# Chromatin run-on reveals nascent RNAs that differentiate normal and malignant brain tissue

Tinyi Chu[1,2], Edward J. Rice[1,3], Gregory T. Booth[4], H. Hans Salamanca[5], Zhong Wang[1], Leighton J. Core[6], Sharon L Longo[7], Robert J. Corona[8], Lawrence S. Chin[7], John T. Lis[4], Hojoong Kwak[4,*], and Charles G. Danko[1,3,*]

[1] Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

[2] Graduate field of Computational Biology, Cornell University, Ithaca, NY 14853.

[3] Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

[4] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853.

[5] Department of Anesthesiology, SUNY Upstate Medical University, Syracuse, NY 13224.

[6] Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut, USA

[7] Department of Neurological Surgery, SUNY Upstate Medical University, Syracuse, NY 13224.

[8] Department of Pathology, SUNY Upstate Medical University, Syracuse, NY 13224.

[*] **Address correspondence to:**

Charles G. Danko, Ph.D.
Baker Institute for Animal Health
Cornell University
Hungerford Hill Rd.
Ithaca, NY 14853
Phone: (607) 256-5620
E-mail: dankoc@gmail.com

Hojoong Kwak, M.D., Ph.D.
Dept. Molecular Biology & Genetics
Cornell University
316 Biotechnology Bldg
Ithaca, NY 14853
Phone: (607) 255-7920
E-mail: hk572@cornell.edu

## Main text:

**Non-coding elements in our genomes that play critical roles in complex disease are frequently marked by highly unstable RNA species[1–4]. Sequencing nascent RNAs attached to an actively transcribing RNA polymerase complex can identify unstable RNAs[5–10], including those templated from gene-distal enhancers (eRNAs)[10–14]. However, nascent RNA sequencing techniques remain challenging to apply in some cell lines and especially to intact tissues, limiting broad applications in fields such as cancer genomics and personalized medicine. Here we report the development of chromatin run-on and sequencing (ChRO-seq), a novel run-on technology that maps the location of RNA polymerase using virtually any frozen tissue sample, including samples with degraded RNA that are intractable to conventional RNA-seq. We used ChRO-seq to develop the first maps of nascent transcription in 23 human glioblastoma (GBM) brain tumors and patient derived xenografts. Remarkably, >90,000 distal enhancers discovered using the signature of eRNA biogenesis within primary GBMs closely resemble those found in the normal human brain, and diverge substantially from GBM cell models. Despite extensive overall similarity, 12% of enhancers in each GBM distinguish normal and malignant brain tissue. These enhancers drive regulatory programs similar to the developing nervous system and are enriched for transcription factor binding sites that specify a stem-like cell fate. These results demonstrate that GBMs largely retain the enhancer landscape associated with their tissue of origin, but selectively adopt regulatory programs that are responsible for driving stem-like cell properties.**

We developed Chromatin Run-On and sequencing (ChRO-seq), a new method to map RNA polymerases in whole cells or tissue samples (Fig. 1a). ChRO-seq avoids the difficulty in obtaining nuclei by optimizing a run-on reaction in easily pelleted chromatin fractions that contain enzymatically active RNA polymerase (see Methods). The run-on incorporates a biotinylated nucleotide triphosphate (NTP) substrate into the existing nascent RNA that provides a high-affinity tag used to enrich nascent transcripts >10,000-fold[5]. The biotin group effectively 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 single-nucleotide resolution for the polymerase active site when sequenced from the 3' end of the RNA[7].

We validated our approach by performing matched ChRO-seq and PRO-seq experiments in the human Jurkat T-cell leukemia line and observed highly correlated levels of RNA polymerase in the bodies of mRNA encoding genes (R= 0.98; Fig. 1b, Supplementary Fig. 1) and sites of promoter-proximal paused RNA polymerase II (Pol II) (R= 0.96; Fig. 1c). The microRNA MIR181 locus illustrates the advantages of ChRO-seq compared with other molecular assays (Fig. 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 (Fig. 1d), providing deep insight into genome regulation that could not be resolved from publicly available data even after the integration of three separate molecular assays (RNA-seq, DNase-I-seq, and H3K27ac ChIP-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[15], 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, which achieved the lowest possible RNA integrity number (RIN = 1.0, **Supplementary Fig. 2**), thus precluding the application of RNA-seq methods optimized for degraded samples (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 RNA polymerases and a mix of biotinylated-NTP and normal NTPs to extend degraded nascent RNA transcripts (**Fig. 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[16], run-on samples achieved a longer RNA length distribution that was better suited for mapping unique reads within the human genome (**Fig. 1e**). Degrading RNA by pretreatment of chromatin with RNases in Jurkat T-cells was used to verify that leChRO-seq produced maps of transcription that are highly similar to those obtained using ChRO-seq (**Supplementary Fig. 3a-b**). Thus, leChRO-seq allows the robust interrogation of archival tissue samples which cannot be analyzed using standard genomic tools.

Mapping transcriptional enhancers and their effect on gene expression using ChRO-seq has clear implications for understanding the basis of complex diseases. We therefore set out to establish the utility of these tools in directly analyzing clinical isolates. To complement the archival GBM, we collected ChRO-seq data from nonmalignant brain tissue and from 22 additional GBM isolates: primary glioblastomas from 19 additional patients and passages of three patient derived xenografts (PDX) (**Supplementary Table 1**). We sequenced each GBM to an average depth of 33 million reads per sample that were uniquely mapped to the hg19 reference genome (10-150M reads/ sample). leChRO-seq data was highly correlated between replicates, comprised of separate biopsies isolated from nearby regions, available for GBM-88-04 and for the nonmalignant brain (**Supplementary Fig. 3c-f**). Histopathology analysis of GBM-15-90 revealed hallmarks of a highly aggressive, grade IV malignant astrocytoma (**Supplementary Fig. 4a-d**). Furthermore, ChRO-seq data revealed 3- to 84-fold changes in the transcription of GBM driver genes relative to nonmalignant brain tissue, such as EGFR (**Fig. 2a**), which are previously reported inside somatic copy number alterations in GBM[17,18].

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 most similar to the other three parts of the tumor (**Supplementary Fig 5**), suggesting that the tumor originated within the core and grew outward radially, as expected.

We focused initial analyses on gene and ncRNA transcription within the primary GBMs. Analysis of annotated genes demonstrated that GBMs from our cohort represent each of the four molecular subtypes previously reported in GBM[19] (**Fig. 2b,Supplementary Fig. 6**). Though most tumors primarily shared expression patterns with one dominant molecular subtype, several tumors in our cohort were similar with multiple subtypes, especially those matching classical and mesenchymal signatures, consistent with reports of cellular heterogeneity within the same tumor[20]. Globally, differential gene transcription analysis identified 1,343 transcriptional changes that distinguish all primary GBMs from multiple replicates of the nonmalignant brain sample ($p$ < 0.01, DESeq2[21]). Genes undergoing transcriptional differences in all GBMs were enriched in biological processes related to cell cycle (p = 5.20E-05, Bonferroni corrected Fisher's exact test), DNA replication / metabolic processes (p = 3.07E-02 and 7.02E-03, Bonferroni corrected Fisher's exact test), and developmental processes (p = 2.79E-02, Bonferroni corrected Fisher's exact test). 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. 2c, Supplementary Fig. 7**). We also discovered multiple ncRNAs whose transcription levels were changed reproducibly across tumors, including the primary transcription unit encoding MIR29A, AC016831.7, and PVT1 (**Fig. 2c, blue**), which confer growth advantages to U87 glioblastoma cells[22–25].

Both active promoters and enhancers, collectively called transcriptional regulatory elements (TREs), have a characteristic pattern of divergently-oriented paused RNA polymerase. Genome-wide run-on transcription assays are therefore a sensitive way to discover the location and activity of TREs, which overlap those defined by acetylation of histone 3 at lysine 27 (H3K27ac)[11–13]. 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 to identify TREs. The dREG-HD algorithm improves the resolution of dREG[12] by imputing smoothed DNase-I-seq signal intensity, and identifies sites initiating transcriptional activity with 80% sensitivity at >90% specificity (**Supplementary Fig. 8**). dREG-HD recovered the nucleosome depleted region in histone modification ChIP-seq and MNase-seq data (**Supplementary Fig. 9**), demonstrating that it has substantially higher resolution compared with our dREG tool alone.

The vast majority (96%) of TREs identified by dREG-HD in our 20 primary GBM samples were DNase-I hypersensitive in at least one of the 216 reference tissues analyzed by ENCODE or Epigenome Roadmap[26,27] (**Fig. 3a-b**). Rare enhancers provide a unique "fingerprint" for quantitatively evaluating the similarity between two samples[28], and could be used to define the relationship between tumors and normal tissue. We therefore developed a strategy that compares active enhancer landscapes obtained using dREG-HD with DHSs in 921 public datasets representing 216 reference tissues (see **Online Methods**). Our strategy consistently discovered the expected cell lines (**Supplementary Fig. 10**), even identifying the expected genotype (GM12878) among all lymphoblastoid cell lines as the most similar to GM12878 PRO-seq data (**Supplementary Fig. 10b**). Remarkably, primary GBM samples have enhancer landscapes that are highly similar to normal brain reference samples, especially to samples derived from the cerebral cortex, and dissimilar to three different well-established GBM cell models (**Fig. 3c, Supplementary Fig. 11**). In GBM-15-90, for instance, 86% of TREs were shared with primary brain tissue (**Supplementary Fig. 12**), which was greater similarity than observed in either GBM cell lines (62% TRE identity) or *in vitro* cultured primary brain cells (75%).

We asked whether contamination of the GBM with normal brain tissue explained the extensive similarity with normal brain reference samples. To rigorously evaluate this hypothesis 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[29], 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 and highly different from glioma model cell lines (**Fig. 3c**). Consistent with this result, GBM cell lines were recognized correctly when a simulated mixture contained as little as 10-20% of any one of the three glioma lines and 80-90% normal brain tissue (see Online Methods, **Supplementary Fig. 13**). Normal brain contamination would therefore have to be higher than 80% to explain the observed differences between primary tumor and cell culture models, much more than the 15% estimated for a typical GBM[30]. Thus, our results suggest that transformed GBM cells largely retain the enhancer signature associated with their cell of origin.

Two models might explain differences in enhancer profiles between primary and cultured GBM cells. Differences might reflect either evolutionary changes in TREs 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 reliable model for many of the transcriptional features associated with primary tumors. That primary astrocytes 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.

We hypothesized that TREs which were transcribed in tumor tissue, but were not DHSs in the normal brain samples, control the malignant phenotype of the tumor. Such tumor-associated TREs (taTREs) comprised 12% of TREs in each tumor (range: 2-24%, **Fig. 4a, Supplementary Table 2**). In contrast to TREs in the normal brain (nbTREs), the majority of these taTREs were distal to annotated transcription start sites, even those that were recurrently discovered across the majority of clinical samples in our cohort of primary GBMs (**Fig. 4A, green line**), suggesting that those which are functional are most likely distal enhancers. A small number of taTREs were recurrently activated across multiple tumors. For example, a taTRE ~12 kb downstream of the gene encoding the Engrailed 2 homeobox (EN2) was recurrently activated in 13/ 23 tumors, likely reflecting the importance of this TRE in GBM biology (**Supplementary Fig. 15**).

We developed a statistical test to identify which tissues shared unexpectedly high overlap with taTREs identified in each tumor (**Supplementary Table 3**) (see **Online Methods**). Unbiased hierarchical clustering of the taTREs between 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. 4b, Supplementary Fig. 16**). taTREs most strongly 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. 4b**, 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 (**Fig. 4b**). We emphasize that activation of these separate transcriptional regulatory programs may reflect gene expression changes in subsets of cells within the tumor. Whereas 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[31], overlap with astrocytes, endothelial cells, fibroblasts, or osteoblasts may capture tumor cells that have trans-differentiated into these lineages[32,33].

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 binding motifs[34]. We identified POU domain containing and SOX family transcription factors enriched in taTREs in the stem-like regulatory program of 57-78% of tumors (**Fig. 4c and Supplementary Fig. 17**). To verify that these enrichments reflect bona-fide binding of the predicted transcription factors, we obtained ChIP-seq data from cultured glioma neurospheres[31]. As predicted, taTREs in the stem-like program were enriched in both ChIP-seq reads and peak calls for both POU3F2 and SOX2 (**Supplementary Fig. 18 and 19**). The differentiated support cell program was highly enriched for binding of activating protein 1, a heterodimer of the transcription factors FOS and JUN (JUND motif shown), as well as a motif strongly resembling heat shock factor 1 (HSF1) (**Fig. 4c**). We also discovered several transcription factors that were enriched in taTREs in both fetal and differentiated regulatory programs, including *STAT*, *SRF*, and *MEOX2*. Many of these motifs were enriched in only a subset of our cohort of tumors, typically 40-60%, suggesting that these transcription factors may contribute to inter-tumor heterogeneity in transcriptional regulation. 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.

Mapping transcriptional enhancers and their effect on gene expression has clear implications for understanding the molecular basis of complex disease and designing targeted therapies. The introduction of ChRO-seq extends the analysis of nascent transcription to virtually any sample that maintains the integrity of protein-DNA interactions – even those whose RNA is degraded. Surprisingly, ChRO-seq revealed that malignant brain tissue largely retains enhancers that were DNase-I hypersensitive in the tissue of origin. A rare population of ectopic enhancers resembled a stem-cell like regulatory program with particularly strong overlap compared with fetal tissues isolated from the nervous system, as well as overlap with differentiated support cells. Our observations are the first to identify regulatory enhancers involved in the stem-like transformation directly from genomic analysis of primary tumors. Using this information we identified key transcription factors that are likely to play a role in the specification of each regulatory program. Notably, several of our predictions were consistent with other lines of experimental evidence, especially for transcription factors involved in maintaining stem-like properties of tumor propagating cells[31]. Our strategy can now be deployed to identify transcriptional regulatory programs that contribute to a myriad of solid tumors and other tissues which have proven challenging to study using existing epigenomic tools.

# Online 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% $CO_2$. 1x$10^6$ 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 and approval by the Institutional Review Board at SUNY Upstate Hospital, Syracuse, NY. 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 [35,36]. 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.

**Table of key reagents in chromatin isolation**

| Chemicals | SOURCE | IDENTIFIER |
|---|---|---|
| RPMI-1640 | Corning | 10-040-CV |
| Fetal Bovine Serum (FBS) – Premium, Heat-Inactivated | Atlanta Biologicals | S11195H |
| 100X Penicillin/Streptomycin Antibiotic | Corning | 30-002-CI |
| 50 mg/ml Gentamicin Antibiotic | Corning | 30-005-CR |
| $MgAc_2$ | | |
| SUPERase In RNase Inhibitor | Life Technologies | AM2694 |
| Complete, EDTA-Free Protease Inhibitor Cocktail Tablet | Roche | 11 873 580 001 |
| **Equipment** | **SOURCE** | **IDENTIFIER** |
| The Cellcrusher Tissue Pulverizer | Cellcrusher Limited | n/a |
| accuSpin Micor 17R Benchtop Centrifuge | Fisher Scientific | 13-100-676 |
| Diagenode Bioruptor | Diagenode | |
| **Experimental Models: Cell Lines** | **SOURCE** | **IDNETIFIER** |
| Jurkat | ATCC | TIB-152 |

| Experimental Models: Tissues | SOURCE | IDNETIFIER |
|---|---|---|
| Human Glioblastoma | SUNY Upstate Medical Center | n/a |

### Chromatin isolation

The chromatin isolation was based on work first described in ref[37]. 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 MgAc2 , 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.

**Table of Key Reagents in ChRO-seq**

| Chemicals | SOURCE | IDENTIFIER |
|---|---|---|
| 10 mM Biotin-11-CTP | Perkin Elmer | NEL542001EA |
| 10 mM Biotin-11-UTP | Perkin Elmer | NEL543001EA |
| Ribonucleotide Solution Set | NEB | N0450S |
| SUPERase In RNase Inhibitor | Life Technologies | AM2694 |
| Trizol LS | Life Technologies | 10296-010 |
| Trizol | Life Technologies | 15596-026 |
| Chloroform | Fisher | BP1145 1 |
| GlycoBlue | Ambion (Thermo Fisher) | AM9515 |
| T4 RNA Ligase 1 (ssRNA | NEB | M0204L |

| | | |
|---|---|---|
| Ligase) | | |
| RNA 5' Pyrophosphohydrolase (RppH) | NEB | M0356S |
| T4 Polynucleotide Kinase (PNK) | NEB | M0201L |
| 10 mM Adenosine 5'-Triphosphate (ATP) | NEB | P0756L |
| SuperScript III Reverse Transcriptase | Life Technologies | 18080044 |
| 100 mM Deoxynucleotide (dNTP) Solution Set | NEB | N0446S |
| Q5 High-Fidelity DNA Polymerase | NEB | M0491L |
| **Adapters & Primers** | **SOURCE** | **SEQUENCE** |
| Reverse 3' RNA Adaptor (Rev 3 – 6N) | IDT | /5Phos/NNNNNNGAUCGUCGGACUGUAC AACUCUGAAC<br>/3InvdT/ (Note: 6N's not in the original design) |
| Reverse 5' RNA adaptor (Rev5) | IDT | 5' CCUUGGCACCCGAGAAUUCCA 3' |
| RNA PCR Primer 1 (RP1) | IDT | 5' – AATGATACGGCGACCACCGAGATCTAC CGTTCAGA<br>GTTCTACAGTCCGA - 3' |
| RNA PCR Primer, Index 1 (RPI1) | IDT | 5' - CAAGCAGAAGACGGCATACGAGAT***CGT GAT***GTGACTGGAG<br>TTCCTTGGCACCCGAGAATTCCA - 3'<br>(Bar Code Index #1 underlined) |
| **Equipment** | **SOURCE** | **IDENTIFIER** |
| Micro Bio-Spin P-30 Gel Columns, Tris Buffer, RNase-free | Bio-Rad | 732-6250 |
| Hydrophilic Streptavidin Magnetic Beads | NEB | S1421S |

| | | |
|---|---|---|
| Mini-Tube Rotator | Fisher Scientific | 05-450-127 |
| MagneSphere Technology Magnetic Separation Stand | Promega | Z5342 |
| accuSpin Micro 17R Benchtop Centrifuge | Fisher Scientific | 13-100-676 |

### 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[38]. Briefly 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[38]. 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 of ChRO-seq and leChRO-seq sequencing reads*

We used our publicly available pipeline to align ChRO-seq and leChRO-seq data (https://github.com/Danko-Lab/utils/tree/master/proseq). Some libraries were prepared using adapters which contained a molecule-specific unique identifier (first 6 bp sequenced; denoted in **Table 2**), and for these we removed PCR duplicates using PRINSEQ lite [39]. Adapters were trimmed from the 3' end of remaining reads using cutadapt with a 10% error rate [40]. Reads were mapped with BWA [41] 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 [42] and the bedGraphToBigWig program in the Kent Source software package [43]. Downstream data analysis was performed using the bigWig software package, available from: https://github.com/andrelmartins/bigWig. All data processing and visualization was done in the R statistical environment [44].

### *Gene transcription activity quantification for ChRO-seq and leChRO-seq*

We quantified transcription activity of ChRO-seq and leChRO-seq data using gene annotations (GENCODE v25 lift 37). 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*

Transcription 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 (RPM normalized reads count+1). The similarity between each sample was measured by Spearman's rank correlation, and clustered using single link clustering. The subtype score was calculated by Pearson correlation with the centroid of corresponding subtype reported by (*n = 23*).

### *Differential expression analysis (DESeq2)*

Transcription activity of genes in each primary GBM / non-malignant brain were quantified by the methods described above. Patients clustered in each dominant subtype were treated as biological replicates (**Fig. 2b**). Two technical replicates of non-malignant brain were used as control. Differential expression analysis was conducted using deSeq2 (Love et al., 2014) and differentially expressed genes were defined as those with a false discovery rate (FDR) less than 0.05.

### *dREG-HD*

*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 [12], 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 e1071 R package, which is based on the libsvm SVM implementation.

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. 8b-c**).

*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[45]. 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 [27].

## ***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 [46] and Hotspot. 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.11** and **15**, 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. All dREG-HD runs were done at the stringent 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 samples (**Fig. 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. 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. To correct this bias, we normalized 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 [47], 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.

### *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 if 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).

### *Simulation of normal brain-contaminated sample*

To simulate a DHS dataset that mimics GBM cell lines with contamination from normal brain tissues, we created 9 pairs of cell / tissue combinations from 3 GBM cell lines (A172, H54 and M059J) and 3 normal brain tissues (globus pallidus, midbrain and frontal cortex). For each sample, mapped sequencing reads from either of the available biological replicates were pooled, and sampled at the frequency indicated to generate a range of contamination with normal brain, as indicated in **Supplementary Fig. 15**. After the simulated sample had been generated, DHSs were called using MACS2, and mutual information was calculated between the simulated mixture and all other samples as described above.

### *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 nonmalignant 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.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 $10^5$ 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 $p(X_{null} > x_{observed}) \leq 1/10^4$. 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*

*Defining genomic regions for motif enrichment comparison.* taTREs from the group indicated in the **Fig.4a** (positive set) were compared against normal brain TRE (background set). Normal brain TREs (nbTRE) were constructed from the dREG-HD sites that intersect with active DHSs peaks in the adult normal brain. 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 primary GBM and PDX [34]. We focused on 1,964 human transcription factor binding motifs from the CisBP database [48] and clustered similar motifs using an affinity propagation algorithm into 622 clusters separately for each sample.  For each cluster, we selected the transcription factor with the highest expression value (measured from ChRO/leChRO-seq) in that sample to represent motifs for each cluster. After selecting motifs in each sample, we merged the set of motifs chosen in at least one sample. When scanning genomic regions of interest, we used TFBSs having a $\log_e$-odds score ≥10 in positive and background sets, with scores obtained by comparing each representative motif model to a third-order Markov background model. Motif enrichment was tested using Fisher's exact test with a Bonferroni correction for multiple hypothesis testing. To account for potential bias resulted from difference in GC-content between positive and background sets, we ran statistical test on 100 independently subsampled GC-matched dREG-HD regions, and summarized the corrected p values by false discovery rate (FDR) using the fdrtool package in R [49], and fold of enrichment across background sets by the median across samples. Motifs that show enrichment in one of the 23 samples (all taTRE against all nbTRE) and were robust to changes in the GC matched background set FDR<0.02 were chosen for downstream analysis.

*Summarizing motif enrichment statistics across patients.* To summarize motif enrichments across 23 patients in our cohort for each transcription module, we looked for the direction of change (i.e. enriched or depleted), and reported the percentage of patients and mean fold change across patients. To define the direction of change, we separately counted the number of patients with significantly enriched or depleted (FDR<0.2) change, and let the major trend represent the sign of the majority. In **Fig.4c**, we reported motifs that show a significant change in at least 50% of patients in at least one transcriptional regulatory module.

# References:

1. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308,** 1149–1154 (2005).
2. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–187 (2010).
3. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
4. Quinodoz, S. & Guttman, M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. *Trends Cell Biol.* **24,** 651–663 (2014).
5. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
6. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).
7. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339,** 950–953 (2013).
8. Mayer, A. *et al.* Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161,** 541–554 (2015).
9. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).
10. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352,** 1225–1228 (2016).
11. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).
12. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12,** 433–438 (2015).
13. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
14. Azofeifa, J. G. & Dowell, R. D. A generative model for the behavior of RNA polymerase. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw599
15. Cai, H. & Luse, D. S. Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. *J. Biol. Chem.* **262,** 298–304 (1987).
16. Choder, M. & Aloni, Y. RNA polymerase II allows unwinding and rewinding of the DNA and thus maintains a constant length of the transcription bubble. *J. Biol. Chem.* **263,** 12994–13002 (1988).
17. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
18. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155,** 462–477 (2013).
19. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17,** 98–110 (2010).
20. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).
21. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
22. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* (2016). doi:10.1126/science.aah7111
23. Xi, Z. *et al.* Overexpression of miR-29a reduces the oncogenic properties of glioblastoma stem cells by downregulating Quaking gene isoform 6. *Oncotarget* **8,** 24949–24963 (2017).

24. Ma, Y. *et al.* PVT1 affects growth of glioma microvascular endothelial cells by negatively regulating miR-186. *Tumour Biol.* **39,** 1010428317694326 (2017).
25. Zhao, D. *et al.* Heat shock protein 47 regulated by miR-29a to enhance glioma tumor growth and invasion. *J. Neurooncol.* **118,** 39–47 (2014).
26. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
27. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
28. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154,** 888–903 (2013).
29. Tentler, J. J. *et al.* Patient-derived tumour xenografts as models for oncology drug development. *Nat. Rev. Clin. Oncol.* **9,** 338–350 (2012).
30. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6,** 8971 (2015).
31. Suvà, M. L. *et al.* Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell* **157,** 580–594 (2014).
32. Ricci-Vitiani, L. *et al.* Tumour vascularization via endothelial differentiation of glioblastoma stem-like cells. *Nature* **468,** 824–828 (2010).
33. Ricci-Vitiani, L. *et al.* Mesenchymal differentiation of glioblastoma stem cells. *Cell Death Differ.* **15,** 1491–1498 (2008).
34. Wang, Z., Martins, A. L. & Danko, C. G. RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw338
35. Canute, G. W. *et al.* Hydroxyurea accelerates the loss of epidermal growth factor receptor genes amplified as double-minute chromosomes in human glioblastoma multiforme. *Neurosurgery* **39,** 976–983 (1996).
36. Eller, J. L., Longo, S. L., Hicklin, D. J. & Canute, G. W. Activity of anti-epidermal growth factor receptor monoclonal antibody C225 against glioblastoma multiforme. *Neurosurgery* **51,** 1005–13; discussion 1013–4 (2002).
37. Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol. Cell. Biol.* **14,** 7219–7225 (1994).
38. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11,** 1455–1476 (2016).
39. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
42. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
43. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14,** 144–161 (2013).
44. Team, R. D. C. R: A language and environment for statistical computing. In R Foundation for Statistical Computing. (2010).
45. Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* **50,** 212–222 (2013).
46. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
47. Hastie, T., Mazumder, R., Lee, J. & Zadeh, R. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *arXiv [stat.ME]* (2014).
48. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152,** 327–339

(2013).

49. Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24,** 1461–1462 (2008).

## Acknowledgements:

We thank M. Viapiano for valuable comments on the manuscript and other members of the Danko, Kwak, and Lis labs for valuable discussions. This research was supported, in part, by US National Institutes of Health (NIH) grants HG009309 to C.G.D. and H.K. and GM25232 to JTL and by the Walbridge Foundation for Brain Cancer Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Walbridge Foundation.

## Author Contributions:

TC, ZW, and CGD analyzed the data. EJR, GB, and HK performed molecular experiments. HK conceived of a chromatin run-on, with input from LJC and JTL. HHS selected tumors for analysis from the GBM tissue bank. RJC and HHS completed the pathologic analysis. LSC and HHS dissected GBM-15-90 brain tissue. SLL runs the GBM tissue bank and performed the murine xenograft experiments. Data collection and analysis was supervised by CGD. The manuscript was written by CGD and TC, with input from the other authors.

## Competing financial interests:

The authors declare no competing financial interests.

## Author information:

All ChRO-seq and leChRO-seq data is currently being deposited into the database of genotypes and phenotypes (dbGaP). Processed data will be deposited into Gene Expression Omnibus in hg19 and hg38 coordinate systems. All data analysis scripts and custom software will be distributed publicly on GitHub.

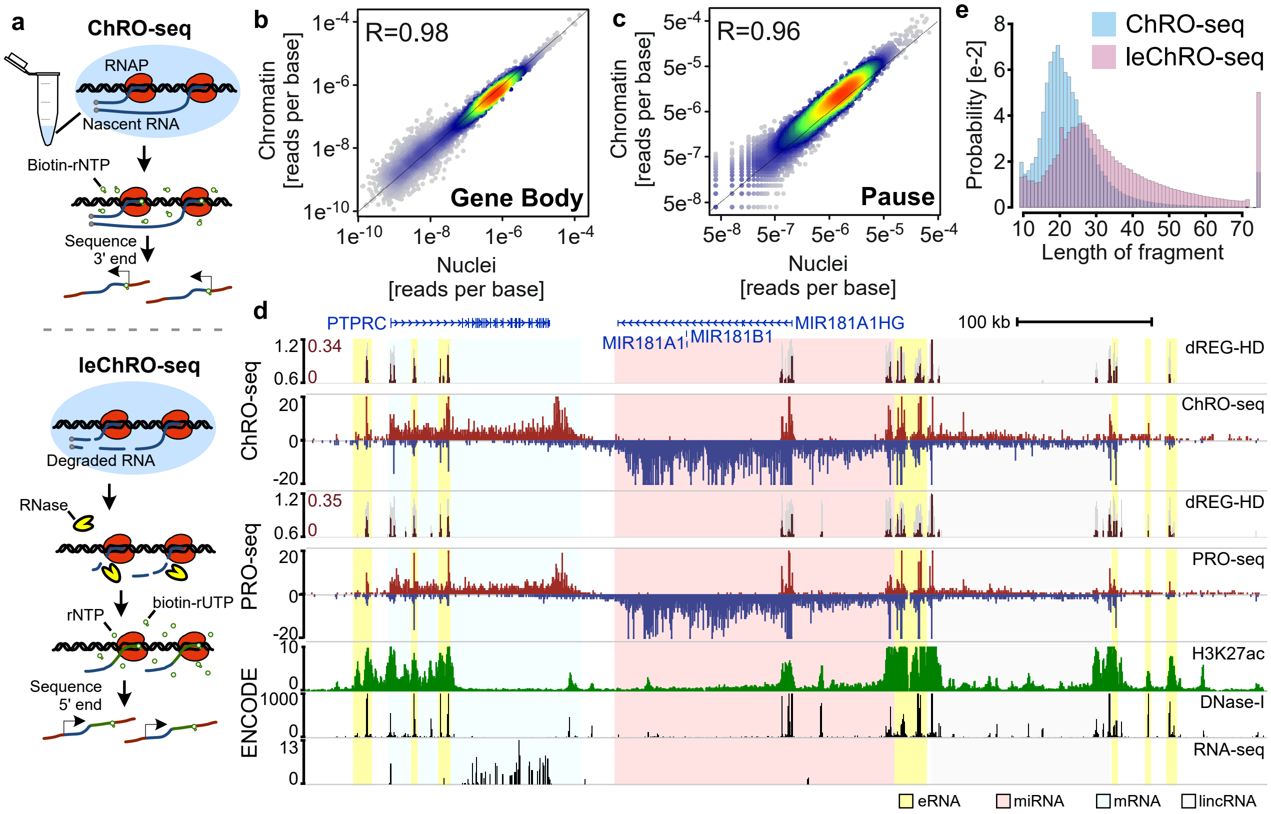

**Fig. 1. ChRO-seq and leChRO-seq measure primary transcription in isolated chromatin.** (a) Isolated chromatin is incubated with biotinylated rNTPs, purified by streptavidin beads, and sequenced from the 3' end. leChRO-seq degrades existing RNA, extends nascent transcripts an average of 100 bp, and sequences RNAs from the 5' end. (b and c) Comparison between matched ChRO-seq and PRO-seq in 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 imputed DNase-I hypersensitivity signal (dark red). (e) The distribution of read lengths from ChRO-seq (blue) and leChRO-seq (pink) in a 30 year old primary GBM.

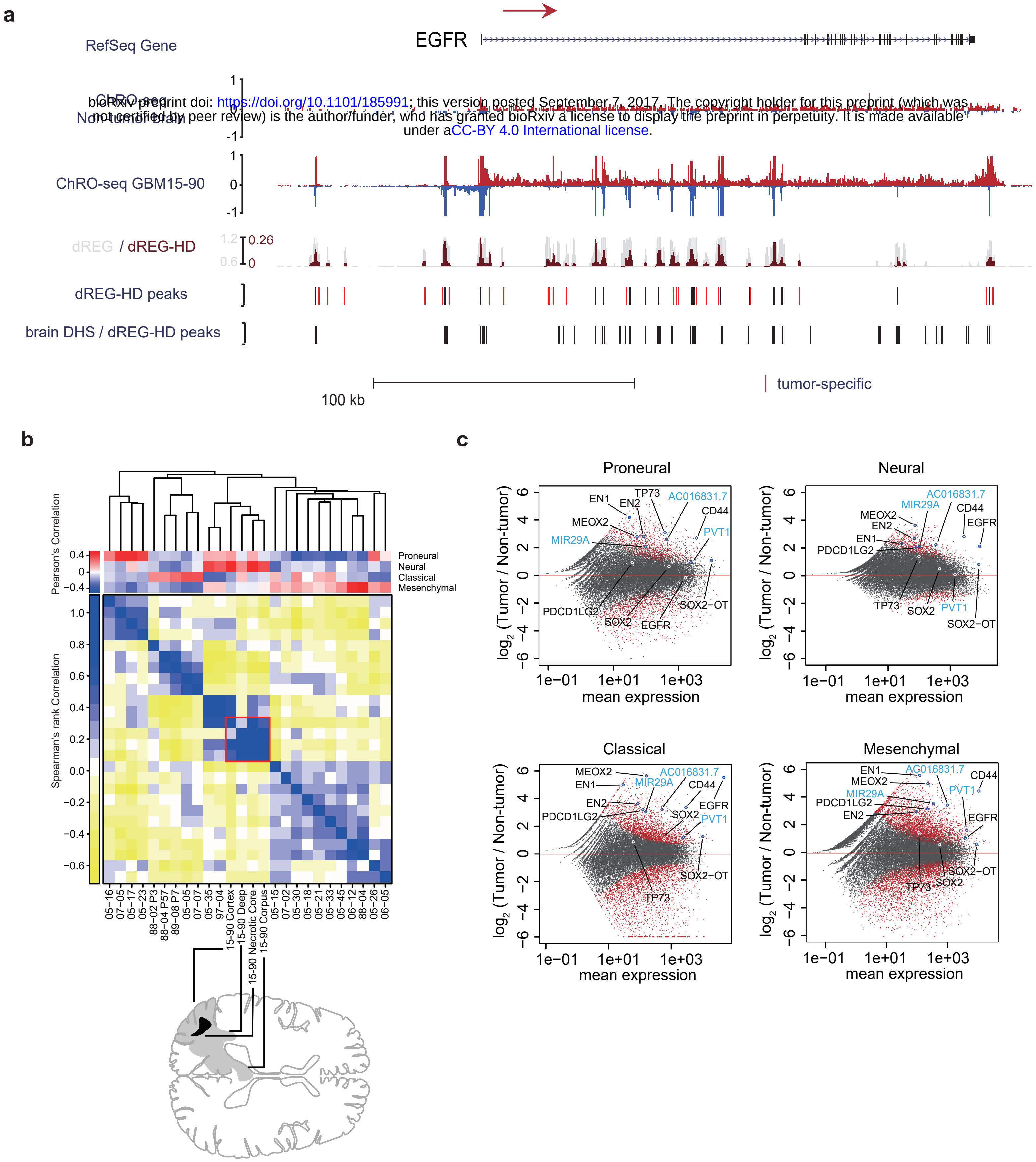

**Fig. 2. ChRO-seq detects transcription in primary human glioblastomas. (a)** RPM normalized ChRO-seq signal at the EGFR locus in nonmalignant brain (top) and GBM1 (center). dREG (gray) and dREG-HD (dark red) signals are shown for GBM-15-90 (track 3). dREG-HD sites that are not DHSs in adult brain reference samples are highlighted in red (track 4). DHSs in 6 adult brain reference samples and dREG-HD peaks from the nonmalignant brain sample (track 5). **(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 in 20 primary GBMs representing 840 signature genes. 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. **(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.

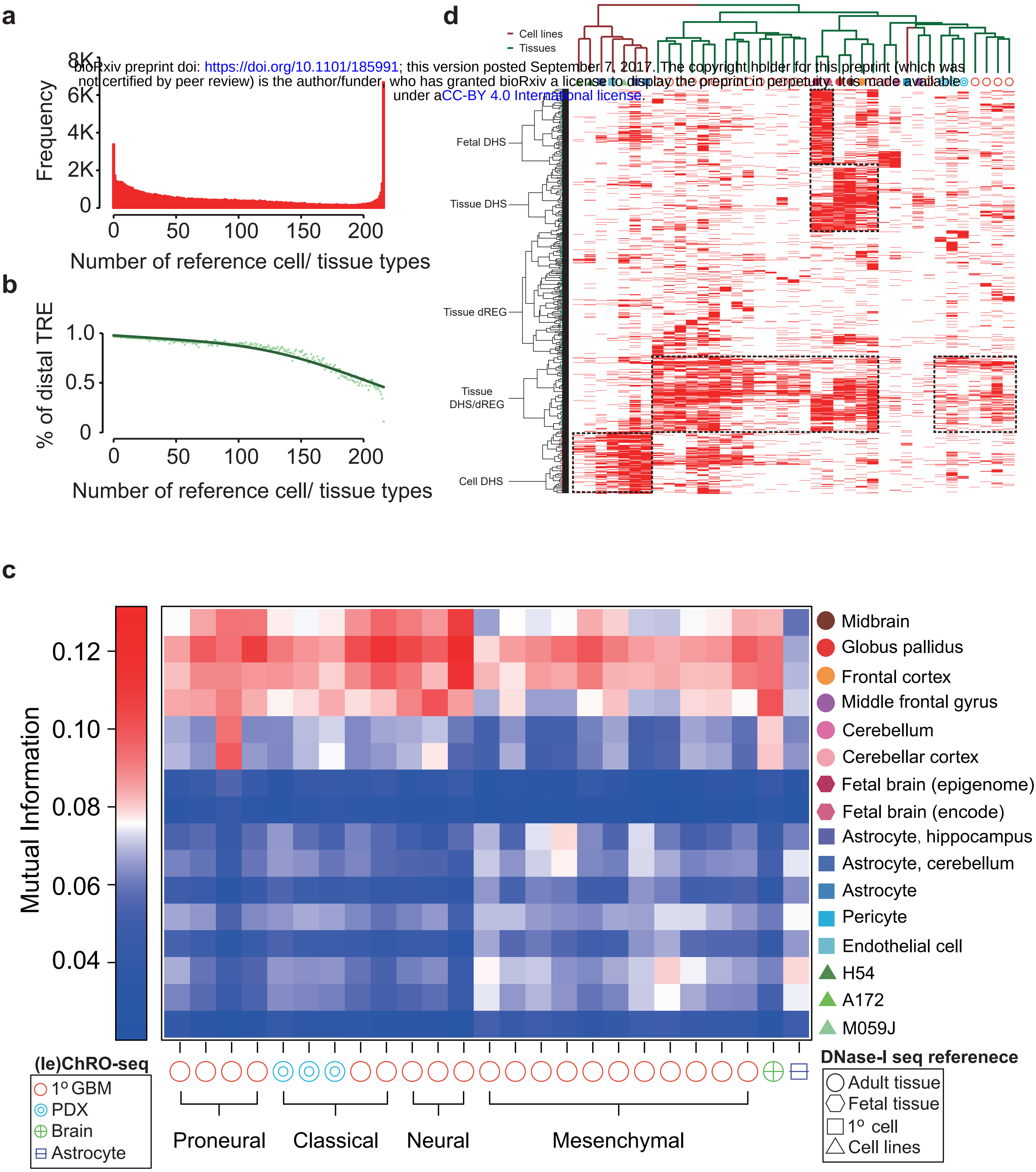

**Fig. 3. Comparison between TREs in primary GBM / PDX and reference DHSs. (a)** Histogram representing the number of reference samples that have a DHS overlapping each dREG-HD site found in any of the 23 primary GBM / PDX samples. **(b)** Percentage of TREs >1kb from the nearest GENCODE transcription 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 TRE. Activate TREs are marked in red; inactive ones are in white.

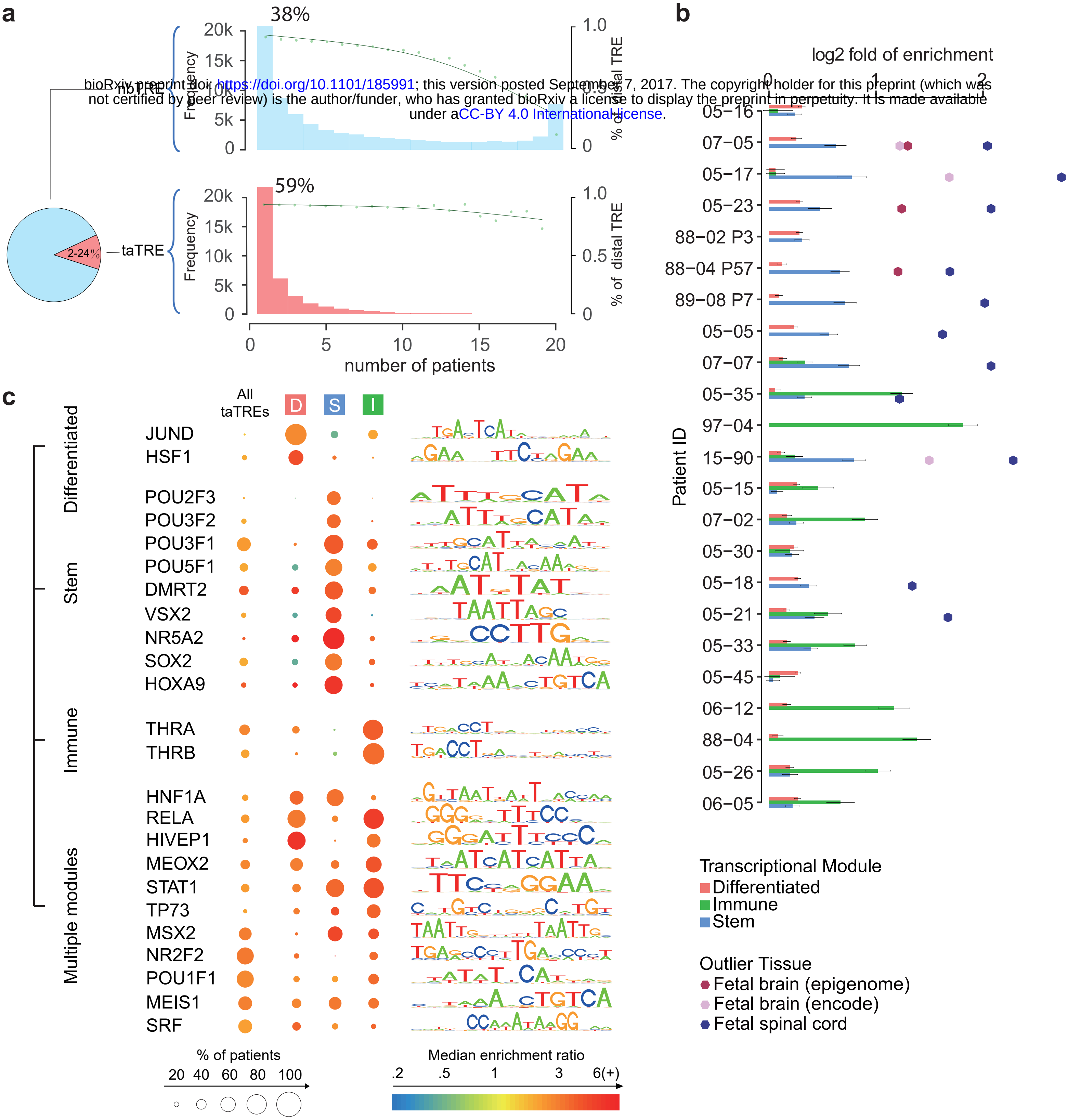

**Fig. 4. Tumor associated TREs (taTREs) activate three regulatory programs. (a)** 2 to 24% of TREs in GBM samples are not found in normal adult brain tissues. The histograms show the distribution of the number of primary GBM patients (out of 20) in which each taTRE is active. The percentage of TREs >1kb from the nearest transcription start site (distal) is shown in green dots. **(b)** Barplots show the fold enrichment of reference tissues in the corresponding GBM. Reference samples were grouped into three clusters, representing stem-like (blue), immune (green), and differentiated (pink) regulatory programs. Error bars represent the standard error. Outliers with 6 times the standard error are highlighted. **(c)** Transcription factor binding motifs enriched in TREs in the indicated regulatory program compared with normal brain. Motifs are divided into four categories on the basis of their enrichment: differentiated program motifs, stem program-specific motifs, immune program-specific motifs, and those enriched in both multiple regulatory programs. The percentage of patients found significantly enriched/depleted for each regulatory program is represented by the radius of the circle and enrichment (red) or depletion (blue) are represented by the color.