Stochastic Gene Expression during Lineage Specification of Single T Helper Lymphocytes

by
Miaoqing Fang

B.S. in Life Sciences with Honors
National University of Singapore, 2006

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the
Massachusetts Institute of Technology

July 2012
[SEPTEMBER 2012]

Signature of author : _____
Department of Biological Engineering
July 31, 2012

Certified by : _____
Alexander van Oudenaarden
Professor of Physics and Biology
Thesis Supervisor

Certified by : _____
Harvey Lodish
Professor of Biological Engineering and Biology
Thesis Supervisor

Accepted by : _____
Forest White
Professor of Biological Engineering
Chairman of Graduate Program

Members of the Thesis Committee voting in favor of the defense:

Arup Chakraborty
Professor of Chemistry and Biological Engineering

Hidde Ploegh
Professor of Biology

Note on prior publications

Elements of this thesis were from the following publications:

M. Fang, H. Xie, S. Dougan, H. Ploegh and A. van Oudenaarden. Stochastic cytokine expression induces mixed T cell states. Manuscript under review.

D. Hebenstreit, **M. Fang**, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology* 7:497 (2011).

They are reused here under copyright agreements that allow reuse thematerial in a thesis or dissertation.

Stochastic Gene Expression during Lineage Specification of Single T Helper Lymphocytes

by
Miaoqing Fang

Submitted to the Department of Biological Engineering
on July 31 2012, in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy in Biological Engineering

Abstract

The adaptive immune system is an extraordinarily diverse inventory comprised of highly specialized cells, the differentiation of which requires numerous lineage specifications at various developmental stages. The precise control of immune cell differentiation and the delicate balance of their population composition are crucial for effective protection against infectious environmental agents, without triggering autoimmune responses or allergies. It is therefore important to understand at the molecular level in individual cells how lineage commitment is regulated. I explored the heterogeneous gene expression during the lineage specification of single T helper cells, by quantitatively measuring mRNA and protein levels. I have discovered a paradigm of cell lineage specification governed by the signaling interplay between extracellular cues and intracellular transcriptional factors, where the strength of extracellular signaling dominates over the intracellular signaling components. In the presence of extracellular cues, T helper cells stochastically acquire any intermediate Th1/Th2 states. The states of T helper cells can be gradually tuned by depriving availability of extracellular cytokines, which are produced stochastically by a small subpopulation of cells. When extracellular cues are removed, the weak intracellular signaling network reveals its effect, leading to classic mutual exclusion of antagonistic transcriptional factors.

Thesis supervisor: Alexander van Oudenaarden
Title: Professor of Physics and Biology

For Xuefeng

Acknowledgements

This thesis would only have been possible with many people to whom I am indebt.

First, I would like to thank Alexander van Oudenaarden, my thesis advisor and the world's leading figure in the field of systems biology, for his guidance and support, and the freedom to pursue my research interests. I have always been impressed by his scientific instincts and vision in research directions. I feel very lucky to be able to learn from him and the talented people in his lab. Although having many professional and personal commitments, Alexander always finds time to discuss my research projects, providing guidance and encouragements. I was very moved when Alexander edited my manuscript at the expense of his personal time in late evenings and early mornings. I feel proud of myself to be part of the van Oudenaarden pedigree.

The members of the van Oudenaarden lab have been very supportive. I am especially grateful to Bernardo Pando. Bernardo is an extremely intelligent and knowledgeable person who I always admire and have learned a lot from. Bernardo taught me the mathematical tools for analyzing biological systems, introduced me to the van Oudenaarden lab, and collaborated with me on the mathematical simulations of the lineage specification of T cell differentiation. I feel very lucky to be the contemporary of Arjun Raj, the inventor of smFISH by short probes, which is the underpinning method to address the scientific problems of my interest. I thank him for his generosity in teaching me smFISH and discussing my project. I am indebt to Sandy Klemm for listening to my research ideas and providing his insightful comments. I am also grateful to all the

brilliant minds who have offered me with comments and criticism on my research, especially Jeff Gore, Gregor Neuert, Shankar Mukherji, Jeroen van Zon, and Nikolai Slavov. I have also fostered precious friendships with my lab mates, especially Qiong Yang, Dong hyun Kim, Magda Bienko, Nicola Crosetto, Yannan Zheng, Lenny Teytelman, Anna van Oudenaarden, and Ni Ji, who have made our lab an enjoyable environment to work in. Ya Lin and Annalisa Pawlosky very generously gave me lots of encouragement, listened to my distress, and shared my happiness – I always enjoy talking to them and thank them for their kindness.

I am grateful to Hidde Ploegh, one of the world's renowned immunologists, for his guidance, insights and criticism. It is regretful that due to time constraints and technical difficulties, I am unable to pursue every aspect of his proposed research directions. I would also like to thank Stephanie Dougan, for providing me with mice tissues for my research work. I would also like to express my heartfelt gratitude to my collaborators, Sarah A. Teichmann and Daniel Hebenstreit, at MRC Laboratory of Molecular Biology in UK, for very inspiring and productive collaborations.

I sincerely thank Harvey Lodish for being a very influential mentor, both in my professional and personal life. I deeply admire Harvey for his unparalleled achievements in every aspect of life, from academia to industry, from judicature to policy-making, from teaching to parenting. I am also grateful towards Arup Chakraborty and Chris Burge for serving on my thesis committee and providing me with invaluable comments.

During my years at MIT, I have enjoyed invaluable friendships with my classmates at Biological Engineering, especially Lily Jeng, Joy Rimchala, Adriene Li, Andrew Khoo, and Robbie Barbero. I thank them for doing problem sets in the "dungeon" together, sharing the fun and hardships in graduate school.

I sincerely thank Berge Englert for being an inspiring mentor and a caring friend. I deeply admire Berge for his mathematical virtuoso, wide scope of knowledge, and generosity in treating people.

My heartfelt gratitude goes to my husband Huangming Xie, who inspired me to pursue a PhD degree when I was an undergraduate. Throughout my years at MIT, He discussed research ideas and provided technical help on my research projects. More importantly, I thank him for his love towards me and our son Lekang.

Finally, I thank my mum Xuefeng Fang, for whom I have existed. I thank her for her love, support, and disciplinary actions that have made me a better person. I will always love her, to whom this thesis is dedicated.

CONTENT

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 1

Introduction

Mammals consist of many distinct types of highly specialized cells, the differentiation of which requires numerous lineage specifications at various developmental stages. In the developmental paradigm, a progenitor cell is capable of differentiating into several lineages. The precise control of progenitor cell differentiation is crucial for achieving a delicate balance in composition of the differentiated cell populations. Commitment to a specific cell fate hinges on the regulation of a single or a handful of master regulators, which are often transcription factors. Given the appropriate signals, which can be extracellular cues such as cytokines, these master regulators orchestrate the expression of a set of effector genes and repression of the genes associated with alternative cell fates.

1.1 Stochastic Gene Expression in Eukaryotic cells

However, gene expression is a fundamentally stochastic process, because noise in transcription and translation can lead to cell-to-cell variations in mRNA and protein levels even in genetically identical cells (Raj and van Oudenaarden, 2008). Studies on gene expression in eukaryotes indicate that gene expression is noisy (Fig 1.1), because transcription occurs in bursts. This can be attributed to that the gene transitions between an inactive and active state (Becskei et al., 2005; Raj et al., 2006; Raser and O'Shea, 2004; Warren et al., 2006), or other possible mechanisms such as the formation of pre-initiation complexes at the promoter region of the DNA and multiple transcription events facilitated by RNA polymerase (Blake et al., 2006; Blake et al., 2003; Raj and van Oudenaarden, 2008). Given the noisy nature of gene expression and the important goal of achieving a precise composition of various lineages of differentiated cells, it is interesting to examine at the molecular level in individual progenitor cells the expression levels of the master regulators.

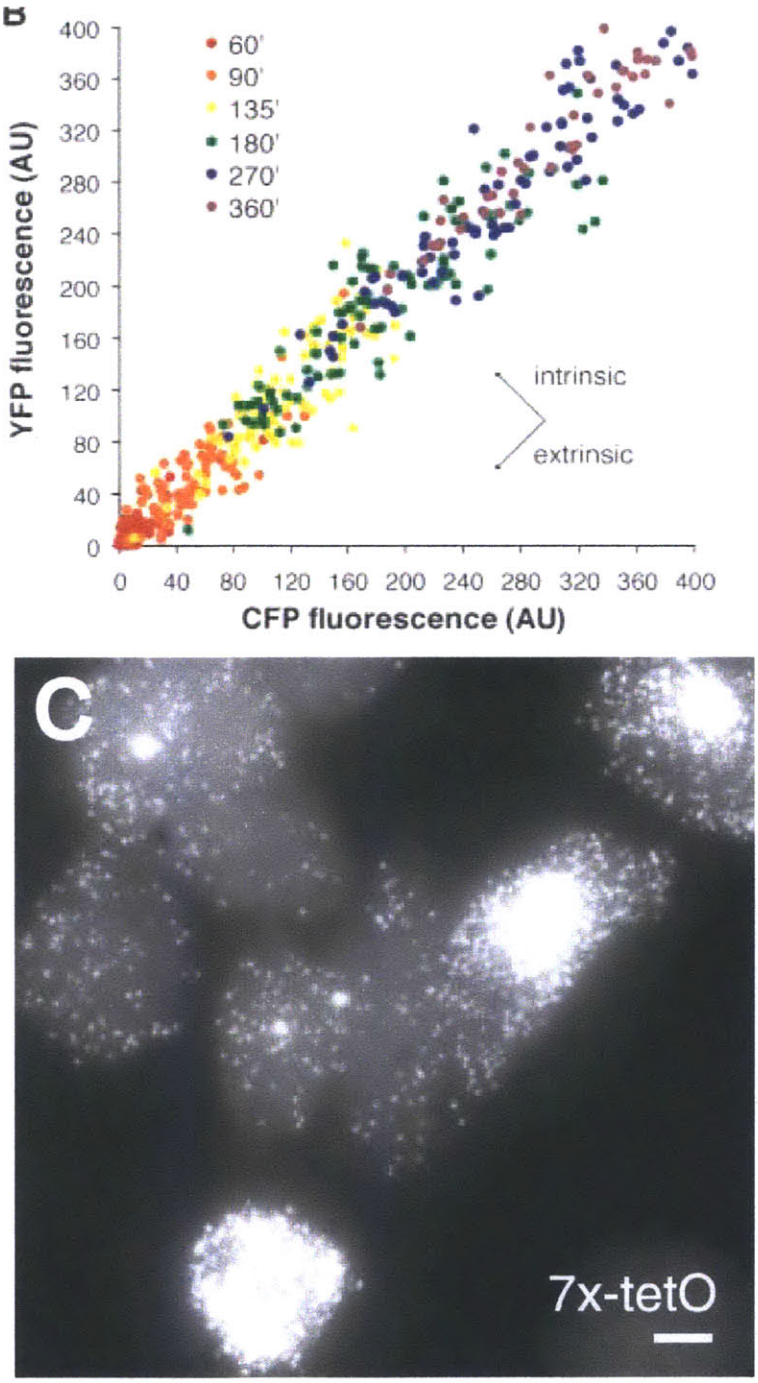


Fig. 1. Stochastic gene expression in eukaryotic cells. The upper panel shows the scatter plot of YFP and CFP, driven by the same promoters on different chromosomes in individual yeast cells, grown under the same condition (Raser and O'Shea, 2004). The

lower panel shows the heterogeneous gene expression in mammalian cell subjected to the same culture environment (Raj and van Oudenaarden, 2008).

To study transcript levels quantitatively in individual cells has been a challenging problem until the past two decades with the invention of novel detection tools that detect single mRNA molecules, such as MS2-GFP method (Beach et al., 1999; Bertrand et al., 1998; Golding et al., 2005), single molecule FISH (smFISH) (Femino et al., 1998; Raj et al., 2006; Raj et al., 2008), single-cell RT-PCR (Bengtsson et al., 2005; Warren et al., 2006), and molecular beacons (Tyagi and Kramer, 1996; Vargas et al., 2005). In this thesis, we deployed a novel smFISH technique for imaging individual mRNA molecules in fixed cells. This method probes each mRNA species with 20 or more short, singly labeled oligonucleotide probes that are about 20-mers in lengths (Fig. 1.2). Simultaneous binding of the probe set to each mRNA molecule results in a diffraction-limited fluorescent spot by fluorescence microscopy, which can be computationally identified using a log filter. By labeling each probe set with a different fluorophore with non-overlapping absorption and emission spectra, we can simultaneous detect multiple mRNA species in single fixed cells. Since this method offers single-molecule resolution, it is more sensitive than conventional quantitative RT-PCR, which relies on exponential signal amplification and thus performs poorly at resolving differences of less than two folds. In addition, single-molecule mRNA FISH is compatible with quantitative immunofluorescence, enabling concurrent quantification of mRNA and protein levels in individual cells. This will enable us to question how many transcripts of the genes of interests are expressed in individual cells and what the correlation between each mRNA and protein species is in individual cells. We can then examine the heterogeneity of mRNA and protein levels in progenitor cells at various time points during their differentiation.

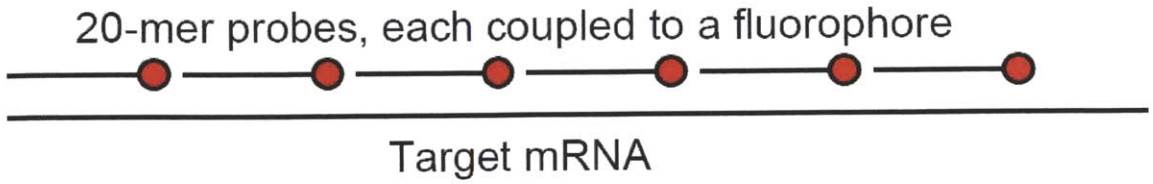


Fig1.2. mRNA FISH with single molecule resolution. This method probes each mRNA species with 20 or more short, singly labeled oligonucleotide probes that are about 20-mers in lengths. Simultaneous binding of a probe set, which typically consists of at least 20 different oligonucleotide probes, to each mRNA molecule results in a diffraction-limited fluorescent spot under fluorescence microscope.

1.2 The Road to finding a suitable model system of cell differentiation

To select a model cell differentiation system, I have tested a few systems. First I started with mesenchymal stem cells, which can differentiate into a variety of cell types, including osteoblasts, chondrocytes and adipocytes (Rosen and Spiegelman, 2000). My plan was to track the expression of master transcription factors for each lineage. I first tested the feasibility of this model system by inducing the mesenchymal stem cells towards the adipose lineage, by adding exogenously added cues such as dexamethasone. The mesenchymal stem cells accumulated fat droplets and acquired the phenotypic features of adipocytes. However, I then realized these cells are not amenable to microscopic imaging. First, extremely high cell confluence was required to differentiate mesenchymal stem cells to adipocytes, resulting cells stacking on top of each other (Fig. 1.3). Secondly, the fat droplets have sharp circular boundaries on the microscopic images, making cell segmentation algorithm confused with real cell boundaries. Thirdly, the fat droplets are fluorescent over a large of spectra under the fluorescent microscopic imaging, resulting in high background noise that mask the real fluorescent signals from single molecule FISH.

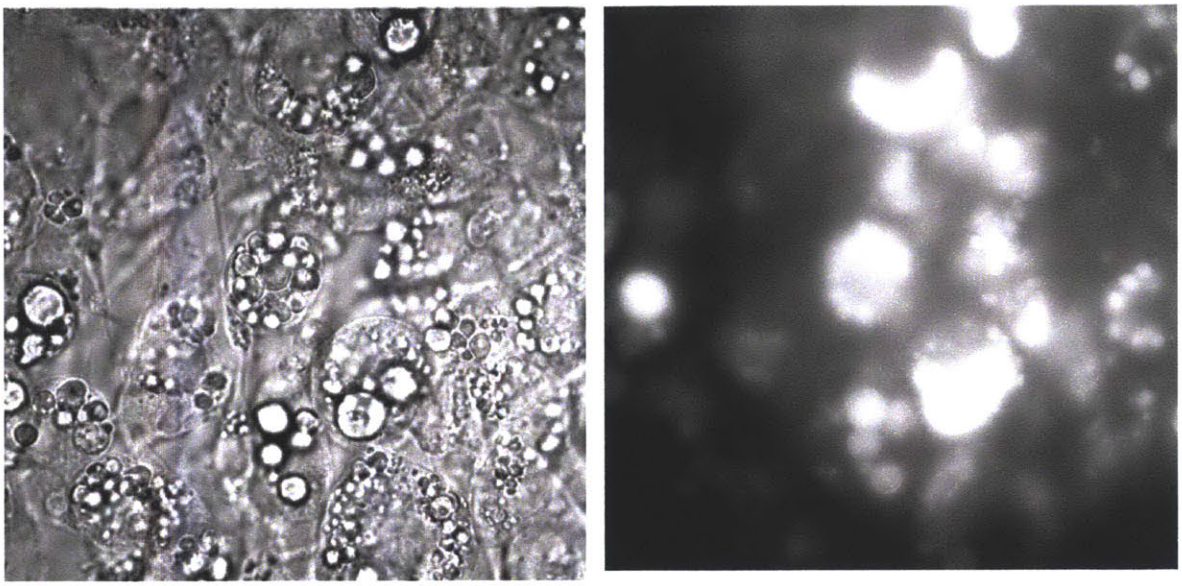


Fig. 1.3. Differentiation of mesenchymal stem cells towards the adipose lineage. The left panel is the bright-field image of the cells, showing accumulation of fat droplets. The right panel is a fluorescent image, showing that fat droplets have strong fluorescence, rendering single-molecule mRNA FISH infeasible in these cells.

1.3 Differentiation of CD4 T helper cells

I continued to examine several types of progenitor cells and nailed down to the naive $CD4^+$ T helper cells, because of its important role in adaptive immunity and technical feasibility to culture and image these cells. The naive $CD4^+$ T helper cells are capable of differentiating into Th1, Th2, Th17, induced regulatory T cells (iTreg) and follicular T cells (fTh). The classical dichotomy of the Th1 versus Th2 is well-established. Th1 lineage, characterized by secretion of hallmark cytokine interferon-γ (IFNγ), is essential for eradicating intracellular pathogens, primarily by activating natural killer (NK) cells and cytotoxic $CD8^+$ T cells that can kill pathogen infected cells and secreting cytokines such as IFNγ to hinder further pathogen entry into cells (Szabo et al., 2000). In contrast, Th2 lineage, characterized by secretion of IL-4, is essential for eliminating extracellular pathogens, primarily by activating B cells to secret antibodies that sequester pathogens or neutralize toxins.

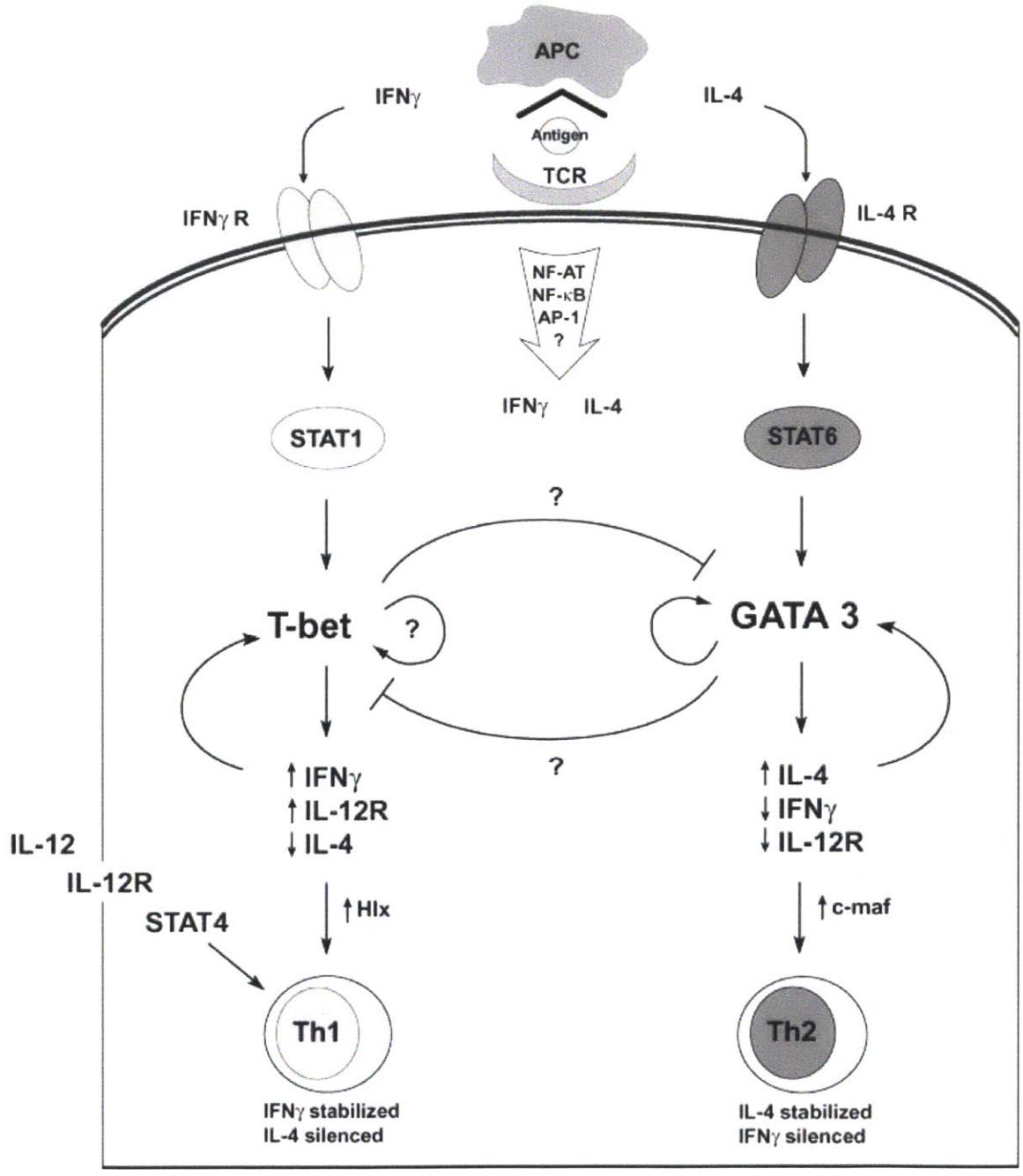


Fig 1.2. Dogmatic view on signaling network during Th1/Th2 differentiation (Szabo, 2003).

Tbet, encoded by *Tbx21*, is the master transcription factor of Th1 differentiation(Szabo et al., 2000), whereas Gata3 is the master transcription factor of Th2 differentiation (Zhang et al., 1997; Zheng and Flavell, 1997) (Fig. 1A). *Tbx21* and *Gata3* expressions are postulated to be mutually exclusive in individual cells (Lohning et al., 2002; Mariani et al., 2004; Murphy and Reiner, 2002; Zhou et al., 2009), owing to positive feedback loops and cross inhibitions. These regulatory networks consist of two types: one that depends on cytokine signaling and the other that is independent of extracellular cytokines and involves only the intracellular players such as transcription factors. Specifically, Tbet activates *Ifng* (Djuretic et al., 2007), and binding of extracellular IFNγ to its receptor triggers STAT1 signaling and induces expression of *Tbx21* (Leonard and O'Shea, 1998). In addition, Tbet induces its own expression in an IFNγR/STAT1 independent manner, possibly through autoinduction and interaction with the transcription factor Hlx (Mullen et al., 2002). Similarly, Gata3 activates *Il4* (Jenner et al., 2009; Tykocinski et al., 2005), and binding of extracellular IL4 to its receptor triggers STAT6 signaling and induces the expression of *Gata3* (Kaplan et al., 1996; Shimoda et al., 1996; Takeda et al., 1996). In addition, Gata3 binds the *Stat6* promoter, leading to a positive feedback independent of extracellular IL4 (Jenner et al., 2009). Furthermore, *Gata3* can also be autoinduced in an IL4R/STAT6 independent manner, possibly by binding its own promoter or enhancer, or mediated by intermediate factors such as c-maf (Ouyang et al., 2000). For cross inhibition, Tbet silences *Il4* (Djuretic et al., 2007), and Gata3 silences *Ifng* (Chang and Aune, 2007; Schoenborn et al., 2007). In addition, Tbet blocks the functions of Gata3 through direct protein-protein interactions between the two transcription factors (Hwang et al., 2005). It has been proposed that small random fluctuations in gene expression can set Tbet or Gata3 level above a threshold required for maintaining subsequent high expression of one transcription factor while silencing the other (Callard, 2007; Chang and Aune, 2007; Schoenborn et al., 2007; Szabo et al., 2003; Yates et al., 2004). However, this notion is largely supported by conjectures based on the current understanding of Th signaling networks and mathematical simulations.

1.4 T cell antigen receptor and its associated kinases

The T cell antigen receptors (TCR) are responsible for recognizing specific antigens presented by major histocompatibility complex (MHC) molecules, forming the basis for the specificity of T cell immunity. Specifically, TCR on CD4 T cells recognizes antigens presented by MHC class II molecules. Being a heterodimer, in 95% of T cells, TCR consist of α/β chains, whereas the remaining 5% consist of γ/δ chains. The CD4 T cells under study in this thesis bear α/β TCRs. TCR by itself is not a signal transducer. Instead, it is associated with the CD3 (cluster of difference 3) protein complex, which contains an immunoreceptor tyrosine-based activation motif (ITAM) useful for signaling. In mammals, CD3 consists of four peptide chains: one CD3γ chain, one CD3δ chain, and two CD3ε chains. Taken together, the TCR-CD3 complex is a hexameric complex.

The TCR signaling pathway consists of proximal signaling, including phosphorylation of the invariant signaling protein CD3 and early signaling molecules such as kinases, calcium-mediated signaling, which leads to release of intracellular Ca^{2+} stores and influx of extracellular Ca^{2+}, and GTP Ras-mitogen-activated protein kinase (MAPK) signaling (Fig. 1.3) (Morris and Allen, 2012; Smith-Garvin et al., 2009). Activation of CD3 is dependent on the affinity between TCR and peptide-MHC complex (pMHC). High affinity TCR-pMHC interactions may be sufficient for signaling, whereas TCR-pMHC interactions with lower affinities depend on coreceptors for signaling. TCR complex in CD4 T cells is associated with CD4 (in cytotoxic T cells, it is associated with CD8 coreceptor), which recruits kinase Lck to activate CD3.

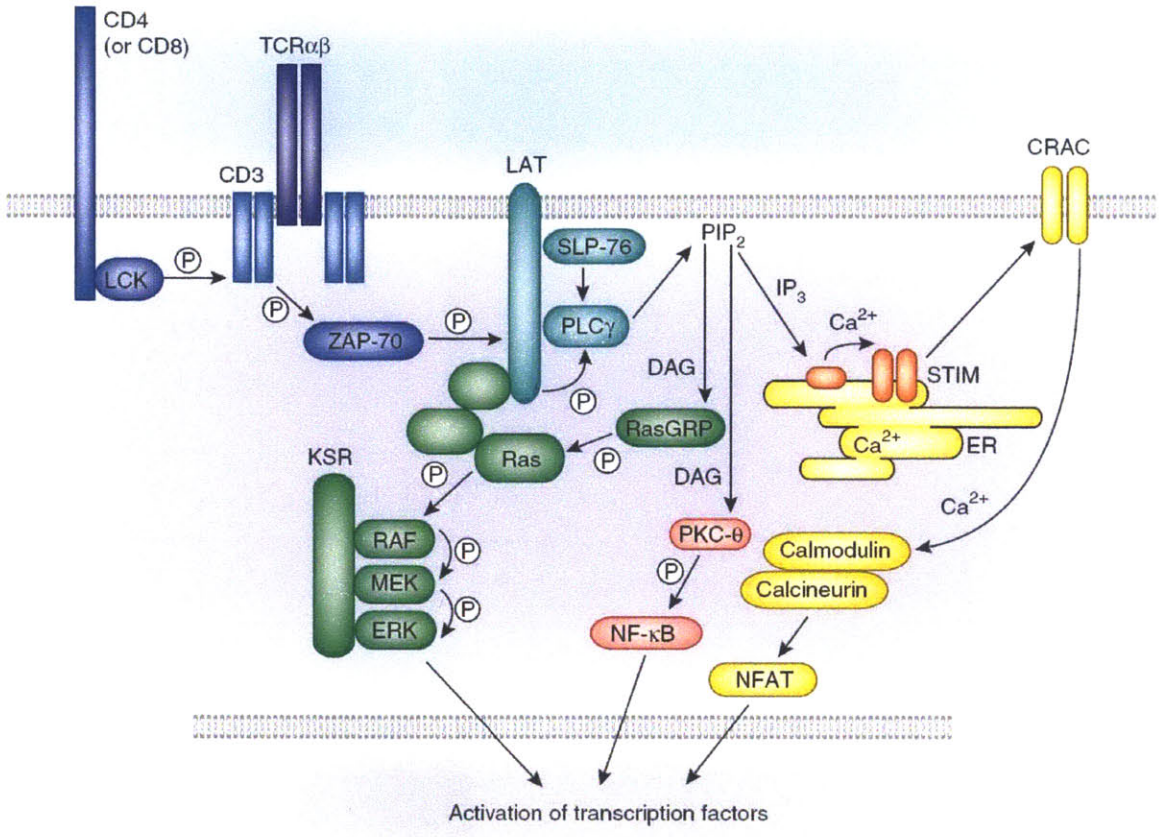


Fig. 1.3. TCR signaling pathways. When TCR recognizes ligands presented by MHC molecule, its associated CD3 triggers a signaling cascade that involves the phosphorylation of proximal TCR components (blue), signaling by the Ras-Erk pathway (green), activation of the transcription factor NF-κB (pink) by PKC-θ, and Ca^{2+} flux – mediated signaling (yellow). These pathways activate transcription factors that mediate a variety of T cell developmental and effector programs (Morris and Allen, 2012). In naive CD4 T cells, these pathways leads to expression of *Tbx21* and *Gata3*, as shown in the later chapters of this thesis.

1.5 Advantages and caveats of studying CD4 T cell differentiation in cell culture

Because our goal is to study stochastic gene expression during CD4 T cell differentiation, we have to ensure that each T cell receives signals of identical strength at the stage of CD3 signaling, with no upstream variations. We decided to culture CD4 T cells on cell culture dishes coated with anti-CD3 antibodies, which leads to clustering of CD3 molecules and thus signaling. This method normalizes a large number of external factors. First, the culture media is uniform, ensuring that each cell is exposed to the identical extracellular cues with no biases in the cytokine milieu, as shown in the data presented in the following chapters. Secondly, since the culture well is uniformly coated with anti-CD3 antibodies, the signaling strength in each cell does not have a spatial dependence. Thirdly, signaling through CD3 directly bypass the need for TCR-pMHC interaction, avoiding variable signaling strengths as an outcome of the diverse TCR repertoire with variable affinities to a specific antigen.

Under physiological conditions, signaling by CD3 is elicited from TCR-pMHC interaction in an affinity-dependent manner. However, in the cell culture used in this study, CD3 signaling is elicited from clustering of CD3 by anti-CD3 antibodies coated on the surface of the cell culture dish. As a result, the downstream signaling strength in cell cultures may be significantly different from that under physiological conditions, failing to capture the TCR-pMHC-affinity-dependent feature of CD3 activation *in vivo*. As a result, differentiation of CD4 T cells *in vivo* is expected to be a more variable process among individual cells than our results on CD4 T cell cultures.

A method that can potentially address the artificiality of anti-CD3 antibody mediated cell culture is to co-culture CD4 T cells with antigen presenting cells (APC). However, this method can result in heterogeneity in CD3 signaling strengths, because activation of CD4 T cells depends on cell-cell contact with APCs, which are not equally available to every T cell in the culture.

1.6 Overview of the thesis

In this thesis, we quantified both mRNA transcript and protein levels in single $CD4^+$ T helper cells upon activation and explored their heterogeneous cell fate decisions. In Chapter 2, we quantified the number of transcripts of five different genes in differentiated Th2 cells. We found that all genes had Fano factors (σ^2/μ) larger than 4, indicating that they had super-Poisson variation (a Poisson random variable would have $\sigma^2/\mu = 1$) and therefore burst-like transcription (Raj et al., 2006). In Chapter 3, we quantified mRNA and protein levels during the early differentiation of naive CD4 T helper (Th) cells into Th1 versus Th2 states. Surprisingly, we observed ubiquitous high-level co-expression of antagonistic transcription factors in individual cells. The expression of these transcription factors can be gradually tuned by extracellular cytokines, which are produced stochastically by a small subpopulation of cells. Upon inhibition of cytokine signaling, we observed the classic mutual exclusion of antagonistic transcription factors, thus revealing a weak intracellular network otherwise overruled by the strong signals that emanate from extracellular cytokines. Chapter 4 concludes our discoveries on stochastic gene expression during lineage specification of single T helper cells, and provides perspectives on future research directions.

CHAPTER 2

Stochastic Gene Expression in Differentiated Single Th2 Cells

This work was published in Molecular Systems Biology 7:497 (2011), in collaboration with Teichmann group at the MRC Laboratory of Molecular Biology in Cambridge, the United Kingdom. The paper was titled "RNA sequencing reveals two major classes of gene expression levels in metazoan cells", authored by Daniel Hebenstreit, Miaoqing Fang, Muxin Gu, Varodom Charoensawan, Alexander van Oudenaarden and Sarah A Teichmann.

My contribution to this work is to conceive and perform the smFISH experiment, perform image analysis, and write the manuscript.

2.1 Abstract

The expression level of a gene is often used as a proxy for determining whether the protein or RNA product is functional in a cell or tissue. Therefore, it is of fundamental importance to understand the global distribution of gene expression levels, and to be able to interpret it mechanistically and functionally. Here we use RNA sequencing of mouse Th2 cells, coupled with a range of other techniques, to show that all genes can be separated, based on their expression abundance, into two distinct groups: one group comprising of lowly expressed and putatively non-functional mRNAs, and the other of highly expressed mRNAs with active chromatin marks at their promoters. These observations are confirmed in many other microarray and RNA-sequencing datasets of metazoan cell types.

Key words: expression levels/RNA-seq/ChIP-seq/RNA-FISH/bimodal

2.2 Introduction

Expression level is frequently used as a way of characterizing gene function, by Northern blotting, PCR, microarrays, and, more recently, RNA-sequencing (Wang et al., 2009a) (RNA-seq). Therefore, it is a central issue in molecular biology to know how many transcripts are expressed in a cell at what levels. This question was studied very early in the history of molecular biology using methods such as reassociation kinetics (Hastie and Bishop, 1976), which indicated the existence of distinct abundance classes, and recently, we pointed out that separate peaks are visible in the abundance distributions of a number of microarray data sets (Hebenstreit et al., 2011). At the same time, microarrays or RNA-seq data have been described as displaying broad, roughly lognormal distributions of expression levels with no clear separation into discrete classes (Hoyle et al., 2002; Lu and King, 2009; Ramskold et al., 2009). There are several reasons for this: many samples are heterogeneous in terms of cell type (Hebenstreit and Teichmann, 2011) or are based on a previous generation of less sensitive microarrays, many are from unicellular organisms rather than animals, and finally, data processing and plotting methods can obscure the presence of distinct abundance classes. Here, we provide experimental and computational support for two overlapping major mRNA abundance classes. Our findings hold for metazoan datasets including human, mouse and *Drosophila* sources.

2.3 Results and Discussion

We initially based our analysis on murine Th2 cells (Zhu et al., 2010) as these cells can be obtained in large quantities *ex vivo* and can be prepared as a pure and homogeneous cell population. Furthermore, there is a well characterized set of genes whose proteins are known to be expressed and functional in Th2 cells, as well as a set of genes known to be not expressed in these cells (Table 2.S1 lists the genes we used in our study, Figure 2.S1 shows expression of two marker proteins in the cells).

We generated Th2 poly(A)+ RNA-seq data for two biological replicates and calculated gene expression levels using the standard measure of Reads Per Kilobase per Million (RPKM) (Mortazavi et al., 2008) (Table 2.S2 gives the number of reads and mappings we obtained). The expression levels of the biological replicates are highly correlated (r^2 = 0.94, Figure 2.S2). We then calculated the mean RPKMs of the two samples for all genes and $\log_2$ transformed these values.

Displaying the distribution of all gene expression levels as a kernel density estimate (KDE) reveals an interesting structure: the majority of genes follow a normal distribution which is centered at a value of ~4 $\log_2$ RPKM (~16 RPKM), while the remaining genes form a shoulder to the left of this main distribution (Figure 2.1A, solid line). This was conserved under different KDE bandwidths (Figure 2.S3, left panel) or different histogram representations (Figure 2.S3, right panel). As genes with zero reads cannot be included on the log scale, we prepared an alternative version of Figure 2.1A where we assigned low RPKM values to these. This helps to illustrate the fraction of zero read genes (Figure 2.S4). As a comparison, we studied microarray data for the same cell type from a recent publication (Wei et al., 2009). The correlation between the microarray and the RNA-seq data was very good and highly statistically significant (Pearson r^2 = 0.83, Spearman ρ = 0.84, Figure 2.S5). Surprisingly, displaying the distribution of microarray expression levels results in a clearly bimodal distribution (Figure 2.1B). Again, the appearance of the distribution was insensitive to the KDE bandwidth choice or histogram bin size (Figure 2.S6). The bimodality was conserved when alternative normalization and processing schemes were used, independent of KDE bandwidths (Figure 2.S7).

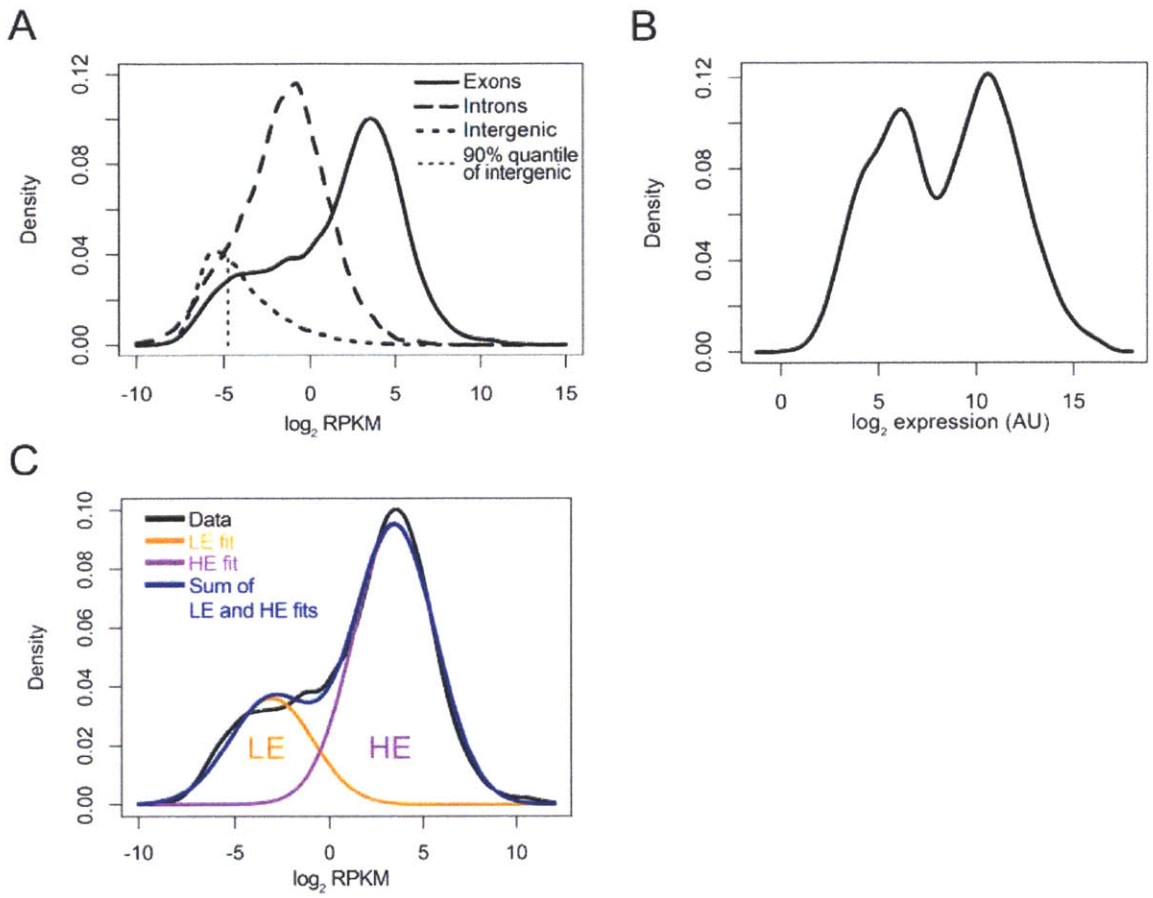


Figure 2.1. Distribution of gene expression levels. (A) Kernel density estimates of RPKM distributions of RNA-seq data within exons, introns and intergenic regions as indicated. The fragments used to estimate intron and intergenic RPKM were based on randomizations using the same length distribution as the exonic parts of genes. The 90% quantile of the intergenic distribution is indicated. (B) Kernel density estimate of expression level distribution of microarray data (Wei et al., 2009). (C) Expectation maximization based curve fitting of RNA-seq data of (A).

Visual inspection of both microarray and RNA-seq data thus reveals two overlapping main components of the distribution of gene expression levels. Quantifying this by curve fitting confirms a good fit to two distributions: the goodness-of-fit (measured by Akaike Information criterion, AIC (Akaike, 1974), Bayesian Information Criterion, BIC (Schwarz, 1978) or Likelihood ratio tests (Casella and Berger, 2001)) shows strong increases for both microarray and RNA-seq data when two-component models are fit by expectation-maximization (compared to single- or more-component models) (Figure 2.S8). We designate the two groups of genes as the lowly expressed (LE) and highly expressed (HE) genes (Figure 2.1C), because we will present evidence below that the LE genes are expressed rather than simply being experimental background. Our findings are not limited to Th2 cells and hold for virtually all recently published metazoan RNA-seq datasets (e.g. (Marioni et al., 2008; Mortazavi et al., 2008; Mudge et al., 2008; Wang et al., 2008), Figure 2.S9 and (Cloonan et al., 2008), Figure 2.S10A, B) and all microarray data sets (e.g. (Cui et al., 2009), GNF Atlas 3 (Lattin et al., 2008), (Chintapalli et al., 2007), Figure 2.S11) we have studied. The existence of further, minor groups of genes cannot be excluded, but is not clear at this point due to the diverse curve-fitting results for the different datasets if higher-order (more than two components) models are considered.

The difference between the microarray and RNA-seq distributions is explained by the fact that the microarrays yield a signal for all genes, part of which is due to cross-hybridization of oligo-nucleotide probes if the gene is not strongly expressed. RNA-seq on the other hand yields a signal for a gene only if at least one sequencing read is found. The accuracy of RNA-seq is biased towards longer and more highly expressed genes, e.g. 5 % of all genes account for 50 % of all reads in our data as well as in other datasets (Bullard et al., 2010; Mortazavi et al., 2008; Oshlack and Wakefield, 2009).

To explore how this accuracy bias affects the shape of the LE distribution, we studied the RNA-seq detection limit. We first plotted the number of genes with zero reads as a function of the total number of reads (taking subsets of the total reads). The number of genes without reads decreases slowly, with no change in slope and hence no indication of reaching a plateau. Even at a total of 25 million reads, ~30% of all genes are undetected (Figure 2.2A). We further estimated the numbers of genes remaining

undetected at each expression level by assuming Poisson-distributed read numbers (Jiang and Wong, 2009) and determining the expected frequency of zeroes. This confirms the sensitivity drop at the lower end of the LE peak (Figure 2.2B). Extrapolating the numbers of expressed genes including the undetected ones reveals an emerging LE peak (Figure 2.2B). Thus the smaller portion of LE genes in the RNA-seq data compared to the microarray data is at least partially due to the RNA-seq detection limit, although this only becomes a problem for genes at less than ~ -3 to -4 $\log_2$ RPKM. It should be noted that these low expression levels correspond to an absence of transcripts in the majority of cells, as we demonstrate further below.

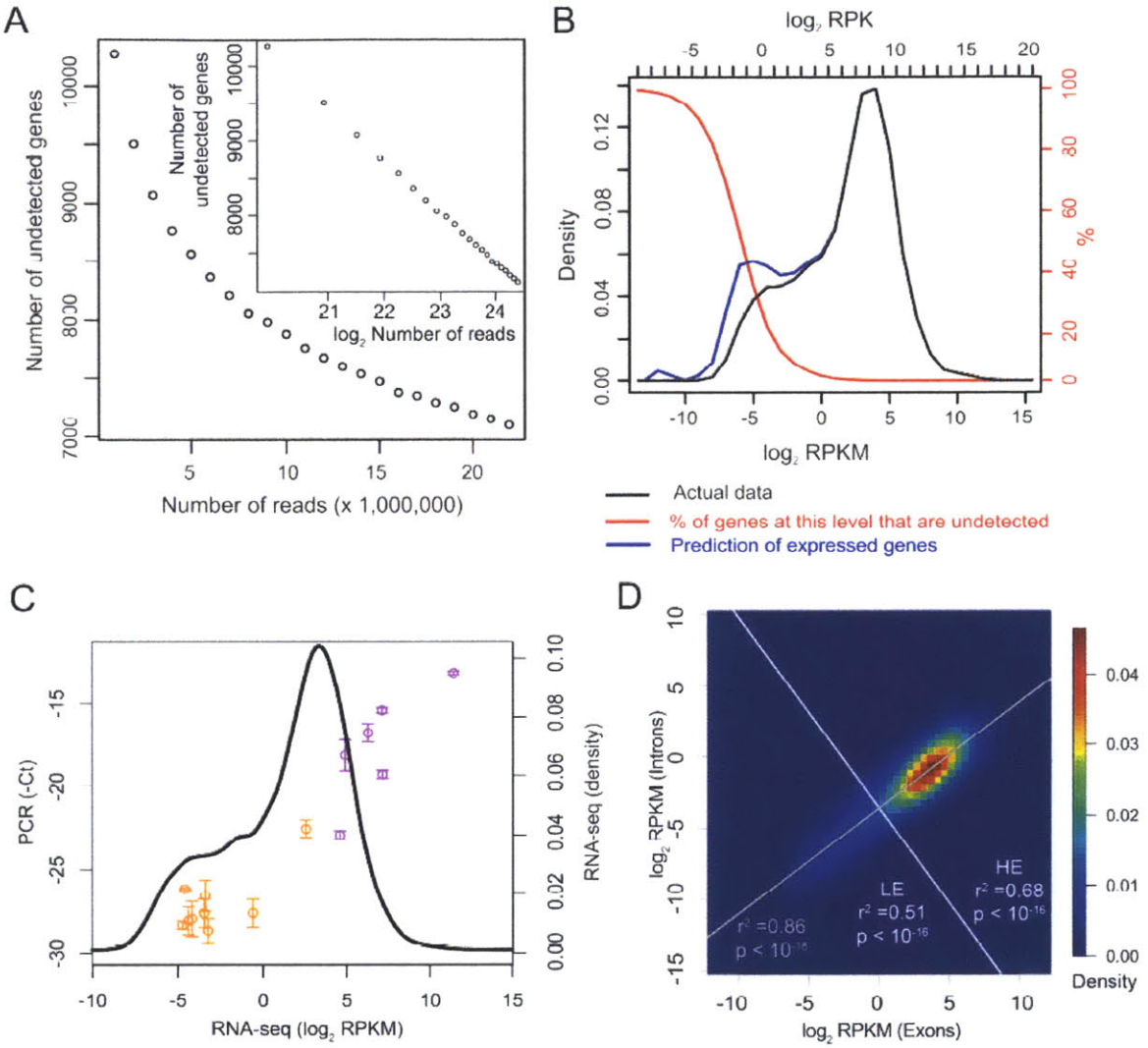


Figure 2.2. Sensitivity of RNA-seq. (A) Detection of genes in dependency of the total read numbers on linear scale and $\log_2$ scale (inset). Random subsets of the total reads for the two RNA-seq replicates were taken and the number of genes with zero reads were plotted *versus* the total read numbers used. The Figure 2.represents an average of five independent subsets for each data point. (B) Prediction of genes remaining undetected due to Poisson statistics underlying RNA-seq. The theoretically expected fraction of genes remaining undetected (red, y-axis on the right side of the Figure in red) was determined for each expression level and was used to infer from the binned (small ticks on top indicate the bins) actual expression data (black) the expressed genes including the undetected ones (blue). In addition to the RPKM scale, the reads per kilobase (RPK)

scale (without normalization to the total number of mapped reads) is shown (on top), which was used for the calculation of the (integer-) Poisson statistic and which, in contrast to the RPKM scale, depends on the total number of sequencing reads. (C) RT-PCR for the genes listed in Table 2.S1. The RNA-seq expression levels of the genes are plotted *versus* the negative threshold cycles (C_t) of the PCRs. The plot is overlaid (with the same x-axis scaling) upon the kernel density estimate of the RNA-seq expression level distribution (black line) to show the positions of the genes in the total expression distribution. Genes either in the LE peak of the RNA-seq distribution or which have been previously characterized as not expressed in Th2 cells are shown in orange. Genes known to be expressed are shown in purple. Error bars indicate standard error of mean from three independent biological replicates. Please refer to Tables S1 and 2.S6 for details of genes and PCR primers. (D) Correlation of RPKM within exons and introns from RNA-seq data of Figure 2.1A. Correlation and significance of correlation were calculated for the whole distribution (gray) or for LE and HE genes separately. Division into LE and HE was performed along a line (white) perpendicular to a fitted trendline (gray), centered at Exon RPKM = 1. The data points are shown as 2D kernel density estimate.

To further confirm that the LE genes correspond to low expression and not experimental noise, we performed realtime PCRs. We tested amplification by exon spanning primers of a set of genes that are known to be expressed or not expressed in Th2 cells, plus five random genes that we detected between -3.7 and -5 $\log_2$ RPKM in the RNA-seq experiment (Table 2.S1). We were able to successfully PCR-amplify all genes with high specificity. The expressed genes map to the HE peak, while almost all unexpressed genes map to the LE peak, if we align the PCR results with the microarray/RNA-seq data (Figure 2.2C).

We also tested the extent to which genomic DNA can be detected in our polyA-purified mRNA sample, as proposed by Ramskold et al (Ramskold et al., 2009) as a means of quantifying experimental background. We randomly selected intergenic fragments with the same length distribution as genes, 10 kb away from genes. The resulting RPKM distribution contains a high number of zero-RPKM fragments (79 %) while the majority of non-zero fragments peaks slightly left of the LE shoulder (Figure 2.1A). The 90 % quantile of this intergenic background distribution is at -4.97 $\log_2$ RPKM, which means that we can be quite confident (with probability > 90 %) that genes with an RPKM value above this level are truly expressed rather than representing experimental background noise (Figure 2.1A). Further, the overlap between the intergenic and the normalized LE fit is small (Figure. 2.S12). We cannot rule out that detection of intergenic DNA corresponds to transcription as well, which would make the case for transcription of LE genes even stronger.

Analysis of the strand-specific mRNA-sequencing data of ES cells of Cloonan et al (Cloonan et al., 2008) yields similar conclusions. The poly(A)-purification protocol selects for reads antisense to genes (the antisense reads correspond to mRNA). In the distribution of 'sense' reads (corresponding to antisense transcripts in genic regions), more than 50 % of genic regions have zero reads. This distribution is unimodal and shifted by ~ 2 $\log_2$ RPKM with respect to the LE distribution, and overlaps almost perfectly with the distribution of reads in intergenic regions (Figure 2.S10A).

We next determined the distribution of RPKM within introns, again using fragments with the same length distribution as transcripts. (Please note that our intronic read densities are not enriched at 5' or 3' ends of the intronic regions (Figure 2.S13).)

The resulting intronic distribution is significantly higher than the intergenic background (two-sided Wilcoxon rank sum test, $p < 2.2 \times 10^{-16}$) and peaks at roughly -1 $\log_2$ RPKM (Figure 2.1A). Introns thus have one- to two orders of magnitude lower read density than exons. This suggests that we are detecting incompletely processed transcripts at a low but significant and uniform level across all the whole range of transcript abundances.

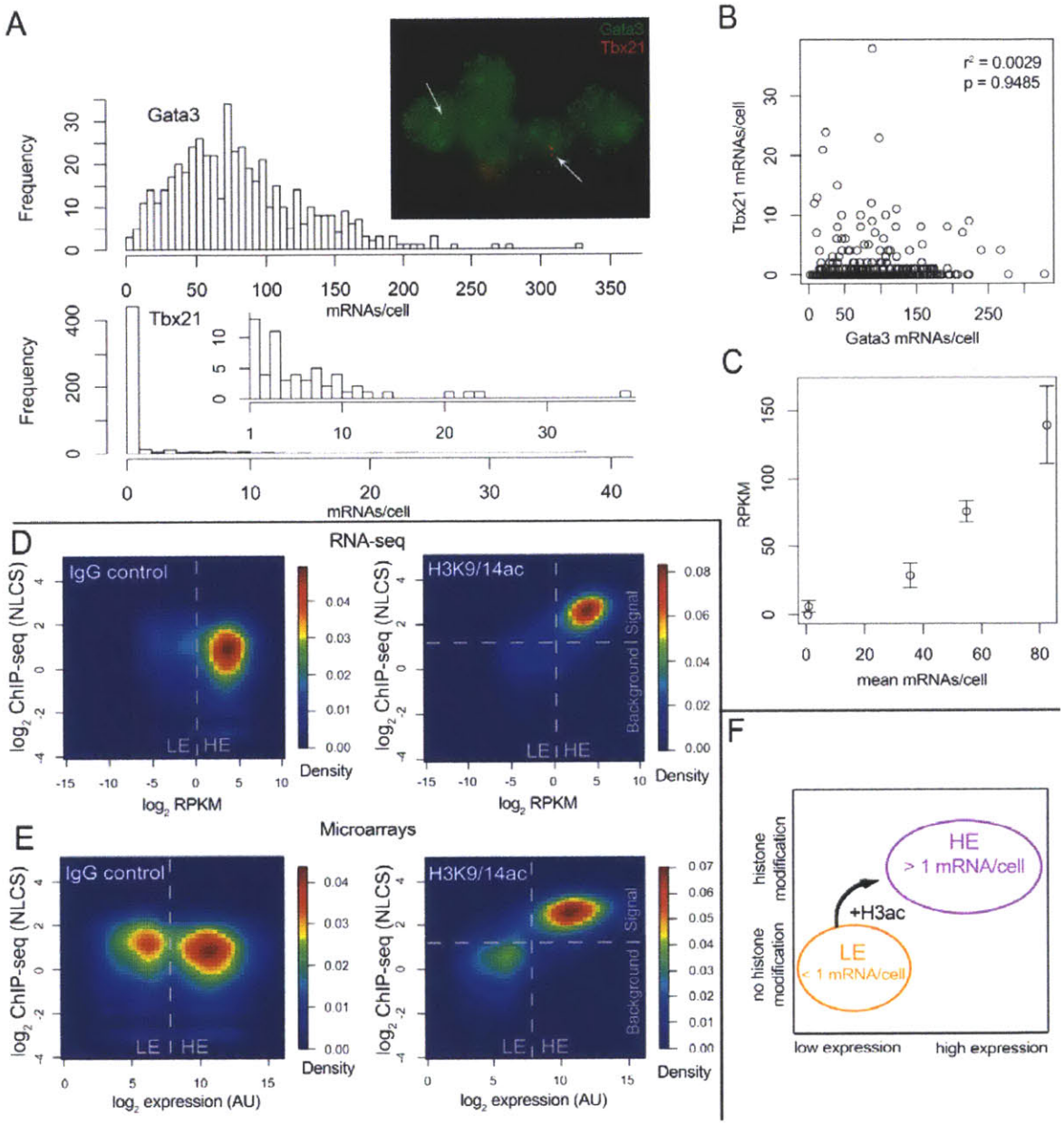


Figure 2.3. (A) Distribution of mRNA numbers among single cells. Histograms for Gata3 and Tbx21 (with an inset histogram starting from 1 instead of 0 to better illustrate higher expressions) and a sample fluorescence microscopy image are shown. Tbx21 transcripts are marked with white arrows to ease identification. (B) Correlation between Gata3 and Tbx21 expression. Correlation coefficient and significance are inset. (C) Plot of mean mRNA numbers per cell *versus* RNA-seq RPKM of five genes. Error bars indicate SEM from two RNA-seq biological replicates. (D-E) 2D kernel density estimates of gene

expression level vs. ChIP-seq signal for each gene for RNA-seq (D) and microarray (E) data. Divisions between background and signal for the ChIP-seq component were determined by curve fitting with the software EpiChIP (Hebenstreit and Teichmann, 2011) and are indicated. Divisions between LE and HE groups of genes are indicated. (F) Scheme summarizing the results.

Since introns are one- to two orders of magnitude longer than exons, introns should be detecTable 2.with roughly the same accuracy as exons, if the full-length set of introns of a gene is used. If we plot the RPKM in exonic regions *versus* the RPKM in intronic regions for each gene, there is significant correlation ($r^2 = 0.86$, $p < 2.2 \times 10^{-16}$) across the whole spectrum of expression levels. Calculating the correlation for lowly and highly expressed genes separately yields only slightly lower correlations among LE genes compared to HE genes, and both correlations are highly significant (Figure 2.2D). This provides evidence that confirms that LE genes are transcribed rather than experimental background: there would not be such a high correlation between introns and exons, particularly in the low abundance region, if their detection were due to noise.

We next studied gene expression using a single cell approach by performing single molecule RNA-FISH (Raj et al., 2008) for five genes that are expressed at different levels according to the literature and our RNA-seq data. The distributions of mRNA numbers per cell were very broad for expressed genes (e.g. Gata3), while low mRNA numbers from 'not-expressed' genes (e.g. Il2) were still detected (Figure 2.3A). All genes had Fano factors (σ^2/μ) larger than 4, indicating that they had extra-Poisson variation (a Poisson random variable would have $\sigma^2/\mu = 1$) and therefore burst-like transcription (Raj and van Oudenaarden, 2009) (Table 2.S3). Importantly, cells expressing Tbx21 were not anti-correlated with cells expressing Gata3 (Figure 2.3B), meaning that we do not have a sub-population of Th1 cells in our Th2 cell populations. This further demonstrates that LE expression is not due to a contaminating cell type, as the same cells express groups of genes at HE and others at LE levels.

Since the RPKM as measured by RNA-seq should be proportional to the mean mRNA numbers per cell, we can use the RNA-FISH results to estimate how our RPKM values translate into mRNA numbers. We find that one RPKM corresponds to an average of roughly one transcript per cell in our Th2 data set (Figure 2.3C). Please note that the value of one RPKM/one transcript on average per cell serves as an estimate only as it is based on a limited number of data points. See Figure 2.S14 for log transformed versions of Figure 2.3A-C.

It should be noted that the two groups of genes at high versus low expression levels cannot result from a mixture of different cell types. Mixing of different cell types

leads to gene expression levels for each gene that are an average across cell types. Hence such distributions will become more unimodal, not less so (following the central limit theorem).

To study the nature of the LE and HE groups in more detail, we prepared Th2 ChIP-seq data for the activating H3K9/14 acetylation histone modification (Roh et al., 2005; Wang et al., 2009b) (H3K9/14ac) and one IgG control. We calculated the histone modification level at each gene by identifying a globally enriched window around the transcription start sites of genes, and using reads in this window as a measure of each gene's modification level, normalized by the total reads (giving the normalized locus specific chromatin state, NLCS, as used in (Hebenstreit et al., 2011)). Thus we were able to plot histone modification levels of each gene against expression levels from the RNA-seq or microarray data using a heatmap representation (Figure 2.3D, RNA-seq, Figure 2.3E, microarrays). Figure 2.S15 is an alternative version of this figure, where we randomly assigned low RPKM values to the zero-read genes.

This strikingly confirms the two groups of gene expression levels, as there is a very good agreement between LE genes and absence of histone marks on one hand, and HE genes and presence of H3K9/14ac marks on the other hand (Figure 2.3D-E). This is seen for both the microarrays as well as the RNA-seq data. This extends previous findings of the relationship between H3K9/14ac and transcriptional activation by revealing an on/off-type of correlation between this histone mark and the LE/HE groups of genes. It should be noted that there is a very weak correlation within the LE and HE groups. The strongest correlation is within the RNA-seq HE group with a correlation coefficient $r^2 = 0.29$ in log space and $r^2 = 0.097$ on linear space.

Since the LE group of genes is still expressed at low levels and contains at least five genes that are characterized as not expressed and non-functional in Th2 cells, it seems likely that the HE group of genes represents the active and functional transcriptome of cells. This is supported by SILAC proteomics data (Graumann et al., 2008) which is available for the embryonic stem cell data we presented earlier (Figure 2.S10) and which indicates protein expression of HE genes only (Figure 2.S10C). The tight correlation recently observed between RNA and protein levels in three mammalian cell lines also supports this (Lundberg et al., 2010).

Gene ontology (GO) analysis of LE and HE genes in the Th2 cells supports the notion that HE comprises the functional transcriptome, as many T cell specific processes (e.g. GO:0050863, GO:0045582, GO:0042110) and housekeeping processes are enriched (Table 2.S4). On the other hand, many GO terms referring to differentiation of other celltypes (e.g. ear development GO:0043583, neuron fate commitment GO:0048663) are enriched among the LE set of genes (Table 2.S5).

In conclusion, our data shows that two large groups of genes can be discriminated based on the distribution of expression levels. RNA-FISH indicates that the boundary between the groups is found at an expression level of roughly one transcript per cell. In addition, H3K9/14ac marks are associated with the promoters of highly expressed genes only (Figure 2.3F). It thus seems likely that the LE/HE groups reflect different transcription kinetics depending on the chromatin state or *vice versa*. The LE group is likely to correspond to 'leaky' expression, producing non-functional transcripts. The majority of LE genes are expressed at less than one copy per cell on average, and it would be interesting to know whether such stochastic expression has any function, e.g. in cell differentiation, or any deleterious effects. There may be a trade-off between the cost of repressing expression entirely and unwanted consequences of stochastic expression.

Regulation of gene expression is mostly mediated by transcription factor binding events at promoters and enhancers, e.g. (Heintzman et al., 2009). Often, differential regulation induces only small changes in expression levels, probably serving to fine-tune expression and shifting genes within the HE group. Our data suggests that in addition to this, there is a key decision about whether a gene becomes "switched on" and expressed which coincides with a boost in both transcription and H3K9/14ac histone modification.

2.4 Materials and methods

Th2 cell differentiation culture

Spleens of C57BL/6 mice aged from 7 weeks to 4 months were removed and softly homogenized through a nylon mesh. The medium used throughout the cell cultures was IMDM supplemented with 10 % FCS, 2 μM L-glutamine, penicillin, streptomycin and 50 μM β-mercaptoethanol. Cells were washed twice and purified by a Ficoll density gradient centrifugation. CD4+CD62L+ cells were isolated by a two-step MACS purification using the CD4+CD62L+ T Cell Isolation Kit II (Miltenyi Biotec). Cells were seeded into 24 well plates that had been coated with a mix of anti-CD3 (1 μg/ml, clone 145-2C11, eBioscience) and anti-CD28 (5 μg/ml, clone 37.51, eBioscience) antibodies overnight, at a density of 250,000 cells/ml and a total volume of 2 ml. The following cytokines and antibodies, respectively, were added to the Th2 culture: recombinant murine IL-4 (10 ng/ml, R&D Systems), neutralizing IFN-γ (5 μg/ml, Sigma). Cells were cultured for 4 to 5 days at 37 °C, 5 % CO_2. After this, cells were taken away from the activation stimulus, diluted 1:2 in fresh medium containing the same cytokine concentration as before. After two to three days of resting time, cells were directly crosslinked in formaldehyde for preparing ChIP-seq samples. For FACS stainings, cells were restimulated with phorbol dibutyrate and ionomycin (both used at 500 ng/ml, both from Sigma) for four hrs in the presence of Monensin (2 μM, eBioscience) for the last two hrs after the resting phase. For Realtime PCRs, the cells were lysed in Trizol.

FACS staining

After restimulation, cells were washed in PBS and fixed overnight in IC fixation buffer (eBioscience). Staining for intracellular transcription factor expression was carried out according to the eBioscience protocol, using Permeabilization buffer (eBioscience), and

the following antibodies: anti-GATA3-Alexa647 (one test, TWAJ, eBioscience), anti-Tbx21-PE (1/400, clone eBio4B10, eBioscience). Stained cells were analysed on a FACSCalibur (BD Biosciences) flow cytometer using Cellquest Pro and FlowJo software.

Realtime PCR

RNA of ~10^6 cells was isolated with Trizol (Invitrogen) according to the manufacturer's protocol. cDNA was produced using Superscript III reverse transcriptase (Invitrogen), following the protocol supplied by the manufacturer. The cDNA was subjected to realtime PCR, using the SYBR green PCR master mix (Applied Biosystems) and a 7900 HT Real-Time PCR system (Applied Biosystems). The threshold cycles (C_t) were determined. The primer sequences used are listed in Table 2.S6 and were mostly obtained from 'Primerbank' (http://pga.mgh.harvard.edu/primerbank/) (Spandidos et al., 2010).

RNA-seq data generation

poly-(A)+ RNA was purified from ~500,000 cells using the Oligotex kit (Qiagen). The manufacturer's protocol was slightly modified to include additional final elution steps resulting in a larger volume. After precipitation of RNA to concentrate it, 1st and 2nd strand cDNA synthesis was performed using the Just cDNA kit (Stratagene), skipping the blunting step and directly proceeding to PCI extraction. Quality of the cDNA was tested by realtime PCR for a housekeeping gene. After this, the cDNA was sonicated for a total of 45 min using the Diagenode Bioruptor at maximum power settings, cycling 30 sec sonications with 30 sec breaks. After precipitation, the sample was processed using the ChIP-seq sample prep kit (Illumina) with a slightly modified protocol (PCR before gel extraction). Sequencing for 36 or 41 bp was carried out on an Illumina GAII genome

analyzer. The data was deposited at Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), accession number GSE28666.

RNA-seq data processing

Reads were mapped to the mouse genome (mm9) with Bowtie (Langmead et al., 2009) using the command options -m 1 --best –strata --solexa1.3-quals, and were assigned to exons of RefSeq genes using custom perl scripts. We used the gene symbol as the primary identifier. Table 2.S2 shows the numbers of mapped reads. We further generated a library of splice junctions based on RefSeq genes, mapped unmapped reads to these and added the numbers of hits to the genes. The numbers of mapped reads per gene were corrected for mapability based on the 'CRG' tracks of the UCSC genome browser. RPKM were then calculated. In the case that multiple splice variants existed, the most highly expressed one was selected as representative for a gene's expression level. For generating the RPKM distributions of intergenic regions, we considered regions with a distance of at least 10 kb to any RefSeq or Ensembl gene. The distribution was based on random fragments of the same length distribution as gene lengths. Mapability was accounted for, and the randomization was performed twenty times. The same procedure was followed for determining the read distribution within introns (of RefSeq genes). To test for a possible RPKM bias in 5' or 3' ends of intronic regions, the introns of each gene were lined up. If the intronic region was at least 6 kb in total, RPKM were separately determined for the most 5' 2 kb, for the 2 kb in the center and for the most 3' 2 kb. The full-length of introns was used (for the sake of higher sensitivity) for plotting RPKM of introns *versus* exons (as in Figure 2.2D). A trend line was calculated based on a least squares fit of the $\log_2$-transformed data. Division into LE and HE was made along a line perpendicular to the trendline, crossing at Exon RPKM = 1. Correlations and significances calculated were based on Pearson's product moment correlation coefficient.

We prepared alternative versions of Figure 2.1A and Figure 2.3D, where we assigned a random $\log_2$ RPKM value derived from a Normal distribution with $\mu = -12$ and $\sigma = 1$ to each gene without sequencing reads (Figure 2.S4 and 2.S15).

Integration of the RNA-seq data with mircoarray- and histone-modification data was based on gene symbols.

The RNA-seq data of (Cloonan et al., 2008) was downloaded from the NCBI short read archive (http://www.ncbi.nlm.nih.gov/sra/), accession number SRX003912. The reads were mapped to mm9 in colorspace format using Bowtie with similar settings as above. The mapped reads were separated into those sense and those antisense to RefSeq genes and processed similarly as described above. Read distributions in intergenic regions were determined as described above for our data.

RNA-seq data from (Mudge et al., 2008) was downloaded from GEO, accession number GSE12297. We used the processed data for 'Cerebellar cortex 40 Control' directly and performed no further calculations except log transformation and kernel density estimation. The RNA-seq data for 'skeletal muscle' from (Wang et al., 2008) was downloaded from GEO (accession number GSE12946). We used the data that was mapped to the human genome (hg18), assigned it to RefSeq genes, and processed it similarly as described above. We further downloaded RNA-seq data from (Marioni et al., 2008) from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra/). The data for human liver tissue was used (accession numbers SRX000571 and SRX000604). The two files were concatenated, mapped to the human genome (hg18) with Bowtie and processed further as described above. Finally, RNA-seq data for mouse brain (Mortazavi et al., 2008) was downloaded from SRA (accession numbers SRX000350 and SRX001866). As described above, the two files were concatenated, mapped to the mouse genome (mm9) with Bowtie and processed further.

Kernel density estimation

Gene expression distributions were displayed as kernel density estimates in most cases.

These were calculated using the function '*density()*' of the freely available statistical software package 'R' (http://www.r-project.org/). We used default settings of this function unless stated otherwise. This means a Gaussian kernel and that the 'bandwidth equals 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (corresponding to Silverman's "rule of thumb", ((Silverman, 1986), page 48, eqn (3.31)) unless the quartiles coincide when a positive result will be guaranteed' (R manual). For 2D kernel density estimations we used the function '*kde2d()*' of the R library 'MASS' with the default bandwidth and a Gaussian kernel. This bandwidth is calculated based on a variation of above formula for the 1D case, where the factor 1.06 instead of 0.9 is used. Densities were estimated at 50 grid points in either direction and displayed as heatmaps.

RNA-seq data sensitivity analysis

The RNA-seq detection limit was explored by two different approaches. Firstly, random subsets of different sizes were taken from the total reads we generated. The number of genes that remained undetected (zero reads) were plotted as a function of the subset size. The subsetting was performed five times for each subset-size and the average number of zero-read genes was determined.

As a second approach, we determined the expected number of zero-read genes depending on the expression level. To this end, we calculated the expected number of reads for each gene in dependency of the expression level (as reads per kilobase, RPK, instead of RPKM which includes normalization by the total number of mapped reads) and gene-length (the length distribution of all genes was used). The expected read number is generally assumed to be Poisson-distributed (Jiang and Wong, 2009) and can be used as an estimator of the single parameter of a Poisson distribution, λ, which is equal to mean and variance of the distribution. Studying the probability density function of a Poisson distribution for a certain λ reveals the expected frequency of zeros, which corresponds to genes of a certain length that remain undetected at a certain RPK despite

being expressed. Assuming an equal distribution of gene lengths at all expression levels, we could thus sum up the proportion of zero read genes for all gene lengths and thus obtain the total expected portion of undetected genes for all RPK levels. For instance, at RPK = 1 we would expect two sequencing reads for a gene that is 2 kb long and one read for a 1 kb gene (giving the same expression level). Since the actual read numbers vary according to a Poisson distribution, not all genes that are expressed at that level will have exactly one or two reads, respectively, but some will have more and some none at all. The Poisson distribution gives the expected portion of zeros, which would be 37 % for the 1 kb gene and 13.5 % for the 2 kb gene. Thus, if we detect 150 1 kb genes and 250 2 kb genes at RPK = 1, we can estimate that a further 127 (= 150/(1 – 0.37) – 150 + 250/(1 – 0.135) – 250) genes of the same lengths are expressed at the same level but remain undetected.

We further used above calculation to estimate how the distribution of expression levels is affected by the sensitivity of RNA-seq. To this end, we binned the actual expression distribution into bins of size 1 on the $\log_2$ RPK scale and extrapolated the number of expressed genes by adding the inferred number of undetected genes to each bin.

Microarray data

Microarray data (Th2) of (Wei et al., 2009) were downloaded from GEO, accession number GSE14308. Either normalized (by the authors) microarray data was used (Figures 2.1B, 2.S5, 2.S6 and 2.S8), in which case present (P) and absent (A) calls of the probesets were ignored, or custom processing schemes were applied to the raw data (Figure 2.S7 and S8). The mean of the two replicates of the microarray data was calculated for each probeset and was $\log_2$-transformed. These values were then linked to RefSeq genes based on the Affymetrix MOE430 2.0 annotations of build 27. If more than one probeset was mapping to a gene, the probeset with the highest intensity was chosen as representative of the gene's expression level.

We further downloaded microarray data for murine bone cells from the GNF Mouse GeneAtlas V3 ((Lattin et al., 2008); GEO, GSE10246) and processed them as described above. Similarly, the processed microarray data for two replicates of human Cd133+ cells (Cui et al., 2009) was downloaded from GEO, accession number GSE12646 and processed (using Affymetrix build 28 annotations for the Affymetrix U133A chip). Finally, we downloaded from GEO (accession number GSE7763) microarray data for *Drosophila* eye tissue from the FlyAtlas (Chintapalli et al., 2007). We mapped the probesets to genes using Affymetrix probe annotations (build 28) for GeneChip Drosophila Genome 2.0 and processed the data the same way as the other datasets.

Curve fitting

Curve fitting and/or clustering of the data into LE and HE sets by expectation maximization was performed on the $\log_2$ transformed RNA-seq or microarray data using the R library 'Mclust'. The log likelihood values output by Mclust were used to calculate AIC (Akaike, 1974), BIC (Schwarz, 1978) and likelihood ratio statistics (Casella and Berger, 2001). The latter were calculated for the model with n components as the null model and the one with $n+1$ components as the alternative model ($0 < n < 9$). We approximated the test statistics with X^2 distributions and calculated the p-values with R.

SILAC data

Processed SILAC data for murine embryonic stem cells was downloaded from the supplementary material of (Graumann et al., 2008). Using UCSC Table 2.browser, we linked the protein expression data to the RNA-seq data of (Cloonan et al., 2008) by referencing the RefSeq protein ID provided by Graumann et al to the gene symbol which

we used as gene identifier for the RNA-seq data. A protein was regarded as expressed if it had a non-zero 'MS intensity' value.

GO analysis

Genes were clustered into LE and HE subsets by expectation maximization using the R library Mclust. Enrichment analysis of 'process' GO terms was performed with the Generic Gene Ontology (GO) Term Finder (http://go.princeton.edu/cgi-bin/GOTermFinder) (Boyle et al., 2004) using the combined LE/HE set of genes as the custom background. Bonferroni-adjusted p-values were used.

Single molecule fluorescence in situ hybridization

We performed single-molecule FISH on the Th2 cells and counted the mRNAs in individual cells as described previously (Raj et al., 2008). Briefly, harvested Th2 cells were fixed with 3.7% formaldehyde for 10 minutes, washed twice with PBS, and permeabilized in 70% ethanol. For hybridization, the samples were resuspended in 100 μl of hybridization solution containing labeled DNA probes in 2xSSC, 1 mg/ml BSA, 10mM VRC, 0.5 mg/ml Escherichia coli tRNA and 0.1 g/ml dextran sulfate, with 10 to 25% formamide, which varies for different probes, and incubated overnight at 30°C. The next day, the samples were washed twice by incubating in 1 ml of wash solution consisting of 10 to 25% formamide and 2xSSC for 30 minutes. The sequences of the probes are available upon request.

Image acquisition

The samples were resuspended in glucose oxidase anti-fade solution, which contains 10 mM Tris (pH 7.5), 2xSSC, 0.4% glucose, supplemented with glucose oxidase and catalase. Then 8 μl of cell suspension were sandwiched between two coverglasses, and mounted on a glass slides using a silicone gasket. Images were taken with a Nikon TE2000 inverted fluorescence microscope equipped with a 100x oil-immersion objective and a Princeton Instruments camera using MetaMorph software (Molecular Devices, Downington, PA). Stacks of images were taken automatically with 0.4 microns between the z-slices.

Image analysis

To segment the cells, a marker-guided watershed algorithm was used. Briefly, cell boundaries were obtained by running an edge detection algorithm on the bright-field image of the cells. To generate markers, the centroid of the region enclosed by individual cell boundaries is computed. A marker-guided watershed algorithm was then run on the distance transformation of the cell boundaries, using the markers located within the cell boundaries (Figure 2.S16). The resultant cell segmentation image was then manually curated for occasional mis-segmentations.

To quantify the number of RNA molecules in each cell, a log filter was run over each optical slice of an image stack to enhance signals. A threshold was taken on the resultant image stack to pick up mRNA spots. The locations of mRNA spots were then taken to be the regional maximum pixel value of each connected region (Figure 2.S17). The number of mRNA spots located within the cell boundaries of an individual cell was thus quantified.

ChIP-seq data analysis

We used murine Th2 cell data for the H3K9/14ac histone modification and an IgG control from (Hebenstreit and Teichmann, 2011) (available on GEO, accession number GSE23092). The reads were mapped to the mouse genome (mm9) using Bowtie as for the RNA-seq analysis. Further steps of the analysis were performed using the software EpiChIP (http://epichip.sourceforge.net/index.html) (Hebenstreit et al., 2011). Briefly, the mapped reads were assumed to be the ends of 200 bp long fragments following the XSET method (Pepke et al., 2009). Then EpiChIP was used to identify an optimal sequence window with respect to gene coordinates for analysis of the histone modification stati at all (RefSeq) genes. The resulting window of -400 to +807 bp at transcriptional start sites was used to quantify the ChIP-seq signal for each gene (the area below the peaks within this window) which was normalized by the total (genomewide) area to yield the "normalized locus specific chromatin signal" (NLCS)(Hebenstreit et al., 2011). These values were $\log_2$ transformed and displayed against the RNA-seq or microarray expression levels as two dimensional density estimations. The threshold separating background from signal was determined with the curve fitting function of EpiChIP. For the alternative version of Figure 2.3D (Figure 2.S15), we assigned a random $\log_2$ RPKM value derived from a Normal distribution with $\mu = -3$ and $\sigma = 1$ to each gene without ChIP-seq sequencing reads.

Acknowledgments

We would like to thank Guilhem Chalancon and Joseph Marsh for reading the manuscript and making valuable suggestions, Ines de Santiago and Ana Pombo for helpful and interesting discussions, Lucy Colwell for her role in establishing a fruitful collaboration, and Jonathon Howard for reminding us of the importance of absolute numbers.

Contributions

Experiments, with the exception of RNA-FISH, were carried out by DH. RNA-FISH staining and image processing were carried out by MF. Computational analyses were carried out by DH, with contributions from MG and VC. DH and SAT wrote the manuscript with contributions from MF and AVO.

2.5 Supplementary Information

Supplementary figures

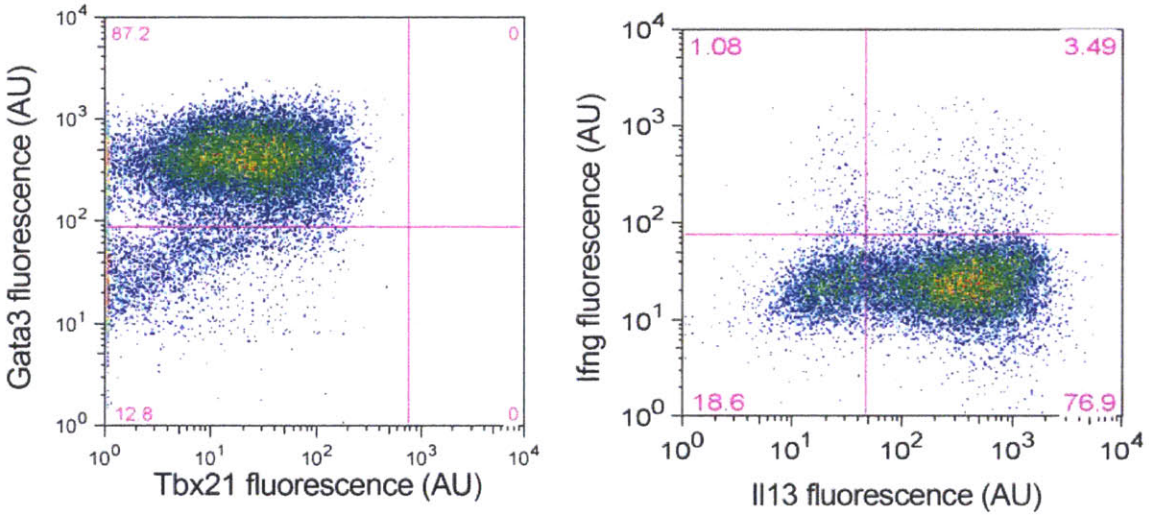


Figure 2.S1. Th2 cells were stained by intracellular staining with anti-Gata3, anti-Tbx21, anti-Ifng, and anti-Il13 antibodies and analyzed by FACS. Gata3 and Il13 are markers of Th2 differentiation, so a high proportion of Gata3 and Il13 expressing cells indicates a high level of Th2 homogeneity in the cell population. Tbx21 and Ifng are markers of Th1 cells, and are shown as a control. Each dot represents a single cell with fluorescence intensities for the two antibody stains on the x- and y-axes. Overlapping dots change color to indicate the density of cells at that point. The purple lines separate the plots into four regions each, depending on whether cells are expressing or the proteins or not. ~80 to 90% purity was routinely achieved, indicating successful Th2 differentiation.

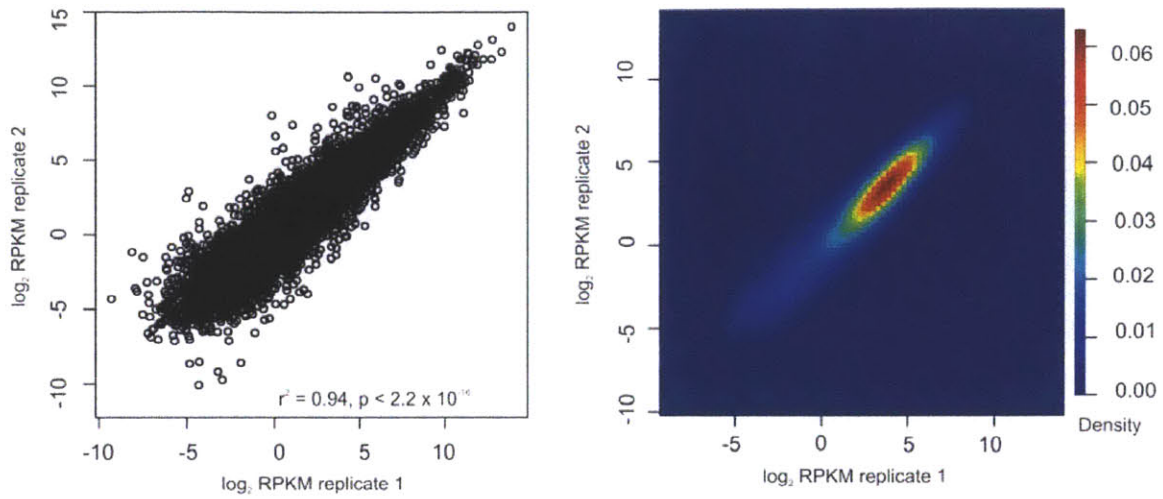


Figure 2.S2. Correlation between two RNA-seq replicates. A scatter plot (left) and a 2-D kernel density estimation are shown (right). Correlation coefficient and significance of correlation are inset in the left panel.

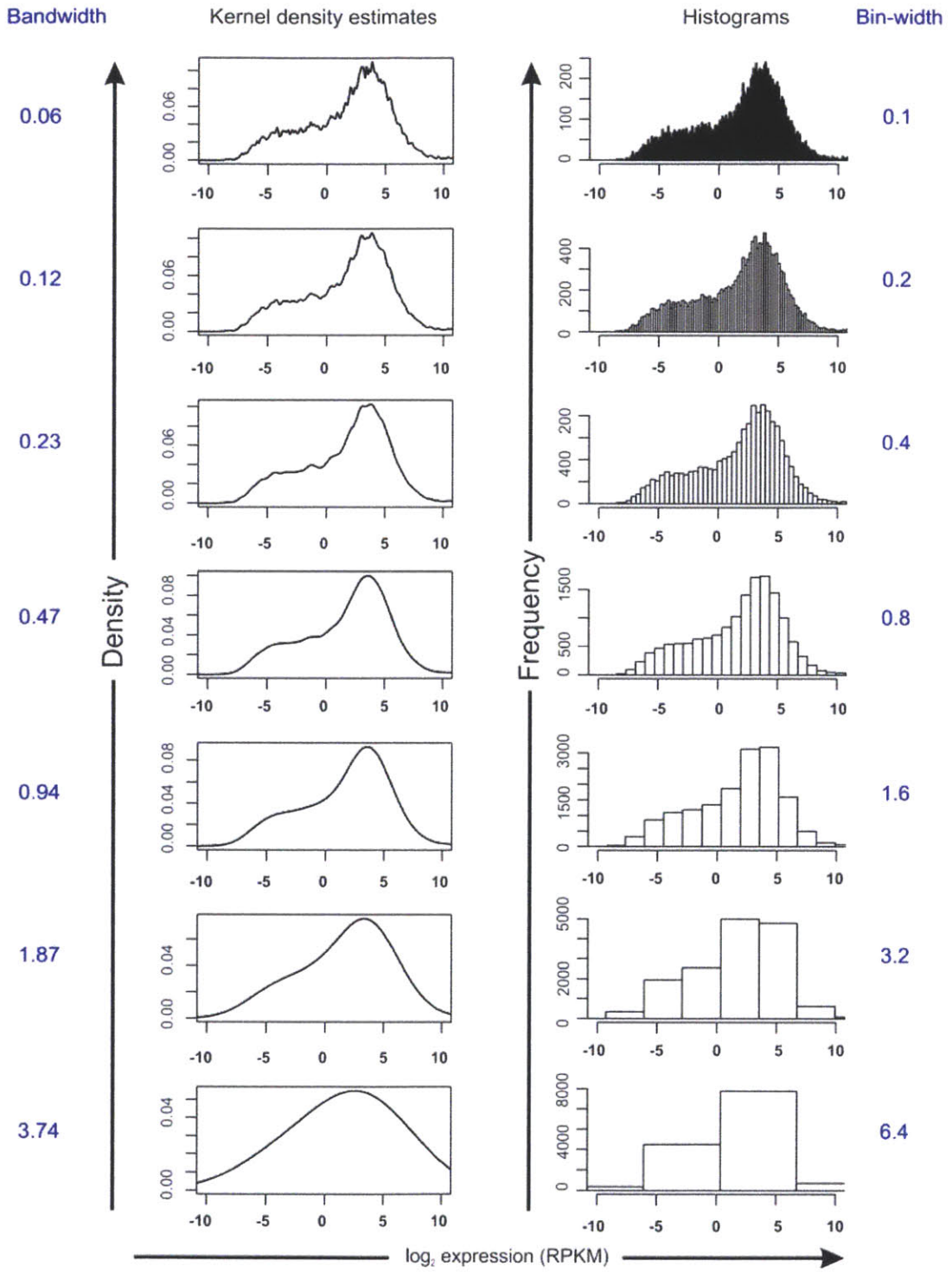


Figure 2.S3. Examples of how different visualization methods affect the appearance of the RNA-seq data. The left panel corresponds to kernel density estimates (KDE). To demonstrate that the structure of the data is conserved under different settings, the

bandwidth (corresponding to the standard deviation of the Gaussian kernel) was increased in 2-fold steps from top to bottom (blue, left side). The bandwidth in the center corresponds to Silverman's 'rule of thumb'. The right panel shows histograms with different bin-sizes (indicated in blue on the right side). The structure of the data is conserved if the bin-size is less than the distance between the two peaks.

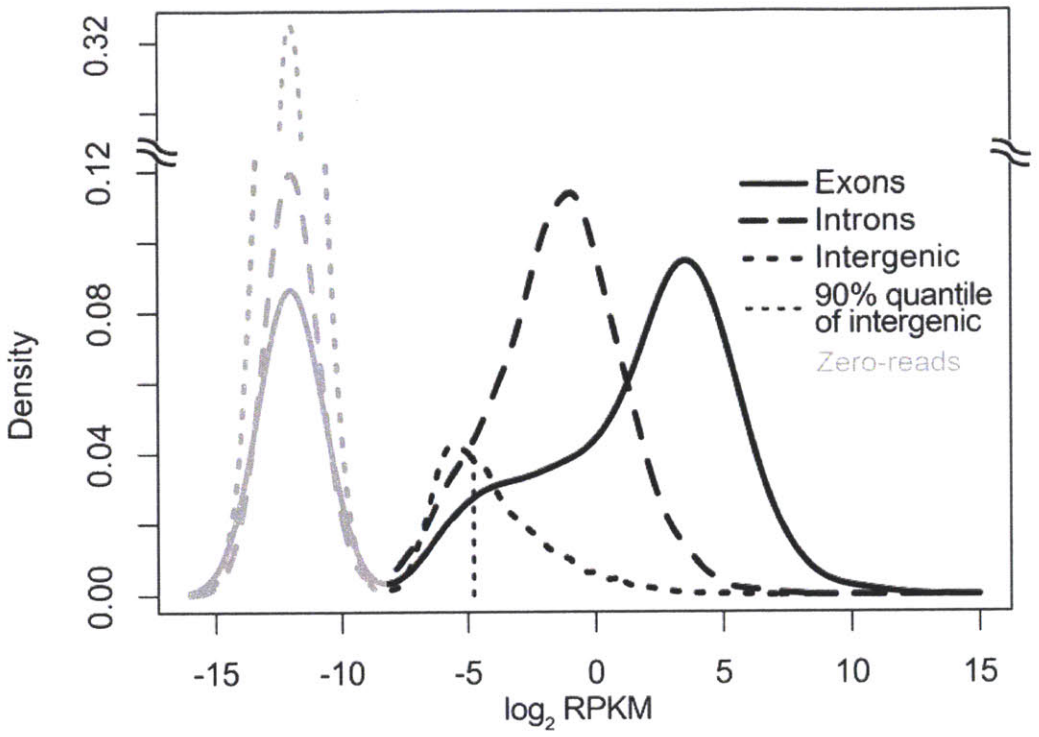


Figure 2.S4. Kernel density estimates of RPKM distributions of RNA-seq data within exons, introns and intergenic regions as in Figure 2.1A. To indicated the fractions of fragments/genes with zero reads (grey), they were assigned random RPKM values, drawn from a normal distribution with mean = -12 and standard-deviation = 1 on the $\log_2$ scale.

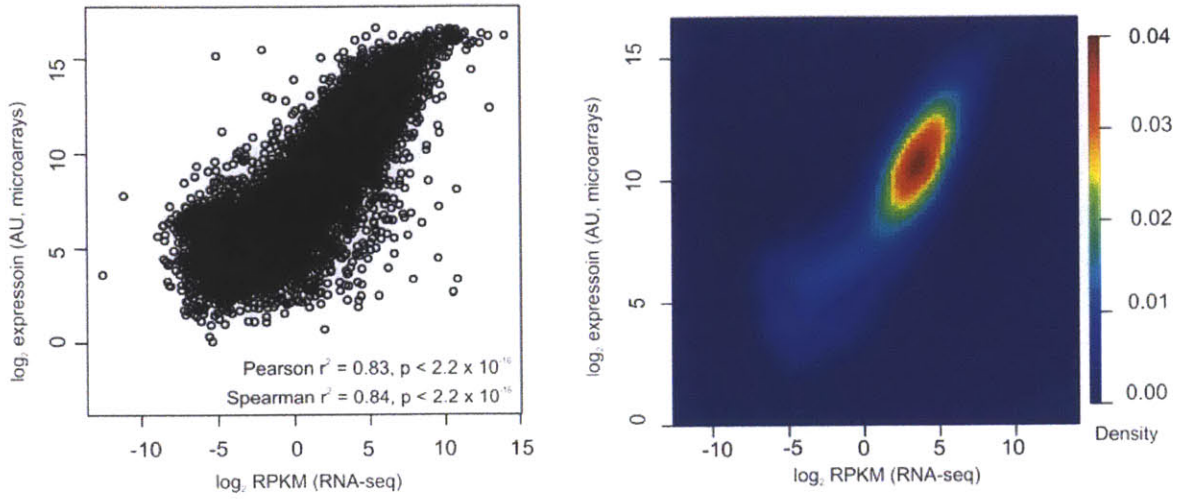


Figure 2.S5. Correlation between RNA-seq and microarray data (Wei et al., 2009). A scatter plot (left) and a 2-D kernel density estimation are shown (right). Correlation coefficients and significance of correlations are inset in the left panel.

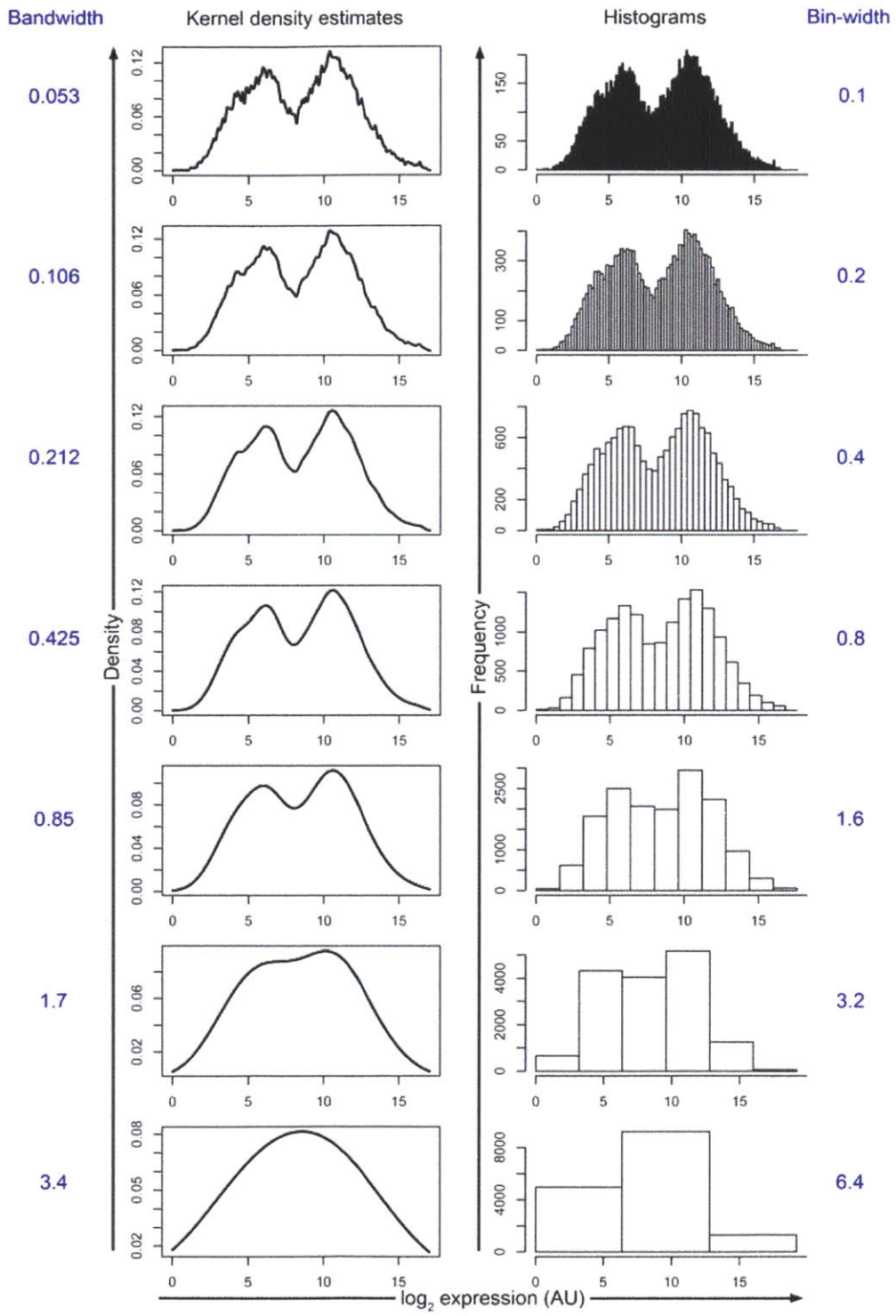
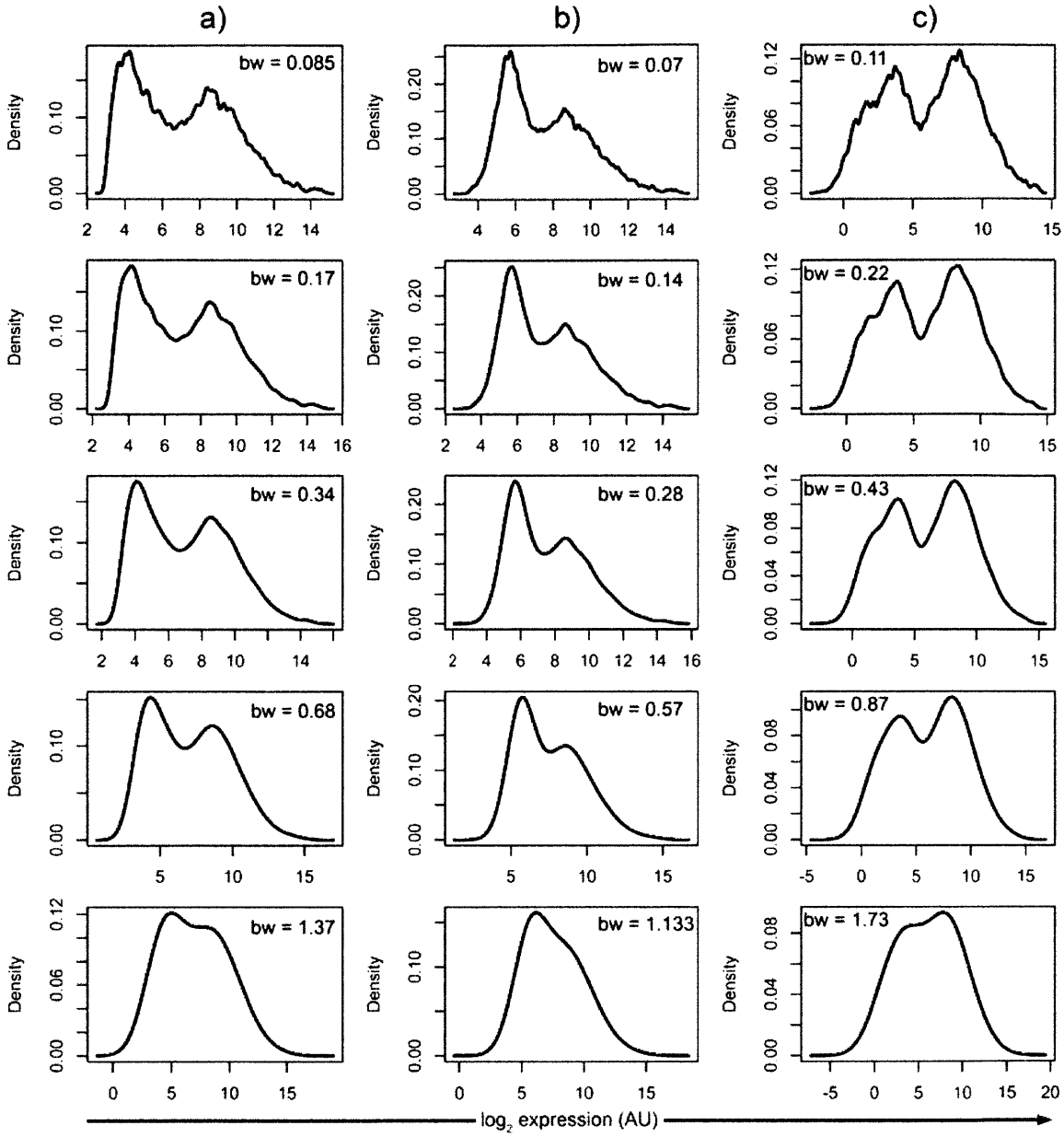


Figure 2.S6. Examples of how different visualization methods affect the appearance of the microarray data ((Wei et al., 2009). The left panel corresponds to kernel density estimates (KDE). To demonstrate that the structure of the data is conserved under

different settings, the bandwidth (corresponding to the standard deviation of the Gaussian kernel) was increased in 2-fold steps from top to bottom (blue, left side). The bandwidth in the center corresponds to Silverman's 'rule of thumb'. The right panel shows histograms with different bin-sizes (indicated in blue on the right side). Bimodality is conserved if the bin-size is less than the distance between the two peaks.



	a)	b)	c)
Background correction	RMA	MAS	MAS
Normalization	Quantile	Quantile	QSpline
PM correction	PM only	PM only	MAS
Summarization	Median polish	avgdiff	Median polish

Figure 2.S7. Examples for three further processing schemes in addition to MAS5 used in the main text. The raw data of (Wei et al., 2009) were processed by schemes a), b), and c) as indicated in the Table 2.and on top of the figure. PM, perfect match, RMA, robust

multi-chip average, MAS, microarray suite (Affymetrix). See the R Vignette of the 'affy' library for explanations of the individual methods and algorithms. Kernel density estimates (KDE) of the gene expression level distributions are shown. To demonstrate that the structure of the data is conserved under different KDE settings, the bandwidth (corresponding to the standard deviation of the Gaussian kernel) was increased in 2-fold steps from top to bottom (the bandwidth is given as 'bw =' in blue). The bandwidth in the center corresponds to Silverman's 'rule of thumb'.

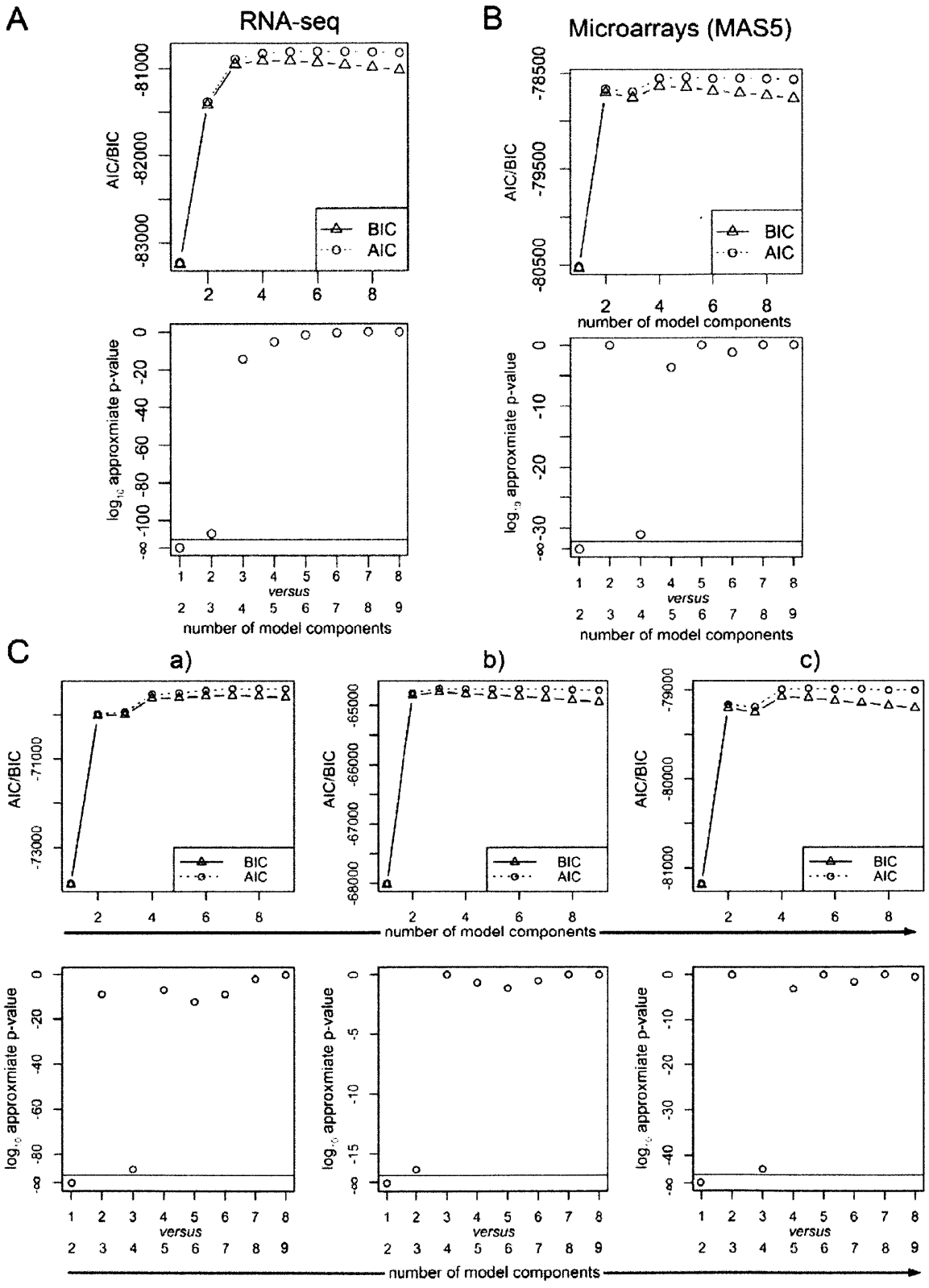

A
RNA-seq
AIC/BIC
BIC
AIC
number of model components
versus
B
Microarrays (MAS5)
C
a)
b)
c)

Figure 2.S8. Goodness-of-fit tests for mixture models of one- to nine lognormal components fit to our RNA-seq data (A) and the microarray data of (Wei et al., 2009) (B, C) by expectation maximization. Tests for data normalized by MAS5 (B), as used in the main text, and by the three alternative normalization methods (C) as demonstated in Figure 2.S7 (a), b) and c)) are shown as indicated. The tests used were the Akaike Information criterion (AIC), the Bayesian information criterion (BIC), and a likelihood ratio test. For the latter, we compared each model to the next more complex one in terms of components. We numerically calculated the $\log_{10}$ p-values based on a χ^2 distribution. In the case that the numerical p-value was zero, we included it on the log scale as $-\infty$.

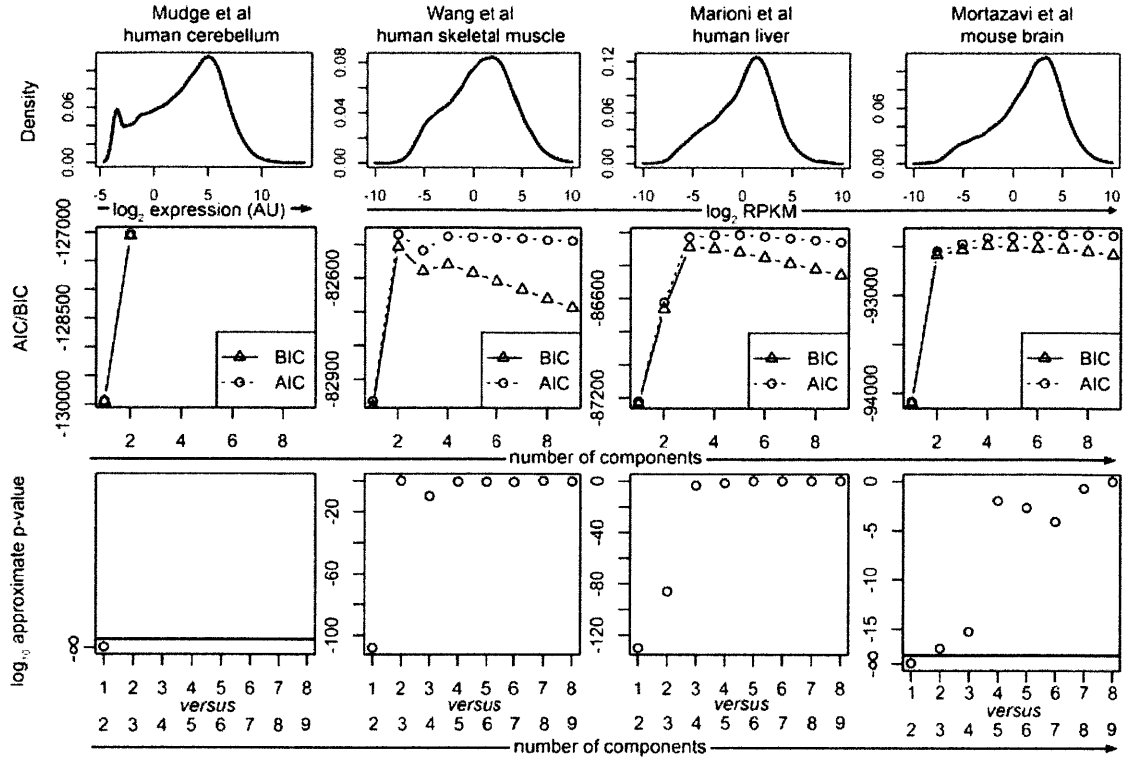


Figure 2.S9. Kernel density estimates (KDE) and goodness-of-fit test for four additional RNA-seq datasets (Marioni et al., 2008; Mortazavi et al., 2008; Mudge et al., 2008; Wang et al., 2008). The KDE are shown on top using a Gaussian kernel and a bandwidth corresponding to Silverman's 'rule of thumb'. All distributions exhibit a shoulder on the left side. The goodness-of-fit tests used were the Akaike Information criterion (AIC), the Bayesian information criterion (BIC), and a likelihood ratio test. For the latter, we compared each model to the next more complex one in terms of components. We numerically calculated the $\log_{10}$ p-values based on a χ^2 distribution. In the case that the numerical p-value was zero, we included it on the log scale as $-\infty$.

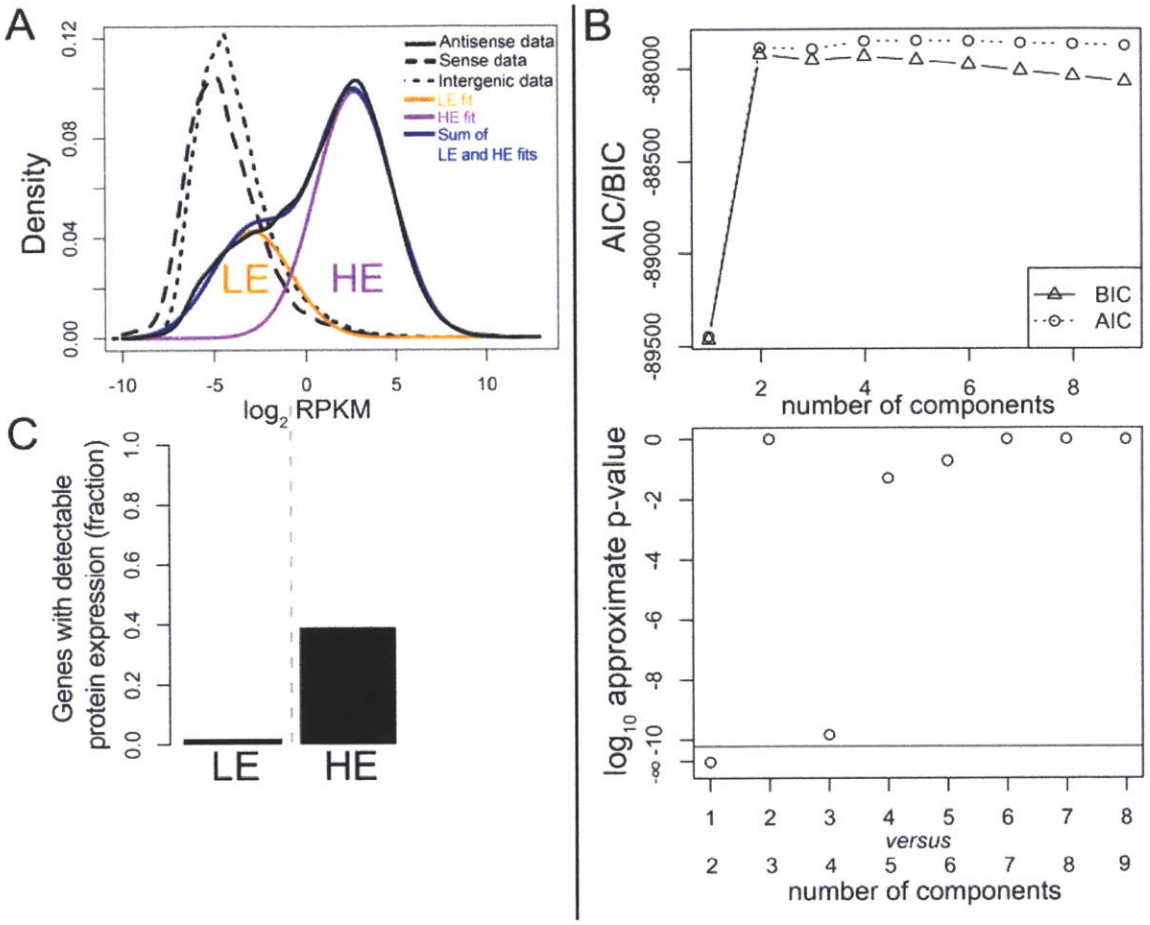


Figure 2.S10. LE and HE groups in RNA-seq data of murine embryonic stem cells from (Cloonan et al., 2008). (A) The kernel density estimates (KDE) of expression levels are shown separately for genes in sense or antisense with reads mapping to them, since the data was prepared in a strand-specific manner (reads antisense to genes are selected by the experimental protocol), and for intergenic regions as indicated. The KDE use a Gaussian kernel and a bandwidth corresponding to Silverman's 'rule of thumb' (see Materials and Methods). Curve fitting was carried out as described for Figure 2.1C. (B) Plots of AIC, BIC and p-values of likelihood ratio tests as goodness-of-fits indicator for one- to nine-component normal distribution mixture models as described in Figure 2.S8 and S9. (C) Genes were separated into LE and HE sets based on the expectation-maximization based curve fittings. SILAC protein expression data of murine embryonic stem cells (Graumann et al., 2008) was used to determine the fraction of genes that are expressed as proteins for the LE and HE sets separately.

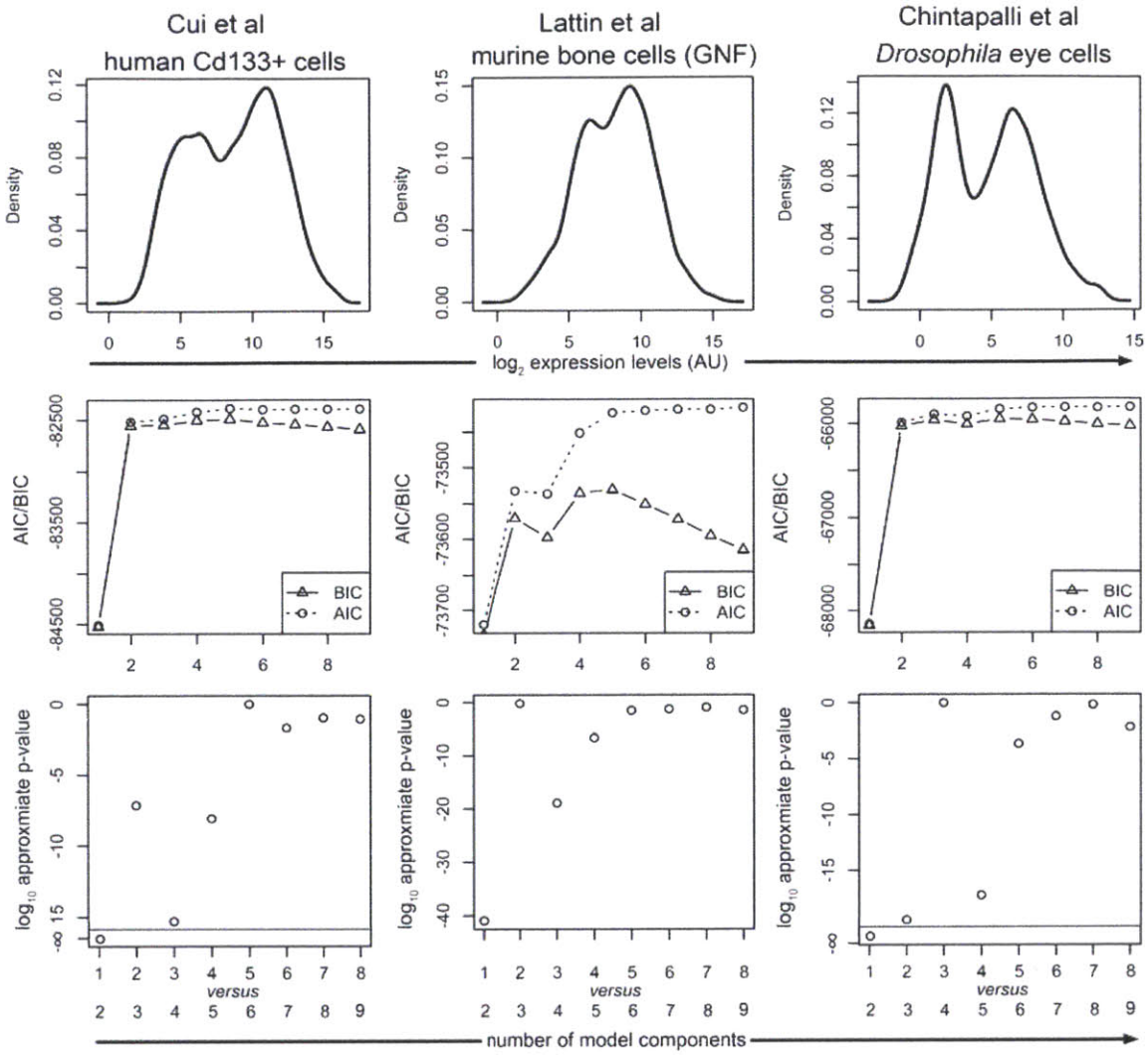


Figure 2.S11. Kernel density estimates (KDE) and goodness-of-fit test for three additional microarray datasets (Chintapalli et al., 2007; Cui et al., 2009; Lattin et al., 2008). The KDE are shown on top using a Gaussian kernel and a bandwidth corresponding to Silverman's 'rule of thumb'. All distributions exhibit bimodality. The goodness-of-fit tests used were the Akaike Information criterion (AIC), the Bayesian information criterion (BIC), and a likelihood ratio test. For the latter, we compared each model to the next more complex one in terms of components. We numerically calculated the $\log_{10}$ p-values based on a χ^2 distribution. In the case that the numerical p-value was zero, we included it on the log scale as $-\infty$.

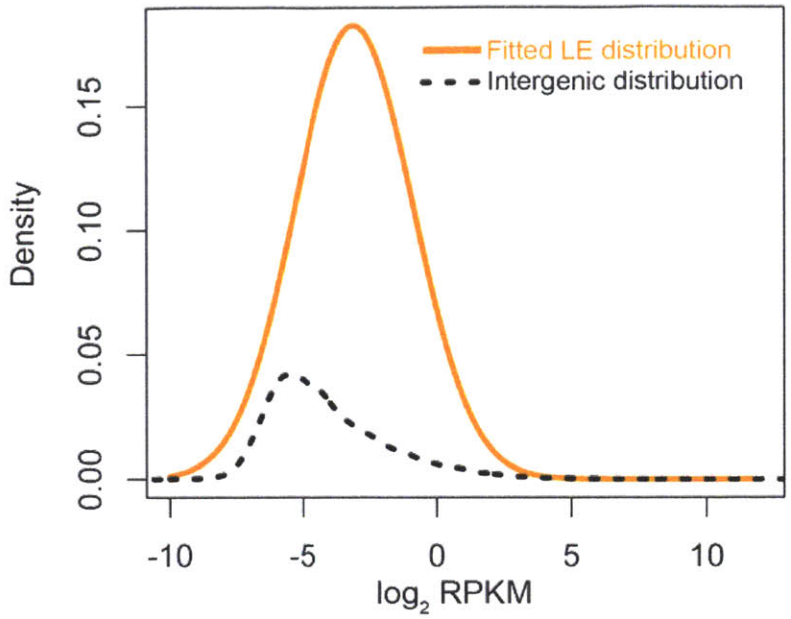


Figure 2.S12. Distributions of RPKM for LE genes and intergenic regions. The fragments used to estimate intergenic RPKM were based on randomizations using the same length distribution as the exonic parts of genes. The area under the LE distribution is normalized to one (in contrast to Figure 2.1A where it is part of the total RPKM distribution within exons). The area under the intergenic distribution is less than one because of the fragments with zero reads (please see Figure 2.S4).

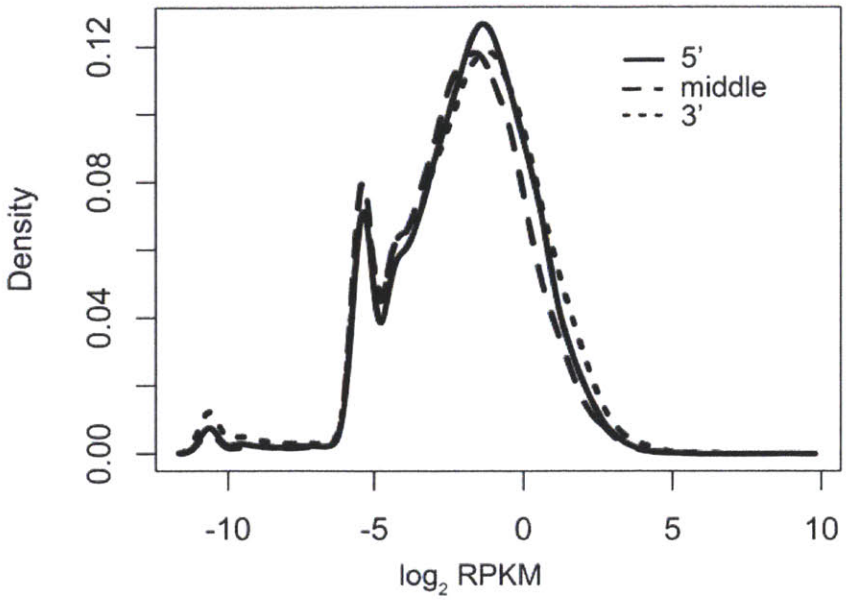


Figure 2.S13. No RPKM bias in 5' or 3' ends of intronic regions. Introns of each gene were lined up. If the intronic region was at least 6 kb in total, RPKM were determined for the most 5' 2 kb, for the 2 kb in the center and for the most 3' 2 kb. The $\log_2$ RPKM distributions for all selected genes are shown and are almost identical.

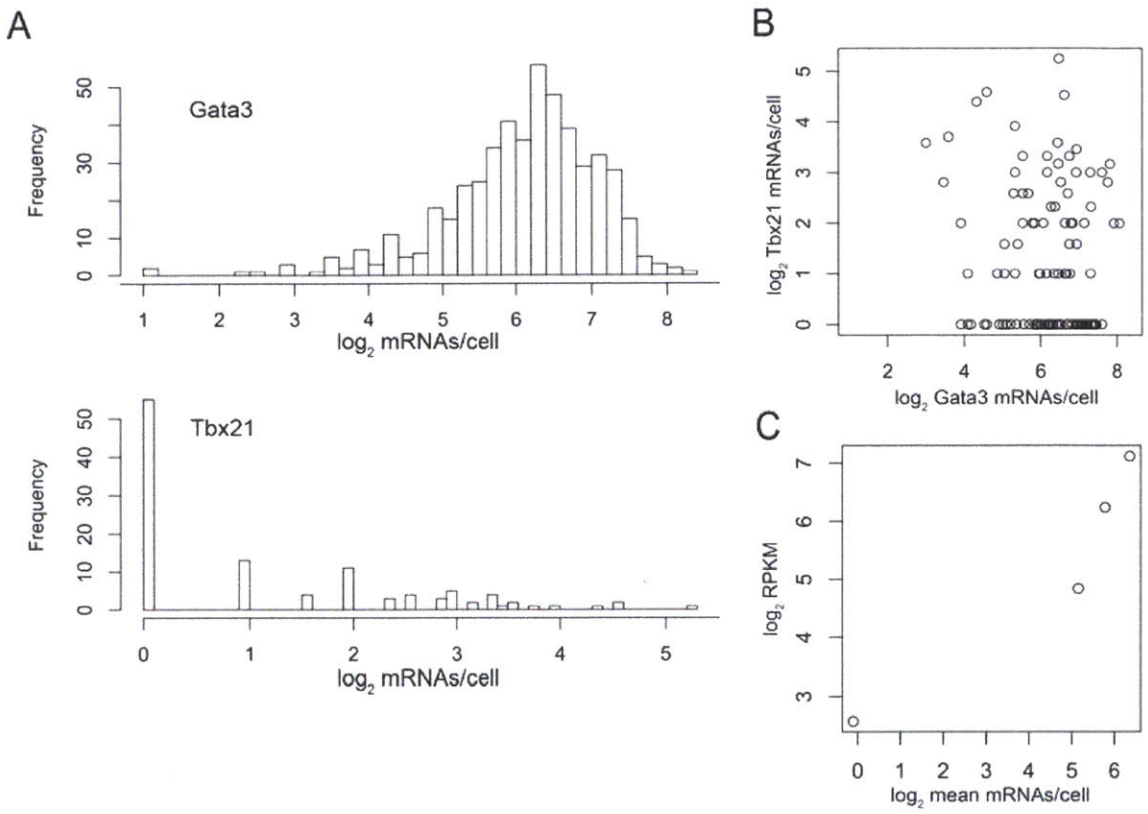


Figure 2.S14. Log_2 transformed plots of Figure 2.3A, B and C.

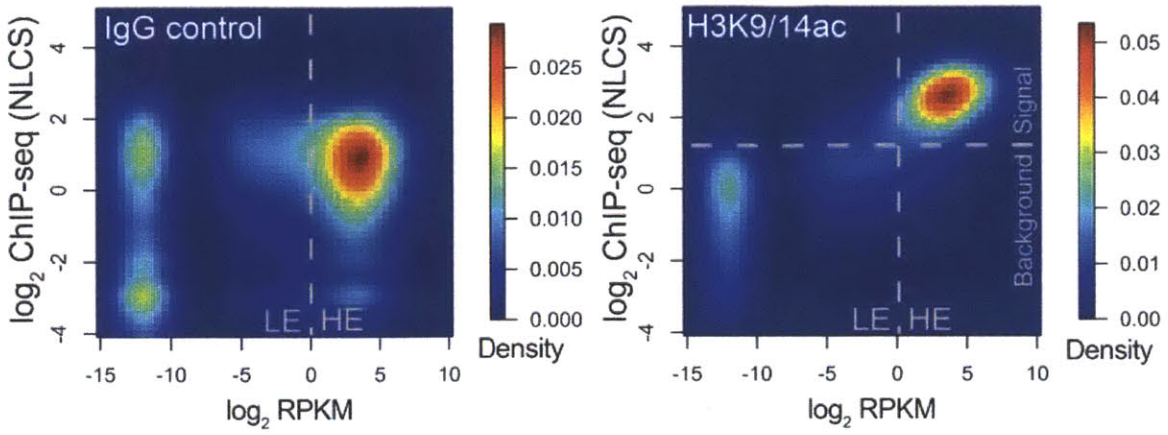


Figure 2.S15. 2D kernel density estimates of RNA-seq gene expression level vs. ChIP-seq signal for each gene as in Figure 2.3D. To indicate the fractions of fragments/genes with zero RNA-seq or ChIP-seq reads, random RPKM value were assigned to them, drawn from normal distributions with mean = -12 or mean = -3, respectively, and standard-deviations = 1 (in both cases) on the $\log_2$ scale. These genes appear as additional blobs with respect to Figure 2.3D.

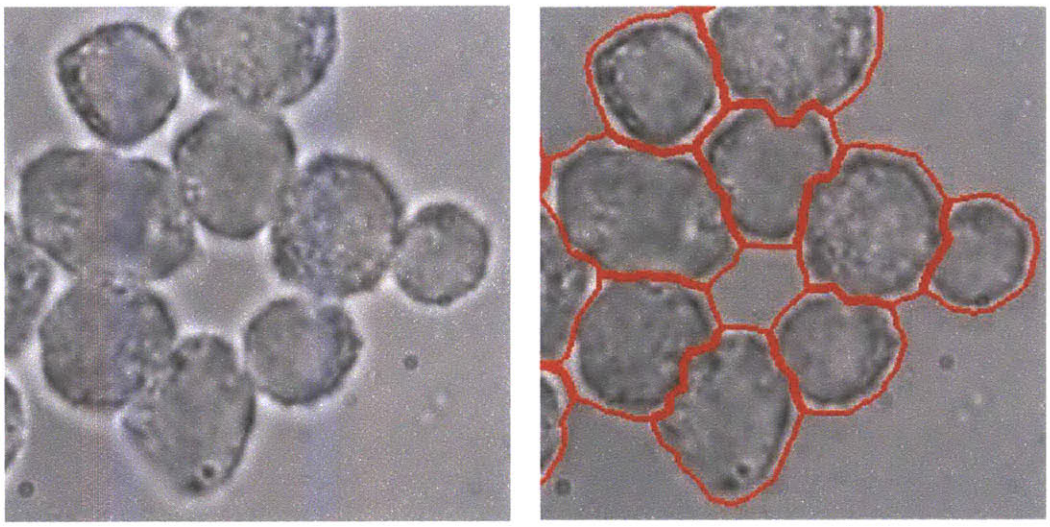


Figure 2.S16. Segmentation of cells using bright-field images. The left panel is a bright-field image of the cells. The right panel is the segmented image.

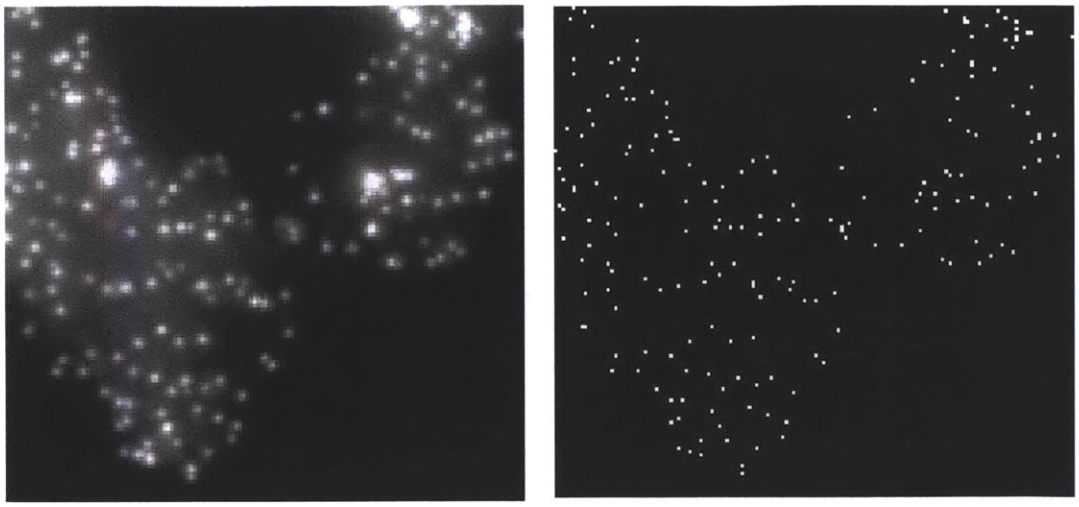


Figure 2.S17. Analysis of mRNA spots. The left panel is a fluorescent maximum Z-projection image showing Gata3 transcripts in Th2 cells. The right panel is processed binary image showing each individual mRNA transcript as a single bright pixel.

Supplementary tables

Gene symbol	Expressed in Th2 cells (literature)?	Expressed in Th2 cells (our RNA-seq)?	Used in FACS stain?	Amplified in PCR?	Used in RNA-FISH?
Arbp	Yes (house keeping gene used as PCR control, e.g. (Hebenstreit et al., 2008))	Yes		Yes	
Cd4	Yes (Zhu et al., 2010)	Yes		Yes	Yes
Gata3	Yes (Zhu et al., 2010)	Yes	Yes	Yes	Yes
Il13	Yes (Zhu et al., 2010)	Yes	Yes	Yes	
Il4	Yes (Zhu et al., 2010)	Yes		Yes	
Il7r	Yes (Gregory et al., 2007)	Yes		Yes	Yes
Tbx21	No (Zhu et al., 2010)	Yes	Yes	Yes	Yes
Ifng	No (Zhu et al., 2010)	Yes (LE)	Yes	Yes	
Il17a	No (Zhu et al., 2010)	Yes (LE)		Yes	
Il2	No (Malek, 2008)	No		Yes	Yes
Rorc	No (Zhu et al., 2010)	Yes (LE)		Yes	
Pgf		Yes (LE)		Yes	
Ptprg		Yes (LE)		Yes	
Wdfy3		Yes (LE)		Yes	
Ripply3		Yes (LE)		Yes	
Glp1r		Yes (LE)		Yes	

Table 2.S1. Genes examined in this study.

Sample	Read length	Total reads	Unique reads mapped to genome	Reads mapped to exons	Reads mapped to splice junctions
Replicate 1	41 bp	16,445,455	11,366,694	9,040,864	1,168,912
Replicate 2	36 bp	26,408,070	8,913,202	6,420,356	670,093

Table 2.S2. RNA-seq sequencing read statistics.

Gene symbol	Median	Mean	Stdev	Fano factor
Cd4	39	54.86	67.83	83.88
Gata3	75	82.56	48.41	28.39
Il2	0	0.68	1.64	4.00
Il7r	24	35.55	36.89	38.29
Tbx21	0	0.93	3.15	10.64

Table 2.S3. Single Molecule RNA-FISH statistics of five genes.

Gene symbol	fwd	rev	Exon spanning?	Junctions binding?
Arbp	AATCTCCAGAGGCACCATTG	ACCCTCCAGAAAGCGAGAGT	Yes	No
Cd4	AAGGGGCATGGGAGAAAGGAT	AAGGTCACTTTGAACACCCAC	Yes	Yes
Gata3	CCCTCCGGCTTCATCCTCT	CTGCACCTGATACTTGAGGC	No	
Il13	CCTGGCTCTTGCTTGCCTT	GGTCTTGTGTGATGTTGCTCA	No	
Il17a	CTCCAGAAGGCCCTCAGACTAC	AGCTTTCCCTCCGCATTGACACAG	Yes	No
Il2	TGAGCAGGATGGAGAATTACAGG	TGTTGTCAGAGCCCTTTAGTTTT	Yes	Yes
Il7r	TATGTGGGGCTCTTTTACGAGT	GCCTCGGCTTTAACTATTGTGT	Yes	Yes
Ifng	ATGAACGCTACACACTGCATC	CCATCCTTTTGCCAGTTCCTC	Yes	No
Pgf	TCTGCTGGGAACAACTCAACA	GTGAGACACCTCATCAGGGTAT	Yes	Yes
Ptprg	AGTCAGTCCGAGGGACAATTC	GGTGGCGTAGTCAAGGAGC	Yes	Yes
Rorc	CCGCTGAGAGGGCTTCAC	TGCAGGAGTAGGCCACATTACA	Yes	Yes
Tbx21	TTTCCAAGAGACCCAGTTCATTG	ATGCGTACATGGACTCAAAGTT	Yes	Yes
Wdfy3	CCACCATCGGGTTCATTAACA	GTGGGACAGAGATGCCTATGT	Yes	No
Ripply3	GGCCCGAAAGTTCCATTCCA	CTCCCGATGTGTGTTGGTCT	Yes	Yes
Glp1r	ACGGTGTCCCTCTCAGAGAC	ATCAAAGGTCCGGTTGCAGAA	Yes	No

Table 2.S6. Primer sequences.

CHAPTER 3

Stochastic Cytokine Expression Induces Mixed T Cell States

3.1 Abstract

During eukaryotic development, the induction of lineage-specific transcription factors typically drives differentiation of multipotent progenitor cells, while repressing that of alternate lineages. We explored the early differentiation of naive CD4 T helper (Th) cells into Th1 versus Th2 states by counting single transcripts in individual cells. Contrary to the current dogma of mutually exclusive expression of antagonistic transcription factors, we observed their ubiquitous co-expression in individual cells, at high levels that are distinct from basal level co-expression during lineage priming(Arinobu et al., 2007; Rothenberg, 2007). The expression of these transcription factors can be gradually tuned by extracellular cytokines, which are produced stochastically by a small subpopulation of cells. Upon inhibition of cytokine signaling, we observed the classic mutual exclusion of antagonistic transcription factors, thus revealing a weak intracellular network otherwise overruled by the strong signals that emanate from extracellular cytokines. These results suggest that during the early differentiation process CD4 T cells stochastically acquire a mixed Th1/Th2 state, biased by extracellular cytokines.

3.2 Introduction

A multipotent progenitor cell can differentiate into a particular lineage by turning on the expression of a lineage-specific transcription factor, which coordinates the expression of a defined set of target genes. Numerous examples of such toggle switch-like cell fate decisions have been observed in the differentiation of hematopoietic cells(Rothenberg, 2007). For example, common myeloid progenitor cells differentiate into granulocyte-monocyte progenitor versus megakaryocyte-erythrocyte progenitor cells based on expression of PU.1 versus Gata1(Arinobu et al., 2007); naive CD4 T cells differentiate into Th1 versus Th2 driven by the expression of Tbet or Gata3(Ouyang et al., 1998; Szabo et al., 2000; Szabo et al., 2003; Zheng and Flavell, 1997). Antagonistic transcription factors are therefore believed to be expressed exclusively in the pertinent cell types, or co-expressed at basal levels in hematopoietic progenitors prior to commitment to "prime" the cells for rapid deployment of transcription factors to execute a particular lineage program(Laiosa et al., 2006). For instance, common myeloid progenitors can co-express low levels of PU1 and GATA1 during lineage priming(Hu et al., 1997), though their expression is mutually exclusive in fully committed state(Laiosa et al., 2006). In the previous studies, high concentrations of cytokines were added to the culture media to bias the cellular decision process towards one particular cell fate(Arinobu et al., 2007; Ouyang et al., 1998; Shaffer et al., 2002; Szabo et al., 2000). To study the plasticity of the early Th1/Th2 decision, we sought to avoid this bias by exploring the spontaneous differentiation of naive CD4 T cells in the absence of exogenously added cytokines.

Tbet, encoded by *Tbx21*, is the master transcription factor of Th1 differentiation associated with production of the hallmark cytokine IFNγ(Szabo et al., 2000), whereas Gata3 is the master transcription factor of Th2 differentiation associated with IL4 production(Zheng and Flavell, 1997). In terminally differentiated individual CD4 T cells, the expression of *Tbx21* and *Gata3* is mutually exclusive(Murphy and Reiner, 2002; Zhou et al., 2009). This is usually attributed to positive feedback loops and cross-inhibitory interactions in the regulatory network (Fig. 3.3.1a). This network consists of two types of interactions: those that depend on cytokine signaling and those that are

cytokine-independent and involve only intracellular players including transcription factors. Specifically, Tbet activates *Ifng*(Djuretic et al., 2007), and extracellular IFNγ can induces *Tbx21* via receptor signaling(Leonard and O'Shea, 1998). Tbet also induces itself independently of signaling via cytokine receptors(Mullen et al., 2002). Similarly, Gata3 activates *Il4*(Jenner et al., 2009; Tykocinski et al., 2005) and extracellular IL4 can induce *Gata3*(Takeda et al., 1996). Furthermore, *Gata3* can be autoinduced independently of signaling via cytokine receptors(Jenner et al., 2009; Ouyang et al., 2000). Finally, Tbet silences *Il4*(Djuretic et al., 2007), Gata3 silences *Ifng*(Chang and Aune, 2007; Schoenborn et al., 2007), and Tbet blocks the function of Gata3 through direct protein-protein interactions(Hwang et al., 2005), leading to cross-inhibitory interactions.

To quantify the number of *Tbx21* and *Gata3* transcripts in activated CD4 T cells, we isolated total $CD4^+$ cells from C57BL/6 mice. CD4 cells were then activated by culturing them in wells coated with anti-CD3 and anti-CD28 antibodies, in the absence of polarizing cytokines or neutralizing antibodies against cytokines, such that CD4 T cells would choose their cell fates without being biased. We performed single-molecule fluorescent *in situ* hybridization (smFISH)(Raj et al., 2008) combined with immunofluorescence to quantify transcripts and protein levels in individual cells (Supplementary Fig. 3.3.1-2).

3.3 Results and Discussions

Without artificially imposed Th1- or Th2-biasing cues, naive CD4 T cells, essentially expressing zero copies of *Tbx21* and *Gata3* transcripts, turned on expression of both *Tbx21* and *Gata3* simultaneously in individual cells, not in a mutually exclusive fashion as current models would predict (Fig. 3.3.1b-d). Distinct from basal co-expression in lineage priming, co-expression of *Tbx21* and *Gata3* are at high levels, such that the mean number of *Gata3* transcript per cell at 48 h is comparable to fully differentiated Th2 cells(Hebenstreit et al.). In addition, the expression levels of *Tbx21* and *Gata3* under non-biased condition are comparable to that treated with polarizing conditions as previously described(Djuretic et al., 2007). High-level co-expression of *Tbx21* and *Gata3* in individual cells is a robust phenomenon observed over a large range of seeding cell density (Supplementary Fig. 3.3.3). Mutant cells that lack a functional *Il4* or *Ifng* gene and therefore exclusively differentiate towards Th1 or Th2 fate respectively(Dalton et al., 1993; Kuhn et al., 1991), display a very different behavior. *Tbx21* and *Gata3* are expressed in a mutually exclusive manner (Fig. 3.3.1e,f, Supplementary Fig. 3.3.4). Importantly the expression levels are similar to wild-type cells (Fig. 3.3.1c). Interestingly, the median stoichiometry between *Tbx21* and *Gata3* expression was 1:1 until 24 h after activation, but *Gata3* levels continued to increase after 24 h while *Tbx21* levels decreased (Supplementary Fig. 3.3.5). As activation time increases, the culture system presumably accumulates more Th2-favoring cytokines. Since most of the significant changes in gene expression occurred within this 48 h period, we focused our analysis on this window in subsequent experiments.

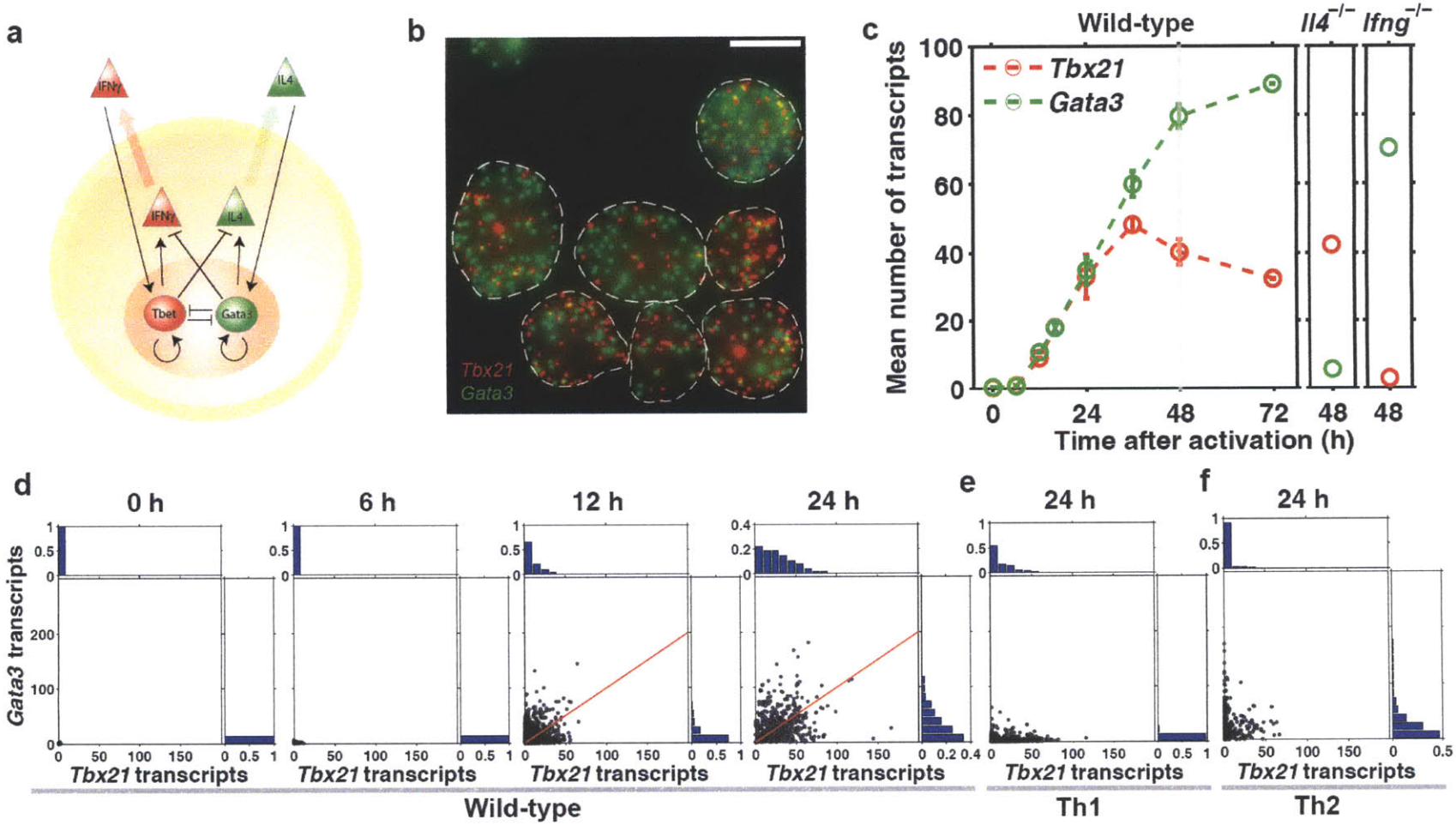


Figure 3.1 *Tbx21* and *Gata3* are transcribed simultaneously in individual CD4 T cells.
(a) Current gene regulatory network proposed to govern Th1/Th2 lineage specification. (b) Visualization of single transcripts of *Tbx21* (red) and *Gata3* (green) in individual CD4 T cells 24 h after activation. White dashed lines are boundaries of individual cells. Scale bar is 10 μm. (c) Mean counts of *Tbx21* and *Gata3* transcripts per cell as a function of activation time. (d) Scatter plots of *Tbx21* and *Gata3* transcripts in individual cells, with marginal distributions. The red line is the median line that divides data points into halves. Individual cells do not show mutual exclusion of *Tbx21* and *Gata3* expression. (e, f) Scatter plots of *Tbx21* and *Gata3* transcripts in CD4 T cells treated with Th1-polarizing (e) and Th2-polarizing (f) conditions at 24 h. Error bars are s.e.m. of replicate experiments.

To demonstrate that transcript counts serve as a good proxy for protein levels, we performed immunofluorescence against Tbet or Gata3 simultaneously with smFISH. Transcript counts and protein levels showed strong correlations in individual CD4 T cells, with a Pearson's correlation coefficient R of 0.59 ($p < 10^{-44}$) for Tbet and 0.85 ($p < 10^{-84}$) for Gata3 (Supplementary Fig. 3.3.6). In addition, translational efficiency, measured by the ratio of immunofluorescence intensity over transcript count, remained constant as a function of activation time (Supplementary Fig. 3.3.7).

We postulated that ubiquitous *Tbx21* and *Gata3* co-expression must be associated with both Th1 and Th2 cytokines produced by CD4 T cells upon activation(Schmitz et al., 1994), since no cytokines were supplied exogenously. We thus investigated the expression of *Ifng* and *Il4* in individual CD4 T cells. Current understanding of the gene regulatory network that governs Th1/Th2 lineage specification would predict that *Ifng* or *Il4* transcripts would be proportional to *Tbx21* or *Gata3* levels in individual cells. Surprisingly, we observed that *Ifng* and *Il4* were expressed only in a rare cell population. While the vast majority of cells were in the OFF state and contained essentially zero copies of *Ifng* or *Il4* transcripts, the rare ON cells expressed up to more than 1000 transcripts (Fig. 3.3.2a,b, Supplementary Fig. 3.3.8). In cells expressing more than 200 transcripts, we could not resolve individual mRNA molecules. Instead, we extrapolated the number of transcripts from the linear relationship between the total fluorescence and number of transcripts using cells with fewer transcripts (Supplementary Fig. 3.3.9). There was a weak positive correlation between *Tbx21* and *Ifng* expression (R = 0.15, $p < 10^{-6}$), or between *Gata3* and *Il4* expression (R = 0.32, $p < 10^{-11}$) (Fig. 3.3.2c). There was no negative correlation between *Gata3* and *Ifng* expression (R = 0.06, $p = 0.04$), or between *Tbx21* and *Il4* expression (R = 0.26, $p < 10^{-9}$) (Supplementary Fig. 3.3.10). In addition, cellular *Tbx21* and *Gata3* levels do not depend on the distance from cytokine producing cells (Fig. 3.3.2a, Supplementary Fig. 3.3.11), indicating that diffusion of cytokine is not rate-limiting and results in a well-mixed milieu. Cells that express high number of cytokine transcripts also contained high levels of cytokine protein as detected by immunofluorescence. Transcriptionally inactive cells did not contain measurable levels of cytokine protein (Supplementary Fig. 3.3.12).

To ensure that the rare cytokine producing cells were not non-naive $CD4^+$ T cells, such as Natural Killer T (NKT) cells, we analyzed the expression of *Klrb1c*, which encodes the NKT cell marker NK1.1, and did not observe any NK1.1-expressing cells (Supplementary Fig. 3.3.13). To ensure that the $CD4^+$ T cells we isolated did not contain effector memory cells, we analyzed CD44 levels using immunofluorescence. There was no significant positive correlation between cytokine expression and CD44 levels in activated cells (Supplementary Fig. 3.3.14). In addition, CD44 levels in naive T cells were low (Supplementary Fig. 3.3.14).

Taken together, we conclude that a rare naive CD4 T cell population stochastically turns on *Ifng* or *Il4* independently of Tbet or Gata3 levels. These rare cells secrete cytokines into their surroundings and instruct other cells to ubiquitously express *Tbx21* and *Gata3*, and may thus play a pioneer role in determining the differentiation outcome of the entire cell population.

While naive CD4 T cells contain essentially zero copies of cytokine transcripts, the fraction of *Ifng* expressing cells increased from 0 to 16 h and decreased moderately afterwards, whereas the fraction of IL4-producing cells increased monotonously (Fig. 3.3.2d). This pattern is consistent with the trend of the mean *Tbx21* and *Gata3* counts per cell (Fig. 3.3.1c). Coupled with the absence of a correlation between cytokine and transcription factor transcript counts in individual cells, the general trend of an increasing fraction of cytokine producing cells indicates that initial cytokine expression is stochastic in individual cells. It is worth noting that at the population level, the fraction of cells transcribing *Ifng* and *Il4* still positively correlates with the means of *Tbx21* and *Gata3* transcripts over time (correlation coefficient = 0.35, $p = 0.044$ between *Tbx21* and *Ifng*; correlation coefficient = 0.98, $p = 1.6 \times 10^{-4}$ between *Gata3* and *Il4*). Cytokine expression becomes ubiquitous as differentiation proceeds, consistent with gene locus modifications mediated by transcription factors over a longer time scale(Ansel et al., 2006; Hegazy et al.; Hofer et al., 2002; Ouyang et al., 2000).

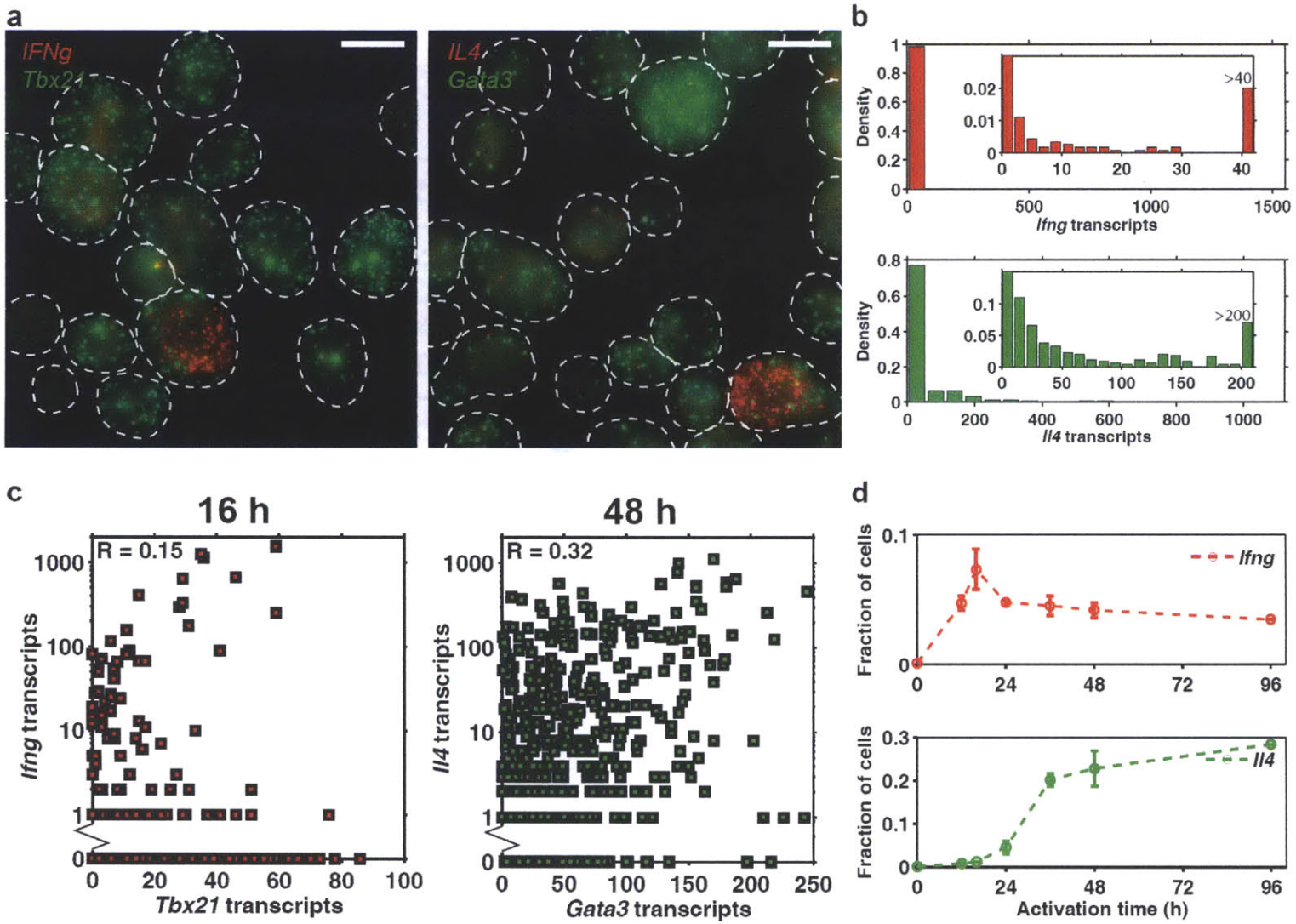


Figure 3.2 *Ifng* and *Il4* are expressed in a rare cell population and their levels show no significant correlation with *Tbx21* and *Gata3* transcripts.

(a) Visualization of single transcripts of *Tbx21* and *Ifng*, and *Gata3* and *Il4* in individual CD4 T cells at 48 h. All scale bars are 10 μm. (b) Distribution of *Ifng* and *Il4* transcripts in individual CD4 T cells, with inset diagrams to better illustrate the fraction of cells that express non-zero copies of cytokines. (c) Scatter plots show a weak positive correlation between *Tbx21* and *Ifng* expression, or between *Gata3* and *Il4* expression. (d) Fraction of cells that express *Ifng* (defined as > 20 transcripts) and that expressing *Il4* (defined as > 50 transcripts) as a function of activation time. Error bars are s.e.m. of replicate experiments.

We then revisited the signaling network governing Th1/Th2 lineage specification during early CD4 T cell differentiation. Given that *Ifng* is stochastically transcribed in a rare pioneer cell population independently of *Tbx21* and *Gata3* levels, induction of *Ifng* by Tbet and repression by Gata3 do not apply to early stages of CD4 T cell differentiation, and a similar situation applies to *Il4*. Since *Tbx21* and *Gata3* are expressed simultaneously in individual cells without mutual exclusion, we postulated that the strength of receptor signaling mediated by cytokines must dominate over the intracellular network, which alone would lead to mutually exclusive expression of *Tbx21* and *Gata3*.

To demonstrate that the strength of cytokine signaling is dominant, we manipulated the amount of cytokine molecules available to cells by adding neutralizing antibodies. In the presence of saturating amounts of anti-IFNγ and anti-IL4 we recapitulated the expression patterns of *Tbx21* and *Gata3* in *Ifng*$^{-/-}$ or *Il4*$^{-/-}$ cells respectively (Fig. 3.3.1e,f), strongly suggesting that this depletion strategy was specific (Supplementary Fig. 3.3.15). Adding an antibody against the Th1 cytokine IL12 had no effect on *Tbx21* or *Gata3* expression (Supplementary Fig. 3.3.16). Downregulation of the appropriate transcription factor could be modulated depending on the amount of neutralizing antibody (Fig. 3.3.3a). To quantitatively interpret data, we converted the *Tbx21-Gata3* scatter plot into polar coordinates of (r, θ) such that a small θ means that a cell is Th1-skewed, and a θ close to $\pi/2$ means Th2-skewed (Fig. 3.3.3b). Converting the data for cells at 24 h in the absence of exogenously added antibodies into polar coordinates shows that θ follows an approximately uniform distribution, a hallmark of lacking mutual exclusion (Fig. 3.3.3c, Supplementary Fig. 3.3.17). In other words, in the absence of any exogenous polarizing cues, CD4 T cells during early differentiation occupy any intermediate cell states between Th1 and Th2 with equal probability. As the concentration of anti-IFNγ increases, the distribution of θ shifts towards $\pi/2$ (more Th2-like), whereas when the concentration of anti-IL4 increases, the distribution of θ shifts towards 0 (more Th1-like) (Fig. 3.3.3d, Supplementary Fig. 3.3.18-19).

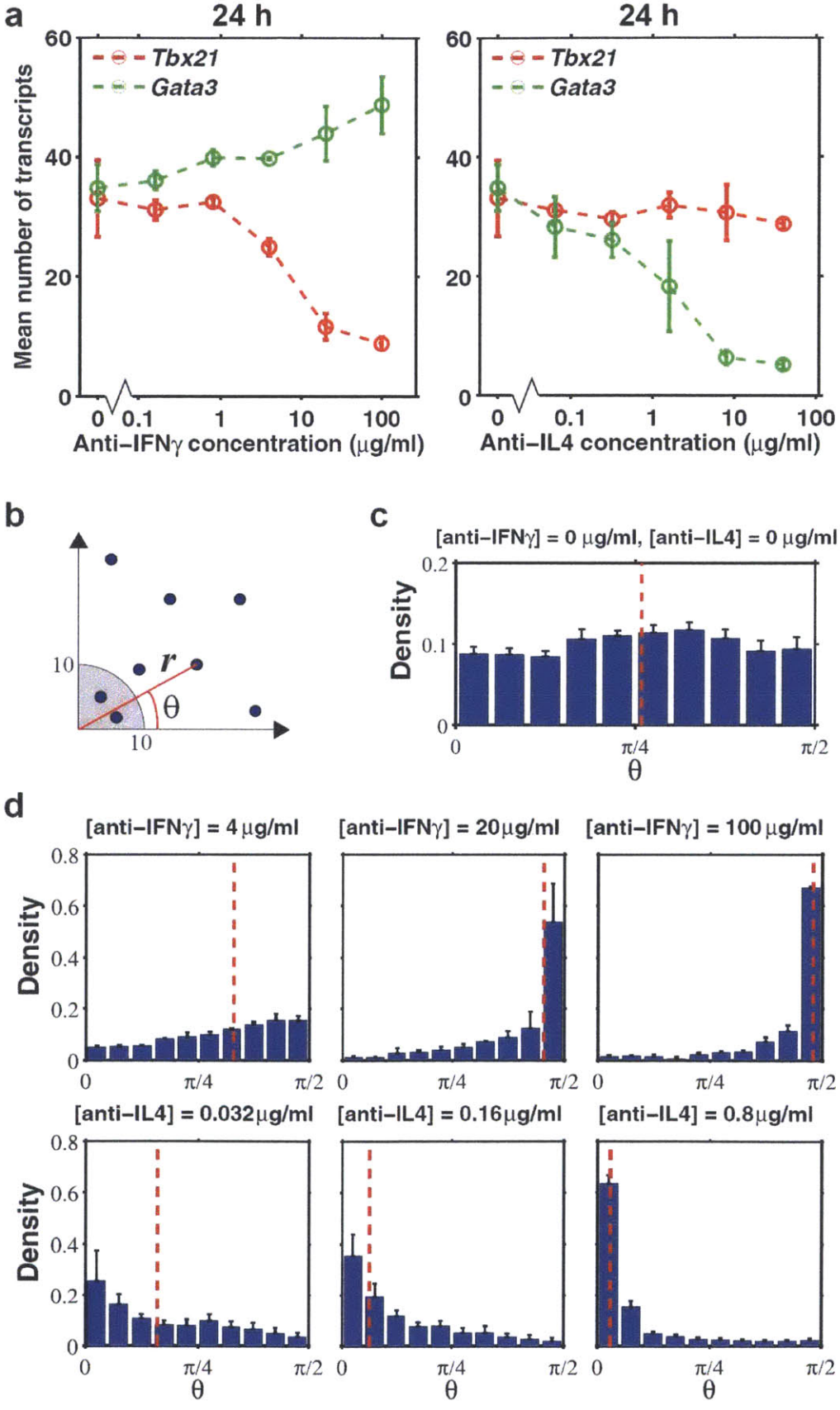

a
24 h
24 h
Tbx21
Gata3
Mean number of transcripts
60
40
20
0
0 0.1 1 10 100
Anti–IFNγ concentration (μg/ml)
Anti–IL4 concentration (μg/ml)
b
r
θ
10
10
c
[anti–IFNγ] = 0 μg/ml, [anti–IL4] = 0 μg/ml
Density
0.2
0.1
0
0 π/4 π/2
θ
d
[anti–IFNγ] = 4 μg/ml
[anti–IFNγ] = 20 μg/ml
[anti–IFNγ] = 100 μg/ml
[anti–IL4] = 0.032 μg/ml
[anti–IL4] = 0.16 μg/ml
[anti–IL4] = 0.8 μg/ml
Density
0.8
0.6
0.4
0.2
0
0 π/4 π/2
θ

Figure 3.3 Depriving cells of IFNγ and IL4 downregulates *Tbx21* and *Gata3* respectively. (a) As the concentration of anti-IFNγ antibody increases, the mean number of *Tbx21* transcripts per cell decreases, while that of *Gata3* transcripts remains constant. The reverse is observed upon addition of anti-IL4 antibody. (b) Conversion of *Tbx21-Gata3* scatter plot into polar coordinates. Small θ: Th1-like state; large θ: Th2-like state. For subsequent analysis, we excluded cells with $r < 10$ (shaded region), because θ is not robust against small fluctuations in the number of transcripts in these cells. (c) Distribution of θ for cells treated with no cytokine-neutralizing antibodies is uniform, using the same data as Fig. 3.3.1c. (d) Distribution of θ indicates that as concentration of anti-IFNγ antibody increases, the cells adopt larger θ (Th2-like state). The reverse is observed upon addition of anti-IL4 antibody. Red lines show the medians of the θ. All data shown are from cells at 24 h. Error bars are s.e.m. of replicate experiments.

Our results suggest that the role of extracellular cytokine signaling in specifying lineage choice is to upregulate the corresponding transcription factor, rather than to repress that of the alternate lineage. We can then explain the ubiquitous co-expression of *Tbx21* and *Gata3*: when CD4 T cells are exposed to both IFNγ and IL4 secreted by the rare cytokine producing cells, they upregulate both *Tbx21* and *Gata3*. The key to the absence of mutually exclusive expression of *Tbx21* and *Gata3* is that cytokine signaling must predominate over the self-activation of Tbet and Gata3 as well as mutual repression between Tbet and Gata3. This suggests that expression of *Tbx21* and *Gata3* is maintained at high levels by extracellular cytokine cues, with comparatively minimal effects from the intracellular signaling network (Fig. 3.3.4a). Therefore, our model of early CD4 T cell fate specification proposes that CD4 T cells are bathed in a cocktail of well-mixed cytokine molecules produced by the rare pioneer cells, thus simultaneously inducing the expression of *Tbx21* and *Gata3* in individual CD4 T cells ubiquitously (Fig. 3.3.4b).

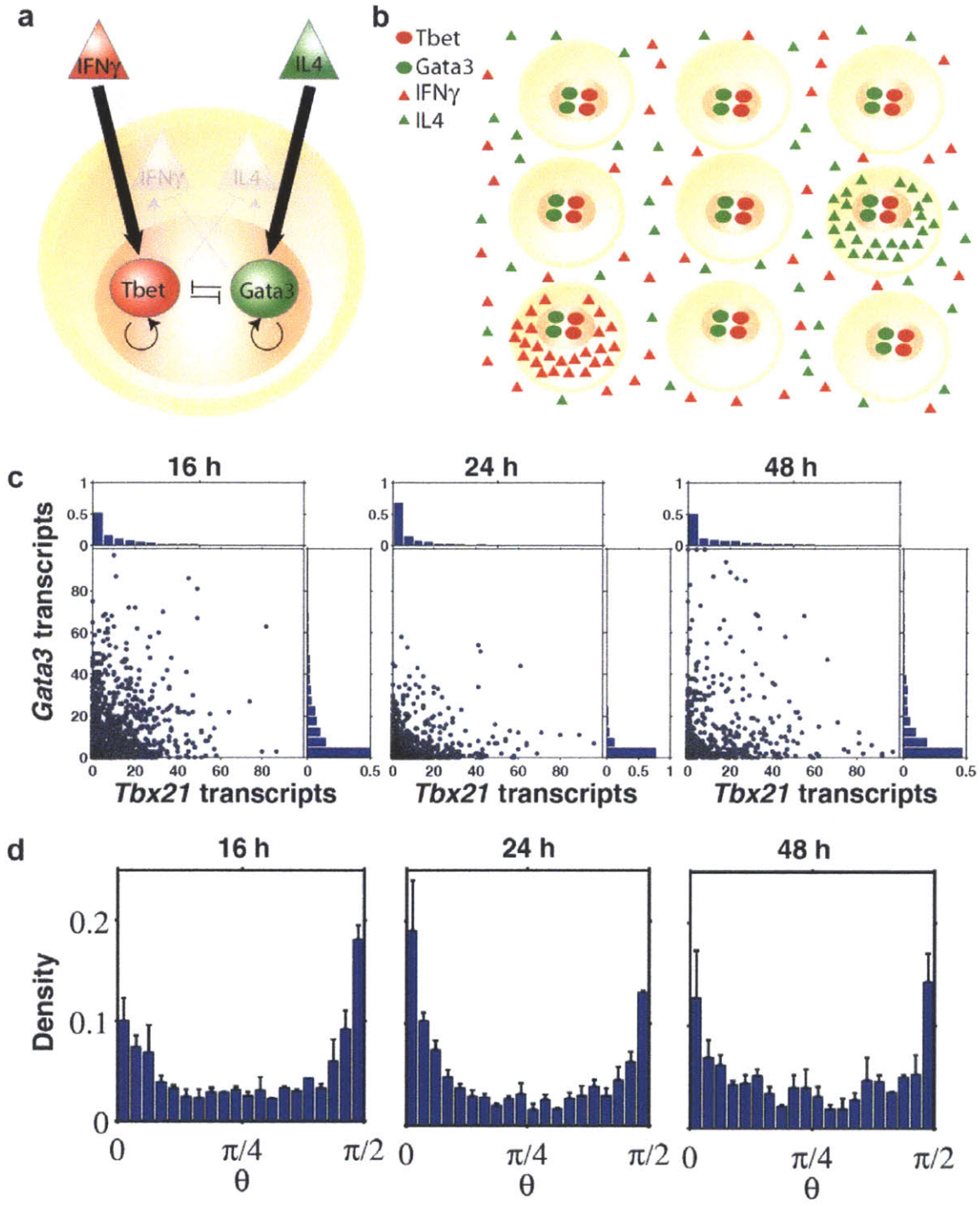


Figure 3.4 Elimination of IFNγ and IL4 leads to mutually exclusive expression of *Tbx21* and *Gata3*.

(a) Our model of the signaling network that governs Th1/Th2 lineage specification. The thickness of arrows indicates the strength of interaction. The intracellular signaling

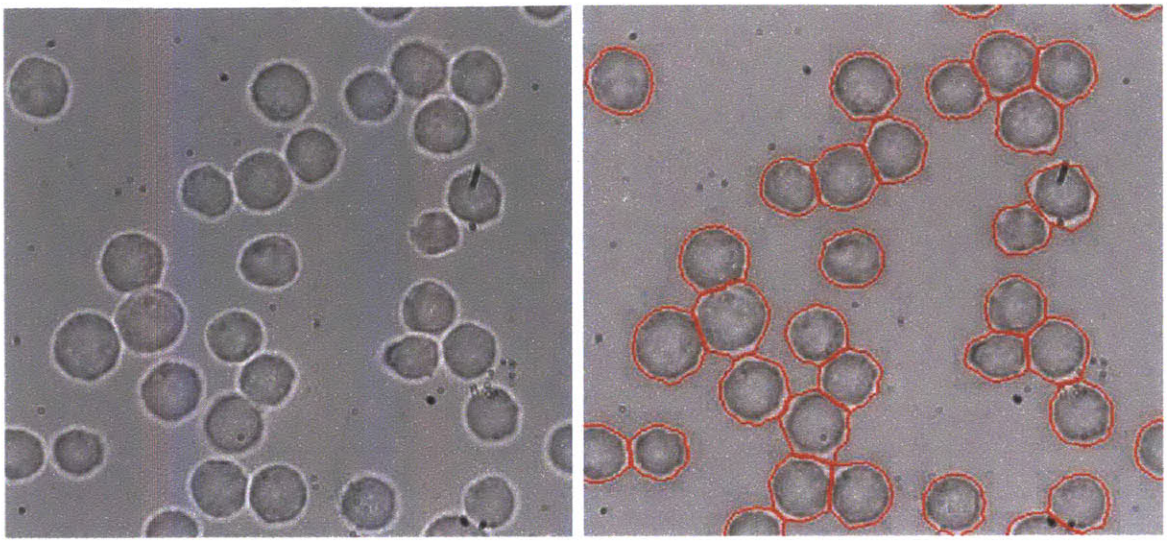
network consists of all the interactions depicted in thin arrows. (b) Illustration of the CD4 T cell population during early activation. (c) Scatter plots showing downregulation and mutual exclusion of *Tbx21* and *Gata3* transcripts in individual cells treated with both anti-IFNγ and anti-IL4 antibodies. (d) Distribution of θ shows that vast majority of cells adopt either very large or small θ (same data as in Fig. 3.3.4a). By two-sample Kolmogorov-Smirnov goodness-of-fit test, distribution of θ for cells under IFNγ and IL4 deprivation are significantly different from untreated cells, $p < 10^{-11}$ at 16 h, $p < 10^{-19}$ at 24 h, $p < 10^{-54}$ at 48 h. Error bars are s.e.m. of replicate experiments.

According to our model, we hypothesized that elimination of extracellular IFNγ and IL4 would leave only the intracellular signaling networks intact and should result in mutually exclusive expression of *Tbx21* and *Gata3* in individual cells. To verify this, we added both anti-IFNγ and anti-IL4. We tested multiple combinations of different concentrations of anti-IFNγ and anti-IL4 antibodies to find an optimum where the median of θ was close to $\pi/4$ (exactly in the middle of Th1 and Th2). Under such conditions, *Tbx21* and *Gata3*, in addition to being downregulated, were expressed in a mutually exclusive manner, such that the majority of cells lay near either along the *Tbx21* or *Gata3* axis on the scatter plot and the distribution of θ has higher density near 0 and $\pi/2$(Fig. 3.3.4c,d). Thus we observed that under IFNγ and IL4 deprivation, only the comparatively weak intracellular signaling components that consist of autoactivation of Tbet and Gata3 as well as their mutual repression are functional, leading to mutually exclusive expression of *Tbx21* and *Gata3*. Interestingly, after 24 hours of activation, cells co-expressing *Tbx21* and *Gata3* can still adopt mutually exclusive expression, if switched to IFNγ and IL4 deprivation (Supplementary Fig. 3.3.20).

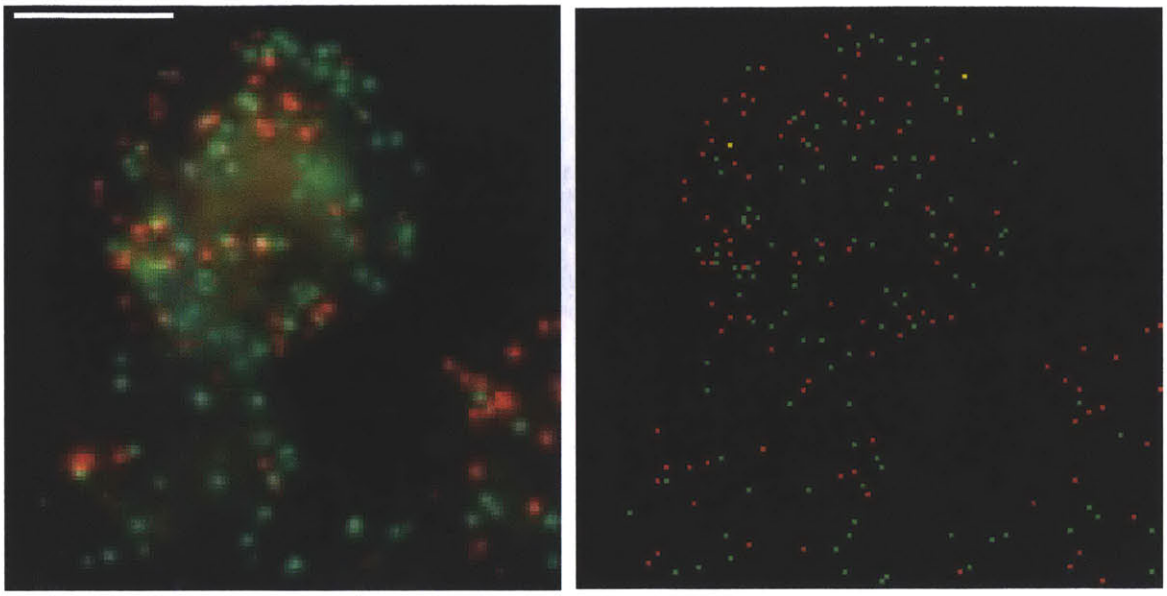
3.4 Conclusion

Using CD4 T cells as a model of cell fate specification, we found ubiquitous co-expression of antagonistic transcription factors during the early stages of CD4 T cell differentiation. Specifically, *Tbx21* and *Gata3* are co-expressed at high levels when both Th1- and Th2-favoring cytokines – IFNγ and IL4 respectively – are available, or mutually exclusively expressed when cells are deprived of both cytokines. Strikingly, activation and cross-inhibition of *Ifng* and *Il4* expression appear to be decoupled from *Tbx21* and *Gata3* levels in individual cells (Fig. 3.3.4a). Instead, *Ifng* and *Il4* are expressed by a rare population, which does not appear to be a contaminating NKT or memory CD4 T cell population. We therefore postulate that these naive CD4 T cells stochastically turned on expression of *Ifng* or *Il4* and translate protein molecules ahead of the bulk population. These cytokine producing cells, though rare, can direct the entire cell population into assuming one particular cell fate. Our data also indicate that signaling strength evoked by extracellular cytokines can override intracellular signaling networks. It would be interesting to explore if these types of stochastic strategies are shared by other cell types *in vitro* and *in vivo*.

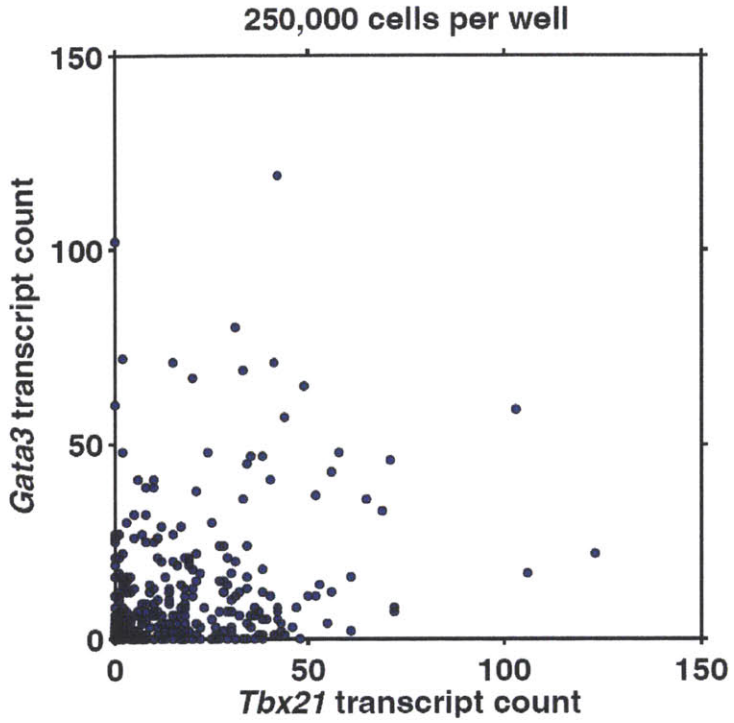
3.5 Supplementary Information



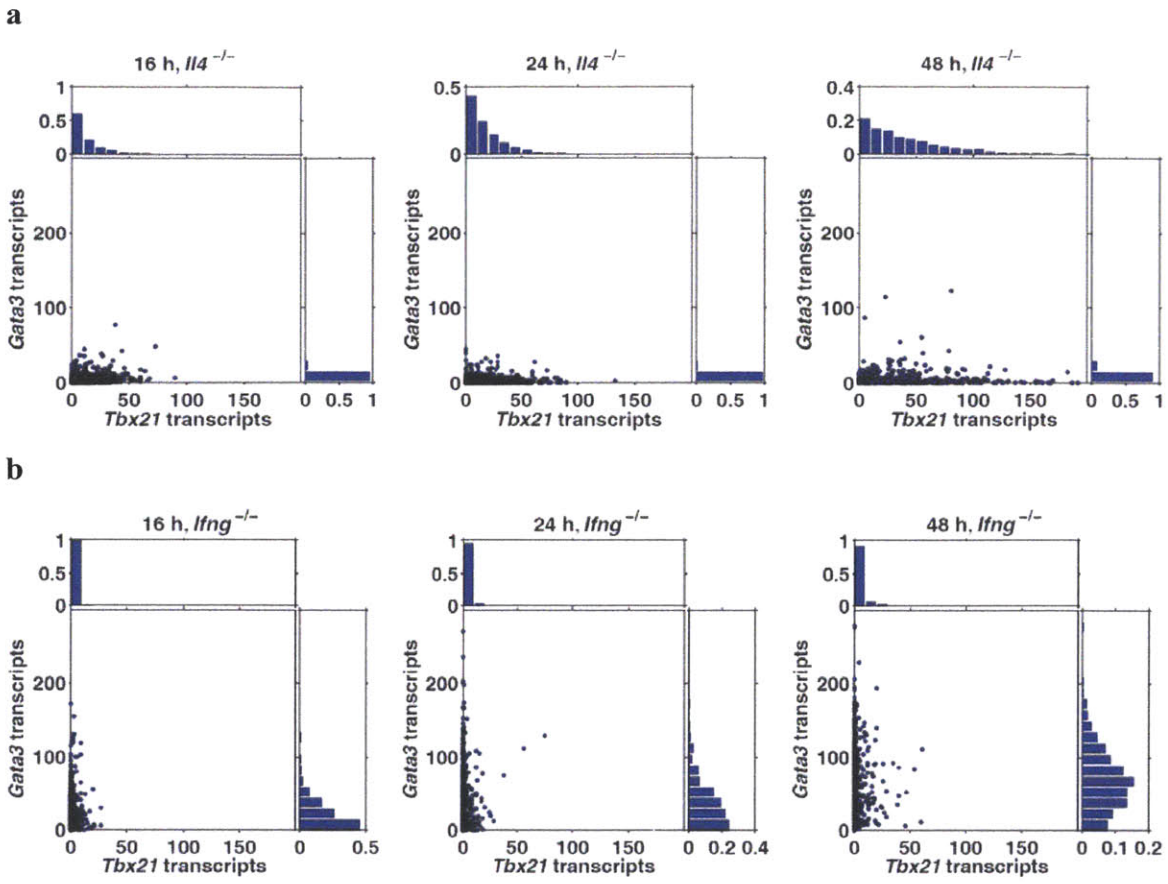
Supplementary Fig. 3.3.1. Segmentation of cells using bright-field images. The left panel is a bright-field image of cultured Th cells. The right panel is the segmented image, using custom software written in MATLAB.



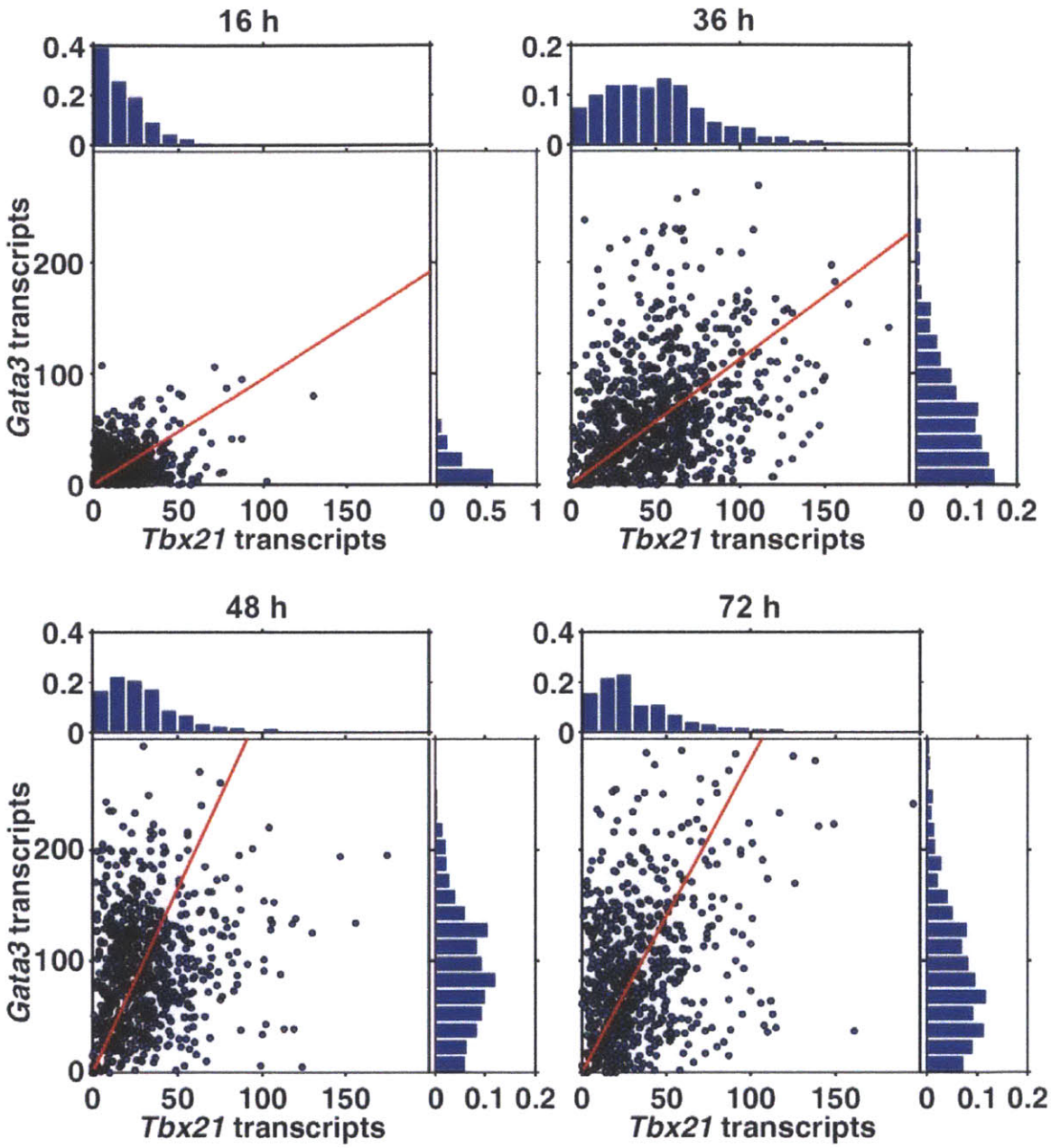
Supplementary Fig. 3.3.2. Image analysis of mRNA spots. The left panel is a fluorescent image showing *Tbx21* (red) and *Gata3* (green) transcripts in Th cells. The right panel is the processed image showing each individual mRNA transcript as a single bright red or green pixel. Scale bar is 10 μm.



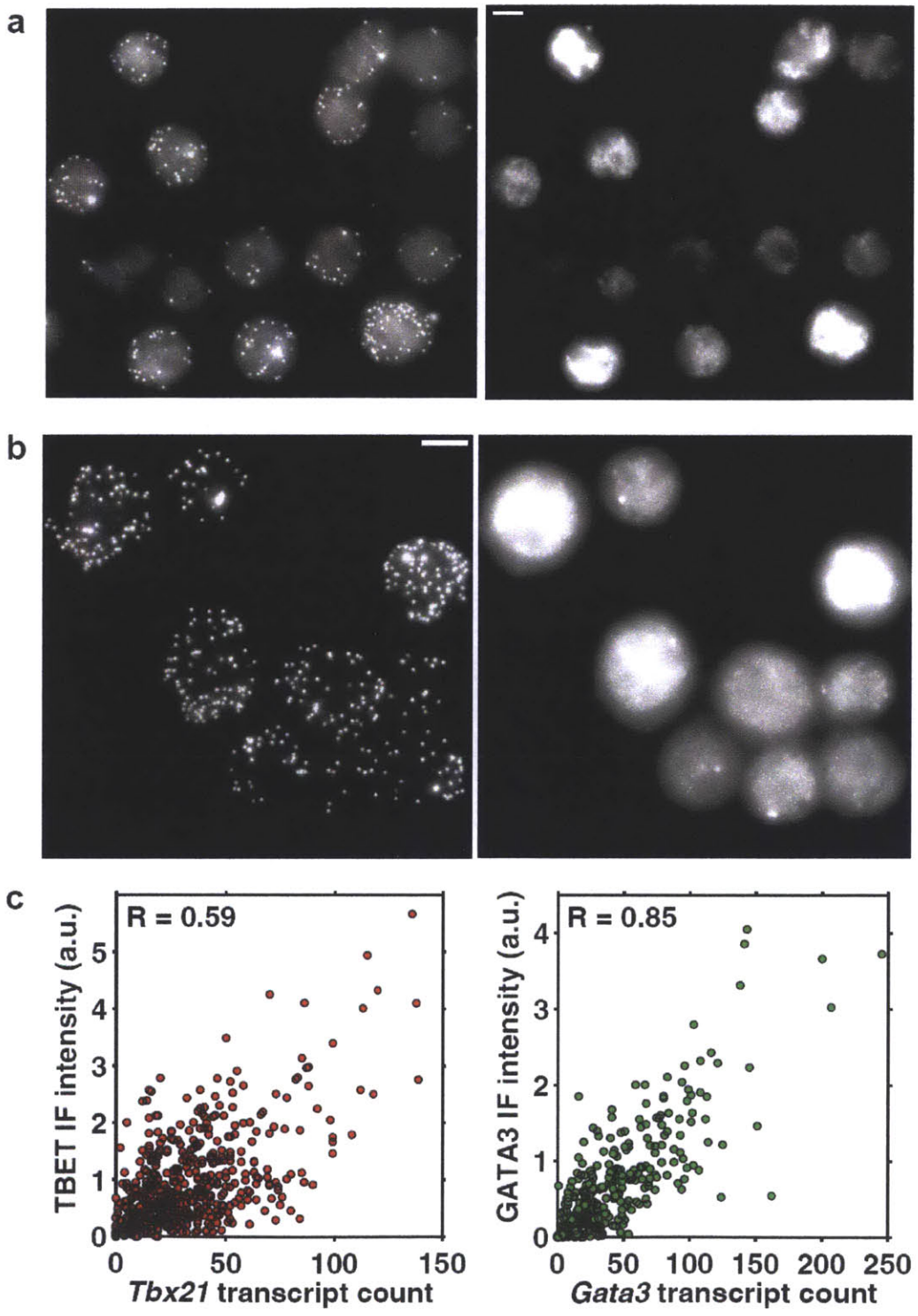
Supplementary Fig. 3.3.3. Scatter plots of *Tbx21* and *Gata3* transcripts in cell cultures of 250,000 cells per well at 24 hours. The cell density in this experiment is 4 times lower than that used in other experiments at 1,000,000 cells per well. It shows that the co-expression of *Tbx21* and *Gata3* transcripts in individual cells is robust over a range of cell densities.



Supplementary Fig. 3.3.4. Scatter plots of *Tbx21* and *Gata3* transcripts in individual cells of *Il4*$^{-/-}$ (**a**) and Ifng$^{-/-}$ (**b**) mice, with marginal distributions at 16 h, 24 h and 48 h. The expression of *Gata3* is downregulated in *Il4*$^{-/-}$ mice. The expression of *Tbx21* is downregulated in *Ifng*$^{-/-}$ mice.

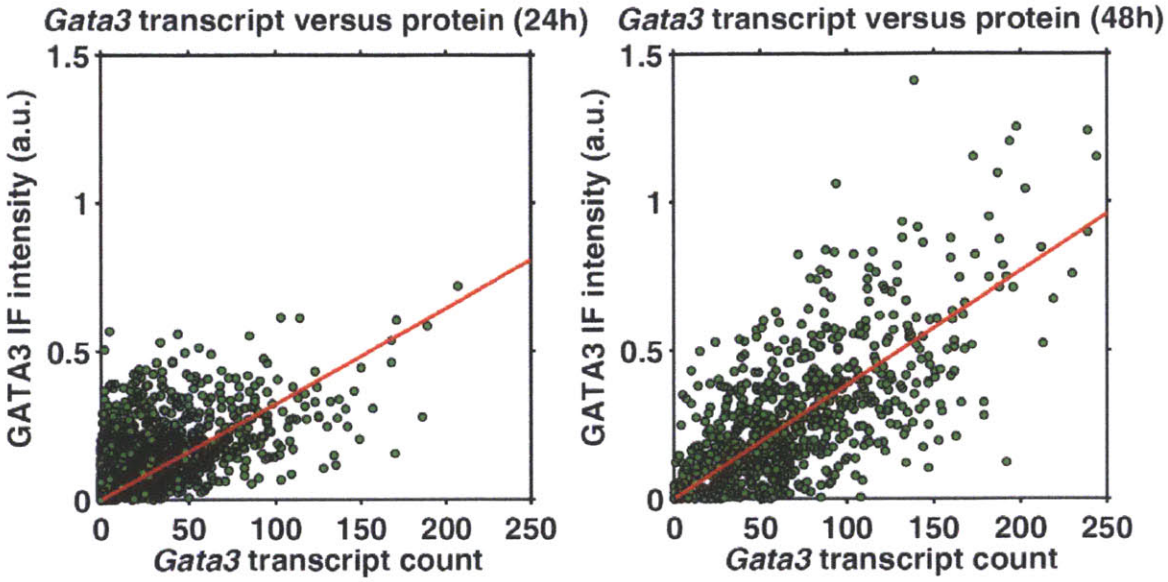


Supplementary Fig. 3.3.5. Scatter plots of *Tbx21* and *Gata3* transcripts in individual cells, with marginal distributions. The red line divides the data set into two equal halves. The data show that no mutual exclusion of *Tbx21* and *Gata3* expression is observed in individual cells. The slope of the red line increases from 24 h to 48 h (compare with **Fig. 3.1d**), indicating the ratio of *Gata3* : *Tbx21* increases from 24 h to 48 h.

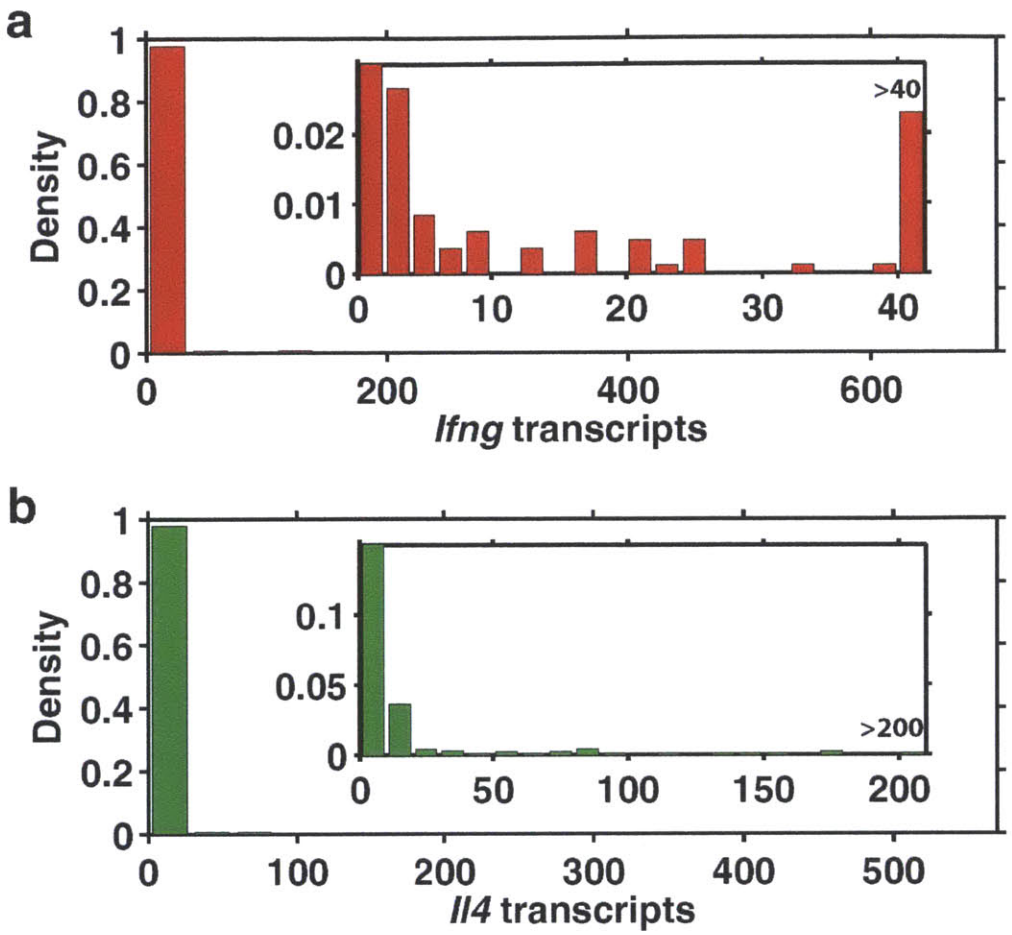


Supplementary Fig. 3.6. Visualization of single *Tbx21* (**a**) and *Gata3* (**b**) transcripts by mRNA-FISH (left of each panel) simultaneously with protein levels by immunofluorescence (right of each panel) in individual Th cells at 24 hours after

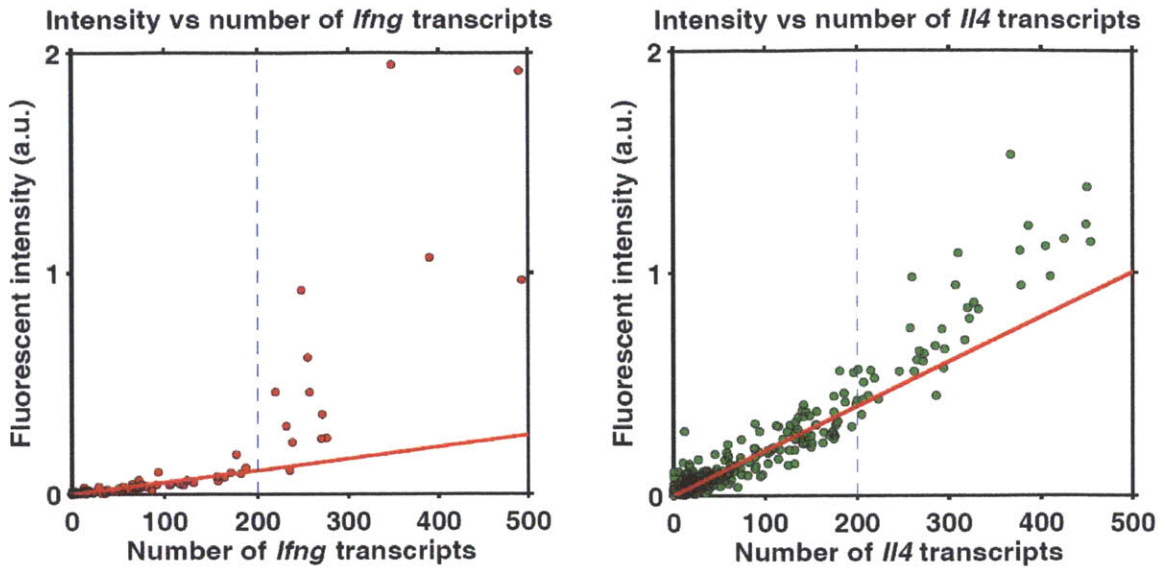
activation. Scale bar is 10 μm. Panel (**c**) is scatter plots showing that transcript counts and protein levels have strong correlations for Tbet and Gata3 in individual Th cells at 24 h, with Pearson's correlation coefficient of 0.59 ($p < 1 \times 10^{-44}$) for Tbet and 0.85 ($p < 1 \times 10^{-84}$) for Gata3.



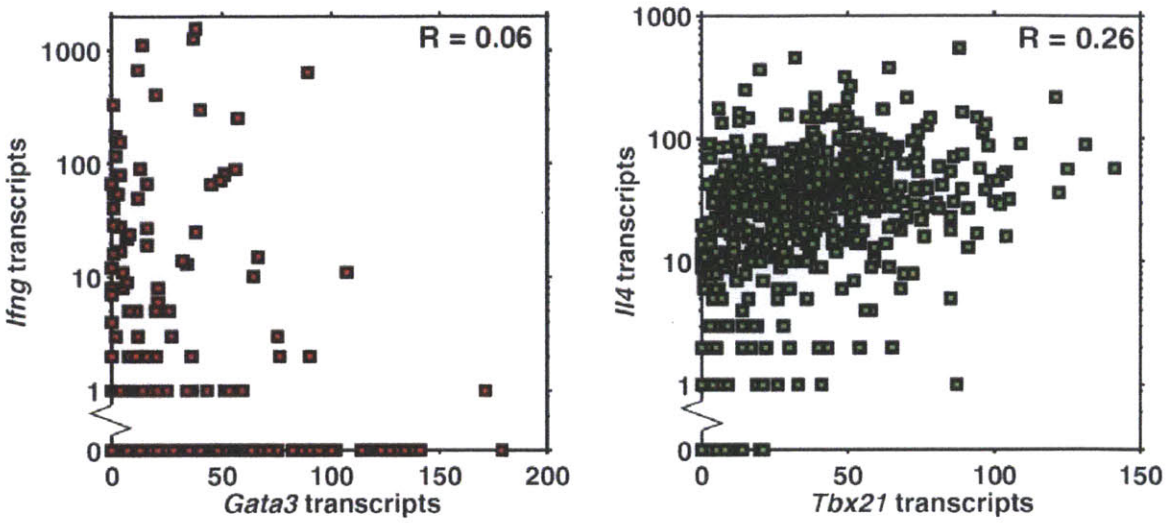
Supplementary Fig. 3.7. GATA3 immunofluorescence intensity versus *Gata3* transcript counts for cells at 24 hours (left) and 48 hours (right) after activation. The red line is the least square fit of the data. The slope of 24-hour data is 0.0032; that of 48-hour data is 0.0038. The two experiments were performed on the same day with the same reagents and same microscope with same exposure time. This result shows that translational efficiency, indicated by the ratio of immunofluorescence intensity over transcript counts, remain constant as a function of activation time.



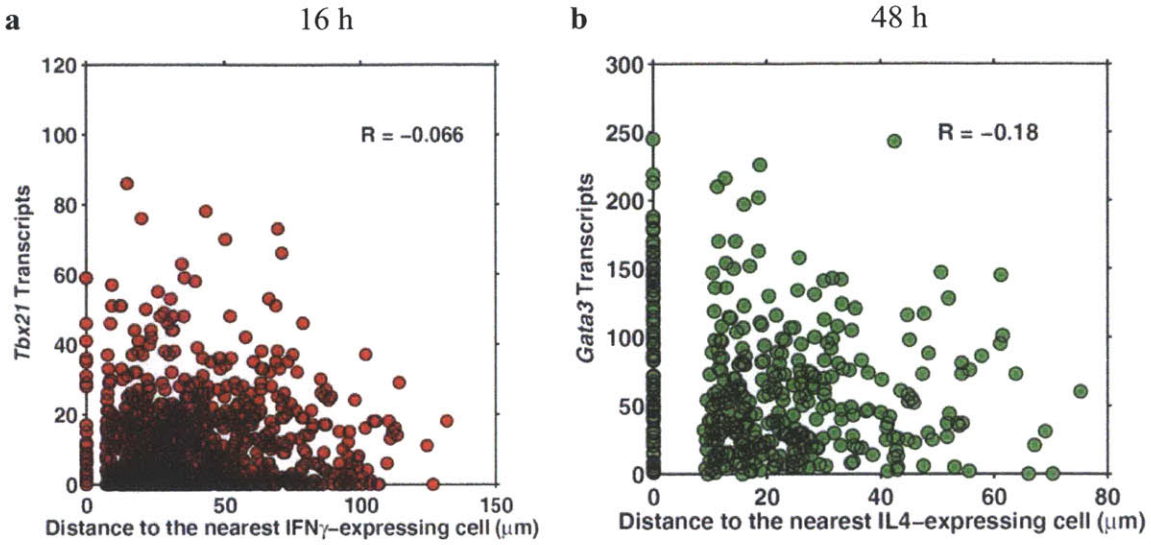
Supplementary Fig. 3.8. Fraction of cytokine-expressing cells at 24 hours, in a control experiment that uses CD4 T cells purified by negative selection (MACS $CD4^+$ T cell isolation kit II), in contrast to CD4 T cells purified by positive selection by $CD4^+$ microbeads used in other experiments. Panel (**a**) shows the probability density of cells expressing *Ifng* transcripts; Panel (**b**) shows the probability density of cells expressing *Il4* transcripts. We have shown that cultures of cells selected by negative selection also give rise to rare cells that stochastically express *Ifng* and *Il4* at high levels. Therefore, rare cytokine-expressing cells observed in the **Fig. 3.2a,b** are not an artifact of positive selection by $CD4^+$ microbeads.



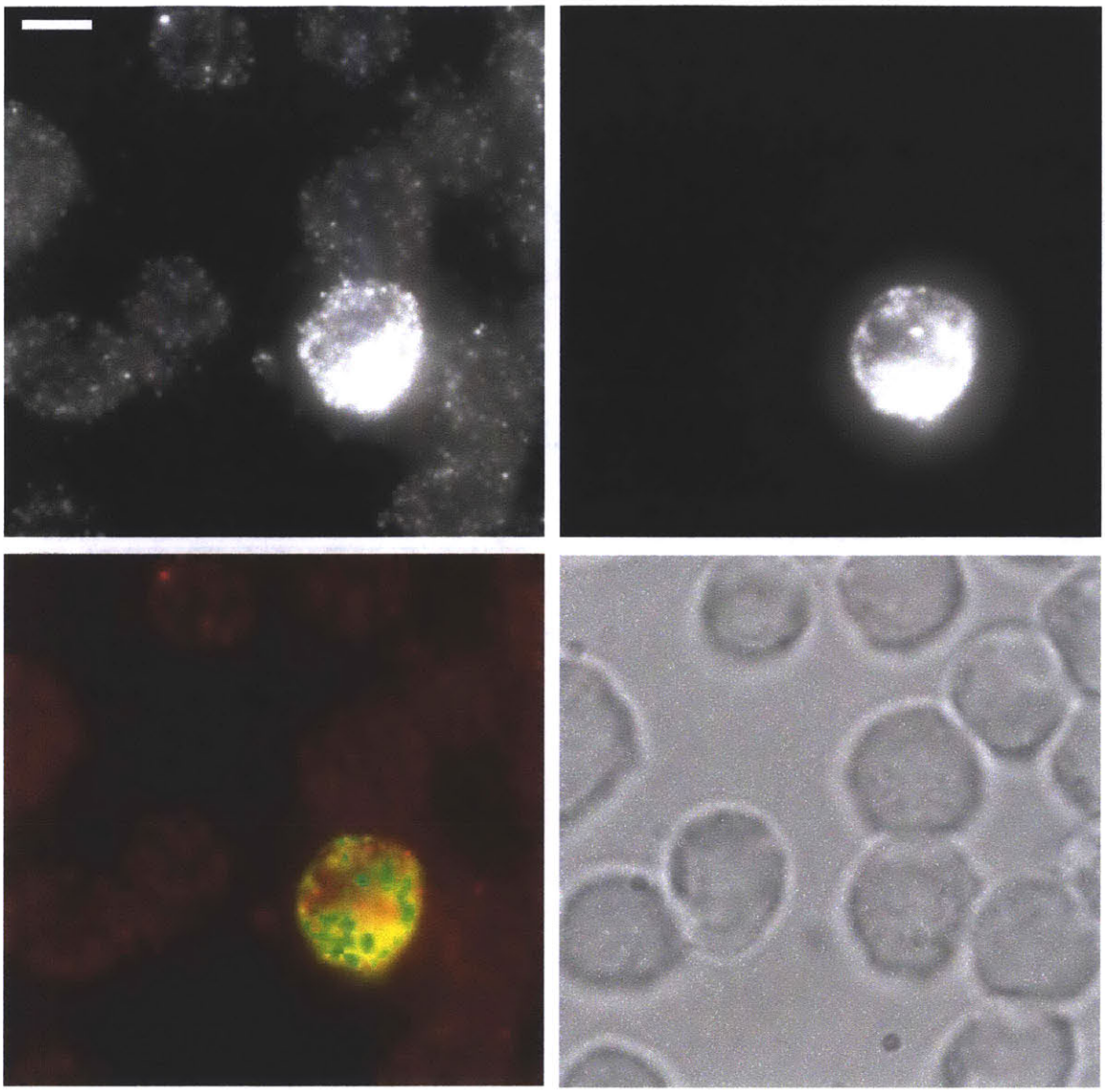
Supplementary Fig. 3.9. Linear relationship exists between total fluorescent intensity of FISH and the computed mRNA transcripts in cells expressing fewer than 200 transcripts. For the *Ifng* plot excluding points with more than 200 computed mRNA transcripts, Pearson's correlation coefficient = 0.86, $p = 5 \times 10^{-24}$; for the *Il4* plot excluding points with more than 200 computed mRNA transcripts, Pearson's correlation coefficient = 0.90, $p = 4 \times 10^{-99}$. We can then extrapolate of the number of transcripts in highly expressing cells using the slope of the linear fit for cells expressing fewer than 200 transcripts.



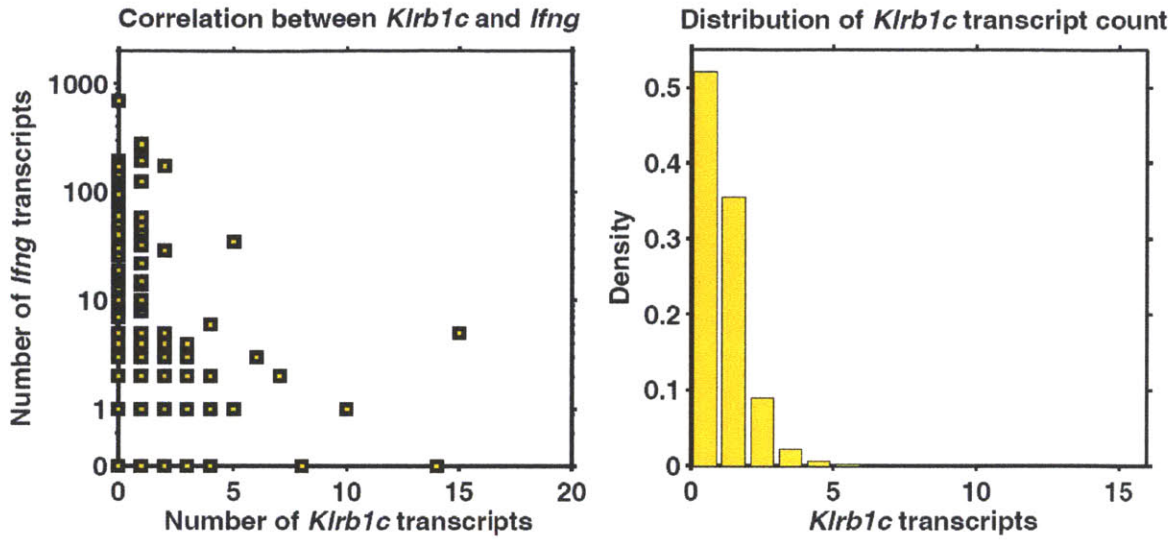
Supplementary Fig. 3.10. Scatter plots showing that there is no negative correlation between *Gata3* and *Ifng* expression, with Pearson's correlation coefficient = 0.06, $p = 0.04$, and that there is no negative correlation between *Tbx21* and *Il4* expression, with Pearson's correlation coefficient = 0.26, $p < 1 \times 10^{-9}$.



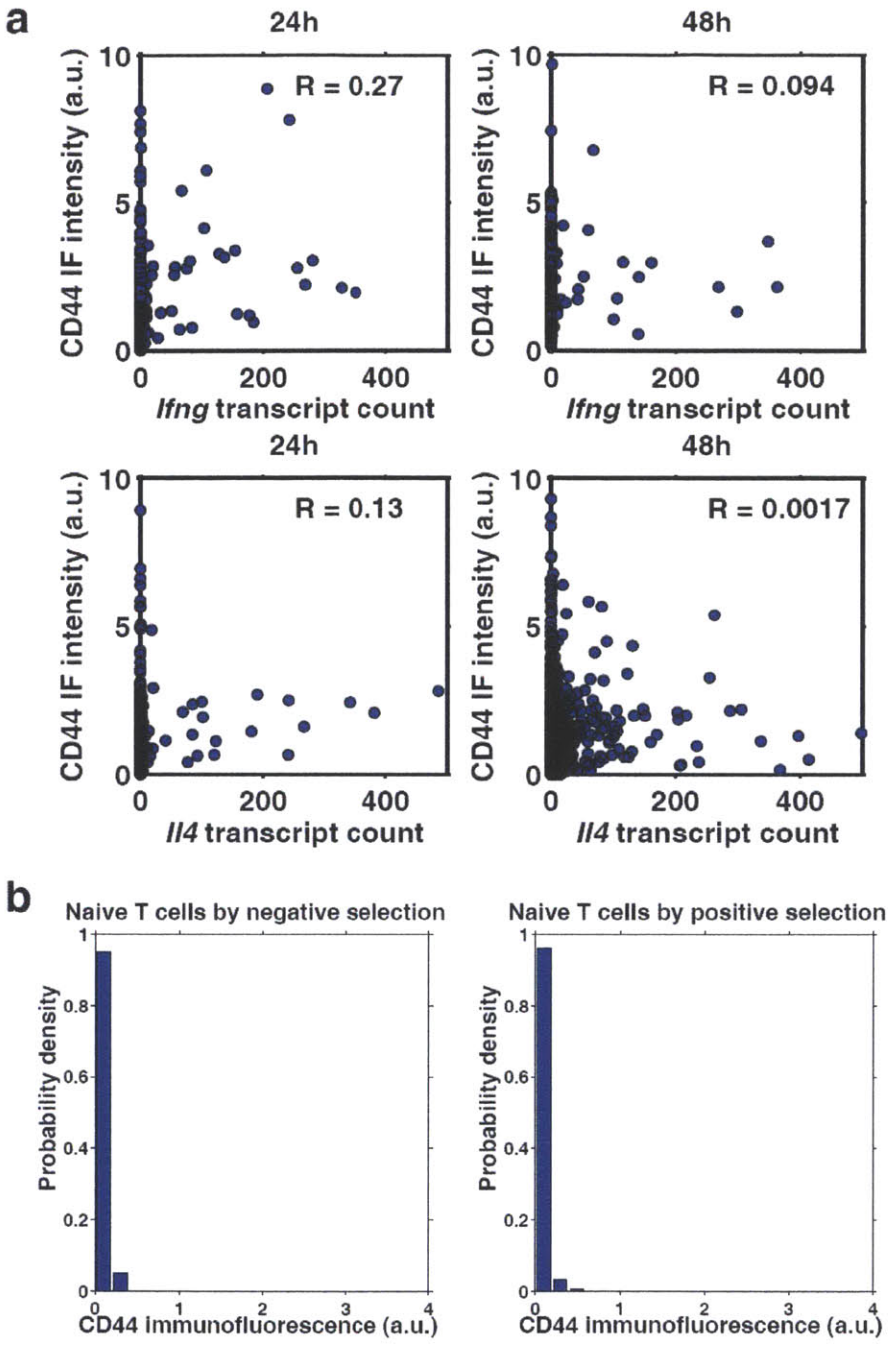
Supplementary Fig. 3.11. The Scatter plot of *Tbx21* (**a**) and *Gata3* (**b**) transcripts in individual cells versus the distance to the nearest *Ifng*-expressing (**a**) or *Il4*-expressing cell (**b**), which is defined as containing more than 20 transcripts of cytokines. The position of each cell is computed as its centroid. It shows that the expression level of *Tbx21* and *Gata3* does not correlate with the distance from the near cytokine-expressing cell. Therefore, diffusion of cytokines from the source cells is not rate limited on the time scale of *Tbx21* and *Gata3* expression. Note that cells at 0 μm for the distance axis are the cytokine-expressing cells. Absence of cells between 0 μm and 7 μm is attributed to the fact that cell diameter is 7 μm, because cells are not overlapping during imaging.



Supplementary Fig. 3.12. Immunofluorescence together with single-molecule FISH on IFNγ shows that only cells expressing *Ifng* transcripts contain IFNγ protein. Cytokine secretion was inhibited for 1 hour to allow cytokine accumulation in these cells before harvesting. The top left panel is immunofluorescence image; the top right panel is single-molecule FISH image; the bottom left panel is the merge of immunofluorescence and single-molecule FISH; the bottom right panel is the bright field image. Scale bar is 10 μm.

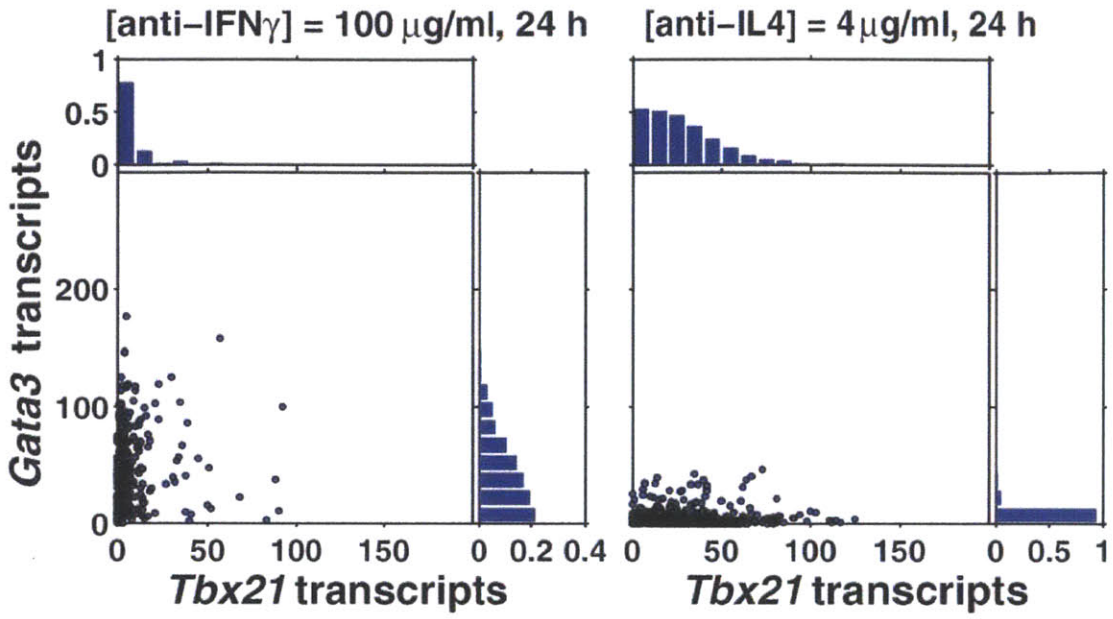


Supplementary Fig. 3.13. The left panel is the scatter plot of *Ifng* and *Klrb1c* transcripts showing that there is no significant positive correlation between *Ifng* and *Klrb1c*, Pearson's correlation coefficient = 0.095, $p = 0.001$, at 16 hours after activation; the right panel shows the distribution of *Klrb1c* transcripts, indicating that *Klrb1c* expression is essentially OFF in all cells. Because *Klrb1c* encodes the marker NK1.1 for NKT cells, the cells expressing *Ifng* are not NKT cells that are not removed during magnetic sorting.

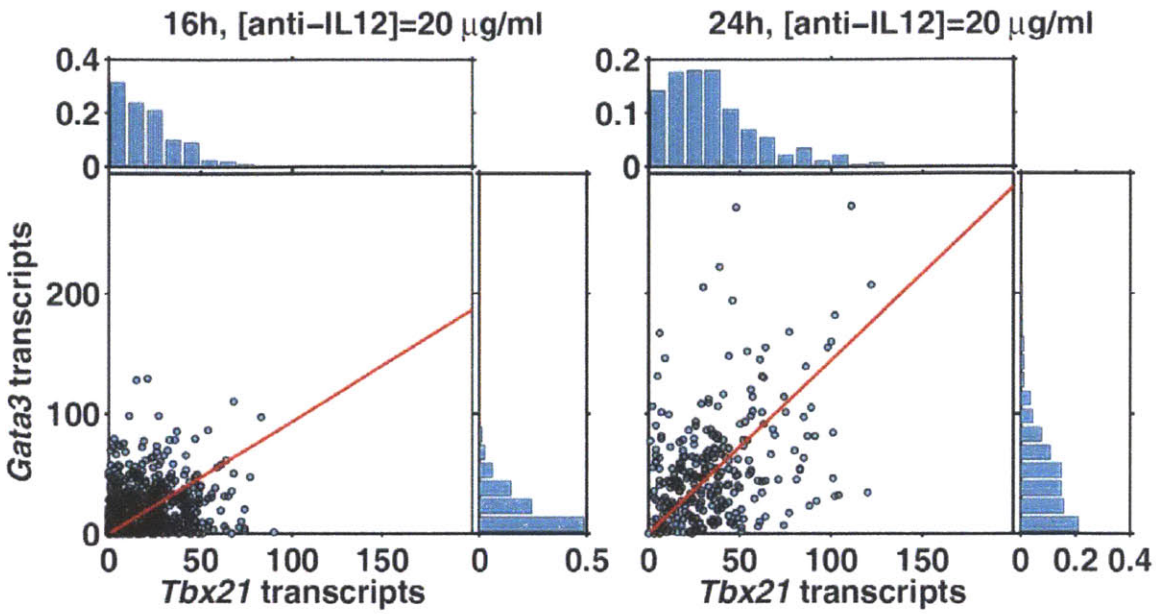


Supplementary Fig. 3.14. Cytokine-expressing cells are not memory T cells. (**a**) Scatter plot of CD44 immunofluoscence versus the number of *Ifng* or *Il4* transcripts shows that there is no significant positive correlation between CD44 levels and *Ifng* (correlation coefficient = 0.27, $p = 4.9 \times 10^{-15}$ at 24 hours; correlation coefficient = 0.094, $p = 0.054$ at 48 hours) or *Il4* expression (correlation coefficient = 0.13, $p = 1.4 \times 10^{-4}$ at 24 hours; correlation coefficient = 0.0017, $p = 0.96$ at 48 hours). *Cd44* is a marker of memory T

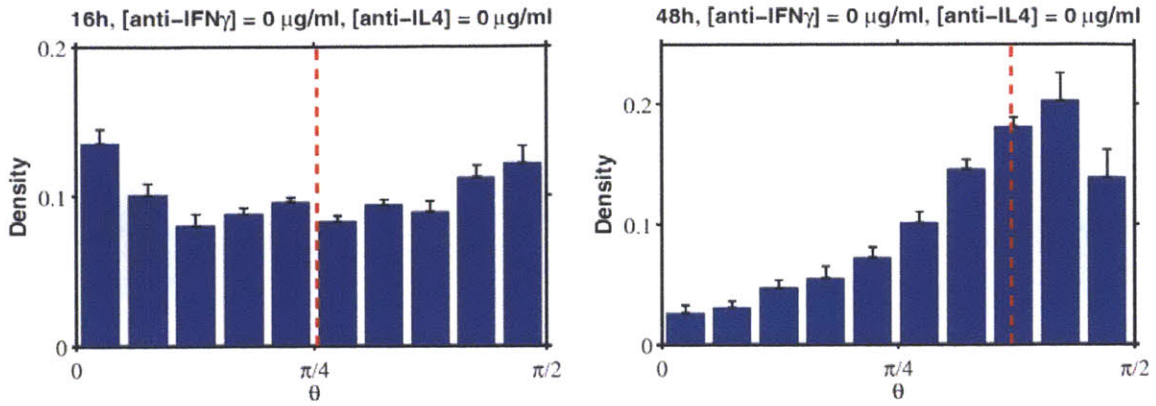
cells. Because cytokine-expressing cells do not preferentially express high levels of *Cd44* transcripts, they are not contaminating memory T cells that are not removed during magnetic sorting. (**b**) Probability density plot of CD44 immunofluorescence of naive T cells isolated by positive selection ($CD4^+$ microbeads) or depletion (MACS $CD4^+$ T cell isolation kit II). It shows that T cells isolated by positive selection, as used ubiquitously in this paper, are similar to T cells isolated by depletion, have low CD44 levels, and do not contain memory cells that are $CD44^+$.



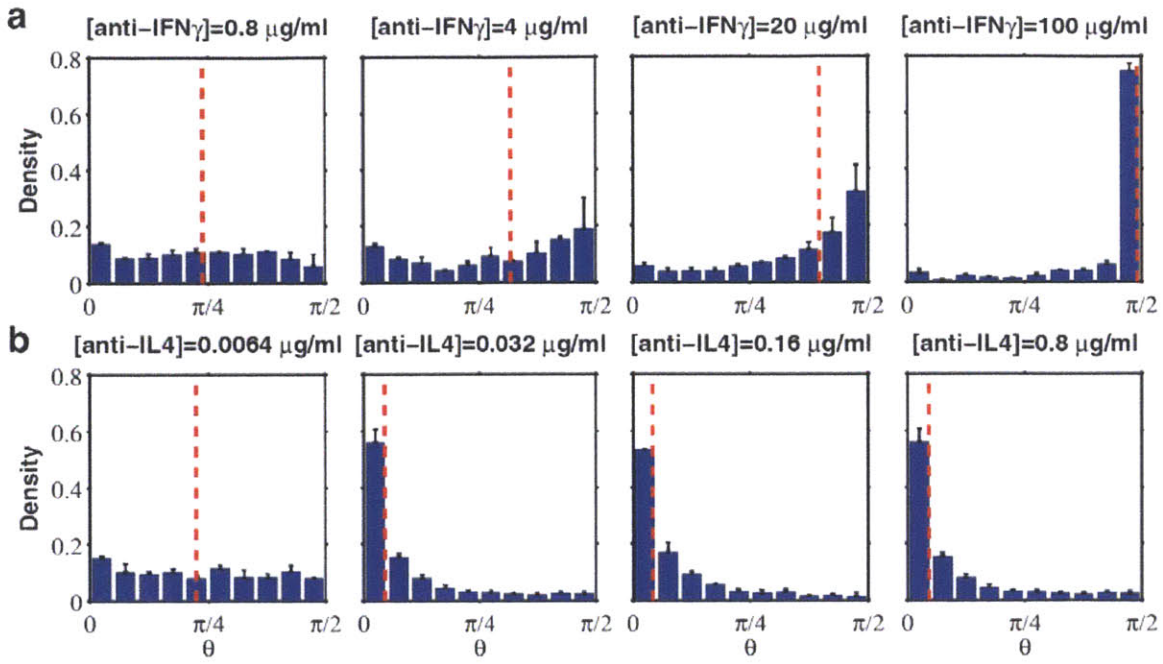
Supplementary Fig. 3.15. Scatter plots with and marginal distributions showing that IFNγ antibody downregulates *Tbx21*, and IL4 antibody downregulates *Gata3* at 24 h.



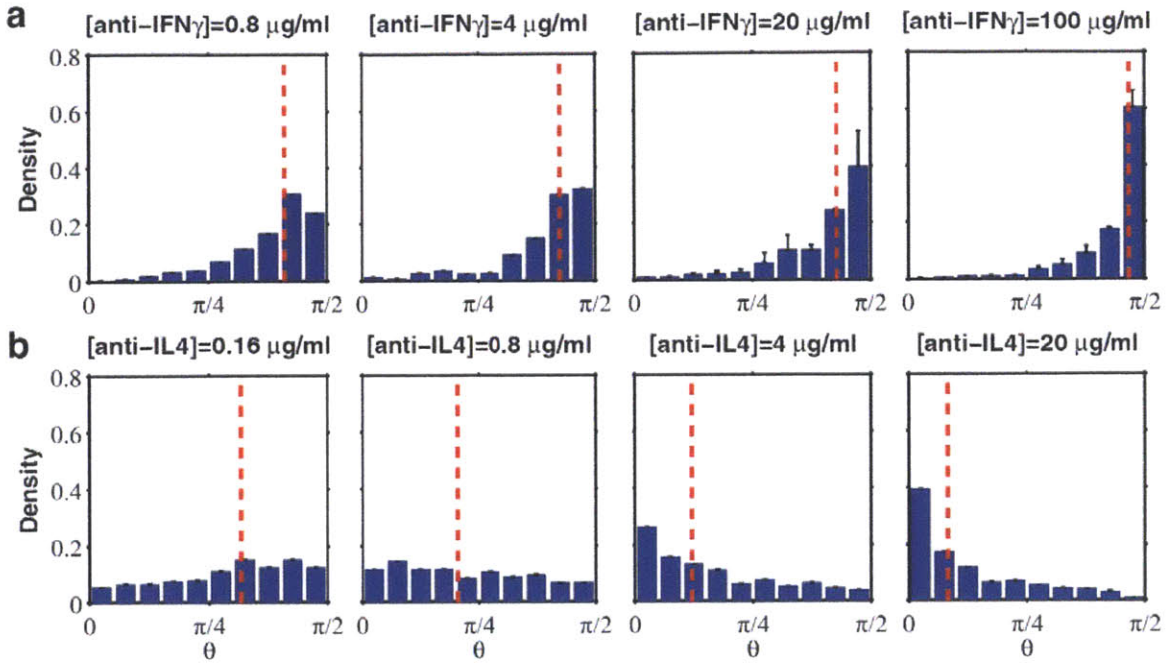
Supplementary Fig. 3.16. Scatter plots and marginal distributions of *Tbx21* and *Gata3* transcripts in individual cells treated with IL12 antibody, with the red line divides data points into halves. The left panel shows cells 16 hours after activation; the right panel shows cells 24 hours after activation. The result shows that anti-IL12 has no effect on the expression of *Tbx21* during early differentiation of Th cells.



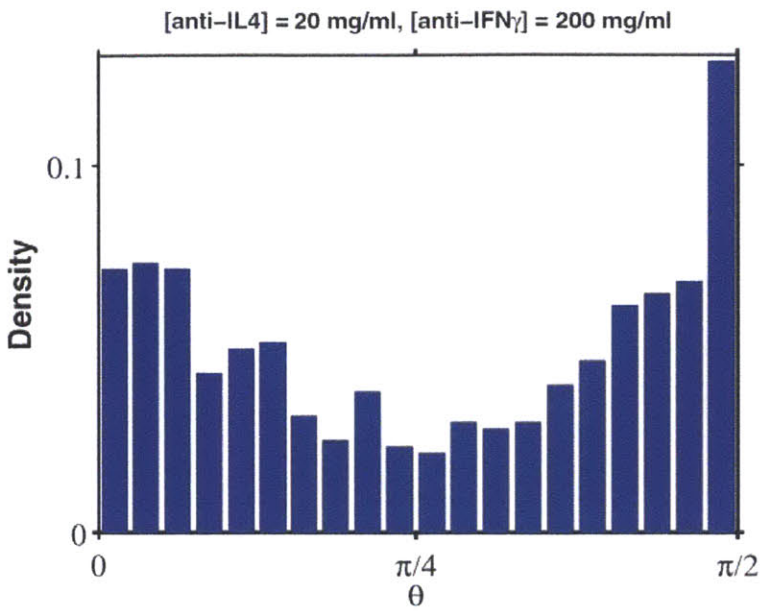
Supplementary Fig. 3.17. Distribution of θ in the absence of neutralizing antibodies. The left panel is 16 hours after activation, where θ follows a uniform distribution. The right panel is 48 hours after activation, where θ is skewed towards $\pi/2$, indicating cells become more Th2-like.



Supplementary Fig. 3.18. Distribution of θ at 16 hours after activation. Panel (**a**) shows that as concentration of anti-IFNγ antibody increases, the cells adopt larger θ. Panel (**b**) shows that as concentration of anti-IL4 antibody increases, the cells adopt smaller θ. Red lines are the medians of the θ distribution.



Supplementary Fig. 3.19. Distribution of θ at 48 hours after activation. Panel (**a**) shows that as concentration of anti-IFNγ antibody increases, the cells adopt larger θ. Panel (**b**) shows that as concentration of anti-IL4 antibody increases, the cells adopt smaller θ. Red lines are the medians of the θ distribution.



Supplementary Fig. 3.20. Distribution of θ at 48 hours, where cells were not treated with any polarizing antibodies for the first 24 hours, followed by the addition of both anti-IFNγ and anti-IL4 antibodies at 24 h. It shows that vast majority of cells adopt either very large or small θ, adopting either a Th1-like or Th2-like cell fate.

Materials and Methods

Strains of mice used

Experiments on wildtype cells were from C57BL/6 mice; experiments on *Il4*$^{-/-}$ cells were from B6.129P2-*Il4*^{tm1Cgn}/J mice; experiments on *Ifng*$^{-/-}$ cells were from B6.129S7-*Ifng*^{tm1Ts}/J mice. C57BL/6, *Ifng*$^{-/-}$ and *Il4*$^{-/-}$ mice were obtained from Jackson labs. All animals were housed at the Whitehead Institute for Biomedical Research and were maintained according to guidelines approved by the Massachusetts Institute of Technology (MIT) Committee on Animal Care.

Cell culture

Spleens and lymph nodes of mice aged from 6 weeks to 2 months were removed, suspended in PBS supplement with 2% FCS, and gently homogenized through a nylon mesh. Red blood cells were lysed with ammonium chloride solution (StemCell Technologies). $CD4^+$ cells were isolated by MACS purification using the CD4 microbeads (Miltenyi Biotec) in all experiments except those that explicitly mentioned negative selection. In experiments where cells were selected by depletion, MACS $CD4^+$ T cell isolation kit II was used. The medium used throughout the cell cultures was RPMI supplemented with 10% FCS, 2 mM L-glutamine, 1% penicillin and streptomycin.

Cells were seeded into 8-well Lab-tek 1.0 coverglass chambers that had been coated with a mixture of anti-CD3 (15 μg/ml, clone 17A2) and anti-CD28 (15 μg/ml, clone 37.51) antibodies for at least 3 hours, at 1,000,000 cells per well in a total volume of 0.5 ml, except one control experiment that explicitly mentioned 250,000 cells per well. The following neutralizing antibodies were used: IFNγ antibody (clone R4-6A2), IL4 antibody (clone BVD4-1D11) and IL12 antibody (clone C17.8). Cells were cultured at 37°C, 5% CO_2. The first refresh of culture media occurred at 48 hours, after which media was refreshed every 24 hours. In experiments with Th1 polarization, 10 ng/ml IFNγ and IL12 and 10 μg/ml anti-IL4 antibodies were supplemented in the media; in experiments

with Th2 polarization, 10 ng/ml IL4 and 10 μg/ml anti-IFNγ antibodies were supplemented in the media.

Single-molecule fluorescence in situ hybridization (smFISH)

We performed smFISH on the T cells and counted the mRNAs in individual cells as described previously(Hebenstreit et al.; Raj et al., 2008). Harvested T cells were fixed in PBS buffer with 3.7% formaldehyde for 10 minutes. After fixation, the cells were washed twice with PBS, permeabilized in 70% ethanol for at least two hours, and stored at 4°C. The T cells were hybridized in the same glass chamber as cell culture. After the 70% ethanol was aspirated, the samples were washed in a solution of 25% formamide and 2×SSC for 5 minutes. After the wash buffer was aspirated, 100 μl of hybridization solution containing labeled DNA probes in 2×SSC, 1 mg/ml BSA, 10 mM VRC, 0.5 mg/ml Escherichia coli tRNA and 0.1 g/ml dextran sulfate, with 25% formamide was added to the sample and incubated overnight at 30°C. The next day, the samples were washed twice by adding 1 ml of wash solution consisting of 25% formamide and 2×SSC. For each wash, the sample was incubated in wash solution for 30 minutes. Then, the sample was resuspended in 2×SSC buffer. The sequences of FISH probes are available upon request.

Immunofluorescence

To simultaneously visualize mRNA and protein levels in cells, we performed immunofluorescence after FISH protocol. The cells were incubated with 2×SSC, 0.2% Triton X-100, 5 mg/ml BSA and fluorescent antibodies for 3 h at 4°C. Where a secondary antibody is required, the samples were incubated with 2×SSC, 0.2% triton X-100, 5 mg/ml BSA and the secondary antibody for 1 h at 4°C. The cells were then washed by incubating with 2×SSC, 0.2% triton X-100, 5 mg/ml BSA for 1 h at 4°C. Tbet antibody is clone 4B10; Gata3 antibody is clone L50-823; IFNγ antibody is polyclonal (AMC4034,

Invitrogen) and a secondary goat-anti-rabbit antibody (A11034, Invitrogen) is used. We test multiple IL4 antibodies for immunofluorescence, but none of them gave satisfactory signal to noise ratio.

Image acquisition

For imaging, the samples were soaked in glucose oxidase (glox) anti-fade solution, which contains 10 mM Tris (pH 7.5), 2×SSC, 0.4% glucose, supplemented with glucose oxidase and catalase. A coverslip was put over the sample. All images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a 100X oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). Stacks of images were taken automatically with 0.4 microns between the z-slices.

Image analysis

To segment the T cells, a marker-guided watershed algorithm was used. Briefly, cell boundaries were obtained by running an edge detection algorithm on the bright-field image of the cells. To generate markers for watershed algorithm, the centroid of the region enclosed by individual cell boundaries is computed. A marker-guided watershed algorithm is then run on the distance transformation of the cell boundaries, using the markers located within the cell boundaries. The resultant cell segmentation image is then manually curated for occasional mis-segmentations.

To quantify the number of RNA molecules in each cell, a log filter is run over each optical slice of the image stack to enhance signals. A threshold is taken on the resultant image stack to pick up mRNA spots. The locations of mRNA spots are then taken to be the regional maximum pixel value of each connected region. The number of mRNA spots located within the cell boundaries of an individual cell can thus be quantified.

To quantify fluorescence signal in each cell, an optical slice corresponding to the central plane of the cells is analyzed. For each image, which covers up to 100 correctly segmented cells, the mean fluorescence per pixel of each cell is computed. The minimum of mean fluorescence is taken to be the background. Then for each cell in the image, the total fluorescence of the cell is computed as the sum of the fluorescence at each pixel subtracting the background. If this value is negative, zero is used instead.

CHAPTER 4

Conclusion and Future Work

In this thesis, I explored the heterogeneous gene expression in single cells during the differentiation of T helper cells. I have discovered a paradigm of cell lineage specification governed by the signaling interplay between extracellular cues and intracellular transcriptional factors, where the strength of extracellular signaling dominates over the intracellular signaling component. In the presence of extracellular cues for both lineages, naive T helper cell co-express *Tbx21* and *Gata3* at high levels, stochastically acquiring any intermediate Th1/Th2 states. The states of T helper cells can be gradually tuned by depriving availability of extracellular cytokines, which are produced stochastically by a small subpopulation of cells. In this model, the rare cytokine-expressing cells act as leaders and can secret cytokines to instruct the whole cell population express the appropriate transcription factors ubiquitously. When the cytokines are removed with neutralizing antibodies, cells down-regulate the expression of the corresponding transcription factor, thus biasing towards cells states that are closer to the alternative lineage. When extracellular cues are removed, the weak intracellular signaling network reveals its effect, leading to classic mutual exclusion of antagonistic transcriptional factors.

Looking forward, many intriguing questions remain to be answered. What is the generality of the paradigm we have discovered in T helper cells? How relevant is our model to T helper cell differentiation *in vivo*? Are the rare cytokine-expressing cells the main providers of cytokines cues, compared to antigen presenting cells? How does cytokine micro-environment affect T helper cell differentiation *in vivo*? How can cells achieve the appropriate Th1/Th2 response to pathogens, given the significant heterogeneity in gene expression levels among individual cells?

4.1 Other CD4 T helper cell lineages – Th17 and iTreg

In addition to Th1 and Th2 lineages, naive CD4 T cells are capable of differentiating into Th17 and induced regulatory T cells (iTreg). Therefore, it is interesting to explore the gene expression pattern of master transcription factors governing Th17 and iTreg lineages to investigate whether these antagonistic transcription factor are also co-expressed at high levels. The master transcription factor governing Th17 lineage specification is RORγT, encode by the gene *Rorc2* in mice; the master transcription factor governing iTreg lineage specification is Foxp3.

We performed single-molecule FISH on both untreated cells and that treated with transforming growth factor β1 (TGFβ1), which is a cytokine that pushes cells towards the Th17-iTreg paradigm in contrast to the Th1-Th2 paradigm. We found that *Rorc2* and *Foxp3* are co-expressed at high levels in some cells, while a significant portion of cells express only *Rorc2* or *Foxp3* at high levels.

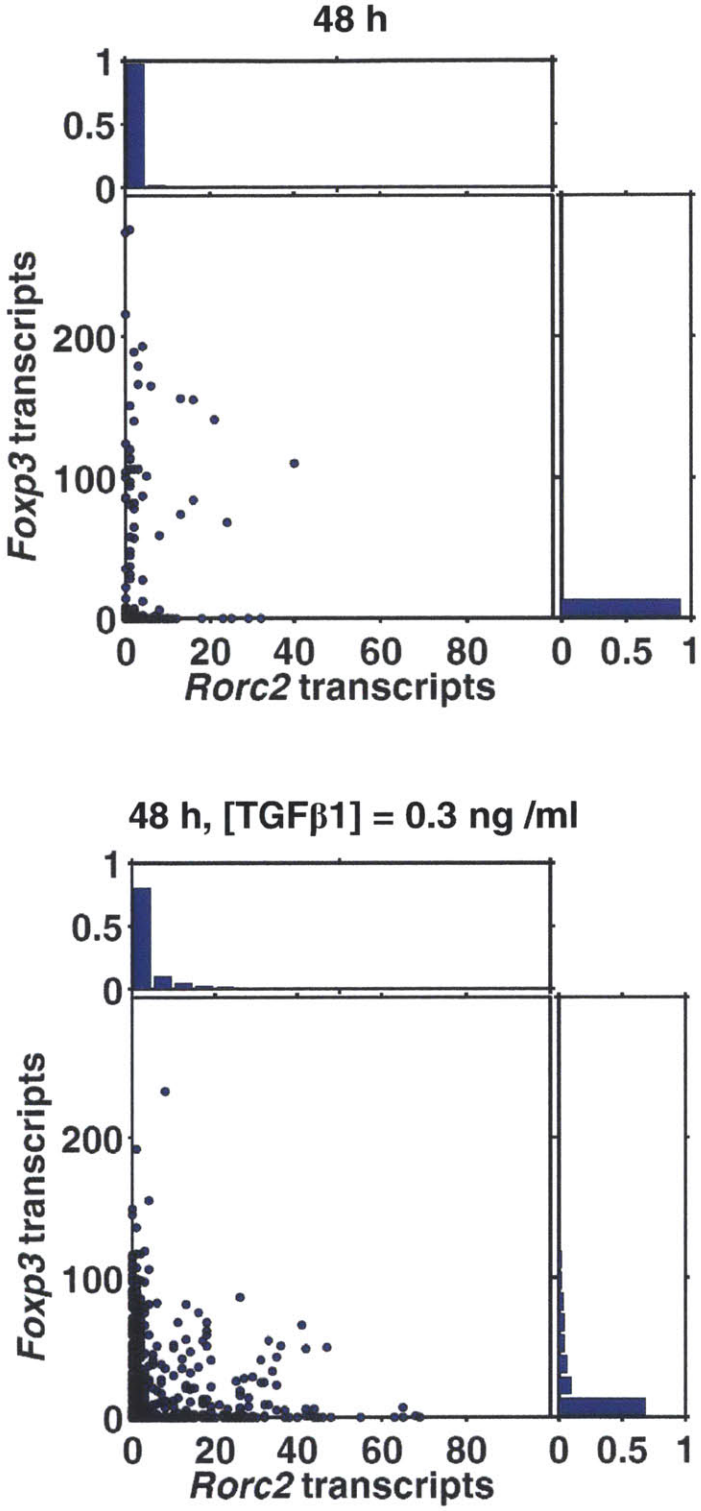
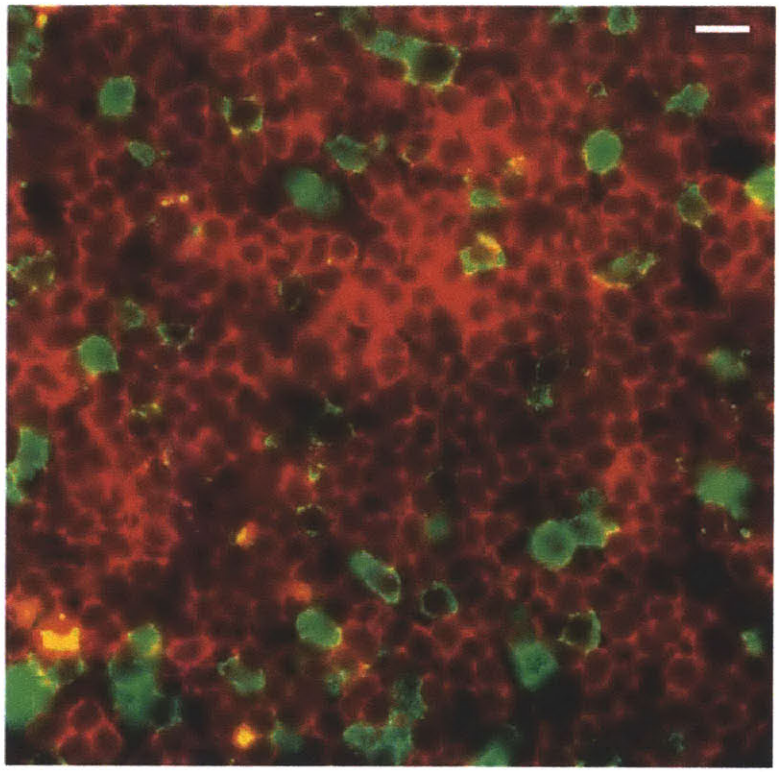


Fig 4.1. Scatter plots with marginal distribution of *Foxp3* and *Rorc2* expression in CD4 T helper cells. The upper panel shows cells untreated with any cytokines; the upper panel shows cells untreated with 0.3 ng/ml of TGFβ1.

4.2 T cell differentiation *in vivo*

To explore T helper cell differentiation *in vivo*, one experimental setup would be to examine the challenged lymph nodes by utilizing fixed tissue sections. Specifically, one can inject one of the mouse footpads with pathogen and adjuvant complex, while leaving the contralateral footpad untreated, then isolate the popliteal lymph nodes on both sides, section the fixed the lymph nodes, and perform smFISH and immunofluorescence for analysis. We explore the technical feasibility of imaging lymph nodes and our preliminary data show that lymph node tissues are amenable to smFISH and immunofluorescence (Figure 4.1 and 4.2).



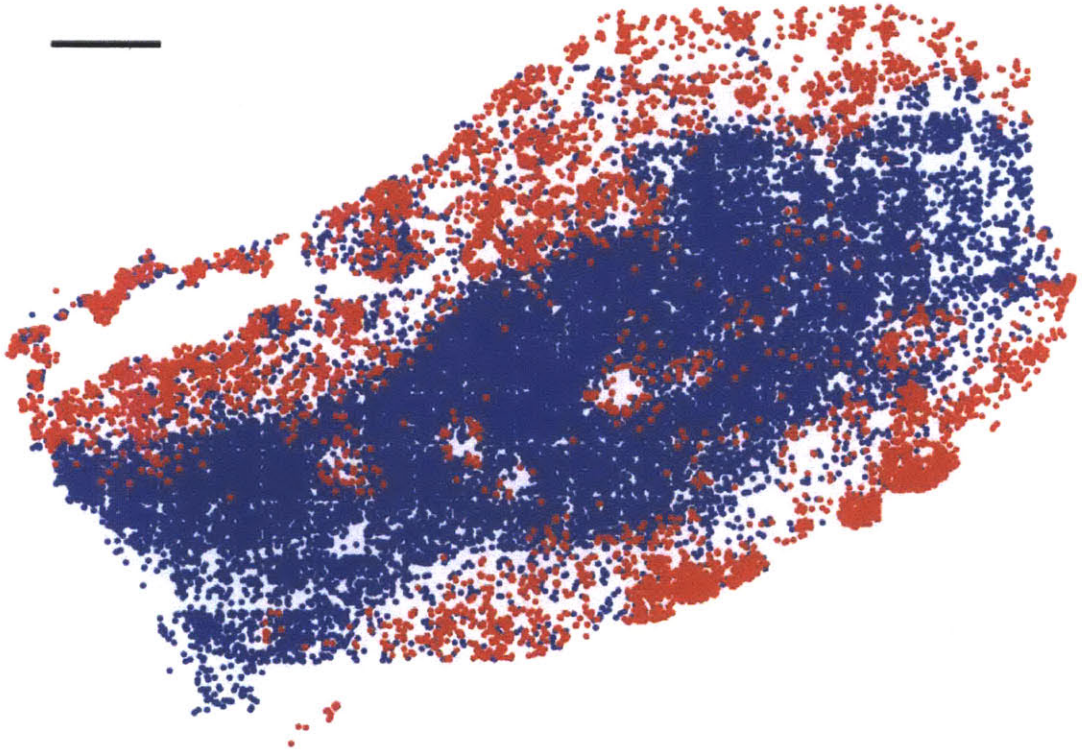


Figure 4.1. Immunofluorescence of the lymph node section. The upper panel shows an image in the cortex region probably inside a germinal center. The cells were labeled with anti-B220 (red) and anti-CD3 (green) antibodies. Anti-B220 labels B-cells and anti-CD3 labels T cells. The scale bar is 10μm. The lower panel shows a processed map of the lymph node, with each red dot indicating a single cell marked by anti-B220 and each blue dot indicating a single cell marked by anti-CD3. The scale bar is 200μm.

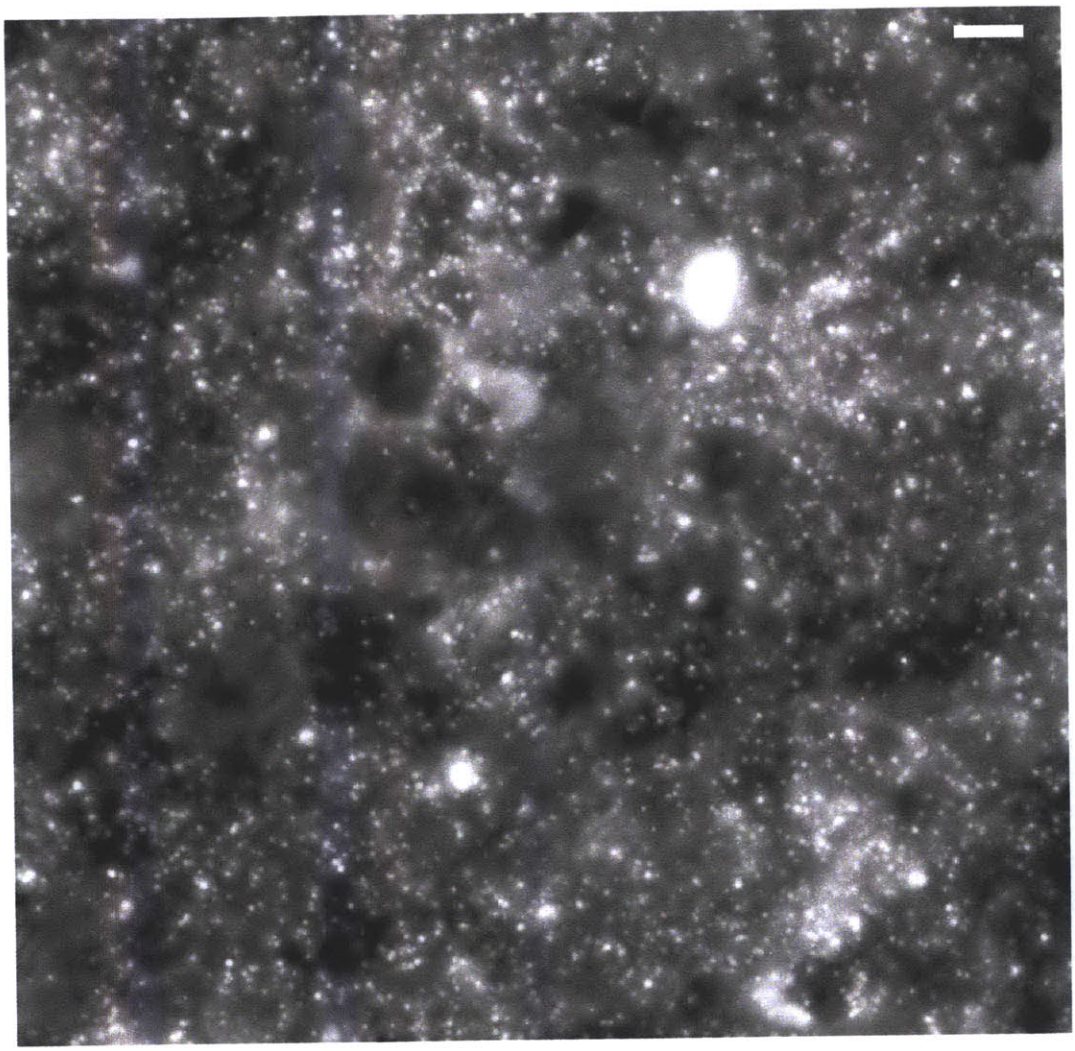


Figure 4.2. smFISH image of thymus tissue section, labeled with probes that detects *CD8α*. The scale bar is 10μm.

Although studying lymph node tissues with smFISH and immunofluorescence is technically feasible, to quantitatively understand gene expression during T helper cell differentiation *in vivo* remains as a challenging endeavor. First, problem is being able to nail down which cells amongst billions of cells in the lymph nodes are the T helper cells of interest. Only 20% of the all the cells in the lymph nodes are $CD4^+$ T cells. Since T

cells have diverse repertoire of TCR that can recognize different epitopes, only a very small subset of T helper cells will respond to pathogens presented. From our preliminary data, it is extremely hard to identify the T helper cells of interest in the ocean of other irrelevant cell types. This problem might be partially solvable by using a TCR transgenic line which has uniform TCR on $CD4^+$ T cells, so that theoretically 20% of the all the cells become relevant. We explored this approach in OT-II mice, but failed to see significant upregulation of *Tbx21* and *Gata3*. Second, given higher autofluorescence in tissue sections than in cell cultures, the noise in the quantification methods is significantly increased. In addition, cells can be sliced through in tissue sections, making segmentation of cell boundaries inaccurate. As a result, it will be impossible to generate clean single-cell data by studying tissue sections. Third, it is much more technically challenging to manipulate the system. For example, it is almost impossible to ensure the $CD4^+$ T helper cells of interest are exposed to controlled dosages of cytokine or neutralizing antibodies against cytokines by intravenous injections, because of problems with metabolism, degradation, interactions with components of the blood stream, and inefficient transfusion into the lymph nodes. Fourth, it is technically impossible to synchronize the cells to initiate differentiation, because not $CD4^+$ T helper cell receives the signaling cue at the same time. Despite the challenges of studying lineage specification *in vivo*, fixed tissue sections remain as a valuable avenue for studying lineage specification under physiological conditions, though probably not with a goal of single-cell resolution.

REFERENCE

Akaike, H. (1974). New Look at Statistical-Model Identification. Ieee Transactions on Automatic Control *Ac19*, 716-723.

Ansel, K.M., Djuretic, I., Tanasa, B., and Rao, A. (2006). Regulation of Th2 differentiation and Il4 locus accessibility. Annu Rev Immunol *24*, 607-656.

Arinobu, Y., Mizuno, S., Chong, Y., Shigematsu, H., Iino, T., Iwasaki, H., Graf, T., Mayfield, R., Chan, S., Kastner, P., *et al.* (2007). Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. Cell Stem Cell *1*, 416-427.

Beach, D.L., Salmon, E.D., and Bloom, K. (1999). Localization and anchoring of mRNA in budding yeast. Curr Biol *9*, 569-578.

Becskei, A., Kaufmann, B.B., and van Oudenaarden, A. (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat Genet *37*, 937-944.

Bengtsson, M., Stahlberg, A., Rorsman, P., and Kubista, M. (2005). Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res *15*, 1388-1392.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S.M., Singer, R.H., and Long, R.M. (1998). Localization of ASH1 mRNA particles in living yeast. Mol Cell *2*, 437-445.

Blake, W.J., Balazsi, G., Kohanski, M.A., Isaacs, F.J., Murphy, K.F., Kuang, Y., Cantor, C.R., Walt, D.R., and Collins, J.J. (2006). Phenotypic consequences of promoter-mediated transcriptional noise. Mol Cell *24*, 853-865.

Blake, W.J., M, K.A., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633-637.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics *20*, 3710-3715.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics *11*, 94.

Callard, R.E. (2007). Decision-making by the immune response. Immunol Cell Biol *85*, 300-305.

Casella, G., and Berger, R.L. (2001). Statistical Inference, 2nd edn (Duxbury Press).

Chang, S., and Aune, T.M. (2007). Dynamic changes in histone-methylation 'marks' across the locus encoding interferon-gamma during the differentiation of T helper type 2 cells. Nat Immunol *8*, 723-731.

Chintapalli, V.R., Wang, J., and Dow, J.A. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet *39*, 715-720.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., *et al.* (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods *5*, 613-619.

Cui, K., Zang, C., Roh, T.Y., Schones, D.E., Childs, R.W., Peng, W., and Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell *4*, 80-93.

Dalton, D.K., Pitts-Meek, S., Keshav, S., Figari, I.S., Bradley, A., and Stewart, T.A. (1993). Multiple defects of immune cell function in mice with disrupted interferon-gamma genes. Science *259*, 1739-1742.

Djuretic, I.M., Levanon, D., Negreanu, V., Groner, Y., Rao, A., and Ansel, K.M. (2007). Transcription factors T-bet and Runx3 cooperate to activate Ifng and silence Il4 in T helper type 1 cells. Nat Immunol *8*, 145-153.

Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of single RNA transcripts in situ. Science *280*, 585-590.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-time kinetics of gene activity in individual bacteria. Cell *123*, 1025-1036.

Graumann, J., Hubner, N.C., Kim, J.B., Ko, K., Moser, M., Kumar, C., Cox, J., Scholer, H., and Mann, M. (2008). Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. Mol Cell Proteomics *7*, 672-683.

Gregory, S.G., Schmidt, S., Seth, P., Oksenberg, J.R., Hart, J., Prokop, A., Caillier, S.J., Ban, M., Goris, A., Barcellos, L.F., *et al.* (2007). Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet *39*, 1083-1091.

Hastie, N.D., and Bishop, J.O. (1976). The expression of three abundance classes of messenger RNA in mouse tissues. Cell *9*, 761-774.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S.A. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst Biol *7*, 497.

Hebenstreit, D., Giaisi, M., Treiber, M.K., Zhang, X.B., Mi, H.F., Horejs-Hoeck, J., Andersen, K.G., Krammer, P.H., Duschl, A., and Li-Weber, M. (2008). LEF-1 negatively controls interleukin-4 expression through a proximal promoter regulatory element. J Biol Chem *283*, 22490-22497.

Hebenstreit, D., Gu, M., Haider, S., Turner, D.J., Lio, P., and Teichmann, S.A. (2011). EpiChIP: gene-by-gene quantification of epigenetic modification levels. Nucleic Acids Res *39*, e27.

Hebenstreit, D., and Teichmann, S.A. (2011). Analysis and simulation of gene expression profiles in pure and mixed cell populations. Phys Biol *8*, 035013.

Hegazy, A.N., Peine, M., Helmstetter, C., Panse, I., Frohlich, A., Bergthaler, A., Flatz, L., Pinschewer, D.D., Radbruch, A., and Lohning, M. Interferons direct Th2 cell reprogramming to generate a stable GATA-3(+)T-bet(+) cell subset with combined Th2 and Th1 cell functions. Immunity *32*, 116-128.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., *et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Hofer, T., Nathansen, H., Lohning, M., Radbruch, A., and Heinrich, R. (2002). GATA-3 transcriptional imprinting in Th2 lymphocytes: a mathematical model. Proc Natl Acad Sci U S A *99*, 9364-9368.

Hoyle, D.C., Rattray, M., Jupp, R., and Brass, A. (2002). Making sense of microarray data distributions. Bioinformatics *18*, 576-584.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev *11*, 774-785.

Hwang, E.S., Szabo, S.J., Schwartzberg, P.L., and Glimcher, L.H. (2005). T helper cell fate specified by kinase-mediated interaction of T-bet with GATA-3. Science *307*, 430-433.

Jenner, R.G., Townsend, M.J., Jackson, I., Sun, K., Bouwman, R.D., Young, R.A., Glimcher, L.H., and Lord, G.M. (2009). The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. Proc Natl Acad Sci U S A *106*, 17876-17881.

Jiang, H., and Wong, W.H. (2009). Statistical inferences for isoform expression in RNA-Seq. Bioinformatics *25*, 1026-1032.

Kaplan, M.H., Schindler, U., Smiley, S.T., and Grusby, M.J. (1996). Stat6 is required for mediating responses to IL-4 and for development of Th2 cells. Immunity *4*, 313-319.

Kuhn, R., Rajewsky, K., and Muller, W. (1991). Generation and analysis of interleukin-4 deficient mice. Science *254*, 707-710.

Laiosa, C.V., Stadtfeld, M., and Graf, T. (2006). Determinants of lymphoid-myeloid lineage diversification. Annu Rev Immunol *24*, 705-738.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lattin, J.E., Schroder, K., Su, A.I., Walker, J.R., Zhang, J., Wiltshire, T., Saijo, K., Glass, C.K., Hume, D.A., Kellie, S., *et al.* (2008). Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res *4*, 5.

Leonard, W.J., and O'Shea, J.J. (1998). Jaks and STATs: biological implications. Annu Rev Immunol *16*, 293-322.

Lohning, M., Richter, A., and Radbruch, A. (2002). Cytokine memory of T helper lymphocytes. Adv Immunol *80*, 115-181.

Lu, C., and King, R.D. (2009). An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems. Bioinformatics *25*, 2020-2027.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenas, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol *6*, 450.

Malek, T.R. (2008). The biology of interleukin-2. Annu Rev Immunol *26*, 453-479.

Mariani, L., Lohning, M., Radbruch, A., and Hofer, T. (2004). Transcriptional control networks of cell differentiation: insights from helper T lymphocytes. Prog Biophys Mol Biol *86*, 45-76.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res *18*, 1509-1517.

Morris, G.P., and Allen, P.M. How the TCR balances sensitivity and specificity for the recognition of self and pathogens. Nat Immunol *13*, 121-128.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Mudge, J., Miller, N.A., Khrebtukova, I., Lindquist, I.E., May, G.D., Huntley, J.J., Luo, S., Zhang, L., van Velkinburgh, J.C., Farmer, A.D., *et al.* (2008). Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. PLoS ONE *3*, e3625.

Mullen, A.C., Hutchins, A.S., High, F.A., Lee, H.W., Sykes, K.J., Chodosh, L.A., and Reiner, S.L. (2002). Hlx is induced by and genetically interacts with T-bet to promote heritable T(H)1 gene induction. Nat Immunol *3*, 652-658.

Murphy, K.M., and Reiner, S.L. (2002). The lineage decisions of helper T cells. Nat Rev Immunol *2*, 933-944.

Oshlack, A., and Wakefield, M.J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol Direct *4*, 14.

Ouyang, W., Lohning, M., Gao, Z., Assenmacher, M., Ranganath, S., Radbruch, A., and Murphy, K.M. (2000). Stat6-independent GATA-3 autoactivation directs IL-4-independent Th2 development and commitment. Immunity *12*, 27-37.

Ouyang, W., Ranganath, S.H., Weindel, K., Bhattacharya, D., Murphy, T.L., Sha, W.C., and Murphy, K.M. (1998). Inhibition of Th1 development mediated by GATA-3 through an IL-4-independent mechanism. Immunity *9*, 745-755.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat Methods *6*, S22-32.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. PLoS Biol *4*, e309.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods *5*, 877-879.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226.

Raj, A., and van Oudenaarden, A. (2009). Single-molecule approaches to stochastic gene expression. Annu Rev Biophys *38*, 255-270.

Ramskold, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol *5*, e1000598.

Raser, J.M., and O'Shea, E.K. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811-1814.

Roh, T.Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev *19*, 542-552.

Rosen, E.D., and Spiegelman, B.M. (2000). Molecular regulation of adipogenesis. Annu Rev Cell Dev Biol *16*, 145-171.

Rothenberg, E.V. (2007). Cell lineage regulators in B and T cell development. Nat Immunol *8*, 441-444.

Schmitz, J., Thiel, A., Kuhn, R., Rajewsky, K., Muller, W., Assenmacher, M., and Radbruch, A. (1994). Induction of interleukin 4 (IL-4) expression in T helper (Th) cells is not dependent on IL-4 from non-Th cells. J Exp Med *179*, 1349-1353.

Schoenborn, J.R., Dorschner, M.O., Sekimata, M., Santer, D.M., Shnyreva, M., Fitzpatrick, D.R., Stamatoyannopoulos, J.A., and Wilson, C.B. (2007). Comprehensive epigenetic profiling identifies multiple distal regulatory elements directing transcription of the gene encoding interferon-gamma. Nat Immunol *8*, 732-742.

Schwarz, G. (1978). Estimating Dimension of a Model. Annals of Statistics *6*, 461-464.

Shaffer, A.L., Lin, K.I., Kuo, T.C., Yu, X., Hurt, E.M., Rosenwald, A., Giltnane, J.M., Yang, L., Zhao, H., Calame, K., *et al.* (2002). Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. Immunity *17*, 51-62.

Shimoda, K., van Deursen, J., Sangster, M.Y., Sarawar, S.R., Carson, R.T., Tripp, R.A., Chu, C., Quelle, F.W., Nosaka, T., Vignali, D.A., *et al.* (1996). Lack of IL-4-induced Th2 response and IgE class switching in mice with disrupted Stat6 gene. Nature *380*, 630-633.

Silverman, B.W. (1986). Density Estimation (London, Chapman and Hall).

Smith-Garvin, J.E., Koretzky, G.A., and Jordan, M.S. (2009). T cell activation. Annu Rev Immunol *27*, 591-619.

Spandidos, A., Wang, X., Wang, H., and Seed, B. (2010). PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. Nucleic Acids Res *38*, D792-799.

Szabo, S.J., Kim, S.T., Costa, G.L., Zhang, X., Fathman, C.G., and Glimcher, L.H. (2000). A novel transcription factor, T-bet, directs Th1 lineage commitment. Cell *100*, 655-669.

Szabo, S.J., Sullivan, B.M., Peng, S.L., and Glimcher, L.H. (2003). Molecular mechanisms regulating Th1 immune responses. Annu Rev Immunol *21*, 713-758.

Takeda, K., Tanaka, T., Shi, W., Matsumoto, M., Minami, M., Kashiwamura, S., Nakanishi, K., Yoshida, N., Kishimoto, T., and Akira, S. (1996). Essential role of Stat6 in IL-4 signalling. Nature *380*, 627-630.

Tyagi, S., and Kramer, F.R. (1996). Molecular beacons: probes that fluoresce upon hybridization. Nat Biotechnol *14*, 303-308.

Tykocinski, L.O., Hajkova, P., Chang, H.D., Stamm, T., Sozeri, O., Lohning, M., Hu-Li, J., Niesner, U., Kreher, S., Friedrich, B., *et al.* (2005). A critical control element for interleukin-4 memory expression in T helper lymphocytes. J Biol Chem *280*, 28177-28185.

Vargas, D.Y., Raj, A., Marras, S.A., Kramer, F.R., and Tyagi, S. (2005). Mechanism of mRNA transport in the nucleus. Proc Natl Acad Sci U S A *102*, 17008-17013.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

Wang, Z., Gerstein, M., and Snyder, M. (2009a). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics *10*, 57-63.

Wang, Z., Schones, D.E., and Zhao, K. (2009b). Characterization of human epigenomes. Curr Opin Genet Dev *19*, 127-134.

Warren, L., Bryder, D., Weissman, I.L., and Quake, S.R. (2006). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc Natl Acad Sci U S A *103*, 17807-17812.

Wei, G., Wei, L., Zhu, J., Zang, C., Hu-Li, J., Yao, Z., Cui, K., Kanno, Y., Roh, T.Y., Watford, W.T., *et al.* (2009). Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. Immunity *30*, 155-167.

Yates, A., Callard, R., and Stark, J. (2004). Combining cytokine signalling with T-bet and GATA-3 regulation in Th1 and Th2 differentiation: a model for cellular decision-making. J Theor Biol *231*, 181-196.

Zhang, D.H., Cohn, L., Ray, P., Bottomly, K., and Ray, A. (1997). Transcription factor GATA-3 is differentially expressed in murine Th1 and Th2 cells and controls Th2-specific expression of the interleukin-5 gene. J Biol Chem *272*, 21597-21603.

Zheng, W., and Flavell, R.A. (1997). The transcription factor GATA-3 is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells. Cell *89*, 587-596.

Zhou, L., Chong, M.M., and Littman, D.R. (2009). Plasticity of CD4+ T cell lineage differentiation. Immunity *30*, 646-655.

Zhu, J., Yamane, H., and Paul, W.E. (2010). Differentiation of effector CD4 T cell populations (*). Annu Rev Immunol *28*, 445-489.