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

Tinyi Chu1,2, Edward J. Rice1,3, Gregory T. Booth4, H. Hans Salamanca5, Zhong Wang1, Leighton J. Core6, Sharon L. Longo7, Robert J. Corona6, Lawrence S. Chin7, John T. Lis4,4* and Hojoong Kwak4,4* and Charles G. Danko1,3*

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

Human genomes encode a wealth of functional elements that have critical roles in the molecular basis of disease. RNAs serve as markers for a surprisingly diverse group of functional elements, indicating the expression level of protein-coding genes (messenger RNAs), as well as the location of enhancers and other non-coding regulatory elements that transcribe short and rapidly degraded non-coding RNAs (ncRNAs)1–4. However, the discovery of ncRNA species, particularly of enhancer-templated RNAs that are characteristic of distal regulatory elements1–3, has proved challenging. Most ncRNAs are not represented in RNA-seq data, owing to the rapid degradation rates of most ncRNAs by the nuclear exosome complex5–7. Chromatin immunoprecipitation and sequencing (ChIP-seq) for RNA polymerase II is of limited value because it has a poor signal-to-noise ratio, which obscures less abundant RNA species8. Likewise, assays that measure nucleosome accessibility, such as DNase-seq9 (DNase I digestion and sequencing) and ATAC-seq10 (assay for transposase-accessible chromatin with high-throughput sequencing), are poor sources of information about transcriptional activity because they identify open chromatin regions irrespective of activity, and they do not provide critical information about mRNAs such as gene expression levels or transcript boundaries.

Recent studies have shown that sequencing of nascent RNAs attached to an actively transcribing RNA polymerase complex is an effective strategy for discovering coding and ncRNAs11–18. Nascent RNA-seq techniques, such as precision nuclear run-on sequencing (PRO-seq)11, provide markedly higher sensitivity for detecting short-lived ncRNAs than techniques which measure total RNA. Thus, PRO-seq and related assays provide a rich source of information about multiple layers of regulatory control, enabling simultaneous measurements of transcription at protein-coding genes and the discovery of active regulatory elements, including enhancers12–14. Cancers are a particularly attractive target for nascent RNA sequencing techniques because cancer is a disease of gene regulation15. In most cancers, somatic changes to DNA sequence affect oncogenic or tumor suppressive pathways16–18. In some cases, somatic mutations affect the core transcriptional machinery directly19, motivating the use of assays that measure the localization of RNA polymerase II (Pol II) directly. Somatic mutations initiate secondary changes in gene expression that are in turn responsible for initiating changes in cell morphology and behavior that are characteristic of malignancy. For this reason, gene expression signatures from RNA-seq and other assays have proved effective as biomarkers, denoting cancer subtypes that are associated with progression and survival. However, which genes undergo regulatory changes in cancer, and in particular the identity of key transcription factors that encode the malignant behaviors of cancer cells by their effect on target genes, remain poorly defined.

Nascent RNA sequencing techniques remain challenging to apply in some cell lines, and in particular to intact clinical isolates derived from patients with cancer. Here, we introduce a new protocol called chromatin run-on and sequencing (ChRO-seq). ChRO-seq produces similar maps of transcription to PRO-seq in cell lines, but it can also be applied to solid tissue samples, including those in which RNA is highly degraded. We used ChRO-seq to analyze 24 human glioblastoma (GBM) brain tumors, patient-derived xenografts (PDXs) and a primary non-malignant brain sample, and...
Results

Run-on assays in solid tissue. We developed ChRO-seq, a method to map RNA polymerase in cell or tissue samples (Fig. 1a). The primary challenge of using PRO-seq is often to obtain nuclei that are suitable for a run-on reaction. We therefore developed an alternative method that relies on fractionating insoluble chromatin, including engaged Pol II\(^{30}\) (see Methods). Insoluble chromatin was re-suspended by sonication and used as the input sample for a run-on reaction (Fig. 1a). The run-on was designed to incorporate a biotinylated nucleoside triphosphate (rNTP), and sequences RNAs from the 5’ end. PRO-seq discovered the primary transcript encoding MIR181 as annotated genes was within the range of variation observed in PRO-seq data from the same cell line (Supplementary Fig. 1). In contrast, we noted differences in the pause peak and in transcription levels beyond the polyadenylation site compared with data from mammalian native elongating transcript sequencing (mNET-seq) and genome-wide sequencing of nascent RNA (Nascent-seq), two other chromatin-based RNA sequencing assays\(^{14,28,29}\) (Supplementary Note 1). ChRO-seq and PRO-seq produced highly correlated levels of RNA polymerase in the bodies of mRNA encoding genes (R = 0.98; Fig. 1b). Likewise, signal for paused Pol II was highly correlated across the 5’ ends of annotated genes (R = 0.96; Fig. 1c), and pause levels in our test ChRO-seq library were within the range of variation observed using nuclei (Supplementary Fig. 2). The microRNA miR-181 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 enhancer-templated RNAs that were not discovered using RNA-seq.

Because RNA prepared from archival tissues is often highly degraded, such samples are poor candidates for genome-wide transcriptome analysis using RNA-seq. The RNA polymerase–DNA complex is more stable than RNA\(^{30}\), suggesting that engaged polymerases

---

**Fig. 1 | ChRO-seq and leChRO-seq measure primary transcription in isolated chromatin.** a. Cartoon depicts ChRO-seq (top) and leChRO-seq (bottom). Isolated chromatin is resuspended into solution, incubated with biotinylated ribonucleoside tri-phosphate (rNTPs), purified by streptavidin beads and sequenced from the 3’ end. leChRO-seq degrades existing RNA, extends nascent transcripts and an average of 100 bp using a mixture of rNTPs and biotinylated uridine triphosphate (rUTP), and sequences RNAs from the 5’ end.

b. Comparison of matched ChRO-seq and PRO-seq data in 41,478 reference sequence (RefSeq) annotated gene bodies (b) or at the peak of paused Pol II (c). d. Comparison of ChRO-seq, PRO-seq, and H3K27ac (histone H3 at lysine 27) ChIP-seq, Dnase-seq and RNA-seq data. dREG-HD shows the raw signal for dREG (gray) and dREG-HD (dark red). The background shading shows the type of RNA produced at each position. e. The distribution of read lengths from ChRO-seq (blue) and leChRO-seq (pink) in a GBM that was stored in a tissue bank for 30 years.

through this technique we have gained new insights into the molecular etiology of GBM.
may provide an avenue for producing new RNAs in archived samples. We obtained a primary GBM (grade IV, ID number GBM-88-04) that was stored in a tissue bank for 30 years. Analysis confirmed that RNA was highly degraded in this sample (RNA integrity number (RIN) = 1.0, Supplementary Fig. 3), thus precluding the application of RNA-seq (which requires an RIN of 2–4). To measure gene expression in this sample, we devised length-extension ChrRO-seq (leChrRO-seq), a variant of ChrRO-seq that uses transcriptionally engaged Pol II and a mix of biotinylated NTP and normal NTPs to extend degraded nascent RNA transcripts (Fig. 1a). 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 polynucleotide exit channel11, whereas run-on samples achieved a longer RNA length distribution that was better suited for mapping unique reads within the human genome (Fig. 1c). Although RNA degradation could, in principle, destabilize RNA polymerase, we nevertheless observed that leChrRO-seq produced maps of transcription that correlated with those obtained using ChrRO-seq and PRO-seq, suggesting that leChrRO-seq accurately measures gene expression and Pol II pausing (Supplementary Figs. 1a, 2, 4). Thus, leChrRO-seq enables the robust interrogation of archival tissue samples that cannot be analyzed using standard genomic tools.

Maps of transcription in primary GBMs. To demonstrate how ChrRO-seq can provide insights into complex disease, we obtained ChrRO-seq or leChrRO-seq data from 20 primary GBMs, three PDXs, and a non-malignant brain (Fig. 2a and Supplementary Table 1). Histopathology showed hallmarks of grade IV malignant astrocytoma in all GBMs (for example, GBM-15-90, Supplementary Fig. 5). We sequenced ChrRO-seq data from each GBM to an average depth of 33 million uniquely mapped reads per sample (10–150 million reads per sample). We confirmed that data collected from biopsies isolated from nearby regions (technical replicates) were highly correlated (Supplementary Fig. 4c–f and Supplementary Note 2).

To gain further insight into how transcription changes in malignant tissue, we analyzed transcription within annotated protein-coding genes and ncRNAs. GBMs from our cohort represent each of the four previously reported molecular subtypes21 (Fig. 2b and Supplementary Fig. 6). Although most tumors have transcription patterns characteristic of one dominant molecular subtype, several tumors in our cohort were similar to multiple subtypes, particularly those matching neural and mesenchymal signatures, consistent with reports of cellular heterogeneity within the same tumor13–19 (Fig. 2b). We identified 2,381 protein-coding genes and 1,123 ncRNAs that were differentially transcribed across all 20 primary GBMs relative to replicates of the non-malignant brain (P < 0.05, false discovery rate (FDR)-corrected Wald test, DESeq220) (Supplementary Table 2). Differentially transcribed genes had notable enrichments in biological processes related to cell cycle, DNA replication and metabolic rate (FDR)-corrected Wald test, DESeq236) (Supplementary Table 2).

TREs define three distinct regulatory programs activated in GBM tissue. TREs that were active in tumor tissue, but were not DHSs in any of the available adult brain reference samples, may contribute to the malignant phenotype of the tumor. Such tumor-associated TREs (tTREs) comprised 2–24% of TREs in each tumor (Supplementary Figs. 15 and 16 and Supplementary Table 4). We developed a statistical test to identify tissues that shared unexpectedly high overlap with tTREs identified in each tumor that controls for DHS scarcity (Supplementary Table 4) (see Methods). Hierarchical clustering of the tTREs among enriched cell or tissue types identified three regulatory programs that were enriched in the primary GBMs; one resembling a stem-like regulatory program, one associated with
Fig. 2 | ChRO-seq detects transcription in primary human glioblastomas. **a**, Reads per million-normalized ChRO-seq signal at the epidermal growth factor receptor (EGFR) locus in nonmalignant brain (top) and GBM-15-90 brain tissues (center). dREG (gray) and dREG-HD (dark red) signals are shown for GBM-15-90. dREG-HD peaks that are not DHSs in adult brain reference samples are highlighted in red. DHSs in six adult brain reference samples and dREG-HD peaks from the non-malignant brain sample. **b**, Upper matrix, subtype scores for each patient, calculated using the Pearson's correlation with the centroid of gene expression of the corresponding subtype. Lower matrix: Spearman's rank correlation over subtype signature genes among 20 primary GBMs. The red square denotes four regions dissected from GBM-15-90. The sample order is based on single-link hierarchical clustering of the lower matrix, shown by the dendrogram. In total, 838 genes were used to calculate the correlation coefficients. **c**, Differential gene transcription of primary GBMs in each subtype compared with non-malignant brain. Genes of interest are highlighted. Long non-coding RNAs are highlighted in blue. PCD1LG2, programmed cell death 1 ligand 2; SNHG1, small nucleolar RNA host gene 1.
differentiated support cells, and a cluster of immune cells (Fig. 4a and Supplementary Fig. 17). taTREs frequently overlapped DHSs ($P < 0.0001$ bootstrap test) in fetal tissues of the nervous system ($2.3–6.6$-fold enrichment in 11 out of 23 GBMs), in particular of the spinal cord and brain, which were derived from the neuroectoderm (Fig. 4a, see ‘Outlier tissue’). We also found evidence of enrichment in other developmental tissues; for example, embryonic stem cells and other fetal tissues from a variety of germ layers, and for a number of terminally differentiated support cell lineages, including astrocytes, endothelial cells, fibroblasts and osteoblasts. Regulatory programs were partially correlated with previously defined molecular subtypes in GBM (Fig. 4b and Supplementary Note 4).

To identify transcription factors that may be involved in maintaining each regulatory program, we classified the taTREs in each tumor sample into regulatory programs based on the cell or tissue types in which they overlap a DHS, and searched for enriched transcription factor binding motifs ($P = 2.66 \times 10^{-5} = [0.05/1,882]$ in at least one patient, Fisher’s exact test, rtfbsdb, v.0.4.0). As we were limited in our ability to distinguish between paralogous transcription factors that share similar DNA binding specificities, we clustered motifs into 14 distinct groups, each associated with multiple transcription factors that promotes stem-like properties that are observed in a population of GBM cells. Similarly, overlap with astrocytes, endothelial cells, fibroblasts or osteoblasts may capture tumor cells that have transdifferentiated into these lineages. Notably, these two signatures were detected in PDX samples as well as primary GBMs, demonstrating that these signatures reflect transcriptional diversity in malignant cells.

Fig. 3 | Comparison of TREs in primary GBM or PDXs 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 or PDX samples. b, Percentage of TREs >1kb from the nearest GENCODE transcription start site. c, Mutual information between TREs in the indicated GBM and the reference sample. d, Clustering of reference samples with primary GBM or PDXs based on the activation of TREs. Active TREs are in red; inactive TREs are in white.
Fig. 4 | Tumor-associated TREs activate three regulatory programs.

**a**. Boxplots showing the log2-fold enrichment of reference tissues enriched in the corresponding GBM. Reference samples enriched in each patient were grouped into three regulatory programs, called stem (blue, \( n = 24 \)), immune (green, \( n = 5 \)) and differentiated (pink, \( n = 21 \)). Box plots show the 25th percentile (left side of box), median (central bar) and 75th percentile (right side of box). Whiskers represent minimum and maximum values. Outlier tissues are indicated in the legend.

**b**. Regulatory modules are correlated with GBM molecular subtypes. The radius of the circle represents the \( P \) value (two-sided Fisher’s exact test) of enrichment of the indicated regulatory programs in subtype-biased TREs. The color represents the magnitude of enrichment (red) or depletion (blue). The number of subtype-biased TREs in each comparison (panels **b** and **c**) is shown in Supplementary Tables 3 and 4.

**c**. Transcription factor binding motifs enriched in TREs of the immune (I), stem (S) and differentiated (D) regulatory program compared with TREs active in the normal brain. All motifs shown were significantly enriched following Bonferroni adjustment of the threshold \( P \) value in at least one patient (\( P < 2.66 \times 10^{-5} = [0.05 / 1,882] \)). The Spearman’s rank correlation heatmap (left) shows the correlation in DNA binding sites matching each motif. The radius of the circle represents the median \( P \) value across patients and the color represents the magnitude of enrichment (red) or depletion (blue). AP-1, activating protein 1; NF-κB, nuclear factor-κB family.
Direct inference of transcription factor regulatory activities in GBMs. The gene-regulatory trans activities that a transcription factor has on its complement of bound TREs can be regulated by multiple transcriptional and post-transcriptional mechanisms. Although in some cases a transcription factor is controlled predominantly by the abundance of its protein, many require a subsequent step such as post-transcriptional activation of the protein product to regulate target genes (Fig. 6a). We asked whether we could distinguish between these two broad regulatory activities by using ChRO-seq, and using an integrative analysis incorporating both ChRO-seq and publicly available mRNA sequencing (mRNA-seq) data.

In the simplest mode of regulation, the gene-regulatory activity of a transcription factor is determined by the abundance of its protein, which can be correlated with the transcriptional activity of its gene and the abundance of its mRNA. To detect this type of regulatory activity, we asked whether motifs enriched in active TREs of each subtype correspond to changes in Pol II density on the primary transcription unit encoding any one of the transcription factors that recognize the corresponding binding motif. In some cases, we observed transcriptional changes in the transcription factor coding gene in the same subtype in which we also observed motif enrichment (Fig. 6b,c and Supplementary Fig. 24, b,c). For example, ChRO-seq signal in the gene body encoding the transcriptional activator CEBPB (CCAAT/enhancer-binding protein B) increased by 4.88-fold in mesenchymal tumors (Fig. 6b), consistent with a 2.43-fold enrichment of its corresponding motif in mesenchymal upregulated TREs (Fig. 5a). Likewise, we found several cases in which mRNA encoding each transcription factor was correlated with the expression of its putative target genes across GBMs to a greater extent than expected based on a null model that controls for molecular subtype (Fig. 6c; see Methods).

We devised a strategy to estimate which transcription factors have gene-regulatory activities that are regulated by transcriptional or post-translational mechanisms. Focusing on the 25 unique motifs enriched in upregulated TREs that are associated with multiple transcription factors, we found evidence of correlated changes for 8 motifs in ChRO-seq data and for 16 in mRNA-seq data (Fig. 6b,c). Several of these correlations were weak in magnitude, which may be consistent with gene-regulatory activities controlled by multiple regulatory mechanisms for these transcription factors. We conservatively identified at least six transcription factors, including TEAD, GATA, HSF and NF-kB, that had low correlations with their putative targets in RNA-seq and no evidence of transcriptional changes in ChRO-seq. These transcription factors seem to be regulated primarily at a post-transcriptional level in GBM.

Transcription factors control group survival-associated genes in mesenchymal GBMs. Known molecular subtypes of GBM do not correlate with survival, presenting a motivation to identify new classifiers that may have prognostic value. We reasoned that the activity of transcription factors that control transcriptional heterogeneity among patients with GBM may control biological functions associated with survival. To determine whether gene-regulatory activities of transcription factors may be useful in predicting clinical outcomes, we compared the hazard ratios at putative target genes of each subtype-biased expression subtypes and their putative target genes.

Transcription factors that control GBM subtype. Transcriptional heterogeneity among GBMs is established primarily by the differential activity of transcription factors. To identify transcription factors that are involved, we focused on TREs with evidence of expression changes among the four previously described molecular subtypes (P < 0.01, FDR-corrected Wald test, Table 5). We identified 38 binding motif clusters with extremely strong evidence of enrichment in active TREs with biased transcription in any subtype (P < 2.66 × 10⁻⁵ = [0.05/1,882], Fisher's exact test, Fig. 5a). Enriched motifs passing our stringent multiple testing correction threshold were most common in the mesenchymal and neural subtypes, and several of these motifs, including those recognized by NF-kB and C/EBP family, are supported by existing literature. In addition, we identified a number of novel motifs that correlate with subtype-biased expression, including RAR, SRF, SOX family and FOX family members.

Next, we set out to identify target genes that are regulated by each transcription factor in GBM cells. We assumed that molecular subtypes described in the current literature do not completely describe the full range of heterogeneity among GBMs. To identify motifs that correlate to heterogeneity and that are only weakly correlated with the known molecular subtypes, we relaxed our statistical cutoff to a more permissive threshold at which we expected substantially higher sensitivity at an acceptable false discovery rate (P < 0.05, nominal Fisher's exact test, Supplementary Fig. 21, see Methods). We identified bound occurrences of each enriched motif using heuristics that provide substantial performance improvements over existing high-resolution tools. Motif occurrences were connected with the closest two annotated genes sharing similar subtype bias within 50 kb (Fig. 5b, Supplementary Figs. 22, 23), using fairly stringent heuristics to limit false discovery rates (see Methods). We validated target genes by confirming that genes sharing a common transcription factor were more highly correlated across 174 primary GBMs than expected based on randomly selected genes sharing the same subtype specificity (Fig. 5c, Supplementary Fig. 24a and Supplementary Note 5). We also confirmed that predicted target genes of STAT1 and NF-kB were enriched for changes after activating those specific transcription factors (Supplementary Note 5). Thus, we have identified transcription factors that contribute to major GBM expression subtypes and their putative target genes.
and NF-κB family transcription factors (Supplementary Fig. 26). Only one of these transcription factors, C/EBP, was encoded by a gene (CEBPB) that was associated with survival at the mRNA level (Supplementary Fig. 27), consistent with the correlation of the gene-regulatory activity of C/EBP with the abundance of its mRNA (Fig. 6b). NF-κB activity was correlated to radioresistance in GBMs, and in
this case its activity was shown to be regulated post-transcriptionally by monitoring the phosphorylated state of the RELA subunit48. This provides further support for the association of a second of survival-associated transcription factors on their own (a significant enrichment; \( P = 0.0006 \), Fisher's exact test), including CCL20 (Supplementary Fig. 29a) and ADM (Fig. 7d), \( P < 0.05 \), chi-squared test) (Supplementary Table 6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 29a,c). Together these 26 genes formed a prognostic signature that strongly predicted overall survival (Fig. 7e). Moreover, differences in survival among these genes were not driven by IDH1 (encoding isocitrate dehydrogenase 1) status (Supplementary Fig. 30).

Gene ontology analysis found that targets of all three transcription factors identified here with clinical outcomes on their own (a significant enrichment; \( P = 0.0006 \), Fisher's exact test), including CCL20 (Supplementary Fig. 29a) and ADM (Fig. 7d), \( P < 0.05 \), chi-squared test) (Supplementary Table 6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 29a,c). Together these 26 genes formed a prognostic signature that strongly predicted overall survival (Fig. 7e). Moreover, differences in survival among these genes were not driven by IDH1 (encoding isocitrate dehydrogenase 1) status (Supplementary Fig. 30).

Gene ontology analysis found that targets of all three transcription factors identified here with clinical outcomes on their own (a significant enrichment; \( P = 0.0006 \), Fisher's exact test), including CCL20 (Supplementary Fig. 29a) and ADM (Fig. 7d), \( P < 0.05 \), chi-squared test) (Supplementary Table 6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 29a,c). Together these 26 genes formed a prognostic signature that strongly predicted overall survival (Fig. 7e). Moreover, differences in survival among these genes were not driven by IDH1 (encoding isocitrate dehydrogenase 1) status (Supplementary Fig. 30).

Gene ontology analysis found that targets of all three transcription factors identified here with clinical outcomes on their own (a significant enrichment; \( P = 0.0006 \), Fisher's exact test), including CCL20 (Supplementary Fig. 29a) and ADM (Fig. 7d), \( P < 0.05 \), chi-squared test) (Supplementary Table 6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 29a,c). Together these 26 genes formed a prognostic signature that strongly predicted overall survival (Fig. 7e). Moreover, differences in survival among these genes were not driven by IDH1 (encoding isocitrate dehydrogenase 1) status (Supplementary Fig. 30).

Gene ontology analysis found that targets of all three transcription factors identified here with clinical outcomes on their own (a significant enrichment; \( P = 0.0006 \), Fisher's exact test), including CCL20 (Supplementary Fig. 29a) and ADM (Fig. 7d), \( P < 0.05 \), chi-squared test) (Supplementary Table 6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 29a,c). Together these 26 genes formed a prognostic signature that strongly predicted overall survival (Fig. 7e). Moreover, differences in survival among these genes were not driven by IDH1 (encoding isocitrate dehydrogenase 1) status (Supplementary Fig. 30).

Gene ontology analysis found that targets of all three transcription factors identified here with clinical outcomes on their own (a significant enrichment; \( P = 0.0006 \), Fisher's exact test), including CCL20 (Supplementary Fig. 29a) and ADM (Fig. 7d), \( P < 0.05 \), chi-squared test) (Supplementary Table 6). High expression of both genes was associated with high risk regardless of subtype assignment, indicating that survival association of these transcription factors was not simply driven by enrichment in the mesenchymal subtype (Supplementary Fig. 29a,c). Together these 26 genes formed a prognostic signature that strongly predicted overall survival (Fig. 7e). Moreover, differences in survival among these genes were not driven by IDH1 (encoding isocitrate dehydrogenase 1) status (Supplementary Fig. 30).
**Fig. 7 | Transcription factors control survival-associated pathways in GBM.**

a. Scatter plot showing the \(-\log_{10}\) two-sided Wilcoxon rank-sum test P value comparing the distribution of hazard ratios of target genes for each transcription factor and two groups of non-target control genes (see Methods). The radius of the circle denotes the \(-\log_{10}\) P value of association between transcription factor messenger RNA levels and survival. Color denotes the loge of the hazard ratio at higher mRNA levels. The dotted red line represents the Bonferroni-adjusted \(\alpha\) and unique to (center, \(n = 62\)) three transcription factors, and for mesenchymal marker genes (right, \(n = 161\)). Mean hazard ratios are shown by white dots and standard deviations are shown by bars. P values were calculated using a two-sided Wilcoxon rank-sum test. b. Venn diagram showing overlap between the target genes of the three indicated transcription factors. c. Violin plot shows the log hazard ratios for target genes shared among (left, \(n = 26\)) and unique to (center, \(n = 62\)) three transcription factors, and for mesenchymal marker genes (right, \(n = 161\)). Mean hazard ratios are shown by white dots and standard deviations are shown by bars. P values were calculated using a two-sided Wilcoxon rank-sum test. d. Browser track of ADM shows the average of reads per million-normalized iChRO-seq signals and dREG-HD scores in mesenchymal (MES, \(n = 3\)) and non-MES (\(n = 8\)) GBMs. MES-biased TREs and motif positions are highlighted in blue. e. Kaplan–Meier plot showing overall survival for 196 patients with high and low average expression levels 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.

Factors were enriched for immune system process and stress responses (\(P < 0.000001\), FDR-corrected Fisher’s exact test, Supplementary Table 7). Taken together, our analysis suggests that C/EBP, RAR, and NF-κB work in concert to activate a shared regulatory program that controls inflammatory processes and correlates with poor clinical outcomes in GBM.

**Discussion**

Nascent transcription is a promising approach for studying the molecular basis of complex disease because unstable RNAs provide deep insights into multiple stages of gene regulation. ChRO-seq can map nascent transcription in many samples that maintain the integrity of protein–DNA interactions, even those in which RNA is highly degraded. ChRO-seq has the potential to be used across the biomedical sciences to analyze regulatory programs that contribute to solid tumors and other tissues that have proved challenging to study using existing molecular tools.

Our analysis of 20 primary tumors provides several insights into transcriptional regulatory programs in malignant tissue. First, we report that enhancers in malignant tissue are surprisingly similar to DHSs in the tissue of origin. This finding suggests that regulatory programs in GBM seem to work within the confines of chromatin architecture that is established in the initiating cell. Regulatory programs are also similar to normal brain in PDXs, demonstrating that tumor-initiating cells are able to reconstitute a diverse cell environment that bears surprising similarity to primary brain tissue. Yet malignant cell behaviors are specified by cancer cells despite this similarity. This may be explained by our discovery of a rare population of ectopic enhancers that resembled fetal tissues isolated from the nervous system, immune cells, and differentiated tumor cells.
Our observations are consistent with models of tumorigenesis in which tumor cells reactivates regulatory programs that are similar in some respects to earlier developmental stages\textsuperscript{10}. These regulatory signatures derived from rare ectopic enhancers may have important prognostic value that can be exploited in future studies.

Our study highlights how transcription factors can be responsible for coordinated changes in the expression of groups of genes that contribute to expression heterogeneity among tumors. ChRo-seq, like other run-on technologies\textsuperscript{55}, provides substantial information about the regulatory activities of transcription factors on chromatin that is independent of transcription factor mRNA expression levels.

In support of our general approach, transcription factor candidates activating TREs in the stem-like regulatory program were similar to those reported previously to be sufficient for initiating tumors in a murine host\textsuperscript{41}. In addition, we used ChRo-seq data to identify transcription factors correlated with gene expression patterns in each GBM subtype.

We report three transcription factors, C/EBP, RAR and NF-κB, whose target genes are systematically correlated with poor clinical outcomes. Our work adds new transcription factors to the current literature, as well as additional support for the role of C/EBP in driving mesenchymal transformation\textsuperscript{11}. NF-κB was previously associated with resistance to radiotherapy and involvement in mesenchymal transformation in GBMs\textsuperscript{46}. Our present work builds on these studies, indicating that NF-κB activation has an unambiguous influence on clinical outcomes. In addition, we found evidence that a third transcription factor, RAR, drives regulatory programs that contribute to survival in GBMs. Notably, post-transcriptional mechanisms seem responsible for activating two of these three transcription factors; NF-κB and RAR. Thus, insights reported here were possible only because ChRo-seq is a more direct indicator of transcription factor activity than other tools applied previously in GBM. As the pharmacology for targeting diverse transcription factor families develops, the transcription factors reported here, as well as our strategies for finding them, should become more useful in nominating targeted therapies.

URLs. ChRo-seq/leChRo-seq alignment pipeline, https://github.com/Danko-Lab/utils/tree/master/proseq; dREG-HD implementation, https://github.com/Danko-Lab/dREG-HD; tTarget implementation, https://github.com/Danko-Lab/tTarget; bigwig software package, https://github.com/andremlarins/bigWig; The Cancer Genome Atlas (TCGA) microarray data, https://tcga-data.nci.nih.gov/docs/publications/gbm_exp/unifiedScaled.txt

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0244-3.

Received: 2 October 2017; Accepted: 21 August 2018; Published online: 22 October 2018

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. Uiltsky, 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. De Santa, F. et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PloS Biol. 8, e1000384 (2010).
6. Preker, P. et al. RNA exosome depletion reveals transcription upstream of active human promoters. Science 322, 1851–1854 (2008).
7. Andersson, R. et al. Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat. Commun. 5, 5336 (2014).
8. Core, L. J. et al. Defining the status of RNA polymerase at promoters. Cell Rep. 2, 1025–1035 (2012).
9. Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, 311–322 (2008).
10. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin. DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
11. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 318, 1845–1848 (2008).
12. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469, 368–373 (2011).
13. 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).
14. Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell 161, 541–554 (2015).
15. Nojima, T. et al. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. Cell 161, 526–540 (2015).
16. Schwab, B. et al. TT-seq maps the human transient transcriptome. Science 337, 1225–1228 (2012).
17. 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).
18. Scruggs, B. S. et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. Mol. Cell 58, 1101–1112 (2015).
19. Danko, C. G. et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat. Methods 12, 433–438 (2015).
20. Azolefka, J. G. & Dowell, R. D. A generative model for the behavior of RNA polymerase. Bioinformatics 33, 227–234 (2016).
21. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
22. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional addiction in cancer. Cell 168, 629–643 (2017).
23. Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. Science 321, 1807–1812 (2008).
24. Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. Cell 155, 462–477 (2013).
25. Mohan, M., Cin, C., Guest, E. & Shilatifard, A. Licensed to elongate: a molecular mechanism for MLL-based leukaemogenesis. Nat. Rev. Cancer 10, 721–728 (2010).
26. Wuerin, 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).
27. 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).
28. Khodor, Y. L. et al. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev. 25, 2502–2521 (2011).
29. Menet, J. S., Rodriguez, J., Abruzzi, K. C. & Rosbash, M. Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. eLife 1, e00011 (2012).
30. Sai, 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).
31. Choder, M. & Aloni, Y. RNA polymerase II allows unwinding and re-winding of the DNA and thus maintains a constant length of the transcription bubble. J. Biol. Chem. 263, 12994–13002 (1988).
32. Verhaak, R. G. W. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDEGFRA, IDH1, EGFR, and NF1. Cancer Cell. 17, 98–110 (2010).
33. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014).
34. Wang, Q. et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer Cell 32, 42–56.e6 (2017).
35. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068 (2008).
36. 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).
37. Liu, S. J. et al. CRISPRI-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 355, eaah7111 (2016).
38. 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).
39. Ma, Y. et al. PVT1 affects growth of glioma microvascular endothelial cells by negatively regulating miR-186. Tumour Biol. 39, 1010428317694326 (2017).
40. 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).
41. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).
42. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
43. Tentler, J. J. et al. Patient-derived tumour xenografts as models for oncology drug development. Nat. Rev. Clin. Oncol. 9, 338–350 (2012).
44. Suvà, M. L. et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. Cell 157, 580–594 (2014).
45. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes.
46. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome.
47. Wang, Z., Martins, A. L. & Danko, C. G. RTFBSDB: an integrated framework for transcription factor binding site analysis. Bioinformatics 32, 3024–3026 (2016).
48. Bhat, K. P. L. et al. Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma. Cancer Cell 24, 331–346 (2013).
49. Carro, M. S. et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463, 318–325 (2010).
50. Danko, C. G. et al. Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. Nat. Ecol. Evol. 2, 537–548 (2018).
51. Luo, X., Chae, M., Krishnakumar, R., Danko, C. G. & Kraus, W. L. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. BMC Genomics 15, 155 (2014).
52. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083–1087 (2016).
53. Wang, M., Zhao, Y. & Zhang, B. Efficient test and visualization of multi-set intersections. Sci. Rep. 5, 16923 (2015).
54. Stergachis, A. B. et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell 154, 888–903 (2013).
55. Azoteifia, J. G. et al. Enhancer RNA profiling predicts transcription factor activity. Genome Res. 28, 334–344 (2018).

Acknowledgements
We thank M. Viapiano and P. Sethupathy for valuable comments on the manuscript and other members of the Danko, Kwak and Lis laboratories 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 J.T.L.), and by the Walbridge Foundation for Brain Cancer Research. T.C thanks the Croucher Foundation for the Croucher Scholarships for Doctoral Study (2013). 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
T.C., Z.W. and C.G.D. analyzed the data. E.J.R., G.B. and H.K. performed molecular experiments. H.K. conceived the chromatin run-on technique, with input from L.J.C. and J.T.L. H.H.S. selected tumors for analysis from the GBM tissue bank. R.J.C. and H.H.S. completed the pathologic analysis. L.S.C. and H.H.S. dissected GBM-15-90 brain tissue. S.L.L. ran the GBM tissue bank and performed the murine xenograft experiments. Data collection and analysis was supervised by C.G.D. The manuscript was written by C.G.D. and T.C., with input from the other authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0244-3.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to H.K. or C.G.D.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2018
Methods

Cell culture. Jurkat cells were grown in RPMI-1640 medium supplemented with 10% fetal bovine serum, 1x penicillin streptomycin antibiotic, 0.125 mg/ml gentamicin antibiotic at 37 °C, 5% CO₂. 1x10⁶ cells were centrifuged at 780g for 5 min at 4 °C. The medium was removed and the cells were resuspended with 1x PBS, and centrifuged again before PBS was removed.

Tissue collection and preparation. GBM-derived cells were prepared from freshly biopsied human tumors obtained with patient consent. The sample collection was approved by the Institutional Review Board of SUNY Upstate Hospital, Syracuse, NY, USA, and followed all relevant ethical regulations. The non-tumor brain sample was dissected from the brain of a patient with epilepsy, with also informed consent and institutional review board approval. To establish patient-derived xenografts, small pieces of freshly resected gliomas were implanted subcutaneously in the flank of athymic nude (nu/nu) mice (Envigo) and serially passaged (mouse-to-mouse) 3 times for Dnase-seq and leChRO-seq. Afterwards, a total of 10 mice were used to generate the data presented. Cell culture. We used our publicly available pipeline (see URLs) to align ChRO-seq and leChRO-seq data. Some libraries were prepared using adapters that contained a molecule-specific unique identifier (first 6 bp sequenced; denoted in Supplementary Table 1), and for these we removed PCR duplicates using PRINSEQ lite⁴. Adapters were trimmed from the 3’ end of remaining reads using cutadapt⁶ with a 10% error rate. Reads were mapped with the Burrows-Wheeler Aligner (BWA)⁷ 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 that 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 were converted to wig format using BedTools⁸ and the bedgraph tool in the Kent Source software package⁹. Downstream data analysis was performed using the bigwig package (v.0.2.9; see URLs). All data processing and visualization was carried out in R statistical environment¹⁰-¹³.

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

Molecular subtype classification. Transcriptional activity of characteristic genes for each GBM subtype were identified by applying the methods described above. Read counts from each sample were normalized by reads per million total reads, followed by log transformation of pseudocount-adjusted data (reads per million-normalized reads +1). The transformed read counts were centered to a mean of zero for each gene. The similarity between each sample was measured by Spearman’s rank correlation, and clustered using single-link clustering. The similarity of each sample to molecular subtypes was calculated using Pearson’s correlation with the centroid of the corresponding subtype.

Differential expression analysis for annotated genes. Patients with gene expression patterns matching each of the four reported molecular subtypes were treated as biological replicates (Fig. 3b and Supplementary Table 3). Two technical replicates of non-malignant brain were used as control. Differential expression analysis was conducted using DESeq2 (v.2.10.0) and differentially expressed genes were defined as those with a FDR of less than 0.05.

Comparison of TREs with DNase I hypersensitive sites. dREG-HD. dREG-HD was run using the default settings. A complete description of dREG-HD can be found in Supplementary Note 6.

Data processing for calling Dnase I hypersensitive sites and dREG-HD sites. We reprocessed all Dnase-seq data and identified 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 MACS2⁶² and Hotspot. To group DHSs for each cell and tissue type with high confidence, we took the union of peaks (using the bedtools ‘merge’ function) from biological replicates followed by intersecting peaks called by Hotspot and MACS2. Lastly, peaks resulting from intersection might be too narrow and hence might be missed during downstream intersection operations, we expanded all short peaks (<150 bp) to 150 bp from the peak center. Analyses involving individual replicates (Supplementary Fig. 1) use only peaks called by MACS2.

ChRO-seq and leChRO-seq data were mapped to hg19 as described above. A dREG score threshold was set at 0.7 to generate dREG peak regions. These dREG regions were used as input for dREG-HD. dREG-HD runs were under the same parameters as those with a FDR of less than 0.05. Conducted using DESeq2 (v.1.20.0) and differentially expressed genes were defined as those with a FDR of less than 0.05.

Comparison of TREs with DNase I hypersensitive sites. dREG-HD. dREG-HD was run using the default settings. A complete description of dREG-HD can be found in Supplementary Note 6.
Mutual information heatmap and clustering analysis. To examine the similarity between the dREG-HD sites in each query sample and DHSs in each reference sample (Fig. 3c), we computed the pairwise mutual information using a pair of dREG-HD and DHSs (as described above) on the same sample defined by merged peaks among all samples included in the analysis. Information on clustering samples based on mutual information using a similar algorithm as described previously9, as in Supplementary Fig. 14, can be found in Supplementary Note 7.

TRE clustering analysis. We analyzed the activation pattern across TREs using the same definition of sample space as described in the mutual information analysis (above). We assigned two states to each TRE, active if it intersected a dREG-HD peak or a DHS and inactive otherwise. The Jaccard distance was used to quantify the similarity between two samples or between two potential TREs. Clustering across samples (columns) and across TREs (rows) was done using wardD2 in R. 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 six brain-related reference samples (16 samples in total).

taTRE enrichment test and clustering into regulatory programs. taTREs were defined as TREs from primary GBM or PDXs that do not intersect with any dREG-HD peaks from our non-malignant brain control or with DHSs found in normal brain tissues. 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 interactions between each reference sample was significantly higher than expected by chance. We generated a null model by assigning the reads per kilobase per million mapped reads (RPKM) normalized TCGA RNA-seq data from 174 patients with GBM, and used the approach that we used to associate transcription factors with target genes, we compared the co-expression of target genes to that of background non-target genes.

Motif enrichment analysis of tumor-associated TREs and subtype-biased TREs. Defining subtype-biased TREs. To search for TREs that differentially activated or repressed transcription in each subtype, we rely on measuring the change of the nascent RNA in the TRE regions. We merged dREG-HD sites called using the relaxed setting across 23 samples. We summed up the read counts of 10ChR-seq and ChR-seq of each merged dREG-HD site extended by 250bp from the center. TREs in patients of the subtype of interest (Supplementary Table 5) were compared against those of the remaining three subtypes. Differential expression analysis was conducted using DESeq22, and subtype-biased TREs are defined as those differentially transcribed with a FDR of less than 0.01.

Defining genomic regions for motif enrichment comparison. We compared motif enrichment in the positive set (the sequence set in which the motifs are found) with a GC-content-matched background control set. In the taTRE motif enrichment, we used the group indicated in Supplementary Fig. 15 as the positive set, and dREG-HD sites that intersect with active DHSs in the normal brain as the background. For subtype-biased TRE motif enrichment analyses we used upregulated or downregulated subtype-biased TREs as the positive set and TREs that did not show significant differential transcription (FDR, DESeq2, P > 0.1) as the background set. For the positive and background sets, we selected the center of peaks and then extended by 150bp from the center. We subsampled background peaks to construct >2,500 GC-content-matched TREs before scanning for motif enrichment.

Motif enrichment analysis. We used the R package rfbPsdb (v=0.4.0) to search for motifs that show enrichment in each primary GBM. We focused on 1,882 human transcription factor binding motifs from the CisBP database58. When scanning genomic regions of interest, we used transcription factor binding sites with a log-odds score of ≥ 2 in positive and background sets, with scores obtained by comparing motif each model to a second-order Markov background model. Motif enrichment was tested using Fisher’s exact test. To account for potential bias resulting from the difference in GC content between positive and background sets, we ran a statistical test on 50 independently subsampled GC-matched dREG-HD regions, and summarized the P values and the fold enrichment across background sets using the median across samples.

We refined motifs discovered using several heuristics, as follows: the motif was enriched (with a fold enrichment of > 1); the enrichment was robust to changes in the GC-matched background set (median P < 2.66 x 10^{-4} = [0.05 / 1.882]); the positive sets have at least 10 sites with log-odds scores ≥ 2; the transcription factor was transcribed (for subtype-biased TREs). In the subtype-biased TRE analysis, we required the one with the most significant enrichment in all taTREs over TREs found in the normal brain samples. In the subtype-biased TRE analysis, we used all motifs meeting the enrichment criteria and heuristics described above.

Motif clustering by genomic position. As we are not able to rigorously distinguish between paralogous transcription factors that share similar DNA binding specificities, we developed a method of clustering motifs based on their occurrences in the genome. We first scanned motifs enriched over genomic regions defined by the positive set. In clustering motifs enriched in taTREs, we used the taTREs merged over 20 primary GBMs as the positive set; for motifs enriched in subtype-biased TREs, we used the corresponding subtype-biased taTREs in which the motifs were enriched as the positive set. We defined the presence of transcription factor binding sites for loci (stand-specific) as a log-odds score of ≥ 2 in positive and background sets, and all other log-odds scores as an absence, with scores obtained as described in the section on motif enrichment analysis of taTREs. The Spearman’s rank order correlation coefficients were computed for each pair of transcription factors, based on their presence or absence pattern across transcription factor binding sites of all motifs of interest. Heatmaps were generated using agglomerative hierarchical clustering using ward.D2 in R.

Validation of regulation between transcription factors and target genes. Association of transcription factors with target genes. We associated transcription factors with target genes by first identifying target TREs, and then searching for target genes based on the location of these TREs. To identify target TREs, we scanned relaxed dREG-HD all GBM regions, extended by 150bp from the TRE center. We identified occurrences of each subtype-biased transcription factor binding motif in subtype-biased TREs that had a log-odds score ≥ 2 compared with a second-order Markov model. The Markov model was fit using on DNA sequence in the TRE as the background. The subset of TREs with potential binding sites for each transcription factor are referred to as query TREs. We used stringent heuristics to link the query TREs to target genes to minimize false positive targets. TREs were linked to putative target genes if the annotated transcription start site of the genes was both of the first two closest to the query TRE and within 50kb, and if the gene was differentially transcribed (P < 0.05, FDR-corrected Wald test, DESeq2) in the same direction as the query TRE.

Defining the background set of non-target genes. We defined background non-target genes as those transcription factors that were distal from (≥ 0.5 Mbp) the query TRE, which showed similar changes in transcription as target genes to control subtype. We required non-target genes to have a transcription start site ≥ 0.5 Mbp from the closest query TRE. To match changes in transcription between target and non-target genes, we subsampled (without replacement) one-half of the genes away from query TREs and differentially transcribed (P < 0.05) in the same direction as target genes, such that the distribution of log, of fold change in transcription was insignificant (two-sided Wilcoxon rank-sum test, P > 0.2). Validation of association between transcription factors and target genes. To validate the approach that we used to associate transcription factors with target genes, we computed the gene-expression to that of background non-target genes. Specifically, we used the reads per kilobase per million mapped reads (RPKM) normalized TCGA RNA-seq data from 174 patients with GBM, and used the Spearman’s rank correlation to measure the degree of co-expression. To avoid the potential co-expression that might be artificially enriched in target genes owing to a higher chance of being located in adjacent positions of the genome, we masked the
correlation coefficients between adjacent genes. We computed the significance of higher co-expression in target genes using the one-sided Wilcoxon rank-sum test.

Quantifying the association between the transcription level of transcription factors and their target genes. We used the RPKM-normalized TCGA RNA-seq data from 174 patients with GBM. We used Spearman’s rank correlation to measure the relationship between the transcription level of transcription factors and the putative target genes. We compared the difference between the distribution of correlation coefficients for target and non-target genes using a two-sided Wilcoxon rank-sum test.

Identification of transcription factors driving survival-associated programs. For each subtype-biased transcription factor, we identified the target genes as described above, and compared the hazard ratio of the target genes with that of non-target genes. We defined two background sets based on non-target genes: the closest genes whose transcription start site was within 50 kb of the query TRE, but whose transcription was unchanged across the samples representing that subtype (P > 0.2; Fig. 7a, x axis); and genes differentially transcribed (P < 0.05) in the same direction as target genes, whose transcription start sites were 0.5 Mb away from the closest query TRE (Fig. 7a, y axis). We downloaded a table of mRNA expression values obtained using a microarray for 11,861 genes, and survival time in days for 196 patients with GBM from TCGA. We computed the hazard ratio of each gene by fitting a Cox proportional hazards regression model for the survival time of patients with the highest 25% expression level over those with the lowest 25%. We used the Wilcoxon rank-sum test to compare the distribution of hazard ratios of target genes and background genes, and derived a two-tailed P value for each background set.

The hazard ratio for individual transcription factors in Fig. 7a and Supplementary Fig. 27a–c and target genes of survival-related transcription factors in Fig. 7e and Supplementary Figs. 27df and 30, were determined using the same regression model. The difference was that instead of using the upper and lower quartiles as the cutoff, we reported the hazard ratio at the threshold between the 0.1 quantile and 0.9 quantile that gave the largest difference between survival times. We tested this difference using a two-sided chi-squared test. This ensured that we reported the largest possible difference in survival time for each individual gene.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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

Data availability
All ChRO-seq and leChRO-seq data can be downloaded from the database of Genotypes and Phenotypes (dbGaP) under accession phs001646.v1.p1. Data collected from Jurkat T cells are available in the Gene Expression Omnibus under accession GSE117832.

References
56. 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).
57. 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, 1005–1013 (2002). discussion 1013–1014.
58. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864 (2011).
59. Martin, M. Catadap removes adapter sequences from high-throughput sequencing reads. EMBO Journal 17, 10–12 (2011).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
62. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. Brief. Bioinform. 14, 144–161 (2013).
63. R Development Core Team R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2010).
64. Wang, Z., Chu, T., Choate, L. A. & Danko, C. G. Rgtsvm: support vector machines on a GPU in R. Preprint at arXiv [stat.ML] (2017). https://arxiv.org/abs/1706.03544.
65. 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).
66. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome. Biol. 9, R137 (2008).
67. Hastie, T., Mazumder, R., Lee, J. & Zadeh, R. Matrix completion and low-rank SVD via fast alternating least squares. Preprint at arXiv [stat.ME] (2014). https://arxiv.org/abs/1410.2596.
68. Jolma, A. et al. DNA-binding specificities of human transcription factors. Cell 152, 327–339 (2013).
Experimental design

1. Sample size
   Describe how sample size was determined.
   No sample size calculation was performed. 20 tumors were selected to provide sufficient statistical power to identify large changes between different tumors, and matches or exceeds many recently published studies. In analyses that require additional statistical power, we also integrated our 20 sample ChRO-seq dataset with publicly accessible RNA-seq or microarray data.

2. Data exclusions
   Describe any data exclusions.
   No data were excluded from our analysis.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   All attempts at replication were successful.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   The present study is of a retrospective cohort of samples that were taken from a tissue bank. There was no interventional component in the present study. Therefore randomization is not applicable.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   Blinding was not relevant to our study because all samples represent primary GBMs, PDXs, or a normal brain reference sample. Statistical analyses in this study used these class labels to identify differences between groups. Therefore there was no need to blind the data analyst.

6. Statistical parameters
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

   - The exact sample size ($n$) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
   - A statement indicating how many times each experiment was replicated.
   - The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section).
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons.
   - The test results (e.g. $p$ values) given as exact values whenever possible and with confidence intervals noted.
   - A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range).
   - Clearly defined error bars.

   See the web collection on statistics for biologists for further resources and guidance.
Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study. We used several standard software packages, including: BWA (0.7.13-r1126), samtools (1.8), the Kent utilities suite (v350), Bedtools (v2.27.1-ge5ad7e4-dirty), R (3.5.0), DESeq2 (1.20.0), RTFBSDB (0.4.0), PRINSEQ-lite (0.20.4), and bedops (2.4.15).

We also used a lot of custom software for all analyses in this study. In all cases the methods are described in the Online methods section. Code will be distributed following publication on the Danko lab GitHub page (https://github.com/danko-lab).

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The Nature Methods guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Brain samples were consumed when collecting data for this study.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

b. Describe the method of cell line authentication used.

N/A.

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Small pieces of GBM were implanted subcutaneously in the flank of athymic nude (nu/nu) mice (Harlan Laboratories/ Envigo, Indianapolis, IN). Mice were female, aged 7-8 weeks, and weigh 20-22 grams.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Human subjects were suffering from glioblastomas. No other information is available from this cohort of subjects.