UNIVERSITY OF CALIFORNIA, SAN DIEGO

Enhancers and Transcriptional Regulation in CD4+ T Cells

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

in

Bioinformatics and Systems Biology

by

Karmel Alon Allison

Committee in charge:

Professor Christopher K. Glass, Chair
Professor Stephen M. Hedrick, Co-Chair
Professor John T. Chang
Professor Terry Gaasterland
Professor Bing Ren

2015

The Dissertation of Karmel Alon Allison is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2015

DEDICATION

To my husband, David, who loved, supported, and funded me throughout this journey,

to my son, Ezra, who was a very good little boy and let Mommy work,

and to God, who is and does all things.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Professor Christopher Glass for his support, guidance, and wisdom as the chair of my committee. I am always impressed by the incredible breadth and depth of information you can keep track of, and the percentage of times when your first guess is exactly correct!

I would also like to thank Professor Stephen Hedrick for his advice, for access to his mental database of all things T cell, and for his rather wry sense of humor.

I would like to thank everyone in the Glass Lab—I stood on the shoulders of all of your hard work to complete this dissertation. In particular, I thank Casey Romanoski, Minna Kaikkonen, and Verena Link for their help, friendship, and guidance; and Nathan Spann for helping me fix many fumbles at the bench.

I thank Leslie Van Ael and Valerie Alon for their assistance in and patience with formatting this dissertation.

Chapter One, in full, is a reprint of the material as it appears in *Nucleic Acids Research*: Allison, Karmel A; Kaikkonen, Minna U; Gaasterland, Terry; Glass, Christopher K, 2014. The dissertation author is the primary author of this paper.

Chapter Two, in full, is currently being prepared for submission for publication: Allison, Karmel A; Stone, Erica L; Collier, Jana G; Gosselin, David; Troutman, Ty Dale; Hedrick, Stephen M; Glass, Christopher K. The dissertation author is the primary author of this paper.

VITA

2006 **Bachelor of Arts, University of California, Berkeley**
Highest Distinction in General Scholarship (Summa Cum Laude)
Regents' and Chancellor's Scholar
Phi Beta Kappa Honors Society

2015 **Doctor of Philosophy, University of California, San Diego**
NIH Ruth Kirschstein National Research Service Award (F31)
Achievement Rewards for College Scientists (ARCS) Scholar
NSF Graduate Research Fellowship Program Honorable Mention
NIDDK Keystone Symposia Scholarship
University of Eastern Finland International Doctorate Visitor Grant

PUBLICATIONS

Allison KA, Glass CK. Macrophage activation as a model system for understanding enhancer transcription and eRNA function. Springer. 2015.

Davis-Turak JC, **Allison KA**, Shokhirev MN, Ponomarenko P, Tsimring LS, Glass CK, Johnson TL, Hoffmann A. Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. *Nucleic Acids Res*. 2015 Jan 30;43(2):699-707.

Allison KA, Kaikkonen MU, Gaasterland T, Glass CK. Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic Acids Res*. 2014 Feb;42(4):2433-47.

Heinz S, Romanoski CE, Benner C, **Allison KA**, Kaikkonen MU, Orozco LD, Glass CK. Natural genetic variation perturbs collaborative transcription factor binding required for enhancer selection and function. *Nature*. 2013 Oct 13.

Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, **Allison KA**, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, Glass CK. Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Mol Cell*. 2013 Aug 8;51(3):310-25.

ABSTRACT OF THE DISSERTATION

Enhancers and Transcriptional Regulation in CD4+ T Cells

by

Karmel Alon Allison

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2015

Professor Christopher K. Glass, Chair
Professor Stephen M. Hedrick, Co-Chair

High-throughput sequencing has given us unprecedented insight into the regulatory networks that govern enhancer selection and transcription in mammalian cells, but many open questions remain as to how the mechanics of transcriptional regulation correspond to biological outputs such as gene expression and downstream

signaling. In this dissertation, I address the nature of enhancer selection and transcriptional regulation in the context of CD4+ T cell signaling in two parts. The first study describes an algorithm and database that together enable the use of Global Run-On Sequencing (GRO-Seq) data, an experimental data type that reveals the kinetics of transcription in the nucleus, for high dimensional analysis of enhancers and other transcriptional regulatory units. The tool developed allows for both the quantification of nascent RNA and the integration of other sequencing data types into analysis of GRO-Seq data, thus facilitating the use of GRO-Seq as an experimental assay of transcriptional behavior. The second study looks at enhancers and transcriptional regulation in a particular biological context: activation of CD4+ T cell by ligands of varying affinity. Using flow cytometry as well as several high-throughput sequencing methods, I found that CD4+ T cells reflect the strength of T Cell Receptor signaling both at the population level and the single-cell level, resulting in graded gene expression profiles for a subset of genes crucial for CD4+ T cell activation. Together, these studies represent an advance in our understanding of enhancer biology, particularly in the context of CD4+ T cell activation.

INTRODUCTION

Transcription in mammalian cells is a highly regulated process, with the expression of any given gene requiring the integration of multiple signaling cues[1]. Enhancers are one component of the transcriptional regulatory network. Unlike gene promoters, enhancers are regions of the genome that are able to regulate transcription from a distance, acting to increase or decrease transcriptional activity at target genes by looping into close proximity with promoters and recruiting components of transcriptional complexes[2]. Many individual enhancers have been identified since the first cellular enhancer was discovered over 30 years ago[3], but with the advent of high-throughput sequencing, our ability to locate and characterize enhancers has increased tremendously. Crucially, Heintzman *et al*[4] described a particular pattern of histone methylation that acts as a signature for enhancers genome-wide, making it possible to map the enhancer landscapes of a variety of cells and organisms by looking for regions with high enrichment of H3K4 mono- and di-methylation (H3K4me1/2) and low enrichment of H3K4 tri-methylation (H3K4me3).

This genome-wide mapping of enhancers associated with specific chromatin signatures has led to the recognition that enhancers are distinct from cell type to cell type[5], even when gene expression is not. These enhancers serve as important binding sites for many transcription factors (TFs), and thus they provide an essential means by which cell-type-specific gene expression programs are established and maintained[5-7]. The precise mechanisms by which enhancers are selected and activated in each cell type remain to be determined, but recent studies propose a hierarchical model in which

lineage determining TFs bind to their specific motifs, thereby opening chromatin to form primed enhancers that can subsequently be bound and activated by signal-dependent TFs in a cell-type-specific manner[8,9]. This model, in which lineage-specific factors are responsible for defining a large proportion of available enhancers, explains how similar signaling pathways can result in different downstream binding profiles for widely expressed signaling factors like Nuclear Factor kappa B (NF-kB)[10] and the Glucocorticoid Receptor (GR)[11].

Notably, surveying the enhancer landscape by ChIP-Sequencing for histone marks has proven a remarkably sensitive assay for differences between cell types. Initial reports detailed differences between cells derived from different tissues[5,7], but recent studies have found many differences between the enhancer repertoires of closely related types of cells, such as Th1 and Th2 CD4+ T cells[12], and even between the same cells in different conditions, such as untreated and lipopolysaccharide-treated macrophages[13,14].

These differences in enhancer landscapes between cells raise an interesting possibility: we could target these cell-type-specific and context-dependent enhancers in order to modulate gene expression within a particular subtype of cells. Such a targeted modulation of gene expression, called enhancer therapy[15], would be valuable in treating tissue-specific autoimmune diseases such as type 1 diabetes, where recognition of self-antigen precipitates the T cell mediated destruction of beta cells in the pancreas[16]. Any therapies that address these pathogenic immune responses to self must find a balance between allowing the disease to progress uncontrolled and

systemic immune suppression. In the case of type 1 diabetes, for example, Cyclosporine A universally suppresses lymphocytes and thus effectively reverses autoimmunity initially[17], but can lead to frequent infections and renal toxicity[18] at doses high enough to maintain suppression[19]. The ideal treatment would affect only the autoreactive cells that are inappropriately responding to beta cells[20]. The promise of enhancer therapy thus becomes extremely powerful, as we could limit the effect of treatment to the subset of pathogenic cells, leaving the rest of the immune system unharmed.

The complex regulatory behavior and therapeutic potential of enhancers make them an important frontier of genomic research. In the following two chapters, I explore enhancer functionality from two directions. First, I present an algorithm and database architecture designed to aid the analysis of a unique type of RNA sequencing data, which allows us to measure expression levels of enhancer RNA (eRNA) as well as gene transcripts genome-wide. Second, I present an application of enhancer analysis, and demonstrate how understanding the enhancer landscape of CD4+ T cells informs our understanding of the behavior of the cell population.

Part 1: GRO-Sequencing for nascent transcripts

In Chapter One, I address a high-throughput sequencing method derived from RNA-Sequencing called Global Run-On sequencing (GRO-Seq)[21]. Briefly, GRO-Seq takes advantage of a nuclear run-on reaction to tag nascent RNAs as they are assembled by RNA Polymerase II. These tagged nascent transcripts are then

sequenced, giving a real-time picture of transcription within the cell. Whereas RNA sequencing measures expression levels of stable, spliced RNA species, GRO-Seq returns data on rates of active transcription, both of coding and non-coding RNA species. For example, GRO-Seq has been used to characterize transcription antisense to gene promoters genome-wide[21], to identify previously unannotated long intergenic non-coding RNA[22], and to parameterize models of co-transcriptional splicing at the 3' ends of genes[23]. At enhancers, GRO-Seq was used to demonstrate extensive bidirectional transcription at enhancers genome-wide[22]. The transcripts generated at enhancers, called enhancer RNA (eRNA), prove remarkably sensitive measures of enhancer activation: they are responsive to estrogen treatment in MCF7 cells[22], are correlated with increased binding of transcription factors FoxA1 and the androgen receptor in LNCaP cells[24], and are tightly linked to nearby gene expression changes in response to pro-inflammatory treatment of macrophages[13]. Further emphasizing the importance of GRO-Seq in understanding transcriptional activity at enhancers, a systematic analysis of the relationship between enhancer marks and enhancer activity using a logistic regression model comprising twenty-four histone marks plus GRO-Sequencing data found that eRNA was the single most predictive indicator of enhancer activity, with eRNA synthesis as measured by GRO-Seq being significantly associated with increased expression of nearby genes in multiple cell types[25].

Thus, GRO-Seq offers a unique window into the dynamics of enhancer selection and activation. However, GRO-Seq presents new challenges for data analysis. Unlike traditional RNA-Sequencing data, the regions of interest derived from

GRO-Seq are largely unannotated. Thus, it proves difficult to determine based on the sequenced short reads what constitutes a single, contiguous transcript in the original cellular context. Prior to Vespucci[26], the algorithm described in Chapter One, analyses of GRO-Seq experiments often relied on counting tags around regions of interest such as gene bodies, promoters, or enhancers identified by histone marks[13,21,24]. A Hidden Markov Model for transcript identification was developed[22], but it was optimized for recovery of RefSeq[27] genes and microRNAs, and was not well not suited for analysis of GRO-Seq transcripts at enhancers.

The algorithm presented in Chapter One is designed to address both annotated genes and smaller non-coding regions simultaneously by mapping contiguous regions of transcription into Euclidean space and stitching over gaps in short read data in a context-dependent manner. The algorithm is integrated into a relational database architecture that further aids analysis of GRO-Seq data by allowing the user to execute arbitrary queries over both GRO-Seq and other data types such as ChIP-Sequencing data, an advancement over the prevailing standard of storing sequence data in flat files that are difficult to integrate across data type and experiment. Thus, the algorithm and database system described significantly increase our ability to analyze GRO-Sequencing data, giving us increased insight into the dynamics of the enhancer networks that are key to gene expression regulation.

Part 2: Analyzing the enhancer landscape in CD4+ T cell activation

While Chapter One describes a method for improving our ability to assay and analyze enhancers genome-wide, Chapter Two focuses on an application of enhancer analysis to gain insight into a particular biological system. CD4+ T cells are an important component of the adaptive immune response to infection. They respond to antigen presented in the peripheral lymphoid tissues, and subsequently can differentiate into several different effector phenotypes[28]. While CD4+ T cells express a wide variety of signaling-responsive receptors, the key receptor for the recognition of antigen is the T cell receptor (TCR). The TCR of a CD4+ T cell binds to peptides presented by antigen presenting cells (APCs) such as dendritic cells and macrophages[29,30]. APCs patrolling a variety of tissues internalize proteins and process them to yield short sequences of amino acids that are presented on the cell surface, bound to molecules encoded by the Major Histocompatibility Complex (MHC)[31]. The activation of CD4+ T cells via the TCR, then, depends on the ternary complex formed by the TCR, the MHC, and the peptide presented in the MHC[32], and requires that a given TCR can precisely fit the particular combination of peptide and MHC being presented[33].

Robust adaptive immune responses thus rely on a set of available TCRs that can recognize a wide variety of pathogenic peptide-MHC (pMHC) complexes but not the self-antigens that are also routinely processed and presented in MHC molecules[34]. The TCR is generated in a manner analogous to immunoglobulin generation in B cells, such that precursor cells in the thymus undergo a series of gene segment rearrangements that result in a great diversity of unique TCRs[28,35]. The possibility that

T cells are generated with TCRs that recognize self-antigen is mitigated by an affinity-based maturation process in the thymus that deletes cells with self-reactive TCRs[36]. The process of semi-random TCR generation followed by thymic selection results in on the order of 10^6 unique TCR sequences in the human body[37]. The traditional model of T cell signaling held that each one of these TCRs recognized a single pMHC complex[38,39]. However, given the number of amino acid sequences that must be recognized for robust immunity, the traditional model has been revised to suggest that at least some subset of the approximately 10^6 unique TCRs are cross-reactive to varying degrees, and able to respond to a set of peptide-MHC complexes[34,40,41].

Each of the three components of the interaction can therefore lead to changes in signaling efficiency in the CD4+ T cell. TCR sequence[38,42], genetic differences in MHC[32,43], and the peptide being presented[44] all have effects on the kinetics of the pMHC-TCR interaction. Many of the kinetic parameters of this interaction have been measured or computed[45-47]: association time, dissociation time, interaction half-life, and binding constant (K_D)[48-54]; dwell time of the pMHC-TCR interaction[48,53]; enthalpy, entropy, heat capacity, and free energy[53,54]; peptide-MHC complex binding affinity[46,55-57]; adhesion frequency[58]; density of complex formation[49]; and relative localization of complexes to the center of the immunological synapse[59]. Nonetheless, the precise relationship between the kinetics of the pMHC-TCR interaction and the strength of TCR signaling are still controversial, with any given kinetic parameter unable to explain all observations of relative signaling strength downstream of the TCR[45,47,48,53]. We could harmonize the differing results by saying that the relevance of

each of the kinetic parameters depends on the TCR, pMHC, and cellular context in question, but advances in our ability to measure kinetic parameters at naturally occurring pMHC-TCR interactions will likely be required before this question can be addressed in full[49,58].

The measurement of and relevance of these kinetic parameters are confounded by the fact that, in addition to the affinity of the pMHC-TCR interaction, the frequency of the interaction affects downstream signaling strength[45-47,60]. Increasing the dose of an antigenic peptide increases signaling output of the T cell population, but this effect is highly dependent on the affinity of the pMHC-TCR interaction[32,43,53,59,61-64]. The precise relationship between dose and affinity is still controversial, however; some models indicate dose and affinity are non-interchangeable parameters on the strength of the pMHC-TCR interaction[60,65], but another shows that dose and affinity are additive at least with respect to cell division time[66].

Regardless of the kinetic details, the affinity and frequency of the pMHC-TCR interaction have significant effects on downstream signaling in the T cell. (Although "affinity" is sometimes used to refer specifically to dissociation rate over association rate [k_{off}/k_{on}][45], I use "affinity" here to refer collectively to the strength of signaling, regardless of which kinetic parameters are involved. This is also sometimes referred to as the "potency" of TCR signaling.) High-affinity binding of the pMHC-TCR complex results in inflammatory responses at a lower concentration of antigen, increased Interleukin 2 (IL2) production, increased IFNγ production, and increased proliferation[32,43,44,56,61,67-69], whereas lower affinity interactions can lead to incomplete

phosphorylation of downstream signaling complexes[50,70], anergy[68,70-73], TCR antagonism[50,51,55,61,63], unstable helper phenotype[74], reduced cytolytic activity[75], impaired memory formation[65,76], or thymic escape of autoreactive cells[77]. Different groups report different phenotypes resulting from low-affinity engagement depending on the TCR and the experimental conditions, but in each case distinct T cell phenotypes emerge dependent on the affinity or dosage of TCR engagement.

Downstream of the TCR, recognition of the pMHC complex is translated into multiple signaling pathways. The TCR itself is a heterodimer with a transmembrane domain containing multiple immunoreceptor tyrosine-based activation motifs (ITAMs)[78]. Upon TCR stimulation, two Src family tyrosine kinases, Lck and Fyn, phosphorylate the tyrosine residues of the ITAMs[79], resulting in the recruitment of Zap70, a tyrosine kinase[80]. Zap70 can phosphorylate a number of substrates, including LAT and SLP76, two adapter proteins that build a larger molecular complex called the signalosome. The signalosome in turn activates PLCγ1, which is upstream of the protein kinase C (PKC) and Ras pathways, as well as of calcium mobilization. The primary PKC in T cells, PKCθ, initiates signaling that leads to the translocation of the pro-inflammatory transcription factor NF-kB into the nucleus[78]. Ras activation begins a series of sequential phosphorylations that activate the mitogen-activated protein kinase kinase (MEK) and subsequently the extracellular-signal-regulated kinase (ERK), culminating in the activation of the AP-1 transcription factor family. AP-1 comprises heterodimers assembled from proteins of the Fos, Jun, and ATF transcription factor families[81], and requires both TCR and co-stimulatory signaling[82]

for activation. The last of the three key pro-inflammatory transcription factors, NFAT, sits downstream of the increase in intracellular calcium, as the calcium-calmodulin-regulated phosphatase calcineurin dephosphorylates NFAT, allowing it to translocate to the nucleus and activate target genes[78,83].

The effect of the affinity and frequency of the pMHC-TCR interaction on the TCR signaling cascade has been studied extensively in thymocytes, T cell precursors that develop in the thymus. In thymocytes, where signaling above a certain threshold leads to negative selection of cells and signaling below a certain threshold leads to death by neglect, the T-cell-specific protein Themis enforces the threshold for positive selection by recruiting the negative regulator SHP1 in response to low-affinity signaling but not high-affinity signaling[84]. Similarly, thymocytes leverage the son of sevenless (SOS) positive feedback loop upstream of the Ras pathway to force commitment to TCR signaling once initiated, resulting in localization of phosphorylated ERK and other Ras pathway members to the cell membrane in positively-selected thymocytes but not negatively-selected thymocytes[85-87].

Themis and SOS are critical for suppressing low-affinity signals and amplifying high-affinity signals in thymocytes, resulting in thresholding behavior, but it is unclear to what extent these pathways are relevant to peripheral T cell signaling[88]. Loss of Themis does not have the same deleterious effects on peripheral T cells as it does on thymocytes[84], and SOS has been shown by one group to be disposable for peripheral T cell activation[89]. Nevertheless, TCR signaling even in peripheral T cells is traditionally thought to be a digital process[90], meaning that signaling downstream of

the TCR is either all-on or all-off. In other words, a given T cell must either be committed to a full response or to no response. This switch-like behavior can be observed in both thymocytes and peripheral T cells via extracellular markers such as CD69[85,87], ERK pathway component localization[85-87], NF-kB activation[91], NFAT localization[92,93], proliferation[94], and cytokine production[93,95]. According to the digital model of T cell activation, any observed differences between high-affinity and low-affinity activation states would be attributed to greater frequencies of T cells responding at the population level, rather than per-cell variability[94-97], since any given T cell can only be all-on or all-off. Still, some aspects of the TCR response have been described as analog, or varying in proportion to the strength of signaling. These analog responses can be observed beginning with the most proximal signaling steps downstream of the TCR. With low-affinity ligands, the ITAMs of the CD3ζ chain are incompletely phosphorylated[50,70,85,98,99], resulting in partial Zap70 activation[85,86]. Further, low-affinity interactions are more dependent on recruitment of the tyrosine kinase Lck, perhaps due to the incomplete phosphorylation of intracellular ITAMs[88,100]. High-affinity interaction results in increased intracellular calcium concentrations[101], higher expression of the transcription factor IRF4[102,103], and decreased cell division time[66]. It is unclear how these analog components of the TCR response fit in to a digital model.

In Chapter Two, I address the controversy between digital and analog TCR responses by subjecting CD4+ T cells to pMHC interactions of differing affinities and at different doses. I use a well-characterized model system for studying the affects of

ligand affinity on CD4+ T cell activation, the transgenic AND mouse. The AND mouse is a mouse strain with a transgenic TCR such that CD4+ T cells all carry the same TCR[104]. This TCR recognizes the pigeon cytochrome c (PCC), along with a set of similar cytochrome c peptides from other organisms[42-44]. The kinetic parameters of the TCR's interaction with many cytochrome c peptides as presented by MHC have been measured and described[56,67]. By presenting the AND TCR with peptides of varying affinity at several doses, I show first that there is room for an analog response within the digital TCR model, such that for any given peptide and dose, the CD4+ T cell is all-on or all-off, but the degree to which the T cell is activated in the all-on response varies in an analog fashion with respect to the strength of TCR signaling. Further, I compare enhancer landscapes and gene expression profiles downstream of the TCR, giving new insight into the global chromatin landscape downstream of the TCR. Surprisingly, the activation kinetics of the CD4+ T cell populations yield graded gene expression profiles, but the changes in gene expression are achieved by leveraging a pre-existing enhancer landscape. Thus, potential enhancer therapies directed at CD4+ T cells could target enhancer activation, rather than enhancer establishment, to inhibit pro-inflammatory signaling.

Together, the studies presented in these two chapters represent advances in our understanding of enhancers as regulatory elements as well as our understanding of the genomics of T cell biology.

REFERENCES

1 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).

2 Ong, C. T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics* **12**, 283-293, doi:10.1038/nrg2957 (2011).

3 Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729-740 (1983).

4 Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* **39**, 311-318, doi:10.1038/ng1966 (2007).

5 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., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M. & Ren, B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:10.1038/nature07829 (2009).

6 Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

7 Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M. & Pennacchio, L. A. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).

8 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

9 Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D. & Glass, C. K. Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487-492, doi:10.1038/nature12615 (2013).

10 Barish, G. D., Yu, R. T., Karunasiri, M., Ocampo, C. B., Dixon, J., Benner, C., Dent, A. L., Tangirala, R. K. & Evans, R. M. Bcl-6 and NF-kappaB cistromes mediate opposing regulation of the innate immune response. *Genes & development* **24**, 2760-2765, doi:10.1101/gad.1998010 (2010).

11 Biddie, S. C., John, S., Sabo, P. J., Thurman, R. E., Johnson, T. A., Schiltz, R. L., Miranda, T. B., Sung, M. H., Trump, S., Lightman, S. L., Vinson, C., Stamatoyannopoulos, J. A. & Hager, G. L. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular cell* **43**, 145-155, doi:10.1016/j.molcel.2011.06.016 (2011).

12 Vahedi, G., Takahashi, H., Nakayamada, S., Sun, H. W., Sartorelli, V., Kanno, Y. & O'Shea, J. J. STATs shape the active enhancer landscape of T cell populations. *Cell* **151**, 981-993, doi:10.1016/j.cell.2012.09.044 (2012).

13 Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C. & Glass, C. K. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**, 310-325, doi:10.1016/j.molcel.2013.07.010 (2013).

14 Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S. & Natoli, G. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157-171, doi:10.1016/j.cell.2012.12.018 (2013).

15 Lam, M. T., Cho, H., Lesch, H. P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M. U., Kim, A. S., Kosaka, M., Lee, C. Y., Watt, A., Grossman, T. R., Rosenfeld, M. G., Evans, R. M. & Glass, C. K. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**, 511-515, doi:10.1038/nature12209 (2013).

16 Delovitch, T. L. & Singh, B. The nonobese diabetic mouse as a model of autoimmune diabetes: immune dysregulation gets the NOD. *Immunity* **7**, 727-738 (1997).

17 Stiller, C. R., Dupre, J., Gent, M., Jenner, M. R., Keown, P. A., Laupacis, A., Martell, R., Rodger, N. W., von Graffenried, B. & Wolfe, B. M. Effects of cyclosporine immunosuppression in insulin-dependent diabetes mellitus of recent onset. *Science* **223**, 1362-1367 (1984).

18 Sibley, R. K., Rynasiewicz, J., Ferguson, R. M., Fryd, D., Sutherland, D. E., Simmons, R. L. & Najarian, J. S. Morphology of cyclosporine nephrotoxicity and acute rejection in patients immunosuppressed with cyclosporine and prednisone. *Surgery* **94**, 225-234 (1983).

19 Bougneres, P. F., Landais, P., Boisson, C., Carel, J. C., Frament, N., Boitard, C., Chaussain, J. L. & Bach, J. F. Limited duration of remission of insulin dependency in children with recent overt type I diabetes treated with low-dose cyclosporin. *Diabetes* **39**, 1264-1272 (1990).

20 Clemente-Casares, X., Tsai, S., Huang, C. & Santamaria, P. Antigen-specific therapeutic approaches in Type 1 diabetes. *Cold Spring Harbor perspectives in medicine* **2**, a007773, doi:10.1101/cshperspect.a007773 (2012).

21 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848, doi:10.1126/science.1162228 (2008).

22 Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome research* **23**, 1210-1223, doi:10.1101/gr.152306.112 (2013).

23 Davis-Turak, J. C., Allison, K., Shokhirev, M. N., Ponomarenko, P., Tsimring, L. S., Glass, C. K., Johnson, T. L. & Hoffmann, A. Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. *Nucleic acids research* **43**, 699-707, doi:10.1093/nar/gku1338 (2015).

24 Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G. & Fu, X. D. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394, doi:10.1038/nature10006 (2011).

25 Zhu, Y., Sun, L., Chen, Z., Whitaker, J. W., Wang, T. & Wang, W. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic acids research* **41**, 10032-10043, doi:10.1093/nar/gkt826 (2013).

26 Allison, K. A., Kaikkonen, M. U., Gaasterland, T. & Glass, C. K. Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic Acids Res* **42**, 2433-2447, doi:10.1093/nar/gkt1237 (2014).

27 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research* **40**, D130-135, doi:10.1093/nar/gkr1079 (2012).

28 Murphy, K. P., Travers, P., Walport, M. & Janeway, C. *Janeway's Immunobiology*. (Garland Science, 2008).

29 Chen, X. & Jensen, P. E. The role of B lymphocytes as antigen-presenting cells. *Archivum immunologiae et therapiae experimentalis* **56**, 77-83, doi:10.1007/s00005-008-0014-5 (2008).

30 Hume, D. A. Macrophages as APC and the dendritic cell myth. *Journal of immunology* **181**, 5829-5835 (2008).

31 Blum, J. S., Wearsch, P. A. & Cresswell, P. Pathways of antigen processing. *Annual review of immunology* **31**, 443-473, doi:10.1146/annurev-immunol-032712-095910 (2013).

32 Heber-Katz, E., Schwartz, R. H., Matis, L. A., Hannum, C., Fairwell, T., Appella, E. & Hansburg, D. Contribution of antigen-presenting cell major histocompatibility complex gene products to the specificity of antigen-induced T cell activation. *The Journal of experimental medicine* **155**, 1086-1099 (1982).

33 Rudolph, M. G., Stanfield, R. L. & Wilson, I. A. How TCRs bind MHCs, peptides, and coreceptors. *Annual review of immunology* **24**, 419-466, doi:10.1146/annurev.immunol.23.021704.115658 (2006).

34 Sewell, A. K. Why must T cells be cross-reactive? *Nature reviews. Immunology* **12**, 669-677, doi:10.1038/nri3279 (2012).

35 Clevers, H., Alarcon, B., Wileman, T. & Terhorst, C. The T cell receptor/CD3 complex: a dynamic protein ensemble. *Annual review of immunology* **6**, 629-662, doi:10.1146/annurev.iy.06.040188.003213 (1988).

36 Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nature reviews. Immunology* **9**, 833-844, doi:10.1038/nri2669 (2009).

37 Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H. & Carlson, C. S. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**, 4099-4107, doi:10.1182/blood-2009-04-217604 (2009).

38 Jerne, N. K. The somatic generation of immune recognition. *European journal of immunology* **1**, 1-9, doi:10.1002/eji.1830010102 (1971).

39 Jerne, N. K. The Natural-Selection Theory of Antibody Formation. *Proceedings of the National Academy of Sciences of the United States of America* **41**, 849-857 (1955).

40 Mason, D. A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology today* **19**, 395-404 (1998).

41 Holland, C. J., Rizkallah, P. J., Vollers, S., Calvo-Calle, J. M., Madura, F., Fuller, A., Sewell, A. K., Stern, L. J., Godkin, A. & Cole, D. K. Minimal conformational plasticity enables TCR cross-reactivity to different MHC class II heterodimers. *Scientific reports* **2**, 629, doi:10.1038/srep00629 (2012).

42 Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I. & Davis, M. M. Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins. *Nature* **308**, 153-158 (1984).

43 Hedrick, S. M., Matis, L. A., Hecht, T. T., Samelson, L. E., Longo, D. L., Heber-Katz, E. & Schwartz, R. H. The fine specificity of antigen and Ia determinant recognition by T cell hybridoma clones specific for pigeon cytochrome c. *Cell* **30**, 141-152 (1982).

44 Solinger, A. M., Ultee, M. E., Margoliash, E. & Schwartz, R. H. T-lymphocyte response to cytochrome c. I. Demonstration of a T-cell heteroclitic proliferative response and identification of a topographic antigenic determinant on pigeon cytochrome c whose immune recognition requires two complementing major histocompatibility complex-linked immune response genes. *The Journal of experimental medicine* **150**, 830-848 (1979).

45 Corse, E., Gottschalk, R. A. & Allison, J. P. Strength of TCR-peptide/MHC interactions and in vivo T cell responses. *Journal of immunology* **186**, 5039-5045, doi:10.4049/jimmunol.1003650 (2011).

46 Schwartz, R. H. T-lymphocyte recognition of antigen in association with gene products of the major histocompatibility complex. *Annual review of immunology* **3**, 237-261, doi:10.1146/annurev.iy.03.040185.001321 (1985).

47 Stone, J. D., Chervin, A. S. & Kranz, D. M. T-cell receptor binding affinities and kinetics: impact on T-cell activity and specificity. *Immunology* **126**, 165-176, doi:10.1111/j.1365-2567.2008.03015.x (2009).

48 Govern, C. C., Paczosa, M. K., Chakraborty, A. K. & Huseby, E. S. Fast on-rates allow short dwell time ligands to activate T cells. *Proceedings of the*

National Academy of Sciences of the United States of America **107**, 8724-8729, doi:10.1073/pnas.1000966107 (2010).

49 Huppa, J. B., Axmann, M., Mortelmaier, M. A., Lillemeier, B. F., Newell, E. W., Brameshuber, M., Klein, L. O., Schutz, G. J. & Davis, M. M. TCR-peptide-MHC interactions in situ show accelerated kinetics and increased affinity. *Nature* **463**, 963-967, doi:10.1038/nature08746 (2010).

50 Kersh, G. J., Kersh, E. N., Fremont, D. H. & Allen, P. M. High- and low-potency ligands with similar affinities for the TCR: the importance of kinetics in TCR signaling. *Immunity* **9**, 817-826 (1998).

51 Lyons, D. S., Lieberman, S. A., Hampl, J., Boniface, J. J., Chien, Y., Berg, L. J. & Davis, M. M. A TCR binds to antagonist ligands with lower affinities and faster dissociation rates than to agonists. *Immunity* **5**, 53-61 (1996).

52 Newell, E. W., Ely, L. K., Kruse, A. C., Reay, P. A., Rodriguez, S. N., Lin, A. E., Kuhns, M. S., Garcia, K. C. & Davis, M. M. Structural basis of specificity and cross-reactivity in T cell receptors specific for cytochrome c-I-E(k). *Journal of immunology* **186**, 5823-5832, doi:10.4049/jimmunol.1100197 (2011).

53 Aleksic, M., Dushek, O., Zhang, H., Shenderov, E., Chen, J. L., Cerundolo, V., Coombs, D. & van der Merwe, P. A. Dependence of T cell antigen recognition on T cell receptor-peptide MHC confinement time. *Immunity* **32**, 163-174, doi:10.1016/j.immuni.2009.11.013 (2010).

54 Krogsgaard, M., Prado, N., Adams, E. J., He, X. L., Chow, D. C., Wilson, D. B., Garcia, K. C. & Davis, M. M. Evidence that structural rearrangements and/or flexibility during TCR binding can contribute to T cell activation. *Molecular cell* **12**, 1367-1378 (2003).

55 De Magistris, M. T., Alexander, J., Coggeshall, M., Altman, A., Gaeta, F. C., Grey, H. M. & Sette, A. Antigen analog-major histocompatibility complexes act as antagonists of the T cell receptor. *Cell* **68**, 625-634 (1992).

56 Rogers, P. R., Grey, H. M. & Croft, M. Modulation of naive CD4 T cell activation with altered peptide ligands: the nature of the peptide and presentation in the context of costimulation are critical for a sustained response. *Journal of immunology* **160**, 3698-3704 (1998).

57 Baumgartner, C. K., Ferrante, A., Nagaoka, M., Gorski, J. & Malherbe, L. P. Peptide-MHC class II complex stability governs CD4 T cell clonal selection. *Journal of immunology* **184**, 573-581, doi:10.4049/jimmunol.0902107 (2010).

58 Huang, J., Zarnitsyna, V. I., Liu, B., Edwards, L. J., Jiang, N., Evavold, B. D. & Zhu, C. The kinetics of two-dimensional TCR and pMHC interactions determine T-cell responsiveness. *Nature* **464**, 932-936, doi:10.1038/nature08944 (2010).

59 Cemerski, S., Das, J., Locasale, J., Arnold, P., Giurisato, E., Markiewicz, M. A., Fremont, D., Allen, P. M., Chakraborty, A. K. & Shaw, A. S. The stimulatory potency of T cell antigens is influenced by the formation of the immunological synapse. *Immunity* **26**, 345-355, doi:10.1016/j.immuni.2007.01.013 (2007).

60 Edwards, L. J. & Evavold, B. D. T cell recognition of weak ligands: roles of signaling, receptor number, and affinity. *Immunologic research* **50**, 39-48, doi:10.1007/s12026-011-8204-3 (2011).

61 Alexander, J., Snoke, K., Ruppert, J., Sidney, J., Wall, M., Southwood, S., Oseroff, C., Arrhenius, T., Gaeta, F. C., Colon, S. M. & et al. Functional consequences of engagement of the T cell receptor by low affinity ligands. *Journal of immunology* **150**, 1-7 (1993).

62 Evavold, B. D., Sloan-Lancaster, J., Hsu, B. L. & Allen, P. M. Separation of T helper 1 clone cytolysis from proliferation and lymphokine production using analog peptides. *Journal of immunology* **150**, 3131-3140 (1993).

63 La Face, D. M., Couture, C., Anderson, K., Shih, G., Alexander, J., Sette, A., Mustelin, T., Altman, A. & Grey, H. M. Differential T cell signaling induced by antagonist peptide-MHC complexes and the associated phenotypic responses. *Journal of immunology* **158**, 2057-2064 (1997).

64 Vanguri, V., Govern, C. C., Smith, R. & Huseby, E. S. Viral antigen density and confinement time regulate the reactivity pattern of CD4 T-cell responses to vaccinia virus infection. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 288-293, doi:10.1073/pnas.1208328110 (2013).

65 Keck, S., Schmaler, M., Ganter, S., Wyss, L., Oberle, S., Huseby, E. S., Zehn, D. & King, C. G. Antigen affinity and antigen dose exert distinct influences on CD4 T-cell differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 14852-14857, doi:10.1073/pnas.1403271111 (2014).

66 Marchingo, J. M., Kan, A., Sutherland, R. M., Duffy, K. R., Wellard, C. J., Belz, G. T., Lew, A. M., Dowling, M. R., Heinzel, S. & Hodgkin, P. D. T cell signaling. Antigen affinity, costimulation, and cytokine inputs sum linearly to

amplify T cell expansion. *Science* **346**, 1123-1127, doi:10.1126/science.1260044 (2014).

67 Rogers, P. R. & Croft, M. Peptide dose, affinity, and time of differentiation can contribute to the Th1/Th2 cytokine balance. *Journal of immunology* **163**, 1205-1213 (1999).

68 Sloan-Lancaster, J., Evavold, B. D. & Allen, P. M. Induction of T-cell anergy by altered T-cell-receptor ligand on live antigen-presenting cells. *Nature* **363**, 156-159, doi:10.1038/363156a0 (1993).

69 Tubo, N. J., Pagan, A. J., Taylor, J. J., Nelson, R. W., Linehan, J. L., Ertelt, J. M., Huseby, E. S., Way, S. S. & Jenkins, M. K. Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. *Cell* **153**, 785-796, doi:10.1016/j.cell.2013.04.007 (2013).

70 Sloan-Lancaster, J., Shaw, A. S., Rothbard, J. B. & Allen, P. M. Partial T cell signaling: altered phospho-zeta and lack of zap70 recruitment in APL-induced T cell anergy. *Cell* **79**, 913-922 (1994).

71 Sloan-Lancaster, J., Evavold, B. D. & Allen, P. M. Th2 cell clonal anergy as a consequence of partial activation. *The Journal of experimental medicine* **180**, 1195-1205 (1994).

72 Tsitoura, D. C., Gelder, C. M., Kemeny, D. M. & Lamb, J. R. Regulation of cytokine production by human Th0 cells following stimulation with peptide analogues: differential expression of TGF-beta in activation and anergy. *Immunology* **92**, 10-19 (1997).

73 Tsitoura, D. C., Holter, W., Cerwenka, A., Gelder, C. M. & Lamb, J. R. Induction of anergy in human T helper 0 cells by stimulation with altered T cell antigen receptor ligands. *Journal of immunology* **156**, 2801-2808 (1996).

74 Nagaoka, M., Hatta, Y., Kawaoka, Y. & Malherbe, L. P. Antigen signal strength during priming determines effector CD4 T cell function and antigen sensitivity during influenza virus challenge. *Journal of immunology* **193**, 2812-2820, doi:10.4049/jimmunol.1401358 (2014).

75 Schmid, D. A., Irving, M. B., Posevitz, V., Hebeisen, M., Posevitz-Fejfar, A., Sarria, J. C., Gomez-Eerland, R., Thome, M., Schumacher, T. N., Romero, P., Speiser, D. E., Zoete, V., Michielin, O. & Rufer, N. Evidence for a TCR affinity threshold delimiting maximal CD8 T cell function. *Journal of immunology* **184**, 4936-4946, doi:10.4049/jimmunol.1000173 (2010).

76 Knudson, K. M., Goplen, N. P., Cunningham, C. A., Daniels, M. A. & Teixeiro, E. Low-affinity T cells are programmed to maintain normal primary responses but are impaired in their recall to low-affinity ligands. *Cell reports* **4**, 554-565, doi:10.1016/j.celrep.2013.07.008 (2013).

77 Koehli, S., Naeher, D., Galati-Fournier, V., Zehn, D. & Palmer, E. Optimal T-cell receptor affinity for inducing autoimmunity. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 17248-17253, doi:10.1073/pnas.1402724111 (2014).

78 Huang, Y. & Wange, R. L. T cell receptor signaling: beyond complex complexes. *The Journal of biological chemistry* **279**, 28827-28830, doi:10.1074/jbc.R400012200 (2004).

79 Love, P. E. & Hayes, S. M. ITAM-mediated signaling by the T-cell antigen receptor. *Cold Spring Harbor perspectives in biology* **2**, a002485, doi:10.1101/cshperspect.a002485 (2010).

80 Chan, A. C., Iwashima, M., Turck, C. W. & Weiss, A. ZAP-70: a 70 kd protein-tyrosine kinase that associates with the TCR zeta chain. *Cell* **71**, 649-662 (1992).

81 Murphy, T. L., Tussiwand, R. & Murphy, K. M. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nature reviews. Immunology* **13**, 499-509, doi:10.1038/nri3470 (2013).

82 Rincon, M. & Flavell, R. A. AP-1 transcriptional activity requires both T-cell receptor-mediated and co-stimulatory signals in primary T lymphocytes. *The EMBO journal* **13**, 4370-4381 (1994).

83 Crabtree, G. R. & Olson, E. N. NFAT signaling: choreographing the social lives of cells. *Cell* **109 Suppl**, S67-79 (2002).

84 Fu, G., Casas, J., Rigaud, S., Rybakin, V., Lambolez, F., Brzostek, J., Hoerter, J. A., Paster, W., Acuto, O., Cheroutre, H., Sauer, K. & Gascoigne, N. R. Themis sets the signal threshold for positive and negative selection in T-cell development. *Nature* **504**, 441-445, doi:10.1038/nature12718 (2013).

85 Daniels, M. A., Teixeiro, E., Gill, J., Hausmann, B., Roubaty, D., Holmberg, K., Werlen, G., Hollander, G. A., Gascoigne, N. R. & Palmer, E. Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling. *Nature* **444**, 724-729, doi:10.1038/nature05269 (2006).

86 Prasad, A., Zikherman, J., Das, J., Roose, J. P., Weiss, A. & Chakraborty, A. K. Origin of the sharp boundary that discriminates positive and negative

selection of thymocytes. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 528-533, doi:10.1073/pnas.0805981105 (2009).

87 Das, J., Ho, M., Zikherman, J., Govern, C., Yang, M., Weiss, A., Chakraborty, A. K. & Roose, J. P. Digital signaling and hysteresis characterize ras activation in lymphoid cells. *Cell* **136**, 337-351, doi:10.1016/j.cell.2008.11.051 (2009).

88 Morris, G. P. & Allen, P. M. How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature immunology* **13**, 121-128, doi:10.1038/ni.2190 (2012).

89 Warnecke, N., Poltorak, M., Kowtharapu, B. S., Arndt, B., Stone, J. C., Schraven, B. & Simeoni, L. TCR-mediated Erk activation does not depend on Sos and Grb2 in peripheral human T cells. *EMBO reports* **13**, 386-391, doi:10.1038/embor.2012.17 (2012).

90 Coward, J., Germain, R. N. & Altan-Bonnet, G. Perspectives for computer modeling in the study of T cell activation. *Cold Spring Harbor perspectives in biology* **2**, a005538, doi:10.1101/cshperspect.a005538 (2010).

91 Kingeter, L. M., Paul, S., Maynard, S. K., Cartwright, N. G. & Schaefer, B. C. Cutting edge: TCR ligation triggers digital activation of NF-kappaB. *Journal of immunology* **185**, 4520-4524, doi:10.4049/jimmunol.1001051 (2010).

92 Marangoni, F., Murooka, T. T., Manzo, T., Kim, E. Y., Carrizosa, E., Elpek, N. M. & Mempel, T. R. The transcription factor NFAT exhibits signal memory during serial T cell interactions with antigen-presenting cells. *Immunity* **38**, 237-249, doi:10.1016/j.immuni.2012.09.012 (2013).

93 Podtschaske, M., Benary, U., Zwinger, S., Hofer, T., Radbruch, A. & Baumgrass, R. Digital NFATc2 activation per cell transforms graded T cell receptor activation into an all-or-none IL-2 expression. *PloS one* **2**, e935, doi:10.1371/journal.pone.0000935 (2007).

94 Au-Yeung, B. B., Zikherman, J., Mueller, J. L., Ashouri, J. F., Matloubian, M., Cheng, D. A., Chen, Y., Shokat, K. M. & Weiss, A. A sharp T-cell antigen receptor signaling threshold for T-cell proliferation. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E3679-3688, doi:10.1073/pnas.1413726111 (2014).

95 Huang, J., Brameshuber, M., Zeng, X., Xie, J., Li, Q. J., Chien, Y. H., Valitutti, S. & Davis, M. M. A single peptide-major histocompatibility complex ligand triggers digital cytokine secretion in CD4(+) T cells. *Immunity* **39**, 846-857, doi:10.1016/j.immuni.2013.08.036 (2013).

96 Zikherman, J. & Au-Yeung, B. The role of T cell receptor signaling thresholds in guiding T cell fate decisions. *Current opinion in immunology* **33C**, 43-48, doi:10.1016/j.coi.2015.01.012 (2015).

97 Butler, T. C., Kardar, M. & Chakraborty, A. K. Quorum sensing allows T cells to discriminate between self and nonself. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 11833-11838, doi:10.1073/pnas.1222467110 (2013).

98 Kersh, E. N., Kersh, G. J. & Allen, P. M. Partially phosphorylated T cell receptor zeta molecules can inhibit T cell activation. *The Journal of experimental medicine* **190**, 1627-1636 (1999).

99 Kersh, E. N., Shaw, A. S. & Allen, P. M. Fidelity of T cell activation through multistep T cell receptor zeta phosphorylation. *Science* **281**, 572-575 (1998).

100 Vidal, K., Daniel, C., Hill, M., Littman, D. R. & Allen, P. M. Differential requirements for CD4 in TCR-ligand interactions. *Journal of immunology* **163**, 4811-4818 (1999).

101 Irvine, D. J., Purbhoo, M. A., Krogsgaard, M. & Davis, M. M. Direct observation of ligand recognition by T cells. *Nature* **419**, 845-849, doi:10.1038/nature01076 (2002).

102 Man, K., Miasari, M., Shi, W., Xin, A., Henstridge, D. C., Preston, S., Pellegrini, M., Belz, G. T., Smyth, G. K., Febbraio, M. A., Nutt, S. L. & Kallies, A. The transcription factor IRF4 is essential for TCR affinity-mediated metabolic programming and clonal expansion of T cells. *Nature immunology* **14**, 1155-1165, doi:10.1038/ni.2710 (2013).

103 Nayar, R., Schutten, E., Bautista, B., Daniels, K., Prince, A. L., Enos, M., Brehm, M. A., Swain, S. L., Welsh, R. M. & Berg, L. J. Graded levels of IRF4 regulate CD8+ T cell differentiation and expansion, but not attrition, in response to acute virus infection. *Journal of immunology* **192**, 5881-5893, doi:10.4049/jimmunol.1303187 (2014).

104 Kaye, J., Hsu, M. L., Sauron, M. E., Jameson, S. C., Gascoigne, N. R. & Hedrick, S. M. Selective development of CD4+ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature* **341**, 746-749, doi:10.1038/341746a0 (1989).

CHAPTER ONE

Vespucci: a system for building annotated databases of nascent transcripts

Global run-on sequencing (GRO-seq) is a recent addition to the series of high throughput sequencing methods that enables new insights into transcriptional dynamics within a cell. However, GRO-sequencing presents new algorithmic challenges, as existing analysis platforms for ChIP-seq and RNA-seq do not address the unique problem of identifying transcriptional units *de novo* from short reads located all across the genome. Here, we present a novel algorithm for *de novo* transcript identification from GRO-sequencing data, along with a system that determines transcript regions, stores them in a relational database, and associates them with known reference annotations. We use this method to analyze GRO-sequencing data from primary mouse macrophages, and derive novel quantitative insights into the extent and characteristics of non-coding transcription in mammalian cells. In doing so, we demonstrate that Vespucci expands existing annotations for mRNAs and lincRNAs by defining the primary transcript beyond the polyadenylation site. In addition, Vespucci generates assemblies for un-annotated non-coding RNAs such as those transcribed from enhancer-like elements. Vespucci thereby provides a robust system for defining, storing, and analyzing diverse classes of primary RNA transcripts that are of increasing biological interest.

INTRODUCTION

High-throughput sequencing has opened up a new window into transcriptional biology and the complex regulatory networks that define RNA and DNA interactions. Global run-on sequencing (GRO-seq) [1] is a recent addition to the series of sequencing-based methods that holds particular promise for understanding real-time transcriptional behavior. GRO-seq captures a point-in-time snapshot of active transcription genome-wide and returns data on the position, length, and orientation of nascent transcripts.

This sequencing technique is now being employed to inspect the nature of transcriptional regulation in a number of experimental conditions [1-4]. The capture of nascent transcripts in each of these conditions reveals a variety of RNA species beyond the standard set derived from genes encoding proteins and microRNAs, including enhancer RNA (eRNA), long intergenic RNA (lincRNA) [2], and promoter-associated RNA [1,5]. GRO-seq thus offers unprecedented insight into the generation of a vast repertoire of non-coding transcripts that are of potential functional significance.

The data collected, however, is both immense and unique; each experiment yields tens of millions of strand-specific short RNA reads across the entire genome. This new sequencing method presents a new algorithmic challenge, as the peak-calling and exonic RNA identification techniques developed for other sequencing methods do not address the particular output of GRO-seq. Unlike ChIP-seq, peaks are not the primary unit of output, and, unlike RNA-seq, nascent transcripts can be anywhere, so

relying on previously annotated regions such as NCBI Reference Sequence (RefSeq) [6] or microRNA genes is insufficient.

To take full advantage of this novel data, regions beyond existing annotations must be considered. Units of transcription must be inferred *de novo* from the short read output of GRO-sequencing experiments. Existing analysis of GRO-seq data relies largely on adaptations of RNA-sequencing analysis techniques, with expression levels calculated from tag counts over gene bodies, promoters, or other explicitly defined genomic regions [1-4,7]. New transcripts can be identified using software such as Cufflinks [8], but these rely on assumptions optimized for spliced RNA. For example, Cufflinks is optimized for paired-end reads; expects uniform density for a given transcript (whereas GRO-seq can reveal pausing and other biologically-relevant deviations from uniformity); and aims to accommodate large gaps (introns) in reads that result from splicing rather than from transcriptional breaks. In short, Cufflinks and similar exon-focused algorithms are not suited to distinguish between the sorts of small and closely spaced regulatory elements that GRO-sequencing reveals.

Hah *et al.* have developed a Hidden Markov Model (HMM) for identification of regions of transcription specifically within GRO-seq data [2]. The software demarcates transcripts using a two-state model, calling regions either "transcribed" or "un-transcribed," and thus is able to identify transcripts from GRO-sequencing short reads *de novo*. However, the HMM is optimized to accurately retrieve transcript boundaries as defined by RefSeq, resulting in the loss or merging of many of the shorter, non-coding RNA transcripts that GRO-sequencing reveals. Further, because

the software relies on flat files for processing and storage, it is difficult to integrate the called transcripts with other types of genomic data, including expression levels from each individual GRO-sequencing experiment and co-occurring peaks from ChIP-sequencing data.

Here, we provide an algorithm for *de novo* identification of unified transcripts from GRO-seq data, along with an implementation that determines transcript regions, stores them in a relational database, and associates them with known reference annotations according to two-dimensional genomic overlap. Crucially, this method captures transcript boundaries as defined by RefSeq while maintaining the ability to identify non-coding RNAs at a high resolution and even retaining information about relative transcript abundance. Further, transcript identification feeds into a database that makes downstream integration of other datatypes feasible.

Using this system, we were able to gain new insight into the types of nascent RNAs being generated inside primary murine macrophages. While the ENCODE Project has begun the process of characterizing mature RNA species [9], there is very little known about the extent and distribution of nascent RNAs, which, unlike mRNAs observed in traditional RNA-sequencing, include a number of transient RNA species that nonetheless play roles in the regulation of gene expression [1,10-13]. Of particular interest are the vast numbers of non-coding RNAs recently found to be derived from transcription of active enhancers [14,15]. The finding that at least some of these eRNAs contribute to enhancer function provides impetus for developing computational tools to define the sites of initiation of these species and their length. Importantly, while the

start and termination sites of transcripts related to mRNA-encoding genes and lincRNAs have for the most part been established by conventional RNA sequencing studies, this information is virtually non-existent for eRNAs. Furthermore, the ENCODE consortium estimates that the human genome contains hundreds of thousands of enhancers [16], the majority of which are selected in a cell-specific manner. Therefore each GRO-seq experiment in a new cell type results in the identification of tens of thousands of previously unannotated eRNAs that are derived from transcription of cell specific enhancers. To address this challenge, we developed Vespucci as a computational method to systematically and quantitatively define discreet nascent transcripts from short sequencing reads obtained in GRO-seq experiments. By tuning parameters for specific types of transcripts, Vespucci returns accurate calls for primary mRNAs, while also deconvoluting complex patterns of transcription from enhancer-rich regions of the genome. Using Vespucci, we provide evidence that many nascent mRNA transcripts extend well beyond RefSeq annotated termination sites. In addition, Vespucci predicts approximately twice as many non-coding transcripts as were identified by other systems like the Hah et al. HMM. These findings demonstrate the value of Vespucci in integrating disparate data types in order to characterize the variety of RNA species observed.

The Python and PostgreSQL code, as well as a pre-loaded Amazon AMI, have been made available for implementation and expansion by interested researchers.

METHODS

Technical details

The current implementation allows sample types and database schema to be split easily by cell type, such that the merging of transcripts is confined to a single cell type. Still, GRO-sequencing runs from multiple cell types can be easily merged together if desired.

The current codebase assumes a PostgreSQL 9.2 database installation; Python 2.7+; and Django 1.2+ with psycopg2 for database access. The codebase is hosted on Github at https://github.com/karmel/vespucci, and includes scripts to build both the transcript and the annotation databases. Instructions are included within the repository. In addition, a pre-loaded Amazon EC2 small instance image is available with instructions at https://github.com/karmel/vespucci.

Cell culture

Primary cells were isolated from 6-8 week-old C57Bl/6 mice. All studies were conducted in accordance with the UCSD Institutional Animal Care and Use Committee. Thioglycollate-elicited macrophages were isolated by peritoneal lavage 3-4 days following peritoneal injection of 2.5 ml thioglycollate. Cells were plated in RPMI medium 1640 and 10% fetal bovine serum, washed after adherence and again fed with fresh medium. The following day fresh medium containing 0.5% fetal bovine serum was added to the cells and serum starvation was carried overnight

GRO-seq library preparation

Briefly, GRO-sequencing takes advantage of a nuclear run-on reaction to incorporate tagged UTP into ongoing transcript synthesis by RNA polymerase. RNAs that incorporate the tagged nucleotides can subsequently be extracted and sequenced, producing a genome-wide library of nascent RNAs. Thus, in contrast to traditional RNA-sequencing, in which mature, stable RNAs are collected, GRO-sequencing returns short read data for RNAs in the act of being transcribed.

Global run-on [1] and library preparation for sequencing [17] were done as described. The protocol was performed as described in Wang *et al*. [3]

Previously published GRO-seq data

Four of the replicates were previously published under GSE48759 [4]. The GEO Accession codes are: GSM1183906 - GSM1183908 and GSM1183914. The MCF-7 GRO-seq data is available under GSE27463 [2], Accession codes GSM678535 - GSM678540; and GSE45822 [12], Accession codes GSM1115995 - GSM1115998.

Read mapping and ChIP-Seq data analysis

Reads were mapped to the mm9 genome using Bowtie2 [18] with the default alignment options (specifically, the command bowtie2 –no-unal –x).

H3K4me1 and input data were taken from GSE21512 [7], Accession codes GSM537986 and GSM537988. MCF-7 H3K4me2 data and input were taken from GSE24166 [19], Acession codes GSM594606 and GSM594608. Peaks were called using HOMER [7] using the command findPeaks and the options -nfr -style histone.

Case study counts

The SQL queries used to generate the counts for the analysis of transcription in macrophages are included as a Supplementary File.

RESULTS

Defining a transcript

Principles

As with most next-generation sequencing-based methods, GRO-seq relies on short (35 - 100 base-pair) reads. For the purposes of this study, we assume each individual read is mapped to the canonical genome of the organism in question with a standard aligner such as Bowtie [20]. Any given uniquely-mappable read, then, can be placed into one-dimensional space with a definitive coordinate consisting of chromosome, strand, start of read, and end of read (Figure 1a).

The location of each short read does not alone describe the relevant units of transcription in the genome; overlapping sets of short reads must be computationally merged so that they represent the extents of biologically relevant transcripts, which here we take to mean linear segments of DNA that are transcribed into continuous RNA sequences by RNA polymerase II [9]. Once merged into continuous units, the count of short reads mapping to a given unit (transcript) can be used to approximate the relative expression level of the transcript [21].

A primary challenge in any short read RNA sequencing application is determining how to merge the fragments into unified transcripts. Each type of sequencing presents unique challenges in this regard; in the case of mRNA-seq, for example, methods have been developed that are designed to identify exon junctions [8,22]. GRO-seq reads, in contrast, are expected to extend through intragenic regions, and further are expected to exist widely both in intergenic regions and in regions antisense to annotated transcripts [1].

There are numerous patterns of GRO-seq data that are important to identify computationally. For example, promoter-associated RNA transcripts are generated at the promoters of genes, antisense to the genic transcript [1] (Figure 1b). Any algorithm addressing GRO-seq needs to identify these RNAs as distinct units, overlapping with but not part of either genic transcripts or nearby eRNAs. Similarly, eRNAs are generated bi-directionally at enhancers [23], and any algorithm must identify each strand of eRNA as a separate but contiguous unit (Figure 1c).

Notably, some transcripts appearing in GRO-seq data seem "obvious" to separate when viewed in the UCSC browser, as with Figures 1b and 1c. However, these cases are the minority, and, further, any such "obvious" separation is *ad hoc* and risks inconsistency when performed manually; in Figure 1d, for example, most observers would not separate the transcript, but existing annotation data from RefSeq indicates there is in fact an important boundary corresponding to a coding sequence. Thus, it is useful to have an algorithmic interpretation that provides a standardized

analysis and additionally can appeal to existing annotation data if available (Figure 1d).

If there are no existing annotations from which to scaffold the current transcript identification, the algorithm must have a standard means of interpreting the short read data that is likely to reflect the biological reality of the transcriptional data. To this end, the present implementation makes several assumptions based on the expected behavior of RNA polymerase: first, that regions that are tiled without gaps by short reads from a single sequencing run are most likely continuously transcribed; and second, that gaps corresponding with a great disparity of read counts per base-pair likely represent breaks in the path of RNA polymerase, with differently regulated transcripts on either side. (Figure S1a shows a schematic of how differential density might yield separate transcripts, and Figure 1e shows an example of this in real GRO-sequencing data, where transcription along each strand in the displayed region is split into two separate transcripts due to differential coverage.)

These two principles— that overlapping reads should be merged and that disparity in the density of reads may warrant separation of otherwise close transcripts— motivate the design of the algorithm described below.

Implementation

Given the size of GRO-seq datasets, with every sample yielding at least tens of millions of reads, any algorithm must be implemented within a framework that is easily maintained and extended, and it must process data quickly enough to be useful

in a laboratory setting. Further, the transcript identification system must be architected such that new samples can be incrementally added to the full set of data without requiring re-processing of all data. To this end, we have developed a Python codebase that processes mapped short read files from GRO-seq experiments into continuous transcript units, determines relative expression levels on a sample-by-sample basis, and stores the data in a PostgreSQL relational database that allows for complex coordinate-based queries over the transcriptional data.

The procedure relies on two key parameters:

- DENSITY_MULTIPLIER: scaling factor to relate density to base-pairs (see step 4 below; default: 10,000). Intuitively, this is the number of base-pairs over which density is considered, so that a difference in one tag per DENSITY_MULTIPLIER base-pairs equates to a one base-pair gap in genomic distance.
- MAX_EDGE: maximum allowed distance in two-dimensional space between proto-transcripts to be stitched together (see step 5 below; default: 500).

The selection of values for these two parameters depends heavily on the desired use case. The larger the value of DENSITY_MULTIPLIER, the more density matters as compared to distance in base-pairs, and the larger the value of MAX_EDGE, the more likely distant transcripts are to be merged into single units. Thus, if the user desires to focus on large transcripts and genes, she might choose a low value for DENSITY_MULTIPLIER and a large value for MAX_EDGE. On the

other hand, if the user desires to focus on small transcripts and ncRNA, she might choose a high value for DENSITY_MULTIPLIER and a small value for MAX_EDGE. The selection of the default values of these parameters, and the values used for the data in this study, are discussed below under 'Parameter Selection.'

Once these parameters are set, the processing of reads into transcripts proceeds as follows for each strand of each chromosome (Figure 2a):

1. Given a mapped tag file (in BAM or SAM [24] format), each tag is reduced to its genomic coordinates and loaded into a database table. The tables are designed such that the dataset for each sample is stored in a separate table.
2. Once loaded, the tags from a single sample are merged, but no analysis of density is attempted. Although individual tag boundaries are not maintained in the merged format, the count of reads and number of gaps between the reads that are merged are tracked for expression level comparisons later.
3. The set of unified transcripts from a single sample are then merged with transcripts from all existing samples.
4. Using the stored tag counts and the genomic coordinates of the merged read, each proto-transcript is mapped as a horizontal line segment in two dimensional space, with the start and end serving as the coordinates along the x-axis, and the density (tags in all runs per base-pair) as the coordinate along the y-axis. Density is scaled by a parameter

(DENSITY_MULTIPLIER); a higher multiplier increases the relative importance of density as compared to position.

5. A second stage of merging begins over the two-dimensional space according to the algorithm described below and the MAX_EDGE parameter. At this stage, several optimization checks filter out proto-transcripts that are likely noise, such as those that have fewer than one tag per sample on average.
6. Transcripts are associated with annotation databases as described in Part 2 below.
7. Transcripts are scored. Any scoring algorithm could be implemented here, but currently two are included:
    - A standard reads per kilobase per million tags (RPKM) score assignment
    - A custom, length-sensitive algorithm:

$$\text{Score} = \text{RPKM} * \log_{100}[\max(1, \text{length} - 200)]$$

This custom score has several modifications as compared to RPKM that make it more sensitive to certain kinds of transcripts:

1. Very short transcripts (less than 200bp) are set to a score of zero. This reduces noise from overlaps of several reads that get stitched together, and from technical artifacts.

2. Long transcripts are handicapped. There are many long transcripts with low levels of transcription that are nonetheless interesting (Figure S1b). RPKM alone has a tendency to decrease with transcript length (Figure S2a), and thus it is difficult to filter out short, noisy transcripts without losing long transcripts for which we accept lower levels of transcription. To address this problem, the custom score scales the RPKM by the log of the length of the transcript, thereby leveling out the scores of very long transcripts (Figure S2b). We use a logarithm with base 100 here in order to ensure that the score scales only minimally over the extremely wide range of transcript lengths.

Note that the choice of using the RPKM or the custom score depends largely on use case; if transcripts less than 200bp long are of particular interest, as might be the case if one were studying pause-release mechanisms using GRO-seq [25], then it would be advisable to use unmodified RPKM instead of the custom score.

When the processing is complete, the derived transcripts can be easily queried, annotated, and associated using a relational database for storage (Figure 2b).

Algorithm

In order to stitch together continuous proto-transcripts in a density-aware manner, we first map the proto-transcripts in two-dimensional space: along the x-axis

is the start and stop, in zero-indexed base-pairs of the read, and the y-axis represents the mean density of short reads over all samples (Figure 2a(4)). In order to relate base-pairs distance to read density, the density is scaled by the DENSITY_MULTIPLIER. This graphical representation allows us to define the distance between any two transcripts as simply the Euclidian distance in this density-base-pair plane (Figure 2a(5)).

Arguably, it is not necessary to relate density and base-pair distance in this manner, and an alternate distance formula could consider separate thresholds for the differences in densities and positions of two proto-transcripts. However, such a distance function would not account for the biologist's intuition that the closer two transcripts are, the more likely they are to be a single unit, even if there is a density difference that might at a greater positional difference warrant separation of two transcripts. In other words, the difference allowed between densities of two proto-transcripts when merging is dependent upon the base-pair distance between the two, and thus considering the Euclidian distance is preferable to a binary threshold that treats density and base-pairs independently.

With a distance function thus specified, we can define our algorithm for merging proto-transcripts into transcripts:

1. We define a graph in which every node is a proto-transcript, and a pair of nodes is connected by an edge if and only if the distance between the two proto-transcripts is less than or equal to the parameter MAX_EDGE.
2. Connected components in this graph represent merged proto-transcripts.

3. Merged proto-transcripts can then be recast as intervals spanning the minimal base-pair start and the maximal base-pair end. Overlapping intervals are merged.

At the end of this procedure, we have produced a set of continuous, non-overlapping transcripts that can be stored, annotated, and so on, as seen in Figure 2a(5) – Figure 2a(7).

A naive algorithm would be quadratic, comparing every node to every other. However, in practice, nodes are ordered, and it is only necessary to consider nodes within a distance of MAX_EDGE. Thus, the algorithm can be practically implemented in linear time with respect to the number of proto-transcripts. In the current implementation, we take advantage of the geometric query space in PostgreSQL to limit the search for neighboring proto-transcripts to a distance of MAX_EDGE.

Annotating a transcript

Using known RefSeq

The procedure described above can proceed naively– that is, based entirely on two-dimensional distance between transcripts and without awareness of existing annotations. In practice, it is useful for the implementation of the algorithm to respect the existing boundaries of genes as annotated by RefSeq, as this allows tag counts and computed expression values to be relevant in the context of the existing literature on gene body based expression comparisons.

Thus, the current implementation of the algorithm makes two important allowances for RefSeq genes. In the first, the allowed distance between proto-transcripts that is traversed during the two-dimensional merging in step 5 can be increased within the boundaries of known RefSeq genes such that gaps are more likely to be covered within genes. This extra allowance increases the likelihood that long, low-expression transcripts are recognized as single units rather than a series of small, gapped transcripts.

The second heuristic applied to the identification of previously annotated transcripts addresses the continuation of transcription past the traditional transcription termination site. In GRO-seq data, we see clearly that transcription does not always stop at the point corresponding to the annotated gene end, but rather continues on for some distance (Figure 1d). In these cases, we may want to be able to compare GRO-seq expression counts in genes to the measurements made in previous RNA expression studies, and thus force a separation between tags falling within RefSeq boundaries and those that extend beyond the boundaries, even if the signal is continuous according to the general rules of merging outlined above. Vespucci can be configured either to force the transcript to be segmented according to RefSeq boundaries (lower blue track in Figure 1d), so that comparisons can be made to more traditional expression data, or to assemble the transcript regardless of the annotated RefSeq boundaries (upper blue track in Figure 1d), so that the full nascent transcript can be analyzed. In the case where segmentation along RefSeq boundaries is forced, the post-gene transcript is

linked via an index to its preceding gene transcript, and we have used this option in Vespucci in the case study analyses below.

Annotation from known databases

In addition to segmentation according to known annotations like RefSeq, it is useful to be able to associate the known annotations with the transcripts that overlap in genomic space. Thus, the current implementation includes logic not only to define transcripts according to RefSeq boundaries, but also to associate RefSeq identifiers with overlapping transcripts strand-specifically. Similarly, we provide logic and data to make associations with non-coding RNA as identified by the Functional RNA Project [26].

Arbitrary data types

The representation of transcripts in terms of genomic coordinates gives power beyond associating with existing annotations. Arbitrary data types, such as peaks identified in individual ChIP-sequencing experiments, repeat regions, conservation scores, and ESTs could all be represented in terms of genomic coordinates and used to annotate transcripts, either within the existing framework, or *ad hoc*. There are many examples included in Supplemental SQL Queries that collectively demonstrate that the power of the current system is its ability integrate expression data across many samples with multiple types of annotative or associative data based on genomic location and distance quickly and easily.

A case study– transcription in macrophages

Transcriptional profiling accomplished by the ENCODE project has revealed that about three-quarters of the genome is transcribed across fifteen human cell lines [9]. Using GRO-seq data from five biological replicates, we analyzed the characteristics of transcription in murine thioglycollate-elicited macrophages using Vespucci, with the intent of characterizing the extent of transcription in a particular primary cell type under unstimulated conditions.

Parameter selection

In order to optimize the selection of the MAX_EDGE and DENSITY_MULTIPLIER parameters, we took advantage of previously published 5'-GRO-seq data [10], which identifies nascent RNA with a 5' 7-methylguanylated cap. The 5'-GRO-seq method thus produces peaks that identify transcription initiation sites of nascent RNAs genome-wide. The data available was in RAW 264.7 cells, which are a macrophage cell line, and thus were expected to be compatible with the primary macrophage GRO-seq data. Because 5'-GRO-seq identifies transcript initiation sites, we would expect transcripts identified by Vespucci to have maximally one 5'-GRO-seq peak; having more than one would be an indicator that the Vespucci transcript had merged together multiple separate units. Conversely, having zero 5'-GRO-seq peaks within a Vespucci transcript could indicate that noise was falsely assembled into a transcript, that a continuous transcript was divided into many transcripts, or that the

two sequencing techniques differ in sensitivity. We desired therefore to select parameters that would maximize the rate at which identified transcripts corresponded with exactly one 5'-GRO-seq peak. Further, in order to avoid advantaging parameters that achieved this higher rate by greatly reducing the total number of transcripts, we added a penalty for the rate at which transcripts were identified with more than one 5'-GRO-seq peak. The resultant metric, which we labeled the Initiation Recapture Rate (IRR), is defined as:

$$\text{IRR} = \frac{(\text{transcripts with one 5'-GRO peak} - \text{transcripts with more than one 5'-GRO peak})}{\text{total transcripts}}$$

We then tested values of MAX_EDGE in the range of 100 – 5,000, and found that the maximum IRR was achieved at a MAX_EDGE of 500 (Figure S1c). Then, holding MAX_EDGE constant at 500, we tested values of DENSITY_MULTIPLIER in the range of 1,000 – 100,000, and found that the maximum IRR was achieved at 10,000 (Figure S1d). Thus, we selected a MAX_EDGE of 500 and a DENSITY_MULTIPLIER of 10,000 for the current study and as the default parameters. Notably, as discussed below, these values perform well when used with human MCF-7 data as well, implying that the currently selected values are applicable to a variety of experimental datasets.

Identification of RNA species

We proceeded to analyze the murine macrophage GRO-seq data with a MAX_EDGE of 500 and a DENSITY_MULTIPLIER of 10,000. Using these values, we see 11% of the sense strand (294,363,940/2,620,345,972 bp) and 11% of the

antisense strand (282,540,749 /2,620,345,972 bp) being actively transcribed in basal conditions.

These regions of transcription across the genome can then be inspected further. Using the unstimulated data, the total number of transcripts passing the minimal threshold to progress from proto-transcripts into the secondary transcript database is 84,076; of these, 34,743 (41%) had a score (as defined by the custom method described in step 7 of the procedure above) of at least 1. The score threshold best suited for analysis depends heavily on intent; if the user is interested in transcripts that are transient or have low expression, setting a lower threshold at the risk of introducing some noise may be advised. On the other hand, the Hidden Markov Model described by Hah *et al.* [2] resulted in only 22,893 transcripts in a human cell line; if the user desires a comparable high-threshold analysis with annotated regions making up ~50% of the transcripts identified, a higher score threshold can be used.

Of these ~35,000 transcripts, only 8,742 (25%) overlapped with RefSeq genes such that the gene was at least half transcribed. A further 1,573 (5%) overlapped with RefSeq genes (same-strand) but covered less than half of the gene. 8,079 (23%) transcripts overlapped with annotated ncRNA (Figure 3a).

The remaining 21,916 transcripts (63%) were not annotated by RefSeq or the ncRNA.org database. These unannotated transcripts comprised a large proportion of the total. Of the unannotated set of transcripts, 12,042 (55%) were within 1 kilobase (kb) of a RefSeq transcript, of which 5,216 (43%) were specifically within 1kb of an active RefSeq transcription start site (TSS), antisense the RefSeq transcript, and thus

warranted labeling as *promoter-associated RNA*. 2,955 transcripts (25%) were antisense of transcribed RefSeq transcript bodies; these include intragenic enhancers and long ncRNA (Figure S2c).

Of the 21,916 unannotated transcripts, 9,874 (45%) were greater than 1kb away from any RefSeq transcript, and were thus labeled as *distal transcripts*. It has been established that enhancer elements in the genome are marked by unique histone methylation patterns [27,28]— namely, high levels of H3K4me1 and H3K4me2 but low levels of H3K4me3— and further are actively transcribed, generating transcripts (eRNAs) [23]. To assign putative labels for distal transcripts, peaks called by HOMER [7] from H3K4me1 ChIP-sequencing in unstimulated macrophages were loaded into the database and queried. 6,211 (63%) of the distal transcripts overlapped with H3K4me1 peaks and were labeled *eRNA*.

The remaining 3,663 transcripts— 37% of distal transcripts and 13% of all transcripts— had no label. Closer inspection of this subset revealed that 867 of the unannotated distal transcripts were within 2kbp of a H3K4me1 peak. Interestingly, many of these appeared to be regions of transcription between clusters of enhancers (Figure 3b), or enhancer-associated RNA extending far past the range of the histone mark (Figure 3c). Taken together, these results imply that the amount of transcription attributable to enhancers is greater than currently accounted for by analyses looking only at regions directly overlapping associated histone marks.

In addition to these general categories of unannotated transcripts, there were some transcripts in this remainder set that were intriguing anomalies. For example,

there was a 100kbp+ region directly downstream of RefSeq gene Gm14461 on chromosome 2 that exhibited active transcription, but was entirely unannotated by RefSeq, ncRNA.org, or known mouse expressed sequence tags (ESTs) (Figure 3d). Transcription throughout this region was not continuous (Figure 3d, inset), and further there were several stretches of repeats that prohibited unique mapping of tags. Thus, the region was segmented into numerous blocks of transcription. Nonetheless, the identification of such regions demonstrates the importance of closer inspection of GRO-sequencing data for the purposes of finding uncharacterized transcripts, as well as the value of the database described here in building these types of transcripts from merged units.

Transcription does not stop at RefSeq termination sites

Particularly interesting to us was the set of transcripts that continued past the annotated 3' ends of RefSeq transcripts. Closer inspection of this subset revealed that the vast majority (7,346; 84%) of the ~8,700 expressed RefSeq transcripts did not terminate at the annotated Transcription Termination Site (TTS), but instead continued for some length afterward (Figure 1d). The expression levels of these *post-gene RNAs* were well correlated with the RefSeq transcripts they followed (Figure 4a), but the lengths of the post-gene RNAs were not determined by the lengths of the associated RefSeq transcripts (Figure S3a) or the expression level of the associated RefSeq transcripts (Figure 4b, S3b).

We next sought to determine why 16% of the expressed RefSeq transcripts did not continue past the annotated TTS. Of the 1,396 RefSeq transcripts with no associated post-gene transcripts, 157 (11%) were labeled as rRNA rather than mRNA by RefSeq. The remaining mRNA had significantly lower expression levels than the set of RefSeq transcripts with post-gene RNA (Figure 4c, S3c), and indeed transcription of these genes often did not reach the annotated TTS at all (Figure 4d).

Given this difference in expression level, we next filtered the set of ~8,700 expressed RefSeq transcripts down to the set of 6,913 mRNA transcripts that don't stop before the annotated 3' end of the gene (84% of the 8,231 expressed RefSeq mRNAs). Remarkably, 6,305 (91%) of this set had associated post-gene RNA, indicating that the annotated TTSs of RefSeq genes greatly underestimate the extent of RNA transcription at these sites.

Confirming results in human cells

In order to confirm the extensibility of the results obtained in the macrophage data, we used Vespucci to analyze human GRO-sequencing data from MCF-7 cells from two separate studies [2,12]. We used the same parameter values in order to ensure that the default values selected were not applicable only to murine data. In the human cell line, a higher percentage of transcripts were unannotated than in the mouse cells (Figure S2d). The distribution of types of unannotated transcripts was surprisingly similar between the two MCF-7 cell studies (Figure S2e, left versus right panels). Fewer transcripts were called as eRNA as compared to the murine data; this is most

likely due to the fact that there is relatively little histone data available in MCF-7 cells, and the publicly available H3K4me2 data used here [19] was less deep than the mouse H3K4me1 data used above. A larger fraction of the human unannotated transcripts remained unassigned to a known category of RNA. Manual inspection of these transcripts revealed that many overlapped with LINE, SINE, and LTR elements identified by the RepeatMasker database (Repeat Library 20120124, accessed at http://www.repeatmasker.org). We used Vespucci to annotate the remaining transcripts that overlapped with LINE elements, and found that more than half of the remaining transcripts occurred at LINE elements. This corroborated recent reports of widespread transcription at retrotransposons being associated with oncogenesis [29-31]. As a whole, these results indicated that both the default parameter values set in Vespucci and the analysis performed for mouse macrophages above could be repeated in data from multiple cell types, species, and labs.

Benchmarking Vespucci

RefSeq Benchmarking

The two exceptions made for RefSeq annotations noted above allow for consistency with the widely maintained standard of counting tags over RefSeq regions. We compared the shared set of RefSeq transcripts identified by Vespucci to those identified by the *analyzeRNA* method available in the HOMER [7] software package (Figure 5a), and found a high degree of correlation (r = .92). There are two systematic differences that account for the tag count discrepancies between the two

datasets: HOMER sums tags for each RefSeq transcript separately, whereas Vespucci stitches over overlapping genes and isoforms and assigns the total tag count for the longest joined transcript to each associated RefSeq transcript (Figures 5b and 5c); and HOMER does not require continuity across long transcripts, and consequently counts tags that are missed by Vespucci when genes are too sparse to be adequately stitched together (Figure 5d). Notably, these discrepancies primarily affect transcripts that are difficult to interpret using GRO-sequencing data, as it is unclear how to divide up tag counts across overlapping transcripts (Figures 5b and 5c) or long, low-level transcripts (Figure 5d). Thus, with these two exceptions made for RefSeq transcripts, Vespucci produces transcripts comparable to existing annotations and methods of analysis.

Benchmarking against a Hidden Markov Model

Hah *et al*. [2] describe an HMM that determines regions of transcription from GRO-seq data using a two-state model. In order to assess the relevance of the transcripts found by Vespucci to those found by the Hah *et al*. HMM, we trained an HMM using the macrophage GRO-sequencing data described above. Parameters were optimized based on the prescribed procedure for the HMM, which relies on the sum of two errors: (1) the fraction of RefSeq transcripts that are broken apart by the called GRO-seq transcripts (Figure S4a), and (2) the fraction of GRO-seq transcripts that merge two or more RefSeq transcripts (Figure S4b). Using these criteria, the minimum summed error (12.7%) was achieved with a negative log transition probability of 100 and a shape parameter of 5. These parameters were used in the model compared

directly to Vespucci, though similar results were achieved using the parameters selected by the optimization performed by Hah *et al*. (200 and 5, respectively).

We calculated the Vespucci's summed error according to the procedure used for the Hah *et al*. HMM, and found the error to be 1.8%, or one-sixth of the comparable error for the HMM. This lower error serves to underline the advantage of using RefSeq boundaries to inform transcript identification, as it prevents Vespucci from breaking apart or merging together known regions of transcription. Notably, if Vespucci is used with the same parameters but without any prior knowledge of RefSeq regions, the summed error is about three times that of the optimized Hah *et al.* HMM (37.2%). This highlights both the advantage Vespucci gains by integrating with existing databases, and the fact that the default parameter set is designed to avoid over-merging unannotated regions of transcription. If a user desires to retrieve RefSeq transcripts without prior knowledge of RefSeq, a larger MAX_EDGE parameter may be used to achieve a lower error. With the macrophage data and no assumption of RefSeq boundaries, Vespucci with a MAX_EDGE parameter of 5,000 yielded a summed error rate of 12.2% (Figure S4c), just below that achieved with the Hah *et al*. HMM.

Given the designed use of Vespucci, the real question of performance comes with transcript calling over unannotated regions of transcription. Whereas Vespucci identifies 24,428 transcripts above a threshold score of 1 that do not overlap same-strand with a RefSeq transcript, the HMM identifies 9,374. Closer inspection of this discrepancy reveals that the HMM is more likely to merge together transcripts

Vespucci calls as separate (Figure 5e), and less likely to call transcripts when GRO-seq expression levels are low (Figure 5f). Further, we compared the calls made by Vespucci and the HMM to the previously published 5'-GRO-seq data [10], and Vespucci more accurately captured the multiplicity of short transcripts associated with distinct transcripts than the HMM (Figures 5e and 5f). To quantify this merging or missing of transcripts by the HMM, we calculated the same two error rates described above for the HMM as compared to Vespucci, and found that 35.4% (7,026) of the HMM transcripts are broken up by Vespucci transcripts, whereas 1.3% (451) of Vespucci transcripts are broken up by HMM transcripts. Notably, changing the parameters of the HMM might result in higher sensitivity identification of transcripts, but only at the expense of reliable calling of RefSeq genes.

DISCUSSION

GRO-sequencing reveals transcriptional dynamics at a genome-wide scale and thus has the power to give unique and novel insight into the regulation of cellular processes. Taking full advantage of this new data source requires combining disparate data sets and identifying within them transcripts of interest. The system introduced here makes this possible by providing a framework for analyzing GRO-sequencing data at both a general level and in great detail. Further, Vespucci allows for easy integration of many different types of sequencing data, which, when taken together, greatly increase the information gained from each single data type.

In this study, we apply Vespucci to annotate nascent RNA transcripts defined by GRO-sequencing data obtained from primary mouse macrophages and a human breast cancer cell line. This analysis yields a comprehensive list of contiguous nascent transcription units derived from both promoters and enhancers throughout the genome. The Vespucci output provides genomic location, score, nearest gene, and expression level in various sequencing runs of interest. By enabling the quantification of GRO-sequencing data, we add it to the set of sequencing-based methods that can be reliably leveraged to investigate a wide array of biological questions. In the current study, we demonstrate the use of Vespucci to identify novel transcripts of interest, such as the long non-coding transcript near Gm14461, and characterize the length and expression values of enhancer-associated RNAs. As each cell type contains a specific complement of enhancers that specify its identity and functional potential, Vespucci will be a valuable tool for annotation of cell-specific eRNAs. In addition, Vespucci quantifies the extent of nascent transcription beyond the annotated 3' ends of genes defined by the site of polyadenylation. This information may be useful in evaluating mechanisms and regulation of transcriptional termination.

One shortfall of the current system is that the parameter defining acceptable gap distance between reads associated with the same transcript must be set heuristically, dependent upon the needs of the user. Ideally, the parameters to identify transcriptional units from reads would be set to minimize errors against a gold standard of transcriptional units. At this time, no such standard for GRO-sequencing data exists. In the current study, we were able to approximate a gold standard using 5'-

GRO-sequencing data, and thus with Vespucci we hope to take the first step towards defining such a gold standard by providing a method and a framework for transcript identification.

Vespucci extends beyond GRO-sequencing data, too; once the database is set up, it is straightforward to add data from ChIP-sequencing runs, external databases, known motifs, single nucleotide polymorphisms, or any other data of interest that can be expressed within genomic coordinate space. In the current study, we demonstrated the integration of data on retrotransposons with the use of LINE data in analyzing MCF-7 cells. Similarly, one might integrate data on repeat regions and mappability; it is possible to load in genomic coordinates of regions of the genome that preclude uniquely mapped reads, and then allow the merging of transcripts to automatically ignore those regions. In the current implementation, we do not include this functionality, as it was found to yield too many spurious results. However, if a particular application prefers inclusiveness in transcript merging, automatically covering across repeat regions can be incorporated into the system. This is just one example of the extensibility of the system, demonstrating that Vespucci allows for integration of many types of genomic data and sequencing samples, making more feasible analyses that cut across the whole breadth of samples and datasets available to a lab.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–4 and Supplementary SQL Query File 1.

FUNDING

This work was supported by the National Institutes of Health [CA17390-01 to C.K.G.]. Funding for open access charge: National Institutes of Health.

ACKNOWLEDGEMENTS

We thank Drs. Vineet Bafna, Casey Romanoski, Chris Benner, Roy Ronen, and David Allison for their helpful comments in reviewing this manuscript.

Chapter One, in full, is a reprint of the material as it appears in *Nucleic Acids Research*: Allison, Karmel A; Kaikkonen, Minna U; Gaasterland, Terry; Glass, Christopher K, 2014. The dissertation author is the primary author of this paper.

Figure 1.1: GRO-sequencing reveals transcriptional dynamics in great detail, but can be difficult to interpret.

A. Short reads from GRO-sequencing experiments (red) can be mapped back to the reference genome (black) and assigned a genomic coordinate that includes chromosome, strand, starting base-pair, and ending base-pair.

B. Promoter-associated RNA (paRNA) overlaps with the 5' end of the Tmbim6 gene, antisense to the gene itself. The blue bar indicates the transcript that has been identified for Cstb itself, and the leftmost green bar shows the extent of the paRNA.

C. Enhancer RNA (eRNA) appears in GRO-sequencing samples (top track) as bi-directional transcripts centered on the binding sites of transcription factors (middle track) and marked by H3K4me1 (bottom track).

D. Transcription can continue past the 3' ends of annotated RefSeq transcripts, making the exact boundary of relevant transcripts difficult to identify. At the Mmp12 locus, manual interpretation could lead to differing interpretations of where to mark transcript boundaries, either including or excluding the run-off at the 3' end of the gene, and thus it is important to have a consistent, algorithmically determined interpretation. Here, we show that Vespucci is able to either respect the RefSeq boundary (lower blue track), or to identify the entire nascent transcript (upper blue track).

E. Neighboring transcription regions can have very different read densities. Two transcripts are identified along the sense strand, denoted by the blue bars at the top. These two transcripts are close in terms of base-pair distance (363bp apart), but they differ in terms of read-per-base-pair densities, and therefore are kept as two separate transcripts.

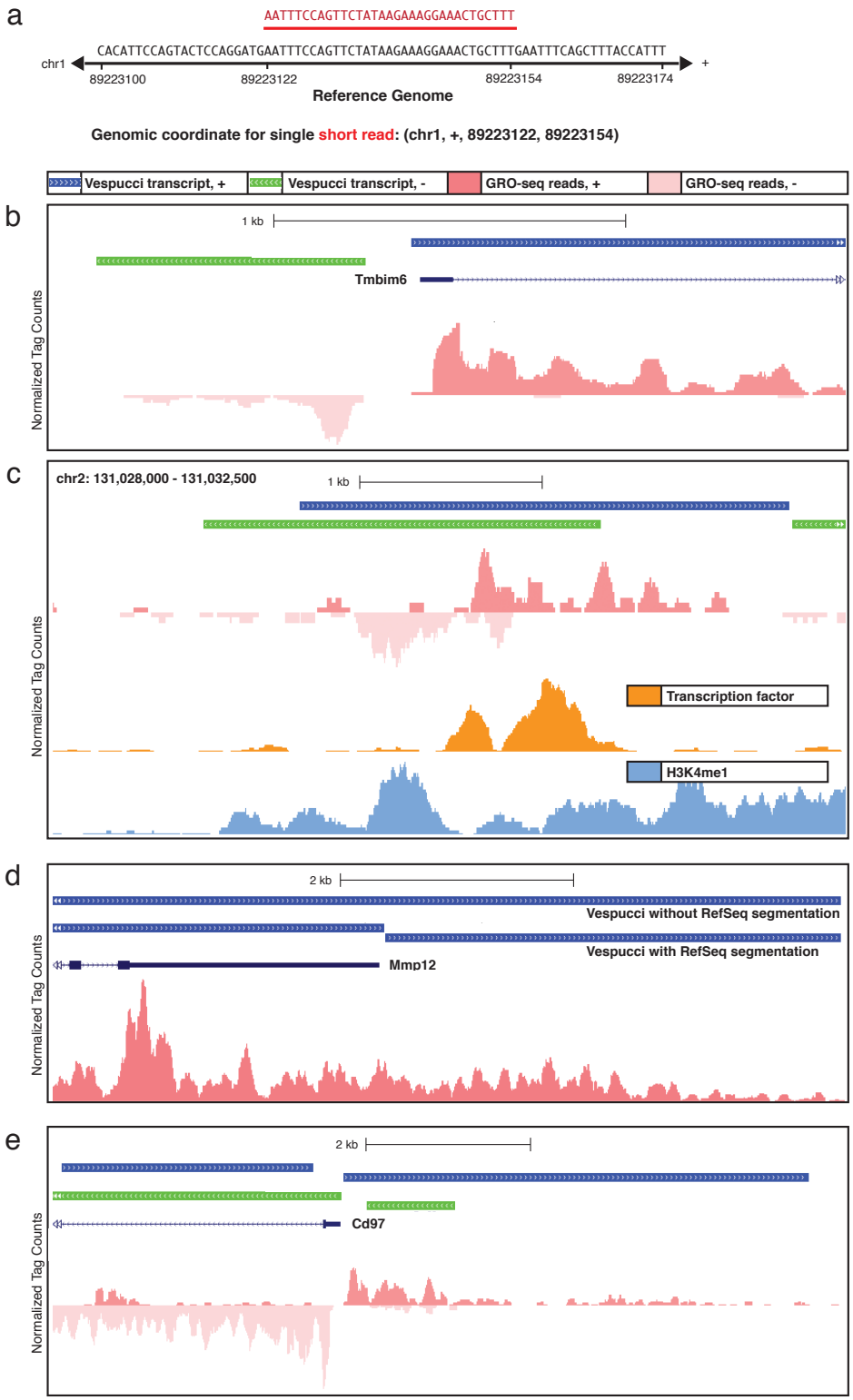


Figure 1 - Glass

Figure 1.2: Stepwise procedure for assembly of transcripts by Vespucci.

A. (1) Each sample, mapped to the reference genome, is reduced to its genomic coordinates and loaded into a separate database table. (2) Short reads from a single run are merged (separated by chromosome). (3) The merged proto-transcripts from each individual run are merged with proto-transcripts from other runs. The number of tags from each different run is stored. (4) The proto-transcripts from (3) are plotted in two-dimensional space, with location in base-pairs along the x-axis and the density in tags per base-pair along the y-axis. Note that the density is scaled according to a parameter, DENSITY_MULTIPLIER, that defines the relationship between the two units of measurement (base-pairs and tags per base-pair). (5) The proto-transcripts in two-dimensional space are then merged according to a MAX_EDGE parameter that operates as the maximal allowed Euclidian distance from the rightmost edge of each transcript. The merged transcript here is considered a continuous unit of transcription by Vespucci. (6) These transcripts can then be associated with known RNA species from RefSeq and ncRNA databases based on genomic coordinates. (7) Transcripts are then scored according to a custom algorithm or reads per kilobase per million (RPKM).
B. Database schema showing the Vespucci transcript table structure, major columns, and related entities. An asterisk indicates a has-many relationship, and ID fields contain references to related tables.

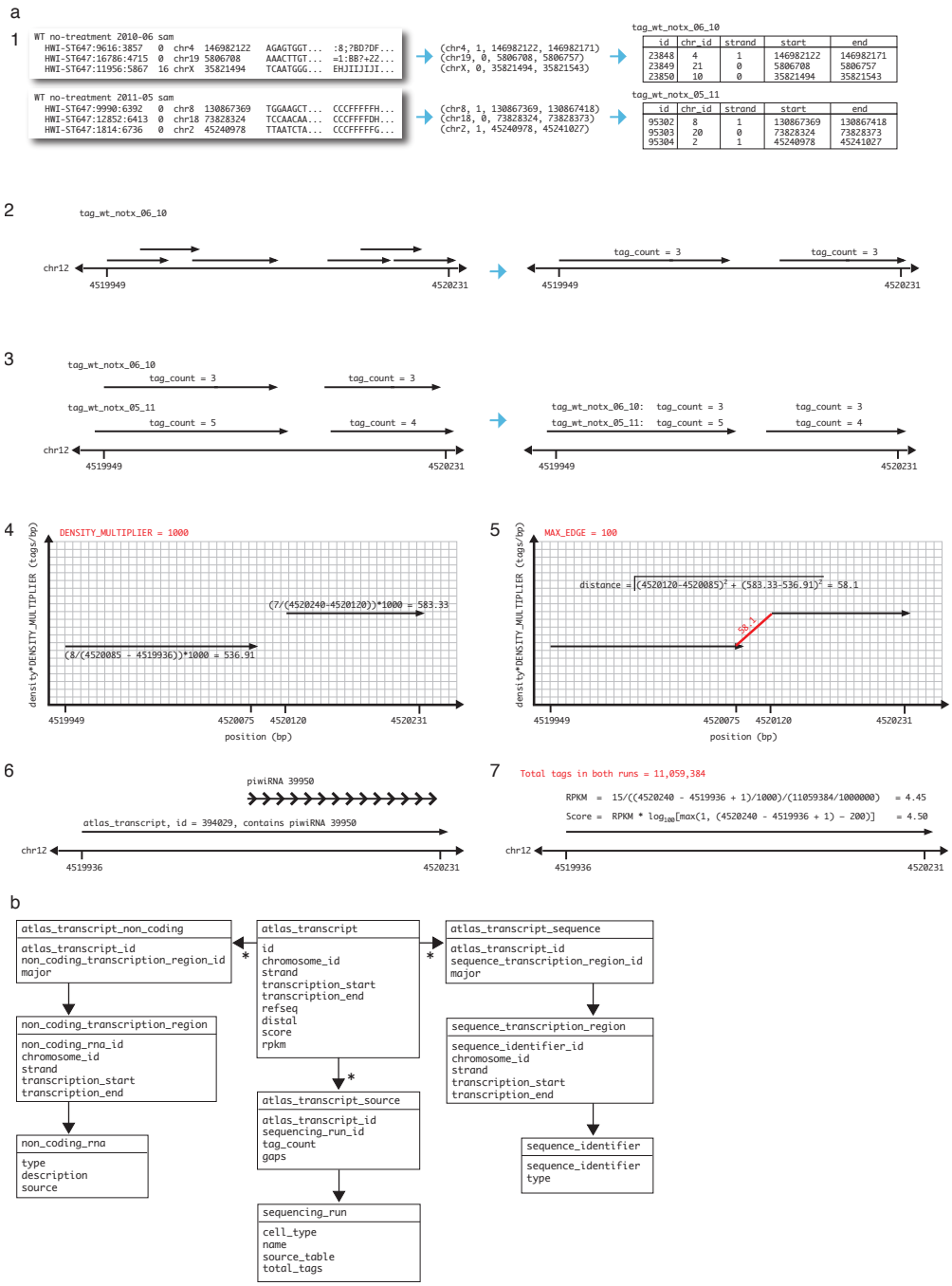


Figure 2 - Glass

Figure 1.3: Vespucci enables the identification and quantification of numerous RNA species in macrophages.

A. Using a score threshold of 1, the great majority (63%, left panel) of transcripts identified are not associated with known RefSeq genes or ncRNA. Of the unannotated set (right panel), more than one half are proximal to RefSeq genes, with the remainder being distal.

B. Transcripts are interspersed not only overlapping with the enhancer histone mark H3K4me1, but also between enhancers, indicating that complex regulatory regions undergo a great deal of active transcription spread over many kilobases.

C. Similarly, transcripts can extend a long distance beyond identifying histone marks at enhancers, with this itergenic region showing low levels of H3K4me1 and GRO-seq signal extending along a single strand for more than 5kb beyond an identified H3K4me1 peak.

D. Vespucci identifies a long, unannotated transcript downstream of Gm14461. Vespucci does not merge the entire region, but, with a gap parameter of 100bp, separates it into several long regions with many shorter regions interspersed throughout. Closer inspection (inset) shows that the boundaries determined by Vespucci reflect real discontinuities in the GRO-seq signal that will require further study to interpret. H3K4me1 is shown on the lower track to indicate that this transcript is methylated at the 5’ end, much as a protein coding gene would be.

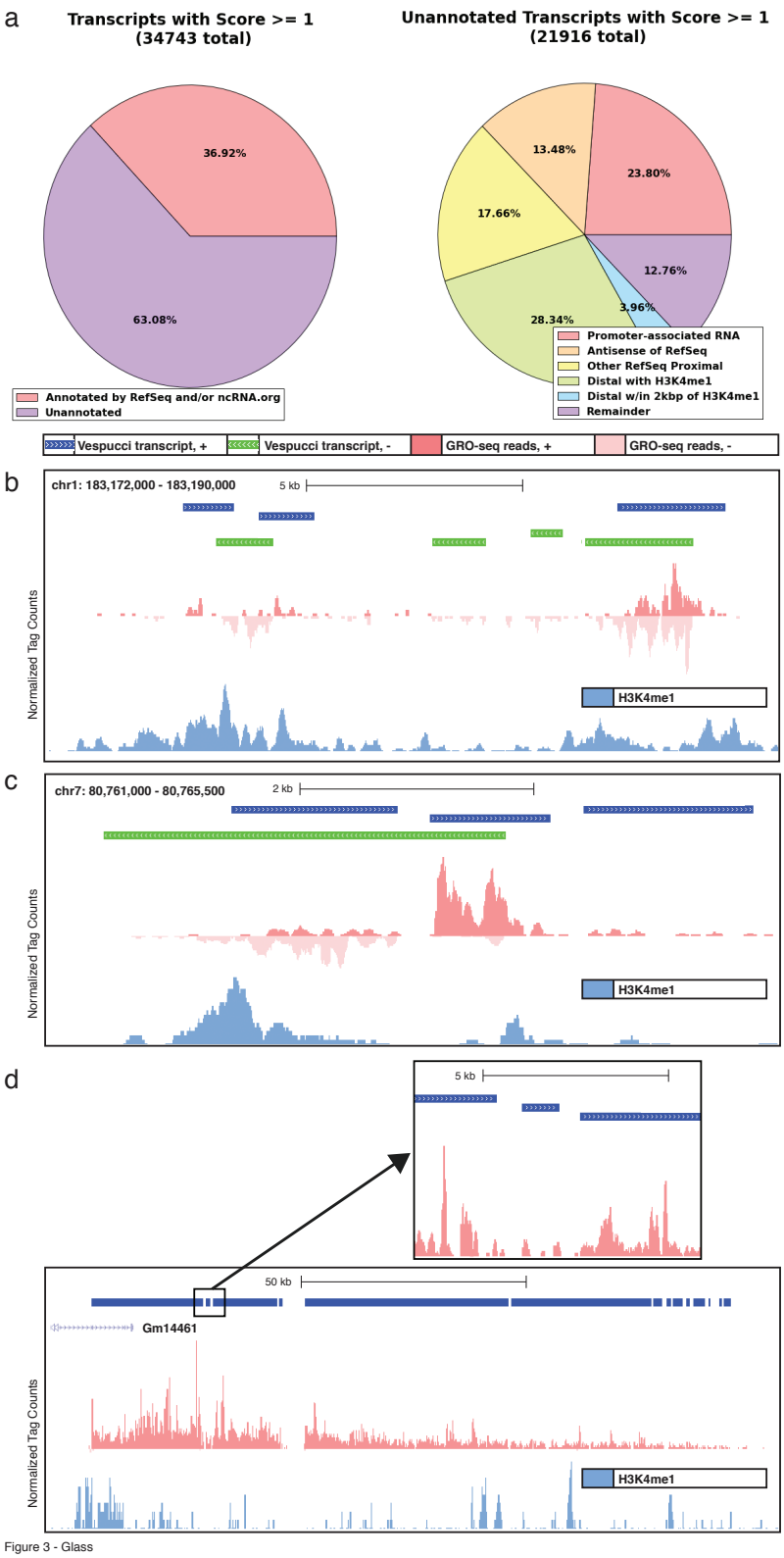


Figure 3 - Glass

Figure 1.4: Transcription continues past the annotated 3' ends of most genes.

A. The expression levels of transcripts immediately following the 3' ends of RefSeq sequences are correlated with those of the preceding RefSeq transcripts as measured with Vespucci scores.

B. The length that transcription carries past the 3' end has a weak but positive correlation with the expression level of the preceding RefSeq transcript as measured with Vespucci scores.

C. 16% of RefSeq transcripts are not found to have post-gene RNA according to Vespucci. These RefSeq transcripts tend to have much lower expression levels as measured with Vespucci scores than the 84% of transcripts that do continue past their annotated 3' ends.

D. In addition to having very low expression levels, many of the RefSeq transcripts without post-gene RNA are notable in that the transcript called by Vespucci does not reach the annotated 3' end of the gene, as is the case with the Ube2w gene here.

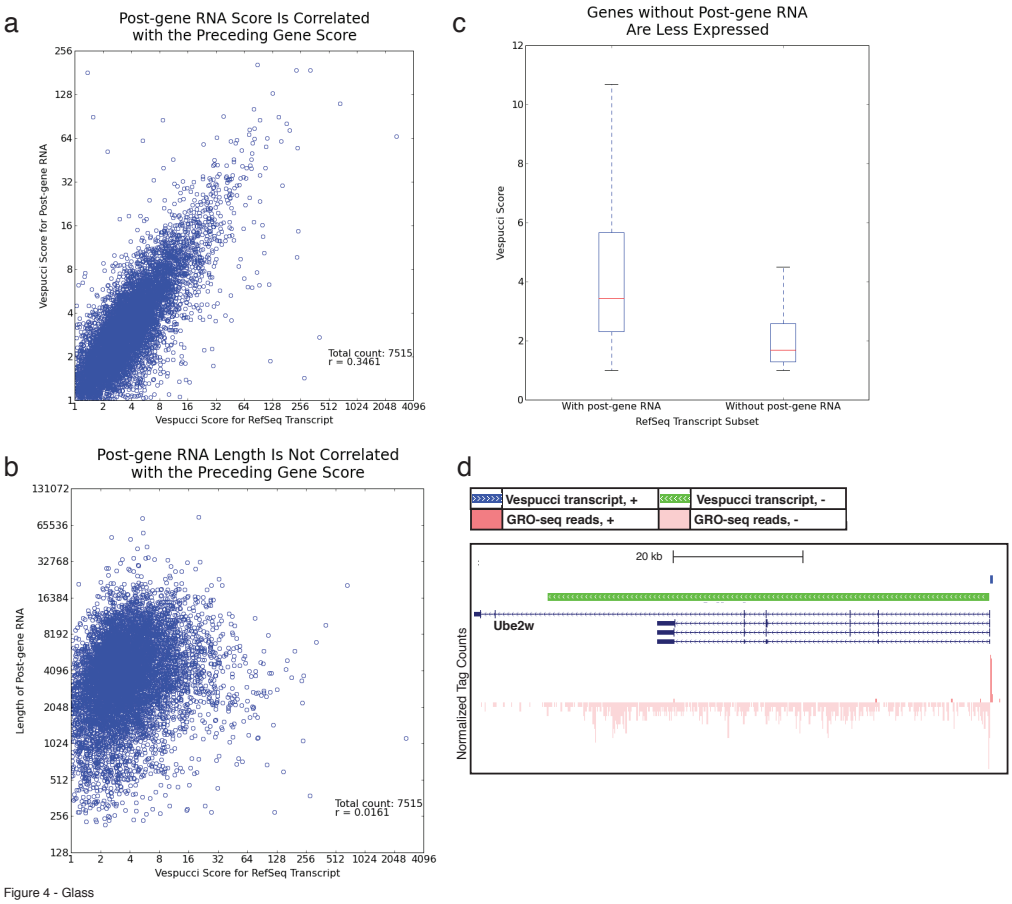


Figure 4 - Glass

Figure 1.5: **Vespucci retrieves RefSeq expression levels without losing non-coding RNAs.**

A. RefSeq identifiers can be used to compare the tag counts determined by Vespucci at RefSeq genes to the tag counts determined by the HOMER software, which uses a gene-centric approach to sum GRO-seq tags over known genes.

B. The correlation between tag counts is generally very good, with deviations from the diagonal attributable to three primary categories of transcripts: Vespucci does not segment transcripts at alternative isoforms, but returns the tags for the whole transcript for each contained isoform. In contrast, HOMER tallies tags within the precise boundaries of each isoform, resulting in discrepancies between the two methods at shorter isoforms, such as the Spp1 gene seen here;

C. as with multiple isoforms, overlapping genes are not segmented by Vespucci, and the tag count for the entire transcript covering Macf1 is associated with the short gene that is overlapping, D830031N03Rik; and

D. genes that have very few, dispersed tags that cannot be adequately merged yield several smaller transcripts according to Vespucci, whereas HOMER implicitly joins them and counts all that fall along the body of the gene regardless of continuity of transcription.

E. The HMM described by Hah *et al*. identifies transcripts using a two-state model that calls regions transcribed (black bars) or untranscribed. The HMM identifies many fewer transcripts than Vespucci, in part because it merges together transcripts called as distinct by Vespucci. Here, three pairs of bi-directional RNAs that are identified as two single units by the HMM. The bottom track shows data from previously published 5'-GRO-seq, a method that detects nascent RNA with a 5' 7-methylguanylated cap. This method identifies start sites of nascent RNAs genome-wide. The data here, from RAW macrophages, shows that Vespucci captures more accurately the separately initiated transcripts.

F. Similarly, some transcripts are called by Vespucci at expression levels too low for the HMM. Here, a paRNA is identified by Vespucci but not the HMM. The bottom track again shows 5'-GRO-seq from RAW macrophages, where the paRNA start site can be clearly seen.

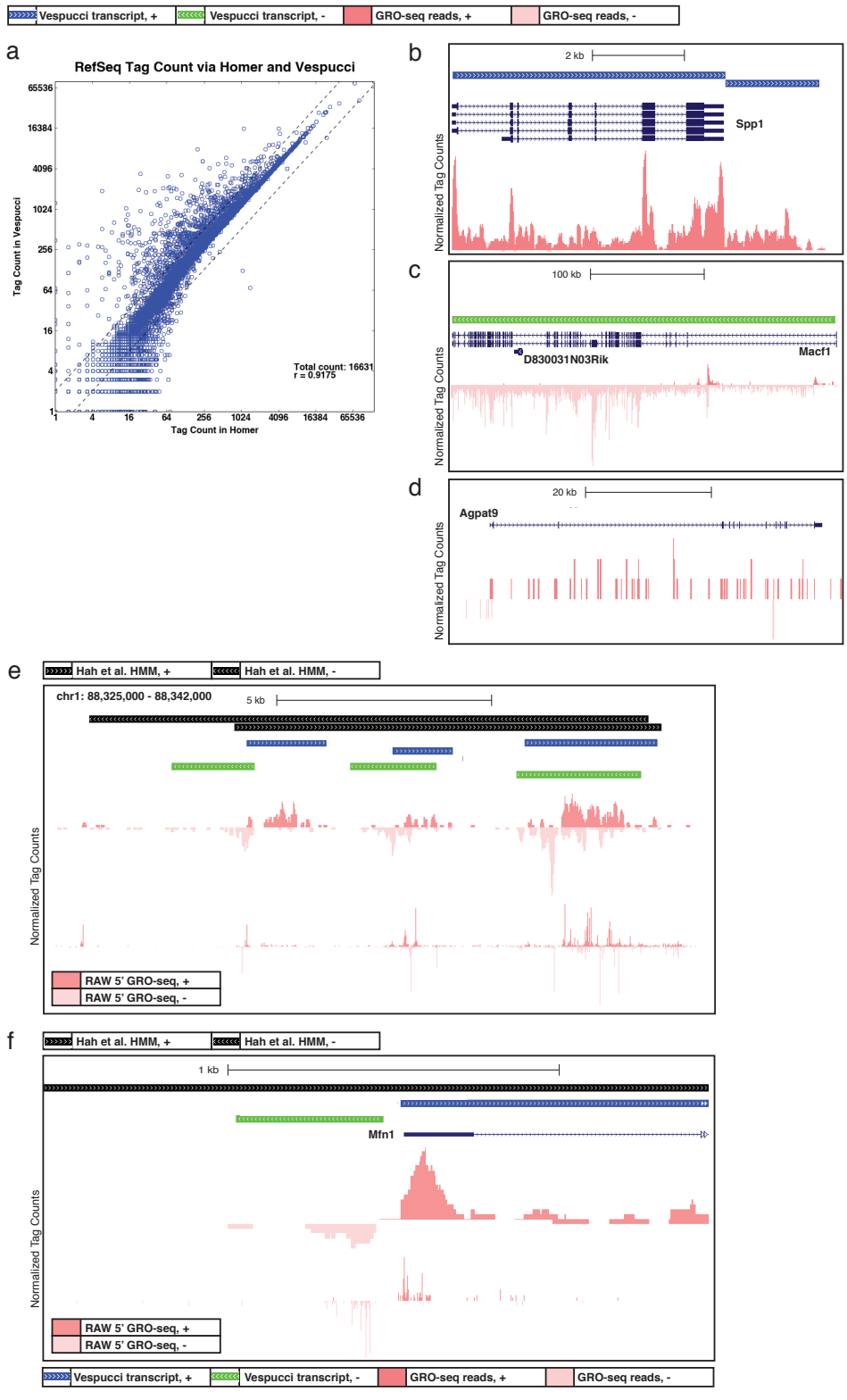


Figure 5 - Glass

Figure 1.S1: Principles of Vespucci analysis.

A. Schematic of read pile-ups, showing regions that are continuously tiled on the left and right. These two regions have different densities, with the right having a higher read-per-base-pair value than the left. These differences in density can advise whether to consider two separate tiled regions as two separately regulated transcripts or as a single unit of transcription.

B. Vespucci takes into account known RefSeq annotations when merging transcripts. This allows the unification of genic transcripts that exhibit numerous gaps due to low expression levels but are known to be continuous, as is the case with the Meaf6 gene here. The blue bar demarcating the transcript generated extends the full length of the gene despite gaps in tags.

C. In order to optimize parameters for the algorithm, we make use of 5'-GRO-seq in RAW macrophages. This data identifies transcription initiation sites genome-wide, and thus ideally there would be only one start site per Vespucci transcript. We thus define an Initiation Recapture Rate (IRR) that measures the extent to which Vespucci aligns with the 5'-GRO-seq data at different parameter settings. In murine macrophages, Vespucci achieved a maximal IRR with a MAX_EDGE of 500

D. (continued from c) and a DENSITY_MULTIPLIER of 10,000.

Figure S1 - Glass

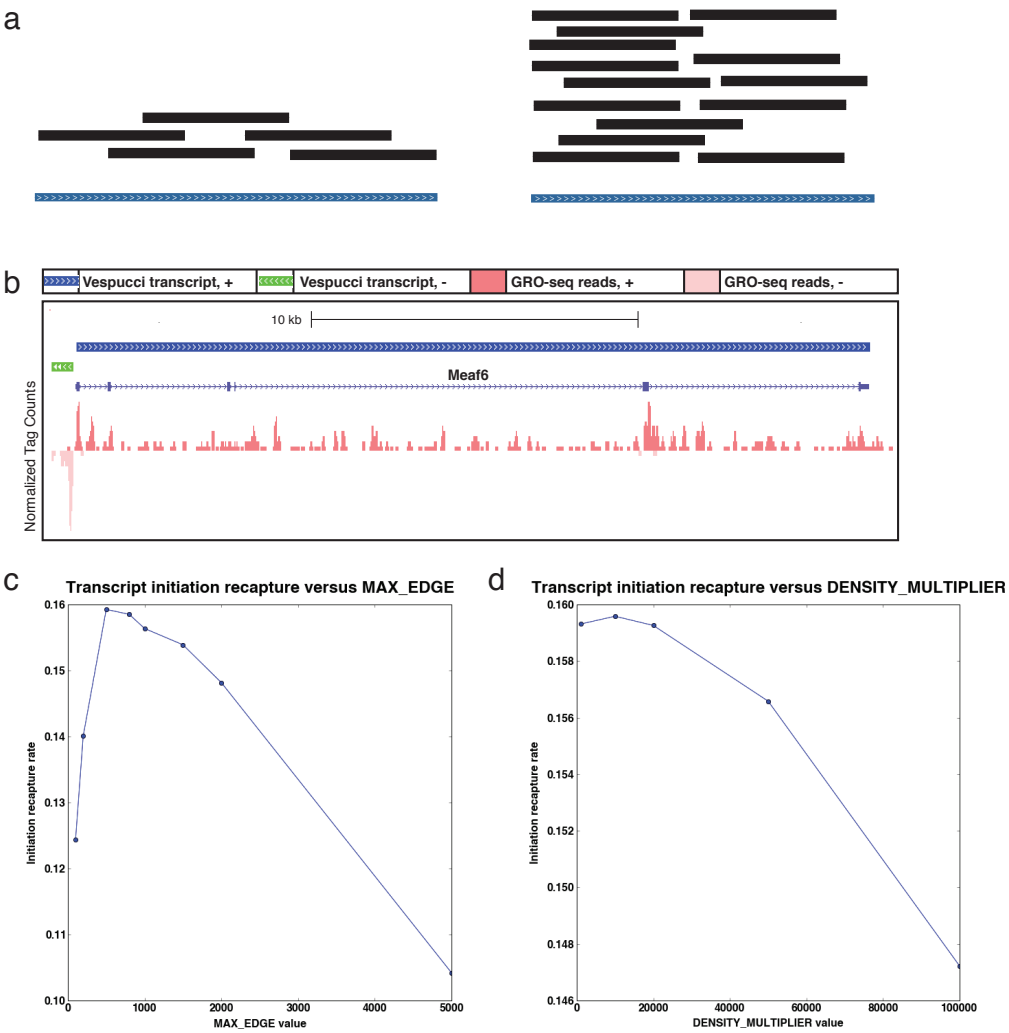


Figure 1.S2: Vespucci enables the identification and quantification of numerous RNA species in macrophages.

A. RPKM tends to negatively weight long transcripts relative to short transcripts, and thus shows a downward slope in relation to transcript length.

B. The Vespucci score, on the other hand, is based on the log of the length of the transcript for longer transcripts, which prevents some of the longer transcripts from falling below whatever scoring threshold is set.

C. Vespucci identifies several different types of transcripts antisense to RefSeq genes, including intragenic enhancers (left) that co-localize with the enhancer mark H3K4me1; and long non-coding RNA species (right) that overlap annotated genes.

D. In order to demonstrate the extensibility of Vespucci to other species, cell types, and experimental hands, we used Vespucci to analyze previously published MCF-7 data from two separate studies. In (D), a smaller fraction of transcripts identified in each of the MCF-7 studies were annotated by the RefSeq or ncRNA databases as compared to murine macrophages.

E. Closer inspection of the unannotated regions demonstrates that Vespucci identifies surprisingly reproducible RNA species in the two MCF-7 studies. A smaller proportion of MCF-7 unannotated transcripts than in the murine macrophages was marked by enhancer-related histone marks, though this may be an artifact of the depth of ChIP-seq data available in these MCF-7 cells. Notably, a large proportion of distal transcripts not otherwise marked overlapped with LINEs, SINEs, and other repeat-rich elements. This finding corroborates previously published research that suggests oncogenesis involves the widespread transcription of retrotransposons.

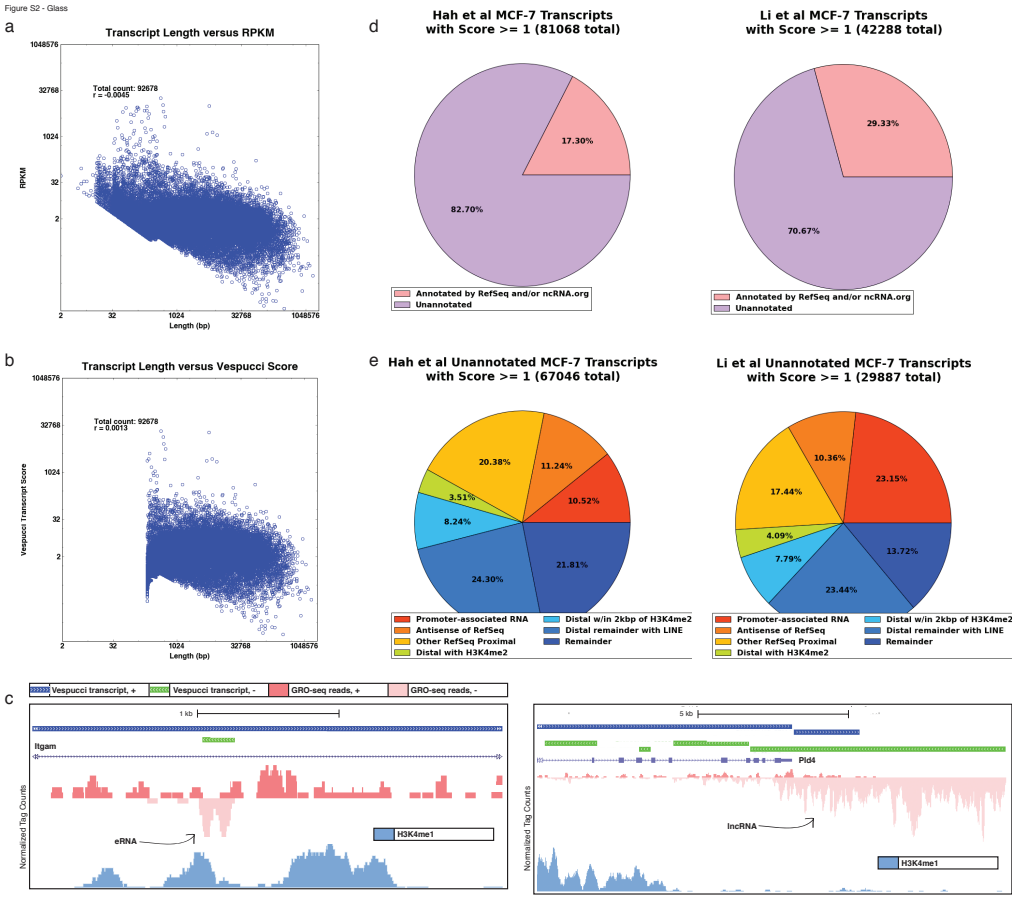


Figure S2 - Glass
a
Transcript Length versus RPKM
Total count: 92678
r = -0.0045
RPKM
Length (bp)
b
Transcript Length versus Vespucci Score
Total count: 92678
r = 0.0013
Vespucci Transcript Score
Length (bp)
c
Vespucci transcript, +
Vespucci transcript, -
GRO-seq reads, +
GRO-seq reads, -
1 kb
Itgam
eRNA
H3K4me1
Normalized Tag Counts
5 kb
Pid4
lncRNA
H3K4me1
Normalized Tag Counts
d
Hah et al MCF-7 Transcripts with Score >= 1 (81068 total)
17.30%
82.70%
Annotated by RefSeq and/or ncRNA.org
Unannotated
Li et al MCF-7 Transcripts with Score >= 1 (42288 total)
29.33%
70.67%
Annotated by RefSeq and/or ncRNA.org
Unannotated
e
Hah et al Unannotated MCF-7 Transcripts with Score >= 1 (67046 total)
20.38%
11.24%
3.51%
10.52%
8.24%
21.81%
24.30%
Promoter-associated RNA
Antisense of RefSeq
Other RefSeq Proximal
Distal with H3K4me2
Distal w/in 2kbp of H3K4me2
Distal remainder with LINE
Remainder
Li et al Unannotated MCF-7 Transcripts with Score >= 1 (29887 total)
10.36%
23.15%
17.44%
4.09%
7.79%
13.72%
23.44%
Promoter-associated RNA
Antisense of RefSeq
Other RefSeq Proximal
Distal with H3K4me2
Distal w/in 2kbp of H3K4me2
Distal remainder with LINE
Remainder

Figure 1.S3: Transcription continues past the annotated 3' ends of most genes.

A. The lengths of transcripts immediately following the 3' ends of RefSeq sequences are not correlated with the lengths of the preceding RefSeq transcripts.

B. The length that transcription carries past the 3' end does not correlate well with the expression level of the preceding RefSeq transcript as measured with RPKM. c. 13% of RefSeq transcripts are not found to have post-gene RNA according to Vespucci. These RefSeq transcripts tend to have much lower expression levels as measured with RPKM than the 87% of transcripts that do continue past their annotated 3' ends.

Figure S3 - Glass

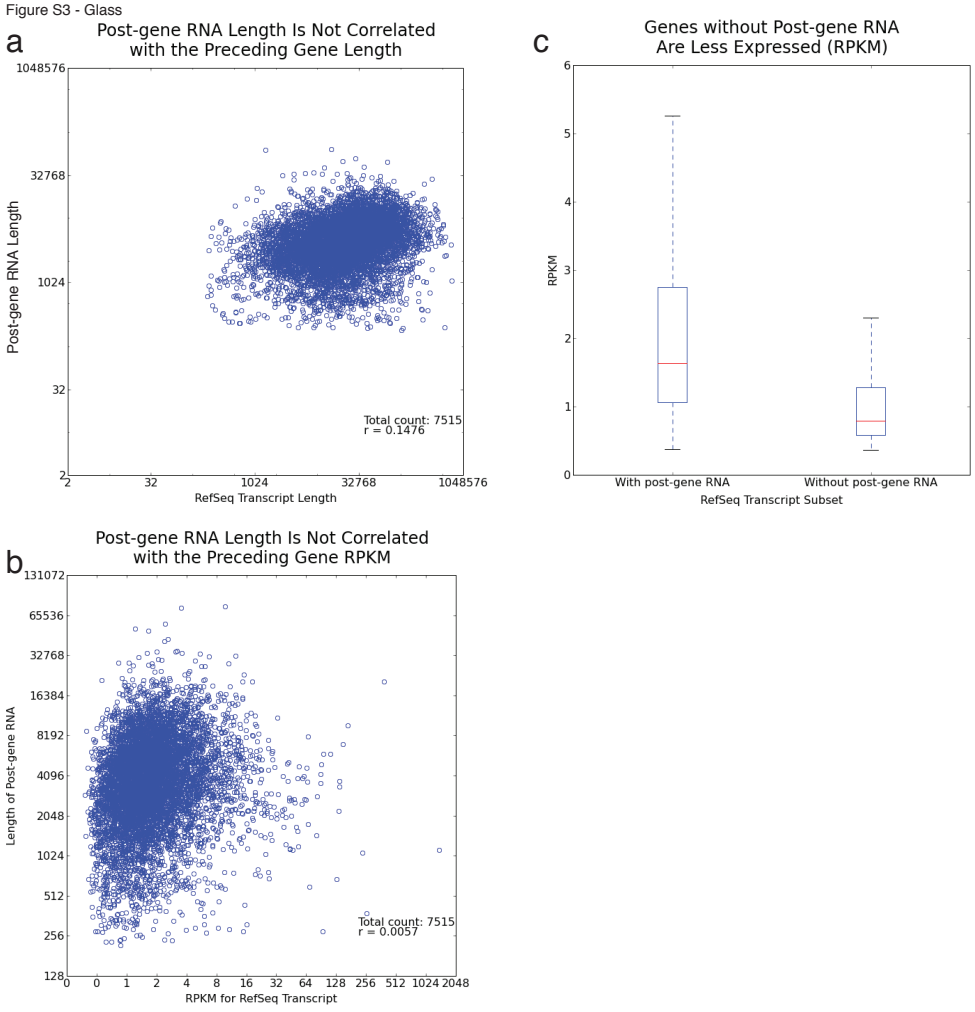


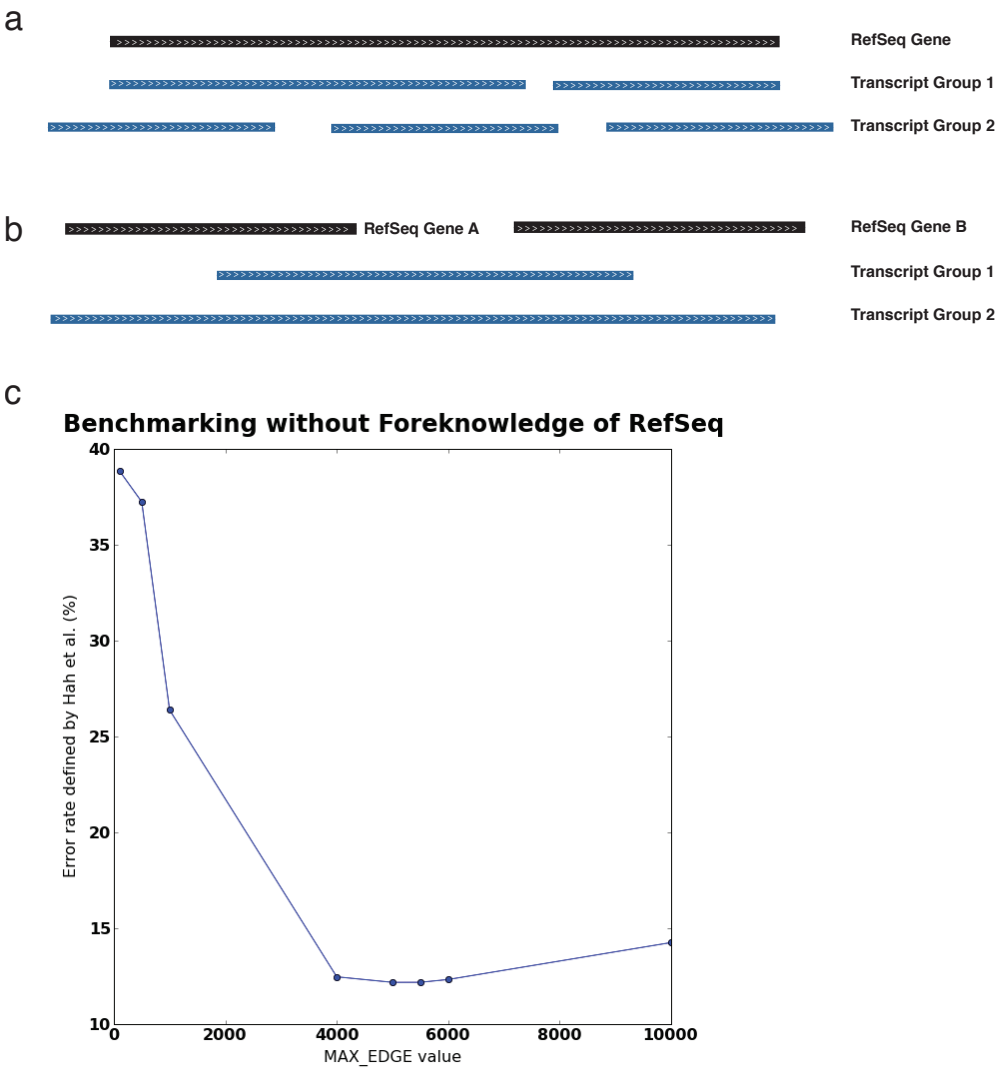
Figure 1.S4: Hah *et al*. measure two types of error.

A. In order to optimize parameters for an HMM, Hah *et al*. measure the fraction of RefSeq transcripts that are broken apart by the called transcripts. Here, transcript groups 1 and 2 break apart a RefSeq gene, and would each increment the measured error by 1/(number of RefSeq transcripts).

B. The second type of error Hah *et al*. measure is the fraction of called transcripts that merge together RefSeq genes. Here, transcript groups 1 and 2 run together RefSeq Gene A and RefSeq Gene B, and would each increment the measured error by 1/(number of called transcripts).

C. Comparing Vespucci to RefSeq transcripts using the error defined by Hah *et al.* advantages Vespucci because, in the default configuration, Vespucci has foreknowledge of RefSeq boundaries when defining transcripts. Here, we built transcript datasets using Vespucci without any RefSeq transcript awareness. Even with no foreknowledge of RefSeq boundaries, Vespucci achieved an error rate equivalent to that of the Hah *et al.* HMM with a MAX_EDGE of 5,000.

Figure S4- Glass



REFERENCES

1 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).

2 Hah, N., Danko, C. G., Core, L., Waterfall, J. J., Siepel, A., Lis, J. T. & Kraus, W. L. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622-634 (2011).

3 Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G. & Fu, X.-D. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390-394 (2011).

4 Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C. & Glass, C. K. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51**, 310-325, doi:10.1016/j.molcel.2013.07.010 (2013).

5 Kaikkonen, M. U., Lam, M. T. Y. & Glass, C. K. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc Res* **90**, 430-440 (2011).

6 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**, 130-135 (2012).

7 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).

8 Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).

9 Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E.,

Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R. & Gingeras, T. R. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

10 Lam, M. T. Y., Cho, H., Lesch, H. P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M. U., Kim, A. S., Kosaka, M., Lee, C. Y., Watt, A., Grossman, T. R., Rosenfeld, M. G., Evans, R. M. & Glass, C. K. Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* **498**, 511-515 (2013).

11 Lai, F., Orom, U. A., Cesaroni, M., Beringer, M., Taatjes, D. J., Blobel, G. A. & Shiekhattar, R. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501, doi:10.1038/nature11884 (2013).

12 Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H. S., Glass, C. K. & Rosenfeld, M. G. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520, doi:10.1038/nature12210 (2013).

13 Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Oude Vrielink, J. A., Elkon, R., Melo, S. A., Leveille, N., Kalluri, R., de Laat, W. & Agami, R. eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell* **49**, 524-535, doi:10.1016/j.molcel.2012.11.021 (2013).

14 Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. & Greenberg, M. E. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187, doi:10.1038/nature09033 (2010).

15 De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C. L. & Natoli, G. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* **8**, e1000384, doi:10.1371/journal.pbio.1000384 (2010).

16 Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. & Snyder, M. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

17 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-223 (2009).

18 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

19 Tsai, W. W., Wang, Z., Yiu, T. T., Akdemir, K. C., Xia, W., Winter, S., Tsai, C. Y., Shi, X., Schwarzer, D., Plunkett, W., Aronow, B., Gozani, O., Fischle, W., Hung, M. C., Patel, D. J. & Barton, M. C. TRIM24 links a non-canonical histone signature to breast cancer. *Nature* **468**, 927-932, doi:10.1038/nature09542 (2010).

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

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

22 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015 (2010).

23 Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. & Greenberg, M. E. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).

24 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

25 Ji, X., Zhou, Y., Pandit, S., Huang, J., Li, H., Lin, C. Y., Xiao, R., Burge, C. B. & Fu, X. D. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell* **153**, 855-868, doi:10.1016/j.cell.2013.04.028 (2013).

26 Mituyama, T., Yamada, K., Hattori, E., Okida, H., Ono, Y., Terai, G., Yoshizawa, A., Komori, T. & Asai, K. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res* **37**, 89-92 (2009).

27 He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., Mieczkowski, P., Lieb, J. D., Zhao, K., Brown, M. & Liu, X. S. Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42**, 343-347 (2010).

28 Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

29 Kaer, K. & Speek, M. Retroelements in human disease. *Gene* **518**, 231-241, doi:10.1016/j.gene.2013.01.008 (2013).

30 Mandal, A. K., Pandey, R., Jha, V. & Mukerji, M. Transcriptome-wide expansion of non-coding regulatory switches: evidence from co-occurrence of Alu exonization, antisense and editing. *Nucleic Acids Res* **41**, 2121-2137, doi:10.1093/nar/gks1457 (2013).

31 Tang, R. B., Wang, H. Y., Lu, H. Y., Xiong, J., Li, H. H., Qiu, X. H. & Liu, H. Q. Increased level of polymerase III transcribed Alu RNA in hepatocellular carcinoma tissue. *Molecular carcinogenesis* **42**, 93-96, doi:10.1002/mc.20057 (2005).

CHAPTER TWO

Affinity and Dose of TCR Engagement Yield Proportional Enhancer and Gene Activity in CD4+ T Cells

Affinity and dose of T cell receptor (TCR) interaction with antigens govern the magnitude of CD4+ T cell responses, but questions remain regarding the quantitative translation of TCR engagement into downstream signals. We find that while the response of CD4+ T cells to antigenic stimulation is bimodal, activated cells exhibit analog responses proportional to signal strength. Gene expression output reflects TCR signal strength, providing a signature of T cell activation. Expression changes rely on a pre-established enhancer landscape and quantitative acetylation at AP-1 binding sites. Finally, we show that graded expression of activation genes depends on ERK pathway activation, suggesting that an ERK-AP-1 axis translates TCR signal strength into proportional activation of enhancers and genes essential for T cell function.

INTRODUCTION

The question of how the T cell receptors (TCRs) of CD4+ T cells respond to ligands of differing affinities and concentrations with such remarkable specificity is of great interest to the study of immunity. The TCR binds to antigen presented by the molecules encoded by the Major Histocompatibility Complex (MHC) such that strength of the TCR signal in response to a particular peptide-MHC complex (pMHC) is dependent on all three components– the antigenic peptide, the MHC, and the TCR itself[1]. Variations in signal efficiency are thus caused by the generated TCR sequence[2, 3], genetic differences in MHC[1, 4], and the peptide being presented[5]. Even small differences in the number or affinity of these pMHC-TCR interactions are read by the TCR and have important consequences for the nature and extent CD4+ T cell activation; high-affinity interactions lead to inflammatory responses at a lower concentration of antigen, increased Interleukin 2 (IL2) and IFNγ production, and increased proliferation[1, 4, 5, 6, 7, 8, 9, 10], whereas lower affinity interactions can lead to incomplete phosphorylation of downstream signaling complexes[11, 12], anergy[9, 12], or TCR antagonism[6, 11]. The precise result of low-affinity engagement varies with experimental conditions, but in each case, a cellular phenotype distinct from high-affinity engagement is produced.

There exists a well-characterized model system for studying the affects of ligand affinity on CD4+ T cell activation: the AND mouse is a strain with a transgenic CD4+ T cell TCR[13]. This TCR recognizes pigeon cytochrome *c* (PCC) along with synthetic and species-variant cytochrome *c* oligopeptides[5, 7, 8]. Notably, though many

of the peptides differ from PCC by a single amino acid, the effects of TCR recognition of the peptides vary greatly. Kinetic parameters and cytokine output of the interaction with many cytochrome *c* peptides and their analogues have been described[7, 8]. Differences in microcluster formation at the membrane have likewise been described[14].

These variable responses to ligands of differing affinity are especially interesting in the context of the digital TCR response. TCR responses have been characterized as digital[15]—that is, signaling downstream of the TCR is either all-on or all-off, such that a given T cell must either be committed to a full response or to no response. Previous work has established this switch-like behavior as observable in terms of extracellular markers such as CD69[16, 17], ERK pathway component localization[16, 17, 18], NF-kB activation[19], NFAT localization[20, 21], cell-cycle entry[22], and cytokine production[21, 23]. As a result, differences in the magnitude of responses to ligands of varying affinity would be attributed to greater frequencies of T cells responding at the population level, rather than per-cell variability[22, 23, 24, 25]. Still, some aspects of the TCR response have been described as analog, or varying in proportion to the strength of signaling: CD3zeta chain phosphorylation[11, 12, 17, 26, 27]; Zap70 activation[17, 18]; intracellular calcium concentrations[28]; expression of the transcription factor IRF4[29, 30]; and cell division time[31]. It is unclear how these analog components of the TCR response fit in to a digital model.

Both the ability of the TCR to discriminate with high resolution between ligands and the digital nature of TCR signaling have been extensively studied at the

level of signaling. Downstream of the TCR, a number of signaling pathways govern the molecular response to engagement. AKT and PKCθ interact at the cell membrane and jointly serve to induce nuclear translocation of the pro-inflammatory transcription factor NF-kB, which in turn is able to activate target genes[32]. AP-1, which comprises homo- or heterodimers assembled from proteins of the Fos, Jun, and ATF transcription factor families[33], requires both TCR and co-stimulatory signaling[34], and is usually activated by the Ras/Raf/Mek/Erk pathway[35, 36]. At least four feedback loops have been identified in thymocytes and peripheral T cells downstream of the TCR[15, 37]. Collectively, these feedback loops serve to enforce a digital response by either dampening sub-threshold signaling or amplifying above-threshold signaling, resulting in T cell responses that are all-off or all-on respectively.

Despite these insights into the signaling pathways downstream of TCR activation, there is little known about the transcriptional programs that determine the distinct phenotypes resulting from high- versus low-affinity stimulation. In this study, we address the question of affinity at the level of the chromatin. We take advantage of the PCC system to assess the effects of varying the dose and affinity of peptide presentation to CD4+ T cells on enhancer formation and gene expression, giving us a genome-wide picture of how TCR signaling is able to achieve such highly specific responses despite its digital signaling pattern. We find first that the digital/analog dichotomy is too simple, and instead CD4+ T cells respond to ligands of varying dose and affinity by modulating both the frequency of responding cells and the level of activation of responding cells at a single cell level. In other words, activation markers

are analog with respect to the strength of TCR signaling when comparing across doses and affinities, but for any single dose and affinity, the overall signaling pattern is digital for the population of cells. We next show that at the population level, the combined effects of analog precision and increasing frequency of responder cells produce gene expression patterns that directly reflect the strength of TCR signaling for a set of activation signature genes. These gene expression patterns can be used to assess CD4+ T cell activation status, and we develop a tool for ranking arbitrary CD4+ T cell populations by activation score. Underlying these gene expression patterns, we find that the enhancer landscape is largely pre-existing, such that TCR engagement results in activation of primed enhancers rather than through selection of new enhancers. Finally, we show that the graded activation score and the expression of activation signature genes are dependent on the amount of phosphorylated ERK activity downstream of the TCR. Together, these results suggest that the degree of ERK activation translates the analog TCR signal resulting from varying the dose and affinity of TCR engagement into downstream gene expression programs.

RESULTS

The TCR response is analog: quantitative responses to signal strength

In order to understand the effects of the digital TCR response on the transcriptional landscapes of CD4+ T cells, we first sought to characterize the "on-state" of the TCR response. CD4+ T cells and CD11c+ antigen presenting cells

(APCs) were isolated from AND transgenic mice (Figures S1A, S1B) and co-cultured for 24 hours with one of a panel of previously described[7, 8] peptides at several different doses. Cell activation was then measured at a single cell level using flow cytometry. As expected, for each given peptide and dose, CD4+ T cells followed a digital pattern, appearing either all-on or all-off according to extracellular activation markers such as CD69 (Figure 1A) and CD25 (Figure S1C). Increasing the peptide dose or affinity significantly increased the percent of activated cells in the population (Figure 1B, 1C). However, when we compared across peptides and concentrations, it was clear that the activation level of the on-state cells was not "all on." Gating on CD69+ cells, each different peptide and different dose of a peptide achieved a different amount of CD69 per cell (Figure 1D). Gating on CD25+ cells yields similar results, with varying amounts of CD25 expressed per cell dependent on both the dose and the affinity of the stimulation (Figure 1E). Thus, while under a given condition the CD4+ T cells were either on or off as per classical digital signaling, when comparing across a panel of conditions, the activation level of the on-state cells is analog with respect to the strength of the TCR signal.

Gene expression is graded genome-wide

In order to understand the effect of this variability genome-wide, we selected two peptides—the low-affinity K99A and the high-affinity PCC—and sequenced mRNA from CD4+ T cells exposed to both a low and a high concentration. We compared the gene expression profiles at 24 hours across five conditions (no-peptide;

low-dose, low-affinity (10uM K99A); high-dose, low-affinity (100uM K99A); low-dose, high-affinity (0.1uM PCC); and high-dose, high-affinity (10uM PCC)), four out of five of which displayed some degree of activation as measured by extracellular markers such as CD69 or CD25. We used principal component analysis (PCA) to determine the primary axes of variation across the approximately 3,000 genes that were expressed above ten reads per kilobase per million (RPKM) and at least two-fold different between any two conditions. A single principal component explained more than 99% of the variance in gene expression changes (Figure S2A).

The first principal component ranks the samples according to what would be expected based on TCR signal strength and extracellular markers such as CD69 and CD25 (Figure 2A). As PC1 captured the gene expression changes concomitant with increasing activation, we extracted the most positive 10% and most negative 10% of the genes along PC1 to determine which genes were important for the axis. The 10% of genes contributing the most positive signal to PC1 were increasing in a generally graded manner with TCR signal strength across the samples, and the 10% of genes contributing the most negative signal were decreasing (Figure 2B; two-tailed p-values based on permutation testing of 2.9e-11 and 1.8e-28, respectively).

Collectively, the most indicative 10% of genes for PC1 provide a multidimensional signal for ranking the samples in one dimension according to TCR signal strength and activation state of the CD4+ T cell; we therefore call these genes "activation signature genes." Looking more closely at the genes in this group, we see many well-documented immune response genes such as Tbx21 (Tbet), Stat1, and TNF

(Figure 2C), all of which increase in a graded manner along with TCR signal strength at the population level. IRF4, previously reported to increase in expression in an analog manner downstream of the TCR on a per-cell basis[29, 30], was also among the activation signature genes, and showed the same graded response pattern at the population level across the conditions (Figure 2D). Gene Ontology analysis yielded several enriched gene categories among the activation signature, with the most highly enriched group being protein biosynthesis and translation genes (Figure 2E) such as Etf1 and Eif3a (Figure S2B). Protein biosynthesis has been previously shown to be increased upon T cell activation, and here we see that many of the genes involved in increasing translational activity are themselves upregulated in a manner that is proportional to the level of activation across the population. Another enriched ontological category was molecular chaperone genes that are responsible for protein folding and unfolding, including six of the Cct family of chaperones (for example, Cct2: Figure S2C) and five heat shock family members (for example, Hsph1) that increased with TCR signal strength.

The graded increase of activation signature genes at the population level corresponded with single-cell increases in CD69 and CD25, but it was unclear whether analog levels of activation signature genes were widespread. In order to determine whether activation signature genes increased on a per-cell basis in more cases, we determined the per-cell protein levels of a panel of genes from the activation signature set with flow cytometry. Not all increases in mRNA levels were reflected at the level of protein (Figure S2D), but for those that were, the increases in mRNA resulted in

increases both in the frequency of cells responding (data not shown) and in protein levels on a per-cell basis. This exemplified by the expression of Tbet (Figure 2F), IRF4 (Figure 2G), CD200 (Figure S2E), Ly6a (Figure S2F), and TNFSF11 (RANKL; Figure S2G). Thus, the graded increase in expression of activation signature genes at the population level is a reflection of both frequency of responding cells and incremental increases in expression levels on a single-cell level for at least a subset of genes.

An activation score ranks CD4+ T cell samples by activation status

Given that PC1 was able to distinguish between the five conditions according to activation state, we extracted the genes from the top and bottom of PC1 that were consistent across replicates to use as a general-purpose activation score able to correctly rank the five conditions by TCR signal (Figure 3A). We compared samples from several publicly available datasets, and the activation score was able to quantitatively rank conditions within a given experiment set such that activated and naïve CD4+ T cells could be distinguished and further that the effects of various genetic perturbations of CD4+ T cell responses could be observed. For example, the activation score correctly recapitulated the findings that polarized helper subsets of CD4+ T cells were more pro-inflammatory than unstimulated cells or induced and natural regulatory T cells[38] (Figure 3B); that at the population level plate-bound anti-CD3 and anti-CD28 induced stronger activation signals than APCs plus antigen[39] (Figure 3C); that costimulation was important for achieving higher activation states

but checkpoint inhibitors could block this effect[40] (Figure 3D); that knockout of Trim28, a molecule necessary for optimal IL2 production, diminished CD4+ T cell activation status[41] (Figure S3A); and that acute LCMV infection produced more robust activation in CD4+ T cells than chronic infection[42] (Figure S3B).

In order to test the value of the activation score, we used it to rank naïve CD4+ T cells from 39 inbred mouse strains[43] (Figure 3E). The activation score quantified the variability in the isolated CD4+ T cells according to activation status, revealing that the genetic differences between the strains yielded different levels of activity even under unstimulated conditions. As would be predicted by known strain phenotypes, C57Bl/6 cells were more activated than most strains, while BALB/c mice were less activated than most strains. The lupus-prone MRL strain and the type 1 diabetes-prone NOD strain had cells that ranked as relatively activated, whereas the type 1 diabetes-resistant NON strain had a relatively low activation score.

The strain with the highest activation score, DBA/2, had top-quartile expression of more than half of the activation signature genes ($p = 7.4e\text{-}30$ by chi-squared test). These included a number of immune effectors such as IRF4, CD25, IL2Rb, Nfkb1 (p105/p50), and Nr4a1 (Nur77), as well as 17 of 32 genes from the protein biosynthesis group and 5 of 12 genes from the molecular chaperone group. Differences in the immune phenotypes of the DBA/2 strain, such as resistance to malaria, have been largely attributed to B cell-dependent mechanisms[44], but the activation score here indicates that naïve CD4 T cells from DBA/2 mice are skewed toward an activated phenotype.

Three of four wild-derived strains had low activation scores: CAST, MSM, and WSB mice. All three of these wild-derived strains had bottom-quartile expression of the immune effectors Irf1, Irf8, Stat1, Nfkb1, and Tnf, indicating that these CD4+ T cells possess a less inflammatory gene expression profile under homeostatic conditions.

Thus, the activation score serves as a widely applicable and quantitative measure of CD4+ T cell activity, and can be used to assess the relative activation status of a variety of CD4+ T cell samples. We have developed a publically available, open source tool (see Materials and Methods) to facilitate the scoring and ranking of datasets by interested parties.

Pre-existing enhancers are leveraged to activate genes

In order to better understand the changes in genome-wide expression patterns that occurred with TCR stimulation, we compared enhancer landscapes with and without stimulation. We first performed ChIP-sequencing for H3K4me2, a marker of primed and active promoters and enhancers[45, 46], across the five conditions. By and large, the H3K4me2-marked regions across the five conditions were very similar, with a comparison of tag counts associated with specific genomic regions under no peptide or 1 μM PCC illustrated in Figure 4A. Though there were some enhancers showing at least two-fold change in H3K4me2 tag counts across conditions, these regions were at the lower end of the tag count range and therefore differences were not significant

(Figure 4A, red points). Thus, the gene expression and phenotypic changes seen after activation were not due to selection of new signal-dependent enhancers.

At promoters, H3K4me2 marks were shared across the five conditions, but activation signature genes showed spreading of the H3K4me2 mark along the body of the gene after TCR stimulation. In contrast, H3K4me2 peaks were narrow and focal for the untreated condition at many of these genes. This effect can be seen at CD69 (Figure S4A) and IRF4 (Figure S4B), resulting in a global increase of the ratio of gene body tags to promoter tags at activation signature genes (Figure S4C) but not genes in the bottom 10% of PC1 (Figure S4D). This implies that the process of activating these genes subsequent to TCR stimulation induces deposition of the dimethyl mark along the body of the genes as they are transcribed.

Motif analysis reveals lineage-determining and signal-dependent TFs

We used *de novo* motif finding[47] to identify lineage-determining transcription factors (LDTFs), also known as pioneer factors or master regulators, which establish cell-type-specific enhancer landscapes, and determine the available open chromatin for subsequent binding of signal-dependent transcription factors (SDTFs)[48, 49, 50, 51, 52]. The top motif was an ETS motif (Figure 4B), capable of being bound by a number of ETS factors that are expressed in CD4+ T cells, including Ets1, Ets2, and Elf1[53]. These enhancers tend to be shared across similar cells as well as thymic T cell precursors[47, 54]. Similarly, Runx factors play an important role in T cell development[55],

and correspondingly the Runx family motif was highly enriched among primed enhancers.

Several known motifs for SDTFs were also enriched among the H3K4me2-marked enhancers (Figure 4C), including an Interferon Regulatory Factor (IRF) motif. Although IRFs respond to interferon signaling[56], and would not be expected to be active in unstimulated cells[33, 57, 58], it is possible that the IRF motif is a "memory" of states of activation during the development of CD4+ T cells, and indeed IRF motifs can be found in several related cell types and multiple stages of thymocyte development (Figure S4E), suggesting that the primed enhancers in naïve CD4+ T cells are predisposed to act as binding sites for key SDTFs[47, 54, 59, 60, 61, 62]. Similarly, an AP-1 motif and an NF-κB motif were significantly enriched in primed enhancers (Figure 4C), corresponding with the fact that TCR signaling greatly increases activity of both of these transcription factors[32, 34].

Given that H3K4me2-marked regions were not substantially changed across the five conditions, we next performed ChIP-sequencing for H3K27Ac, a marker for active enhancers[63], under a stimulated condition (1uM PCC) and the unstimulated condition. In contrast to H3K4me2, a substantial portion of enhancers exhibited increases in the H3K27Ac activation mark (Figure 4D). The union set of enhancers was enriched for a similar set of motifs as the primed enhancers (Figure S4F, S4G). Enhancers that became more active with TCR engagement were highly enriched for both AP-1 and NF-κB motifs (Figure 4E), and were more likely to be proximal to activation signature genes than would be expected at random (p-value = 2.0e-20 by

chi-squared test). These enhancers included, for example, those upstream of Il2ra (CD25; Figure 4F) and CD69 (Figure 4G).

To investigate whether there was a quantitative relationship between TCR signal strength and enhancer activation, we performed independent ChIP-Seq for H3K27ac in response to both peptides at low and high concentrations. Given the prevalence of the AP-1 motif in the signal-responsive enhancers, we analyzed H3K27Ac tag counts at AP-1 binding sites genome-wide using publicly available ChIP-Sequencing data from *in vitro* activated TH17 cells[58]. There was an increase in H3K27Ac deposition at AP-1 binding sites that reflected the graded strength of TCR signaling (BATF shown in Figure 4H; other AP-1 family members in Figures S4H and S4I), indicating that AP-1 binding sites became more active in a graded manner corresponding to increasing TCR signaling. Graded changes in H3K27Ac were much less pronounced at CTCF binding sites, which occur at enhancers but are also more broadly distributed and play roles in establishing boundary elements.

Super-enhancers prime signaling genes

We next looked at changes in super-enhancers[64, 65, 66] upon activation using the H3K27Ac mark. Most super-enhancers (approximately 450 out of 700 total) identified were shared by both the unstimulated and stimulated conditions. GO analysis of genes nearby the shared super-enhancers showed enrichment for leukocyte activation genes and Pleckstrin homology genes (Figure 5A), indicating that super-enhancers in CD4+ T cells prime genes important for inflammatory signaling. These basally-primed

super-enhancers included regions near key T cell genes such as Ets1 (Figure 5B), Runx1 (Figure S5A), Ctla4/Icos/CD28 (Figure 5C), and IRF4 (Figure S5B). Notably, even though many of the super-enhancers exist prior to stimulation, super-enhancers near activation signature genes show an increase in H3K27Ac signal subsequent to TCR signaling (Figure 5D).

118 of the 568 super-enhancers identified after TCR stimulation were not identified as super-enhancers in the unstimulated condition. The super-enhancers that required TCR signaling were enriched for leukocyte activation genes (Benjamini-Hochberg adjusted p-value = 4.6e-7) crucial for T cell activation, including Batf (Figure 5E), Il2ra (CD25, Figure 5F), Tbx21 (Tbet, Figure 5G), Lag3 (Figure S5C), and Stat5b (Figure S5D).

ERK translates TCR signal strength downstream

The Ras/Raf/Mek/Erk pathway downstream of the TCR activates the AP-1 transcription factor family through a series of phosphorylation events and transcriptional induction of immediate-early genes[33, 34]. Given the fact that the AP-1 motif was enriched at enhancers showing increasing activity and the fact that AP-1 binding sites saw increasing H3K27Ac deposition, we sought to determine whether AP-1 and the ERK pathway were relevant to the increasing expression of activation signature genes across the conditions. We first compared the level of phosphorylated ERK (p-ERK) in each condition using flow cytometry, and found that, like CD69 and CD25, the amount of p-ERK in the p-ERK+ cells varied on a per-cell basis in each

condition, increasing with TCR signal strength (Figure 6A). As the increasing levels of p-ERK paralleled the general pattern of expression of the activation signature genes, we assessed binding frequencies of AP-1 factors in the gene promoters of the activation signature genes as compared to the bottom 10% of PC1 genes, and found that activation signature genes showed a significantly higher frequency of AP-1 binding (Figure 6B).

ERK pathway activation and AP-1 binding therefore seemed to correlate well with the graded profile of activation signature genes and the increasing activation score across the samples. In order to determine whether the gradual increase in ERK pathway activation was causal in translating TCR signaling into gradual increases in the expression of activation signature genes, we pretreated the CD4+ T cells with a low-dose MEK inhibitor (MEKi). MEK inhibition upstream of ERK was capable of suppressing p-ERK activity entirely, and titration of MEKi yielded intermediate levels of p-ERK on a per-cell basis (Figure S6A).

Low-dose MEK inhibition decreased the levels of the extracellular signaling marker CD69 (Figure S6B), which was in the top 10% of PC1, but this suppression was not universal, as CD4, an example of a gene not in the top 10% of PC1, was not significantly affected (Figure S6C). In order to see if this preferential suppression of activation signature genes was widespread, we performed RNA-seq on the five conditions after pretreatment with MEKi at IC50 (0.5 uM). If TCR signaling strength upstream of pERK yields graded levels of ERK that are in turn essential for the graded levels of activation response genes, then reduction of pERK levels should move each

sample downwards in activation score, such that the high-dose, high-affinity case looks like the low-dose, high-affinity; the low-dose, high-affinity looks like the high-dose, low-affinity; and so on.

Accordingly, MEK inhibition at IC50 decreased expression of activation signature genes (Figures 6C, S6D), but, as with extracellular expression of CD69, this effect was selective; expression of genes in the bottom 10% was increased or unchanged (Figures 6D, S6E). Using the activation score to assess total T cell activation status, we found that MEKi shifted each sample down in score (Figure 6E), as would be expected if the graded levels of pERK seen with each condition were prescriptive of the activation status of the condition. Thus, graded levels of pERK downstream of the TCR help to translate the analog activation signal to graded levels of enhancer activity and gene expression genome-wide (Figure 6F).

DISCUSSION

Understanding how CD4+ T cells respond to ligands of different doses and affinities is critical to understanding the nature of the adaptive immune response to both pathogens and self. Here, we have shown that the traditional model of a purely digital TCR response is too simple; on a per-cell basis, stronger TCR signals result in higher levels of phosphorylated ERK, a proportional increase in enhancer acetylation, and quantitative increases in activation markers such as CD69 and CD25 (Figure 6F). As a result of both these single-cell differences and the increasing frequency of respondent cells, varying the dose or the affinity of the pMHC-TCR interaction results

in a gene expression profile that is graded corresponding to increasing strength of TCR signaling. Notably, the predominance of PC1 and the graded gene expression patterns together indicate that dose and affinity are not interpreted separately downstream of the TCR, but rather overall signaling strength sets the level of activation across the population.

Ranking genes along a primary axis of variation allowed us to extract a set of activation signature genes that increase in a graded fashion at the population level proportionally to TCR signal strength, and further to establish an activation score that can rank arbitrary CD4+ T cell samples by the strength of signaling. The data presented here therefore gives us a greater understanding of the CD4+ T cell response to ligands of varying concentrations and affinities, and informs our understanding of the CD4+ T cell response under many conditions.

Significant differences in primed enhancers have been demonstrated under several stimulating conditions in macrophages, and demonstrate the ability of cells to quickly remodel chromatin to initiate particular gene expression programs[46, 67]. Surprisingly, we did not find significant changes in the primed enhancer landscape upon TCR activation in CD4+ T cells. Similarly, even the more labile activation mark H3K27Ac was largely similar across conditions, with many activation genes marked as super-enhancers even before stimulation. While it remains to be seen whether non-TCR signaling pathways or polarizing conditions induce more dramatic changes, the data presented here indicates that the CD4+ T cell enhancer landscape is largely pre-established, with subsets of H3K4me2-marked enhancers increasing in activity, but

little in the way of *de novo* enhancer establishment. This finding helps to explain the speed and plasticity of the CD4+ T cell response[68]—if all enhancers are primed basally, and many are even activated basally, then pro-inflammatory transcription factors can bind at established enhancers and initiate new gene expression programs with minimal additional transcriptional machinery.

Both the frequency of AP-1 binding and the level of pERK correlate with the strength of TCR signaling and the graded expression of activation signature genes. At least one of the feedback loops leading to digital TCR signaling, the son of sevenless (SOS) positive feedback loop, exists upstream of ERK, and it has been shown in thymocytes that pERK signaling is digital[16, 17, 18]. Our analog results for pERK can be interpreted to support the notion that TCR signaling in thymocytes functions differently than TCR signaling in mature T cells. Notably, both Themis and SOS, two key components of digital signaling in thymocytes, do not seem to be critical to mature T cell signaling[69, 70].

The graded levels of pERK in CD4+ T cells prove important for downstream enhancer and gene activity. We have here established a mechanistic link between the level of ERK signaling and the expression patterns of activation signature genes by using an inhibitor of MEK, upstream of ERK. Low-dose MEK inhibition selectively decreased expression of the activation signature genes such that the activation score under the inhibited conditions was incrementally decreased. This indicates that the analog levels of pERK seen on a per-cell basis are translated at a population level into increased enhancer and gene activity, and that “turning down” pERK levels selectively

diminishes the activation status of the cells. This finding is of particular interest in light of the clinical availability of numerous RAF, MEK, and ERK inhibitors[71, 72]. Our findings suggest that low-dose ERK pathway inhibition could be used to selectively decrease the activity of activation signature genes in CD4+ T cells, achieving low-level immunosuppression without killing T cells or completely removing their ability to respond to TCR signaling. Further, the effect of MEK inhibitors on CD4+ T cells raises questions about the immunosuppressive effects of using MEK inhibitors in cancer treatment, especially as current clinical trials combine MEK inhibitors with checkpoint-blockade inhibitors[71, 73].

In sum, this study makes use of a unique model system to dissect the transcriptional responses of CD4+ T cells to increasing strength of signaling, and demonstrates that analog levels of pERK within the context of digital TCR signaling flow downstream to result in graded gene expression profiles and enhancer landscapes that can be used to characterize CD4+ T cell signaling at large.

MATERIALS AND METHODS

Mice

AND mice on a B10.BR background were received from Dr. Michael Croft[7, 8] and bred in a specific pathogen free facility. All animal experiments were in compliance with the ethical standards set forth by UC San Diego's Institutional Annual Care and Use Committee (IUCAC).

Cells

Spleens were extracted and manually digested. CD11c+ cells were isolated using Miltenyi Biotec Inc. (San Diego, CA) MACS magnetic cell separation with positive selection for CD11c (CD11c, Biolegend, cat. no. 117304). Subsequently, the CD11c- splenic fraction was used to negatively select for naïve CD4+ T cells using the Miltenyi MACS system with the following antibodies: CD11c (Biolegend, cat. no. 117304); CD45R (eBioscience, cat. no. 13-0452-86); CD11b (eBioscience, cat. no. 13-0112-86); CD25 (eBioscience, cat. no. 36-0251-85); CD49b (eBioscience, cat. no. 13-5971-85); CD69 (eBioscience, cat. no. 13-0691-85); CD8a (eBioscience, cat. no. 13-0081-86); Ly-6G (eBioscience, cat. no. 13-5931-86); MHC class II (eBioscience, cat. no. 13-5321-85); TER-119 (eBioscience, cat. no. 13-5921-85. CD11c+ and CD4+ cells were cultured at a ratio of 1:2 in 96-well round-bottom plates for 24 hours, 108 hours (for proliferation assay), or 3.5 hours (for ERK phosphorylation staining). Peptides were added at indicated concentrations with the CD11c+ and CD4+ cells in DMEM supplemented with 10% Fetal Bovine Serum. For sequencing experiments, CD4+ cells were re-isolated from the culture using the Miltenyi MACS system and the same set of antibodies as above less CD25 and CD69. For phospho-ERK staining, whole splenic cells were used, rather than purified CD11c+ and CD4+ cells.

Peptides

Peptides were ordered from Peptide 2.0 (Chantilly, VA) with the following amino acid sequences[7, 8]:

PCC – KAERADLIAYLKQATAK

K99A – KAERADLIAYLAQATAK

Y97K – ANERADLIAKLKQATK

K99E – ANERADLIAYLEQATK

MCC – ANERADLIAYLKQATK

Lyophilized peptides were resuspended in water, and added at the indicated concentrations to the cell cultures. Unstimulated CD4s received an equivalent amount of water alone.

Flow cytometry

Flow cytometry was performed on a LSR II and LSR Fortessa, both from BD Biosciences (San Jose, CA). Cells were stained as per manufacturers' protocols with the following antibodies: CD4-APC (eBioscience, cat. no. 17-0042-83); CD4-PE-Cyanine7 (BioLegend, cat. no. 116016); CD69-FITC (eBioscience, cat. no. 11-0691-82); CD25-PE (eBioscience, cat. no. 12-0251-82); Valpha11-FITC (BD Pharmingen, cat. no. 553222); Vbeta3-PE (BD Pharmingen, cat. no. 553209); CD11c-PE-Cyanine7 (eBioscience, cat. no. 25-0114-82); IRF4-PerCP-Cy5 (eBioscience, cat. no. 46-9858-80); Tbet-PE (Santa Cruz Biotechnology, cat. no. SC-21749); CD122-PE (BioLegend, cat. no. 105905); Ly6a-APC-Cyanine7 (BioLegend, cat. no. 108125); CD200-APC (BioLegend, cat. no. 123809); TNFSF11-APC (BD Biosciences, cat. no. 560296); phospho-ERK-Alexa Fluor 488 (Cell Signaling, cat. no. 4344S). Live/dead staining was performed using Fixable Aqua (Life Technologies, cat. no. L34957; or Biolegend,

cat. no. 423102). Cells were gated on CD4+, Aqua- cells. For phosphor-ERK staining, permeabilization was performed using BD Phosflow Perm Buffer III (cat. no. 558050) and BD Fix Buffer I (cat. no. 557870). Analysis was performed with FlowJo v10.6 (Tree Star; Ashland, OR).

Sequencing

Prior to sequencing, CD4+ T cells were separated from the co-cultured cells using Miltenyi MACS negative selection as described above for the initial culturing. ChIP-sequencing for H3K4me2 and H3K27Ac in the 1uM peptide treatments was performed as described[74], with the following modifications: sodium butyrate was used to inhibit de-acetylation; and three RIPA and three LiCl washes were performed instead of five and one. ChIP-sequencing for H3K4me2 and H3K27Ac across the five conditions was performed as described[75]. RNA-sequencing was performed as described, with minor modifications[76].

ChIP-sequencing antibodies used were: H3K4me2 (Millipore, cat. no. 07-030) and H3K27Ac (Abcam, cat. no. ab4729 and Active Motif, cat. no. 39135).

MEK inhibitor treatment

CD4+ T cells, isolated as described above, were pre-treated with 0.5uM Promega U0126 (cat. no. V1121) for thirty minutes at 37ºC. CD11c+ cells and peptides were subsequently as indicated and cultured in the presence of the inhibitor for 24 hours.

Analysis

ChIP-sequencing reads were mapped to the mm10 genome using Bowtie2[77], and RNA-sequencing reads were mapped using STAR[78]. Default allowed error rates were used, and only uniquely mapping reads were used in downstream analysis. Initial processing of aligned data and peak calling was performed using Homer[47]. IDR analysis[79] for ChIP-sequencing replicates was performed using the homer-idr package[80]. Vespucci[81] was used for counting AP-1 tags in gene regions.

Gene Ontology analysis was done with the DAVID Gene Functional Classification tool using the default background[82].

Super-enhancers were called using Homer[47], which follows the published procedure[64, 65, 66] by first stitching together peaks into larger regions and then sorting regions by normalized H3K27Ac tag count. Region scores are plotted against rank, and a threshold is defined by finding the point at which the tangent to the plotted rank-scores is one. Regions past that threshold are called super-enhancers.

Underlying data sets, including RPKM values and peaks, as well as code for all analyses described is publicly available at https://github.com/karmel/and-tcr-affinity. Analyses were performed using iPython Notebook[83]. Clustering and PCA was performed using the scikit-learn package[84]. For full execution details and parameters, please see the code in the Github repository linked here.

Activation signature scores

To generate the list of genes used in the activation signature score, we separately ran Principal Component Analysis on two replicates of RNA-Seq data and also the combined expression data from both replicates. Genes with an RPKM less than 100 in the No Peptide condition or a standard deviation greater than 20% of the No Peptide expression level across replicates were then omitted from the target set of genes. Remaining genes were sorted along PC1, and genes that were in the top ten percent in all three data sets (215) or the bottom ten percent in all three data sets (137) were included in the set of activation signature score genes used in analysis.

To compute the activation signature score, we take the dot product of the values of genes along PC1 in the combined RNA-Seq data set and the mean-centered expression levels for those genes for each sample in an experimental data set, yielding a single scalar score for each experiment. The scores across samples are then scaled by the max score, ensuring values are in the range of [-1, 1].

The activation signature score tool is described and downloadable here: https://github.com/karmel/and-tcr-affinity/tree/master/andtcr/rna/activationscore

Public data

Publicly available datasets used for the analyses in Figure 3 and Figure S3 is available from GEO with the following Accession Codes: GSE14308, GSE32224, GSE41866, GSE42276, GSE54938, and GSE60337. AP-1 binding data is from GSE39756. For figure S5, the following datasets were used: GSE56456, GSE31233, GSE40463, GSE21365, GSE56098, GSE21512.

Accession codes

Raw and processed data are provided in the Gene Expression Omnibus (GEO) under accession number GSE69545.

CONTRIBUTIONS

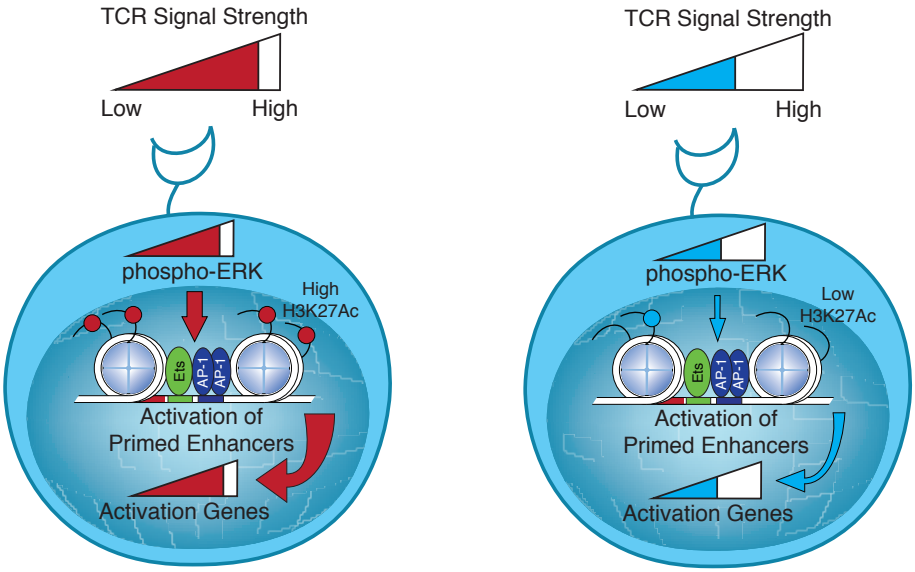
KAA, CKG, and SMH designed the studies. KAA performed experiments and analyzed data. ELS performed experiments and advised study design. JGC, DG, and TDT performed experiments.

ACKNOWLEDGMENTS

The authors would like to thank Leslie Van Ael and David Allison for assistance with preparation of the manuscript. These studies were primarily supported by NIH grants DK091183, CA17390, DK063491, R01-AI103440, and the San Diego Center for Systems Biology (GM085764). KAA was supported by F31-AI112269 (NIAID); ELS was supported by K01-DK095008 (NIDDK); TDT was supported by T32-CA009523 (NIH); and DG was supported by a Canadian Institutes of Health Research Fellowship. The authors declare no conflict of interest.

Chapter Two, in full, is currently being prepared for submission for publication: Allison, Karmel A; Stone, Erica L; Collier, Jana G; Gosselin, David; Troutman, Ty Dale; Hedrick, Stephen M; Glass, Christopher K. The dissertation author is the primary author of this paper.

Figure 2.0: Graphical abstract



Allison 2015
Graphical abstract

Figure 2.1: Both frequency of responding cells and per-cell activation levels increase with increasing signal strength.

A. Purified AND T cells and CD11c+ APCs were co-cultured in the presence of indicated peptides at the indicated concentrations for 24h. Flow cytometry was then used to phenotype the CD4+ T cells. The histograms show CD69 expression of CD4+ T cells. The annotated bar indicates the gate used to identify CD69+ CD4 cells in subsequent figures.

B. Gating on CD4+ cells as in 1A, there is a bimodal distribution of CD25 expression resulting from activating levels of high-affinity (PCC) or low-affinity (K99A) peptides. The annotated bar indicates the gate used to identify CD25+ cells.

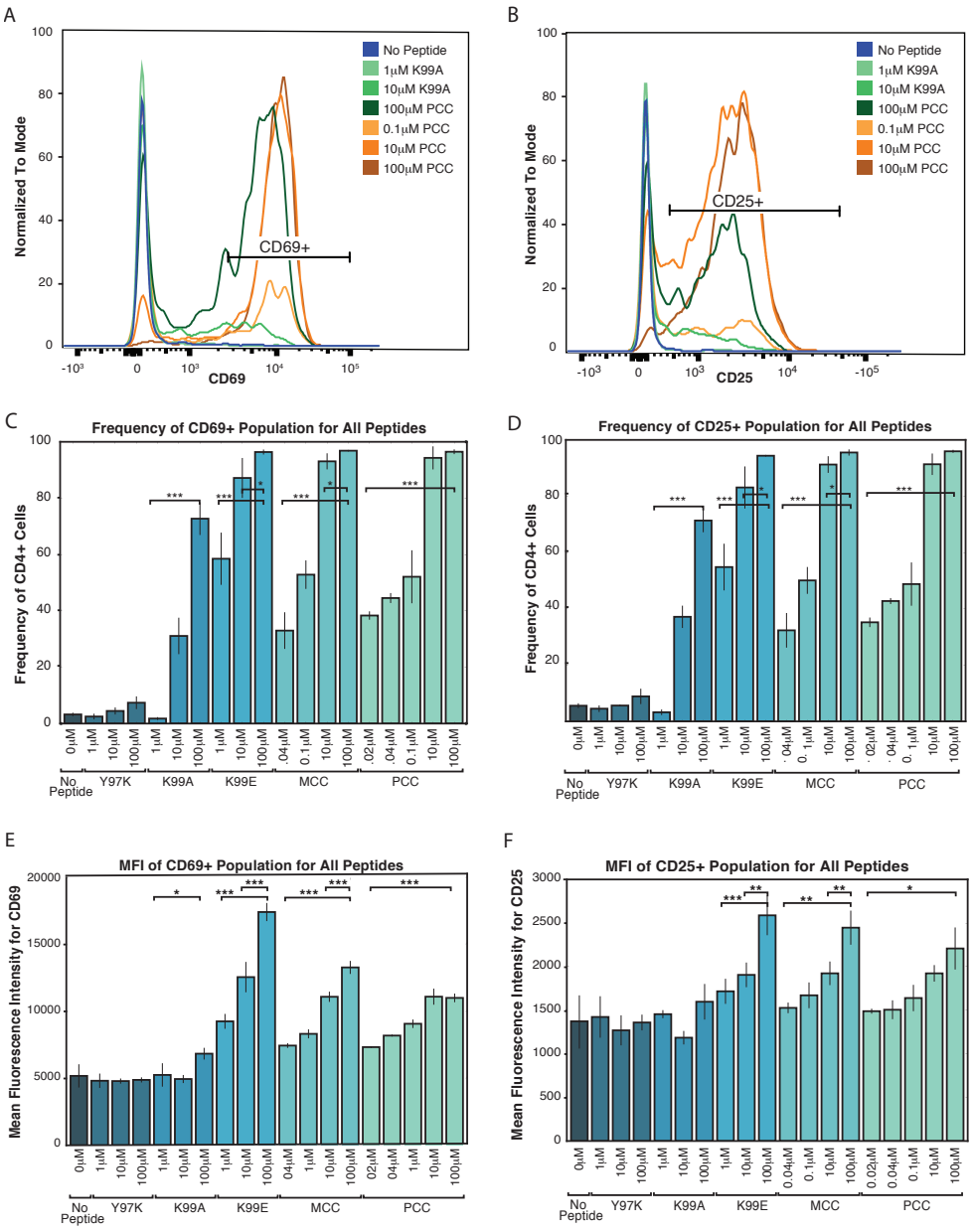
C. The percent of CD4+ cells that are CD69+ (using gate shown in 1A) varies with the peptide presented and concentration of the indicated peptide.

D. The percent of CD4+ cells that are positive for the activation marker CD25 varies with both peptide and dose.

E. Gating on CD4+ CD69+ cells (as shown in 1A), the geometric mean fluorescence intensity (MFI) of CD69 per cell in each condition varies.

F. The geometric MFI of CD25, gated on CD4+ CD25+ cells (as shown in 1C), varies with peptide and dose.

(P-values based on Student's t test; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.)



Allison 2015
Figure 1

Figure 2.2: RNA-Sequencing reveals graded expression of activation signature genes.

A. Principal component analysis (PCA) of the approximately 3,200 genes that changed between any two samples reveals that the primary axis of variation (PC1, shown along the x-axis) orders the five conditions by increasing TCR signal strength: No Peptide; low-dose, low-affinity (10uM K99A); high-dose, low-affinity (100uM K99A); low-dose, high-affinity (0.1uM PCC); and high-dose, high-affinity (10uM PCC).

B. After ordering the ~3,200 genes used for PCA by their contribution to PC1, we extracted the top 10%—that is, the ~320 genes contributing most positively to a sample's PC1 value—and the bottom 10%—that is, the ~320 genes contributing most negatively to a sample's PC1 value. Each group displays a clear trend, with the top 10% increasing in expression as signal strength increases, and the bottom 10% decreasing in expression. Each blue line represents a gene, with reads per kilobase per million (RPKM) normalized from 0 to 1 across the five conditions. Significance was determined using permutation testing, where the mean difference between genes in the No Peptide sample as compared to 10uM PCC was normally distributed over randomly generated groups of genes. This normal distribution was compared to the top 10% and bottom 10% genes to generate a p-value.

C. Genes in the top 10% of PC1, termed activation signature genes, include many genes previously identified as important to CD4+ T cell activation, such as Tbx21 (Tbet), Stat1, and Tnf. Reads per kilobase per million (RPKM) for each increases with increasing signaling strength.

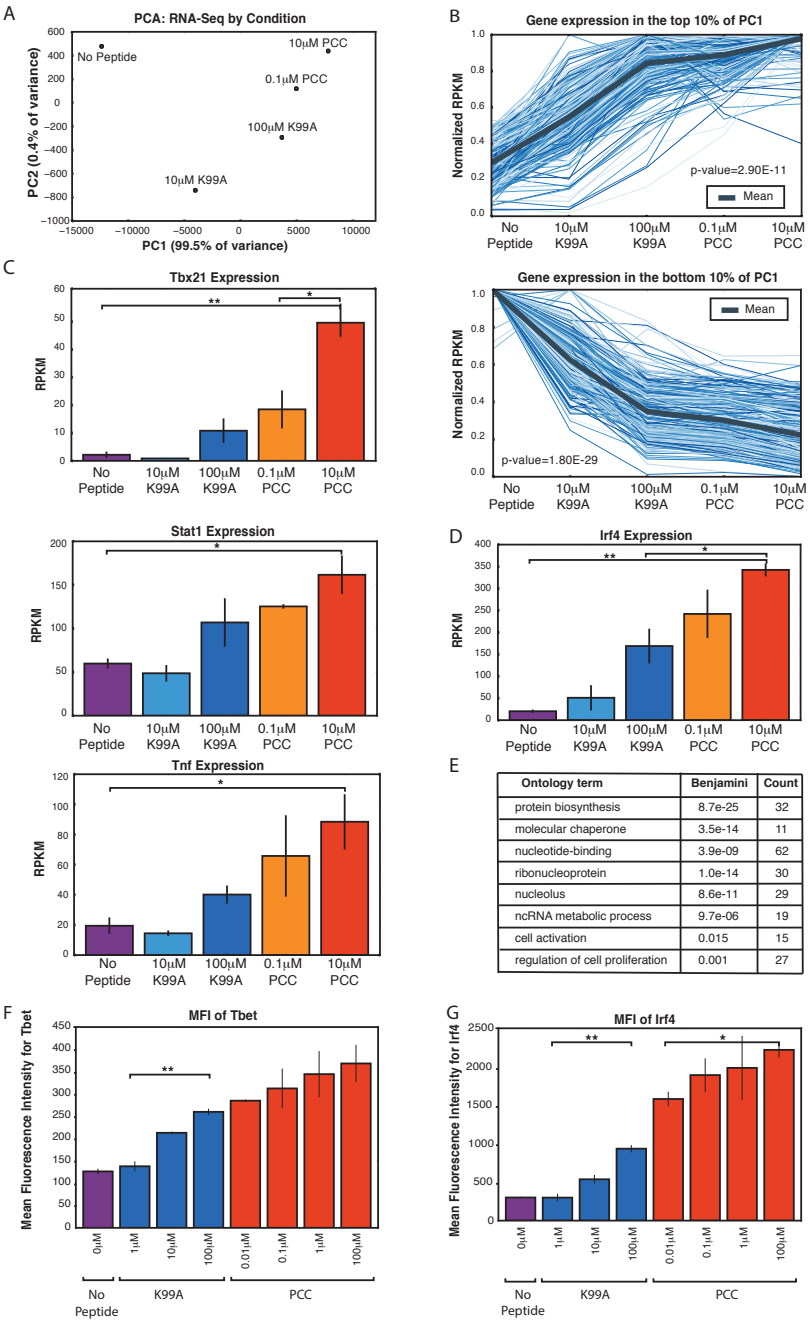
D. Irf4, a transcription factor previously shown to be more highly expressed with increasing TCR affinity, shows graded expression across the five conditions.

E. Gene Ontology (GO) analysis of activation signature genes shows enrichment for protein biosynthesis and molecular chaperone genes. P-values shown are Benjamini-Hochberg adjusted p-values.

F. As measured by flow cytometry, the geometric MFI of Tbet in CD4+ cells increases on a per-cell basis with increasing signal strength.

G. Similarly, per-cell protein levels of IRF4 increase with increasing signal strength when measured with flow cytometry.

(P-values based on Student's t test; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.)



Ontology term	Benjamini	Count
protein biosynthesis	8.7e-25	32
molecular chaperone	3.5e-14	11
nucleotide-binding	3.9e-09	62
ribonucleoprotein	1.0e-14	30
nucleolus	8.6e-11	29
ncRNA metabolic process	9.7e-06	19
cell activation	0.015	15
regulation of cell proliferation	0.001	27

Allison 2015
Figure 2

Figure 2.3: PC1 can be used to rank arbitrary CD4+ T cell data sets.

A. An activation score derived from the top and bottom genes along PC1 ranks the five conditions according to TCR signaling strength. The score correctly captures that 100uM K99A and 0.1uM PCC are very similar in activation status. Note that the score ranks samples within an experiment, but is not an absolute metric for comparing across experiment groups.

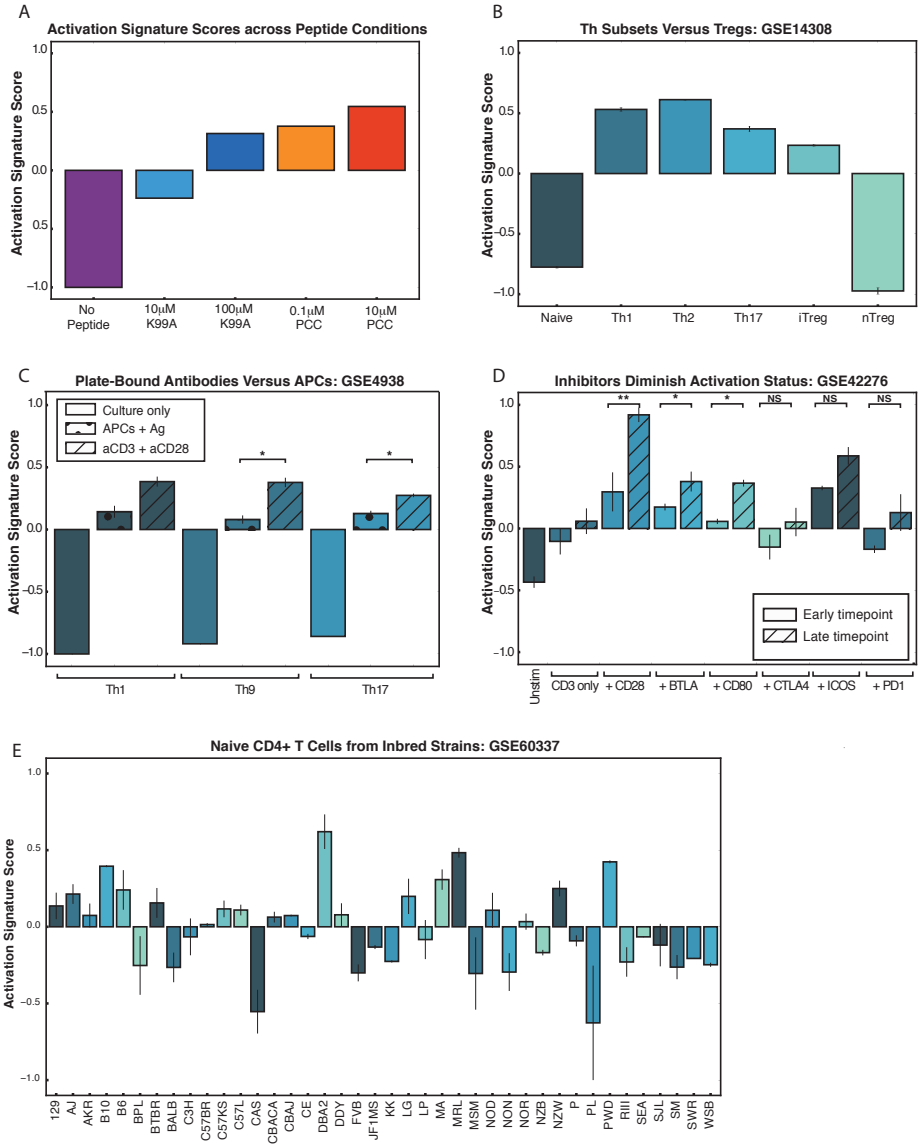
B. The activation score can be used to compare arbitrary CD4+ T cell data sets. Here, activation scores were calculated for microarrays from helper T cell subsets, and they demonstrate that naïve cells and native regulatory T cells (nTregs) are less classically activated than Th1, Th2, and Th17 polarized subsets.

C. The activation score captures the fact that plate-bound anti-CD3 and anti-CD28 stimulation of helper T cell subsets results in stronger signaling than antigen as presented by APCs.

D. CD4+ T cells were subjected to a variety of stimulatory or inhibitory treatments: anti-CD3 alone, or anti-CD3 with anti-CD28, anti-BTLA, anti-CD80, anti-CTLA4, anti-ICOS, or anti-PD1. Gene expression profiles at early (1h and 4h) and late (20h and 48h) time points yield activation scores in line with the characterization of CD28, BTLA, CD80, and ICOS as co-stimulatory, and CTLA4 and PD1 as inhibitory. Although it might be expected that anti-CD80 would have an inhibitory effect, these results are in line with the conclusions from the originally published analysis.

E. Naïve CD4+ splenocytes were isolated from 39 mouse strains. Using the PC1-derived activation score, we can rank the CD4+ cells from each strain as either more or less activated under basal conditions. Using the activation score, we recapitulate the finding that C57Bl6 mice have more pro-inflammatory cells than BALB/c mice. The highest scoring strain, DBA/2, shows top-quartile expression of immune effectors as well as protein biosynthesis genes.

(P-values based on Student's t test; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.)



Allison 2015
Figure 3

Figure 2.4: Primed enhancers are pre-existing, but gain activation markers with treatment.

A. Comparing primed enhancers marked by H3K4me2 peaks reveals strong correlation between untreated and treated samples. Normalized tag counts in the No Peptide condition are plotted against those in a 1uM PCC condition, with red dots coloring those that are more than two-fold up-regulated in the 1uM PCC condition. The up-regulated enhancers are both few in number and low in tag count.

B. *De novo* motif finding identifies lineage-determining transcription factor (LDTF) motifs among primed enhancers shared by the five conditions. An ETS motif is most prominent, and a RUNX motif is likewise highly enriched over the randomly selected background. Both ETS and RUNX factors play important roles in T cell development.

C. Among primed enhancers shared by all five conditions, including the untreated condition, pro-inflammatory transcription factor motifs are enriched. An IRF family motif, AP-1 family motif (represented by BATF), and NFkB motif (represented by REL) are all significantly enriched among shared enhancers marked by H3K4me2.

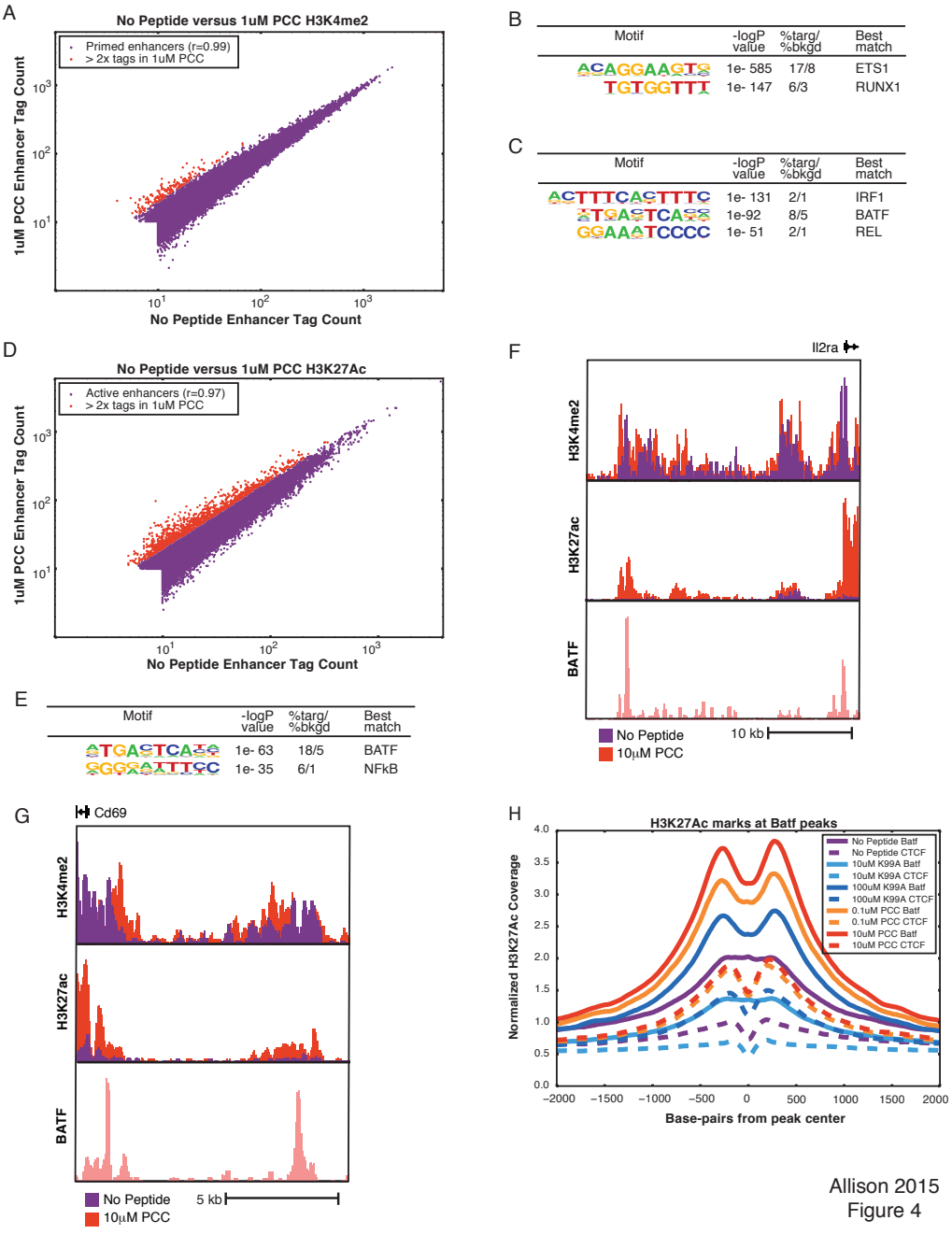
D. Comparing H3K27Ac tag counts at enhancers in No Peptide as compared to 1uM PCC treatment reveals that many enhancers see increasing H3K27Ac deposition upon stimulation. Points in red indicate greater than two-fold increase in tags upon treatment.

E. Enhancers that are more active upon stimulation, as determined by greater than two-fold H3K27Ac tags in 1uM PCC treatment as compared to No Peptide, are enriched for pro-inflammatory transcription factor motifs. BATF, an AP-1 family member, and NFkB are most prominent.

F. Enhancers that are more active with stimulation are enriched near activation signature genes, as can be seen with this enhancer upstream of the activation signature gene Il2ra (CD25).

G. Enhancers upstream of the activation signature gene CD69 show an increase in H3K27Ac deposition upon treatment with 1uM PCC.

H. Genome-wide, deposition of H3K27Ac, a marker of transcription factor activity, reflects increasing TCR signal strength at the binding sites of AP-1 family members, including Batf.



Allison 2015
Figure 4

Figure 2.5: Super-enhancers prime T cell activation genes.

A. Gene Ontology (GO) analysis of genes nearest to the ~450 super-enhancers shared by treated and untreated conditions show enrichment for T cell activation genes. P-values shown are Benjamini-Hochberg adjusted p-values.

B. H3K27Ac marks a large super-enhancer around the lineage-determining transcription factor Ets1 in both the No Peptide and 1uM PCC conditions. The super-enhancer spans the ~600 kbp region shown.

C. The ~400 kbp super-enhancer region encompassing Cd28, Ctla4, and Icos is marked by H3K27Ac in both treated and untreated conditions.

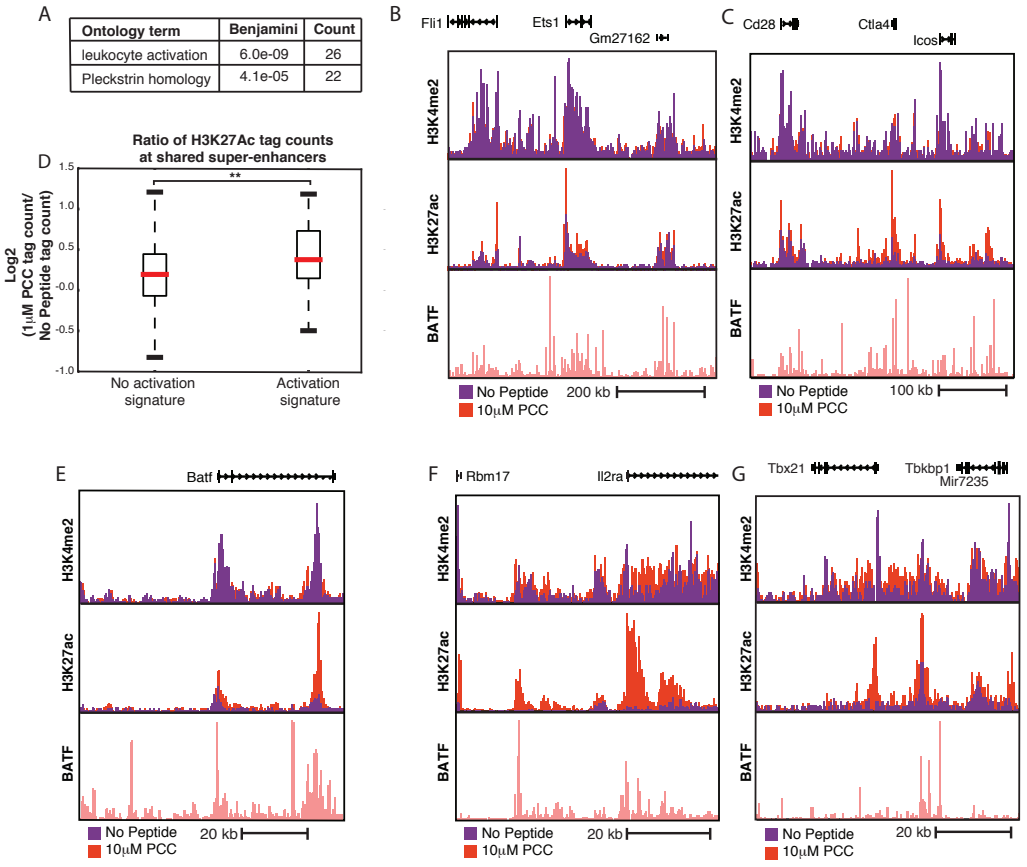
D. Despite being heavily marked by H3K27Ac in both untreated and treated conditions, shared super-enhancers near activation signature genes show a significant gain in H3K27Ac tags in response to stimulation as compared to the shared super-enhancers not proximal to activation signature genes. In other words, basally primed super-enhancers near activation signature genes see significant increases in activity upon stimulation, correlating with increased gene expression at the activation signature genes.

E. Some regions of H3K27Ac deposition required TCR stimulation to pass the super-enhancer threshold, as can be seen here at the ~60 kbp region encompassing Batf, an AP-1 family member. While H3K27Ac is clearly present under basal conditions, there is a substantial increase in enhancer activity upon treatment with 10uM PCC.

F. Il2ra (CD25) shows increased enhancer activity and formation of a super-enhancer in the treated condition.

G. Similarly, the region surrounding Tbx21 (Tbet) shows substantial increases in activity subsequent to stimulation, resulting in the formation of a super-enhancer.

(P-values based on Student's t test; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.)



Allison 2015
Figure 5

Figure 2.6: ERK signaling translates TCR signal strength into graded gene expression.

A. ERK phosphorylation is a measure of ERK pathway activity. Flow cytometry for phospho-ERK after 3.5 hours of co-culturing shows that, on a per-cell basis, increasing signal strength yields increasing levels of phospho-ERK among CD4+ phospho-ERK+ cells.

B. ERK pathway activation is upstream of the transcription factor AP-1. ChIP-sequencing tags for four AP-1 family members (Batf, cJun, JunB, and JunD) in Th17 cells shows that there is an enrichment for AP-1 binding near the promoters (plus or minus 1,000 bp from the TSS) of activation signature genes (top 10% of PC1) as compared to all genes or the genes in the bottom 10% of PC1.

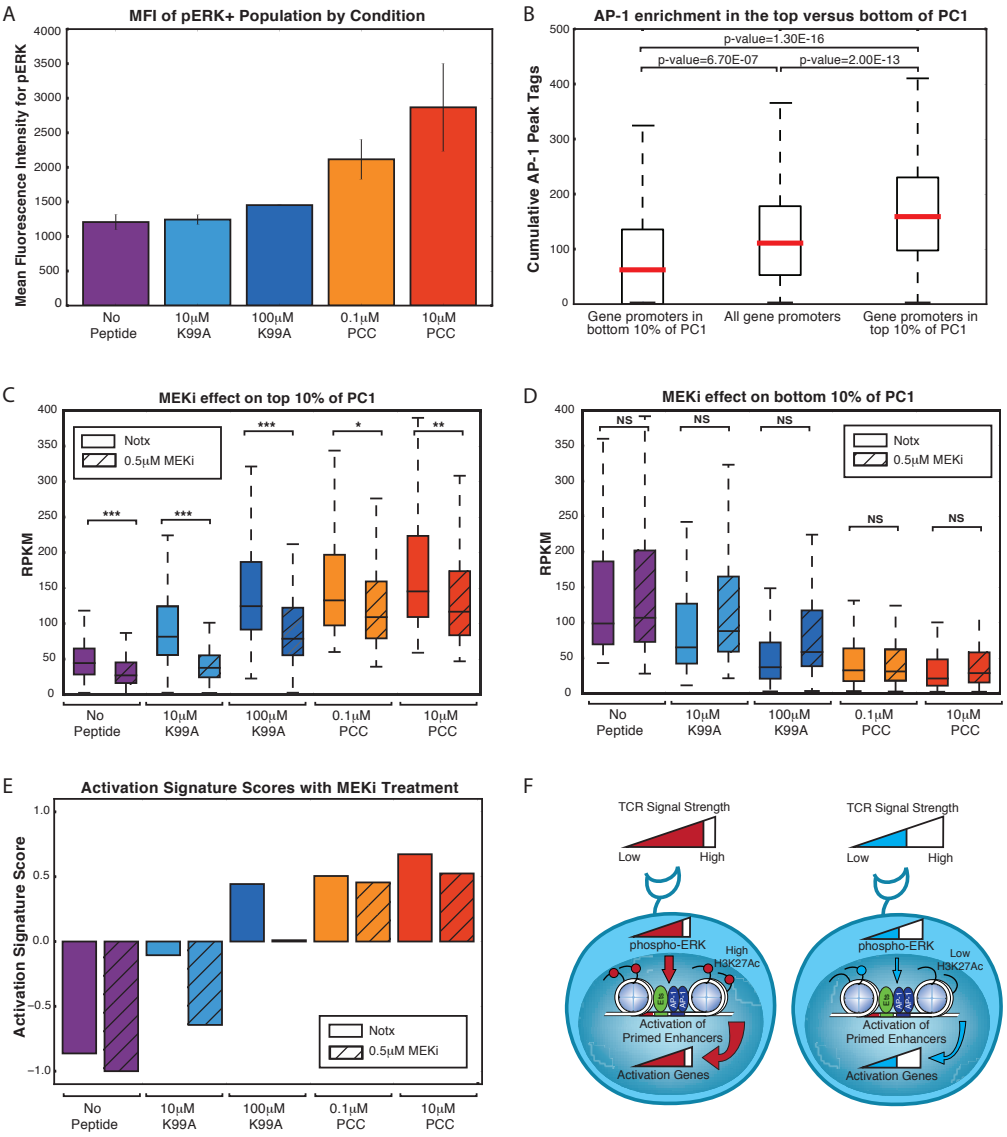
C. A MEK inhibitor that dampens signaling upstream of the ERK pathway preferentially diminishes expression of activation signature genes, as seen in the fact that the RPKM of genes in the top 10% of PC1 is significantly reduced with treatment.

D. The reduction of RPKM seen with the activation signature genes is not a general effect, as the RPKM of the genes in the bottom 10% of PC1 are not significantly affected by MEK inhibitor treatment.

E. We quantified the effect of MEK inhibitor treatment using the activation signature score. Treatment with the MEK inhibitor reduces the activation signature score for all samples.

F. Schematic of the ERK-AP-1 axis. See text for details.

(P-values based on Student's t test; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.)



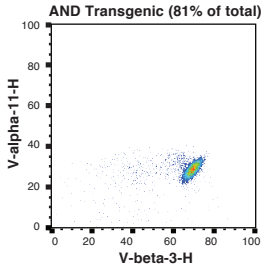
Allison 2015
Figure 6

Figure 2.S1: CD4+ T cells and APCs were purified from AND mice.

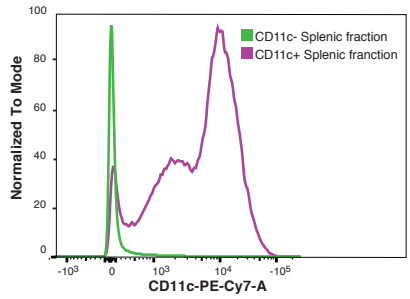
A. The AND mouse TCR includes the V alpha 11 chain and the V beta 3 chain. The great majority of naïve CD4+ T cells extracted from AND mouse spleens expressed this pair of TCR chains. Naïve CD4+ T cells were isolated using negative selection with the Miltenyi MACS system, as described in the methods section.

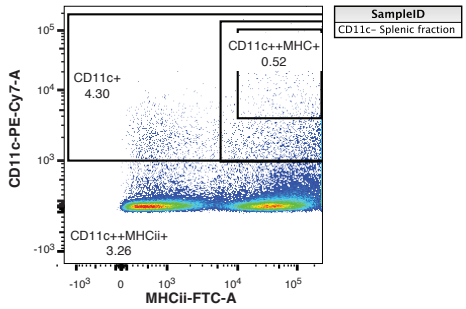
B. APCs for peptide presentation were extracted from mouse spleens using positive selection for CD11c. The extracted cells were largely positive for both CD11c and MHC class II molecules.

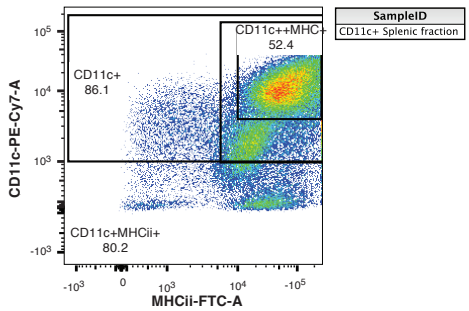
A



B







Allison 2015
Figure S1

Figure 2.S2: RNA-Sequencing reveals graded expression of activation signature genes.

A. 99.5% of the variance between the five conditions could be explained with a single principal component.

B. Eif3a, a protein biosynthesis gene, increased in expression with increasing TCR signaling strength.

C. Cct2, a molecular chaperone gene, increased in expression with increasing TCR signaling strength.

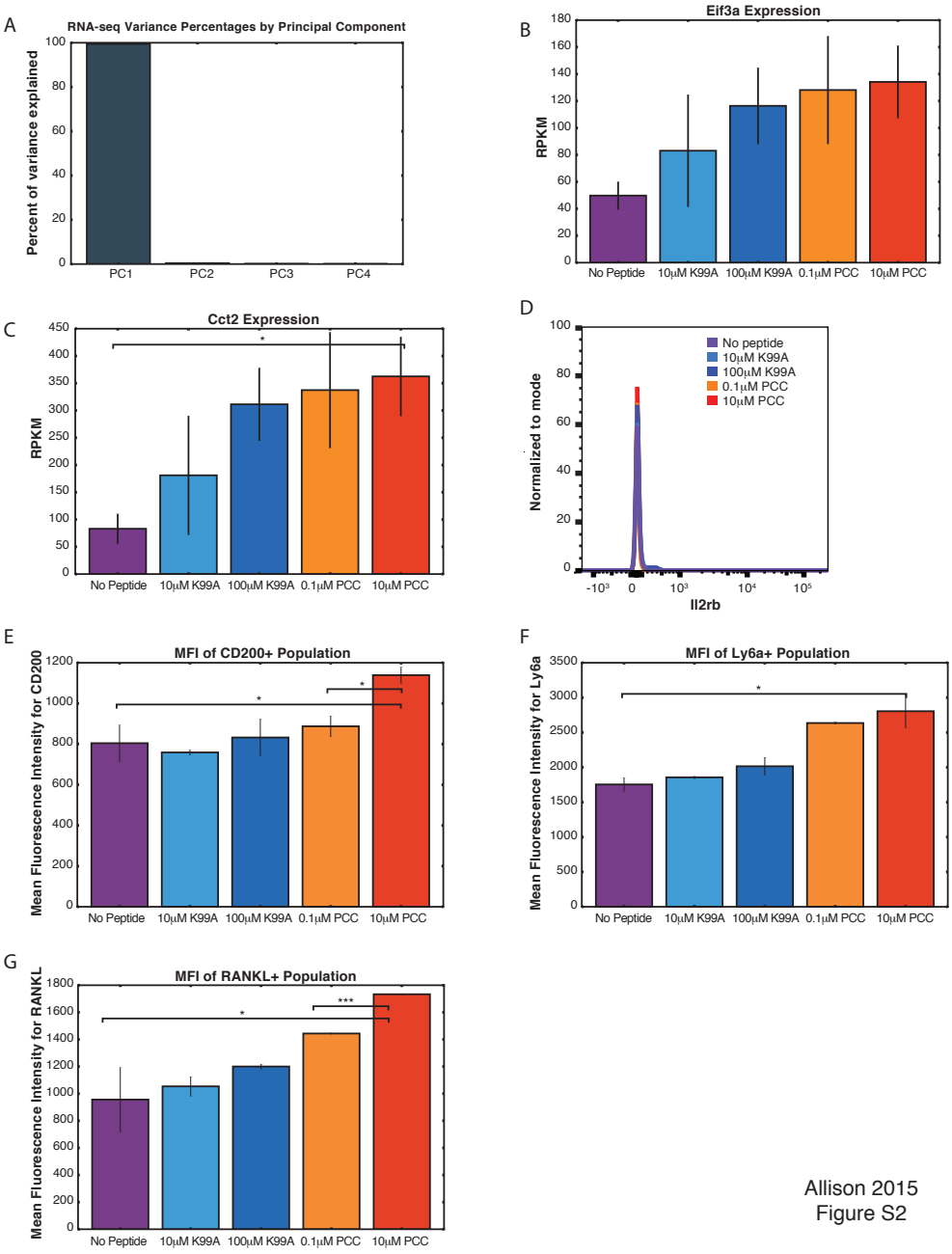
D. Not all gene expression changes are reflected in extracellular protein levels. Il2rb, an activation signature gene, does not significantly change when measured with extracellular flow cytometry.

E. CD200, an activation signature gene, shows a graded gene expression pattern across the five conditions, and this is reflected by extracellular presentation of the CD200 molecule as measured by flow cytometry. This plot shows the geometric MFI of CD4+ CD200+ cells across the conditions.

F. Similarly, Ly6a increases in both population gene expression levels and single-cell protein levels as measured by flow cytometry. This plot shows the geometric MFI of CD4+ Ly6a+ cells across the conditions.

G. The receptor TNFSF11 (RANKL) shows a graded increase in gene expression that is reflected in the per-cell levels of extracellular expression of the protein. This plot shows the geometric MFI of CD4+ RANKL+ cells across the conditions.

H. (P-values based on Student's t test; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.)

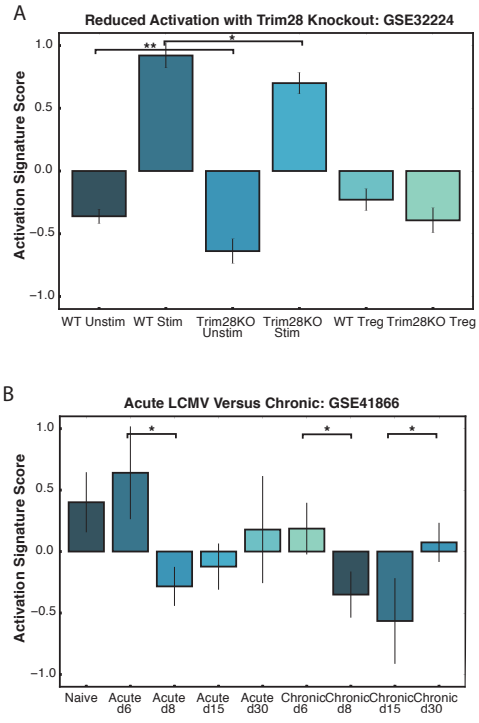


Allison 2015
Figure S2

Figure 2.S3: PC1 can be used to rank arbitrary CD4+ T cell data sets.

A. The activation signature score was used to quantify CD4+ T cell activation status under Trim28 knockout conditions. Loss of Trim28 resulted in a lowering of the activation signature score, corroborating the previously reported results that the Trim28-deficient cells produced less IL2.

B. Acute LCMV infection results in a higher activation signature score for CD4+ T cells at early time-points, marking the peak of infection before cells begin to turn off activation programs. The activation signature score does not reach such a high level in a chronic infection model.



Allison 2015
Figure S3

Figure 2.S4: Primed enhancers are pre-existing, but gain activation markers with treatment.

A. Despite the similarity of enhancer profiles across conditions, the H3K4me2 mark "spreads" from the transcription start site along the body of genes at a subset of genes. This subset is significantly enriched for activation signature genes such as CD69, shown here.

B. Irf4, another activation signature gene, shows a similar spreading of dimethyl along the body of the gene after treatment.

C. Normalized tag counts along the first 4,000 bp of all activation signature genes show that, overall, the No Peptide condition shows a narrow peak at the transcription start site, but this peak is smoothed out as the dimethyl mark spreads along the body of the gene in the treated samples.

D. The spread of dimethyl seen at activation signature genes is not present at genes in the bottom 10% of PC1, where all five conditions show a similar pattern of dimethyl deposition across the first 4,000 bp of the genes.

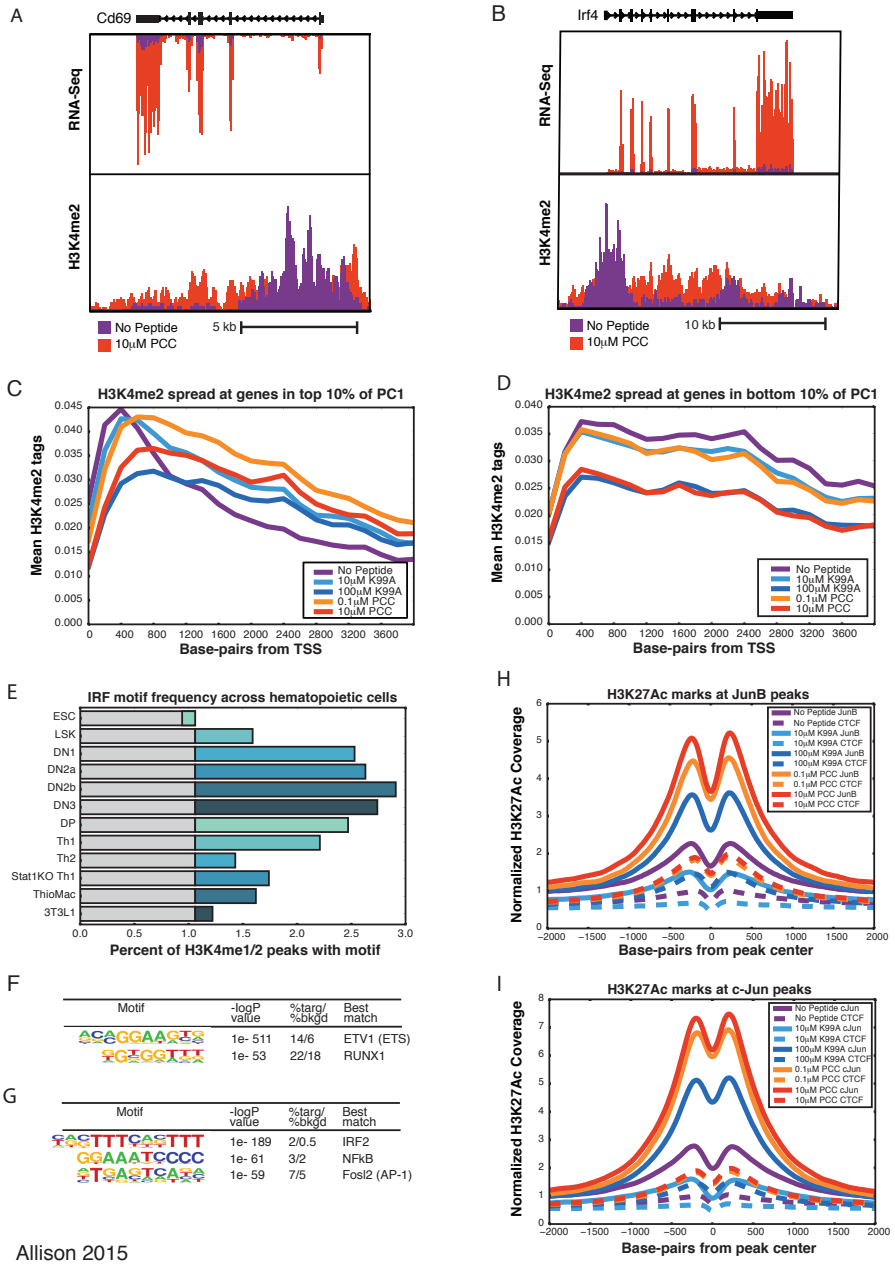
E. The Interferon Response Family (IRF) motif is enriched in enhancers from many cells in the T cell lineage, with motif frequency peaking in thymocytes (DN1 through DP). The grey portion of the bars represents mean background enrichment of the motif, and the colored section of the bar shows the difference in enrichment between the cell type indicated and the background. From top to bottom, the cells indicated are: Embryonic stem cells; $Lin^{-}Sca\text{-}1^{+}c\text{-}Kit^{+}$ (LSK) hematopoietic progenitor cells; fetal liver derived double negative 1 thymocytes ; fetal liver derived double negative 2a thymocytes ; fetal liver derived double negative 2b thymocytes; double negative 3 thymocytes; double positive thymocytes; Th1 polarized CD4+ T cells; Th2 polarized CD4+ T cells; Th1 polarized CD4+ T cells from a Stat1 knockout model; thioglycollate-elicited macrophages; and the adipocyte-derived 3T3L1 cell line.

F. *De novo* motif finding identifies lineage-determining transcription factor (LDTF) motifs among H3K27Ac-marked enhancers shared by the five conditions. As with the primed enhancers, the ETS family motif and a RUNX motif are highly enriched.

G. Among activated enhancers shared by all five conditions, pro-inflammatory transcription factor motifs are enriched. As in primed enhancers, IRF, NFkB, and AP-1 motifs are all enriched in the shared H3K27Ac peaks.

H. As seen with Batf, deposition of H3K27Ac, a marker of transcription factor activity, reflects increasing TCR signal strength at the genome-wide binding sites of the AP-1 family member JunB.

I. Similarly, deposition of H3K27Ac reflects increasing TCR signal strength at the genome-wide binding sites of the AP-1 family member cJun.



Allison 2015
Figure S4

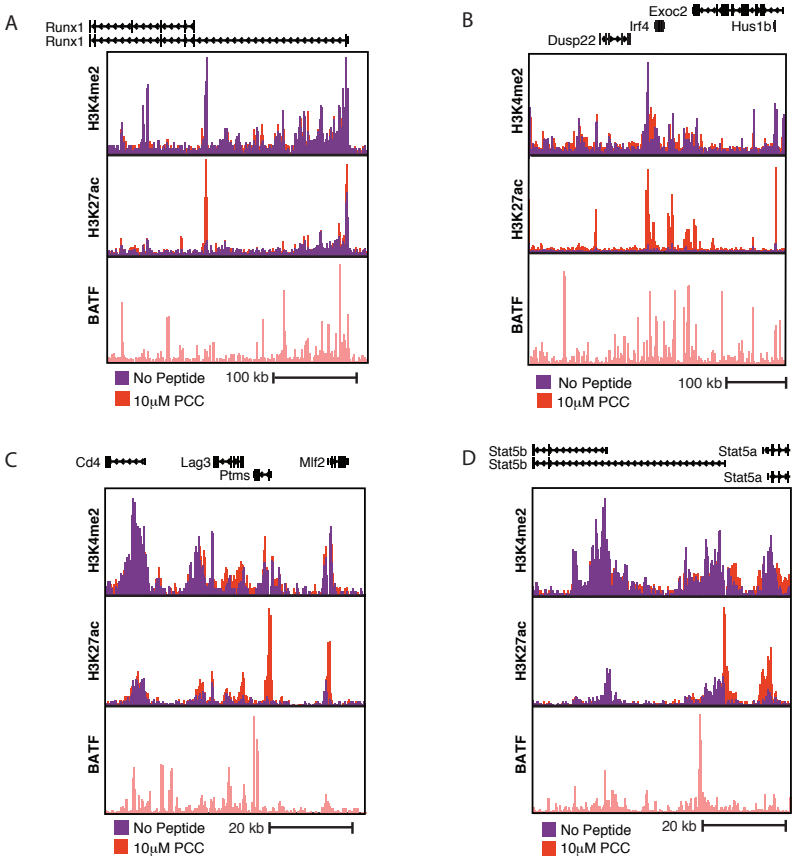
Figure 2.S5: Super-enhancers prime T cell activation genes.

A. A basally-primed super-enhancer encompasses the Runx1 locus.

B. Despite being up-regulated in response to treatment, the Irf4 locus has a super-enhancer even under untreated conditions.

C. Lag3, a negative regulator of T cell signaling that is up-regulated with treatment, sees increased enhancer activity and formation of a super-enhancer in treated conditions.

D. Stat5b, a transcription factor important for T cell signaling, shows increased enhancer activity and formation of a super-enhancer in the treated condition.



Allison 2015
Figure S5

Figure 2.S6: ERK signaling translates TCR signal strength into graded gene expression.

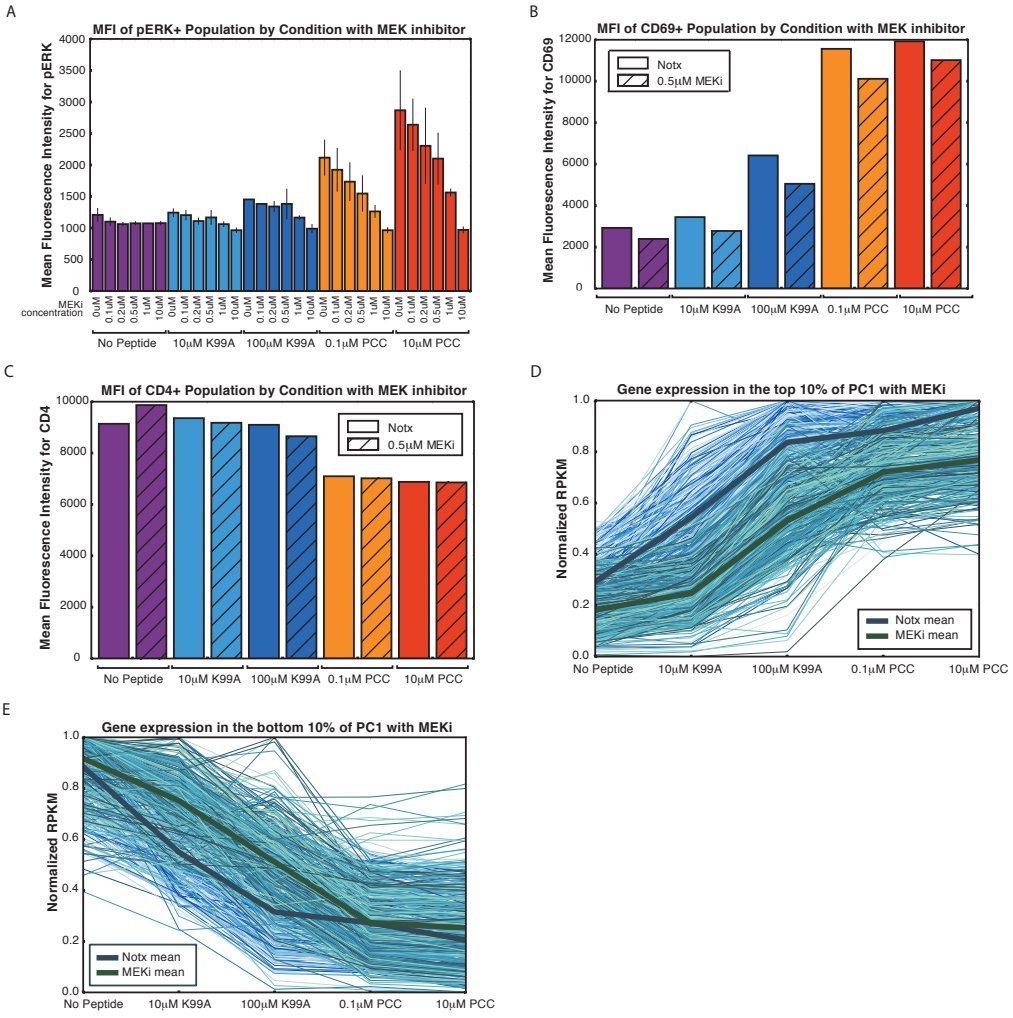
A. Per-cell levels of phospho-ERK as measured by flow cytometry are analog with respect to the dosage of MEK inhibitor treatment. For each peptide condition, increasing the concentration of the MEK inhibitor gradually reduces the geometric MFI of phospho-ERK in the treated cells.

B. MEK inhibitor treatment at 0.5uM (IC50) results in the preferential reduction of activation signature genes. CD69, one of the activation signature genes, reflects this decrease in expression level at the protein level, as measured by extracellular flow cytometry.

C. In contrast to CD69, CD4 does not show a change in expression level upon treatment with the MEK inhibitor.

D. For the genes in the top 10% of PC1, MEK inhibitor treatment (green lines) results in a decrease of expression as compared to the uninhibited samples (blue lines). Each line plots the normalized RPKM of one gene across the five samples, either with the MEK inhibitor (green) or without (blue).

E. In contrast to the top 10%, the genes in the bottom 10% of PC1 are not significantly affected by MEK inhibitor treatment, and, if anything, increase in expression level rather than decrease.



Allison 2015
Figure S6

REFERENCES

1 Heber-Katz, E., Schwartz, R. H., Matis, L. A., Hannum, C., Fairwell, T., Appella, E. & Hansburg, D. Contribution of antigen-presenting cell major histocompatibility complex gene products to the specificity of antigen-induced T cell activation. *The Journal of experimental medicine* **155**, 1086-1099 (1982).

2 Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I. & Davis, M. M. Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins. *Nature* **308**, 153-158 (1984).

3 Jerne, N. K. The somatic generation of immune recognition. *European journal of immunology* **1**, 1-9, doi:10.1002/eji.1830010102 (1971).

4 Hedrick, S. M., Matis, L. A., Hecht, T. T., Samelson, L. E., Longo, D. L., Heber-Katz, E. & Schwartz, R. H. The fine specificity of antigen and Ia determinant recognition by T cell hybridoma clones specific for pigeon cytochrome c. *Cell* **30**, 141-152 (1982).

5 Solinger, A. M., Ultee, M. E., Margoliash, E. & Schwartz, R. H. T-lymphocyte response to cytochrome c. I. Demonstration of a T-cell heteroclitic proliferative response and identification of a topographic antigenic determinant on pigeon cytochrome c whose immune recognition requires two complementing major histocompatibility complex-linked immune response genes. *The Journal of experimental medicine* **150**, 830-848 (1979).

6 Alexander, J., Snoke, K., Ruppert, J., Sidney, J., Wall, M., Southwood, S., Oseroff, C., Arrhenius, T., Gaeta, F. C., Colon, S. M. & et al. Functional consequences of engagement of the T cell receptor by low affinity ligands. *Journal of immunology* **150**, 1-7 (1993).

7 Rogers, P. R. & Croft, M. Peptide dose, affinity, and time of differentiation can contribute to the Th1/Th2 cytokine balance. *Journal of immunology* **163**, 1205-1213 (1999).

8 Rogers, P. R., Grey, H. M. & Croft, M. Modulation of naive CD4 T cell activation with altered peptide ligands: the nature of the peptide and presentation in the context of costimulation are critical for a sustained response. *Journal of immunology* **160**, 3698-3704 (1998).

9 Sloan-Lancaster, J., Evavold, B. D. & Allen, P. M. Induction of T-cell anergy by altered T-cell-receptor ligand on live antigen-presenting cells. *Nature* **363**, 156-159, doi:10.1038/363156a0 (1993).

10 Tubo, N. J., Pagan, A. J., Taylor, J. J., Nelson, R. W., Linehan, J. L., Ertelt, J. M., Huseby, E. S., Way, S. S. & Jenkins, M. K. Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. *Cell* **153**, 785-796, doi:10.1016/j.cell.2013.04.007 (2013).

11 Kersh, G. J., Kersh, E. N., Fremont, D. H. & Allen, P. M. High- and low-potency ligands with similar affinities for the TCR: the importance of kinetics in TCR signaling. *Immunity* **9**, 817-826 (1998).

12 Sloan-Lancaster, J., Shaw, A. S., Rothbard, J. B. & Allen, P. M. Partial T cell signaling: altered phospho-zeta and lack of zap70 recruitment in APL-induced T cell anergy. *Cell* **79**, 913-922 (1994).

13 Kaye, J., Hsu, M. L., Sauron, M. E., Jameson, S. C., Gascoigne, N. R. & Hedrick, S. M. Selective development of CD4+ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature* **341**, 746-749, doi:10.1038/341746a0 (1989).

14 Varma, R. C., T.; Padhan, K.; Muller, J. in *Keystone Symposium on Advances in the Knowledge and Treatment of Autoimmunity* (Whistler, British Columbia, Canada, 2013 April 4 – 9).

15 Coward, J., Germain, R. N. & Altan-Bonnet, G. Perspectives for computer modeling in the study of T cell activation. *Cold Spring Harbor perspectives in biology* **2**, a005538, doi:10.1101/cshperspect.a005538 (2010).

16 Das, J., Ho, M., Zikherman, J., Govern, C., Yang, M., Weiss, A., Chakraborty, A. K. & Roose, J. P. Digital signaling and hysteresis characterize ras activation in lymphoid cells. *Cell* **136**, 337-351, doi:10.1016/j.cell.2008.11.051 (2009).

17 Daniels, M. A., Teixeiro, E., Gill, J., Hausmann, B., Roubaty, D., Holmberg, K., Werlen, G., Hollander, G. A., Gascoigne, N. R. & Palmer, E. Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling. *Nature* **444**, 724-729, doi:10.1038/nature05269 (2006).

18 Prasad, A., Zikherman, J., Das, J., Roose, J. P., Weiss, A. & Chakraborty, A. K. Origin of the sharp boundary that discriminates positive and negative selection of thymocytes. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 528-533, doi:10.1073/pnas.0805981105 (2009).

19 Kingeter, L. M., Paul, S., Maynard, S. K., Cartwright, N. G. & Schaefer, B. C. Cutting edge: TCR ligation triggers digital activation of NF-kappaB. *Journal of immunology* **185**, 4520-4524, doi:10.4049/jimmunol.1001051 (2010).

20 Marangoni, F., Murooka, T. T., Manzo, T., Kim, E. Y., Carrizosa, E., Elpek, N. M. & Mempel, T. R. The transcription factor NFAT exhibits signal memory during serial T cell interactions with antigen-presenting cells. *Immunity* **38**, 237-249, doi:10.1016/j.immuni.2012.09.012 (2013).

21 Podtschaske, M., Benary, U., Zwinger, S., Hofer, T., Radbruch, A. & Baumgrass, R. Digital NFATc2 activation per cell transforms graded T cell receptor activation into an all-or-none IL-2 expression. *PloS one* **2**, e935, doi:10.1371/journal.pone.0000935 (2007).

22 Au-Yeung, B. B., Zikherman, J., Mueller, J. L., Ashouri, J. F., Matloubian, M., Cheng, D. A., Chen, Y., Shokat, K. M. & Weiss, A. A sharp T-cell antigen receptor signaling threshold for T-cell proliferation. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E3679-3688, doi:10.1073/pnas.1413726111 (2014).

23 Huang, J., Brameshuber, M., Zeng, X., Xie, J., Li, Q. J., Chien, Y. H., Valitutti, S. & Davis, M. M. A single peptide-major histocompatibility complex ligand triggers digital cytokine secretion in CD4(+) T cells. *Immunity* **39**, 846-857, doi:10.1016/j.immuni.2013.08.036 (2013).

24 Zikherman, J. & Au-Yeung, B. The role of T cell receptor signaling thresholds in guiding T cell fate decisions. *Current opinion in immunology* **33C**, 43-48, doi:10.1016/j.coi.2015.01.012 (2015).

25 Butler, T. C., Kardar, M. & Chakraborty, A. K. Quorum sensing allows T cells to discriminate between self and nonself. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 11833-11838, doi:10.1073/pnas.1222467110 (2013).

26 Kersh, E. N., Kersh, G. J. & Allen, P. M. Partially phosphorylated T cell receptor zeta molecules can inhibit T cell activation. *The Journal of experimental medicine* **190**, 1627-1636 (1999).

27 Kersh, E. N., Shaw, A. S. & Allen, P. M. Fidelity of T cell activation through multistep T cell receptor zeta phosphorylation. *Science* **281**, 572-575 (1998).

28 Irvine, D. J., Purbhoo, M. A., Krogsgaard, M. & Davis, M. M. Direct observation of ligand recognition by T cells. *Nature* **419**, 845-849, doi:10.1038/nature01076 (2002).

29 Man, K., Miasari, M., Shi, W., Xin, A., Henstridge, D. C., Preston, S., Pellegrini, M., Belz, G. T., Smyth, G. K., Febbraio, M. A., Nutt, S. L. & Kallies, A. The transcription factor IRF4 is essential for TCR affinity-mediated

metabolic programming and clonal expansion of T cells. *Nature immunology* **14**, 1155-1165, doi:10.1038/ni.2710 (2013).

30 Nayar, R., Schutten, E., Bautista, B., Daniels, K., Prince, A. L., Enos, M., Brehm, M. A., Swain, S. L., Welsh, R. M. & Berg, L. J. Graded levels of IRF4 regulate CD8+ T cell differentiation and expansion, but not attrition, in response to acute virus infection. *Journal of immunology* **192**, 5881-5893, doi:10.4049/jimmunol.1303187 (2014).

31 Marchingo, J. M., Kan, A., Sutherland, R. M., Duffy, K. R., Wellard, C. J., Belz, G. T., Lew, A. M., Dowling, M. R., Heinzel, S. & Hodgkin, P. D. T cell signaling. Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. *Science* **346**, 1123-1127, doi:10.1126/science.1260044 (2014).

32 Huang, Y. & Wange, R. L. T cell receptor signaling: beyond complex complexes. *The Journal of biological chemistry* **279**, 28827-28830, doi:10.1074/jbc.R400012200 (2004).

33 Murphy, T. L., Tussiwand, R. & Murphy, K. M. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nature reviews. Immunology* **13**, 499-509, doi:10.1038/nri3470 (2013).

34 Rincon, M. & Flavell, R. A. AP-1 transcriptional activity requires both T-cell receptor-mediated and co-stimulatory signals in primary T lymphocytes. *The EMBO journal* **13**, 4370-4381 (1994).

35 Murphy, L. O. & Blenis, J. MAPK signal specificity: the right place at the right time. *Trends in biochemical sciences* **31**, 268-275, doi:10.1016/j.tibs.2006.03.009 (2006).

36 Schade, A. E. & Levine, A. D. Cutting edge: extracellular signal-regulated kinases 1/2 function as integrators of TCR signal strength. *Journal of immunology* **172**, 5828-5832 (2004).

37 Feinerman, O., Germain, R. N. & Altan-Bonnet, G. Quantitative challenges in understanding ligand discrimination by alphabeta T cells. *Molecular immunology* **45**, 619-631, doi:10.1016/j.molimm.2007.03.028 (2008).

38 Wei, G., Wei, L., Zhu, J., Zang, C., Hu-Li, J., Yao, Z., Cui, K., Kanno, Y., Roh, T. Y., Watford, W. T., Schones, D. E., Peng, W., Sun, H. W., Paul, W. E., O'Shea, J. J. & Zhao, K. Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells. *Immunity* **30**, 155-167, doi:10.1016/j.immuni.2008.12.009 (2009).

39 Tan, C., Wei, L., Vistica, B. P., Shi, G., Wawrousek, E. F. & Gery, I. Phenotypes of Th lineages generated by the commonly used activation with anti-CD3/CD28 antibodies differ from those generated by the physiological activation with the specific antigen. *Cellular & molecular immunology* **11**, 305-313, doi:10.1038/cmi.2014.8 (2014).

40 Wakamatsu, E., Mathis, D. & Benoist, C. Convergent and divergent effects of costimulatory molecules in conventional and regulatory CD4+ T cells. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 1023-1028, doi:10.1073/pnas.1220688110 (2013).

41 Chikuma, S., Suita, N., Okazaki, I. M., Shibayama, S. & Honjo, T. TRIM28 prevents autoinflammatory T cell development in vivo. *Nature immunology* **13**, 596-603, doi:10.1038/ni.2293 (2012).

42 Doering, T. A., Crawford, A., Angelosanto, J. M., Paley, M. A., Ziegler, C. G. & Wherry, E. J. Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. *Immunity* **37**, 1130-1144, doi:10.1016/j.immuni.2012.08.021 (2012).

43 Mostafavi, S., Ortiz-Lopez, A., Bogue, M. A., Hattori, K., Pop, C., Koller, D., Mathis, D., Benoist, C. & Immunological Genome, C. Variation and genetic control of gene expression in primary immunocytes across inbred mouse strains. *Journal of immunology* **193**, 4485-4496, doi:10.4049/jimmunol.1401280 (2014).

44 Bakir, H. Y., Tomiyama-Miyaji, C., Watanabe, H., Nagura, T., Kawamura, T., Sekikawa, H. & Abo, T. Reasons why DBA/2 mice are resistant to malarial infection: expansion of CD3int B220+ gammadelta T cells with double-negative CD4- CD8- phenotype in the liver. *Immunology* **117**, 127-135, doi:10.1111/j.1365-2567.2005.02273.x (2006).

45 He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., Mieczkowski, P., Lieb, J. D., Zhao, K., Brown, M. & Liu, X. S. Nucleosome dynamics define transcriptional enhancers. *Nature genetics* **42**, 343-347, doi:10.1038/ng.545 (2010).

46 Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C. & Glass, C. K. Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Molecular cell* **51**, 310-325, doi:10.1016/j.molcel.2013.07.010 (2013).

47 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple combinations of lineage-

determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

48 Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., Blecher-Gonen, R., Bornstein, C., Amann-Zalcenstein, D., Weiner, A., Friedrich, D., Meldrim, J., Ram, O., Cheng, C., Gnirke, A., Fisher, S., Friedman, N., Wong, B., Bernstein, B. E., Nusbaum, C., Hacohen, N., Regev, A. & Amit, I. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Molecular cell* **47**, 810-822, doi:10.1016/j.molcel.2012.07.030 (2012).

49 Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D. & Glass, C. K. Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487-492, doi:10.1038/nature12615 (2013).

50 Mullen, A. C., Orlando, D. A., Newman, J. J., Loven, J., Kumar, R. M., Bilodeau, S., Reddy, J., Guenther, M. G., DeKoter, R. P. & Young, R. A. Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell* **147**, 565-576, doi:10.1016/j.cell.2011.08.050 (2011).

51 Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994-1004, doi:10.1016/j.cell.2012.09.045 (2012).

52 Trompouki, E., Bowman, T. V., Lawton, L. N., Fan, Z. P., Wu, D. C., DiBiase, A., Martin, C. S., Cech, J. N., Sessa, A. K., Leblanc, J. L., Li, P., Durand, E. M., Mosimann, C., Heffner, G. C., Daley, G. Q., Paulson, R. F., Young, R. A. & Zon, L. I. Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* **147**, 577-589, doi:10.1016/j.cell.2011.09.044 (2011).

53 Anderson, M. K., Hernandez-Hoyos, G., Diamond, R. A. & Rothenberg, E. V. Precise developmental regulation of Ets family transcription factors during specification and commitment to the T cell lineage. *Development* **126**, 3131-3148 (1999).

54 Zhang, J. A., Mortazavi, A., Williams, B. A., Wold, B. J. & Rothenberg, E. V. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* **149**, 467-482, doi:10.1016/j.cell.2012.01.056 (2012).

55 Wong, W. F., Kohu, K., Chiba, T., Sato, T. & Satake, M. Interplay of transcription factors in T-cell differentiation and function: the role of Runx. *Immunology* **132**, 157-164, doi:10.1111/j.1365-2567.2010.03381.x (2011).

56 Ozato, K., Tailor, P. & Kubota, T. The interferon regulatory factor family in host defense: mechanism of action. *The Journal of biological chemistry* **282**, 20065-20069, doi:10.1074/jbc.R700003200 (2007).

57 Glasmacher, E., Agrawal, S., Chang, A. B., Murphy, T. L., Zeng, W., Vander Lugt, B., Khan, A. A., Ciofani, M., Spooner, C. J., Rutz, S., Hackney, J., Nurieva, R., Escalante, C. R., Ouyang, W., Littman, D. R., Murphy, K. M. & Singh, H. A genomic regulatory element that directs assembly and function of immune-specific AP-1-IRF complexes. *Science* **338**, 975-980, doi:10.1126/science.1228309 (2012).

58 Li, P., Spolski, R., Liao, W., Wang, L., Murphy, T. L., Murphy, K. M. & Leonard, W. J. BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature* **490**, 543-546, doi:10.1038/nature11530 (2012).

59 Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T. & Wysocka, J. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell stem cell* **14**, 838-853, doi:10.1016/j.stem.2014.04.003 (2014).

60 Mikkelsen, T. S., Xu, Z., Zhang, X., Wang, L., Gimble, J. M., Lander, E. S. & Rosen, E. D. Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156-169, doi:10.1016/j.cell.2010.09.006 (2010).

61 Mishra, B. P., Zaffuto, K. M., Artinger, E. L., Org, T., Mikkola, H. K., Cheng, C., Djabali, M. & Ernst, P. The histone methyltransferase activity of MLL1 is dispensable for hematopoiesis and leukemogenesis. *Cell reports* **7**, 1239-1247, doi:10.1016/j.celrep.2014.04.015 (2014).

62 Vahedi, G., Takahashi, H., Nakayamada, S., Sun, H. W., Sartorelli, V., Kanno, Y. & O'Shea, J. J. STATs shape the active enhancer landscape of T cell populations. *Cell* **151**, 981-993, doi:10.1016/j.cell.2012.09.044 (2012).

63 Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A. & Jaenisch, R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).

64 Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A., Hoke, H. A. & Young, R. A. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

65 Loven, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I. & Young, R. A. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:10.1016/j.cell.2013.03.036 (2013).

66 Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I. & Young, R. A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

67 Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S. & Natoli, G. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157-171, doi:10.1016/j.cell.2012.12.018 (2013).

68 Zhou, L., Chong, M. M. & Littman, D. R. Plasticity of CD4+ T cell lineage differentiation. *Immunity* **30**, 646-655, doi:10.1016/j.immuni.2009.05.001 (2009).

69 Fu, G., Casas, J., Rigaud, S., Rybakin, V., Lambolez, F., Brzostek, J., Hoerter, J. A., Paster, W., Acuto, O., Cheroutre, H., Sauer, K. & Gascoigne, N. R. Themis sets the signal threshold for positive and negative selection in T-cell development. *Nature* **504**, 441-445, doi:10.1038/nature12718 (2013).

70 Warnecke, N., Poltorak, M., Kowtharapu, B. S., Arndt, B., Stone, J. C., Schraven, B. & Simeoni, L. TCR-mediated Erk activation does not depend on Sos and Grb2 in peripheral human T cells. *EMBO reports* **13**, 386-391, doi:10.1038/embor.2012.17 (2012).

71 Zhao, Y. & Adjei, A. A. The clinical development of MEK inhibitors. *Nature reviews. Clinical oncology* **11**, 385-400, doi:10.1038/nrclinonc.2014.83 (2014).

72 Samatar, A. A. & Poulikakos, P. I. Targeting RAS-ERK signalling in cancer: promises and challenges. *Nature reviews. Drug discovery* **13**, 928-942, doi:10.1038/nrd4281 (2014).

73 Vella, L. J., Pasam, A., Dimopoulos, N., Andrews, M., Knights, A., Puaux, A. L., Louahed, J., Chen, W., Woods, K. & Cebon, J. S. MEK inhibition, alone or in combination with BRAF inhibition, affects multiple functions of isolated

normal human lymphocytes and dendritic cells. *Cancer immunology research* **2**, 351-360, doi:10.1158/2326-6066.CIR-13-0181 (2014).

74 Gilfillan, G. D., Hughes, T., Sheng, Y., Hjorthaug, H. S., Straub, T., Gervin, K., Harris, J. R., Undlien, D. E. & Lyle, R. Limitations and possibilities of low cell number ChIP-seq. *BMC genomics* **13**, 645, doi:10.1186/1471-2164-13-645 (2012).

75 Gosselin, D., Link, V. M., Romanoski, C. E., Fonseca, G. J., Eichenfield, D. Z., Spann, N. J., Stender, J. D., Chun, H. B., Garner, H., Geissmann, F. & Glass, C. K. Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. *Cell* **159**, 1327-1340, doi:10.1016/j.cell.2014.11.023 (2014).

76 Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P. & Brutnell, T. P. A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. *PloS one* **6**, e26426, doi:10.1371/journal.pone.0026426 (2011).

77 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

78 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

79 Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J. & Snyder, M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813-1831, doi:10.1101/gr.136184.111 (2012).

80 karmel. homer-idr. doi:10.5281/zenodo.11619 (2014).

81 Allison, K. A., Kaikkonen, M. U., Gaasterland, T. & Glass, C. K. Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic acids research* **42**, 2433-2447, doi:10.1093/nar/gkt1237 (2014).

82 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

83 Perez, F. & Granger, B. E. IPython: a System for Interactive Scientific Computing. *Computing in Science and Engineering* **9**, 21-29 (2007).

84 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).