Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme

Tinyi Chu, Edward J. Rice, Gregory T. Booth, H. Hans Salamanca, Zhong Wang, Leighton J. Core, Sharon L Longo, Robert J. Corona, Lawrence S. Chin, John T. Lis, Hojoong Kwak, and Charles G. Danko

Supplementary Note 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 decreases 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. 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 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. 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. 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. 2b**). Transcription in the core was situated between the other three parts of the tumor in a principal component analysis (PCA) (**Supplementary Fig. 5**), consistent with a model in which the tumor originated within the core and grew outward radially.

Supplementary Note 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. 3d, Supplementary Fig. 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 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 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 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. 5c; Supplementary Fig. 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. 22**). Moreover these changes in the nearest two genes explained many of the markers defined in microarray studies(Verhaak et al. 2010) (**Supplementary Fig. 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 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 4**). We fit an epsilon-support vector regression model using the Rgtsvm R package(Z. Wang et al. 2017).

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. 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.

λ=({number of bp in dREG peak} / {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 @ 0.17 FDR), and one stringent condition (knots ratio=1350 and background threshold=0.026) that optimizes for low FDR (sensitivity=0.79 @ 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 ref(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(ENCODE Project Consortium 2012).

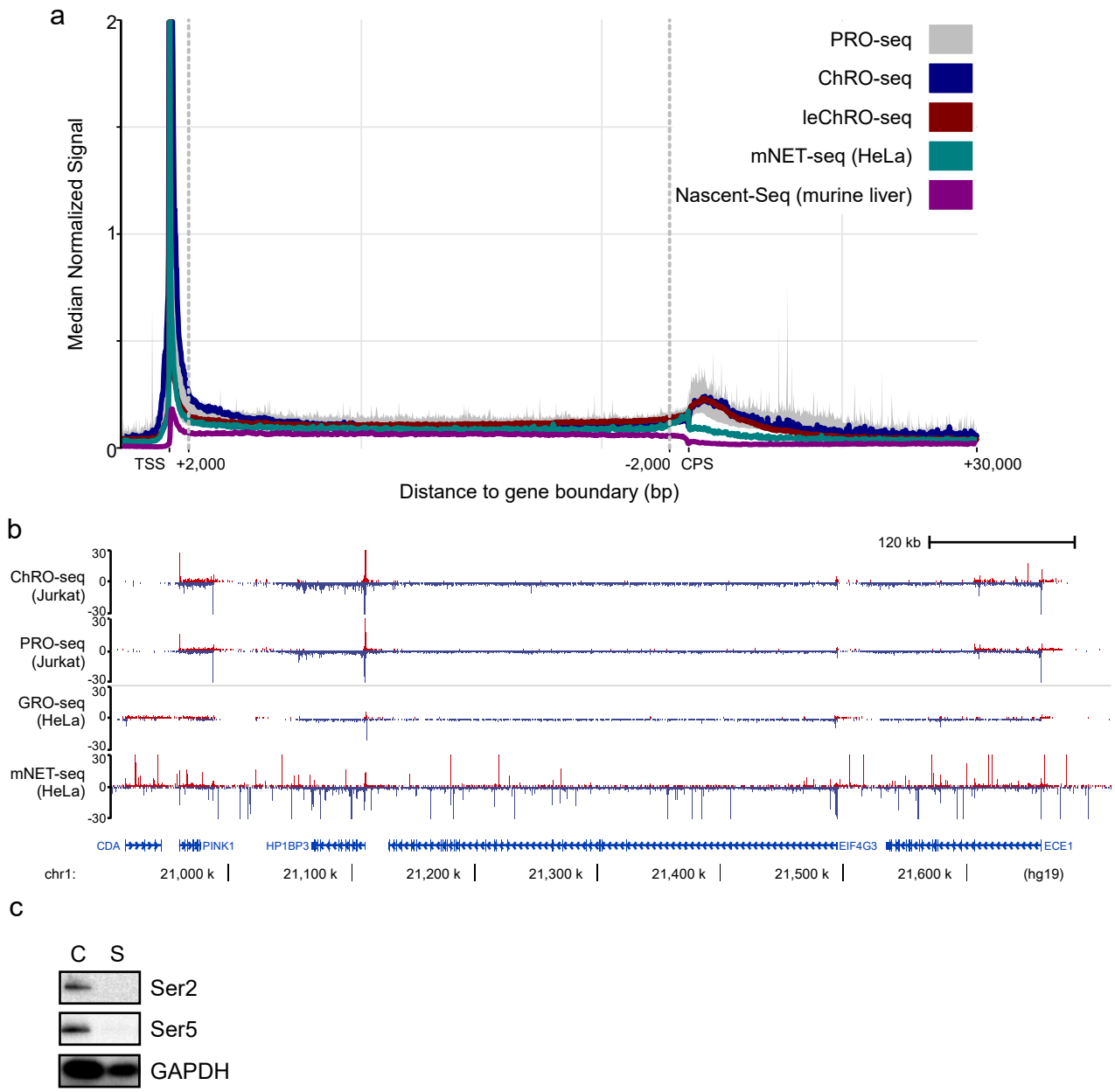
Supplementary Note 7: Description of the dREG-HD method

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.

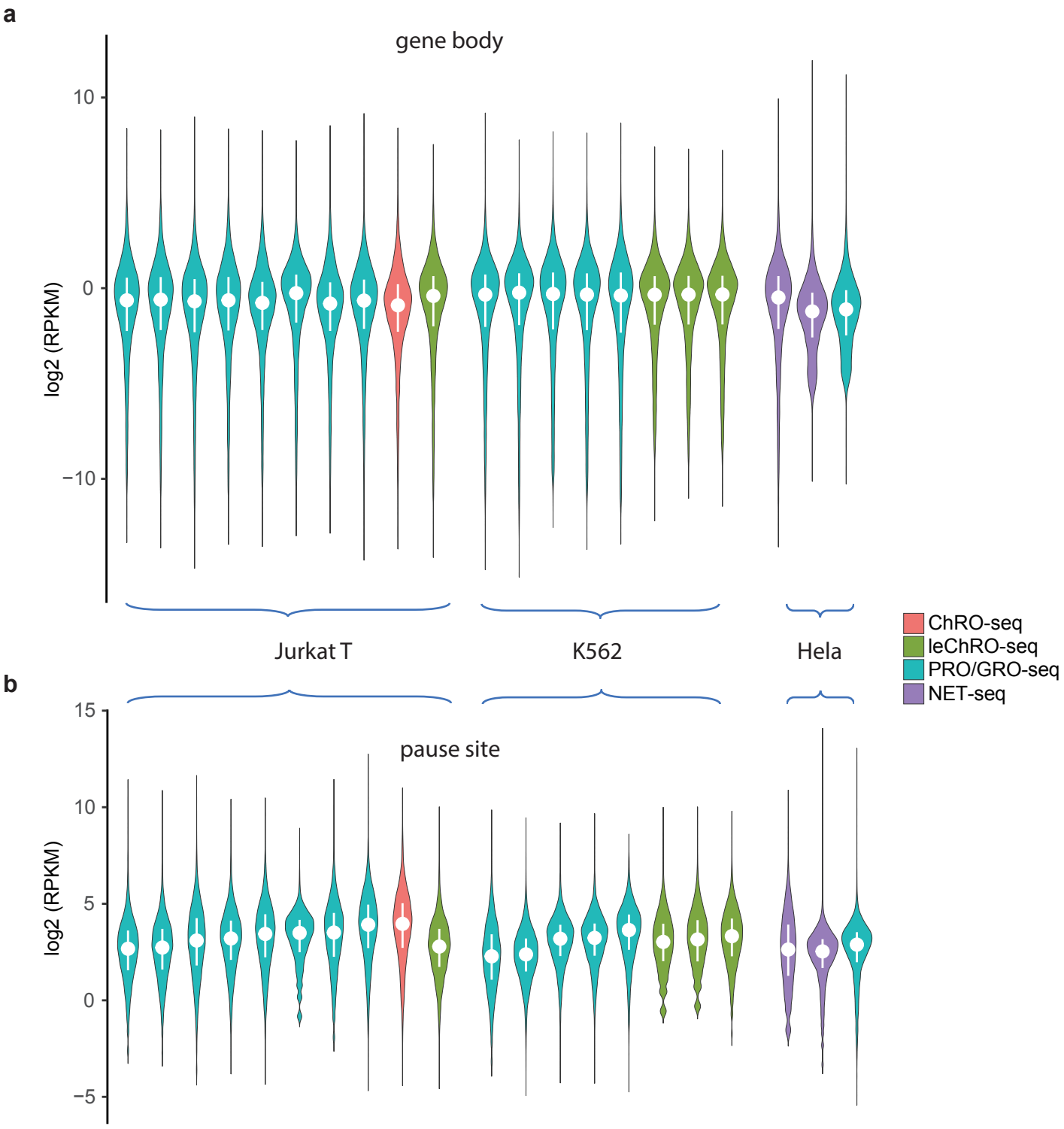
Among multiple samples normalizing the mutual information metric is more complicated. We devised an approach that was used in **Supplementary Fig. 14**. We consider a square matrix with rows and columns representing each sample. Each element in this matrix represents the mutual information between a 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. 2014), 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
```

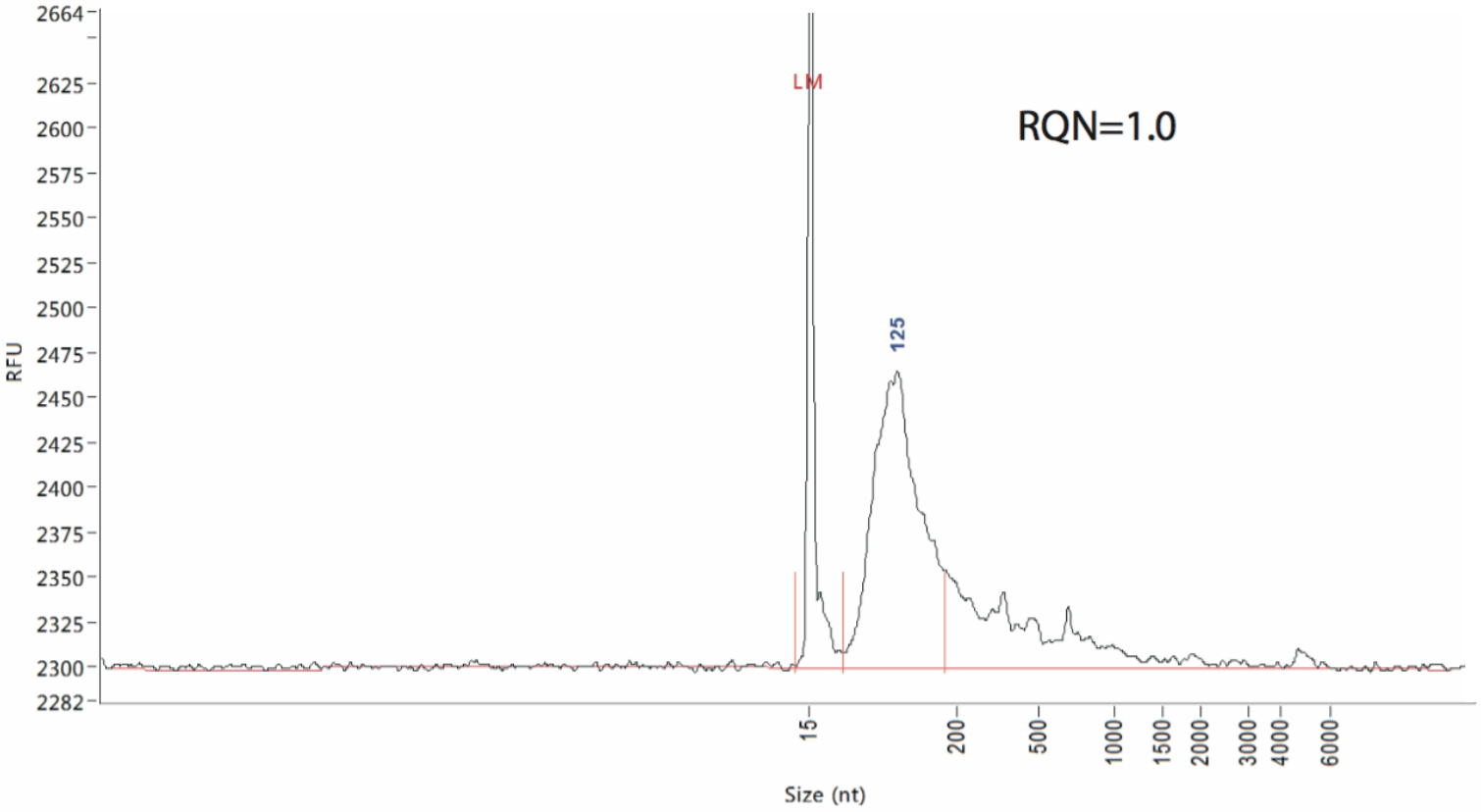
The convergence criterion was defined as the maximum of the absolute value of element-wise difference between matrix returned from previous two consecutive runs. Although there is no mathematical guarantee of convergence, this approach converged typically after four cycles with the datasets that we used.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



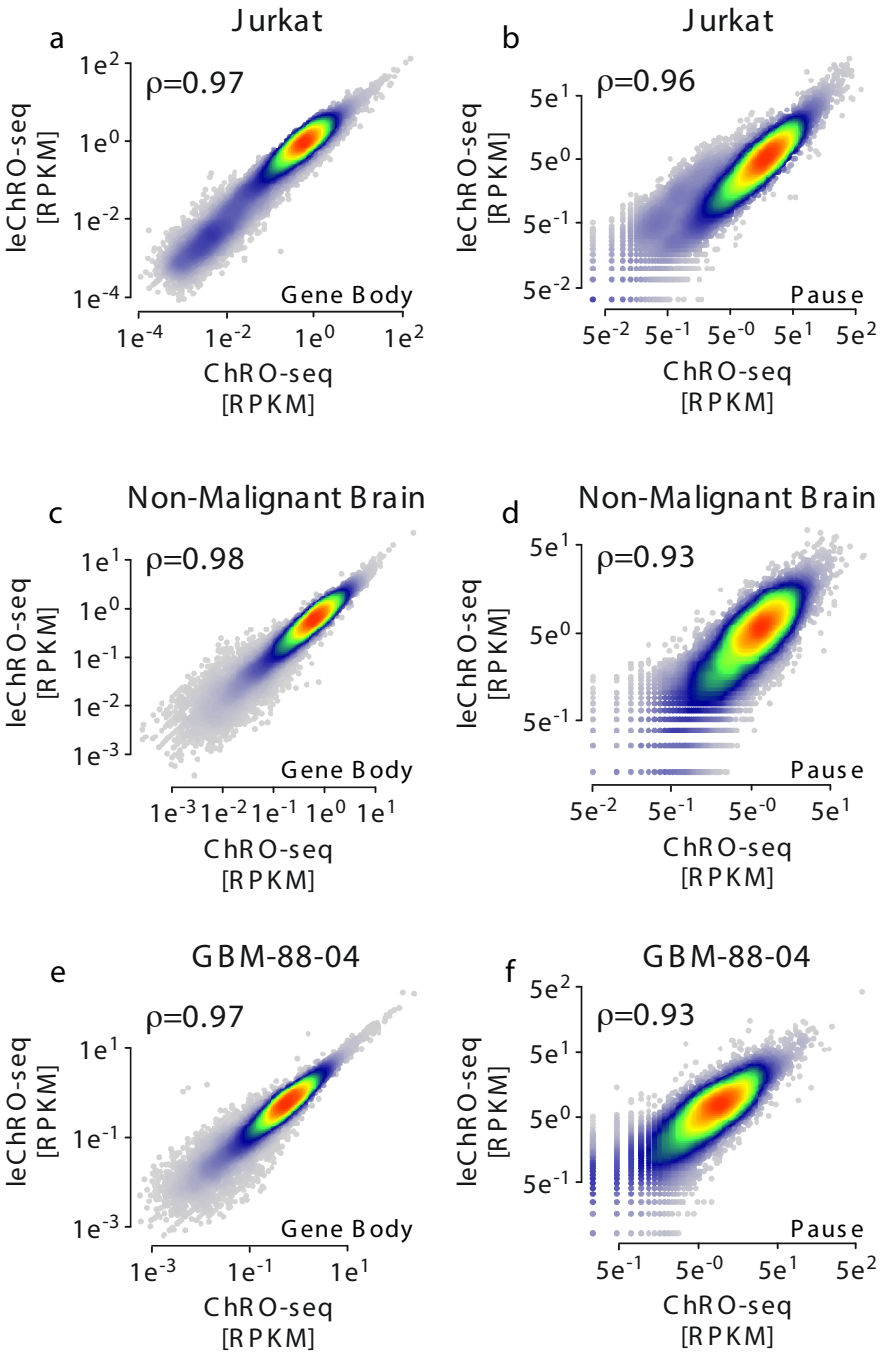
Supplementary Fig. 1. Differences between ChRO-seq and other run-on assays. (a) Length-normalized meta plots show the median signal across 8,403 active gene bodies using PRO-seq (gray), ChRO-seq (blue), leChRO-seq (red), mNET-seq (teal), and Nascent-Seq (purple). **(b)** The genome browser shows the 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.



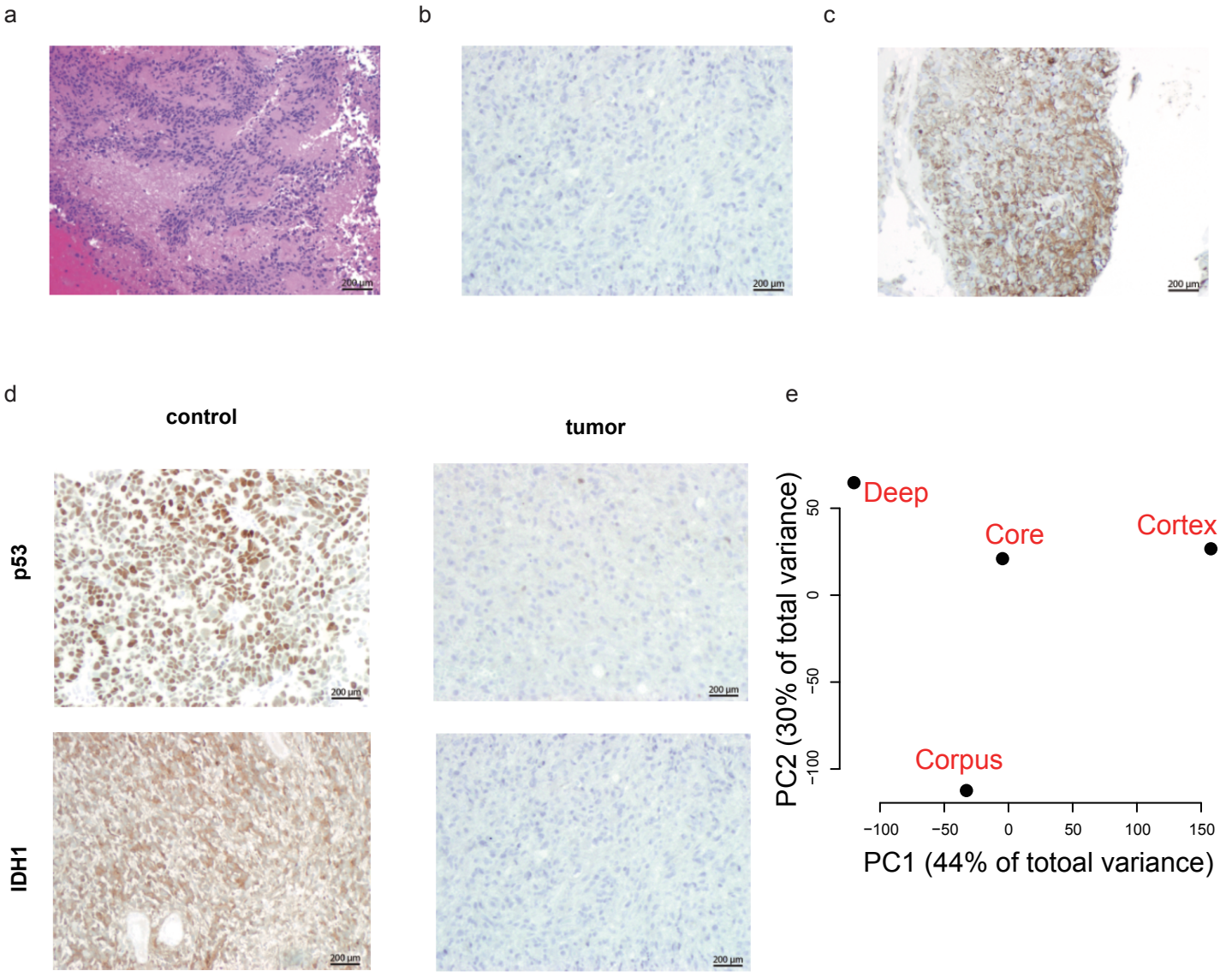
Supplementary Fig. 2. Distribution of signal intensity in the gene body and pause. Violin plot shows the distribution of $\log_2$ of reads per kilobase per million mapped (RPKM) on (**a**) gene body (N=37,184) and (**b**) pause site (N=37,184). Plots are grouped by cell type and colored by the method. White dots represent the means, while the bars represent standard deviations.



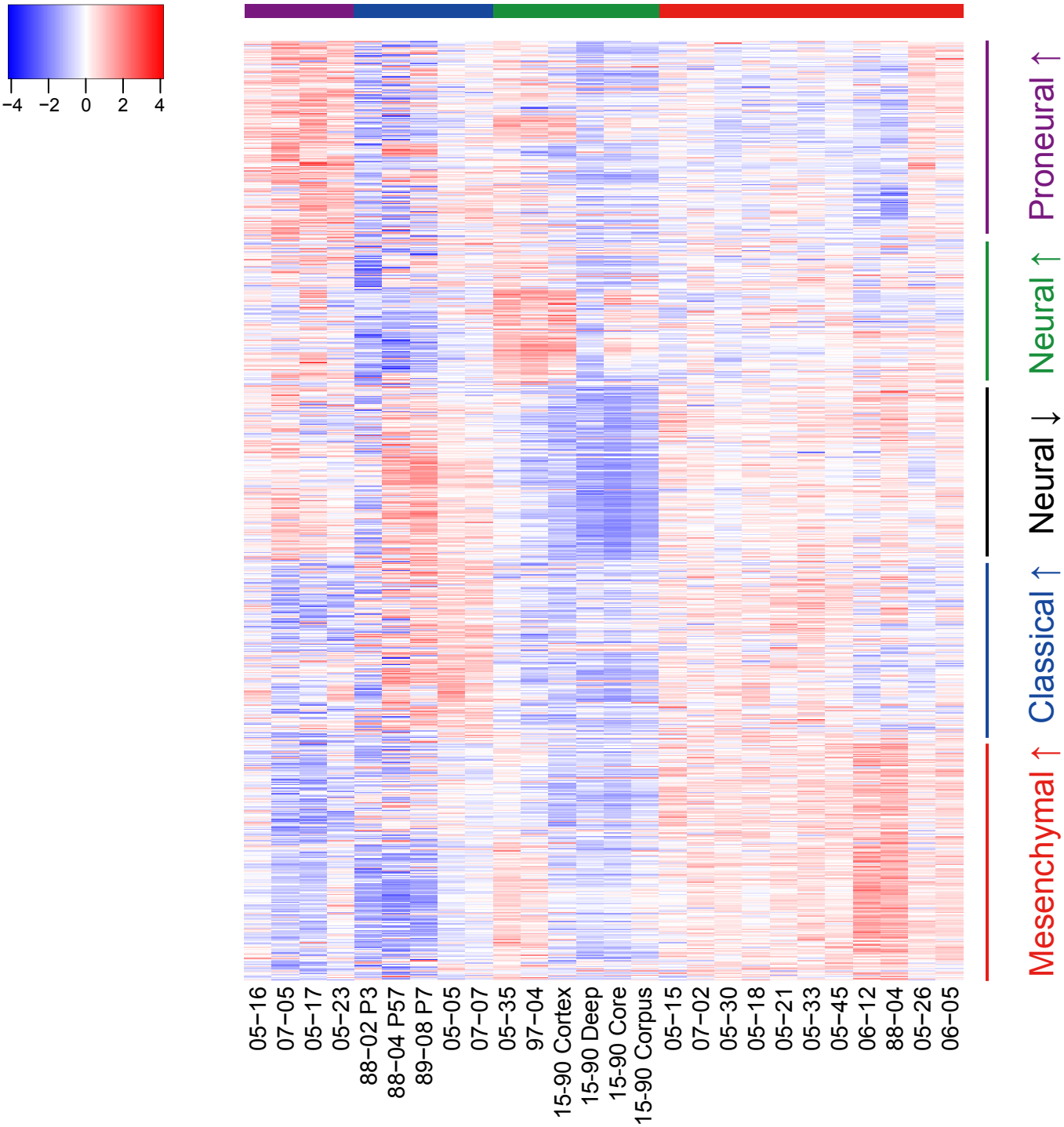
Supplementary Fig. 3. Bioanalyzer analysis of RNA isolated from GBM-88-04. The plot reported by the Bioanalyzer software shows the size of RNA isolated from GBM-88-04 in units of nucleotides (nt, X-axis) as a function of the relative fluorescence units (RFU, Y-axis). RNA Quality Number (RQN = 1) shown 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.



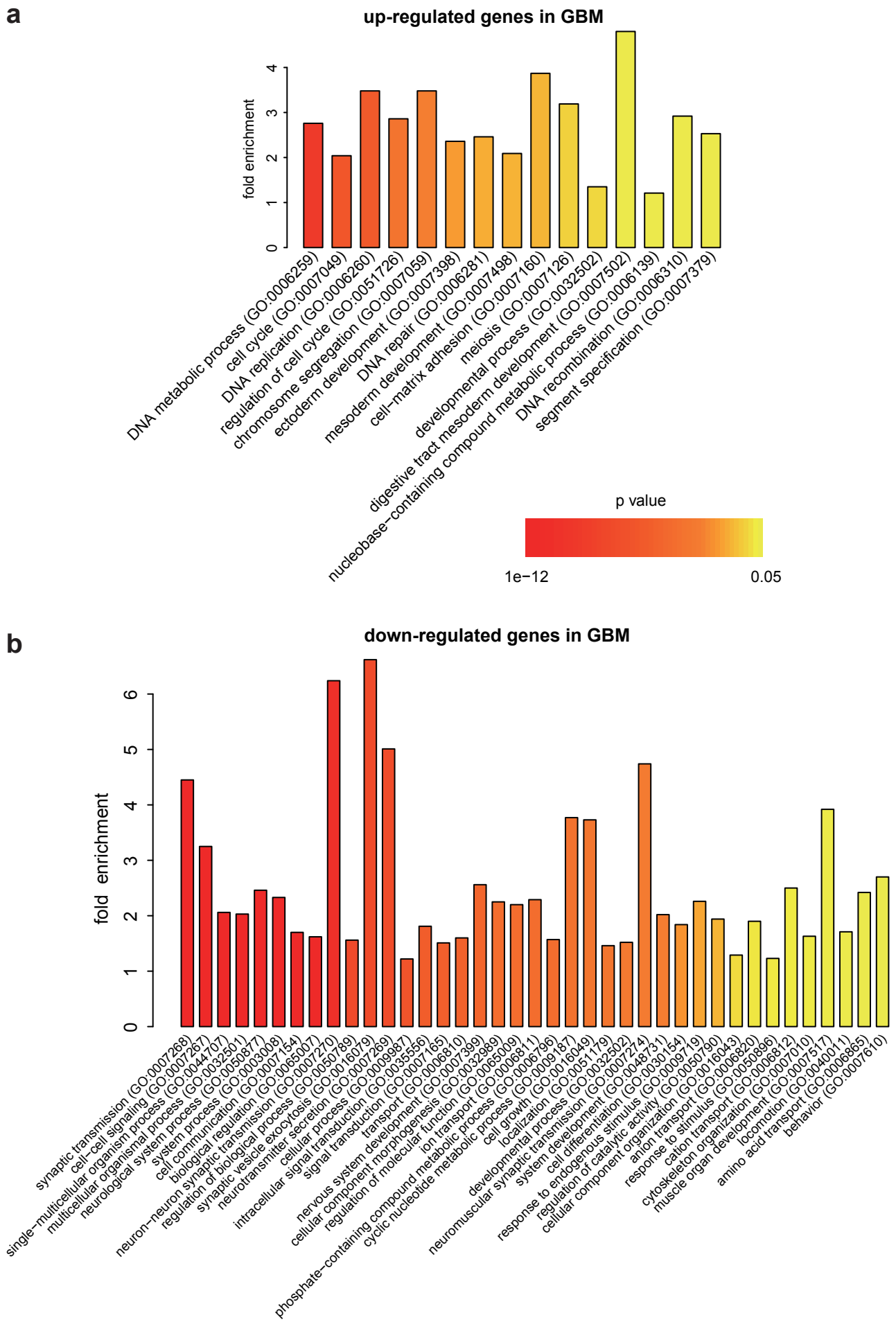
Supplementary Fig. 4. Correlation between ChRO-seq and leChRO-seq. (**a**-**f**) Scatterplots show the density of reads mapping in the gene bodies (+1000 to gene end) (**a**, **c**, **e**) or in the promoter proximal pause near the transcription start site (**b**, **d**, **f**) of 41,478 RefSeq genes. 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.



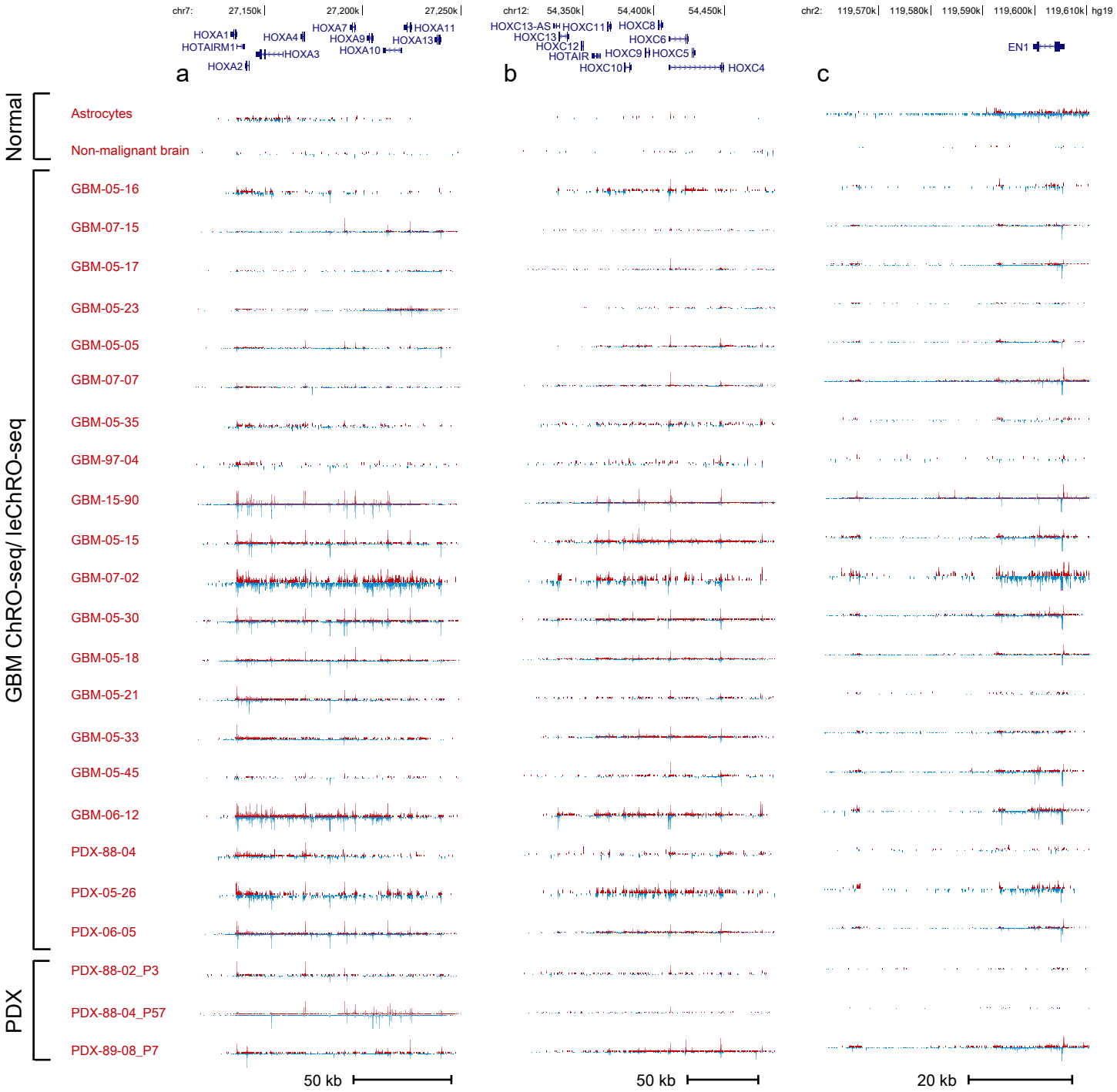
Supplementary Fig. 5. Brain biopsies display immunohistochemical markers of high grade glioma in GBM-15-90. (**a**) Pseudopalisading borders with necrotic centers. (**b**) IDH1 staining is negative. (**c**) GFAP is stained as positive. (**d**) Additional markers of high grade glioma between the tumor 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).



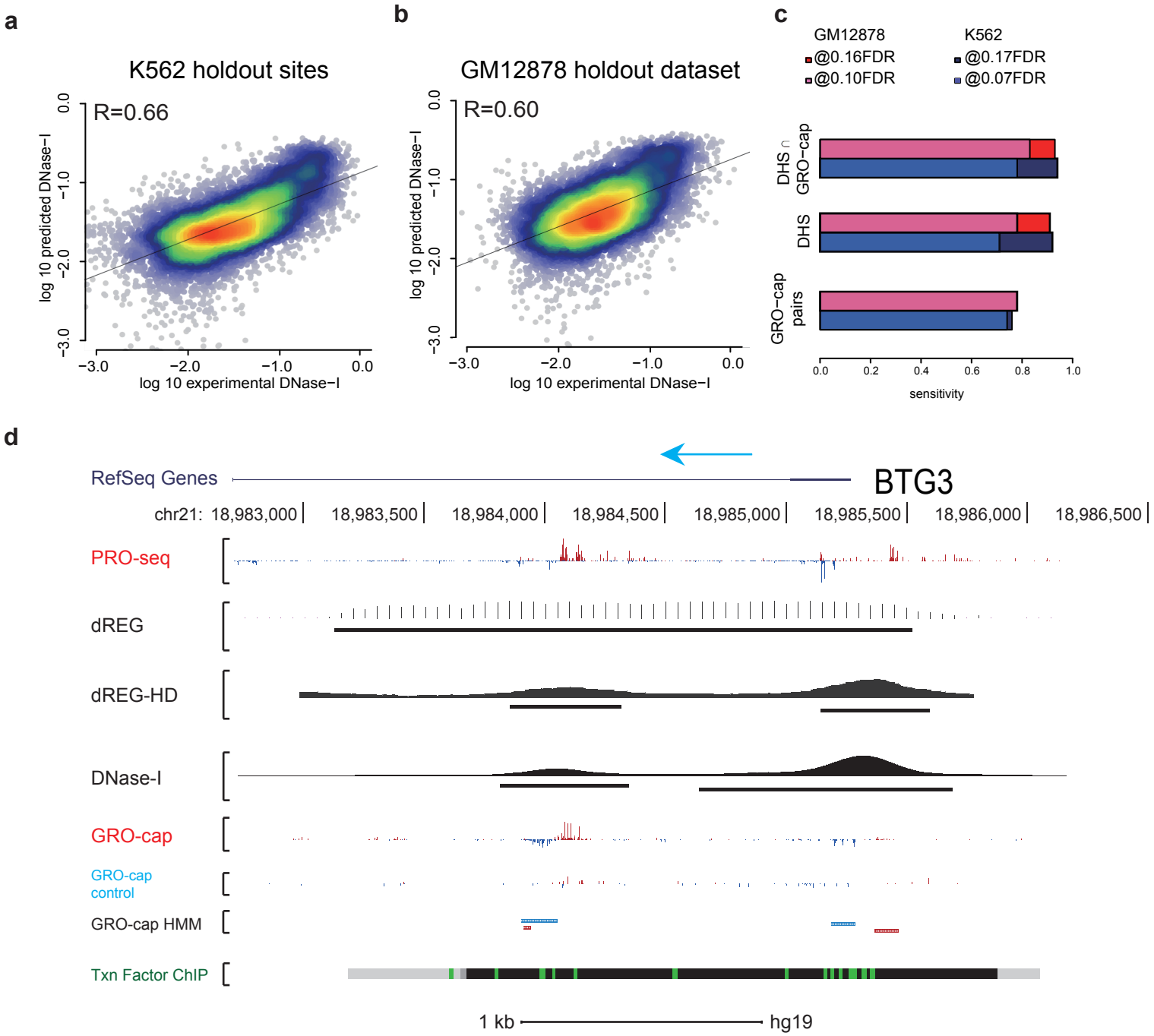
Supplementary Fig. 6. Expression of molecular subtype predictor genes in primary GBM / PDX samples. Heatmap shows the expression of 838 genes relevant for classifying among the four known molecular subtypes of glioblastoma. Red colors indicate higher transcription activity and blue colors indicate lower activity. Samples are ordered based on subtype.



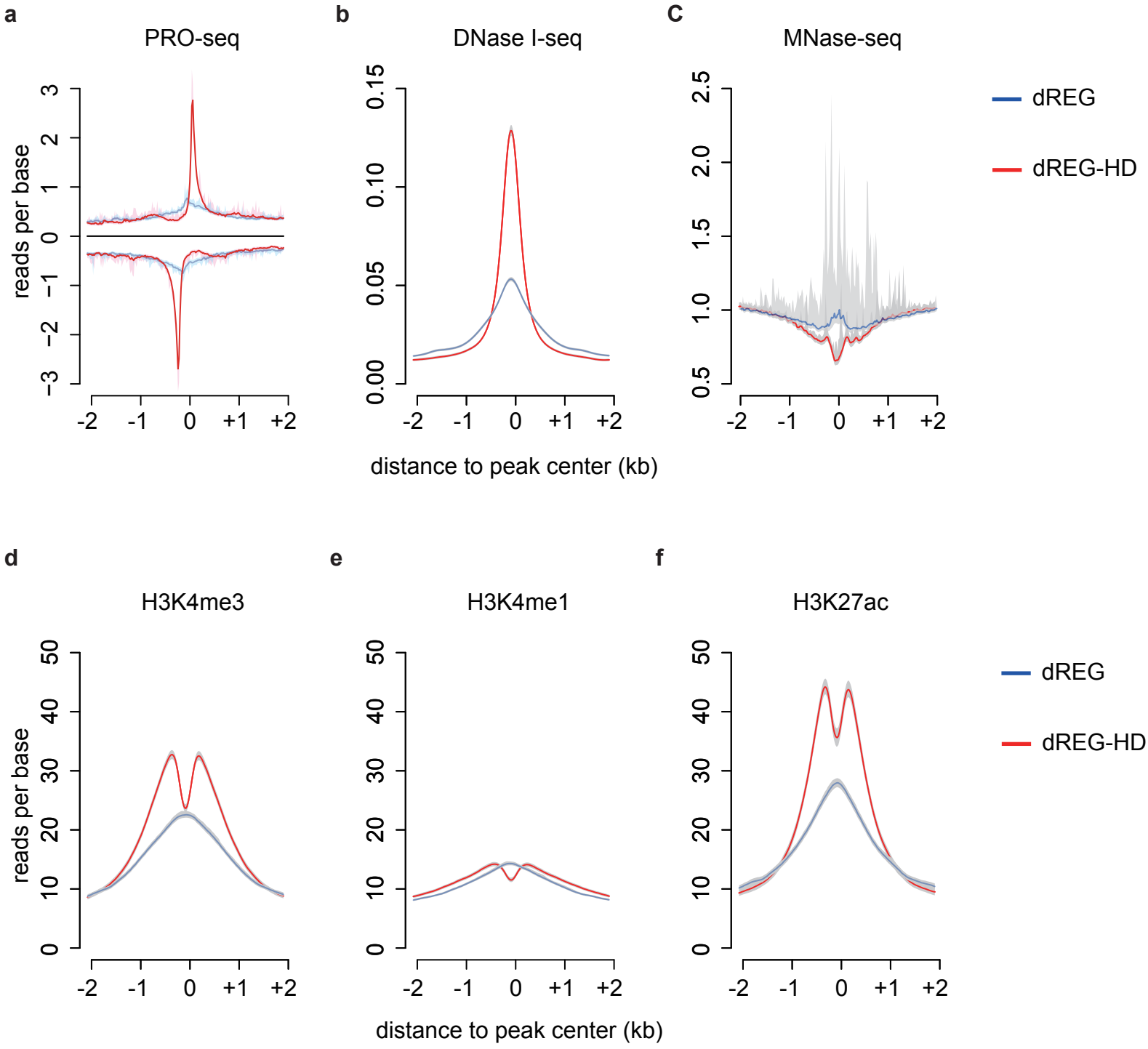
Supplementary Fig. 7. Gene ontology analysis of differentially expressed genes in GBM compared to non-malignant brain tissue. Barplot shows the the gene ontogoly terms enriched for genes up-regulated in GBM (**a**, N=2,018) and down-regulated in GBM (**b**, N=1,486). Ontology groups are ordered by statistical 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.



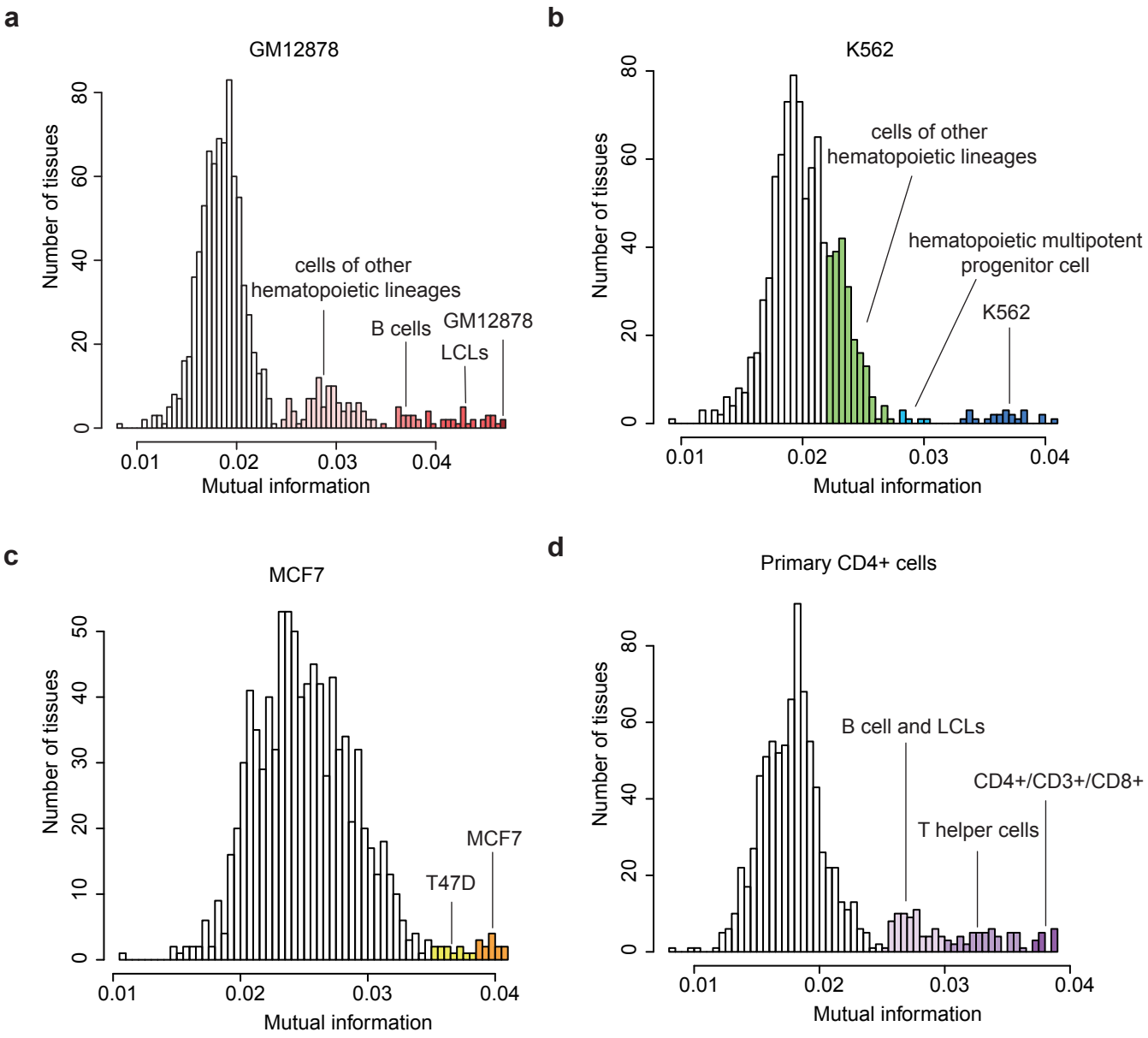
Supplementary Fig. 8. HOXA, HOXC, and EN1 loci show strong differential expression in primary GBM and PDX. Browser tracks of ChRO-seq signal in primary GBM, PDX, cultured astrocyte, and non-malignant brain samples, DNase-I hypersensitivity in normal adult and fetal brain tissues, and H3K27ac peaks in 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.



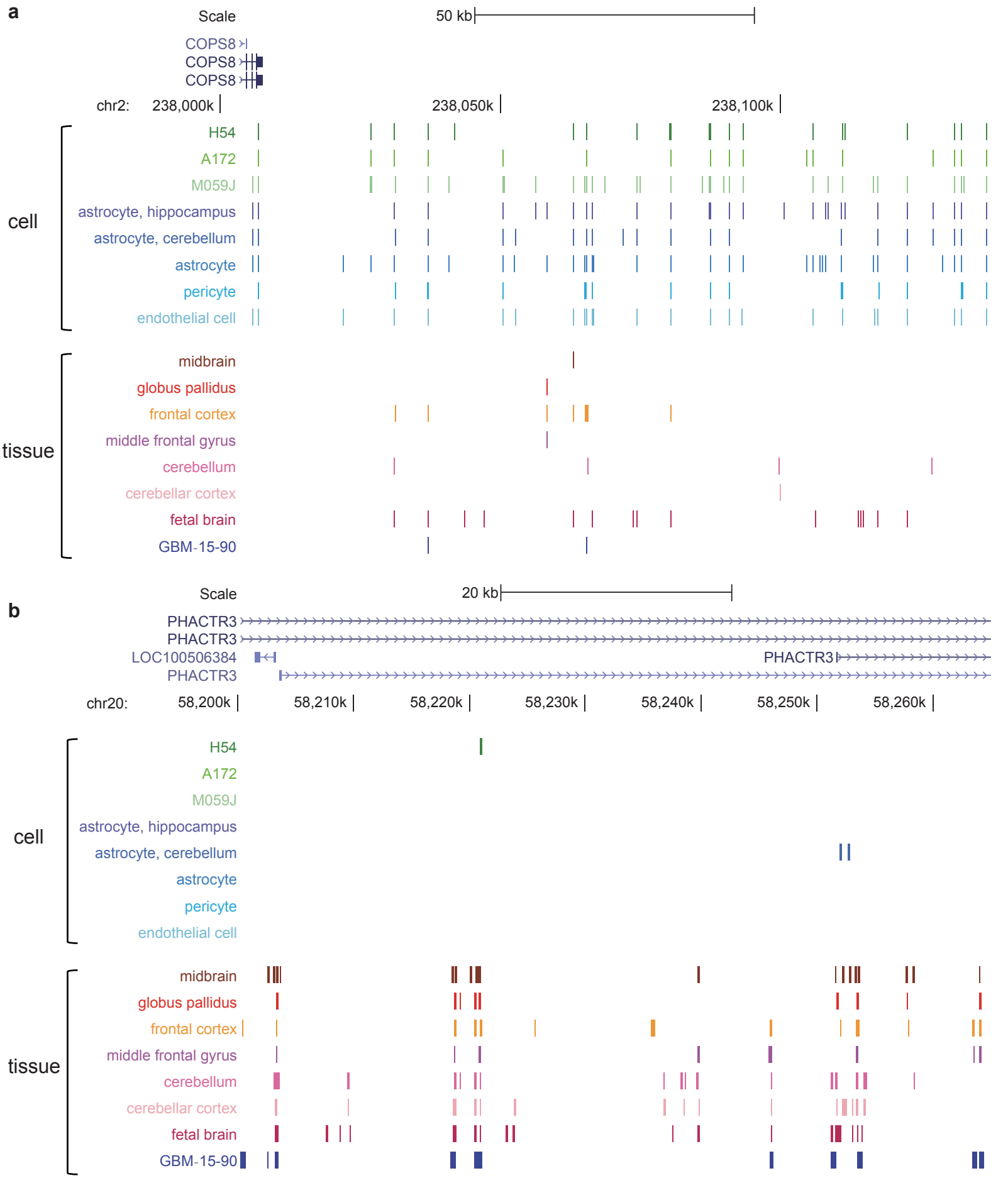
Supplementary Fig. 9. dREG-HD refines TRE predictions by imputing DNase-I hypersensitivity. (**a** and **b**) Density scatter plots show a comparison between predicted and experimental DNase-I hypersensitivity signals in K562 holdout sites that were not used during training (**a**, N=303,068) and a complete 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.



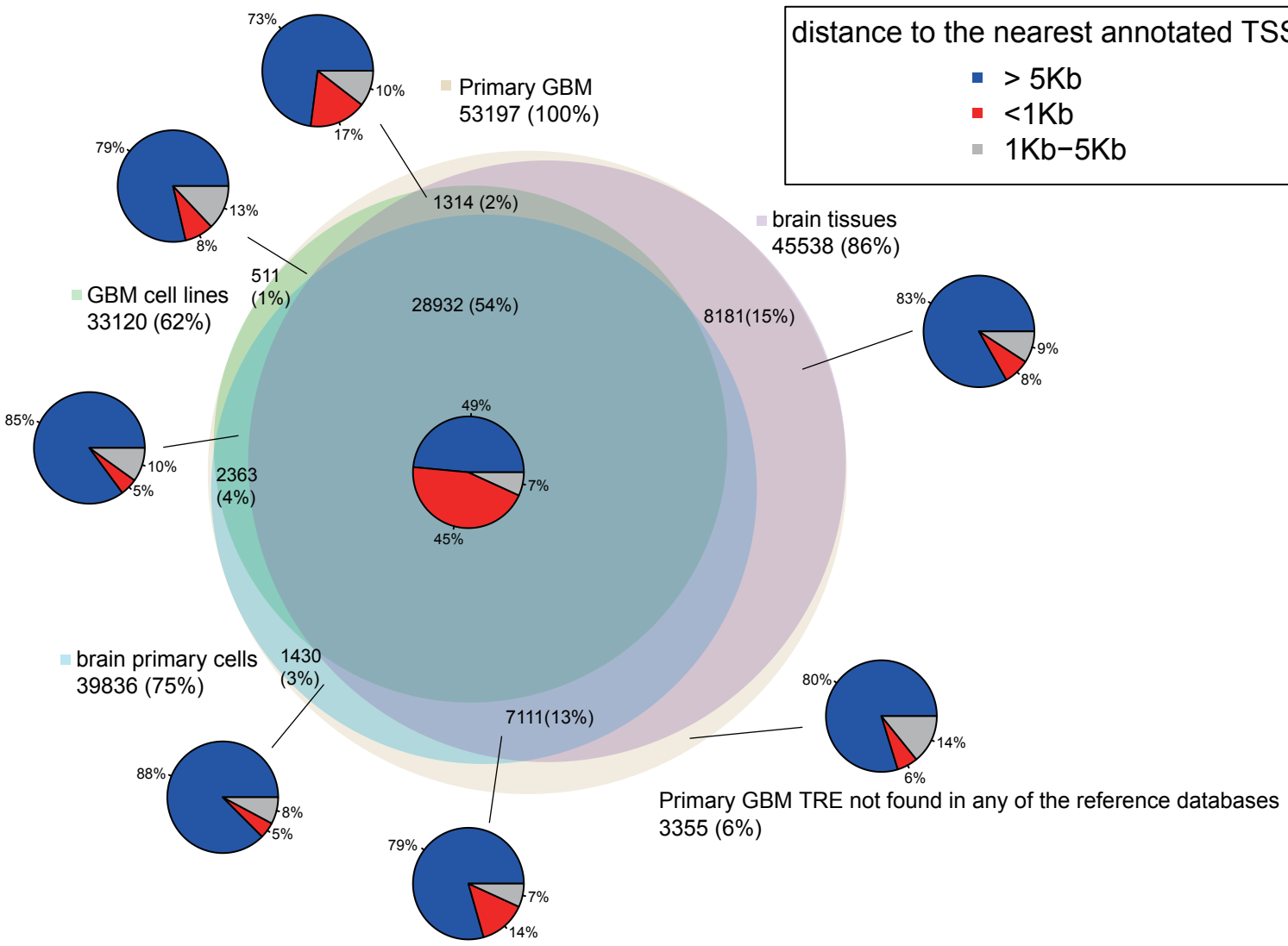
Supplementary Fig. 10. Metaplots for PRO-seq, chromosome accessibility, and histone modifications that marks active TREs. Signals of the indicated mark over dREG and dREG-HD regions are shown in blue and red, respectively. Shadows marks the 25 and 75 percentiles of 1000 samples of 10% of the data (see methods).



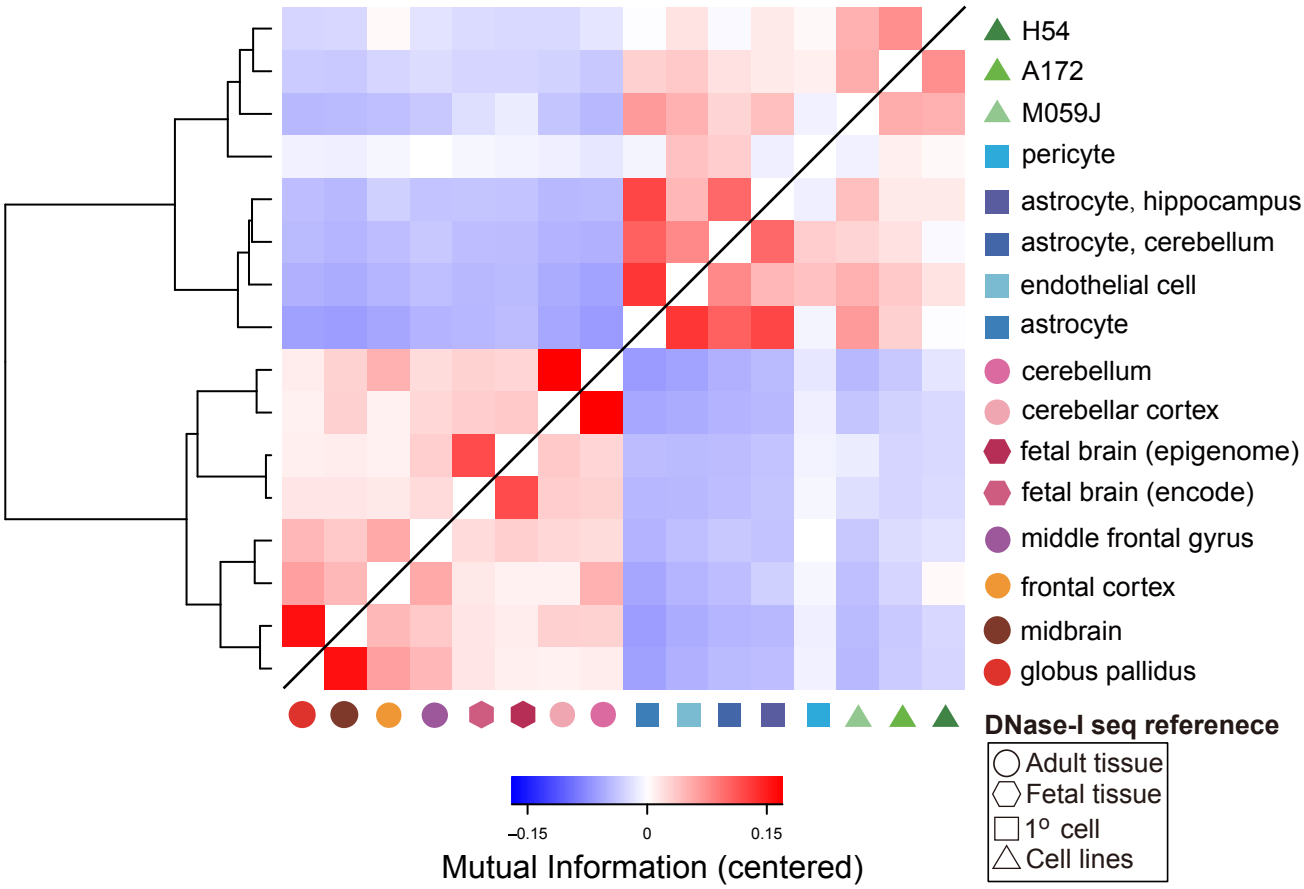
Supplementary Fig. 11. Mutual information is an accurate similarity measure for TREs. Histogram represents the mutual information between dREG-HD sites identified using PRO-seq or GRO-seq data and DHSs from 921 public DNase-I-seq experiments and in the indicated sample (**a**:GM12878, **b**:K562, **c**:MCF-7, **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.



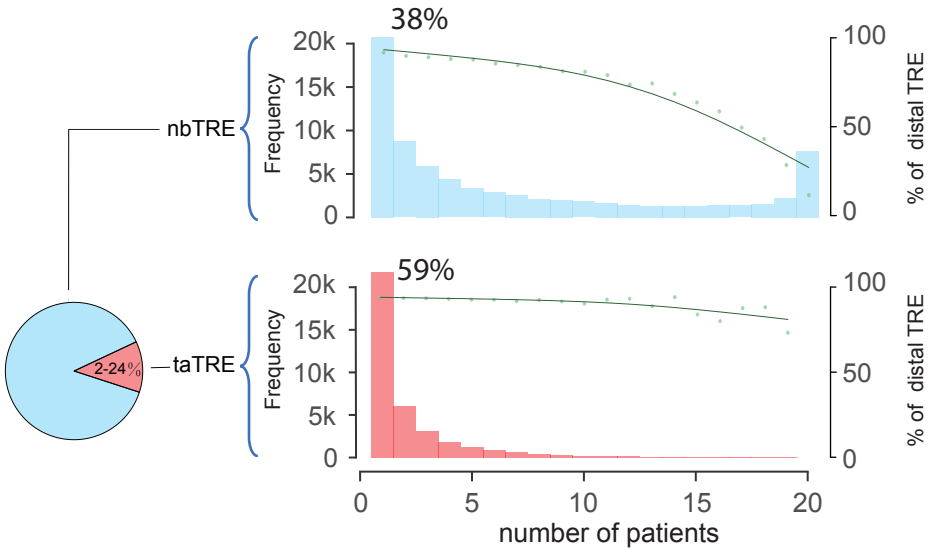
Supplementary Fig. 12. DNase-I hypersensitive sites with differences between brain tissues and cultured brain cells. (**a**) Locus near the COPS8 gene that shows consecutive activation of TREs in cultured brain cells but not in normal brain tissues. (**b**) Locus near PHACTR3 gene that shows activation of TREs in primary brain tissues but not in cultured brain cells.



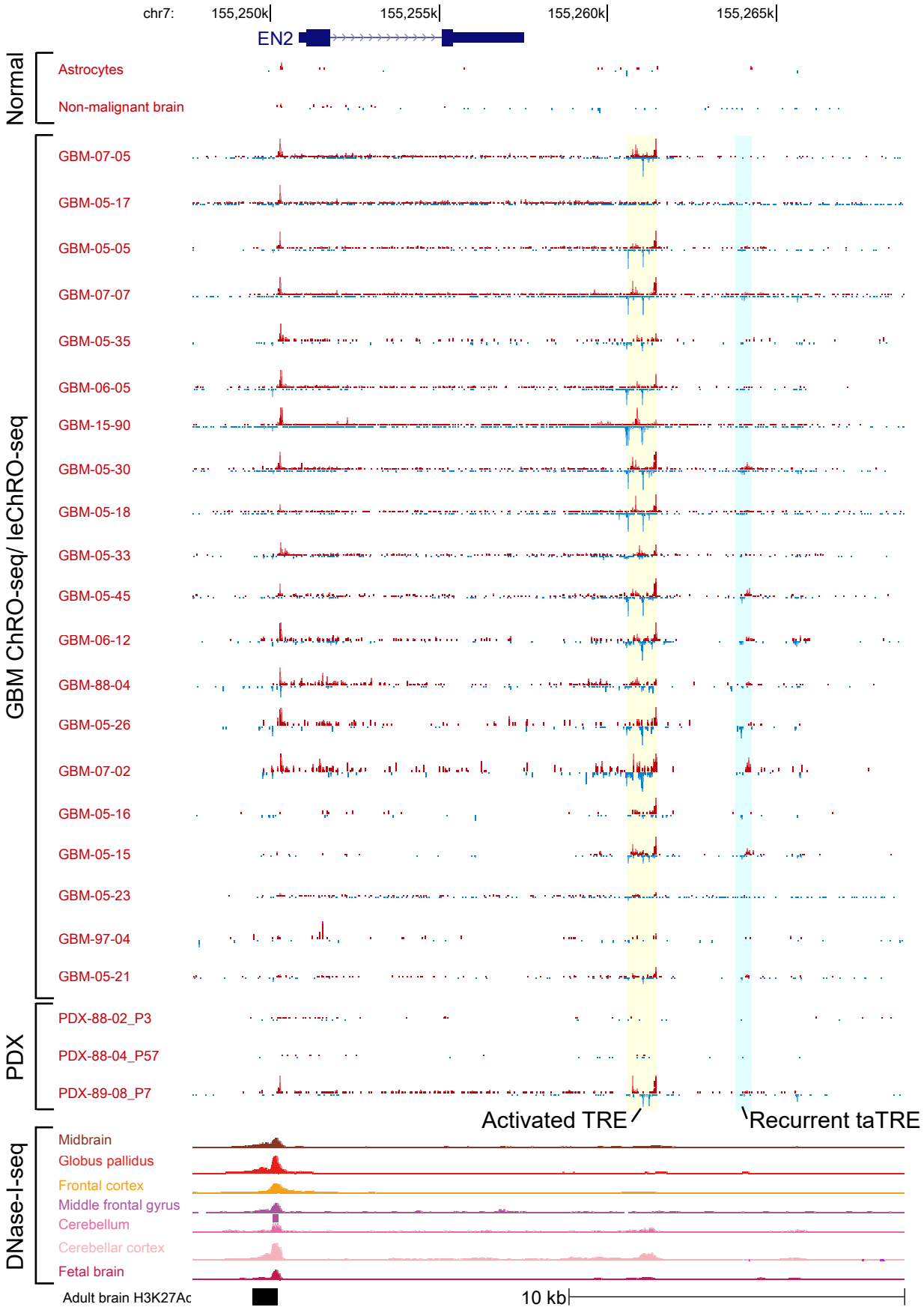
Supplementary Fig. 13. Venn diagram showing similarity in TREs between primary GBMs, normal brain tissue, and primary brain cells grown in tissue culture. Venn diagram denotes the overlap between TREs found in GBM-15-90 and normal brain (pink), GBM cell line models (green), or primary brain cells that 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).



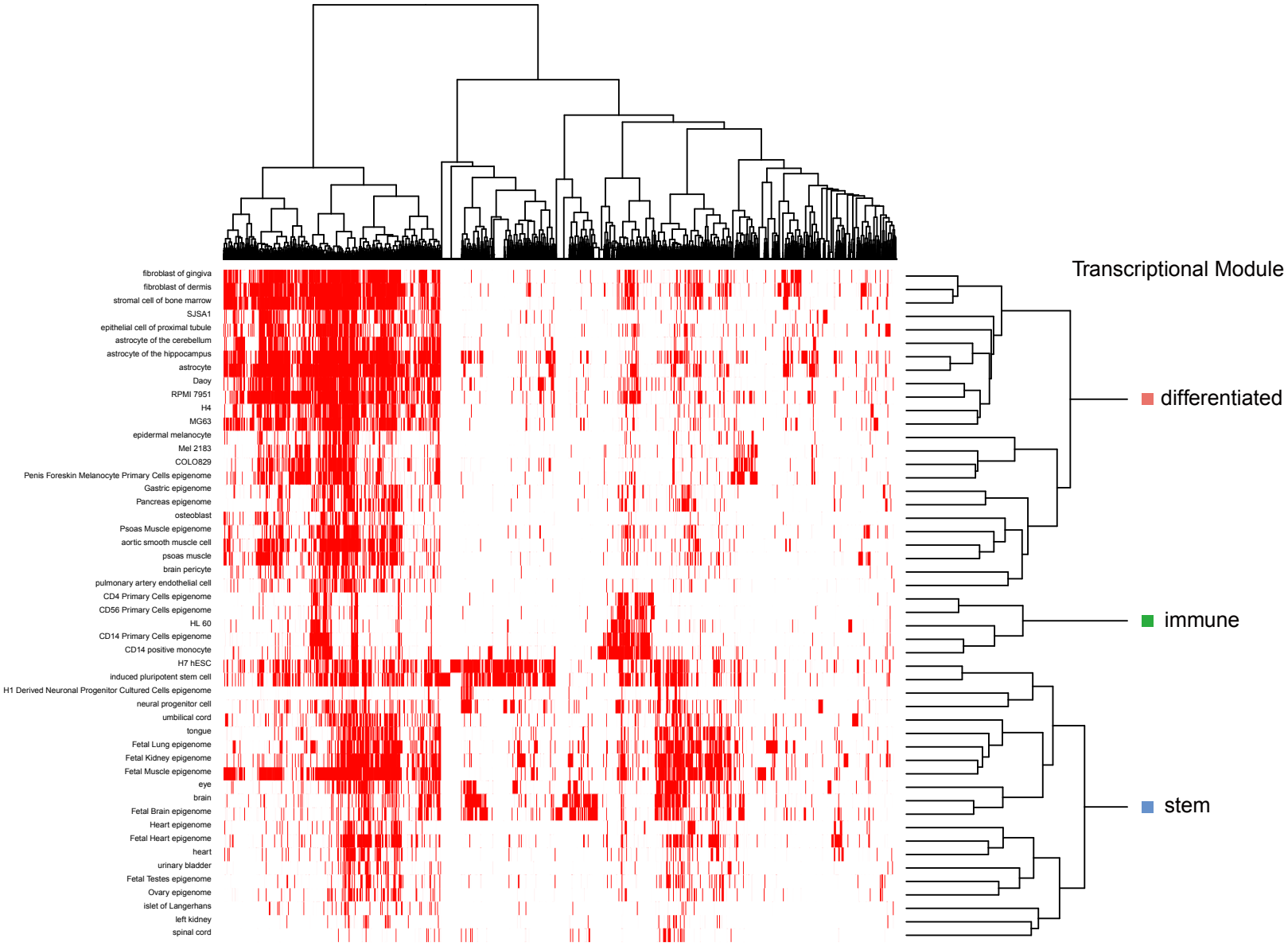
Supplementary Fig. 14. Pairwise mutual information among TREs from brain-related reference DHSs centered by the mean of each sample. Heatmap shows the centered mutual information between the indicated samples. Sample order was selected by hierarchical clustering using the algorithm described in Supplementary Note 7.



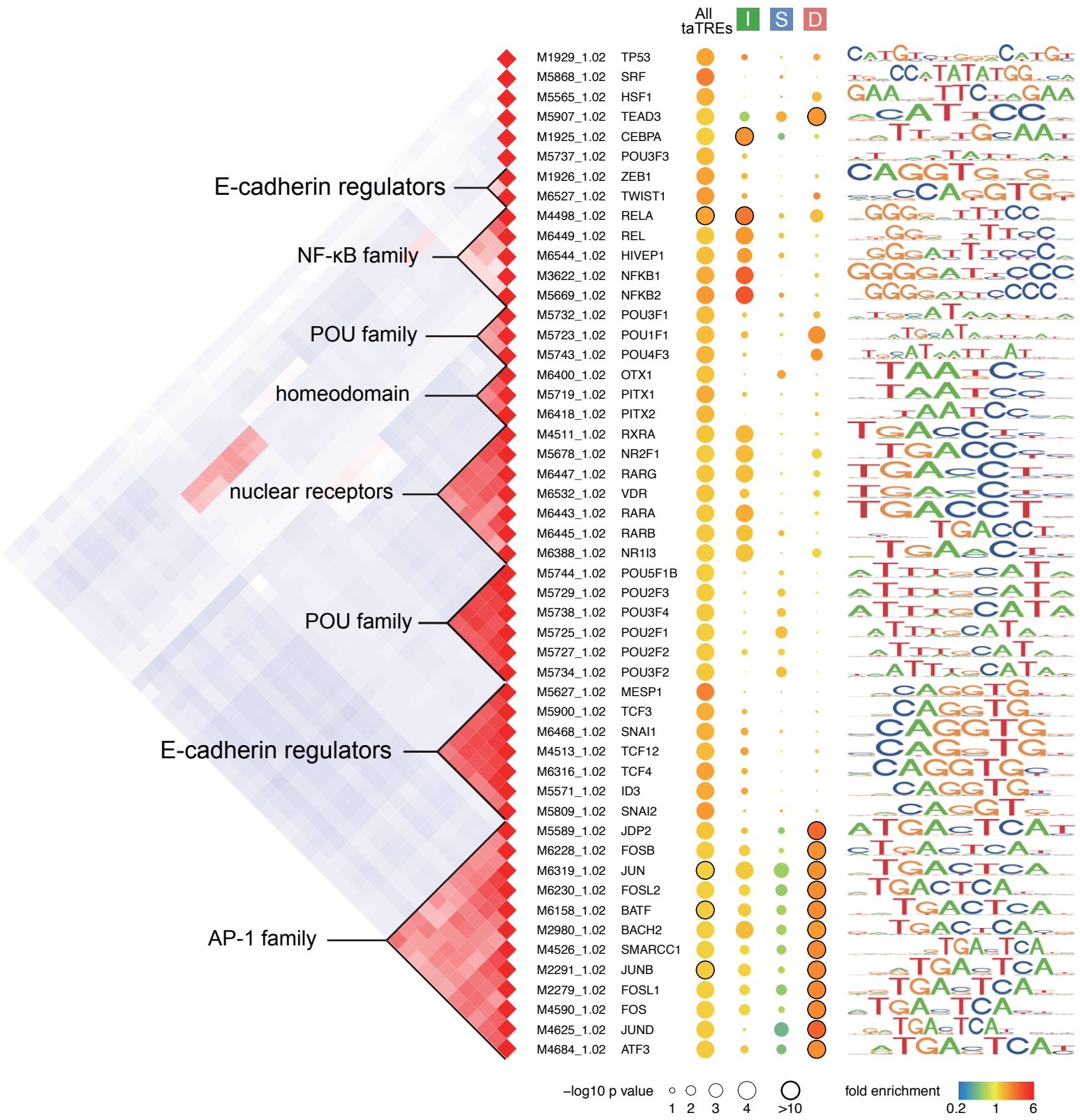
Supplementary Fig. 15. Distribution of the frequency across GBM patients of normal brain and taTREs. Histograms show the distribution of the number of primary GBM patients (out of 20) in which each TRE is active. 2 to 24% of TREs in GBM samples are not found in normal adult brain tissues. The percentage of TREs >1kb from the nearest transcription start site (distal) is shown in green dots.



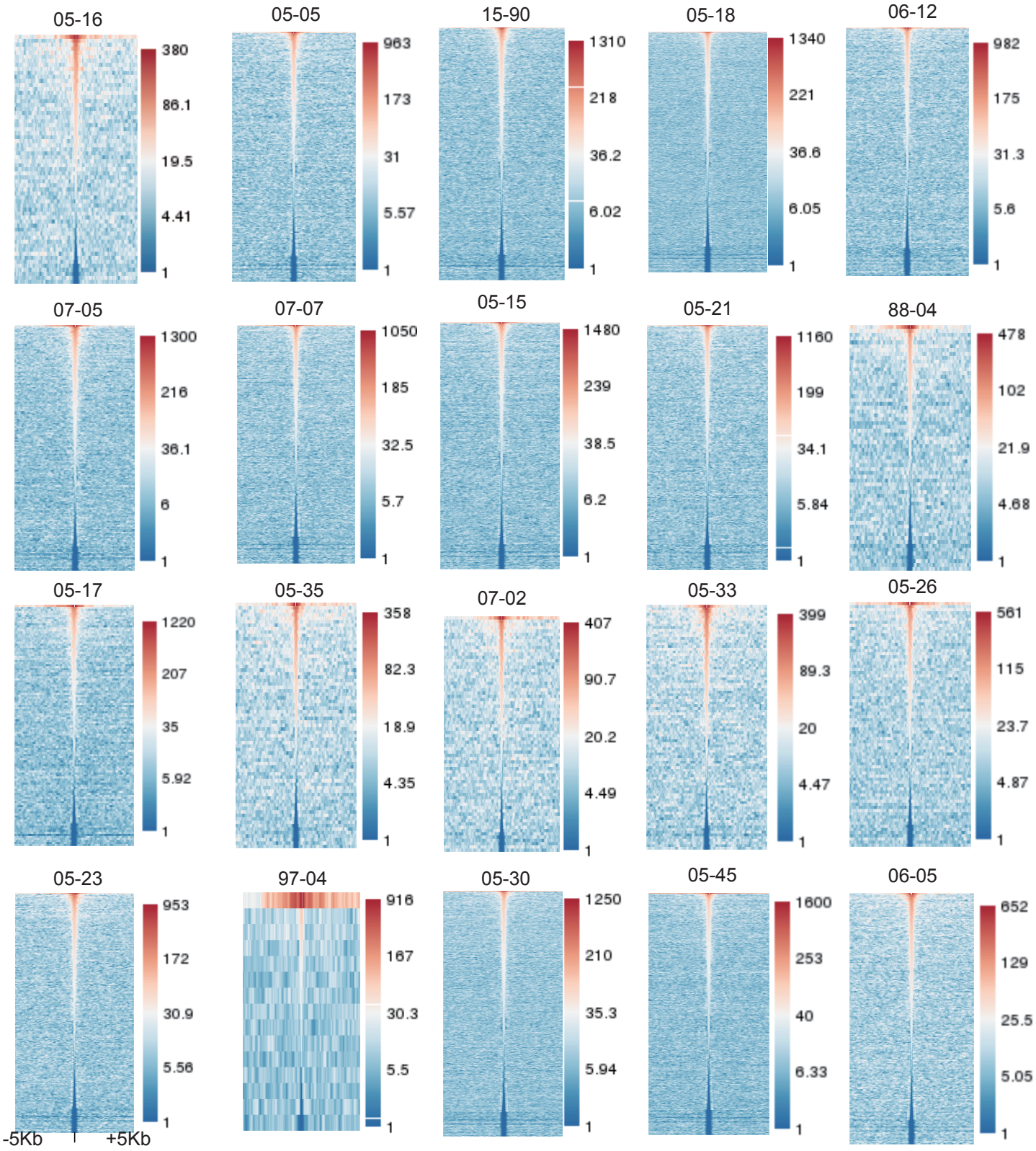
Supplementary Fig. 16. EN2 locus show strong differential expression and activation of taTREs in GBM. Browser tracks of ChRO-seq signal in primary GBM and PDX, normal astrocyte and non-malignant brain samples, DNase-I hypersensitivity and in normal adult and fetal brain tissues, and H3K27ac peaks in 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.



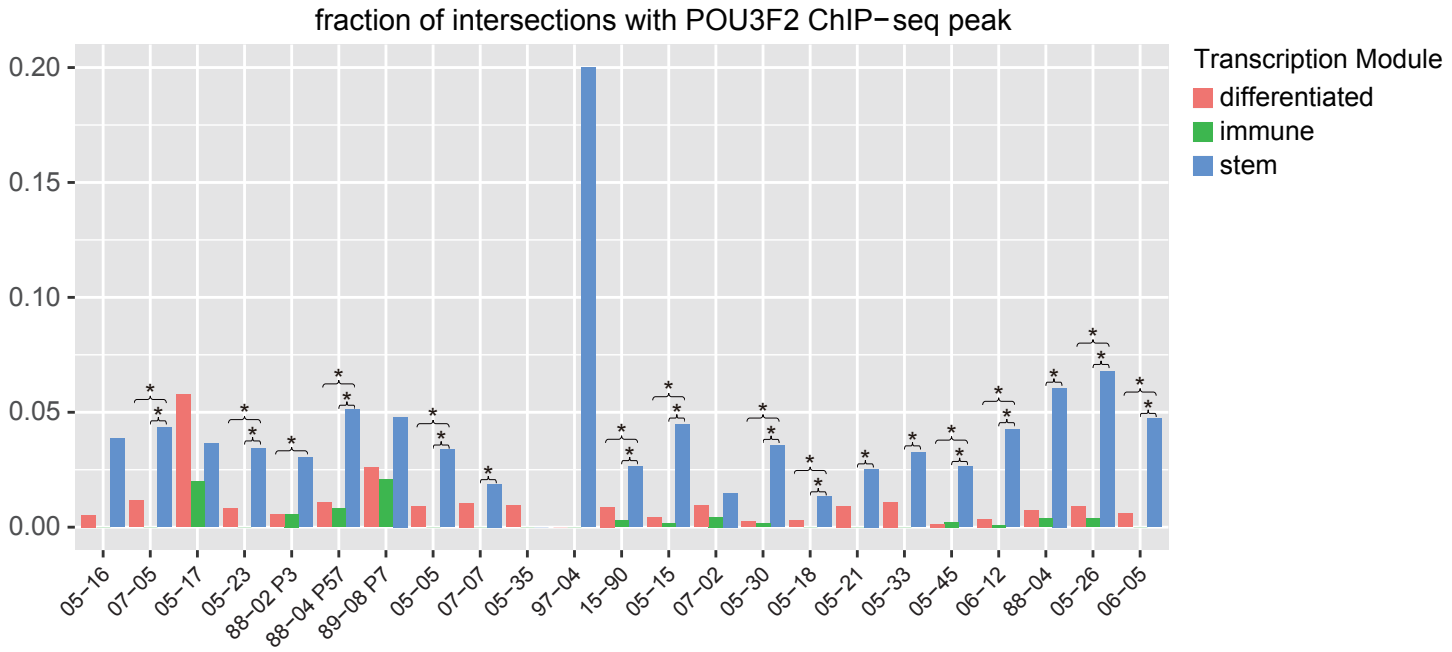
Supplementary Fig. 17. Clustering of taTREs-enriched reference samples. Clustering of reference samples enriched for taTREs based on the activation of TREs. Active TREs are marked in red; inactive ones are in white. Row dendrograms are cut down to three trees, each corresponding to the indicated transcriptional regulatory program (i.e., stem- or fetal-like, immune, and differentiated).



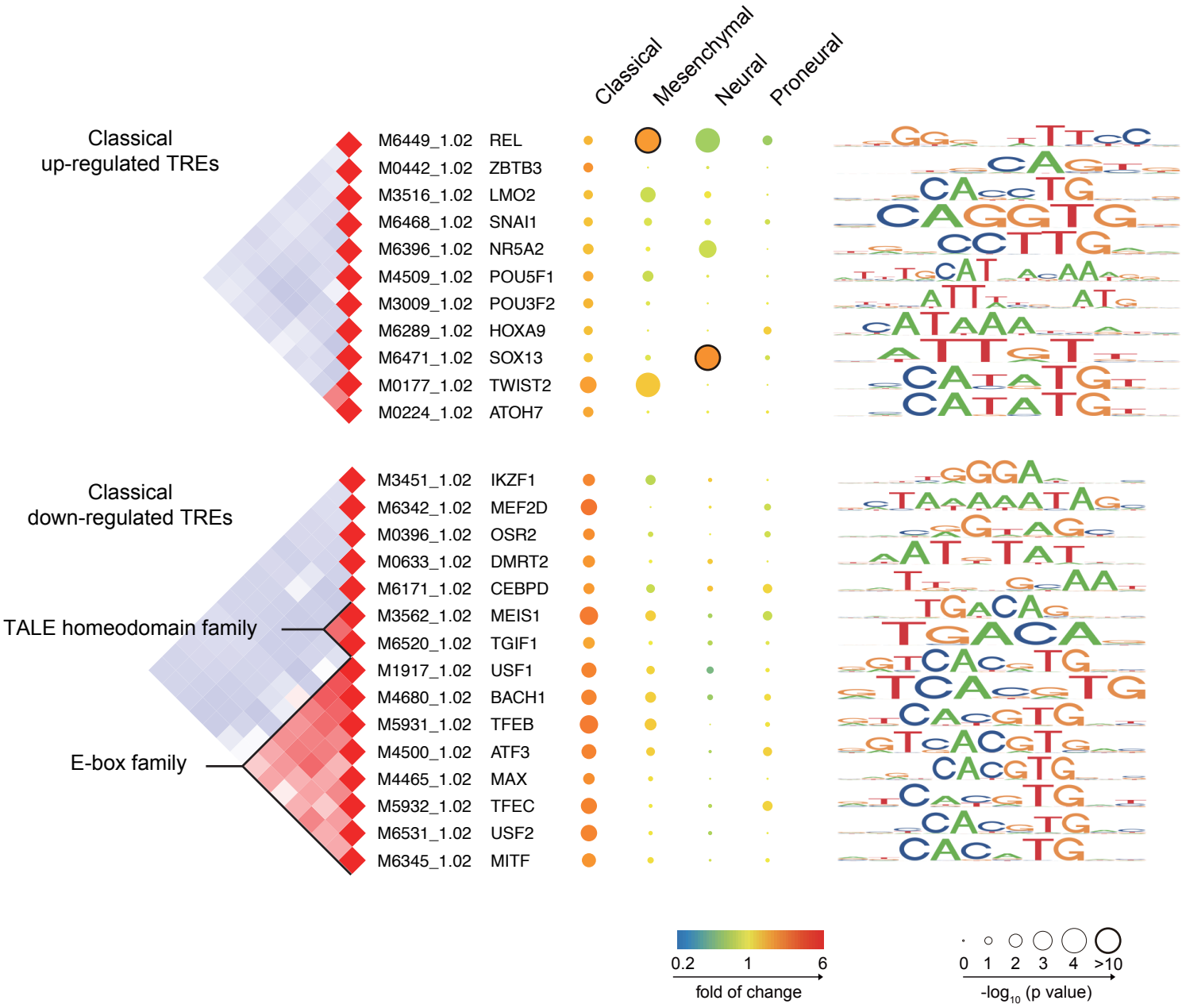
Supplementary Fig. 18. Transcription factor binding motifs enriched in TREs in the indicated regulatory program compared with normal brain. Transcription factor binding motifs enriched in TREs that are members of the immune (I), stem (S), or differentiated (D) regulatory program (top) compared with TREs 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 3.

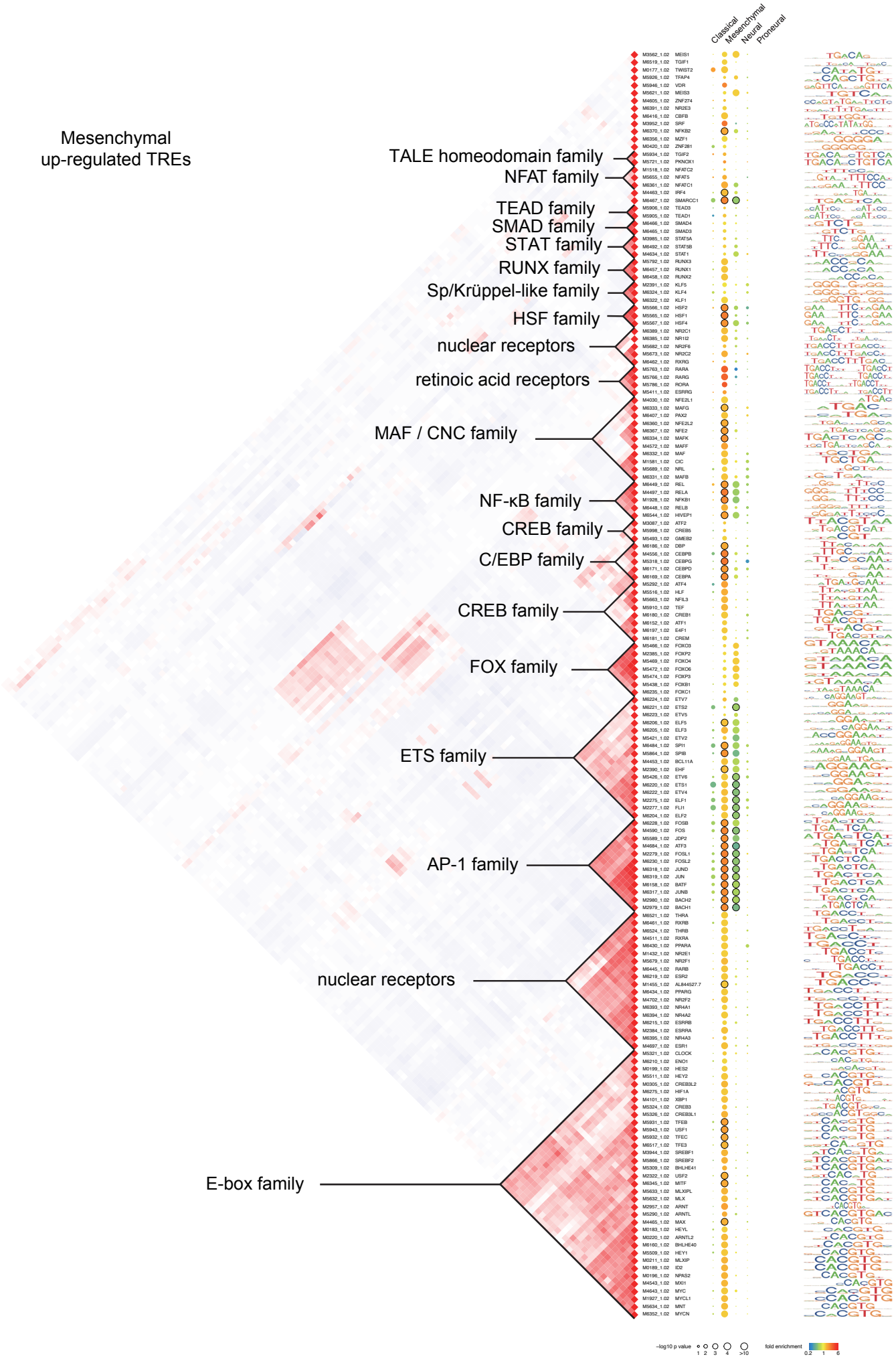


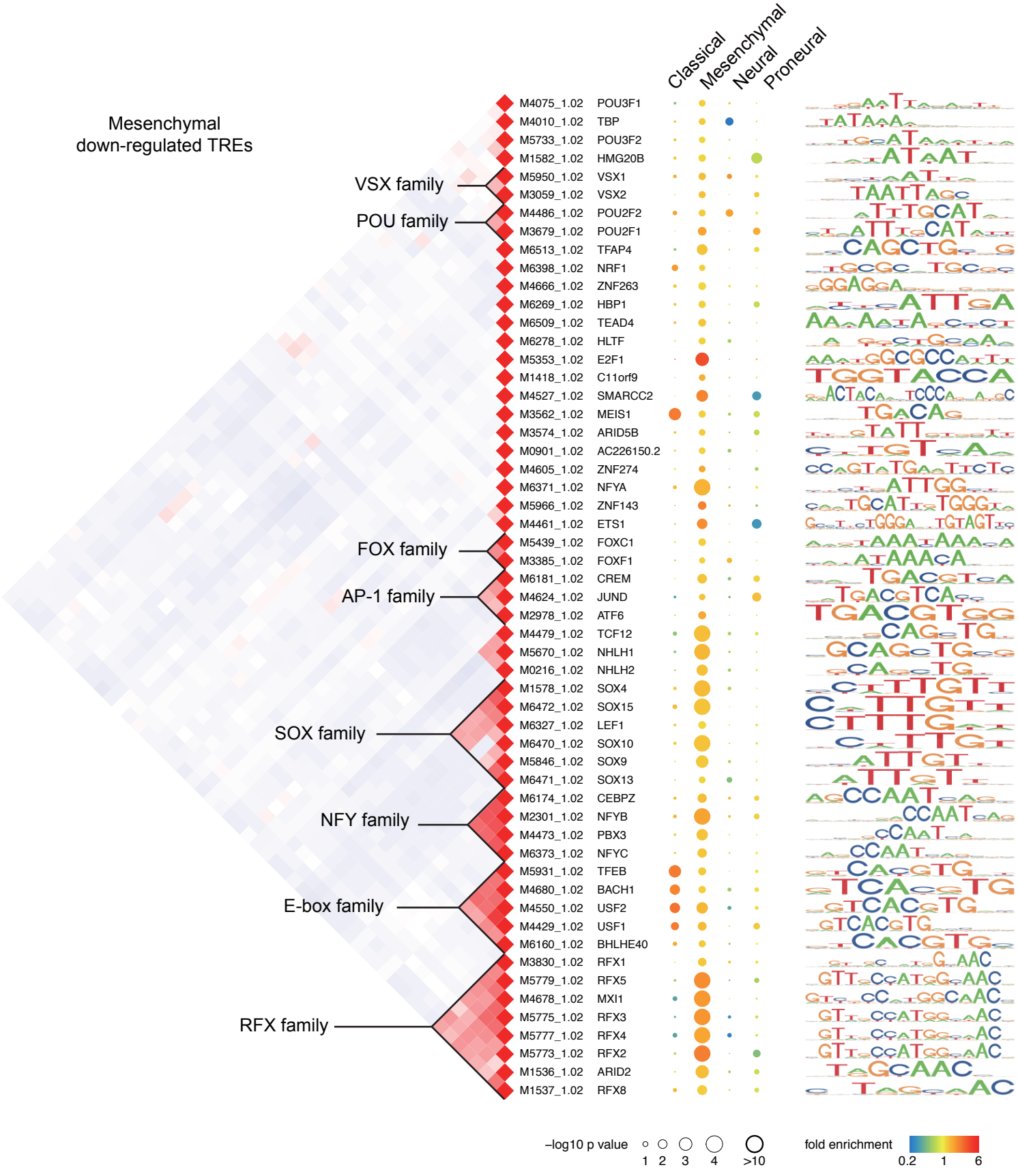
Supplementary Fig. 19. taTREs show enrichment of POU3F2 binding in tumor propagating cells. Heatmaps show ChIP-seq signals for POU3F2 in tumor propagating cells ±5kb surrounding the center of taTREs. Data was from (Suvà et al. 2014). Rows were ordered by the sum of ChIP-seq signals. Plots are made using the R pheatmap package (Kolde 2015).



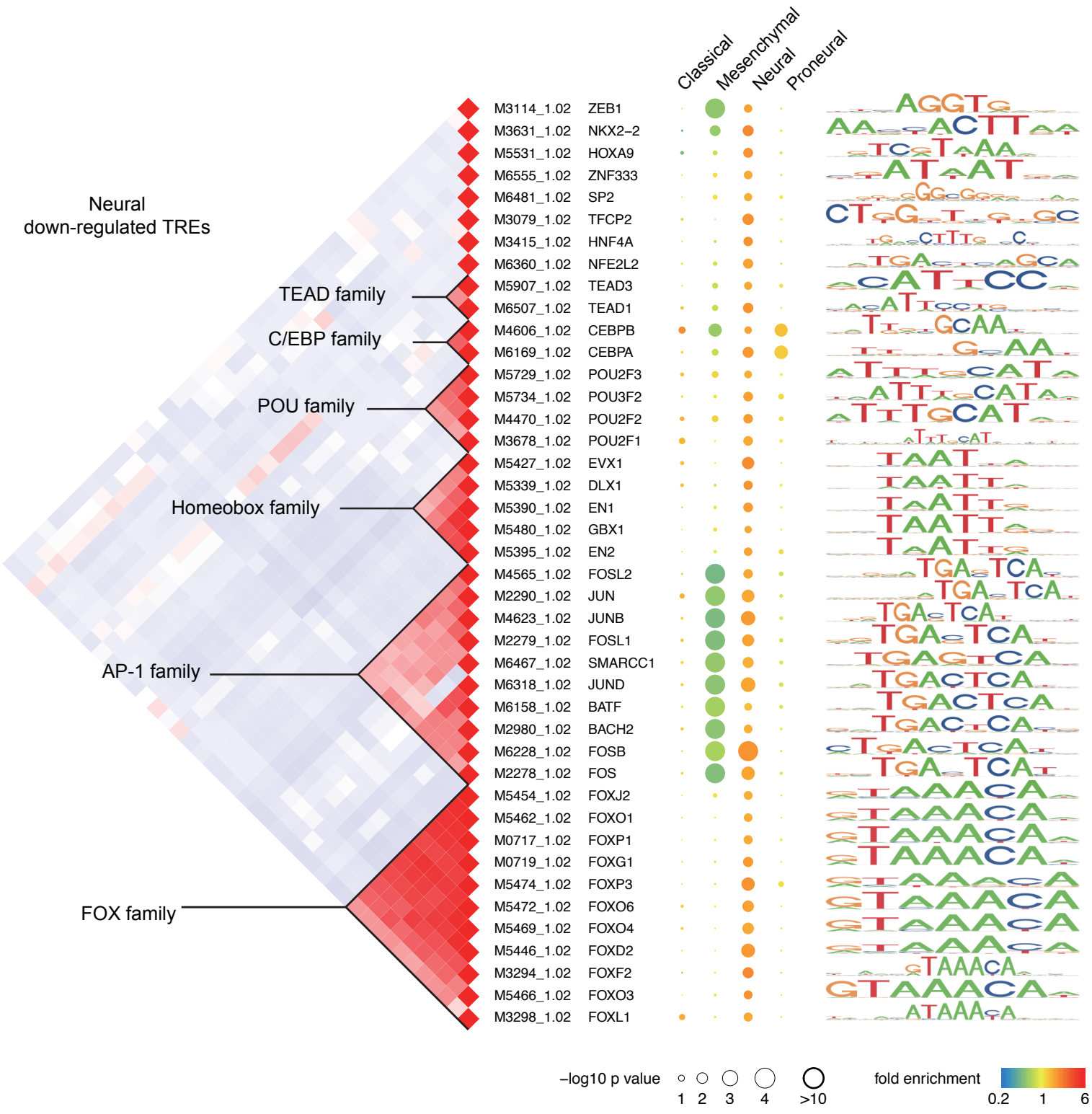
Supplementary Fig. 20. Stem program taTREs enriched for POU3F2 ChIP-seq peaks. The height of bars shows the fraction of POU3F2 ChIP-seq peaks that intersect with taTRE in each of the primary GBM / PDX samples. taTREs from differentiated and stem programs are colored in red and green respectively. Primary 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.

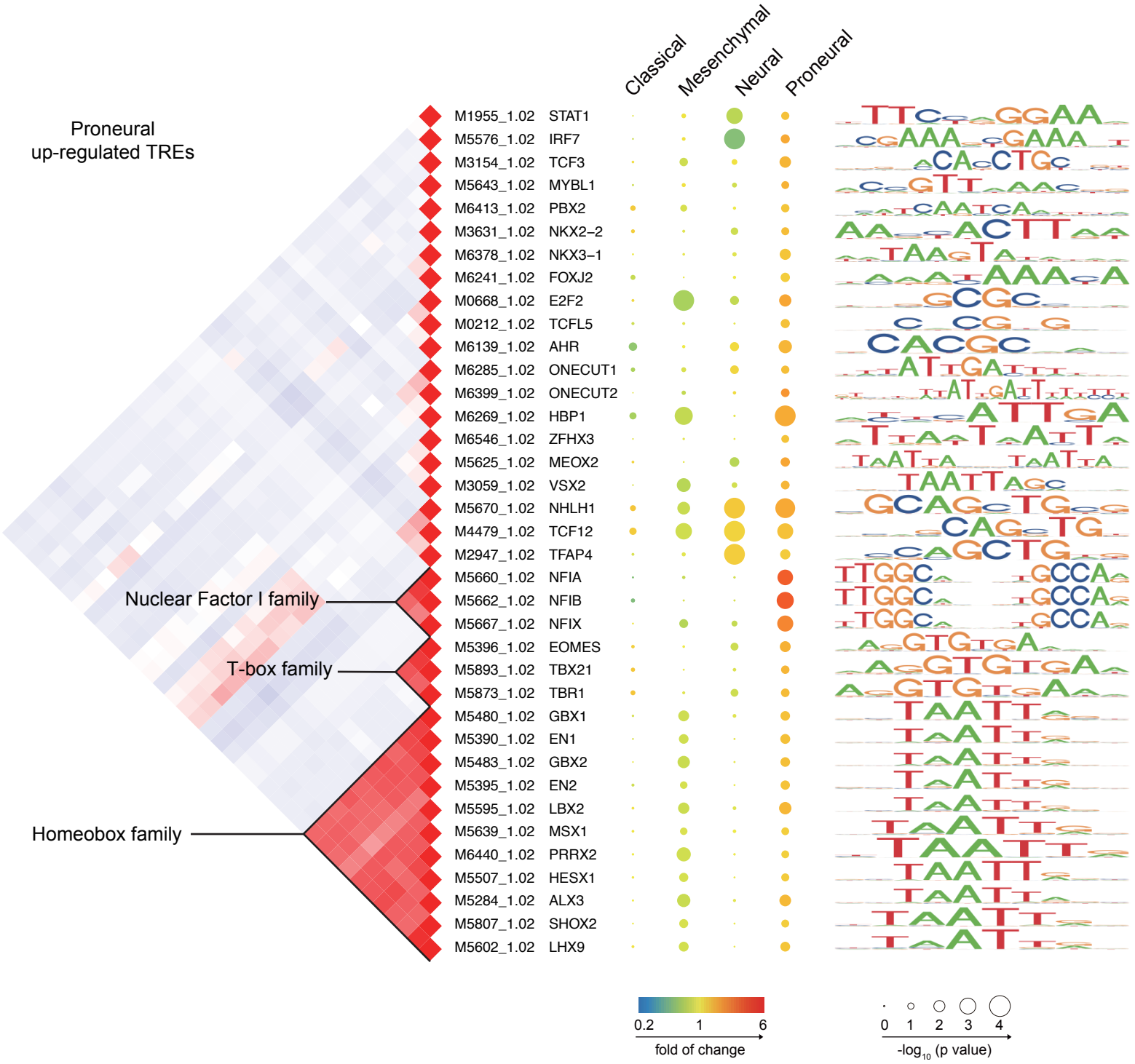

Classical
Mesenchymal
Neural
Proneural
Classical
up-regulated TREs
M6449_1.02 REL
M0442_1.02 ZBTB3
M3516_1.02 LMO2
M6468_1.02 SNAI1
M6396_1.02 NR5A2
M4509_1.02 POU5F1
M3009_1.02 POU3F2
M6289_1.02 HOXA9
M6471_1.02 SOX13
M0177_1.02 TWIST2
M0224_1.02 ATOH7
Classical
down-regulated TREs
TALE homeodomain 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)

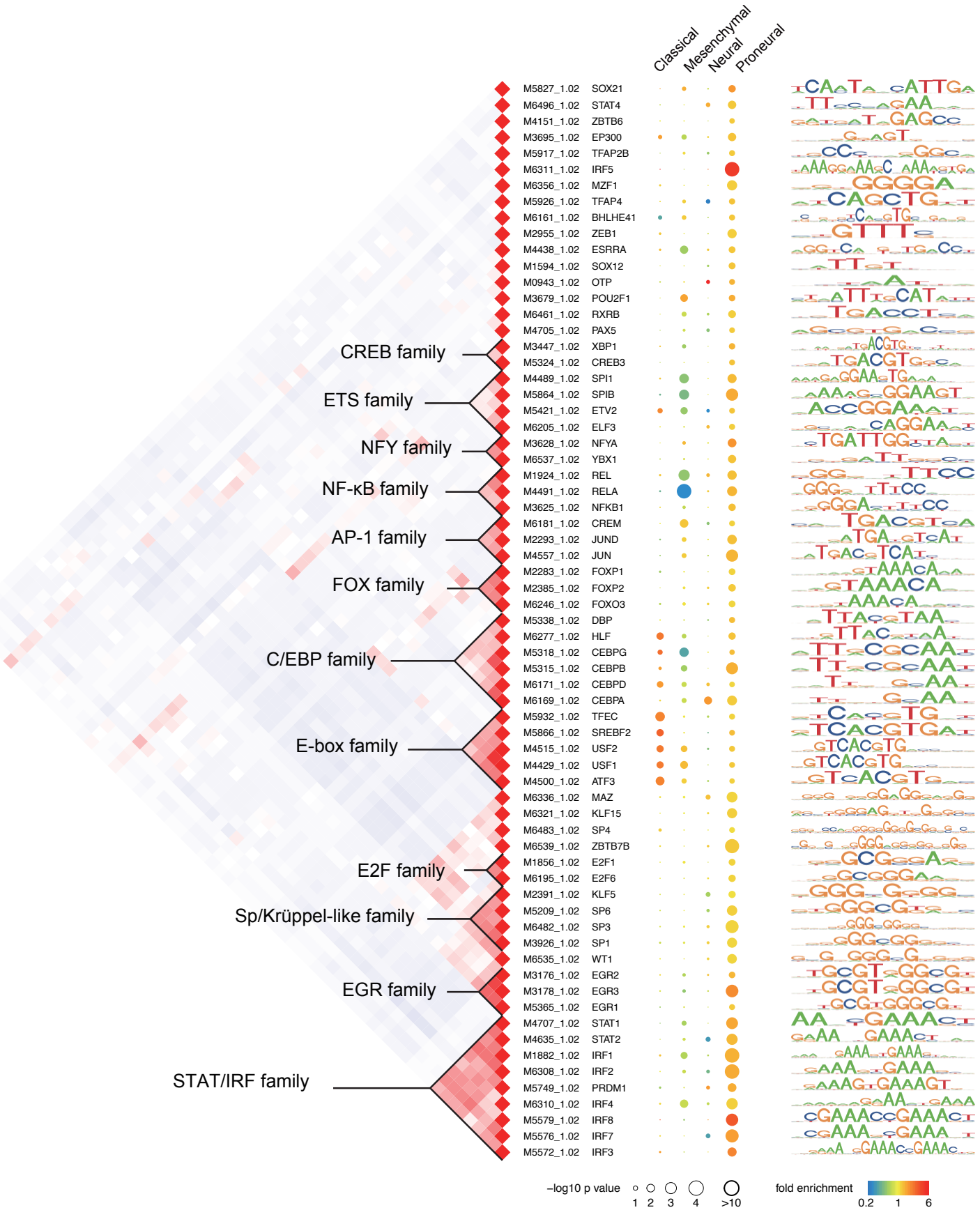

Mesenchymal
up-regulated TREs
Classical
Mesenchymal
Neural
Proneural
TALE homeodomain family
NFAT family
TEAD family
SMAD family
STAT family
RUNX family
Sp/Krüppel-like family
HSF family
nuclear receptors
retinoic acid receptors
MAF / CNC family
NF-κB family
CREB family
C/EBP family
CREB family
FOX family
ETS family
AP-1 family
nuclear receptors
E-box family
−log10 p value
1 2 3 4 >10
fold enrichment
0.2 1 6


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

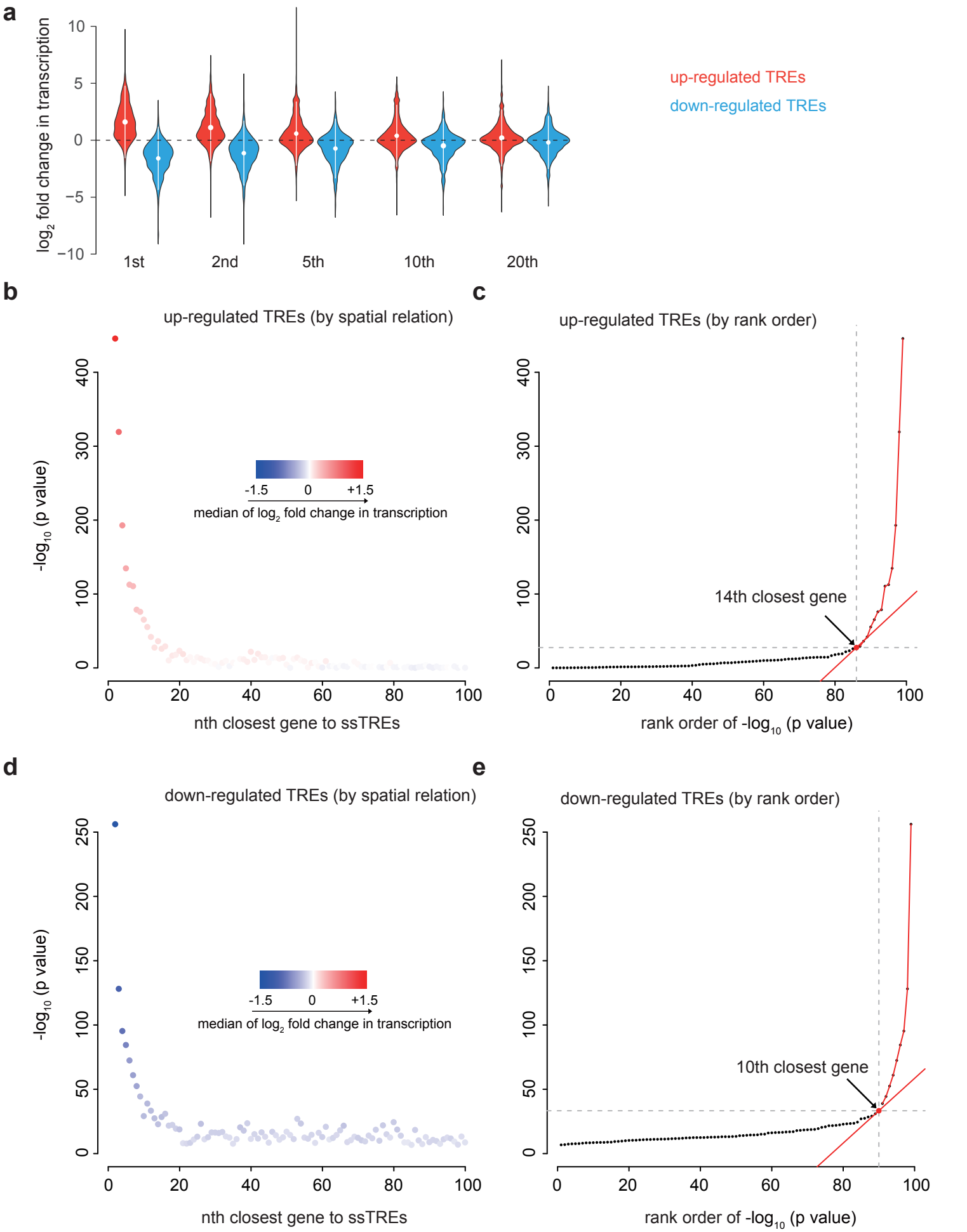

Neural
up-regulated TREs
Classical
Mesenchymal
Neural
Proneural
M6149_1.02 ARID5B
M6285_1.02 ONECUT1
M1582_1.02 HMG20B
M6509_1.02 TEAD4
M6473_1.02 SOX17
M1901_1.02 RREB1
M1418_1.02 C11orf9
M6139_1.02 AHR
M4574_1.02 REST
M3022_1.02 CUX1
M4840_1.02 FEZF1
M6516_1.02 TCF3
M3359_1.02 NR3C1
M6463_1.02 SMAD1
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
GATA family
SOX family
TALE homeodomain family
FOX family
SOX family
TCF/LEF family
NFY family
FOX family
–log10 p value
1 2 3 4 >10
fold enrichment
0.2 1 6


Neural
down-regulated TREs
TEAD family
C/EBP family
POU family
Homeobox family
AP-1 family
FOX family
Classical
Mesenchymal
Neural
Proneural
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

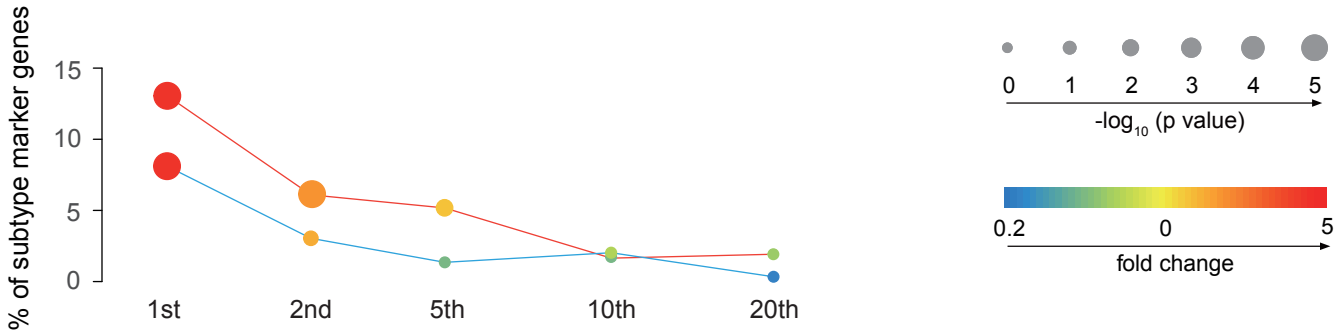

Proneural
up-regulated TREs
Classical
Mesenchymal
Neural
Proneural
M1955_1.02 STAT1
M5576_1.02 IRF7
M3154_1.02 TCF3
M5643_1.02 MYBL1
M6413_1.02 PBX2
M3631_1.02 NKX2–2
M6378_1.02 NKX3–1
M6241_1.02 FOXJ2
M0668_1.02 E2F2
M0212_1.02 TCFL5
M6139_1.02 AHR
M6285_1.02 ONECUT1
M6399_1.02 ONECUT2
M6269_1.02 HBP1
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
Nuclear Factor I family
T-box family
Homeobox family
0.2
1
6
fold of change
0
1
2
3
4
-log10 (p value)



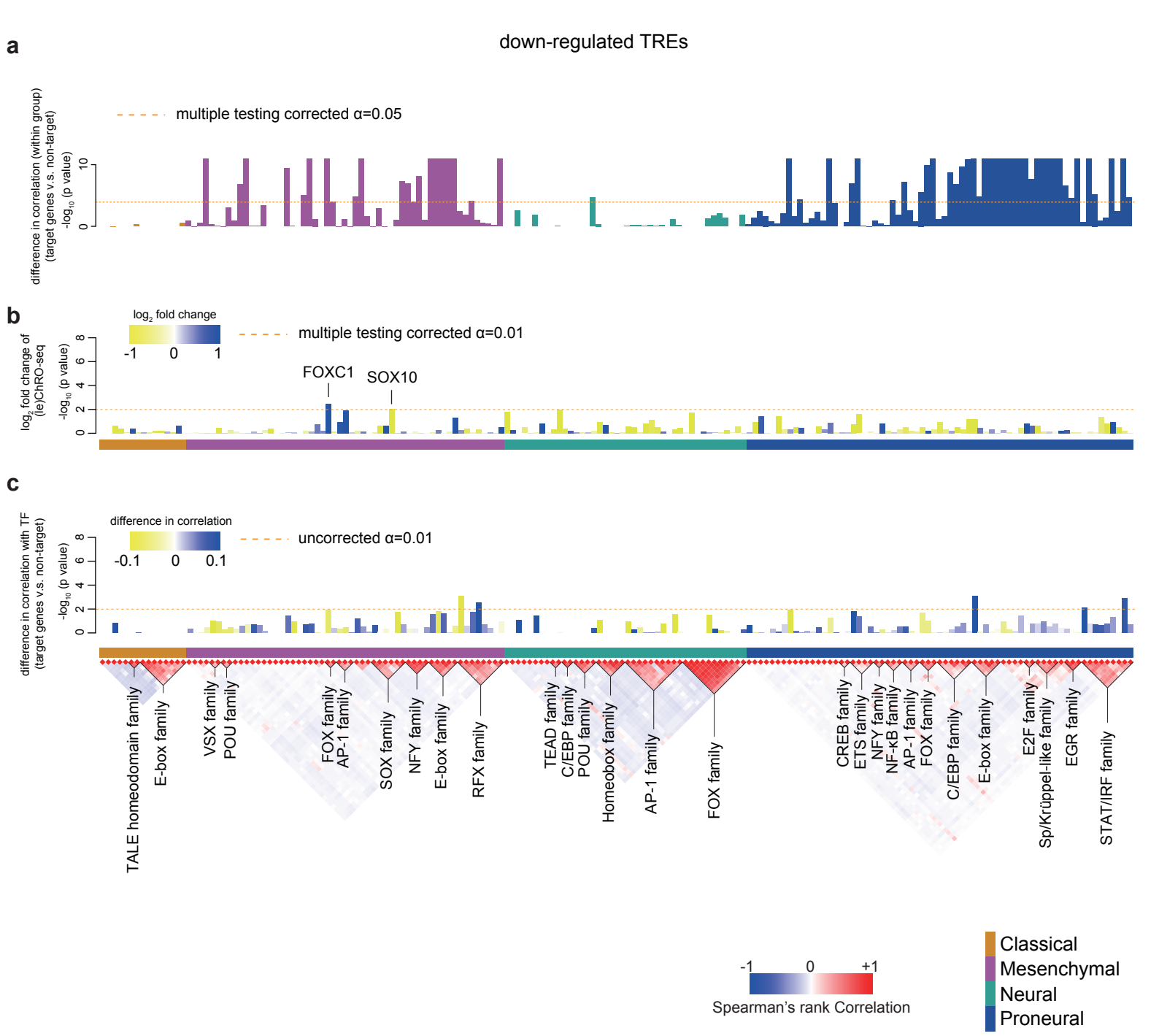
Supplementary Fig. 21. Transcription factor binding motifs enriched in TREs up-regulated or down-regulated in each known molecular subtype. Transcription factor binding motifs enriched in TREs that were up- or down-regulated in the indicated subtype. The Spearman's rank correlation heatmap (left) shows 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 4.



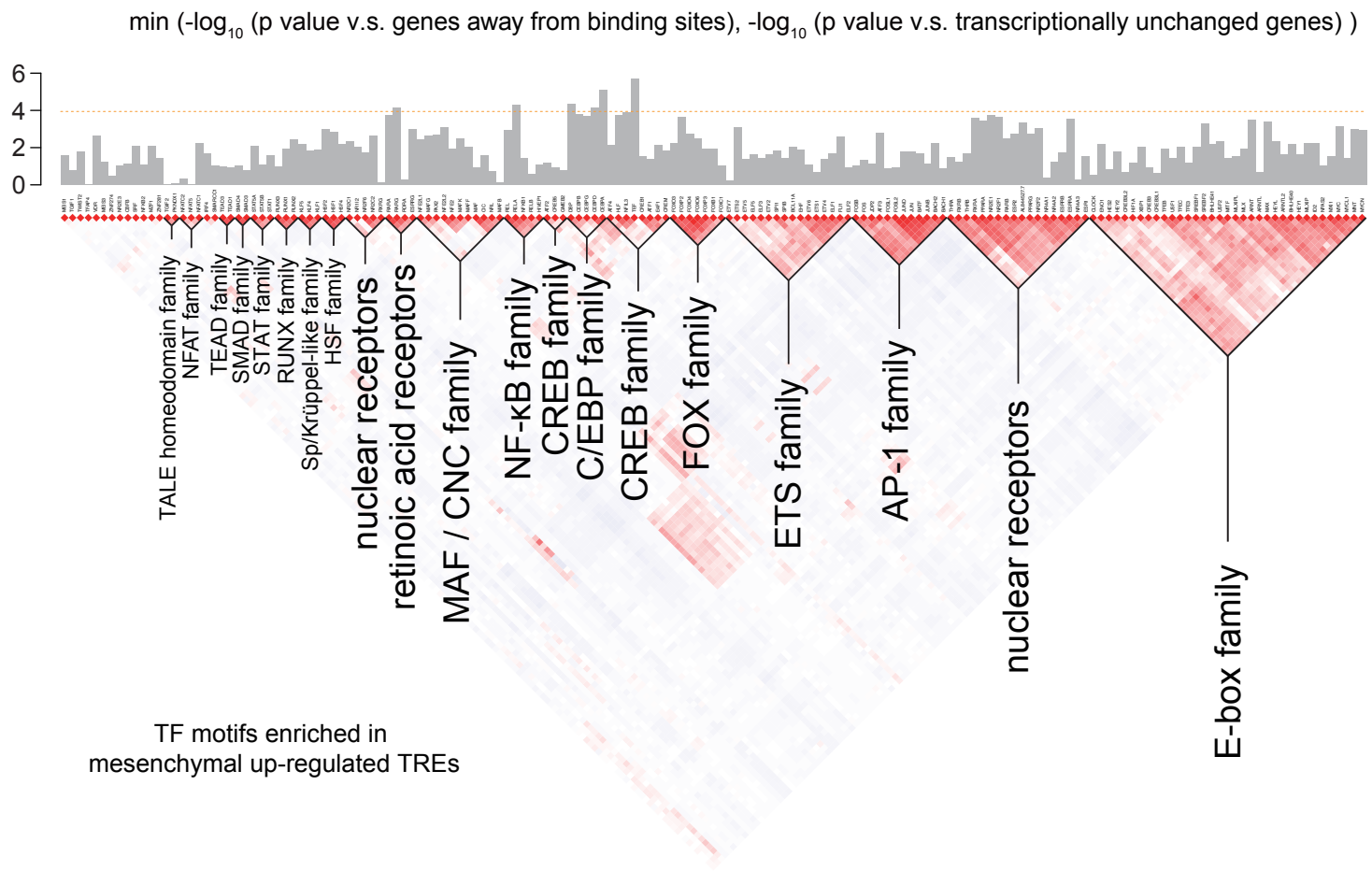
Supplementary Fig. 22. Subtype-biased TREs correlate with the transcription of nearby genes. (**a**) Violin plots show the distribution of $\log_2$ fold change in the transcription of n th closest genes to TREs that were up (red, N=4,960) or down (blue, N=1,815) -regulated in any subtype. White dots 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).



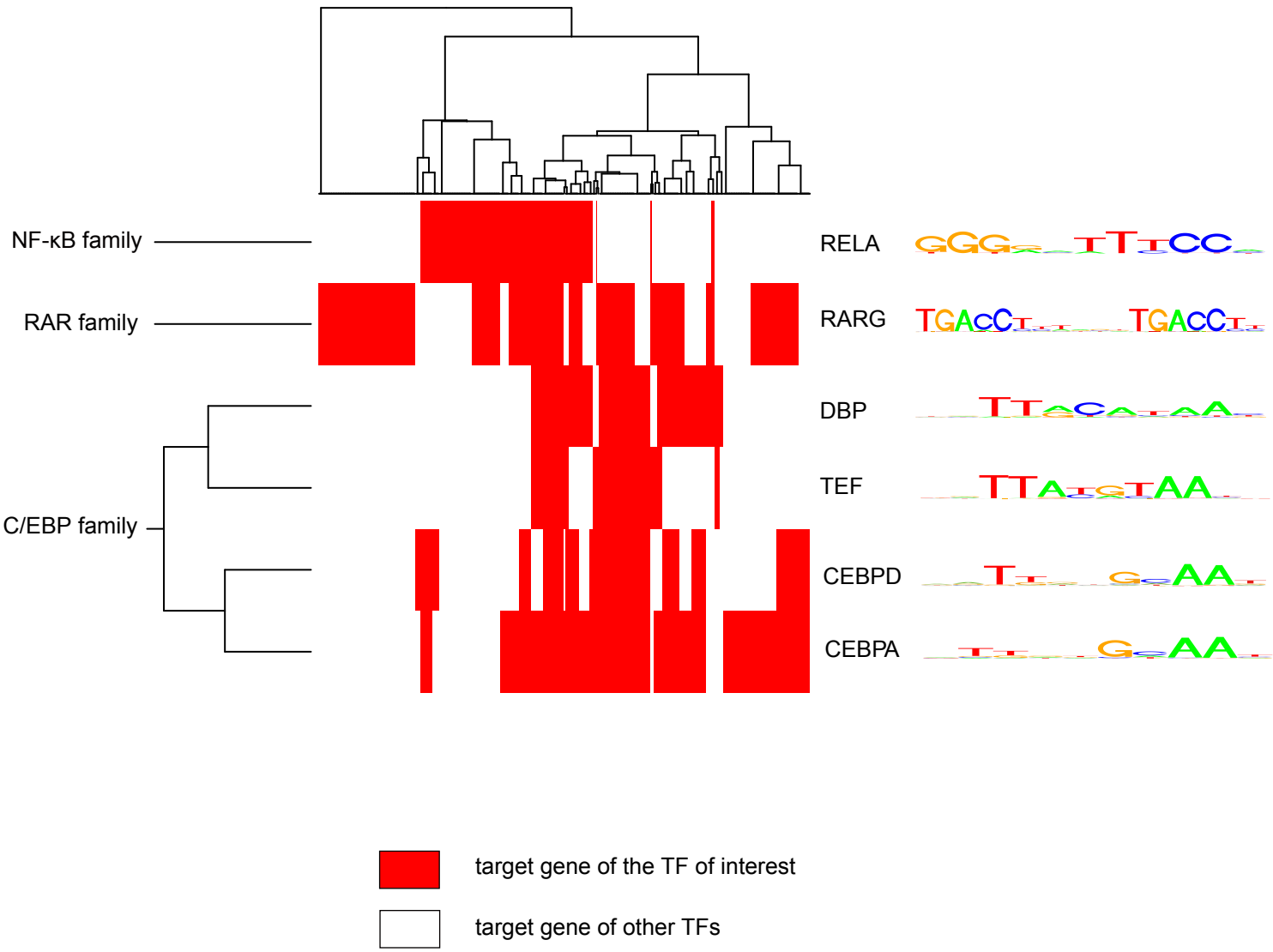
Supplementary Fig. 23. Subtype-biased TREs are near a large proportion of subtype specific genes. Line chart show the percentage of subtype marker genes (Y-axis) positioned n genes from the closest subtype-biased TREs. Separate lines are shown for up (red, N=4,960) or down (blue, N=1,815) -regulated 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.



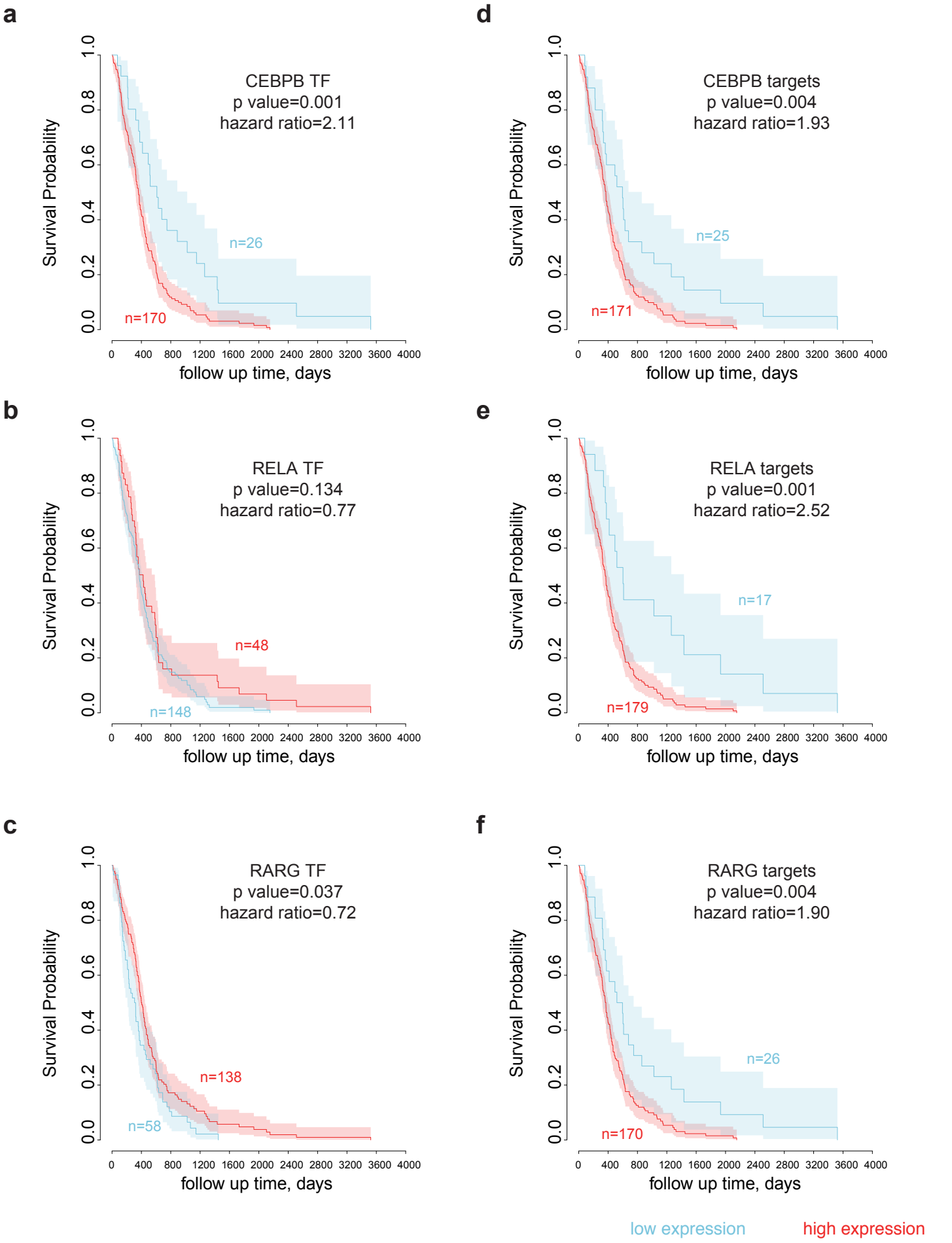
Supplementary Fig. 24. Barplots show the relationship between transcription factors enriched over TREs down-regulated in each subtype and their putative target genes. (**a**) Barplots show the -log$_{10}$ Wilcoxon rank sum of p value of having higher correlation among 174 TCGA patients between target 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



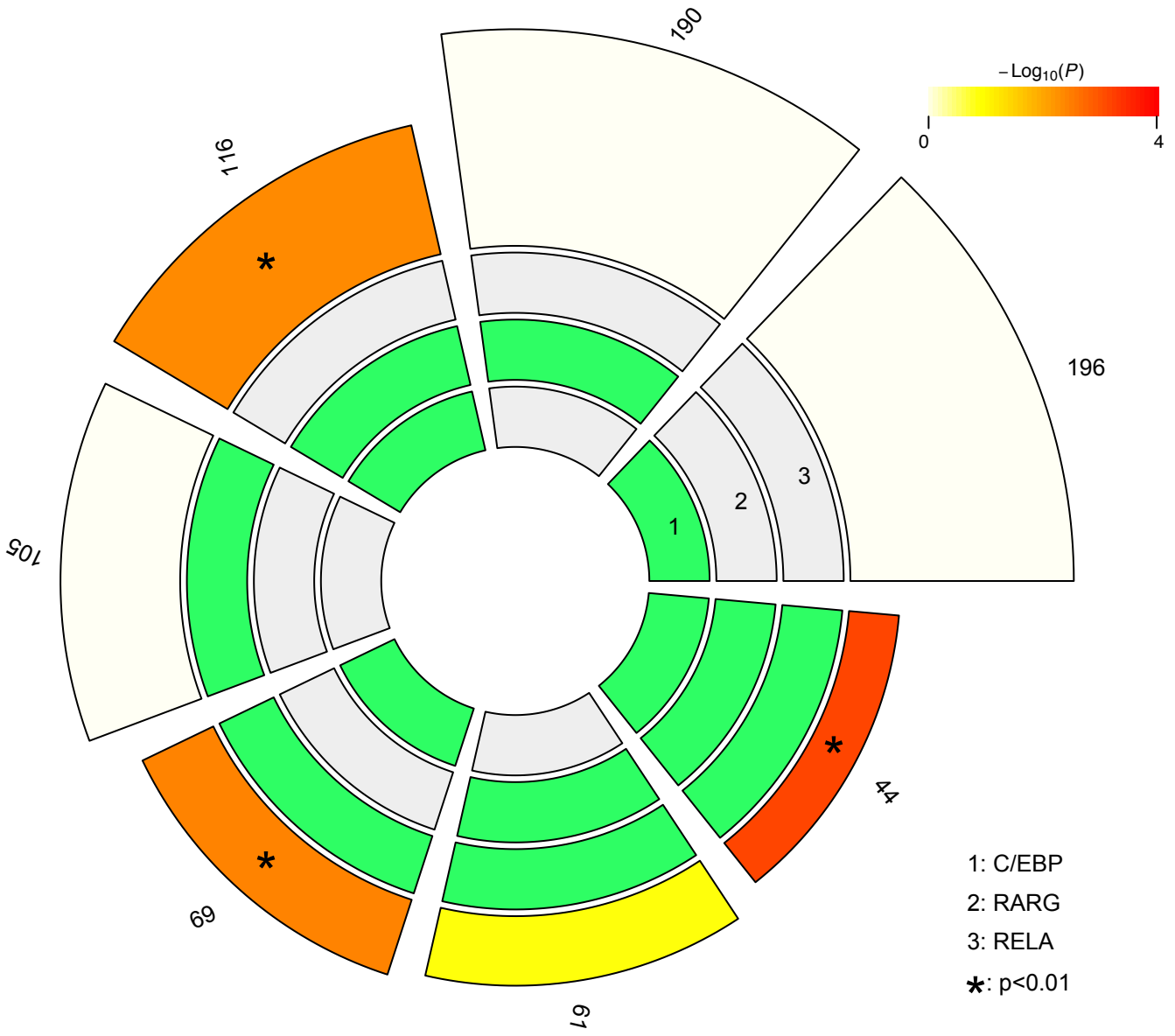
Supplementary Fig. 25. Barplots show transcription factor binding motifs controlling survival-related genes in mesenchymal GBMs. The minimum of the two -$\log_{10}$ p values on the x-axis and y-axis of figure 7a (two-sided Wilcoxon rank sum test) are plotted by the order of motifs cluster. In total, 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.



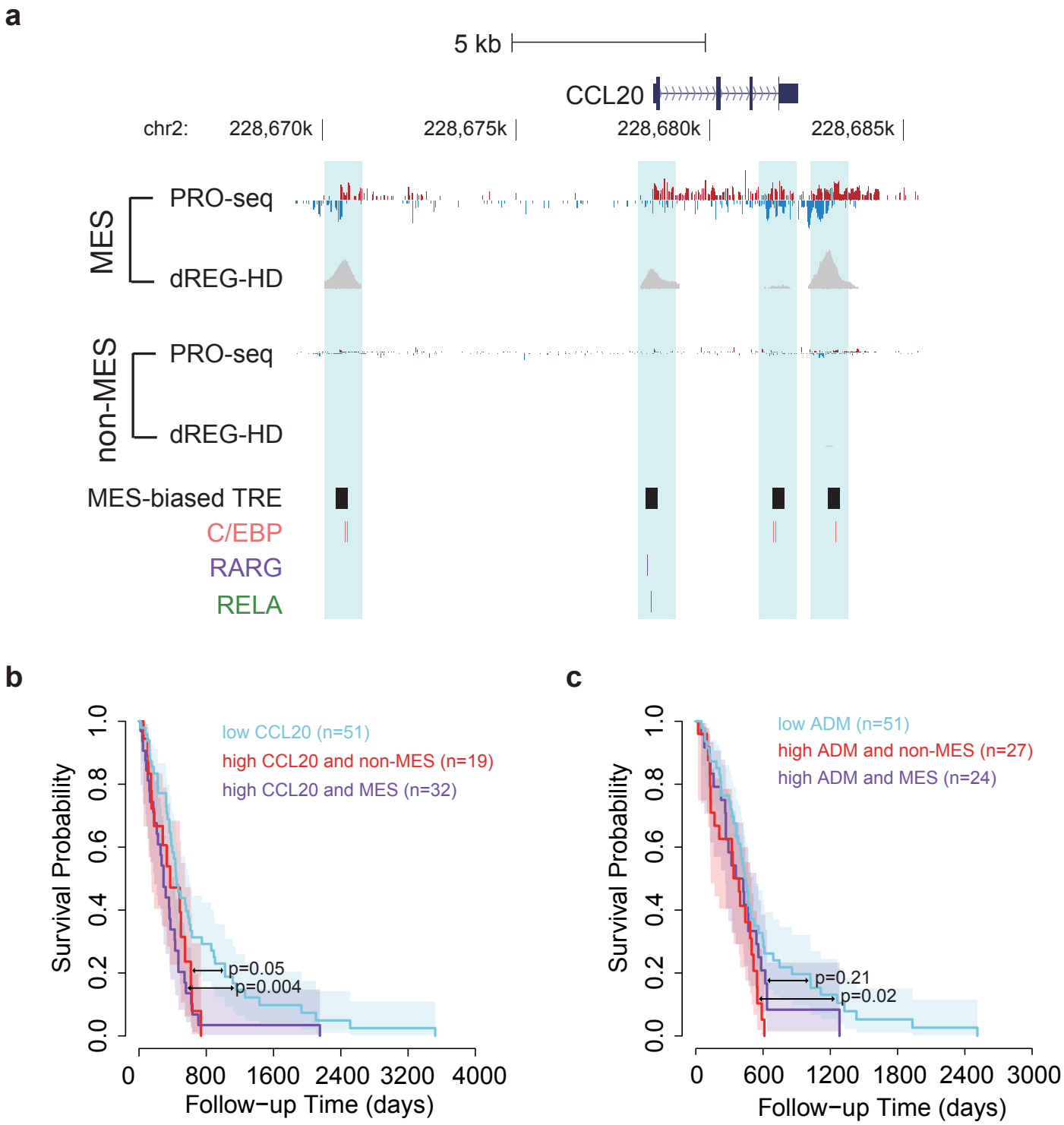
Supplementary Fig. 26. Heatmap shows the clustering of target genes of six transcription factors with significant survival association. Hierarchical agglomerative clustering groups target genes of one or more transcription factor. Red indicates the target gene belongs to the putative targets of the corresponding transcription factor and white indicates otherwise.



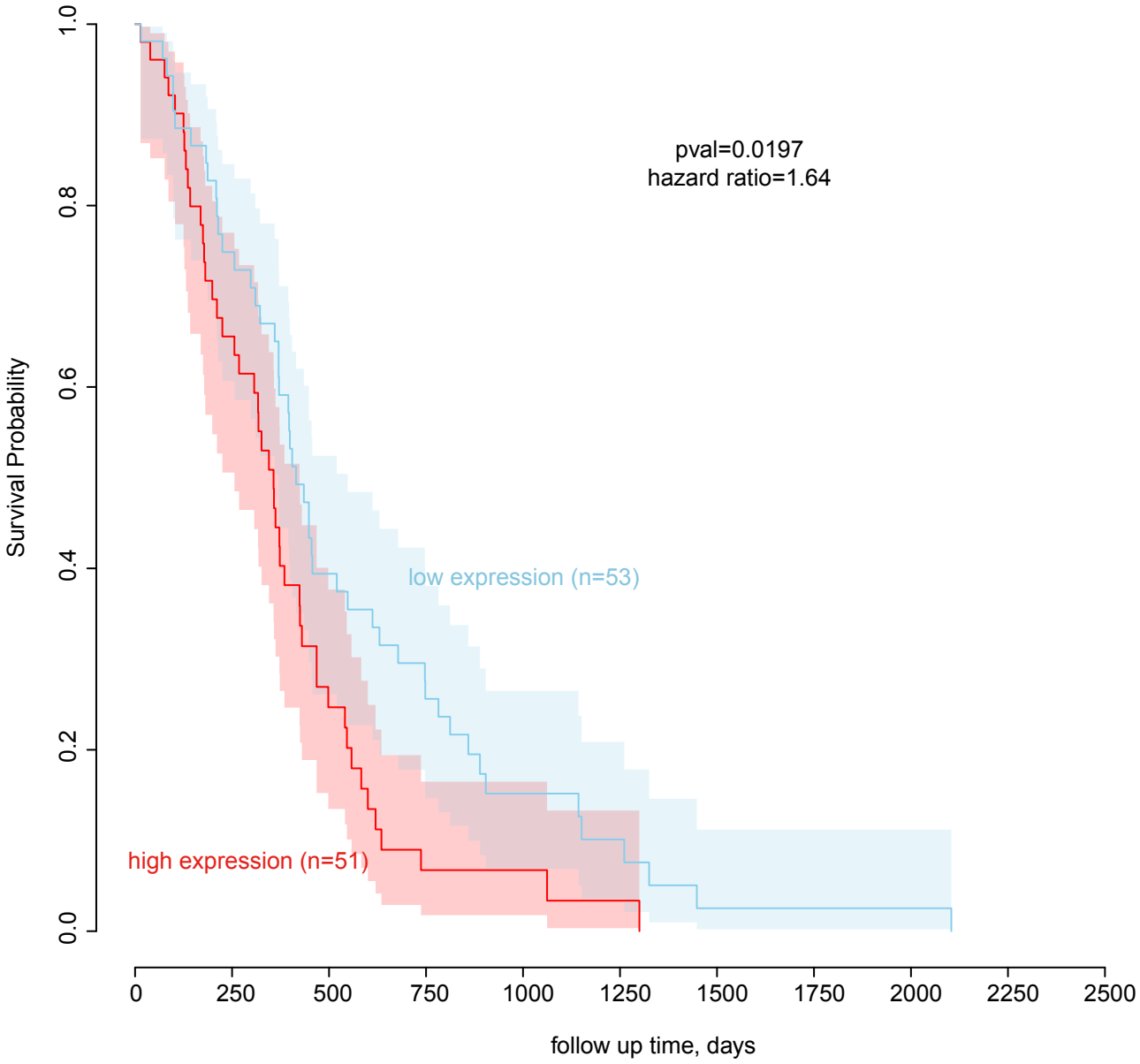
Supplementary Fig. 27. Kaplan–Meier plots show the difference in survival between patients with different expression levels of transcription factors (a-c) and of their corresponding target genes (d-f). P values and hazard ratios were calculated by comparing patients of higher expression level (red) with 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. 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. 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. 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.

Supplementary Table Legends:

Supplementary tables are provided in the associated attached Excel worksheet document. Please see this document for each supplementary table included with this paper.

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

Supplementary Table 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 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.

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

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

Supplementary Table 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 gene is not measured by the microarray data.

Supplementary Table 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).

Supplementary References

Bhat, Krishna P. L., Veerakumar Balasubramaniyan, Brian Vaillant, Ravesanker Ezhilarasan, Karlijn Hummelink, Faith Hollingsworth, Khalida Wani, et al. 2013. “Mesenchymal Differentiation Mediated by NF-κB Promotes Radiation Resistance in Glioblastoma.” *Cancer Cell* 24 (3): 331–46.

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* 155 (2): 462–77.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. “Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses.” *Science* 351 (6277): 1083–87.

Core, Leighton J., Joshua J. Waterfall, and John T. Lis. 2008. “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.” *Science* 322 (5909): 1845–48.

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* 50 (2): 212–22.

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* 12 (5): 433–38.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.

Hastie, Trevor, Rahul Mazumder, Jason Lee, and Reza Zadeh. 2014. “Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares.” *arXiv [stat.ME]*. arXiv. http://arxiv.org/abs/1410.2596.

Khodor, Yevgenia L., Joseph Rodriguez, Katharine C. Abruzzi, Chih-Hang Anthony Tang, Michael T. Marr 2nd, and Michael Rosbash. 2011. “Nascent-Seq Indicates Widespread Cotranscriptional Pre-mRNA Splicing in Drosophila.” *Genes & Development* 25 (23): 2502–12.

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* 339 (6122): 950–53.

Luo, Xin, Minho Chae, Raga Krishnakumar, Charles G. Danko, and W. Lee Kraus. 2014. “Dynamic Reorganization of the AC16 Cardiomyocyte Transcriptome in Response to TNFα Signaling Revealed by Integrated Genomic Analyses.” *BMC Genomics* 15 (1): 155.

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* 161 (3): 541–54.

Schwalb, Björn, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. 2016. “TT-Seq Maps the Human Transient Transcriptome.” *Science* 352 (6290): 1225–28.

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* 17 (1): 98–110.

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* 32 (1): 42–

56.e6.
Wang, Zhong, Tinyi Chu, Lauren A. Choate, and Charles G. Danko. 2017. “Rgtsvm: Support Vector Machines on a GPU in R.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1706.05544.