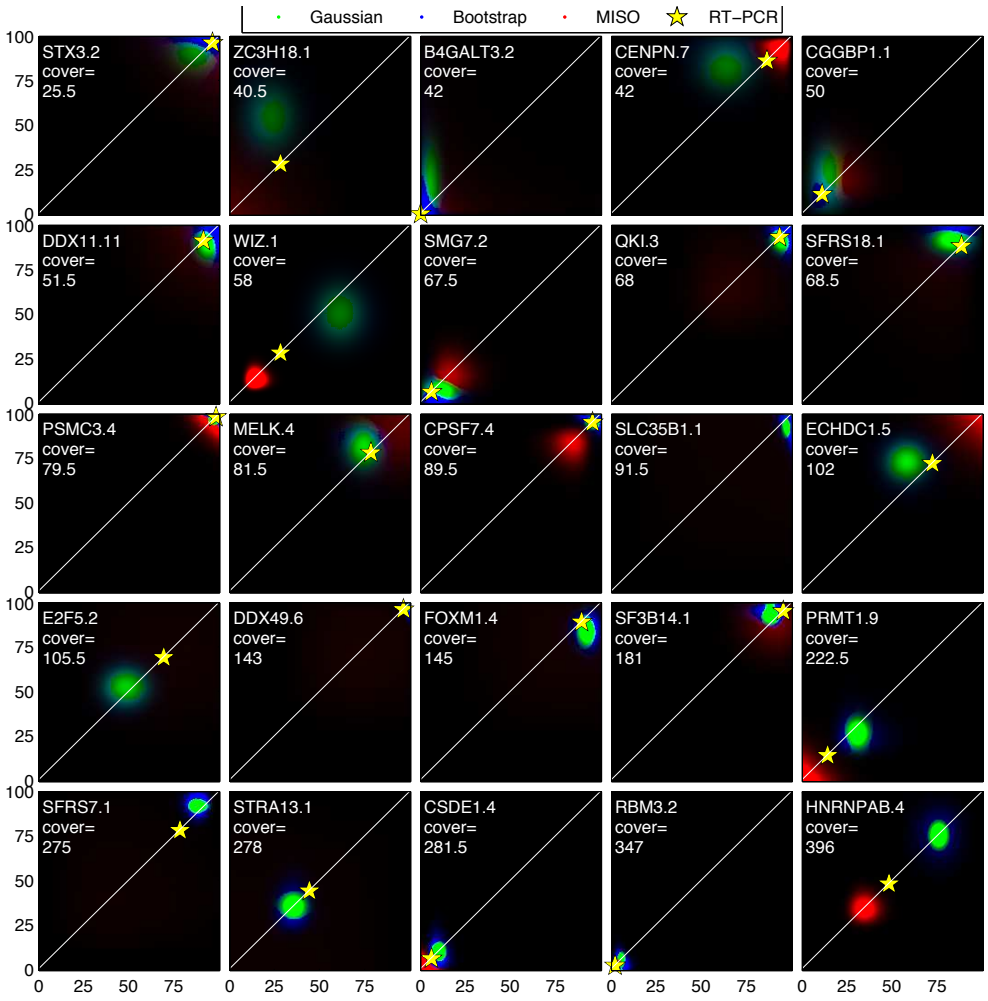


Figure 2.4: Constellation plot of the estimated PSI distributions from one vs. another half of the positions in each cassette event. The distribution of PSI along the x-axis, $P_x(\Psi)$ over the range (0-100%) is estimated from a random half of the positions and the distribution on the y-axis $P_y(\Psi)$ comes from the remaining half of the positions. The distributions are color-coded according to their methods. The intensity of each pixel $(x, y) = (a, b)$ corresponds to the product of the distributions $P_x(\psi = a) * P_y(\psi = b)$. In regions where the distributions for different methods overlap, the one with the higher probability is shown and the rest are suppressed. Each white diagonal marks the region of perfect agreement for both distributions. The yellow star along each diagonal is placed at the x- and y-coordinate matching the PSI value determined by RT-PCR for the event whose name and cover are printed in white font. This figure is best viewed in color.

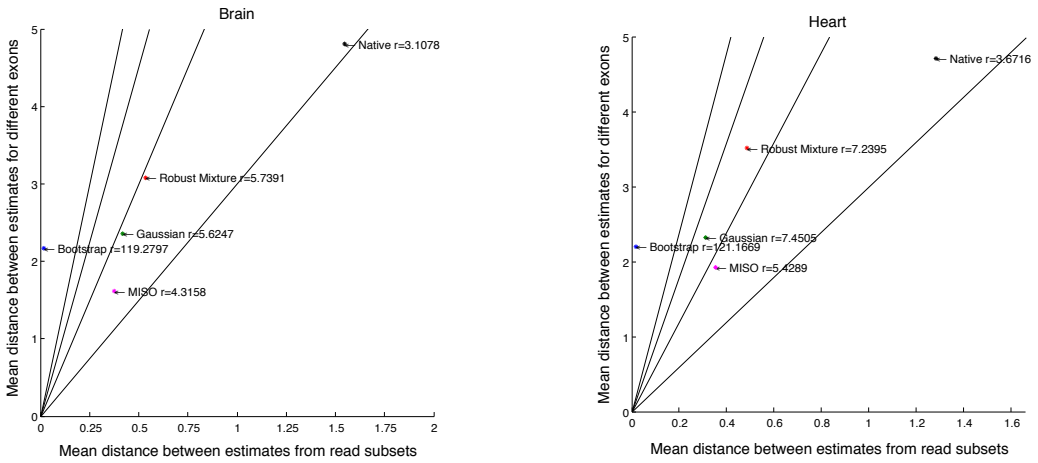


Figure 2.5: Plots of the consistency ratio between inter- and intra-exon divergence in the estimated PSI distributions for five of the methods in two human tissues. The PSI estimates were generated for a random half of the positions in each junction and compared to the PSI estimate from the other half within the same exon and between different exons. More consistent methods have a higher consistency ratio.

Chapter 3

Adenosine to Inosine RNA editing in C. elegans

3.1 The dsRBP and inactive editor, ADR-1, utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome

Inadequate adenosine-to-inosine editing of noncoding regions occurs in disease, often uncorrelated with ADAR levels, underscoring the need to study deaminase independent control of editing. C. elegans have two ADAR proteins, ADR-2 and the theoretically catalytically inactive ADR-1. Using high-throughput RNA sequencing of wild-type and adr mutant worms, we expanded the repertoire of C. elegans edited transcripts over 5-fold and confirmed that ADR-2 is the only active deaminase in vivo. Despite lacking deaminase function, ADR-1 affects editing of over 60 adenosines within the 3' UTRs of 16 different mRNAs. Furthermore, ADR-1 interacts directly with ADR-2 substrates, even in the absence of ADR-2; and mutations within its dsRNA binding domains abolished both binding and editing regulation. We conclude that ADR-1 acts as a major regulator of editing by binding ADR-2 substrates in vivo and raises the possibility that other dsRNA binding proteins, including the inactive human ADARs, regulate RNA editing by

deaminase-independent mechanisms.

3.2 Introduction

RNA editing is a posttranscriptional process that introduces changes in RNA sequences and structures (Gott and Emeson, 2000). The most prevalent form of RNA editing in metazoa is the hydrolytic deamination of adenosine (A) to inosine (I) (Nishikura, 2010). Adenosine deaminases that act on RNA (ADARs) bind to double-stranded regions of RNA and catalyze this type of editing (Goodman et al., 2012; Savva et al., 2012). Although RNA editing was initially thought to be restricted to a few select mRNAs in the central nervous system, it is now clear that adenosine deamination is widespread, with current estimates of 400,000-1,000,000 A-to-I editing events in the human transcriptome (Ramaswami et al., 2013). Adenosine and inosine have different base-pairing properties; therefore, editing alters RNA structure (Bass, 2002). Furthermore, as inosine is recognized as guanosine by cellular machinery, RNA editing can modify splice sites, alter the amino acid encoded by a codon and redirect miRNAs and siRNAs to new targets (Hundley and Bass, 2010; Rosenthal and Seeburg, 2012). As the extent of RNA editing varies during development and between cell types (Wahlstedt et al., 2009), this type of modification dynamically regulates gene expression (Tan et al., 2009). The molecular diversity generated by ADARs is most pronounced in the brain transcriptome (Blow et al., 2004; Paul and Bass, 1998). Consistent with this, deletion of ADARs in lower organisms, such as C. elegans and Drosophila, results in behavioral defects (Palladino et al., 2000; Tonkin et al., 2002), indicating that RNA editing is required for proper neuronal function. Furthermore, alterations in editing levels have been observed in a number of neuropathological diseases, including epilepsy, depression, amyotrophic lateral sclerosis, and brain tumors (Farajollahi and Maas, 2010; Tariq and Jantsch, 2012). In both development and disease, ADAR expression levels do not directly correlate with the extent of editing (Maas et al., 2001; Wahlstedt et al., 2009), implying that other mechanisms exist to regulate ADAR-mediated RNA editing. Both alternative splicing (Lai et

al., 1997; Rueter et al., 1999) and post-translational modification (Desterro et al., 2005) of ADARs generate less active variants of ADARs. Likewise, editing activity can be inhibited by sequestration of ADAR in the nucleolus (Sansam et al., 2003) or enhanced by proteins that promote nuclear localization of ADARs (Marcucci et al., 2011; Ohta et al., 2008). In addition to proteins that directly regulate ADARs, it has recently been demonstrated that both the local RNA structure (Daniel et al., 2012) and RNA binding protein (RBP) landscape of individual transcripts (Tariq et al., 2013) regulate ADAR activity. To date, none of these mechanisms have been linked to reduced RNA editing activity in disease (Orlandi et al., 2012). Furthermore, it is unlikely that the regulators of specific transcripts will play a key role in the global hypoediting of transcripts observed in many human cancers and neurological diseases. To identify mechanisms that could decrease global RNA editing levels, we focused on the role of catalytically inactive ADAR family members. The C. elegans genome encodes two proteins with the common ADAR family domain structure (ADR-1 and ADR-2). However, ADR-1 lacks several key amino acids required for deaminase activity. Worms lacking the adr-2 gene, have no detectable editing of the six known edited endogenous mRNAs (Tonkin et al., 2002), suggesting that ADR-2 is the catalytically active ADAR protein in worms. However, initial studies of worms lacking adr-1 revealed alterations in the editing efficiency of all six endogenous mRNAs examined (Tonkin et al., 2002). In addition, recent deep sequencing of C. elegans small RNAs identified over 30 small RNAs that are edited in vivo, and each have altered editing levels in worms lacking adr-1 (Warf et al., 2012). These prior observations suggest ADR-1 regulates editing. However, it is also possible that background mutations in the strains lacking adr-1 contribute to alterations in editing or that loss of adr-1 indirectly affects editing by ADR-2. To directly address these concerns, we developed a quantitative assay to measure in vivo editing levels of worms expressing adr-1 transgenes. About 40% of adenosines within three known edited mRNAs were affected by loss of adr-1. Furthermore, using a combination of high-throughput RNA sequencing of transgenic worms and probabilistic modeling we were able to identify 48 novel edited transcripts and demonstrate that loss of adr-1 affects editing of at least

half of these newly identified ADAR targets. Using an RNA immunoprecipitation (RIP) assay, we demonstrate that ADR-1 directly binds to known editing targets in vivo, that disrupting this binding alters editing of the mRNAs, and that ADR-1 and ADR-2 co-occupy transcripts in vivo. In summary, we demonstrate that catalytically inactive ADR-1 acts as a global regulator of editing by binding to target mRNAs and modulating the accessibility of ADR-2 for target adenosines.

3.3 Results

3.3.1 ADR1 significantly alters RNA editing of multiple mRNA targets

To determine the ability of ADR-1 to directly regulate RNA editing in vivo, we established a quantitative assay to measure changes in editing in worms lacking adr-1 and then tested if these changes were rescued by an ADR-1 transgene. First, we examined the editing levels at 50 individual adenosines within three known edited mRNAs: C35E7.6, lam-2 and pop-1. These three mRNAs were chosen based on the diversity in their cellular function and length of the double-stranded 3 UTR, which range from 517 to 1423 nucleotides. RNA was isolated from three independent biological replicates of wild-type and adr-1(-) adult worms. After reverse transcription, PCR amplification and Sanger sequencing, the editing efficiency was quantitatively measured using the Bio-Edit program. Technical replicates of the editing assay suggest that percent editing at each site can be determined with less than 1% error (Figure 3.1A), which is consistent with recently published data on the accuracy of measuring editing efficiency by Sanger sequencing (Eggington et al., 2011). Of the 50 edited adenosines, we observed statistically significant differences in editing levels between wild-type and adr-1(-) worms at 22 individual sites (Figure 3.1A). The bulk of the statistically significant sites (91%) had decreased editing, ranging from 3-35%, in the absence of adr-1.

To demonstrate that these sites are directly regulated by ADR-1, a 3X FLAG tagged genomic version of adr-1 was re-introduced to adr-1(-) worms by

microinjection. Importantly, this transgenic worm rescues a known adr-1 dependent effect on neuronal protein expression (Hundley et al., 2008), indicating that the transgene expresses functional ADR-1 protein (Figure 3.1B). As the transgenic worms express FLAG-ADR-1 from an extrachromosomal array that is transmitted to progeny at a high frequency, but not 100%, a neuronal GFP marker was co-injected and flow cytometry was used to purify worms containing the ADR-1 transgene. In addition, to reduce effects of developmental timing on editing efficiency all worms analyzed were also sorted by size to obtain only young adults. The quantitative editing assay showed that FLAG-ADR-1 significantly restored editing to 15 of the 22 editing sites altered in adr-1(-) worms (Figure 3.1B). It is important to note, that editing changes in the FLAG-ADR-1 are not a general phenomenon, as editing sites that are not affected by the loss of adr-1 are not altered by the transgene (Figure 3.1C). The 15 ADR-1 regulated sites include both adenosines that have increased and decreased editing in the absence of adr-1. Together, these data indicate that ADR-1 alters editing of multiple transcripts, but the effects vary depending upon the individual adenosines examined.

3.3.2 ADR1 binds directly to ADR2 target mRNAs in vivo

As the effects of adr-1(-) on editing are site specific, we hypothesized that ADR-1 is capable of regulating editing by utilizing two dsRNA binding domains (dsRBDs) to bind to potential editing substrates and alter accessibility of ADR-2 to particular nucleotides in the target mRNA. To determine if ADR-1 could bind ADR-2 editing targets in vivo, we developed an RNA-immunoprecipitation (RIP) assay for ADR-1. As a previously generated polyclonal antibody to ADR-1 was incapable of immunoprecipitating ADR-1 efficiently, the 3x FLAG-tagged ADR-1 transgenic worm was utilized. To measure specific binding of ADR-1 to target mRNAs in vivo, we compared immunoprecipitates (IPs) from FLAG-ADR-1 and adr-1(-) worms that were subjected to UV irradiation (Fig 2A). The IP samples were treated with Proteinase K to degrade FLAG-ADR-1 and release ADR-1 associated RNAs into the supernatant. RNA was extracted from the IP supernatant, reverse transcribed and quantified using real-time PCR (qRT-PCR).

Primers that amplify the three mRNAs tested in Figure 3.1, produced 3-15 fold more cDNA in the FLAG-ADR-1 IPs compared to the adr-1(-) IPs (Fig 2B). In contrast, an mRNA that does not contain dsRNA, gpd-3, is not enriched, indicating that, in vivo, ADR-1 specifically binds to these double-stranded ADR-2 target mRNAs.

As these three mRNAs have both adenosines that are inhibited and enhanced by ADR-1, these data support the hypothesis that ADR-1 modulates editing via a direct interaction with dsRNA. However, in order to regulate editing, ADR-1 needs to bind to the dsRNA before it is edited. To test this possibility, we performed the RIP assay in cells expressing FLAG-ADR-1, but lacking adr-2 and RNA editing. FLAG-ADR-1 was expressed and immunoprecipitated to a similar level in the presence and absence of adr-2 (Fig 2C). Compared to the adr-1(-) worms, all three ADAR target mRNAs were enriched to a similar extent in the FLAG-ADR-1 IPs in the presence and absence of adr-2 (Fig 2D), indicating that binding of ADR-1 to known edited mRNAs is independent of ADR-2. Furthermore, as these mRNAs have no detectable editing in adr-2(-) worms, we conclude that ADR-1 binds unedited mRNAs in the cell.

3.3.3 ADR1 alters RNA editing via binding to dsRNA

Our results indicate that ADR-1 binds to mRNAs that are targets for editing by ADR-2 in vivo. To determine if this binding is required for the ability of ADR-1 to alter editing efficiency in vivo, we created mutations in the dsRNA binding domains (dsRBDs) of C. elegans ADR-1 and examined the effects on endogenous RNA editing. A patch of lysine residues, referred to as the KKxxK motif (K=lysine, x=any amino acid), is required for binding of dsRNA binding proteins to dsRNA (Ramos et al., 2000; Ryter and Schultz, 1998). Mutation of the lysine residues to glutamate (E) and alanine (A) disrupts binding of human ADARs to dsRNA (Valente and Nishikura, 2007). To disrupt ADR-1 dsRNA binding, the KKxxK motif was mutated to EAxxA within both of the dsRBDs (referred to as the ds1+2 mutant) (Fig 3A). Similar to the aforementioned wild-type ADR-1, the ds1+2 mutant was also 3XFLAG tagged and reintroduced in the adr-1(-) back-

ground. The FLAG-ADR-1 ds1+2 mutant protein is expressed in the transgenic worms to about the same level as transgenic wild-type FLAG-ADR-1 (Fig 3B). To test whether these mutations disrupt ADR-1 binding to dsRNA, the RIP assay was performed with the ds1+2 mutant. In contrast to wild-type ADR-1, the ds1+2 mutant IPs were not enriched for the ADR-2 editing targets (Fig 3C). Thus, the ds1+2 mutant has defects in mRNA binding in vivo.

To determine if ADR-1 binding to target mRNAs influences editing efficiency, we compared in vivo editing levels of the FLAG-ADR-1 worms to the FLAG-ADR-1 ds1+2 mutant at the 15 sites that were identified as significantly regulated by ADR-1 (Figure 3.1B). As ADR-1 primarily promotes editing within these target mRNAs, most of the sites exhibit decreased editing in the absence of adr-1, with the exception of nucleotide 631 of lam-2, which has increased editing in adr-1(-) worms (Figure 3.1A). The ADR-1 ds1+2 mutant failed to significantly restore editing to 11 of these 15 sites, including nucleotide 631 of lam-2 (Fig 3D). Thus, ADR-1 binding to target mRNAs is required both for its ability to promote and inhibit editing of known edited mRNA targets in vivo.

3.3.4 Binding of dsRNA by ADR1 regulates editing of transcripts

Our data indicates that ADR-1 binding to target mRNAs alters editing of specific adenosines in vivo. To understand the impact of ADR-1 across the transcriptome, we conducted strand-specific RNA-sequencing (RNA-Seq) of RNA from wild-type (N2), adr-1(-), adr-2(-), FLAG-ADR-1 and FLAG-ADR-1 ds1+2 mutant adult worms and compared the nucleotide changes amongst the strains and the published C. elegans genomic sequence (WS220,ce10) (Fig 4A). To distinguish true RNA editing events from single nucleotide polymorphisms (SNPs), we removed annotated SNPs using Illuminas iGenomes collection. Unannotated variants were further addressed by performing RNA-Seq on RNA from adr-1(-);adr-2(-) worms and identifying all single-nucleotide variants (118,651 SNV) between the adr-1(-);adr-2(-) RNA (which lacks all A-to-I editing) and the C. elegans genome. These variants were subtracted from all other RNA-seq datasets. A Bayesian inverse

probability model was then adapted (Li et al., 2008) to identify high-confidence A-to-I editing sites from the RNA-seq data, where a confidence value based on the number of reads is associated with each predicted site. Empirically, we found that a confidence threshold of 0.995 produced the largest number of predicted sites in all strains: 59 sites in N2, 141 sites in adr-1(-), 71 sites in FLAG-ADR-1, 102 sites in FLAG-ADR-1 ds1+2 mutant, while identifying the lowest number of edits in the adr-2(-) strain (6 sites) that we presumed represented false positives (Table 3.1).

Of the 270 unique high confidence editing sites that were identified, but not present in adr-2(-) worms (Table 3.1), 250 sites are novel editing events that occur within 48 different transcripts; the remaining 20 high confidence sites were located within the previously identified ADAR targets C35E7.6, lam-2 and rncs-1 (Morse et al., 2002; Morse and Bass, 1999). The majority (71%) of these candidate-editing events occur within non-coding regions of the genome (Fig 4B). Strikingly, the vast majority of editing events occurred in 3' UTRs, consistent with the hypothesis that A-to-I editing controls gene expression by altering regulatory motifs in these regions. Interestingly, regions of the genome that encode for transposons were the second most highly identified (18%) category of editing events. In addition, we did identify 11 potential editing sites in coding regions of 8 different mRNAs. As editing events in the coding region of C. elegans mRNAs had not previously been identified, this suggests that similar to mammalian and Drosophila ADARs, C. elegans ADARs may also perform site selective editing in vivo.

Although ADARs target dsRNA of any sequence, the extent of editing at a particular site depends on the neighboring nucleotide context (Wahlstedt and Ohman, 2011). Using the Two Sample Logo software (Vacic et al., 2006), the 270 candidate editing sites had an over-representation of A both immediately 5 and 3 to the edited adenosine, whereas both G and C are under-represented at the positions 5 to the edited adenosine and C is under-represented 3 to the edited adenosine (Fig 4C). Both in vitro biochemical studies and transcriptome-wide RNA-Seq data indicates that human ADARs have a similar 5 preference. However human ADARs tend to favor a G at the 3 position to the edited adenosine (Lehmann and Bass,

2000; Riedmann et al., 2008). It is important to note, that because of overlapping specificities of mammalian ADARs, human transcriptome-wide datasets apply to editing by both human ADAR1 and ADAR2. However, as C. elegans ADR-2 is responsible for deamination of all of the RNA-Seq sites, our data provides the first in vivo nucleotide preferences of a single ADAR acting primarily at noncoding regions.

To validate the potential editing sites, Sanger sequencing editing assays were performed for 9 novel edited transcripts (Figure 3.2). Importantly, 50 of the 53 predicted sites were verified by Sanger sequencing, suggesting the false discovery rate of the pipeline is approximately 5.7%. However, in addition to the 50 editing sites identified from the RNA-Seq analysis, the Sanger sequencing of these 9 novel transcripts revealed 179 additional editing sites (Table 3.2), indicating that our probabilistic model is capable of identifying highly edited transcripts.

To determine if ADR-1 affected editing in the transcriptome, the editing efficiency of the 270 high confidence editing sites was quantified using a novel Bayesian model. To ensure accurate quantification, we processed all the RNA-Seq reads through the bioinformatics pipeline described above (Fig 4A), with one exception: read filter 5d was relaxed from requiring an edit site to be 25 nt away from each end down to a less-stringent 5 nt and required a minimum of 5 reads for a site in a given strain. With these criteria, we were able to quantify editing of over 100 sites for each of the four strains, with any two strains having an overlap of between 72-105 editing sites (Figure 3.3A-D). Pairwise comparison of the editing sites identified from the four RNA-Seq data sets indicated that the editing efficiency is most consistent between the wild type and FLAG-ADR-1 strains (Fig 4D, Figure 3.3E-G). This is consistent with the Sanger sequencing data of known editing sites and provides further evidence that the FLAG-ADR-1 transgene is capable of restoring editing to the adr-1(-) strain at most sites. As over-two thirds of the wild-type and FLAG-ADR-1 sites fell within one standard deviation (12%) of the regression line on the scatter plot, we used this threshold to categorize our newly identified sites into ADR-1 and non-ADR-1 regulated (Table 3.3). As multiple RNA-Seq studies have shown that determination of editing levels tends

to increase with read coverage (Bahn et al., 2012; Lee et al., 2013), it is important to note that similar results (greater than 80% overlap) were obtained when we estimated the error of editing at each site upon the read density at a given site in each strain (Table 3.3), suggesting that the editing percent thresholds for ADR-1 regulated and non-regulated sites are accurate. Comparison of editing levels at the 81 sites common between wild-type and adr-1(-) RNA-Seq datasets revealed that over half (56%) of the edited adenosines have altered editing levels in the absence of adr-1 (Table 3.3). Interestingly, 44 of these 45 sites are located within the 3 UTRs of 13 novel edited transcripts that we identified. This data is consistent with our quantitative Sanger sequencing analysis of the 3 UTRs of known ADAR targets (Fig 1A). In addition, at 38 of these ADR-1 regulated sites we were able to quantify editing levels for both the FLAG ADR-1 and FLAG-ADR-1 ds1+2 RNA-Seq datasets. Editing levels at 13 sites located within the 3 UTRs of 8 newly identified ADAR target mRNAs were dependent upon dsRNA binding by ADR-1 (Fig 4E). Together these transcriptome-wide studies indicate that ADR-1 regulates editing of specific adenosines within the 3 UTRs of the majority of C. elegans edited mRNAs and dsRNA binding is required for this function.

3.3.5 ADR1 and ADR2 co-occupy multiple transcripts in vivo

At present it is unclear how ADR-1 binding to mRNAs affects editing by ADR-2. It is possible that ADR-1 and ADR-2 heterodimerize in the cell to edit certain transcripts, whereas other transcripts are edited by ADR-2 alone. Alternatively, it is possible that ADR-1 and ADR-2 interact on the same transcripts, but regulate editing in an adenosine-specific manner. To gain insight into these possibilities, we examined the wild-type and FLAG-ADR-1 RNA-Seq datasets to determine whether editing at ADR-1 regulated adenosines occurred on the same reads as edited adenosines that are not affected by loss of adr-1. For most of the novel transcripts that are edited in the 3 UTR (9/12), editing was observed at both adenosines affected by adr-1 and non-regulated sites, within the same 75 nt read (Table 3.3).

To provide further evidence that ADR-1 and ADR-2 associate on common targets in vivo, we immunoprecipitated FLAG-ADR1 and tested for the presence of ADR-2 with an ADR-2 specific antibody (Fig 4F). ADR-2 was present in the FLAG-ADR-1 IPs, but not the FLAG-ADR-1 ds1+2 mutant IPs or IPs from worms lacking adr-1 (Fig 4G). Consistent with an RNA-dependent interaction of ADR-1 and ADR-2, IPs of wild-type ADR-1 treated with RNase also resulted in reduced ADR-2 co-immunoprecipitation (Figure 3.5). Together, these data suggest that ADR-1 and ADR-2 interact on transcripts in vivo, but are not likely to heterodimerize independent of target mRNAs.

3.4 Discussion

In this study, we have demonstrated that C. elegans ADR-1 utilizes its double-stranded RNA binding function to regulate A-to-I editing levels in vivo. Using a high-throughput RNA sequencing approach coupled to probabilistic modeling, we were able to expand the number of known ADAR target mRNAs five-fold, as well as provide the first transcriptome-wide evidence that ADR-1 is a catalytically inactive member of the ADAR family. Furthermore, using both our extensive Sanger sequencing analysis of ADAR targets and quantification of transcriptome-wide RNA-Seq data, we demonstrate that ADR-1 regulates editing efficiency of specific adenosines within most ADAR target 3 UTRs.

We propose that ADR-1 regulates editing by binding to target mRNAs and altering accessibility of ADR-2 for specific adenosines. Multiple recent studies support the idea that the RNA binding protein landscape of ADAR target mRNAs affects editing levels (Bhogal et al., 2011; Chen, 2013; Garncarz et al., 2013; Tariq et al., 2013). However, in most of these studies, the RNA binding activity of the regulators was not shown to be required for A-to-I regulatory activity and these regulators were all single-stranded RNA binding proteins that altered editing of specific coding editing events. In contrast, we demonstrate that ADR-1 binds to several target mRNAs via its dsRNA binding domains, and that this binding is required for regulation of editing. This dsRNA binding activity would allow ADR-

1 to interact with nearly all the same targets as ADR-2, thus allowing it to serve a more global role in regulating editing within long double-stranded regions. As dsRBDs are the second most abundant RNA recognition motif (Stefl et al., 2010), it is unlikely that this regulatory role is limited to C. elegans ADR-1. Consistent with this, 20% of our newly discovered edited transcripts overlap with recently identified targets of another dsRNA binding protein (dsRBP), C. elegans Staufen (LeGendre et al., 2013) (Table 3.1).

Our Sanger sequencing and transcriptome-wide analyses suggest that the regulatory role of ADR-1 is specific to certain adenosines (Fig 1A, Table 3.3). Although dsRBPs are generally presumed to lack sequence specificity (Tian et al., 2004), recent structural data suggests ADARs recognize specific nucleotides within a dsRNA target (Stefl et al., 2010). Our RIP assay indicates that ADR-1 binds to the lam-2 and pop-1 mRNAs to a similar extent in the presence and absence of adr-2 (Fig 2D). Thus, at least for certain edited mRNAs, ADR-1 does not compete with ADR-2 for binding sites in vivo. Consistent with this, the majority of the ADR-1 regulated sites identified in both the RNA-Seq datasets and the Sanger analysis have enhanced editing in the presence of adr-1 (Fig 1A, Table 3.3), suggesting that ADR-1 functions primarily to promote ADR-2 editing, not compete with ADR-2 for target adenosines. As editing is not required for ADR-1 to bind these mRNAs, we postulate that, in vivo, ADR-1 first binds to target mRNAs and then either alters binding of ADR-2 to specific regions and/or regulates the catalytic activity of ADR-2 (See Graphical abstract). Interestingly, it was recently demonstrated that human ADAR1 binding to mRNAs creates binding sites for another RNA binding protein, HuR, which results in increased RNA stability of HuR-ADAR1 bound transcripts (Wang et al., 2013). Similar to what was demonstrated for human ADAR1-HuR, we detected an in vivo interaction between wild-type ADR-1 and ADR-2, but not the ADR-1 ds1+2 mutant, which is consistent with ADR-1 and ADR-2 interacting on target mRNA. Interestingly, it has previously been suggested that human ADAR homodimerization on dsRNA is required for efficient editing in vitro (Jaikaran et al., 2002). Although our evidence indicates that ADR-1 utilizes dsRNA binding to regulate editing by ADR-2, it is possible that this regulatory

function is due to effects of ADR-1 on expression of other RNA binding proteins, that in turn alter ADR-2 accessibility to target mRNAs. Future work aimed at both identifying ADR-1 and ADR-2 binding sites on mRNAs in vivo and determining the impact of ADR-1 on ADR-2 editing efficiency on target mRNAs in vitro will be needed to determine if there is a correlation between binding site specificity and regulation of specific sites. In summary, our results indicate that ADR-1 utilizes dsRNA binding to regulate A-to-I editing across the C. elegans transcriptome. These studies not only suggest a potential biological function for the catalytically inactive ADARs present in humans, but also unveil a potential mechanism for other dsRBPs, such as Staufen, to regulate RNA editing levels.

3.5 Experimental Procedures

3.5.1 Maintenance of worm strains and Transgenics

Worm strains were maintained by growth on NGM plates seeded with Escherichia coli OP50. Transgenic worm lines were generated by microinjection. A detailed description of the injections and transgenic strains is given in the Extended Experimental Procedures.

3.5.2 RNA Isolation and Editing Assays

Total RNA was isolated using Trizol (Invitrogen). RNA was further treated with Turbo DNase (Ambion) and then isolated using the RNA Easy Extraction kit (Qiagen). Editing assays were performed using Thermoscript RT (Invitrogen) for reverse transcription and PFX Platinum DNA Polymerase (Invitrogen) for PCR amplification with gene-specific primers (Table 3.4). PCR products were gel purified and subjected to Sanger sequencing. Editing was quantified using the program BioEdit, which quantifies adenosine and guanosine peak heights. For all editing assays, negative controls were conducted without Thermoscript RT to ensure that all DNA subjected to Sanger sequencing resulted from cDNA amplification.

3.5.3 Strand-specific RNA sequencing

Strand specific mRNA sequencing libraries were prepared as described previously (Parkhomchuk et al., 2009). Libraries were normalized to 2nM and sequenced for SE76 cycles on either HiSeq2000 (adr-1(-);adr-2(-)) or Illumina GAII (all other strains).

3.5.4 Bioinformatics Pipeline

To achieve accurate identification of editing sites, we combined filters from existing pipelines (Chen, 2013; Lee et al., 2013; Levanon et al., 2004; Ramaswami et al., 2012) in a strand-specific manner. Accurate quantification was performed by extending the existing Bayesian method for genomic variant calling used in the 1000 Genomes project (Li et al., 2008) with a custom-designed prior on the editing % (Figure 3.4). In addition to leveraging established considerations with regards to read sequencing and alignment errors (Kleinman and Majewski, 2012; Lin et al., 2012; Pickrell et al., 2012) our approach benefits greatly from using the adr-1(-);adr-2(-) strain as a powerful filter for unannotated variants to maintain low false positive rates while confidently identifying RNA editing sites. Detailed steps of the pipeline and the Bayesian method for variant calling are described in the Extended Experimental Procedures.

3.5.5 RNA Immunoprecipitation (RIP) Assay

After washing with IP Buffer (50mM HEPES, pH 7.4; 70mM K-Acetate, 5mM Mg-Acetate, .05% NP-40 and 10% glycerol), worms were subjected to 3J/cm2 of UV radiation using the Spectrolinker (Spectronics Corp.) and stored at -80C. To obtain cell lysates, frozen worms were ground with a mortar and pestle on dry ice. After thawing, the lysate was centrifuged to remove insoluble material and the protein concentration was measured with Bradford reagent (Sigma). Five micrograms of extract was added to anti-Flag magnetic beads (Sigma) that were washed with wash buffer (WB: 0.5M NaCl, 160mM Tris-HCl pH 7.5). After incubation for 1 hour at 4C, the beads were washed with ice-cold WB, resuspended in low

salt WB (0.11M NaCl), 1?l RNasin (Promega) and 0.5?l of 20mg/ml proteinase K (Sigma) and incubated at 42C for 15 minutes to degrade protein and release bound RNA. Protein samples were subjected to SDS-PAGE and western blotting with a FLAG antibody (Sigma). RNA samples were isolated as described above. Following DNase treatment, qRT-PCR for known editing targets was performed as previously described (Hundley et al., 2008).

3.5.6 Flow Cytometry

Flow cytometry was conducted at the IUB Flow Cytometry Core Facility by a dedicated technician using the COPAS Select (Union Biometrica) large particle sorter. Parameters were adjusted manually to select either only adult worms for non-transgenic strains or adult worms expressing GFP for transgenic lines.

3.6 Acknowledgements

This chapter is adapted from M Washburn*, **B Kakaradov***, B Sundararaman, E Wheeler, S Hoon, G Yeo, H Hundley. The dsRBP and inactive editor ADR-1 utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome. Cell Reports. (2014). The dissertation author was a joint first author of this paper, and was responsible for all computational research.

3.7 Figures

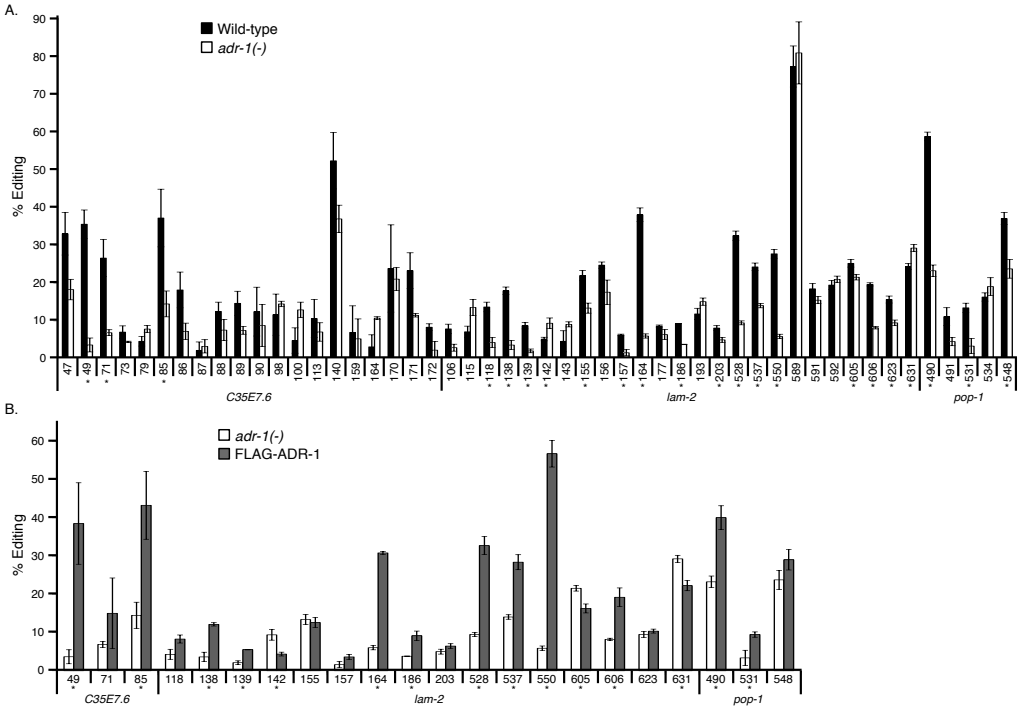


Figure 3.1: (A and B) Editing levels at individual nucleotides within the 3 UTRs were measured for 3 biological replicates. Error bars represent standard error of the mean (SEM). Significant changes ($p < 0.05$) in editing levels between (A) wild-type and adr-1(-) or (B) adr-1(-) and FLAG-ADR-1 are marked with an asterisk.

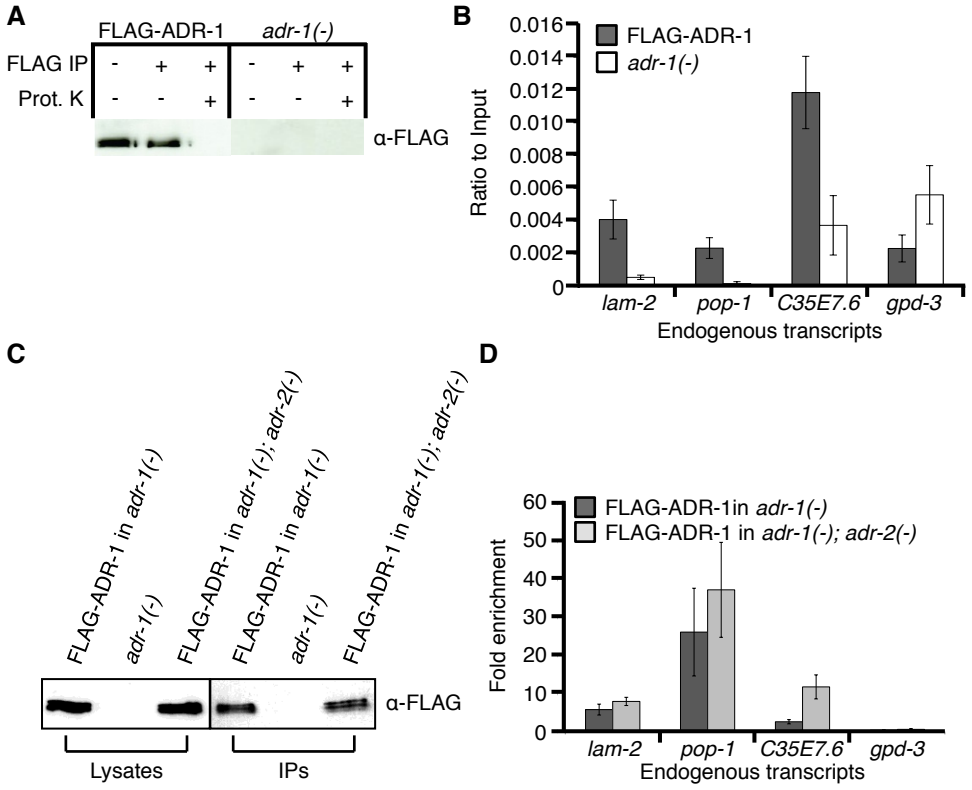


Figure 3.2: (A) Lysates from the indicated worm lines were subjected to FLAG IP and treatment with Proteinase K (Prot. K). A portion of the untreated lysate (IP-, Prot. K-), IP (IP+, Prot. K-) and beads after Prot. K treatment (IP+, Prot. K-) were subjected to immunoblotting for the FLAG epitope. (B) cDNA levels for the indicated endogenous mRNAs were measured using qRT-PCR. Values from the IP samples of FLAG-ADR-1 in adr-1(-) and the negative control adr-1(-) were divided by their respective input levels. Error bars represent SEM for three biological replicates. (C) Lysates from the indicated worm lines were subjected to immunoprecipitation with magnetic FLAG resin. A portion of the input lysate and IPs were subjected to immunoblotting for the FLAG epitope. (D) cDNA levels for the indicated endogenous mRNAs were measured using qRT-PCR. The ratios of the cDNAs present in the IP samples of the indicated strains were divided by their respective input levels and normalized to the negative control adr-1(-) to give a fold enrichment. Error bars represent SEM for three biological replicates.

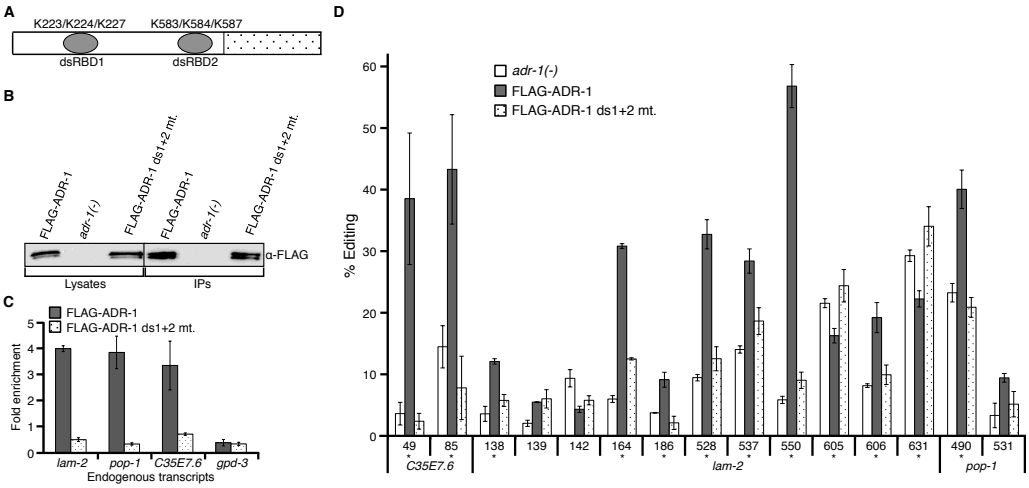


Figure 3.3: (A) Schematic of ADR-1 protein with dsRBDs (grey ovals) and deaminase domain (patterned rectangle). Lysine (K) residues mutated are indicated above each dsRBD. (B) FLAG Immunoblotting of lysates and IPs of the indicated strains. (C) Ratio of the cDNA present in the IP samples divided by the input cDNA levels for the indicated strains were divided by the IP:input ratio of the adr-1(-) worms. Error bars represent SEM for three biological replicates. (D) Calculated percent editing in the indicated strains for the endogenous mRNAs of C35E7.6, lam-2 and pop-1. Error bars represent SEM of 3 biological replicates. Significant changes ($p < 0.05$) in editing levels between FLAG-ADR-1 and FLAG-ADR-1 ds1+2 mutant are marked with an asterisk.

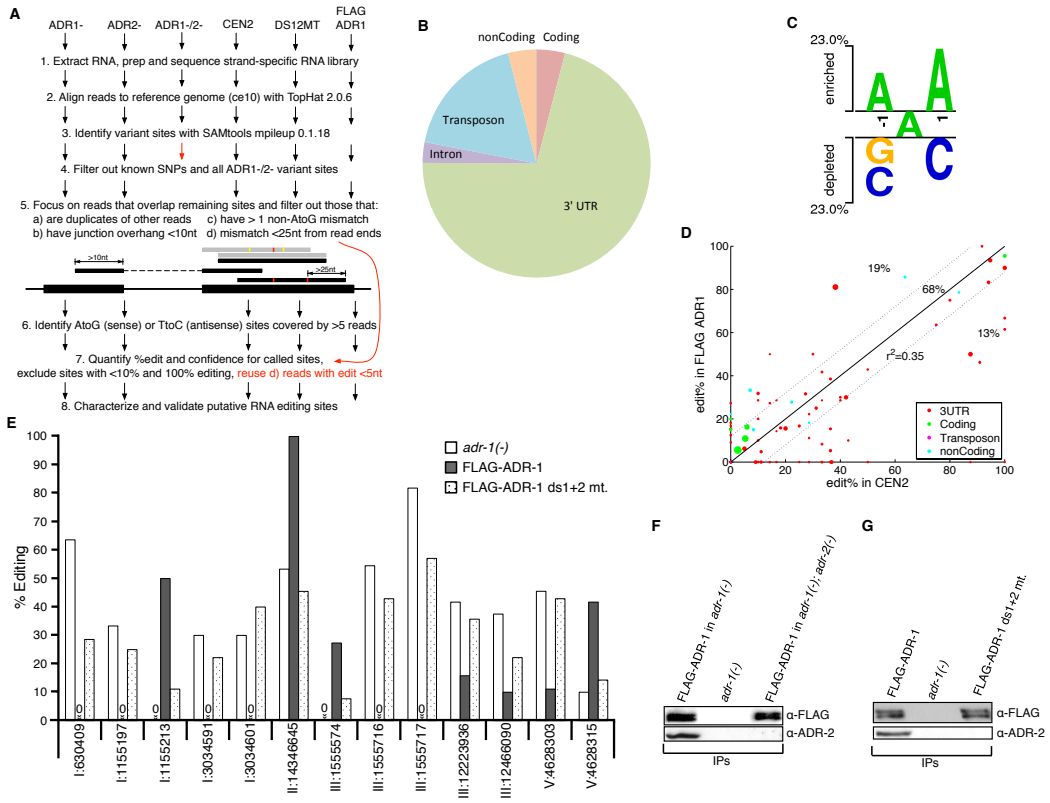


Figure 3.4: (A) Bioinformatics strategy depicting the major steps for processing strand-specific RNA-seq data into A-to-I sites for each strain. (B) Distribution of identified RNA editing sites within annotated transcriptome regions. (C) Nucleotide preferences for the 270 candidate editing sites were calculated compared to a randomized control. Enriched and depleted nucleotides are shown above and below the axis, respectively. The level of conservation is represented by letter height. Logos were generated using a t-test with $p < .005$ and no Bonferroni correction. (D) Scatter plots of percent editing of quantified sites that overlap in the wildtype (CEN2) and FLAG-ADR-1 datasets. The r^2 fit to the $y = x$ line (black diagonal). The margin (dotted line) between no-change and differentially-edited sites equals 12 units of change in the edit % (one standard deviation). (E) Editing levels for 13 sites from the RNA-seq data where editing levels between adr-1(-) and FLAG-ADR-1 and between FLAG-ADR-1 and FLAG-ADR-1 ds1+2 mutant were greater than 12% (Table S3). Adenosines that had no observed editing are marked with a zero above the x-axis. (F and G) Immunoblotting analysis of FLAG IPs from the indicated strains. IPs were performed as previously stated except worms were not subjected to UV-crosslinking and only light salt washes were employed.

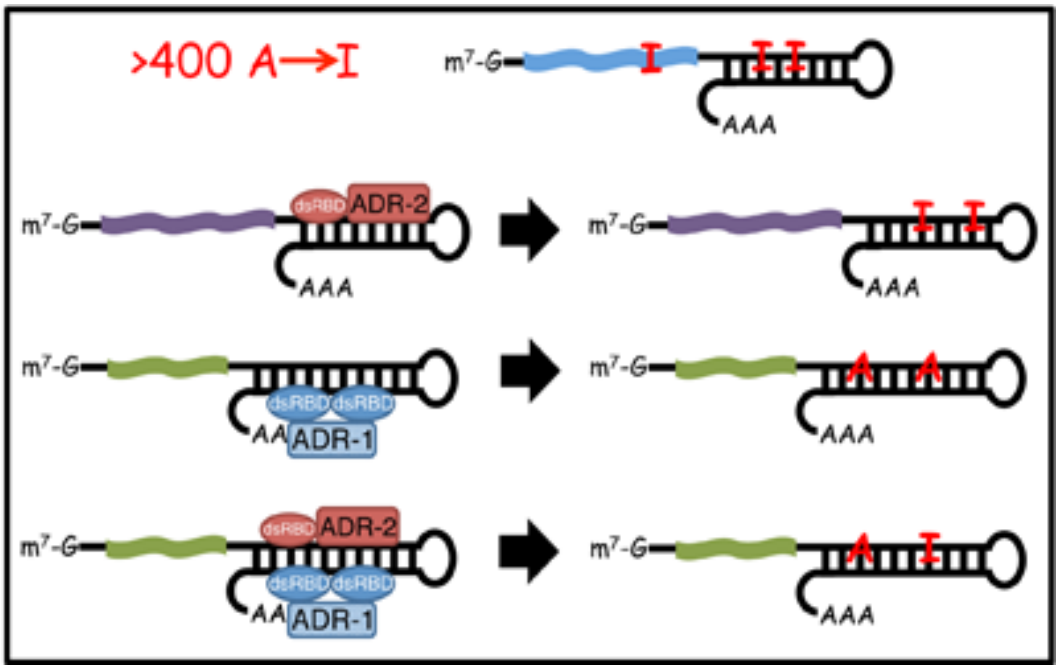


Figure 3.5: Identification of > 400 novel A-to-I editing sites, primarily within noncoding regions. ADR-1 regulates editing of specific adenosines within 3 UTRs of diverse transcripts. ADR-1 regulates RNA editing by directly binding to ADR-2 target mRNAs. ADR-1 and ADR-2 do not form heterodimers, but co-occupy transcripts in vivo.

Chapter 4

Single-cell analysis reveals asymmetric T cell specification during adaptive immunity

T lymphocytes responding to microbial infection give rise to effector cells that mediate acute host defense and memory cells that provide long-lived immunity, but the fundamental question of when and how these cells arise remains unresolved. Here we combine single-cell gene expression analyses with machine-learning approaches to trace the transcriptional roadmap of individual $CD8^+$ T lymphocytes throughout the course of an immune response *in vivo*. Gene expression signatures predictive of eventual fates could be discerned as early as the first T lymphocyte division and may be influenced by asymmetric partitioning of the interleukin-2 receptor during mitosis. These findings underscore the importance of single-cell analyses in understanding fate determination and provide new insights into the specification of divergent lymphocyte fates early during an immune response to microbial infection.

4.1 Early specification of CD8$^+$ T lymphocyte fates during adaptive immunity revealed by single-cell gene expression analyses

During a microbial infection, responding T lymphocytes give rise to two distinct classes of cellular progeny, effector cells that provide acute host defense and long-lived memory cells that provide durable immunity [1]. Terminally differentiated, short-lived effector cells (T_{sle}) can be identified phenotypically by high expression of the lectin-like receptor (KLRG1) and low expression of the interleukin 7 receptor (IL-7R)[2]. At least two distinct subsets of long-lived memory cells, central memory (T_{cm}) and effector memory (T_{em}), have been described and can be distinguished on the basis of their proliferative capacity, cytotoxicity, anatomic localization and expression of certain homing and chemokine receptors, including L-selectin (CD62L) and CCR7[3,4].

Prior studies using single-cell adoptive transfer and genetic barcoding approaches[5,6] have elegantly demonstrated that a single nave CD8$^+$ T lymphocyte can give rise to more than one fate, and importantly, is capable of generating all of the diverse cellular fates necessary for an immune response. The process by which a single activated T lymphocyte yields effector- and memory-fated progeny and the timing at which these differentiation pathways begin to diverge, however, remain unresolved. One possibility is that the progeny of an activated nave CD8$^+$ T lymphocyte progress along a linear differentiation path, initially becoming effector cells, with a subset of these cells later acquiring the memory fate[1,7,8]. An alternative possibility is that the first CD8$^+$ T cell division *in vivo* is asymmetric[9,10], thereby enabling lymphocyte fates to diverge early during an immune response owing to unequal inheritance of certain determinants, such as the interferon γ (IFN-γ) receptor and the T-box transcription factor, T-bet.

Tracing individual lymphocytes sequentially as they differentiate *in vivo* might distinguish whether lymphocytes progress along a linear differentiation pathway[1,7,8] or diverge early during an immune response. While genomic profiling studies have begun to elucidate the transcriptional networks that control lymphocyte

fate specification[11-13], these studies have been based on analyses of bulk cellular populations, making it impossible to discern cell fate decisions made by individual T cells. Recent technological advances that have coupled microfluidics technologies with high-throughput qRT-PCR analyses have enabled detailed analyses of cell fate decisions in *C. elegans* development, induced stem cell reprogramming and cancer biology[14-17]. Here, we applied single-cell gene expression profiling to investigate the ontogeny of effector and memory $CD8^+$ T lymphocytes during a microbial infection *in vivo*, uncovering evidence for heterogeneity in gene expression within individual lymphocytes early after the initiation of an adaptive immune response.

4.2 Results

4.2.1 Single-cell gene expression analyses of $CD8^+$ T lymphocytes *in vivo*

In order to delineate the hierarchy and mechanism of $CD8^+$ T cell differentiation during an adaptive immune response at the single-cell level, we employed an experimental system that allowed us to interrogate the gene expression of individual $CD8^+$ T lymphocytes throughout the course of a microbial infection *in vivo*. $CD8^+$ T cells transgenic for the OT-1 T cell receptor that recognize a specific ovalbumin epitope were adoptively transferred into wild-type recipient mice. Mice were infected intravenously 24 hours later with recombinant *Listeria monocytogenes* bacteria expressing ovalbumin (Lm-OVA) and $CD8^+$ T cells were sorted throughout the course of infection for single-cell analysis (**Fig. 4.1**). In addition, we selected for analysis terminally differentiated short-lived effector cells (T_{sle}, $KLRG1^{hi}IL\text{-}7R^{lo}$)[2], putative memory precursor cells (T_{mp}, $KLRG1^{lo}IL\text{-}7R^{hi}$)[2], and central memory (T_{cm}, $CD44^{hi}CD62L^{hi}$) and effector memory (T_{em}, $CD44^{hi}CD62L^{lo}$)[3,4] cells (**Fig. 4.1**).

Quantitative real-time PCR analysis was performed using Fluidigm 96.96 Dynamic Arrays, enabling simultaneous measurement of expression for 96 genes

in 96 individual cells . Among the 94 gene targets (**Table 4.1**) we selected for analysis were transcriptional regulators previously reported to influence $CD8^+$ T lymphocyte differentiation[18-25]; cytokines, chemokines, and their receptors[19]; and molecules associated with tissue homing and survival[19].

Table 4.1: 94 selected gene targets grouped according to their function.

Class	**Genes**
Apoptosis	Bnip2, Bnip3l, Casp3, Casp9, Cflar, Pdcd1
Cytokine/ chemokine receptors	Ccr5, Ccr6, Ccr7, Cxcr3, Ifngr1, Ifnar1, Il2ra, Il2rb, Il7r, Klrc1, Klrg1, Tnfrsf1, Tnfrsf9
Cytokines, chemokines, granzyme	Ccl3, Ccl4, Ccl5, Cxcl10, Gzmb, Ifng, Il2, Il3, Lif, Xcl1
Polarity/proteasome	Prkcz, Psmb1, Psmb7
Housekeeping	Actb, Gapdh, Rn18s, Rpl35
Signaling, proliferation, self-renewal	Bag2, Bmi1, Bmp2, Cd28, Cd4, Cd44, Cd69, Cd8a, Grap2, Hk2, Lag3, Lgals1, Mapk3, Mapk8, Mapk14, Mela, Mtor, Myc, Ptprc, RelA, Sema7a, Serpinb6b, Serpinb9, Setd7, Sell, Thy1
Transcription factors	Atf1, Bcl11b, Bcl6, Bhlhe40, Eomes, Foxo1, Foxo3, Foxp1, Foxp3, Gata3, Hopx, Id2, Id3, Irf4, Irf8, Klf2, Lef1, Nfatc1, Nfatc2, Prdm1, Rel, Runx1, Runx2, Stat1, Stat4, Tbx21, Tcf3, Tcf7, Tcf12, Tox, Zeb2, Zfp281

After excluding failed reactions, expression data from 1,300 single cells were retained for in-depth analyses . Because expression of "housekeeping" genes has been shown to vary substantially across cell types and states of differentiation[26], the expression of each gene of interest was utilized without normalization for all of the analyses performed herein.

We used principal component analysis (PCA) to visualize the expression data globally. PCA is an unsupervised dimensionality reduction method that we used to project the data into 2 dimensions by its coordinates in the first two principal components (PC1 and PC2) that account for the largest variations in the data. These PCs are linear combinations of the 94 original genes. PCA revealed that nave, T_{sle}, T_{em}, and T_{cm} cells are clustered distinctly (**Fig. 4.2a**).

Expression of *Sell* and *Tcf7*, which encode the trafficking molecule CD62L and the transcription factor TCF-1, distinguished nave from T_{sle} cells, consistent with previous findings[2,4]. Though T_{sle} cells formed a distinct cluster, these cells were projected closest to T_{em} cells (**Fig. 4.2a**), suggesting related gene expression profiles that may underlie some of their functional similarities, such as cytotoxicity and secretion of pro-inflammatory cytokines[27]. This clustering was driven primarily by *Zeb2*, a transcription factor expressed in T_{sle} cells[12]. In addition, T_{em} and T_{cm} cells occupied distinct clusters, with higher expression of *Tcf7*, *Il2rb*, *Il7r*, *Cxcr3* and *Sell* mRNA in T_{cm} cells and higher expression of *Zeb2* mRNA in T_{em} cells accounting for the variance between these memory cell populations. Some of the disparities observed at the transcriptional level were confirmed at the protein level (**Fig. 4.2b**), supporting the finding that T_{cm} and T_{em} cells are molecularly distinct. The higher expression of *Il7r* and *Tcf7*, regulators of T lymphocyte survival and longevity[25,28], that we observed in T_{cm} cells may underlie the superior capacity of these cells to persist *in vivo*[29]. Putative memory precursor (T_{mp}) cells did not form a distinct cluster but overlapped with T_{sle}, T_{em} and T_{cm} cells (**Fig. 4.2c**). These results suggested that putative T_{mp} cells are molecularly heterogeneous, raising the possibility that this population may not represent memory precursor cells, but instead may be comprised of "mature" memory and terminally differentiated effector cells. Together these findings suggest that T_{sle}, T_{cm}, and T_{em} cells, but not putative T_{mp} cells, exhibit similar gene expression profiles at the single-cell level.

4.2.2 Molecular heterogeneity at the single-cell level early after infection

To assess whether single responding $CD8^+$ T cells comprised distinct clusters early after infection, we analyzed the gene expression profiles of individual $CD8^+$ T cells isolated throughout the course of infection (**Fig. 4.3a**). PCA revealed substantial heterogeneity among cells isolated early after infection (division 1 and day 3) compared to cells isolated at later timepoints (day 5 and day 7, T_{sle}, T_{cm}, and T_{em}). The first two principal components captured 17% of the variance

in our dataset, slightly lower than that previously observed[15], likely a reflection of a higher degree of heterogeneity in lymphocytes during differentiation and the greater number of genes analyzed in our study. In agreement with our findings using PCA, an alternative unsupervised method, t-distributed Stochastic Neighbor Embedding analysis[30], was also performed and showed similar results . To test whether the heterogeneity observed using data from single cells could be recapitulated using data from bulk cells, we formally compared the analyses using data derived from single versus bulk populations (**Fig. 4.3a**). We found that the heterogeneity we observed at the single-cell level within putative T_{mp} cells and cells isolated early after infection was not apparent in the bulk analysis (**Fig. 4.3a**), thus illustrating the power and necessity of using a single-cell approach.

To further evaluate the degree of heterogeneity within and between cell populations at each time point, which was not previously possible using bulk analysis, we applied the Jensen-Shannon Divergence (JSD) metric, a non-parametric, model-free measure of similarity between two empirical probability distributions. In general, the intra-population JSD was lowest in nave cells and highest in cells isolated early after infection (**Fig. 4.3b**) . We observed that the intra-population JSD decreased as a function of time following infection, with the notable exception of putative T_{mp} cells (**Fig. 4.3b**). These cells exhibited a high degree of intra-population divergence, consistent with the apparent heterogeneity of these cells by PCA (**Fig. 4.3a**). Comparing JSD pair-wise between all cell populations (nave, division 1, days 3, 5, and 7 post-infection, T_{mp}, T_{sle}, T_{cm}, and T_{em}) yielded similar observations, with the greatest divergence found between cells isolated early versus late after infection (**Fig. 4.3b** . Importantly, the inter-population JSD metric was not affected by group size . Together these results demonstrate that $CD8^+$ T lymphocytes responding to microbe exhibit substantial molecular heterogeneity at the single-cell level early after infection that diminishes with time.

4.2.3 Distinct transcriptional signatures early after infection

We hypothesized that the heterogeneity observed within lymphocytes early post-infection might reflect distinct gene expression patterns that are predictive of more differentiated cells. We reasoned that supervised classifiers trained on relatively well-defined, differentiated cellular fates, such as sorted T_{cm} and T_{sle} cells, could be utilized to assess whether cells isolated early post-infection might be fated towards specific $CD8^+$ T lymphocyte subsets. Boosted decision trees[31] were chosen over other classification frameworks with similar performance characteristics because the learned trees are easily interpretable. A decision tree that was built from the data consisted of several predictive rules that compare the expression of *Ptprc, Sell,* and *Ccl5* to thresholds learned from that data to decide whether a cell is more T_{cm}- or T_{sle}-like (**Fig. 4.4a**). Ensembles of decision trees were trained with RobustBoost[32] to generate a binary classifier that achieved misclassification error of approximately 4% in leave-one-out cross validation which was split evenly when distinguishing between T_{cm} versus T_{sle} cells (**Fig. 4.4b** . The classifier revealed that *Sel1* and *Il7r* were among the most predictive genes whose high expression accurately described T_{cm} cells, whereas the lack of their expression, along with high expression of *Zeb2*, defined T_{sle} cells (**Fig. 4.4c**). Application of the classifier to cells isolated at days 5 and 7 post-infection revealed that 49% and 57% of total CD8+ T cells at these timepoints were more like T_{sle} than T_{cm} cells (**Fig. 4.4d**), consistent with the expected percentages of T_{sle} cells at days 5 and 7 post-infection[2].

We next asked whether the classifier could discern the fates of responding lymphocytes isolated early during an immune response. It has been previously suggested that asymmetric $CD8^+$ T lymphocyte division yields immune synapse-proximal ("proximal") and synapse-distal ("distal") daughter cells that are differentially fated[6], raising the possibility that these cells might already exhibit distinct gene expression patterns that are predictive of their eventual fates as early as the first cell division. To test this possibility, putative proximal and distal daughter cells, which can be distinguished by their relative abundance of CD8 and CD11a[9],

were sorted and analyzed. The classifier revealed that most proximal daughter cells more closely resembled T_{sle} cells, while most distal daughter cells more closely resembled T_{cm} cells (**Fig. 4.4d**), suggesting that these cells may indeed be differentially fated.

As further evidence that proximal and distal daughter cells display unique molecular patterns that might drive their distinct fates, we observed that these cells exhibited a pronounced disparity in the expression of genes associated with the effector or memory fates (**Fig. 4.4e**). Certain genes associated with the memory fate in $CD8^+$ T cells, including *Eomes*, *Sell*, *Il7r*, *Il2rb*, *Tcf7*, *Id3*, and *Bcl6*[18,19,21,24,25], were more highly expressed in distal daughter cells. Conversely, certain genes associated with terminally differentiated effector cells, such as *Tbx21*, *Prdm1*, and *Grzmb*[19,20,22], were only detected in proximal, but not distal daughter cells. While it remains possible that the gene expression patterns of early lymphocytes might change as the cells continue to differentiate, together these results are indicative of distinct molecular patterns, suggestive of a possible predisposition towards different fates, within cells that may have undergone an asymmetric division *in vivo*.

4.2.4 Predicting temporal expression of key orchestrators of $CD8^+$ T cell fates

Having determined that the gene expression patterns of less differentiated cells could be utilized to predict their eventual fates, we next sought to develop a simple generative model of $CD8^+$ T lymphocyte fate specification that would capture key genes involved in each step of the differentiation pathway of an individual nave cell. In contrast to the classifiers we trained on sort-purified cells to discriminate between differentiated cellular fates (T_{cm} vs. T_{sle}), we used a Hidden Markov Model (HMM) trained on lymphocytes representing intermediate states of differentiation (division 1, day 3, day 5) between the nave state and the differentiated fates (**Fig. 4.5a**). HMMs have been applied to sequential and time-series analyses in diverse fields and have been particularly useful for modeling "hidden", unobserved states during biological processes[33,34]. HMMs not only capture static

expression profiles between subpopulations at a particular stage, but can also detect dynamic expression changes responsible for the transitions between them. To construct a temporal paradigm of T lymphocyte fate specification *in vivo*, we first defined 6 linear and 12 divergent HMMs representing possible hypothetical states (pre-T_{sle}, pre-memory) into which an individual nave T lymphocyte could transition through prior to differentiating into three observed fates (T_{sle}, T_{cm}, and T_{em}). To evaluate each HMM, all possible paths were analyzed for each individual cell . Incorporating the single-cell measurements obtained serially within $CD8^+$ T lymphocytes differentiating *in vivo*, we calculated the likelihood of each of the possible differentiation paths for each defined linear or divergent HMM . To determine both the significance and robustness of each HMM model, we randomly varied the initial values of the transition matrices by 10% and computed the log likelihood for each iteration. Our results showed that the divergent models generally outperformed the linear models, and an early divergent model was identified as the most likely pathway (**Fig. 4.5b**). The performance of this final model was further evaluated by random ordering of the population labels of the cells as well as the associated expression values. Importantly, the likelihood of the best model was significantly ($p=0.00034$) higher than the likelihood for shuffled data, showing that the model robustly indicated that an activated $CD8^+$ T lymphocyte gives rise to cells that transition through either a hypothetical pre-T_{sle} or pre-memory state. Pre-T_{sle} cells can undergo further differentiation to acquire the T_{sle} fate, whereas pre-memory cells can further diverge to give rise to T_{cm} or T_{em} cells. Together these findings suggest that an early divergent model may be the most likely pathway underlying lymphocyte fate specification *in vivo*.

We analyzed the changes in expression of all 94 genes during each of these five unique transitions: nave to pre-T_{sle}, nave to pre-memory, pre-T_{sle} to T_{sle}, pre-memory to T_{cm}, and pre-memory to T_{em} (**Fig. 4.5b,c**) . This analysis revealed both shared and unique molecular features of each transition. The nave to pre-T_{sle} and nave to pre-memory transitions, for example, were both associated with increased expression of *Lgals1*. Notably, however, the nave to pre-T_{sle} transition was associated with higher *Il2ra* and lower *Cxcr3*, *Sell*, and *Tcf7* expression than the

nave to pre-memory transition, raising the possibility that these genes might influence whether a cell proceeds along the pathway towards terminal differentiation or self-renewal. Like the early transitions from the nave state, the pre-memory to T_{cm} and pre-memory to T_{em} transitions exhibited certain shared molecular regulators, including increased expression of *Ccl5* and decreased expression of *Foxo1* and *Cxcr3*. However, the pre-memory to T_{cm} transition was uniquely associated with increased expression of *Tcf7*, *Il7r* and *Sell*. By contrast, the pre-T_{sle} to T_{sle} transition was associated with increased *Ccl5* and decreased *Il2ra*, *Il2rb*, and *Foxo1*. Together these results provide evidence for temporal expression patterns of key genes that influence the fates of $CD8^+$ T lymphocytes responding to microbial infection *in vivo*.

4.2.5 Asymmetric partitioning of IL2Rα in is associated with distinct cellular fates

The prediction, raised by our temporal model, that *Ilr2a* might represent an early molecular switch promoting the pathway towards terminal differentiation was intriguing in light of recent work suggesting a role for IL-2 signaling in $CD8^+$ T lymphocyte differentiation[35-39]. To determine how early in effector- versus memory-fated lineages a possible disparity in *IL2ra* could be detected, we used flow cytometry to examine the expression of IL-2Rα in $CD8^+$ T cells that had undergone their first division *in vivo* in response to microbial infection. We observed that differential abundance of IL-2Rα on the cell surface distinguished two populations of 1^{st} daughter cells (**Fig. 4.6a**) and that IL-2Rα abundance was inversely correlated with CD62L expression (**Fig. 4.6a**), which is highly expressed in T_{cm} cells. Furthermore, cells with higher expression of IL-2Rα also exhibited an increased capacity for IFN-γ and granzyme B production, characteristic of effector cells (**Fig. 4.6b**).

To test the hypothesis that the amount of IL-2Rα expression conferred a distinct predisposition towards the effector or memory lineages, we sorted IL-2Rα^{hi}CD62L^{lo} or IL-2Rα^{lo}CD62L^{hi} cells that had undergone their first division *in vivo*. Cells were then adoptively transferred into recipient wild-type mice that

had been infected 48 hours previously with Lm-OVA. We tracked the progeny of adoptively transferred cells at multiple time points throughout the course of the primary response and found that the progeny of both IL-2Rα^{hi}CD62L^{lo} and IL-2Rα^{lo}CD62L^{hi} cells were detectable following infection . Notably, however, the progeny of the transferred IL-2Rα^{lo} cells exhibited a 4-fold increased capacity to give rise to CD62L^{hi} central memory cells, compared to the progeny of transferred IL-2Rα^{hi} cells (**Fig. 4.6c**). To confirm functionally that these cells were indeed memory lymphocytes, we tested their ability to respond to microbial re-challenge. Recipient mice were re-challenged with Lm-OVA at day 50 after primary infection. We observed a 10-fold increased expansion by the progeny of transferred CD8$^+$ T cells in recipient mice that had received IL-2Rα^{lo}CD62L^{hi} cells compared to mice that received IL-2Rα^{hi}CD62L^{lo} cells (**Fig. 4.6d**), suggesting that these cells exhibit a differential capacity to give rise to memory lymphocytes.

Because certain cytokine and immune receptors can undergo unequal partitioning during cell division[9], we hypothesized that asymmetric segregation of IL-2Rα and CD62L during mitosis might provide a mechanism underlying their differential abundance on daughter cells that had undergone their first division *in vivo*. We used an experimental system that has previously allowed us to examine T cells preparing for their first division in response to a microbe[**9**]. OT-1 CD8$^+$ T cells were labeled with CFSE and adoptively transferred into recipient mice that were infected 24 hours previously with Lm-OVA. Undivided donor CD8$^+$ T cells were isolated by flow cytometry at 36 hours after transfer and examined by confocal microscopy. We observed that IL-2Rα and CD62L exhibited a pronounced asymmetric distribution in cells that were preparing for division (**Fig. 4.6e**). Taken together, these results suggest that the asymmetric segregation of IL-2Rα and CD62L during the first CD8$^+$ T lymphocyte division *in vivo* may influence the transcriptional profiles of the nascent daughter cells and their eventual fates.

4.3 Discussion

Recent advances in high-throughput single-cell gene expression profiling have enabled their utilization in such diverse fields as embryonic development, hematopoiesis, stem cell reprogramming and cancer biology[14-17]. These advances, coupled with computational modeling approaches, enabled us to investigate, on a level of molecular detail not previously possible, the ontogeny of effector and memory lymphocytes during a microbial infection *in vivo*. We find evidence for considerable heterogeneity in gene expression within individual $CD8^+$ T lymphocytes early after the initiation of a microbial infection. Importantly, we demonstrate that this heterogeneity cannot be revealed using traditional bulk population analyses and that many of the computational analyses performed herein, including JSD, classifier and HMM, are possible only with data derived from single cells. These observations provide a compelling argument for the integration of single-cell approaches into future studies of immune cell fate specification.

Using sequential single-cell gene expression measurements within activated lymphocytes during the course of a microbial infection *in vivo*, we constructed a temporal model that enables us to predict the timing and changes in the expression of key genes within individual lymphocytes as they transition from the nave state towards each of several cellular fates. We provide experimental evidence supporting an important prediction of this temporal model– that differential expression of IL-2Rα may reflect one of the earliest molecular determinants influencing the memory versus effector fate decision. Moreover, we demonstrate that unequal partitioning of IL-2Rα during the first asymmetric division *in vivo* may result in its disparate abundance in daughter lymphocytes, potentially contributing to their acquisition of distinct gene expression profiles and cellular fates.

Along with prior evidence that other critical signaling molecules, such as IFN-γR, can be unequally partitioned[9], these results suggest that asymmetric segregation of cytokine receptors during lymphocyte division may result in increased IL-2 and IFN-γ signaling encountered by proximal daughter cells relative to distal daughters. As IL-2 has previously been shown to induce *Prdm1* and repress *Bcl6* and *Il7ra*[37,38], while IFN-γ is known to induce *Tbx21*[40,41], differential cy-

tokine signaling encountered by proximal and distal daughter cells may initiate a pre-effector or pre-memory gene expression program, respectively, consistent with our experimental observations and with prior work showing that cells that receive prolonged IL-2 signals acquire characteristics of terminally differentiated effector cells[37]. Continued changes in gene expression patterns, influenced by environmental signals, may enable lymphocytes to continue along distinct pathways towards terminal differentiation or self-renewal.

Recent reports describing additional subsets of memory T lymphocytes, however, raise the possibility that the effector or central memory lineages may not be the exclusive fate choices adopted by the progeny of IL-2Rα^{hi}CD62L^{lo} and IL-2Rα^{lo} CD62L^{hi} cells. Tissue-resident memory T cells[42,43] do not circulate and instead remain in the peripheral tissues after pathogen clearance, while so-called "effector-phenotype" memory T cells share certain phenotypic characteristics with terminally differentiated effector cells and mediate robust immune protection in certain infectious settings despite exhibiting poor proliferative recall responses[44]. Indeed, the progeny of IL-2Rα^{hi}CD62L^{lo} cells appear to give rise to a population of lymphocytes that, while poorly proliferative in response to microbial rechallenge, persist *in vivo*, reminiscent of effector-phenotype memory cells. Thus, it remains possible that the first cellular division, in addition to mediating a divergence of the effector and memory fates, may also facilitate the specification of distinct memory cell subset fates.

Although the generation of long-lived memory lymphocytes is an essential feature of an adaptive immune response, the fundamental question of when and how these cells arise has remained controversial. Resolving whether lymphocytes progress along a linear differentiation pathway, or diverge early during an immune response, owing to asymmetric cell division, necessitated tracing individual lymphocytes as they undergo differentiation *in vivo*. By interrogating the gene expression patterns of individual lymphocytes during an immune response to microbial pathogen, we have been able to reconstruct the lineage path of single lymphocytes as they differentiate *in vivo*. This approach has yielded new insights underlying lymphocyte fate specification and provides new evidence supporting an

early divergence of lymphocyte fates, via asymmetric division, during an adaptive immune response to a microbial infection. More broadly, we anticipate that single-cell gene expression approaches undertaken by investigators across scientific disciplines, along with ever-improving advances in such technologies as single-cell RNA sequencing[45,46] and single-cell mass cytometry[47], will continue to provide unprecedented molecular insights into cell fate specification in diverse biological settings, including immunity, development, and cancer.

4.4 Methods

4.4.1 Mice

All animal work was done in accordance with Institutional Animal Care and Use Guidelines of the University of California, San Diego. All mice were housed in specific pathogen-free conditions prior to use. Wild-type C57/BL6J mice were purchased from the Jackson Laboratory and OT-1 TCR transgenic mice recognizing ovalbumin peptide SIINFEKL (residues 257-264)/K^b were used.

4.4.2 Adoptive cell transfers and infections

5 x 10^3 OT-1 CD45.1$^+$ CD8$^+$ T cells were adoptively transferred into congenic wild-type CD45.2 recipients, followed by infection intravenously one day later with 5 x 10^3 colony-forming units (CFU) of *Listeria monocytogenes* expressing full-length chicken ovalbumin (Lm-OVA). Splenocytes were isolated from recipient mice at 5, 7, or 45 days post-infection. To isolate cells at 3 days post-infection, 2 x 10^4 OT-1 CD8$^+$ T cells were adoptively transferred. To isolate cells that had undergone their first division, 2 x 10^6 OT-1 CD8$^+$ T cells were first labeled with carboxyfluorescein diacetate succinimidyl ester (CFSE) prior to adoptive transfer and recipient mice were sacrificed at 48 hours post-infection. Cells were stained with fluorochrome-labeled antibodies against CD8, CD44, CD4, CD11b, CD11c, and F4/80, and sorted on a MoFlo (Beckman Coulter) or FACS Aria II (BD Biosciences) flow cytometer.

4.4.3 Microbead-based enrichment

Magnetic bead-based enrichment was performed as previously described[48]. Single cell suspensions were prepared from infected mice that had received OT-1 $CD8^+$ T cells, stained with PE-conjugated anti-CD45.1 antibody, washed, stained with anti-PE magnetic microbeads (Miltenyi Biotec), and enriched through a magnetic column. Cells were then stained and sorted as described above.

4.4.4 Lymphocyte fate tracking experiments

Splenocytes from infected recipient mice that had received CFSE-labeled OT-1 $CD8^+$ T cells were stained with fluorochrome-conjugated antibodies against CD8, CD62L, and IL-2Rα. Cells that had undergone their first division (represented as the second brightest CFSE peak) were electronically gated, and IL-$2R\alpha^{hi}CD62L^{lo}$ or IL-$2R\alpha^{lo}CD62L^{hi}$ cells were sorted. 350 cells of each phenotype were adoptively transferred into separate infection-matched $CD45.2^+$ wild-type recipient mice. The progeny of transferred $CD45.1^+$ T cells were monitored throughout the primary response by serial bleeding. At 50 days post-infection, recipient mice were re-challenged with 5×10^5 CFU of Lm-OVA and expansion of the progeny of donor $CD45.1^+$ T cells tracked in the peripheral blood.

4.4.5 Antibodies and flow cytometry

The following antibodies were used: CD8a (53-6.7), CD45.1 (A20), CD62L (MEL-14), KLRG1 (2F1), IFN-γ (XMG1.2), CD44 (1M7), IL-2Rα (PC61), Vα2 (B20.1), CD4 (RM4-5), B220 (RA3-6B2), CD11b (M1/70), CD11c (N418), F4/80 (BM8), IL-7R (A7R34), and F(ab′)2 anti-rabbit anti-IgG and were obtained from Biolegend or eBioscience. Rabbit anti-TCF-1 (C63D9) antibody was obtained from Cell Signaling Technology. Anti-human PE-conjugated Granzyme B (GB11) was obtained from Life Technologies. For intracellular detection of IFN-γ, $CD8^+$ T cells were stimulated *ex vivo* with 0.25 ng/ml SIINFEKL in the presence of Brefeldin A (Sigma) for 4 hours at 37C; cells were fixed in 4% paraformaldehyde (Electron Microscopy Services) and permeabilized prior to staining. All samples

were analyzed on an Accuri C6 or FACS Canto (BD Biosciences).

4.4.6 Single-cell gene expression assays

Inventoried TaqMan assays (Life Technologies) were pooled to a final concentration of 0.2X for each of the 94 gene expression assays. Single $CD8^+$ T cells were sorted directly into RT-PreAmp Master Mix (Life Technologies) containing the pooled assays. Cell lysis, sequence-specific RT, and sequence-specific amplification of cDNA were performed as previously described[14], and analyzed in 96.96 Dynamic Arrays on a BioMark system (Fluidigm). Ct values were calculated from the BioMark system software. Cells in which both *Actb* and *Rn18s* mRNA expression were detected were retained for further analyses.

4.4.7 Statistical analysis

For statistical analysis, the Kolmogorov-Smirnov test was used for model-free comparisons involving two groups (**Figs. 4.2b, 4.6b,c,d**). Differences at $P < 0.05$ were considered significant.

4.4.8 T lymphocyte confocal microscopy

Immunofluorescence of T cells was performed as previously described[9] with the following antibodies: anti-β-tubulin (Sigma); anti-IL-2Rα(PC61.5), anti-CD62L (MEL14) (eBioscience); and anti-mouse Alexa Fluor 488 and anti-rat Alexa Fluor 647 (Life Technologies). DAPI (Life Technologies) was used to detect DNA. Cells undergoing cytokinesis were identified by dual nuclei and pronounced cytoplasmic cleft by brightfield. Acquisition of image stacks was performed as previously described[9] using a FV1000 laser scanning confocal microscope (Olympus). The volume of 3D pixels (voxels) containing the designated receptor fluorescence was quantified within each nascent daughter in cytokinetic cells as previously described[9] using ImageJ software.

4.4.9 Data and pre-processing

The log expression of each gene g was computed as follows $\log E_{g,c} = 40 - Ct_{g,c}$ where c is the cell and $Ct_{g,c}$ is the Ct value obtained from the Biomark (Fluidigm). Cells c' with undefined Ct values ($Ct_{g,c'} = 999$) for both g =*Rn18s* and g = *Actb*, or cells c'' with at least $60 \leq \sum_{g=1}^{94} 1\{E_{g,c''} \leq 0\}$ unexpressed genes were also removed from our analyses. The remaining "good" cells in each population were deemed sufficient for all subsequent analyses since they exceed the number of free parameters for any supervised model by a factor of at least 5.

4.4.10 Principal component analysis (PCA)

We used principal component analysis (PCA) to reduce dimensionality of the data with a linear transformation and projected the data X from its original 94 dimensions down to the first two principal components. PCA was performed in Matlab using the function `pca`. In order to visualize the clustering of populations, we projected the cells from their original 94-gene space to the first two principal components of X. Each principal component, also known as eigen-gene, captures some percentage of the total variance in X proportional to its corresponding eigen value in the singular value decomposition of X. The first two eigen-genes have the largest eigen values. In order to visualize the contribution of each original dimension to these eigen-genes, we project the 94 unit vectors on to the 2D space spanned by the principal components. These projections combine into the scatter & spike plots in Figs. 2a, 2c, 3a.

4.4.11 T-distributed stochastic neighborhood embedding (tSNE)

To confirm our unsupervised clustering results, in addition to PCA we have also performed t-distributed Stochastic Neighborhood Embedding (tSNE)[29] , which is one of the most powerful dimensionality reduction methods, on our dataset. tSNE is specifically designed for visualization of high-dimensional data and has been shown to capture more useful variance and more complex clustering

patterns in the data by attempting to preserve the distances between datapoints from high to low dimension without any prior assumptions on the distribution of the data. In contrast, PCA only captures linear relationships between genes and principal components and assumes a single homoscedastic (spherical) Gaussian distribution for the entire dataset. The results of tSNE are shown in Supplementary Fig. 2.

4.4.12 Jensen-Shannon divergence

To quantify the differences between the populations and heterogeneity within each population, we use the Jensen-Shannon Divergence (JSD), a symmetric version of the Küllback-Liebler (KL) divergence which is a parameter- and model-free metric of the distance between empirical distributions. Given two sets of experimental measurements, $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$, such as expression profiles for individual cells from the T_{cm} vs T_{em} populations (in this case $x_i \in \mathbb{R}^{94}$), we use the JSD to characterize the distance between the two empirical distributions P_x and P_y implied by the T_{cm} and T_{em} cells, respectively.

$$JS(P_x, P_y) = \frac{1}{2}KL(P_x\|M) + \frac{1}{2}KL(P_y\|M) \tag{4.1}$$

$$KL(P\|M) = \int_z P(z) \ln \frac{P(z)}{M(z)} \tag{4.2}$$

where $M = (P_x + P_y)/2$ is an equal mixture of the two distributions and the KL divergence can be approximated over discretized histograms of its two input distributions

$$P_x \approx \hat{P}_x(i) = \int_i^{i+1} P_x dx \quad \text{and} \quad P_y \approx \hat{P}_y(i) = \int_i^{i+1} P_y dy$$

This is the common form of JSD, which does not take into account the group sizes m and n. In lieu of using the more general form which allows for arbitrary re-weighting of the contribution from each distribution, we randomly sub-sampled the larger group and concluded that the common form we used is not sensitive to

group size differences when those sizes are within a factor of 2, i.e.

$$\min(m, n) \geq \max(m, n)/2$$

We interpret each cell's expression profile as a sample from a 94-dimensional empirical distribution of its population. Expression values for each of the 94 genes is discretized in the same bins, so we can simply add the single-dimensional JSD between the two populations for each gene. Moreover, we can identify the most and least differentially expressed genes between the two populations which need not match the PCA results exactly since the JSD analysis does not make the simple linear modeling assumption that PCA does. Finally, to quantify the heterogeneity within a single population, we partitioned it in half randomly and measure the JSD between the two halves. Averaging this intra-population JSD for multiple random partitions gives an estimate of the true variation in each population.

This approach is more principled than a previous application[14] of JSD to measure single cell diversity which arbitrarily converts each cell's expression profile into a separate probability distribution over RNA molecules. This is a misrepresentation of the BioMark's output which does not distribute a fixed budget of expression units over the 94 genes of interest but rather measures the doubling times for each PCR primer, and can be justified only for single cell RNA-seq experiments where comparable numbers of reads are produced for each cell.

4.4.13 Rationale for approach to supervised analysis of gene expression data

PCA and other unsupervised dimensionality-reduction methods aid in understanding the structure of a cell population. However, these determinations are performed by visual inspection. Suppose we are given a heterogeneous (unsorted) population of cells X'. In order to classify a new cell, i.e. to identify which subpopulation it belongs to, we could co-cluster the new samples with existing labeled data in X. This approach is suboptimal for two reasons: efficiency and accuracy. The co-clustering approach is not efficient because in order to classify even one

new cell x' in X', we need to re-run PCA on the original data X extended by a single row x'. More importantly the accuracy of this approach depends not only on the quality of X, but also on that of x', which we are trying to assess. Suppose that some of the new samples in X' contain bad or noisy readings that are not filtered by our criteria for X. Then the variance inherent in X' will eclipse the useful structure observed in X and the co-clustering result will be unrelated to, or even worse, counter to the original clustering of X. To resolve both of these problems, we decided on a supervised analysis which learns to distinguish between sub-populations of the labeled data X in the form of boosted classifiers and applies the classifiers to the remainder of the cells in X.

4.4.14 Robust boosting

We used RobustBoost[31] to train an ensemble of decision trees at depths ≤ 20. We chose boosting over other classification frameworks because the models that are learned are easily interpretable. For example, the Alternating Decision Tree (ADTree)[30] for the T_{cm} vs T_{sle} classifier consists of simple rules where the expression of *Ptprc*, *Sell*, and *Ccl5* are compared to thresholds learned from the data. The classifier's confidence is measured by the margin of each prediction (see red bars in Fig. 4d). We evaluate the performance of the classifier by its prediction accuracy in leave-one-out cross-validation, where the m classifiers $b_1, b_2, \ldots, b_m$ are each trained on a different subset of $m-1$ cells. Each classifier b_i is tested on cell x_c, which corresponds to the c^{th} row of the data matrix X, after being trained on the remaining cells $X_{-c} = \{x_1, x_2, \ldots, x_{c-1}, x_{c+1}, \ldots, x_m\}$. This cross-validation produces a group of m classifiers that provides an estimate of the generalization error $\epsilon = \sum_{c=1}^{m} b_c(x_c)/m$ on the validation examples. This also generates an overall margin $\gamma = \sum_{c=1}^{m} \gamma_c$ on the training examples by tallying the predictions of $m-1$ informed and 1 uninformed classifier for each of the m cells, where $\gamma_c = \|b_{1\ldots m}(x_c) - l_c\|$ and l_c is the label of cell c (in this case $l_c = -1$ means T_{cm} and $l_c = 1$ means T_{sle}).

4.4.15 Temporal model of CD8^+ T cell differentiation

Akin to the Heisenberg uncertainty principle, the problem of observing a cell's gene expression is that we must modify (i.e. destroy) the cell in order to observe its gene expression. While not a concern in the single-cell analysis of static populations, this is a limitation in capturing the dynamics of tracing the lineage of the cell. We propose a statistical modeling approach to overcome this limitation with approximate single-cell histories sampled from the available time-series gene expression data[32] . Briefly, we constructed hypothetical differentiation paths and trained a hidden markov model (HMM) on the resulting expression time-courses. Starting from each Naïve cell, we sampled cells in successively more mature stages whose expression profiles satisfy an ensemble of predictors for one of the terminal fates, matched these samples in the early differentiation stages (day 1 and day 3), connected both ends of each path, and finally estimated the transition and observation parameters of a 6-state HMM in order to learn the state-to-state transition probabilities and in-state mixture components that capture the dynamics of gene expression in the hypothetical histories.

Input data

To capture the temporal structure of T cell differentiation in our time-course gene expression data from single cells, we developed a semi-supervised method based on the fate classifier predictions in early heterogeneous populations from Fig. 4e and on the expression profiles of putative pre-memory and pre-effector cells sort-purified from Tmp and T_{sle} populations in day 5 after infection. Then, we constructed hypothetical differentiation histories of single cells starting from the Naïve population, going through an intermediate stage and ending in one of the three terminal fates: T_{cm}, T_{em}, or T_{sle}. To approximate the real distribution of proliferation transitions between these stages, we used 1,000 bootstrap samples from each sub-population stringed along one of the three main paths according to their classifier scores. This resulted in an empirical distribution over early transitions: Naïve $\rightarrow$ pre-T_{sle}, Naïve $\rightarrow$ pre-Memory; and another distribution over late transitions: pre-memory $\rightarrow$ T_{cm}, pre-Memory $\rightarrow$ T_{em}, pre-T_{sle} $\rightarrow$ T_{sle}. The

early transitions were connected to the late transitions by cells at the intermediate states as shown in Fig. 5a.

Model structure

Since the differentiation dynamics of individual proliferating T cells are not yet well described, we used an HMM to model the data because of its simple, yet powerful structure which decouples uncertainty in the lineage reconstruction (state transitions) from measurement noise (observations/emissions). We constructed a HMM with 6 states: Naïve, pre-memory, T_{cm}, T_{em}, pre-T_{sle}, T_{sle}, to capture the signal in each empirical distribution from our temporal approximation input. Each state emits gene expressions from a mixture of two 94-dimensional Gaussians with full covariance matrices.

Due to concerns over our model's sensitivity to initialization, we constructed 18 biologically plausible differentiation pathways (6 sequential and 12 bifurcating, whose structures are shown in Supplementary Figs. 8 and 9, respectively) and fixed the transition parameters to the corresponding adjacency matrix of each structure in turn. Using the learning algorithm described below, we calculated the posterior log likelihood of each pathway. To address any further concerns over the robustness of these results, we reinitialized each structure twice more with 10% random noise drawn from the `Uniform[0,1]` distribution, which also ensured that there are no zero-probability transitions between any two states.

Transition parameters

For a cell c in state f, the probability of transitioning to state t is $T^c_{f,t}$. We assume that other cells whose expression profile in state f is similar to that of cell c will have similar differentiation potential and in particular have similar probability of transitioning to state t. This assumption lets us share the parameters $T_{f,t} = P(f \to t)$ which gives the probability of any cell in state f to proliferate to state t.

Observation parameters

Due to the bimodal nature of the violin plots in Fig. 4f, we model the observed expression x of cell c in state i, as a mixture model of two Gaussians with 94-dimensional means μ_i^c and η_i^c, and 94×94 full covariance matrixes Σ_i^c and Ξ_i^c. Like the transitions, parameter sharing between cells allows us to simplify the observation parameters, so we can write down the observation model:

$$P(x|s = i) \propto a_i \mathcal{N}(x; \mu_i, \Sigma_i) + b_i \mathcal{N}(x; \eta_i, \Xi_i)$$

Learning algorithm

First, we initialized the model parameters to their prior distributions. Specifically, the transitions $P(f \rightarrow t)$ were initialized to the matrix $T^0_{f,t}$ shown in Fig. 5b. The emission parameters for the Naïve, $\mathrm{T_{sle}}$, $\mathrm{T_{cm}}$, and $\mathrm{T_{em}}$ states were initialized to the maximum likelihood fit for a mixture of two Gaussians to the empirical histograms of gene expression for the respective population. The emission parameters in the intermediate states: pre-memory and pre-$\mathrm{T_{sle}}$ were fit to the empirical histograms accumulated over all intermediate states. The transition parameters were fixed throughout the duration of each learning run, but were randomized with up to 10% noise as detailed above.

Finally, we optimized the parameters of the HMM using the Expectation Maximization algorithm implemented in pmtk3, the probabilistic modeling toolkit for Matlab/Octave [33]. The learned emission parameters were used to identify the genes whose relative expression changed the most during each transition, as shown in (**Fig. 4.5b**) and summarized in (**Fig. 4.5c**). While we did not learn the transition probabilities, we did re-sample them from 18 plausible structures and picked the most likely structure whose transition matrix is shown in Supplementary Fig. 10 and whose adjacency graph is on the far right of Supplementary Fig. 9. To determine the most likely structure, we calculated the posterior likelihood of each model (in each of its 3 re-initializations) and compared them visually using box plots in Supplementary Figs. 8 and 9. To further gauge the statistical significance of the best model, we randomly shuffled the input data 20 times and

built a background distribution of the resulting log-likelihoods. This background approached a normal distribution with mean log-likelihood of -1.13e+06, which is 2.53 standard deviations worse than the average log-likelihood of the best model, -1.02e+06.

4.5 Acknowledgements

This chapter is adapted from J Arsenio*, **B Kakaradov***, PJ Metz, SH Kim, GW Yeo, JT Chang. Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. Nature Immunology. (2014). The dissertation author was a joint first author of this paper, and was responsible for all computational research.

4.6 Figures

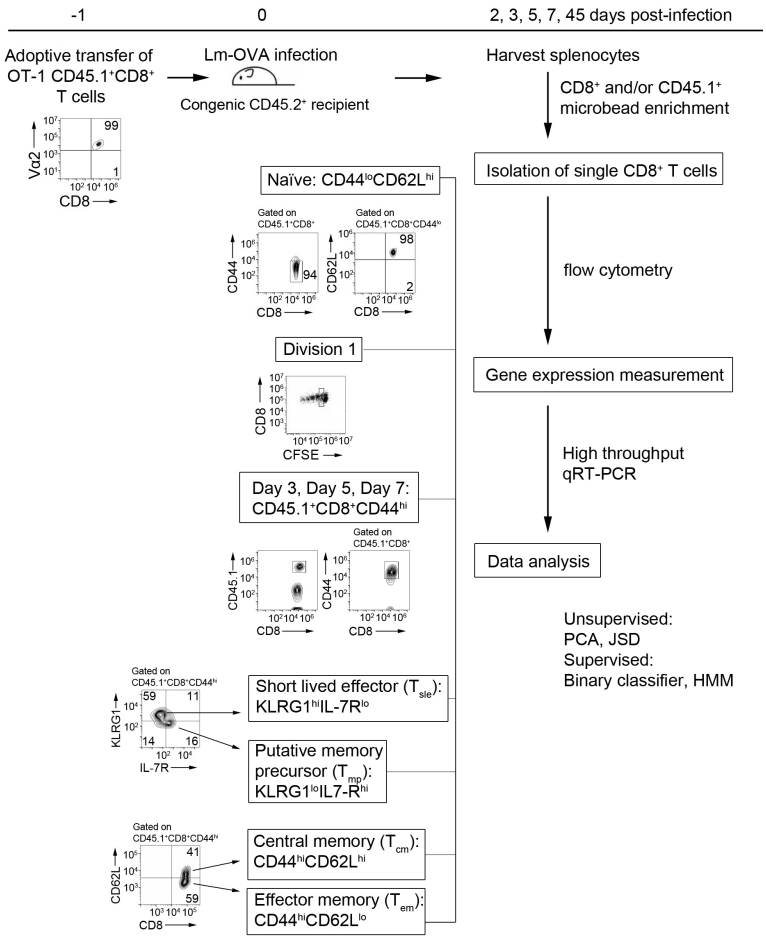


Figure 4.1: Gating strategy and experimental approach for single-cell gene expression analyses of $CD8^+$ T cell subsets isolated from uninfected (nave, $CD8^+CD44^{lo}CD62L^{hi}$) or CD45.2 recipient mice infected with Lm-OVA 24h after intravenous adoptive transfer of unlabeled or CFSE-labeled $CD45.1^+$OT-1 $CD8^+$ T cells. $CD8^+$ T cell subsets were isolated at various time points post-infection: division 1 ($CD8^+CD45.1^+CD44^{hi}$ cells within 2^{nd} brightest CFSE peak); days 3, 5, and 7 post-infection; day 7 T_{sle} ($CD8^+CD45.1^+CD44^{hi}KLRG1^{hi}IL\text{-}7R^{lo}$), day 7 putative T_{mp} ($CD8^+CD45.1^+CD44^{hi}KLRG1^{lo}IL\text{-}7R^{hi}$), day 45 T_{cm} ($CD8^+CD45.1^+CD44^{hi}CD62L^{hi}$), and day 45 T_{em} ($CD8^+CD45.1^+CD44^{hi}CD62L^{lo}$). Data are representative of three experiments. Data analysis approaches included unsupervised Principal Component analysis (PCA), and Jensen-Shannon Divergence (JSD), and supervised binary classifier and Hidden Markov Model (HMM).

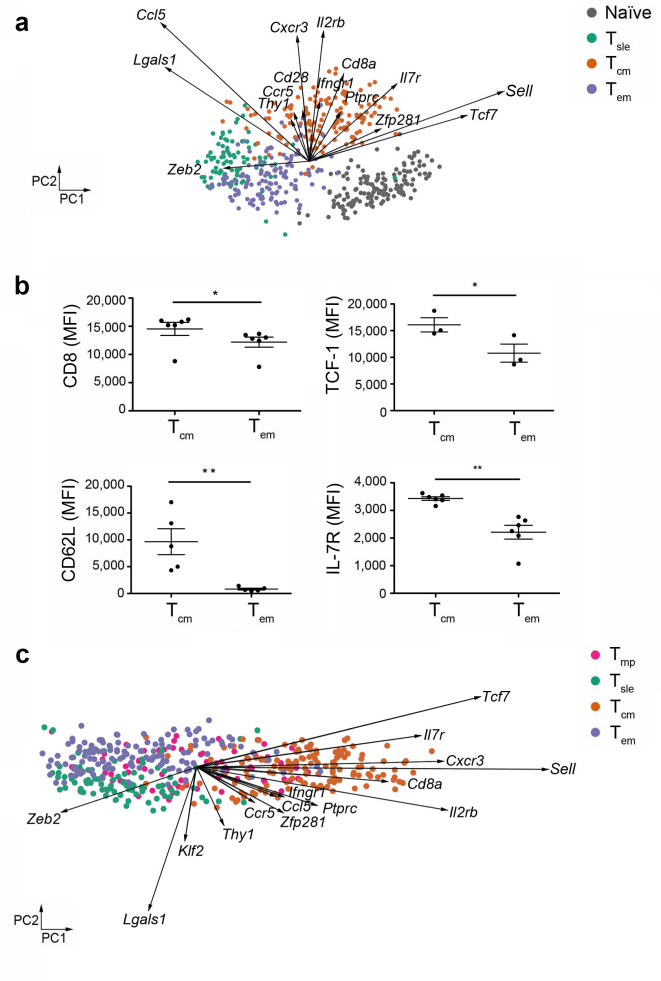


Figure 4.2: Effector and memory CD8$^+$ T lymphocyte subsets are molecularly distinct on a single-cell level. (**a**) Principal component (PC) projections (PC1, horizontal axis; PC2, vertical axis) of single-cell gene expression data derived from individual lymphocytes from the indicated populations. Each circle represents an individual cell of the indicated population: nave (gray), T_{sle} (green), T_{cm} (orange), and T_{em} (purple) cells. Each vector emanating from the origin represents an individual gene. PC1 and PC2 account for 11% and 9% of the variance, respectively. (**b**) Mean fluorescence intensity (MFI) of CD8 (*Cd8a*), TCF-1 (*Tcf7*), CD62L (*Sell*), and IL-7R (*Il7r*) protein expression in T_{cm} and T_{em} cells, assessed by flow cytometry. * $P < 0.05$, ** $P < 0.01$ (Kolmogorov-Smirnov test). Data are representative of two experiments with at least 3 mice in each experiment (error bars, s.e.m.). (**c**) PC projections of single-cell gene expression data derived from individual lymphocytes from the indicated populations: T_{mp} (pink), T_{sle} (green), T_{cm} (orange), and T_{em} (purple) cells. Each vector emanating from the origin represents an individual gene. PC1 and PC2 account for 11% and 6% of the variance.

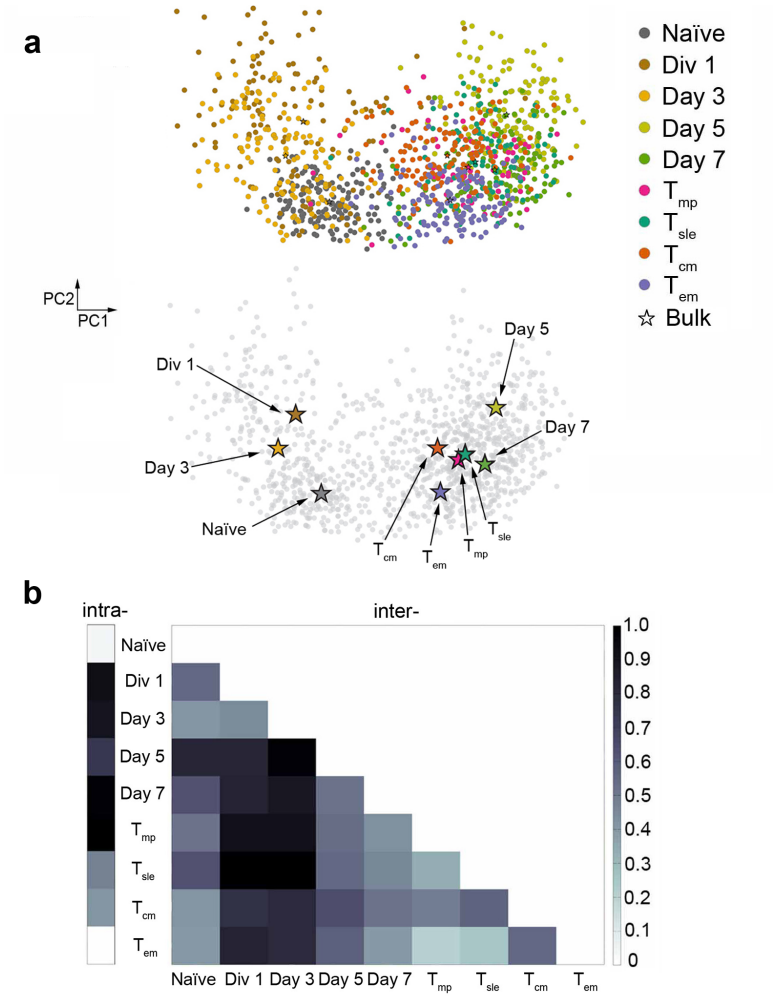


Figure 4.3: Early heterogeneity of gene expression exhibited by individual CD8$^+$ T lymphocytes during an immune response. (**a**) Projections of single-cell gene expression data derived from individual lymphocytes from the indicated populations (top). Each circle represents an individual cell of the indicated population representing: nave (gray), division 1 (brown), day 3 (yellow), day 5 (light green), day 7 (green), T_{mp} (pink), T_{sle} (teal), T_{cm} (orange), and T_{em} (purple) cells. PC1 and PC2 account for 10% and 7% of the variance, respectively. Analysis derived from pooled "bulk" samples from each experimental condition, shown as colored stars (bottom). Stars filled with each color represent "bulk" nave (gray), division 1 (brown), day 3 (yellow), day 5 (light green), day 7 (green), T_{mp} (pink), T_{sle} (teal), T_{cm} (orange), and T_{em} (purple) cells with grayed-out single-cell data points in the background for clarity. (**c**) Intra- (left) and inter-population (right) Jensen-Shannon Divergence (JSD) of mean gene expression within and between the indicated CD8$^+$ T cell populations is shown.

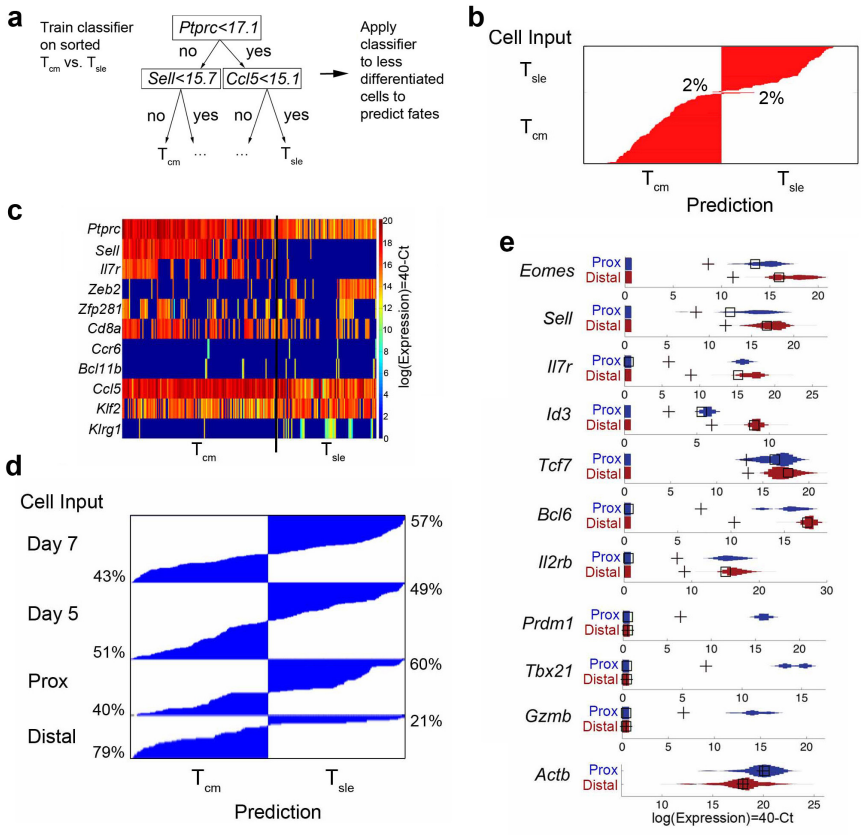


Figure 4.4: Classifier analysis predicts eventual fates of individual CD8$^+$ T lymphocytes. (**a**) Decision tree built from the data consisting of several predictive rules comparing expression of *Ptprc, Ccl5,* and *Sell* to decide whether a cell is more T_{cm}- or T_{sle}-like; two terminal nodes labeled "..." depict a continuation of the decision tree. (**b**) Predictions by the classifier on sort-purified T_{cm} and T_{sle} cells that were cross-validated during training. Horizontal red lines indicate the voting margin for each individual cell and internal confidence of the classifier's prediction for that cell; percentages indicate rate of misclassification of a T_{sle} as T_{cm} and of a T_{cm} as T_{sle}. (**c**) Binary classifier trained to distinguish between a pair of differentiated cell fates (T_{cm} vs. T_{sle}). Single vertical lines along the x-axis represent each individual sort-purified T_{cm} or T_{sle} cell and its expression of each gene. (**d**) Individual CD8$^+$ T cells (horizontal blue lines) from the indicated populations (cells isolated at day 5 or 7 post-infection; proximal ("prox") or distal daughter ("distal") cells at the first division) were interrogated by the classifier and predictions were sorted by confidence from the most T_{cm}-like to most T_{sle}-like cells. Percentages indicate proportion of cells predicted to be more T_{cm}-like (left) or T_{sle}-like (right) within each cell population. (**e**) Violin plots showing expression levels of the indicated genes by first division proximal (blue) and distal daughter (red) cells. Black crosses and squares represent mean and mode values, respectively.

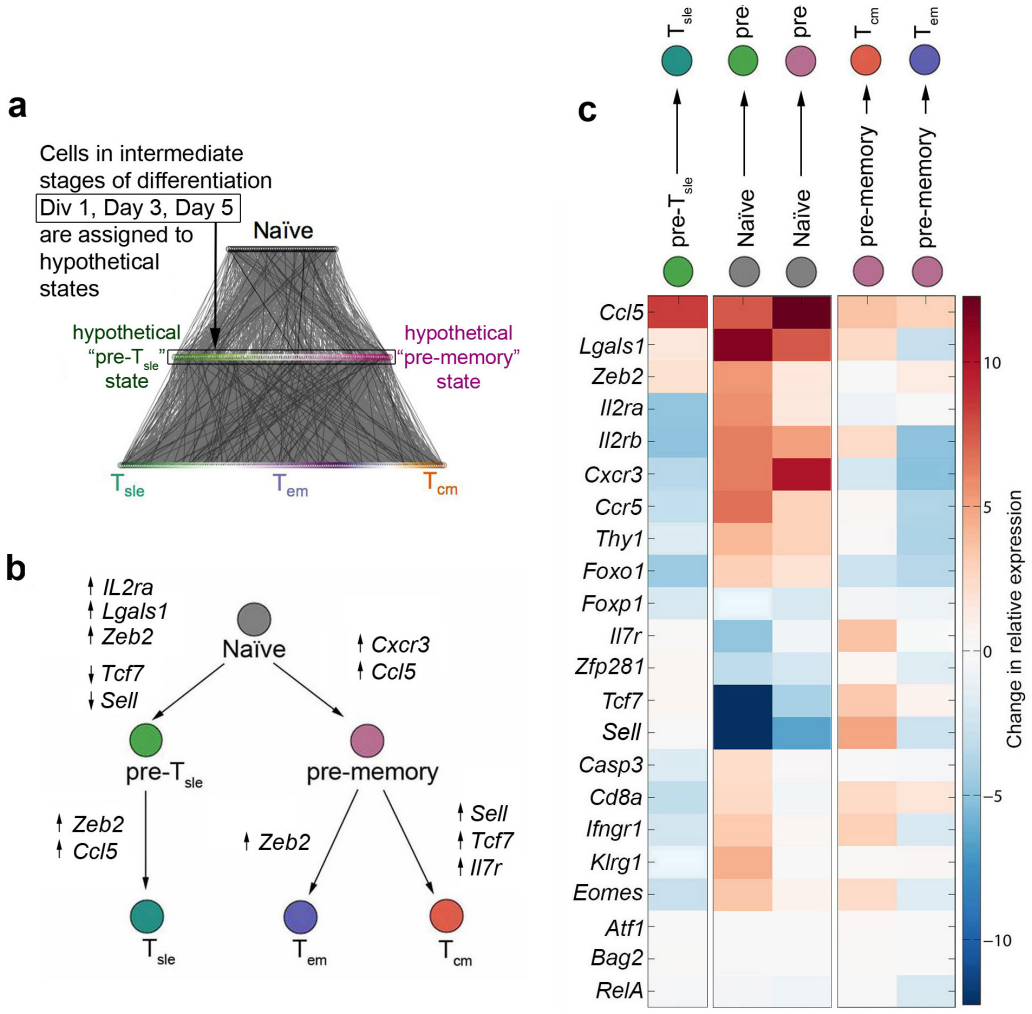


Figure 4.5: Temporal model predicts the differentiation paths of individual $CD8^+$ T lymphocytes. (**a**) Cells in early states of differentiation (division 1, day 3, day 5) were ranked by their T_{sle}- or memory-like expression profiles (green to purple gradient in middle row). Cells were then linked to sorted nave (black top row) and sorted T_{sle}, T_{em} and T_{cm} cells (green to purple to orange gradient in bottom row) in a random fashion, forming hypothetical differentiation paths (black lines) that were analyzed with a Hidden Markov Model. (**b**) Most likely model of $CD8^+$ T lymphocyte differentiation with key gene expression changes associated with each of 5 unique transitions: nave to pre-T_{sle} , nave to pre-memory, pre-T_{sle} to T_{sle}, pre-memory to T_{cm}, and pre-memory to T_{em}. Colored circles represent each cell state or fate. (**c**) Summary of key changes in gene expression during each transition phase predicted by temporal model of $CD8^+$ T lymphocyte differentiation.

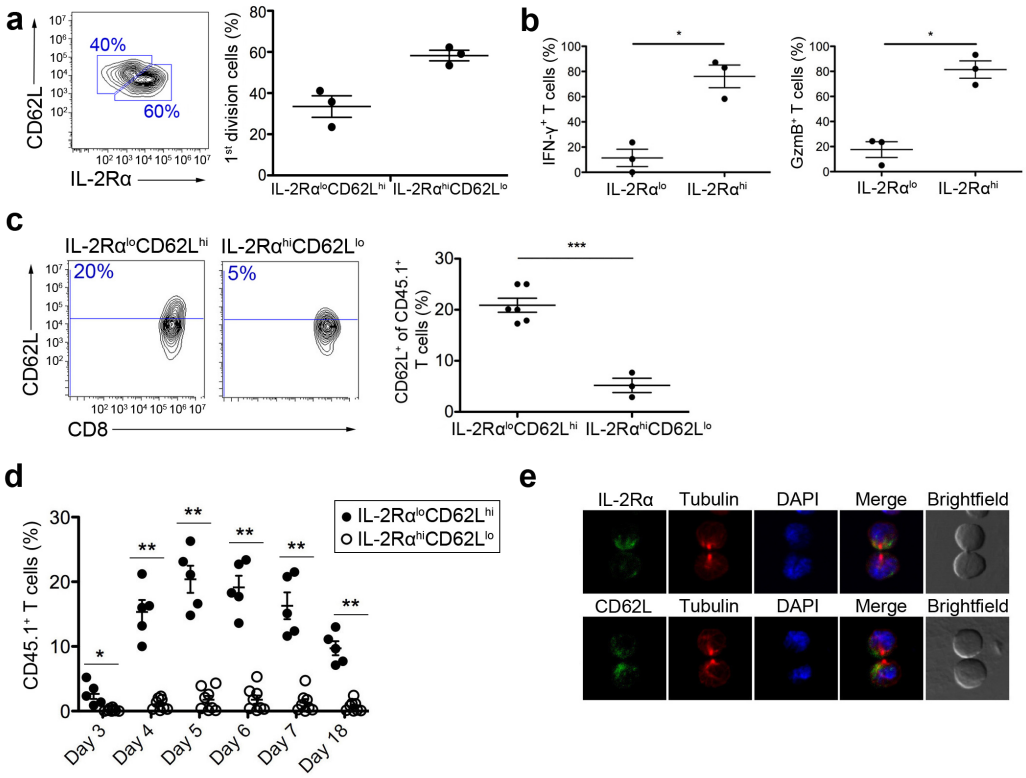


Figure 4.6: Asymmetric segregation of IL-2Rα during T lymphocyte division influences the eventual fates of the daughter cells. (**a**) IL-2Rα and CD62L expression (left) by OT-1 CD8$^+$ T cells undergoing their first division following adoptive transfer into recipients and subsequent infection with Lm-OVA 24h later. Frequencies of IL-2Rα^{lo}CD62L^{hi} and IL-2Rα^{hi}CD62L^{lo} cells (right); each circle represents an individual mouse and lines indicate the mean. (**b**) Frequencies of IFN-γ and granzyme B expression by IL-2Rα^{lo}CD62L^{hi} and IL-2Rα^{hi}CD62L^{lo} cells as in (**a**). (**c**) CD62L expression (left) on d49 post-infection by CD45.1$^+$CD8$^+$ T cells in CD45.2$^+$ mice (n=13) that had been previously challenged with Lm-OVA and injected with sort-purified 1st division IL-2Rα^{lo}CD62L^{hi} or IL-2Rα^{hi}CD62L^{lo} cells 48h later. Frequencies of CD62L$^+$ cells (right); each circle represents an individual mouse and lines indicate the mean. (**d**) Expansion of CD45.1$^+$CD8$^+$ T cells, assessed by serial bleeding, in mice depicted in (**c**) that were subsequently re-challenged with Lm-OVA at d50 post-primary infection. (**e**) Morphology of IL-2Rα or CD62L (green), β-tubulin (red), and DNA (blue), assessed by confocal microscopy, in sorted OT-1 CD8+ T cells undergoing their first division following adoptive transfer into LM-OVA-infected recipients. Asymmetric segregation of IL-2Rα and CD62L was observed in 60% (n=96) and 62% (n=74) of cells, respectively. Data are representative of 2 (**c**, **d**) or 3 experiments (**a, b, e**); error bars represent s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P = 0.0002$ (Kolmogor-Smirnov test).

Chapter 5

Discussion and Future Directions

5.1 Biology and statistics in the era of big data

The problem with single-hypothesis investigation is that when your hypothesis turn out to be incorrect, often that is the end of the project. Even if the incorrectness of the hypothesis is somehow significant by itself and may be noted in a publication, the data collected to invalidate it will not usually be useful in other studies. In contrast, with high-throughput technologies, it is now possible to design multi-hypothesis experiments from the start. Even if the initial motivation for collecting that data is a particular hypothesis and it turns out to be incorrect, a high-throughput dataset can enable both its creators and other investigators to check many other hypotheses when the dataset is shared with the broader scientific community. However, there is a limit to the size of the class of hypotheses that a single experiment can entertain. This limit is imposed by the technological and budget limitations. Practically, there is a limit to the amount of resources such as reagents and time that can be devoted to a single experiment, no matter how broad-reaching its results might be. Therefore, most high-throughput technologies are designed to trade accuracy for scale. For example, in the contrast between RNA-seq and RT-PCR, the former has much larger scale and is seen as an exploratory method, while the latter is a lot more accurate and is seen primarily as a validation method. That is true, until the introduction of targeted sequencing which can adjust the tradeoff between scale and accuracy more smoothly. Now,

the original methods are perceived simply as two extremes on the spectrum of accuracy-scale settings. Novel iterative experiment designs that start on the exploratory end of that spectrum and work toward the accuracy end (**Fig. 5.1**) are already being adopted by many labs, including ours. They allow a smoother transition from general exploration to focused investigation, which is especially relevant for single-cell transcriptomic studies.

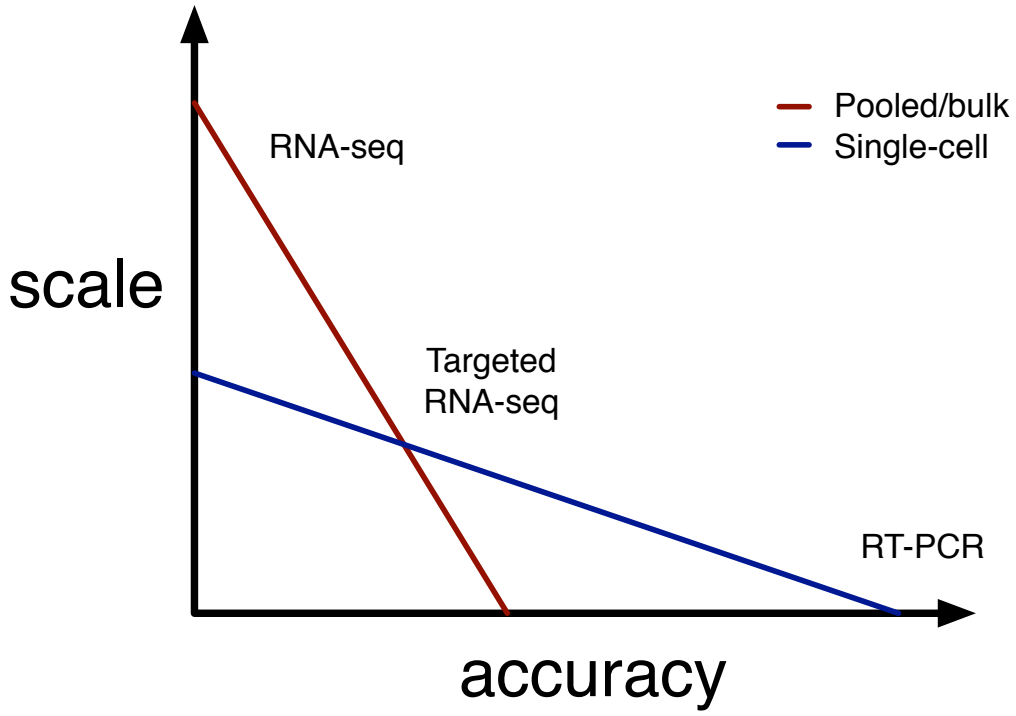


Figure 5.1: Approximate tradeoff curve depicts the smooth transition between ledge scale, less accurate assays versus smaller-scale but more accurate assays. Tradeoff curve is steeper for pooled/bulk samples (red) than for single-cell samples (blue) because the former can utilize high-coverage libraries representing a wider range of transcripts pooled from multiple cells, but the inherent averaging of pooling decreases the maximum possible accuracy of pooled-cell assays. On the other hand, the scale of single-cell experiments is still limited to hundreds of cells at a time by the current microfluidic technology; however, single-cell RT-PCR provides the most accurate measurement of gene expression possible with current tools.

Just as the properties of experimental design are becoming more flexible, so is the intellectual border in computational and scientific training for biologists. In the past 8 years, interdisciplinary collaborations in which I have participated have

gotten cozier in both physical and intellectual space. During my first collaboration, I re-analyzed proteomics data which had already been published independently of the original authors and only sent them predictions after months of computational experimentation. Accordingly, the turn-around time for the follow-up publication was 3 years. In the first half of my PhD career, I worked in a closer collaboration with monthly check-in meetings on a mixed dataset of published and novel results. The total time to publication was halved to one and a half years. I currently work in a hybrid wet/dry lab and have tight-knit collaborations with experimental biologists with weekly meetings where I can participate not only the the post-production data analysis, but also in the experimental design. This enabled me to work on two projects simultaneously, each about 1 year long. There is a clear trend which converges on a logical conclusion–biological research will be increasingly computational/statistical not only in its analysis but also in its design. Encapsulating the full experimental loop into a single mind or at least a single tight-knit group cuts down on communication overhead and standardizes the stages for each scientific project so that any of the participants will eventually play the role of experiment design, data generation and analysis, and hypothesis evaluation. As new biologists are trained to be increasingly computationally and statistically literate, eventually the specialized role of bioinformatician/biostatistician will hopefully disappear as it has for computational physicists (now simply called physicists). However, simply combining twice the knowledge into the mind of a single biology expert is not only logistically challenging, but also risks producing "jack-of-all-trades" generalists who may not have deep enough understanding of either field to advance the state of the science. To mitigate this risk, we must borrow robust abstraction and automated testing techniques from the computing hardware field, where a software developer relies on the expertise of language and compiler designers, who rely on the expertise of electronic hardware engineers, who in turn rely on the expertise of materials scientists to enable their increasingly powerful designs without increasing their complexity. Similar semi-automated systems will be developed in biology for prioritizing experimental conditions, evaluating hypotheses, and order-ranking validations. Large multi-national consortia such as ENCODE

have already started to realize this, and some of their most-computationally gifted members are shifting their focus away from data analysis improvements with diminishing returns to choosing how to fill out their sparsely-populated experimental matrix (of condition-by-assay) in an optimal way from information-theoretic and resource-practical viewpoints.

5.2 From read-only to read+write biology

Virtually all of the analysis methods presented in this dissertation were designed for a read-only experimental paradigm in which the study design is already or mostly completed by biologists and data are either already or in the process of being generated, so the only thing left for a computational expert is to analyze said data and make novel predictions about the system. I call this analytical paradigm read-only bioinformatics because the main information only flows from the wet bench, through the sequencing center, to the compute cluster. There are many great discoveries have been made in this mode of operation. Note that exclusively human scientists were involved with the first stage, experimental design, and last stage, result interpretation and validation, of these studies. However, there is a a lot to be gained from closing this experimental loop and partially automating even these steps which have been reserved for humans. For example, in addition to ranking its predictions by confidence, a computational system can automatically design the validation studies, perform some of them, check how far off its predictions were and re-analyze the original data based on this new domain knowledge and adjust its predictions, coming full circle and perhaps repeating for a few more iterations until it *appears* to match the human intuition of which results are scientifically valid and interesting. Eventually, automated systems will be able to optimize and perform much of the experimental design and I call this experimental paradigm read+write bioinformatics, and hope/believe that it is the future of our field.

This may seem like overly-optimistic dreaming by a computational biologist not fully aware of the practical difficulties and physical limitation of actual experiments. However, after spending a very educational and humbling two months in

the wet lab trying to make simple RT-PCR work, I can guarantee that my optimism is not due to oversimplifying the unfamiliar. On the contrary, I believe that my optimism is necessary for us to make the next leap in progress. Our current read-only paradigm is akin to observing a lightly-biased coin come up with more successes than failures, but in order to accelerate the rate of success, we have to scale the rate of coin flipping accordingly. Instead of restricting ourselves to passive observation, a much better approach is to actively increase the success bias of the coin. Several statistical frameworks such as active learning and Bayesian optimization already exist that readily take into account both observational and interventional experiments. On the laboratory side of biology, there are two fields which have already stated building high-throughput interventional tools and moving in this direction. They are genome editing and synthetic genomics.

5.2.1 Synthetic Genomics

Synthetic genomics is a relatively new field that engineers organisms to acquire a specific property, gain a specific function, or produce a specific compound by explicitly re-programming their genomes. It is distinguished from agricultural practices of breeding animals or plans with desirable properties in that the generic changes are pre-selected, monitored, and tightly controlled. Most successfully, it has been applied to modify or optimize microbial metabolism for the production of biofuels or other valuable compounds. One of the most significant and popular achievements in synthetic genomics was the creation of a 'new' organism, *micoplasma laboratorium*, whose genome was synthesized by scientists at the J. Craig Venter Institute based on one of the simplest existing microbes, *micoplasma genitalium*. Even though I had no expertise in synthetic genomics at the time, I quickly realized the potential for this approach not only in designing custom genomes for the synthetic genomics community but also for organizing useful modules of genes based on similarity along the phenotype of interest. There are two roadblocks to using synthetic genomics as the "write" part of the design-data-analyze-validate-intervene experimental loop. First, the size of synthetic genomes currently achievable is only suitable for prokaryotic organisms, while most of the

interesting transcriptomic variations (such as splicing and editing) are only present in eukaryotes. Second, to reach the full potential of this technology, molecular biologists will need to construct a combinatorial number of synthetic genomes with every RBP binding motif or splice/edit site modified in combination with others. This is truly not achievable, limiting the utility of synthetic genomics for transcriptome studies to reporter systems and tags to be inserted via plasmids and activated conditionally by environmental factors for all cells in a given condition.

5.2.2 Genome Editing

Genome editing is a very new field that enables molecular biologists to make certain modifications to specific loci in existing genomes. The locus complementarity is achieved through DNA-binding domains and sequence-recongnition molecules on proteins in one of three main families: Zinc fingers, TALENs, or CRIPRs. Discovered in that order, these genome recognition systems evolved as anti-intrusion adaptations in bacteria. In the hands of molecular bioengineers, they have become increasingly more effective at editing and specific in targeting the intended loci. The CRISPR/Cas9 system is the latest and most efficient for genome editing in general, not only because of its increased span (it can target any sequence of nucleotides, unlike zinc fingers and TALENs), but also because of it's increased specificity to that sequence through the use of a complimentary guide RNA. Genome editing is relevant to transcriptomics not only as the interventional part of the experimental cycle, but also as the potentially therapeutic part of translating disease studies into medicine. For example, genome editing can be used to rescue the RNA editing deficiency in aging neurons of ALS patients by mutating the GRIA2 Q/R site directly. Because the CRISPR/Cas9 system acts *in vivo*, and infecting a patient's brain with lentivirus is not yet considered safe, the most likely method of delivering the benefits of genome editing to the clinic is through the use of patient-derived induced pluripotent stem (iPS) cells. Because patient-derived iPS cells contain the original genome, they can be directed to neural lineages to model the particular disease, and to screen classic drugs, and novel therapies such as permanent genome editing or transient siRNA-, shRNA-,

or Cas9-mediated gene silencing.

5.2.3 Open questions on RNA editing

How essential is RNA editing in the brain, really? As mentioned in Chapter 1, ADAR proteins are essential. Knocking out either of them results in abnormalities for worms, or embryonic lethal in mammals. RNA editing of well-known targets is also essential. However, if the well known GRIA2 Q/R site is manually edited by mutating the genome, and both ADAR genes are knocked out, the mutation seem to rescue the The resultant mutant/knockout mice survive into adulthood with mild but peculiar behavioral

5.3 Disease Diagnostics

RNA biology is worth pursuing not only for a fundamental understanding of life, but also for discovering was to better that life. Even more exciting than exploring the regulation of RNA expression, splicing, and editing, is determining the causes and effects of their misregulation in disease. With increased knowledge of the working system comes the potential to understand its faults, and a promise to engineer a cure for them.

Appendix A

Open thoughts on science

Asking the right questions has never gone out of style, but how do we define right *a priory*? In hindsight, it's easy to see which questions were right because the background was complete enough, the experimental methods were perfected enough and the question had a large enough audience that its answer made a significant impact. These are a lot of aspects to get just right, without fully understanding the scientific, technological, and even political forces that govern a given contribution. It gets so complicated that a lot of smart people, including seasoned professors throw up their hands and sometimes attribute to luck all of the aspects they cannot account for, much less control for. However, this is a very unscientific approach and I would like to propose a more principled way of exploring this issue.

How could we define these intuitive metrics of "rightness" and write an objective function in these terms. To keep it simple, let's restricts ourselves to the set of predictions from one of our existing high-thoughput dataset analyses. The questions we can ask are limited to: 1) is this a valid RNA edit or splice event, and 2) does it vary between healthy and diseased samples. We can ask these two questions of every predicted edit/splice event and validate the answer with RT-PCR. For thousands of edit sites and tens of thousands of splice events, this is prohibitively expensive. So, we usually draw an arbitrary confidence cutoff and validate the top 100 most-confident predictions.

Appendix B

RNA editing math

B.1 Introduction

To understand the functional roles of ADAR1 and ADAR2, RNA editing sites must be identified and quantified accurately with ssRNA-seq data from each of the single- and double-knockout strains and compared to wild type worms. We achieve accurate identification by combining filters from existing pipelines [**?**, **?**, **?**, **?**] in a strand-specific manner and accurate quantification by extending the existing Bayesian method for genomic variant calling used in the 1000 Genomes project [**?**]. In addition to leveraging established considerations with regards to read sequencing and alignment errors [**?**], our approach benefits greatly from using the ADR1-/ADR2- double-knockout strain as a powerful filter for unannotated variants to maintain low false positive rates while confidently identifying RNA editing sites.

B.2 Pipeline description

1. ssRNAseq: The ADR1-/2- double-knockout sample was sequenced on one lane of Illumina's HiSeq 2000 yielding 216 million single-end 76nt reads. Each other sample was sequenced on a lane of Illumina GAII yielding between 37 and 42 million reads of the same type.

2. Mapping: Sequenced reads were mapped to the C. elegans reference genome (ce10, WS220) with the spliced aligner TopHat (version 2.0.6) allowing only uniquely-mapped reads with up to two mismatches each with command line options -Mx 1 and -N 2.

3. variant calling: sites with RNA-DNA differences were identified by SAMtools mpileup (version 0.1.18) tallying up to 1000 alignments per site. Additional command line options used were -D -I and -g.

4. Site filters: Annotated SNPs were obtained from Illumina's iGenomes collection for C. elegans (ce10) and unannotated variants were extracted from the ADR1-/2- double-knockout strain. These genomic variants were filtered from the putative sites in all other strains reducing the number

5. Read filters: Each read aligned to one the remaining putative sites was filtered out if: a) it was a suspected PCR duplicate, according to SAMtools rmdup (version 0.1.18) b) it had a junction overhang of less than 10nt according to its SAMtools CIGAR string c) it had more than one non-A2G or non-C2T mismatch or any short indel, according to its SAMtools MD tag. d) it had a mismatch less than 25nt away from either end of the read (this was changed to 5nt in the relaxed version used for quantification)

6. Identify sites: Putative RNA editing sites were identified from A2G variants on the sense strand and T2C variants on the antisense strand that were covered by more than 5 reads which passed the filters in step 5, including the stringent 25nt threshold for filter 5d).

7. Quantify sites: The extent of editing at each site and our confidence in that prediction were quantified by a novel extension of the classical Bayesian model used for genomic variants, which is described in more detail in the next section.

8. To increase the accuracy and confidence of our predictions, we used additional reads from the relaxed version of filter 5d) that overlap the sites identified in step 6. Moreover, we dropped sites which exhibited editing in 100% of the

reads (suggesting a genomic variant not filtered out by step 4) and those with very low editing (less than 10%), which would have been hard to distinguish from sequencing errors.

9. The predicted RNA editing sites from each strain were characterized according to their position in annotated genic regions (introns, exons, 3'/5' UTRs, etc.) and according to their overlap with other strains. Finally, 50 out of over 400 sites predicted in the ADR1- or CEN2 strains were validated by Sanger sequencing.

B.3 Details of Bayesian quantification model

Also known as the "inverse probability model" in the SNP calling community [**?**], a Bayesian model for identifying DNA polymorphisms from error-prone sequencing data has been shown to perform favorably to other discrete and discriminative models [**?**, **?**]. In general, the power of a Bayesian approach is its combination of prior knowledge and observed data into a posterior estimate. The prior knowledge encodes general domain-specific information like biases in the sequencing technology, while the observed data contain signals specific to editing sites in a particular experiment. In this exposition, we will use a simple context-independent prior for all editing sites, which consist of pseudo-counts of edited and non-edited reads: β and α, respectively. For sequence alignments in particular, the benefit of a Bayesian approach is that even low-coverage regions can give reasonable posterior estimates of the editing efficiency with low confidence, while high-coverage regions will give very accurate posterior estimates with high confidence.

For example, consider two candidate editing sites: site L has low coverage and site H has high coverage. Let the number of reads from edited (g) and unedited (a) transcripts containing those sites be: $g_L = 1, a_L = 9$ for site L and $g_H = 10, a_H = 90$ for site H. The observed counts suggest that both sites are edited with 10% efficiency, but we are inclined to believe that site H really is edited while site L is not and its single edited read could have easily been produced by a sequencing

error. While filter-based approaches require manual fine-tuning to be able to filter out site L while keeping site H, the Bayesian approaches will simply have a lot more confidence that site H is edited. To formalize the notion of confidence, we introduce a latent binary variable γ which indicates whether a nucleotide is edited $\gamma = 1$ or not $\gamma = 0$. Given a prior belief in the occurrence of RNA editing $P(\gamma_S) = \frac{\beta\gamma_S + \alpha(1-\gamma_S)}{\alpha+\beta}$ at a particular site S (which is currently site-independent but can be extended to differ depending on the genomic context or read position of S), and the likelihood of observing the RNA-seq reads at site S conditioned on the hypothesis of editing $P(a, g|\gamma_S = 1)$ versus no editing $P(a, g|\gamma_S = 0)$ which captures the probability of a sequencing error ϵ, Bayesian models for DNA-RNA differences use the "inverse probability" rule to produce a posterior belief on whether site S is edited or not:

$$P(\gamma_S|a, g) = \frac{P(\gamma_S)P(a, g|\gamma_S)}{P(a, g|\gamma = 0) + P(a, g|\gamma = 1)} = \frac{1}{\epsilon^a + (1-\epsilon)^g} \begin{cases} \alpha(1-\epsilon)^g & \text{if } \gamma_S = 0 \\ \beta\epsilon^a & \text{if } \gamma_S = 1 \end{cases} \tag{B.1}$$

Thus, instead of relying on a stringent threshold on the coverage to identify editing sites or completely excluding particular genomic loci such as splice junctions, we will compare our confidence in the editing hypothesis $P(\gamma_S = 1|a, g)$ to that of the no-editing hypothesis $P(\gamma_S = 0|a, g)$. A convenient way to measure the difference in these two hypotheses as a particular genomic site S is to take their log-ratio, which causes the partition function $P(a, b) = \epsilon^a + (1-\epsilon)^g$ to cancel out from top and bottom:

$$LLR(a, g) = \log \frac{P(\gamma_S = 1|a, g)}{P(\gamma_S = 0|a, g)} = \log \frac{\alpha(1-\epsilon)^g}{\beta\epsilon^a} \tag{B.2}$$

This measure depicted by the heatmap in (**Fig. B.1**) has the desirable property of extracting the maximum confidence from the coverage at a given editing site. However, LLR alone is not sufficient to accept or reject either hypothesis in the way p-values are often used and misused [**?**]. However, it is very useful in ranking different sites in order of relative confidence that editing occurs at each.

Given a ranked list of potentially edited sites, this approach still requires a cutoff in order to make actual predictions subject to validation. However, compared to the multiple thresholds for each filter in pipeline-based approaches, it is easier to manually pick or learn this parameter from training data. We tried three confidence cutoffs (0.95, 0.995, and 0.999) and chose the 0.995 based on two factors: the number of sites predicted in the ADR1- and CEN2 strands (141 and 59, respectively) was sufficiently large, but the number of sites in the ADR2- strand remained relatively low (only 6).

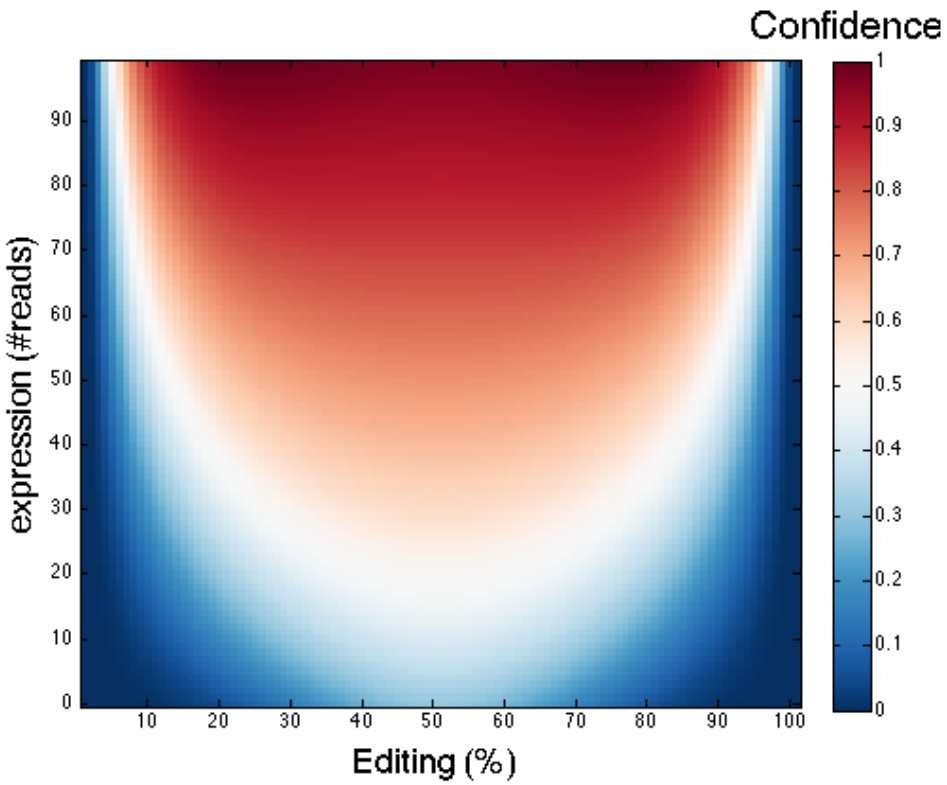


Figure B.1: A heat map of the prior distribution that captures the confidence at various any valid combination of expression (y-axis) and editing % (x-axis)

Bibliography

Afkarian, M. et al. (2002). T-bet is a STAT1-induced regulator of IL-12R expression in naive CD4+ T cells. Nat Immunol 3, 549-57.

Ahmed, R., Gray, D. (1996). Immunological memory and protective immunity: understanding their relation. Science 272, 54–60

Bahn, J.H., Lee, J.H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res 22, 142-150.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ: Deciphering the splicing code. Nature 2010.

Bass, B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu. Rev. Biochem. 71, 817-846.

Bendall, S.C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332, 687-96 (2011).

Best, J.A. et al. Transcriptional insights into the CD8(+) T cell response to infection and memory T cell formation. Nat Immunol 14, 404-12 (2013).

Bhogal, B., Jepson, J.E., Savva, Y.A., Pepper, A.S., Reenan, R.A., and Jongens, T.A. (2011). Modulation of dADAR-dependent RNA editing by the Drosophila fragile X mental retardation protein. Nat Neurosci 14, 1517-1524.

Blow, M., Futreal, P.A., Wooster, R., and Stratton, M.R. (2004). A survey of RNA editing in human brain. Genome Res. 14, 2379-2387.

Beerenwinkel, N., Drton, M. (2007). A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. Biostatistics 8, 53-71.

Buchholz, V.R. et al. (2013). Disparate individual fates compose robust CD8+ T cell immunity. Science 340, 630-5

Buganim, Y. et al. (2012). Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. Cell 150, 1209-22

Bulla, J. & B., I. Stylized facts of financial time series and hidden semi-Markov models. Computational Statistics & Data Analysis 51, 2192-2209 (2006).

Chang, J.T. et al. Asymmetric T lymphocyte division in the initiation of adaptive immune responses. Science 315, 1687-91 (2007).

Chang, J.T. et al. Asymmetric proteasome segregation as a mechanism for unequal partitioning of the transcription factor T-bet during T lymphocyte division. Immunity 34, 492-504 (2011).

Chen, L. (2013). Characterization and comparison of human nuclear and cytosolic editomes. Proc Natl Acad Sci U S A 110, E2741-2747.

Dalerba, P. et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol 29, 1120-7 (2011).

Daniel, C., Veno, M.T., Ekdahl, Y., Kjems, J., and Ohman, M. (2012). A distant cis acting intronic element induces site-selective RNA editing. Nucleic Acids Res 40, 9876-9886.

Davison A, Hinkley D: Bootstrap methods and their application. Cambridge Univ Pr 1997.

Desterro, J.M., Keegan, L.P., Jaffray, E., Hay, R.T., O'Connell, M.A., and Carmo-Fonseca, M. (2005). SUMO-1 modification alters ADAR1 editing activity. Mol Biol Cell 16, 5115-5126.

Eggington, J.M., Greene, T., and Bass, B.L. (2011). Predicting sites of ADAR editing in double-stranded RNA. Nat Commun 2, 319.

Farajollahi, S., and Maas, S. (2010). Molecular diversity through RNA editing: a balancing act. Trends Genet 26, 221-230.

Feau, S., Arens, R., Togher, S. & Schoenberger, S.P. Autocrine IL-2 is required for secondary population expansion of CD8(+) memory T cells. Nat Immunol 12, 908-13 (2011).

Freund, Y., Mason, Llew. The alternating decision tree learning algorithm. Proc. 16th International Conference on Machine Learning, 124-133 (1999).

Freund, Y.B., L, Littman, M. Linear Separation, Drifting Games and Boosting. (2009).

Garncarz, W., Tariq, A., Handl, C., Pusch, O., and Jantsch, M.F. (2013). A high-throughput screen to identify enhancers of ADAR-mediated RNA-editing. RNA Biol 10, 192-204.

Gebhardt, T. et al. Memory T cells in nonlymphoid tissue that provide enhanced local immunity during infection with herpes simplex virus. Nat Immunol 10, 524-30 (2009).

Gerlach, C. et al. One naive T cell, multiple fates in CD8+ T cell differentiation. J Exp Med 207, 1235-46 (2010).

Goodman, R.A., Macbeth, M.R., and Beal, P.A. (2012). ADAR proteins: structure and catalytic mechanism. Curr Top Microbiol Immunol 353, 1-33.

Gott, J.M., and Emeson, R.B. (2000). Functions and mechanisms of RNA editing. Annu. Rev. Genet. 34, 499-531.

Guo, G. et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell 18, 675-85 (2010).

Hansen KD, Brenner SE, Dudoit S: Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 2010, 38(12):e131–e131, [http://nar.oxfordjournals.org/content/38/12/e131.abstract].

Hundley, H.A., and Bass, B.L. (2010). ADAR editing in double-stranded UTRs and other noncoding RNA sequences. Trends Biochem Sci 35, 377-383.

Hundley, H.A., Krauchuk, A.A., and Bass, B.L. (2008). C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes. RNA 14, 2050-2060.

Ichii, H. et al. Role for Bcl-6 in the generation and maintenance of memory CD8+ T cells. Nat Immunol 3, 558-63 (2002).

Jaikaran, D.C., Collins, C.H., and MacMillan, A.M. (2002). Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site. J Biol Chem 277, 37624-37629.

Joshi, N.S. et al. Inflammation directs memory precursor and short-lived effector CD8(+) T cell fates via the graded expression of T-bet transcription factor. Immunity 27, 281-95 (2007).

Kalia, V. et al. Prolonged interleukin-2Ralpha expression on virus-specific CD8+ T cells favors terminal-effector differentiation in vivo. Immunity 32, 91-103 (2010).

Kaech, S.M., Cui, W. Transcriptional control of effector and memory CD8+ T cell differentiation. Nat Rev Immunol 12, 749-61 (2012).

Kaech, S.M., Hemby, S., Kersh, E. & Ahmed, R. Molecular and functional profiling of memory CD8 T cell differentiation. Cell 111, 837-51 (2002).

Kallies, A., Xin, A., Belz, G.T. & Nutt, S.L. Blimp-1 transcription factor is required for the differentiation of effector CD8(+) T cells and memory responses. Immunity 31, 283-95 (2009).

Katz Y, Wang ET, Airoldi EM, Burge CB: Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Meth 2010, 7(12):1009–1015, [http://dx.doi.org/10.1038/nmeth.1528].

Kleinman, C.L., and Majewski, J. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 335, 1302; author reply 1302.

Lai, F., Chen, C.X., Carter, K.C., and Nishikura, K. (1997). Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. Mol Cell Biol 17, 2413-2424.

Lee, J.H., Ang, J.K., and Xiao, X. (2013). Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. RNA 19, 725-732.

LeGendre, J.B., Campbell, Z.T., Kroll-Conner, P., Anderson, P., Kimble, J., and Wickens, M. (2013). RNA targets and specificity of Staufen, a double-stranded RNA-binding protein in Caenorhabditis elegans. J Biol Chem 288, 2532-2545.

Lehmann, K.A., and Bass, B.L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry 39, 12875-12884.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. 22, 1001-1005.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18, 1851-1858.

Lighvani, A.A. et al. T-bet is rapidly induced by interferon-gamma in lymphoid and myeloid cells. Proc Natl Acad Sci U S A 98, 15137-42 (2001).

Lin, W., Piskol, R., Tan, M.H., and Li, J.B. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 335, 1302; author reply 1302.

Lu, R., Neff, N.F., Quake, S.R. & Weissman, I.L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat Biotechnol 29, 928-33 (2011).

Maas, S., Patt, S., Schrey, M., and Rich, A. (2001). Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. Proc Natl Acad Sci U S A 98, 14687-14692.

Marcucci, R., Brindle, J., Paro, S., Casadio, A., Hempel, S., Morrice, N., Bisso, A., Keegan, L.P., Del Sal, G., and O'Connell, M.A. (2011). Pin1 and WWP2 regulate GluR2 Q/R site RNA editing by ADAR2 with opposing effects. EMBO J 30, 4211-4222.

Masopust, D. et al. Dynamic T cell migration program provides resident memory within intestinal epithelium. J Exp Med 207, 553-64 (2010).

Masopust, D., Kaech, S.M., Wherry, E.J. & Ahmed, R. The role of programming in memory T-cell development. Curr Opin Immunol 16, 217-25 (2004).

Masopust, D., Vezys, V., Marzo, A.L. & Lefrancois, L. Preferential localization of effector memory cells in nonlymphoid tissue. Science 291, 2413-7 (2001).

Moon, J.J. et al. Tracking epitope-specific T cells. Nat Protoc 4, 565-81 (2009).

Morse, D.P., Aruscavage, P.J., and Bass, B.L. (2002). RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. Proc Natl Acad Sci U S A 99, 7906-7911.

Morse, D.P., and Bass, B.L. (1999). Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. Proc. Natl. Acad. Sci. U. S. A. 96, 6048-6053.

Mortazavi A, , Wold B: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 2008, [http://dx.doi.org/10.1038/nmeth.1226].

Nicolae M, Mangul S, Mandoiu I, Zelikovsky A: Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms for Molecular Biology 2011, 6, [http://www.almob.org/content/6/1/9].

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem 79, 321-349.

Obar, J.J. & Lefrancois, L. Early signals during CD8 T cell priming regulate the generation of central memory cells. J Immunol 185, 263-72 (2010).

Ohta, H., Fujiwara, M., Ohshima, Y., and Ishihara, T. (2008). ADBP-1 regulates an ADAR RNA-editing enzyme to antagonize RNA-interference-mediated gene silencing in Caenorhabditis elegans. Genetics 180, 785-796.

Olson, J.A., McDonald-Hyman, C., Jameson, S.C. & Hamilton, S.E. Effector-like CD8(+) T cells in the memory population mediate potent protective immunity. Immunity 38, 1250-60 (2013).

Orlandi, C., Barbon, A., and Barlati, S. (2012). Activity regulation of adenosine deaminases acting on RNA (ADARs). Mol Neurobiol 45, 61-75.

Palladino, M.J., Keegan, L.P., O'Connell, M.A., and Reenan, R.A. (2000). dADAR, a Drosophila double-stranded RNA-specific adenosine deaminase is highly developmentally regulated and is itself a target for RNA editing. RNA 6, 1004-1018.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37, e123.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 2008.

Paul, M.S., and Bass, B.L. (1998). Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. EMBO J 17, 1120-1127.

Pearce, E.L. et al. Control of effector CD8+ T cell function by the transcription factor Eomesodermin. Science 302, 1041-3 (2003).

Pickrell, J.K., Gilad, Y., and Pritchard, J.K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 335, 1302; author reply 1302.

Pipkin, M.E. et al. Interleukin-2 and inflammation induce distinct transcriptional programs that promote the differentiation of effector cytolytic T cells. Immunity 32, 79-90 (2010).

Ramaswami, G., Lin, W., Piskol, R., Tan, M.H., Davis, C., and Li, J.B. (2012). Accurate identification of human Alu and non-Alu RNA editing sites. Nat Methods.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L.P., Deng, P., O'Connell, M.A., and Li, J.B. (2013). Identifying RNA editing sites using RNA sequencing data alone. Nat Methods 10, 128-132.

Ramos, A., Grunert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J 19, 997-1009.

Riedmann, E.M., Schopoff, S., Hartner, J.C., and Jantsch, M.F. (2008). Specificity of ADAR-mediated RNA editing in newly identified targets. RNA 14, 1110-1118.

Roberts A, Trapnell C, Donaghey J, Rinn J, Pachter L: Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 2011, 12(3), [http://genomebiology.com/2011/12/3/R22].

Rosenthal, J.J., and Seeburg, P.H. (2012). A-to-I RNA editing: effects on proteins key to neural excitability. Neuron 74, 432-439.

Rueter, S.M., Dawson, T.R., and Emeson, R.B. (1999). Regulation of alternative splicing by RNA editing. Nature 399, 75-80.

Rutishauser, R.L. et al. Transcriptional repressor Blimp-1 promotes CD8(+) T cell terminal differentiation and represses the acquisition of central memory T cell properties. Immunity 31, 296-308 (2009).

Ryter, J.M., and Schultz, S.C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J 17, 7505-7513.

Saltzman AL, Pan Q, Blencowe BJ: Regulation of alternative splicing by the core spliceosomal ma- chinery. Genes and Development 2011, 25:373 – 384.

Sallusto, F., Lenig, D., Forster, R., Lipp, M. & Lanzavecchia, A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature 401, 708-12 (1999).

Sansam, C.L., Wells, K.S., and Emeson, R.B. (2003). Modulation of RNA editing by functional nucleolar sequestration of ADAR2. Proc Natl Acad Sci U S A 100, 14018-14023.

Sarkar, S. et al. Functional and genomic profiling of effector CD8 T cell subsets with distinct memory fates. J Exp Med 205, 625-40 (2008).

Savva, Y.A., Rieder, L.E., and Reenan, R.A. (2012). The ADAR protein family. Genome Biol 13, 252.

Schluns, K.S., Kieper, W.C., Jameson, S.C. & Lefrancois, L. Interleukin-7 mediates the homeostasis of naive and memory CD8 T cells in vivo. Nat Immunol 1, 426-32 (2000).

Shalek, A.K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236-40 (2013).

Srivastava S, Chen L: A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Research 2010, [http://nar.oxfordjournals.org/cgi/content/abstract/gkq670v1].

Stefl, R., Oberstrass, F.C., Hood, J.L., Jourdan, M., Zimmermann, M., Skrisovska, L., Maris, C., Peng, L., Hofr, C., Emeson, R.B., et al. (2010). The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. Cell 143, 225-237.

Stemberger, C. et al. A single naive CD8+ T cell precursor can develop into diverse effector and memory subsets. Immunity 27, 985-97 (2007).

Szabo, S.J. et al. Distinct effects of T-bet in TH1 lineage commitment and IFN-gamma production in CD4 and CD8 T cells. Science 295, 338-42 (2002).

Tan, B.Z., Huang, H., Lam, R., and Soong, T.W. (2009). Dynamic regulation of RNA editing of ion channels and receptors in the mammalian nervous system. Mol Brain 2, 13.

Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377-82 (2009).

Tariq, A., Garncarz, W., Handl, C., Balik, A., Pusch, O., and Jantsch, M.F. (2013). RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation. Nucleic Acids Res 41, 2581-2593.

Tariq, A., and Jantsch, M.F. (2012). Transcript diversification in the nervous system: a to I RNA editing in CNS function and disease development. Front Neurosci 6, 99.

Tian, B., Bevilacqua, P.C., Diegelman-Parente, A., and Mathews, M.B. (2004). The double-stranded-RNA-binding motif: interference and much more. Nat Rev Mol Cell Biol 5, 1013-1023.

Tonkin, L.A., Saccomanno, L., Morse, D.P., Brodigan, T., Krause, M., and Bass, B.L. (2002). RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. EMBO J. 21, 6025-6035.

Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009, 25(9):1105–1111.

Trapnell C, Williams BA, , Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell di↵erentiation. Nat Biotech 2010.

Turro E, Su SY, Goncalves A, Coin L, Richardson S, Lewin A: Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biology 2011, 12(2), [http://genomebiology. com/2011/12/2/R13].

Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22, 1536-1537.

Valente, L., and Nishikura, K. (2007). RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions. J Biol Chem 282, 16054-16061.

van der Maaten, L.J.P., G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9, 2579-2605 (2008).

Wahlstedt, H., Daniel, C., Enstero, M., and Ohman, M. (2009). Large-scale mRNA sequencing determines global regulation of RNA editing during brain development. Genome Res 19, 978-986.

Wahlstedt, H., and Ohman, M. (2011). Site-selective versus promiscuous A-to-I editing. Wiley Interdiscip Rev RNA 2, 761-771.

Wang, I.X., So, E., Devlin, J.L., Zhao, Y., Wu, M., and Cheung, V.G. (2013). ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression. Cell Rep 5, 849-860.

Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009, [http://dx.doi.org/10.1038/nrg2484].

Warf, M.B., Shepherd, B.A., Johnson, W.E., and Bass, B.L. (2012). Effects of ADARs on small RNA processing pathways in C. elegans. Genome Res 22, 1488-1498.

Warren, L., Bryder, D., Weissman, I.L. & Quake, S.R. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc Natl Acad Sci U S A 103, 17807-12 (2006).

Wherry, E.J. et al. Lineage relationship and protective immunity of memory CD8 T cell subsets. Nat Immunol 4, 225-34 (2003).

Williams, M.A., Tyznik, A.J. & Bevan, M.J. Interleukin-2 signals during priming are required for secondary expansion of CD8+ memory T cells. Nature 441, 890-3 (2006).

Yang, C.Y. et al. The transcriptional regulators Id2 and Id3 control the formation of distinct memory CD8+ T cell subsets. Nat Immunol 12, 1221-9 (2011).

Zhou, X. et al. Differentiation and persistence of memory CD8(+) T cells depend on T cell factor 1. Immunity 33, 229-40 (2010).