BAYESIAN STATISTICAL INFERENCE FOR TUMOR MICROENVIRONMENT COMPOSITIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Tin Yi Chu

December 2019

BAYESIAN STATISTICAL INFERENCE FOR TUMOR MICROENVIRONMENT COMPOSITIONS

Tin Yi Chu, Ph. D.

Cornell University 2019

The complex interaction between tumor and its microenvironment is essential for oncogenesis, survival and growth of tumor. These interactions allow tumor to uptake nutrient from environment and evade from immune surveillances. Understanding these interactions is fundamental to the design of immunotherapies and other targeted therapies. Advances in sequencing technologies have enabled measurement of gene transcription and regulation across large cohorts of cancer patients and also down to the single cell resolution. In this work, using glioblastoma (GBM) as a model system, I present the bioinformatic characterization of tumors and their microenvironment, and the statistical models towards an unsupervised and automated way of understanding the compositions.

The first part describes the new sequencing method, Chromatin Run-on Sequencing (ChRO-seq), and its use in characterizing the transcription regulatory landscape in primary glioblastoma. Taking advantage of the ability for ChRO-seq to quantify nascent RNAs directly from solid tissues, I developed bioinformatic tools called dREG-HD to map the genome-wide positions of transcription regulatory elements (TREs) based on their nascent RNA patterns, which formed the basis for quantifying the enhancer activity. As ChRO-seq also enables simultaneous quantification of transcription activity of genes, I developed the tool tfTarget to map the network formed between transcription factor, TREs and target genes. Using tfTarget I identified tumor-associated transcription modules and regulatory networks

associated with known GBM subtypes. More importantly, I identified three transcription factors from the immune module that negatively correlated with patient survival. This work showed that ChRO-seq is a powerful tool for understanding transcription regulation in complex diseases, highlighting the clinical importance of tumor microenvironment in GBM.

The second part develops a Bayesian statistical model for understanding the tumor compositions using bulk sample RNA-seq and/or ChRO-seq collected from large patient cohorts in conjunction with prior knowledge learned from the single cell RNA-seq and/or ATAC-seq data collected from normal and tumor tissues. This model is expected to address the following questions of central importance in cancer biology. First, what transcription pathways are ectopically regulated in tumor patients, and to what extent in each patient? Secondly, what are the cell type compositions in the tumor microenvironment of each patient? Lastly, do any of pathways or the cells present in the microenvironment interact among each other? Answers to these questions shall provide insights into new druggable targets through modulating tumor microenvironment.

BIOGRAPHICAL SKETCH

Tin Yi Chu was born in 1989, in Suzhou city, China. He became a naturalized citizen of Hong Kong at the age of 13. He attended the Suzhou No. 10 Middle School. Upon graduation, he attended the Chinese University of Hong Kong (CUHK). He received B.Sc. in biochemistry from CUHK in 2013 with the first-class honors. As an undergraduate student, he received trainings in structural biology and biochemistry. He independently solved the crystal structure of ribosomal maturation factor RimP and published his work in the Journal of Biological Chemistry.

He was then accepted to the Computational Biology program at Cornell University and received the Croucher Scholarship for Doctoral Study. In 2014, he joined the lab of Dr. Charles Danko at Baker Institute of Animal Health. While doing his dissertation work Tin Yi developed his specialties in statistical modeling and inference and their applications in cancer biology. He published his work in cancer epigenomics of glioblastoma as the first author in the paper “Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme”, which was highlighted in the journal of Neuro-Oncology. He also developed various bioinformatic tools for analyzing Run-on sequencing data, including dREG-HD and tfTarget. Towards the end of his graduate training, Tin Yi developed the statistical model TED to study tumor microenvironment compositions.

This thesis is dedicated to my family for their unconditional support.

ACKNOWLEDGMENTS

First I would like to acknowledge Dr. Thorsten Joachims and Dr. Praveen Sethupathy for agreeing to serve as members of my thesis committee. Their suggestions and support are invaluable for my career. I would like to acknowledge Dr. John Lis, Dr. Hojoong Kwak, and my ex-committee member Dr. Kristy Richards for their support in my graduate study. I would also like to acknowledge members of the Danko lab for all their help and support for the past six years. In particular I would like to thank Ed Rice for his invaluable contributions to the GBM project, and Zhong Wang for his patient mentorship in programming-related issues and insightful discussions for the development of dREG-HD, without which I would not be here today. Most importantly, I would like to thank my PI, Dr. Charles Danko for being extremely supportive while rigorous in scientific trainings, without which I would not have matured as a scientist.

TABLE OF CONTENTS

Table of Contents

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1

CHROMATIN RUN-ON AND SEQUENCING MAPS THE TRANSCRIPTIONAL REGULATORY LANDSCAPE OF GLIOBLASTOMA MULTIFORME

1.1 Abstract

The human genome encodes a variety of poorly understood RNA species that remain challenging to identify using existing genomic tools. We developed chromatin run-on and sequencing (ChRO-seq) to map the location of RNA polymerase using virtually any input sample, including samples with degraded RNA that are intractable to RNA-seq. We used ChRO-seq to map nascent transcription in primary human glioblastoma (GBM) brain tumors. Whereas enhancers discovered in primary GBMs resemble open chromatin in the normal human brain, rare enhancers activated in malignant tissue drive regulatory programs similar to the developing nervous system. We identified enhancers that regulate genes characteristic of each known GBM subtype, identified transcription factors that drive them, and discovered a core group of transcription factors that control the expression of genes associated with clinical outcomes. This study reveals the transcriptional basis of GBM and introduces ChRO-seq to map regulatory programs contributing to complex diseases.

1.2 Introduction

Our genomes encode a wealth of functional elements that play critical roles in the molecular basis of disease. RNAs serve as a marker for a surprisingly diverse group of functional elements, revealing the expression level of protein coding genes (mRNAs), as well as the location of enhancers and other non-coding regulatory elements which transcribe short and rapidly degraded non-coding RNAs (ncRNA)(Cheng et al. 2005; Chen et al. 2010; Ulitsky and Bartel 2013; Quinodoz and Guttman 2014; de Santa et al.

2010). However, the discovery of ncRNA species, especially of enhancer-templated RNAs (eRNAs) characteristic of distal regulatory elements(Kim et al. 2010; de Santa et al. 2010), has proven challenging. Most ncRNAs are not represented in RNA-seq data, owing to the rapid degradation rates of most ncRNAs by the nuclear exosome complex(Preker et al. 2008; Andersson, Refsing Andersen, et al. 2014). Chromatin immunoprecipitation and sequencing (ChIP-seq) for RNA polymerase II is of limited value because it has a poor signal-to-noise ratio which obscures less abundant RNA species(Core et al. 2012). Likewise, assays that measure nuclease accessibility, such as DNase-I-seq9 and ATAC-seq(Buenrostro et al. 2013), are poor sources of information about transcriptional activity because they identify open chromatin regions irrespective of activity, and do not measure critical sources of information about mRNAs such as gene expression levels or transcript boundaries.

Recent studies have shown that sequencing nascent RNAs attached to an actively transcribing RNA polymerase complex is an effective strategy for discovering coding and ncRNAs(Core, Waterfall, and Lis 2008; Churchman and Weissman 2011; Kwak et al. 2013; Mayer et al. 2015; Nojima et al. 2015; Schwalb et al. 2016; Core et al. 2014; Scruggs et al. 2015). Nascent RNA-seq techniques, such as Precision Run-On and Sequencing (PRO-seq)(Kwak et al. 2013), provide significantly higher sensitivity in detecting short-lived ncRNAs. Thus, PRO-seq and related assays provide a rich source of information about multiple layers of regulatory control, enabling simultaneous measurements of transcription at protein-coding genes and the discovery of active regulatory elements, including enhancers(Danko et al. 2015; Azofeifa and Dowell 2017; Andersson, Gebhard, et al. 2014).

Cancers are a particularly attractive target for nascent RNA sequencing techniques because cancer is a disease of gene regulation(Bradner, Hnisz, and Young 2017). In most cancers, somatic changes to DNA sequence affect oncogenic or tumor suppressive pathways(Parsons et al. 2008; Brennan et al. 2013). In some cases somatic mutations affect the core transcriptional machinery directly(Mohan et al. 2010), motivating the use of assays that directly measure the localization of Pol II. Somatic mutations initiate secondary changes in gene expression that are responsible for initiating changes in cell morphology and behavior that are characteristic of malignancy. For this reason, gene expression signatures from RNA-seq and other assays have proven effective as biomarkers, denoting cancer subtypes that are associated with progression and survival. However, which genes undergo regulatory changes in cancer, and especially the identity of key transcription factors that encode the malignant behaviors of cancer cells by their effect on target genes, remain poorly defined.

Nascent RNA sequencing techniques remain challenging to apply in some cell lines and especially to intact clinical isolates derived from cancer patients. Here we introduce a new chromatin-based run-on protocol, called Chromatin Run-On and Sequencing (ChRO-seq). ChRO-seq produces similar maps of transcription to PRO-seq in cell lines, but can also be applied to solid tissue samples, even those in which RNA is highly degraded. We used ChRO-seq to analyze 24 human glioblastoma multiforme (GBM) brain tumors, patient derived xenografts (PDXs), and a primary non-malignant brain sample, revealing new insights into the molecular etiology of GBM.

1.3 Results

1.3.1 Run-on assays in solid tissue

We developed Chromatin Run-On and sequencing (ChRO-seq), a new method to map RNA polymerase in cell or tissue samples (Fig. 1.1a). The primary challenge faced when using PRO-seq is often obtaining nuclei that are suitable for a run-on reaction. We therefore developed an alternative method which relies on fractionating insoluble chromatin, including engaged RNA polymerase II (Pol II)(Wuarin and Schibler 1994) (see Methods). Insoluble chromatin was re-suspended by sonication and used as input to a run-on reaction (Fig. 1.1a). The run-on was designed to incorporate a biotinylated nucleotide triphosphate (NTP) substrate into the existing nascent RNA that provides a high-affinity tag used to enrich nascent transcripts. The biotin group prevents the RNA polymerase from elongating after being incorporated into the 3' end of the nascent RNA when performed in the absence of normal NTPs, thus enabling up to single-nucleotide resolution for the polymerase active site(Kwak et al. 2013; Mahat et al. 2016).

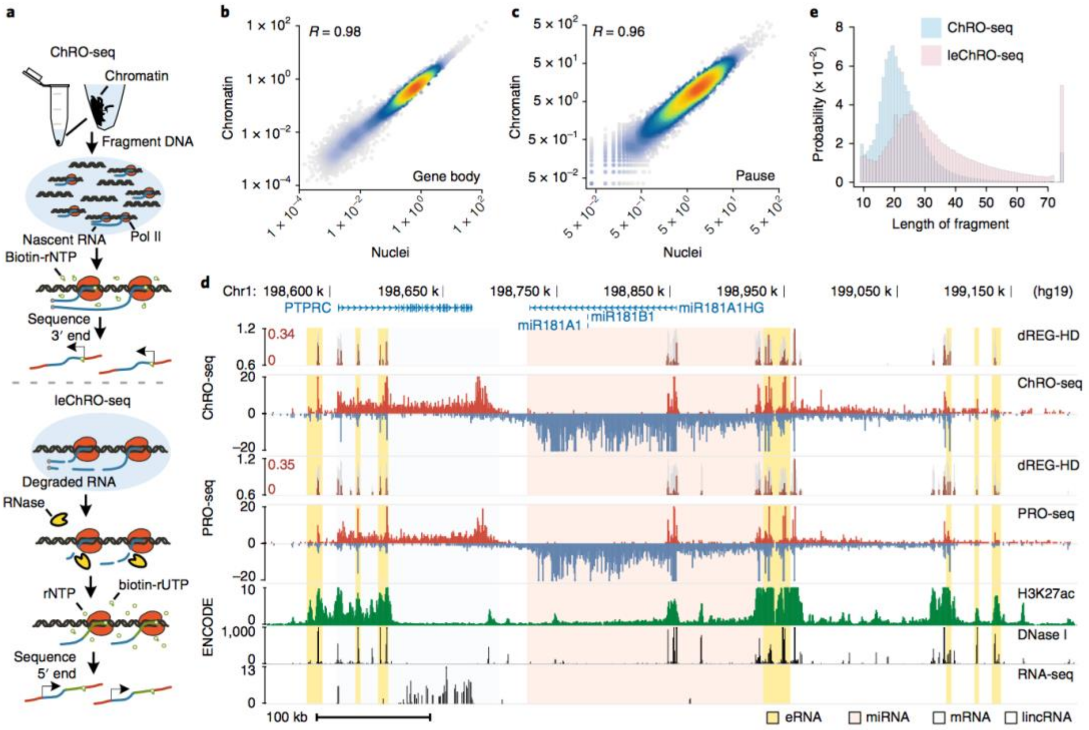


Fig. 1. 1 ChRO-seq and leChRO-seq measure primary transcription in isolated chromatin.

(**a**) Isolated chromatin is resuspended into solution, incubated with biotinylated rNTPs, purified by streptavidin beads, and sequenced from the 3' end. leChRO-seq degrades existing RNA, extends nascent transcripts an average of 100 bp, and sequences RNAs from the 5' end. (**b** and **c**) Comparison between matched ChRO-seq and PRO-seq in 41,478 RefSeq annotated gene bodies (**b**) or at the peak of paused Pol II (**c**). (**d**) Comparison between ChRO-seq (top three tracks), PRO-seq (center), and H3K27ac ChIP-seq, DNase-I-seq, and RNA-seq (bottom). dREG-HD shows the raw signal for dREG (gray) and dREG-HD signal (dark red). The shaded background shows the type of RNA produced at each position (**e**) The distribution of read lengths from ChRO-seq (blue) and leChRO-seq (pink) in a 30 year old primary GBM.

We performed matched ChRO-seq and PRO-seq experiments in the human Jurkat T-cell leukemia line, in which both nuclei and chromatin could be obtained. Median ChRO-seq signal across annotated genes was within the range of variation observed in PRO-seq data from the same cell line (Supplementary Fig. 1.1). In contrast, we noted differences in the pause peak and transcription past the polyadenylation site compared with mNET-seq and Nascent-seq, two other chromatin-based RNA sequencing assays(Mayer et al. 2015; Khodor et al. 2011; Menet et al. 2012) (Supplementary Note 1.1). ChRO-seq and PRO-seq produced highly correlated levels of RNA polymerase in the bodies of mRNA encoding genes (R= 0.98; Fig. 1.1b). Likewise, signal for paused Pol II was highly correlated across the 5' ends of annotated genes (R= 0.96; Fig. 1.1c), and pause levels in our test ChRO-seq library were within the range of variation observed using nuclei (Supplementary Fig. 1.2). The microRNA MIR181 locus illustrates the advantages of ChRO-seq compared with other molecular assays (Fig. 1.1d). Notably, both ChRO-seq and PRO-seq discovered the primary transcript encoding MIR181 as well as dozens of eRNAs that were not discovered using RNA-seq.

Because RNA prepared from archival tissues is often highly degraded, such samples are poor candidates for genome-wide transcriptome analysis using RNA-seq. The RNA polymerase-DNA complex is more stable than RNA(Cai and Luse 1987), suggesting that engaged polymerases may provide an avenue for producing new RNAs in archived samples. We obtained a primary glioblastoma multiforme (GBM) (grade IV, ID# GBM-88-04) that was stored in a tissue bank for 30 years. Bioanalyzer analysis confirmed that RNA was highly degraded in this sample (RIN = 1.0, Supplementary Fig. 1.3), thus precluding the application of RNA-seq (requires RIN of 2-4). To measure gene expression in this sample, we devised length extension ChRO-seq (leChRO-seq), a variant of ChRO-seq that uses transcriptionally-engaged Pol II and a mix of biotinylated-NTP and normal NTPs to extend degraded nascent RNA transcripts (Fig. 1.1a). Whereas libraries prepared without an extended run-on had a median insert size of 20 bp, precisely the length of RNA protected from degradation by the polymerase exit channel(Choder and Aloni 1988), run-on samples achieved a longer RNA length distribution that was better suited for mapping unique reads within the human genome (Fig. 1.1e). Although RNA degradation could, in principal, destabilize RNA polymerase, we nevertheless observed that leChRO-seq produced maps of transcription that were correlated with those obtained using ChRO-seq and PRO-seq, suggesting that leChRO-seq accurately measures gene expression and pausing (Supplementary Fig. 1.1a, 1.2, 1.4a). Thus, leChRO-seq allows the robust interrogation of archival tissue samples which cannot be analyzed using standard genomic tools.

1.3.2 Maps of transcription in primary GBMs

To demonstrate how ChRO-seq can provide insights into complex disease, we obtained ChRO-seq or leChRO-seq data from 20 primary glioblastomas, three patient derived xenografts (PDX), and a non-malignant brain (Fig. 1.2a; Supplementary Table 1.1). Histopathology revealed hallmarks of grade IV malignant astrocytoma in all GBMs (e.g., GBM-15-90, Supplementary Fig. 1.5). We sequenced ChRO-seq data from each GBM to an average depth of 33 million uniquely mapped reads per sample (10-150M reads/ sample). We confirmed that data collected from biopsies isolated from nearby regions (technical replicates) were highly correlated (Supplementary Fig. 1.4c-f, Supplementary Note 1.2).

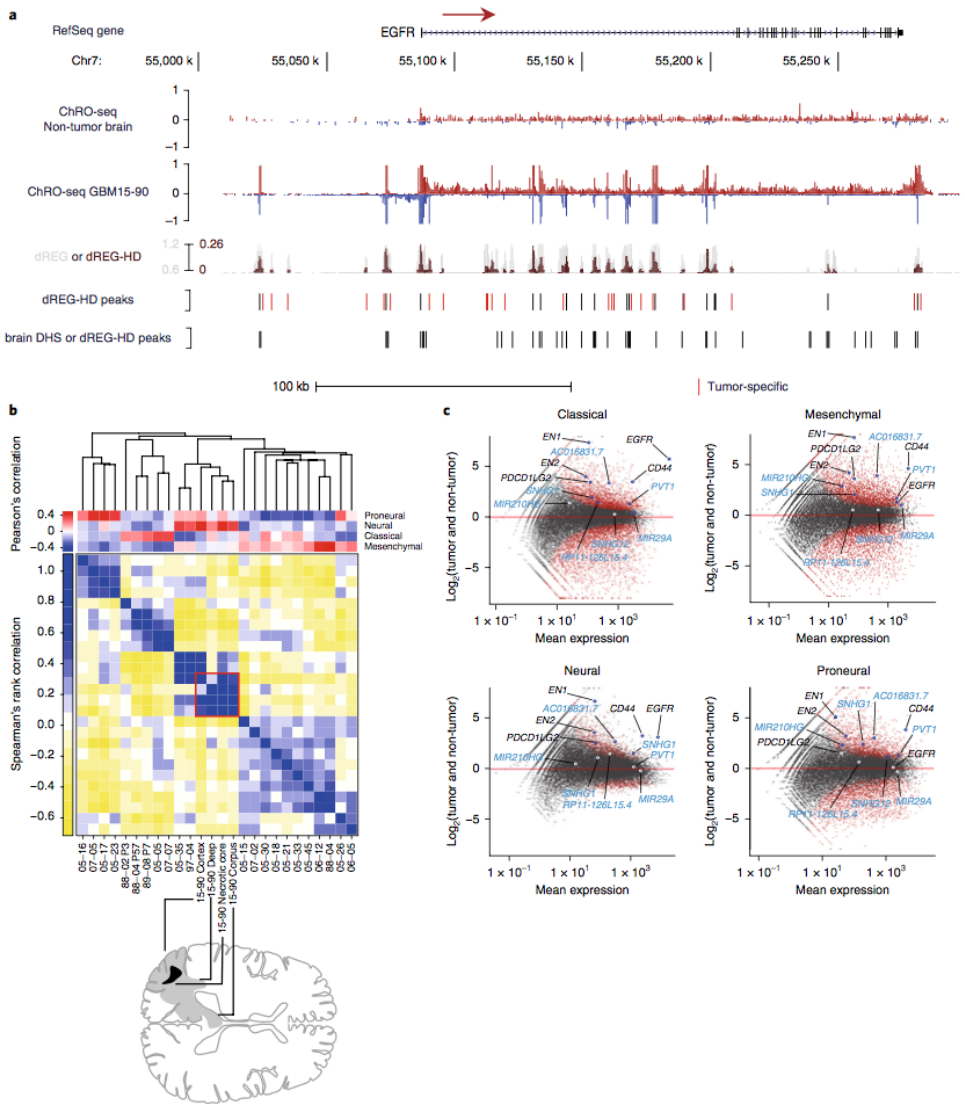


Fig. 1. 2 ChRO-seq detects transcription in primary human glioblastomas.

(**a**) RPM normalized ChRO-seq signal at the EGFR locus in nonmalignant brain (top) and GBM-15-90 (center). dREG (gray) and dREG-HD (dark red) signals are shown for GBM-15-90. dREG-HD peaks that are not DHSs in adult brain reference samples are highlighted in red. DHSs in 6 adult brain reference samples and dREG-HD peaks from the non-malignant brain sample. (**b**) Upper matrix: subtype scores for each patient, calculated by Pearson's correlation with the centroid of gene expression of corresponding subtype. Lower matrix: Spearman's rank correlation over subtype signature genes among 20 primary GBMs. Red square denotes four regions dissected from GBM-15-90. Sample order is based on single-link hierarchical clustering of the lower matrix, shown by the dendrogram. In total, 838 genes were used for calculating the correlation coefficients. (**c**) Differential gene transcription of primary GBMs in each subtype compared with non-malignant brain. Genes of interest are highlighted. lncRNAs are highlighted in blue.

To gain further insight into how transcription changes in malignant tissue, we analyzed transcription within annotated protein-coding genes and non-coding RNAs. GBMs from our cohort represent each of the four previously reported molecular subtypes(Verhaak et al. 2010) (Fig. 1.2b, Supplementary Fig. 1.6). Though most tumors have transcription patterns characteristic of one dominant molecular subtype, several tumors in our cohort were similar to multiple subtypes, especially those matching neural and mesenchymal signatures, consistent with reports of cellular heterogeneity within the same tumor(Patel et al. 2014; Q. Wang et al. 2017) (Fig. 1.2b). We identified 2,381 protein-coding genes and 1,123 ncRNAs that were differentially transcribed across all 20 primary GBMs relative to replicates of the non-malignant brain ($p < 0.05$, False discovery rate [FDR] corrected Wald test, DESeq2(Love, Huber, and Anders 2014)) (Supplementary Table 1.2). Differentially transcribed genes had notable enrichments in biological processes related to cell cycle, DNA replication / metabolic processes, development (up-regulated in the tumor), and nervous system homeostasis (down-regulated) (Supplementary Fig. 1.7). For example, multiple transcription factors with a role specifying nervous system development were expressed more highly in nearly all tumors, including the HOX gene clusters and engrailed-1 and 2 (EN1 and EN2) (Fig. 1.2c; Supplementary Fig. 1.8). Notably, we discovered several differentially transcribed long non-coding RNAs (lncRNAs) that confer growth advantages to U87 glioblastoma cells(Liu et al. 2017; Xi et al. 2017; Ma et al. 2017; Zhao et al. 2014) (e.g., AC016831.7, PVT1, SNHG1, etc. Fig. 1.2c; Supplementary Table 1.3). Taken together, our analysis of ChRO-seq data identified transcriptional changes in both genes and lincRNAs that were shared between GBMs in our cohort.

1.3.3 GBM enhancers retain signatures of normal brain tissue

Active transcriptional regulatory elements (TREs), including promoters and enhancers, have a characteristic pattern of RNA polymerase initiation that allows their discovery using ChRO-seq data(Kim et al. 2010; de Santa et al. 2010; Core et al. 2014; Danko et al. 2015; Azofeifa and Dowell 2017; Andersson, Gebhard, et al. 2014). We developed a novel algorithm to identify the precise location of active TREs, called dREG-HD, which takes PRO-seq or ChRO-seq data as input and identifies TREs that are similar to the subset of DNase-I hypersensitive sites (DHSs) that exhibit local transcription initiation. The dREG-HD algorithm improved the resolution of dREG19 by imputing smoothed DNase-I-seq signal intensity, and identified sites initiating transcriptional activity with 80% sensitivity at >90% specificity (Supplementary Fig. 1.9). dREG-HD recovered the nucleosome depleted region in histone modification ChIP-seq and MNase-seq data (Supplementary Fig. 1.10), demonstrating that it had substantially higher resolution compared with dREG alone.

The vast majority (96%) of TREs identified by dREG-HD in each primary GBM sample were DHSs in at least one of the 216 reference tissues analyzed by ENCODE or Epigenome Roadmap(Roadmap Epigenomics Consortium et al. 2015; Dunham et al. 2012). However, most DHSs were discovered in only a few of the tissues in the reference dataset (Fig. 1.3a) and were distal (>1 kb) to annotated transcription start sites (Fig. 1.3b), suggesting that many reflect the activity of cell-type specific distal enhancers in the tumor. Rare distal TREs (henceforth referred to as “enhancers”) provide a unique “fingerprint” for quantitatively evaluating the similarity between two samples, and could be used to define the relationship between tumors and normal tissue.

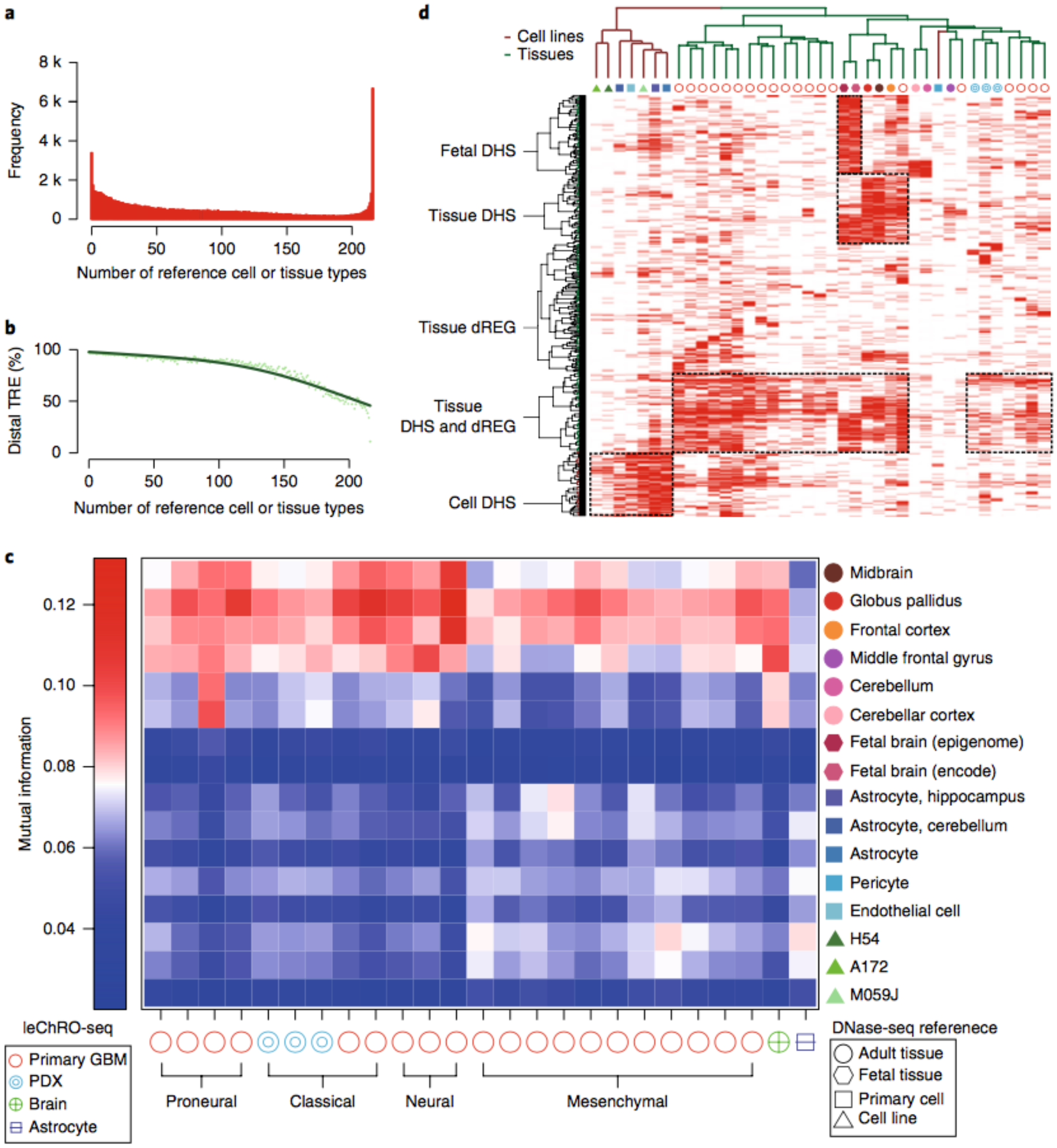


Fig. 1. 3 Comparison between TREs in primary GBM / PDX and reference DHSs.

(**a**) Histogram representing the number of reference samples that have a DHS overlapping each dREG-HD site found in any of the 23 primary GBM / PDX samples. (**b**) Percentage of TREs >1kb from the nearest GENCODE transcription start site. (**c**) Mutual information between TREs in the indicated GBM and reference sample. (**d**) Clustering of reference samples with primary GBM / PDX based on the activation of TREs. Activate TREs are marked in red; inactive ones are in white.

We developed a strategy that compares active enhancer landscapes obtained using dREG-HD with DHSs across all public datasets (see Methods). Our strategy consistently discovered the expected cell lines (Supplementary Fig. 1.11), even identifying the expected genotype (GM12878) among all lymphoblastoid cell lines as the most similar to GM12878 PRO-seq data (Supplementary Fig. 1.11a). Using unique enhancers to “fingerprint” primary GBM samples revealed enhancer landscapes that were highly similar to normal brain reference samples compared to other reference tissues (Fig. 1.3c, Supplementary Fig. 1.12). In GBM-15-90, for instance, 86% of TREs were shared with primary brain tissue, which was greater similarity than observed in either GBM cell lines (62% TRE identity) or in vitro cultured primary brain cells (75%) (Supplementary Fig. 1.13). Clustering TREs in several brain related cell types suggested that differences between primary tumors and GBM cell lines were caused by differences in the tumor microenvironment (Fig. 1.3D; Supplementary Fig. 1.14; Supplementary Note 1.3).

To evaluate whether contamination of the GBM with normal brain tissue explained the extensive similarity with normal brain reference samples, we used leChRO-seq data from three PDXs, in which primary GBMs were grown in a murine host. In PDXs, murine cells replace both normal tissue and stroma(Tentler et al. 2012), and can be distinguished from tumor cells based on species-specific differences in DNA sequence. Mutual information ranked all PDX samples as similar to the normal human brain compared with all other samples (Fig. 1.3c). Thus we conclude that primary GBM cells are more similar to their cell of origin than may have been anticipated based on prior analysis of cell models.

1.3.4 TREs define three distinct regulatory programs activated in GBM tissue

TREs that were active in tumor tissue, but were not DHSs in any of the available adult brain reference samples, are strong candidates for contributing to the malignant phenotype of the tumor. Such tumor-associated TREs (taTREs) comprised 2-24% of TREs in each tumor (Supplementary Fig. 1.15, 1.16, Supplementary Table 1.4). We developed a statistical test to identify tissues which shared unexpectedly high overlap with taTREs identified in each tumor that controls for DHS scarcity (Supplementary Table 1.4) (see Online Methods). Hierarchical clustering of the taTREs among significant cell types revealed three regulatory programs that were enriched in the primary GBMs; one resembling a stem-like regulatory program, one associated with differentiated support cells, and a cluster of immune cells (Fig. 1.4a, Supplementary Fig. 1.17). taTREs significantly ($p < 1e-4$, bootstrap test) overlapped DHSs in fetal tissues of the nervous system (2.3-6.6-fold enrichment in 11/ 23 GBMs), especially spinal cord and brain, two fetal tissues derived from the neuroectoderm (Fig. 1.4a, see “Outlier tissues”). We also found evidence for enrichment in additional developmental tissues, for example embryonic stem cells and other fetal tissues from a variety of germ layers, and for a number of terminally differentiated support cell lineages including astrocytes, endothelial cells, fibroblasts, and osteoblasts. Regulatory programs were partially correlated with previously defined molecular subtypes in GBM (Fig. 1.4c; Supplementary Note 1.4). We emphasize that activation of these separate transcriptional regulatory programs may reflect gene expression changes in subsets of cells within the tumor. Overlap between taTREs and fetal brain tissue likely reflects the activation of a regulatory program that promotes stem-like properties observed in a population of GBM cells(Suvà et al. 2014). Similarly, overlap with astrocytes, endothelial cells, fibroblasts, or osteoblasts may capture tumor cells that have trans-differentiated into these lineages(Lucia Ricci-Vitiani et al. 2010; L. Ricci-

Vitiani et al. 2008). Notably, these two signatures were detected in PDX samples as well as primary GBMs, demonstrating that these signatures reflect transcriptional diversity in malignant cells.

To identify transcription factors involved in maintaining each regulatory program, we classified the taTREs in each tumor sample into regulatory programs based on their cell type overlap, and searched for enriched transcription factor binding motifs ($p < 0.05 / 1882$ in at least one patient, Fisher's exact test, Rtfbsdb(Z. Wang, Martins, and Danko 2016)). As we were limited in our ability to distinguish between paralogous transcription factors that share similar DNA binding specificities, we clustered motifs into 14 distinct groups, each associated with multiple transcription factors that may contribute to differences in expression (Fig. 1.4b). Many of these motifs showed mutually exclusive enrichment in the three regulatory programs (Fig. 1.4b; Supplementary Fig. 1.18), supporting the hypothesis that each regulatory program is a transcriptionally distinct program mediated by a different group of transcription factors. We identified POU domain containing transcription factors enriched in taTREs in the stem-like regulatory program. As predicted, taTREs in the stem-like program were enriched in both ChIP-seq reads and peak calls for POU3F2 in cultured glioma neurospheres(Suvà et al. 2014) (Supplementary Fig. 1.19, 1.20). The differentiated support cell program was highly enriched for binding of activating protein 1 (AP-1), a heterodimer of the transcription factors FOS and JUN, a motif resembling heat shock factor 1 (HSF1), and the TEAD family (Fig. 1.4b). The immune program was enriched for C/EBP family (C/EBPA), NF-κB family (RELA), and the Retinoic Acid Receptor family (RARA), in agreement with reports that at least two of these factors play an important role in inflammatory responses in GBM(Bhat et al. 2013; Carro et al. 2010). Taken together, we have identified taTREs that correlate with complex behaviors intrinsic to malignant cells, for instance the stem-like regulatory program that was shared with neuroectodermal tissue, and identified candidate transcription factors that contribute to each behavior.

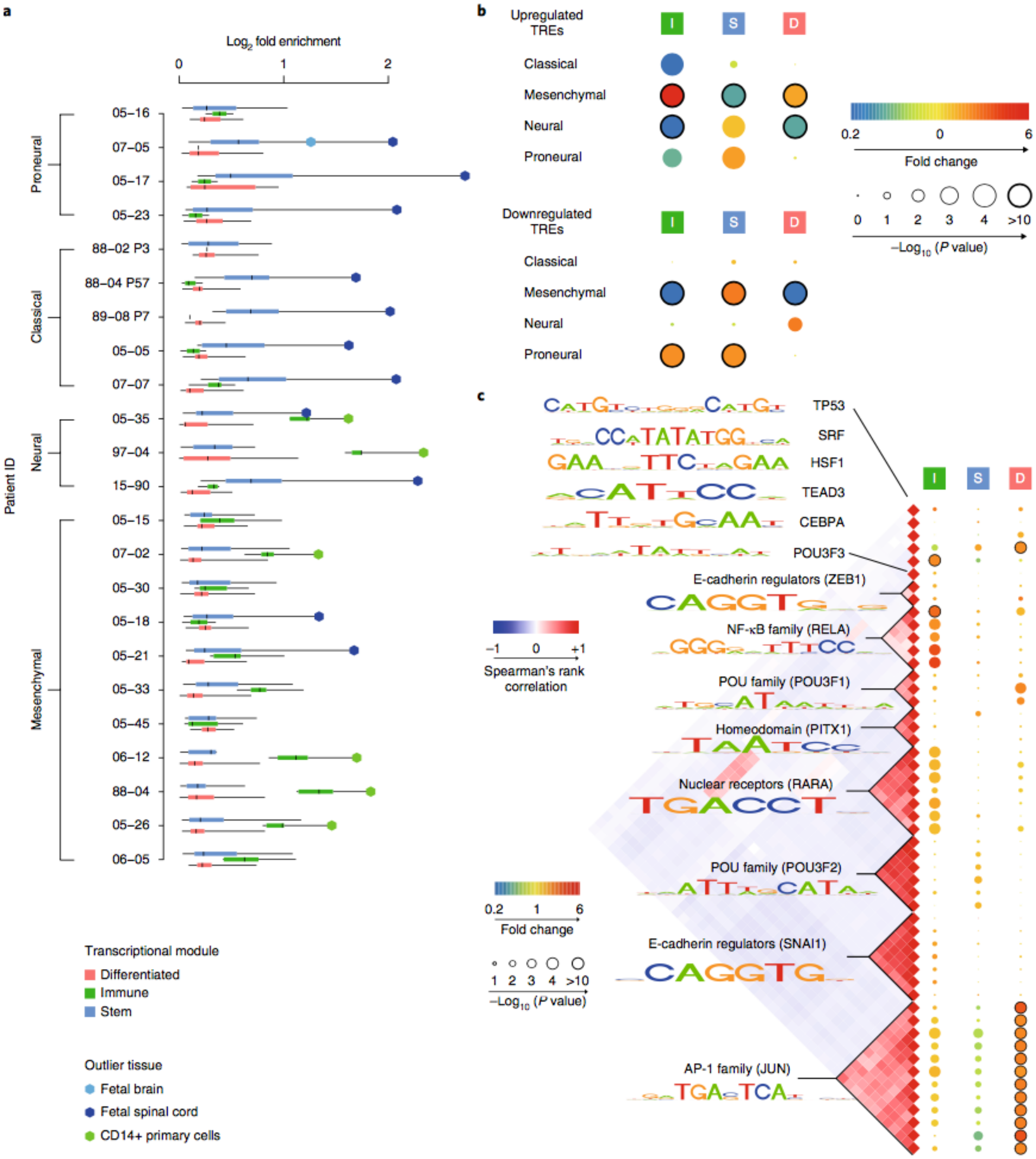


Fig. 1. 4 Tumor associated TREs (taTREs) activate three regulatory programs.

(**a**) Boxplots show the $\log_2$ fold enrichment of reference tissues enriched in the corresponding GBM. Reference samples enriched in each patient were grouped into three regulatory programs, called stem (blue, n= 24), immune (green, n= 5), and differentiated (pink, n= 21). Box plots show the 25th percentile (bottom of box), median (central bar), and 75th percentile (top of box). Whiskers represent minimum and maximum values. Outlier tissues are indicated in the legend. (**b**) Transcription factor binding motifs enriched in TREs of the immune (I), stem (S), and differentiated (D) regulatory program compared with TREs active in the normal brain. All motifs shown were significantly enriched following Bonferroni adjustment of the threshold p value in at least one patient ($p < 0.05 / 1882$, two-sided Fisher's exact test). The Spearman's rank correlation heatmap (left) shows the correlation in DNA binding sites matching each motif. The radius of the circle represents the median *p* value across patients is and the color represents the magnitude of enrichment (red) or depletion (blue). (**c**) The radius of the circle represents the *p* value (two-sided Fisher's exact test) of enrichment of the indicated regulatory programs in subtype-biased TREs. The color represents the magnitude of enrichment (red) or depletion (blue). Number of subtype-biased TREs in each comparison (panels b and c) is shown in Supplementary Table 1.3 and 1.4.

1.3.5 Transcription factors controlling GBM subtype

Transcriptional heterogeneity among GBMs is established in large part by the differential activity of transcription factors. To identify transcription factors that are involved, we focused on TREs with evidence of expression changes among the four previously described molecular subtypes ($p < 0.01$, FDR corrected Wald test, DESeq2) (Supplementary Table 1.5). We identified 38 binding motif clusters with extremely strong evidence of enrichment in active TREs with biased transcription in any subtype ($p < 0.05 / 1882$, Fisher's exact test, Fig. 1.5a). Significantly enriched motifs passing our stringent multiple testing correction threshold were most common in the mesenchymal and neural subtypes, in which several had previous support in the literature, including those recognized by nuclear factor-κB (NF-κB) family and CCAAT/Enhancer Binding Protein (C/EBP) family enriched in TREs up-regulated in mesenchymal tumors(Bhat et al. 2013; Carro et al. 2010). Additionally, we identified numerous novel motif associations that correlate with subtype-biased expression including, for instance, RARA, SRF, SOX-family, and FOX-family.

Next we set out to identify target genes regulated by each transcription factor in GBM cells. First, we assume that molecular subtypes described in current literature do not completely describe the full range of heterogeneity among GBMs. To identify motifs contributing to heterogeneity that are only weakly correlated with the known molecular subtypes, we relaxed our statistical cutoff to a more permissive threshold at which we expected substantially higher sensitivity at an acceptable false discovery rate ($p < 0.05$, nominal Fisher's exact test, Supplementary Fig. 1.21, see Methods). We identified bound occurrences of each enriched motif using heuristics that provide substantial performance improvements over existing high-resolution tools(Danko et al. 2018). Motif occurrences were connected with the closest two annotated genes sharing similar subtype-bias within 50 kb (Fig. 1.5b), using fairly stringent heuristics to limit false discovery rates (see Methods). We validated target genes by confirming that genes sharing a common transcription factor were more highly correlated across 174 primary GBMs(Brennan et al. 2013) than expected based on randomly selected genes sharing the same subtype specificity (Fig. 1.5c; Supplementary Note 1.5). Thus, we have identified transcription factors contributing to major GBM expression subtypes and their putative target genes.

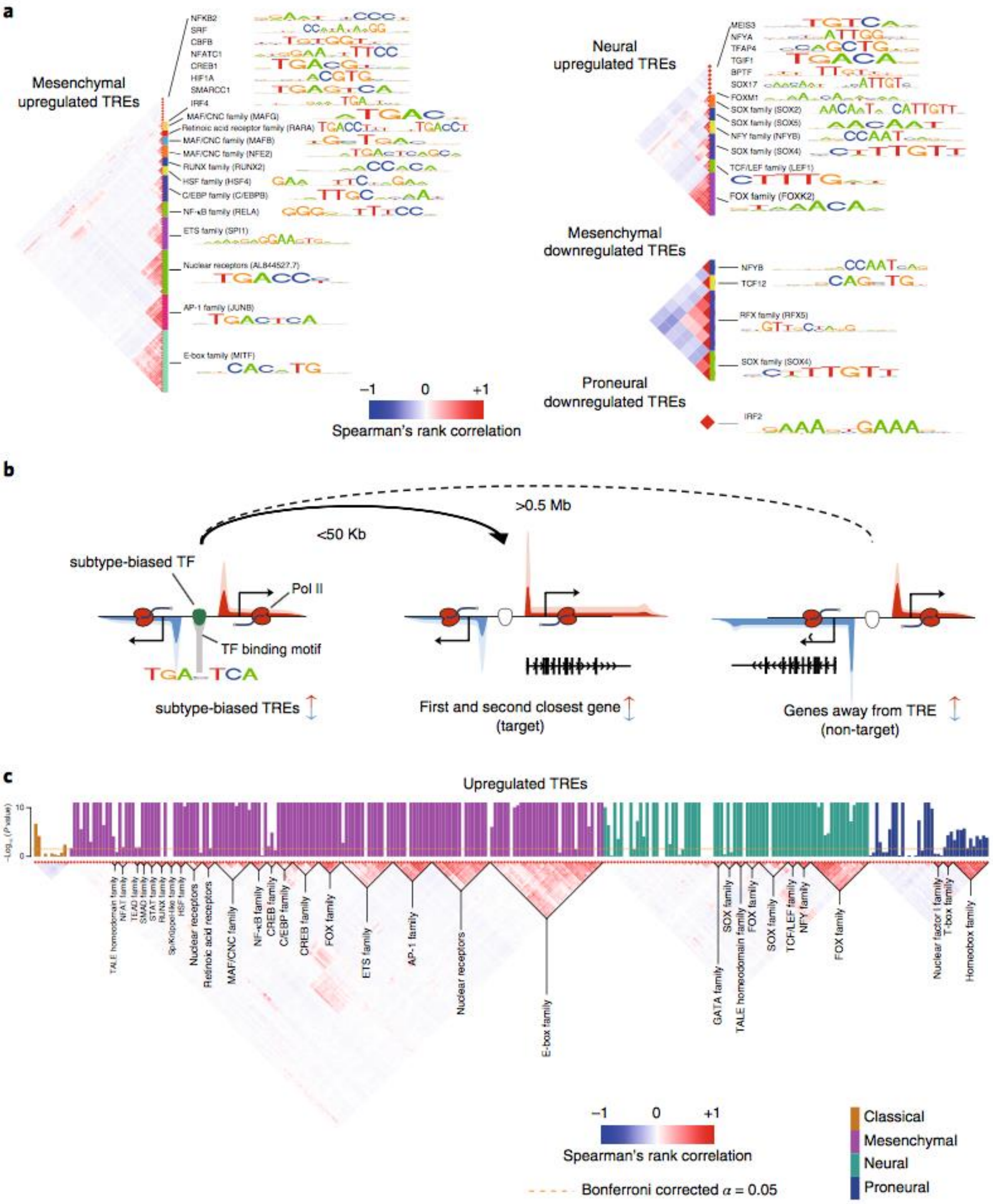

a
Mesenchymal upregulated TREs
NFKB2
SRF
CBFB
NFATC1
CREB1
HIF1A
SMARCC1
IRF4
MAF/CNC family (MAFG)
Retinoic acid receptor family (RARA)
MAF/CNC family (MAFB)
MAF/CNC family (NFE2)
RUNX family (RUNX2)
HSF family (HSF4)
C/EBP family (C/EBPB)
NF-κB family (RELA)
ETS family (SPI1)
Nuclear receptors (AL844527.7)
AP-1 family (JUNB)
E-box family (MITF)
−1
0
+1
Spearman's rank correlation
Neural upregulated TREs
MEIS3
NFYA
TFAP4
TGIF1
BPTF
SOX17
FOXM1
SOX family (SOX2)
SOX family (SOX5)
NFY family (NFYB)
SOX family (SOX4)
TCF/LEF family (LEF1)
FOX family (FOXK2)
Mesenchymal downregulated TREs
NFYB
TCF12
RFX family (RFX5)
SOX family (SOX4)
Proneural downregulated TREs
IRF2
b
>0.5 Mb
<50 Kb
subtype-biased TF
Pol II
TF binding motif
TGA–TCA
subtype-biased TREs
First and second closest gene
(target)
Genes away from TRE
(non-target)
c
Upregulated TREs
$-\log_{10}$ (P value)
10
0
TALE homeodomain family
NFAT family
TEAD family
SMAD family
STAT family
RUNX family
Sp/Krüppel-like family
HSF family
Nuclear receptors
Retinoic acid receptors
MAF/CNC family
NF-κB family
CREB family
C/EBP family
CREB family
FOX family
ETS family
AP-1 family
Nuclear receptors
E-box family
GATA family
SOX family
TALE homeodomain family
FOX family
SOX family
TCF/LEF family
NFY family
FOX family
Nuclear factor I family
T-box family
Homeobox family
−1
0
+1
Spearman's rank correlation
Bonferroni corrected α = 0.05
Classical
Mesenchymal
Neural
Proneural

Fig. 1. 5 Transcription factors influencing transcriptional heterogeneity in GBM.

(**a**) Transcription factor binding motifs enriched in TREs that were up- or down-regulated in the indicated subtype. All motifs shown were significantly enriched following Bonferroni adjustment of the threshold *p* value ($p < 0.05 / 1882$, two-sided Fisher's exact test; sample size shown in Supplementary Table 4). The Spearman's rank correlation heatmap (left) shows the correlation in motif recognition. Families of transcription factors and their representative motifs are highlighted. (**b**) Cartoon illustrating heuristics used to identify target genes of subtype-specific transcription factor and for defining non-target (control) genes. Changes in transcription of both target and non-target genes are of the same direction as that of subtype-biased TREs. Target genes are the 1st and 2nd genes within 50 Kb of the TRE. Non-target genes are at least 0.5 Mb away. (**c**) Barplots show the $-\log_{10}$ Wilcoxon rank sum *p* value of having higher correlations among target genes of each transcription factor binding motif than a control set (columns; N=174 TCGA patients with RNA-seq data available). Barplots are colored by subtype in which they were found to be enriched ($p < 0.05$, two-sided Fisher's exact test). The Spearman's rank correlation between the binding sites of each motif is shown (bottom). Transcription factor families are indicated below the plot. The dotted line shows the Bonferroni adjusted threshold for the between-target validation experiment.

1.3.6 Direct inference of transcription factor regulatory activities in GBMs

The gene-regulatory “trans” activities that a transcription factor has on its complement of bound TREs can be regulated by multiple transcriptional and post-transcriptional mechanisms. While some transcription factors are controlled predominantly by the abundance of its protein, many require a subsequent step such as post-transcriptional activation of the protein product to regulate target genes (Fig. 1.6a). We asked whether we could distinguish between these two broad regulatory activities by using ChRO-seq, and using an integrative analysis incorporating both ChRO-seq and publicly available mRNA-seq data.

In the simplest mode of regulation, the gene-regulatory activity of a transcription factor is determined by the abundance of its protein, which can be correlated with the transcriptional activity of its gene and the abundance of its mRNA. To detect this type of regulatory activity, we asked whether motifs enriched in active TREs of each subtype correspond to changes in Pol II density on the primary transcription unit encoding any one of the transcription factors that recognize the corresponding binding motif. In some cases, we observed transcriptional changes in the transcription factor coding gene in the same subtype in which we also observed motif enrichment (Fig. 1.6b-c; Supplementary Fig. 1.24b). For instance, ChRO-seq signal in the gene body encoding the transcriptional activator CEBPB increased by 4.88-fold in mesenchymal tumors (Fig. 1.6b), consistent with a 2.43-fold enrichment of its corresponding motif in mesenchymal upregulated TREs (Fig. 1.5a). Likewise, we found several cases in which mRNA encoding each transcription factor was correlated with the expression of its putative target genes across GBMs to a greater extent than expected based on a null model that controls for molecular subtype (Fig. 1.6c; see Methods).

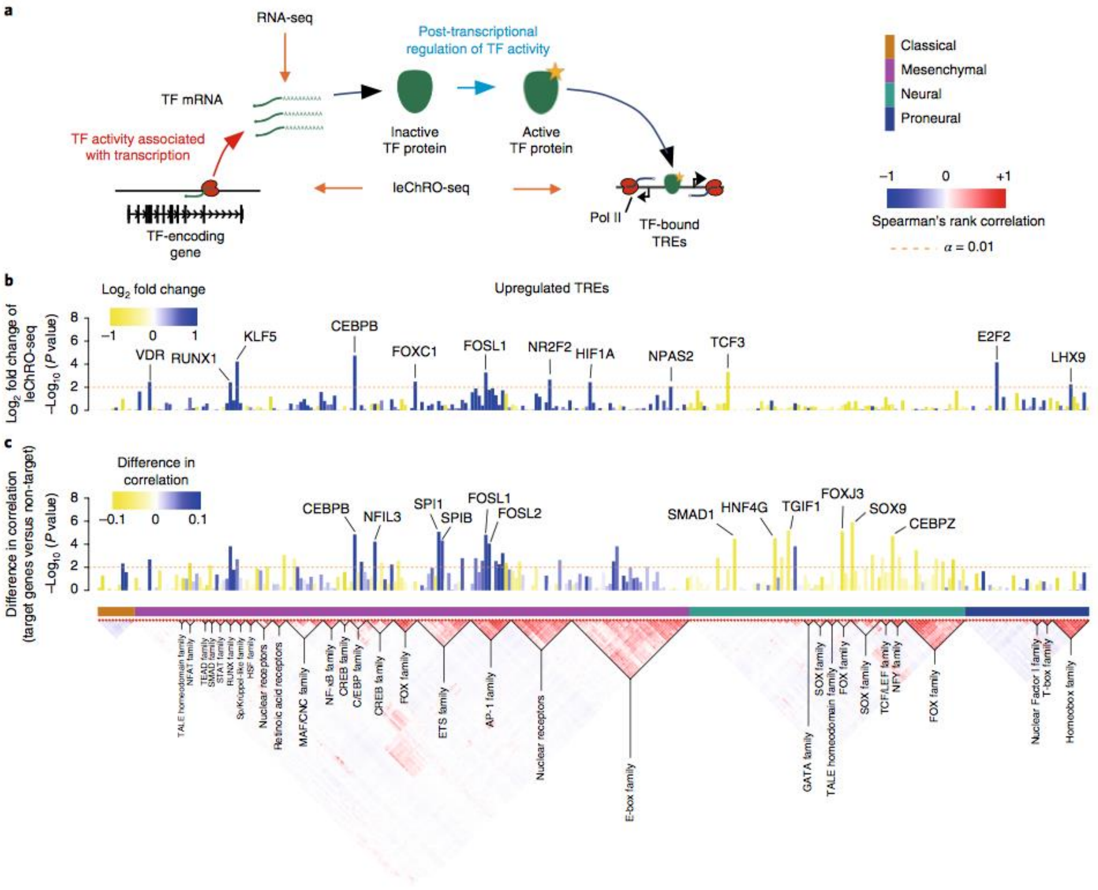


Fig. 1. 6 Regulatory activities of transcription factors are controlled by transcription and post-transcriptional mechanisms in GBM.

(**a**) The cartoon illustrates the stages at which transcription factor activities can be regulated and the corresponding signals detected by RNA-seq and (le)ChRO-seq. The activity of some transcription factors correlates predominantly with the abundance of its protein. Many transcription factors require post-transcriptional activation of the protein product before regulating target genes. (**b**) Barplot shows the FDR corrected -log$_{10}$*p* value (DESeq2, Wald test, n= 2 [classical] or 3 [other subtypes]) representing changes in Pol II abundance detected by (le)ChRO-seq on the gene encoding the indicated transcription factor. The level of upregulation (blue) and downregulation (yellow) in the subtype indicated by the colored boxes (below the barplot) is shown by the color scale. The dashed line shows the the FDR corrected α at 0.01. (**c**) Barplot shows the -log$_{10}$ two-sided Wilcoxon rank sum test *p* value denoting differences in the distribution of correlations between the mRNA encoding the indicated transcription factor and either target or non-target control genes. The blue/ yellow color scale represents the median difference in correlation between target and non-target genes over 174 mRNA-seq samples. The dashed line shows the uncorrected α at 0.01.

We devised a strategy to estimate which transcription factors have gene-regulatory activities that were regulated by transcriptional or post-translational mechanisms. Focusing on the 25 unique motifs enriched in up-regulated TREs that are associated with multiple transcription factors, we found evidence of correlated changes in ChRO-seq data for eight and in mRNA-seq for 16 (Fig. 1.6b-c). Several of these correlations were weak in magnitude, which may be consistent with gene-regulatory activities controlled by multiple regulatory mechanisms for these transcription factors. We conservatively identified at least six transcription factors, including TEAD, GATA, HSF, and NF-kB, which had low correlations with their putative targets in RNA-seq and no evidence of transcriptional changes in ChRO-seq. These transcription factors were regulated primarily at a post-transcriptional level in GBM.

1.3.7 Transcription factors control groups of survival-associated genes in mesenchymal GBMs

Known molecular subtypes of GBM do not correlate with survival(Verhaak et al. 2010), presenting a motivation to identify new classifiers that may have prognostic value. We hypothesized that the activity of transcription factors which control transcriptional heterogeneity among GBM patients may control biological functions correlated with survival. To determine whether gene-regulatory activities of transcription factors may be useful in predicting clinical outcomes, we compared the hazards ratio at putative target genes of each subtype specific binding motif. We analyzed two sets of non-target control genes: 1) The nearest annotated transcription start site (within 50 kb) of each subtype-specific TRE that was not changed in that subtype, and 2) Differentially transcribed genes in the same subtype that were not identified as targets, because the transcription start site was >0.5Mb away from the nearest putative binding site. Our analysis identified six transcription factors significantly associated with poor clinical outcomes, all in mesenchymal tumors ($p < 0.05 / 432$, Wilcoxon, Fig. 1.7a, Supplementary Fig. 1.25), which we clustered into three unique DNA binding specificities (RAR, C/EBP family, and RELA [NF-κB] Supplementary Fig. 1.26). Only one of these transcription factors, C/EBP, was associated with survival at the mRNA level (Supplementary Fig. 1.27), consistent with the gene-regulatory activity of C/EBP family correlating with the abundance of its mRNA (Fig. 1.6b). RELA activity was correlated to radio-resistance in GBMs, and in this case its activity was shown to be regulated post-transcriptionally by monitoring the phosphorylated state of the RELA protein(Bhat et al. 2013), providing an additional source of support for a second of the transcription factors identified here associated with clinical outcomes. In addition, we also identified RAR, which to our knowledge has not been linked to survival in GBM.

Surprisingly all three survival associated transcription factors regulated overlapping sets of putative target genes. Of four different combinations in which multiple transcription factors could regulate overlapping targets, three were more common than expected ($p < 0.01$; super exact test(M. Wang, Zhao, and Zhang 2015); Fig. 1.7b; Supplementary Fig. 1.28), including 44 target genes that were shared among all three transcription factors. Target genes shared among all three transcription factors had significantly higher hazard ratios than unique target genes (Fig. 1.7c,d, $p = 1.1e-3$, Wilcoxon). Of the 26 shared targets for which hazards ratios were available, all were negatively correlated with survival, and eight were significantly associated with clinical outcomes on their own (a significant enrichment [$p = 6e-4$, Fisher's exact test]), including CCL20 (Supplementary Fig. 1.29a) and ADM (Fig. 1.7d), ($p < 0.05$, Chi-squared test) (Supplementary Table 1.6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 1.29b-c). Moreover, differences in survival among these genes were not driven by IDH1 status (Supplementary Fig. 1.30). Gene ontology analysis found that targets of all three transcription factors were enriched for immune system process and stress responses ($p < 1e-5$, false discovery rate (FDR) corrected Fisher’s exact test, Supplementary Table 1.7). Taken together, our analysis suggests that C/EBP, RARG, and NF-κB work in concert to activate a shared regulatory program that controls inflammatory processes and correlates with poor clinical outcomes in GBM.

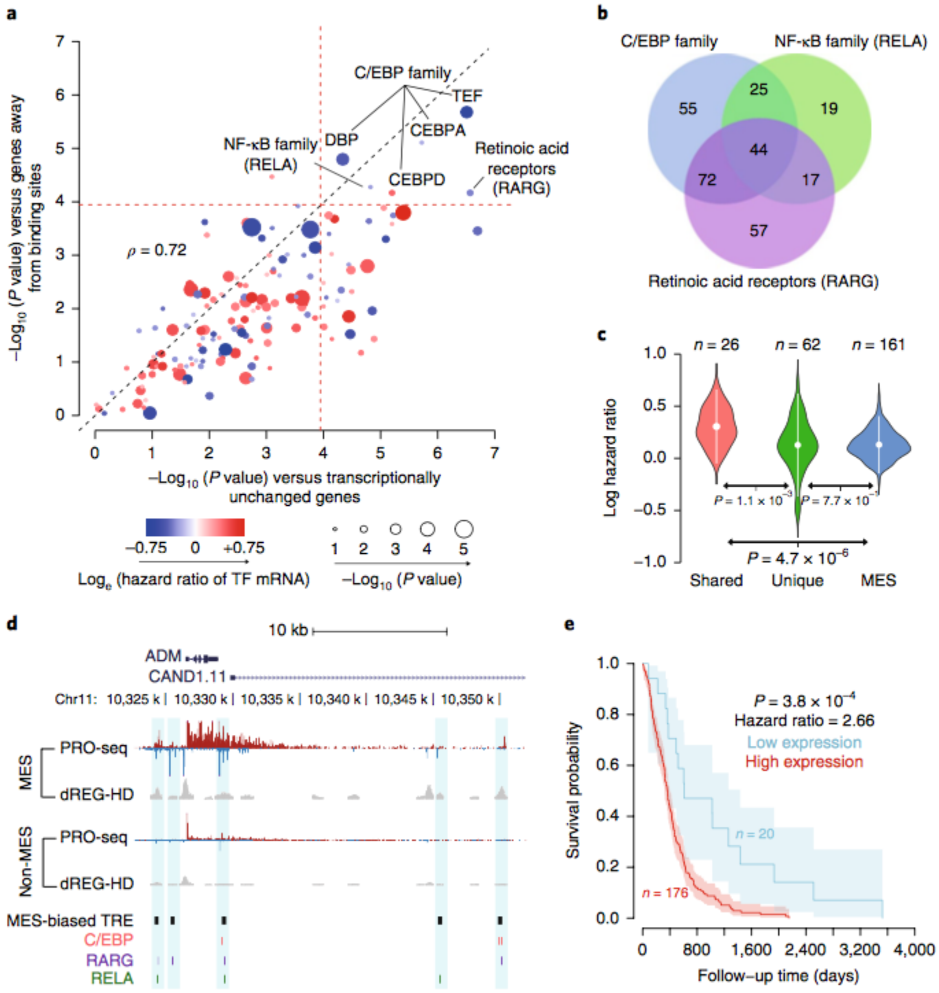


Fig. 1. 7 Transcription factors control survival associated pathways in GBM.

(**a**) Scatter plot shows the -$\log_{10}$ two-sided Wilcoxon rank sum test *p* value comparing the distribution of hazards ratios of target genes for each transcription factor and two groups of non-target control genes (see Methods). The radius of the circle denotes the -$\log_{10}$ *p* value of association between transcription factor mRNA levels and survival. Color denotes the $\log_e$ of the hazard ratio at higher mRNA levels. The dotted red line represents the Bonferroni adjusted α threshold (0.05/ 432). (**b**) Venn diagram shows overlap between the target genes of the three indicated transcription factors. (**c**) Violin plot shows the $\log_e$ hazard ratios for target genes shared among (left, N=26) and unique to (center, N=62) three transcription factors, and for mesenchymal marker genes (right, N=161). Mean hazard ratios are shown by white dots and standard deviations are shown by bars. P values were calculated by a two-sided Wilcoxon rank sum test. (**d**) Browser track of ADM shows the average of RPM normalized (le)ChRO-seq signals and dREG-HD scores in mesenchymal (MES, n= 3) and non-MES (n= 8) GBMs. MES-biased TREs and motif positions are highlighted in blue. (**e**) Kaplan–Meier plot shows overall survival between 196 patients with high and low average expression level of 26 shared target genes. The cutoff was determined based on the minimum *p* value in the difference between survival time using a two-sided Chi-squared test. Shaded regions mark the 95% confidence interval.

1.4 Discussion

Nascent transcription is a promising approach for studying the molecular basis of complex disease because unstable RNAs provide deep insights into multiple stages of gene regulation. ChRO-seq maps nascent transcription in virtually any sample that maintains the integrity of protein-DNA interactions – even those in which RNA is highly degraded. ChRO-seq has important applications throughout the biomedical sciences in analyzing regulatory programs that contribute to solid tumors and other tissues which have proven challenging to study using existing molecular tools.

Our analysis of 20 primary tumors revealed several insights into transcriptional regulatory programs in malignant tissue. First, we report that enhancers in malignant tissue were surprisingly similar to DHSs in the tissue of origin. This finding suggests that regulatory programs in GBM often work within the confines of chromatin architecture that is established in the initiating cell. Regulatory programs were also similar to normal brain in PDXs, demonstrating that tumor initiating cells are able to reconstitute a diverse cell environment that bares surprising similarity to primary brain tissue. Yet how are malignant cell behaviors specified by cancer cells despite this similarity? We found a rare population of ectopic enhancers that resembled fetal tissues isolated from the nervous system, immune cells, and differentiated tumor cells. Our observations are consistent with models of tumorigenesis in which tumor cells reactivate regulatory programs that were similar in some respects to an earlier developmental stage(Stergachis et al. 2013). These regulatory signatures derived from rare ectopic enhancers may have important prognostic value that can be exploited in future studies.

Our study highlights how transcription factors are responsible for coordinated changes in the expression of groups of genes that contribute to expression heterogeneity among tumors. ChRO-seq, like other run on technologies(Azofeifa et al. 2018), provides substantial information about the regulatory activities of transcription factors on chromatin that is independent of transcription factor expression levels. In support of our general approach, transcription factor candidates activating TREs in the stem-like regulatory program were similar to those reported previously to be sufficient for initiating tumors in a murine host(Suvà et al. 2014). Additionally, we used ChRO-seq data to identify transcription factors that establish differences in gene expression characteristic of reported GBM subtypes.

We report three transcription factors, C/EBP, RAR, and NF-κB, whose target genes were systematically correlated with poor clinical outcomes. Our work adds new transcription factors to the current literature, as well as additional support for the role of C/EBP in driving mesenchymal transformation(Carro et al. 2010). NF-κB was previously associated with resistance to radiotherapy and involvement in mesenchymal transformation in GBMs48. Our present work builds on these studies to show that NF-κB activation has an unambiguous influence on clinical outcomes. Additionally, we found evidence that a third transcription factor, RAR, drives regulatory programs that contribute to survival in GBMs. Notably, post-transcriptional mechanisms are responsible for activating two of these three transcription factors, NF-κB and RAR. Thus insights reported here were possible only because ChRO-seq is a more direct indicator of transcription factor activity than other tools previously applied in GBM. As the pharmacology for targeting diverse transcription factor families develops, the transcription factors reported here, as well as our strategies for finding them, will become more useful in nominating targeted therapies.

1.5 Methods

Cell culture. Jurkat cells were grown in RPMI-1640 supplemented with 10% fetal bovine serum, 1X Penicillin/Streptomycin Antibiotic, 0.125 mg/ml Gentamicin Antibiotic at 37$_{o}$C, 5% CO$_{2}$. 1x10$_{6}$ cells were centrifuged at 700 x g 4$_{o}$C 5 min. The media was removed, and the cells were rinsed with 1X PBS, centrifuged, and PBS was removed.

Tissue collection and preparation. Glioblastoma-derived cells were prepared from freshly biopsied human tumors obtained with patient consent. Sample collection was approval by the Institutional Review Board at SUNY Upstate Hospital, Syracuse, NY, and followed all relevant ethical regulations. The non-tumor brain sample was dissected from the brain of an epileptic patient, also with informed consent and IRB approval. To establish patient-derived xenografts, small pieces of freshly resected gliomas were implanted subcutaneously in the flank of athymic nude (nu/nu) mice (Harlan Laboratories / Envigo, Indianapolis,IN) and serially passaged (mouse-to-mouse) 3 times for PDX-UMU88-02, 7 times for PDX-UMU89-08, and 57 times for PDX-88-04 p57, as previously described(Canute et al. 1998; Eller et al. 2002). All mouse work was approved by the SUNY Upstate IACUC and followed all relevant ethical regulations. To prepare chromatin pellets tissue samples were pulverized in a cell crusher. The Cellcrusher was chilled in liquid nitrogen. Frozen glioblastoma tissue (~ 100 mg) was placed in the Cellcrusher, the pestle is placed into the Cellcrusher, and the pestle was stuck with the mallet until the tissue was fractured into a fine powder.

Chromatin isolation. The chromatin isolation was based on work first described in (Wuarin and Schibler 1994). For chromatin (ChRO) isolation from cultured cells or tissue we added 1 ml of 1x NUN Buffer (0.3 M NaCl, 1M Urea, 1% NP-40, 20 mM

HEPES, pH 7.5, 7.5 mM MgCl2, 0.2 mM EDTA, 1 mM DTT, 20 units/ml RNase Inhibitor (Life Technologies # AM2694), 1X Protease Inhibitor Cocktail (Roche # 11 873 580 001)). Samples were vigorously vortexed for one minute. An additional 500 µl of appropriate NUN Buffer was added to each sample and vigorously vortexed for an additional 30 seconds. For length extension chromatin (leChRO) isolation from cultured cells or tissue we added 1 ml of 1x NUN Buffer, as described previously, spiked with 50 units/ml RNase Cocktail Enzyme Mix (Ambion # 2286) in place of the RNase Inhibitor. The samples were incubated on ice for 30 minutes with a brief vortex every 10 minutes. Samples were centrifuged at 12,500 x g at 4$_o$C for 30 minutes after which the NUN Buffer was removed from the chromatin pellet. The chromatin pellet was washed with 1 ml 50 mM Tris-HCl, pH 7.5 supplemented with 40 units/ml RNase Inhibitor (Life Technologies # AM2694), centrifuged at 10,000 x g, 4$_o$C, for 5 minutes, and buffer discarded. The chromatin was washed two additional times. After washing, 100 µl of chromatin storage buffer (50mM Tris-HCl, pH 8.0, 25% Glycerol, 5mM MgAc$_2$, 0.1mM EDTA, 5mM DTT, 40 units/ml RNase Inhibitor) was added to each sample. The samples were loaded into the Bioruptor and sonicated using the following conditions: power setting on high, cycle time of ten minutes with cycle durations of 30 seconds on and 30 seconds off. The sonication was repeated up to 3 times as needed to get the chromatin pellet into suspension. Samples were stored at -80$_o$C.

Chromatin Run-On and sequencing (ChRO-seq) library preparation. After chromatin isolation, the chromatin run-on and sequencing library prep closely followed the methods described previously(Mahat et al. 2016). Chromatin from 1x10$_6$ Jurkat T-cells or 10-100 mg of primary glioblastoma or 100 mg of PDX in 100 µL chromatin storage buffer was mixed with 100 µL of 2x chromatin run-on buffer (10 mM Tris-HCl pH 8.0, 5 mM MgCl$_2$,1 mM DTT, 300 mM KCl, 400 µM ATP (NEB #

N0450S), 40 μM Biotin-11-CTP (Perkin Elmer # NEL542001EA), 400 μM GTP (NEB # N0450S), 40 μM Biotin-11-UTP (Perkin Elmer # NEL543001EA), 0.8 units/μl SUPERase In RNase Inhibitor (Life Technologies # AM2694), 1% Sarkosyl (Fisher Scientific # AC612075000)). The run-on reaction was incubated at 37$_o$C for 5 minutes. The reaction was stopped by adding Trizol LS (Life Technologies # 10296-010) and pelleted with GlycoBlue (Ambion # AM9515) to visualize the RNA pellet. The RNA pellet was resuspended in DEPC treated water and heat denatured at 65$_o$C for 40 seconds. In ChRO-seq, we digested RNA by base hydrolysis in 0.2N NaOH on ice for 8 minutes, which ideally yields RNA lengths ranging from 40 – 100 bases. This step was excluded from leChRO-seq. Nascent RNA was purified by binding streptavidin beads (NEB # S1421S) and washed as described(Mahat et al. 2016). RNA was removed from beads by Trizol and followed by the 3' adapter ligation (NEB # M0204L). A second bead binding was performed followed by a 5' de-capping with RppH (NEB # M0356S). The 5' end was phosphorylated using PNK (NEB # M0201L) followed by a purification with Trizol (Life Technologies # 15596-026). A 5' adapter was then ligated onto the RNA transcript. A third bead binding was then followed by a reverse transcription reaction to generate cDNA (Life Technologies # 18080-044). cDNA was then amplified (NEB # M0491L) to generate the ChRO-seq libraries which were prepared based on manufacturer's' protocol (Illumina) and sequenced using Illumina NextSeq500 at the Cornell University Biotechnology Resource Center.

Mapping ChRO-seq and leChRO-seq sequencing reads. We used our publicly available pipeline to align ChRO-seq and leChRO-seq data. Some libraries were prepared using adapters which contained a molecule-specific unique identifier (first 6 bp sequenced; denoted in Supplementary Table 1.1), and for these we removed PCR duplicates using PRINSEQ lite(Schmieder and Edwards 2011). Adapters were trimmed from the 3’ end of remaining reads using cutadapt with a 10% error rate(Martin 2011). Reads were mapped with BWA(H. Li and Durbin 2010) to the human reference genome (hg19) plus a single copy of the Pol I ribosomal RNA transcription unit (GenBank ID# U13369.1). The location of the RNA polymerase active site was represented by a single base which denotes the 3’ end (ChRO-seq) or 5’ end (leChRO-seq) of the nascent RNA, which corresponds to the position on the 5’ or 3’ end of each sequenced read respectively. Mapped reads converted to bigWig format using BedTools(Quinlan and Hall 2010) and the bedGraphToBigWig program in the Kent Source software package(Kuhn, Haussler, and James Kent 2013). Downstream data analysis was performed using the bigWig software package. All data processing and visualization was done in the R statistical environment(R Development Core Team 2011).

Gene transcription analyses. Gene transcription activity quantification for ChRO-seq and leChRO-seq. We quantified transcriptional activity using gene annotations from GENCODE v25 lift 37, expect for the cross-comparison with TCGA RNA-seq data where we used GENCODE v22 lift 37 to match the annotation of TCGA data. We counted reads in the interval between 500 bp downstream of the annotated transcription start site to the end of the gene for comparisons. This window was selected to avoid counting reads in the pause peak near the transcription start site. We limited analyses to gene annotations longer than 1,000 bp in length.

Molecular subtype classification. Transcriptional activity of characteristic genes for each GBM subtype (n = 23) were quantified by the methods described above. Reads count from each sample are normalized by reads per million total reads count, followed by log2 transformation of pseudo count adjusted data (RPM normalized reads count+1). The transformed read count was centered to a mean of zero for each gene. The similarity between each sample was measured by Spearman's rank correlation, and clustered using single link clustering. The similarity of each sample to molecular subtypes(Verhaak et al. 2010) were calculated using Pearson's correlation with the centroid of corresponding subtype.

Differential expression analysis (DESeq2) for annotated genes. Patients clustered in each dominant molecular subtype were treated as biological replicates (Fig. 1.2b and Supplementary Table 1.3). Two technical replicates of non-malignant brain were used as control. Differential expression analysis was conducted using DESeq2(Love, Huber, and Anders 2014) and differentially expressed genes were defined as those with a false discovery rate (FDR) less than 0.05.

Comparison of TREs with DNase-I hypersensitive sites. dREG-HD. dREG-HD was run using the default settings. A complete description of dREG-HD can be found in Supplementary Note 1.6.

Data processing for calling DNase-I hypersensitive sites and dREG-HD sites. We reprocessed all DNase-I-seq data and identified DNase-I hypersensitive sites (DHSs) using a uniform pipeline. We retrieved mapped reads from either ENCODE or Epigenome roadmap projects aligned to hg19. We called peaks in individual biological replicates, 921 samples in total, using MACS2(Zhang et al. 2008) and Hotspot(John et al. 2011). To group DHSs for each cell and tissue type with high confidence, we took the union of peaks (bedtools merge) from biological replicates followed by intersecting peaks called by Hotspot and MACS2. Lastly since peaks resulted from intersection may be too narrow and hence become missed during downstream intersection operations, we expanded all short peaks (<150bp) to 150bp from the peak center. Analyses involving individual replicates, in Supplementary Fig. 1.11, use only peaks called by MACS2.

ChRO/leChRO-seq data was mapped to hg19 as described above. dREG score was thresholded at 0.7 to generate dREG peak regions for dREG-HD run. dREG-HD runs were done at the stringent condition, except for analysis of subtype biased TREs, where we used dREG-HD sites called at relaxed condition.

Mutual information analysis. We used mutual information to compare the similarity between TREs observed in any pair of DHS or dREG-HD datasets. DHSs or dREG-HD peaks of sample involved in the comparison were merged in order to construct a sample space in which two or more samples would be compared. Each dataset was then summarized as a random variable, represented by a zero-one vector in which each element represents a TREs in the sample space, and takes a value of 1 if it intersects with that peak and 0 otherwise. We calculated the mutual information between two random variables, X and Y, using the formula below:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Comparison between tumor and reference brain tissues and cell lines. We selected brain-related samples from uniformly processed DHSs and categorized the reference dataset by sample origin, namely normal adult brain tissues (globus pallidus, midbrain, frontal cortex, middle frontal gyrus, cerebellum and cerebellar cortex), primary brain cells (astrocyte of the hippocampus, astrocyte of the cerebellum, and normal human astrocytes), and GBM cell lines (A172, H54 and M059J).

Mutual information heatmap and clustering analysis. To compare the similarity between the dREG-HD sites in each query samples and DHSs in each reference sample (Fig. 1.3c), we computed the pairwise mutual information between each pair of dREG-HD and DHSs (as described above) on the sample space defined by merged peaks among all samples included in the analysis. Information on clustering samples based on mutual information, as in Supplementary Fig. 1.14, can be found in Supplementary Note 1.7.

TRE clustering analysis. We analyzed the activation pattern across TREs, using the same definition of sample space described in the mutual information analysis (above). We assigned two states to each TRE, active if intersected dREG-HD/ DHS, and inactive otherwise. The Jaccard distance was used to quantify the similarity between two samples or between two potential TREs. Clustering across samples (columns) and across TREs (rows) was done using ward.D2 method. To reduce the influence of noise on the clusters, we limited analysis to TREs that were activated in at least two query samples but less than 6 brain-related reference samples (16 samples in total).

taTRE enrichment test and clustering into regulatory programs. taTREs were defined as TREs from primary GBM / PDX that do not intersect with any dREG-HD peaks from our non-malignant brain control nor with DHSs found in normal brain tissues (including globus pallidus, midbrain, frontal cortex, middle frontal gyrus, cerebellum and cerebellar cortex). These taTREs represent a stringent subset enriched for TREs associated with the malignant phenotypes observed in brain tumors. dREG-HD sites or DHSs that overlapped with ENCODE consensus hg19 blacklist regions were excluded from analysis.

The majority of taTREs intersected DHSs in one or more reference ENCODE and

Epigenome Roadmap samples (Fig. 1.3a). We devised a statistical test to determine whether the observed number of intersections with each reference sample is significantly higher than expected by chance. We generated a null distribution by sampling DHSs with replacement from all TREs found in reference samples, controlling for the distribution of uniqueness (i.e., the number reference samples which each taTRE intersects) of taTREs from a particular GBM / PDX. The simulation was run for 105 times for each sample, each simulation drawing the same number of taTREs observed in that sample. We selected tissues with a stringent statistical significance cutoff of p(Xnull > xobserved) ≤ 1/104. Reference samples that showed significant enrichment in at least one third of (≥8) GBM or PDX were chosen as taTRE-associated references for downstream analysis.

In total 50 significant taTRE-enriched reference samples were clustered by methods described in the TRE clustering analysis section. Fold of enrichment was calculated as the $x_{observed}$ / E[X_{null}]. The dendrogram was cut down to three clusters. DHS regions that show up in more than half of reference samples in each cluster were picked as representative DHS driving a regulatory program that is characteristic for that cluster. taTREs overlapping these representative DHSs unique to each cluster were selected for downstream analysis.

Motif enrichment analysis of tumor-associated TREs and subtype-biased TREs. Defining subtype-biased TREs. To search for TREs that differentially activated or repressed in each subtype, we rely on measuring the change of the nascent RNA in the TRE regions. We merged dREG-HD sites called using the relaxed setting across 23 samples. We summed up the reads count of leChRO/ChRO-seq of each merged dREG-HD sites extended by 250bp from the center. TREs in patients of the subtype of interest (Supplementary Table 1.5) were compared against those of the rest three subtypes. Differential expression analysis was conducted using DESeq2(Love, Huber, and Anders 2014), and subtype-biased TREs are defined as those differentially transcribed with a false discovery rate (FDR) less than 0.01.

Defining genomic regions for motif enrichment comparison. We examined motif enrichment in the positive set compared with a GC-content matched background control set. In the taTRE motif enrichment, we used the group indicated in Supplementary Fig. 1.15 as the positive set, and dREG-HD sites that intersect with active DHSs in the normal brain as the background. For subtype-biased TRE motif enrichment analyses we used up or down-regulated subtype-specific TREs as the positive set and TREs that did not show significant differential transcription (FDR DESeq2 p > 0.1) as the background set. For the positive and background sets we selected the center of peaks and then extended by 150bp from the center. We subsampled background peaks to construct >2,500 GC-content matched TREs before scanning for motif enrichment.

Motif enrichment analysis. We used the R package rtfbsdb to search for motifs that show enrichment in each primary GBM(Z. Wang, Martins, and Danko 2016). We focused on 1,882 human transcription factor binding motifs from the CisBP database(Jolma et al. 2013). When scanning genomic regions of interest, we used TFBSs having a $\log_e$-odds score ≥7 in positive and background sets, with scores obtained by comparing each representative motif model to a second-order Markov background model. Motif enrichment was tested using Fisher’s exact test. To account for potential bias resulted from difference in GC-content between positive and background sets, we ran statistical test on 50 independently subsampled GC-matched dREG-HD regions, and summarized the p values and the fold enrichment across background sets by the median across samples.

We refined motifs discovered using several heuristics, as follows: 1) The motif was enriched (i.e., with a fold enrichment greater than 1), 2) the enrichment was robustly significant to changes in the GC matched background set (median $p < 0.05/ 1882$), 3) the positive sets have at least 10 sites with $\log_e$-odds score ≥7, 4) the transcription factor was transcribed (for subtype-biased TREs). In the subtype-biased TRE analysis for up-regulated TREs we required at least 2 ChRO/leChRO-seq reads in the gene body in all samples of either the subtype of interest or other three subtypes.

Summarizing motif enrichment statistics across patients (taTREs analysis only). Motifs in the taTRE analysis that were enriched in at least one primary GBM (all taTRE against all normal brain TRE) were chosen for downstream analysis. The enrichment statistics of three regulatory modules-taTREs were also summarized by median over the patients that show significant enrichment for the motif. Lastly, for each transcription factors with multiple motif IDs, we reported the one with the most significantly enrichment in all taTREs over nbTREs. In the subtype-biased TRE analysis, we used all motifs meeting the enrichment criteria and heuristics described above.

Motif clustering by genomic position. Because we are not able to rigorously distinguish between paralogous transcription factors that share similar DNA binding specificities, we developed a method of clustering motifs based on their occurrences in the genome. We first scanned motifs enriched over genomic regions defined by the positive set. In clustering motifs enriched in taTREs, we used the taTREs merged over

20 primary GBMs as the positive set; for motifs enriched in subtype biased TREs, we used the corresponding subtype biased TRE in which the motifs were enriched as the positive set. We defined the presence of TFBSs for loci (strand-specific) having a $\log_e$-odds score ≥7 in positive and background sets, and absence otherwise, with scores obtained by the method described in the section Motif enrichment analysis of taTRE. The Spearman's rank order correlation coefficients were computed for each pair of transcription factors, based on their presence/absence pattern across TFBSs of all motifs of interest. Heatmaps were generated using agglomerative hierarchical clustering using the ward.D2 method.

Validation of regulation between transcription factors and target genes.

Associating transcription factors to target genes. We associated transcription factors to target genes by first identifying its target TREs, and then search for target genes based on location of these TREs. To identify target TREs, we scanned "relaxed dREG-HD all GBM" regions, extended by 150bp from the center, using itself as the second-order Markov background model. For each subtype-specific transcription factor, we defined its binding sites as 1) subtype-biased TREs that undergo differentially transcription in the same subtype, and 2) have a $\log_e$-odds score ≥7 for at least one corresponding motif ids that also showed enrichment ($p<0.05$). This subset of TREs represents the potential binding and regulating sites of the TF of interest, referred to as query TREs. We use stringent heuristics link the query TREs to target genes in order to reduce false positive links. TREs were linked to putative target genes if: 1) the annotated transcriptional start site of the genes is the first two closest to the query TRE and within 50kb, and 2) the gene is differentially transcribed (FDR corrected DESeq2 $p < 0.05$) in the same direction as the query TRE.

Defining the background set of non-target genes. We defined background non-target genes of each transcription factor as those distal from (>0.5 Mb) the query TRE, but which show similar changes in transcription as that of target genes (to control for subtype). We required non-target genes had a transcription start site >0.5Mb from the closest query TRE. To match changes in transcription between target and non-target genes, we subsampled half of the genes away from query TREs and differentially transcribed ($p<0.05$) in the same direction as that of target genes without replacement, such that the distribution of $\log_2$ of fold change in transcription was insignificant (two-sided Wilcoxon $p > 0.2$).

Validation of association between transcription factors and target genes. To validate of our approach associating transcription factors to target genes, we compared the co-expression of target genes to that of background non-target genes. Specifically, we used the RPKM normalized TCGA RNA-seq data from 174 GBM patients downloaded from, and used the Spearman's rank correlation to measure the degree of co-expression. To avoid the potential co-expression that might be artificially enriched in target genes due to higher chance of being located in adjacent positions of the genome, we masked the correlations coefficients between adjacent genes. We computed the significance for target genes to have higher co-expression using one-sided Wilcoxon rank-sum test.

Quantifying the association between the transcription level of transcription factors and its target genes. We used the RPKM normalized TCGA RNA-seq data from 174 GBM patients, and used the Spearman's rank correlation to measure the monotonic relation between the transcription level of transcription factors and the putative target genes. We compared the difference between the distribution of correlation coefficients for target and non-target genes using the Wilcoxon rank-sum test and derive the two-sided p value.

Identification of transcription factors driving survival-associated programs. For each subtype-specific transcription factor, we identified the target genes as described above, and compared the hazard ratio of the target genes with that of non-target genes. We defined two sets of background based on non-target genes: 1) the closest genes whose transcription start site was also within 50 kb to the query TRE, but whose transcription unchanged across the samples representing that subtype ($p > 0.2$, Fig. 1.7a, x axis), and 2) genes differentially transcribed ($p < 0.05$) in the same direction as target genes, whose transcription start sites were 0.5Mb away from the closest query TRE (Fig. 1.7a, y axis). The clinical data, the scaled mRNA abundance level of 11,861 genes across 200 GBM patients (196/200 with information of survival days available), and unified over three microarray platforms, was downloaded from TCGA(Verhaak et al. 2010). We computed the hazard ratio of each gene by fitting a Cox proportional hazards regression model for survival time of patients with expression level in upper 25% of transcription levels over those with lower 25%. This ensures that all genes were fit for the regression model using the same balanced number of patients. We used the Wilcoxon test to compare the distribution of hazard ratios of target genes and background genes, and derived a two-sided p values for each background set.

The hazard ratio of analysis for individual transcription factors in Fig. 1.7a and Supplementary Fig. 1.27a-c, and target genes of survival-related transcription factors in Fig. 1.7e, Supplementary Fig. 1.27d-f and 1.30, were determined by the same regression model. The difference was that, instead of using the upper and lower quartiles as the cutoff, we reported the hazard ratio at the threshold between 0.1 quantile and 0.9 quantile that gave the largest difference between survival times. This difference was calculated by two-sided p value from Chi-squared test. This ensured that we reported the largest possible difference in survival time for each individual gene.

URLs: ChRO-seq/ leChRO-seq alignment pipeline: https://github.com/Danko-Lab/utils/tree/master/proseq
dREG-HD implementation: https://github.com/Danko-Lab/dREG.HD
tfTarget implementation: https://github.com/Danko-Lab/tfTarget
bigWig software package: https://github.com/andrelmartins/bigWig
TCGA Microarray data: https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/unifiedScaled.txt

Code availability: Custom scripts for the tfTarget package can be downloaded form: https://github.com/Danko-Lab/tfTarget. dREG-HD can be obtained from: https://github.com/Danko-Lab/dREG.HD.

Data availability: All ChRO-seq and leChRO-seq data can be downloaded from the database of genotypes and phenotypes (dbGaP) under accession number phs001646.v1.p1.

1.6 Supplementary Notes

Supplementary Note 1.1: Comparison between ChRO-seq and other chromatin-based RNA-seq assays

ChRO-seq draws its intellectual heritage from other run-on and sequencing assays(Kwak et al. 2013; Core, Waterfall, and Lis 2008) and from assays that sequence RNA from a chromatin fractionation, such as Nascent-seq(Khodor et al. 2011) and variations of mammalian NET-seq (mNET-seq)(Mayer et al. 2015). Compared with other chromatin-based RNA-seq assays, ChRO-seq includes a run-on reaction to incorporate an affinity tag that is specific to engaged RNA polymerase. This design has a number of advantages compared with other chromatin-based assays. In particular, the biotin tag stringently selects for engaged and transcriptionally competent RNA polymerase, allowing high-quality data even in cases where there is significant contamination from cytoplasmic RNAs, and depleting for highly abundant chromatin associated small RNAs. We expected these advantages to decrease the variability of the assay and provide a higher confidence that each read represents engaged RNA polymerase.

We used metagene plots that normalize gene length and compared the median profiles obtained across annotated genes among all assays. Median ChRO-seq and leChRO-seq signal across annotated genes was within the range of variation observed in PRO-seq data from the same cell line, and differed to varying degrees compared to Nascent-seq and mNET-seq (Supplementary Fig. 1.1a). Among these assays, Nascent-seq was the largest outlier. Nascent-seq was depleted for signal associated with a paused Pol II that was picked up by all other assays, likely because of a stringent size selection of 200-

300 bp after fragmentation that omits short fragments associated with a paused RNA polymerase. Pol II is known to continue transcribing for 5-20 kb after polyadenylation cleavage before transcription termination and these profiles are captured in PRO-seq data(Schwalb et al. 2016). PRO-seq and ChRO-seq show extensive signal for transcription past the polyadenylation site, whereas the signal in both Nascent-seq and mNET-seq drops quickly after the polyadenylation site. There may be a variety of reasons for these differences, including size selection, computational filtering steps(Mayer et al. 2015) and other factors.

In addition to differences in the average profile, mNET-seq has large numbers of reads aligning to specific regions (or “spikes”) within the gene body that are not visible on the average profiles (Supplementary Fig. 1b). Spikes are absent from ChRO-seq data, indicating that they are not associated with transcriptionally competent RNA polymerase, or that polymerase is sufficiently backtracked that signals are not detected in a run-on reaction.

Supplementary Note 1.2: Intra-tumor heterogeneity

We evaluated the concordance of ChRO-seq by analyzing separate slabs of tissue available from the same patient for the normal brain sample and GBM-88-04. In all cases, ChRO-seq data produced reasonably concordant estimates of Pol II both in the bodies and at the 5’ ends of annotated genes (Supplementary Fig. 1.4c-f). To evaluate intra-tumor heterogeneity, we performed intraoperative MRI guided neuronavigation techniques to dissect GBM-15-90 tissue from four tumor regions (Fig. 1.2b) corresponding to the inner mass with necrotic center (core), an area deep within the tumor mass inferior to the necrotic area (deep), a site proximal to the cortical surface superior to the necrotic site (cortex), and an actively infiltrating area at the genu of the posterior corpus callosum (corpus). ChRO-seq libraries in the four GBM regions tested were remarkably highly correlated, especially when compared to inter-tumor heterogeneity (Fig. 1.2b). Transcription in the core was situated between the other three parts of the tumor in a principal component analysis (PCA) (Supplementary Fig. 1.5h), consistent with a model in which the tumor originated within the core and grew outward radially.

Supplementary Note 1.3: Tumor microenvironment explains enhancer differences between primary and *in vitro* tissue cultures

Two models might explain differences in enhancer profiles between primary and cultured GBM cells. Differences might reflect either evolutionary changes as cancer cells adapt to *in vitro* tissue culture conditions, or differences in the cellular microenvironment between tissue culture and primary tumors. To distinguish between these two models, we used TREs to cluster 20 primary GBMs, 3 PDXs, 8 normal brain tissues, 3 GBM cell lines, and 5 brain-related primary cell types which were dissociated from the brain and grown *in vitro* for a limited number of passages. This analysis supported two major clusters, one composed of normal brain and tumor tissues grown *in vivo* and the other of cells grown *in vitro* (Fig. 1.3d, Supplementary Fig. 1.14). Notably, PDX samples clustered with the primary brain samples, demonstrating that PDXs are a reasonably accurate model for many of the transcriptional features associated with primary tumors. That primary brain cells passaged for a limited duration in tissue culture clustered with the GBM models strongly implicates the microenvironment in causing differences in the enhancer landscape of cells.

Supplementary Note 1.4: Comparison between regulatory programs and molecular subtypes

We asked how the stem, immune, and differentiated regulatory programs relate to previously described molecular subtypes in GBM. We used ChRO-seq signal to identify 6,775 TREs that were differentially transcribed in 2-3 primary GBMs most characteristic of each molecular subtype relative to samples representing the other three subtypes ($p < 0.01$, DESeq2; Supplementary Table 1.4). We compared subtype-biased TREs with those in the stem, immune, and differentiated regulatory program. TREs upregulated in mesenchymal GBMs were enriched 6-fold in the immune regulatory program ($p < 1e-10$, Fisher's exact test; Fig. 4c), consistent with the mesenchymal subtype having higher numbers of tumor infiltrating immune cells (Bhat et al. 2013; Q. Wang et al. 2017). TREs up-regulated in neural and proneural GBMs were enriched in signatures in the stem-like program (Fig. 4c). Nevertheless, TREs in the stem, immune, and differentiated regulatory programs did not always correlate with molecular subtype. For instance, two of the neural tumors in our cohort had a substantial immune regulatory program, and several mesenchymal tumors were strongly enriched for a stem-like program (Fig. 4a). Thus, the three regulatory programs discovered on the basis of rare enhancer fingerprints were distinct from previously reported subtypes, motivating correlations between these clusters and clinical outcomes once larger cohorts of tumors are analyzed using ChRO-seq.

Supplementary Note 1.5: Validation of motifs and target genes contributing to subtype heterogeneity

To validate motifs and predicted target genes, we used the expectation that genes which share a common transcription factor should have expression levels that are more highly correlated with one another across tumors. We analyzed an independent RNA-seq dataset from a cohort of 174 primary GBMs(Brennan et al. 2013). Among the 304 transcription factors enriched in any subtype we noted a significantly stronger correlation between putative target genes for 235 (77%) compared with randomly selected genes matched for similar subtype specificity (Fig. 1.5c; Supplementary Fig. 1.24a). Furthermore, in two cases (NF-κB and STAT1), we found PRO-seq or RNA-seq data following activation of a signaling pathway targeting that transcription factor(Luo et al. 2014; Chuong, Elde, and Feschotte 2016). Despite both published experiments occurring in a different cell type and environmental context, we nevertheless found predicted targets to be 3.0-fold (NF-κB; $p < 3.0e-21$, Fisher's exact test) and 6.9-fold (STAT1, $p = 1.9e-11$, Fisher's exact test) enriched in genes responding in these experiments. Finally, as expected, changes in transcription of TREs correlated with nearby genes, and were strongest for the nearest 1-2 genes from each TRE (Supplementary Fig. 1.22). Moreover these changes in the nearest two genes explained many of the markers defined in microarray studies (Verhaak et al. 2010) (Supplementary Fig. 1.23). Thus, we have identified transcription factors contributing to major GBM expression subtypes, and a set of putative target genes of each transcription factor.

Supplementary Note 1.6: Description of the dREG-HD method

Overview. We trained an epsilon-support vector regression (SVR) model that maps PRO-seq, GRO-seq, or ChRO-seq data to smoothed DNase-I-seq intensity values.

Because dREG reliably identifies the location of transcribed TREs that are enriched for DHSs (Danko et al. 2015), with its primary limitation being poor resolution, we limited the training and validation set to dREG sites. The SVR was trained to impute DNase-I values of the positions of interest based on its input PRO-seq data. The trained SVR can then be used to predict DNase-I-seq signal intensities in any cell type for which PRO-seq data is available. To identify the location of transcribed DNase-I hypersensitive sites (DHSs) we developed a heuristic method to identify local maxima in imputed DNase I-seq data. A detailed description of these tools is provided in the following sections. The source code for the R package of dREG-HD is available from https://github.com/Danko-Lab/dREG.HD.git.

Training the dREG-HD support vector regression model. PRO-seq data was normalized by the number of mapped reads and was summarized as a feature vector consisting of ±1800 bp surrounding each site of interest. In total, 113,568 sites were selected, and were divided into 80% for training and 20% for validation. Parameters for the feature vector (e.g., window size) were selected by maximizing the Pearson correlation coefficients between the imputed and experimental DNase-I score over the holdout validation set used during model training (Supplementary Table 1.4). We fit an epsilon-support vector regression model using the Rgtsvm R package(Z. Wang, Martins, and Danko 2016).

We optimized several tuning parameters of the model during training. We tested various kernels, including linear, Gaussian, and sigmoidal. Only the Gaussian kernel was able to accurately impute the DNase-I profile. Experiments optimizing the window size and number of windows revealed that feature vectors with the same total length but different step size result in similar performance on the validation set, but certain combinations with fewer windows achieved much less run time in practice.

The feature vector we selected for dREG-HD used non-overlapping windows of 60bp in size and 30 windows upstream and downstream of each site, and resulted in near-maximal accuracy and short run times on real data. To make imputation less sensitive to outliers, we scaled the normalized PRO-seq feature vector during imputation such that its maximum value is within the 90th percentile of the training examples. This adjustment makes the imputation less sensitive to outliers and improves the correlation and FDR by 4% and 2%, respectively.

The optimized model achieved a log scale Pearson correlation with experimental DNase-I seq data integrated over 80bp non-overlapping windows within dREG regions of 0.66 at sites held out from the K562 dataset on which dREG-HD was trained and 0.60 in a GM12878 GRO-seq dataset that was completely held out during model training and parameter optimization (Supplementary Fig. 1.9).

Curve fitting and peak calling. The imputed DNase-I values were subjected to smoothing and peak calling within each contiguous dREG region. To avoid effects on the edge of dREG regions, we extended dREG sites by ±200bp on each side before peak calling. We fit the imputed DNase-I signal using smoothing cubic spline. We defined a parameter, the knots ratio, to control the degree to which curve fitting smoothed the dREG-HD signal. The degree of freedom (λ) of curve fitting for each extended dREG region was controlled by knots ratio using the following formula.

$$\lambda=(\{\text{number of bp in dREG peak}\} \ / \ \{\text{knots ratio}\}) + 3$$

This formulation allowed the equivalent degrees of freedom to increase proportionally to the length of the dREG peak size, but kept the value of the knots ratio higher than a cubic polynomial.

Next we identified peaks in the imputed dREG-HD signal, defined as local maxima in the smoothed imputed DNase-I-seq profiles. We identified peaks using a set of heuristics. First, we identify local maxima in the dREG-HD signal by regions with a first order derivative of 0. The peak is defined to span the entire region with a negative second order derivative. Because dREG-HD is assumed to fit the shape of a Guassian, this approach constrains peaks to occur in the region between ±σ for a Gaussian-shaped imputed DNase-I profile. We optimized curve fitting and peak calling over two parameters: 1) knots ratio and 2) threshold on the absolute height of a peak. Values of the two parameters were optimized over a grid to achieve a balance between sensitivity and false discovery rate (FDR). We chose two separate parameter combinations: one ‘relaxed’ set of peaks (knots ratio=397.4, and background threshold=0.02) that optimizes for high sensitivity (sensitivity=0.94 at 0.17 FDR), and one stringent condition (knots ratio=1350 and background threshold=0.026) that optimizes for low FDR (sensitivity=0.79 at 0.07FDR).

Validation metric and genome wide performance. We used genomic data in GM12878 and K562 cell lines to train and evaluate the performance of dREG-HD genome-wide. Specificity was defined as the fraction of dREG-HD peaks calls that intersect with at least one of the following sources of genomic data: Duke DNase-I peaks, UW DNase-I peaks, or GRO-cap HMM peaks. Sensitivity was defined as the fraction of true positives, or sites supported by all three sources of data that also overlapped with dREG. To avoid creating small peaks by an intersection operation, all data was merged by first taking a union operation and then by finding sites that are covered by all three data sources. All dREG-HD model training was performed on K562 data. Data from GM12878 was used as a complete holdout dataset that was not used during model training or parameter optimization.

Metaplots for dREG and dREG-HD. Metaplots were generated using the bigWig package for R with the default settings. This package used a subsampling approach to find the profile near a typical site, similar to (Danko et al. 2013). Our approach samples 10% of the peaks without replacement. We take the center of each dREG-HD site and sum up reads by windows of size 20bp for total of 2000 bp in each direction. The sampling procedure is repeated 1000 times, and for each window the 25% quartile (bottom of gray interval), median (solid line), and 75% quartile (top of tray interval) were calculated and displayed on the plot. Data from all plots were generated by the ENCODE project(Dunham et al. 2012).

Supplementary Note 1.7: Description of the dREG-HD method

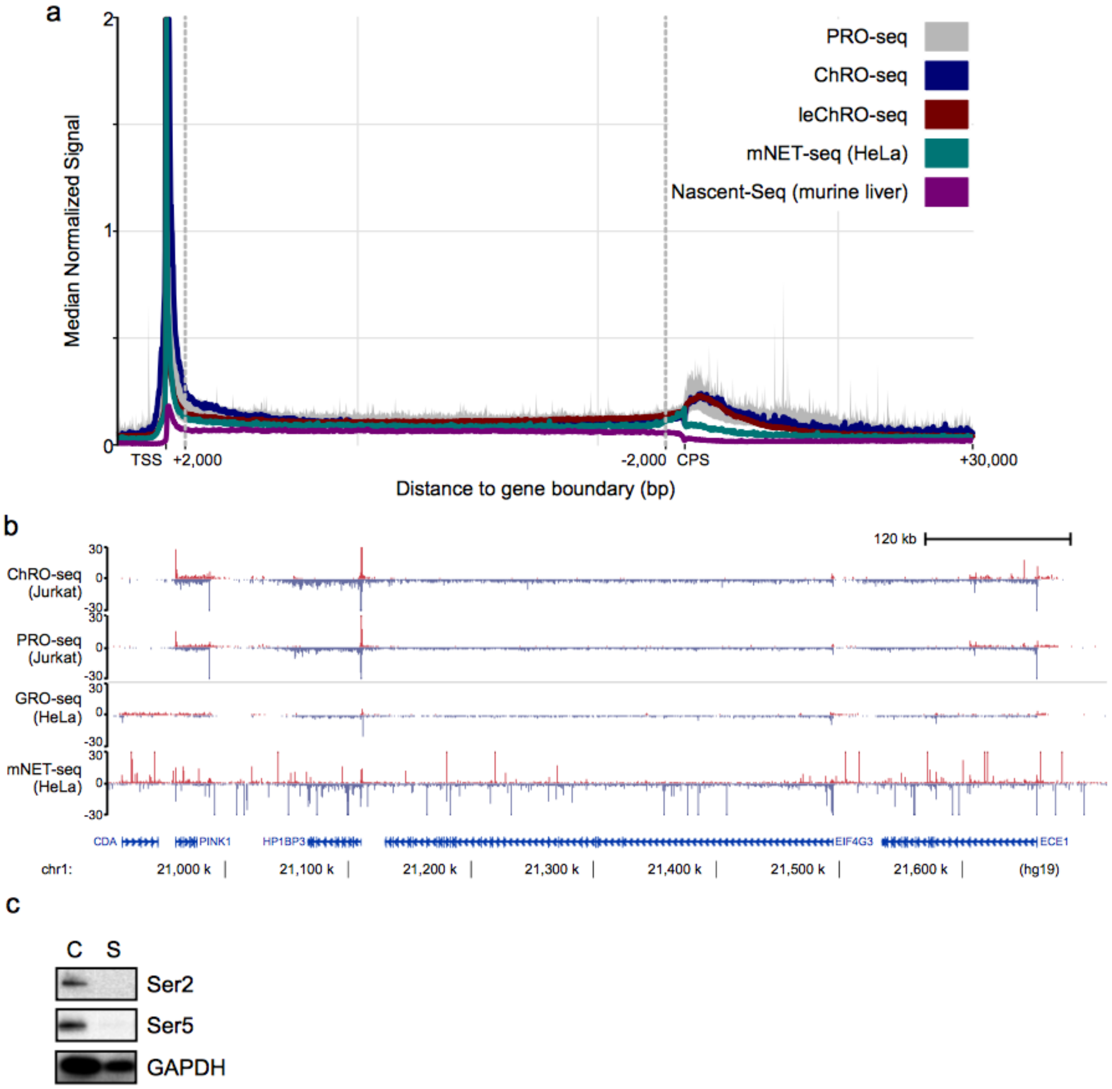
We noted a systematic bias in the distribution of mutual information across query samples that appeared to reflect data quality and sequencing depth in either ChRO-seq or DNase-I-seq data. We devised a strategy to correct for this bias when clustering samples. Our strategy effectively normalizes the mutual information of each query sample with respect to the sum of mutual information for that query sample.

Among multiple samples normalizing the mutual information metric is more complicated. We devised an approach that was used in Supplementary Fig. 1.14. We consider a square matrix with rows and columns representing each sample. Each element in this matrix represents the mutual information between a pair of samples. Our objective is to center the mutual information across each row or column while preserving the rank order and range of mutual information. We accomplished this using the following algorithm, which is similar to (Hastie et al. 2015), but guarantees symmetry:

```
#matrix centering algorithm
WHILE convergence criterion does not meet
    FOR i from 1 to number of columns
        current mean<-mean of ith column
        ith row <- ith row - current mean
        ith column <- ith column - current mean
    END FOR
END WHILE
```

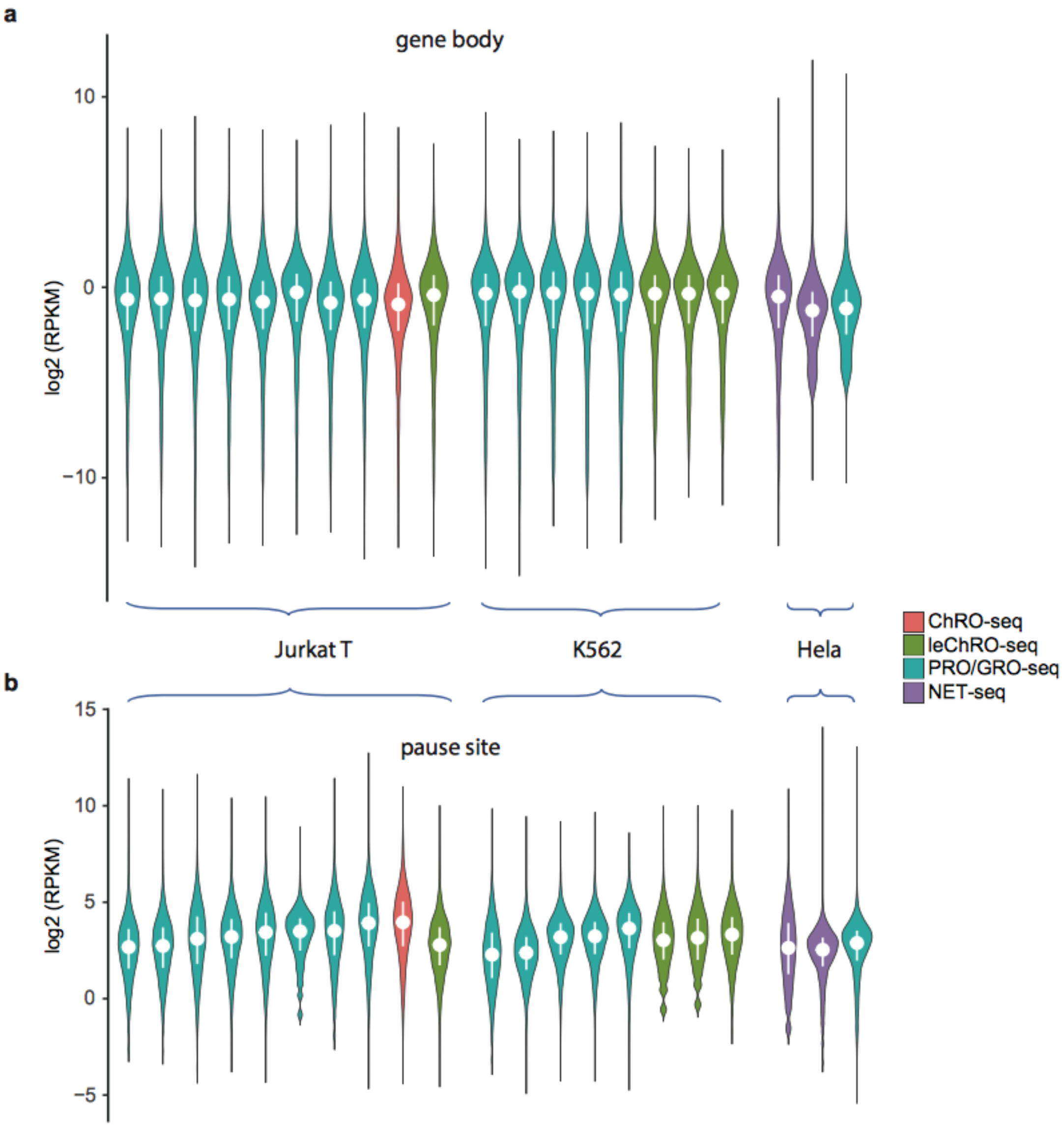
The convergence criterion was defined as the maximum of the absolute value of element-wise difference between matrix returned from previous two consecutive runs. Although there is no mathematical guarantee of convergence, this approach converged typically after four cycles with the datasets that we used. After centering the matrix was clustered using the ward.D2 clustering algorithm implemented in the heatmap function in R.

1.7 Supplementary Figures



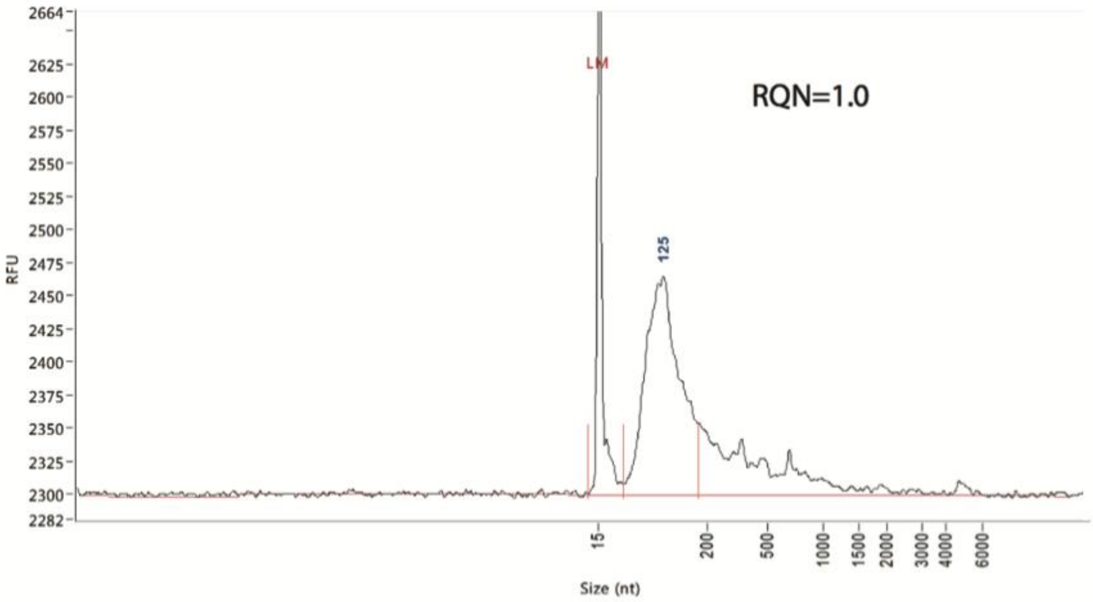
Supplementary Fig. 1. 1 Differences between ChRO-seq and other run-on assays.

(a) Length-normalized meta plots show the median signal across 8,403 active gene bodies using PRO-seq (gray), ChRO-seq (blue), leChRO-seq (red), mNET-seq (teal), and Nascent-Seq (purple). **(b)** The genome browser shows the signal near the EIF4G3 gene locus in ChRO-seq, PRO-seq, GRO-seq, and mNET-seq. **(c)** Western blot showing GAPDH and two active forms of Pol II, defined as phosphorylated serine 2 (ser2) and serine 5 (ser5) in the carboxy-terminal domain, in the chromatin (C) and supernatant (S) fractions.



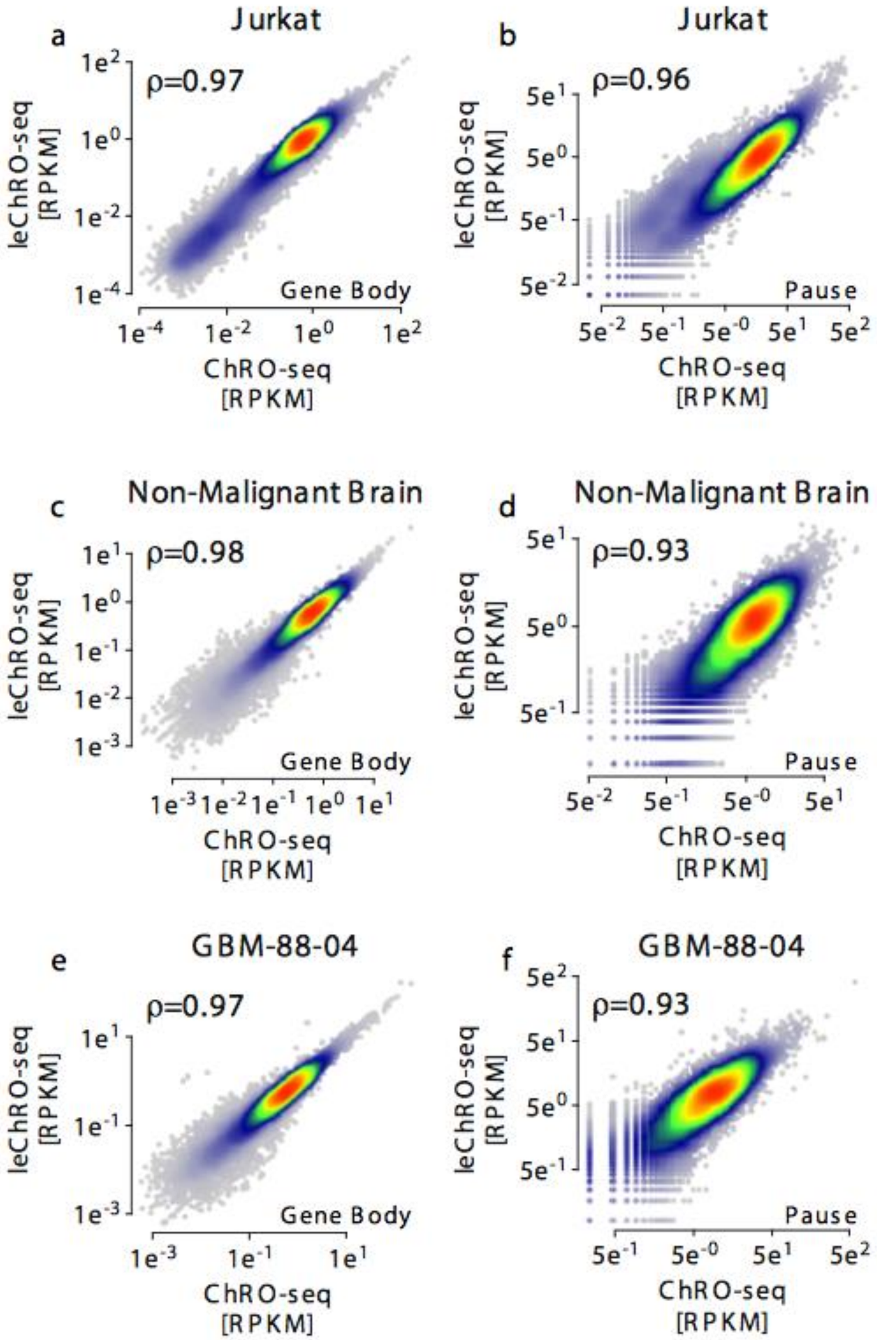
Supplementary Fig. 1. 2 Distribution of signal intensity in the gene body and pause.

Violin plot shows the distribution of $\log_2$ of reads per kilobase per million mapped (RPKM) on (**a**) gene body (N=37,184) and (**b**) pause site (N=37,184). Plots are grouped by cell type and colored by the method. White dots represent the means, while the bars represent standard deviations.



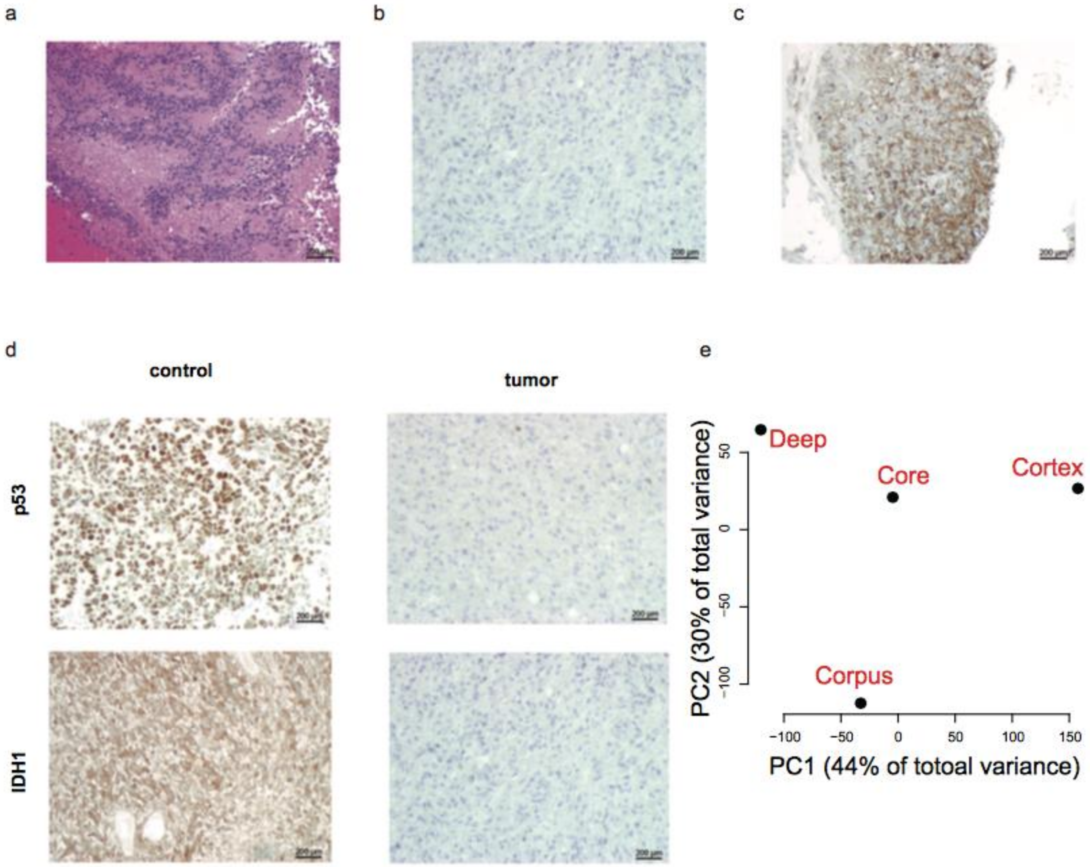
Supplementary Fig. 1. 3 Bioanalyzer analysis of RNA isolated from GBM-88-04.

The plot reported by the Bioanalyzer software shows the size of RNA isolated from GBM-88-04 in units of nucleotides (nt, X-axis) as a function of the relative fluorescence units (RFU, Y-axis). RNA Quality Number (RQN = 1) shown in the trace denotes extensive RNA degradation. The mode of the distribution of RNA sizes is shown (125 nt). The Bioanalyzer analysis was performed once.



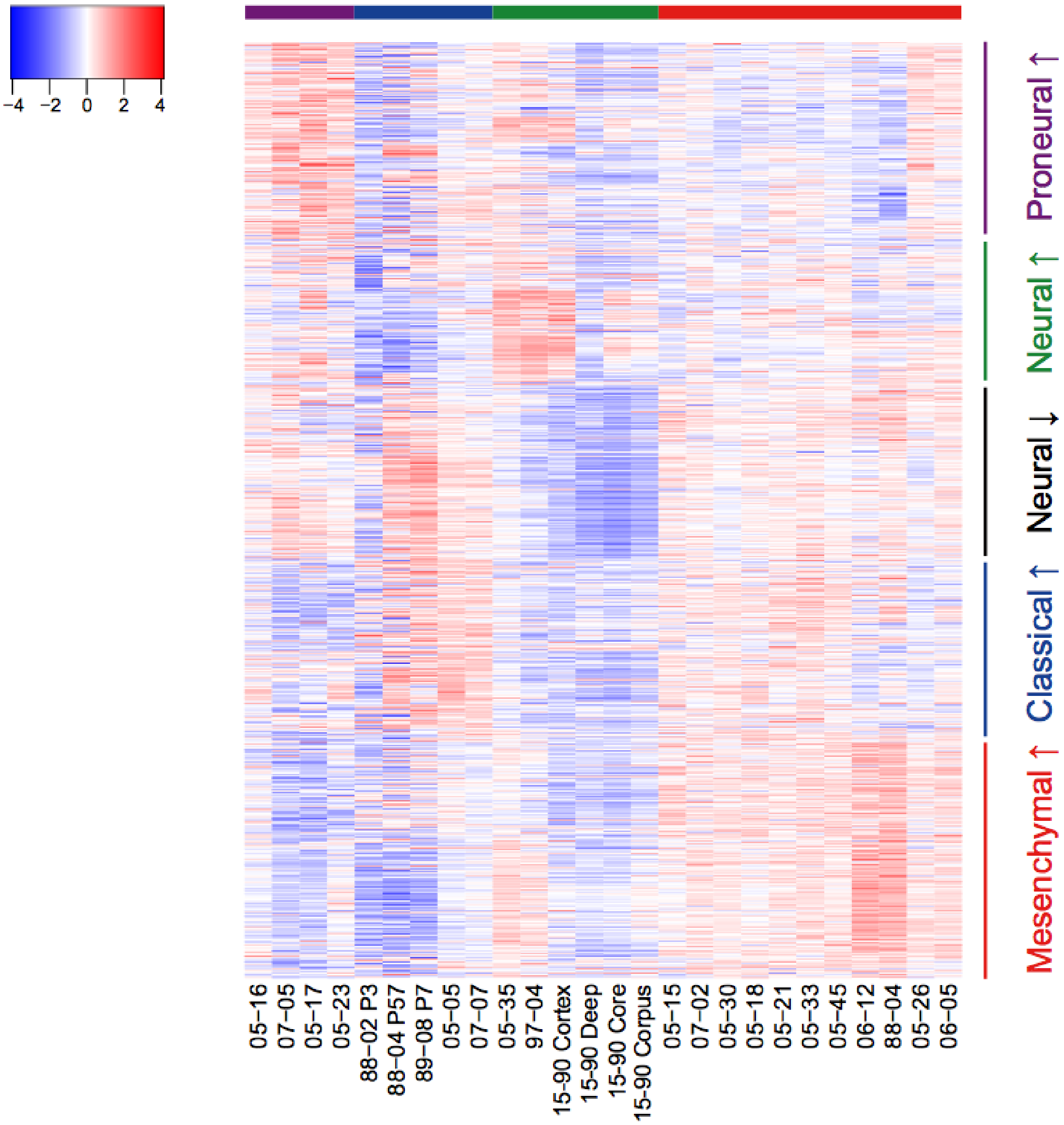
Supplementary Fig. 1. 4 Correlation between ChRO-seq and leChRO-seq.

(**a-f**) Scatterplots show the density of reads mapping in the gene bodies (+1000 to gene end) (**a, c, e**) or in the promoter proximal pause near the transcription start site (**b, d, f**) of 41,478 RefSeq genes. All axes are in units of reads per kilobase per million mapped (RPKM). Spearman's rank correlation (ρ) is shown in each plot. The color scale denotes the density of points.



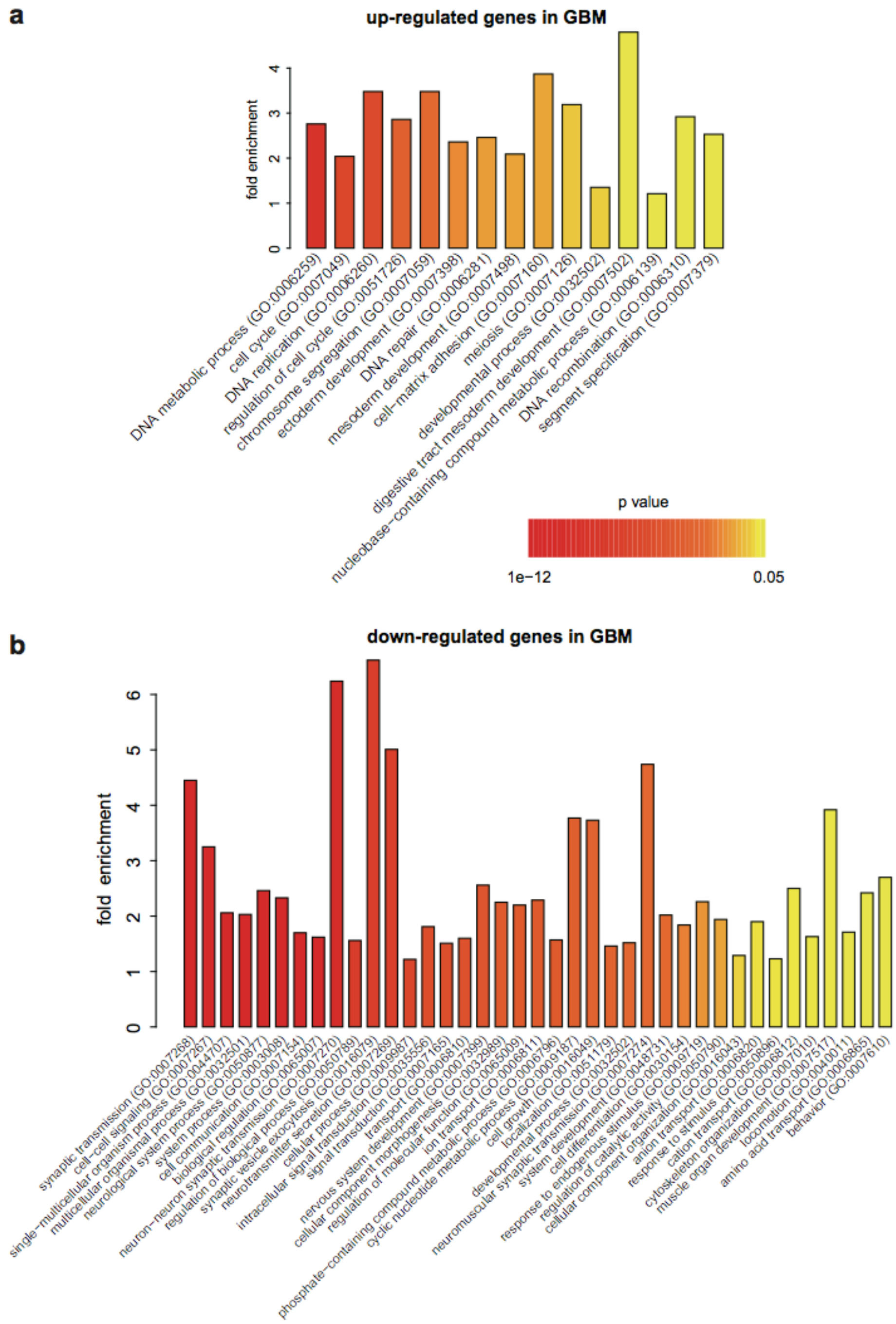
Supplementary Fig. 1. 5. Brain biopsies display immunohistochemical markers of high grade glioma in GBM-15-90.

(**a**) Pseudopalisading borders with necrotic centers. (**b**) IDH1 staining is negative. (**c**) GFAP is stained as positive. (**d**) Additional markers of high grade glioma between the tumor include p53-/- and IDH-/- using an IDH-1 positive glioblastoma as a positive control. All images are representative views from a single patient (GBM-15-90). All scale bars represent 200 µm(**e)** Principal component analysis of transcription in the four tumor regions dissected from GBM-15-90 (N of genes=23,961).



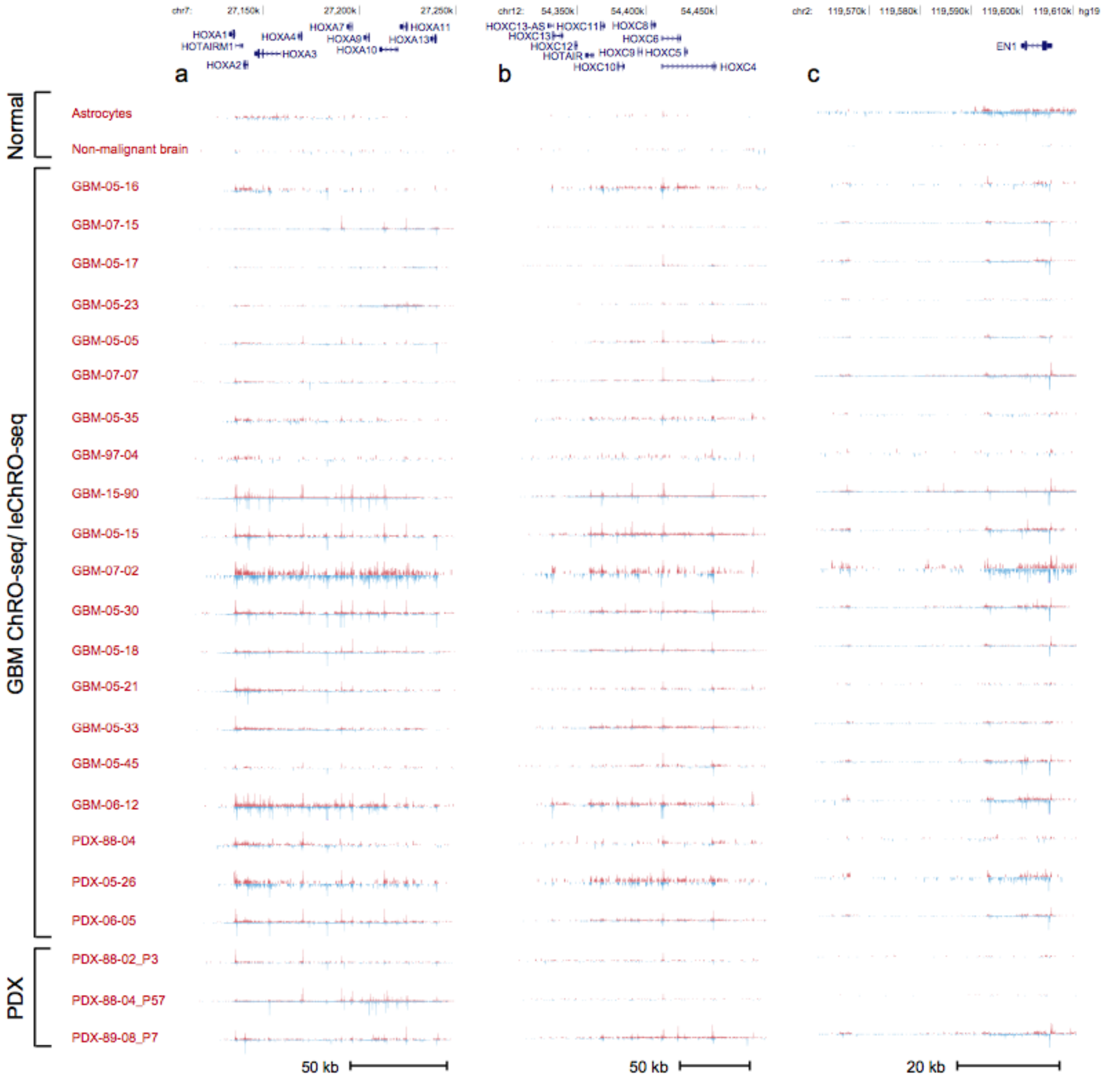
Supplementary Fig. 1. 6 Expression of molecular subtype predictor genes in primary GBM / PDX samples.

Heatmap shows the expression of 838 genes relevant for classifying among the four known molecular subtypes of glioblastoma. Red colors indicate higher transcription activity and blue colors indicate lower activity. Samples are ordered based on subtype.



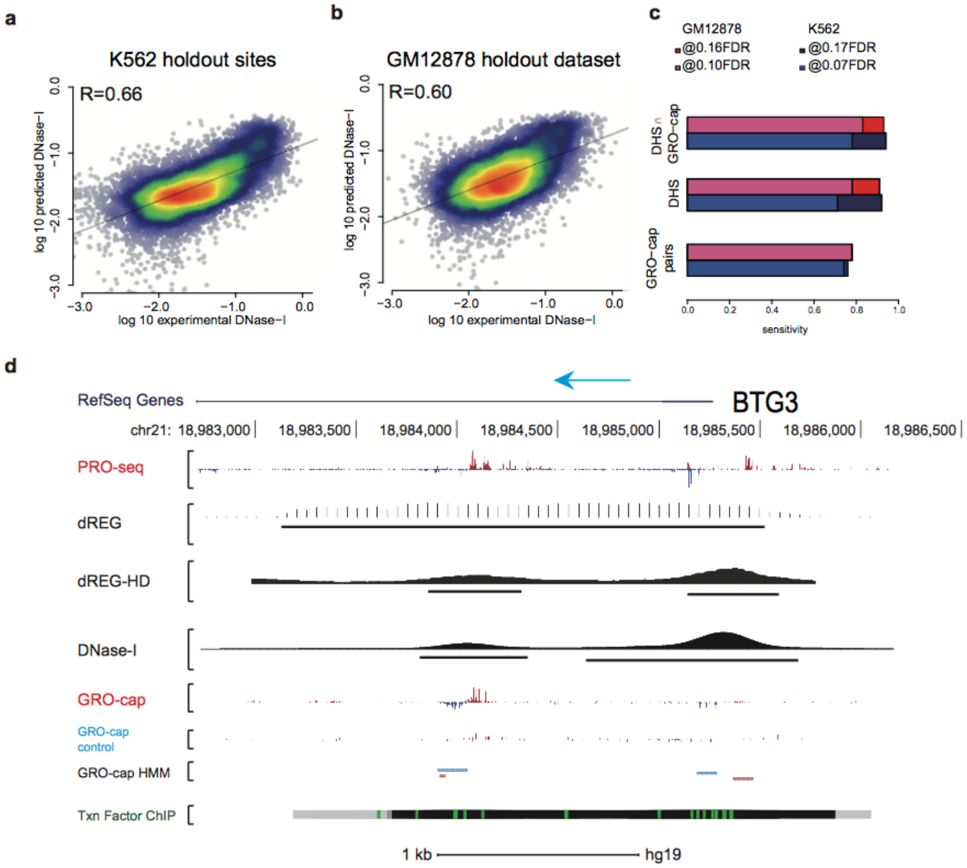
Supplementary Fig. 1. 7 Gene ontology analysis of differentially expressed genes in GBM compared to non-malignant brain tissue.

Barplot shows the the gene ontogoly terms enriched for genes up-regulated in GBM (**a**, N=2,018) and down-regulated in GBM (**b**, N=1,486). Ontology groups are ordered by statistical significance of enrichment and colored by their *p* values (two-sided Fisher's Exact with FDR multiple test correction). The height of each bar indicates the fold enrichment of the indicated gene ontology term.



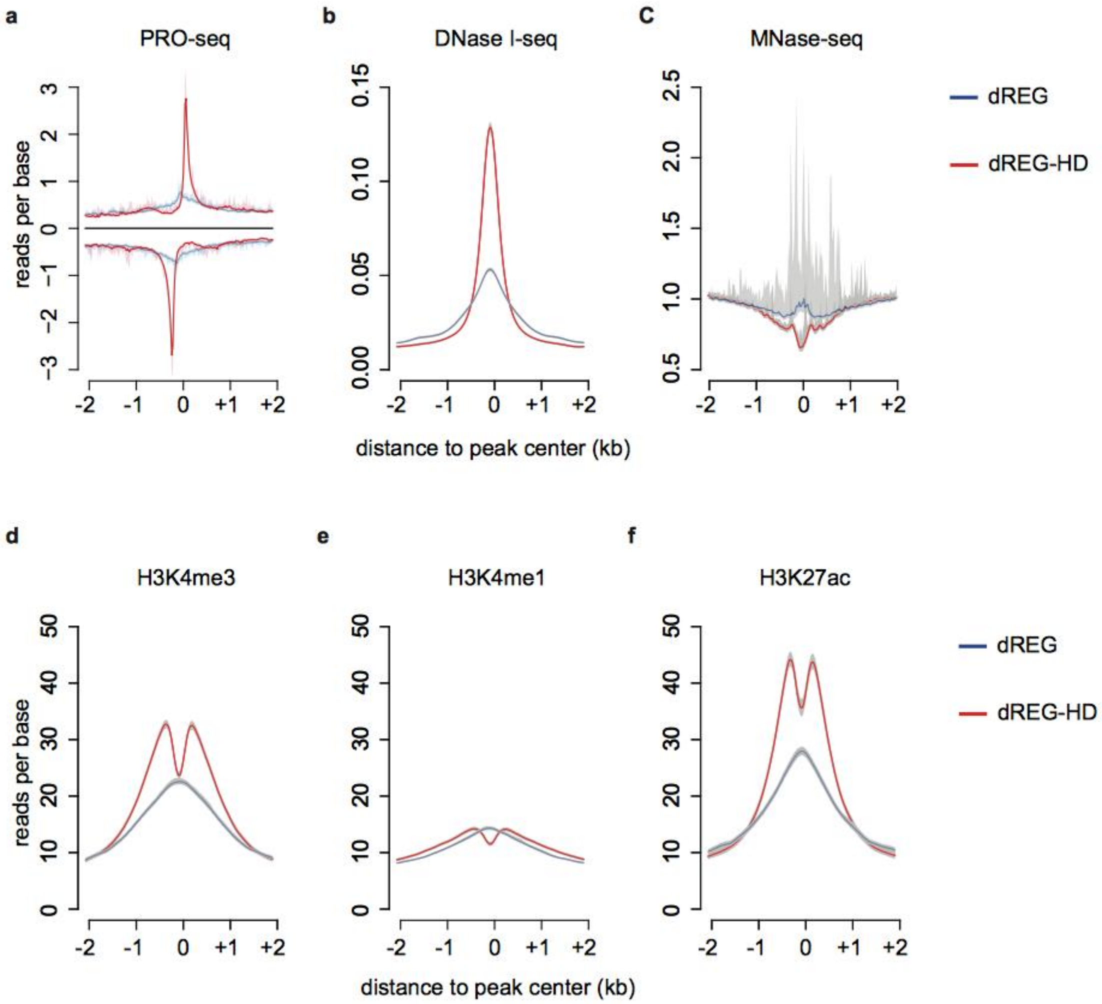
Supplementary Fig. 1. 8 HOXA, HOXC, and EN1 loci show strong differential expression in primary GBM and PDX.

Browser tracks of ChRO-seq signal in primary GBM, PDX, cultured astrocyte, and non-malignant brain samples, DNase-I hypersensitivity in normal adult and fetal brain tissues, and H3K27ac peaks in normal adult brain tissues near (**a**) HOXA, (**b**) HOXC, and (**c**) EN1 loci. ChRO-seq signal signals are normalized by RPM, and summarized by the mean+whiskers function for display. DNase-I hypersensitivity signal is summed across bigWig files of biological replicates from the ENCODE source.



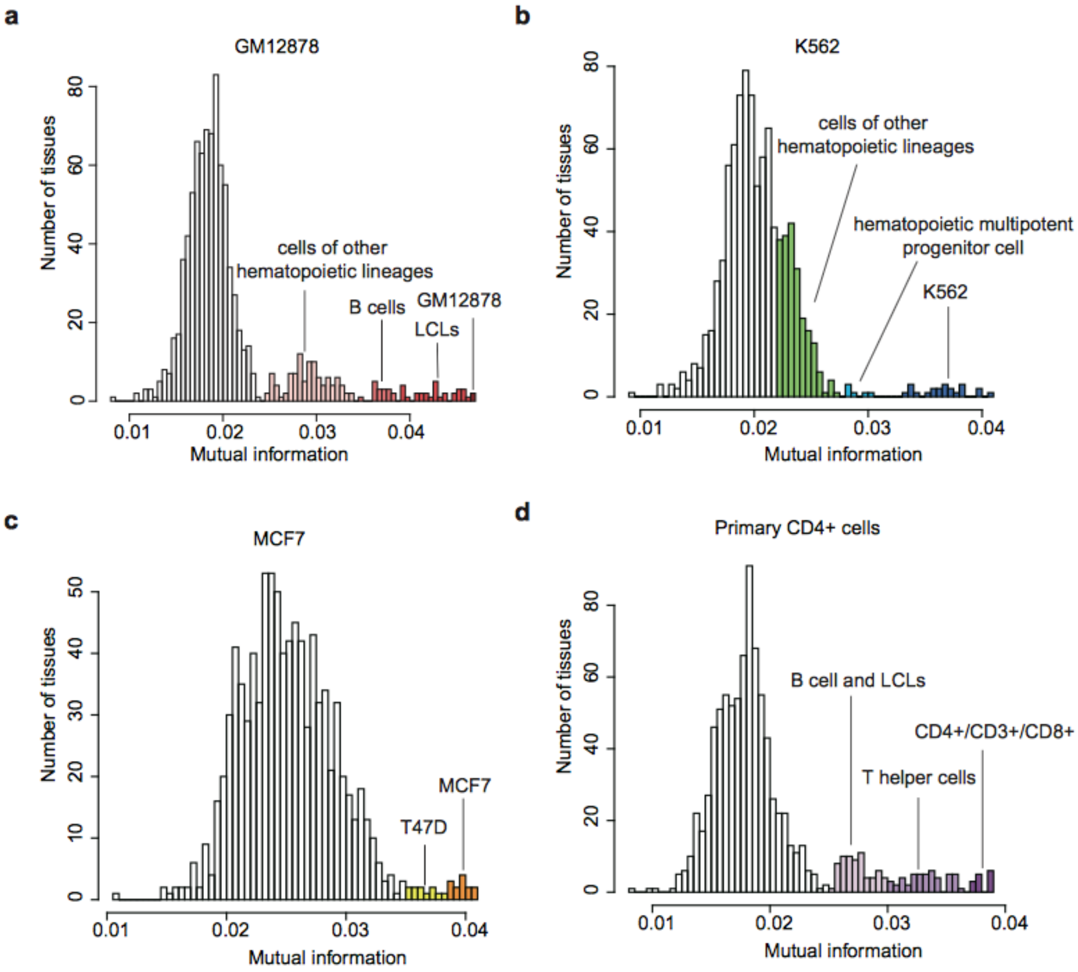
Supplementary Fig. 1. 9 dREG-HD refines TRE predictions by imputing DNase-I hypersensitivity.

(**a** and **b**) Density scatter plots show a comparison between predicted and experimental DNase-I hypersensitivity signals in K562 holdout sites that were not used during training (**a**, N=303,068) and a complete holdout dataset in GM12878 (**b**, N=448,128). Points represent the sum of DNase-I hypersensitivity signals for non-overlapping 80bp windows. (**c**) Sensitivity of dREG-HD to detect DHSs that intersect dREG regions, paired GRO-cap HMM peaks, and the intersection of DHSs and GRO-cap pairs. Prediction in K562 and GM12878 are colored in blue and red respectively. The sensitivity analyzed under 'relaxed' dREG-HD setting was colored in dark red/blue, and those under 'stringent' setting were colored in light red/blue. The expected false discovery rate of the 'relaxed' and 'stringent' settings are indicated above the barplot. (**d**) Browser track of a region near the transcription start site of *BTG3* in K562 cells. From top to bottom tracks represent: 1) RefSeq genes showing the transcription start site of *BTG3*; 2) PRO-seq colored in red (forward) and blue (reverse); 3) dREG scores and peaks; 4) dREG-HD scores and peaks; 5) DNase-I hypersensitivity signal and peaks; 5) GRO-cap reads. 6) The no-TAP control experiment matched to GRO-cap signal; 7) Transcription start sites identified using the GRO-cap signal; 8) Potential transcription factor binding detected by ENCODE ChIP-seq. Peak calls are colored in gray and black and the best match to a transcription factor binding motif is colored in green.



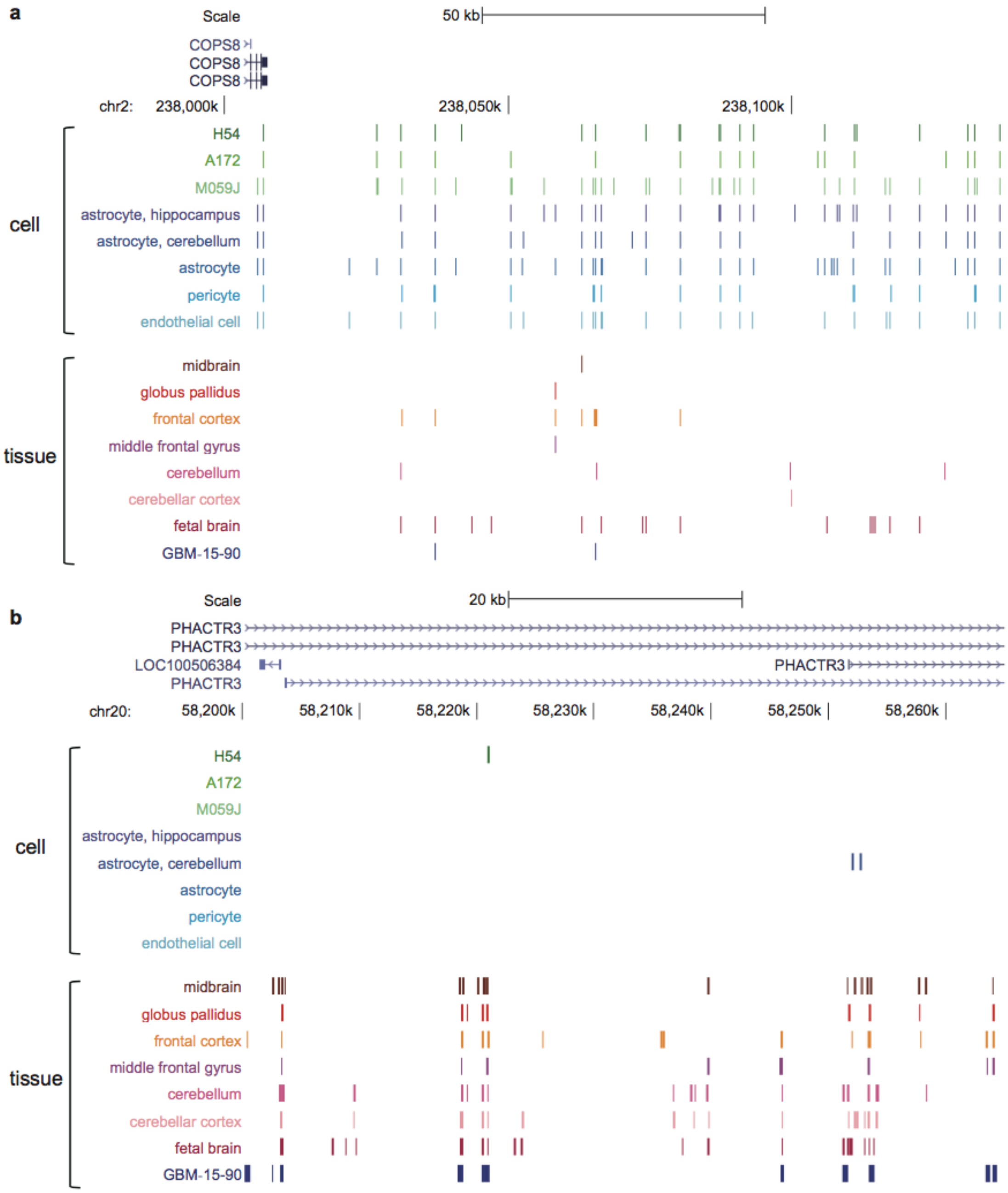
Supplementary Fig. 1. 10 Metaplots for PRO-seq, chromosome accessibility, and histone modifications that marks active TREs.

Signals of the indicated mark over dREG and dREG-HD regions are shown in blue and red, respectively. Shadows marks the 25 and 75 percentiles of 1000 samples of 10% of the data (see methods).



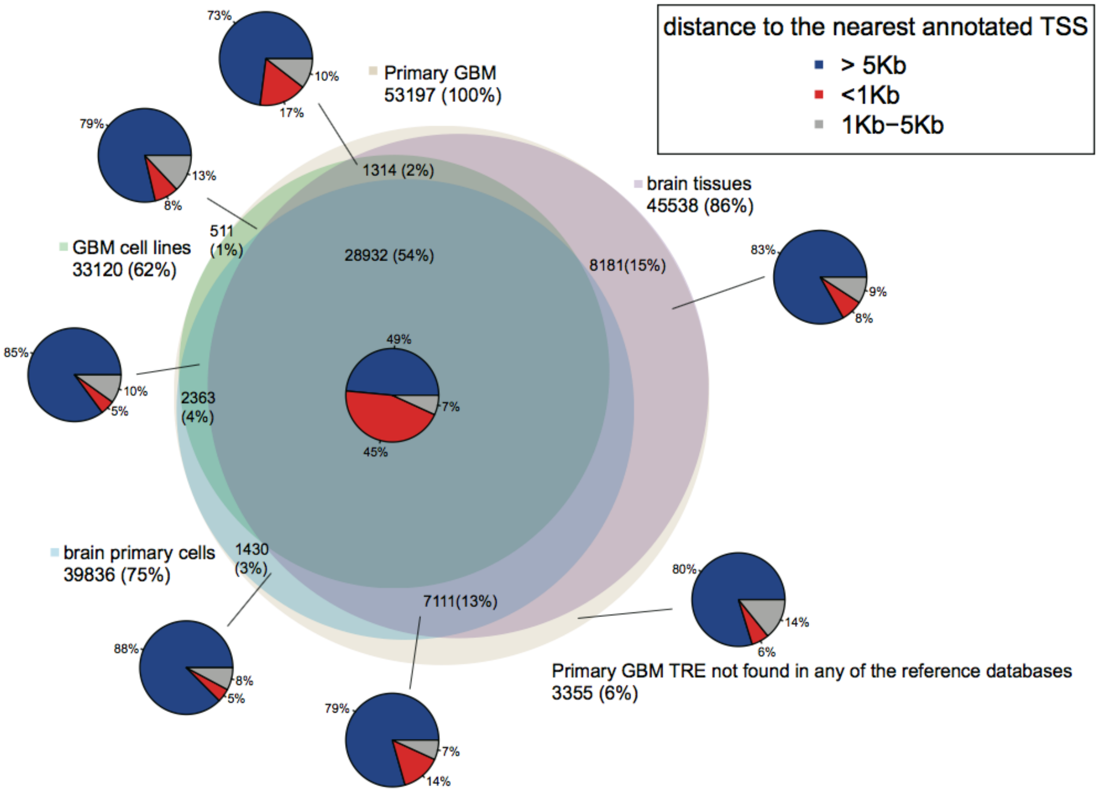
Supplementary Fig. 1. 11 Mutual information is an accurate similarity measure for TREs.

Histogram represents the mutual information between dREG-HD sites identified using PRO-seq or GRO-seq data and DHSs from 921 public DNase-I-seq experiments and in the indicated sample (**a**:GM12878, **b**:K562, **c**:MCF-7, **d:**human primary CD4+ T-cells). In all cases, mutual information selects the sample that was most similar in the reference DHS data, including those of the same or similar cell types, are highlighted.



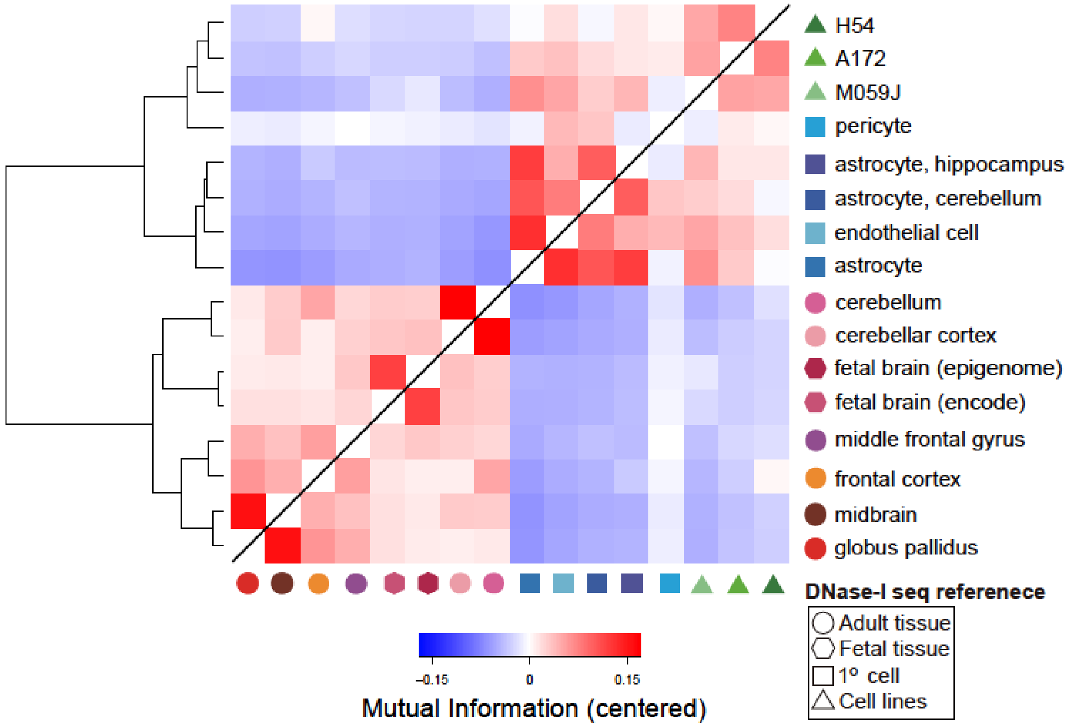
Supplementary Fig. 1. 12 DNase-I hypersensitive sites with differences between brain tissues and cultured brain cells.

(**a**) Locus near the *COPS8* gene that shows consecutive activation of TREs in cultured brain cells but not in normal brain tissues. (**b**) Locus near *PHACTR3* gene that shows activation of TREs in primary brain tissues but not in cultured brain cells.



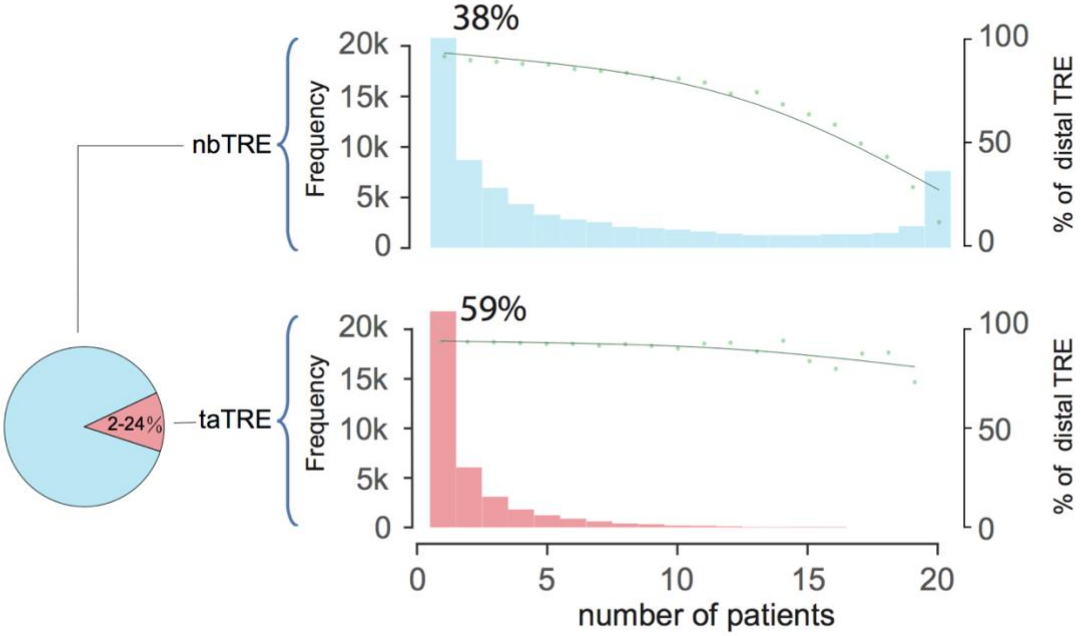
Supplementary Fig. 1. 13 Venn diagram showing similarity in TREs between primary GBMs, normal brain tissue, and primary brain cells grown in tissue culture.

Venn diagram denotes the overlap between TREs found in GBM-15-90 and normal brain (pink), GBM cell line models (green), or primary brain cells that were dissociated from normal brain tissue and grown in culture for a limited number of passages (teal). For each overlap, the number and fraction of TREs is shown. Pie charts denote the fraction of TREs that are >5kb from the nearest annotated transcription start site (blue), <1kb (red), or between 1kb-5kb (gray).



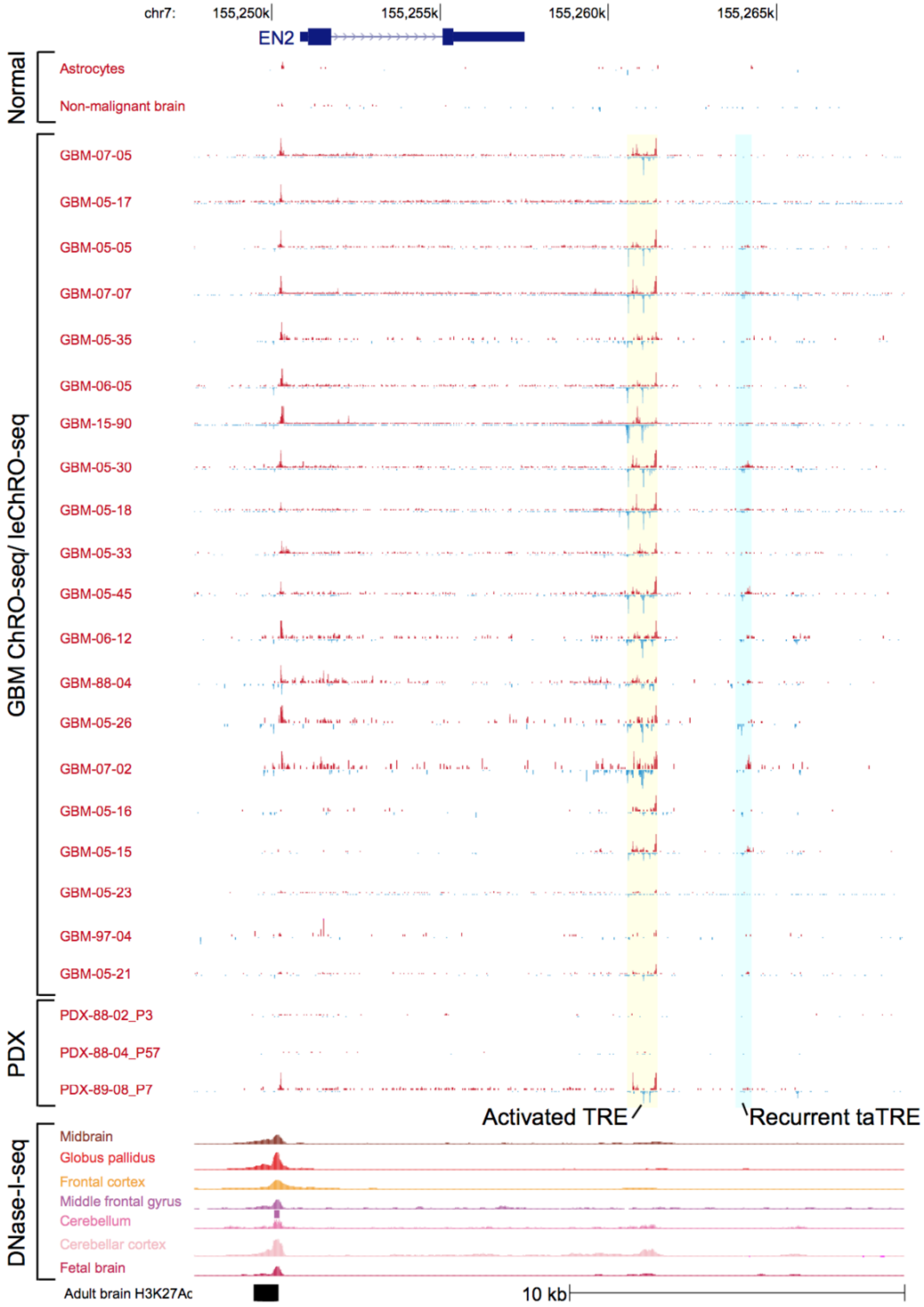
Supplementary Fig. 1. 14 Pairwise mutual information among TREs from brain-related reference DHSs centered by the mean of each sample.

Heatmap shows the centered mutual information between the indicated samples. Sample order was selected by hierarchical clustering using the algorithm described in Supplementary Note 1.7.



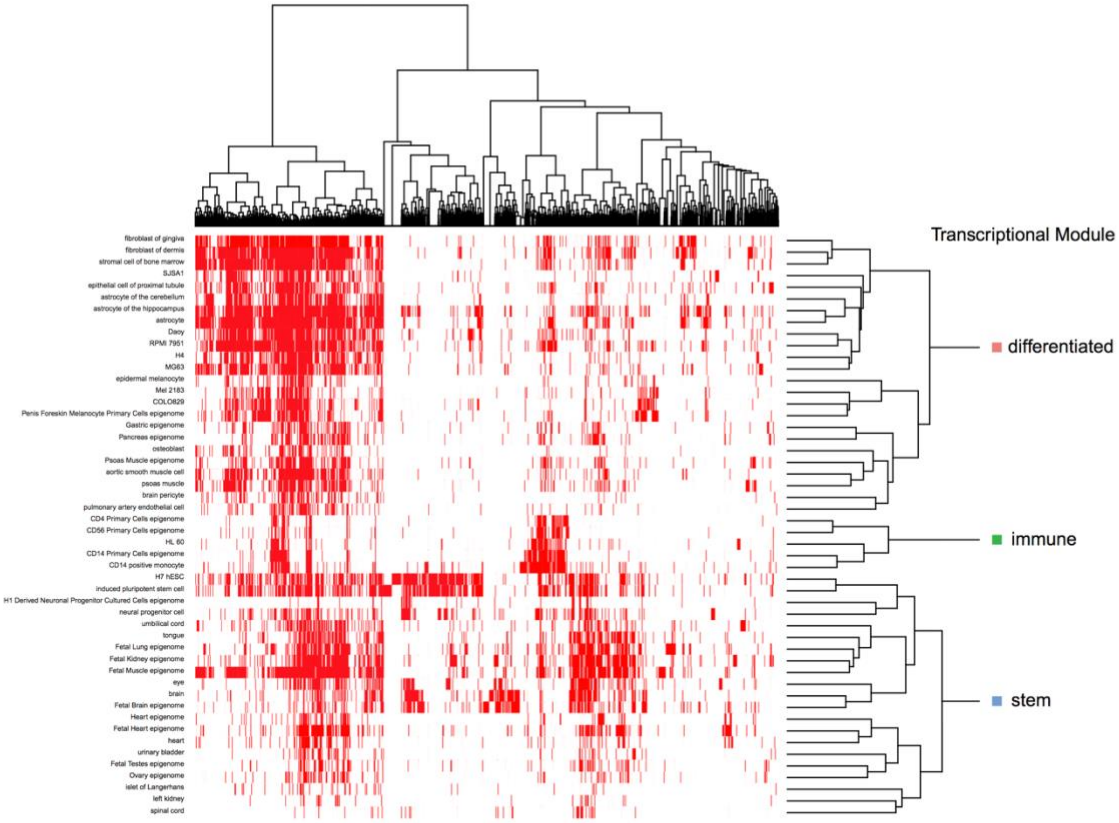
Supplementary Fig. 1. 15 Distribution of the frequency across GBM patients of normal brain and taTREs.

Histograms show the distribution of the number of primary GBM patients (out of 20) in which each TRE is active. 2 to 24% of TREs in GBM samples are not found in normal adult brain tissues. The percentage of TREs >1kb from the nearest transcription start site (distal) is shown in green dots.


chr7:
155,250k
155,255k
155,260k
155,265k
EN2
Normal
Astrocytes
Non-malignant brain
GBM ChRO-seq/ leChRO-seq
GBM-07-05
GBM-05-17
GBM-05-05
GBM-07-07
GBM-05-35
GBM-06-05
GBM-15-90
GBM-05-30
GBM-05-18
GBM-05-33
GBM-05-45
GBM-06-12
GBM-88-04
GBM-05-26
GBM-07-02
GBM-05-16
GBM-05-15
GBM-05-23
GBM-97-04
GBM-05-21
PDX
PDX-88-02_P3
PDX-88-04_P57
PDX-89-08_P7
Activated TRE
Recurrent taTRE
DNase-I-seq
Midbrain
Globus pallidus
Frontal cortex
Middle frontal gyrus
Cerebellum
Cerebellar cortex
Fetal brain
Adult brain H3K27Ac
10 kb

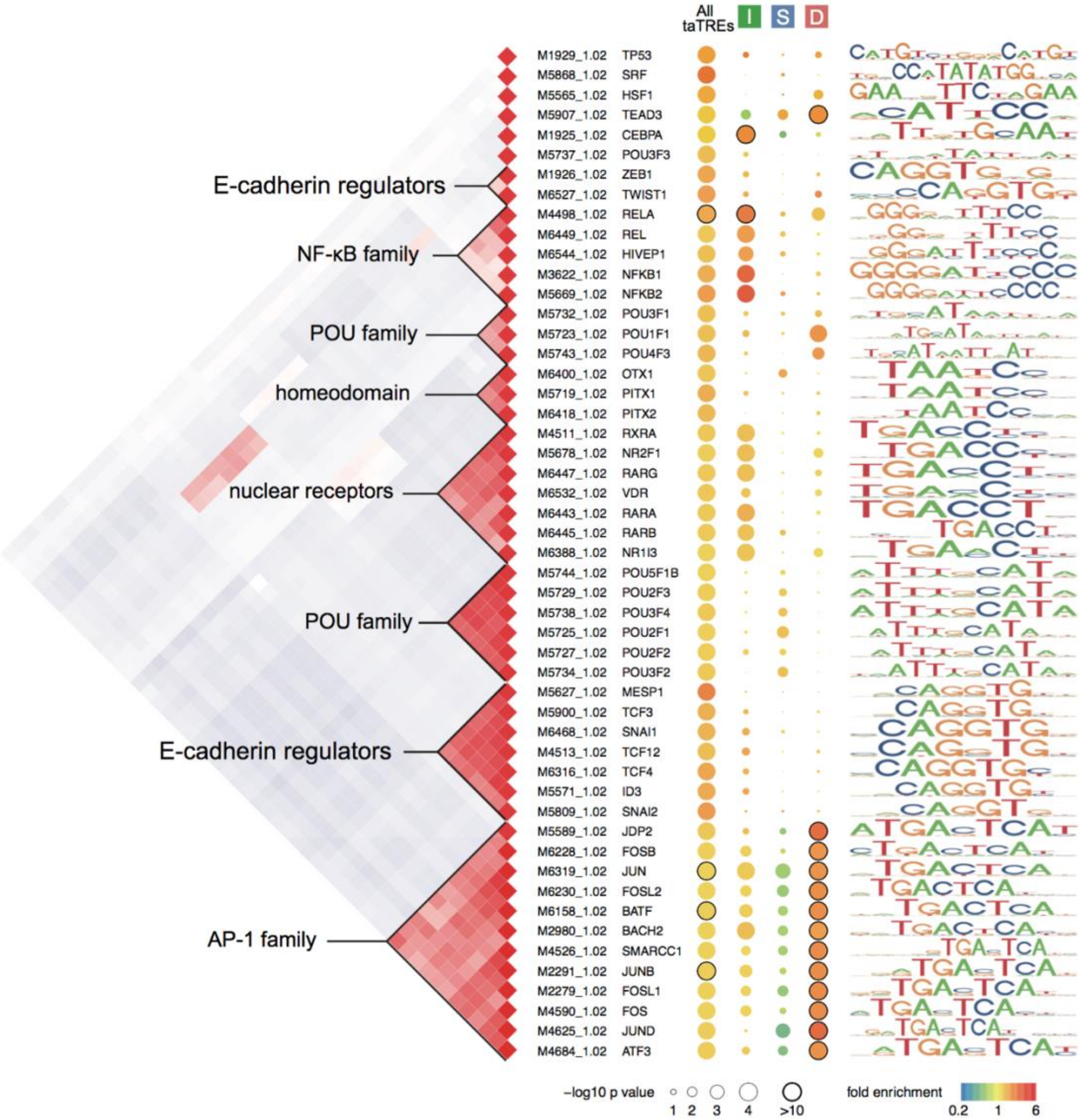
Supplementary Fig. 1. 16 EN2 locus show strong differential expression and activation of taTREs in GBM.

Browser tracks of ChRO-seq signal in primary GBM and PDX, normal astrocyte and non-malignant brain samples, DNase-I hypersensitivity and in normal adult and fetal brain tissues, and H3K27ac peaks in normal adult brain tissues near the *EN2* gene. taTREs that are activated in GBM samples are highlighted in blue. The yellow bar highlights a TRE that is highly active in GBM but not in non-malignant brain. Although it is DNase-I hypersensitive in some of adult brain tissues, it is not associated with the active transcription marker H3K27ac in any of the normal adult brain tissue.



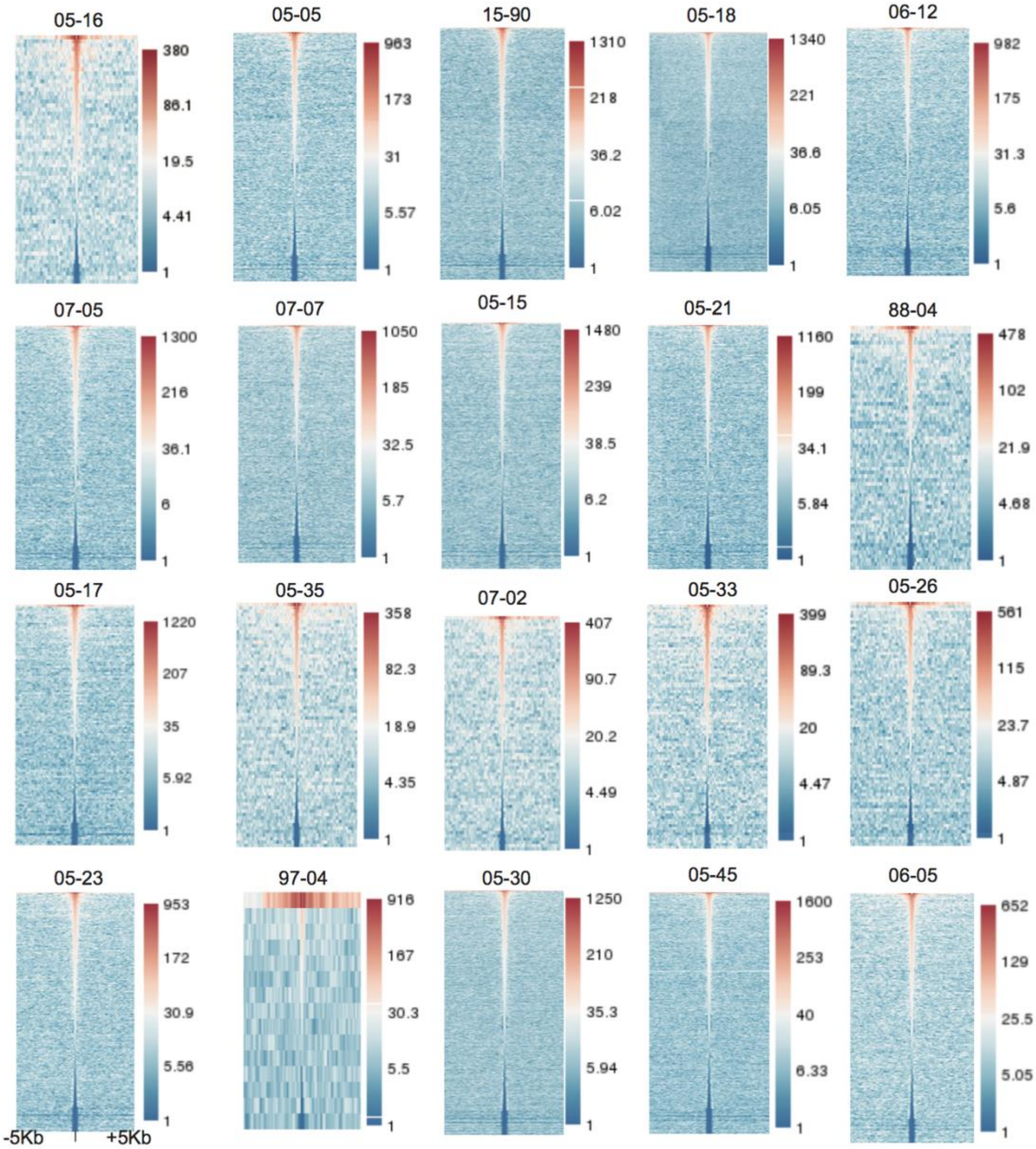
Supplementary Fig. 1. 17 Clustering of taTREs-enriched reference samples.

Clustering of reference samples enriched for taTREs based on the activation of TREs. Active TREs are marked in red; inactive ones are in white. Row dendrograms are cut down to three trees, each corresponding to the indicated transcriptional regulatory program (i.e., stem- or fetal-like, immune, and differentiated).



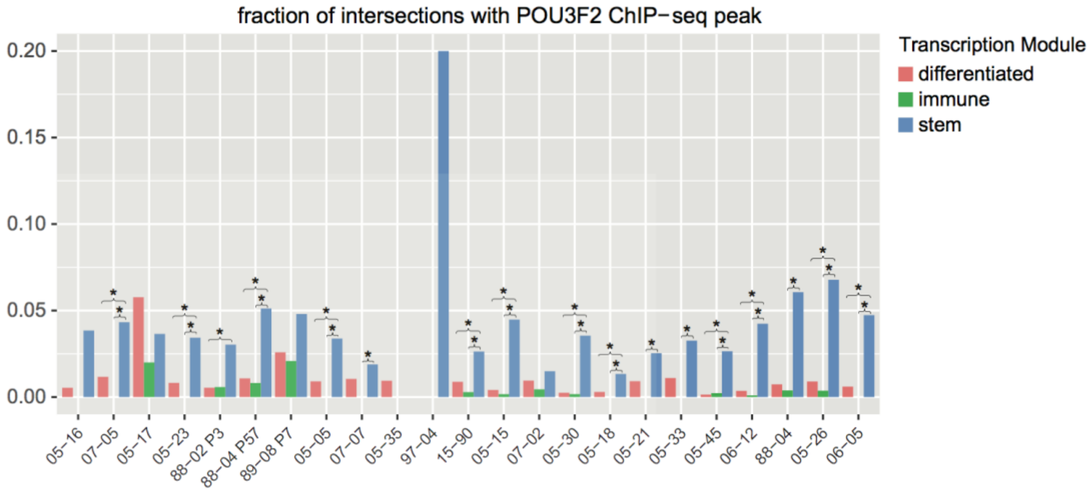
Supplementary Fig. 1. 18 Transcription factor binding motifs enriched in TREs in the indicated regulatory program compared with normal brain.

Transcription factor binding motifs enriched in TREs that are members of the immune (I), stem (S), or differentiated (D) regulatory program (top) compared with TREs active in the normal brain. Spearman's rank correlation (heatmap, left) shows the correlation in DNA sequence recognition motif. Families of transcription factor and their representative motifs are highlighted. The median *p* value across patients significantly enriched/depleted (unadjusted $p < 0.05$, two-sided Fisher's exact test) in taTREs for each motif (right) are represented by the radius of the circle and enrichment (red) or depletion (blue) are represented by the color. The number of taTREs in each test is shown in Supplementary Table 1.3.



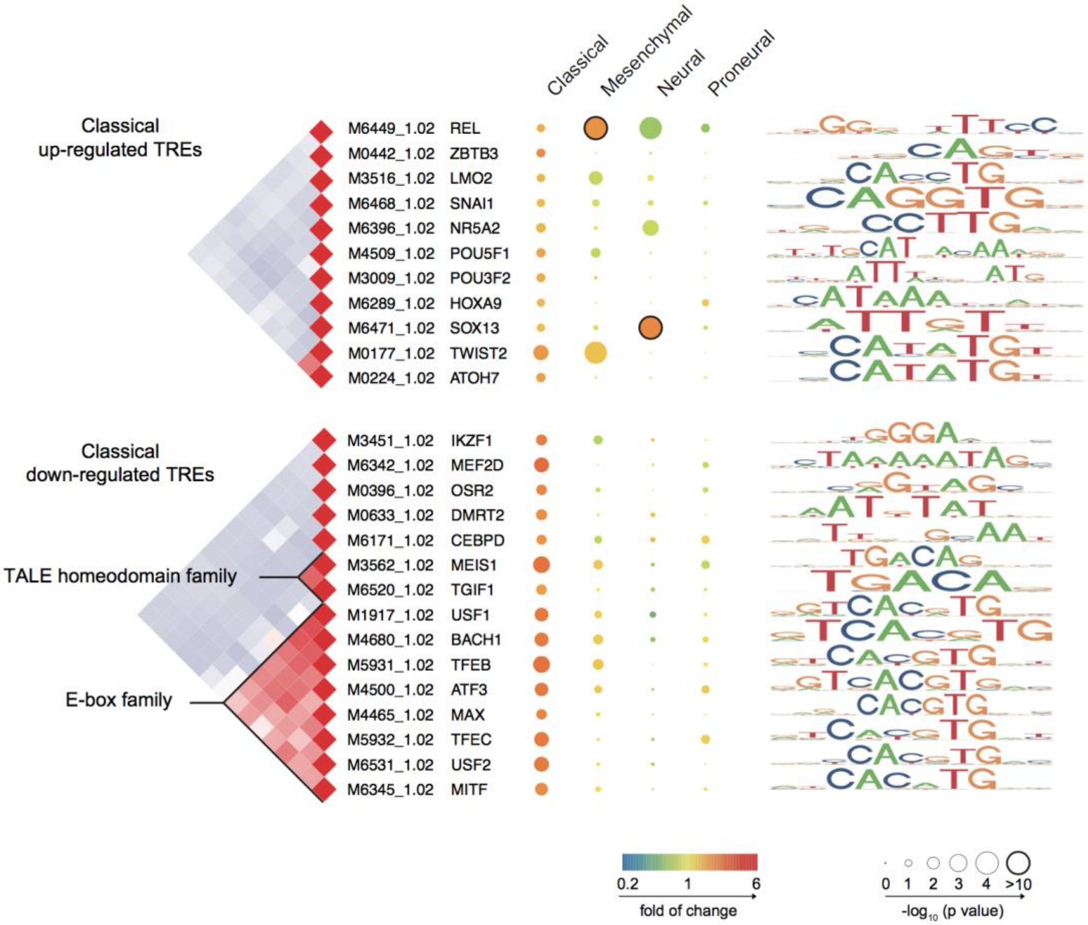
Supplementary Fig. 1. 19 taTREs show enrichment of POU3F2 binding in tumor propagating cells.

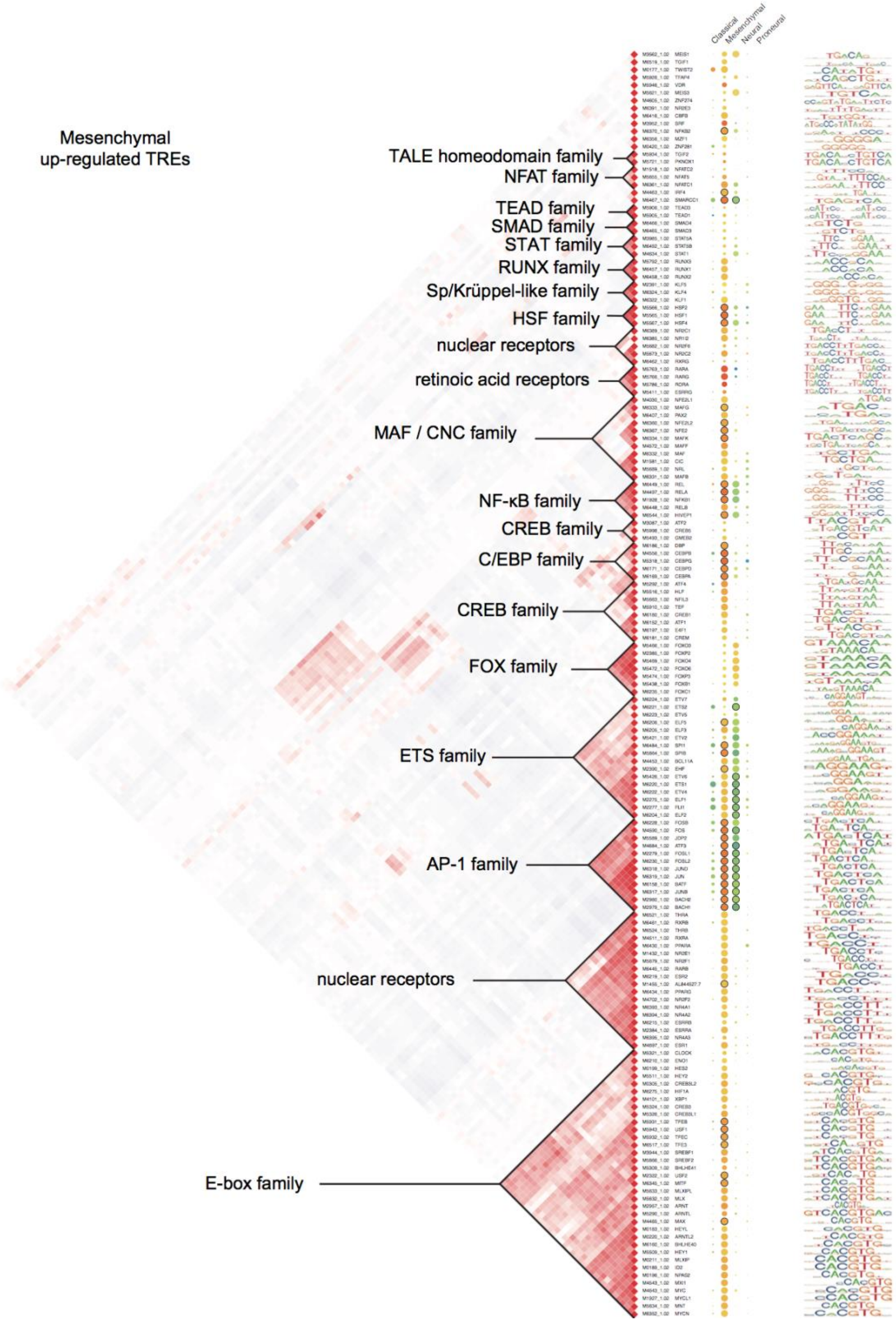
Heatmaps show ChIP-seq signals for POU3F2 in tumor propagating cells ±5kb surrounding the center of taTREs. Data was from (Suvà et al. 2014). Rows were ordered by the sum of ChIP-seq signals. Plots are made using the R pheatmap package (Kolde 2015).

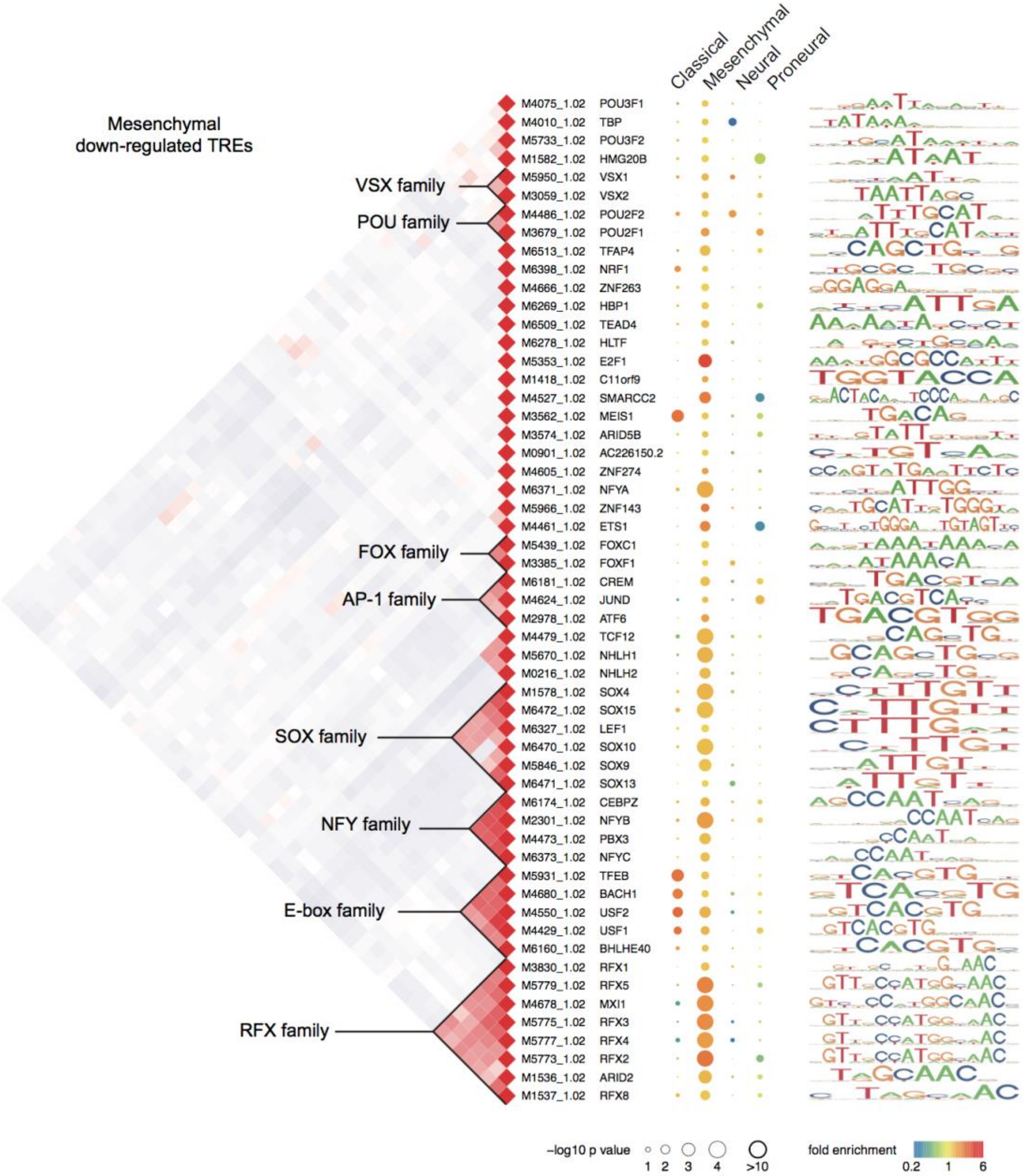


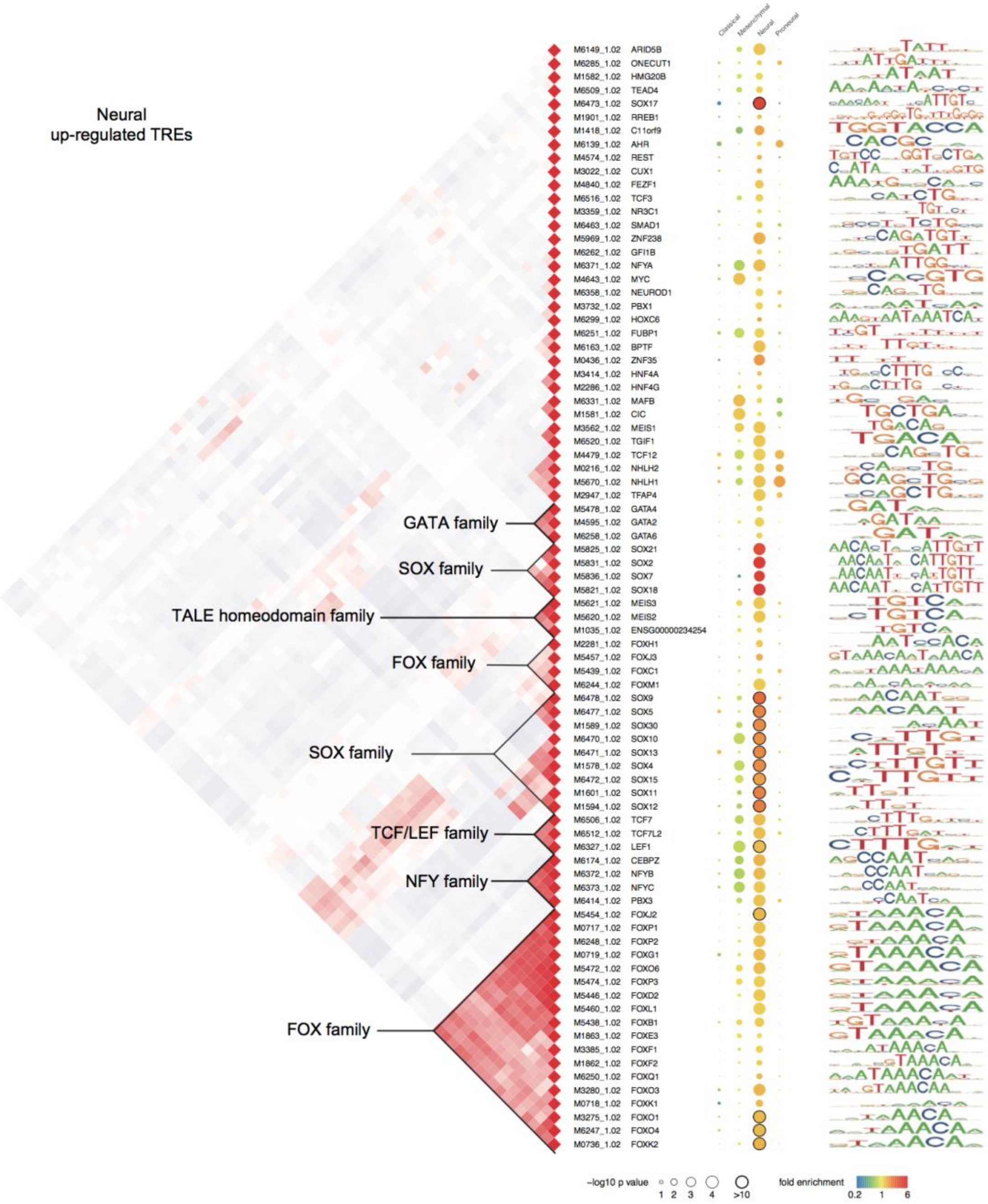
Supplementary Fig. 1. 20 Stem program taTREs enriched for POU3F2 ChIP-seq peaks.

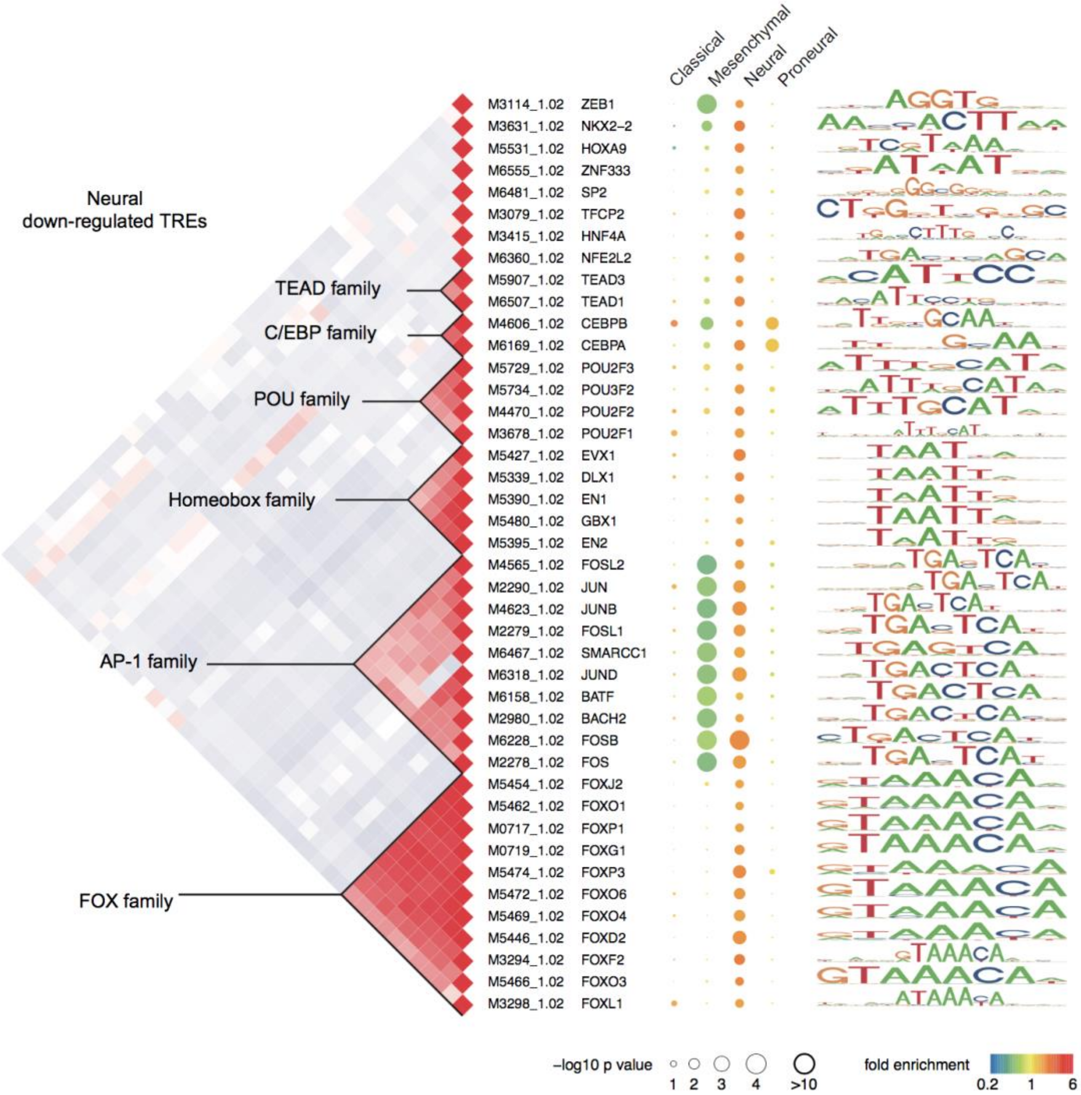
The height of bars shows the fraction of POU3F2 ChIP-seq peaks that intersect with taTRE in each of the primary GBM / PDX samples. taTREs from differentiated and stem programs are colored in red and green respectively. Primary GBM / PDX samples in which ChIP-seq peaks were enriched in stem program taTREs are marked by an asterisk (unadjusted $p < 0.05$, one-sided Fisher's exact test). Sample size for POU3F2 ChIP-seq peaks overlapped with each module: differentiated: mean=3.1, sd=1.5; stem: mean=5.8, sd=3.5; immune: mean=0.5, sd=0.5.

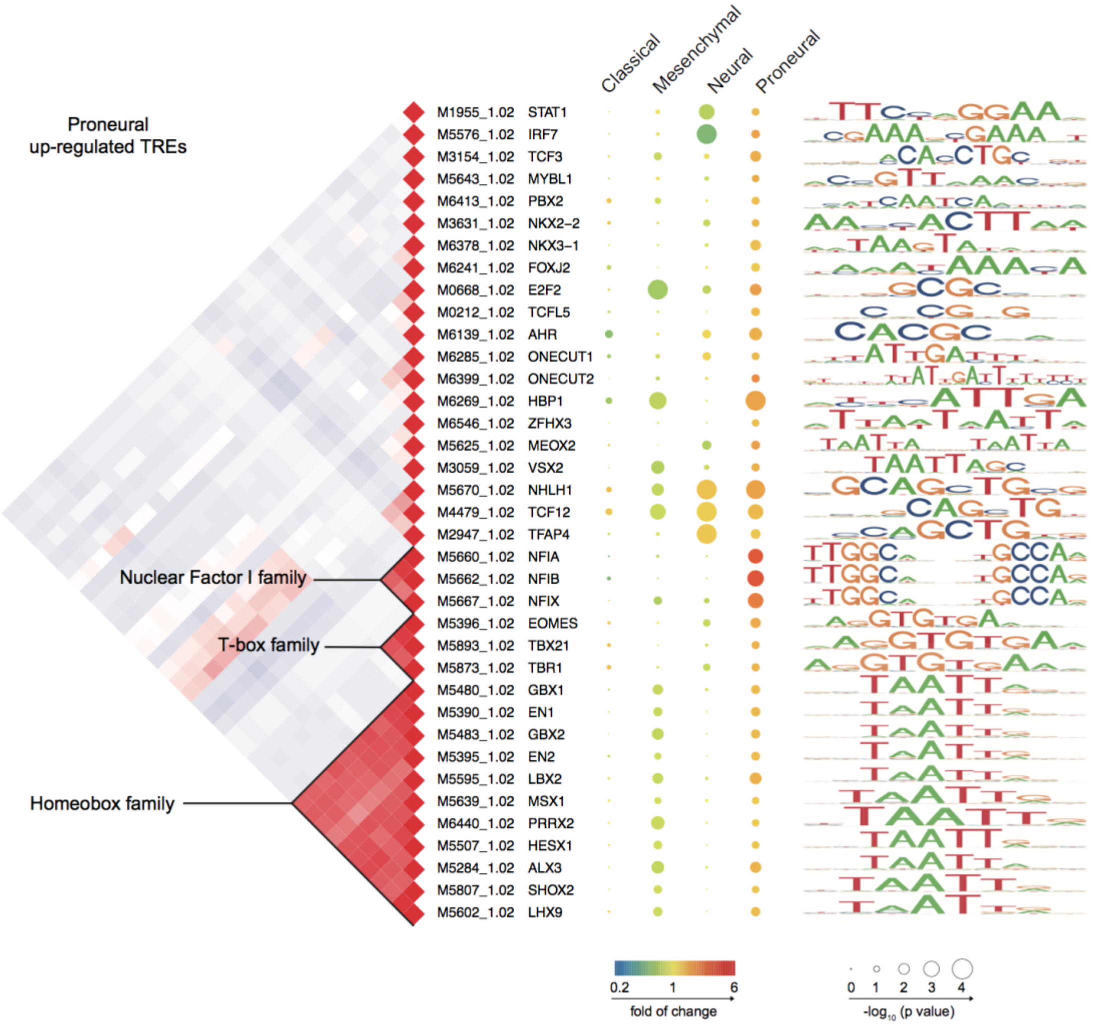

Classical
Mesenchymal
Neural
Proneural
Classical up-regulated TREs
M6449_1.02 REL
M0442_1.02 ZBTB3
M3516_1.02 LMO2
M6468_1.02 SNAI1
M6396_1.02 NR5A2
M4509_1.02 POU5F1
M3009_1.02 POU3F2
M6289_1.02 HOXA9
M6471_1.02 SOX13
M0177_1.02 TWIST2
M0224_1.02 ATOH7
Classical down-regulated TREs
TALE homeodomain family
E-box family
M3451_1.02 IKZF1
M6342_1.02 MEF2D
M0396_1.02 OSR2
M0633_1.02 DMRT2
M6171_1.02 CEBPD
M3562_1.02 MEIS1
M6520_1.02 TGIF1
M1917_1.02 USF1
M4680_1.02 BACH1
M5931_1.02 TFEB
M4500_1.02 ATF3
M4465_1.02 MAX
M5932_1.02 TFEC
M6531_1.02 USF2
M6345_1.02 MITF
0.2 1 6
fold of change
0 1 2 3 4 >10
-log10 (p value)

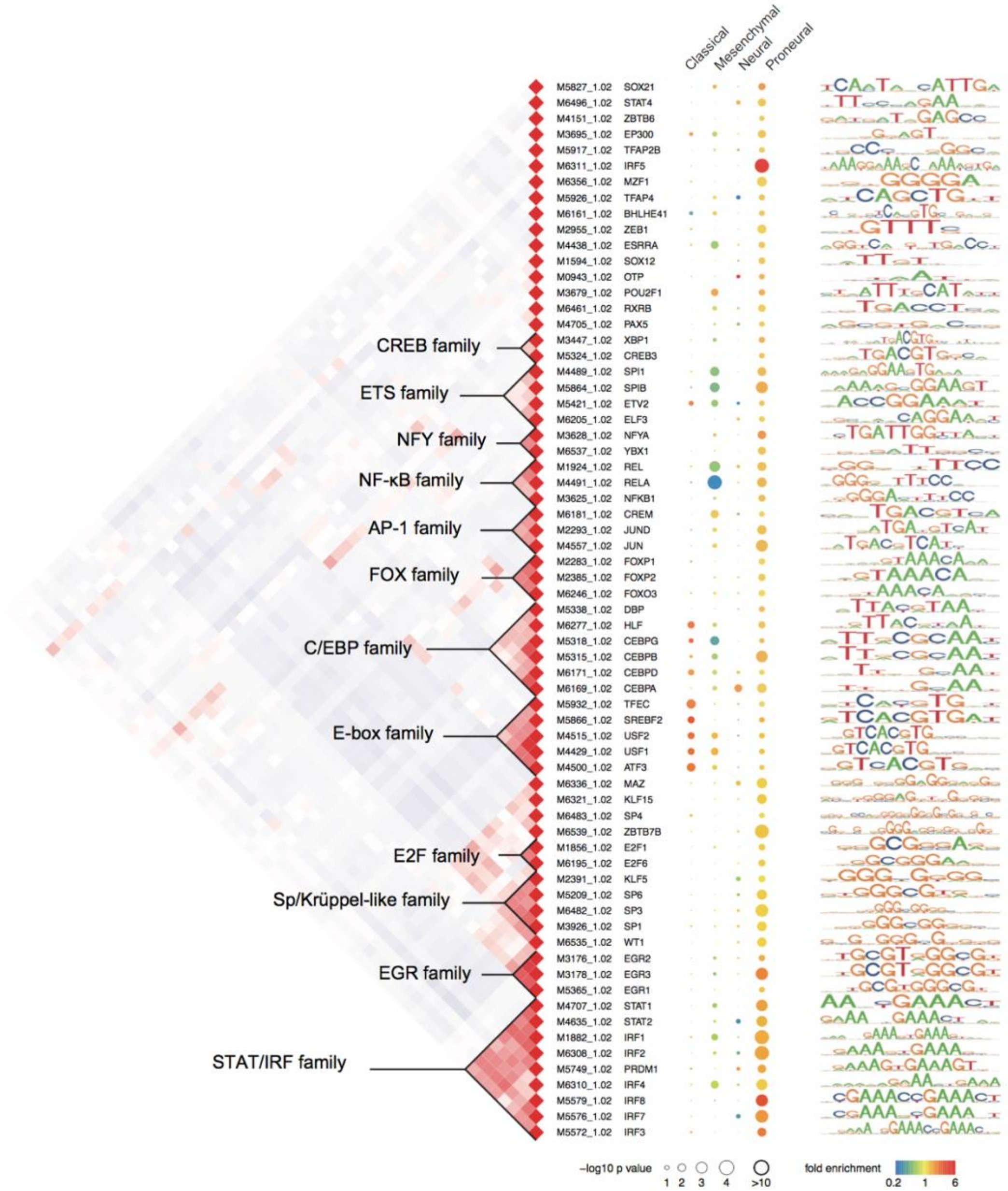

Mesenchymal
up-regulated TREs
Classical
Mesenchymal
Neural
Proneural
TALE homeodomain family
NFAT family
TEAD family
SMAD family
STAT family
RUNX family
Sp/Krüppel-like family
HSF family
nuclear receptors
retinoic acid receptors
MAF / CNC family
NF-κB family
CREB family
C/EBP family
CREB family
FOX family
ETS family
AP-1 family
nuclear receptors
E-box family


Mesenchymal down-regulated TREs
VSX family
POU family
FOX family
AP-1 family
SOX family
NFY family
E-box family
RFX family
Classical
Mesenchymal
Neural
Proneural
M4075_1.02 POU3F1
M4010_1.02 TBP
M5733_1.02 POU3F2
M1582_1.02 HMG20B
M5950_1.02 VSX1
M3059_1.02 VSX2
M4486_1.02 POU2F2
M3679_1.02 POU2F1
M6513_1.02 TFAP4
M6398_1.02 NRF1
M4666_1.02 ZNF263
M6269_1.02 HBP1
M6509_1.02 TEAD4
M6278_1.02 HLTF
M5353_1.02 E2F1
M1418_1.02 C11orf9
M4527_1.02 SMARCC2
M3562_1.02 MEIS1
M3574_1.02 ARID5B
M0901_1.02 AC226150.2
M4605_1.02 ZNF274
M6371_1.02 NFYA
M5966_1.02 ZNF143
M4461_1.02 ETS1
M5439_1.02 FOXC1
M3385_1.02 FOXF1
M6181_1.02 CREM
M4624_1.02 JUND
M2978_1.02 ATF6
M4479_1.02 TCF12
M5670_1.02 NHLH1
M0216_1.02 NHLH2
M1578_1.02 SOX4
M6472_1.02 SOX15
M6327_1.02 LEF1
M6470_1.02 SOX10
M5846_1.02 SOX9
M6471_1.02 SOX13
M6174_1.02 CEBPZ
M2301_1.02 NFYB
M4473_1.02 PBX3
M6373_1.02 NFYC
M5931_1.02 TFEB
M4680_1.02 BACH1
M4550_1.02 USF2
M4429_1.02 USF1
M6160_1.02 BHLHE40
M3830_1.02 RFX1
M5779_1.02 RFX5
M4678_1.02 MXI1
M5775_1.02 RFX3
M5777_1.02 RFX4
M5773_1.02 RFX2
M1536_1.02 ARID2
M1537_1.02 RFX8
−log10 p value
1 2 3 4 >10
fold enrichment
0.2 1 6


Neural
up-regulated TREs
GATA family
SOX family
TALE homeodomain family
FOX family
SOX family
TCF/LEF family
NFY family
FOX family
Classical
Mesenchymal
Neural
Proneural
M6149_1.02 ARID5B
M6285_1.02 ONECUT1
M1582_1.02 HMG20B
M6509_1.02 TEAD4
M6473_1.02 SOX17
M1901_1.02 RREB1
M1418_1.02 C11orf9
M6139_1.02 AHR
M4574_1.02 REST
M3022_1.02 CUX1
M4840_1.02 FEZF1
M6516_1.02 TCF3
M3359_1.02 NR3C1
M6463_1.02 SMAD1
M5969_1.02 ZNF238
M6262_1.02 GFI1B
M6371_1.02 NFYA
M4643_1.02 MYC
M6358_1.02 NEUROD1
M3732_1.02 PBX1
M6299_1.02 HOXC6
M6251_1.02 FUBP1
M6163_1.02 BPTF
M0436_1.02 ZNF35
M3414_1.02 HNF4A
M2286_1.02 HNF4G
M6331_1.02 MAFB
M1581_1.02 CIC
M3562_1.02 MEIS1
M6520_1.02 TGIF1
M4479_1.02 TCF12
M0216_1.02 NHLH2
M5670_1.02 NHLH1
M2947_1.02 TFAP4
M5478_1.02 GATA4
M4595_1.02 GATA2
M6258_1.02 GATA6
M5825_1.02 SOX21
M5831_1.02 SOX2
M5836_1.02 SOX7
M5821_1.02 SOX18
M5621_1.02 MEIS3
M5620_1.02 MEIS2
M1035_1.02 ENSG00000234254
M2281_1.02 FOXH1
M5457_1.02 FOXJ3
M5439_1.02 FOXC1
M6244_1.02 FOXM1
M6478_1.02 SOX9
M6477_1.02 SOX5
M1589_1.02 SOX30
M6470_1.02 SOX10
M6471_1.02 SOX13
M1578_1.02 SOX4
M6472_1.02 SOX15
M1601_1.02 SOX11
M1594_1.02 SOX12
M6506_1.02 TCF7
M6512_1.02 TCF7L2
M6327_1.02 LEF1
M6174_1.02 CEBPZ
M6372_1.02 NFYB
M6373_1.02 NFYC
M6414_1.02 PBX3
M5454_1.02 FOXJ2
M0717_1.02 FOXP1
M6248_1.02 FOXP2
M0719_1.02 FOXG1
M5472_1.02 FOXO6
M5474_1.02 FOXP3
M5446_1.02 FOXD2
M5460_1.02 FOXL1
M5438_1.02 FOXB1
M1863_1.02 FOXE3
M3385_1.02 FOXF1
M1862_1.02 FOXF2
M6250_1.02 FOXQ1
M3280_1.02 FOXO3
M0718_1.02 FOXK1
M3275_1.02 FOXO1
M6247_1.02 FOXO4
M0736_1.02 FOXK2
–log10 p value
1 2 3 4 >10
fold enrichment
0.2 1 6

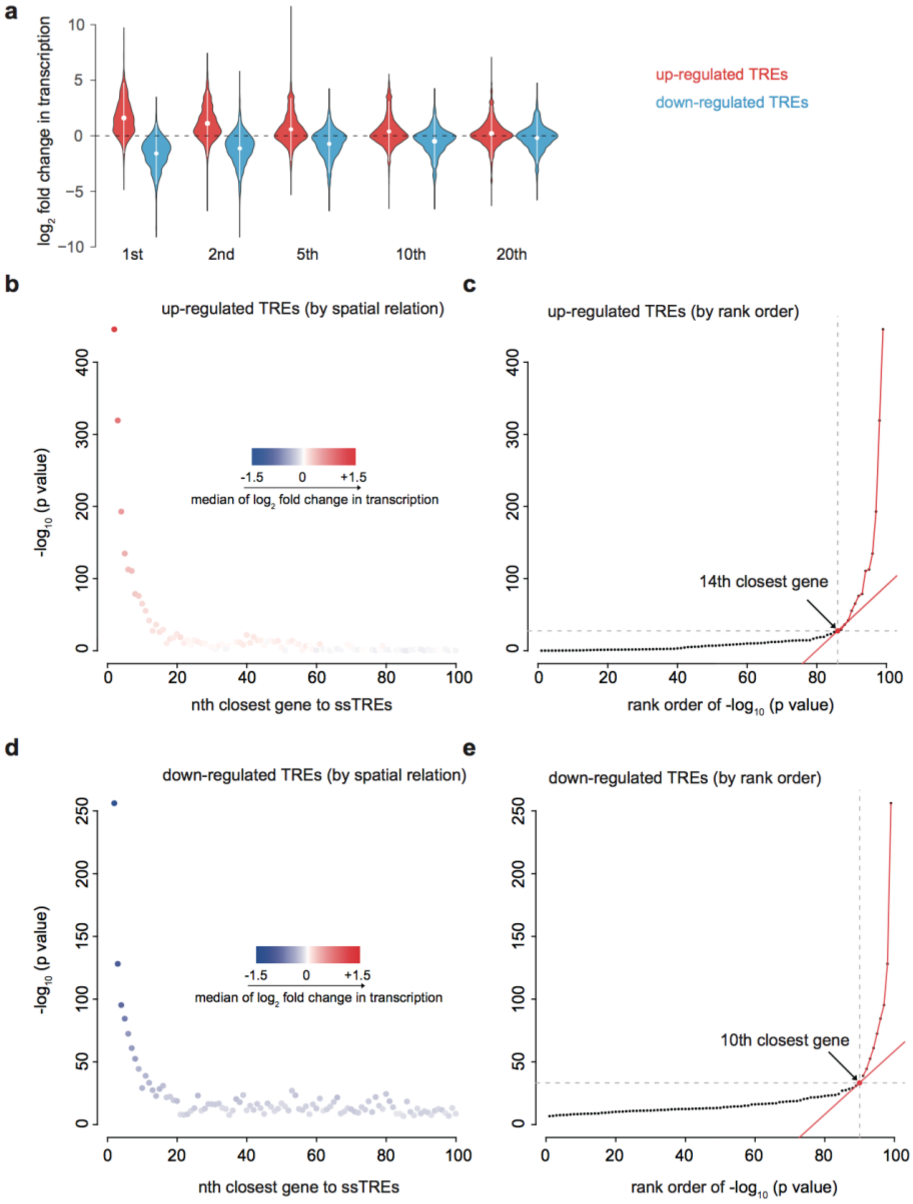

Classical
Mesenchymal
Neural
Proneural
Neural
down-regulated TREs
TEAD family
C/EBP family
POU family
Homeobox family
AP-1 family
FOX family
M3114_1.02 ZEB1
M3631_1.02 NKX2-2
M5531_1.02 HOXA9
M6555_1.02 ZNF333
M6481_1.02 SP2
M3079_1.02 TFCP2
M3415_1.02 HNF4A
M6360_1.02 NFE2L2
M5907_1.02 TEAD3
M6507_1.02 TEAD1
M4606_1.02 CEBPB
M6169_1.02 CEBPA
M5729_1.02 POU2F3
M5734_1.02 POU3F2
M4470_1.02 POU2F2
M3678_1.02 POU2F1
M5427_1.02 EVX1
M5339_1.02 DLX1
M5390_1.02 EN1
M5480_1.02 GBX1
M5395_1.02 EN2
M4565_1.02 FOSL2
M2290_1.02 JUN
M4623_1.02 JUNB
M2279_1.02 FOSL1
M6467_1.02 SMARCC1
M6318_1.02 JUND
M6158_1.02 BATF
M2980_1.02 BACH2
M6228_1.02 FOSB
M2278_1.02 FOS
M5454_1.02 FOXJ2
M5462_1.02 FOXO1
M0717_1.02 FOXP1
M0719_1.02 FOXG1
M5474_1.02 FOXP3
M5472_1.02 FOXO6
M5469_1.02 FOXO4
M5446_1.02 FOXD2
M3294_1.02 FOXF2
M5466_1.02 FOXO3
M3298_1.02 FOXL1
−log10 p value
1 2 3 4 >10
fold enrichment
0.2 1 6


Classical
Mesenchymal
Neural
Proneural
Proneural
up-regulated TREs
Nuclear Factor I family
T-box family
Homeobox family
M1955_1.02 STAT1
M5576_1.02 IRF7
M3154_1.02 TCF3
M5643_1.02 MYBL1
M6413_1.02 PBX2
M3631_1.02 NKX2–2
M6378_1.02 NKX3–1
M6241_1.02 FOXJ2
M0668_1.02 E2F2
M0212_1.02 TCFL5
M6139_1.02 AHR
M6285_1.02 ONECUT1
M6399_1.02 ONECUT2
M6269_1.02 HBP1
M6546_1.02 ZFHX3
M5625_1.02 MEOX2
M3059_1.02 VSX2
M5670_1.02 NHLH1
M4479_1.02 TCF12
M2947_1.02 TFAP4
M5660_1.02 NFIA
M5662_1.02 NFIB
M5667_1.02 NFIX
M5396_1.02 EOMES
M5893_1.02 TBX21
M5873_1.02 TBR1
M5480_1.02 GBX1
M5390_1.02 EN1
M5483_1.02 GBX2
M5395_1.02 EN2
M5595_1.02 LBX2
M5639_1.02 MSX1
M6440_1.02 PRRX2
M5507_1.02 HESX1
M5284_1.02 ALX3
M5807_1.02 SHOX2
M5602_1.02 LHX9
0.2
1
6
fold of change
0
1
2
3
4
-log10 (p value)



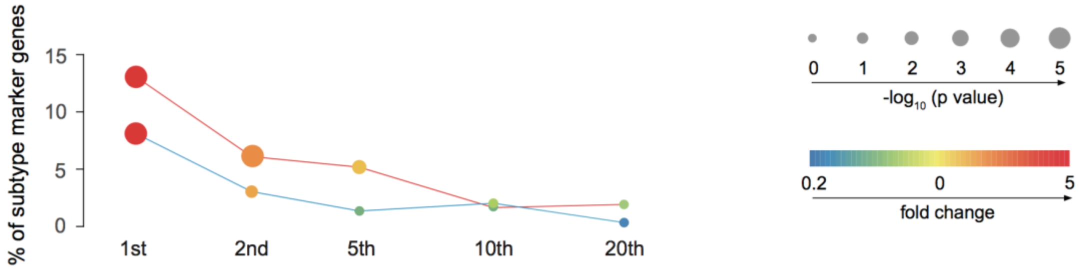
Supplementary Fig. 1. 21 Transcription factor binding motifs enriched in TREs up-regulated or down-regulated in each known molecular subtype.

Transcription factor binding motifs enriched in TREs that were up- or down-regulated in the indicated subtype. The Spearman's rank correlation heatmap (left) shows the correlation in DNA binding sites matching each motif. Families of transcription factors and their representative motifs are highlighted. Right: Enrichment of transcription factor binding motifs in TRE with biased transcription in the indicated subtype. The unadjusted *p* values (two-sided Fisher's exact test) of motifs are represented by the radius of the circle, and enrichment (red) or depletion (blue) are represented by the rainbow color scale. The number of subtype-biased TREs in each group is shown in Supplementary Table 1.4.



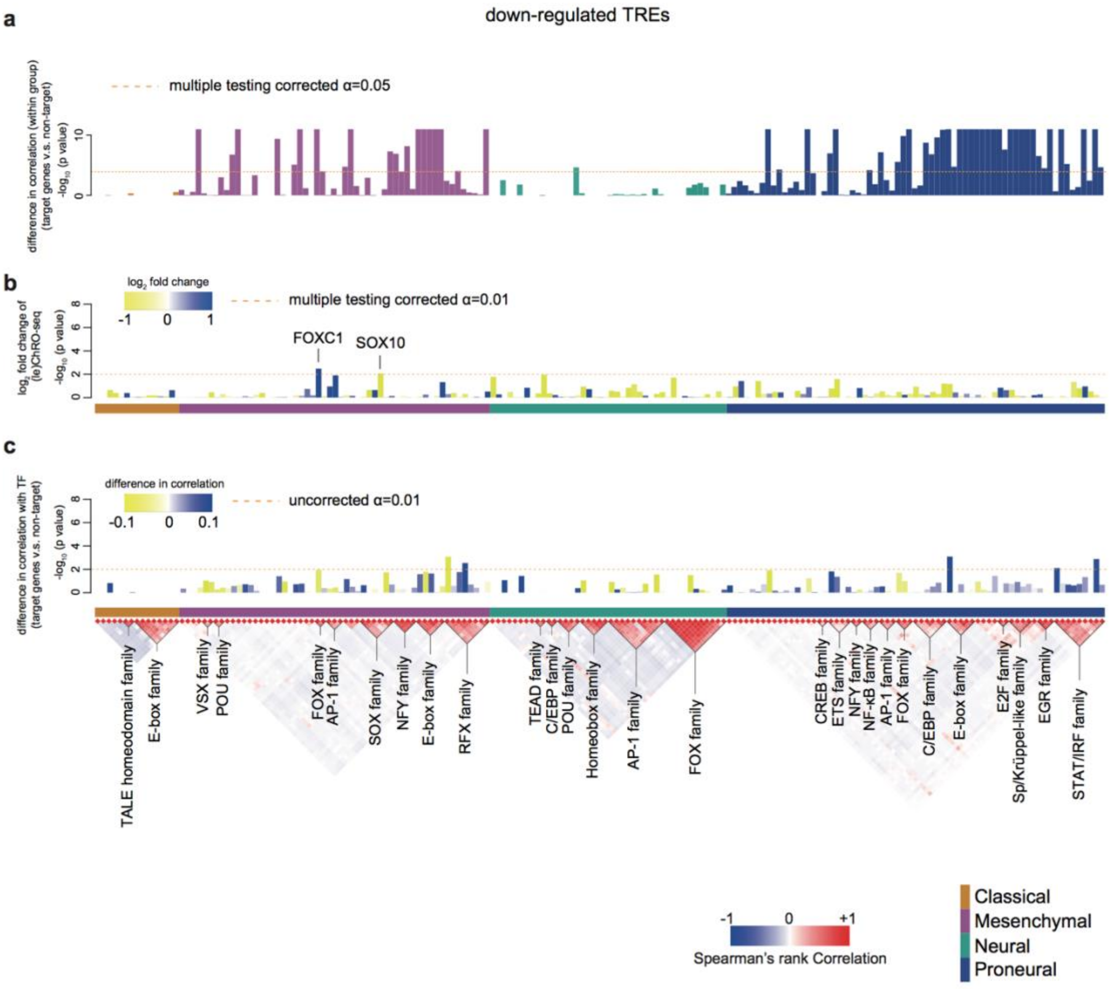
Supplementary Fig. 1. 22 Subtype-biased TREs correlate with the transcription of nearby genes.

(**a**) Violin plots show the distribution of $\log_2$ fold change in the transcription of n th closest genes to TREs that were up (red, N=4,960) or down (blue, N=1,815) -regulated in any subtype. White dots represent the means, while the bars represent standard deviations. (**b** and **d**) Scatter plots show the -$\log_{10}$ two-sided t-test *p* value testing the null hypothesis that the $\log_2$ fold change is equal to zero as a function of nth closest gene to the subtype-biased TRE. Separate plots are shown for up (**b**, N=4,960) or down (**d**, N=1,815) -regulated gene/ TRE pairs. Median $\log_2$ fold change in transcription is represented using red and blue color scale. (**c** and **e**) The rank-ordered version of (**c**) and (**d**) show outliers in change of transcription determined at the inflection point (marked by red).



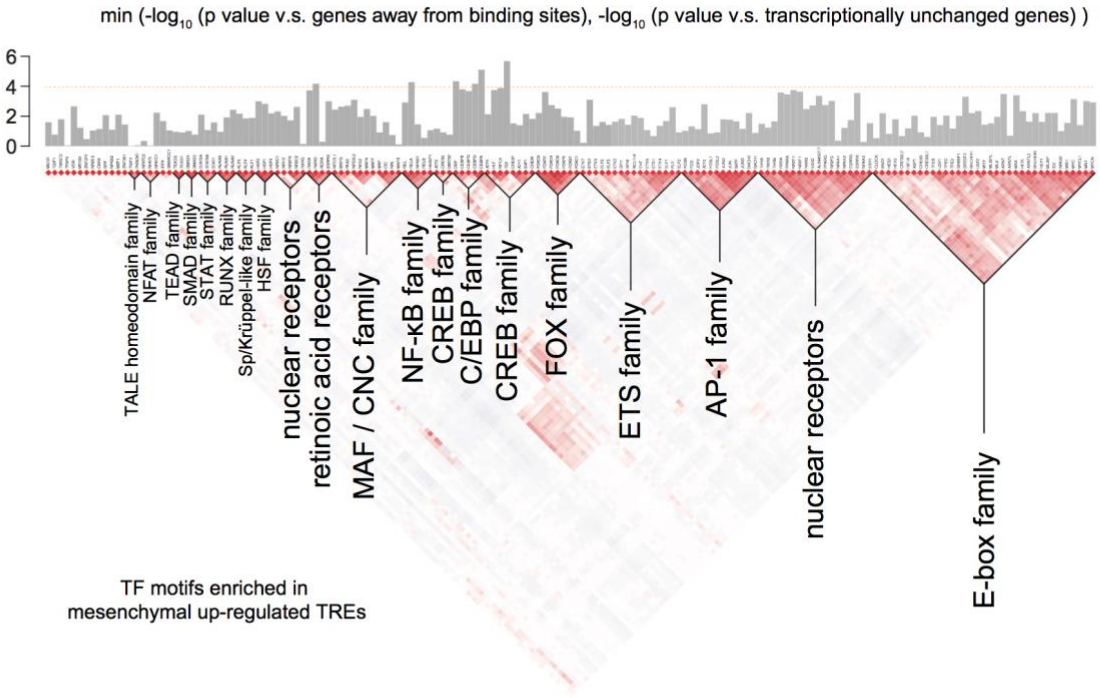
Supplementary Fig. 1. 23 Subtype-biased TREs are near a large proportion of subtype specific genes.

Line chart show the percentage of subtype marker genes (Y-axis) positioned *n* genes from the closest subtype-biased TREs. Separate lines are shown for up (red, N=4,960) or down (blue, N=1,815) -regulated gene/ TRE pairs. The enrichment (red) or depletion (blue) over the expected number of genes is represented by the color, and the unadjusted *p* values of two-sided Fisher's exact test for enrichment is represented by the radius of the circle.



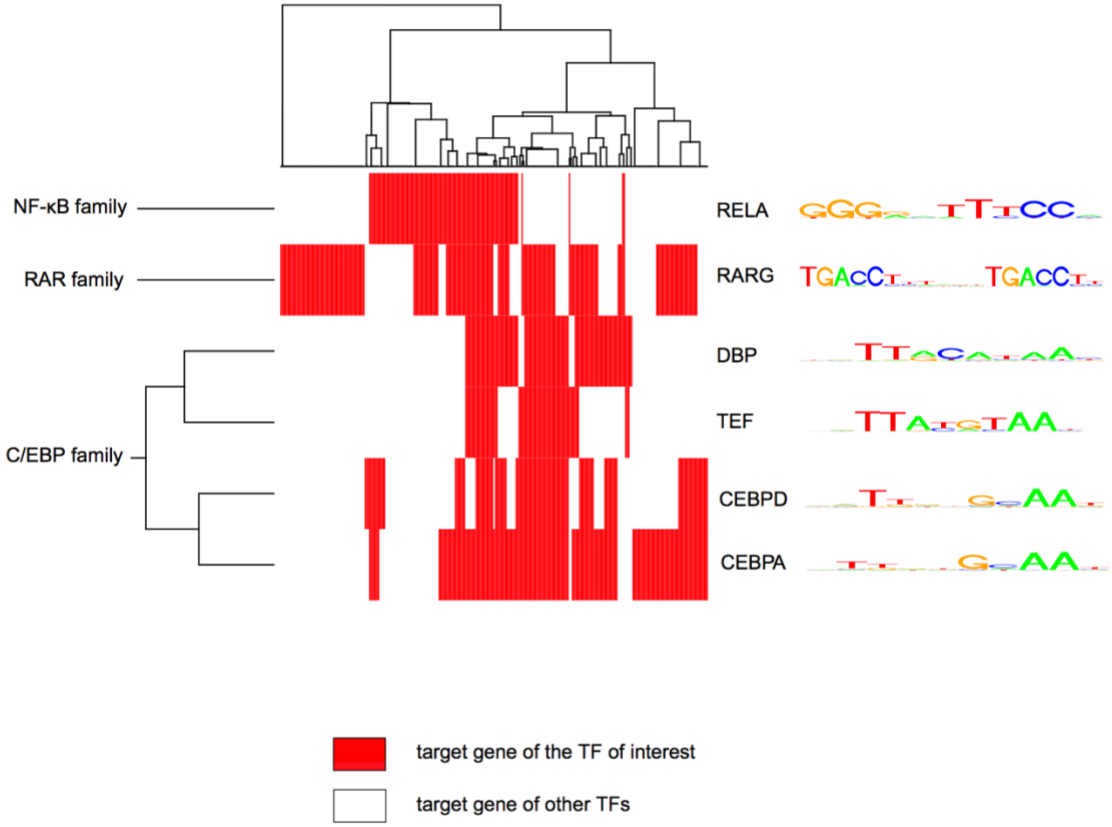
Supplementary Fig. 1. 24 Barplots show the relationship between transcription factors enriched over TREs down-regulated in each subtype and their putative target genes.

(**a**) Barplots show the -$\log_{10}$ Wilcoxon rank sum of *p* value of having higher correlation among 174 TCGA patients between target genes for each transcription factor compared with a control set. Barplots are colored by subtype in which they were found to be enriched (unadjusted $p <$ 0.05, two-sided Fisher's exact test). (**b**) Barplot shows the FDR corrected -$\log_{10}$ *p* value (DESeq2, Wald test, n= 2 [classical] or 3 [other subtypes]) representing changes in Pol II abundance detected by (le)ChRO-seq on the gene encoding the indicated transcription factor. The level of upregulation (blue) and downregulation (yellow) in the subtype indicated by the colored boxes (below the barplot) is shown by the color scale. The horizontal color bar below the barplot indicates the corresponding subtype in which the motif shows enrichment in the downregulated TREs. The dashed line shows the the FDR corrected α at 0.01. (**c**) Barplot shows the -$\log_{10}$ two-sided Wilcoxon rank sum test *p* value denoting differences in the distribution of correlations between the mRNA encoding the indicated transcription factor and either target or non-target control genes. The blue/ yellow color scale represents the median difference in correlation between target and non-target genes over 174 mRNA-seq samples. The dashed line shows the uncorrected α at 0.01



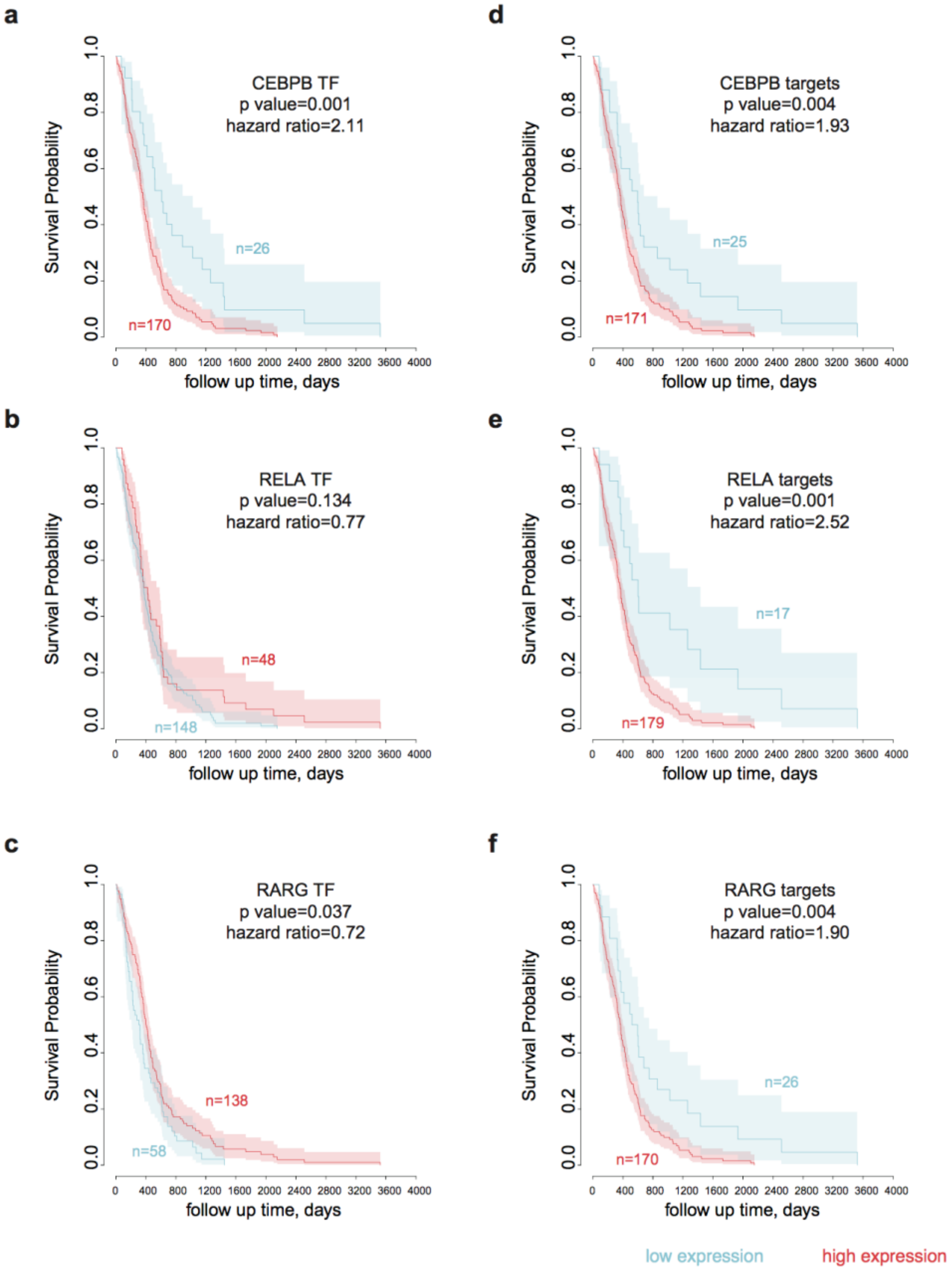
Supplementary Fig. 1. 25 Barplots show transcription factor binding motifs controlling survival-related genes in mesenchymal GBMs.

The minimum of the two -log$_{10}$ *p* values on the x-axis and y-axis of Fig 1.7a (two-sided Wilcoxon rank sum test) are plotted by the order of motifs cluster. In total, 196 TCGA patients with microarray data and survival information were used to calculate the hazard ratio. The dotted red line represents the Bonferroni adjusted α value at 0.05.



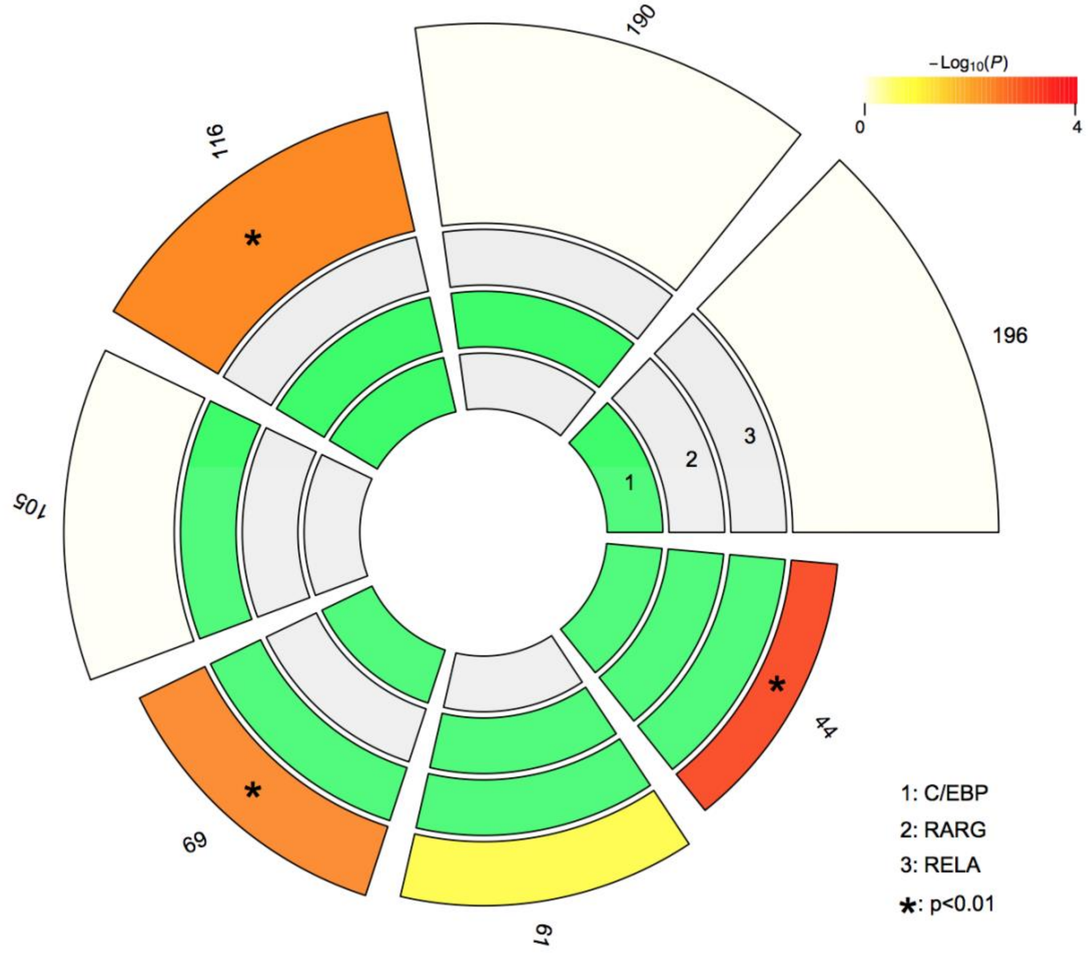
Supplementary Fig. 1. 26 Heatmap shows the clustering of target genes of six transcription factors with significant survival association.

Hierarchical agglomerative clustering groups target genes of one or more transcription factor. Red indicates the target gene belongs to the putative targets of the corresponding transcription factor and white indicates otherwise.



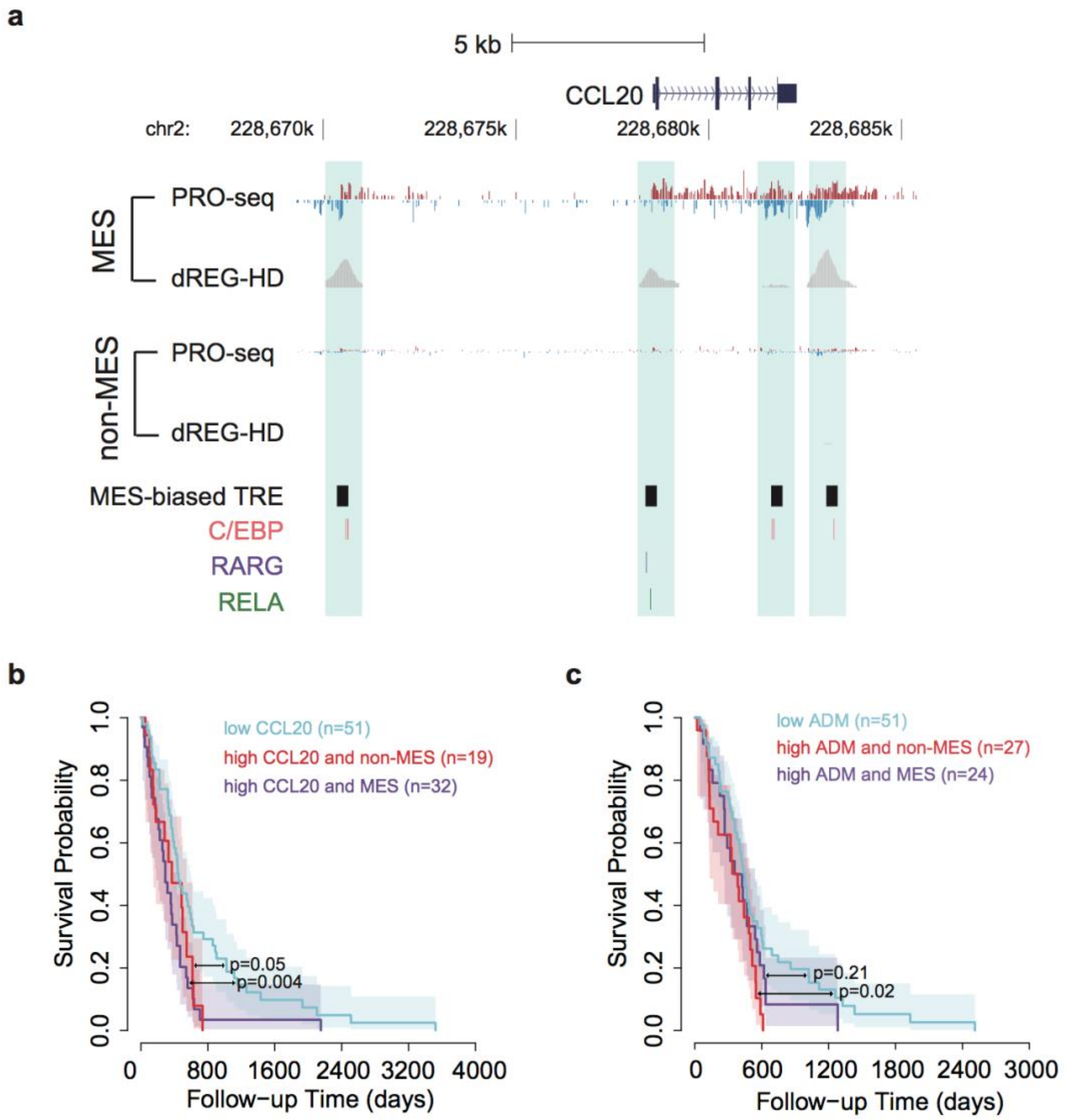
Supplementary Fig. 1. 27 Kaplan–Meier plots show the difference in survival between patients with different expression levels of transcription factors (a-c) and of their corresponding target genes (d-f).

P values and hazard ratios were calculated by comparing patients of higher expression level (red) with those of lower expression level (blue) across 196 patients. The mean expression level was used to represent target genes of each transcription factor. The optimum cutoff of mean expression level was determined by minimizing the *p* values (two-sided Chi-squared test) between survival time. Shaded regions mark the 95% confidence interval of each group.



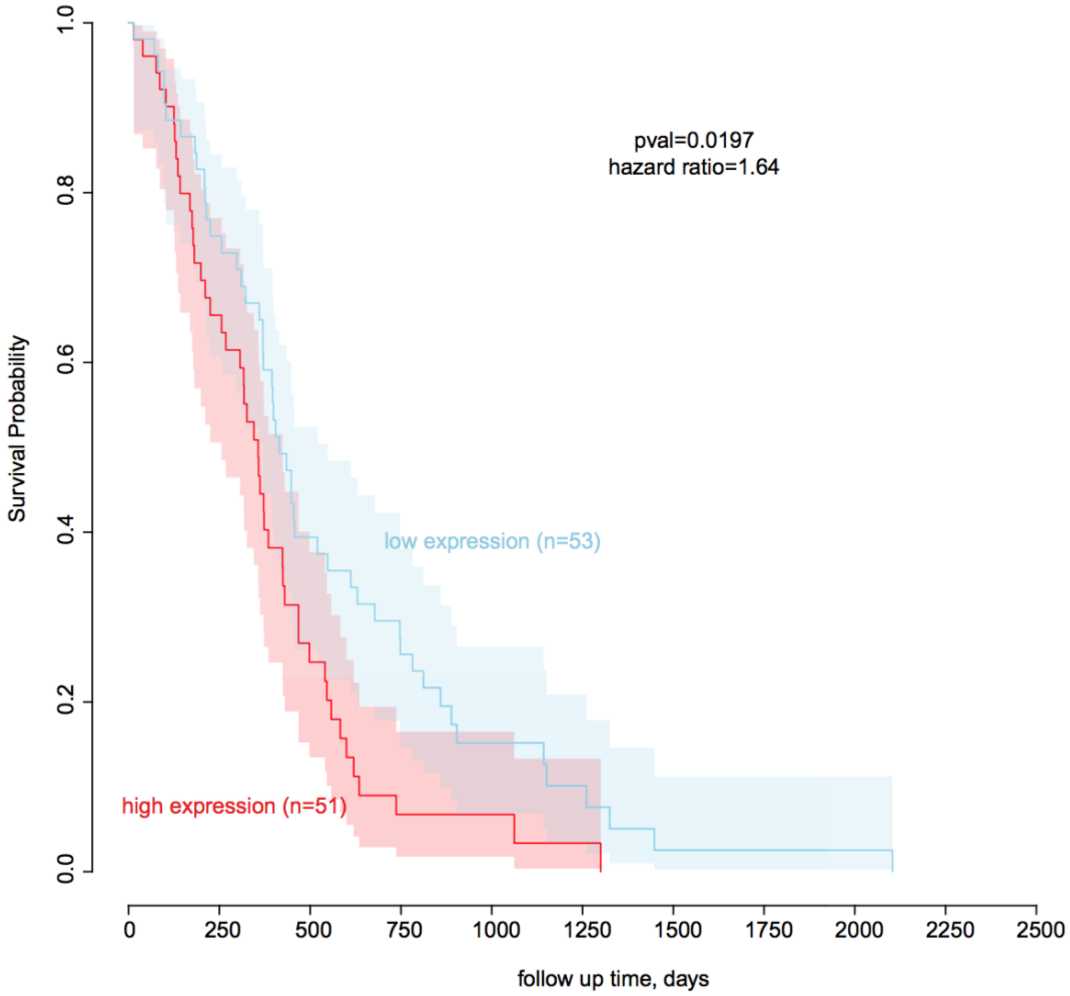
Supplementary Fig. 1. 28 Concentric circles visualize the enrichment of overlapping between target genes of C/EBP, RARG, and NF-κB/RELA.

The first three inner circles indicate the combination of transcription factors (C/EBP, RARG, and NF-κB/RELA) regulating each target gene. The outer circle is filled by a color scale representing the -$\log_{10}$ of *p* value (one-sided super exact test) of the overlap compared with random assignment among 362 genes in proximity to mesenchymal-biased TREs and up-regulated in mesenchymal GBM subtype. In total, 289 genes from three transcription factors were involved in the test. The exact number of each combination is shown on the outermost sector. Statistically significant overlap (one-sided super exact test, unadjusted $p < 0.01$) is marked by an asterisk.



Supplementary Fig. 1. 29 The Browser track of CCL20 and Kaplan–Meier plots of CCL20 and ADM.

(**a**) Browser track of the locus encoding the *CCL20* gene shows the average of RPM normalized (le)ChRO-seq signals and dREG-HD scores in mesenchymal (n= 3) and non-mesenchymal (n= 8) GBMs. Mesenchymal-biased TREs are highlighted in blue. Positions of MES-biased TRE and motifs of C/EBP, RARG, and NF-κB/RELA transcription factors are shown on the bottom. (**b** and **c**) Kaplan–Meier plots show survival rate for patients with 1) lower quartile CCL20 (b) or ADM (c) expression level (light blue), 2) upper quartile expression level of tumors in the non-mesenchymal subtype (red), and 3) upper quartile gene expression level for tumors in the mesenchymal subtype (purple). P values were calculated using a two-sided Chi-squared test. Shaded regions mark the 95% confidence interval of each group.



Supplementary Fig. 1. 30 Kaplan–Meier plot shows survival rate of IDH wild-type patients.

Kaplan–Meier plot shows overall survival between 104 IDH1 wild-type patients with high and low average expression level of 26 shared target genes. The cutoff was determined based on the minimum *p* value in the difference between survival time using a two-sided Chi-squared test. Shaded regions mark the 95% confidence interval.

1.8 Supplementary Tables

Supplementary Table 1. 1 Technical information for all samples used in the experiment.

	Run-on method	Sequencing Depth (number of total mappable reads)	# dREG sites	# dREG-HD sites (0.1 FDR)	# dREG-HD sites (0.16 FDR)	Unique molecular index barcode
K562	PRO-seq	374946808	54933	24434	43826	
GM12878	GRO-seq	105936649	59300	31026	47280	
MCF-7	PRO-seq	134344736	45112	24044	37782	
Human Naive T cell	PRO-seq	96447847	25513	22138	32186	
Nonmalignant brain-all	ChRO-seq / leChRO-seq	26618420	22005	26684	20359	
UMU94-13_Chr (normal_brain_ATCACG)	ChRO-seq	19012439				
UMU94-13_leChr (NB_Deep)	leChRO-seq	7605981				X
Human astrocytes	leChRO-seq	13124347	18851	17276	23682	
GBM-05-16	leChRO-seq	12026819	21490	25259	19950	X
GBM-07-05	leChRO-seq	33477124	32647	40065	29420	X
GBM-05-17	leChRO-seq	37121068	33150	40099	28699	X
GBM-05-23	leChRO-seq	38750059	37637	45178	32574	X
GBM-05-05	leChRO-seq	31814017	37146	44211	32874	X
GBM-07-07	leChRO-seq	53694384	44676	52255	37165	X
GBM-05-35	leChRO-seq	18865872	26219	31269	24524	X

GBM-97-04	leChRO-seq	10593102	16014	17872	14968	X
GBM-15-90 (all)	ChRO-seq	150711728	67812	53197	80953	
GBM-15-90 Core	ChRO-seq	38115236				
GBM-15-90 Corpus	ChRO-seq	38896761				
GBM-15-90 Deep	ChRO-seq	36289294				
GBM-15-90 Cortex	ChRO-seq	37410437				
GBM-05-15	leChRO-seq	21731177	36089	44682	33388	X
GBM-07-02	leChRO-seq	10303802	20144	24138	18988	X
GBM-05-30	leChRO-seq	26041476	48728	59064	44908	X
GBM-05-18	leChRO-seq	34744863	54982	65604	49345	X
GBM-05-21	leChRO-seq	27920964	37450	44699	33950	X
GBM-05-33	leChRO-seq	25874121	24782	30388	23024	X
GBM-05-45	leChRO-seq	21405989	40558	48592	37792	X
GBM-06-12	leChRO-seq	19037300	32594	40524	30744	X
GBM-88-04 (all)	ChRO-seq / leChRO-seq	22477998	20981	26621	19551	X
GBM-88-04_leChr1 (ROS1_RNase)	leChRO-seq	11982314				X
GBM-88-04_Chr (ROS1_primary_d2e7)	ChRO-seq	4236986				
GBM-88-04_leChr2	leChRO-	6258698				

(GBM_primary)	seq					
GBM-05-26	leChRO-seq	11581790	22825	27083	21509	X
GBM-06-05	leChRO-seq	32896583	38268	46223	35007	X
PDX-88-02_P3	leChRO-seq	29292587	41525	37769	50183	X
PDX-89-08_P7	leChRO-seq	18944268	22675	27496	20462	X
PDX-88-04_P57	ChRO-seq	119662422	34901	42500	28063	

Supplementary Table 1. 2 Differentially transcribed genes across all 20 primary GBMs relative to technical replicates of the non-malignant brain detected using DESeq2.

The first 7 columns show the information of the annotated genes. The log2FoldChange shows the $\log_2$ of ratio in transcription, measured as primary GBM patients (n=20) over non-malignant brain (n=2). The padj shows the FDR-corrected p values (Wald test). Genes with padj<0.05 were shown.

Supplementary Table 1. 3 Differentially transcribed genes across each GBM subtype relative to technical replicates of the non-malignant brain detected using DESeq2.

The first 7 columns show the information of the annotated genes. The last eight columns show the $\log_2$ fold change and adjusted p values for each of the four subtypes. Subtypename.log2FoldChange shows the $\log_2$ of ratio in transcription, measured as the GBM of the given subtype (n= 2 [classical] or 3 [other subtypes]) over non-malignant brain (n=2). The Subtypename.padj shows the FDR-corrected p values (Wald test) for the change of transcription in the given subtype. Genes with padj<0.05 in at least one subtype were shown.

Due to the length of the list, please refer to the online Supplementary Information https://doi.org/10.1038/s41588-018-0244-3

Supplementary Table 1. 4 The distribution of taTRE in each patient and each transcriptional module.

Sample	No. of total TRE	No. of taTRE	Percentage of taTRE	No. of taTRE in differentiated module	No. of taTRE in stem module	No. of taTRE in immune module
GBM-05-05	32874	4101	0.12	441	207	147
GBM-05-15	33388	5511	0.17	1219	201	596
GBM-05-16	19950	1255	0.06	186	52	40
GBM-05-17	28699	2374	0.08	52	137	50
GBM-05-18	49345	11910	0.24	1687	449	497
GBM-05-21	33950	4703	0.14	545	237	394
GBM-05-23	32574	4556	0.14	488	204	158
GBM-05-26	21509	1567	0.07	222	59	273
GBM-05-30	44908	9046	0.2	1602	310	587
GBM-05-33	23024	1832	0.08	182	92	187
GBM-05-35	24524	1435	0.06	106	59	214
GBM-05-45	37792	6541	0.17	1420	189	440
GBM-06-05	35007	4475	0.13	828	169	498
GBM-06-12	30744	4800	0.16	838	118	1049
GBM-07-02	18988	1455	0.08	210	67	224
GBM-07-05	29420	3661	0.12	342	208	94
GBM-07-07	37165	5126	0.14	286	265	227
GBM-88-02_P3	37769	6668	0.18	739	198	173
GBM-88-04	19551	1297	0.07	278	254	123
GBM-88-04_P57	28063	4156	0.15	136	33	261
GBM-89-08_P7	20462	1994	0.1	116	125	48
GBM-97-04	14968	319	0.02	8	5	49
GBM-15-90	53197	6648	0.12	567	456	349

Supplementary Table 1. 5 The distribution of subtype-biased TRE.

<table>
<tr><td colspan="5">Total dREG-HD called at relaxed condition sites=177729</td></tr>
<tr><td>Subtype</td><td>Patient</td><td>up</td><td>down</td><td>unchanged</td></tr>
<tr><td>Classical</td><td>GBM-05-05, GBM-07-07</td><td>243</td><td>58</td><td>69994</td></tr>
<tr><td>Mesenchymal</td><td>GBM-07-02, GBM-06-12, GBM-88-04</td><td>2174</td><td>732</td><td>134435</td></tr>
<tr><td>Neural</td><td>GBM-05-35, GBM-97-04, GBM-15-90</td><td>2134</td><td>100</td><td>155556</td></tr>
<tr><td>Proneural</td><td>GBM-05-16, GBM-07-05, GBM-05-23</td><td>409</td><td>925</td><td>129648</td></tr>
</table>

Supplementary Table 1. 6 Clinical statistics of the target genes shared by three survival-associated transcription factors.

P value is calculated by two-sided Chi-squared test for the survival days of patient with upper quartile expression (N=51) and lower quartile expression (N=51) of the given gene. Hazard ratio is defined as higher expression / lower expression. NA value indicates that the genes is not measured by the microarray data.

Gene name	P value	Difference in median of survival time (days)	Hazard ratio of high transcription
CXCL3	0.0387	-64	1.57
PPP1R15A	0.0566	-73	1.49
ALOX5AP	0.396	-37.5	1.19
SLC25A37	0.773	-48	1.06
MGAT1	0.159	-50.5	1.35
ZC3H12A	0.106	-67	1.42
LPCAT1	0.291	-100.5	1.25
VPS37C	0.233	-42.5	1.28
DUSP6	0.0151	-137	1.68
SDCBP	0.614	-40	1.11
SLC16A6	0.916	-48	1.02
PAX8	0.633	-46	1.11
ADM	0.0343	-98	1.60
KYNU	0.997	38.5	1.00
SERPINH1	0.146	-73	1.36
SERPINE1	0.289	-52	1.25
UPP1	0.0187	-107	1.66
ITGA5	0.00541	-137	1.81
RHOH	0.228	-67	1.29
TNFAIP3	0.389	-37	1.20
CHI3L1	0.0211	-74.5	1.64
NNMT	0.329	-17	1.23
BCL2A1	0.0396	-86	1.55
CFLAR	0.199	-31	1.31
MYBPH	0.0812	-96.5	1.45

CCL20	0.00237	-149	1.94
CD300LB	NA	NA	NA
RP13-516M14.2	NA	NA	NA
LINC01272	NA	NA	NA
LINC01270	NA	NA	NA
OSM	NA	NA	NA
EAF1	NA	NA	NA
RP11-24F11.2	NA	NA	NA
RBM47	NA	NA	NA
FILIP1	NA	NA	NA
RP11-356I2.4	NA	NA	NA
CCDC71L	NA	NA	NA
NECTIN4	NA	NA	NA
RP11-519G16.3	NA	NA	NA
AC092839.3	NA	NA	NA
PAX8-AS1	NA	NA	NA
AF064858.8	NA	NA	NA
CTB-138E5.1	NA	NA	NA
AC051642.1	NA	NA	NA

Supplementary Table 1. 7 Gene ontology analysis of target genes of three survival-associated transcription factors.

Table shows the fold of enrichment and p value (two-sided Fisher's Exact with FDR multiple test correction) of each gene ontology terms (Sample size: RELA=127; C/EBP=196; RARG=273).

	RELA		C/EBP		RARG	
GO Terms	Fold Enrichment	FDR	Fold Enrichment	FDR	Fold Enrichment	FDR
regulation of phosphorus metabolic process (GO:0051174)	3.83	2.30E-06	3.61	1.16E-11	2.82	4.92E-06
regulation of phosphate metabolic process (GO:0019220)	3.86	3.80E-06	3.49	3.57E-11	2.77	1.09E-05
response to stress (GO:0006950)	2.69	5.94E-06	2.58	2.72E-11	2.6	3.14E-10
regulation of phosphorylation (GO:0042325)	4.01	7.42E-06	3.8	3.86E-11	2.93	1.21E-05
positive regulation of phosphate metabolic process (GO:0045937)	4.57	8.39E-06	4.25	1.84E-10	3.62	1.58E-06
positive regulation of phosphorus metabolic process (GO:0010562)	4.57	9.59E-06	4.25	2.11E-10	3.62	1.72E-06

response to lipopolysaccharide (GO:0032496)	9.38	1.23E-05	7.84	1.13E-08	7.06	1.88E-06
response to cytokine (GO:0034097)	4.58	1.48E-05	3.99	1.26E-08	4.02	1.05E-06
response to molecule of bacterial origin (GO:0002237)	8.84	1.78E-05	7.39	2.36E-08	6.66	3.39E-06
regulation of protein phosphorylation (GO:0001932)	3.84	3.86E-05	3.96	3.11E-11	3.01	1.79E-05
regulation of response to stimulus (GO:0048583)	2.3	4.05E-05	2.13	3.85E-08	1.95	3.04E-05
response to organic substance (GO:0010033)	2.69	5.46E-05	2.61	1.64E-09	2.39	1.87E-06
response to chemical (GO:0042221)	2.25	8.36E-05	2.08	9.65E-08	1.94	2.47E-05
immune system process (GO:0002376)	2.75	8.45E-05	2.8	2.42E-10	2.78	6.58E-09

CHAPTER 2

DISCOVERING TRANSCRIPTIONAL REGULATORY ELEMENTS FROM RUN-ON AND SEQUENCING DATA USING THE WEB-BASED DREG GATEWAY

2.1 Abstract

Transcription is effective mark that can be used to identify the location of active enhancers and promoters, collectively known as transcriptional regulatory elements (TREs). We have recently introduced dREG, a tool for the identification of TREs using run-on and sequencing assays like GRO-seq, PRO-seq, and ChRO-seq. In this protocol, we present step-by-step instructions for running dREG on an arbitrary run-on and sequencing dataset. Users provide dREG with bigWig files representing the location of RNA polymerase in a cell or tissue sample of interest. dREG returns genomic regions that are predicted to be active TREs. Finally, we demonstrate the use of dREG regions in discovering transcription factors controlling response to a stimulus and predict their target genes. Together, this protocol provides detailed instructions for running dREG on arbitrary run-on and sequencing data.

2.2 Introduction

DNA sequence control regions, such as promoters, enhancers, and insulators, collectively known as transcriptional regulatory elements (TREs), are critical components of the genetic regulatory programs of all organisms. TREs can be identified using a family of run-on and sequencing assays that map the location of RNA polymerase, including global run-on and sequencing (GRO-seq), precision run-on and sequencing (PRO-seq), and chromatin run-on and sequencing (ChRO-seq) (Core, Waterfall, and Lis 2008; Chu et al. 2018). The detection of transcriptional regulatory elements by GRO-seq, PRO-seq, and ChRO-seq data (dREG) is a method that can be used to identify the location of transcriptional regulatory elements (Danko et al. 2015; Z. Wang et al. 2019).

In this article we provide a detailed protocol for using dREG to identify TREs in any global run on and sequencing experiment. We provide two separate ways to run dREG: First, we demonstrate the use of the dREG web server, available at (http://dreg.dnasequence.org). Second, we provide a detailed account of the steps that are required to download and run dREG in a user's own computer system. Finally, we provide an example of how the output of dREG can be used to map the location of transcription factors and predict which genes are targets. In summary, this protocol allows researchers to discover the location of active TREs by running dREG on PRO-seq data collected in their own lab.

2.3 Strategic Planning

The input to dREG consists of mapped reads from a GRO-seq, PRO-seq, or ChRO-seq experiment. The quality and quantity of the experimental data are major factors in determining how sensitive dREG will be in detecting TREs. We have found that dREG has a reasonable statistical power for discovering TREs with as few as ~40M uniquely mappable reads, and saturates detection of TREs in well-studied ENCODE cell lines with >80M reads (Z. Wang et al. 2019). To increase the number of reads available for TRE discovery, we typically merge biological replicates to improve our statistical power prior to running dREG. To further improve data quality, our lab makes extensive use of unique molecular identifiers (UMIs) in RNA adapters during library prep, which allow us to identify and remove any PCR duplicates (Mahat et al. 2016; Fu et al. 2014). Typical duplication rates vary due to a variety of factors, including the quality of the input sample, the amount of starting material, and the number of cycles of PCR amplification. These experimental controls and considerations must be considered carefully while planning a PRO-seq experiment.

Once investigators have experimental data in hand, the next step is to produce two bigWig files which represent the position of RNA polymerase on the positive and negative strands. The sequence alignment and processing steps to make the input bigWig files is another major factor influencing the success of dREG. Users can create bigWig files from their own alignment pipeline that are compatible with dREG. However, dREG makes several assumptions about data processing that are critical for success. Critical elements of a bioinformatics pipeline will include:

- **Include a copy of the Pol I transcription unit in the reference genome**. PRO-seq data resolves the location of all four RNA polymerases found in Metazoan cells (Pol I, II, III, and Mt) (Core, Waterfall, and Lis 2008; Kwak et al. 2013; Blumberg et al. 2017; Hah et al. 2011). DNA encoding the Pol I transcription unit is highly repetitive, and is not included in most mammalian reference genomes. Nevertheless, the Pol I transcription unit is a substantial source of reads in a typical PRO-seq experiment (10-30%). Many of these reads will align spuriously to retrotransposed and non-functional copies of the Pol I transcription unit, which can create mapping artifacts (Core, Waterfall, and Lis 2008). To solve this issue, we include a single copy of the repeating DNA that encodes the Pol I transcription unit in the reference genome used to map reads. We use GenBank ID# U13369.1. Including a copy of this transcription unit provides an alternative place for Pol I reads to map, preventing reads from accumulating in Pol I repeats.

- **Trim 3’ adapters, but leave the fragments.** Much of the signal for dREG comes from paused RNA polymerase. RNA polymerase pauses 30-60 bp downstream of the transcription start site (Kwak et al. 2013). Due to this short RNA fragment length, paused reads in most PRO-seq libraries will sequence a substantial amount of adapter. This leads to poor mapping rates in full-length reads. Therefore, it is crucial to remove contaminating 3’ adapters so that paused fragments will map to the reference genome properly.
- **Representing RNA polymerase location using a single base.** PRO-seq measures the location of the RNA polymerase active site, in many cases at nearly single nucleotide resolution. Therefore, it is logical to represent the coordinate of RNA polymerase using the genomic position that best represents the polymerase location, rather than representing the entire read. dREG assumes that each read is represented in the bigWig file by a single base. We have noted poor performance when reads are extended. It is critical that users pass in bigWig files that represent RNA polymerase using a single nucleotide.
- **Data represents unnormalized raw counts.** dREG assumes that data represents the number of individual sequence tags that are located at each genomic position. For this reason, it is critical that input data is not normalized. The dREG server checks to ensure that input data is expressed as integers, and will return an error if this is not the case.

As an alternative to developing their own pipeline, users are also able to use our bioinformatic pipeline for aligning PRO-seq data. Our pipeline produces bigWig files that are compatible with dREG, and can be found at the following URL: https://github.com/Danko-Lab/proseq_2.0. Our PRO-seq pipeline takes single-end or pair-ended sequencing reads (fastq format) as input. The pipeline automates routine pre-processing and alignment steps, including pre-processing reads to remove the adapter sequences and trim based on base quality, and deduplicate the reads if UMI barcodes are used. Sequencing reads are mapped to a reference genome using BWA. Aligned BAM files are converted into bigWig format in which each read is represented by a single base.

To run our pipeline users must first download the pipeline files and install dependencies indicated in the README.md. In addition, users need to provide a path to a BWA index file and the path to the chromInfo file for the genome of choice. After running this pipeline, users should have processed data files in the specified output directory.

Finally, we also provide a tool that converts mapped reads from a BAM file into bigWig files that are compatible with bigWig. This tool is available here: https://github.com/Danko-Lab/RunOnBamToBigWig

We have found that visualizing aligned data in a genome browser prior (e.g., IGV or UCSC) to downstream analysis is a useful way to catch any data quality or alignment issues. Users are directed to the Troubleshooting section for additional information and examples.

2.4 Basic Protocol 1

Finding TREs in PRO-seq data using the dREG web server.

2.4.1 Introductory paragraph

dREG identifies active transcriptional regulatory regions (TRE) based on a pre-trained Support Vector Regression (SVR) model, which can be used to do TRE discovery and peak-calling on GRO-seq, PRO-seq, or ChRO-seq data. In general, running dREG by the web server executes the following steps.

1) Identify informative genomic positions. Loci that are low in PRO-seq reads are pre-filtered and excluded from running peak calling. We select loci for analysis that meet either of the following heuristics: 1) contain more than 3 reads in a 100 bp interval on either strand, or 2) more than 1 reads in 1 kbp interval on both strands. We refer to positions meeting these criteria as “informative positions”.

2) Predicting dREG scores. We used support vector regression (SVR) to score 50 bp intervals along the genome, using a pre-trained SVR model. The PRO-seq profile of each informative position was described using a 360-dimensional feature vector. This feature vector integrates the PRO-seq counts using sliding windows at 5 different scales, and transformed using logistic normalization to better represent their shapes. Extracting the feature vector was done on CPUs, and can be distributed on multiple CPU cores. dREG runs the actual prediction on the GPU, leveraging the power of parallelized computing, and hence greatly improves the efficiency of computing.

3）Calling dREG peaks. We stitch regions of high dREG scores into candidate peaks, and then estimate the probability that these peaks are drawn from the negative set of sites. The final predictions for genomic regions that contain transcription start sites are corrected using the false discovery rate correction for multiple testing and reported to the user.

2.4.2 Necessary Resources

1）Javacript and Cookie-enabled browsers. Currently 3 browsers are recommended: Firefox, Google Chrome, and Safari.

2) Sample data:

bigWig files compatible for running dREG can be downloaded from Gene Expression Omnibus (GEO). Links to the example bigWig files are listed in Table 2.1. For simplicity, we rename each files by removing the GSM/GSE ID, such that GSM2265095_H1-U_plus.bw becomes H1-U_plus.bw, etc. The bigWig files can also be downloaded from

ftp://cbsuftp.tc.cornell.edu/danko/hub/protocol.files/bigWigs.raw .

2.4.3 Protocol steps—Step annotations

1) Map reads to the reference genome and confirm the appropriate format and data quality. dREG makes several assumptions about how RNA polymerase is represented in the input bigWig files that substantially affect the results (see Strategic Planning section). In particular: (1) The location of each read must be represented by a single base that denotes as accurately as possible the location of the RNA polymerase active site, and (2) Data must be unnormalized raw read counts. Users who have not worked with PRO-seq data before can use our alignment pipeline, which is compatible with

GRO-seq, PRO-seq, and ChRO-seq data that is both single or paired-end. Finally, bigWig files can be merged from correlated replicates using the script https://github.com/Danko-Lab/proseq2.0/blob/master/mergeBigWigs.bsh.

2) For the purpose of demonstration, we provide an example using human primary T cells with/without PMA and ionomycin treatment (PI). The sample files are listed above, and can be downloaded as dREG-ready bigWigs files from the GEO database using either ftp/http protocol.

To increase the sensitivity of dREG, users may merge the bigWigs of biological replicates under each experimental condition, i.e. PI treated and untreated, using the mergeBigWigs.bsh. The script is shown below. The merged bigWig files can also be downloaded from ftp://cbsuftp.tc.cornell.edu/danko/hub/protocol.files/bigWigs.merged.

```
$ mergeBigWigs.bsh -c chromInfo.hg19 H-U_plus.bw H1-U_plus.bw H2-U_plus.bw H4-U_plus.bw
$ mergeBigWigs.bsh -c chromInfo.hg19 H-U_minus.bw H1-U_minus.bw H2-U_minus.bw H4-U_minus.bw
$ mergeBigWigs.bsh -c chromInfo.hg19 H-PI_plus.bw H1-PI_plus.bw H2-PI_plus.bw H4-PI_plus.bw
$ mergeBigWigs.bsh -c chromInfo.hg19 H-PI_minus.bw H1-PI_minus.bw H2-PI_minus.bw H4-PI_minus.bw
```

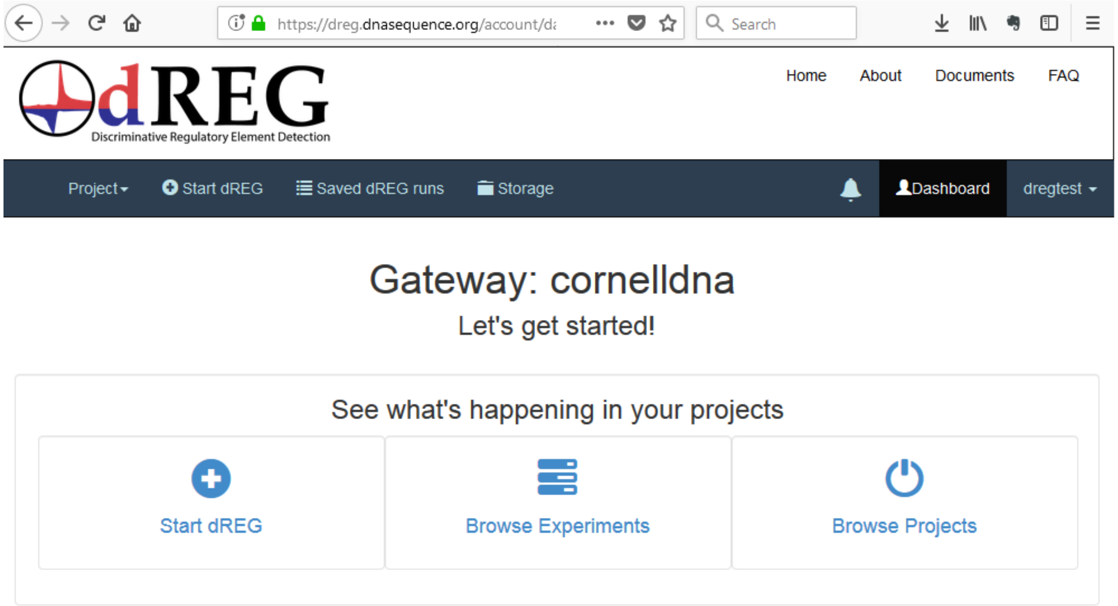
Where the chromInfo.hg19 is a text file that specifies the chromosome size, and can be downloaded and generated from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/ . We will use one pair of merged bigWig files, H-U_plus.bw and H-U_minus.bw, as an example to run dREG. To run the downstream analysis on differential regulation, users need to run dREGs for H-PI_plus.bw and H-PI_minus.bw as well. The output of dREG of these two pairs of files can be downloaded from ftp://cbsuftp.tc.cornell.edu/danko/hub/protocol.files/dREG.output. To avoid wasting unnecessary computing resources on running the examples, users are advised to directly download the results for the examples from the above ftp link, or use their own data of interest.

3) Navigate to the dREG Science Gateway. The dREG Science Gateway can be accessed at http://dREG.dnasequence.org/.

4) Register for an account. The dREG gateway requires users create a new account. Users may register for an account at the homepage of dREG gateway. The dREG gateway will send an email containing the link to activate the account. Please check the spam email in case the registration email is blocked. Under rare circumstances, the activation email can be quarantined by institutional email accounts, which are usually are not delivered to the email box, and hence cannot be found in any email folders, including inbox and spam. If the emails from dREG gateway are found to be undeliverable, please contact your administrator or use another email account, such as Gmail, for registration propose.

Once the registration is completed, users may sign in the account and use the following steps to run a dREG analysis.

5) Once logged in, a dashboard will show up to the user. Click the “Start dREG” icon to create a new dREG analysis.



Gateway: cornelldna

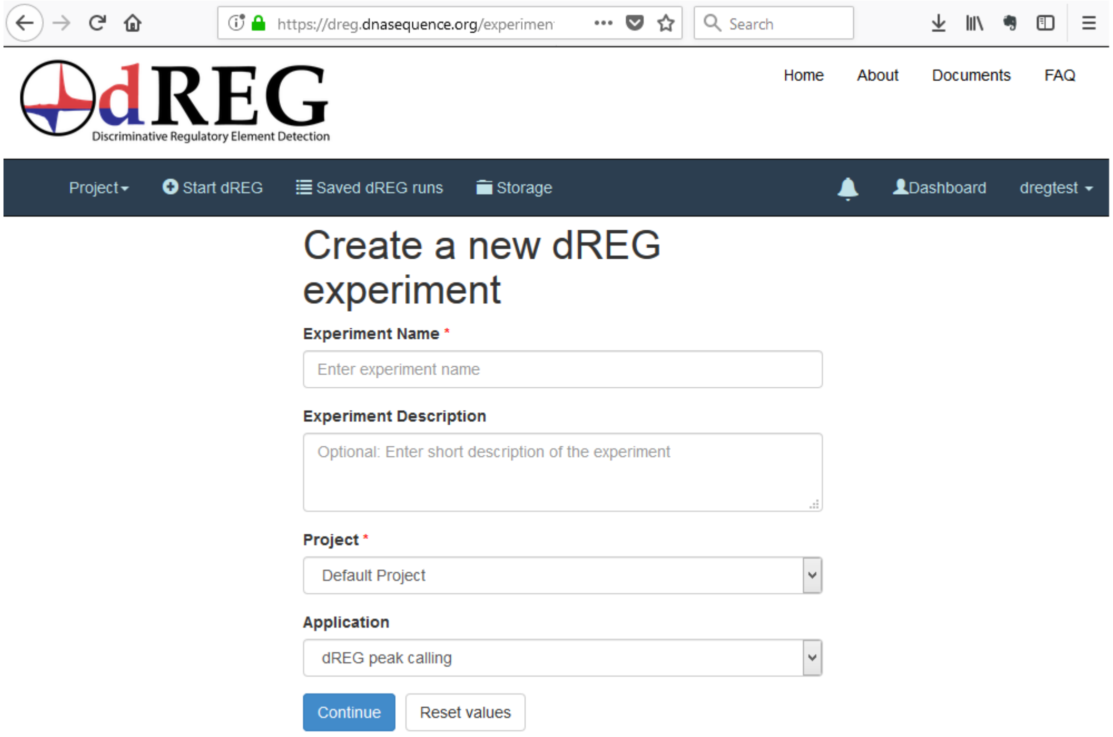
Let's get started!

See what's happening in your projects

Fig. 2. 1 The Dashboard page.

This page is the entry point for dREG peak calling. Select “Start dREG” to launch a new computation experiment for a new PRO-seq data set.

6) The next window (Fig. 2.2) requests information about the dREG experimental design, including the name of the experiment, the project name "Default Project", and other metadata.



Create a new dREG experiment

Experiment Name *

Enter experiment name

Experiment Description

Optional: Enter short description of the experiment

Project *

Default Project

Application

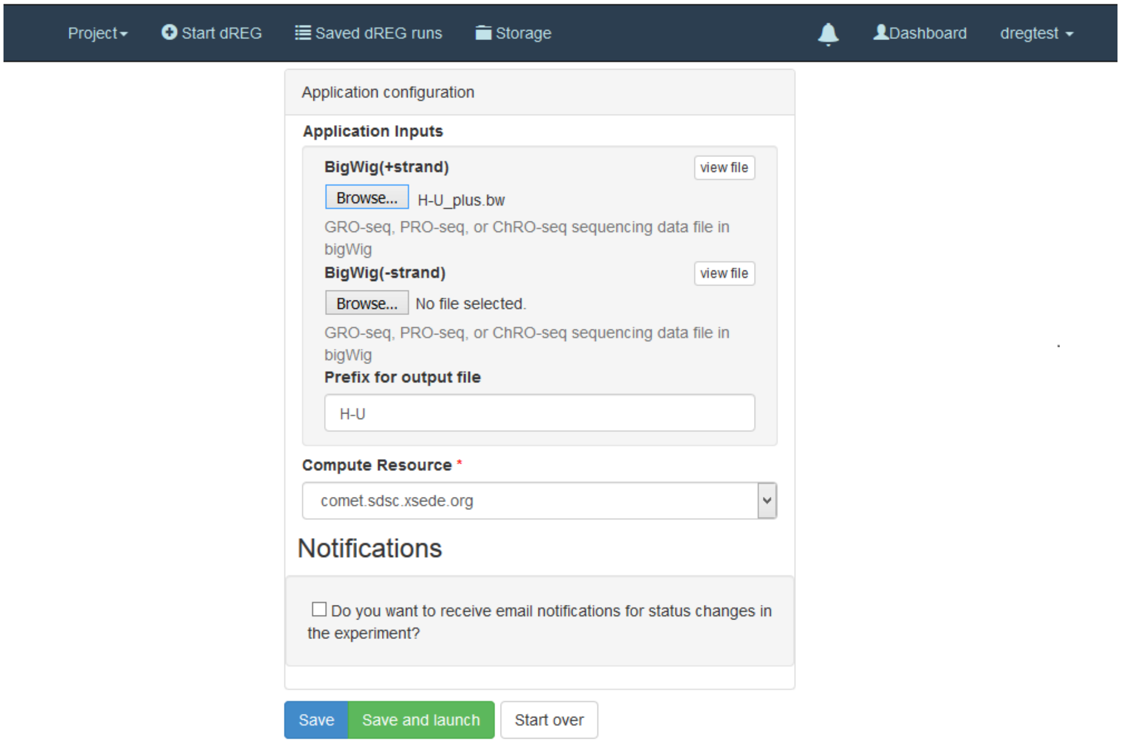
dREG peak calling

Fig. 2. 2 The webpage to launch a new dREG experiment.

Users need to select the dREG peak calling application and input metadata about the experiment. The experiment name can be used to identify the experiment in shared projects.

7) Upload the bigWig files representing the location of RNA polymerase on the plus and minus strand, in which case we use H-U_plus.bw and H-U_minus.bw . Users also need to specify the prefix to the names of the output files, which will be used to label the files that are delivered to the user in the output. Once bigWig files are uploaded onto the server, please click “Save and launch”.

Upon being launched, dREG gateway will submit the computing task to computers in the XSEDE cluster using the Apache Airavata server (Pamidighantam et al. 2016). These processes include i) transfering user-submitted files to storage space of the GPU node, ii) submitting a bash script which specifies the runs of dREG to the GPU node, iii) the execution of bash script is queued, and an notification email will be sent back to Apache Airavata server once the script is executed, iv) once the dREG run is complete, another notification email will be sent to the Apache Airavata server, v) the Apache Airavata server returns the result of dREG run to the user’s web storage and notify the user through user’s email.



Application configuration

Application Inputs

BigWig(+strand)

H-U_plus.bw

GRO-seq, PRO-seq, or ChRO-seq sequencing data file in bigWig

BigWig(-strand)

No file selected.

GRO-seq, PRO-seq, or ChRO-seq sequencing data file in bigWig

Prefix for output file

H-U

Compute Resource *

comet.sdsc.xsede.org

Notifications

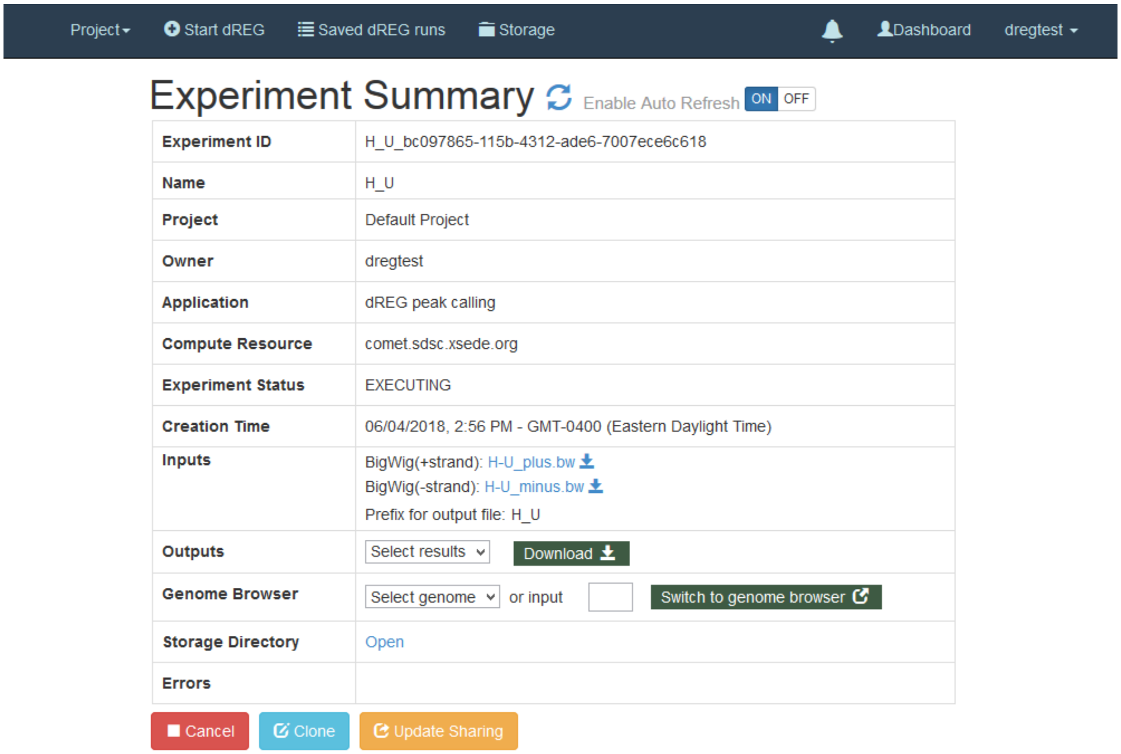
□ Do you want to receive email notifications for status changes in the experiment?

Fig. 2. 3 The Data Upload page.

This page is the second step of starting new dREG experiment.

8) After launching dREG, users will be directed to a summary table of the current task, shown in Fig. 2.4, which lists the cluster address, the status of queue, the input files, and the creation time of the task. If the computing task is returned quickly, it usually means the dREG run was interrupted by errors.

There are two possibilities to this: the use of bigWig input files that do not meet the requirement, such as use of normalized values, and the mapping of whole reads instead of only the end of the reads to the bigWig, or insufficient computing resource of the server (see **Troubleshooting**). Click “Open” under the “Storage Directory” and access the log files to identify any errors. If the project runs normally, users may close the webpage. Each run usually takes 4-12 hours, depending on the queue and the execution time on the GPU node.



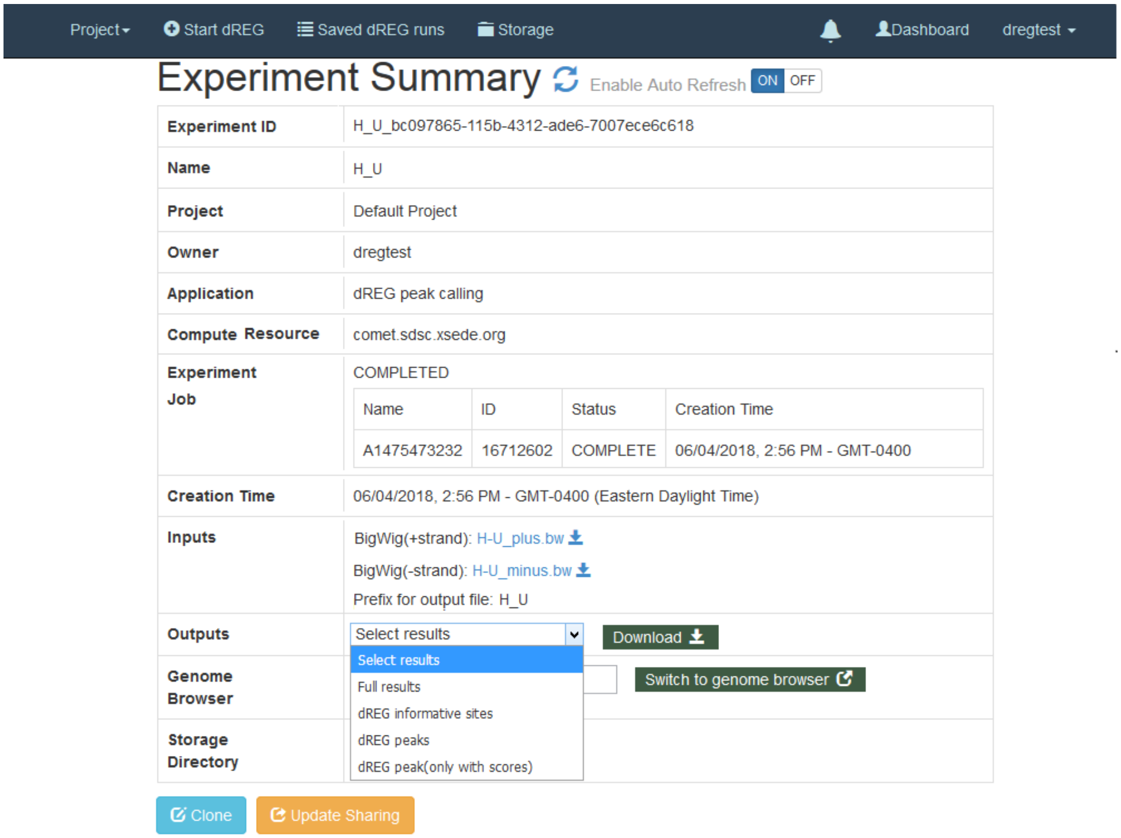
Experiment Summary

Experiment ID	H_U_bc097865-115b-4312-ade6-7007ece6c618
Name	H_U
Project	Default Project
Owner	dregtest
Application	dREG peak calling
Compute Resource	comet.sdsc.xsede.org
Experiment Status	EXECUTING
Creation Time	06/04/2018, 2:56 PM - GMT-0400 (Eastern Daylight Time)
Inputs	BigWig(+strand): H-U_plus.bw BigWig(-strand): H-U_minus.bw Prefix for output file: H_U
Outputs	Select results Download
Genome Browser	Select genome or input Switch to genome browser
Storage Directory	Open
Errors	

Fig. 2. 4 The Experiment Summary page.

This table page lists the basic experimental information, jobs status, and allows the user to download or visualize data once complete.

9) Downloading results. Users will be notified by e-mail when the dREG run is complete. Once dREG is complete, users may log in the dREG gateway and click the “Browse Experiments” icon on the user dashboard to access the dREG results. Users will be directed to “Experiment Summary”, where the results of the dREG run will be available for download onto a local machine (see Guidelines for Understanding Results). Results can also be visualized using the WashU Epigenome Genome browser.



Experiment Summary

Experiment ID	H_U_bc097865-115b-4312-ade6-7007ece6c618
Name	H_U
Project	Default Project
Owner	dregtest
Application	dREG peak calling
Compute Resource	comet.sdsc.xsede.org
Experiment Job	COMPLETED
Creation Time	06/04/2018, 2:56 PM - GMT-0400 (Eastern Daylight Time)
Inputs	BigWig(+strand): H-U_plus.bw BigWig(-strand): H-U_minus.bw Prefix for output file: H_U
Outputs	
Genome Browser	
Storage Directory	

Name	ID	Status	Creation Time
A1475473232	16712602	COMPLETE	06/04/2018, 2:56 PM - GMT-0400

Fig. 2. 5 The Output File list.

The drop-down list shows main 4 results can be downloaded using the download link. In the web storage page, additional files are available for download. To run downstream analysis of differential regulation, download the results to local directory.

10) Visualizing results on the WashU Epigenome Browser (Zhou et al. 2011). To visualize the result, choose / type in the reference genome build in UCSC version numbering that was used to create the bigWig files (e.g., hg38 or hg19 for human, or mm10 for mouse) and then click "Switch to genome browser". The browser will open a new tab that will lead users to the WashU epigenome browser webpage. Both the input bigWig files and results of dREG will be visualized in separate tracks, as shown in Fig. 2.6.

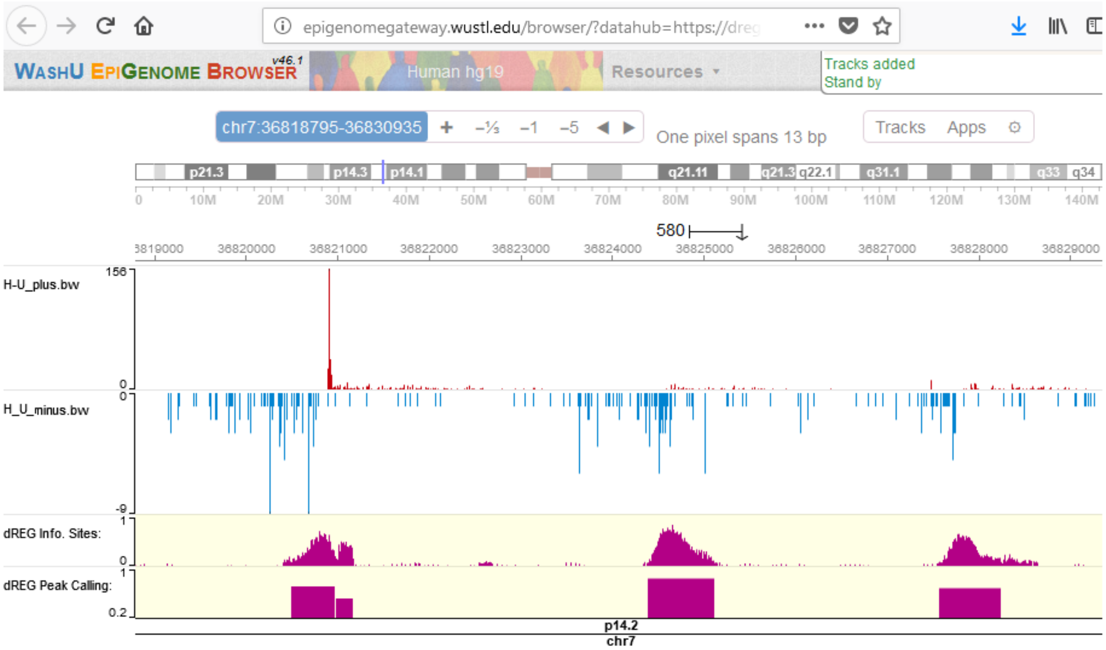


Fig. 2. 6 The Genome Browser page.

The four genome browser tracks show mapped PRO-seq reads in plus strand, mapped reads in minus strand strand, dREG scores for each informative position, and the location of significant peak region (FDR<0.05). Use the zoom buttons at the top line to view dREG peaks near your locus of interest.

11) Once dREG is complete, results will be stored on the server for a period of 30 days.

2.5 Alternate Protocol

Running a local copy of dREG

Many applications may require downloading and running dREG locally. Here we provide a detailed protocol for running a local copy of dREG.

2.5.1 Necessary Resources

Estimates of hardware resources are based on a deeply sequenced (~40-400 M mapped reads) PRO-seq for Human Genome Reference GRCh37d5

Hardware
A Linux computer with at least 128 GB of RAM
8 CPU cores
GPU with 12 GB memory (supports CUDA 6.5 or above)
Disk storage of 1TB
Run-time, 4-12 hrs

Software
(1) Git (https://git-scm.com/download/linux)
(2) bedops (http://bedops.readthedocs.org/en/latest/index.html)
(3) boost library (https://www.boost.org/users/download/)
(4) CUDA 6.5 or above (https://developer.nvidia.com/cuda-toolkit)
(4) R software with the following package:
a) dREG and its dependencies (https://github.com/Danko-Lab/dREG)

bigWig (>= 0.2-9), data.table, e1071, mvtnorm, parallel, rmutil, randomForest, snowfall.

See Support Protocol for installation instructions

b) Rgtsvm and its dependencies (https://github.com/Danko-Lab/Rgtsvm)

bit64, snow, SparseM, Matrix

See Support Protocol for installation instructions

Files

dREG SVR model used for peak calling, it can be downloaded from ftp://cbsuftp.tc.cornell.edu/danko/hub/dreg.models

As of this writing, the most recent model is named asvm.gdm.6.6M.20170828.rdata.

2.5.2 Protocol steps—Step annotations

1) Map reads to the reference genome and confirm the appropriate format and data quality.

2) Run the main dREG application. The main dREG pipeline scores 50 bp intervals along the genome for similarity to a TRE, and generates a BED file with narrow peaks, peak scores, probability, and peak center positions. We provide a bash script which allows users to automatically execute all of the stages of this pipeline. The script is under the dREG directory, and can be configured and run as follows:

1. Set an environment variable for the path to the RData file containing the pre-trained SVM

```
export dREG_MODEL=/your/path/asvm.gdm.6.6M.20170828.rdata
```

2. Run dREG by executing the main bash script: run_dREG.bsh. First define and assign variables required for running the run_dREG.bsh

```
# -- PRO-seq data (plus strand).
# Read counts (not normalized) formatted as a bigWig file.
PLUS_STRAND_BW=H-U_plus.bw


#-- PRO-seq data (minus strand).
# Read counts (not normalized) formatted as a bigWig file.
MINUS_STRAND_BW=H-U_minus.bw

#-- The prefix of the output file.
OUT_PREFIX=H-U

# [optional, default=1]
# CPU cores can be used for feature extraction and peak identification.
CPU_CORES=16


# [optional, default=NA]
# GPU id when multiple GPU cards are available. The first ID is 0.
GPU_ID=0
```

Build the run_dREG.bsh command (this example uses parameters defined above):

```
$ bash run_dREG.bsh\
    $PLUS_STRAND_BW\
    $MINUS_STRAND_BW\
    $OUT_PREFIX\
    $dREG_MODEL\
    $CPU_CORES\
    $GPU_ID
```

The actual time for running run_dREG.bsh depends on the number of informative positions, the number of broad peaks generated from these informative positions, and the speed of the computer on which dREG is run. Due to the large size of the new dREG model, large amounts of intermediate data are generated when running dREG. Users are advised to make sure that they have sufficient amount of free memory, otherwise the dREG process may be killed by the system.

Once dREG exits, it should add 5 main files under the current working directory. These files are described in detail under Guidelines for Understanding Results.

2.6 Basic Protocol 2

Using dREG to identify transcription factors and their downstream target genes

2.6.1 Introductory paragraph

Transcription factors (TFs) are proteins that affect the abundance of RNA polymerase on genes by binding to specific DNA sequence elements in TREs identified using dREG. PRO-seq and related run-on and sequencing assays measure RNA polymerase at both regulatory elements and annotated genes. This information can be used to identify specific groups of TREs regulated by each transcription factor, and predict a set of putative target genes responding to each TF. This information results in predictions for a partial regulatory network connecting TFs to the set of bound TREs, and the potential target genes associated with each binding event (TF-TRE-target gene).

One important task in many biological applications is to identify changes in TF binding between two conditions (e.g. treatment vs. control). Other applications require connecting changes in TF recruitment to the activity of downstream target genes. We have recently developed a strategy to solve both of these problems, and implemented our solution in an R package called tfTarget (Chu et al. 2018). This protocol describes how to use tfTarget to identify the TF-TRE-target gene networks that control differences between groups of samples.

2.6.2 Necessary Resources

Recommend requirements:

A Linux computer with 128 GB of RAM

CPU 16 cores

Data storage of 2TB

Run-time, 1 hr

Minimum requirements:

A Linux computer with 16 GB of RAM

1 core

Data storage of 200 GB

Run-time, 5 hr

Input files:

dREG narrow peaks of two conditions. An example can be downloaded from ftp://cbsuftp.tc.cornell.edu/danko/hub/protocol.files/dREG.output.

bigWig files for PRO-seq data of two conditions, with at least two replicates for each condition. An example can be downloaded from ftp://cbsuftp.tc.cornell.edu/danko/hub/protocol.files/bigWigs.raw .

A gene annotation file, of the same genome assembly of the bigWig files. See 3.2 for details.

The tfs.rdata file containing the TF motif database (required for non-human species). See 3.3 for details.

The 2bit file representing the genome of interest. See 3.4 for details.

Install the tfTarget package and dependencies.

1. Install R, and dependent packages including rphast, rtfbdbs, grid, cluster, apcluster, DESeq2, gplots.

2. Install tfTarget package for R

```
$ git clone https://github.com/Danko-Lab/tfTarget.git
$ cd tfTarget
$ R CMD INSTALL tfTarget
```

Prepare input files.

3. tfTarget works by 1) identifying differentially transcribed genes and TREs, 2) scanning the differentially transcribed TREs and assigning TF motifs to each of them, and 3) tabulate the TF, TRE and genes nearby with the information about differential transcription. Step 1 requires genomic intervals specifying the regions of genes, i.e. the gene annotation file, and TREs, i.e. the dREG regions. Step 2 needs the additional information about the TF motif database, stored in an rdata file.

3.1 Prepare a BED file specifying genomic intervals of TREs (using dREG). Some thought must be put into how to handle separate dREG intervals from multiple separate conditions. We will typically merge dREG regions across different biological conditions, and use these BED regions for downstream analysis (Danko et al. 2018; Chu et al. 2017). The genomic intervals of TREs are in bed3 format. Only the first three columns will be used. Use “cat” instead of “zcat” if the input dREG files are unzipped.

```
$ zcat H-PI.dREG.peak.score.bed.gz   H-U.dREG.peak.score.bed.gz \
| LC_COLLATE=C sort -k1,1 -k2,2n \
| bedtools merge -i stdin > merged.dREG.bed
```

3.2 Prepare the gene annotation files. The gene annotation file should be in bed6 format, i.e. strand specific. This can be prepared from GENCODE or Refseq gtf files. We recommend specifying gene ID and gene name as 4th and 5th columns of the annotation file, which will show up in the output for identification. GENCODE files can be downloaded from https://www.gencodegenes.org/releases/current.html. The script below give an example of downloading the gene annotation gtf file and then converting it to bed6 format. The output file is also available to download at ftp://cbsuftp.tc.cornell.edu/danko/hub/protocol.files/gencode.v19.annotation.bed .

```
$ wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.an
notation.gtf.gz
$ zcat gencode.v19.annotation.gtf.gz \
| awk 'OFS="\t" {if ($3=="gene") {print $1,$4-1,$5,$10,$18,$7}}' \
| tr -d '";' > gencode.v19.annotation.bed
```

3.3 Generate the database of motifs (required only for non-homo sapiens species). The tfTarget package uses motifs predicted in the Cis-BP database (Weirauch et al. 2014), and computes locations using RTFBSDB (Z. Wang, Martins, and Danko 2016). For Homo sapiens, the database of motifs is self-contained in tfTarget package, and will be used by default. For others species, users may use the following command to generate the species.tfs.rdata, which contains the curated transcription factor motifs database for the species of interests. The look-up table for species name can be found from the "species" column (the 1st column) of http://cisbp.ccbr.utoronto.ca/summary.php?by=1&orderby=Species

```
$ R --vanilla --slave --args Mus_musculus < get.tfs.R
```

3.4 Download the reference genome in 2bit format. Reference genome can be found at http://hgdownload.cse.ucsc.edu/downloads.html . For the example, we download the reference genome for hg19.

```
$ wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit
```

Running tfTarget package:

4. tfTarget can be run using the following command “bash run_tfTarget.bsh ...”, with … specifying the parameters of running the tfTarget. In our case, we specify the following parameters. Users may need to change to their own directory correspondingly.

```
$ dreg_path=merged.dREG.bed
$ gene_path=gencode.v19.annotation.bed
$ bigWig_path=/your/path/
$ twoBit_path=/your/path/hg19.2bit
$ ncores=30
$ prefix=tcell
$ query_files="H1-PI_plus.bw H1-PI_minus.bw H2-PI_plus.bw H2-PI_minus.bw H4-PI_plus.bw H4-PI_minus.bw"
$ control_files="H1-U_plus.bw H1-U_minus.bw H2-U_plus.bw H2-U_minus.bw H4-U_plus.bw H4-U_minus.bw"
$ bash run_tfTarget.bsh\
                -TRE.path $dreg_path\
                -gene.path $gene_path\
                -bigWig.path $bigWig_path\
                -2bit.path $twoBit_path\
                -query $query_files\
                -control $control_files\
                -prefix $prefix\
                -ncores $ncores
```

-TRE.path and -gene.path options specify the paths to the bed files of dREG and gene annotations, respectively. -2bit.path specifies the path to the 2bit files. The genomic assembly should be consistent for these three files. -bigWig.path together with -control and -query specify the bigWig files, ordered by plus then minus strand. If -bigWig.path is not present, the current directory will be used as default. -prefix specifies the prefix for all output files. Other parameters are optional, and the details are listed on https://github.com/Danko-Lab/tfTarget.

tfTarget will by default run the complete workflow. tfTarget will identify differentially regulated TF-TRE-gene combinations between the two conditions. Alternatively, users may run only subsets of modules. Use the tag “-deseq” (without argument) to only run DEseq2 on TREs and genes. Use the tag “-rtfbsdb” to only run DEseq2 and then rtfbsdb to identify TF motifs enriched in differentially regulated TREs.

Interpreting the results from tfTarget

A complete tfTarget run will output several .pdf files and .txt files. The TFs enriched in differentially regulated TREs are shown in 2D dot plots grouped in two pdfs (Fig. 2.7), up.motif.pdf and down.motif.pdf. The p values of enrichment/depletion of motifs, calculated by two-sided Fisher's exact test, are represented by the radius of the circle, and enrichment (red) or depletion (blue) are represented by the rainbow color scale.

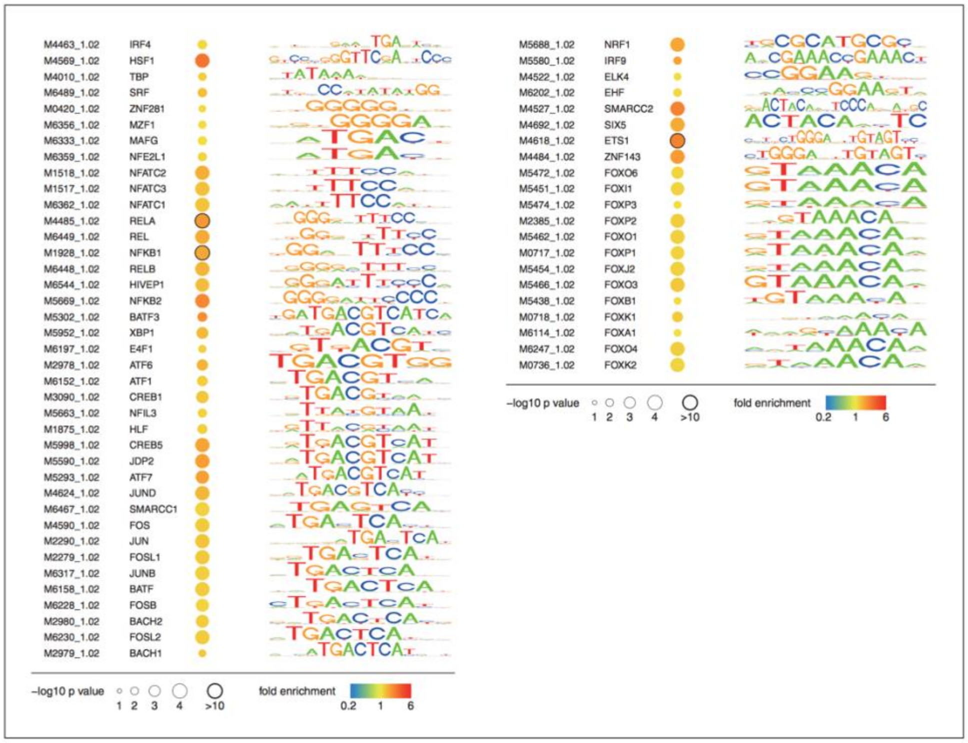
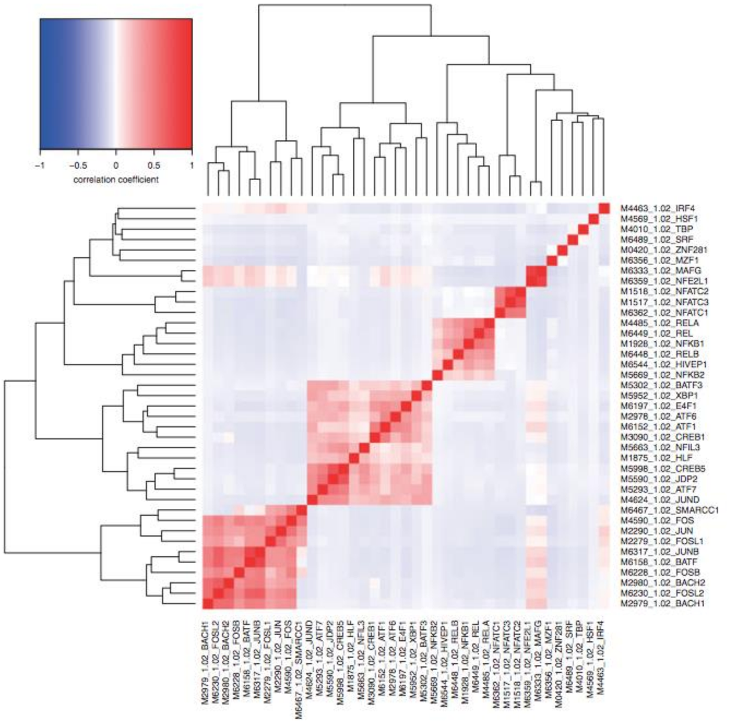


Fig. 2. 7 Motifs enriched in TREs up-regulated (left) and down-regulated (right) in PMA and ionomycin treatment, ordered by motif clusters.

Methods relying on the use of each TFs' position weighted matrix are limited in the ability to distinguish between paralogous transcription factors that share similar DNA binding specificities. To account for that when interpreting the enrichment results, tfTarget generates two additional heatmaps in pdf format that show the relation among motifs by clustering them into distinct groups based on their position in differentially regulated TREs. Note that the ordering of the motifs are consistent between 2D plots and heatmaps. The example of heatmap is shown in Fig 2.8.

The detailed statistics of tfTarget output are provided in three txt files. The ".TRE.deseq.txt" and the ".gene.deseq.txt" file lists DESeq2 statistics for each TRE and gene. Rows with all NA value are genes excluded from DESeq2 runs due to short gene length (<=1Kb).

The ".TF.TRE.gene.txt" file tabulates the relation between TFs, TREs and target genes. The results are subjected to the restriction by distance between TREs and the transcriptional start site of target gene (specified by "-dist" tag, default=50kb), the nth closest gene to the TRE (specified by "-closest.N" tag, default=2), and the p values for genes that showed same direction of log2foldchange as its regulator TRE (specified by "-pval.gene" tag, default=0.05). If needed, the latter two tags can be switched off by specifying "-closest.N off" or "-pval.gene off" to output a more inclusive list of potential target genes.



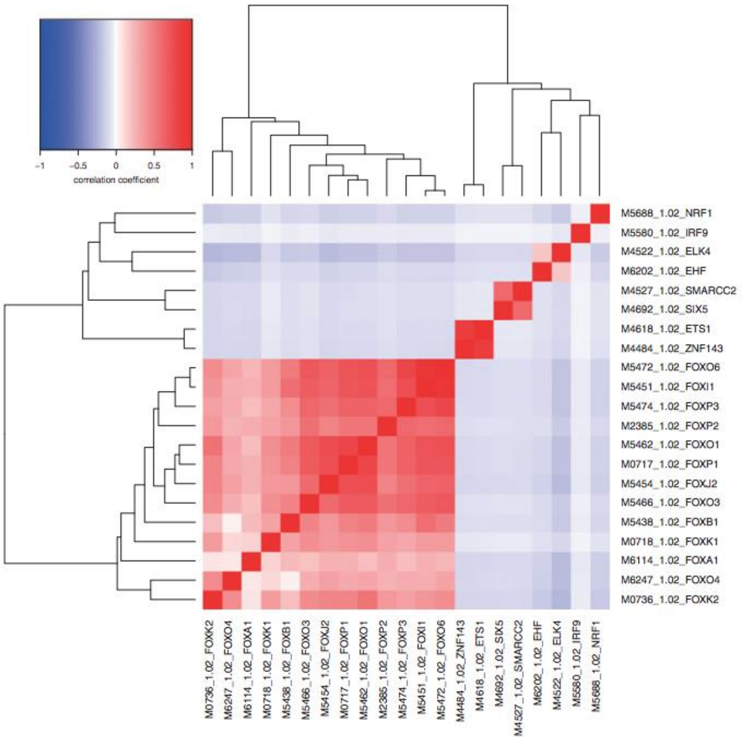


Fig. 2. 8 Heatmap shows clusters of TF motifs enriched in TREs up-regulated (upper)

and down-regulated (lower) in PMA and ionomycin treatment.

2.7 Support Protocol

INSTALLATION OF dREG AND DEPENDENCIES

dREG has been packaged to minimize the complexity of installation. The examples below use the version available at the time of publication. Please see the repositories for up-to-date instructions.

In the following examples, please modify/your/cuda/home and /your/boost/home to appropriate locations. Also, please use the same path to dREG and use this path in all of these steps.

Necessary Resources
Linux-based system with Web access

dREG and Rgtsvm Installation

1. Install R, CUDA, and Boost libraries. Please discuss this with your local systems administrator if you are unsure how to proceed. Make sure you know the path to both the CUDA home and BOOST home directories.

2. Install Rgtsvm package for GPU

```
$ export YOUR_CUDA_HOME=/your/cuda/home
$ export YOUR_BOOST_HOME=/your/boost/home
$ git clone https://github.com/Danko-Lab/Rgtsvm.git
$ cd Rgtsvm
$ make R_dependencies
$ R CMD INSTALL --configure-args="--with-cuda-home=$YOUR_CUDA_HOME --with-boost-home=$YOUR_BOOST_HOME" Rgtsvm
```

3. Install dREG package for R

```
$ git clone https://github.com/Danko-Lab/dREG.git
$ cd dREG
$ make R_dependencies
$ make dreg
```

4. Add the dREG directory to the path environment variable

```
export PATH=/your/dreg/path:$PATH
```

GUIDELINES FOR UNDERSTANDING RESULTS

The results contains the following files compressed using zip format, dREG scores of informative positions (BED format), significant dREG peaks with full information, significant dREG peaks with score only, significant dREG peaks with probability only and raw peaks. Users may either download all files as a whole or individual files separately. Raw data and results will be stored in the web storage space for up to 1 month, and outdated data will be cleaned periodically. Users are advised to download their results in time.

Running dREG will generate 5 main files under the current directory, as follows:

1. $OUT_PREFIX.dREG.infp.bed.gz:
BEDGRAPH file,   includes all informative sites and dREG scores.

2. $OUT_PREFIX.dREG.peak.full.bed.gz:
BED file, reportes all statistically significant peaks under the FDR correction (p-value < 0.05) with information about the peak position, max score, p-value (corrected using the Benjamini and Hochberg (Benjamini and Hochberg 1995) false discovery rate), and peak center.

3. $OUT_PREFIX.dREG.peak.score.bed.gz:
BED file, Significant peaks with dREG score using FDR correction (p-value < 0.05), it is partial of full information.

4. $OUT_PREFIX.dREG.peak.prob.bed.gz:
BED file, Significant peaks with probability using FDR correction (p-value < 0.05), it is partial of full information.

5. $OUT_PREFIX.raw.peak.bed.gz:
BED file, All raw peaks without p-value correction and any filters. This file is only available in the storage directory.

The peak calling script provides the option of outputting additional information for each dREG peak in the file .dREG.peak.full.bed.gz. This information includes the maximum dREG score, the probability of containing the TSS, the position of the peak center. The example is shows as follows.

```
$ zcat H-U.dREG.peak.full.bed.gz | head -
0.0233064860794223 718600
chr1    565610    565820    0.481951465645328   0         565730
chr1    567400    567760    0.899182482973753
        0.00000007889653145096l9 567590
chr1    569770    570140    0.598068431544673   0.000421199513751816
        569960
chr1    713850    714390    1.03941550120052    0         714210
chr1    714410    714780    0.426737839205382   0.00319618563824731
        714580
chr1    718370    718720    0.307836830876394   0.0233064860794223
        718600
```

chr1	723510	723830	0.333111129863476	0	723690
chr1	762570	762800	0.52136896174093	0.0104458702875426	762740
chr1	762820	763230	0.655564547732281	0.000339774400826395	762970
chr1	776390	776730	0.283670427106323	0.0350731200504118	776590

2.8 Commentary

Background Information

DNA sequence control regions, such as promoters, enhancers, and insulators, collectively known as transcriptional regulatory elements (TREs), are critical components of the genetic regulatory programs of all organisms. TREs regulate gene expression by facilitating (or inhibiting) chromatin decompaction, transcription initiation, and the release of RNA polymerase II (Pol II) into productive elongation (Fuda, Ardehali, and Lis 2009). In addition to well-characterized roles in cellular development, dysfunction in TREs also play pivotal roles in a myriad of different disease states (Shlyueva, Stampfel, and Stark 2014; Long, Prescott, and Wysocka 2016). The comprehensive identification of TREs has therefore emerged as a primary challenge in genomic research.

Active TREs recruit RNA polymerase and initiate a local and highly characteristic pattern of transcription initiation (Kim et al. 2010; de Santa et al. 2010; Core et al. 2014; Scruggs et al. 2015). Transcription initiation is a highly specific signal that can be useful for identifying active TREs in a cell type–specific manner (Melgar, Collins, and Sethupathy 2011; Core et al. 2014; Danko et al. 2015; Andersson, Gebhard, et al. 2014; Azofeifa and Dowell 2017). Although first characterized in mammals, initiation appears to mark enhancers in other Metazoan organisms (Henriques et al. 2018; Mikhaylichenko et al. 2018; Rennie et al. 2018). However, the majority of initiation events give rise to highly unstable RNA species that are rapidly degraded by the nuclear exosome complex (Preker et al. 2008; Andersson, Refsing Andersen, et al. 2014). For this reason, methods that measure the production of nascent RNAs on chromatin, such as precision run-on and sequencing (PRO-seq) and related run-on assays, are particularly sensitive experimental methods to detect these transient enhancer-associated RNAs because they measure primary transcription before unstable RNAs are degraded by the exosome (Core et al. 2014).

We have recently introduced a novel computational method called the detection of regulatory elements using GRO-seq, PRO-seq, or ChROseq (dREG) to identify TREs de novo using PRO-seq, GRO-seq, or ChRO-seq data (Danko et al. 2015; Z. Wang et al. 2019). Most recently, we have developed a web-based portal using XSEDE servers to run dREG (Z. Wang et al. 2019). Here we provide a detailed step-by-step tutorial into how to use both the dREG web server and the downloaded dREG software. Finally, we close by providing insights into the downstream applications of these methods for discovering transcription factors responsible for a variety of biological processes.

2.8 Critical Parameters

The quality and quantity of the experimental data are major factors in determining how sensitive dREG will be in detecting TREs. We have found that dREG has a reasonable statistical power for discovering TREs with as few as ~40M uniquely mappable reads, and saturates detection of TREs in well-studied ENCODE cell lines with >75M reads (Z. Wang et al. 2019). To increase the number of reads available for TRE discovery, we typically merge biological replicates to improve our statistical power prior to running dREG.

To further improve data quality, our lab makes extensive use of unique molecular identifiers (UMIs) in RNA adapters during library prep, which allow us to identify and remove any PCR duplicates (Mahat et al. 2016; Fu et al. 2014). Typical duplication rates vary due to a variety of factors, including the quality of the input sample, the amount of starting material, and the number of cycles of PCR amplification. These experimental parameters must be considered carefully while planning a PRO-seq experiment.

2.9 Troubleshooting

The most common problems associated with running dREG can be identified by a careful examination of the input bigWig files using a genome browser (e.g., IGV, WashU, or UCSC). A genome-browser view that shows high-quality PRO-seq data is depicted in Fig. 2.9. Note that the direction of transcription resolved by PRO-seq is largely consistent with gene annotations, and gene bodies tend to have a uniform coverage of reads without excessively large gaps. Notes on identifying several common problems that are likely to be faced by users are listed below:

Poor quality PRO-seq data. Poor quality PRO-seq data is characterized by high numbers of reads at only a handful of genomic locations (Fig. 2.9). Unfortunately, this problem requires re-making new data. Troubleshooting tips for the experimental data are covered elsewhere (Mahat et al. 2016). Users are also encouraged to start with more input material and make use of UMIs in their sequencing adapters, which can help to clean up data that has been amplified for too many cycles (at the expense of sequencing depth).

Extending reads. The location of RNA polymerase in PRO-seq data is naturally represented by a single nucleotide position. dREG assumes that bigWig files will represent RNA polymerase in this manner. The solution to this problem is to remake bigWig files while representing the data using only a single position.

Using normalized counts in bigWig files. dREG assumes that input data will consist of integers (i.e., 0, 1, 2, …), and will return an error if it finds this is not the case. The solution to this problem is to remake bigWig files with raw counts.

Failure to reverse the strand. Many (but not all) PRO-seq protocols sequence from the reverse complement of the tagged RNA, and as a result reads must be reversed prior to downstream analysis. Reversed data is shown in Fig. 2.9. Note that most of the reads aligning within annotated genes is reversed relative to the annotation, and the divergent transcription and pause peak appear on the end (rather than the beginning) of each transcription unit. At the time of this writing, dREG does not detect this issue automatically. The solution to this problem is to remake bigWig files reversing the strand.

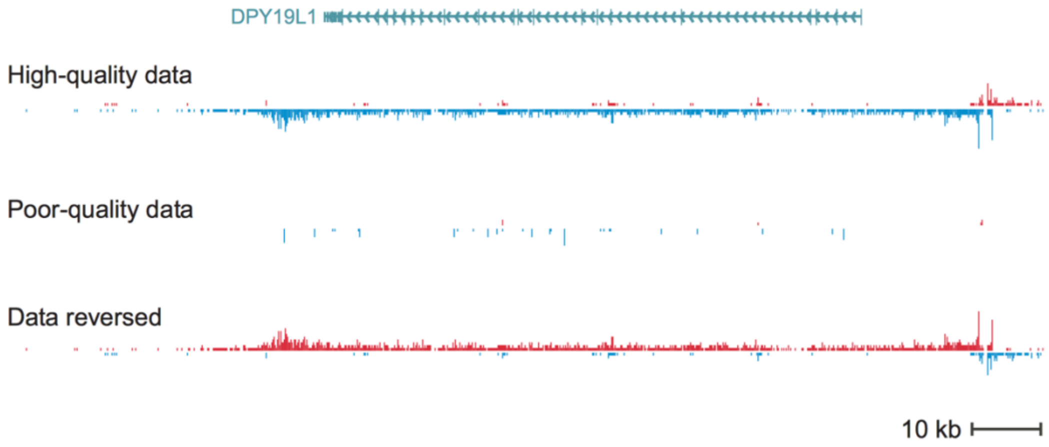


Fig. 2. 9 Genome browser shows high quality PRO-seq data (top), poor quality data (center), and data that was mapped to the reverse strand (bottom).

Table 2. 1 The GEO links to the example files used in the protocol.

Gene Expression Omnibus ID	Sample name	Link	File names
GSM2265095	Human 1 - CD4+ T-cells Untreated	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2265095	GSM2265095_H1-U_plus.bw GSM2265095_H1-U_minus.bw
GSM2265096	Human 1 - CD4+ T-cells PMA+Ionomycin	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2265096	GSM2265096_H1-PI_plus.bw GSM2265096_H1-PI_minus.bw
GSM2265098	Human 2, draw 2 - CD4+ T-cells Untreated	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2265098	GSM2265098_H2-U_plus.bw GSM2265098_H2-U_minus.bw
GSM2265097 and GSM2265099	Human 2, (merged from draw 1 and 2) - CD4+ T-cells PMA+Ionomycin	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85337	GSE85337_H2-PI_plus.bw GSE85337_H2-PI_minus.bw
GSM3021718	Human 4 - CD4+ T-cells Untreated	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3021718	GSM3021718_H4-U_plus.bw GSM3021718_H4-U_minus.bw
GSM3021719	Human 4 - CD4+ T-cells PMA+Ionomycin	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3021719	GSM3021719_H4-PI_plus.bw GSM3021719_H4-PI_minus.bw

CHAPTER 3

JOINT BAYESIAN STATISTICAL MODELING OF TUMOR AND MICROENVIRONMENT

3.1 Abstract

The complex interaction between tumor and its microenvironment is essential for the oncogenesis, survival and growth of the tumor. These interactions allow tumors to uptake nutrients from the environment and evade immune surveillance. Understanding these interactions is fundamental to the design of immunotherapies and other targeted therapies. Advances in sequencing techniques, such as RNA-seq and ATAC-seq, have enabled measurements of gene transcription and regulation across large cohorts of cancer patients down to the single cell resolution. However, single cell assays are still too cumbersome and expensive to scale to hundreds of patients necessary to understand interactions between tumor cells and their microenvironment. In this work I present statistical models called Tumor Microenvironment Deconvolution (TED) that jointly infer the regulation of tumor-specific pathways and the composition of multiple cell types in the tumor microenvironment for each patient from bulk RNA-seq/ATAC-seq data. TED shows high accuracy on both simulated and scRNA-seq glioma data, and significantly outperforms linear regression models.

3.2 Introduction

Tumor growth requires malignant cells to overcome multiple environmental pressures, including escaping detection by the immune system and hijacking body nutrient supplies to promote angiogenesis. To achieve these, tumors often must interact with their unique microenvironment, which is comprised of non-malignant environmental cell types including immune and stromal cells. Within the last decade, immune checkpoint pathways whereby tumor cells escape immune surveillance have been found, targets among which include the cytotoxic T-lymphocyte associated protein 4 (CTLA4), programmed cell death 1 (PD1) protein and the programmed death-ligand 1 (PD-L1) protein. Targeting these molecules by engineered antibodies, known as the immune checkpoint inhibitors, has shown efficacy in multiple cancer types, including melanoma, non-small cell lung cancer, liver cancer, kidney cancer and lymphoma (Wolchok 2015), suggesting a common immune-escape mechanism shared by multiple cancer types. Although current immune checkpoint inhibitors may achieve curative performance, they are often responsive in only a small subset of patients. This suggests that tumors may adopt other unknown mechanisms to escape immune surveillance. Unraveling the interactions between tumor and other cell types in the microenvironment and understanding their heterogeneity is imperative to the discovery of new druggable targets.

Transcription and epigenetic profiles contain information from all of the cell types within a sample. The decreasing cost of high throughput RNA-seq enables scalable transcriptome profiling to hundreds of cancer patients. The signal measured from these bulk tissues (bkRNA-seq) is approximately a weighted sum of the transcription of multiple cell types. Although recent advancement in single cell RNA-seq (scRNA-seq) allows the measurement of transcriptome profile in individual cells, it is still costly, generates relatively sparse signals, and require extensive sample preprocessing. As a result, the existing scRNA-seq datasets of tumor patients are limited to only a few to dozen patients. Therefore, deconvolving cell type compositions from bkRNA-seq is a promising approach to generate sufficient statistical power for understanding the interactions among different cell types.

Current methods of deconvolution fall into two categories, each has their own limitations. Noticeably, none of them can be used to explicitly model the heterogenous expression profile of tumor cells, and hence their inferred data cannot be immediately used to study the interaction between tumor and environmental cells. Most of the methods, except for a few marked below, require marker genes, either manually curated or predetermined from reference expression profiles. Selecting the marker genes is a highly arbitrary task and risks losing signal. It is unclear to what extent do different criteria of marker genes affect the deconvolution. This is particularly problematic since expression profiles from related cell types are not independent, but form tree structures in which related cell types share the bulk of their expression programs. Hence curating the marker genes down to the root will cause significant loss of signal. For example, curating the marker genes of T helper cells, T memory cells and T regulatory cells inevitably forgo the signature genes of CD4+ T cells. As a result, the curated marker genes are often at the scale of as few as several hundred - orders of magnitude sparse than the total number of annotated genes in the genome (~60K, including both coding and non-coding genes). The first type of deconvolution algorithm is known as the complete deconvolution, which explicitly relies on marker genes of each reference cell types to jointly impute the expression profile of the cell type and their proportions based on non-negative matrix factorization. Methods belong to this category include, MMAD, deconf, ssKL and ssFrobenius(Liebner, Huang, and Parvin 2014; Repsilber et al. 2010; Brunet et al. 2004; Lee and Seung 2001). These methods are not viable on real cancer data, due to the heterogeneity among patients, in which curating marker genes for tumor cells are infeasible. Moreover, the use of marker genes, referred to as the “anchors” in the language of non-negative matrix factorization assumes the expression of the gene in only the corresponding cell type and zero in other cell types, which is usually violated in real gene expression profiles.

The second group is referred to as partial deconvolution, which only attempts to infer the percentage of environmental cells. They are built upon either constrained optimization, including EPIC and quanTIseq (Racle et al. 2017; Finotello et al. 2017) which uses non-negativity and sum-to-one constraint to infer the absolute proportion, TIMER (B. Li et al. 2016) which uses non-negativity constraint to infer relative proportions, CIBERSORT (Newman et al. 2015) which uses regularized but unconstrained optimization and may yield uninterpretable negative or greater-than-one proportionalities. CIBERSORT implicitly assumes marginal independence among samples, and hence information shared among samples is not utilized. Approaches of this type are built on the (yet untested) assumption that the marker genes are only expressed in environmental cells, but zero in the tumor cells. Some genes, such as PD-L1, are expressed in both the tumor and immune cells. As a result, the estimators of these approaches are bound to overestimate the fractions of environmental cells. Other methods such as NNML and PERT (Qiao et al. 2012)assumes weights of environmental cells sum up to one, and do not take the proportion of unknown tumor fraction into consideration, and hence can only work to deconvolve the mixture composed of approximately linear combinations of environmental cells. The related methods NNMLnp (Qiao et al. 2012) is the only model, to our knowledge, that explicitly models the expression profile of tumor without the need for marker genes. However, it assumes the existence of only one type of expression profile of unknown tumor fraction across multiple tumor patients, and hence cannot capture the heterogeneity across patients. Moreover, it erroneously assumes the only tumor expression profile is a linear combination of normal cells, which is clearly violated in real tumor samples, where tumor cells often up-regulates large number of oncogenes that are not expressed in normal cells. Lastly, the majority of deconvolution algorithms, with the notable exception of NNML, NNMLnp and PERT, work on log transformed version of the depth/gene length normalized count, which cannot capture the error distribution of integer-valued count data.

3.3 Statistical Model

The statistical model of TED is based on several assumptions about the tumor and the microenvironment. Previous scRNA-seq studies have shown that tumor cells are highly heterogenous, and often clustered by patients when visualized on the tsne plot. In contrast, environmental cells, e.g. immune, stroma, endothelial and other normal cells, are largely clustered by their cell types than by patients. This observation can be explained by the heterogeneous somatic mutations accumulated in the clonal events of the tumor cells, while normal environmental cells, expect for a few genes which are indicative of immune clonal expansion, such as the T cell receptor genes, are devoid of such events. Built upon such observations, TED assumes the expression profiles of environmental cells is preserved across all tumor patients, yet allows those describing tumor cells to be estimated for each patient. Therefore, TED explicitly models the heterogeneity of the tumor cells. It also relies on the use of a complete set of gene expression profiles for the environmental cells, which can be readily derived from scRNA-seq of the whole tumor from a small number of patients. In practice, the environmental cells may also show mild variations and form subtypes due to the differences in tumor microenvironment, e.g. M1 macrophage and M2 macrophage. In case of incomplete observation, components that cannot be accounted for by the references will be absorbed by the tumor fraction and hence inflate the proportions of tumor.

The graphical model of TED is illustrated in Fig. 3.1. Let X denotes the bkRNA-seq measured for N patients on G genes. Each patient n∈{1, ..., N} is sequenced at the depth of R_n reads. The expression profiles of one reference origin of tumor cells and m environmental cells are also observed, denoted by $\varphi \in \mathbb{R}^G$ and $\psi_m \in \mathbb{R}^G$ respectively. We allow φ to undergo K types of perturbations, i.e. log of fold change, denoted by γ_k. The perturbed expression profile, concatenated with ψ, yields the overall expression profile $\Phi \in \mathbb{R}^{G*(K+M)}$. $\theta_n \in \mathbb{R}^{M+K}$denotes the proportion of each cell type in the nth sample. Z denotes the cell type from which each read is generated. α and σ are hyper-parameters of the model. All other variables are learned during model inference.

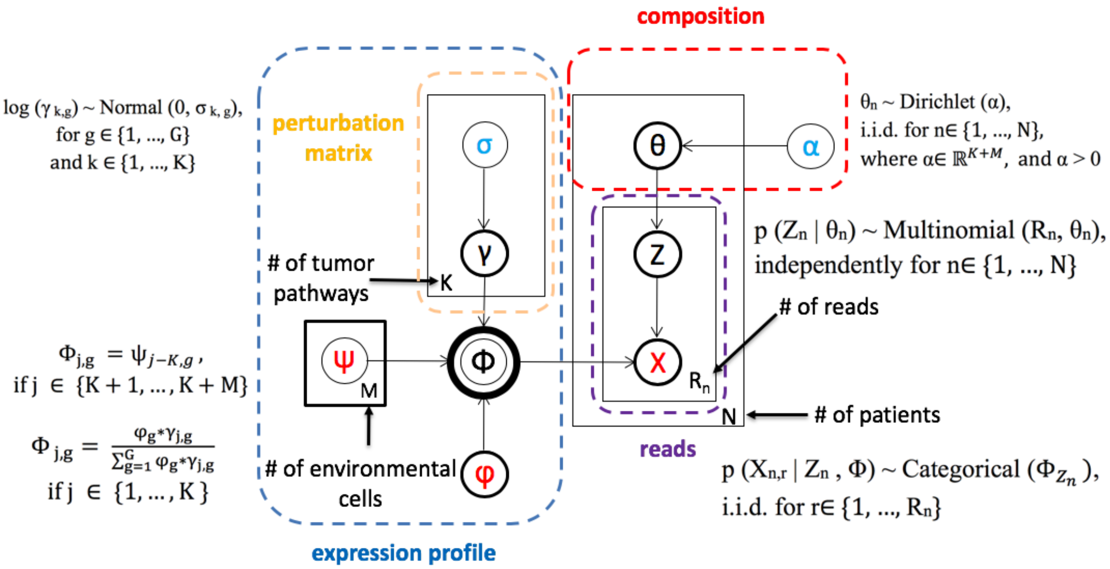

composition
$\log(\gamma_{k,g}) \sim \text{Normal}(0, \sigma_{k,g})$,
for g ∈ {1, ..., G}
and k ∈ {1, ..., K}
perturbation matrix
σ
γ
# of tumor pathways
K
θ
α
$\theta_n \sim \text{Dirichlet}(\alpha)$,
i.i.d. for n∈ {1, ..., N},
where $\alpha \in \mathbb{R}^{K+M}$, and α > 0
Z
$p(Z_n \mid \theta_n) \sim \text{Multinomial}(R_n, \theta_n)$,
independently for n∈ {1, ..., N}
# of reads
$\Phi_{j,g} = \psi_{j-K,g}$,
if j ∈ {K + 1, ..., K + M}
ψ
M
Φ
X
R_n
N
# of patients
$\Phi_{j,g} = \frac{\varphi_g * \gamma_{j,g}}{\sum_{g=1}^{G} \varphi_g * \gamma_{j,g}}$
if j ∈ {1, ..., K}
# of environmental cells
φ
reads
$p(X_{n,r} \mid Z_n, \Phi) \sim \text{Categorical}(\Phi_{Z_n})$,
i.i.d. for r∈ {1, ..., R_n}
expression profile

Fig. 3. 1 The Plate Model of TED.

Hyper-parameters are colored in blue; observed variables are colored in red; latent variables are in black. Double cycle denotes deterministic augmented variables.

The generative process of the TED is as follows.

1. Generate the full expression profiles:

$\log(\gamma_{k,g}) \sim \text{Normal}(0, \sigma_{k,g})$, for $g \in \{1, ..., G\}$ and $k \in \{1, ..., K\}$

$$\begin{cases} \Phi_{j,g} = \frac{\varphi_g * \gamma_{j,g}}{\sum_{g=1}^{G} \varphi_g * \gamma_{j,g}}, & \text{if } j \in \{1, \dots, K\} \\ \Phi_{j,g} = \psi_{j-K,g}, & \text{if } j \in \{K+1, \dots, K+M\} \end{cases}$$

2. Generate proportions for tumor pathways and environmental cells:

$\theta_n \sim \text{Dirichlet}(\alpha)$, i.i.d. for $n \in \{1, ..., N\}$, where $\alpha \in \mathbb{R}^{K+M}$. and $\alpha > 0$.

3. Generate the reads for bkRNA-seq:

$p(Z_n \mid \theta_n) \sim \text{Multinomial}(R_n, \theta_n)$, independently for $n \in \{1, ..., N\}$

$p(X_{n,r} \mid Z_n, \Phi) \sim \text{Categorical}(\Phi_{Z_n})$, i.i.d. for $r \in \{1, ..., R_n\}$

3.4 Model inference

We use a generalized Gibbs-Expectation Maximization (EM) algorithm to get point estimate of the γ by maximizing the log of posterior, marginalizing over other nuisance variables. In the E step we use Gibbs sampling to approximate posterior $p(\theta, Z \mid \gamma_{old}, \varphi, \psi, X ; \alpha)$, since computing the exact distribution is intractable. In the M step we use the conjugate gradient algorithm to numerically maximize the log of posterior $E_{p(Z \mid \gamma^{old}, X)}[\log(p(\gamma, Z \mid \varphi, \psi, X; \sigma))]$.

3.4.1 E step (Gibbs sampling):

The posterior distribution of the E step is closely related to the Latent Dirichlet Allocation model. Considering the conditional independence relations that $p(\theta, Z \mid \gamma, \varphi, \psi, X ; \alpha) = \prod_n p(\theta_n, Z_n \mid \gamma, \varphi, \psi, X_n; \alpha)$, the posterior of each patient $p(\theta_n, Z_n \mid \gamma, \varphi, \psi, X_n; \alpha)$ can be sampled in parallel by multithreading. Therefore, we use Gibbs sampling, to iteratively sample $p(\theta_n \mid Z_n, \gamma, \varphi, \psi, X_n ; \alpha)$ and $p(Z_n \mid \theta_n, \gamma,$

φ, ψ, X_n ; α). Zhu et.al. has shown that by introducing augmented variables $\widetilde{X}_{gn} = \sum_{n=1}^{R_n} I_{\{X_{r,n}=g\}}$, and $\tilde{Z}_{gn,j} = \sum_{\{r:X_{r,n}=g\}} I_{\{Z_{r,n}=j\}}$, the complexity of Gibbs sampling does not scale with the read depth, which enables efficient sampling and memory usage. Due to the conjugacy of Dirichlet and multinomial, the posterior distribution can be read off from the joint distribution, shown as below.

$p(\theta_n \mid \tilde{Z}_n, \gamma, \varphi, \psi, X_n ; \alpha) \sim \text{Dirichlet}\ (\alpha + \sum_{g=1}^{G} \tilde{Z}_{gn})$

$p(\tilde{Z}_{gn} \mid \theta_n, \gamma, \varphi, \psi, X_n ; \alpha) \sim \text{Multinomial}\ (\widetilde{X}_{gn}, \frac{\Phi_{.g} \odot \theta_n}{\sum_{j=1}^{K+M} \Phi_{j.g}\theta_n})$, where $\odot$ denotes elementwise multiplication.

Empirically, the Gibbs chain converged to the stationary distribution fairly quickly. The Gibbs samples are collected after burn-in and thinned to reduce auto-correlations.

3.4.2 M step:

In the M step, the parameter γ is updated to maximize the expectation of the log posterior, with the expectation taken with respect to the Gibbs samples drawn from the E step. Since there is no closed form solution to the posterior, we use the conjugate gradient algorithm to numerically optimize the posterior using the analytical gradient. Also, observing the conditional independence relations

Specifically, the objective function is:

$$\mathrm{E}[\ \log\ (\mathrm{p}(\gamma, \tilde{Z}\ |\varphi, \psi, X\ ;\ \sigma))]$$

$$\propto\ \mathrm{E}[\ \log\ (\mathrm{p}(\ \tilde{Z} \mid \Phi\ ))] + \mathrm{E}[\ \log\ (\mathrm{p}(\gamma \mid \sigma))]$$

$$=\ \mathrm{E}[\ \textstyle\sum_j \sum_n \log\ (\mathrm{p}(\ \tilde{Z}_{.n,j}\ \mid \Phi_j))] + \mathrm{E}[\ \sum_j \log\ (\mathrm{p}(\gamma_j \mid \sigma_j))]$$

$$\propto \textstyle\sum_s \sum_j \sum_n \sum_g \tilde{Z}_{s,gn,j}\ *\ \log(\Phi_{j,g})\ +\ \sum_j \sum_g - \frac{1}{2{\sigma_{j,g}}^2} *\ {\gamma_{j,g}}^2$$

, where s denotes the Gibbs samples.

The derivatives are:

$$\frac{\partial\, \mathrm{E}[\, \log (\mathrm{p}(\gamma,\tilde{\mathrm{Z}}\,|\varphi,\psi,\mathrm{X}\,;\sigma))]}{\partial \gamma_{j,g}} = \sum_s \sum_n \tilde{\mathrm{Z}}_{\mathrm{s,gn,j}} - \sum_s \sum_n \sum_g \tilde{\mathrm{Z}}_{\mathrm{s,gn,j}} * \Phi_{\mathrm{j,g}} + -\frac{1}{\sigma_{j,g}{}^2} * \gamma_{j,g}$$

Observe that the derivative of γ_{j^0} does not depend on other $\gamma_{j\neq j^0}$, each γ_j can be optimized in parallel, with respect to the objective function being

$$\sum_s \sum_n \sum_g \tilde{\mathrm{Z}}_{\mathrm{s,gn},j^0} * \log(\Phi_{j^0,\mathrm{g}}) + \sum_g -\frac{1}{2\sigma_{j^0,g}{}^2} * \gamma_{j^0,g}{}^2$$

3.5 Results

3.5.1 Evaluate on simulated data

In this section, we evaluated TED on data simulated from known parameters. We used bkRNA-seq collected from hemopoietic lineages to better reflect the distribution of expression profiles of real cells, which often shows bimodality after log transformation (Fig. 3.2), and the correlation structures among them.

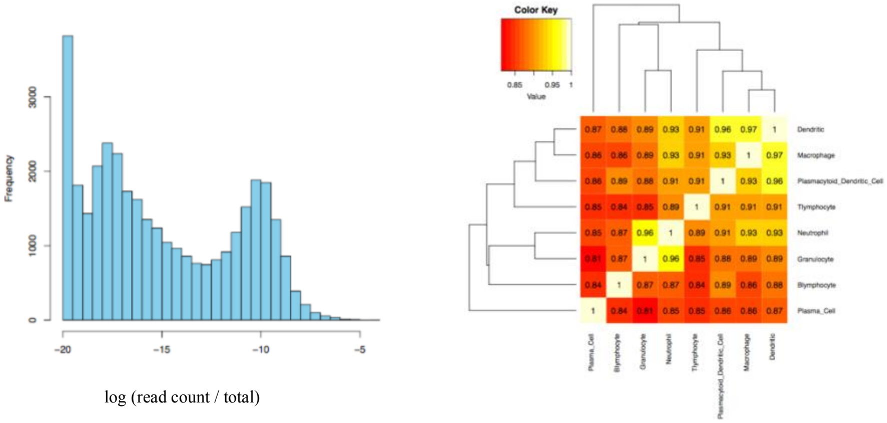


Fig. 3. 2 data distribution of bkRNA-seq data.

Left: histogram illustrates the log of depth-normalized read counts. Right: heatmap shows pairwise spearman rank correlation coefficients between the gene expression profiles in different cell types.

We set alpha=1, sigma=2, and perturbed the expression profile of B lymphocytes (the first column in the data matrix) to generate tumor profiles at K=10. In total, we simulated 50 patients, and the read depth of each patient was drawn from Poisson distribution $\lambda=10_9$. To avoid a vanishing gradient caused by extremely small value of $\Phi_{j,g}$, we resampled the distribution of $\gamma_{j,g}$ to match the range of φ. The EM converged, and successfully recovered the proportions of each cell type as well as the expression profiles of tumor pathways (Fig. 3.3).

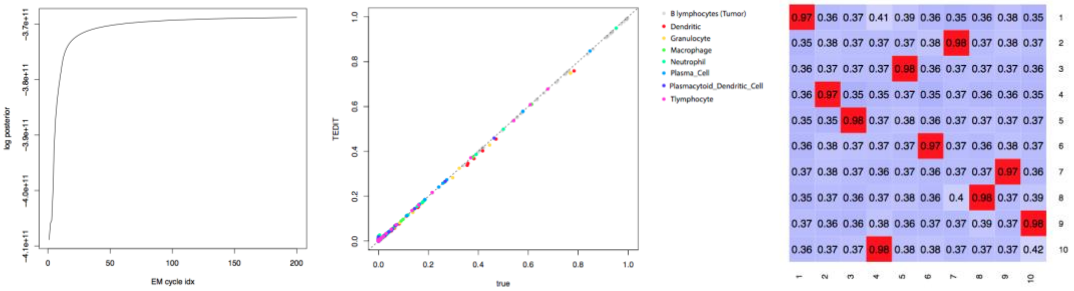


Fig. 3. 3 Performance on Simulated data.

Left: log posterior increment over the EM cycles. Middle: inferred cell types fractions verse ground truth. Right: inferred tumor pathway expression profiles (rows) verses true tumor pathway expression profiles (columns).

3.5.2 Evaluation on glioma scRNA-seq data

To validate on real tumor datasets, we evaluate the performance of TED on bkRNA-seq data simulated by adding up read counts measured using scRNA-seq data. We curated scRNA-seq data from 8 high grade glioma patients, and performed a leave-one-out test. Specifically, we used the φ and ψ derived from 7 training samples, infer over the whole set, and evaluated the performance on the holdout sample. Hyper-parameters are set at sigma=2, alpha=1, K=10. TED learned to optimize the tumor expression profile, and reached high accuracy over the holdout set (Fig. 3.3) after EM converges. TED also showed robust performance at multiple Ks, reaching a comparable performance even at K=30 (r=0.993), and does not seem to overfit the tumor component.

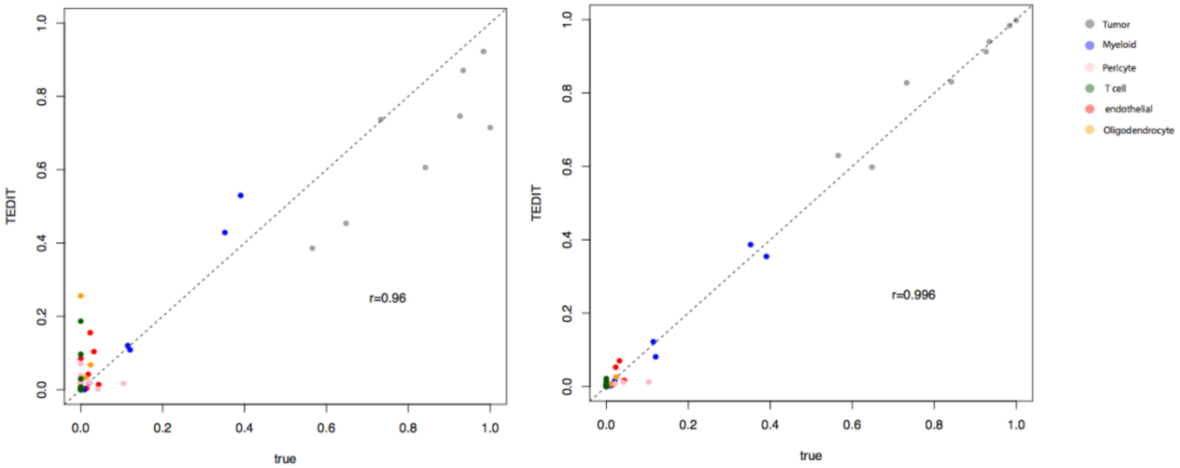
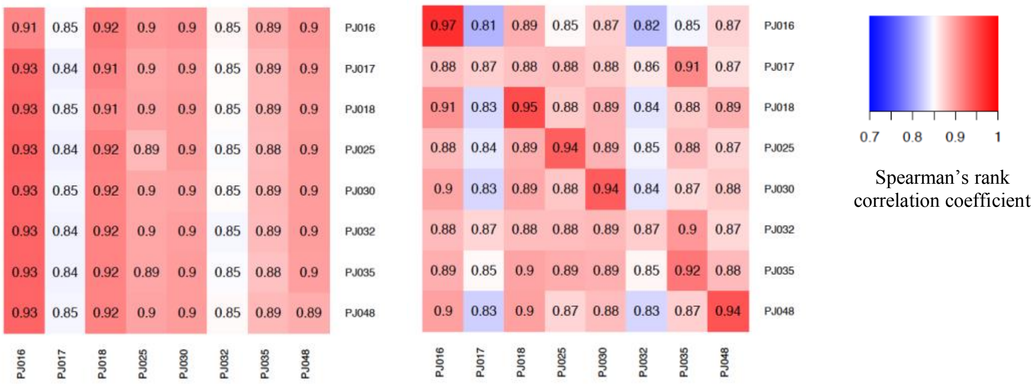


Fig. 3. 4 The expected proportion of different cell types after the 1st E step, i.e., before optimization (left), and at the convergence of EM (right).

TED also learned to accurately recover the expression profiles of the tumor cells in each patient (Fig. 3.4). TED showed interesting properties on the inferred parameter γ describing tumor pathways. At K equal or greater than the number of patients, TED still has the tendency to group patients of similar transcription rather than distributing each patient into an individual γ (Fig. 3.5). We found that two patients PJ032 and PJ017 shared the activation of the 8th tumor pathway. They are the only two patients of mesenchymal subtype which show high myeloid infiltration, and are also the closest on the tsne plot (Fig.1a of (Yuan et al. 2018)). Interestingly one tumor patient PJ016 showed activation of two correlated pathways to slightly different extent is the one shows two distinct subpopulations on the tsne plot (Fig.1a of (Yuan et al. 2018)).



	PJ016	PJ017	PJ018	PJ025	PJ030	PJ032	PJ035	PJ048
PJ016	0.91	0.85	0.92	0.9	0.9	0.85	0.89	0.9
PJ017	0.93	0.84	0.91	0.9	0.9	0.85	0.89	0.9
PJ018	0.93	0.85	0.91	0.9	0.9	0.85	0.89	0.9
PJ025	0.93	0.84	0.92	0.89	0.9	0.85	0.88	0.9
PJ030	0.93	0.85	0.92	0.9	0.9	0.85	0.89	0.9
PJ032	0.93	0.84	0.92	0.9	0.9	0.85	0.89	0.9
PJ035	0.93	0.84	0.92	0.89	0.9	0.85	0.88	0.9
PJ048	0.93	0.85	0.92	0.9	0.9	0.85	0.89	0.89

	PJ016	PJ017	PJ018	PJ025	PJ030	PJ032	PJ035	PJ048
PJ016	0.97	0.81	0.89	0.85	0.87	0.82	0.85	0.87
PJ017	0.88	0.87	0.88	0.88	0.88	0.86	0.91	0.87
PJ018	0.91	0.83	0.95	0.88	0.89	0.84	0.88	0.89
PJ025	0.88	0.84	0.89	0.94	0.89	0.85	0.88	0.87
PJ030	0.9	0.83	0.89	0.88	0.94	0.84	0.87	0.88
PJ032	0.88	0.87	0.88	0.88	0.89	0.87	0.9	0.87
PJ035	0.89	0.85	0.9	0.89	0.89	0.85	0.92	0.88
PJ048	0.9	0.83	0.9	0.87	0.88	0.83	0.87	0.94

Fig. 3. 5 Spearman correlation between the expression profiles of inferred tumor cells (rows) and ground truth (columns).

Tumor profiles are calculated by collapsing tumor cells from each patient. Left: expression profiles learned after the 1st cycle. Right: expression profiles learned at the convergence of EM.

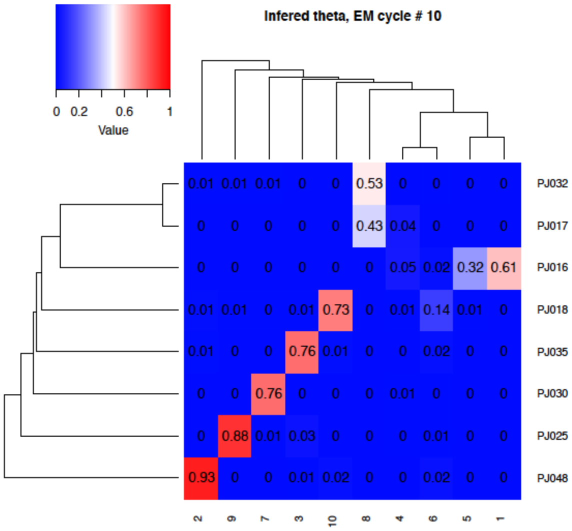


Fig. 3. 6 Heatmap shows the activation of tumor pathways in each patient at K=10.

We compared the performance of TED by benchmarking against linear models. We tried depth normalized and log depth normalized feature vectors under two scenarios, namely with or without the expression of tumor expression profile summed up from the training scRNA-seq data (Fig. 3.6). We observe that in all cases, linear regression severely underestimated the tumor fractions, while overestimate those of rare populations of environmental cells. Regression without the tumor variable has the poorest performance, due to the false assumption that tumor cells have zero expression over all gene dimensions. This is observed but often overlooked in naive regression-based methods, such as CIBERSORT and quanTIseq. Incorporating the tumor variable into the regression model significantly improves the estimates, yet still underestimates the tumor fraction but overestimates other populations. This is due to the inability for a single tumor cell reference to account for the heterogeneities among patients, especially the unobserved patient in the holdout set. Residuals resulted from the deviation from the reference tumor will be absorbed into other cell types and inflate their proportions. Applying the log transform to the feature space is essential for the linear regression approach, as the error in the RNA-seq data usually become more normally distributed after log transformation. Taken together, through actively learning the embeddings of the tumor transcription profile, TED greatly improves the inference of cell fractions.

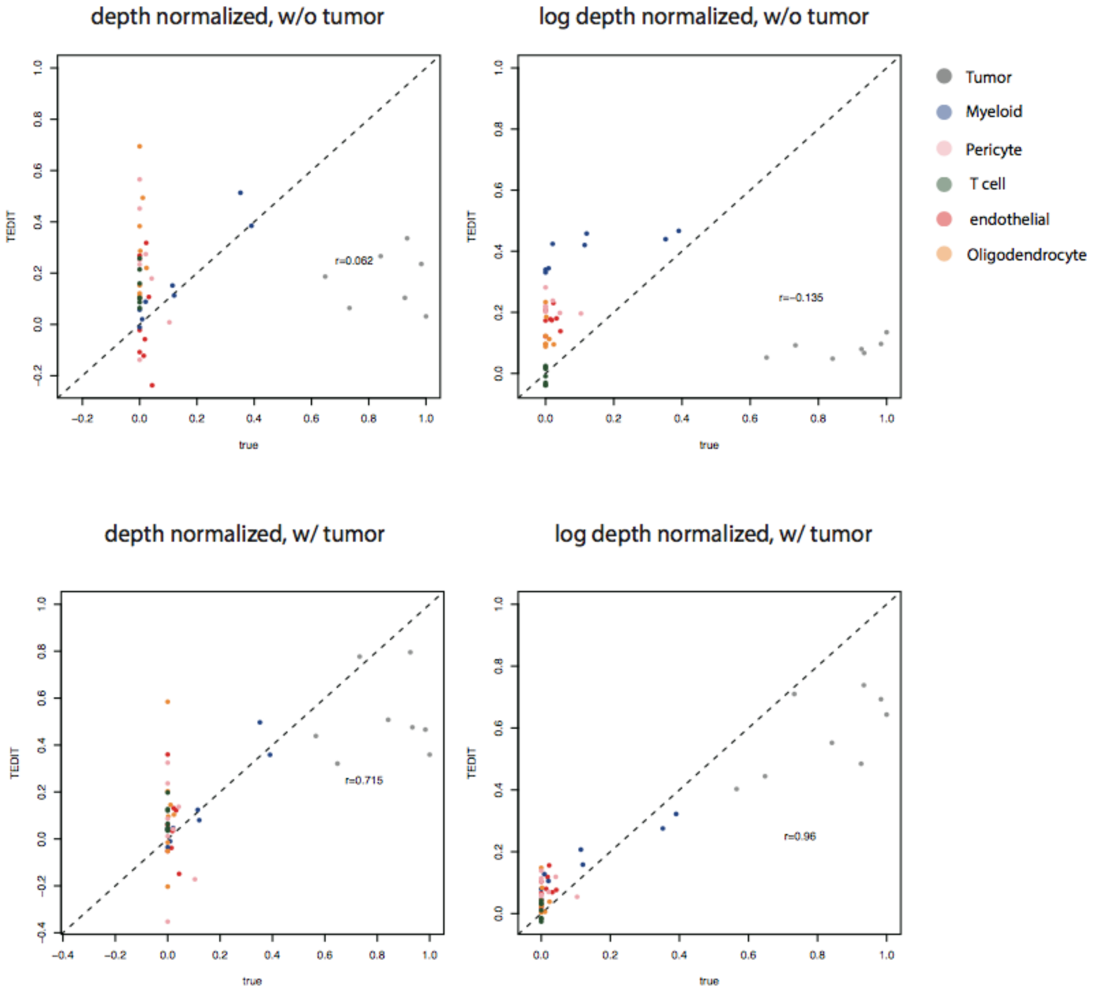


Fig. 3. 7 Cell Fractions Predicted using Linear Regression (least square fit).

3.6 Discussion

We have built the Bayesian statistical model TED which accurately jointly infers the tumor pathways and fractions of environmental cells from raw RNA-seq read counts, without the need for marker genes. TED significantly improves over the regression-based methods, where tumor faction is inevitably underestimated. Considering the enormous amount of bkRNA-seq data collected from large cohorts and across cancer types, TED can be of great potential in uncovering the covariance structure between tumor pathways and the fractions of each environmental cell. This would allow us to answer important questions such as how tumor cells evade, inhibit or even hijack the immune system. Specifically, for glioblastoma, one important, but yet unanswered question, is how cells of myeloid or lymphoid lineage infiltrate the brain tumor from blood, and what role they play in oncogenesis. Conversely, TED can also be used to infer immune expression profiles when the expression profiles of other cell types are held fixed. As bkRNA-seq data is available for large number of cohorts via GTEx, this can have important applications in studying the tissue-specific immune profiles of the tissue residential cells or other immune cells that migrate into the tissue under healthy or disease states.

It is tempting to jointly infer the expression profiles and proportions of both the tumor and environmental cells of interest. However, this may run into identifiability issues, where environmental cells may absorb gene expressed from the tumor cells. Future directions of TED can explore the use of gene-specific variance / co-variance $\sigma 2$, and cell type-specific α estimated using empirical Bayes approach from prior datasets to make the posterior identifiable. TED can incorporate the use of multiple φs, each has a unique biological meaning, e.g. potential tumor cell origins or cells collected at different developmental lineages / time series, allowing the inferred tumor pathway to have richer biological meaning.

3.7 Data Access

The scRNA glioma datasets are curated from GEO (accession ID: GSE103224). The annotations of cell types are kindly provided by Dr. Peter A. Sims through personal correspondence.

CONCLUDING PERSPECTIVE

The two chapters represent top-down and bottom-up frameworks for understanding transcription regulation in tumor and its microenvironment. They rely on functional epigenomic measurements, such as ChRO-seq and RNA-seq, from bulk tumor tissues, and leverage prior knowledge from reference cells of known cell types. Chapter one takes advantage of the highly cell type-specific TREs measured by ChRO-seq, and compared its activation profile with respect to those of reference cells measured by DNase I-hypersensitive sites. These TREs allowed us to map the cell types resembled by the tumor and its microenvironment and get a semi-quantitative enrichment score for each reference cell type. Several transcription factors enriched in the immune transcription modules showed survival association, highlighting the clinical importance of tumor microenvironment in GBM. Chapter two develops TED, a Bayesian statistical model that jointly infers the tumor expression profile and the proportion environmental cells. TED accounts for the heterogeneity of tumor cell by modeling it as a weighted sum of multiple cells with perturbed pathways. On both simulated and scRNA-seq data, TED reaches high accuracy and significantly outperform linear regression, representing a fully automated and quantitative approach for understanding the tumor cell in context of the microenvironment.

Both frameworks have advantages and disadvantages. The top-down framework is easy to comprehend and does not explicitly assume a specific generative model, yet often requires extensive manual intervention. It relies on the use of enhancers which are highly cell type-specific and is unclear to what extent it may generalize to the use of transcription level of genes. The top-down framework also analyzes one sample at a time, and hence information shared cannot be captured. The bottom-up approach is based on a specific model built according to human prior knowledge. The model implicitly extracts cell type-specific signals and does not require the need for manually curating maker genes a priori, which fully utilizes all measured signals. As a result, model training is highly automated, and performs well on transcription abundance of genes. In addition, due to the joint inference over multiple patients, estimates of parameters have the desired shrinkage effect. The main disadvantage is that the training process may be difficult to control, e.g. sensitivity to initialization and susceptibility to local minimum/maximum. The combination of these two approaches, i.e. post-hoc inspection using top-down approach for parameters inferred from modeling, represents an ideal framework for understanding the compositions of tumor and the microenvironment. In addition, TED can be easily modified to allow joint inference using matched transcription and regulatory signals.

The bottleneck of TED is its assumption that environmental cells do not show heterogeneity across patients, which may be violated due to tumor microenvironment interactions. Whereas TED accurately recovers the transcription profile of tumor cells and the proportions of environmental cells, it cannot infer unknown cell types of the environmental cells. In case of the presence of unknown cell types or cell types that deviate from the reference component, TED will absorb their read count into the tumor component and may underestimate the proportions of the reference component. Future directions that incorporate the use of gene-specific hyperparameters may allow joint inference of expression profiles of environmental cells of interest. The fundamental structure of TED separates the modeling of gene regulation and cell type proportions into individual sub-models, allowing more engineering flexibility. For example, a multivariable Gaussian distribution, or even non-parametric Bayes approaches, can be used to model the prior distribution of differential regulation in each cell type, and logistic normal distribution can be used to directly model the covariance structures between proportions of each cell type. Taken together, TED represents a groundbreaking framework that automatically leverages prior knowledges to study tumor and its microenvironment.

BIBLIOGRAPHY

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. "An Atlas of Active Enhancers across Human Cell Types and Tissues." *Nature*. https://doi.org/10.1038/nature12787.

Andersson, Robin, Peter Refsing Andersen, Eivind Valen, Leighton J. Core, Jette Bornholdt, Mette Boyd, Torben Heick Jensen, and Albin Sandelin. 2014. "Nuclear Stability and Transcriptional Directionality Separate Functionally Distinct RNA Species." *Nature Communications*. https://doi.org/10.1038/ncomms6336.

Azofeifa, Joseph G., Mary A. Allen, Josephina R. Hendrix, Timothy Read, Jonathan D. Rubin, and Robin D. Dowell. 2018. "Enhancer RNA Profiling Predicts Transcription Factor Activity." *Genome Research*. https://doi.org/10.1101/gr.225755.117.

Azofeifa, Joseph G., and Robin D. Dowell. 2017. "A Generative Model for the Behavior of RNA Polymerase." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw599.

Bhat, Krishna P.L., Veerakumar Balasubramaniyan, Brian Vaillant, Ravesanker Ezhilarasan, Karlijn Hummelink, Faith Hollingsworth, Khalida Wani, et al. 2013. "Mesenchymal Differentiation Mediated by NF-KB Promotes Radiation Resistance in Glioblastoma." *Cancer Cell*. https://doi.org/10.1016/j.ccr.2013.08.001.

Blumberg, Amit, Edward J. Rice, Anshul Kundaje, Charles G. Danko, and Dan Mishmar. 2017. "Initiation of MtDNA Transcription Is Followed by Pausing, and Diverges across Human Cell Types and during Evolution." *Genome Research*. https://doi.org/10.1101/gr.209924.116.

Bradner, James E., Denes Hnisz, and Richard A. Young. 2017. "Transcriptional Addiction in Cancer." *Cell*. https://doi.org/10.1016/j.cell.2016.12.013.

Brennan, Cameron W., Roel G.W. Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R. Salama, Siyuan Zheng, et al. 2013. "The Somatic Genomic

Landscape of Glioblastoma.” *Cell*. https://doi.org/10.1016/j.cell.2013.09.034.

Brunet, J.-P., P. Tamayo, T. R. Golub, and J. P. Mesirov. 2004. “Metagenes and Molecular Pattern Discovery Using Matrix Factorization.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0308531101.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” *Nature Methods*. https://doi.org/10.1038/nmeth.2688.

Cai, H., and D. S. Luse. 1987. “Transcription Initiation by RNA Polymerase II in Vitro. Properties of Preinitiation, Initiation, and Elongation Complexes.” *Journal of Biological Chemistry*.

Canute, Gregory W., Sharon L. Longo, John A. Longo, Michèle M. Shetler, Thomas E. Coyle, Jeffrey A. Winfield, and Peter J. Hahn. 1998. “The Hydroxyurea-Induced Loss of Double-Minute Chromosomes Containing Amplified Epidermal Growth Factor Receptor Genes Reduces the Tumorigenicity and Growth of Human Glioblastoma Multiforme.” *Neurosurgery*. https://doi.org/10.1097/00006123-199803000-00031.

Carro, Maria Stella, Wei Keat Lim, Mariano Javier Alvarez, Robert J. Bollo, Xudong Zhao, Evan Y. Snyder, Erik P. Sulman, et al. 2010. “The Transcriptional Network for Mesenchymal Transformation of Brain Tumours.” *Nature*. https://doi.org/10.1038/nature08712.

Chen, Vincent B., W. Bryan Arendall, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson, and David C. Richardson. 2010. “MolProbity: All-Atom Structure Validation for Macromolecular Crystallography.” *Acta Crystallographica Section D: Biological Crystallography* 66 (1): 12–21. https://doi.org/10.1107/S0907444909042073.

Cheng, Jill, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, et al. 2005. “Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution.” *Science*. https://doi.org/10.1126/science.1108625.

Choder, M., and Y. Aloni. 1988. “RNA Polymerase II Allows Unwinding and Rewinding of the DNA and Thus Maintains a Constant Length of the Transcription Bubble.” *Journal of Biological Chemistry*.

Chu, Tinyi, Edward J. Rice, Gregory T. Booth, H. Hans Salamanca, Zhong Wang, Leighton J. Core, Sharon L. Longo, et al. 2018. “Chromatin Run-on and Sequencing Maps the Transcriptional Regulatory Landscape of Glioblastoma Multiforme.” *Nature Genetics*. https://doi.org/10.1038/s41588-018-0244-3.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. “Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses.” *Science*. https://doi.org/10.1126/science.aad5497.

Churchman, L. Stirling, and Jonathan S. Weissman. 2011. “Nascent Transcript Sequencing Visualizes Transcription at Nucleotide Resolution.” *Nature*. https://doi.org/10.1038/nature09652.

Core, Leighton J., André L. Martins, Charles G. Danko, Colin T. Waters, Adam Siepel, and John T. Lis. 2014. “Analysis of Nascent RNA Identifies a Unified Architecture of Initiation Regions at Mammalian Promoters and Enhancers.” *Nature Genetics*. https://doi.org/10.1038/ng.3142.

Core, Leighton J., Joshua J. Waterfall, Daniel A. Gilchrist, David C. Fargo, Hojoong Kwak, Karen Adelman, and John T. Lis. 2012. “Defining the Status of RNA Polymerase at Promoters.” *Cell Reports*. https://doi.org/10.1016/j.celrep.2012.08.034.

Core, Leighton J., Joshua J. Waterfall, and John T. Lis. 2008. “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.” *Science*. https://doi.org/10.1126/science.1162228.

Danko, Charles G., Nasun Hah, Xin Luo, André L. Martins, Leighton Core, John T. Lis, Adam Siepel, and W. Lee Kraus. 2013. “Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2013.02.015.

Danko, Charles G., Stephanie L. Hyland, Leighton J. Core, Andre L. Martins, Colin T. Waters, Hyung Won Lee, Vivian G. Cheung, W. Lee Kraus, John T. Lis, and

Adam Siepel. 2015. “Identification of Active Transcriptional Regulatory Elements from GRO-Seq Data.” *Nature Methods*. https://doi.org/10.1038/nmeth.3329.

Danko, Charles G, Lauren A Choate, Brooke A Marks, Edward J Rice, Zhong Wang, Tinyi Chu, Andre L Martins, et al. 2018. “Dynamic Evolution of Regulatory Element Ensembles in Primate CD4+ T Cells.” *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-017-0447-5.

Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, et al. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature*. https://doi.org/10.1038/nature11247.

Eller, Jorge L., Sharon L. Longo, Daniel J. Hicklin, Gregory W. Canute, Franz Buchegger, Yves M. Dupertuis, Nicolas De Tribolet, et al. 2002. “Activity of Anti-Epidermal Growth Factor Receptor Monoclonal Antibody C225 against Glioblastoma Multiforme.” *Neurosurgery*. https://doi.org/10.1097/00006123-200210000-00028.

Finotello, Francesca, Clemens Mayer, Christina Plattner, Gerhard Laschober, Dietmar Rieder, Hubert Hackl, Anne Krogsdam, et al. 2017. “QuanTIseq: Quantifying Immune Contexture of Human Tumors.” *BioRxiv*. https://doi.org/10.1101/223180.

Fu, Glenn K., Weihong Xu, Julie Wilhelmy, Michael N. Mindrinos, Ronald W. Davis, Wenzhong Xiao, and Stephen P. A. Fodor. 2014. “Molecular Indexing Enables Quantitative Targeted RNA Sequencing and Reveals Poor Efficiencies in Standard Library Preparations.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1323732111.

Fuda, Nicholas J., M. Behfar Ardehali, and John T. Lis. 2009. “Defining Mechanisms That Regulate RNA Polymerase II Transcription in Vivo.” *Nature*. https://doi.org/10.1038/nature08449.

Hah, Nasun, Charles G. Danko, Leighton Core, Joshua J. Waterfall, Adam Siepel, John T. Lis, and W. Lee Kraus. 2011. “A Rapid, Extensive, and Transient

Transcriptional Response to Estrogen Signaling in Breast Cancer Cells." *Cell*. https://doi.org/10.1016/j.cell.2011.03.042.

Hastie, Trevor, Rahul Mazumder, Jason D Lee, and Reza Zadeh. 2015. "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares." *The Journal of Machine Learning Research* 16 (1): 3367–3402.

John, Sam, Peter J. Sabo, Robert E. Thurman, Myong Hee Sung, Simon C. Biddie, Thomas A. Johnson, Gordon L. Hager, and John A. Stamatoyannopoulos. 2011. "Chromatin Accessibility Pre-Determines Glucocorticoid Receptor Binding Patterns." *Nature Genetics*. https://doi.org/10.1038/ng.759.

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. "DNA-Binding Specificities of Human Transcription Factors." *Cell*. https://doi.org/10.1016/j.cell.2012.12.009.

Khodor, Yevgenia L., Joseph Rodriguez, Katharine C. Abruzzi, Chih Hang Anthony Tang, Michael T. Marr, and Michael Rosbash. 2011. "Nascent-Seq Indicates Widespread Cotranscriptional Pre-MRNA Splicing in Drosophila." *Genes and Development*. https://doi.org/10.1101/gad.178962.111.

Kim, Tae Kyung, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, et al. 2010. "Widespread Transcription at Neuronal Activity-Regulated Enhancers." *Nature*. https://doi.org/10.1038/nature09033.

Kolde, Raivo. 2015. "Pheatmap : Pretty Heatmaps." *R Package Version 1.0.8*.

Kuhn, Robert M., David Haussler, and W. James Kent. 2013. "The UCSC Genome Browser and Associated Tools." *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbs038.

Kwak, Hojoong, Nicholas J. Fuda, Leighton J. Core, and John T. Lis. 2013. "Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing." *Science*. https://doi.org/10.1126/science.1229386.

Lee, Dd, and Hs Seung. 2001. "Algorithms for Non-Negative Matrix Factorization." *Advances in Neural Information Processing Systems*. https://doi.org/10.1109/IJCNN.2008.4634046.

Li, Bo, Eric Severson, Jean Christophe Pignon, Haoquan Zhao, Taiwen Li, Jesse

Novak, Peng Jiang, et al. 2016. “Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy.” *Genome Biology*. https://doi.org/10.1186/s13059-016-1028-7.

Li, Heng, and Richard Durbin. 2010. “Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btp698.

Liebner, David A., Kun Huang, and Jeffrey D. Parvin. 2014. “MMAD: Microarray Microdissection with Analysis of Differences Is a Computational Tool for Deconvoluting Cell Type-Specific Contributions from Tissue Samples.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt566.

Liu, S. John, Max A. Horlbeck, Seung Woo Cho, Harjus S. Birk, Martina Malatesta, Daniel He, Frank J. Attenello, et al. 2017. “CRISPRi-Based Genome-Scale Identification of Functional Long Noncoding RNA Loci in Human Cells.” *Science*. https://doi.org/10.1126/science.aah7111.

Long, Hannah K., Sara L. Prescott, and Joanna Wysocka. 2016. “Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution.” *Cell*. https://doi.org/10.1016/j.cell.2016.09.018.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology*. https://doi.org/10.1186/s13059-014-0550-8.

Luo, Xin, Minho Chae, Raga Krishnakumar, Charles G. Danko, and W. L. Kraus. 2014. “Dynamic Reorganization of the AC16 Cardiomyocyte Transcriptome in Response to TNFα Signaling Revealed by Integrated Genomic Analyses.” *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-155.

Ma, Yawen, Ping Wang, Yixue Xue, Chengbin Qu, Jian Zheng, Xiaobai Liu, Jun Ma, and Yunhui Liu. 2017. “PVT1 Affects Growth of Glioma Microvascular Endothelial Cells by Negatively Regulating MiR-186.” *Tumor Biology*. https://doi.org/10.1177/1010428317694326.

Mahat, Dig Bijay, Hojoong Kwak, Gregory T. Booth, Iris H. Jonkers, Charles G. Danko, Ravi K. Patel, Colin T. Waters, Katie Munson, Leighton J. Core, and

John T. Lis. 2016. “Base-Pair-Resolution Genome-Wide Mapping of Active RNA Polymerases Using Precision Nuclear Run-on (PRO-Seq).” *Nature Protocols*. https://doi.org/10.1038/nprot.2016.086.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal*. https://doi.org/10.14806/ej.17.1.200.

Mayer, Andreas, Julia Di Iulio, Seth Maleri, Umut Eser, Jeff Vierstra, Alex Reynolds, Richard Sandstrom, John A. Stamatoyannopoulos, and L. Stirling Churchman. 2015. “Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution.” *Cell*. https://doi.org/10.1016/j.cell.2015.03.010.

Melgar, Michael F., Francis S. Collins, and Praveen Sethupathy. 2011. “Discovery of Active Enhancers through Bidirectional Expression of Short Transcripts.” *Genome Biology*. https://doi.org/10.1186/gb-2011-12-11-r113.

Menet, Jerome S., Joseph Rodriguez, Katharine C. Abruzzi, and Michael Rosbash. 2012. “Nascent-Seq Reveals Novel Features of Mouse Circadian Transcriptional Regulation.” *ELife*. https://doi.org/10.7554/eLife.00011.

Mohan, Man, Chengqi Lin, Erin Guest, and Ali Shilatifard. 2010. “Licensed to Elongate: A Molecular Mechanism for MLL-Based Leukaemogenesis.” *Nature Reviews Cancer*. https://doi.org/10.1038/nrc2915.

Newman, Aaron M., Chih Long Liu, Michael R. Green, Andrew J. Gentles, Weiguo Feng, Yue Xu, Chuong D. Hoang, Maximilian Diehn, and Ash A. Alizadeh. 2015. “Robust Enumeration of Cell Subsets from Tissue Expression Profiles.” *Nature Methods*. https://doi.org/10.1038/nmeth.3337.

Nojima, Takayuki, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J. Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J. Proudfoot. 2015. “Mammalian NET-Seq Reveals Genome-Wide Nascent Transcription Coupled to RNA Processing.” *Cell*. https://doi.org/10.1016/j.cell.2015.03.027.

Parsons, D. Williams, Siân Jones, Xiaosong Zhang, Jimmy Cheng Ho Lin, Rebecca J. Leary, Philipp Angenendt, Parminder Mankoo, et al. 2008. “An Integrated Genomic Analysis of Human Glioblastoma Multiforme.” *Science*.

https://doi.org/10.1126/science.1164382.

Patel, Anoop P., Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, et al. 2014. “Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma.” *Science*. https://doi.org/10.1126/science.1254257.

Preker, Pascal, Jesper Nielsen, Susanne Kammler, Søren Lykke-Andersen, Marianne S. Christensen, Christophe K. Mapendano, Mikkel H. Schierup, and Torben Heick Jensen. 2008. “RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters.” *Science*. https://doi.org/10.1126/science.1164096.

Qiao, Wenlian, Gerald Quon, Elizabeth Csaszar, Mei Yu, Quaid Morris, and Peter W. Zandstra. 2012. “PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions.” *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002838.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq033.

Quinodoz, Sofia, and Mitchell Guttman. 2014. “Long Noncoding RNAs: An Emerging Link between Gene Regulation and Nuclear Organization.” *Trends in Cell Biology*. https://doi.org/10.1016/j.tcb.2014.08.009.

R Development Core Team, R. 2011. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7.

Racle, Julien, Kaat de Jonge, Petra Baumgaertner, Daniel E. Speiser, and David Gfeller. 2017. “Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data.” *ELife*. https://doi.org/10.7554/eLife.26476.

Repsilber, Dirk, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F. Black, Joachim Selbig, Shreemanta K. Parida, Stefan H E Kaufmann, and Marc Jacobsen. 2010. “Biomarker Discovery in Heterogeneous Tissue Samples -Taking the in-Silico

Deconfounding Approach." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-11-27.

Ricci-Vitiani, L., R. Pallini, L. M. Larocca, D. G. Lombardi, M. Signore, F. Pierconti, G. Petrucci, N. Montano, G. Maira, and R. De Maria. 2008. "Mesenchymal Differentiation of Glioblastoma Stem Cells." *Cell Death and Differentiation*. https://doi.org/10.1038/cdd.2008.72.

Ricci-Vitiani, Lucia, Roberto Pallini, Mauro Biffoni, Matilde Todaro, Gloria Invernici, Tonia Cenci, Giulio Maira, et al. 2010. "Tumour Vascularization via Endothelial Differentiation of Glioblastoma Stem-like Cells." *Nature*. https://doi.org/10.1038/nature09557.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. "Integrative Analysis of 111 Reference Human Epigenomes." *Nature*. https://doi.org/10.1038/nature14248.

Santa, Francesca de, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia Lin Wei, and Gioacchino Natoli. 2010. "A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers." *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1000384.

Schmieder, Robert, and Robert Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr026.

Schwalb, Björn, Margaux Michel, Benedikt Zacher, Katja Frü Hauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. 2016. "TT-Seq Maps the Human Transient Transcriptome." *Science*. https://doi.org/10.1126/science.aad9841.

Scruggs, Benjamin S., Daniel A. Gilchrist, Sergei Nechaev, Ginger W. Muse, Adam Burkholder, David C. Fargo, and Karen Adelman. 2015. "Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin." *Molecular Cell*.

https://doi.org/10.1016/j.molcel.2015.04.006.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. “Transcriptional Enhancers: From Properties to Genome-Wide Predictions.” *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3682.

Stergachis, Andrew B., Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L. Paige, Benjamin Vernot, et al. 2013. “Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes.” *Cell*. https://doi.org/10.1016/j.cell.2013.07.020.

Suvà, Mario L., Esther Rheinbay, Shawn M. Gillespie, Anoop P. Patel, Hiroaki Wakimoto, Samuel D. Rabkin, Nicolo Riggi, et al. 2014. “Reconstructing and Reprogramming the Tumor-Propagating Potential of Glioblastoma Stem-like Cells.” *Cell*. https://doi.org/10.1016/j.cell.2014.02.030.

Tentler, John J., Aik Choon Tan, Colin D. Weekes, Antonio Jimeno, Stephen Leong, Todd M. Pitts, John J. Arcaroli, Wells A. Messersmith, and S. Gail Eckhardt. 2012. “Patient-Derived Tumour Xenografts as Models for Oncology Drug Development.” *Nature Reviews Clinical Oncology*. https://doi.org/10.1038/nrclinonc.2012.61.

Ulitsky, Igor, and David P. Bartel. 2013. “XLincRNAs: Genomics, Evolution, and Mechanisms.” *Cell*. https://doi.org/10.1016/j.cell.2013.06.020.

Verhaak, Roel G.W., Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, et al. 2010. “Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1.” *Cancer Cell*. https://doi.org/10.1016/j.ccr.2009.12.020.

Wang, Minghui, Yongzhong Zhao, and Bin Zhang. 2015. “Efficient Test and Visualization of Multi-Set Intersections.” *Scientific Reports*. https://doi.org/10.1038/srep16923.

Wang, Qianghu, Baoli Hu, Xin Hu, Hoon Kim, Massimo Squatrito, Lisa Scarpace, Ana C. deCarvalho, et al. 2017. “Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the

Microenvironment.” *Cancer Cell*. https://doi.org/10.1016/j.ccell.2017.06.003.

Wang, Zhong, Tinyi Chu, Lauren A. Choate, and Charles G. Danko. 2019. “Identification of Regulatory Elements from Nascent Transcription Using DREG.” *Genome Research*. https://doi.org/10.1101/gr.238279.118.

Wang, Zhong, André L. Martins, and Charles G. Danko. 2016. “RTFBSDB: An Integrated Framework for Transcription Factor Binding Site Analysis.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw338.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. “Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.” *Cell*. https://doi.org/10.1016/j.cell.2014.08.009.

Wolchok, Jedd D. 2015. “PD-1 Blockers.” *Cell*. https://doi.org/10.1016/j.cell.2015.07.045.

Wuarin, J, and U Schibler. 1994. “Physical Isolation of Nascent RNA Chains Transcribed by RNA Polymerase II: Evidence for Cotranscriptional Splicing.” *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.14.11.7219.

Xi, Zhuo, Ping Wang, Yixue Xue, Chao Shang, Xiaobai Liu, Jun Ma, Zhiqing Li, Zhen Li, Min Bao, and Yunhui Liu. 2017. “Overexpression of MiR-29a Reduces the Oncogenic Properties of Glioblastoma Stem Cells by Downregulating Quaking Gene Isoform 6.” *Oncotarget*. https://doi.org/10.18632/oncotarget.15327.

Yuan, Jinzhou, Hanna Mendes Levitin, Veronique Frattini, Erin C. Bush, Deborah M. Boyett, Jorge Samanamud, Michele Ceccarelli, et al. 2018. “Single-Cell Transcriptome Analysis of Lineage Diversity in High-Grade Glioma.” *Genome Medicine*. https://doi.org/10.1186/s13073-018-0567-9.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nussbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology*. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhao, Dan, Xiaochun Jiang, Chengyun Yao, Li Zhang, Huixiang Liu, Hongping Xia, and Yongsheng Wang. 2014. “Heat Shock Protein 47 Regulated by MiR-29a to

Enhance Glioma Tumor Growth and Invasion." *Journal of Neuro-Oncology*. https://doi.org/10.1007/s11060-014-1412-7.

Zhou, Xin, Brett Maricque, Mingchao Xie, Daofeng Li, Vasavi Sundaram, Eric A. Martin, Brian C. Koebbe, et al. 2011. "The Human Epigenome Browser at Washington University." *Nature Methods*. https://doi.org/10.1038/nmeth.1772.