Analysis of, and software development for, ChIP-Seq and RNA-Seq data

by

Yu-Hsuan Lin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2012

Doctoral Committee:

Professor James Douglas Engel, Co-Chair
Assistant Professor Maureen A. Sartor, Co-Chair
Professor Kerby A Shedden
Assistant Professor Ivan Patrick Maillard
Research Assistant Professor Fan Meng

To my grand parents and parents

Acknowledgements

I would like to thank Dr. Sakie Hosoya-Ohmura for her work in writing our manuscript on the Gata3 enhancer, and Dr. Henriette O'Geen for the initial version of the manuscript relating to chapter 3. I would also like to thank Dr. Lihong Shi for developing the *ex vivo* CD34 culture system in our laboratory and for performing the RNA-Seq experiments that provide the basis for the bioinformatics analyses presented in Chapter 4. I would also thank Yanxiao Zhang for revising the ChIP-Seq analysis pipeline to make it more efficient and to incorporate additional functionality. Finally, I greatly appreciate the advice of Dr. Engel and Dr. Sartor for their great contribution to this thesis.

Table of Contents

List of Figures

Figure

List of Tables

Table

Chapter I

Introduction

A. OVERVIEW

With the advent of new sequencing technologies developed over the past two decades, researchers have completely sequenced the genomes of a number of organisms using automated Sanger sequencing (Metzker, 2010). This "first generation" technology led to the discovery of high similarities of genomic sequences among organisms, and led scientists to question what causes the differences among organisms. By investigating the conserved sequences among species, many of these highly conserved sequence segments turned out to be critical for regulating gene expression. These regulatory elements recruit multiple transcription factors to work in concert in order to direct target gene expression. However, using traditional reporter genes to define the identity, location and activities of these regulatory modules is expensive and extremely time consuming.

Continuously evolving sequencing technology has brought us to next generation sequencing (NGS). NGS can inexpensively generate hundreds of millions of short sequence reads with a single instrument run (Metzker, 2010). Since its introduction, NGS has been successfully applied to interrogate various aspects of cellular status and biological processes on a genome-wide scale, including transcription factor

binding profiles (Johnson, et al., 2007; Robertson, et al., 2007), histone modification status (Barski, et al., 2007; Mikkelsen, et al., 2007), and gene expression level and composition (Cloonan, et al., 2008; Lister, et al., 2008; Mortazavi, et al., 2008; Nagalakshmi, et al., 2008; Wilhelm, et al., 2008). Algorithmic details of the many different ChIP-Seq analysis software tools differ, but all report a score (usually with statistical significance) of the binding strength for a given protein-DNA binding event. However, these analyses do not address the question of whether the binding is functional or not. Moreover, while analysis of replicate variation has been incorporated into RNA-Seq experiments, no similar paradigm has been optimized for ChIP-Seq experiments with biological replicates. Assessment of biological variation is particularly relevant to ChIP-Seq experiments assessing histone modifications, nucleosome placement, or other epigenomic marks that may vary significantly among tissues and individuals.

In this thesis, I have analyzed RNA-Seq and ChIP-Seq data related to hematopoiesis. Hematopoiesis serves as an ideal model to study lineage commitment, specification and development as orchestrated by regulatory transcription factor proteins. Blood cells are replenished daily throughout the lifespan of humans to elicit their basic functions such as oxygen transport and immune protection. All classes of blood cells are derived from a rare pool of hematopoietic stem cells. These hematopoietic stem cells are capable of both self-renewal and differentiation into all mature hematopoietic cell lineages, including T cells and erythrocytes. Transcription factors have long been known to play important roles in these differentiation processes. For

example, GATA-3 is recurrently required during T cell development, and orphan nuclear receptors NR2C2 plus NR2C1 have been shown to play a pivotal role in erythropoiesis by binding to the γ-globin gene promoter and repressing its expression. Ectopic expression of, and malfunction by, key hematopoietic regulatory proteins contribute to multiple blood diseases. For example, improper *Gata3* expression can lead to T cell lymphoma (Nawijn, et al., 2001; van Hamburg, et al., 2008), and sickle-shaped red blood cells are caused by a single amino acid substitution in the β-globin gene. However, the *cis* element that directs *Gata3* transcription during T cell development *in vivo* is still not fully characterized. Similarly, cellular target sequences of NR2C1/NR2C2 are largely unknown but one can imagine how such knowledge would be desirable for developing anti-NR2C2 therapeutics to de-repress fetal γ-globin expression to alleviate the painful symptoms and pathophysiology of sickle cell disease. This thesis work was intended to utilize data generated from high throughput sequencing technologies to help to elucidate the mechanisms of transcriptional regulation in two different hematopoietic lineages and to develop a robust software pipeline for analyzing ChIP-Seq data.

B. THESIS CHAPTER SUMMARY

i. Chapter II summary

The second chapter of my dissertation involves the description of methods and experiments that were designed to predict the genomic position of a *Gata3* enhancer element (a contiguous stretch of DNA nucleotides of undetermined length, but

usually lying between 300 and 1,000 base pairs) that specified its expression in the T cell lineage by integrating DNA sequencing data with phylogenetic conservation scores. *Gata3* is a transcription factor known for its role in cooperating with other transcription factors in an orchestrated way during the development of certain immune cells (T lymphocytes and natural killer cells). Precise control of *Gata3* expression is critical for normal T cell development, as ectopic *Gata3* expression has been shown to be carcinogenic in mice. However, the factors that regulate *Gata3* transcription are largely unknown. In addition, GATA-3 has been shown to play pivotal developmental roles in many tissues and its expression in those distinct cell types is usually controlled by a tissue-specific enhancer, often located quite far from the structural gene. By mapping chromatin sites that are hypersensitive to DNAse I cleavage (identified using multiple high-throughput technologies), together with so-called regulatory sequence potential scores, I helped to predict a syntenic sequence that is conserved in mouse and human that serves as a regulatory element for *Gata3* in T lymphocytes. The predicted DNA sequence lies approximately 280 kilobase pairs 3' to the *Gata3* gene and was confirmed experimentally by Dr. Sakie Hosoya-Ohmura in Dr. Engel's laboratory: the evolutionarily conserved sequence is able to direct *Gata3* expression at multiple stages of T cell development, from immature early T lineage progenitors to mature peripheral T cells. This regulatory function was demonstrated using flow cytometry to show that this *cis* element is able to induce the transcription of a reporter gene *in vivo* (in transgenic mice) at various stages exclusively during T cell development. I performed additional imaging studies to confirm the T cell specificity of this regulatory element for *Gata3* by

analyzing these same transgenic mice visualizing the organs and by whole animal *in vivo* imaging.

ii. Chapter III summary

The third chaper of my research is focused on identifying NR2C2 (human testicular receptor 4, or TR4) genome-wide DNA bound target sequences in four of the human ENCODE (ENCyclopedia Of DNA Elements) consortium cell lines. TR4 belongs to the nuclear receptor superfamily and is referred to as an orphan since it has no currently identified ligand. TR4 was initially identified in testis, but was later shown to be expressed almost ubiquitously, including in erythroid cells. There are two facets to the roles TR4 plays in directing target gene expression. TR4 was demonstrated to be able to regulate target gene expression in liver carcinoma HepG2 cells, while it was also reported to be able to form heterodimers with another closely related family member NR2C1 (or TR2) to function as a repressor of the human fetal γ-globin genes, the genes responsible for mediating oxygen transfer in fetal erythroid cells. The TR2/TR4 heterodimer forms a complex, named direct repeat erythroid-definitive (DRED), that binds to direct repeat 1 (DR1) sequence motifs in the human fetal γ-globin gene promoters. Furthermore, inhibition of TR2/TR4 in mice as well as in human erythroid cells has been shown to lead to fetal γ-globin induction, a condition known to alleviate sickle cell disease (SCD) pathophysiology. Therefore, TR4 serves as an appealing target for the development of SCD therapeutics. Since the identification of TR2/TR4 as a human fetal γ-globin gene repressor, Dr. Engel's laboratory has focused on developing pharmacological

inhibitors of TR2/TR4. Genome-scale identification of target genes regulated by TR4 will be vital to predict and minimize any potential side effects of prescribing anti-TR4 therapeutics to human patients.

High-throughput DNA sequencing technologies coupled with bioinformatics analyses were employed for the first time here to identify all TR4 target genes in four established human cell lines. By comparing the genome-wide binding sites to reference gene locations, I showed that TR4 preferentially binds at gene proximal promoters and within the first exon or intron of the various target genes. These binding sites were enriched for the canonical nuclear receptor binding sequence motif (so called DR, or direct repeat, elements) as well as the ETS transcription factor family binding motif. The results also suggested that TR4 preferentially binds to genes playing crucial roles in RNA transcription and processing. Subsequent bioinformatics analysis results revealed TR4 binding at a subset of targets may be facilitated through the recruitment of ELK4, an Ets transcription factor family member.

iii. Chapter IV summary

The fourth chapter of this thesis work concentrates on characterizing the changing transcriptome dynamics during human CD34+ hematopoietic progenitor cell differentiation. Human hematopoietic progenitor cells can be characterized by cell surface expression of CD34 and can be obtained from various sources, such as bone marrow, fetal liver, umbilical cord and peripheral blood. Human hematopoietic stem cells are entirely within the CD34+ compartment, and therefore this marker is used

clinically for generating repopulating cells for bone marrow transplants. An *ex vivo* CD34+ stem cell differentiation culture system has been described that can produce fully mature human red blood cells when the culture is supplied with appropriate cytokine combinations. This culture system then serves as an ideal model to study gene expression profile changes during normal erythroid development and maturation. Previous microarray studies using differentiating CD34+ cells have provided lists of candidate erythroid regulators for further investigation. However, due to the hybridization-based nature of microarrays, they have limited dynamic range. Here, a bioinformatics analysis was performed to investigate the erythroid cell transcriptome during differentiation using next generation mRNA sequencing technology. By confirming the data quality, I discovered that a group of transcripts, including recognized erythroid factors, are consistently expressed at a high level during differentiation. Functional analysis revealed that this gene cluster is enriched with genes that act in the ribosome pathway and on the translation machinery. This gene group is comprised of a number of known erythroid transcription factors along with other transcription factors that play currently unknown roles in erythropoiesis. Many of the erythroid genes, which were shown to be differentially expressed during CD34 differentiation using array platforms, were also found to be differentially expressed in our experiments, thereby validating our approach. Going further, we also identified potential novel isoforms of known erythroid transcriptional regulators, such as LSD1 and SOX6. Additional lists of novel intergenic and intronic transcripts were also identified and serve as candidates for

further functional analysis. This study will provide an invaluable comprehensive repository of the differentiating erythroid transcriptome.

iv. Chapter V summary

The fifth chapter of my thesis is devoted to the development of a computational pipeline, called PePr (a Peak Prioritization pipeline), to prioritize the potential protein-DNA interaction sites of a transcription factor or histone modification according to their location relative to gene structures. ChIP-Seq has been widely employed to identify *in vivo* protein-DNA interactions or histone modifications on a genome-wide scale. A growing number of software applications have been developed and shown to successfully identify transcription factor binding or histone posttranslational modifications from ChIP-Seq experiments. However, while peak lists reported by different programs tend to agree on strong binding signals, they can vary significantly for presumptive weaker binding sties. This may be partially due to the fact that applications use various background models and statistical distributions; it is also likely compounded by the fact that most methods do not model variation among replicates/samples when applicable. In addition, little effort has been devoted to incorporating external annotation into the peak calling process, though it has the potential to be beneficial. Importantly, none exploit the use of gene structure location relative to peaks.

To address these issues and to more fully exploit the power of replicates and external annotation, we developed PePr, a ChIP-Seq Peak Prioritization pipeline that utilizes a sliding window approach with a negative binomial model to accommodate

variation among biological replicates. While the negative binomial model has been used extensively in the analysis of RNA-seq data, relatively few ChIP-Seq pipelines have exploited this distribution (Feng, et al., 2008; Ji, et al., 2008). We further extended our pipeline by incorporating information regarding the binding profile relative to gene structure to prioritize the ChIP-Seq peaks. PePr was developed to be flexible in that it supports various input file formats as well as to provide options for users to control the behavior of the peak calling process. We compared the performance of PePr with two other peak finding methods, ERANGE and MACS, and tested how well PePr prioritizes binding sites that correspond to functional regulation of gene expression, as determined by RNA-Seq or microarrays. Such a tool will benefit investigators by helping them to identify the most important functional sites, i.e. those that are potentially the most interesting for detailed follow-up, and binding preference relative to gene structure. It will also be important in studies assessing significant differences in histone modification or DNA methylation when comparing groups of individuals.

C. THESIS CONTRIBUTION

In chapter 2, I helped to identify a regulatory element for *Gata3* and confirmed the T cell specificity by *in vivo* imaging. In chapter 3, I analyzed TR4 ChIP-Seq data in the four ENCODE consortium cell lines. Results suggested that TR4 preferentially binds to gene proximal regions in a DR1 sequence motif, and predicted a role for ETS factors in TR4 action. Functional analysis suggested that TR4 target genes are involved in general biological processes. In chapter 4, I analyzed RNA-Seq data

accumulated from differentiating human hematopoietic progenitor cells. Analysis suggested a list of potential novel erythroid regulatory factors, and revealed potential novel isoforms of known erythroid regulatory proteins. In chapter 5, I developed a ChIP-Seq peak prioritization software pipeline, which can model variation among biological replicates and prioritize peaks based on the binding relative to gene structure by incorporating external annotation.

Chapter II

An NK and T Cell Enhancer Lies 280 Kilobase Pairs 3' to the *Gata3* Structural Gene

A. INTRODUCTION

Maturation of T lineage lymphocytes is one of the most clearly defined pathways in all of developmental biology. Immature hematopoietic cells from the bone marrow migrate through the bloodstream to initially populate the thymus. The earliest detectable thymic progenitor (early T lineage progenitors [ETP]) cells differentiate uniquely in the thymic microenvironment through several early stages in which neither the CD4 nor CD8 coreceptors are expressed (double-negative [DN] cells, stages 2 to 4) and thence into cells which express both CD4 and CD8 (double-positive [DP] cells) and finally generate either CD4 single-positive (CD4 SP; $CD4^+$ $CD8^-$) cells or CD8 single-positive (CD8 SP; $CD4^-$ $CD8^+$) cells. $CD4^+$ thymocytes have the potential to differentiate into helper (Th) or regulatory T cells, while CD8 SP cells are programmed to fulfill cytotoxic functions. These mature single-positive thymocytes exit the thymus to execute their defined effector functions after activation in the periphery.

The zinc-finger transcription factor GATA-3 (Ko, et al., 1991; Yamamoto, et al., 1990) is expressed throughout T cell development (Hosoya, et al., 2010), peaking in

abundance in CD4 SP and Th2 cells (David-Fung, et al., 2006; Hendriks, et al., 1999; Hernandez-Hoyos, et al., 2003; Sambandam, et al., 2005; Tydell, et al., 2007; Zhang, et al., 1997; Zheng and Flavell, 1997). GATA-3 function has been shown to be vital for the generation of ETP (Hosoya, et al., 2009), double-negative (DN) 4 stage, and CD4 SP cells (Pai, et al., 2003) and for the differentiation and function of Th2 cells (Pai, et al., 2004; Zhu, et al., 2004). While its expression is critical for normal T cell development, enforced ectopic expression of GATA-3 can have catastrophic consequences (Anderson, et al., 2002; Chen and Zhang, 2001; Nawijn, et al., 2001; Nawijn, et al., 2001; Taghon, et al., 2001; Taghon, et al., 2007), such as causing T cell lymphoma in transgenic mice (Nawijn, et al., 2001; van Hamburg, et al., 2008) and converting DP cells into a premalignant state (van Hamburg, et al., 2008). Additionally, GATA-3 plays a role in the aberrant survival of T lymphoma cells in E2A mutant mice (Xu and Kee, 2007). These results, taken together, suggest that both the timing and abundance of GATA-3 must be exquisitely regulated for proper development of the T cell lineage.

Ours and many other laboratories have sought to define how this key T lymphocyte regulatory protein is itself so precisely modulated during T cell development, but prior studies have failed to conclusively identify *Gata3* transcriptional *cis* elements that are capable of conferring appropriate regulatory properties to this gene *in vivo*. In exploring the transcriptional networks that lead to proper T cell differentiation, potential *trans* upstream regulators of *Gata3* have been proposed (Amsen, et al., 2007; Fang, et al., 2007; Maurice, et al., 2007; Yang, et al., 2009; Yu, et al., 2009).

However, while *Gata3* proximal promoter sequences are capable of activating its transcription in transfection experiments (George, et al., 1994), we report here that neither of the *Gata3* promoters (Asnagli, et al., 2002) is capable of conferring such activity *in vivo*. Since previous studies have not demonstrated a functional requirement for the direct association of any of the proposed epistatic *Gata3* regulators (Notch/CSL, c-Myb, T cell factor 1 [TCF-1], and Dec2) with their cognate binding sites in the *Gata3* promoters through site mutagenesis followed by *in vivo* activity tests, the experiments described here clearly demonstrate that those sequences are insufficient for thymic T cell-specific expression.

We and many others have shown that GATA-3 plays critical roles in quite diverse developmental events (Asselin-Labat, et al., 2007; Grigorieva, et al., 2010; Kaufman, et al., 2003; Kouros-Mehr, et al., 2006; Kurek, et al., 2007; Lim, et al., 2000; Moriguchi, et al., 2006; Tsarovina, et al., 2010) and that *Gata3* expression in those tissues and organs is usually dictated by individual tissue-specific enhancers (George, et al., 1994; Hasegawa, et al., 2007; Lakshmanan, et al., 1998; Lakshmanan, et al., 1999; Lieuw, et al., 1997). Here, we report the identification of a *cis* element that regulates *Gata3* expression during multiple stages of T cell development and that is located far 3′ to the *Gata3* gene. This element induces the transcription of a reporter gene *in vivo* in thymic ETP, natural killer (NK), γδT, CD4 SP, and peripheral $CD4^+$ stages and thus its expression pattern reflects that of endogenous *Gata3*. While additional *cis* elements may be required to fully support appropriate expression of *Gata3* during T cell development, this distant element appears to confer activity at

several of the major developmental transitions that are required for *Gata3* T cell-specific transcriptional control *in vivo*.

B. MATERIALS AND METHODS

i. Mice

Transgenic mice were generated using standard techniques in the University of Michigan Transgenic Animal Model Core or using our own instruments. Bacterial artificial chromosome (BAC) or plasmid DNAs were microinjected into (C57BL/6J × SJL)F2 fertilized oocytes. Transgenic lines were established by crossing onto a CD1 background. GATA-3–enhanced green fluorescent protein (eGFP) fusion cDNA knock-in mice ($Gata3^{g/+}$) were generated previously (Hosoya, et al., 2009) (T. Moriguchi et al., unpublished data). $Gata3^{z/+}$ mice (Hendriks, et al., 1999) and $Tg^{B125.LacZ}$ mice (the genome sequence-revised endpoints are −451 to +211 kb, with respect to the translation start site) were described previously (Hasegawa, et al., 2007; Lakshmanan, et al., 1998; Lakshmanan, et al., 1999). All animal experiments were approved by the University Committee on Use and Care of Animals of the University of Michigan and were performed according to their guidelines (IACUC approval no. 8611).

ii. BAC recombineering

The RPCI-23 C57BL/6J mouse BACs used in this study are described by the following endpoints (±0.5 kbp) relative to the *Gata3* translational start site: 43G17,

+49/+294; 193E6, +128/+330; 263A8, +269/+473. Modification of BAC clones was performed as previously described (Khandekar, et al., 2004; Lee, et al., 2001).

iii. eGFP reporter BAC recombinants

The pGATA-3–eGFP fusion cDNA plasmid containing the genomic *1b* promoter (unpublished data) was digested with NcoI and self-ligated to remove only the *Gata3* cDNA; this is referred to as the p*G3-1b*.eGFP plasmid. p*G3-1b*.eGFP was digested with EcoRI and NotI to excise the fragment containing the *1b* promoter (sequences corresponding to bp −1314 to +1, with respect to the translational initiation site) and eGFP cassette. A p*G3* BAC-targeting vector (unpublished data), which contains an Frt-Neo-Frt selection cassette, the *1b*-LacZ reporter, and two homology arms that were identical to two segments of the *SacBII* gene in the pBACe3.6 vector backbone, was digested with EcoRI and NotI to remove the *1b*-LacZ cassette. The EcoRI-NotI fragment of p*G3-1b*.eGFP and the EcoRI-NotI fragment of the p*G3* BAC-targeting vector were ligated to generate the *1b*.eGFP BAC-targeting plasmid. The resultant plasmid was digested, purified, and used for BAC homologous recombination. The recombinant BAC clones were verified by restriction enzyme digest pattern and Southern blotting (data not shown).

iv. Deletion of *TCE-7.1* from BAC 43G17

Two homology arms that were located immediately adjacent to either end of *TCE-7.1* were amplified by PCR and then subcloned into the pFrtNeo plasmid, which contains the Frt-Neo-Frt cassette (Khandekar, et al., 2004). BAC homologous

recombination was performed using the purified targeting fragment as described previously (Khandekar, et al., 2004). The resultant recombinant BAC (*43G17*Δ*7.1*) was verified by restriction enzyme digestion and Southern blotting (data not shown).

v. Construction of an eGFP reporter plasmid containing *TCE-7.1*

To prepare the *1b*.eGFP reporter plasmid, the Neo cassette was removed from the *1b*.eGFP BAC-targeting plasmid. BAC 43G17 DNA was digested with SalI and KpnI. The 7.1-kbp SalI-KpnI fragment (*TCE-7.1* fragment) was gel purified and cloned into the SalI/KpnI sites of pGEM-4Z. The *TCE-7.1* fragment was verified by restriction digestion pattern and sequencing. To generate the *7.1-1b*.eGFP reporter plasmid, the SalI-KpnI *TCE-7.1* fragment from the pGEM-4Z 7.1-kbp plasmid, the PacI-SalI fragment of the *1b*.eGFP plasmid, and the PacI-KpnI fragment of pNEB193 were ligated together. The resultant 7.1-*1b*.eGFP plasmid was digested with PmeI and KpnI and used for microinjection. To generate *Gata3-1b* promoter-only transgenic mice, the *1b*.eGFP plasmid was digested with PacI and PmeI and used for microinjection.

vi. Flow cytometry

Single-cell suspensions of thymocytes, bone marrow, splenocytes, or peripheral blood were incubated with Fc Block (BD Biosciences). Splenocytes and peripheral blood were hemolyzed using NH_4Cl before incubation with Fc Block. The following antibodies (either from eBioscience or from BD Biosciences) were then applied:

phycoerythrin-cyanine 7-conjugated (PE-Cy7) anti-CD4 (RM4-5), PE–anti-CD4 (H129.19), allophycocyanin (APC)–anti-CD8a (53-6.7), biotin–anti-CD8a (53-6.7), PE–anti-CD44 (IM7), peridinin chlorophyll protein-Cy5.5 (PerCP-Cy5.5)–anti-CD62L (MEL-14), PE–Cy7–anti- CD25 (PC61.5), APC–anti-c-Kit (2B8), PE–anti-CD3e (145-2C11), PE–Cy7–anti-CD3e (145-2C11), biotin–anti-CD3e (145-2C11), APC–anti-γδTCR (T cell receptor) (GL3), biotin–anti-γδTCR (GL3), PE–Cy7–anti-CD19 (1D3), biotin–anti-CD19 (1D3), APC–anti-CD49b (DX5), APC–anti-B220 (RA3-6B2), biotin–anti-B220 (RA3-6B2), APC–eFluor780–anti-Mac1 (M1/70), biotin–anti-Mac1 (M1/70), eFluor450–anti-Gr1 (RB6-8C5), biotin–anti-Gr1 (RB6-8C5), APC–anti-TER119 (TER-119), biotin–anti-TER119 (TER-119), PE–anti-CD71 (R17217), PerCP–Cy5.5–anti-TCRβ (H57-597), biotin–anti-TCRβ (H57-597), PE–anti-CD69 (H1.2F3), biotin–anti-NK1.1 (PK136), biotin–anti-CD11c (N418), PE–Cy7–anti-gamma interferon (anti-IFN-γ) (XMG1.2), APC–anti-interleukin 4 (anti-IL-4) (11B11), PE–Cy7–anti-Sca1 (D7), streptavidin eFluor450. Immature T cells were analyzed as previously described (Hosoya, et al., 2009). The FluoReporter LacZ flow cytometry kit (Molecular Probes) was used to analyze LacZ expression according to the manufacturer's protocol. Cells were analyzed on either FACSCanto II (BD Biosciences) or FACSCalibur (BD Biosciences). Dead cells were excluded by DAPI (4′,6-diamidino-2-phenylindole) or propidium iodide. Acquired data were analyzed using either Weasel (WEHI Biotechnology Centre), FlowJo (Tree Star, Inc.), FACSDiva, or Cell Quest (BD Biosciences) software. The mean fluorescence intensity (MFI) of eGFP was normalized using the LinearFlow green flow cytometry intensity calibration kit (Molecular Probes) in most experiments. These calibration beads

were excited by 488 nm, and fluorescence measurements were performed in the same manner as eGFP measurements in every experiment. A standard curve was generated based on acquired calibration bead data, and eGFP MFI of each sample was normalized using the standard curve.

vii. *In vitro* CD4$^+$ T cell differentiation assay

CD4$^+$ CD25$^-$ splenocytes were purified using the Dynal mouse CD4-negative isolation kit (Invitrogen) in combination with affinity-purified anti-mouse CD25 antibody (PC61.5; eBioscience) and cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum (FBS), 4 mM l-glutamine, 100 U/ml penicillin, 100 μg/ml streptomycin, 50 μM 2-mercaptoethanol, 25 mM HEPES, 0.1 mM nonessential amino acids, and 1 mM sodium pyruvate. Isolated CD4$^+$ cells were stimulated with plate-bound anti-CD3e antibody (10 μg/ml; 145-2C11; BD Biosciences) and anti-CD28 antibody (10 μg/ml; 37.51; BD Biosciences). For nonpolarizing conditions, 10 ng/ml recombinant human IL-2 (PeproTech) was added. For Th1-polarizing condition, 10 μg/ml anti-IL-4 antibody (11B11; BD Biosciences), 10 ng/ml IL-2, and 5 ng/ml recombinant mouse IL-12 (PeproTech) were added. For Th2-polarizing condition, 10 μg/ml anti-IFN-γ antibody (XMG1.2; BD Biosciences), 10 μg/ml anti-IL-12 antibody (C17.8; BD Biosciences), 10 ng/ml IL-2, and 10 ng/ml recombinant mouse IL-4 (PeproTech) were added. On day 4 of culture, cells were diluted. On day 6, cells were restimulated with plate-bound anti-CD3e and anti-CD28 antibodies for 6 h. During the last 2 h, 10 μg/ml brefeldin A (Sigma-Aldrich) was added. Half of the cells were

analyzed using flow cytometry to evaluate eGFP expression, while the other half were fixed with 4% paraformaldehyde, permeabilized with Perm/Wash buffer (BD Biosciences), and used for intracellular staining to confirm differentiation by detecting IFN-γ and IL-4 (data not shown).

viii. *In vivo* and *ex vivo* imaging

Postnatal day 4 mice (anesthetized with isoflurane vapor) or fresh organs from adult mice were analyzed using an IVIS spectrum (Caliper Life Sciences). The excitation and emission wavelengths used in this study were 465 nm and 520 nm, respectively. Acquired data were analyzed using Living Image 4.0 software.

ix. Bioinformatics

Comparisons of mouse genomic sequences with human, dog, and rat were performed using VISTA (http://genome.lbl.gov/vista/index.shtml). The information describing the position of DNase I hypersensitive sites (DHSs) in human $CD4^+$ T cells were obtained from the UCSC genome browser (Boyle, et al., 2008; Crawford, et al., 2006; Crawford, et al., 2004; Crawford, et al., 2006; Xi, et al., 2007).

x. ESTs

The information regarding the two expressed sequence tags (ESTs) that are located within 300 kbp of *TCE-7.1* is as follows: *AK080422*, *Mus musculus* 7-day neonate cerebellum cDNA, RIKEN full-length enriched library, clone ID A730010B06 (http://genome.ucsc.edu/cgi-

bin/hgGene?hgg_gene=uc008ihc.1&hgg_prot=&hgg_chrom=chr2&hgg_start=9512845&hgg_end=9519470&hgg_type=knownGene&db=mm9&hgsid=186400201); and *AK035738*, *Mus musculus* adult male urinary bladder cDNA, RIKEN full-length enriched library, clone ID 9530097M04 (http://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=uc008ihb.1&hgg_prot=&hgg_chrom=chr2&hgg_start=9274116&hgg_end=9369303&hgg_type=knownGene&db=mm9&hgsid=186400201).

C. RESULTS

i. Neither the *Gata3-1a* nor *-1b* promoter confers T cell autonomous expression *in vivo*

Given our previous report that sequences in the *Gata3-1b* (gene-proximal) promoter exerted differential T cell activity in transfection experiments (George, et al., 1994), we first asked whether the same promoter was capable of directing T cell transcription *in vivo*. Transgenic mice were generated in which an eGFP reporter cassette was directed by *1b* promoter sequences (*1b*.eGFP), and expression in T cells was monitored by flow cytometry. Surprisingly, the reporter gene failed to be expressed in T cells of adult transgenic ($Tg^{1b.eGFP}$) mice (Figure 1A). In contrast, mice expressing a germ line eGFP–GATA-3 fusion protein ($Gata3^{g/+}$) (Figure 1A) or a germ line *lacZ* insertion at the *Gata3* initiation codon ($Gata3z^{/+}$) (Figure 1B) both robustly express the reporters in T cells (Hendriks, et al., 1999; Hosoya, et al., 2009). These data demonstrate that the *Gata3-1b* promoter is insufficient to confer T cell-specific transcription *in vivo* and, therefore, that an additional *cis* element(s) is required.

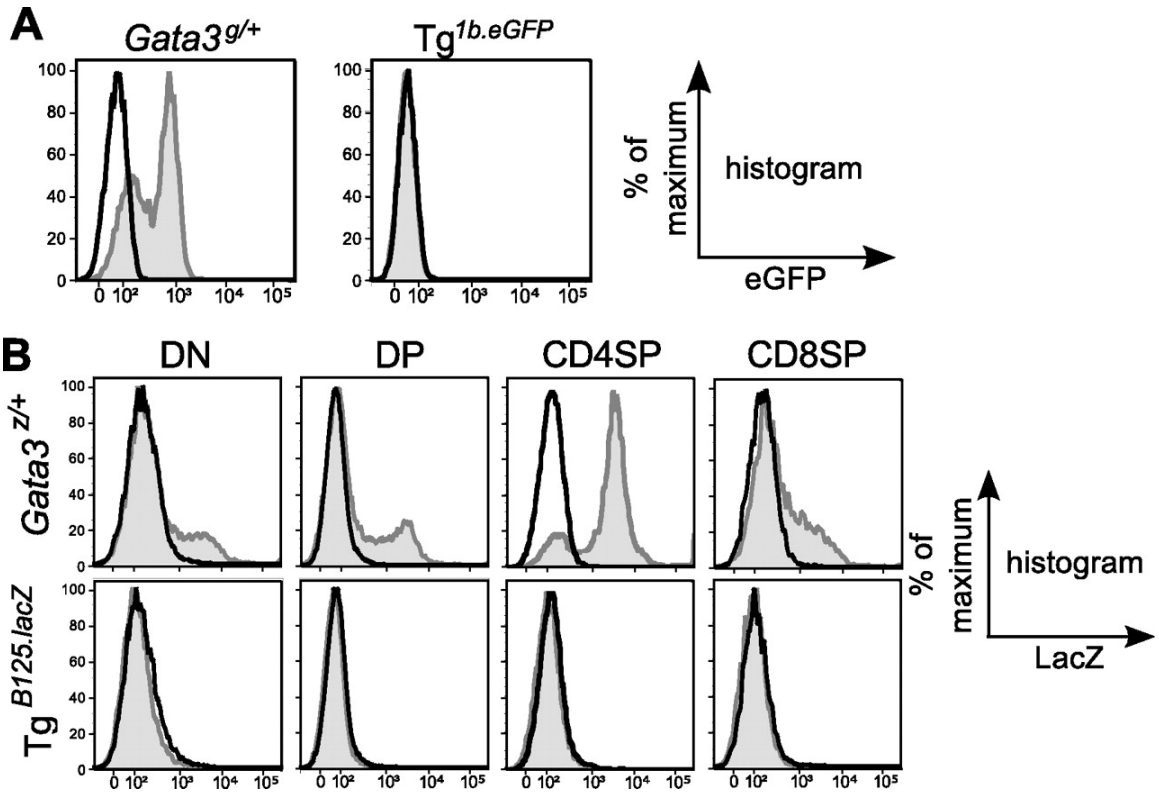


Figure 1. Neither the *Gata3-1b* promoter alone nor a 662-kbp *Gata3/LacZ* YAC containing both *1a* and *1b* promoters recapitulates GATA-3 activity in thymocytes. (A) eGFP expression in CD4 SP thymocytes from *Gata3*$^{g/+}$ mice (left, gray shaded histogram) or Tg$^{1b.eGFP}$ mice (right, gray shaded histogram). The black-line (open) histograms indicate eGFP fluorescence in wild-type thymocytes. Data represent at least three mice of each genotype. (B) LacZ expression in thymocytes stained with anti-CD4 and anti-CD8 antibodies was examined by flow cytometry. Each population was gated as depicted. The gray shaded histograms indicate fluorescein di-β-D-galactopyranoside fluorescence (13) resulting from hydrolysis due to β-galactosidase expression in either *Gata3-lacZ* knock-in (*Gata3*$^{z/+}$) (17) or B125-*lacZ* YAC transgenic (27) mice, while the black-line (open) histograms indicate expression in wild-type mice. These individual data are representative of results from at least three mice of each genotype.

ii. A potent T cell element located far 3′ to *Gata3*

We previously generated B125 yeast artificial chromosome (YAC) *LacZ* reporter transgenic mice harboring 662 kbp of genomic DNA containing the 33-kbp *Gata3* structural gene as well as vast swaths of adjacent 5′ (451-kbp) and 3′ (211-kbp) genomic noncoding sequences (Figure 1B, Tg$^{B125.LacZ}$) (Lakshmanan, et al., 1998;

Lakshmanan, et al., 1999). We compared β-galactosidase expression in the thymocytes of $Tg^{B125.LacZ}$ mice and *LacZ* germ line knock-in ($Gata3^{z/+}$) animals. Surprisingly, *LacZ* expression was not observed in adult $Tg^{B125.LacZ}$ thymocytes (Figure 1B). In contrast, the $Gata3^{z}$ germ line knock-in allele was strongly expressed, most robustly in the CD4 SP population that also abundantly expresses endogenous GATA-3 in thymocytes (Figure 1B). These data demonstrate that even 662 kbp of contiguous genomic sequence, including the gene and both *Gata3* promoters, was insufficient to direct *Gata3* T cell transcription *in vivo* and that the *cis* elements required for T lineage specification must be located beyond the boundaries of that YAC.

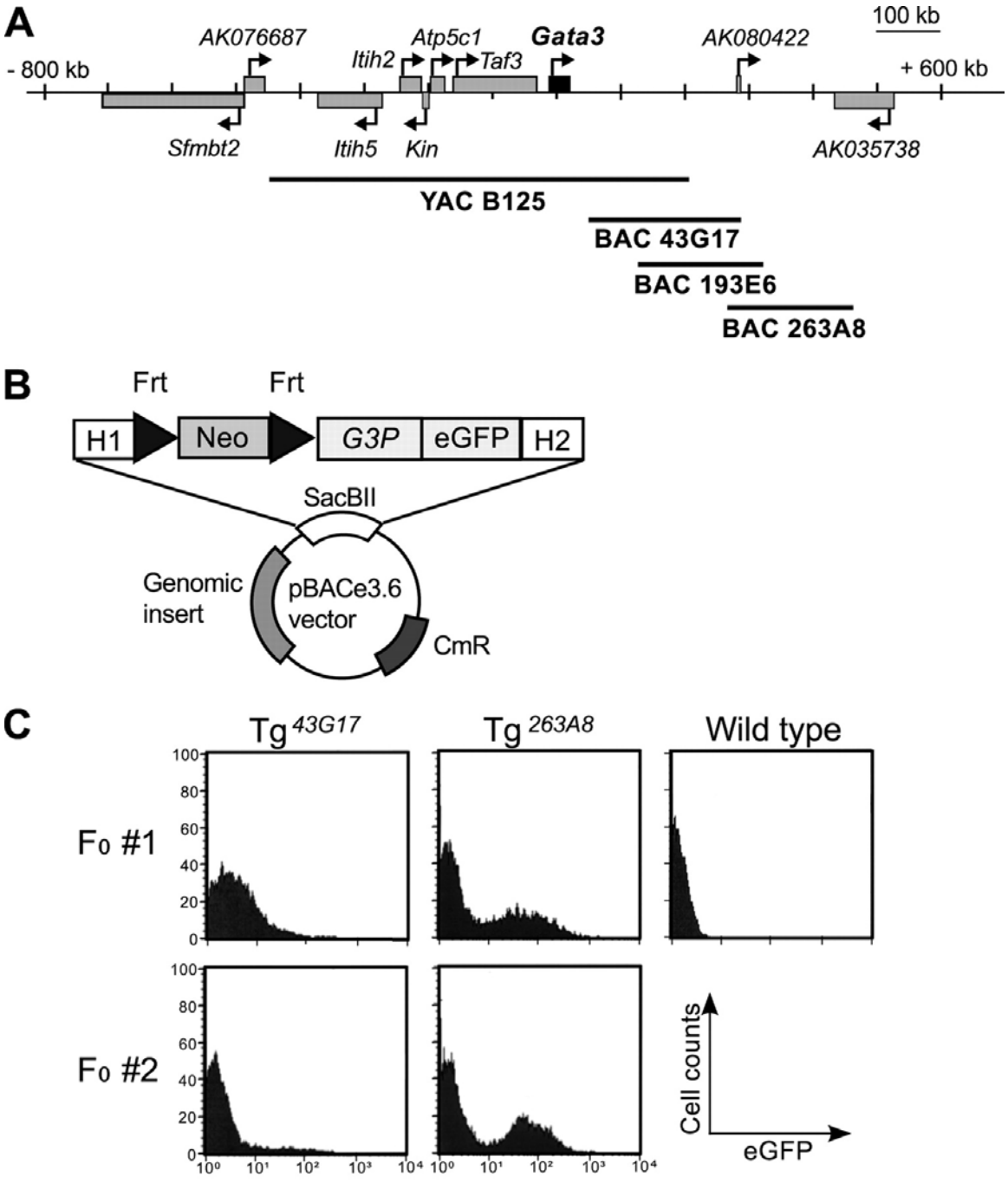


Figure 2. A candidate T lymphocyte enhancer element is located far 3' to the *Gata3* gene. (A) The *Gata3* gene and adjacent genes on mouse chromosome 2. The relative genomic positions of the BAC and YAC clones examined in this study are depicted graphically. (B) Schematic diagram of the targeting cassette used to generate modified BACs. H1 and H2, homology arms; Neo, neomycin resistance gene; *G3P, Gata3* promoter; SacBII, SacBII gene present in the vector backbone of the RPCI-23 mouse BAC library; CmR, chloramphenicol acetyltransferase gene C, founder screening of BAC-trap Tg embryos. eGFP expression in total thymocytes recovered from E18.5 F_0 Tg embryos was analyzed by flow cytometry. The results of two independent F_0 Tg embryos for each BAC clone are shown; in each case, a

fraction of the thymocytes expressed eGFP, except from wild-type embryos.

Next, we began to examine sequences lying even further away using the coupled BAC/transgenic (BAC-trap) assay we developed and exploited previously to identify several distant *Gata2* urogenital enhancers (Khandekar, et al., 2004). Three BACs that overlapped and extended 3′ to the B125 YAC (Figure 2A) were modified by recombineering (Lee, et al., 2001) to insert an eGFP reporter gene directed by the *Gata3-1b* (gene-proximal) promoter (*1b*.eGFP) into each BAC vector backbone (Figure 2B). The *1b* promoter was examined (instead of the more distal *1a* promoter) in these studies since more than 98% of peripheral T cell transcripts initiate from exon *1b* (Yu, et al., 2009). The three recombineered BACs were used to generate founder transgenic animals, and robust eGFP expression was detected in thymocytes from multiple transgenic animals (Figure 2C and data not shown) using all three BACs. Although eGFP expression initially appeared to differ between the different BAC clones, after recovery of multiple founders bearing each clone (and after subsequent analysis of multiple BAC transgenic lines), we concluded that the heterogeneity was due to mosaic expression of the transgenes in the founder transgenic mice. Based on the T cell-directed eGFP responses observed in this founder screen, we immediately focused on the region of overlap between the three BAC clones.

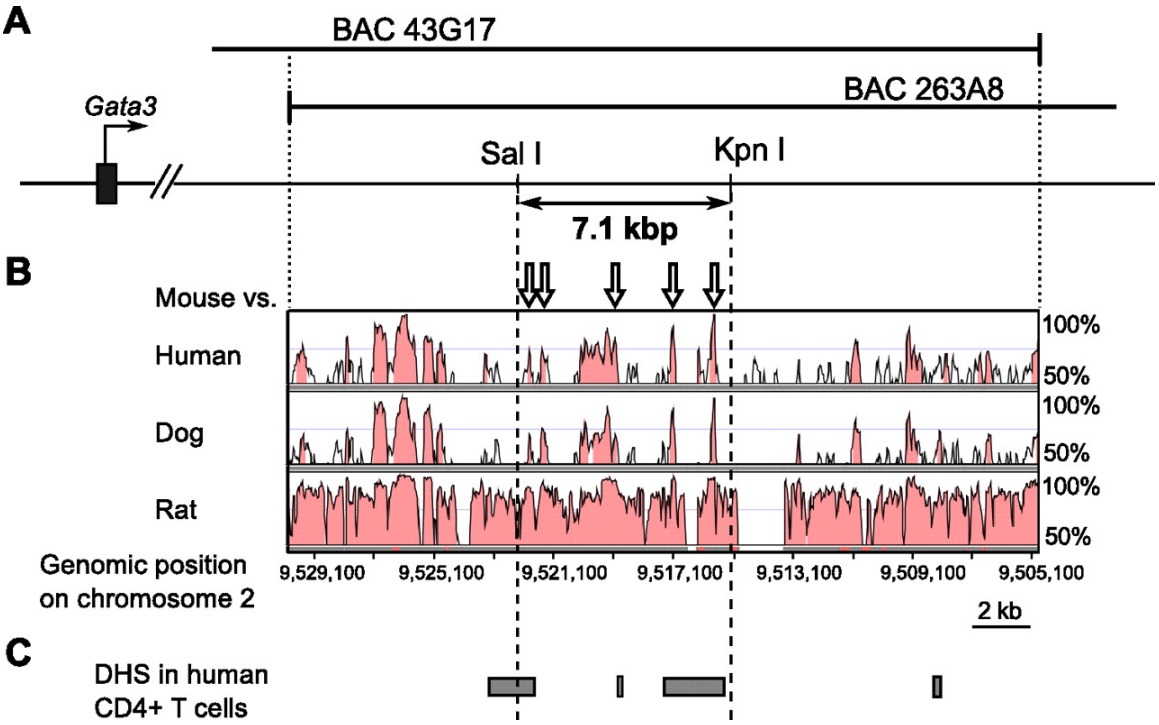


Figure 3. Mapping conserved noncoding sequences (CNS) and DNase I hypersensitive sites (DHS) in the overlap between two BAC clones. (A) The region of overlap (approximately 25 kbp) between BACs 43G17 and 263A8 is depicted. (B) Genomic sequences within the overlap (mouse chromosome 2, 9,530,005 to 9,504,884) were compared with the human, dog, and rat genomes, respectively. CNSs are colored pink. (C) The DHS homologies corresponding to human $CD4^+$ T cells are depicted as gray rectangles. Open arrows in panel B indicate CNSs that were predicted to be potential regulatory elements (i.e., high ESPERR scores).

Preliminary bioinformatic analysis revealed (Figure 3A and B) that multiple species-conserved noncoding sequence (CNS) elements lie within the overlap (approximately 25 kbp) between the RPCI-23 library BACs 43G17 and 263A8. We further informed the analysis by aligning DHS data from primary human $CD4^+$ T cells (Boyle, et al., 2008; Crawford, et al., 2006; Crawford, et al., 2004; Crawford, et al., 2006; Xi, et al., 2007) (Figure 3C); several of the CNS were close to, or overlapped, DHS sites (Figure 3). Given the close relationship between conserved sequence elements and DHS sequences with transcriptional control, we

hypothesized that the most highly conserved CNS overlapping the DHS (shown in Figure 3B) might serve as a *cis* regulatory element that controls T cell-specific *Gata3* transcription.

iii. A 7.1-kbp genomic fragment directs *Gata3* activity at multiple T cell stages

Within the 25-kbp overlapping BAC interval, we initially focused on a 7.1-kbp SalI/KpnI restriction fragment that contained multiple CNS elements and that also had high predicted regulatory sequence potential (evolutionary and sequence pattern extraction through reduced representations [ESPERR]) scores (Taylor, et al., 2006) (data not shown); additionally, this fragment encompassed most of the DHS (Figure 3), and therefore we assigned to it the preliminary designation *TCE-7.1* (7.1-kbp T cell element). First, we asked whether or not this fragment contained T cell enhancer activity by deleting the corresponding region (Δ7.1) via recombineering from the BAC 43G17 clone into which a *1b*.eGFP reporter cassette had already been inserted. We chose to examine BAC 43G17 instead of the two others simply because we first observed transcription of eGFP reporter gene in thymocytes of Tg^{43G17} mice (data not shown). The deletion BAC was then used to generate transgenic mice ($Tg^{43G17\Delta7.1}$). eGFP fluorescence was conspicuously absent in peripheral $CD4^+$ cells in all founder $Tg^{43G17\Delta7.1}$ mice (0/14 transgenic mice expressed eGFP fluorescence) compared to that in the cells of the parental Tg^{43G17} mice (Table 1), indicating that *TCE-7.1* is necessary for direction of reporter gene transcription in T cells. To ask if that same fragment alone was sufficient to enhance T cell transcription, it was linked to the same *1b*.eGFP reporter cassette that was used in the BAC vector

modification recombineering experiments; this reporter was then used to generate founder transgenic mice ($Tg^{7.1\text{-}1b.eGFP}$). We found that eGFP in peripheral $CD4^+$ T cells increased in all founder mice (6/6) bearing *TCE-7.1* linked to the promoter compared to in transgenic mice bearing the *1b* promoter alone (Table 1).

Table 1.
eGFP expression in the peripheral blood of F_0 Tg mice

Transgene	No. of mice with $CD4^+$ $eGFP^+$ cells/no. of mice with Tg (PCR^+)
Tg^{43G17}	7/8
$Tg^{43G17\Delta 7.1}$	0/14
$Tg^{1b.eGFP}$	1[a]/14
$Tg^{7.1\text{-}1b.eGFP}$	6/6

[a]eGFP expression was detected in $CD4^+$ cells and granulocytes ($Gr1^+Mac1^+$) in only one founder $Tg^{1b.eGFP}$ mouse. We concluded that it was ectopic expression since the remaining founder $Tg^{1b.eGFP}$ mice never expressed eGFP in $CD4^+$ cells.

We next established transgenic lines bearing each of these reporters and assayed the lines for when (during T cell development) and where (in which organs) the transgenes were expressed. We initially analyzed thymocytes from multiple established transgenic lines: six lines of $Tg^{7.1\text{-}1b.eGFP}$, four lines of $Tg^{1b.eGFP}$, five lines of Tg^{43G17}, and three lines of $Tg^{43G17\Delta 7.1}$. Thymocytes were electrically gated into DN1 to DN4, DP, CD4SP, and CD8SP stages using anti-CD4, CD8, CD25, and CD44 antibodies; we found that eGFP was expressed in those cells in all lines of both Tg^{43G17} and $Tg^{7.1\text{-}1b.eGFP}$ mice. In contrast, eGFP expression was not observed in the absence of *TCE-7.1* (data not shown). Based on those pilot experiments, we chose two lines of each construct-derived transgenic mouse and examined their T cell expression profiles in detail as described in Materials and Methods. We found that eGFP was expressed in CD4SP $CD69^+$ thymocytes as well as in other stages of both Tg^{43G17} mice and $Tg^{7.1\text{-}1b.eGFP}$ mice. In contrast, eGFP expression was essentially abolished when *TCE-7.1*

was deleted, and expression reverted to the levels observed in nontransgenic mice (Figure 4 A and B and data not shown). Moreover, we found that the pattern of eGFP expression in Tg^{43G17} mice and Tg$^{7.1\text{-}1b.eGFP}$ mice reflected the expression of endogenous *Gata3* during late stages of thymocyte development in which positive selection and CD4 versus CD8 lineage choice occur.

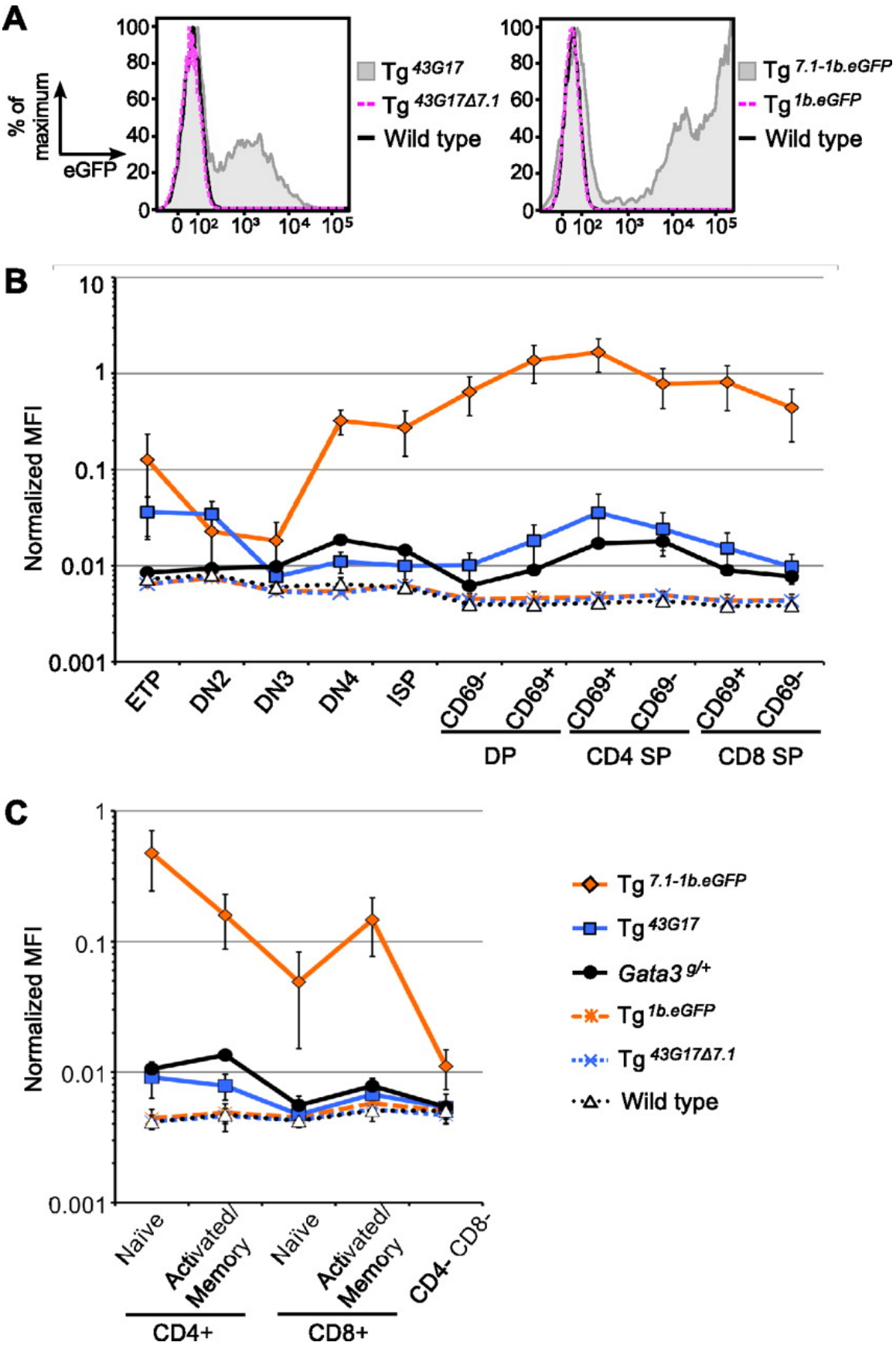

A
% of maximum
eGFP
100
80
60
40
20
0
0 10² 10³ 10⁴ 10⁵
Tg 43G17
Tg 43G17Δ7.1
Wild type
Tg 7.1-1b.eGFP
Tg 1b.eGFP
Wild type
B
Normalized MFI
10
1
0.1
0.01
0.001
ETP
DN2
DN3
DN4
ISP
CD69-
CD69+
CD69+
CD69-
CD69+
CD69-
DP
CD4 SP
CD8 SP
C
Normalized MFI
1
0.1
0.01
0.001
Naive
Activated/ Memory
Naive
Activated/ Memory
CD4- CD8-
CD4+
CD8+
Tg 7.1-1b.eGFP
Tg 43G17
Gata3 g/+
Tg 1b.eGFP
Tg 43G17Δ7.1
Wild type

Figure 4. A 7.1-kbp fragment (*TCE-7.1*) within the BAC overlap directs αβ T cell reporter gene transcription. (A) eGFP expression in CD69^{+} CD4 SP thymocytes from Tg^{43G17}, Tg$^{43G17\,\Delta 7.1}$, Tg$^{7.1\text{-}1b.eGFP}$, Tg$^{1b.eGFP}$, and wild-type mice. Data are representative of results from at least three individual mice of each genotype. (B and C) Normalized mean fluorescence intensity (MFI) of eGFP in each population of thymocytes (B) and splenocytes (C). MFI was normalized using the LinearFlow green flow cytometry intensity calibration kit and presented as percentage of relative fluorescence of calibration beads. Error bars (B and C) denote means ± standard deviations (SD). Two lines of each construct-derived Tg mouse were examined (data not shown), and at least three individual mice of each Tg line were analyzed. Data represent a single line from each construct-derived Tg mouse. Note that all MFI expression data are presented on a log scale. ETP, lineage-negative (Lin^{-}) CD25^{low} c-Kit^{high}; DN2, Lin^{-} CD25^{high} c-Kit^{high}; DN3, Lin^{-} CD25^{high} c-Kit^{low}; DN4, Lin^{-} CD25^{low} c-Kit^{low}; ISP, TCRbeta^{low} CD8 SP; CD69^{-} DP, TCRbeta^{low} CD69^{-} DP; CD69^{+} DP, TCRbeta^{+} CD69^{+} DP; CD4 SP CD69^{+}, TCRbeta^{+} CD69^{+} CD4 SP; CD4 SP CD69^{-}, TCRbeta^{+} CD69^{-} CD4 SP; CD8 SP CD69^{+}, TCRbeta^{+} CD69^{+} CD8 SP; CD8 SP CD69^{-}, TCRbeta^{+} CD69^{-} CD8 SP; Naïve CD4^{+}, CD4^{+} CD62L^{high} CD44^{low}; Activated/Memory CD4^{+}, CD4^{+} CD62L^{low} CD44^{high}; Naïve CD8^{+}, CD8^{+} CD62L^{high} CD44^{low}; Activated/Memory CD8^{+}, CD8^{+} CD62L^{low} CD44^{high}.

Endogenous *Gata3* is induced by T cell receptor (TCR) signaling during positive selection at the DP stage, and this induction requires the activity of transcription factor c-Myb. After DP cells differentiate into intermediate (CD69^{+}) CD4 SP cells, endogenous *Gata3* expression remains high and then gradually diminishes during maturation into mature (CD69^{-}) CD4 SP cells. In contrast, there is no induction of *Gata3* in intermediate (CD69^{+}) CD8 SP cells, and its level of expression diminishes even further as the cells differentiate into mature (CD69^{-}) CD8 SP cells (Hernandez-Hoyos, et al., 2003; Maurice, et al., 2007; Nawijn, et al., 2001). We observed induced expression of the MFI of eGFP after cell differentiation from the preselection (CD69^{-} DP cells) stage into CD69^{+} DP cells, where positive selection has begun, in both Tg^{43G17} and Tg$^{7.1\text{-}1b.eGFP}$ mice. eGFP increased in intermediate CD4 SP cells and declined thereafter in mature CD4 SP cells, as does endogenous *Gata3*. In CD8 SP cells, eGFP declined, again reflecting the endogenous *Gata3* expression pattern (Figure 4B and data not shown). These results suggest that *TCE-7.1* contains *cis*

information required to direct the late stages of *Gata3*-regulated thymocyte development.

In contrast to late thymocyte development, some differences in expression characteristics of the *TCE-7.1* transgenes were detected at earlier stages than the reported expression of endogenous *Gata3* mRNA (David-Fung, et al., 2006; Tydell, et al., 2007). For example, $Tg^{7.1\text{-}1b.eGFP}$ and Tg^{43G17} mice displayed intense eGFP fluorescence at the ETP stage that gradually diminished as they differentiated into the DN2 and DN3 stages (Figure 4B and data not shown). In addition, $Tg^{7.1\text{-}1b.eGFP}$ $CD69^-$ DP thymocytes expressed eGFP more intensely than immature CD8 SP (ISP) cells, while Tg^{43G17} exhibited no difference between those two populations. Neither is true of GATA-3 expression from the endogenous locus (Figure 4B). Taken together, we tentatively conclude that the *cis* element(s) that is required to stimulate *Gata3* transcription at the DN3 stage (David-Fung, et al., 2006; Tydell, et al., 2007) or required to negatively regulate *Gata3* at the ETP stage must be located beyond the boundaries of the 43G17 BAC, while additional regulatory elements that negatively regulate *Gata3* at the $CD69^-$ DP stage or positively regulate it at the ISP stage must exist outside *TCE-7.1* but may be included within the boundaries of BAC 43G17.

We also analyzed splenocytes to examine reporter gene expression in peripheral T cells in greater detail. Both Tg^{43G17} and $Tg^{7.1\text{-}1b.eGFP}$ displayed higher eGFP MFI in naïve $CD4^+$ cells than in naïve $CD8^+$ cells, which is similar to the pattern of

endogenous *Gata3*. Those activities were essentially ablated in the absence of *TCE-7.1* (Figure 4C and data not shown). These results indicate that *TCE-7.1* is critical for transcription in peripheral T cells, although small differences were detectable. For example, eGFP was higher in naïve CD4$^+$ cells than in activated/memory CD4$^+$ cells of both Tg^{43G17} and Tg$^{7.1\text{-}1b.eGFP}$ mice (Figure 4C), which again does not perfectly reflect the *in vivo* changes in GATA-3 that occur during T cell differentiation. In addition, eGFP in naïve and activated/memory CD8$^+$ splenocytes remained high compared to CD4$^-$ CD8$^-$ splenocytes, especially in Tg$^{7.1\text{-}1b.eGFP}$ mice (Figure 4C). Taken together, these results demonstrate that *TCE-7.1* is vital for transcription of a *Gata3* promoter-directed reporter gene in thymocytes and in splenocytes *in vivo*, although the element within *TCE-7.1* may not alone be sufficient to precisely recapitulate all aspects of *Gata3* T cell expression. We therefore speculate that additional regulatory elements that negatively regulate *Gata3* in peripheral CD8$^+$ T cells may exist outside *TCE-7.1* but within the 43G17 BAC.

In order to determine whether *TCE-7.1* contains the element(s) specifying increased *Gata3* transcription in Th2 cells but not in Th1 cells (Zhang, et al., 1997; Zheng and Flavell, 1997), we analyzed the expression of eGFP in CD4$^+$ cells from Tg^{43G17} mice under a variety of cytokine stimulatory conditions (see Materials and Methods). After stimulation, eGFP in CD4$^+$ cells under Th2-polarizing and nonpolarizing conditions was not significantly altered, although the cells did express low levels of eGFP (Figure 5), in keeping with the observed properties of GATA-3 expression *in vivo*. We confirmed that the *in vitro* polarization of these cells was successful by

monitoring cytokine induction (IFN-γ and IL-4) (data not shown). In $CD4^+$ cells recovered from $Tg^{43G17\Delta 7.1}$ mice, eGFP was not observed. In contrast, $Tg^{7.1\text{-}1b.eGFP}$ $CD4^+$ cells cultured under stimulatory conditions somewhat surprisingly displayed greatly reduced eGFP fluorescence under Th2-polarizing or nonpolarizing conditions compared to that of naïve $CD4^+$ cells. Moreover, $Tg^{7.1\text{-}1b.eGFP}$ cells under Th1-polarizing conditions actually displayed higher eGFP expression than under Th2-polarizing or nonpolarizing conditions (Figure 5 and data not shown). Both results contradict the GATA-3 expression characteristics observed *in vivo*. These data suggest that *TCE-7.1*, either in concert with the *Gata3-1b* promoter or even within the context of the entire BAC43G17 clone, does not contain the regulatory element that specifies *Gata3* activation in Th2 cells or in activated $CD4^+$ T cells, although *TCE-7.1* does contain an element that is critical for high-basal-level transcription in stimulated $CD4^+$ cells. Furthermore, a putative *cis* element that must be located outside the boundaries described by *TCE-7.1*, but within the boundaries specified by BAC 43G17, may additionally be required for the repression of *Gata3* in Th1 cells and naïve $CD4^+$ T cells.

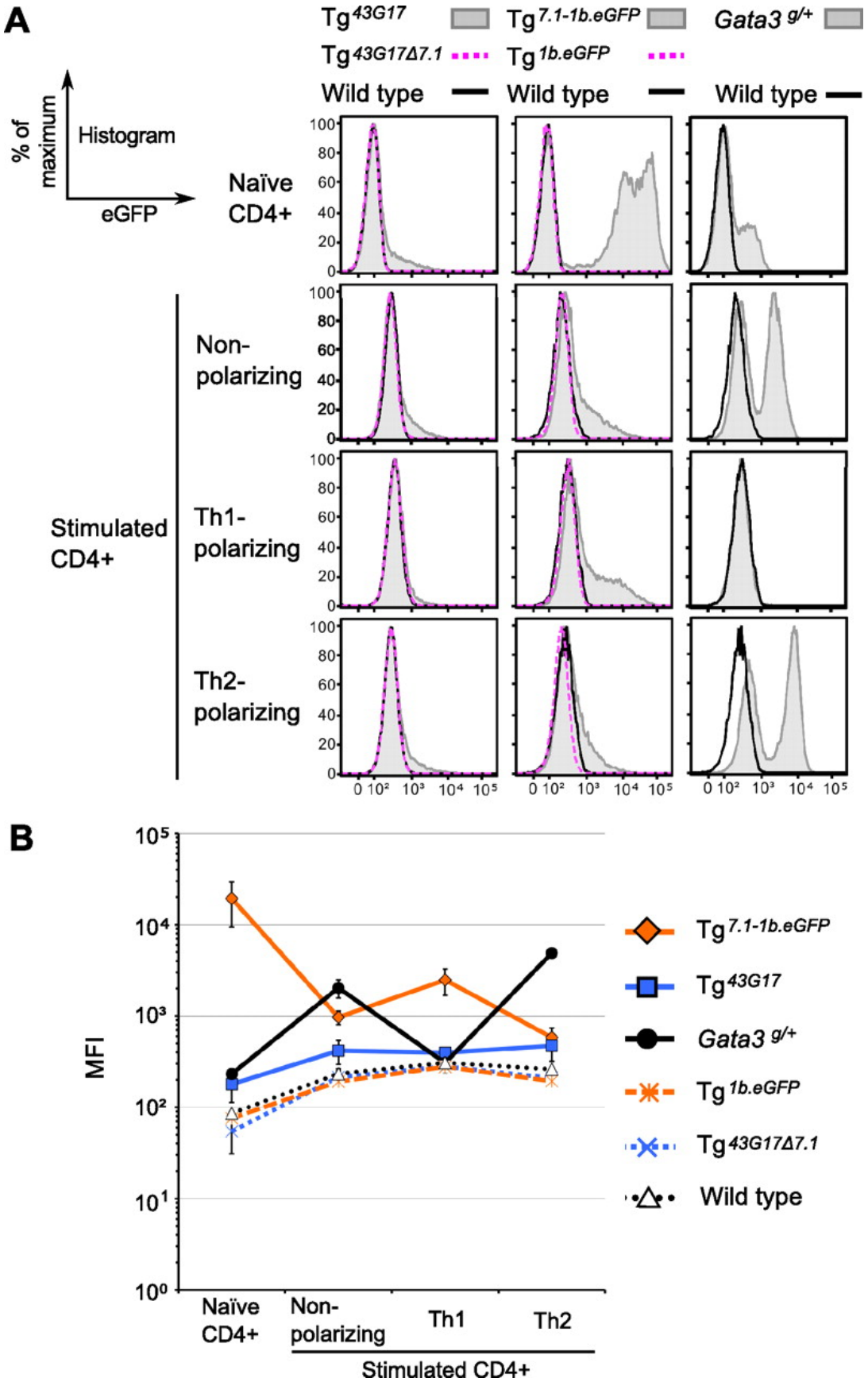

A
Tg43G17
Tg7.1-1b.eGFP
Gata3 g/+
Tg43G17Δ7.1
Tg1b.eGFP
Wild type
Wild type
Wild type
% of maximum
Histogram
eGFP
Naïve CD4+
Stimulated CD4+
Non-polarizing
Th1-polarizing
Th2-polarizing
0 10² 10³ 10⁴ 10⁵
B
MFI
Naïve CD4+
Non-polarizing
Th1
Th2
Stimulated CD4+
Tg7.1-1b.eGFP
Tg43G17
Gata3 g/+
Tg1b.eGFP
Tg43G17Δ7.1
Wild type

Figure 5. *TCE-7.1* directs reporter gene transcription in stimulated CD4$^+$ cells. (A) Histograms of eGFP in naïve CD4$^+$ splenocytes and stimulated CD4$^+$ splenocytes under various conditions. Data are representative of results from at least three individual mice of each genotype. (B) MFI of eGFP in naïve and stimulated CD4$^+$ splenocytes under various conditions. Note that MFI was not normalized using calibration beads in this experiment. Error bars denote means ± SD. Two lines of each construct-derived Tg mouse were examined (data not shown), and at least three individual mice of each Tg line were analyzed. Data represent a single line from each construct-derived Tg mouse. Note that MFI expression data are presented on a log scale.

iv. *TCE-7.1* bears *Gata3* T and NK cell-specific regulatory information

In addition to its well-characterized roles in αβ T cell, sympathoadrenal, kidney, parathyroid, breast epithelial, and epidermal development (Asselin-Labat, et al., 2007; Grigorieva, et al., 2010; Kaufman, et al., 2003; Kouros-Mehr, et al., 2006; Kurek, et al., 2007; Lim, et al., 2000; Moriguchi, et al., 2006; Tsarovina, et al., 2010), GATA-3 has been shown to play critical roles in the generation and maturation of NK cells (Samson, et al., 2003; Vosshenrich, et al., 2006). In addition, GATA-3 is expressed in γδ T cells (Hosoya, et al., 2009) as well as in hematopoietic progenitors (Sambandam, et al., 2005), although its function there is not well understood. We analyzed NK cells, γδ T cells, and hematopoietic progenitors (lineage$^-$ Sca1$^+$ c-Kit^{hi} [LSK]) in multiple transgenic lines to ascertain whether *TCE-7.1* was also active in those related lymphoid lineages.

Somewhat surprisingly, robust eGFP fluorescence was observed in thymic NK cells and γδ T cells in both Tg$^{7.1\text{-}1b.eGFP}$ and Tg^{43G17} mice, while mice bearing the 7.1-kbp deleted BAC Tg$^{43G17\Delta7.1}$ or mice bearing only the *Gata3-1b* promoter failed to express eGFP (Figure 6A). In contrast to GATA-3-expressing cells (e.g., αβ or γδ T cells or NK

cells), other hematopoietic lineages that do not express endogenous GATA-3 also failed to express eGFP, in both $Tg^{7.1\text{-}1b.eGFP}$ and Tg^{43G17} mice (Figure 6A). Furthermore, neither $Tg^{7.1\text{-}1b.eGFP}$ nor Tg^{43G17} express eGFP in the early hematopoietic progenitor compartment (Figure 6B). These results demonstrated that *TCE-7.1* is active in T and NK cells but not in other hematopoietic lineages or progenitors.

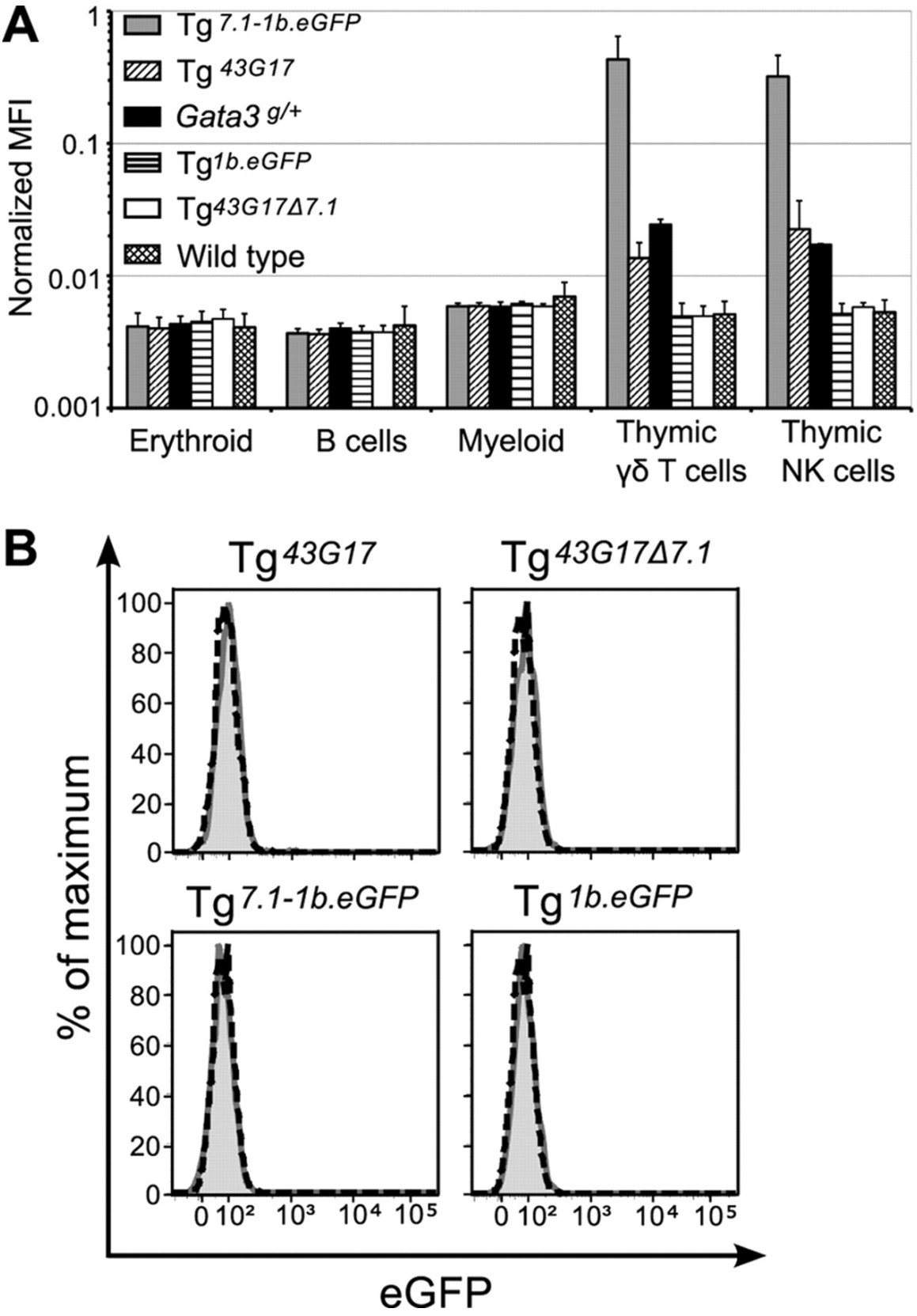


Figure 6. Among hematopoietic cells, *TCE-7.1* confers only NK cell and αβ and γδ T cell

enhancer activity. (A) Normalized MFI of eGFP in erythroid cells (TER119$^+$), B cells (CD19$^+$ B220$^+$ CD3$^-$), and myeloid cells (Gr1$^+$ Mac1$^+$) in the bone marrow, as well as γδ T cells (TCRγδ$^+$) and NK cells (CD3$^-$ CD19$^-$ DX5$^+$) in the thymus are shown. Error bars denote means ± SD. Note that MFI data are presented on a log scale. (B) eGFP expression in bone marrow hema- topoietic progenitors (Lin$^-$ Sca1$^+$ c-Kit^{hi}). The shaded histograms indicate each Tg mouse, while dashed lines indicate wild-type mice. For both panels A and B, two lines of each construct-derived Tg mouse were examined (data not shown), and at least three individual mice of each Tg line were analyzed. Data represent a single line from each construct-derived Tg mouse.

In order to more globally examine whether the eGFP expression conferred by *TCE-7.1* was T cell specific, we examined Tg$^{7.1\text{-}1b.eGFP}$ expression by IVIS Spectrum whole-body *in vivo* imaging. Robust eGFP fluorescence was detected exclusively in the thymi of Tg$^{7.1\text{-}1b.eGFP}$ neonates (Figure 7A). In contrast, other organs in these mice did not display expression above background levels; similarly, neither Tg$^{1b.eGFP}$ nor wild-type mice expressed detectable eGFP (Figure 7A and data not shown). *Ex vivo* imaging of individual adult organs confirmed the conclusions on the living neonatal mice. As shown in Figure 7B, eGFP fluorescence was detected only in the thymi of Tg$^{7.1\text{-}1b.eGFP}$ mice but not in other organs. In agreement with conclusions from the *in vivo* imaging, neither Tg$^{1b.eGFP}$ nor wild-type adult mice displayed detectable eGFP expression (Figure 7B). These results demonstrated conclusively that sequences within *TCE-7.1* direct exclusive T and NK cell-specific transcription of *Gata3*.

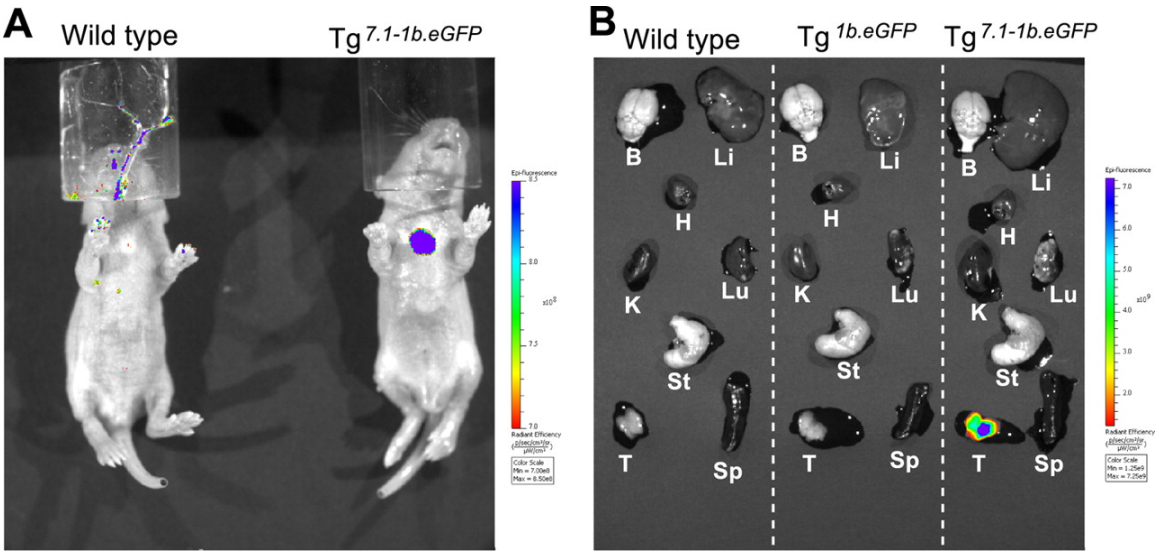


Figure 7. The *TCE-7.1* enhancer is T cell specific. (A) eGFP expression in living P4 mice. Data are representative of multiple pups examined in two independent experiments. (B) eGFP expression in various organs from each genotype of adult mice. Three individual mice of each genotype were analyzed. B, brain; Li, liver; H, heart; K, kidney; Lu, lung; St, stomach; T, thymus; Sp, spleen. In both panels A and B, two lines of both Tg$^{7.1\text{-}1b.eGFP}$ and Tg$^{1b.eGFP}$ mouse were examined (data not shown).

D. DISCUSSION

Here, we report that a 7.1-kbp DNA fragment (abbreviated *TCE-7.1*) located 280 kbp 3′ to the *Gata3* gene contains *cis* information that is critical for the transcription of *Gata3*, both at multiple stages of T cell development and in thymic NK cells. While previous experiments have identified central roles for GATA-3 at multiple stages of T cell development (Hosoya, et al., 2009; Pai, et al., 2004; Pai, et al., 2003; Zhu, et al., 2004), it is expressed at all stages that have been examined, but its level clearly differs markedly between developmental stages (e.g., in CD4 versus CD8 SP cells or in Th1 versus Th2 cells) (Hendriks, et al., 1999; Hernandez-Hoyos, et al., 2003; Zhang, et al., 1997; Zheng and Flavell, 1997). Although several *trans*-acting factors are believed to directly regulate *Gata3* (Amsen, et al., 2007; Fang, et al., 2007;

Maurice, et al., 2007; Yang, et al., 2009; Yu, et al., 2009), a coherent mechanism explaining how this information is integrated to allow differential, stage-specific *Gata3* expression at multiple stages of T cell development *in vivo* has not emerged.

The BAC-trap transgenic assay utilized in this report reiterated the generality of this assay, showing that even a very distant *cis* element, located more than 200 kbp from the structural gene (the amount of information usually borne in a BAC), can be identified, thus revealing the position of a cell-specific enhancer that is capable of conferring transcription to a reporter gene at several discrete developmental stages, from ETP to peripheral $CD4^+$ T cells, *in vivo*. The expression pattern of the reporter was similar to that of endogenous *Gata3* even though differences were documented. Since eGFP expression was not observed in the absence of the enhancer-bearing fragment, we conclude that this element is critical for transcription of the *Gata3* gene in T lymphocytes.

In this study, both Tg^{43G17} and $Tg^{7.1\text{-}1b.eGFP}$ displayed expression profiles that reflected almost perfectly the endogenous *Gata3* expression pattern in late stages of thymocyte development. At the DP stage, the *TCRα* locus begins to rearrange, and subsequently a low level of the TCRαβ complex is expressed. Those DP cells are poised for positive selection, although most of them fail to emerge intact from selection: only a few cells that have an appropriate TCR affinity are positively selected and finally emerge to differentiate into either CD4 SP or CD8 SP cells. Endogenous *Gata3* is induced at the onset of positive selection, and its expression is

controlled by TCR signaling (Hernandez-Hoyos, et al., 2003; Nawijn, et al., 2001). Based on the remarkably similar expression patterns of endogenous *Gata3*, Tg^{43G17}, and Tg$^{7.1\text{-}1b.eGFP}$ during late thymocyte development, we conclude that *TCE-7.1* contains the activity required for development through those stages.

We found that *TCE-7.1* harbors multiple putative transcription factor binding sites through bioinformatic analyses (data not shown). Perhaps not surprisingly, candidate binding sites for many transcription factors that are critical for T cell development can be identified in this region. For example, highly species-conserved sequences contain putative binding sites for transcription factors c-Myb, Runx, E2A, and TCF-1 as well as others. The proto-oncogene c-Myb is required for the induction of *Gata3* following TCR signaling, and the binding of c-Myb to the *Gata3-1b* promoter is detectable in thymocytes (Maurice, et al., 2007). It is important to remember that this same Myb binding site is present in the B125 YAC, and that the YAC-derived transgene was not expressed in thymocytes. Moreover, this binding site was absent in all of the BAC and *7.1-1b.eGFP* reporter constructs examined in this report, but nonetheless the reporter gene was induced in the T cells of both Tg$^{7.1\text{-}1b.eGFP}$ and Tg^{43G17} mice. Since the fragment bearing the T cell enhancer activity contains several putative c-Myb binding sites, it is possible that these enhancer sites participate in *Gata3* induction by c-Myb.

Although expression in both the Tg$^{7.1\text{-}1b.eGFP}$ and Tg^{43G17} mice resembled that of endogenous *Gata3*, whether this element or group of elements constitute a bona fide

cis element for *Gata3* has not been conclusively demonstrated. Only two spliced ESTs other than *Gata3* are located within 300 kbp of *TCE-7.1*. Those two ESTs, *AK080422* and *AK035738*, have been identified as nonprotein-coding mRNAs (Okazaki, et al., 2002). *AK080422* was detected in a mouse neonatal cerebellum cDNA library, while *AK035738* was observed in an adult male mouse urinary bladder cDNA library. The function of those two ESTs is unknown, and their expression in thymocytes has not been detected. Taken together, assigning *TCE-7.1* activity to the *Gata3* gene is likely the most conservative interpretation of these data.

Numerous interesting questions emerge from this study: does a single element within *TCE-7.1* control *Gata3* expression in αβT, γδT, and NK cells, or alternatively do multiple elements within *TCE-7.1* each regulate transcription in those distinct lineages? Do multiple elements, perhaps in different combinations, consort to elicit proper stage-specific *Gata3* activation during T cell development, or do different cofactors, all acting on a single *cis* element within *TCE-7.1*, function at different stages to confer the specificity? Answers to these fascinating questions should be resolved soon by the many groups studying *Gata3* function in T cell transcription.

Finally, the data predict that (one or multiple) sequences within *TCE-7.1* must collaborate with as-yet-undiscovered *cis* elements lying beyond the *TCE-7.1* boundaries to fully recapitulate proper *Gata3* expression in T lymphocytes. *TCE-7.1* clearly contains sequences that are important for directing transcription in both the

T cell and NK cell lineages. Further analysis of *Gata3* gene regulation in T cells should help us to not only understand the complex mechanisms of gene regulation that are used to confer T cell specificity to this regulatory network but perhaps might also shed light on the mechanisms where GATA-3 may play an oncogenic role in leukemia and lymphoma (Nawijn, et al., 2001; van Hamburg, et al., 2008; Xu and Kee, 2007).

Chapter III

Genome-Wide Binding of the Orphan Nuclear Receptor TR4 Suggests Its General Role in Fundamental Biological Processes

A. INTRODUCTION

There are an estimated 1400 site-specific DNA binding factors encoded in the human genome (Vaquerizas, et al., 2009). Although these factors can influence transcription when their binding sites are cloned in front of core promoters, they usually do not function alone. Most often, individual transcription factors collaborate to orchestrate gene expression through combinatorial binding to regulatory regions in chromatin (Farnham, 2009). These regions, termed *cis* modules, thereby activate, repress or otherwise epigenetically modify the transcriptional responses of individual genes. Elucidating the position and activities of individual *cis* modules using reporter genes is time consuming and expensive. With recent advances in DNA sequencing technology, it is now feasible to generate global protein-DNA interaction profiles by chromatin immunoprecipitation (ChIP) followed by ultra-high-throughput sequencing (Park, 2009). *Cis* modules can then often be identified by applying bioinformatics searches for one or more *cis* motifs recognized by unrelated alternative factors near the binding sites of the factor analyzed by ChIP-seq or by the co-localization of bound sites for two or more unrelated different site-specific factors.

Nuclear receptors (NRs) represent a special class of transcription factors that direct target gene transcription in a ligand-dependent fashion. NRs contain a DNA-binding domain that recognizes a specific DNA sequence, as well as a ligand binding domain that renders these factors environmentally-dependent regulators via interaction with distinct cognate ligands (Mangelsdorf, et al., 1995). The great majority of NRs homodimerize or heterodimerize with another NR, and then bind to two copies of a repeated hexanucleotide sequence (called a half-site) separated by variable spacing (Sandelin and Wasserman, 2005). The half-site consensus, AGGTCA, can occur in either orientation and variation from the consensus allows numerous alternative binding sites of (probably) variable affinity (Sandelin and Wasserman, 2005). Based on the number of spacer nucleotides separating the two half-sites and the orientation of the two half-sites relative to each other, NR binding sites have been categorized as direct repeats (DR0 - DR8), everted repeats (ER0 - ER8) or inverted repeats (IR0-IR8) (Sandelin and Wasserman, 2005).

NR2C2 (human testicular receptor 4, TR4, in the older nomenclature) belongs to the nuclear receptor superfamily and is termed an orphan receptor due to the fact that no ligand has been discovered (Lee, et al., 2002; Noy, 2007; Su Liu, 2010). TR4 was initially identified in hypothalamus, prostate, and testis cDNA libraries, but has since been demonstrated to be broadly expressed in many physiological systems (Bookout, et al., 2006; Chang, et al., 1994). For example, TR4 has been shown to activate target gene expression in liver carcinoma HepG2 cells (Lee, et al., 1997). In contrast, in erythroid cells, TR4 can heterodimerize with another closely related

family member (TR2, or NR2C1) and binds to a DR1 (direct repeats with one nucleotide spacer) element to repress target gene transcription (Omori, et al., 2005; Tanabe, et al., 2002; Tanabe, et al., 2007; Tanabe, et al., 2007). The binding affinity of the TR4 homodimer for the DR1 element *in vitro* is equivalent to that of the TR2:TR4 heterodimer (Tanabe, et al., 2007), and TR4 mRNA is more abundant than TR2 in human erythroid cells (Tanabe, unpublished observations). However, the broader physiological functions for, and the *in vivo* genome-wide binding patterns of, this broadly expressed nuclear receptor are obscure. We therefore chose to initially investigate genome wide TR4 binding anticipating that these studies might reveal some common, but also perhaps some tissue-specific, metabolic processes to which this factor contributes.

In this study we investigated the first genome-wide identification of cellular targets of TR4 and preliminary characterization of TR4 *in vivo* binding in multiple cell types, including those in which TR4 has been suggested to be an activator (liver) and cells in which TR4 has been suggested to be a repressor (blood). Using ChIP-seq, we determined TR4 *in vivo* binding in four human ENCODE cell lines: K562 erythroleukemia cells, HepG2 liver carcinoma, HeLa cervical carcinoma, and GM12878 immortalized lymphoblast cells. TR4 binding patterns identified in the four diverse cell lines suggest that this factor controls cell metabolism by binding to the proximal promoter regions that are common to several hundred genes. Motif analysis shows that TR4 strongly prefers a DR1 sequence to all other categories of repeat elements *in vivo*. By integration of TR4 binding data with histone

modification patterns and other genomic structures, we predict, and then experimentally test, putative *cis* modules.

B. RESULTS AND DISCUSSION

i. Identification of genome-wide TR4 binding sites

With no known ligand and few proposed binding sites in mouse and human cell lines (Chen, et al., 2008; Kim, et al., 2005; Liu, et al., 2007; Shyr, et al., 2009), the function of the TR4 orphan nuclear receptor was largely unknown when we began these studies. Previous studies examined its function in different blood cells and found that TR4 bound to the CD36 promoter in macrophages (Xie, et al., 2009) and to the GATA1 enhancer G1HE (Tanabe, et al., 2007) in $CD34^+$ cells, but only after *in vitro* differentiation for 11 days. To further elucidate biological roles for TR4, we set out to identify *in vivo* TR4 binding sites throughout the entire human genome using chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq). We wanted to compare its binding profiles in cells derived from different tissue types. We chose to identify TR4 targets in cell types selected by the ENCODE Consortium (http://www.genome.gov/10005107), including human chronic myelogenous leukemia cells (K562), human cervical carcinoma cells (HeLa), lymphoblastoid cells (GM12878), and hepatocellular carcinoma cells (HepG2). By characterizing its binding in these cell lines, we could compare TR4 binding sites with other transcription factor binding sites and histone marks determined by other ENCODE groups examining these same cell types. We first validated the presence of TR4 protein in these cell lines by Western Blot analysis. We began our ChIP

experiments using the hematopoietic cell line K562 and the liver cell line HepG2, but were unable to confirm TR4 enrichment at targets previously published in the specialized and differentiated hematopoietic cells. Therefore, we initially proceeded without having positive controls for the ChIP assays. We prepared sequencing libraries from ChIP experiments from two independently grown batches of HepG2 cells. Samples were sequenced using the Illumina GA2 platform and ChIP-seq data were analyzed using the Sole-search software (http://chipseq.genomecenter.ucdavis.edu/cgi-bin/chipseq.cgi; [21]). Only sequences that uniquely matched those in the human genome were retained for analysis. 9.7 million sequence reads were obtained from replicate 1 and 8.2 million from replicate 2. Using the Sole-search peak calling program with default settings (FDR 0.0001, alpha value 0.001), 1,547 and 2,246 TR4 binding sites were identified in HepG2 cells for replicate 1 and replicate 2, respectively. 1,243 (80%) of the 1,547 peaks called from replicate 1 were also present in the 2,246 peaks called from replicate 2. This overlap demonstrates good reproducibility between biological replicates. To obtain the final list of 2,672 TR4 binding sites in HepG2 cells, all reads (17.8 million) from both biological replicates were merged. We then performed TR4 ChIP experiments for the other cell types and used standard PCR to confirm enrichment at three sites (TNFIAP1, SCAP, ECSIT) previously identified in HepG2 cells. ChIP-seq libraries were then prepared from two biological replicates using the TR4 antibody resulting in 23 million sequence reads for HeLa cells, 30 million for GM12878 cells and 16 million for K562 cells. 1,767 TR4 binding sites were identified in HeLa cells, 1,180 TR4 binding sites in GM12878 cells and 732 TR4

binding sites in K562 cells; see Figure 8 for the binding patterns of TR4 across the entire chromosome 12 in all four cell types.

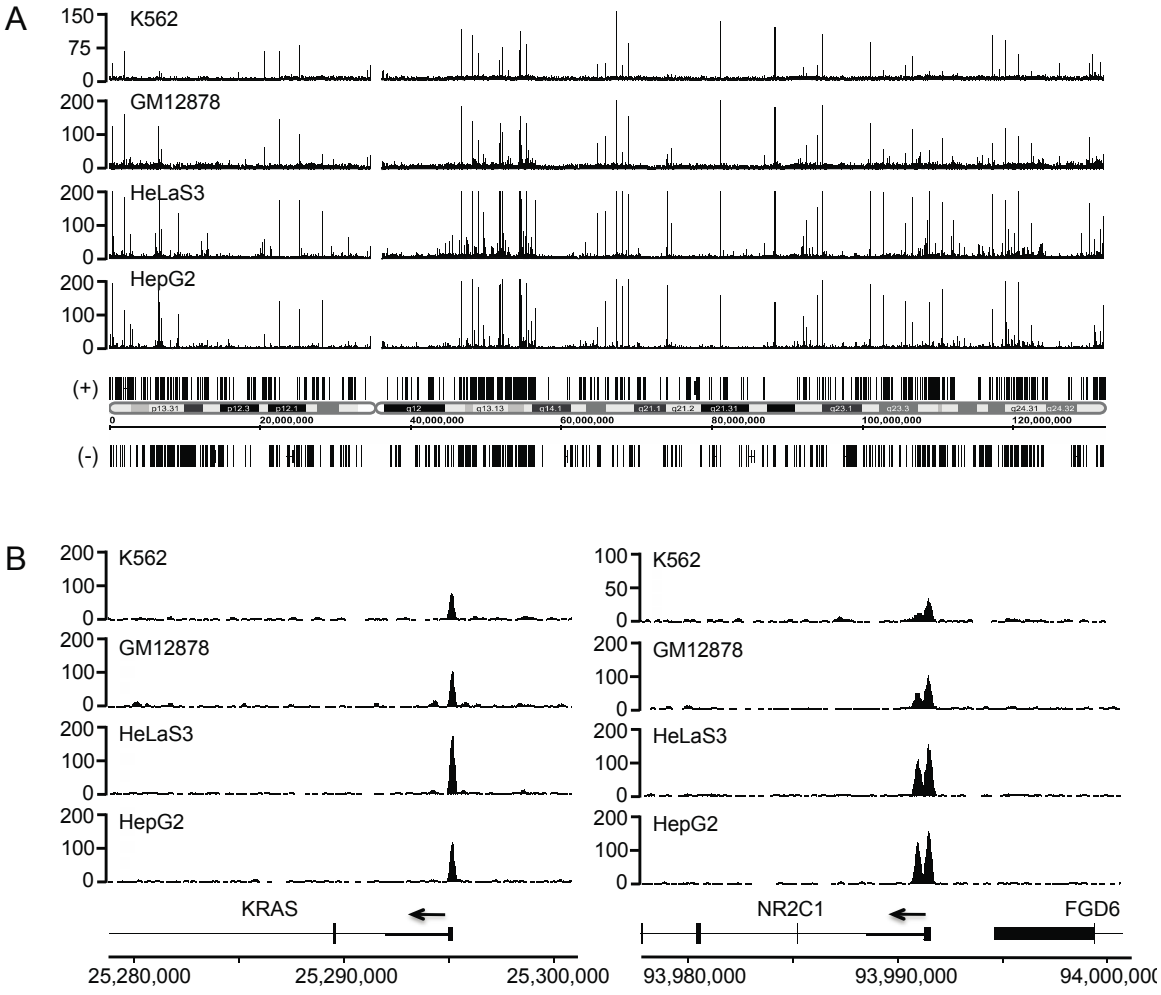


Figure 8. Comparison of TR4 targets in 4 different cell types. ChIP-seq binding patterns of TR4 (NR2C2) from K562, GM12878, HeLa, and HepG2 cells are shown (A) for entire chromosome 12 and (B) for target genes KRAS and TR2 (NR2C1). The number of tags reflecting the ChIP enrichments is plotted on the y axis and chromosomal coordinates (hg18) are shown on the x axis. RefSeq genes are indicated in (+) and (-) orientation. Target genes KRAS and TR2 are in (-) orientation as indicated by the arrows.

The position to which a transcription factor binds relative to the start site of transcription can provide insight into how the factor regulates transcription. For example, E2F family members bind to core promoter regions and are thought to

stimulate transcription by interaction with the basal transcription machinery (Fry, et al., 1999; Xu, et al., 2007). In contrast, other transcription factors, such as GATA1 or TCF4 (TCF7L2), show significant binding to sites often located more than 10 kb away from the gene that they regulate (Blahnik, et al., 2010; Fujiwara, et al., 2009), suggesting that these factors may regulate transcription by looping mechanisms. Although the number of TR4 binding sites varied among the different cell types, location analysis revealed that TR4 preferentially binds close to the transcription start sites of its target genes. The majority of TR4 binding sites (65-82%) is located either in the proximal promoter (up to 2 kb upstream of TSS) or is found within the first exon or first intron of a RefSeq gene. In HeLa cells, 36% of TR4 binding occurred in the proximal promoter and 41% in the gene region, mainly in the first exon or first intron (Figure 9A and 9B). To further characterize TR4 binding sites, TR4 ChIP-seq reads were organized into 100 bp bins relative to the start site of transcription. The distribution of TR4 peaks relative to the transcription start site demonstrated that the majority of TR4 binding occurs between 1 kb upstream and 1 kb downstream of a TSS (Figure 9C). For example, 1,135 (63%) of the 1,767 HeLa binding sites were located within ± 1 kb from a TSS. This preference was also reflected in an elevated median height of peaks near a TSS; the median peak value was 114 for peaks within ± 1 kb of a TSS, but only 50 for peaks outside this range. For the rest of our studies, we therefore focused on the targets found within 1 kb of a TSS. This encompassed 1,154 TR4 binding sites for HeLa, 1,732 for HepG2, 537 for K562 and 535 for GM12878 cells.

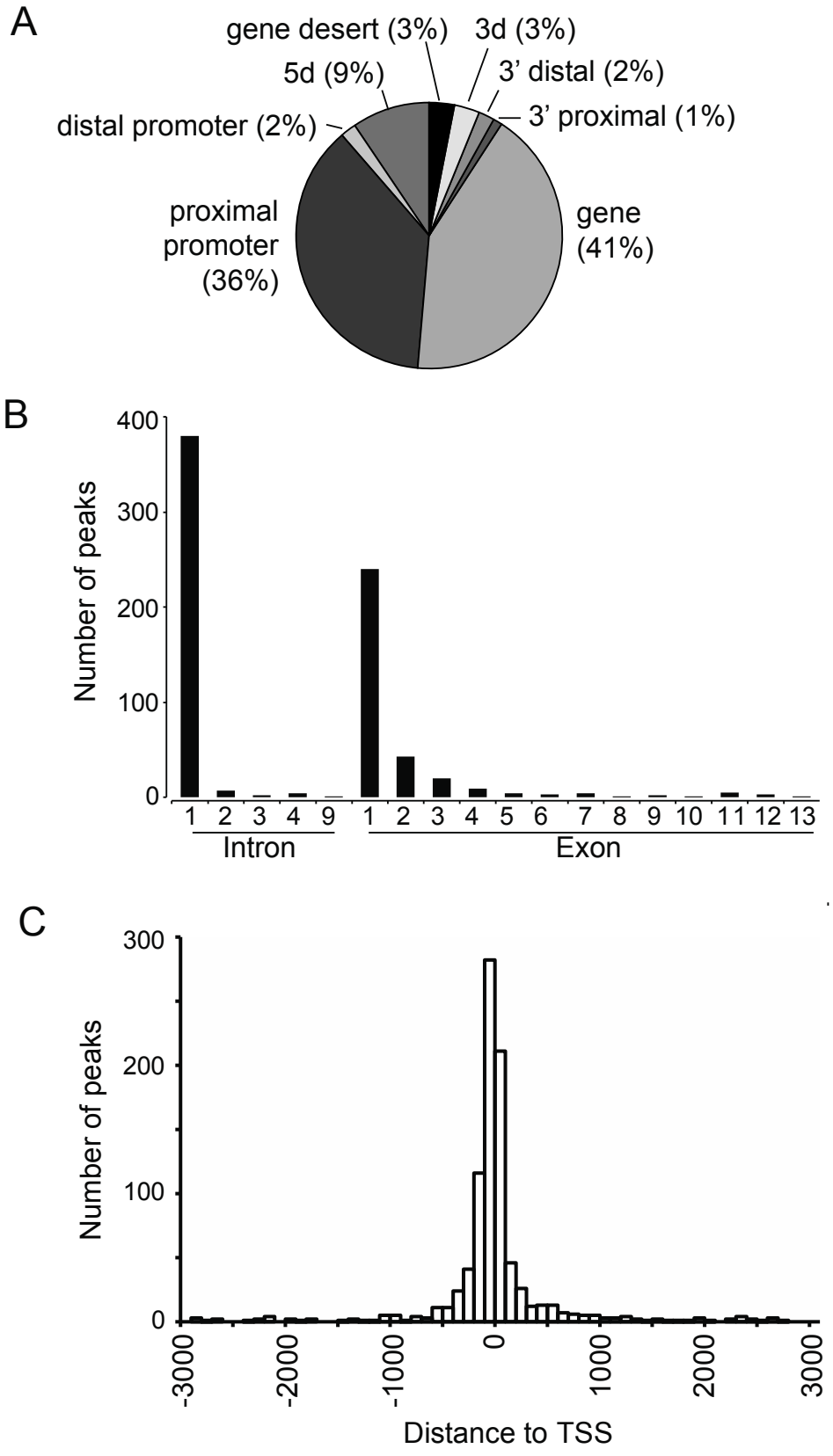


Figure 9. Location analysis of TR4 binding sites in HeLa cells. (A) Shown is a pie chart

indicating the distribution of called TR4 peaks. Categories are based on the distance of the peak to the nearest RefSeq gene: 5 d (10 - 100 kb upstream of TSS), distal promoter (2 - 10 kb upstream of TSS), proximal promoter (<2 kb upstream of TSS), gene (exon or intron), 3' proximal (<2 kb downstream of the last exon), 3' distal (2 - 10 kb downstream of the last exon), 3 d (10 - 100 kb downstream of the last exon), and gene desert (>100 kb from a RefSeq gene). (B) Distribution of peaks found within genes. (C) Histogram showing the distribution of peak distances relative to the transcription start site (TSS) of the nearest gene. Peaks were combined in 100 bp bins.

A significant fraction of TR4 binding sites was shared among cell types (Figure 8B). For example, out of the 537 TR4 binding sites in K562 cells, 504 (94%) are also occupied in HeLa cells, 471 (88%) are also bound in HepG2 cells and 406 (76%) are also bound in GM12878 cells. When comparing 1,157 TR4 binding sites from HeLa with 1,732 from HepG2 cells, we found 922 (80%) were shared TR4 target sites. We next matched the TR4 peaks to the nearest gene. In some cases more than one peak matched to a given gene. As a consequence, the number of TR4 binding sites is slightly higher than the number of target genes. We compared 1,135 TR4 target genes from HeLa, 535 from K562, 530 from GM12878 and 1,688 from HepG2 cells (Figure 10). 532 target genes were shared in at least 3 cell types and 332 target genes were shared among all four cell types. While blood cells shared most of their TR4 targets, liver cells contained the largest number of unique target genes. TR4 may regulate genes important for basic biological processes shared in multiple cell types, while it may play an additional role in regulating cell type specific genes.

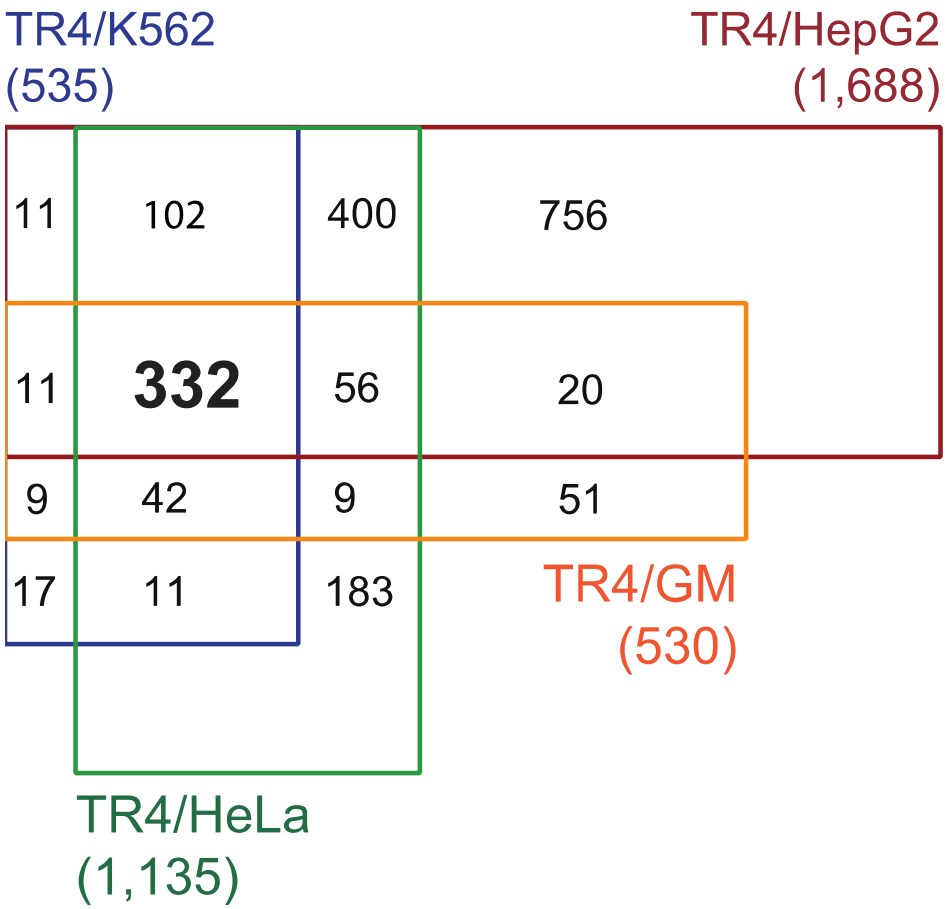


Figure 10. Overlap of TR4 target genes in 4 cell types. A target gene is defined as the nearest gene to a ChIP-seq peak. In some cases a target gene was contained more than one peak. Genome- wide TR4 ChIP-seq has identified 535 target genes in K562, 1,688 in HepG2, 1,135 in HeLa, and 530 in GM12878 cells within ± 1 kb of a transcription start site. 332 genes are identified as common targets in the 4 cell types.

ii. TR4 target genes are involved in fundamental biological processes

As shown above, the majority of TR4 targets are shared between different cell types. To shed light on the common function of genes targeted by TR4, Gene Ontology analysis was performed using ConceptGen (http://conceptgen.ncibi.org/core/conceptGen/index.jsp; (Sartor, et al., 2010)) to identify the functional categories enriched in the overlapping targets in 4 cell types (p-value < 0.05, modified Fisher's exact test). All Entrez Genes were used as

background to determine the significance of over-representation. Categories of TR4 target genes are highly enriched in fundamental biological processes, such as RNA metabolism and protein translation (ribosome) (Figure 11A). In addition, TR4 may also regulate cell type-specific genes. To test this hypothesis, we performed Gene Ontology analysis on genes found in only one cell type. The number of unique target genes in K562, HeLa, and GM12878 cells was not sufficient to perform meaningful Gene Ontology analysis. However when 756 TR4 target genes specific to HepG2 cells were analyzed, we found some unique functional categories (Figure 11B). HepG2 specific target genes were significantly enriched for ubiquitin cycle, nucleosome, chromatin assembly and metabolic processes, particularly those involving organic acid, carbohydrates, and lipids. Interestingly, a few previous studies have suggested a role for TR4 in gluconeogenesis (Liu, et al., 2007). Furthermore, TR4 may exert its function by sensing lipids and the presence of fatty acids was found to enhance cofactor recruitment to TR4 (Tsai, et al., 2009) suggesting an important role for lipids in TR4 function.

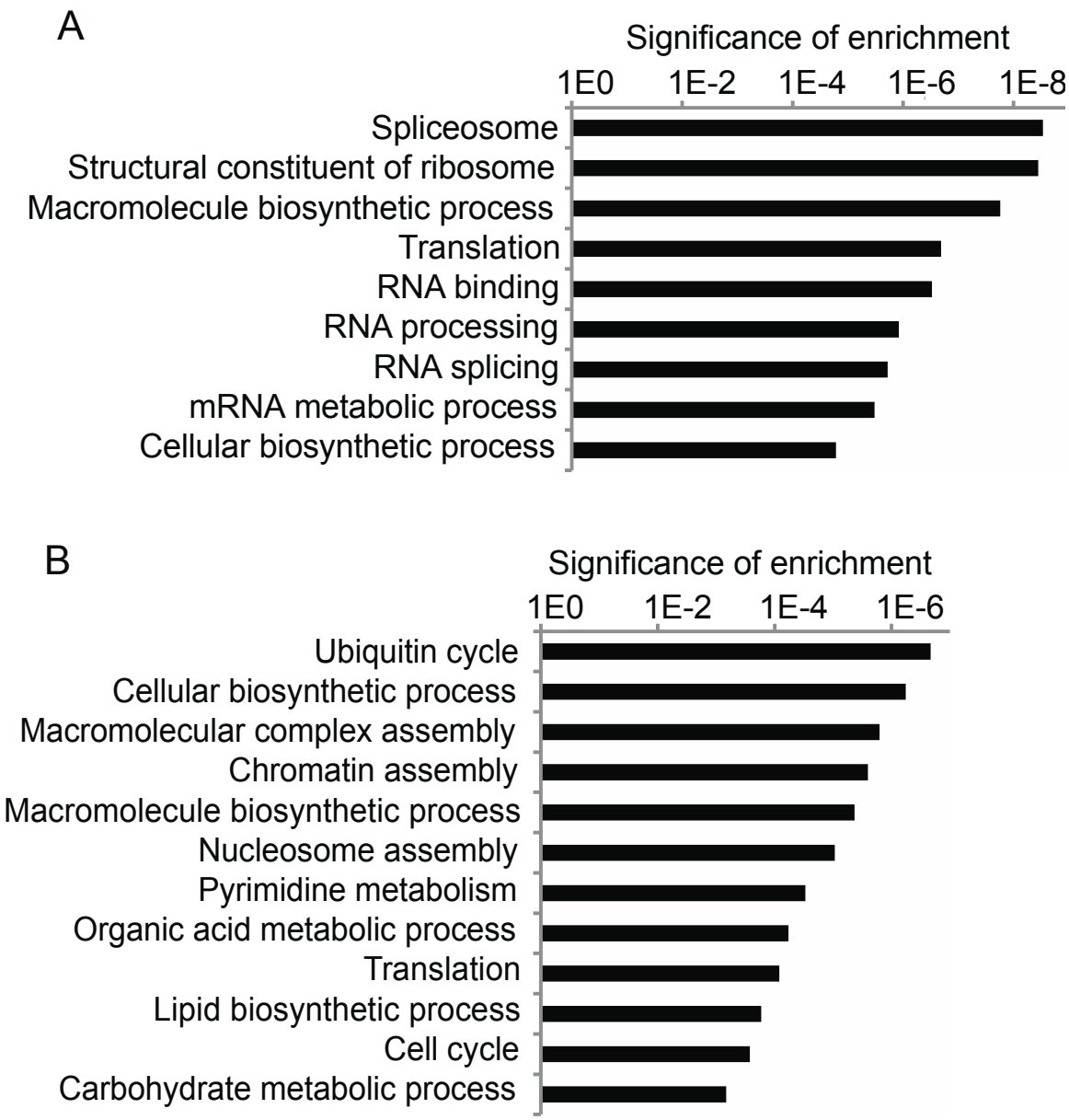


Figure 11. Functional enrichment analysis of TR4 target genes. (A) Targets common to all 4 cell types and (B) targets unique to HepG2 cells. Significantly enriched gene ontology terms for biological processes are shown on the y axis; the x axis represents p-values for each enriched category.

In recent years it has become evident that transcription factors often play dual roles, affecting activation as well as repression of target genes. Previous studies have implicated TR4 in both activation and repression of cellular target genes (Lee, et al.,

2002). TR4 binds to DNA as a homodimer, but preferentially forms heterodimers with the orphan receptor TR2 (Lee, et al., 1998). Recently, a global atlas for transcription factor networks has been assembled based on physical protein-protein interactions using mammalian two hybrid data (Ravasi, et al., 2010). This study identified TR4 (NR2C2), Nuclear Receptor Interacting Protein 1 (NRIP1) (RIP140), and histone deacetylases HDAC3 and HDAC4 as proteins interacting with TR2 (NR2C1). NRIP1 may function as a corepressor or coactivator depending on the interacting protein (White, et al., 2008). Furthermore, post translational modifications of TR4 influence its interaction with cofactors (Huq, et al., 2006). Phosphorylation of TR4 is accomplished by MAP kinases and results in recruitment of NRIP1. On the other hand, dephosphorylated TR4 recruits the coactivator pCAF. We wanted to determine whether TR4 target genes are expressed or silenced. For this purpose, we matched TR4 target genes in HeLa and HepG2 cells (1,135 and 1,688 respectively) to their RNA expression values from Illumina expression arrays (Figure 12). The median expression value of TR4 target genes in HeLa and HepG2 cells (median expression value 535 and 504, respectively) is higher than the median expression value of all genes from the HepG2 expression array (median expression value 219). TR4 target genes are also expressed at higher levels than a set of 3000 randomly selected genes from the HepG2 expression array (median expression value 228). Based on RNA expression analysis, TR4 target genes are generally expressed.

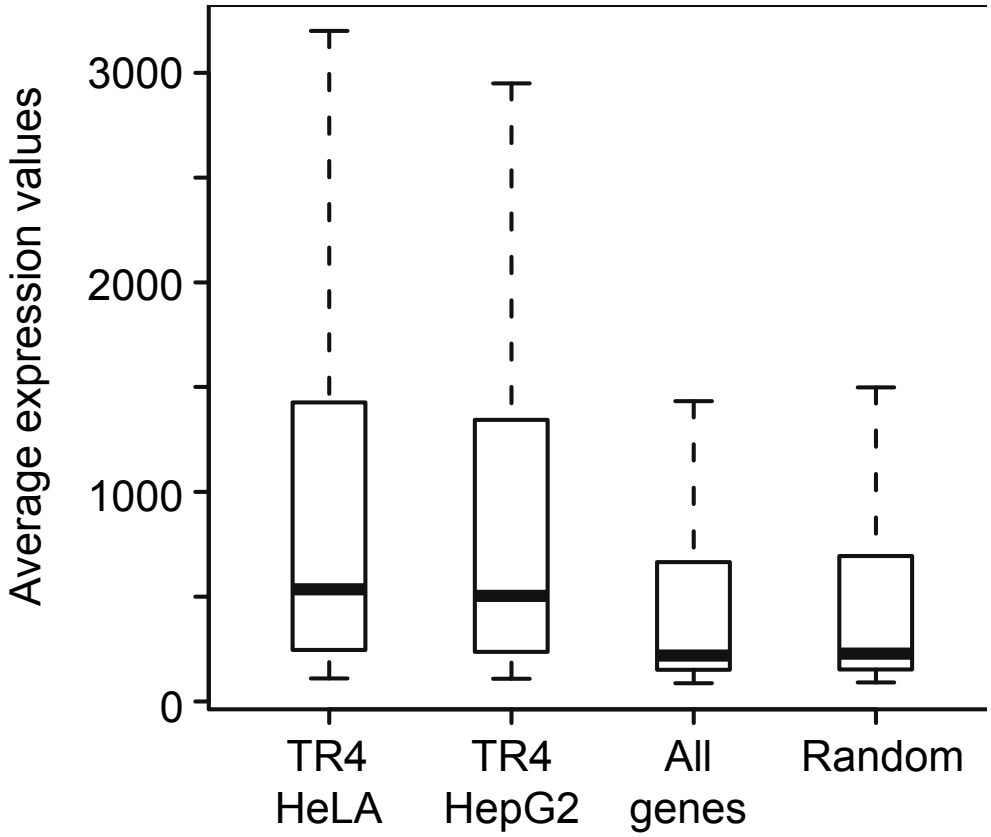


Figure 12. Expression analysis of TR4 target genes. Box-and- whisker diagrams show the range of expression values of TR4 bound genes in HeLa and HepG2 cells in comparison to expression values of all genes present on the HepG2 expression array and to the set of 3000 randomly selected genes. Expression values are plotted on the y axis. The central line in the box-and-whisker plots shows the position of the median, the upper and lower boundaries of the box represent the location of the upper (75th percentile) and the lower (25th percentile) quartiles, respectively. Data outliers are not shown.

The correlation between TR4 binding and expression of target genes suggests that TR4 binds to open accessible chromatin regions. To test this hypothesis, we examined the epigenetic signature at TR4 binding sites using ChIP-seq data of various histone marks in K562 cells. Overlap of TR4 binding sites with histone marks typical for open and repressed chromatin was determined using the gffOverlap tool from Sole-search (http://chipseq.genomecenter.ucdavis.edu/cgi-

bin/chipseq.cgi; (Blahnik, et al., 2010)). A distance of 200 base pairs between peaks was allowed to take nucleosome positioning into account. A remarkable 534 of the 537 TR4 target sites in K562 cells were also occupied by H3K4me3, which is a mark for accessible chromatin. No significant overlap with the repressive chromatin marks H3K27me3 or H3K9me3 was found (2 and 5 peaks, respectively). It has been shown in yeast and also human cells that transcription factors often bind in the linker region between nucleosomes (Lee, et al., 2007; Park, 2009). To determine whether TR4 binding occurs in nucleosome depleted regions, we analyzed sequence tag density for TR4 and H3K4me3 binding relative to the transcription start sites (Figure 13). TR4 binding was highest within 100 base pairs upstream of the TSS while the histone mark H3K4me3 is lowest in this region and reaching maximum where TR4 binding tails off, suggesting predisposition of TR4 binding sites to the linker region.

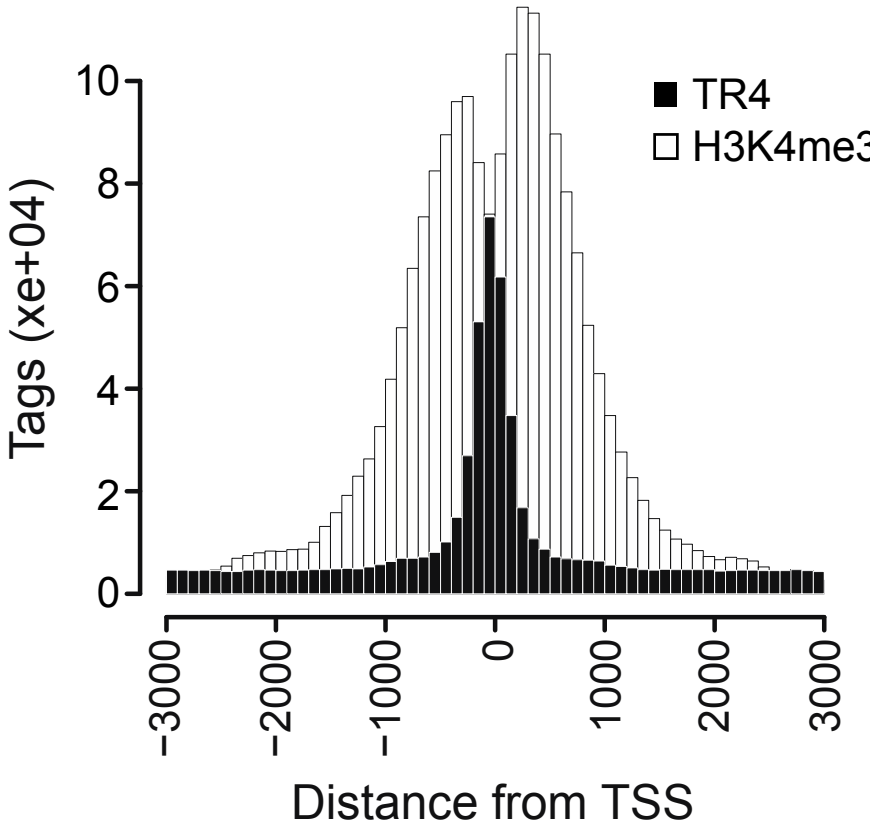
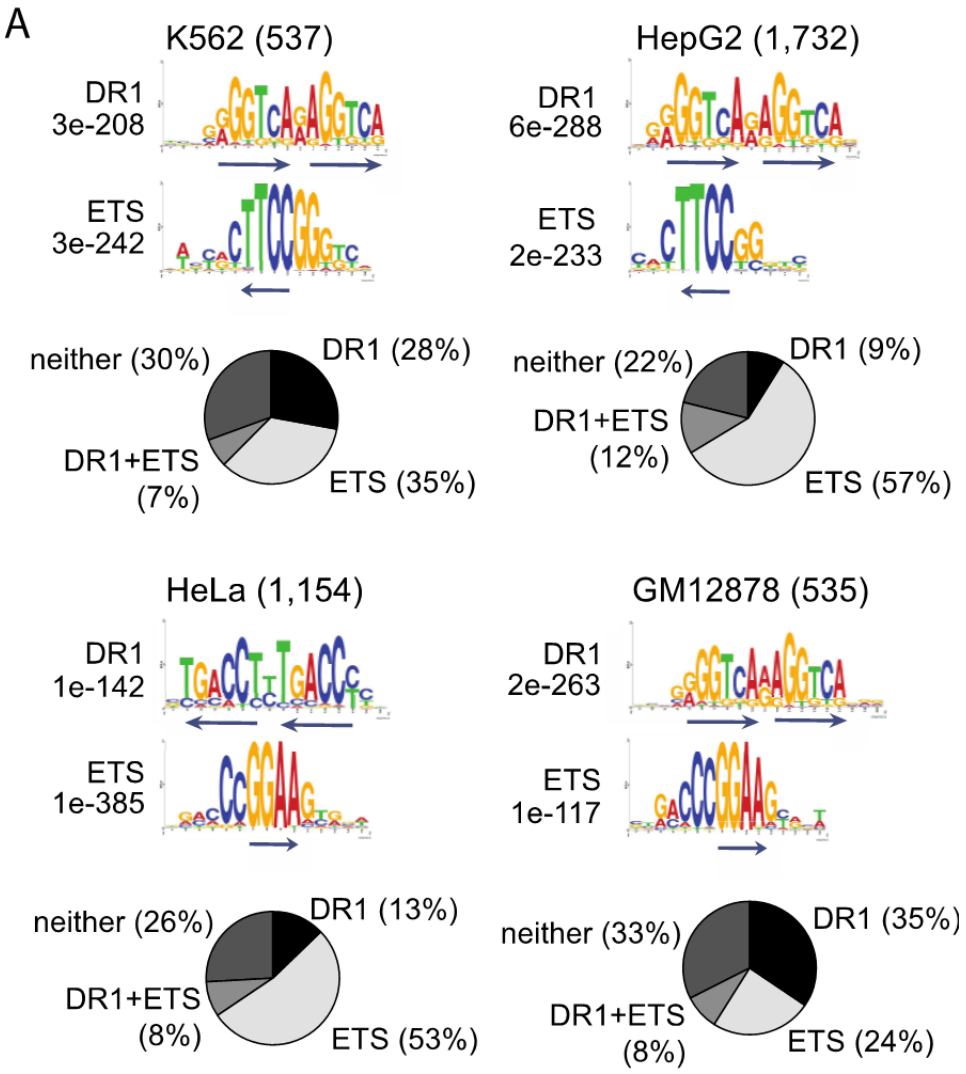


Figure 13. TR4 binding relative to nucleosomes. Positions of the histone mark H3K4me3 and TR4 occupancy are plotted for the 735 genes bound by TR4 in K562 cells. Sequence tags in bins of 100 base pairs are plotted on the y axis; distance to transcription start site is shown on the x axis.

iii. Motif analysis suggests the importance of ETS family members in TR4 action

In vitro experiments have shown that TR4 binds to the direct repeat (DR) of AGGTCA, which is the consensus binding site for a number of nuclear hormone receptors including estrogen receptor alpha and PPAR. Further studies have indicated that TR4 can bind to direct repeats separated by zero to five nucleotides (DR0 - DR5) (Kim, et al., 2005; Lee, et al., 1997; Lee, et al., 1998; Tanabe, et al., 2002). However, all previous studies were performed using *in vitro* assays. We used

the *de novo* motif discovery program MEME to identify motifs overrepresented in TR4 binding sites to determine if TR4 has the same specificity *in vivo*. To allow identification of DR elements and its spacing and flanking nucleotides, the minimum motif length was set between 12 (length of two half sites with no spacing in between) and 20 nucleotides (length of two half sites with up to 8 nucleotides in between). The canonical DR motif with one nucleotide spacing (DR1) was significantly overrepresented in all four cell types with the preferred spacing nucleotide being an A or G (Figure 14A). The canonical DR1 motif accounts for about 150 TR4 binding sites (28% in K562, 9% in HepG2, 13% in HeLa, and 35% in GM12878 cells). Interestingly, the % of peaks having a DR1 motif is much higher in the blood cell lines (K562 and GM12878) than in the other two cell types. The lack of the DR1 motif in the remaining peaks may indicate that TR4 associates with some sites only indirectly by binding to a different transcription factor.


A
K562 (537)
DR1
3e-208
ETS
3e-242
neither (30%)
DR1 (28%)
DR1+ETS
(7%)
ETS (35%)
HepG2 (1,732)
DR1
6e-288
ETS
2e-233
neither (22%)
DR1 (9%)
DR1+ETS
(12%)
ETS (57%)
HeLa (1,154)
DR1
1e-142
ETS
1e-385
neither (26%)
DR1 (13%)
DR1+ETS
(8%)
ETS (53%)
GM12878 (535)
DR1
2e-263
ETS
1e-117
neither (33%)
DR1 (35%)
DR1+ETS
(8%)
ETS (24%)

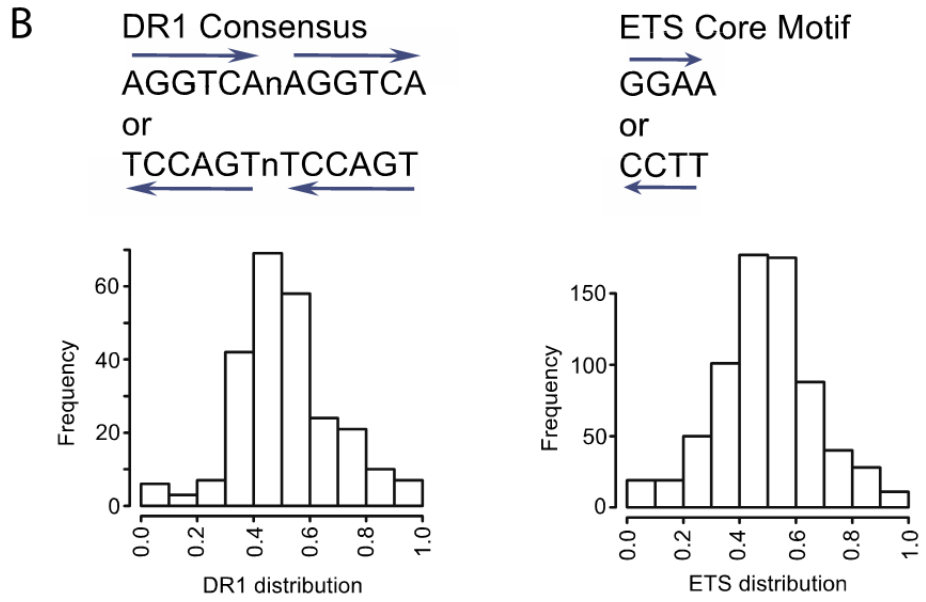

B
DR1 Consensus
AGGTCAnAGGTCA
or
TCCAGTnTCCAGT
ETS Core Motif
GGAA
or
CCTT
Frequency
0
20
40
60
0.0
0.2
0.4
0.6
0.8
1.0
DR1 distribution
Frequency
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
ETS distribution

Figure 14. Motif analysis of TR4 binding sites. (A) Sequences for TR4 binding sites located within 1 kb upstream and downstream of a TSS were retrieved. Significantly overrepresented motifs within TR4 binding sites were identified by MEME. The number of targets is indicated in parenthesis. E-values indicate significance of a given motif. Pie charts show occurrence of DR1 alone, ETS alone, DR1 and ETS, and neither of these motifs within TR4 binding sites. (B) DR1 motif and ETS core motif are depicted in either orientation. Occurrence of DR1 and ETS motifs relative to TR4 peak center in HeLa cells is shown in a histogram. Peak frequency is plotted along the y axis; distance from the peak center is plotted on the x axis. Similar results were obtained with the other 3 cell types, histograms are not shown.

Transcription factors often regulate expression of nearby genes in combination with other transcription factors through complex *cis* regulatory modules (Jin, et al., 2006). Our initial motif analysis revealed the significant recurrence of an ETS motif in addition to the DR1 element. Members of the ETS transcription factor family such as ELK4, E74A, and GABPA recognize the ETS core motif GGAA. Using 13,010 human promoter sequences, the ETS motif has been identified as one of those motifs exhibiting statistically significant clustering near the transcription start site (FitzGerald, et al., 2004). The ETS motif was predominantly found in the promoters of genes with essential cellular functions, such as ribosomal genes, mitochondrial ribosomal genes, basal transcription factor genes and proteosomal genes. The ETS motif is not only found at genes regulating similar processes as TR4 target genes, but also preferentially occurs 100 base pairs upstream of a transcription start site. The ETS motif occurs in a significant portion of TR4 binding sites (35% in K562, 57% in HepG2, 53% in HeLa, and 24% in GM12878 cells). Only about 10% of target genes contain both the DR1 and the ETS motif (Figure 14A). Combining both motifs can account for 67-78% of TR4 peaks (70% in K562, 78% in HepG2, 74% in HeLa, and 67% in GM12878 cells) suggesting a combinatorial role for ETS family members

in TR4 function. Similar results were obtained using other *de novo* motif discovery programs such as NHR-Scan (Sandelin and Wasserman, 2005) and W-ChIPMotifs (Jin, et al., 2009).

It has been postulated that the true binding site for transcription factors should be located under the center of the peak (Valouev, et al., 2008). We analyzed the distribution of both motifs relative to the center of the TR4 binding sites and found that the DR1 as well as the ETS motif are located under the peak center (Figure 14B). The close proximity of these binding sites suggests a *cis* regulatory network involving TR4 and ETS family members.

iv. ETS transcription factor ELK4 co-occupies TR4 target sites

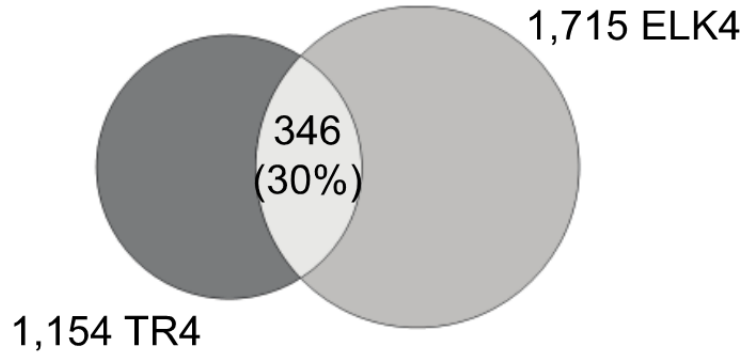
We wanted to test the hypothesis that TR4 and a member of the ETS family co-localize with TR4 *in vivo* using ChIP-seq. Motif analysis implicates the ETS family, but does not provide information as to which family member might bind to TR4 target sites. There is a high degree of functional redundancy between different members of the ETS transcription factors. Comparison of ELK1 and GABPA binding regions revealed redundant as well as unique targets between the two ETS family members (Boros, et al., 2009; Boros, et al., 2009). It has also been shown that ETS transcription factors interact with other transcription factors to regulate gene expression. For example, ELK1 is thought to function through cooperation with the serum response factor SRF (Boros, et al., 2009; O'Donnell, et al., 2008). ChIP-chip analysis showed that 22% of all ELK1 binding regions were also bound by SRF,

while the majority of ELK1 targets is SRF-independent.

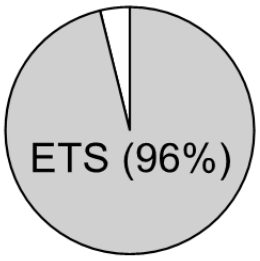
To explore the possibility that ETS transcription factors might cooperate with TR4, we performed ChIP-seq analysis of ELK1 as well as ELK4 in HeLa cells and binding sites were determined using Sole-search. 2,312 ELK4 peaks were identified from 21 million reads and 702 ELK1 peaks were identified from 13 million reads, with 86% of the ELK1 sites also being ELK4 binding sites. When we compared the 1,135 TR4 targets present within 1 kb of a TSS with 1,715 ELK4 targets found within 1 kb of a TSS, a significant overlap of 30% was observed (Figure 15A; see Figure 16A for ChIP-seq binding pattern). To identify the motifs utilized for TR4 recruitment at the 346 TR4 binding sites that are also occupied by ELK4, we performed motif analysis using MEME. The ETS motif was highly overrepresented (E-value 3.3e-310), while the DR1 motif was not (E-value 2.5e+4) (Figure 15B). We have thus identified a TR4-ELK4 *cis* module that accounts for 30% of TR4 binding sites. These sites are characterized by overrepresentation of the ETS motif in 96% of the sites and the lack of a DR1 element typically thought to recruit TR4. Therefore, TR4 does not directly bind to DNA via a DR1 element at these sites, but appears to be recruited through an ETS factor. We also analyzed the localization of binding relative to gene structure and found that TR4 and ELK4 display very similar patterns, with maximum binding between 500 bp upstream and downstream of a transcription start site (Figure 15C). The occurrence of both factors at common binding sites was confirmed by quantitative PCR using independent biological replicates (Figure 16B). Although we experimentally identified a *cis* regulatory module involving ELK4 at

~30% of TR4 binding sites, the ETS core motif was identified using bioinformatics to be within 53% of TR4 binding regions. It is possible that other ETS family members occupy these sites. It has been shown that the ETS family members ELK and GABPA shared half of their binding sites, while the other half were specific for a particular ETS factor (Boros, et al., 2009). Although further studies are needed, it is possible that ELK4 facilitates TR4 binding to promoter regions that do not contain the DR1 motif, suggesting the presence of ELK4 dependent and ELK4 independent modes of TR4 action (Figure 17).

A



B 346 sites bound by TR4 and ELK4
ETS motif: E-value 3e-310



C

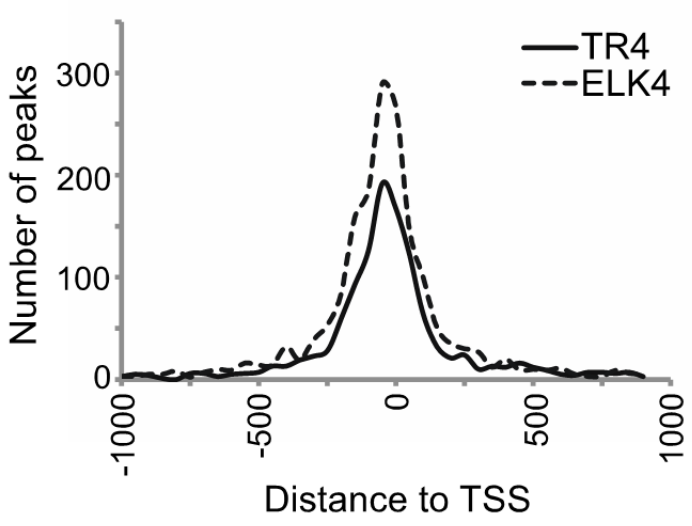


Figure 15. Overlap of TR4 and ELK4 binding sites in HeLa cells. (A) Venn diagram shows the overlap of TR4 and ELK4 binding sites within ± 1 kb of transcription start site. (B) Motif analysis was performed on the 346 sites bound by both factors; the overrepresented ETS motif is shown. Pie chart shows the occurrence of the DR1 motif, ETS motif and neither of these motifs. (C) Histogram shows binding of TR4 and ELK4 relative to the transcription start sites. Binding sites were binned into 50 base pair bins. Number of peaks is shown on the y axis; distance relative to transcription start site is plotted on the x axis.

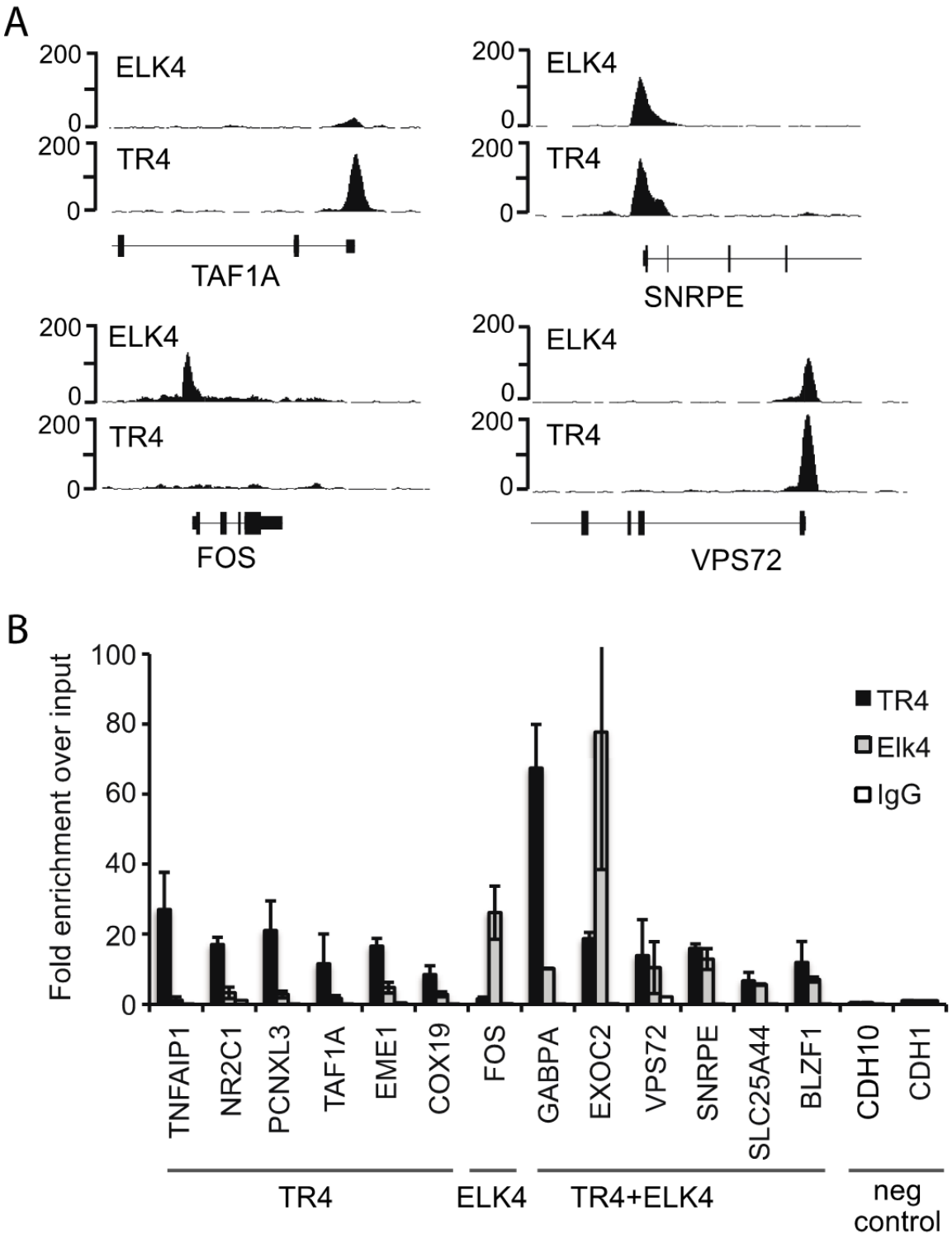


Figure 16. TR4 and ELK4 bind to common target genes. (A) ChIP- seq signal track of TR4 and ELK4 enrichment at common and unique target sites in HeLa cells. TAF1A promoter region is bound by TR4 only; C-FOS promoter region is occupied by ELK4 only, while EXOC2, SNRPE and VPS72 gene promoters are occupied by TR4 and ELK4. Number of sequence tags representing enrichment is plotted on the y axis. (B) ChIP validation of TR4

and ELK4 binding sites using qPCR. Relative enrichment was calculated over input DNA and plotted on the y axis. Each data point represents the average of triplicate ChIP experiments. Rabbit IgG was used as a non-specific control ChIP. Promoter regions tested for ChIP enrichment are shown on the x axis. The C-FOS promoter region is used as a positive control for ELK4 binding, CDH1 and CDH10 promoter regions were used as negative control regions for both, TR4 and ELK4 binding.

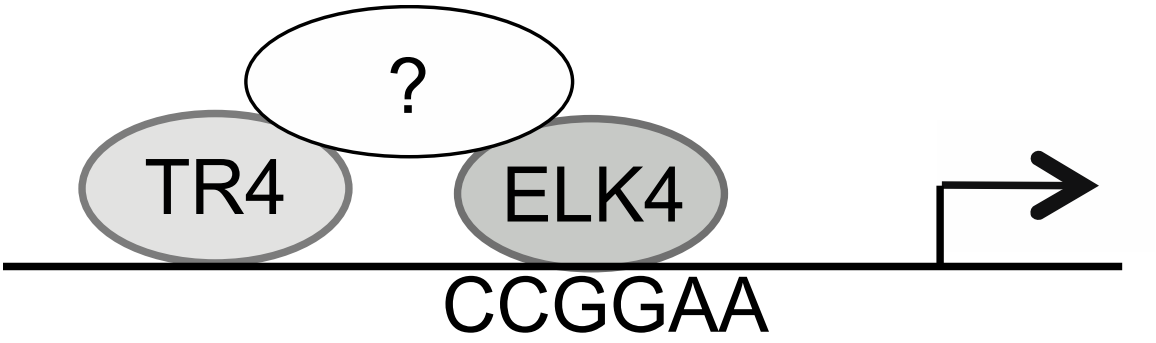


Figure 17. Model of TR4-ELK4 cis module. Gene promoters bound by both transcription factors, TR4 and ELK4, lack the DR1 element, but contain an ETS motif. This suggests that TR4 binding at these sites is facilitated through an ETS family member such as ELK4, possibly with the help of a bridging protein. TR4 may then augment ELK4 binding through non-specific DNA association, as depicted, or by serving as a non-DNA binding scaffold for additional accessory proteins.

C. CONCLUSIONS

While it had been established that TR4 plays a critical role in embryonic development, differentiation and lipid metabolism, the modes by which it functions were previously unclear. To obtain a better understanding of the TR4 modes of action, we used ChIP-seq technology to identify TR4 target genes *in vivo* in multiple cell lines. This allowed us to confirm TR4 binding *in vivo* to the direct repeat of AGGTCA separated by one nucleotide (also known as a DR1 element) at endogenous target sites in all four cell types examined. Using *de novo* motif discovery, we found that the ETS motif CCGGAA was significantly overrepresented in TR4 binding sites, suggesting a role for ETS family members in TR4 action. To confirm the co-occurrence of these two factors *in vivo*, we performed ChIP-seq for the ETS

transcription factor ELK4 and we found that about one third of TR4 target sites were indeed bound by ELK4. Sites that are bound by both factors contain an ETS motif, but lack the DR1 element typically thought to recruit TR4. These data suggest that TR4 may regulate specific subsets of target genes through ETS dependent as well as ETS independent pathways. Future studies will focus on the interdependence of these two transcription factors. Thus our approach of defining genome-wide binding patterns for a factor, followed by motif analysis to suggest possible *cis* modules, and then genome-wide analysis of the putative co-localizing factor has worked well to identify a TR4-ELK4 *cis* module.

Interestingly, we identified TR4 target genes that are common to quite diverse cell types (representatives of blood, liver, and epidermal cells). These genes were involved in fundamental biological processes such as RNA metabolism and protein translation. In addition, TR4 also binds near genes that are highly cell type-specific. For example, in HepG2 cells TR4 binds near genes that are involved in organic acid, lipid and carbohydrate metabolism. TR4 knockout mice show insulin hypersensitivity (Liu, et al., 2007) and TR4 can be induced by certain essential fatty acids resulting in TR4 activation followed by the up-regulation of the apolipoprotein E precursor (ApoE) and cytosolic phosphoenolpyruvate carboxykinase 1 PEPCK gene (Huq, et al., 2006), which is thought to contribute to diabetics-induced hyperglycemia (Gomez-Valades, et al., 2006; Valera, et al., 1994). Knowing the direct TR4 binding sites, it will be an interesting focus of future studies to evaluate the pathways underlying TR4 action and its possible role in metabolic diseases.

D. METHODS

i. Cell culture and crosslinking

K562, HeLa, HepG2, and GM12878 cells for ChIP-seq were grown and crosslinked by the National Cell Culture Center (NCCC) as part of the ENCODE project. K562 and GM12878 cells were grown in RPMI supplemented with 10% fetal bovine serum (FBS), 2 mM L-Glutamine, 100 U/mL penicillin-streptomycin. HeLa and HepG2 cells were grown in DMEM medium supplemented with 10% FBS, 2 mM L-Glutamine, 100 U/mL penicillin-streptomycin. Cells were either processed for RNA isolation or crosslinked 10 minutes at a concentration of 1% formaldehyde, snap frozen and stored at -80C.

ii. Chromatin immunoprecipitation (ChIP) assay and library preparation

ChIP assays and the libraries for Illumina sequencing were prepared as described in detail in O'Geen et al. 2010 (O'Geen, et al., 2010). Briefly, chromatin from 10^8 cells was diluted with 5 volumes IP dilution buffer (50 mM Tris pH7.4, 150 mM NaCl, 1% (v/v) igepal, 0.25% (w/v) deoxycholic acid, 1 mM EDTA pH8) and incubated at 4C over night with either 50 µl of rabbit anti-TR4 antibody (Tanabe, et al., 2007). 300 µl protein A agarose beads were added for 2 hours to capture the immune complexes. Beads were washed three times with IP dilution buffer and once with phosphate-buffered saline. ChIP assays using 20 µl rabbit anti-ELK4 (Santa Cruz Biotechnology sc-13030X) or 20 µl of monoclonal rabbit anti-ELK1 (Epitomics #1277-1) were performed using StaphA cells as described on the Farnham lab web site

(http://www.genomecenter.ucdavis.edu/farnham/pdf/FarnhamLabChIP%20Protocol.pdf). For sequencing experiments, StaphA cells were only blocked with BSA and the preclearing step was omitted. After reversal of crosslinks and RNase treatment, ChIP DNA was purified and used directly for library preparation.

iii. Sequencing and data analysis

Libraries were sequenced using the Illumina GA2 platform by the DNA Technologies Core Facility at the University of California-Davis (http://genomecenter.ucdavis.edu/dna_technologies/). The ChIP-seq data has been deposited in the NCBI Gene Expression Omnibus (accession number GSE24685). In addition, all TR4 ChIP-seq data can be visualized and downloaded from the UCSC browser at http://www.genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=169984430&c=chr9&g=wgEncodeYaleChIPseq. Peaks were called using the Sole-search software with default parameters (FDR0.0001, alpha value 0.001) using sequenced libraries of matched Input DNA for each cell type (Blahnik, et al., 2010). Peak overlap analysis based on chromosomal coordinates as well as location analysis were also performed using the Sole-search software. Gene Ontology analysis was performed using ConceptGen to identify the functional categories enriched in the overlapping targets in 4 cell types. (p-value < 0.05, modified Fisher's exact test). In addition to GO terms, other concepts were tested for significant enrichment in the gene set. All Entrez Genes were used as background to determine the significance of over-representation.

iv. Motif analysis

In vivo binding sequences from TR4 peak files were retrieved from UCSC Genome Database (hg18, March 2006). Unbiased motif analysis was performed using MEME to identify statistically overrepresented motifs in the TR4 peak sequences present in 4 cell types. The following parameters were used "-dna -nmotifs 5 -mod zoops -minw 12 -maxw 20 -maxsize 2000000 -revcomp", which specify the number of motifs to search for, the zoops assumption (zero or one occurrence per peak sequence), the minimum motif length of 12 (length of a repeat element with no spacing between two half sites), the maximum motif length of 20 (length of a repeat element with 8 spacing nucleotides between two half sites), the maximum dataset size of 2,000,000 characters. Sequences were searched in forward and reverse orientation.

v. RNA preparation and Illumina expression arrays

RNA was prepared from three independent cultures of 10^6 HeLa or HepG2 cells using Invitrogen Trizol according to the manufacture's recommendations. The Illumina TotalPrep RNA amplification kit from Ambion (AMIL1791) was used to generate biotinylated, amplified RNA for hybridization with the Illumina Sentrix Expression Beadchips, HumanHt-12. The Sentrix gene expression beadchips used for this study consisted of a 12-array, 2 stripe format comprising approximately 48 k probes/array. In this collection 24,000 probes were from RefSeq sequences and 24,000 from other Genbank sequences (see http://www.illumina.com/pages.ilmn?ID=197 for more details). Arrays were

processed as per manufacturer's instructions, scanned at medium PMT settings as recommended by the manufacturer, and analyzed using Bead Studio Software v. 2.3.41. Data was normalized using the "average" method, which simply adjusts the intensities of two populations of gene expression values such that the means of the populations become equal. Relative expression values were calculated using an algorithm provided by Bead Studio. The expression array data has been deposited in the NCBI Gene Expression Omnibus (accession numbers GSE24419 for HepG2 and GSE19146 for HeLa data).

vi. ChIP assay and quantitative PCR (qPCR)

To confirm targets identified by ChIP-seq, all ChIP assays were performed using StaphA cells. 10^7 cells were used per ChIP experiment and adjusted amounts of the same antibodies and pre-immune serum (rabbit IgG) as described above. Immunoprecipitated DNA was purified and eluted in 50 µl water. 1 µl of ChIP DNA or 3 ng of Input DNA were used for qPCR analysis. Quantitative PCR experiments were performed at least in duplicates, from at least two independent ChIP assays on a Bio-Rad DNA Engine Opticon Real-Time PCR System using SYBR® Green Master PCR Mix (SIGMA) according to the manufacturer's instructions. Results were analyzed relative to input. Each target site was calculated as 2 to the power of the cycle threshold (cT) difference between input DNA and ChIP samples. Enrichments at target sites are compared to negative/unbound control regions CDH1 and CDH10.

Chapter IV

RNA-Seq Analysis of Differentiating CD34+ Cells Suggests Novel Isoforms of Erythroid Regulators

A. INTRODUCTION

Blood is a fluid tissue composed of blood cells and viscous liquid (plasma) that is replenished daily throughout the lifespan of all vertebrate animals. Blood cells exert fundamental functions to sustain human life, such as delivering nutrients and transporting oxygen and CO_2 to and from tissues, respectively, to be exchanged in the lungs. Human adult blood cells are bright red due to oxygenated hemoglobin, which is a tetramer consisting of two α- and two β-globin chains that tetramerize to form adult hemoglobin (HbA). The two adult β-globin peptides are substituted by two γ subunits in fetal hemoglobin (HbF). A missense mutation in the adult β-globin gene causes the mutant HbA (called HbS) to form long head-to-tail polymers that lead to sickled-shaped red blood cells, which are normally smooth, round biconcave discs. These sickled-shaped red blood cells are common in patients with sickle cell disease (SCD) and cause pain, vascular damage, organ morbidity and early death in these patients. Remarkably, increased fetal γ-globin synthesis in sickled red cells can inhibit sickle polymer formation and hence alleviate SCD pathophysiology (Bunn, 1997). Similarly, activating γ-globin chain expression in β-thalassemia patients (who have lower or absent β-chains) is also expected to ameliorate the

severity of the anemia. Taken together, investigating erythropoiesis in a model system that can recapitulate normal erythroid maturation can provide insights into the molecular mechanisms for lineage specification and erythroid development, and hence should be clinically beneficial for the development of therapeutics to treat the hemoglobinopathies.

Human hematopoietic progenitor cells are characterized by surface expression of the CD34 antigen, and these CD34+ progenitor cells can be derived from bone marrow, fetal liver, umbilical cord and mitogen mobilized peripheral blood samples. Human CD34+ hematopoietic progenitors can be induced to undergo erythroid differentiation *in vitro*, and hence serve as an excellent model system to study erythroid lineage maturation and differentiation (Giarratana, et al., 2005). A growing body of studies have utilized this *in vitro* differentiation system to interrogate the transcript expression profiles during erythroid differentiation using high throughput gene expression arrays (Keller, et al., 2006; Merryweather-Clarke, et al., 2011; Peller, et al., 2009; Singleton, et al., 2008; Sripichai, et al., 2009; Tondeur, et al., 2010). These studies have highlighted the validity of transcriptome analysis using the erythroid differentiation system to predict candidate regulatory factors (Keller, et al., 2006) and to identify co-regulation of genes during erythropoiesis (Keller, et al., 2006; Peller, et al., 2009). In addition, the CD34+ culture system was used to demonstrate the effect of cytokines on fetal hemoglobin expression via histone modification and transcription factor levels (Sripichai, et al., 2009). A more focused study also led to the discovery of a mutation in KLF1 that is

associated with rare blood group In(Lu) phenotype (Singleton, et al., 2008). A recent study provided the hematology research community with a rich resource examining the global erythroid gene expression profile during human erythroid progenitor differentiation (Merryweather-Clarke, et al., 2011). Characterization of the erythroid exome suggested that there exists increased alternative splicing in genes involved in cell motility and immune response (Tondeur, et al., 2010). However, array platforms are known to have the dual disadvantages of restricted dynamic range and hybridization inconsistency.

The development of deep sequencing technologies has provided an alternative, unbiased approach for interrogating the human erythroid transcriptome during terminal differentiation. Here, we take advantage of next generation sequencing to characterize the transcriptome dynamics of human hematopoietic progenitor differentiation. We demonstrate reproducibility between replicates, and identify constitutively expressed as well as differentially expressed transcripts during differentiation. We also identify novel candidate alternative isoforms of known erythroid regulators as well as novel transcripts lying within introns or intergenic regions. These data should serve as a valuable resource for further functional analysis of human erythroid differentiation.

B. METHODS

i. RNA sequencing

Two stages of human CD34+ hematopoietic progenitor cells were obtained from the

Fred Hutchinson Cancer Research Center: one stage was from G-CSF-mobilized peripheral blood from healthy adult donors and the second was from umbilical cord blood cells. These adult or fetal cells, respectively, were cultured *ex vivo* and induced to differentiate into the erythroid lineage pathway by the addition of cytokines interleukin-3 (IL-3), stem cell factor (SCF) and erythropoietin (Epo). At Days 4, 8, 11 and 14 during differentiation, cells were collected and cDNA libraries were constructed according to instructions from Illumina. Paired-end mRNA sequencing was subsequently performed at the University of Michigan sequencing core using the Illumina GA2 platform. Each stage representing individual differentiation time points has two biological replicates.

ii. Alignments of sequence reads and evaluation of data quality

TopHat v1.3.3 (Trapnell, et al., 2009) was used to align the sequence reads to the human reference genome (version hg18) with default parameter settings. "-r 40" was used to specify that the mean inner distance between mated pairs is 40 bp. TopHat was run with and without supplying gene model annotation for the differential expression analysis without or with novel gene and transcript discovery, respectively. Pearson correlation of log2-transformed FPKM of known transcript abundance between replicates and between different samples was performed to evaluate the data reproducibility.

iii. Differential expression analysis

When performing differential expression analysis without novel gene and transcript

discovery, transcript abundance was estimated using Cuffdiff, a program included in the Cufflinks v1.2.1 software suite (Trapnell, et al., 2010), by supplying the BAM files obtained from TopHat run with gene model annotation. To identify the novel transcripts or splicing isoforms from our RNA-Seq data, we used Cufflinks v1.2.1 to take output BAM files from TopHat (without supplying gene model annotation) to assemble transcripts *de novo*. The results from replicates were then merged using Cuffmerge, a program in the Cufflinks v1.2.1 package, and then Cuffdiff was used to perform differential expression analysis.

iv. Functional analysis

Functional analysis was performed using ConceptGen (Sartor, et al., 2010) to identify the functional categories enriched in the gene clusters of interest (p-value < 0.05, modified Fisher's exact test). In addition to Gene Ontology (GO) terms, other biological concepts, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (Kanehisa, et al., 2008) and Biocarta pathways, were also tested for significant enrichment among differentially expressed genes. All Entrez Genes in the RNA-Seq analysis annotation were used as background to determine the significance of over-representation.

C. RESULTS AND DISCUSSION

i. Transcriptome dynamics during CD34+ progenitor cell differentiation

Human CD34+ hematopoietic progenitor cells derived from G-CSF-mobilized peripheral blood of healthy adult donors and from umbilical cord blood cells were

cultured and induced to undergo differentiation through the erythroid lineage with hematopoietic cytokines. Cells were then harvested after 4, 8, 11 and 14 days of differentiation. The day 4 (D4) sample is comprised primarily of proerythroblasts. At day 8 (D8), the sample consists of proerythroblasts and basophilic erythroblasts. The day 11 (D11) sample includes basophilic erythroblasts and polychromatic erythroblasts, while the day 14 (D14) sample is composed primarily of orthochromatic erythroblasts and reticulocytes. As the cells progressively go through these terminal differentiation stages, more cells become enucleated as hemoglobin synthesis increases.

Total RNA was extracted from each sample and then sequenced using next generation sequencing technology to characterize the transcriptome dynamics during hematopoietic progenitor differentiation. The total number of reads as well as the number of uniquely mapped reads obtained in each sample within each run is summarized in Tables 2 and 3. The sequence reads were subsequently mapped to human genome version hg18 with TopHat v1.3.3 either with or without genome model annotation. On average, 35 million paired-end sequence tags were obtained from each biological replicate recovered from individual differentiation stages, and 76% of these reads could be uniquely mapped to the reference genome. Transcript abundance, in FPKM units (Fragments Per Kilobase of transcript per Million mapped reads), was then estimated using the Cufflinks v1.2.1 software package. The number of transcripts with FPKM greater than or equal to 0 at each time point is summarized in Table 4.

	# of raw paired reads	# (%) of uniquely mapped read pairs
Adult D8	32,076,766	24,342,115 (75.89%)
Adult D14	35,473,072	23,528,503 (66.33%)
Adult D4	34,325,400	27,080,060 (78.89%)
Adult D8	39,051,628	30,148,239 (77.20%)
Adult D11	36,689,086	27,836,176 (75.87%)
Adult D14	32,042,933	23,740,501 (74.09%)
Fetal D8	40,334,656	31,154,719 (77.24%)
Fetal D14	38,973,369	29,155,021 (74.81%)

Table 2. Raw read summary statistics for replicate 1 (Run183).

	# of raw paired reads	# (%) of uniquely mapped read pairs
Fetal D4	37,655,738	30,202,629 (80.21%)
Fetal D8	33,470,686	26,413,939 (78.92%)
Fetal D11	35,718,063	27,303,936 (76.44%)
Fetal D14	33,738,634	25,309,549 (75.02%)
Fetal D4	35,319,610	27,413,422 (77.62%)
Fetal D11	33,318,145	25,372,665 (76.15%)
Adult D4	43,530,414	34,365,085 (78.95%)
Adult D11	34,203,044	24,341,443 (71.17%)

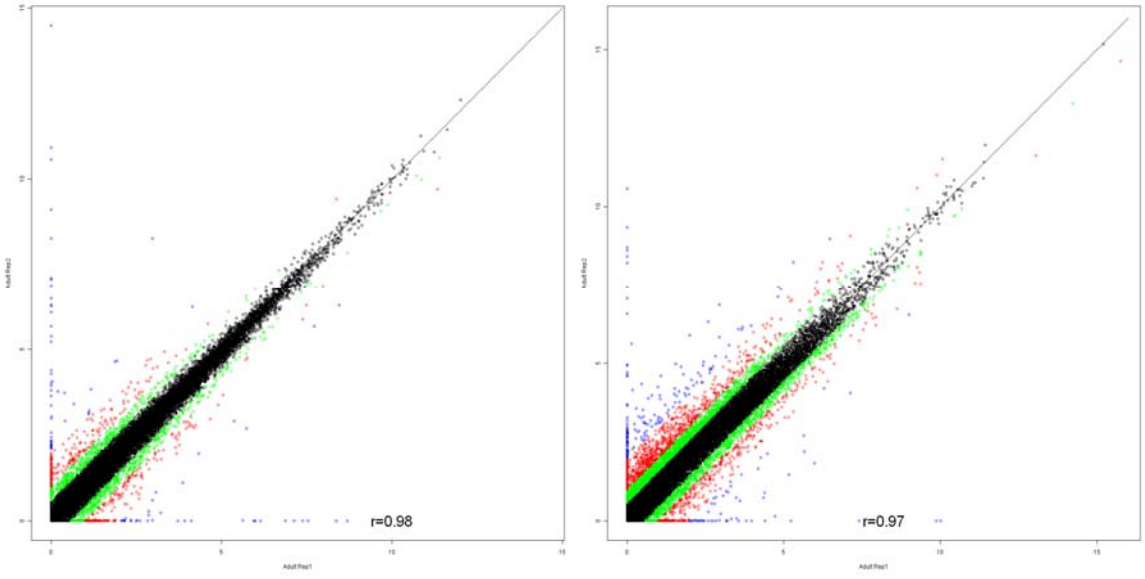
Table 3. Raw read summary statistics for replicate 2 (Run270).

	# of expressed known transcripts (FPKM > 0)	# of non-expressed known transcripts (FPKM = 0)
Adult D4	26,833	11,701

Adult D8	26,158	12,376
Adult D11	25,452	13,082
Adult D14	25,452	13,082
Total (Union)	30,140	16,611

Table 4. Number of expressed (FPKM > 0) and non-expressed (FPKM = 0) transcripts recovered from adult samples at each time point.

To evaluate the data quality in terms of reproducibility, we investigated the correlation between biological replicates and between different samples at each time point. As shown in Figures 18 and 19, Pearson correlation coefficients between biological replicates are higher than the correlation coefficients between different samples. When D4 to D8, to D11, or to D14 samples were compared, the correlation coefficients gradually decreased; however, there was no significant difference observed when comparing transcript abundance between samples at adjacent time points. One of the reasons may be that each pair of consecutive time points was too close to observe significant differences in gene expression changes.



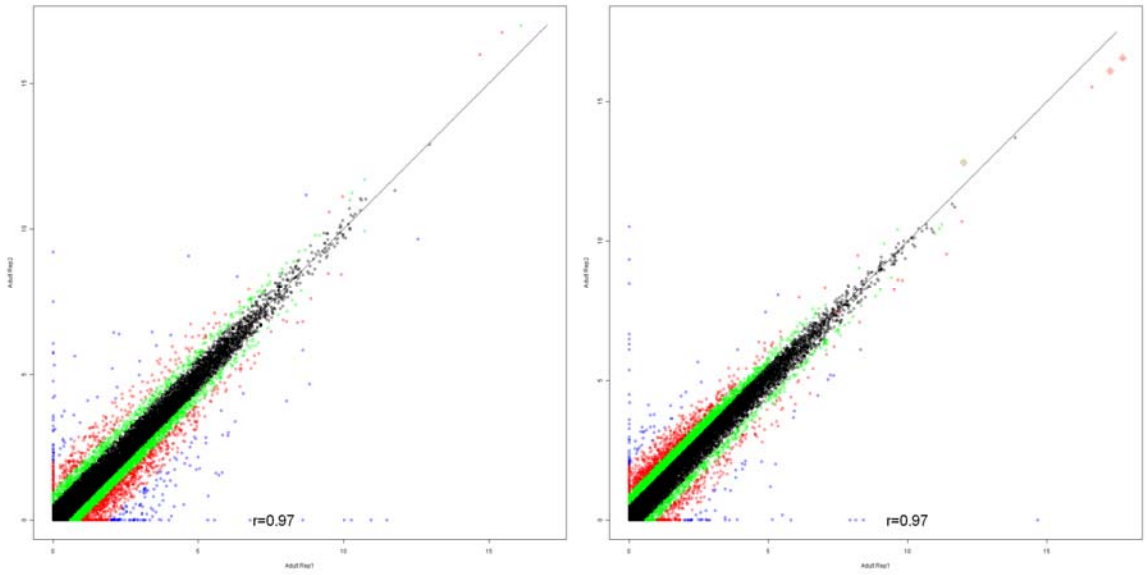


Figure 18. Correlation between biological replicates. The pearson correlation coefficients between replicates are all above 0.97. From top left, top right, bottom left, and bottom right, each represents D4, D8, D11, and D14 replicates. Green points represent transcripts with greater than 1.5 fold but less than 2 fold difference between replicates, whereas red points represent greater than 2 fold but less than 4 fold, and blue points represent greater than 4 fold differences.

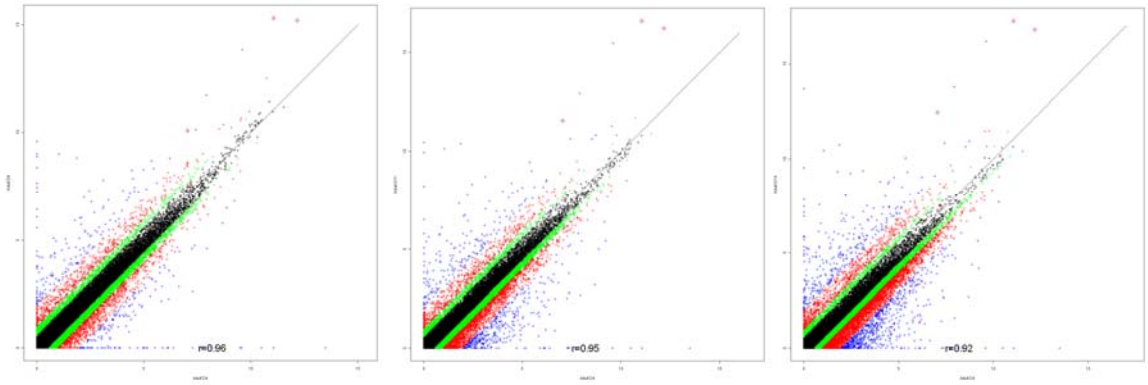


Figure 19. Correlation between D4 and D8, D4 and D11, and D4 and D14 from left to right. The pearson correlation coefficients are 0.96, 0.95, and 0.92 respectively. Green points represent transcripts with greater than 1.5 fold but less than 2 fold difference between replicates, whereas red points represent greater than 2 fold but less than 4 fold, and blue points represent greater than 4 fold differences.

So-called “housekeeping” genes are expressed at relatively constant levels and ubiquitously, as they typically perform fundamentally important cellular metabolic

functions. Similarly, we might expect crucial erythroid regulators to also be relatively abundantly expressed during erythroid progenitor differentiation. Therefore, before identifying the differential transcript expression during hematopoietic progenitor differentiation, we focused first on the transcripts that are consistently expressed at high levels at each progressive stage to identify highly expressed transcripts at each time point. Figure 20 shows the log2-transformed global transcript abundance at the four time points that were examined in this analysis. We define highly expressed transcripts as those whose log2-transformed FPKM is greater than 7, because this resulted in a reasonable number of transcripts for subsequent functional analysis. In summary, 652, 620, 446, and 536 transcripts met this cutoff at D4, D8, D11 and D14, respectively.

Figure 21 shows a heatmap reflecting the expression profile of the most highly expressed transcripts. A core group of 308 genes are consistently expressed at high level (log2(FPKM) > 7) at all 4 stages. As expected, the globin genes and GAPDH are all within this group. Given their abundant expression during erythroid differentiation, we next asked what functions this collection of genes might exert during differentiation. Functional analysis was initiated by identifying the enriched functional annotations within the gene group, which revealed that the ribosome pathway was enriched according to KEGG pathway analysis. Hemoglobin complex and oxygen transport were also among the highly enriched GO terms, as expected since members of the globin gene family are highly expressed throughout erythroid differentiation. These observations were documented in a previous study of

erythroid differentiation (Merryweather-Clarke, et al., 2011). Known erythroid transcriptional regulatory proteins such as KLF1, NFE2, GATA1, GFI1B, and LDB1, were also found to be expressed at a consistently high level. Furthermore, we observed consistently high expression of serine/arginine-rich splicing factor 2 (SRSF2), eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), and ribosomal protein S19 (RPS19) which were also reported in the same previous microarray study (Merryweather-Clarke, et al., 2011). An additional group of transcription factors were also identified among these constitutively highly expressed genes but their specific roles in erythropoiesis remain unclear. For example, the transcription factor E2F2 controls cell cycle progression and has been reported to regulate maturation and terminal cell division during erythropoiesis through GATA-1 binding (Kadri, et al., 2009). However, our data suggested that another E2F transcription factor, namely E2F4, was highly expressed throughout erythroid progenitor differentiation. E2F4 is a transcriptional regulator of the cell cycle program and was reported to play a role in fetal erythropoiesis by promoting cell cycle progression and cellular proliferation (Kinross, et al., 2006). However, its function in adult erythropoiesis remains unclear. Hence it may be worthwhile to further investigate the role of E2F4 in erythropoiesis. Regardless, all of these transcription factors (listed in Table 5) may be interesting potential candidates for further investigation.

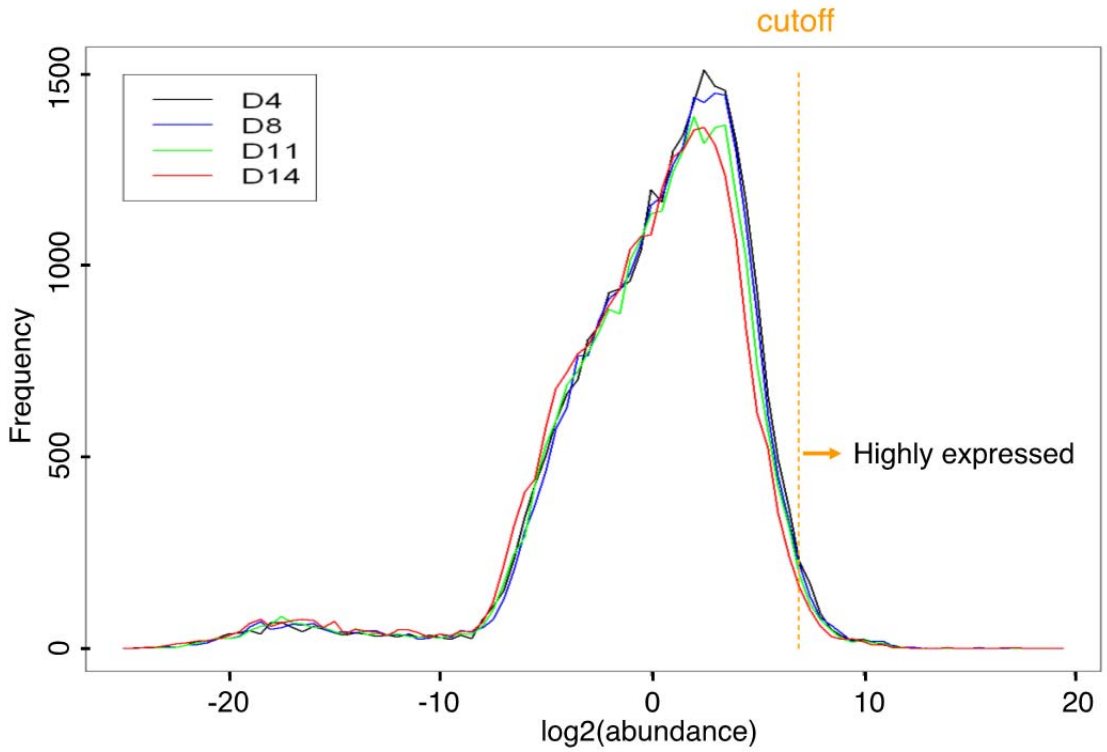


Figure 20. Global transcript abundance during erythroid progenitor differentiation. Orange dashed line indicates the cutoff where all transcripts to the right have log2(FPKM) greater than 7.

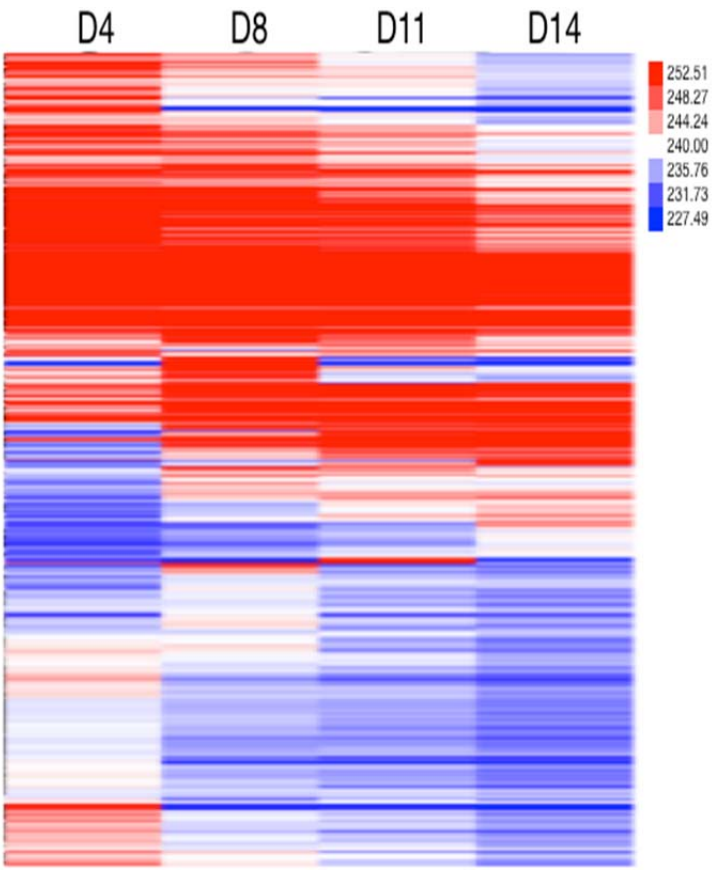


Figure 21. Heatmap showing the expression profile of the most highly expressed transcripts during erythroid progenitor differentiation. The expression values of the most highly expressed transcripts are plotted such that higher expression levels are in red and lower levels are in blue. Some transcripts may have expression level with log2(FPKM) > 7 at only a subset of time points, while others have log2(FPKM) > 7 at all 4 time points.

Gene Symbol	RefSeq ID	D4	D8	D11	D14
AES	NM_001130	181.63	206.62	211.80	172.28
ATF4	NM_182810	238.23	364.07	316.36	323.08
BTF3	NM_001207	317.23	302.66	237.00	203.16
CALR	NM_004343	1111.31	466.80	498.88	396.45
CSNK2B	NM_001320	221.52	204.89	175.19	152.47
DDX5	NM_004396	186.33	170.52	165.40	135.78
E2F4	NM_001950	142.69	174.93	231.92	208.43
EDF1	NM_003792	339.53	282.18	268.23	221.10
GATA1	NM_002049	223.04	281.74	296.64	267.09
GFI1B	NM_004188	212.52	421.92	410.84	389.79
HDGF	NM_004494	241.83	279.62	329.80	268.05
HMGA1	NM_145899	405.81	425.77	360.26	245.20

HMGB2	NM_002129	293.49	537.16	566.18	532.22
HNRNPA2B1	NM_002137	397.30	295.21	292.49	216.30
ILF2	NM_004515	281.22	176.75	176.32	129.26
KLF1	NM_006563	278.30	468.13	571.35	588.29
LDB1	NM_003893	173.69	245.55	195.99	153.05
LYL1	NM_005583	162.77	150.88	168.00	155.17
MAZ	NM_001042539	199.77	253.06	199.74	160.48
MYBL2	NM_002466	172.10	262.60	281.31	244.15
NFE2	NM_001136023	165.12	199.10	252.66	328.53
PFN1	NM_005022	1942.02	1395.24	1233.23	808.87
PHB2	NM_007273	293.31	218.63	199.62	151.52
PTMA	NM_002823	351.71	322.39	137.18	161.52
RAN	NM_006325	564.58	438.59	457.39	343.55
RUVBL2	NM_006666	262.67	251.30	213.95	142.57
SFPQ	NM_005066	242.25	190.45	183.61	131.27
TCEB2	NM_207013	200.79	183.01	169.23	151.74
THOC4	NM_005782	287.34	222.39	229.15	166.93
TRIM28	NM_005762	543.97	413.83	348.79	247.08
UXT	NM_004182	168.36	178.83	156.71	137.53
YBX1	NM_004559	730.72	567.44	606.78	461.84
YWHAH	NM_003405	130.40	189.78	182.49	160.30

Table 5. Highly expressed transcription factors in alphabetical order. A list of transcription factors was compiled from JASPAR database (Bryne, et al., 2008) and Genomatix software suite (Cartharius, et al., 2005). These transcription factors are identified in adult CD34+ progenitor cells as highly expressed at all 4 time points. This table shows the RefSeq mRNA id of the factors along with the FPKMs at each time point.

We next focused on identifying global differentially expressed transcripts during hematopoietic progenitor differentiation. Known transcripts having at least 2-fold change with a FDR < 0.05 between any two consecutive stages were chosen for subsequent detailed analysis. A total of 1806 transcripts were determined to be differentially expressed based on this criterion. Unsupervised hierarchical clustering was then performed on these transcripts to identify co-expressed gene clusters (Figure 22). From the resulting heatmap, we were able to discern transcript clusters that could be divided into six major groups. Transcript expression profiles

for individual clusters are shown in Figure 23. There are two groups of genes corresponding to up (cluster 2) and down (cluster 4) regulated transcripts within the six groups. Members of the globin transcript family fall within cluster 2, and many well known erythroid-restricted transcription factors are notably also found within the cluster (e.g. NFE2, FOXO3, EPOR, BCL6, EPB49, KLF3 and E2F2). This cluster also contains several transcription factors whose roles in erythropoiesis are largely uncharacterized and therefore serve as potentially interesting targets for further investigation. Functional annotation results suggest that oxygen transport, hemoglobin complex, apoptosis, and heme biosynthetic processes were enriched among the transcripts found in cluster 2 (Figure 24). In contrast, downregulated transcripts in cluster 4 are enriched in biological processes such as hemostasis, immune system process, regulation of body fluid level, blood coagulation, mast cell activation, and myeloid leukocyte activation (Figure 25). Cluster 4 also contains previously known negative erythroid regulators or factors whose expression diminishes during erythroid differentiation, such as c-MYB, SPI1/PU.1, FLI1, and RUNX1.

Transcription factor binding sites enriched in the promoters (defined as -450 to +50 around the transcriptional start site, TSS) of genes in each cluster were predicted using Pscan software (Zambelli, et al., 2009). The transcription factors whose binding sites are enriched within the promoter region in clusters 2, 3, and 4 are shown in Figures 26, 27 and 28, respectively. GABPA is the only transcription factor whose binding motif is enriched in the promoter regions of cluster 6 genes. Some

transcription factors whose binding sites are enriched within the promoters are only expressed at modest levels during terminal differentiation. Finally, we also performed this transcription factor binding site analysis for highly expressed transcripts (with log2(FPKM) > 7 at all 4 stages) and the results are shown in Figure 29.

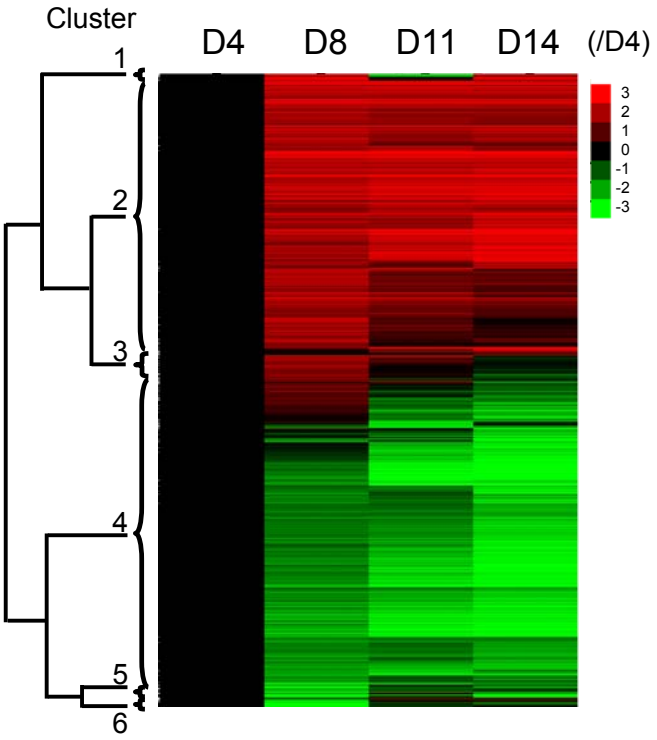


Figure 22. Heatmap showing the transcriptome dynamics during differentiation. The transcripts represented in the heatmap have at least fold change greater than 2 in any pair of consecutive time points with associated FDR less than 0.5. All transcript abundance levels are normalized to their abundance at D4. Therefore the leftmost column representing D4 is black. The red cells represent higher expression level while green cells represent lower expression levels relative to their D4 expression levels.

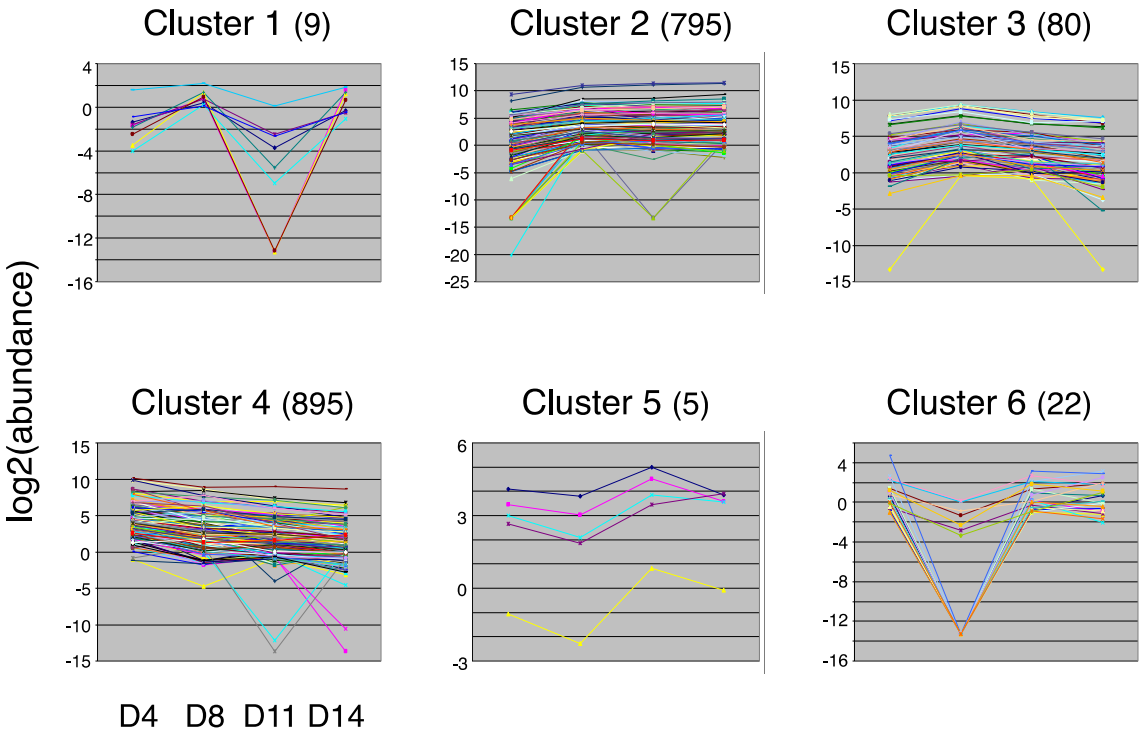


Figure 23. Differential expression pattern of each cluster. The log2-transformed abundance of transcripts in the six clusters shown in previous figures is plotted. The number of transcripts in each cluster is shown in parentheses.

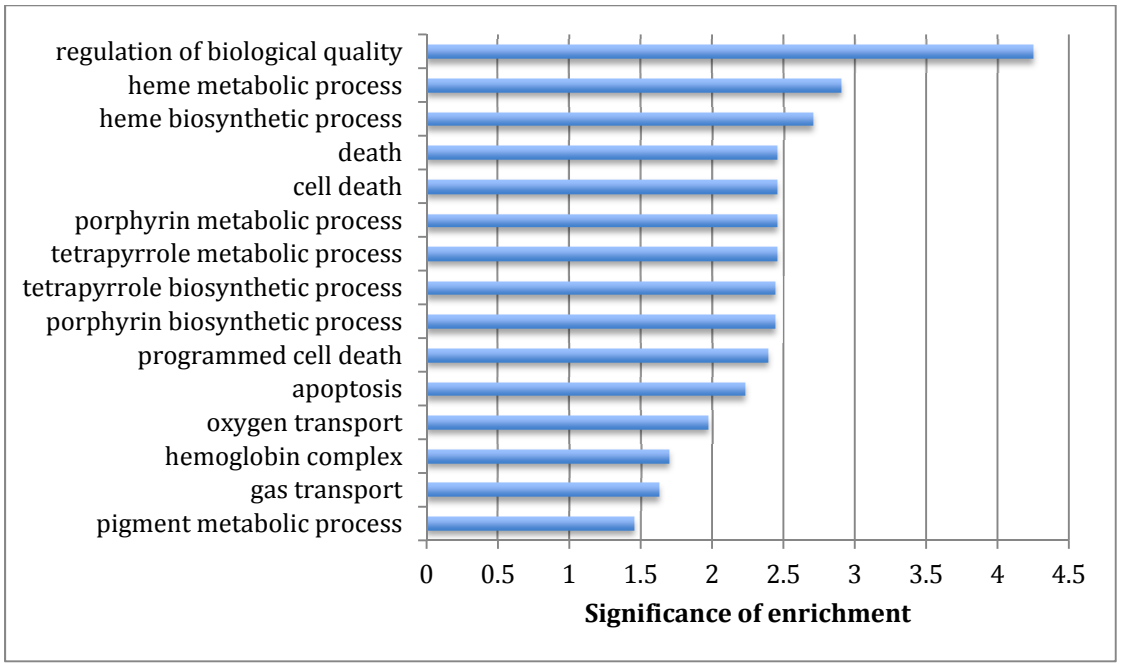


Figure 24. GO analysis of cluster 2. The top 15 enriched GO annotation terms for the genes

in cluster 2 are shown here. The significance of enrichment is defined as –log10(q-value). A q-value of 0.05 was used as the significance cutoff. These represent a cluster of transcripts whose expression levels increased during terminal erythroid differentiation.

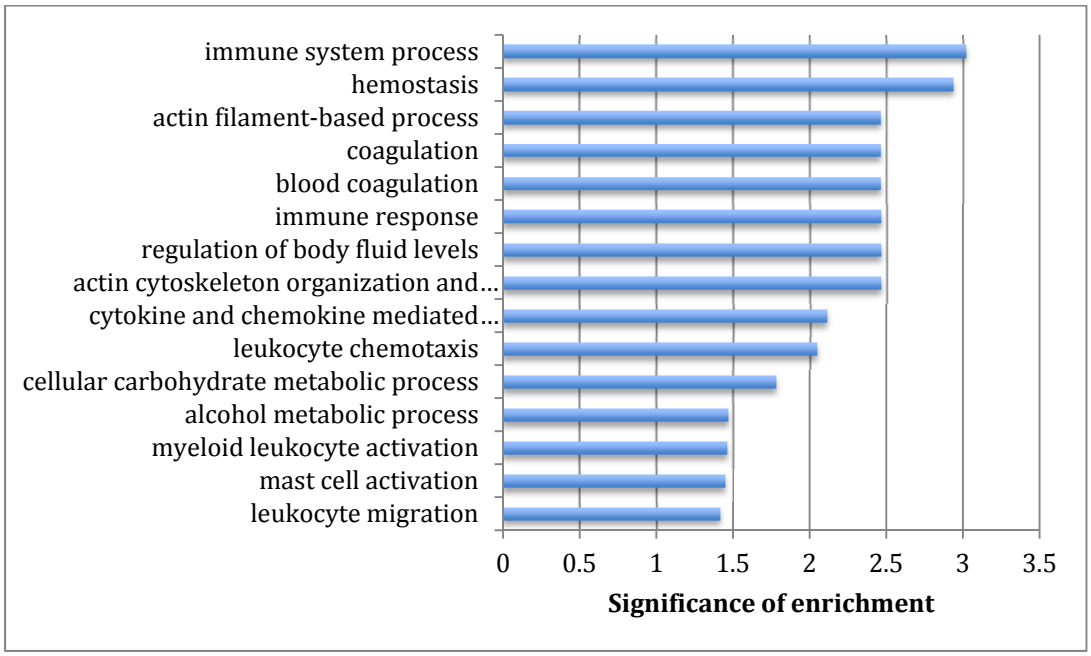


Figure 25. GO analysis of cluster 4. The top 15 enriched GO annotation terms for the genes in cluster 4 are shown here. The significance of enrichment is defined as –log10(q-value). A q-value of 0.05 was used as the significance cutoff. These represent a cluster of transcripts whose expression levels decreased during terminal erythroid differentiation.

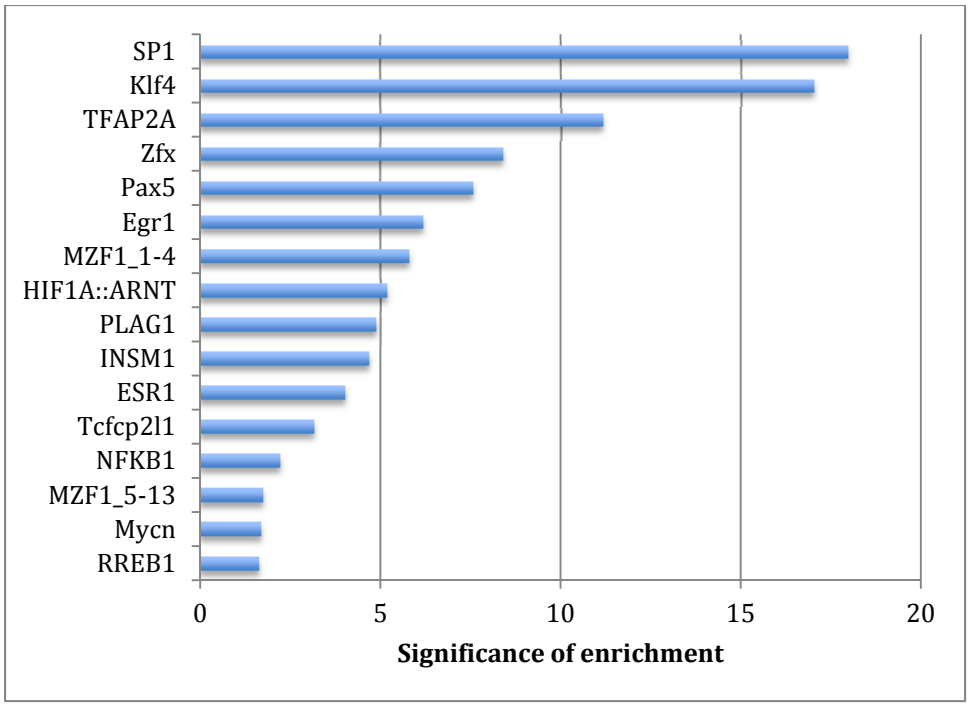


Figure 26. TFBS enrichment for genes in cluster 2. The significance of enrichment is defined as –log10(Bonferroni-corrected p-value). A significance cutoff of 0.05 was applied to the Bonferroni-corrected p-value. The transcription factors shown in this figure represent the factors whose binding sites enriched in the cluster 2 promoters.

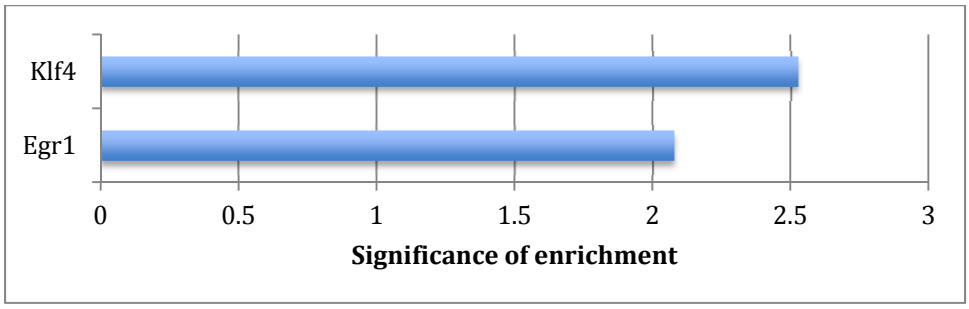


Figure 27. TFBS enrichment for genes in cluster 3. The significance of enrichment is defined as –log10(Bonferroni-corrected p-value). A significance cutoff of 0.05 was applied to the Bonferroni-corrected p-value. The transcription factors shown in this figure represent the factors whose binding sites are enriched within the cluster 3 promoters.

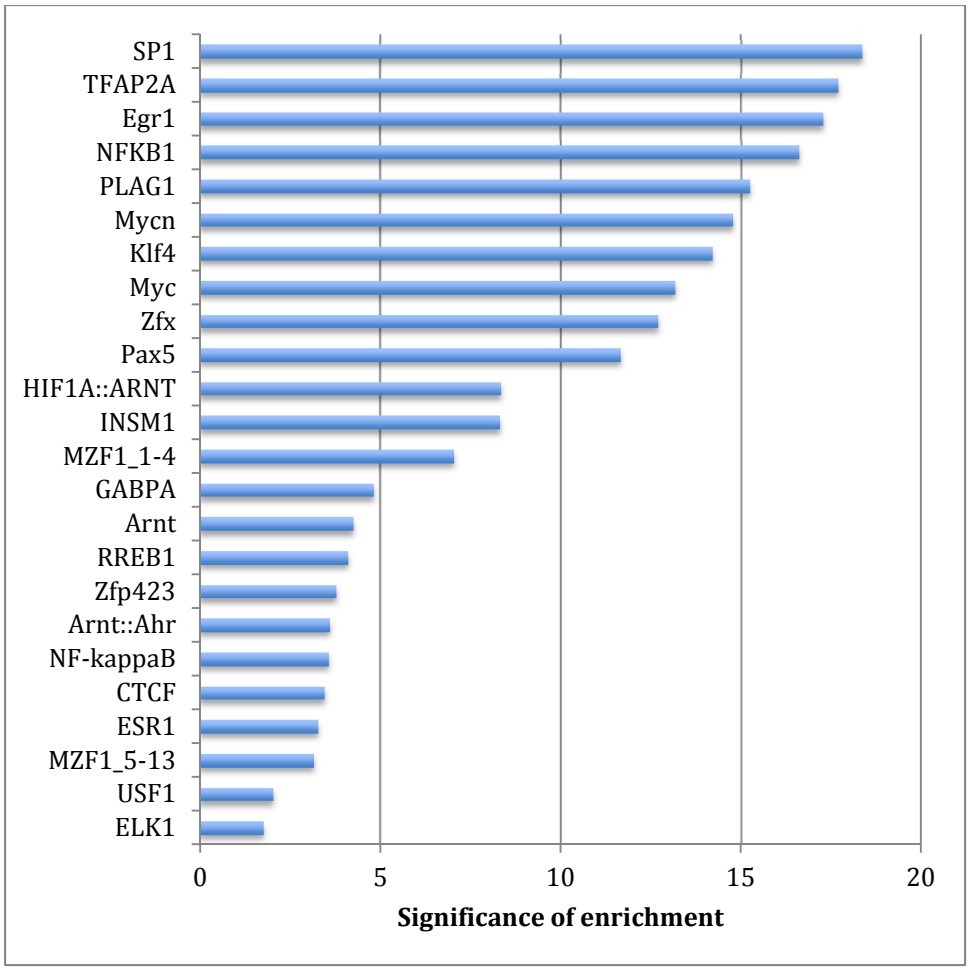


Figure 28. TFBS enrichment for genes in cluster 4. The significance of enrichment is defined as –log10(Bonferroni-corrected p-value). A significance cutoff of 0.05 was applied to the Bonferroni-corrected p-value. The transcription factors shown in this figure represent the factors whose binding sites are enriched within cluster 4 promoters.

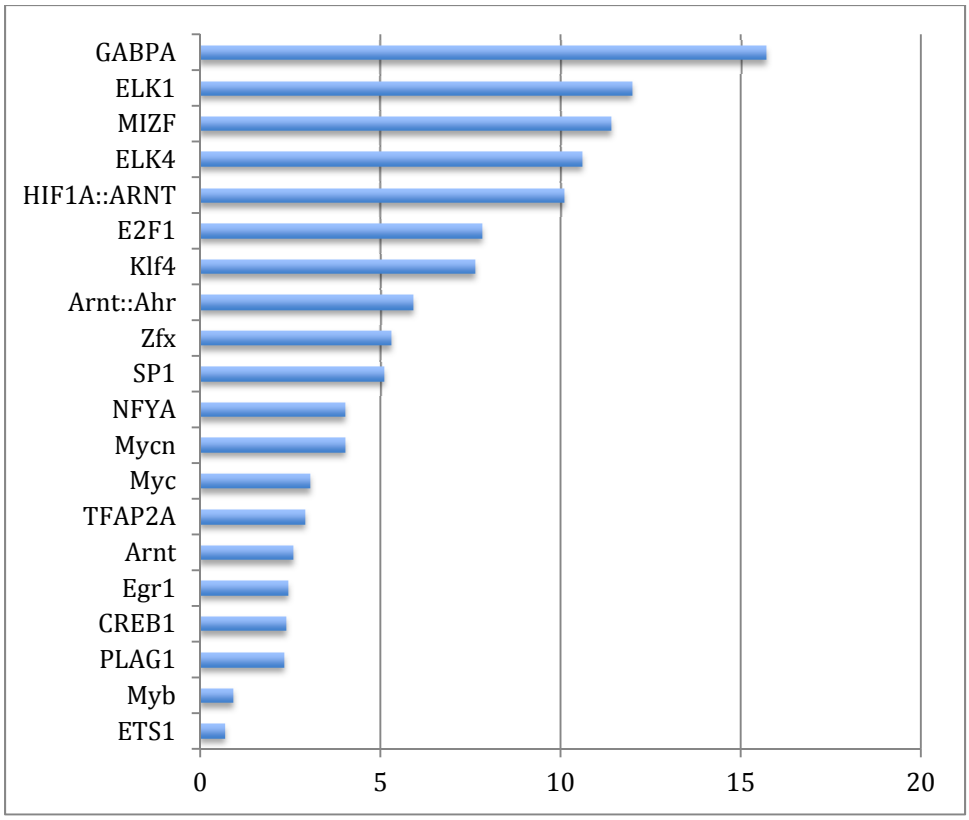


Figure 29. TFBS entichment for genes with log2(FPKM) > 7 in all 4 time points. The significance of enrichment is defined as –log10(Bonferroni-corrected p-value). A significance cutoff of 0.05 was applied to the Bonferroni-corrected p-value.

The hedgehog signaling pathway was first identified in the fruit fly as a major developmental signaling pathway decades ago (Ingham and McMahon, 2001; Nusslein-Volhard and Wieschaus, 1980). The function of the hedgehog pathway in embryonic development is well conserved across species and it has also been shown to play a critical role in cancer development, progression and metastasis (Harris, et al., 2011; Mar, et al., 2011). However, the role hedgehog plays in erythropoiesis is poorly defined and remains controversial (Lim and Matsui, 2010; Mar, et al., 2011). Smoothened (SMO) is a critical component in this signaling pathway (Lim and Matsui, 2010; Mar, et al., 2011) and we found that *SMO* expression diminishes

markedly (from FPKM of 3.34 to 0.01) during erythroid progenitor differentiation, despite the fact that it has been claimed by multiple groups that the loss of *SMO* is dispensable for definitive hematopoiesis (Dierks, et al., 2008; Gao, et al., 2009; Hofmann, et al., 2009). Indeed, one group has reported that the deletion of Smo has an effect on the recovery of spleen stress progenitors from acute anemia (Perry, et al., 2009). This observation is in keeping with the observations derived in our experimental conditions since the CD34+ progenitor cells were cultured *ex vivo* and therefore were likely already differentiating under stress conditions. We speculate that SMO may play a role in erythroid progenitor maturation or differentiation.

ii. Novel isoforms of known erythroid regulators

Previous high throughput studies have been conducted to characterize transcriptome dynamics (Keller, et al., 2006; Merryweather-Clarke, et al., 2011; Peller, et al., 2009; Singleton, et al., 2008; Sripichai, et al., 2009; Tondeur, et al., 2010). However, recently developed parallel sequencing technology made it possible for unbiased mapping and quantifying whole transcriptomes in an unprecedented manner. In addition to being capable of measuring transcript abundance in an unbiased fashion, one of the clear advantages of RNA-Seq is the ability to identify novel splicing events.

After examining the dynamic patterns of transcript abundance during differentiation in the previous section, we then focused on detection of alternative splicing isoforms. Potential novel splice isoforms are identified in this analysis as

assembled transcripts that share at least one splice junction with a known reference transcript. A total of 14,993, 14,316, 14,283, and 14,154 novel splicing isoforms are predicted from the D4, D8, D11 and D14 samples, respectively. To our surprise, among these predicted novel isoforms, 66, 36, 22, and 35 were determined to be specific to the D4, D8, D11, and D14 transcript pools, respectively.

While it was observed that splicing isoform diversity decreases during neural differentiation, this gene level splice isoform analysis was actually performed on the top 500 abundantly expressed genes and hence does not capture the global characteristics (Wu, et al., 2010). We then performed the splice junction analysis on all predicted novel isoforms from known genes, and Figure 30 shows this global view of the unique splice junction counts for each gene that changed during terminal differentiation. 220 known genes have a change in their number of unique splice junctions between any two consecutive stages and hence are deduced to be differentially expressed during differentiation. There are three known erythroid gene regulatory proteins that are predicted to have novel splicing isoforms: KDM1A (LSD1), SOX6, and LMO2.

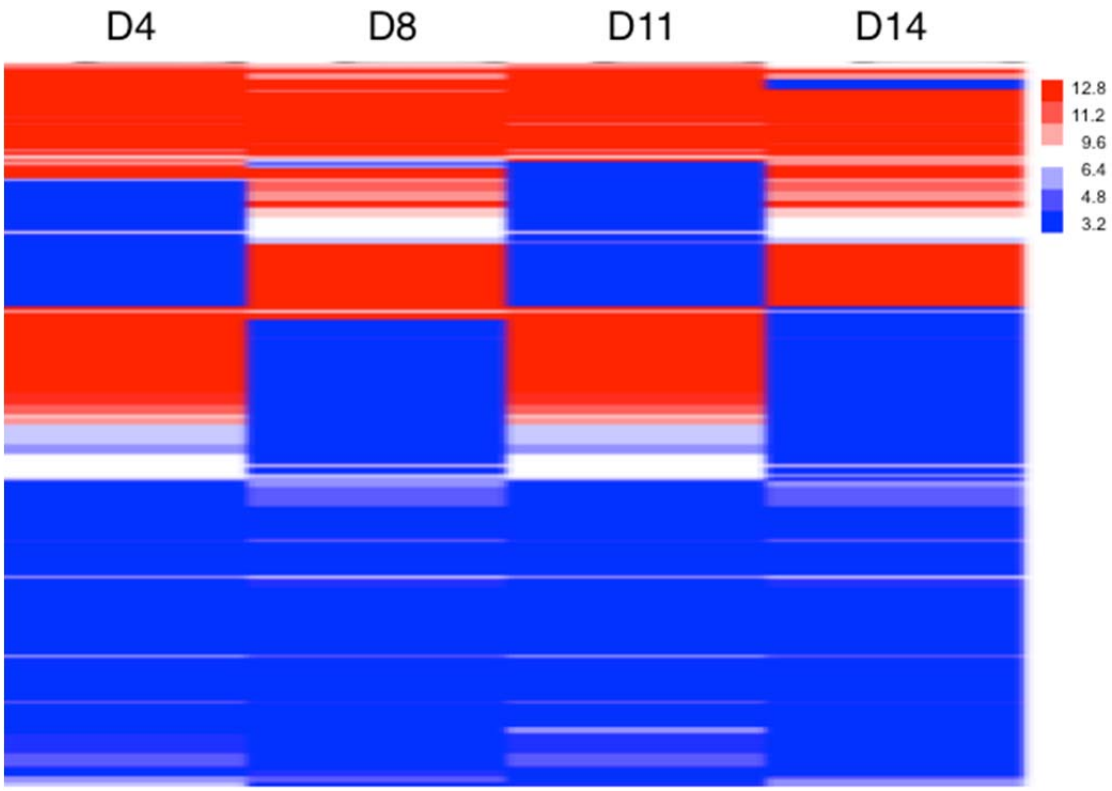


Figure 30. Global view of unique splice junction counts. Each row represents a gene that has a change in its unique splice juctions between adjacent time points. Red represents more unique splicing junction at the time point, whereas the blue represents less unique splicing junction at a time point. For known genes whose unique splicing junction counts change between any two consecutive stages during differentiation, the numbers of unique splice junctions at each time point were extracted and clustered using unsupervised hierarchical clustering to create the above heatmap.

KDM1A (LSD1) was initially identified as a component of the CtBP transcriptional repressor complex (Ballas, et al., 2001; Shi, et al., 2003), but has later been found in other complexes, such as NRD (Tong, et al., 1998), CoREST (You, et al., 2001), and a group of HDAC complexes (Hakimi, et al., 2002; Hakimi, et al., 2003; Humphrey, et al., 2001). The function of the LSD1-CoREST-HDAC complex was originally characterized in non-neuronal and neuronal precursors (Ballas, et al., 2001; Ballas, et al., 2005; Battaglioli, et al., 2002), but it has recently been demonstrated to play a

role in hematopoiesis (Saleque, et al., 2007). shRNA-mediated knock-down of LSD1 was demonstrated to inhibit erythroid differentiation (Saleque, et al., 2007) and to induce γ-globin gene expression in human primary CD34+ progenitor cells (Shi, unpublished data). Together these lines of evidence suggest that LSD1 might be an appealing pharmaceutical target for treating hemoglobinopathies. Surprisingly, a potential novel isoform of LSD1 was identified in differentiating CD34+ cells (Figure 31). In addition, by exploring the UCSC EST database, the structure of the potential novel isoform was compatible with two reported ESTs: DR762029 and CN341829. Therefore, it may be of interest to examine which isoforms are expressed during erythroid progenitor differentiation and to analyze the specific function of that isoform during erythropoiesis.

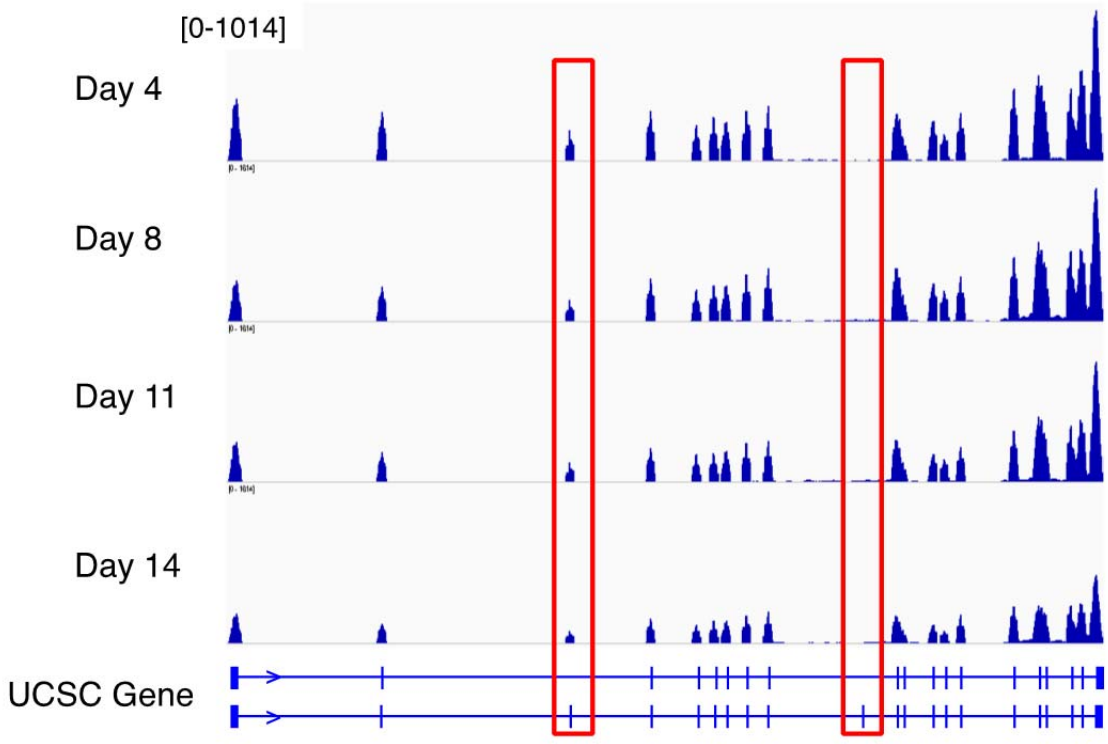


Figure 31. Genomic locus of LSD1. The square bracket indicates the range of the peak height. The pile up of sequence reads marked by red box on the left indicates that the second isoform of LSD1 is expressed. However, the region marked by the red box on the right

disagrees with this isoform. Taken together, these suggest a potential novel isoform of LSD1 being expressed during erythroid differentiation.

SOX6 (aka SRY (sex determining region Y)-box 6) was originally identified in a mouse testis cDNA library, but has since been shown to play an essential role in the development of the central nervous system (Hamada-Kanazawa, et al., 2004; Hamada-Kanazawa, et al., 2004; Stolt, et al., 2006), in chondrogenesis (Ikeda, et al., 2004), and for cardiac and skeletal muscle cell differentiation (Hagiwara, et al., 2000; Hagiwara, et al., 2005). Sox6 was also recently reported to be a crucial regulator of murine definitive erythropoiesis (Dumitriu, et al., 2010; Dumitriu, et al., 2006) and was shown to enhance human erythroid progenitor differentiation (Cantu, et al., 2011). SOX6 can also silence epsilonY globin transcription during murine definitive erythropoiesis (Cohen-Barak, et al., 2007; Yi, et al., 2006). Therefore, SOX6 could also serve as a therapeutic target for sickle cell disease or β-thalassemia and counteract the sickle polymerization by reactivating epsilonY in those patients. In differentiating CD34 erythroid progenitor cells, a potential novel exon was identified in my work (Figure 32) and was supported by a human EST, BU657066.

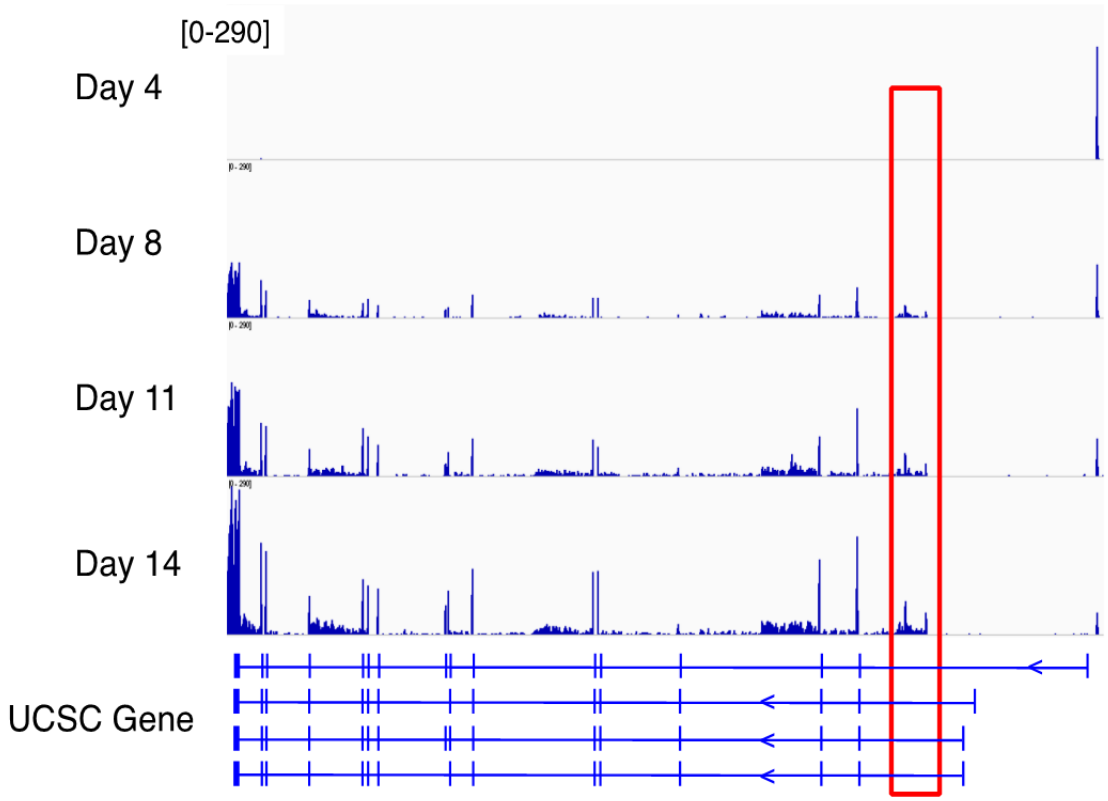


Figure 32. Genomic locus of SOX6. The square bracket indicates the range of the peak height. The peak of sequence reads marked by the red box falls within intronic region of all the isoforms of SOX6. This may suggest an unannotated exon of SOX6 at D8 through D14.

LMO2, LIM domain only 2 (rhombotin-like 1), was formerly called RBTN2, RBTNL1, and TTG-2. It was originally discovered through its involvement in a chromosomal translocation that occurs in some adult T-cell acute leukemias (Boehm, et al., 1991; Foroni, et al., 1992; Royer-Pokora, et al., 1991). The essential role of LMO2 in erythropoiesis was first characterized in a study in which the gene was targeted for inactivation in embryonic stem (ES) cells, and the resulting homozygous *Lmo2* null mutant animals were depleted in yolk sac erythropoiesis (Warren, et al., 1994). Furthermore, a complete block to erythroid development was observed in *Lmo2*$^{-/-}$ ES cells or wild-type ES cells that had been transduced with an anti-LMO2 single-

chain antibody (Nam, et al., 2008; Warren, et al., 1994). However, LMO2 can also act in the opposite way, to also inhibit erythroid progenitor differentiation, when it is overexpressed (Visvader, et al., 1997). While proper control of LMO2 expression is pivotal for normal erythroid development, overexpression of *Lmo2* is more likely to have a dominant negative effect (Terano, et al., 2005). A recent report indicated that a previously unrecognized promoter of *Lmo2* can mediate its expression (Oram, et al., 2010), prompting us to carefully inspect the *Lmo2* locus in differentiating erythroid progenitors. Unexpectedly, a peak with accumulating sequence reads was observed in an intron of the *Lmo2* gene (Figure 33), and *in vivo* utilization of this potential novel exon was also supported in two reported human ESTs: AV759180, AA742325.

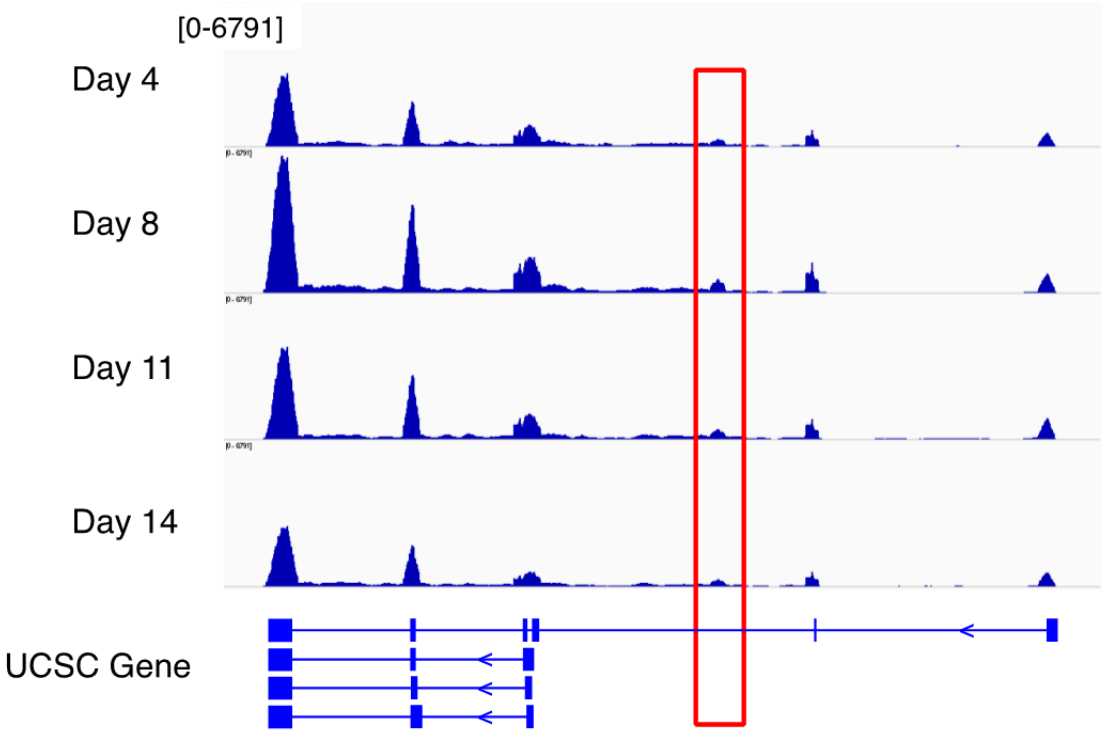


Figure 33. Genomic locus of LMO2. The square bracket indicates the range of the peak height. The small peak of reads marked by the red box falls within the intron of the longest

isoform of LMO2. This may suggest an unannotated exon of LMO2.

iii. Potential novel intergenic/intronic transcripts

One of the advantages of RNA-Seq over array-based platforms comes with the unbiased sequence of poly-A RNAs. Therefore, it is speculated that with this unbiased survey of the transcriptome, we may expect novel transcripts. Here I present a preliminary characterization of the potential novel transcripts falling entirely within introns or intergenic regions. To identify these two categories of novel transcripts we started with the entire list of predicted intergenic and intronic transcripts from the Cufflinks package (5546 intergenic and 3492 intronic transcripts) and then followed several steps to remove likely false positives. We compared the latest human gene annotation (knownGene) from UCSC with our novel transcript list to exclude any predictions that overlap with the latest known genes annotation. We then looked at the characteristics of the known transcripts to set up criteria for refining the list of predicted novel transcripts. The characteristics considered involved the ORF length, number of exons, expression levels, repetitive sequence composition, and homology. A threshold was chosen for each characteristic such that 5% of the known transcripts do not meet the threshold, (5% quantile). The predicted novel intergenic/intronic transcripts considered for further case study are those that met the threshold for all criteria: ORF length > 384 bp, more than 2 exons, less than 30% of repetitive sequences, > 0.1 average conservation score, and expression levels greater than 3.46e-4, 7.01e-5, 0, and 0 FPKM in D4, D8, D11, and D14. Applying these criteria to the predicted novel

transcripts, we obtained a final list of potential novel intergenic (31) and intronic (1) transcripts. An example of a predicted novel intergenic transcript in the final list is shown in Figure 34.

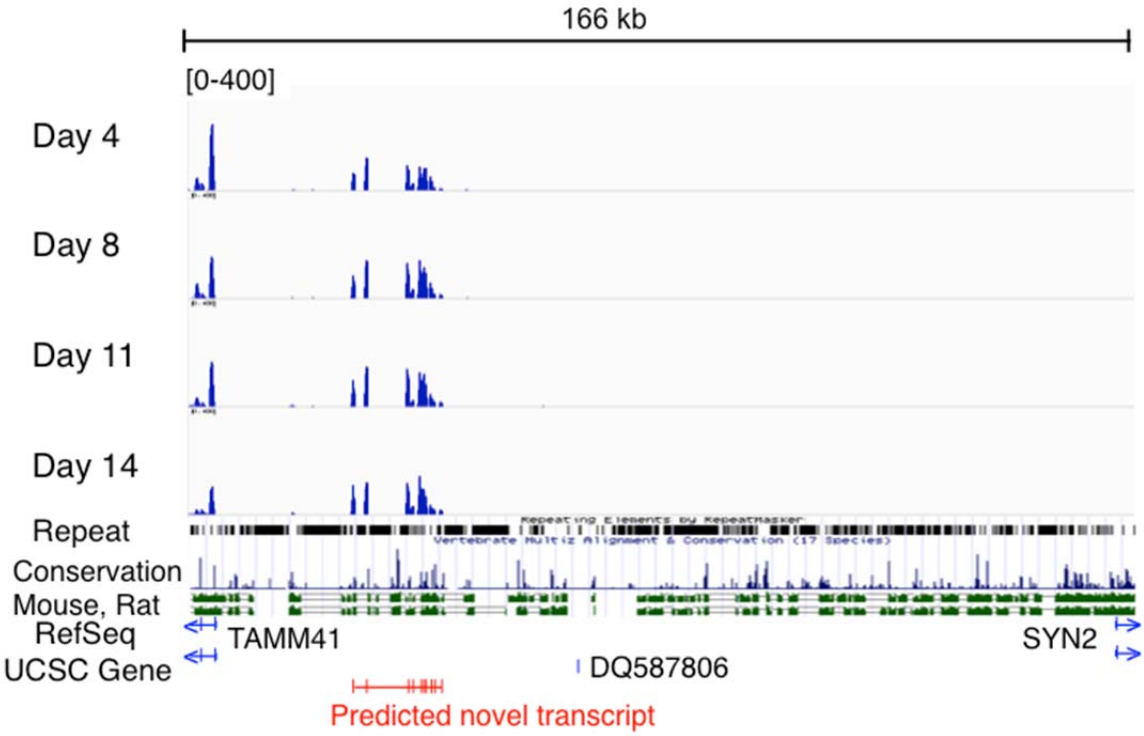


Figure 34. Genomic locus of a potential novel intergenic transcript. The square bracket indicates the range of the peak height. This predicted novel transcript falls within intergenic region between genes TAMM41 and SYN2.

D. CONCLUSIONS

In this report, we have characterized the transcriptome dynamics of human hematopoietic progenitor differentiation using RNA-Seq. Genes highly expressed during erythroid progenitor differentiation were identified, including several novel or recognized but previously unappreciated factors. Differentially expressed transcripts clustered into 6 groups, and several transcription factors were identified whose binding sites were enriched in distinct gene clusters. By comparing the most

immature D4 to mature D14 samples, genes downregulated during differentiation are most enriched in Rho GTPase cell motility transcripts, possibly reflecting their lessening requirement for mobility/homing as the cells mature. In addition, these downregulated genes are also enriched in mitochondrial, endoplasmic reticulum and Golgi compartments. This may reflect the fact that mammalian red blood cells discard these organelles during terminal differentiation. On the other hand, genes upregulated during differentiation are enriched in heme biosynthesis and hemoglobin chaperone pathways. Other enriched GO terms included oxygen binding and oxygen transport activity. Potential novel isoforms of known erythroid regulators LSD1, SOX6 and LMO2 have been identified and have supporting cloned ESTs. In addition, using a well-defined set of criteria, several novel transcripts expressed in these cells were defined. To summarize, these data are consistent with previous studies using microarray platforms, but additionally provide unbiased transcriptome profiling and splicing isoform information. Therefore, these data provide an excellent resource for further focused downstream analysis of hematopoietic progenitor differentiation.

Chapter V

PePr: A ChIP-Seq Peak Prioritization Pipeline for Testing Replicated ChIP-Seq Data and Integrating External Annotations

A. INTRODUCTION

Understanding the flow of genetic information in living cells has been under extensive investigation for decades. Determination of genome-wide binding patterns of DNA-associated proteins and genome-wide profiles of epigenetic marks are central to this information flow. Chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing (ChIP-Seq) has enabled researchers to generate genome-wide *in vivo* interaction maps between DNA-associated proteins and DNA sequences, and achieves higher resolution and comprehensiveness than previously used ChIP-chip experiments (Barski, et al., 2007; Johnson, et al., 2007; Mikkelsen, et al., 2007; Park, 2009; Pepke, et al., 2009; Robertson, et al., 2007).

As the cost of ChIP-Seq experiments decreases we expect that relevant biological replicates will increasingly be studied. ChIP- Seq has sparked the development of several computational methods for detecting the binding sites of the immunoprecipitated factor from ChIP-Seq experiments (Blahnik, et al., 2010; Boyle, et al., 2008; Fejes, et al., 2008; Ji, et al., 2008; Johnson, et al., 2007; Jothi, et al., 2008; Kharchenko, et al., 2008; Mortazavi, et al., 2008; Nix, et al., 2008; Robertson, et al.,

2007; Rozowsky, et al., 2009; Tuteja, et al., 2009; Valouev, et al., 2008; Zang, et al., 2009; Zhang, et al., 2008). The background model used by different peak calling programs differs and has included using the Poisson distribution (Qin, et al., 2010; Zhang, et al., 2008), the binomial distribution (Ji, et al., 2008; Nix, et al., 2008), the negative binomial distribution (Feng, et al., 2008; Ji, et al., 2008), the zero-inflated negative binomial distribution (Rashid, et al., 2011), and kernel density estimation (Valouev, et al., 2008). However, to our knowledge, none of the peak calling programs calculates an estimation of variance among biological replicates. When data from biological replicates exist, the user is often left to concatenate reads from replicates together. Nevertheless, variation among independent biological replicates has been observed to be higher than that among technical replicates (Kaufmann, et al., 2009), and this variation is especially important in certain applications involving epigenomics, such as histone methylation or acetylation. Recently, DBChIP (Liang and Keles, 2012) was developed to identify differential binding from ChIP-Seq data and was able to account for replicate variation when applicable. However, the program works with output from existing peak finding programs and does not take into account this variation during the peak-calling process. Therefore it does not fully address replicate variation when detecting peaks from ChIP-Seq raw sequence tags.

Moreover, the above-mentioned programs report the peaks based on statistical significance using the ChIP-Seq data alone. Given the numerous experimental and computational external annotation data available, surprisingly little effort is devoted

to incorporating this wealth of information. A recently published peak finder, ZINBA, is based on a mixture regression framework and classifies genomic regions into background, enrichment or artificial zero count. ZINBA has shown that the integration of external annotation, such as GC content, may improve peak detection in ChIP-Seq experiments (Rashid, et al., 2011). Therefore, incorporating certain external functional annotations into the ChIP-Seq analysis framework provides additional information that could potentially improve prioritization of binding peaks. Here, we use location of the ChIP-Seq binding relative to a gene as a source of prior information to help better prioritize the ChIP-Seq binding events. Successful identification of functional binding sites will make better use of the ChIP-Seq results for prioritizing downstream investigations or validation studies that link factor binding or histone modifications to transcription.

Here we present a novel ChIP-Seq Peak Prioritization (PePr) pipeline that takes into account the variation among biological replicates from ChIP-Seq experiments. This pipeline has the option to rank peaks based on their global binding pattern by incorporating the location of ChIP enriched regions relative to gene structure. PePr also has the advantage of being flexible enough to analyze either sharp, narrow peaks (e.g. transcription factors) or broad peaks, as is observed for certain histone modifications. Although our method is inherently different from other ChIP-Seq software, we chose two commonly used programs, ERANGE (Johnson, et al., 2007; Mortazavi, et al., 2008) and MACS (Zhang, et al., 2008) to compare with our base method to ensure satisfactory performance. ERANGE calculates fold enrichment

using control data to assess background and does not use any statistical distribution for background modeling. MACS employs a dynamic (local) Poisson distribution to capture local variances in the ChIP-Seq data. Both ERANGE and MACS support sequence reads concatenation from several replicates. We applied these two programs and our method to ChIP-Seq data for the bZIP transcription factor Activating Transcription Factor 4 (ATF4) using ATF4 knockout cells as the control. We then applied our method to ChIP-Seq data for histone 3 lysine 27 trimethylation (H3K27me3) assessed on four squamous cell carcinoma cell lines. The first ChIP-Seq data set, used to evaluate PePr performance, studies the role of ATF4 in response to endoplasmic reticulum (ER) stress and was performed in parallel with RNA-Seq to assess gene expression changes in the same samples. The other ChIP-Seq data, using antibody against H3K27me3, also has gene expression data available for the same cells. We compared the performance of our basic PePr implementation capable of estimating the dispersion factor of the negative binomial distribution to account for the biological variation among replicates (Version 1, V1) with the implementation that further incorporates external annotation (Version 2, V2), and with a basic method that also uses the negative binomial distribution but with concatenation of reads from replicates.

B. METHODS

i. Datasets

a. *ATF4* - ChIP-Seq and RNA-Seq data for ATF4 were used in this study to assess the performance of PePr. Three biological replicates of ChIP-Seq using Illumina Genome

Analyzer with wild-type and ATF4 knock-out mouse embryonic fibroblast (MEF) cells treated with tunicamycin were performed (one lane per sample) and aligned to the mouse reference genome build version mm9 using ELAND. The number of reads and percent aligned for each lane is provided in Table 6. In addition, two biological replicates of RNA-Seq using Illumina Genome Analyzer with wild-type and ATF4 knock-out MEF cells also treated with tunicamycin were performed. (For further details, see supplemental methods.) For RNA-Seq data analysis, Bowtie was employed to align reads to the mouse reference genome (version mm9) plus known splice junctions, created by ERANGE scripts and UCSC known gene models (Langmead, et al., 2009; Mortazavi, et al., 2008). Counts of reads and RPKM values for each gene were determined using ERANGE software and were tested for differential expression in R using the limma package and IBMT method (Sartor, et al., 2006). Testing was performed using log2-read counts normalized to the total number of aligned reads for each sample. We then tested wild-type versus Atf4 -/-. The IBMT method is an empirical Bayesian method that provides improved estimates of variance for experiments with small samples sizes, while taking into account the relationship between variance levels and the total read count. The False Discover Rate (FDR) for each comparison was calculated using the Benjamini-Hochberg method.

b. ***H3K27 trimethylation*** - ChIP-Seq data using antibody against H3K27me3 and Affymetrix Human Genome U133 Plus 2.0 array gene expression data were utilized to evaluate the performance of PePr in analyzing histone modification data. These

data demonstrated a high level of variation among individuals and thus presents a case where we expect accounting for biological variation to result in an improvement; the data also exhibited broader peak width compared to ATF4. ChIP-Seq using two Human Papillomavirus (HPV)-positive (CaSki and UMSCC-47) and two HPV-negative (UMSCC-4 and UMSCC-74A) squamous cell carcinoma cell lines were performed. Cell lines were cultured as previously described (Sartor, et al., 2011), and chromatin immunoprecipitation and library preparation was performed by GenPathway (part of Active Motif, Carlsbad, CA) using a commercial quality antibody specific for H3K27 trimethylation. DNA was amplified according to the Illumina ChIP-Seq library construction protocol, and a region of 250-350 bp was excised from the preparative Agarose gel. Sequencing of the four immunoprecipitated samples and four input DNA samples was performed at the University of Michigan DNA sequencing core using the Illumina HiSeq with 50 base single-end reads. Raw reads were quality checked using FastQC and aligned to the human reference genome build version hg19 using BWA with the default parameters (Li and Durbin, 2009). The number of reads and percent aligned for each lane is provided in Table 7. Affymetrix expression arrays were performed using the same cell lines (available at Gene Expression Omnibus # GSE24089) and were processed and analyzed as previously described (Sartor, et al., 2011).

ii. PePr input formats

Our pipeline accepts multiple formats, including SAM, default Bowtie output, BED format, and direct output from ELAND V2 read aligner with the -multi or -extended

option specified. Control data is required.

iii. PePr preprocessing

a. *Removal of duplicates* - Users have the option either to remove duplicated reads (which may originate from PCR amplification), or to keep all mapped reads. The maximum number of duplicated reads allowed for a single position can be set by the user or determined by a binomial test (Zhang, et al., 2008).

b. *Shift size calculation* - Because reads randomly occur from either the plus or minus DNA strand, shifting sequence tags towards the 3' end by half of the estimated fragment length can improve estimation of the precise protein-DNA interaction site (Park, 2009). For each chromosome, a shift size *d* is determined by maximizing the overlap between reads from forward and reverse strands. The median of the determined shift sizes from all chromosomes is then used to shift all the reads towards their 3' direction by *d*. We found this to provide a stable estimate. The maximum shift size allowed is 300 bp. If the estimated shift size *d* is smaller than 20 or when the estimated variance of the shift size learned from all chromosomes is greater than 0.3, the shift size *d* is set to 100 bp. Users have the option to either use the result of this analysis or provide their own shift size in the next analysis step.

c. *Window size calculation* - A window size, which provides the flexibility to analyze either sharp or broad peak profiles, is then calculated for each chromosome.

First, all reads are assigned to non-overlapping 25-base pair bins. Then, for each chromosome, 100 candidate peak regions are identified iteratively. In each iteration, the bin with the largest number of reads is denoted as the seed for candidate region, and the region is extended to flanking bins that harbor more than 10% of reads of the seed. The width of these regions each provides a single estimate of the window size. The median of the computed window sizes from all chromosomes is then used. Again, we found that this value was sufficiently stable across chromosomes. Users have the option to either use the result of this analysis or provide their own window size in the next step.

d. ***Normalization*** - The number of reads per window is calculated for each sample using a sliding window algorithm for all chromosomes. For this step, users have the option of using no normalization, a simple "scale-up" method that normalizes the total number of reads in each sample to the maximum among all samples, or the iterative, quantile method defined in the edgeR R package (Robinson, et al., 2010).

iv. PePr peak detection

a. ***Dispersion factor calculation*** - Our V1 and V2 pipelines use a negative binomial model, which requires the estimation of a common dispersion factor for all windows and/or a dispersion factor for each window. The negative binomial distribution allows us the flexibility to separately estimate the means and variance levels. Users have the choice of using a common dispersion factor, a local dispersion factor (estimated per window), or a mixed dispersion factor, which is defined as the

geometric mean of each individual window's dispersion factor and the common dispersion. Users also have the option to remove windows with < *n* reads before estimating the common dispersion factor (default: *n*=5). A table illustrating how this parameter affects results on the ATF4 dataset is provided in Table 8. It is assumed that the read count in each window follows a negative binomial distribution, such that $X_i \sim$ N.B. (μ_i, σ), $Y_i \sim$ N.B. ($\gamma_i\mu_i$, σ) where $i = 1, \ldots, G$ is the window in the genome, X_i and Y_i represent the number of reads in the i^{th} window in the genome from the control and the ChIP samples (or group 1 and group 2 samples) respectively, μ_i is the mean read level for window *i* in controls (estimated by, $\hat{\mu}_i$, the average read count in i^{th} window in the genome from control samples), σ is the common dispersion factor (or σ_i for local or mixed dispersion), and γ_i is the fold change between read level in the ChIP'd sample group (group 2) versus the control sample group (group 1) in the i^{th} window. γ_i , μ_i , and σ are estimated using their maximum likelihood estimates. We then test for significant differences in binding between the ChIP and control groups (group 2 versus group 1) by testing the hypothesis H_0: $\gamma_i <= 1$ versus H_1: $\gamma_i > 1$. A histogram depicting the local dispersion estimates for ATF4 and H3K27me3 data can be found in Figure 35.

b. ***P-value and FDR calculation*** - Either the generalized likelihood ratio test (GLRT) or log-gamma estimate with the Wald test (Aban, et al., 2008) may be used in V1, whereas the generalized likelihood ratio test (GLRT) is used in V2. For the Wald Test, a z-score for each window is calculated, and parametric (Benjamini-Hochberg) FDR is then calculated to correct for multiple testing. Empirical (sample swap) FDR

is also calculated. For the GLRT, P_i^1 and P_i^0, the likelihood under the alternative hypothesis (H_1: $\gamma_i > 1$) and null hypothesis (H_0: $\gamma_i \leq 1$) respectively, are defined as described in (Aban, et al., 2008).

Let $\overline{X}_i$ be the average read count in the i^{th} window from control (group 1), and $\overline{Y}_i$ the average read count in the i^{th} window in the genome from ChIP (group 2) sample. Then $\hat{\gamma}_i^1$, the maximum likelihood estimate for γ_i^1 under the alternative hypothesis (H_1: $\gamma_i > 1$) is $\hat{\gamma}_i^1 = \overline{Y}_i / \overline{X}_i$ when there are more reads in the i^{th} window in the ChIP (group 2) sample than in control (KO) sample, and is otherwise restricted to 1. $\hat{\gamma}_i^0$, the maximum likelihood estimate for γ_i^0 under the null hypothesis (H_0: $\gamma_i \leq 1$) is 1 when there are more reads in the i^{th} window in control (group 1) than in ChIP (group 2) samples, and equals to $\hat{\gamma}_i^0$ otherwise.

v. Incorporating peak location relative to gene structure (PePr version 2 pipeline)

Our method for incorporating peak locations relative to gene structure is data dependent. Thus, if no relationship exists between gene structures and binding locations, PePr V2 analysis will adequately capture this. However, if there is a relationship between the binding profile and location relative to a gene, we will capture this relationship, and provide a summary to the user. We define 12 bins as genomic regions relative to gene structure as: 5' UTR, 3' UTR, 0-1kb 5', 1-5kb 5', 5-10kb 5', >10kb, 0-5kb 3', 5-10kb 3', exon1, intron1, other exon, other intron (Figure

36). These bins are derived from the UCSC knownGene table, and genes residing within another gene (whose transcription end site occured before the other gene's transcription end site) were removed to avoid ambiguity when assigning binding to the defined bins. For ATF4 ChIP-Seq, we used mm9 and for H3K27 ChIP-Seq, we used hg19 knownGene file. We define a mixture model with its log-likelihood being:

$$\sum_{i=1}^{G} \log(P_i^0 * (1 - \pi_{b(i)}^1) + P_i^1 * \pi_{b(i)}^1)$$

where P_i^0 is the likelihood of $\gamma_i \leq 1$ under the null hypothesis (H_0) of no functional binding and P_i^1 is the likelihood under the alternative hypothesis (H_1) of $\gamma_i > 1$ for window i. $\pi_{b(i)}^1$ is the marginal probability of binding under H_1 for the i^{th} window in the genome that belongs to bin $b(i)$. $\pi_{b(i)}^1$ is the same as π_{bj}^1 for every window i belonging to bin b_j, $j = 1, \ldots, 12$. After the likelihood calculation, we then calculate initial estimates of π_{bj}^1 by the following formula:

$$\left(\sum_{i:b(i)=b_j} (P_i^1 * \pi_{b(i)}^1) / (P_i^1 * \pi_{b(i)}^1 + P_i^0 * \pi_{b(i)}^0) \right) / Bj$$

where B_j is the total number of windows belonging to bin b_j, $j = 1, \ldots, 12$. We take an iterative approach by iteratively substituting the estimated π_{bj}^1 into the same formula as above to obtain new π_{bj}^1 estimates. As a secondary validation that this method converges appropriately, we also implemented a Metropolis Hastings algorithm, and our results showed that this process converges to that estimated by the Metropolis Hastings algorithm, but with monotone convergence and quicker run time.

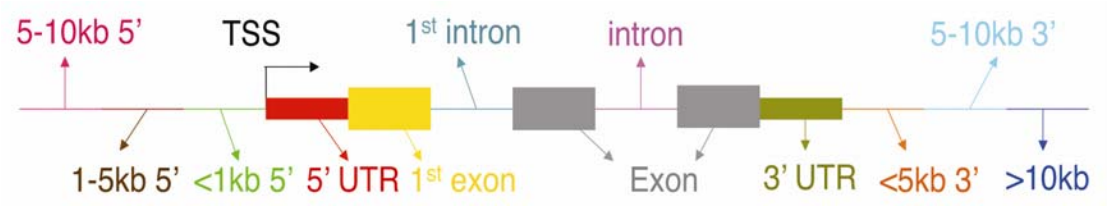


Figure 36. The twelve bins defined according to the gene structure and used in the mixture model of PePr V2.

FDR is then calculated using the Benjamini-Hochberg method from the p-value for specific window *i*. p-values are calculated as follows:

$$P_i^0 * \pi_{b(i)}^0 / (P_i^0 * \pi_{b(i)}^0 + P_i^1 * \pi_{b(i)}^1)$$

Adjacent significant windows with FDR less than the default (0.05) or user-chosen value separated by a gap of a length of window size are merged and reported as a peak along with the peak length, read counts in the peak, p-value, FDR, nearest gene and the bin assignment.

vi. Basic pipeline

For comparison, we developed a basic peak finder that also uses a negative binomial distribution, but that concatenates reads from replicate samples and then performs the above-mentioned processes (shift size and window size calculation). The difference in number of mapped reads between the corresponding windows is calculated. For each window, the number of reads (n_i) is assumed to follow a negative binomial distribution, $n_i \sim$ negative binomial (α, β), and α and β are estimated and the FDR is calculated as described in (Ji, et al., 2008). Briefly, adjusted p-values are calculated using the False discovery rate (FDR) approach, calculated for each window by dividing the number of false positive windows (determined with a

specific height cutoff in the negative binomial modeled background), by the number of bins determined at that same height cutoff in the control-subtracted ChIP sample. Adjacent windows with FDR less than 0.05, or separated by a gap no more than the window size were merged and reported as a peak along with the FDR and read counts in the peak.

vii. ATF4 peak finding analyses

All uniquely mapped reads with up to two mismatches were extracted for downstream peak calling performance evaluation. We evaluated the performance of PePr by applying our basic pipeline, the two versions of PePr, ERANGE, and MACS with the ATF4 dataset described above. In all cases duplicated reads >2 times were removed based on results of the binomial test. For ERANGE, the "-listPeak", "-revbackground", "-nomulti", and "- shift learn" were set. For MACS, "-pvalue=1e-15" was the only parameter used. The sequences of the peaks reported with the above mentioned parameters from all methods were retrieved from UCSC Genome Database (http://www.genome.ucsc.edu, mm9, July 2007).

De novo motif search was subsequently performed using the MEME (Bailey and Elkan, 1994) suite on the sequences with the same parameters. The parameters used were "-dna -nmotifs 3 -mod zoops -maxw 12 -maxsize 20000000 -revcomp", which specify the number of motifs to search for, the zoops assumption (zero or one occurrence per peak sequence), the maximum motif length of 12, and the maximum dataset size of 2,000,000 characters. Sequences were searched in both forward and

reverse orientations using the additive peak sequences of 200 peaks. That is we sorted the peaks in increasing order of FDR or p-value whichever is applicable, then ran MEME on bins of 200 peaks. Because MACS was the only method that reported more than 10000 peaks, we only applied the top 10000 peaks to MEME. The motif found by MEME was then compared to known motifs in JASPAR motif database (Bryne, et al., 2008) using Tomtom (Gupta, et al., 2007). A match is called at significance threshold of E-value (the expected number of false positives) less than 10 using Pearson correlation coefficient as the motif column comparison function. To evaluate the performance of PePr in prioritizing functional peaks, we compared the reported peaks by all methods described above to the parallel ATF4 RNA-Seq data. We first identified the target genes of ChIP-Seq bindings that were also detected by RNA-Seq and then sorted these genes by increasing order of the distance between the binding and the gene's TSS. We then calculated the percentage of genes differentially expressed (greater than 2 fold change and FDR < 0.05) for every 100 genes.

viii. H3K27me3 differential peak finding analyses

To evaluate the performance of our pipeline in calling broad peaks in a two-group comparison, we applied PePr to the histone H3K27 trimethylation ChIP-Seq data described above. To assess the performance of PePr in terms of prioritizing functional peaks, we compared the associated genes of ranked peaks from our basic pipeline and the two versions of PePr to the corresponding H3K27 microarray expression data. Because our preliminary analysis indicated that H3K27me3 marks

in HPV-negative samples tend to be closer to gene TSSs, and HPV was recently shown to disrupt H3K27me3 in keratinocytes (Hyland, et al., 2011), we were interested in regions where HPV-negative samples were marked by H3K27me3, but HPV-positive samples were not, to identify regions lost to HPV. Because H3K27me3 typically represses transcription, we expected these sites to be down-regulated in HPV-negative cells compared to HPV-positive cells. We then examined the correlation between genes associated with a ChIP-Seq peak and significance of differential expression, based on the associated Affymetrix study. Specifically, to test whether there is an enrichment of the overlap between a gene with differential binding within 3kb of its TSS and its differential expression, we sorted the expression data by p-value and then calculated the Fisher's exact test odds ratio for increasing numbers of genes.

C. RESULTS

Here we report the development of a ChIP-Seq analysis software that addresses two aspects in ChIP-Seq analysis. To evaluate the performance of PePr, we analyzed two datasets: ATF4 transcription factor, and H3K27 trimethylation in HPV-positive and HPV-negative squamous cell carcinoma cell lines. Because the advantage of PePr comes from its ability to assess variation among biological replicates in the peak detection process, we expected PePr to have little advantage in the analysis of ATF4, and a greater advantage in the analysis of the H3K27 experiment. Furthermore, the ATF4 dataset represents the type of sharp ChIP-Seq peaks typically observed for transcription factors, whereas the H3K27me3 histone dataset represents a class of

broader peaks. Thus, the use of these datasets demonstrates the applicability of PePr to both types of ChIP-Seq data.

i. Transcription factor analysis results

By applying the three ATF4 ChIP-Seq replicates to the basic implementation, the two versions of PePr, ERANGE (tends to have high specificity; identifies relatively few peaks), and MACS (tends to have high sensitivity; identifies a large number of peaks) (Wilbanks and Facciotti, 2010), different numbers of peaks were reported to be bound by ATF4 *in vivo*. With PePr, a shift size of 40 bp was used, a window size of 300 bp was used, and the common dispersion factor was estimated to be 0.37. Marginal likelihood estimates for each of the twelve bins are provided in Figure 37. We used the option "-shift learn" to direct ERANGE to learn shift size from the first chromosome, but the estimated shift size was 0. We used default parameter settings for MACS; MACS initially estimated the shift size to be 41, but because this is less than its minimum allowed, the default shift size of 100 was automatically used. We initially used 10^{-6} as a p-value cutoff for MACS, and it returned 31200 peaks. We then turned to use a more stringent criterion (p-value of 10^{-15}) for peak calling using MACS. There were 9884 peaks identified by the basic pipeline, and 7364 and 5358 peaks reported by PePr version 1 and 2, respectively, whereas ERANCE and MACS (using a p-value cut-off of 10^{-15}) reported 3345 and 15392 peaks, respectively. A Venn diagram depicting the overlap between PePr V1, ERANGE, and MACS are shown in Figure 38.

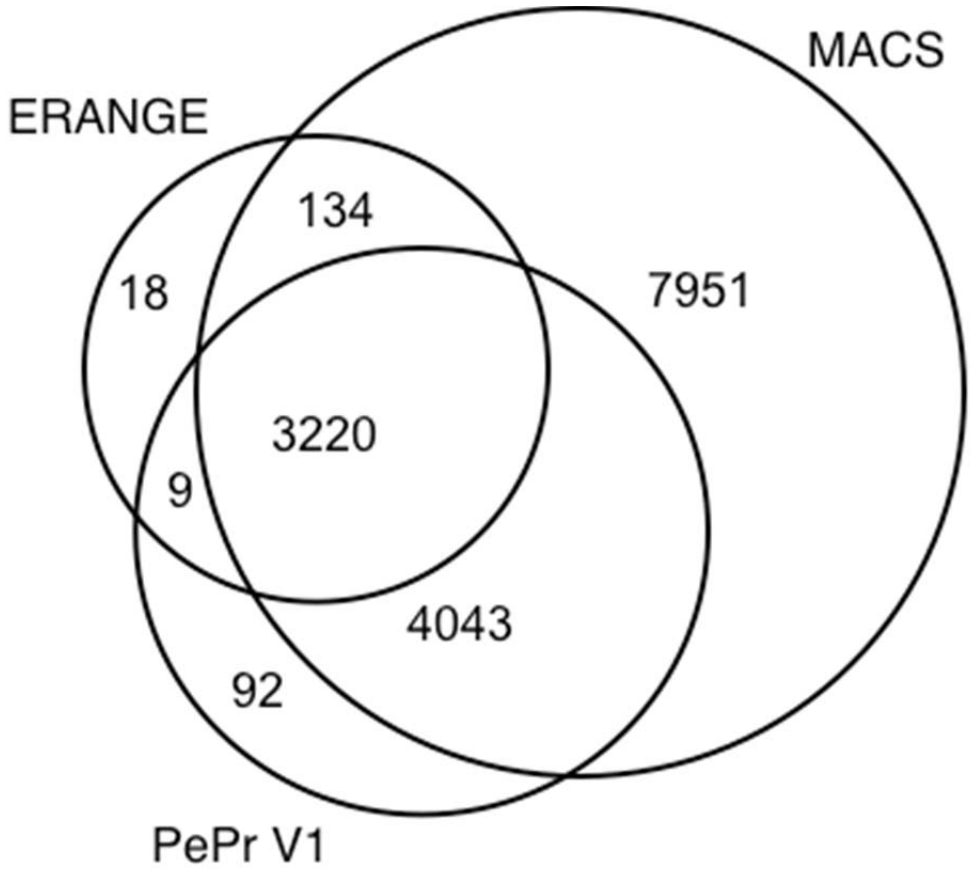


Figure 38. Venn diagram demonstrating the overlap in ATF4 binding peaks between PePr V1, ERANGE, and MACS. Sums of numbers differ slightly from total number of peaks for a method due to some peaks overlapping with more than one peak from another method.

Given the difference in the number of peaks reported by the different software packages, we sought to determine the percentages of the called peaks that represent true binding sites for ATF4. Due to the lack of gold-standard qPCR-verified binding sites for ATF4, we considered ATF4 ChIP-Seq binding sites containing a high confidence canonical ATF4 motif as genuine ATF4 binding sites. We then performed *de novo* motif searches using MEME suite to identify the sequence motifs significantly enriched in the called peaks by all programs sorted by p-value or FDR (Figure 39). Overall, ERANGE-identified peaks have the highest percentage

(97.67%) of peaks containing a motif that matched well to the canonical bZIP transcription factor family motif in JASPAR database, however ERANGE also identified the fewest peaks and the fewest total peaks with a motif. 86.98% of the top 10000 MACS-reported peaks contained the matched motif. The basic implementation, PePr V1, and PePr V2 had 70.75%, 89.35%, and 91.68% of peaks containing the matched motif, respectively. Limiting MACS to the top 7350 peaks (similar to PePr V1), we still find that PePr V1 performs slightly better.

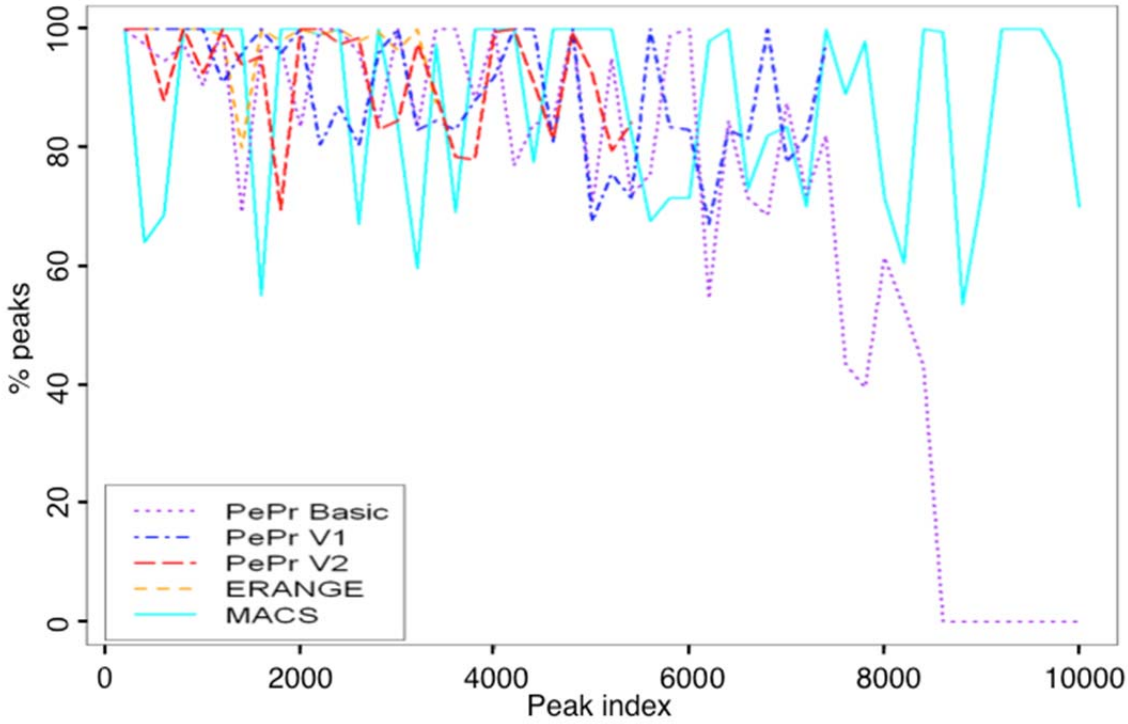


Figure 39. Percentage of peaks containing canonical ATF4 motif. Purple represents PePr basic implementation, blue represents PePr V1, red represents PePr V2, orange represents ERANGE, and cyan represents MACS.

We then evaluated the spatial resolution of the ChIP-Seq analysis methods by inspecting the distance of the peak center to the identified motif. The distribution of

the distances between peak centers to the motif within the peak, when present, is displayed in Figure 40. Although ERANGE had the highest percentage of peaks with an ATF4 motif, the motif is distributed wider across the peak regions. MACS, PePr V1, and PePr V2, all had similar spatial resolution. Our motif analysis of ATF4 suggests that PePr performance on data with little variation, though it is not superior to ERANGE and MACS, is comparable to that of ERANGE and MACS, and with sensitivity and specificity balanced between ERANGE (high specificity) and MACS (high sensitivity).

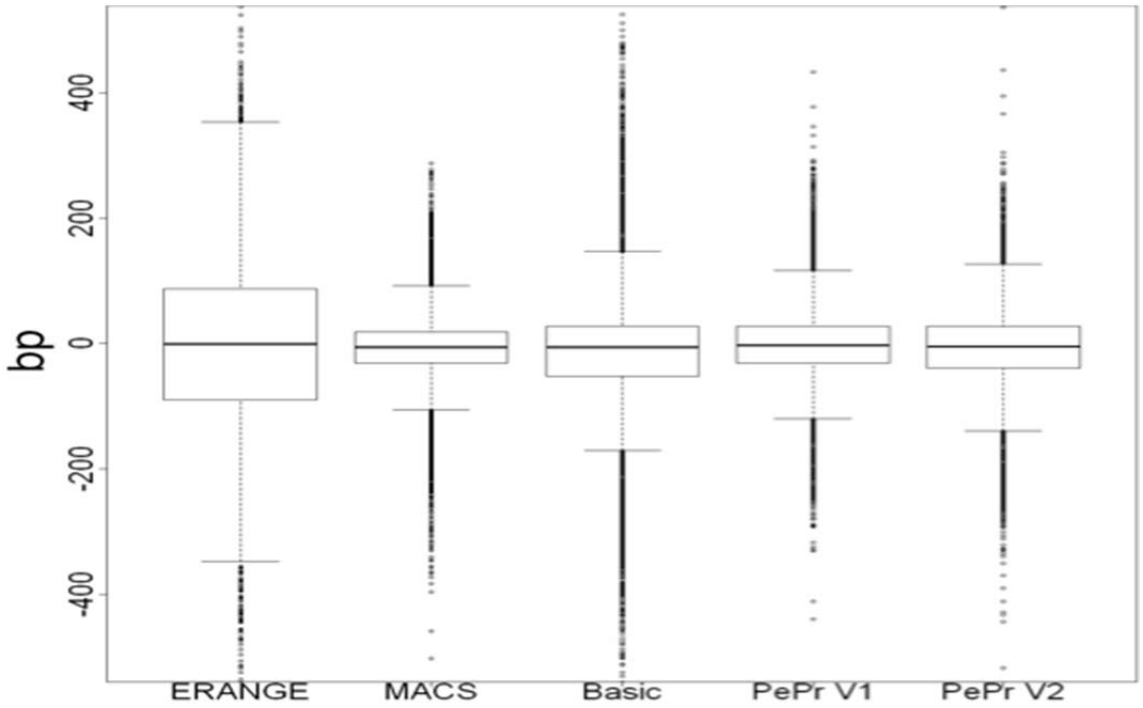


Figure 40. Boxplot showing the spatial resolution of the ChIP-Seq peak finders with ATF4 data. The motifs within the peaks identified by ERANGE are more dispersed than other programs. While MACS has the best spatial resolution in locating the peaks containing a motif, PePr V1 had nearly identical resolution. PePr V2 had the next best resolution.

We next looked at how the identified ChIP-Seq peaks from different programs correlate with target gene differential expression. A ChIP-Seq target is identified as

the closest gene to a peak, and is considered differentially expressed if the expression change is more than 2 fold between ATF4 knock out and wild-type and with FDR less than 0.05. We then sorted the ChIP-Seq targets based on the distance from the TSS to the identifying peak in increasing order. For every 100 genes, we calculated the percentage of genes that are differentially expressed. Barplots showing the percentage of differential expression within every 100 target genes for each program are shown in Figure 41. From these figures we can see that target genes closer to a peak identified by PePr V2 better correlate with differential expression, although methods that identify more peaks tend to show a better association with gene expression in this regard.

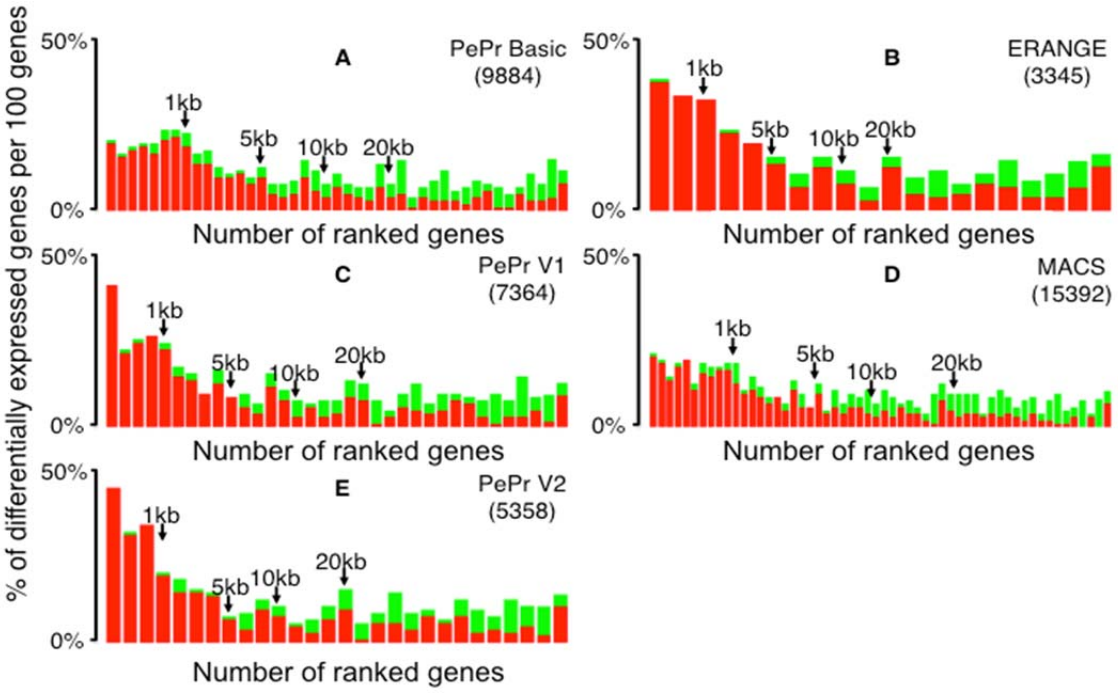


Figure 41. ChIP-Seq binding sites vs. differential expression for the transcription factor ATF4. (A) PePr basic implementation. (B) ERANGE. (C) PePr V1. (D) MACS. (E) PePr V2. The number of peaks identified by each peak finder are listed in parentheses. Each bar represents 100 genes closest to predicted binding site. Overall, PePr V2 tends to identify targets better correlate with differential expression measured by RNA-Seq in the same

samples.

ii. Histone modification analysis results

H3K27me3 ChIP-Seq was performed using two HPV-positive and two HPV-negative squamous cell carcinoma cell lines along with corresponding input controls. PePr estimated the shift size to be 90 bp, the window size to be 888 bp, and common dispersion factor to be 0.53 for comparing HPV(+) and HPV(-) samples. The larger window size compared to the ATF4 data shows the ability of our method to flexibly adjust to broader peak profiles, and the larger dispersion factor is reflective of the higher heterogeneity among samples compared to ATF4. Here we used the mixed dispersion estimate for each window with PePr V1 and V2. Based on a previous report that HPV E7 protein causes a loss of H3K27me3 (Hyland, et al., 2011) and our preliminary analysis that HPV(-) H3K27me3 marks tend to be closer to gene TSSs than HPV(+) marks, we focused on identifying HPV(-) specific peaks (lost due to HPV), and then assessed how the peaks correlate with gene expression differences by HPV status. We applied the H3K27me3 ChIP-Seq to the basic negative binomial pipeline and the two versions of PePr. Consistent with the previous report, PePr V2 reported an enrichment of HPV(-)-specific H3K27me3 just upstream of a TSS and within exons, whereas HPV(+) specific peaks occurred more often in introns and far from genes. Marginal likelihood estimates for each of the twelve bins are provided in Figure 42.

To evaluate how HPV(-) specific peaks identified by PePr correlate with gene expression data, we used HPV(+) as our control for PePr V1 and V2, and we used

our basic pipeline to implement a currently used approach for this type of analysis. For our basic approach, we used the two HPV(-) and two HPV(+) samples each separately with corresponding input control data to identify peaks for each individual. Regions identified in both HPV(-) samples, but neither of the HPV(+) samples was then obtained by comparing the HPV(-) overlapping regions to HPV(+) peak regions, and removing the sites overlapping between the two from the HPV(-) overlapping regions. We next sorted the genes from the microarray dataset by the p-value associated with differential expression, and used Fisher's exact test to test for significant overrepresentation between differentially expressed genes and the genes with a HPV(-) specific H3K27me3 mark within 3 kb of its TSS. We calculated the Fisher's exact test odds ratio for increasing numbers of top ranked genes to assess the level of overlap between genes with an H3K27me3 mark and differential expression. (We used odds ratio instead of the Fisher's exact p-value to avoid a strong dependence on the sample size, since different versions of PePr identified different numbers of peaks.) Results show an overall increase of the odds ratio moving from the basic pipeline to PePr V1 to PePr V2 (Figure 43 and 44). These results suggest that incorporating biological variation and additional annotation into the peak calling process increases the correlation with differential expression.

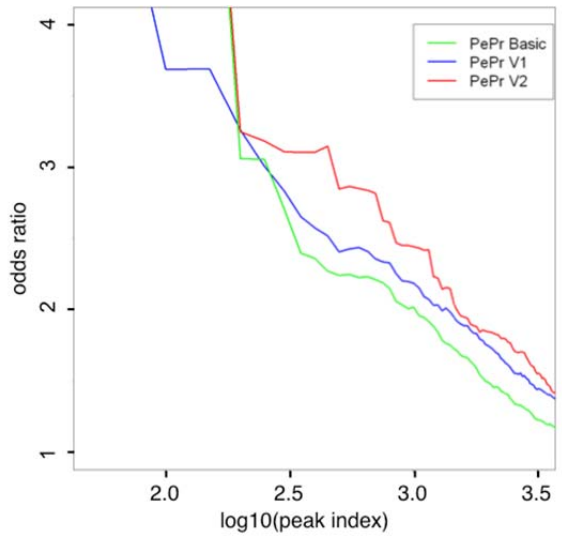


Figure 43. HPV(-) vs (+) odds ratio for enrichment of the overlap between a gene with HPV(-) specific H3K27me3 within 3kb of its TSS and its differential expression limited to up-regulation in HPV(+). Green line represents our basic implementation, blue line represents PePr V1, and red line represents PePr V2.

To illustrate the cause for improvement from the basic pipeline to PePr, we display three regions that the basic pipeline approach identified as significant, but PePr V1 and 2 did not, and three regions vice versa. The regions identified only by the basic, current approach show large variation between the two HPV(-) samples, and/or somewhat smaller (unidentified peaks) in HPV(+) cells (Figures 45 and 46). However, the three regions identified only by PePr V1 and 2 show a large difference between HPV(-) and HPV(+) cells (Figures 47 and 48).

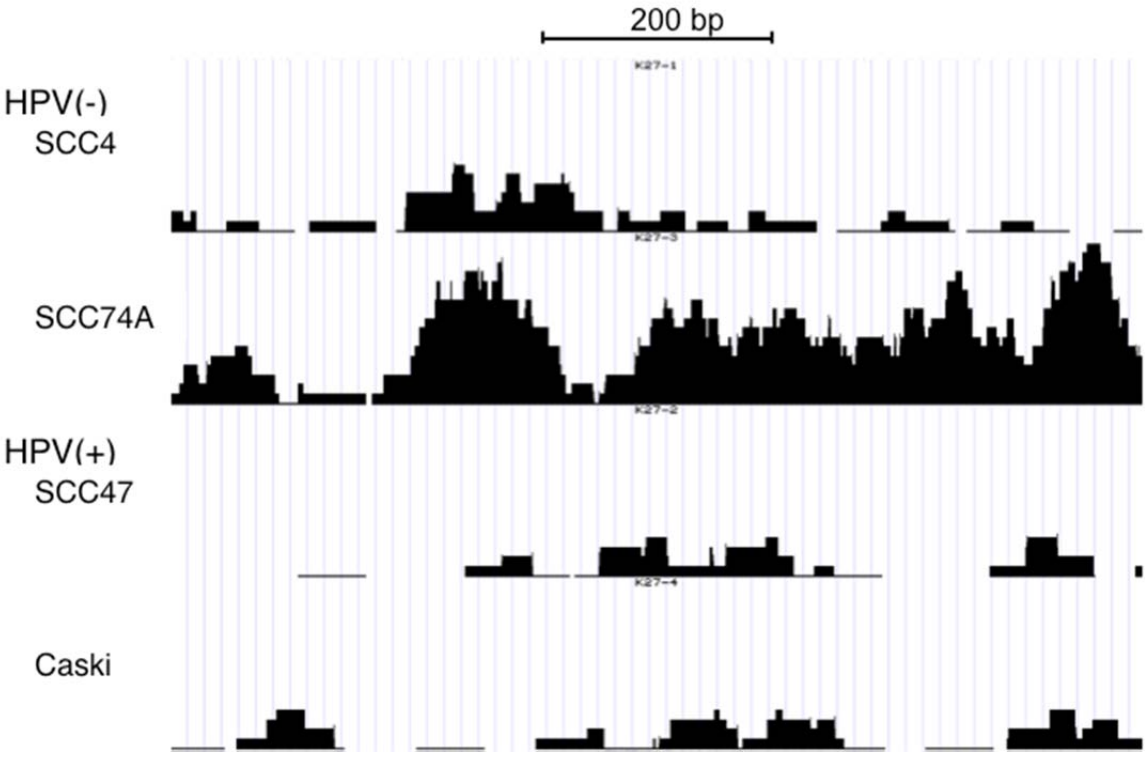


Figure 45. An H3K27me3 peak found by PePr basic implementation but not V1 and V2. Region shown is chr17:26741500-26742341.

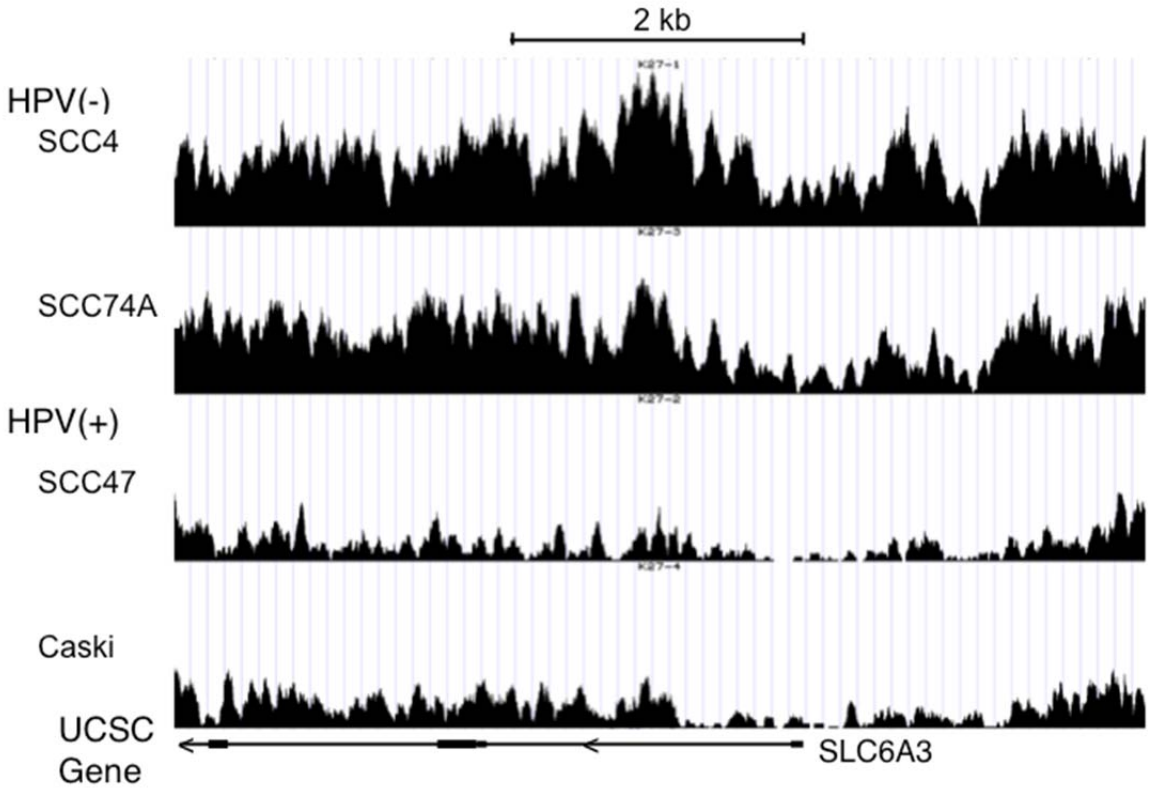


Figure 47. An H3K27me3 peak identified by PePr V1 and V2 but not by the basic implementation. Region shown is chr5:1441224-1447883.

D. CONCLUSIONS

An abundance of ChIP-Seq analysis software programs are available for researchers to identify so-called "peak" regions, indicating genomic regions of interest. However, none of these peak calling programs estimate variation among replicates within the peak calling framework. Here, we introduced our Peak Prioritization pipeline, available to the research community as python scripts. Our newly introduced PePr program not only accounts for variation among biological replicates, but also optionally incorporates location of read-enriched regions relative to gene structure via a mixture model. Demonstrating the use of our PePr pipeline on both transcription factor and histone modification datasets, we illustrated how PePr is

able to incorporate such information and identify peaks correlated with differential expression. Furthermore, PePr can accommodate replicate variation without sacrificing spatial accuracy. By taking into account the location of the binding relative to gene structure, PePr uses properties of the data itself to prioritize peaks better correlated with functional binding. As the cost of ChIP-Seq becomes more affordable, we expect there will be a significant increase in the use of biological replicates in ChIP-Seq, especially in those relating to epigenomics studies. Therefore, based on our current study, we believe PePr will benefit the biomedical research community.

E. SUPPLEMENTAL MATERIAL

i. Supplemental methods

a. Cell culture and MEF generation

Mouse embryonic fibroblasts (MEFs) were generated as described previously (Rutkowski, et al., 2006) , and were then cultured in regular DMEM with 10% FBS (Invitrogen), penicillin-streptomycin, nonessential amino acids, and essential amino acids.

b. Chromatin immunoprecipitation (ChIP)

Atf4+/+ and *Atf4-/-* MEF cells were treated with Tm (2 μg/ml) (Sigma) for 10 hrs, followed by cross-linking and subsequent chromatin immunoprecipitation was performed using anti-ATF4 antibodies (Su and Kilberg, 2008).

ii. Supplemental tables and figures

Table 6. Raw read summary for ATF4 data.

	# of reads	% mapped
WT_Rep1	19,254,443	81.35
KO_Rep1	19,141,706	80.93
WT_Rep2	22,237,936	78.25
KO_Rep2	21,507,605	76.54
WT_Rep3	22,035,854	78.33
KO_Rep3	21,317,951	76.55

Table 7. Raw read summary for H3K27me3 data.

	# of reads	% mapped
HPV(+) SCC47	76,898,730	93.90
HPV(+) CaSki	89,612,434	94.85
HPV(-) SCC4	96,840,383	94.27
HPV(-) SCC74A	85,179,882	93.24

Table 8. The effect of different window sizes and the minimum number of reads/window used for dispersion estimation on the number of peaks identified by PePr V1. Analyses were performed using the GLRT (generalized likelihood ratio test) with the ATF4 dataset. Results using the alternative Wald's test followed a similar trend (data not shown). Values are # peaks / # negative peaks (empirical FDR). Empirical FDR is calculated by dividing the number of negative peaks (via sample swap) by the number of peaks.

cut-off / window size	1	5	10
200	4463/0 (0%)	5668/4 (<0.1%)	10344/16 (0.2%)
300	7364/8 (0.1%)	7878/15 (0.2%)	10297/36 (0.3%)
400	9125/52 (0.6%)	9390/59 (0.6%)	10557/81 (0.8%)
600	11223/121 (1.1%)	11415/125 (1.1%)	11865/115 (1.0%)

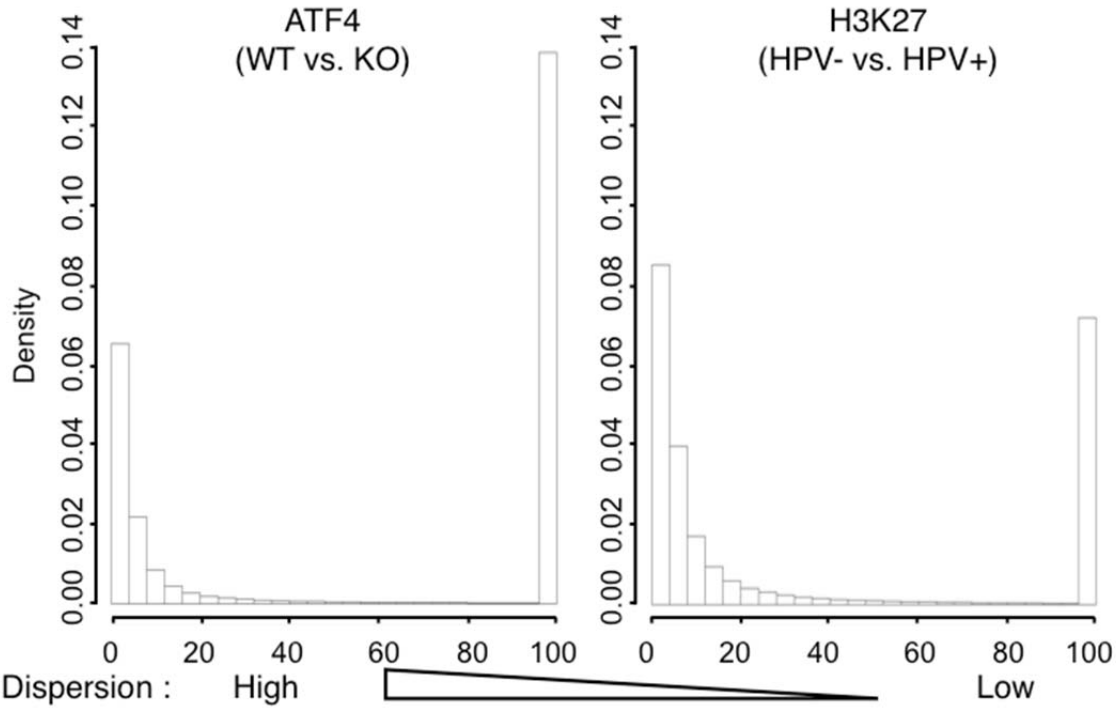


Figure 35. Histograms showing the inverse local dispersion estimates for ATF4 (left) and H3K27me3 (right) data. Dispersion values less than 0.01 are shown grouped at 100.

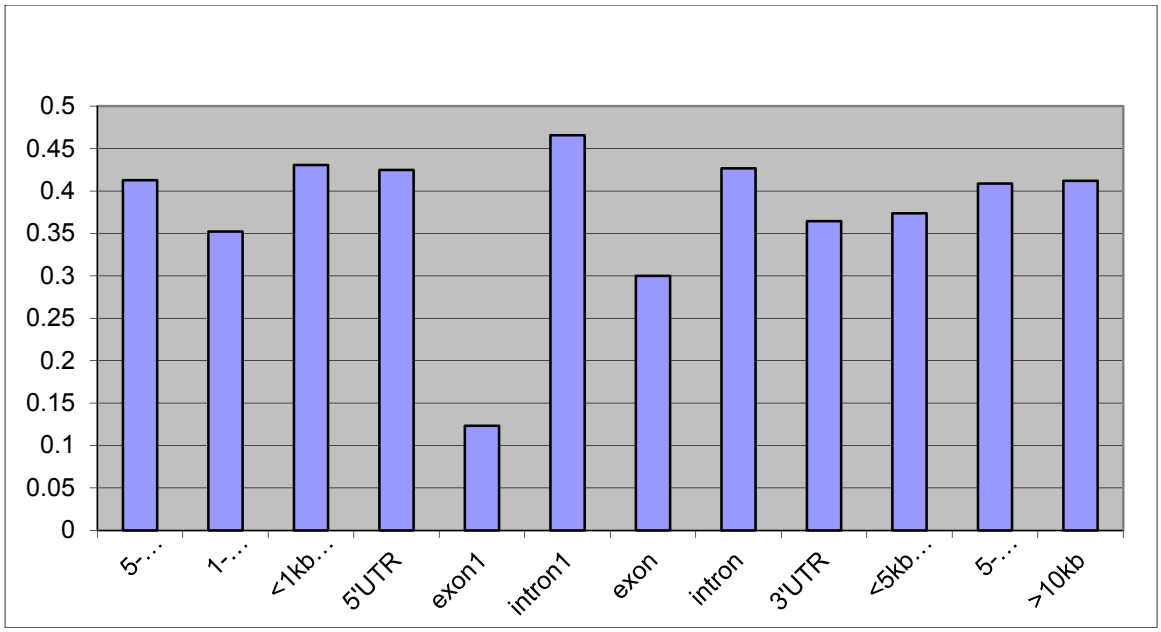


Figure 37. The π_{bj} (marginal likelihood) bin estimates for the twelve bins with ATF4 data. As observed, binding was strongest in the proximal promoter region (<1 kb 5' and 5' UTR) and in introns, although it was not a strong preference towards these regions. Very little binding occurred in first exons.

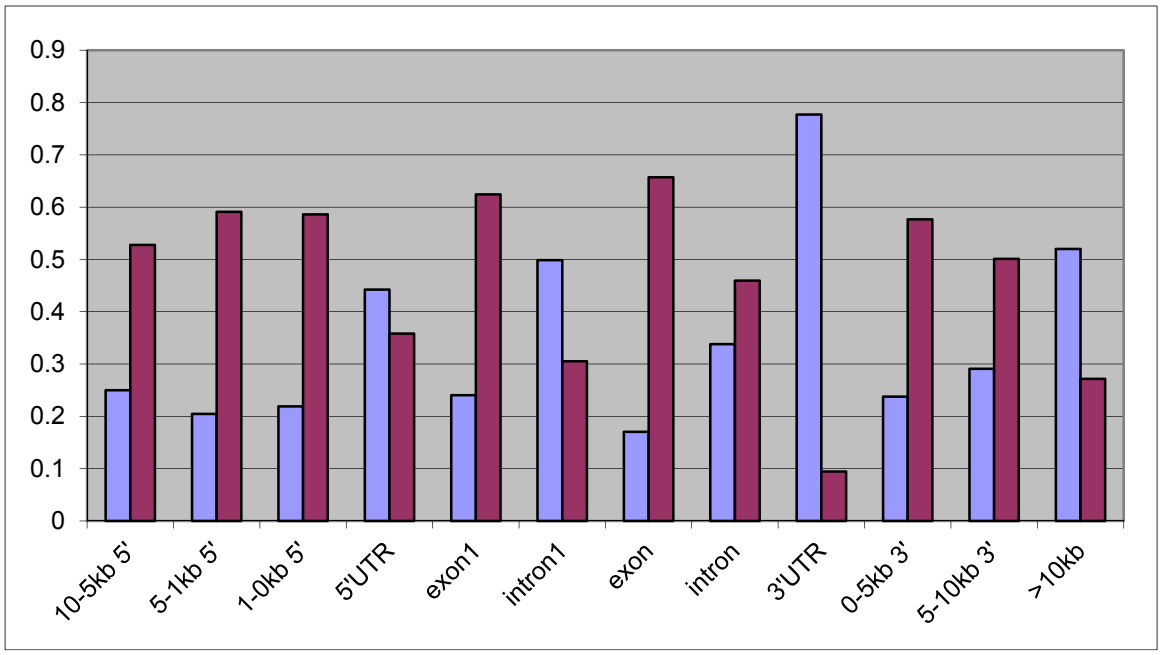


Figure 42. The π_{bj} (marginal likelihood) bin estimates for the twelve bins with H3K27me3 data comparing the HPV(-) to HPV(+) cells. Blue is HPV(+) and purple represents HPV(-). As observed, H3K27 trimethylation is more prominent near transcription start sites and throughout most of the gene body in HPV(-) cells, while it is more prominent in intergenic and certain non-coding regions for HPV(+) cells.

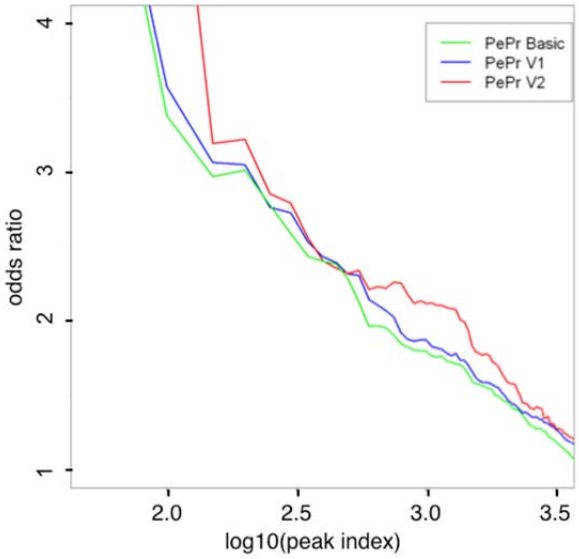


Figure 44. HPV(-) vs HPV(+) odds ratio for enrichment of the overlap between a gene with HPV(-) specific H3K27me3 within 3kb of its TSS and its differential expression (up- or down-regulated). Green line represents our basic implementation, blue line represents PePr V1, and red line represents PePr V2. Consistent with results obtained by restricting to up-regulation in HPV(+) (see Figure 43), overall, PePr V2 identified targets better correlate with differential expression.

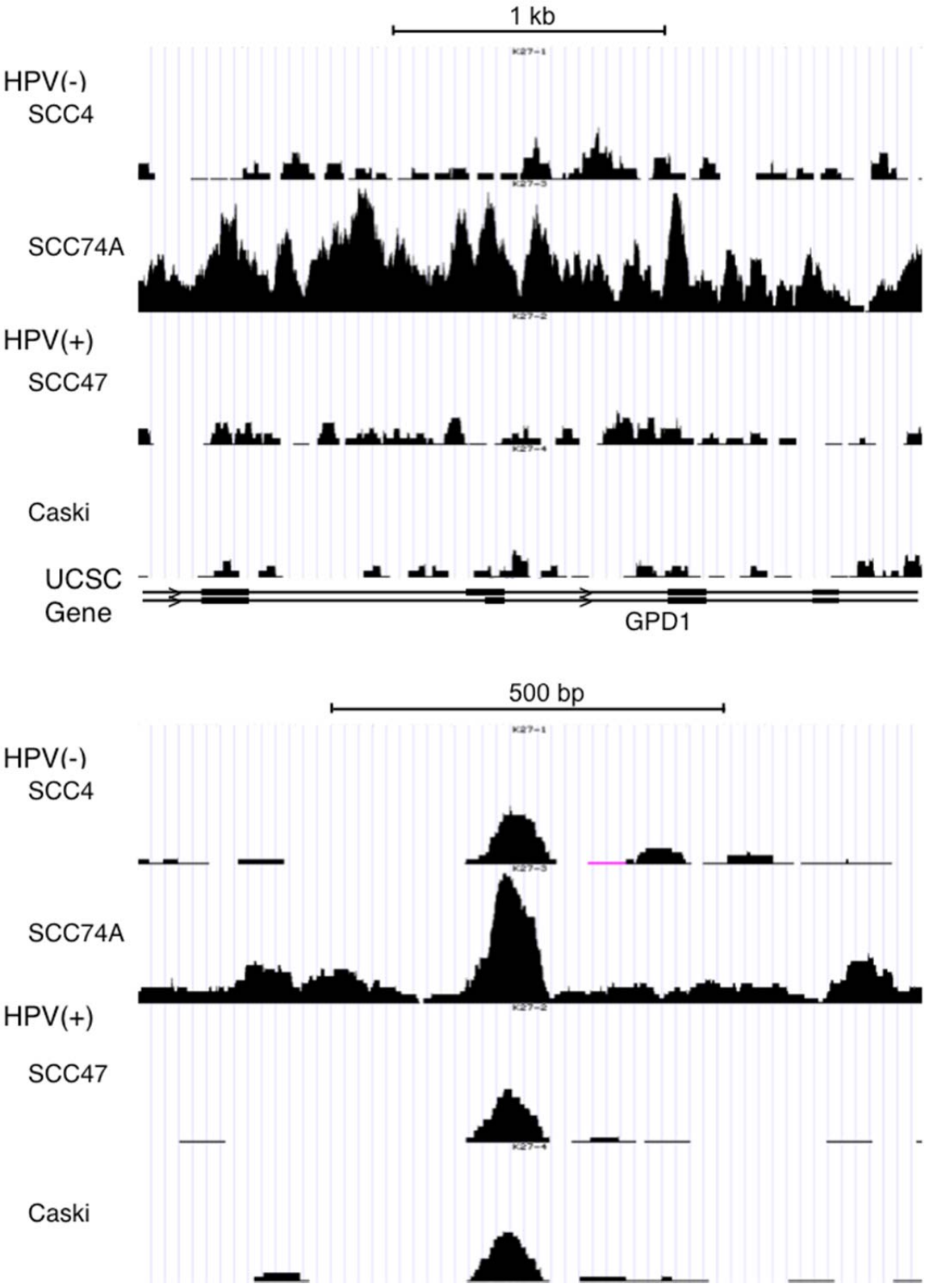


Figure 46. Two H3K27me3 HPV(-) specific peak regions found by the basic, current approach, but not by PePr V1 and V2. A peak was identified in both HPV(-) cell lines, but

neither HPV(+) cell lines. (Top) chr12:50498119-50501003: This figure illustrates that the current approach may detect regions where one HPV(-) cell ine (SCC74A) has nearly the same read profile as one of the HPV(+) cell lines (SCC47) due to discretizing them as (peak/no peak). (Bottom) chrY:330500-331500: Similar to (Top), but with HPV(-) SCC4 being similar to HPV(+) SCC47.

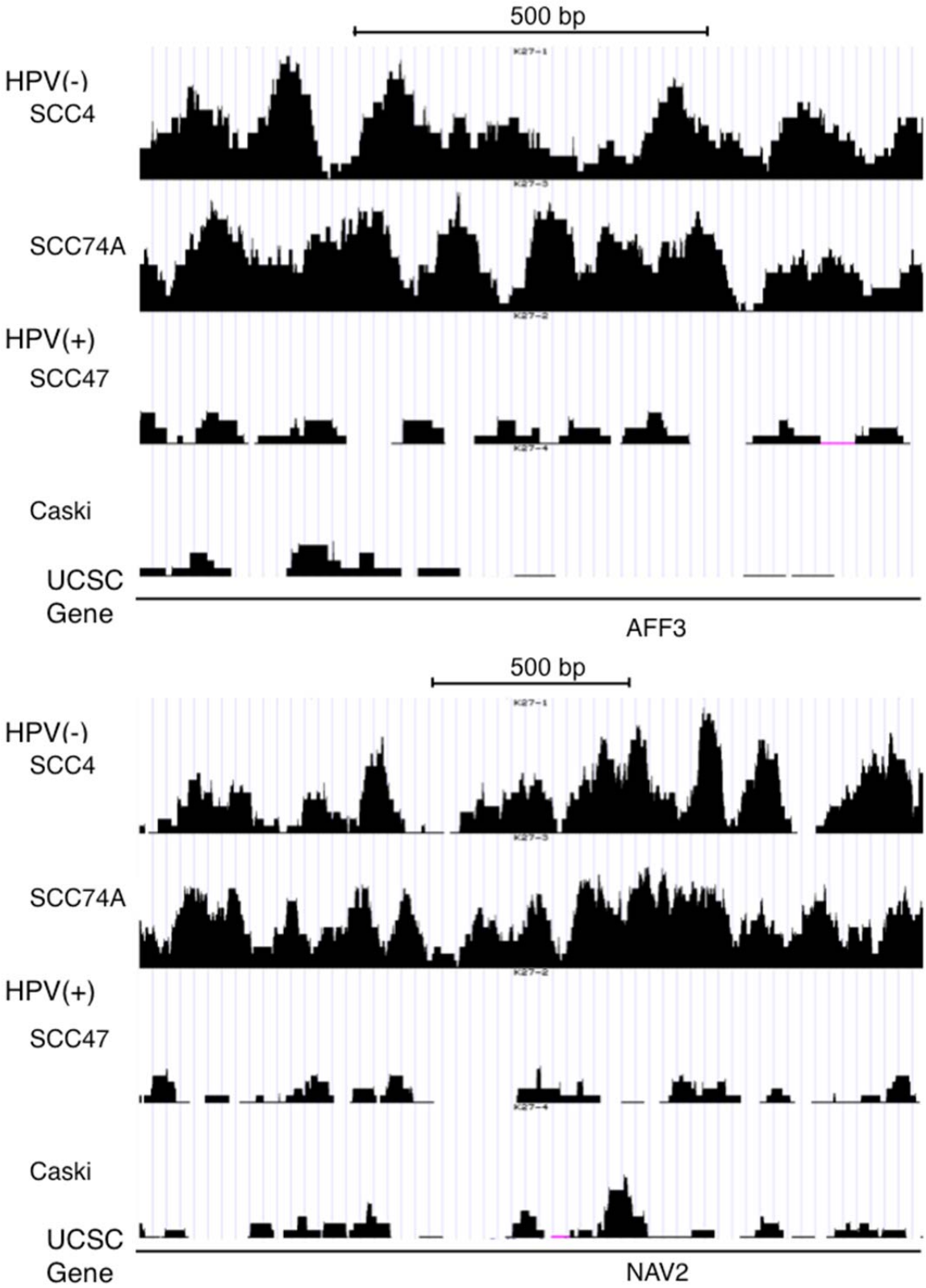


Figure 48. Two H3K27me3 HPV(-) specific peaks found by PePr V1 and PePr V2, but not by using the basic, current approach. (Top) chr2:100218348-100219457 and (Bottom) chr11:19529562-19531559. In both cases, it is clear that the two HPV(-) cell lines have a

significantly stronger peak profile than either HPV(+) cell lines.

Chapter VI

Conclusions

A. SUMMARY OF THESIS WORK

Numerous genome annotations exist to document genomic characteristics. These annotations could serve as helpful auxiliary tools for bioinformatics analyses and computational software development to better refine targets in multiple types of high-throughput experiments. For example, pivotal functional regulatory elements are thought to be conserved through evolution. By identifying highly conserved sequence segments in model organisms, regulatory elements can be better prioritized and studied. In addition, high throughput experiments generate a myriad of data, and incorporating external annotations may help distinguish the desired targets from background noise.

In this thesis work, I first described the utilization of external annotations, such as phylogenetic conservation and human DNAse I hypersensitive sites, to help to predict a 7.1 kb syntenic region that is conserved between human and mouse that ultimately was determined to serve as a *Gata3* enhancer in T lymphocytes. I also performed *in vivo* imaging to confirm its T cell-specific stimulatory activity on a *Gata3* reporter gene.

ChIP-Seq datasets of TR4 in the four ENCODE consortium cell lines were analyzed to characterize its *in vivo* binding preference (gene proximal binding at

DR1 sequence motif), to predict the broader biological process in RNA processing in which it may be involved, and to propose at a *cis* module of TR4 with ETS transcription factor ELK4 can commonly be identified.

From the analysis of RNA-Seq data derived from differentiating human hematopoietic progenitor cells, I confirmed the data reproducibility, and the differential expression patterns of several known erythroid genes were consistent with previous array-based studies. Analysis results also suggested a list of potential novel erythroid regulatory factors, and revealed potential novel isoforms of known erythroid regulatory proteins.

Finally, I developed a ChIP-Seq analysis software to take into account biological variation among relicates. To better orient biologists for downstream experimental design, the software pipeline also incorporates external annotation, specifically the binding relative to gene structure information, to better capture the relationship between physical binding and functional regulatory events.

B. FUTURE DIRECTIONS

i. Chapter II

The 7.1 kbp *Gata3* T cell enhancer is now under intensive investigation and further dissection by Dr. Sakie Hosoya-Ohmura. She continues to refine the position of the minimal sequences required within the 7.1 kbp fragment to confer the T cell-specific stimulatory activity, to determine the factors that bind within this minimal enhancer to confer the stimulatory activity, and to determine whether this enhancer plays a role in T cell lymphoma that can be caused by ectopic *Gata3* expression.

ii. Chapter III

While a significant fraction of *in vivo* TR4 binding sites in the four ENCODE cell lines was observed to also contain an ETS binding motif (to which the ETS factor ELK4 was bound in one of the cell lines), questions regarding the *cis* module involving TR4 and ETS transcription factors remain unanswered. Future work is required to investigate whether other ETS factors also cooperate with TR4 function. An expression level screen of ETS family factors to exclude low abundant ETS factors could be an initial step to address this question by surveying existing expression databases, such as through UCSC Genome Browser.

iii. Chapter IV

While RNA-Seq data from adult CD34+ hematopoietic progenitor cells were analyzed, the difference between the adult CD34 transcriptome and the fetal CD34 transcriptome remain unclear. Comparison between the two would shed light on the maturation and development of red blood cells. In addition, it would be interesting to verify whether the predicted novel isoforms of known erythroid regulatory proteins do play roles in erythropoiesis.

iv. Chapter V

Chapter V introduced a novel peak prioritization pipeline (PePr) for ChIP-Seq analysis. One area for future improvement is to boost the software running time; this could partially be accomplished by utilizing a file containing precomputed

genomic regions for the twelve bins related to gene structure. The benefit from the incorporation of binding relative to gene structure was demonstrated; future work is required to determine to what extent incorporating additional annotations would improve the analysis results for transcription factors and/or epigenomic studies, such as for histone modifications and DNA methylation.

Bibliography

Aban, I.B., Cutter, G.R. and Mavinga, N. (2008) Inferences and Power Analysis Concerning Two Negative Binomial Distributions with An Application to MRI Lesion Counts Data, *Comput Stat Data Anal*, **53**, 820-833.

Amsen, D.*, et al.* (2007) Direct regulation of Gata3 expression determines the T helper differentiation potential of Notch, *Immunity*, **27**, 89-99.

Anderson, M.K.*, et al.* (2002) Definition of regulatory network elements for T cell development by perturbation analysis with PU.1 and GATA-3, *Dev Biol*, **246**, 103-121.

Asnagli, H., Afkarian, M. and Murphy, K.M. (2002) Cutting edge: Identification of an alternative GATA-3 promoter directing tissue-specific gene expression in mouse and human, *J Immunol*, **168**, 4268-4271.

Asselin-Labat, M.L.*, et al.* (2007) Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation, *Nat Cell Biol*, **9**, 201-209.

Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc Int Conf Intell Syst Mol Biol*, **2**, 28-36.

Ballas, N.*, et al.* (2001) Regulation of neuronal traits by a novel transcriptional complex, *Neuron*, **31**, 353-365.

Ballas, N.*, et al.* (2005) REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis, *Cell*, **121**, 645-657.

Barski, A.*, et al.* (2007) High-resolution profiling of histone methylations in the human genome, *Cell*, **129**, 823-837.

Battaglioli, E.*, et al.* (2002) REST repression of neuronal genes requires components of the hSWI.SNF complex, *J Biol Chem*, **277**, 41038-41045.

Blahnik, K.R.*, et al.* (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data, *Nucleic Acids Res*, **38**, e13.

Boehm, T.*, et al.* (1991) The rhombotin family of cysteine-rich LIM-domain oncogenes: distinct members are involved in T-cell translocations to human chromosomes 11p15 and 11p13, *Proc Natl Acad Sci U S A*, **88**, 4367-4371.

Bookout, A.L.*, et al.* (2006) Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network, *Cell*, **126**, 789-799.

Boros, J.*, et al.* (2009) Elucidation of the ELK1 target gene network reveals a role in the coordinate regulation of core components of the gene regulation machinery, *Genome Res*, **19**, 1963-1973.

Boros, J.*, et al.* (2009) Overlapping promoter targeting by Elk-1 and other divergent ETS-domain transcription factor family members, *Nucleic Acids Res*, **37**, 7368-7380.

Boyle, A.P.*, et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome, *Cell*, **132**, 311-322.

Boyle, A.P.*, et al.* (2008) F-Seq: a feature density estimator for high-throughput sequence tags, *Bioinformatics*, **24**, 2537-2538.

Bryne, J.C.*, et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update, *Nucleic Acids Res*, **36**, D102-106.

Bunn, H.F. (1997) Pathogenesis and treatment of sickle cell disease, *N Engl J Med*, **337**, 762-769.

Cantu, C.*, et al.* (2011) Sox6 enhances erythroid differentiation in human erythroid progenitors, *Blood*, **117**, 3669-3679.

Cartharius, K.*, et al.* (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites, *Bioinformatics*, **21**, 2933-2942.

Chang, C.*, et al.* (1994) Human and rat TR4 orphan receptors specify a subclass of the steroid receptor superfamily, *Proc Natl Acad Sci U S A*, **91**, 6040-6044.

Chen, D. and Zhang, G. (2001) Enforced expression of the GATA-3 transcription factor affects cell fate decisions in hematopoiesis, *Exp Hematol*, **29**, 971-980.

Chen, L.M.*, et al.* (2008) Subfertility with defective folliculogenesis in female mice lacking testicular orphan nuclear receptor 4, *Mol Endocrinol*, **22**, 858-867.

Cloonan, N.*, et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing, *Nat Methods*, **5**, 613-619.

Cohen-Barak, O.*, et al.* (2007) Stem cell transplantation demonstrates that Sox6 represses epsilon y globin expression in definitive erythropoiesis of adult mice, *Exp Hematol*, **35**, 358-367.

Crawford, G.E.*, et al.* (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays, *Nat Methods*, **3**, 503-509.

Crawford, G.E.*, et al.* (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites, *Proc Natl Acad Sci U S A*, **101**, 992-997.

Crawford, G.E.*, et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS), *Genome Res*, **16**, 123-131.

David-Fung, E.S.*, et al.* (2006) Progression of regulatory gene expression states in fetal and adult pro-T-cell development, *Immunol Rev*, **209**, 212-236.

Dierks, C.*, et al.* (2008) Expansion of Bcr-Abl-positive leukemic stem cells is dependent on Hedgehog pathway activation, *Cancer Cell*, **14**, 238-249.

Dumitriu, B.*, et al.* (2010) Sox6 is necessary for efficient erythropoiesis in adult mice under physiological and anemia-induced stress conditions, *PLoS One*, **5**, e12088.

Dumitriu, B.*, et al.* (2006) Sox6 cell-autonomously stimulates erythroid cell survival, proliferation, and terminal maturation and is thereby an important enhancer of definitive erythropoiesis during mouse development, *Blood*, **108**, 1198-1207.

Fang, T.C.*, et al.* (2007) Notch directly regulates Gata3 expression during T helper 2 cell differentiation, *Immunity*, **27**, 100-110.

Farnham, P.J. (2009) Insights from genomic profiling of transcription factors, *Nat Rev Genet*, **10**, 605-616.

Fejes, A.P.*, et al.* (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology, *Bioinformatics*, **24**, 1729-1730.

Feng, W.*, et al.* (2008) A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology, *BMC Genomics*, **9 Suppl 2**, S23.

FitzGerald, P.C., *et al.* (2004) Clustering of DNA sequences in human promoters, *Genome Res*, **14**, 1562-1574.

Foroni, L., *et al.* (1992) The rhombotin gene family encode related LIM-domain proteins whose differing expression suggests multiple roles in mouse development, *J Mol Biol*, **226**, 747-761.

Fry, C.J., *et al.* (1999) Activation of the murine dihydrofolate reductase promoter by E2F1. A requirement for CBP recruitment, *J Biol Chem*, **274**, 15883-15891.

Fujiwara, T., *et al.* (2009) Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy, *Mol Cell*, **36**, 667-681.

Gao, J., *et al.* (2009) Hedgehog signaling is dispensable for adult hematopoietic stem cell function, *Cell Stem Cell*, **4**, 548-558.

George, K.M., *et al.* (1994) Embryonic expression and cloning of the murine GATA-3 gene, *Development*, **120**, 2673-2686.

Giarratana, M.C., *et al.* (2005) Ex vivo generation of fully mature human red blood cells from hematopoietic stem cells, *Nat Biotechnol*, **23**, 69-74.

Gomez-Valades, A.G., *et al.* (2006) Overcoming diabetes-induced hyperglycemia through inhibition of hepatic phosphoenolpyruvate carboxykinase (GTP) with RNAi, *Mol Ther*, **13**, 401-410.

Grigorieva, I.V., *et al.* (2010) Gata3-deficient mice develop parathyroid abnormalities due to dysregulation of the parathyroid-specific transcription factor Gcm2, *J Clin Invest*, **120**, 2144-2155.

Gupta, S., *et al.* (2007) Quantifying similarity between motifs, *Genome Biol*, **8**, R24.

Hagiwara, N.*, et al.* (2000) Sox6 is a candidate gene for p100H myopathy, heart block, and sudden neonatal death, *Proc Natl Acad Sci U S A*, **97**, 4180-4185.

Hagiwara, N., Ma, B. and Ly, A. (2005) Slow and fast fiber isoform gene expression is systematically altered in skeletal muscle of the Sox6 mutant, p100H, *Dev Dyn*, **234**, 301-311.

Hakimi, M.A.*, et al.* (2002) A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes, *Proc Natl Acad Sci U S A*, **99**, 7420-7425.

Hakimi, M.A.*, et al.* (2003) A candidate X-linked mental retardation gene is a component of a new family of histone deacetylase-containing complexes, *J Biol Chem*, **278**, 7234-7239.

Hamada-Kanazawa, M.*, et al.* (2004) Sox6 overexpression causes cellular aggregation and the neuronal differentiation of P19 embryonic carcinoma cells in the absence of retinoic acid, *FEBS Lett*, **560**, 192-198.

Hamada-Kanazawa, M.*, et al.* (2004) Suppression of Sox6 in P19 cells leads to failure of neuronal differentiation by retinoic acid and induces retinoic acid-dependent apoptosis, *FEBS Lett*, **577**, 60-66.

Harris, L.G., Samant, R.S. and Shevde, L.A. (2011) Hedgehog signaling: networking to nurture a promalignant tumor microenvironment, *Mol Cancer Res*, **9**, 1165-1174.

Hasegawa, S.L.*, et al.* (2007) Dosage-dependent rescue of definitive nephrogenesis by a distant Gata3 enhancer, *Dev Biol*, **301**, 568-577.

Hendriks, R.W., *et al.* (1999) Expression of the transcription factor GATA-3 is required for the development of the earliest T cell progenitors and correlates with stages of cellular proliferation in the thymus, *Eur J Immunol*, **29**, 1912-1918.

Hernandez-Hoyos, G., *et al.* (2003) GATA-3 expression is controlled by TCR signals and regulates CD4/CD8 differentiation, *Immunity*, **19**, 83-94.

Hofmann, I., *et al.* (2009) Hedgehog signaling is dispensable for adult murine hematopoietic stem cell function and hematopoiesis, *Cell Stem Cell*, **4**, 559-567.

Hosoya, T., *et al.* (2009) GATA-3 is required for early T lineage progenitor development, *J Exp Med*, **206**, 2987-3000.

Hosoya, T., Maillard, I. and Engel, J.D. (2010) From the cradle to the grave: activities of GATA-3 throughout T-cell development and differentiation, *Immunol Rev*, **238**, 110-125.

Humphrey, G.W., *et al.* (2001) Stable histone deacetylase complexes distinguished by the presence of SANT domain proteins CoREST/kiaa0071 and Mta-L1, *J Biol Chem*, **276**, 6817-6824.

Huq, M.D., *et al.* (2006) Modulation of testicular receptor 4 activity by mitogen-activated protein kinase-mediated phosphorylation, *Mol Cell Proteomics*, **5**, 2072-2082.

Hyland, P.L., *et al.* (2011) Evidence for alteration of EZH2, BMI1, and KDM6A and epigenetic reprogramming in human papillomavirus type 16 E6/E7-expressing keratinocytes, *J Virol*, **85**, 10999-11006.

Ikeda, T., *et al.* (2004) The combination of SOX5, SOX6, and SOX9 (the SOX trio) provides signals sufficient for induction of permanent cartilage, *Arthritis Rheum*, **50**, 3561-3573.

Ingham, P.W. and McMahon, A.P. (2001) Hedgehog signaling in animal development: paradigms and principles, *Genes Dev*, **15**, 3059-3087.

Ji, H., *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data, *Nat Biotechnol*, **26**, 1293-1300.

Jin, V.X., *et al.* (2009) W-ChIPMotifs: a web application tool for de novo motif discovery from ChIP-based high-throughput data, *Bioinformatics*, **25**, 3191-3193.

Jin, V.X., *et al.* (2006) A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data--a case study using E2F1, *Genome Res*, **16**, 1585-1595.

Johnson, D.S., *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions, *Science*, **316**, 1497-1502.

Jothi, R., *et al.* (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data, *Nucleic Acids Res*, **36**, 5221-5231.

Kadri, Z., *et al.* (2009) Direct binding of pRb/E2F-2 to GATA-1 regulates maturation and terminal cell division during erythropoiesis, *PLoS Biol*, **7**, e1000123.

Kanehisa, M., *et al.* (2008) KEGG for linking genomes to life and the environment, *Nucleic Acids Res*, **36**, D480-484.

Kaufman, C.K., *et al.* (2003) GATA-3: an unexpected regulator of cell lineage determination in skin, *Genes Dev*, **17**, 2108-2122.

Kaufmann, K.*, et al.* (2009) Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower, *PLoS Biol*, **7**, e1000090.

Keller, M.A.*, et al.* (2006) Transcriptional regulatory network analysis of developing human erythroid progenitors reveals patterns of coregulation and potential transcriptional regulators, *Physiol Genomics*, **28**, 114-128.

Khandekar, M.*, et al.* (2004) Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system, *Mol Cell Biol*, **24**, 10263-10276.

Kharchenko, P.V., Tolstorukov, M.Y. and Park, P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins, *Nat Biotechnol*, **26**, 1351-1359.

Kim, E.*, et al.* (2005) Induction of apolipoprotein E expression by TR4 orphan nuclear receptor via 5' proximal promoter region, *Biochem Biophys Res Commun*, **328**, 85-90.

Kinross, K.M.*, et al.* (2006) E2f4 regulates fetal erythropoiesis through the promotion of cellular proliferation, *Blood*, **108**, 886-895.

Ko, L.J.*, et al.* (1991) Murine and human T-lymphocyte GATA-3 factors mediate transcription through a cis-regulatory element within the human T-cell receptor delta gene enhancer, *Mol Cell Biol*, **11**, 2778-2784.

Kouros-Mehr, H.*, et al.* (2006) GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland, *Cell*, **127**, 1041-1055.

Kurek, D.*, et al.* (2007) Transcriptome and phenotypic analysis reveals Gata3-dependent signalling pathways in murine hair follicles, *Development*, **134**, 261-272.

Lakshmanan, G.*, et al.* (1998) Partial rescue of GATA-3 by yeast artificial chromosome transgenes, *Dev Biol*, **204**, 451-463.

Lakshmanan, G.*, et al.* (1999) Localization of distant urogenital system-, central nervous system-, and endocardium-specific transcriptional regulatory elements in the GATA-3 locus, *Mol Cell Biol*, **19**, 1558-1568.

Langmead, B.*, et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol*, **10**, R25.

Lee, C.H., Chinpaisal, C. and Wei, L.N. (1998) A novel nuclear receptor heterodimerization pathway mediated by orphan receptors TR2 and TR4, *J Biol Chem*, **273**, 25209-25215.

Lee, E.C.*, et al.* (2001) A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA, *Genomics*, **73**, 56-65.

Lee, W.*, et al.* (2007) A high-resolution atlas of nucleosome occupancy in yeast, *Nat Genet*, **39**, 1235-1244.

Lee, Y.F., Lee, H.J. and Chang, C. (2002) Recent advances in the TR2 and TR4 orphan receptors of the nuclear receptor superfamily, *J Steroid Biochem Mol Biol*, **81**, 291-308.

Lee, Y.F.*, et al.* (1997) Identification of direct repeat 4 as a positive regulatory element for the human TR4 orphan receptor. A modulator for the thyroid hormone target genes, *J Biol Chem*, **272**, 12215-12220.

Lee, Y.F., *et al.* (1998) Negative feedback control of the retinoid-retinoic acid/retinoid X receptor pathway by the human TR4 orphan receptor, a member of the steroid receptor superfamily, *J Biol Chem*, **273**, 13437-13443.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.

Liang, K. and Keles, S. (2012) Detecting differential binding of transcription factors with ChIP-seq, *Bioinformatics*, **28**, 121-122.

Lieuw, K.H., *et al.* (1997) Temporal and spatial control of murine GATA-3 transcription by promoter-proximal regulatory elements, *Dev Biol*, **188**, 1-16.

Lim, K.C., *et al.* (2000) Gata3 loss leads to embryonic lethality due to noradrenaline deficiency of the sympathetic nervous system, *Nat Genet*, **25**, 209-212.

Lim, Y. and Matsui, W. (2010) Hedgehog signaling in hematopoiesis, *Crit Rev Eukaryot Gene Expr*, **20**, 129-139.

Lister, R., *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis, *Cell*, **133**, 523-536.

Liu, N.C., *et al.* (2007) Loss of TR4 orphan nuclear receptor reduces phosphoenolpyruvate carboxykinase-mediated gluconeogenesis, *Diabetes*, **56**, 2901-2909.

Mangelsdorf, D.J., *et al.* (1995) The nuclear receptor superfamily: the second decade, *Cell*, **83**, 835-839.

Mar, B.G., *et al.* (2011) The controversial role of the Hedgehog pathway in normal and malignant hematopoiesis, *Leukemia*, **25**, 1665-1673.

Maurice, D.*, et al.* (2007) c-Myb regulates lineage choice in developing thymocytes via its target gene Gata3, *EMBO J*, **26**, 3629-3640.

Merryweather-Clarke, A.T.*, et al.* (2011) Global gene expression analysis of human erythroid progenitors, *Blood*, **117**, e96-108.

Metzker, M.L. (2010) Sequencing technologies - the next generation, *Nat Rev Genet*, **11**, 31-46.

Mikkelsen, T.S.*, et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature*, **448**, 553-560.

Moriguchi, T.*, et al.* (2006) Gata3 participates in a complex transcriptional feedback network to regulate sympathoadrenal differentiation, *Development*, **133**, 3871-3881.

Mortazavi, A.*, et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat Methods*, **5**, 621-628.

Nagalakshmi, U.*, et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science*, **320**, 1344-1349.

Nam, C.H.*, et al.* (2008) An antibody inhibitor of the LMO2-protein complex blocks its normal and tumorigenic functions, *Oncogene*, **27**, 4962-4968.

Nawijn, M.C.*, et al.* (2001) Enforced expression of GATA-3 in transgenic mice inhibits Th1 differentiation and induces the formation of a T1/ST2-expressing Th2-committed T cell compartment in vivo, *J Immunol*, **167**, 724-732.

Nawijn, M.C.*, et al.* (2001) Enforced expression of GATA-3 during T cell development inhibits maturation of CD8 single-positive cells and induces thymic lymphoma in transgenic mice, *J Immunol*, **167**, 715-723.

Nix, D.A., Courdy, S.J. and Boucher, K.M. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks, *BMC Bioinformatics*, **9**, 523.

Noy, N. (2007) Ligand specificity of nuclear hormone receptors: sifting through promiscuity, *Biochemistry*, **46**, 13461-13467.

Nusslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in Drosophila, *Nature*, **287**, 795-801.

O'Donnell, A., Yang, S.H. and Sharrocks, A.D. (2008) MAP kinase-mediated c-fos regulation relies on a histone acetylation relay switch, *Mol Cell*, **29**, 780-785.

O'Geen, H., Frietze, S. and Farnham, P.J. (2010) Using ChIP-seq technology to identify targets of zinc finger transcription factors, *Methods Mol Biol*, **649**, 437-455.

Okazaki, Y.*, et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs, *Nature*, **420**, 563-573.

Omori, A.*, et al.* (2005) Adult stage gamma-globin silencing is mediated by a promoter direct repeat element, *Mol Cell Biol*, **25**, 3443-3451.

Oram, S.H.*, et al.* (2010) A previously unrecognized promoter of LMO2 forms part of a transcriptional regulatory circuit mediating LMO2 expression in a subset of T-acute lymphoblastic leukaemia patients, *Oncogene*, **29**, 5796-5808.

Pai, S.Y., Truitt, M.L. and Ho, I.C. (2004) GATA-3 deficiency abrogates the development and maintenance of T helper type 2 cells, *Proc Natl Acad Sci U S A*, **101**, 1993-1998.

Pai, S.Y.*, et al.* (2003) Critical roles for transcription factor GATA-3 in thymocyte development, *Immunity*, **19**, 863-875.

Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology, *Nat Rev Genet*, **10**, 669-680.

Peller, S.*, et al.* (2009) Identification of gene networks associated with erythroid differentiation, *Blood Cells Mol Dis*, **43**, 74-80.

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

Perry, J.M.*, et al.* (2009) Maintenance of the BMP4-dependent stress erythropoiesis pathway in the murine spleen requires hedgehog signaling, *Blood*, **113**, 911-918.

Qin, Z.S.*, et al.* (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data, *BMC Bioinformatics*, **11**, 369.

Rashid, N.U.*, et al.* (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions, *Genome Biol*, **12**, R67.

Rashid, N.U.*, et al.* (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions, *Genome Biol*, **12**, R67.

Ravasi, T.*, et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man, *Cell*, **140**, 744-752.

Robertson, G.*, et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, *Nat Methods*, **4**, 651-657.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139-140.

Royer-Pokora, B., Loos, U. and Ludwig, W.D. (1991) TTG-2, a new gene encoding a cysteine-rich protein with the LIM motif, is overexpressed in acute T-cell leukaemia with the t(11;14)(p13;q11), *Oncogene*, **6**, 1887-1893.

Rozowsky, J.*, et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls, *Nat Biotechnol*, **27**, 66-75.

Rutkowski, D.T.*, et al.* (2006) Adaptation to ER stress is mediated by differential stabilities of pro-survival and pro-apoptotic mRNAs and proteins, *PLoS Biol*, **4**, e374.

Saleque, S.*, et al.* (2007) Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1, *Mol Cell*, **27**, 562-572.

Sambandam, A.*, et al.* (2005) Notch signaling controls the generation and differentiation of early T lineage progenitors, *Nat Immunol*, **6**, 663-670.

Samson, S.I.*, et al.* (2003) GATA-3 promotes maturation, IFN-gamma production, and liver-specific homing of NK cells, *Immunity*, **19**, 701-711.

Sandelin, A. and Wasserman, W.W. (2005) Prediction of nuclear hormone receptor response elements, *Mol Endocrinol*, **19**, 595-606.

Sartor, M.A.*, et al.* (2011) Genome-wide methylation and expression differences in HPV(+) and HPV(-) squamous cell carcinoma cell lines are consistent with divergent mechanisms of carcinogenesis, *Epigenetics*, **6**, 777-787.

Sartor, M.A., *et al.* (2010) ConceptGen: a gene set enrichment and gene set relation mapping tool, *Bioinformatics*, **26**, 456-463.

Sartor, M.A., *et al.* (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments, *BMC Bioinformatics*, **7**, 538.

Shi, Y., *et al.* (2003) Coordinated histone modifications mediated by a CtBP co-repressor complex, *Nature*, **422**, 735-738.

Shyr, C.R., *et al.* (2009) Roles of testicular orphan nuclear receptors 2 and 4 in early embryonic development and embryonic stem cells, *Endocrinology*, **150**, 2454-2462.

Singleton, B.K., *et al.* (2008) Mutations in EKLF/KLF1 form the molecular basis of the rare blood group In(Lu) phenotype, *Blood*, **112**, 2081-2088.

Sripichai, O., *et al.* (2009) Cytokine-mediated increases in fetal hemoglobin are associated with globin gene histone modification and transcription factor reprogramming, *Blood*, **114**, 2299-2306.

Stolt, C.C., *et al.* (2006) SoxD proteins influence multiple stages of oligodendrocyte development and modulate SoxE protein function, *Dev Cell*, **11**, 697-709.

Su Liu, S.X., Yi-fen Lee and Chawnshang Chang (2010) Physiological Functions of TR2 and TR4 Orphan Nuclear Receptor. In, *Nuclear Receptors: Current Concepts and Future Challenges*. Springer Netherlands, pp. 327-343.

Su, N. and Kilberg, M.S. (2008) C/EBP homology protein (CHOP) interacts with activating transcription factor 4 (ATF4) and negatively regulates the stress-dependent induction of the asparagine synthetase gene, *J Biol Chem*, **283**, 35106-35117.

Taghon, T.*, et al.* (2001) Enforced expression of GATA-3 severely reduces human thymic cellularity, *J Immunol*, **167**, 4468-4475.

Taghon, T., Yui, M.A. and Rothenberg, E.V. (2007) Mast cell lineage diversion of T lineage precursors by the essential T cell transcription factor GATA-3, *Nat Immunol*, **8**, 845-855.

Tanabe, O.*, et al.* (2002) An embryonic/fetal beta-type globin gene repressor contains a nuclear receptor TR2/TR4 heterodimer, *EMBO J*, **21**, 3434-3442.

Tanabe, O.*, et al.* (2007) Embryonic and fetal beta-globin gene repression by the orphan nuclear receptors, TR2 and TR4, *EMBO J*, **26**, 2295-2306.

Tanabe, O.*, et al.* (2007) The TR2 and TR4 orphan nuclear receptors repress Gata1 transcription, *Genes Dev*, **21**, 2832-2844.

Taylor, J.*, et al.* (2006) ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements, *Genome Res*, **16**, 1596-1604.

Terano, T.*, et al.* (2005) Transcriptional control of fetal liver hematopoiesis: dominant negative effect of the overexpression of the LIM domain mutants of LMO2, *Exp Hematol*, **33**, 641-651.

Tondeur, S.*, et al.* (2010) Expression map of the human exome in CD34+ cells and blood cells: increased alternative splicing in cell motility and immune response genes, *PLoS One*, **5**, e8990.

Tong, J.K.*, et al.* (1998) Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex, *Nature*, **395**, 917-921.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105-1111.

Trapnell, C.*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nat Biotechnol*, **28**, 511-515.

Tsai, N.P.*, et al.* (2009) Activation of testicular orphan receptor 4 by fatty acids, *Biochim Biophys Acta*, **1789**, 734-740.

Tsarovina, K.*, et al.* (2010) The Gata3 transcription factor is required for the survival of embryonic and adult sympathetic neurons, *J Neurosci*, **30**, 10833-10843.

Tuteja, G.*, et al.* (2009) Extracting transcription factor targets from ChIP-Seq data, *Nucleic Acids Res*, **37**, e113.

Tydell, C.C.*, et al.* (2007) Molecular dissection of prethymic progenitor entry into the T lymphocyte developmental pathway, *J Immunol*, **179**, 421-438.

Valera, A.*, et al.* (1994) Transgenic mice overexpressing phosphoenolpyruvate carboxykinase develop non-insulin-dependent diabetes mellitus, *Proc Natl Acad Sci U S A*, **91**, 9151-9154.

Valouev, A.*, et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data, *Nat Methods*, **5**, 829-834.

van Hamburg, J.P.*, et al.* (2008) Cooperation of Gata3, c-Myc and Notch in malignant transformation of double positive thymocytes, *Mol Immunol*, **45**, 3085-3095.

Vaquerizas, J.M.*, et al.* (2009) A census of human transcription factors: function, expression and evolution, *Nat Rev Genet*, **10**, 252-263.

Visvader, J.E.*, et al.* (1997) The LIM-domain binding protein Ldb1 and its partner LMO2 act as negative regulators of erythroid differentiation, *Proc Natl Acad Sci U S A*, **94**, 13707-13712.

Vosshenrich, C.A.*, et al.* (2006) A thymic pathway of mouse natural killer cell development characterized by expression of GATA-3 and CD127, *Nat Immunol*, **7**, 1217-1224.

Warren, A.J.*, et al.* (1994) The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development, *Cell*, **78**, 45-57.

White, R.*, et al.* (2008) Role of RIP140 in metabolic tissues: connections to disease, *FEBS Lett*, **582**, 39-45.

Wilbanks, E.G. and Facciotti, M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection, *PLoS One*, **5**, e11471.

Wilhelm, B.T.*, et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, *Nature*, **453**, 1239-1243.

Wu, J.Q.*, et al.* (2010) Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing, *Proc Natl Acad Sci U S A*, **107**, 5254-5259.

Xi, H.*, et al.* (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome, *PLoS Genet*, **3**, e136.

Xie, S.*, et al.* (2009) TR4 nuclear receptor functions as a fatty acid sensor to modulate CD36 expression and foam cell formation, *Proc Natl Acad Sci U S A*, **106**, 13353-13358.

Xu, W. and Kee, B.L. (2007) Growth factor independent 1B (Gfi1b) is an E2A target gene that modulates Gata3 in T-cell lymphomas, *Blood*, **109**, 4406-4414.

Xu, X., *et al.* (2007) A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members, *Genome Res*, **17**, 1550-1561.

Yamamoto, M., *et al.* (1990) Activity and tissue-specific expression of the transcription factor NF-E1 multigene family, *Genes Dev*, **4**, 1650-1662.

Yang, X.O., *et al.* (2009) Requirement for the basic helix-loop-helix transcription factor Dec2 in initial TH2 lineage commitment, *Nat Immunol*, **10**, 1260-1266.

Yi, Z., *et al.* (2006) Sox6 directly silences epsilon globin expression in definitive erythropoiesis, *PLoS Genet*, **2**, e14.

You, A., *et al.* (2001) CoREST is an integral component of the CoREST- human histone deacetylase complex, *Proc Natl Acad Sci U S A*, **98**, 1454-1458.

Yu, Q., *et al.* (2009) T cell factor 1 initiates the T helper type 2 fate by inducing the transcription factor GATA-3 and repressing interferon-gamma, *Nat Immunol*, **10**, 992-999.

Zambelli, F., Pesole, G. and Pavesi, G. (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes, *Nucleic Acids Res*, **37**, W247-252.

Zang, C., *et al.* (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics*, **25**, 1952-1958.

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

Zhang, Y.*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS), *Genome Biol*, **9**, R137.

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

Zhu, J.*, et al.* (2004) Conditional deletion of Gata3 shows its essential function in T(H)1-T(H)2 responses, *Nat Immunol*, **5**, 1157-1165.