# Functional analysis of low-grade glioma genetic variants predicts key target genes and transcription factors

## SUPPLEMENTARY METHODS

### Genotype imputation

Germline genotype data of 513 LGG patients in birdseed format were obtained from the TCGA GDC Data Portal Legacy Archive[1] for imputation. Instead of tumor tissues, genotype data from matched normal tissues were used for all analyses to avoid miscalls from genotyping error and somatic mutations. The genotypes of the tag SNPs measured by Affymetrix human SNP array 6.0 were matched to hg19 coordinates using Affymetrix genome-wide SNP annotation file. Tag SNPs with genotype confidence score > 0.01 were filtered out. Untagged SNPs were imputed and phased from tagged SNPs of 513 LGG patients using the Michigan Imputation Server[2]. We chose Haplotype Reference Consortium (HRC) panel[3] (version r1.1 2016) as the imputation reference panel and Eagle2[4] as the phasing engine. Imputed genotypes were retained if the minor allele frequency (MAF) exceeded 0.005 and estimated imputation accuracy ($R^2$) exceeded 0.4. Then, imputed genotypes were retained if the maximum of the estimated posterior probabilities for genotypes 0/0, 0/1 and 1/1 exceeded 0.9. Here, 0 denotes the reference allele and 1 denotes the alternative allele of the SNPs. We extracted the haplotype of GWAS and exonic SNPs in the target genes using the imputed and phased results from the Michigan Imputation Server[2].

### eQTL linear model

We performed eQTL analysis using the TCGA LGG data to identify candidate target genes associated with the GWAS SNPs. We imputed high-confidence genotypes at the GWAS SNPs and

restricted the eQTL analysis to the genes residing within the 4 Mb window centered at each GWAS SNP. We used the following multivariate linear regression model to assess the association between GWAS SNP's genotype and the gene expression level, while adjusting for gene copy number, tumor site, tumor grade, histological diagnosis, and gender:

$$E_i = \alpha_i + \beta_i \cdot GT + \gamma_i \cdot \overline{CNS_i} + \sum_{j=1}^{4} \theta_{ij} \cdot Cov_j + \varepsilon_i,$$

where $i$ indexes the genes within the 4 Mb window centered at the SNP; $E_i$ = $\log_2$(RSEM+1) denotes the log-transformed gene expression level in RSEM units; $GT \in \{0, 1, 2\}$ denotes the number of alternative alleles of the GWAS SNP with respect to the human reference genome; $\overline{CNS_i}$ is the length-weighted average of tumor copy number segmentation, $log_2(copy\ number/2)$, covering gene $i$; $Cov_j$ represents each of the four covariates included in the eQTL analysis: tumor site, tumor grade, histological diagnosis, and gender, where tumor site ∈ {"supratentorial, frontal lobe", "supratentorial, occipital lobe", "supratentorial, parietal lobe", "supratentorial, temporal lobe", "supratentorial, not otherwise specified"}, tumor grade ∈ {"grade II", "grade III"}, histological diagnosis ∈ {"astrocytoma", "oligodendroglioma", "oligoastrocytoma"} and gender ∈ {"male", "female"}; $\alpha_i$ denotes the intercept; and, $\varepsilon_i$ denotes the error term.

**Epigenomic data**

We obtained fetal brain DNase I hypersensitive sites sequencing (DNase-seq) and histone modification (H3K4me1, H3K4me3 and H3K27ac) chromatin immunoprecipitation followed by sequencing (ChIP-seq) datasets of brain tissues from the Roadmap Epigenomics Mapping Consortium (REMC) database[5]. Primary tumor assay for transposase-accessible chromatin sequencing (ATAC-seq) aligned BAM files of glioma patients were downloaded from the TCGA

Data Portal[1]. Processed ATAC-seq[6], H3K27ac and proximity ligation-assisted ChIP-seq (PLAC-seq) data in oligodendrocytes were obtained from *Nott et al.*[7] (https://genome.ucsc.edu/s/nottalexi/glassLab_BrainCellTypes_hg19). CIC ChIP-seq datasets were obtained from Gene Expression Omnibus accession numbers GSE125166 (human mesenchymal stem cells), GSE75115 (human endometrial stromal cells) and GSE109619 (acute myelogenous leukemia cell line KG1).

For the ATAC-seq data, there were 13 samples in total with the aligned reads (hg38) in BAM format. We extracted the read counts by allele at the rs648044 location using bcftools mpileup[8] option. We considered only the bases with a Phred quality score of at least 20. Out of 13 samples, three were removed because the imputed genotype status of rs648044 was not heterozygous. The genotype status of rs648044 in one of the samples was imputed to be homozygous alternative allele (AA) but was retained because the ATAC-seq reads showed high coverage for both alleles at the SNP location. For each sample, the significance of the skew between the two alleles was evaluated using a binomial test. The resulting *P*-values were then combined using the Fisher's method from the R package metap[9] (Supplementary Table 2).

**TF binding affinity perturbation analysis**

Position weight matrices (PWMs) for transcription factors (TFs) were obtained from the following databases: JASPAR[10], HOCOMOCO Human v10[11], Jolma2013[12], and TRANSFAC[13]. We used the list of LGG SNPs (GWAS + high LD SNPs) and PWMs, together with the motif finding tool FIMO[14], to calculate the binding affinity of TFs interacting with two versions of a sequence harboring the different SNP alleles (see Zhang *et al.*[15] for further details). We also performed a neutral mutation-model permutation test[15] to assess the statistical significance of the difference in

binding affinity scores for a TF between the two alleles of a given SNP. Briefly, the permutation test is based on a null model that a random mutation of any motif position does not alter the binding affinity score. The sampling probability of the location $i$ to be mutated was chosen to be proportional to 1/D($PWM_i$ $|Q$), where D($PWM_i$ $|Q$) is the Kullback-Leibler divergence between the PWM probability vector at location $i$ and the genomic background nucleotide frequency vector $Q \equiv (p_A, p_C, p_G, p_T) = (0.3, 0.2, 0.2., 0.3)$. Given the choice of location $i$ to be mutated, the base were randomly mutated by sampling the substitution nucleotides from $PWM_i$. We chose a permutation test significance level of $P < 0.05$ to identify TFs whose binding might be significantly disrupted by the SNP.

**TF-Target gene correlation analysis**

We evaluated TF-target gene expression correlation coefficients in the three genotype groups (homozygous risk, heterozygous and homozygous non-risk) to further prioritize TFs obtained from TF binding affinity perturbation analysis. We started with a list of putative target genes from eQTL analysis and candidate TFs from motif analysis. For each target gene and TF pair, we calculated Pearson correlation coefficients between the target gene and TF expression values ($\log_2$(RSEM+1) units) stratified into the three genotype groups. We performed both subgroup-specific and combined group ("*IDH*$^{mut}$ only" and triple-positive) analyses. We then prioritized TFs based on the reasoning that the correlation would likely be strongest (weakest) in the homozygous genotype group preserving (disrupting) the TF motif (see Supplementary Fig. 1 in Zhang *et al*.[15]).

***CIC* inactivating mutations**

We acquired the mutation calls from the four somatic variant calling pipelines: MuSE[16],

MuTect2[17], SomaticSniper[18] and VarScan2[19], available in the GDC Data Portal[1]. For each of the four pipelines, we then obtained all *CIC* inactivating mutations based on whether the PolyPhen column contained the "probably_damaging" term or the IMPACT column was "HIGH"[20]. To reduce false positives, we finally retained only those inactivating mutation calls detected by at least two of the four pipelines.

**Electrophoretic Mobility Shift Assay (EMSA)**

EMSA was performed with the mixture of the recombinant MafF protein and four different DNA oligonucleotides: Positive Control (PC), rs648044 locus containing the A allele, rs648044 rs648044 locus containing the G allele, Negative Control (NC). The sequences of the oligonucleotides and recombinant MafF are given below. Two complementary oligonucleotide strands were mixed in a PCR tube and hybridized by increasing the temperature to 98°C and lowering by 5°C every 5 minutes until the temperature reached 4°C using a thermocycler (Mastercycler Personal, Eppendorf). For the binding of MafF, 4 pmole of the hybridized oligonucleotide and 32 pmole of MafF were mixed with 10 mM $MgCl_2$ in T50 buffer (10 mM Tris-Cl, 50 mM NaCl, pH 8.0). For MafF negative samples, the same materials were mixed, except for MafF. The resulting mixtures were incubated at 37°C for 30 minutes. After the incubation, the mixtures were subjected to polyacrylamide gel electrophoresis, fluorescently labeled by SYBR Gold Nucleic Acid Gel Stain (S11494, ThermoFisher) and visualized on a UV illuminator (Dyna Light UV Transilluminator, Labnet International, Inc.).

The recombinant MafF is purchased from Abnova (GST-tag removed, catalog number: H00023764-Q01), and the recombinant MafF sequence is:

MSVDPLSSKALKIKRELSENTPHLSDEALMGLSVRELNRHLRGLSAEEVTRLKQRRRTL

KNRGYAASCRVKRVCQKEELQKQKSELEREVDKLARENAAMRLELDALRGK.

The oligonucleotide sequences representing the rs648044 locus are given below, where the core binding motif of MAFF is highlighted and underlined:

81 bp flanking sequence harboring the rs648044-A allele:

5'-CCTTGCACTGGCACATTCCTGCTGTTTTCTTCTGC**<u>TCAGCA</u>**GAGCCGAACG GCTCTCACTTCCTGGCTAGCTCTGTGTGCT-3'

81 bp flanking sequence harboring the rs648044-G allele:

5'-CCTTGCACTGGCACATTCCTGCTGTTTTCTTCTGC**<u>TCAGCG</u>**GAGCCGAACG GCTCTCACTTCCTGGCTAGCTCTGTGTGCT-3'

The control sequences were designed based on ChIP-seq results (explained in the following section). All oligonucleotides were purchased from Integrated DNA Technologies.

**EMSA experiment MAFF positive control (PC) and negative control (NC) sequences**

We first obtained MAFF ChIP-seq peaks in HepG2, K562 and HelaS3 cell lines from ENCODE (HepG2: ENCFF611VKE; K562: ENCFF864KPF; HelaS3: ENCFF575MOS). We then ranked the peaks by *q*-value in each cell type to obtain top peak regions. We intersected the top peak regions and scanned the sequences using FIMO[14], keeping only the consensus peaks containing a MAFF binding motif. Upon visual inspection of raw ChIP-seq signals of the top remaining candidates, we chose chr14:77423081-77423161 (hg19) as the 81bp-long positive control sequence centered around a MAFF core binding motif (TCAGCA).

For the negative control sequence, we first scanned the 81 bp flanking sequence harboring the rs648044-A allele and obtained all sub-sequences which might partially contain a core MAFF

motif. We then randomly permuted the nucleotides in those subsequences. We checked that the resulting negative control sequence satisfied the following two criteria: (1) there were no more than three adjacent nucleotides of MAFF core binding motif (TCAGCA or its reverse compliment TGCTGA); (2) the GC content of the negative control sequence was approximately the same as the original sequence harboring the rs648044-A allele.

The positive control and negative control sequence are provided below, where the core binding motif of MAFF is highlighted in the positive control:

Positive control sequence:

5'-GTTCCCGCCGCCCGGAGGCTCATTTGTCACCCGC**TGCTGACTCAGCA**CTT CTGCAGAAGGCTTTTCCCTCCGCTTTGGAGG-3'

Negative control sequence:

5'-ATAACACAGAGCTAGCCACGAAGTGAGAGCCGTTCGCCTCGTGGACGGA GAAGAAAACACCCGGAATGTGCCAGTGCAATT-3'

**Cell Culture, RNAi and RNA expression**

The cell line SF10417 (from UCSF) derived from a human *IDH1*$^{R132H}$ mutant, *TERT* promoter-mutant, 1p/19q-codeleted oligodendroglioma was used to assess the effect of MAFF knockdown on *ZBTB16* and *NCAM1* expression. The cells were grown in NeuroCult NS-A media (STEMCELL Technologies) supplemented with L-Glutamine, B27, N2, Sodium Pyruvate and Pen/Strep (Life Technologies) in the presence of growth factors bFGF, EGF (STEMCELL

Technologies), and PDGFAA (PeproTech). Lentivirus were produced with short-hairpin RNA (shRNA) with either a non-target shRNA control vector or a vector designed to reduce the expression of *MAFF* ($n$ = 3 independent constructs, Sigma Mission). Cells were infected with a MOI of 1, followed by selection of transduced cells with puromycin. Populations were subcloned and total RNA was isolated for three independent clones for each of the vectors (Qiagen). First strand cDNA synthesis was completed with SuperScript IV VILO (Invitrogen) and gene expression was measured with TaqMan probes according to manufacture guidelines for genes *18S*, *MAFF*, *ZBTB16* and *NCAM1* (Applied Biosystems).

**TF binding prediction using a convolutional neural network (CNN)**

We trained a CNN using the Keras package[21] and engineered several layers using TensorFlow[22]. The structure of the constructed CNN is illustrated in Fig. 4C. Each input sample consisted of a 1001 x 9 matrix containing the one-hot encoded forward and reverse DNA sequence information and the quantile-normalized DNase-seq signal in the given 1001bp region. The CNN model had one convolutional layer consisting of 40 convolutional filters. All convolutional filters had a size of 12 x 5 and slid on a 1001 x 5 input matrix representing positive strand DNA sequence and DNase-seq signal. To capture the motifs present in negative strands, the 1001 x 5 submatrix representing the corresponding reverse complement DNA sequence and DNase-seq signal was passed through the same set of convolutional filters. We then extracted the maximum of the convolutional layer output from the positive and negative strands, and passed the output through a max pooling layer of size 40 and stride 40. The max pooling layer output was flattened and passed to a fully connected layer of 80 neurons, and then was passed through a fully connected layer of 10 neurons. Finally, the output of the second fully connected layer was passed to a single output

neuron encoding the binding probability of the TF. Rectified linear unit (ReLU) function was used as the activation function throughout the CNN, except for the output layer where we utilized a sigmoid function to restrict the output between 0 and 1. The loss function of the CNN model is Binary Cross-Entropy (CE) loss:

$$CE = \frac{1}{N}\sum_{i=1}^{N}[-t_i \log(\sigma_i) - (1 - t_i)\log(1 - \sigma_i)],$$

where $N$ is the batch size; $t_i \in \{0,1\}$ represents whether the TF truly binds the DNA or not, based on ChIP-seq result; and $\sigma_i$ is the output of the sigmoid function from the last layer. The CNN was trained using Adam optimizer[23] with batch size 1000, and the training was stopped when the validation loss did not decrease for over 100 epochs.

We trained the CNN using SP1 ChIP-seq data from six cell lines/tissue (H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7) and tested its performance using the cell line A549 (Supplementary Table 5). We obtained DNase-seq and ChIP-seq data in the above cell lines from ENCODE or REMC databases, and performed quantile normalization to reduce batch effect. We then collected all optimal ChIP-seq peak regions centered around peak centers to form the positive dataset (102934 samples) and selected an equal number of regions with no ChIP-seq peak to form the negative dataset. To increase the number of training samples and reduce overfitting, we then translated our initial positive and negative datasets by -20 bp, -10 bp, 0 bp, 10 bp, and 20 bp to form the translated positive and negative datasets. After removing the samples that fell into hg19 "blacklist" regions, we obtained a total of 514668 samples for the translated positive dataset and 514634 samples for the translated negative dataset. We then split the translated datasets into training and validation datasets with ratio 80% to 20% (training dataset: 823441 samples; validation dataset: 205861 samples). We tested the performance of the trained CNN using A549 chromosome 1 (chr1) positive and negative datasets (3785 samples for each) and calculated the

receiver operating characteristic (ROC) curve (Supplementary Fig. 7A). Finally, the trained CNN was used to predict the binding affinity of SP1 at rs12225399 using the REMC fetal brain DNase-seq samples as input (Supplementary Table 5).

**Extraction of CNN-learned motif using simulated annealing**

We used the simulated annealing technique[24] to perform probabilistic optimization of CNN-learned motifs over the set of all possible input sequences. The simulated annealing algorithm samples sequences through a discrete-time inhomogeneous Markov chain with transition probabilities determined by a cost function $J$ and a temperature parameter $T$. Specifically, sequences with lower $J$ values are sampled with higher probability, and as the temperature decreases, the ratio of the sampling probability for sequences with lower $J$ values to the sampling probability for sequences with higher $J$ values increases.

To sample sequence $\boldsymbol{x}$ that maximizes the input of the sigmoid function in the last layer of the trained CNN, we used the cost function $J(\boldsymbol{x}) = -y(\boldsymbol{x})$, where $y(\boldsymbol{x})$ denotes the pre-activation of the output neuron. For the trained CNN, we initialized 50 instances of simulated annealing at the 50 elements of the test set (A549 chr1 ChIP-seq dataset) predicted to have the highest pre-activation. The pseudocode for the simulated algorithm is given below; $\boldsymbol{x}_n$ is the input sequence at the $n^{\text{th}}$ iteration, $d$ is the initial temperature, and $N_{iter}$, $N_{sample}$ $N_{interval}$ are the iteration count, sample size and interval size, respectively. At each iteration, we only changed one nucleotide in the sequence, while the quantile-normalized DNase-seq signal was held unchanged.

**Algorithm: Simulated Annealing**

---

**Given**: $\boldsymbol{x_0}$, $J(\boldsymbol{x})$, $d$, $N_{iter}, N_{sample}, N_{interval}$

**for** $n$ in $(1,2,\ldots N_{iter})$

$T = d/\log(n+1)$

## select one base index in the input sequence

$i \sim \text{Unif}(\{0,\ldots,1001\})$,

$\boldsymbol{x}_{proposed} \equiv \boldsymbol{x}_{n-1}$

## replace the nucleotide in the base selected with a randomly sampled nucleotide

$(\boldsymbol{x}_{proposed})_i \sim \text{Unif}(\{A,C,G,T\} - \{(\boldsymbol{x}_{n-1})_i\})$

$u \sim \text{Unif}([0,1])$

**if** $\exp\left(-\frac{J(\boldsymbol{x}_{proposed}) - J(x_{n-1})}{T}\right) > u$ ## Accept

$\boldsymbol{x}_n = \boldsymbol{x}_{proposed}$

**else** ## Reject

$\boldsymbol{x}_n = \boldsymbol{x}_{n-1}$

**return** $\{\ \boldsymbol{x}_n\ :\ (n > N_{iter} - (N_{sample})(N_{interval}))$ **and** $n \bmod N_{interval} = 0\ \}$

---

The initial temperature $d$ is chosen by empirical observation of the distribution of the cost function $J$, in particular, by considering the height of communication of local minima. Although we could not guarantee obtaining global minima at the end of the simulation, we still wanted to ensure that Markov chains could transition from shallow to deeper basins of the cost function. We therefore selected the initial temperature $d$ by estimating the minimum communication height for sequences near the local minima of $J$. First, we supplemented each of the 50 test set inputs (A549 chr1) predicted to have the highest pre-activation of $y(\boldsymbol{x})$ with 1000 inputs sampled uniformly and

without replacement from the union of training, validation, and test sets. To remove the variation in $J(\boldsymbol{x})$ caused by DNase-seq, we fixed the DNase-seq signal of all 1000 samples to the corresponding maximum pre-activation test input DNase-seq signal. Next, for each of the test inputs and the supplemented 1000 samples, we calculated the difference $(J_{i,k} - \min_j (J_{i,j}))$, where $1 \leq k \leq 1001, 1 \leq i \leq 50$, and $J_{i,k}$ denotes the $k^{\text{th}}$ value of $J$ among the 1001 samples associated with the test set input $i$. We then combined the calculated difference values, and ranked these 50000 difference values. We observed a transition from a rapid increase in $J$ for inputs in the lowest percentiles of $J$, to a moderate increase in $J$ for the bulk of the remaining inputs. We thus estimated the fraction of inputs near strong local minima and chose the 1st percentile of the ranked $\cup_{i=1}^{50} \left\{ (J_{i,k} - \min_j (J_{i,j})) : 1 \leq k \leq 1001 \right\}$ as a threshold, yielding 2.24 as the initial temperature. The values of the other parameters were $N_{iter} = 5 \times 10^5$, $N_{sample} = 10^4$ and $N_{interval} = 10$.

Using these parameters, we performed simulated annealing starting from each of the 50 test inputs which were predicted to have the highest pre-activation values. For each of the 50 simulated annealing experiments, we monitored the minimum of $J(\boldsymbol{x})$ across the previous iterations versus the iteration number $n$. After the minimization of $J(\boldsymbol{x})$ stabilized, we recorded the sampled sequences every 10 iterations starting from the 400000th iteration for each simulated annealing experiment. We then chose the experiment with the lowest stable $\min(J(\boldsymbol{x}))$ out of the 50 experiments as our best scenario, and visualized the CNN-learned motif using the recorded sequences through WebLogo[25] 3.

**ALG$^3$ web resource**

WashU genome browser was embedded in our web resource using the source code from their GitHub repository[26,27]. Gencode V19 genes, REMC fetal brain (DNase, H3K4me1) and REMC

dorsolateral prefrontal cortex (H3K4me1, H3K27ac) datasets were obtained from the WashU repositories (Supplementary Table 7). ReMap 2018[28] data containing genome-wide ChIP-seq peaks processed from public and ENCODE databases were downloaded from the ReMap website (Supplementary Table 7).

## SUPPLEMENTARY TABLES

**Supplementary Table 1**: Twenty five LGG GWAS loci analyzed in this study. Reference and alternative alleles in the human reference genome are provided, with the risk allele underlined.

| GWAS SNP ID | Reference allele | Alternative allele |
|---|---|---|
| rs2736100 | C | A |
| rs55705857 | A | G |
| rs4977756 | G | A |
| rs11196067 | A | T |
| rs11599775 | G | A |
| rs648044 | A | G |
| rs498872 | A | G |
| rs12803321 | G | C |
| rs1275600 | T | A |
| rs1801591 | G | A |
| rs77633900 | G | C |
| rs78378222 | T | G |
| rs6010620 | A | G |
| rs4252707 | G | A |
| rs12076373 | G | C |
| rs7572263 | A | G |
| rs11706832 | A | C |
| rs11598018 | C | A |
| rs7107785 | T | C |
| rs10131032 | G | A |
| rs3751667 | C | T |
| rs10069690 | C | T |
| rs75061358 | T | G |
| rs634537 | T | G |
| rs2297440 | T | C |

**Supplementary Table 2**: Allele-specific ATAC-seq read counts covering the GWAS SNP rs648044 and the corresponding two-sided binomial test *P*-values for TCGA-LGG samples. The *P*-values were combined using the Fisher's method.

| **Aliquot barcode** | **A** | **G** | **Binomial *P*-value** |
|---|---|---|---|
| TCGA-P5-A735-01A-31-A617-42-X017-S03 | 47 | 45 | 4.59E-01 |
| TCGA-E1-A7YI-01A-31-A617-42-X020-S08 | 84 | 83 | 5.00E-01 |
| TCGA-DU-5870-02A-21-A646-42-X030-S02 | 61 | 54 | 2.88E-01 |
| TCGA-FG-A4MU-01B-31-A615-42-X016-S10 | 61 | 40 | 2.30E-02 |
| TCGA-P5-A72W-01A-31-A617-42-X018-S01 | 69 | 40 | 3.52E-03 |
| TCGA-DU-6407-02B-21-A645-42-X036-S01 | 112 | 98 | 1.85E-01 |
| TCGA-P5-A72X-01A-31-A617-42-X014-S09 | 8 | 5 | 2.91E-01 |
| TCGA-FG-A4MY-01A-31-A616-42-X015-S02 | 25 | 16 | 1.06E-01 |
| TCGA-F6-A8O3-01A-31-A617-42-X013-S07 | 35 | 33 | 4.52E-01 |
| TCGA-W9-A837-01A-31-A617-42-X019-S03 | 16 | 15 | 5.00E-01 |
| **Total** | 518 | 429 | **1.00E-02 (Fisher's method)** |

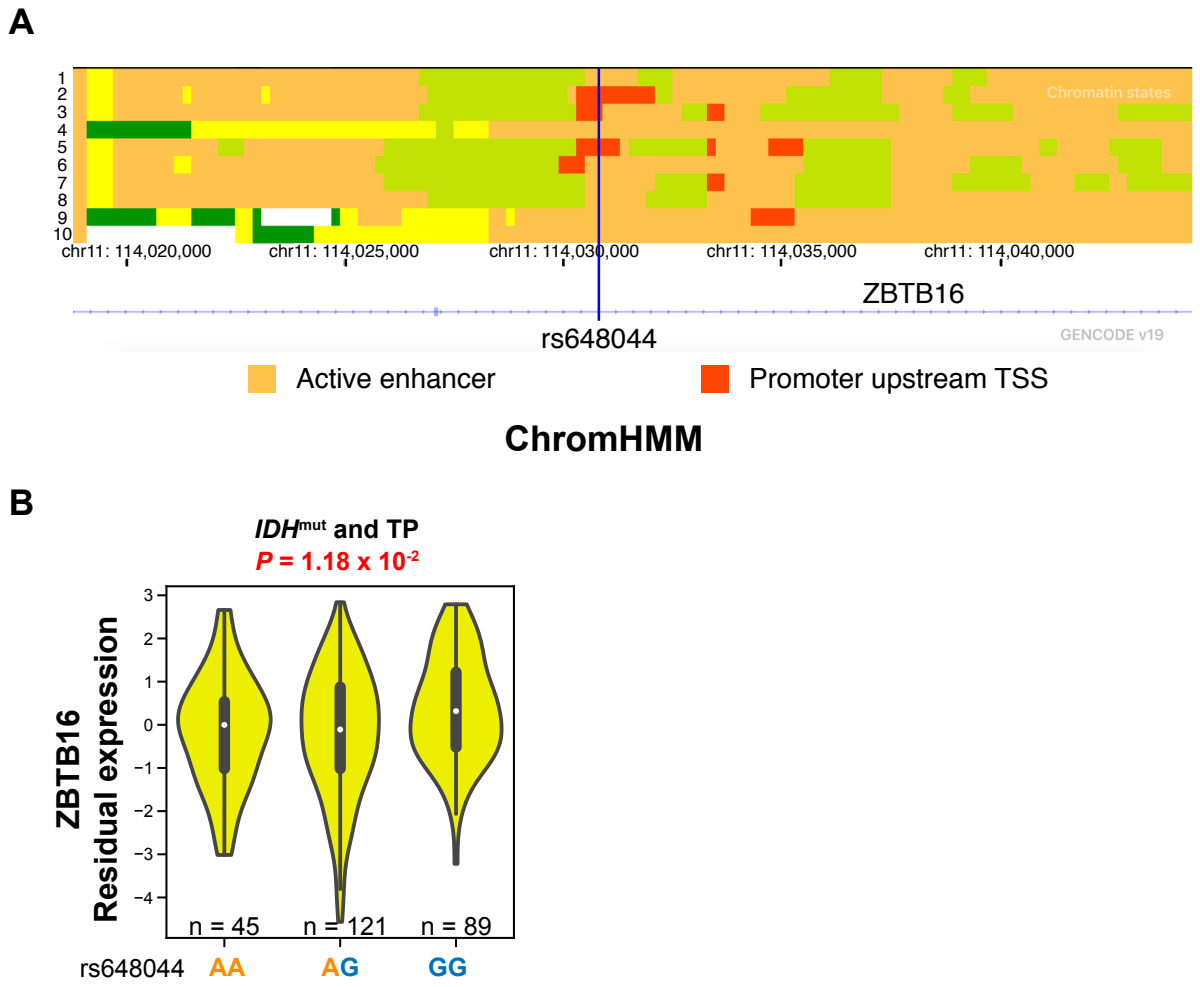
**Supplementary Table 3**: Inactivating mutation status of *CIC* in the triple-positive group gliomas stratified into non-risk (GG) and risk (AA+AG) genotypes of rs648044.

| **CIC status** | **GG** | **AA+AG** |
|---|---|---|
| **Mutant** | 16 | 25 |
| **Wild-type** | 15 | 48 |

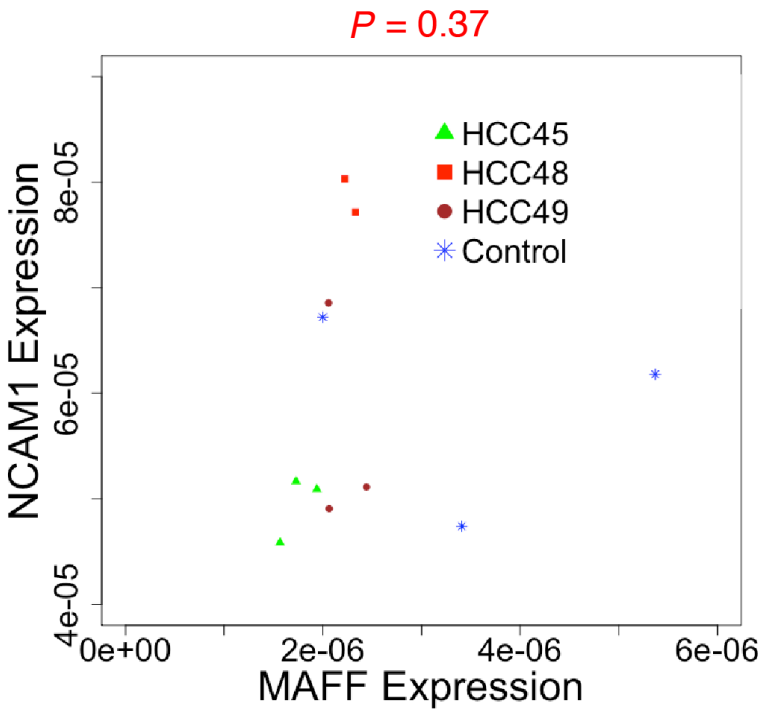
**Supplementary Table 7**: Source and URL for the preloaded tracks in the embedded genome browser of ALG[3].

| **Track name** | **Source** | **URL** |
|---|---|---|
| Gencode v19 genes | GENCODE | https://egg.wustl.edu/d/hg19/gencodeV19.gz |
| ReMap 2018 peaks | ReMap | http://tagc.univ-mrs.fr/remap/index.php |
| Fetal Brain DNase – Male, Female | REMC | https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E081-DNase.fc.signal.bigwig, https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E082-DNase.fc.signal.bigwig |
| Fetal Brain H3K4me1 – Male, Female | REMC | https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E081-H3K4me1.fc.signal.bigwig, https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E082-H3K4me1.fc.signal.bigwig |
| Dorsolateral Prefrontal Cortex – H3K4me1, H3K27ac | REMC | https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E073-H3K4me1.fc.signal.bigwig, https://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidated/macs2signal/foldChange/E073-H3K27ac.fc.signal.bigwig |

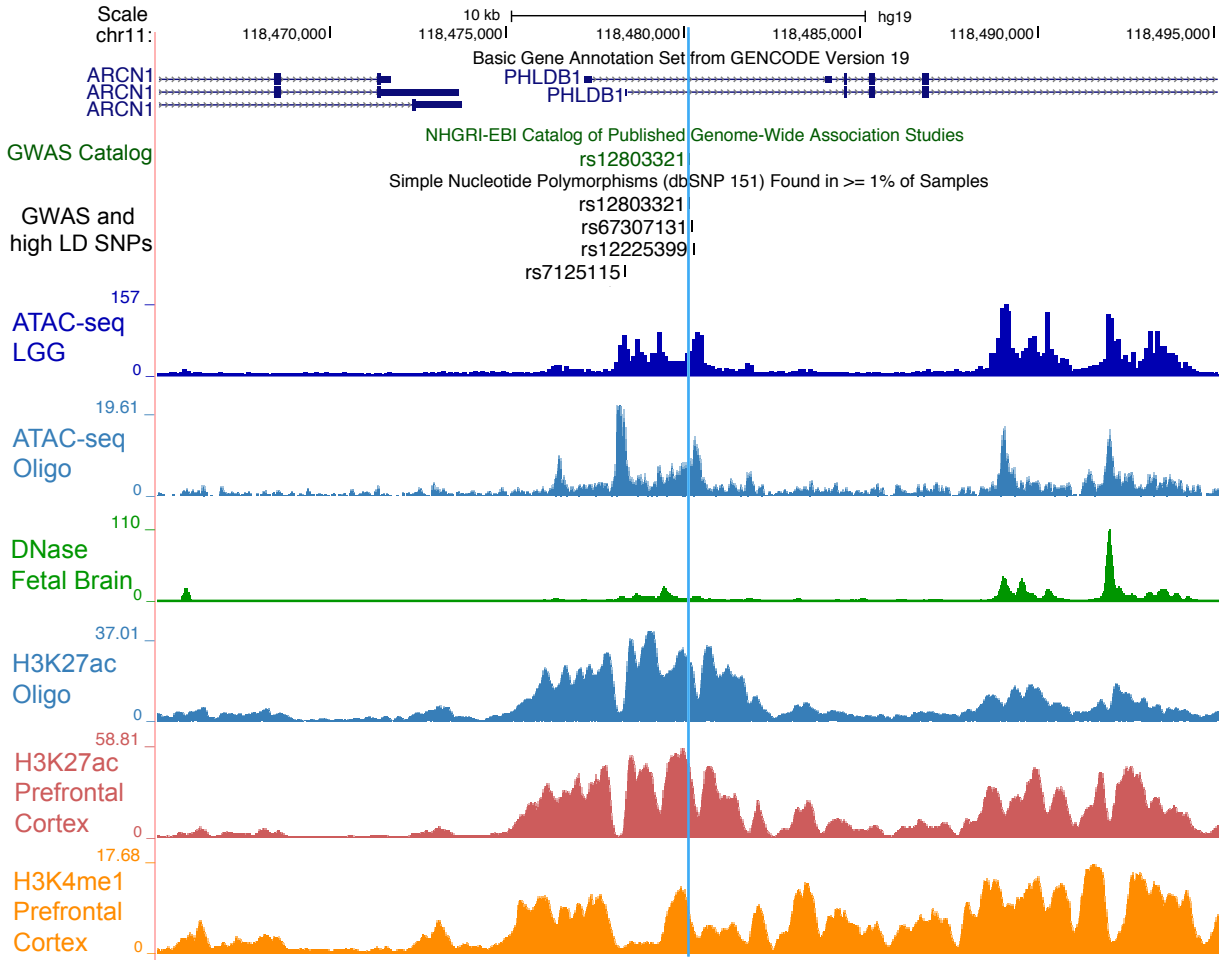
## SUPPLEMENTARY FIGURES



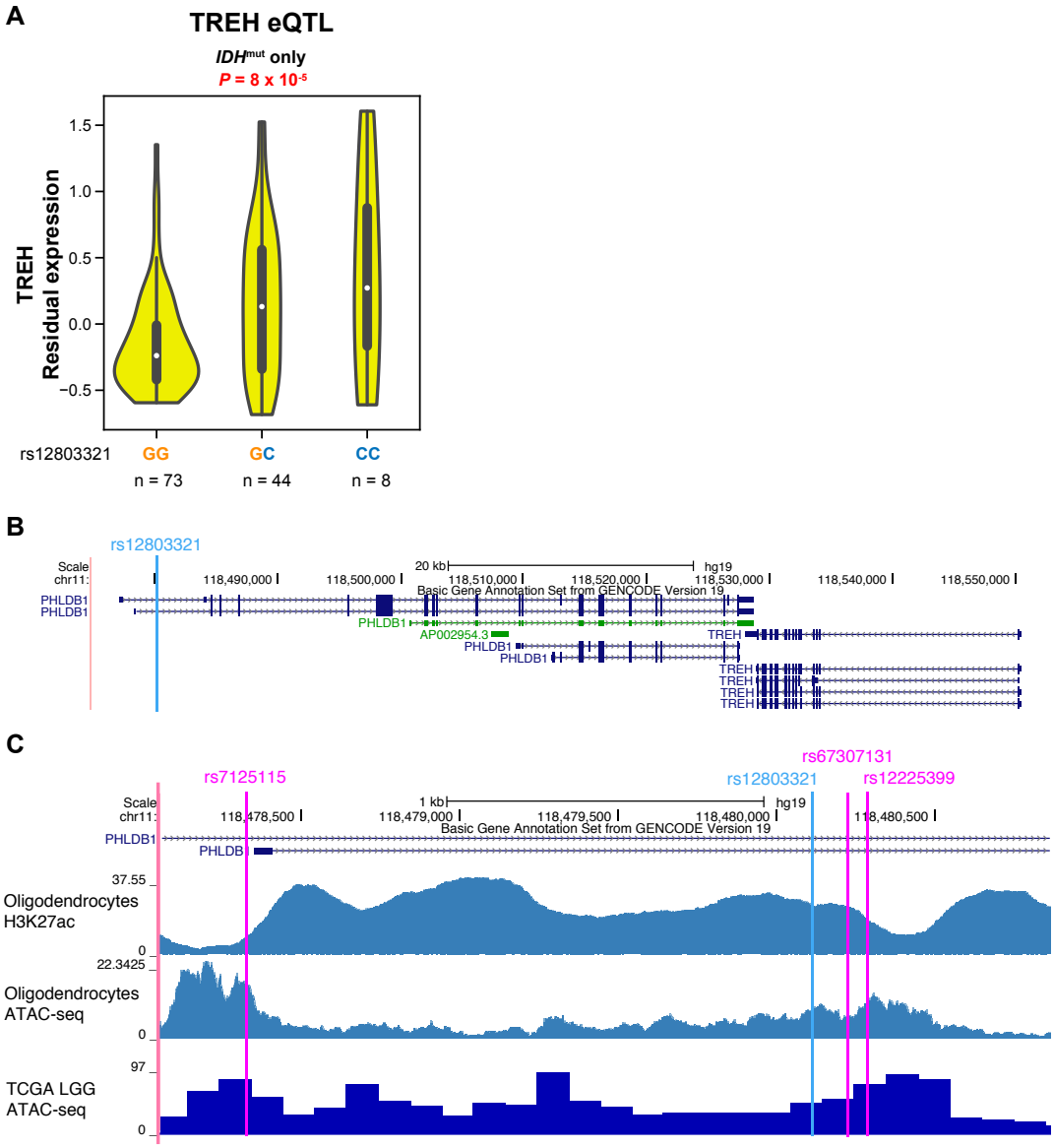

**Supplementary Fig. 1:** The GWAS SNP rs648044, located in an intron of *ZBTB16*, modulates *ZBTB16* mRNA expression. (A) ChromHMM[29] tracks of 10 brain tissue samples from Roadmap Epigenomics Mapping Consortium (REMC) database[5] for the region harboring the GWAS SNP rs648044. The samples are Brain Angular Gyrus, Anterior Caudate, Cingulate Gyrus, Germinal Matrix, Hippocampus Middle, Inferior Temporal Lobe, Dorsolateral Prefrontal Cortex, Substantia Nigra, Fetal Brain Female and Fetal Brain Male. (B) eQTL result for rs648044 and *ZBTB16* in the combined TCGA-LGG "*IDH*$^{mut}$ only" and triple-positive (TP) group. Throughout the text, the risk and non-risk alleles of a SNP are colored orange and blue, respectively.



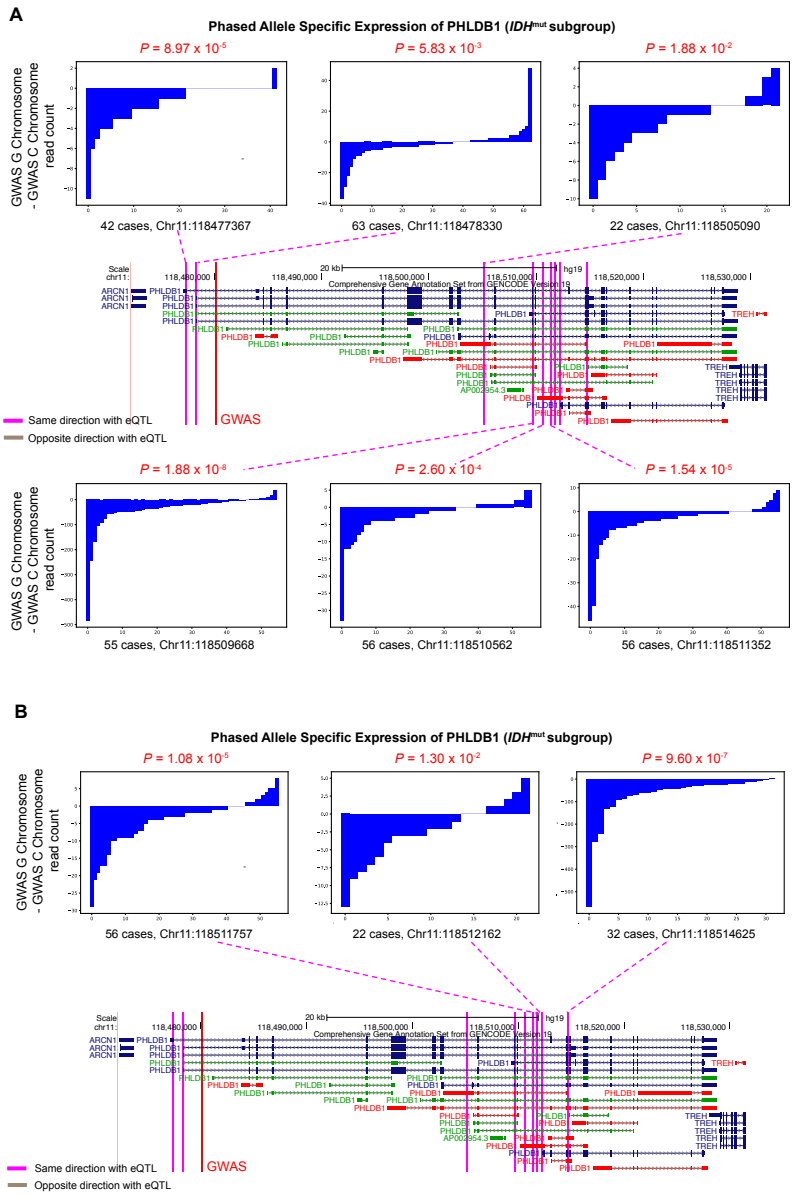

**Supplementary Fig. 2:** The results of MAFF RNAi knockdown experiment showing insignificant effect on *NCAM1* expression. One-sided *t*-test *P*-value between the control group and the combined group of three independent shRNA clones is shown on top of the figure.



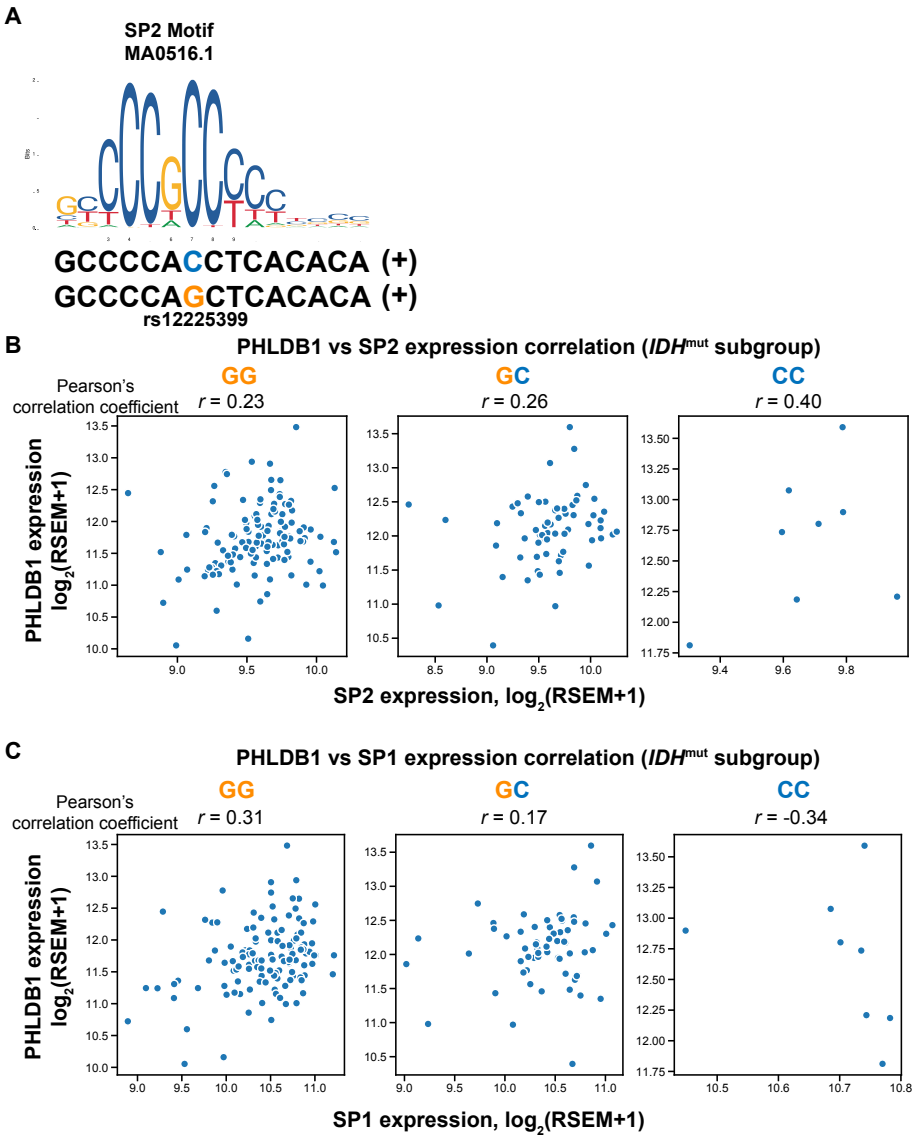

**Supplementary Fig. 3:** The epigenomic landscape of the region harboring the GWAS SNP rs12803321. A snapshot of the *PHLDB1* locus where the GWAS SNP rs12803321 is denoted by a blue vertical line. The top three tracks are: basic gene annotation set from GENCODE version 19, LGG GWAS SNPs from GWAS catalog[30] and SNPs in high LD with rs12803321 from the Single Nucleotide Polymorphism Database[31] (dbSNP 151). The lower epigenomic tracks are: TCGA-LGG ATAC-seq, oligodendrocytes ATAC-seq[7], and REMC data in fetal brain and prefrontal cortex.




**Supplementary Fig. 4:** eQTL analysis for the candidate target gene *TREH*, and epigenomic landscape of the region harboring the GWAS SNP rs12803321. (A) eQTL result for rs12803321 and *TREH* in the TCGA-LGG "*IDH*$^{mut}$ only" subgroup. (B) Gene track showing the GWAS SNP rs12803321 and candidate target genes, *PHLDB1* and *TREH*. rs12803321 is denoted by a blue vertical line. (C) Enlarged view of the epigenomic landscape of the region harboring the GWAS SNP rs12803321 and three high LD SNPs. Tracks from top to bottom are the oligodendrocytes H3K27ac[7], oligodendrocytes ATAC-seq[7] and TCGA-LGG ATAC-seq signals.

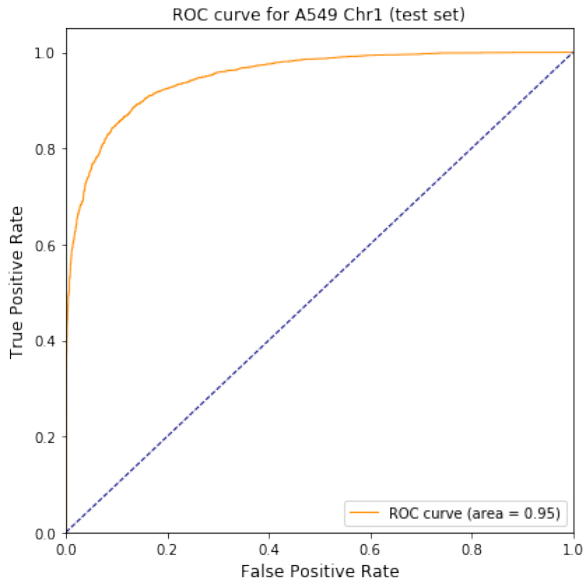



**Supplementary Fig. 5:** (A, B) Phased allele-specific transcription pattern of *PHLDB1* at 9 exonic SNPs in the TCGA-LGG "*IDH*$^{mut}$ only" molecular group. All 9 SNPs show higher transcription emanating from the rs12803321-C haplotype, consistent with the eQTL analysis. For each exonic SNP, RNA-seq read count differences between the two chromosomes harboring the rs12803321-G allele and the rs12803321-C allele are sorted across patients and shown as bar plots. The *P*-value from the Wilcoxon signed-rank sum test is shown at the top of each bar plot. The genomic locations of these 9 exonic SNPs are shown in the GENCODE Version 19 track.

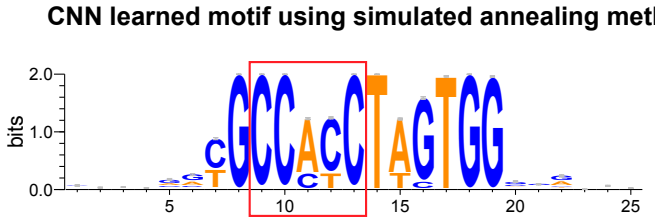



**Supplementary Fig. 6:** The SNP rs12225399 in high LD with the GWAS SNP rs12803321 likely modulates *PHLDB1* expression through perturbing the binding affinity of SP1/SP2. (A) SP2 motif logo MA0516.1 (JASPAR[10]) and two versions of the flanking sequence harboring the rs12225399-C and rs12225399-G alleles. (B) Scatter plots of *SP2* vs. *PHLDB1* expression in the three genotypes of rs12225399 in the TCGA-LGG "*IDH*$^{mut}$ only" subgroup. Spearman's correlation coefficients between *SP2* and *PHLDB1* in the "*IDH*$^{mut}$ only" group are: $\rho = 0.25$ (rs12225399-GG), $\rho = 0.26$ (rs12225399-GC), $\rho = 0.36$ (rs12225399-CC). (C) Scatter plots of *SP1* vs. *PHLDB1* expression in the three genotypes of rs12225399 in the TCGA-LGG "*IDH*$^{mut}$ only" subgroup.

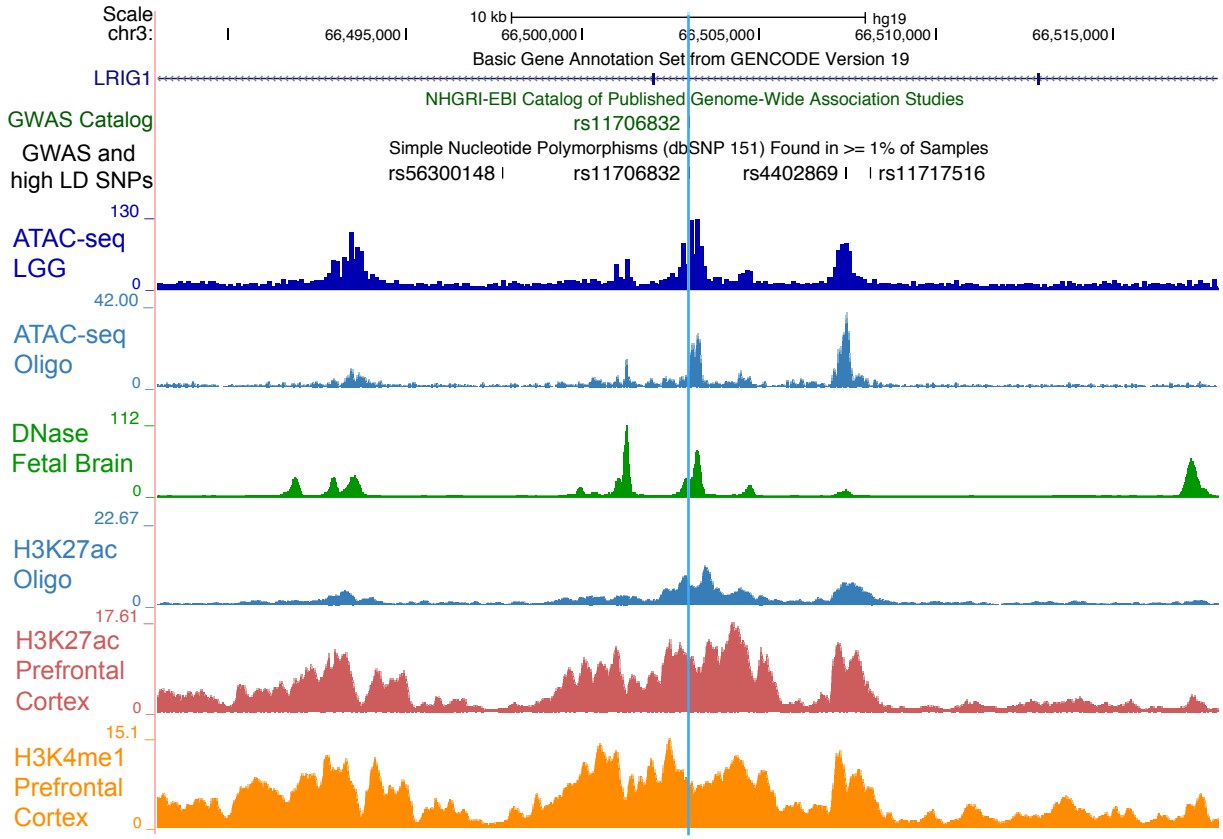
A



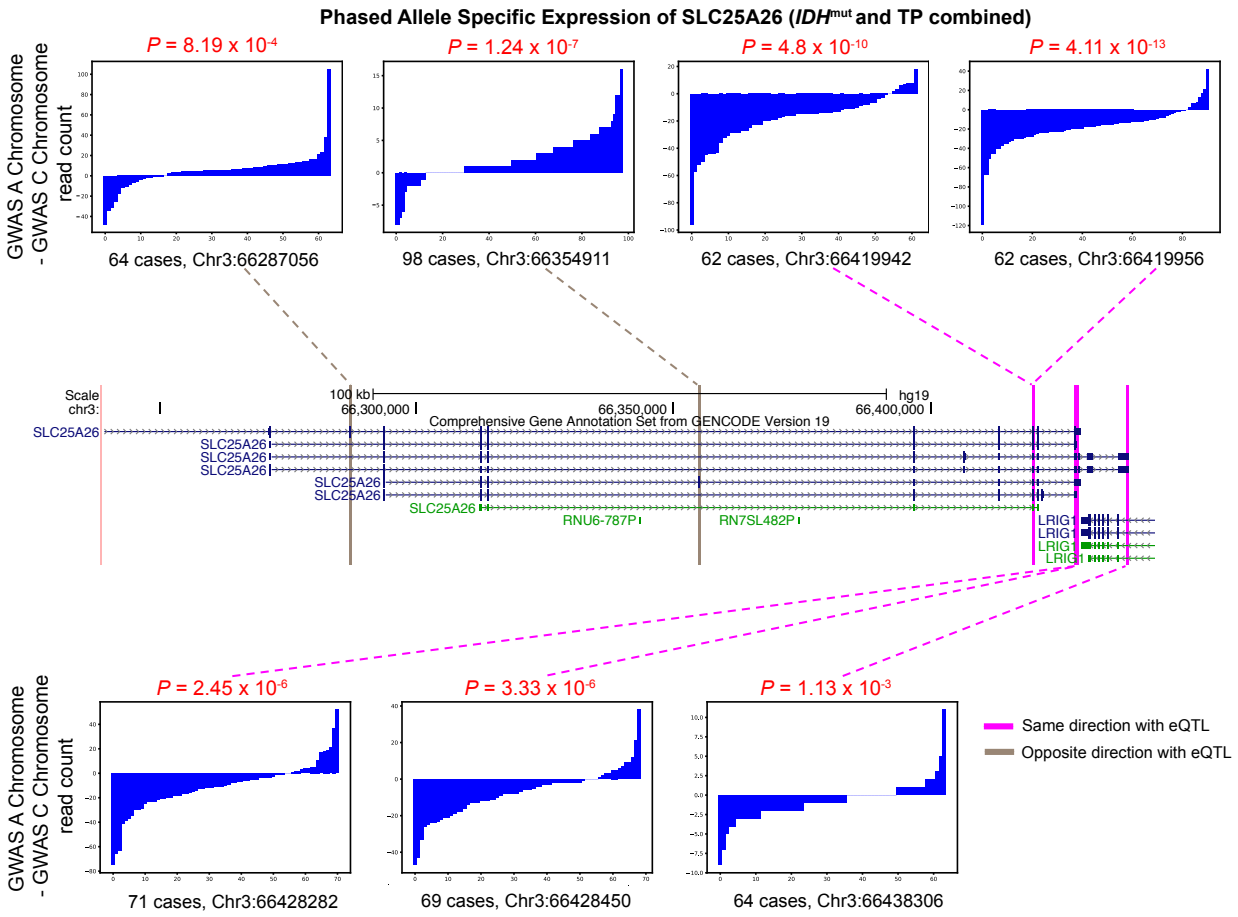

B



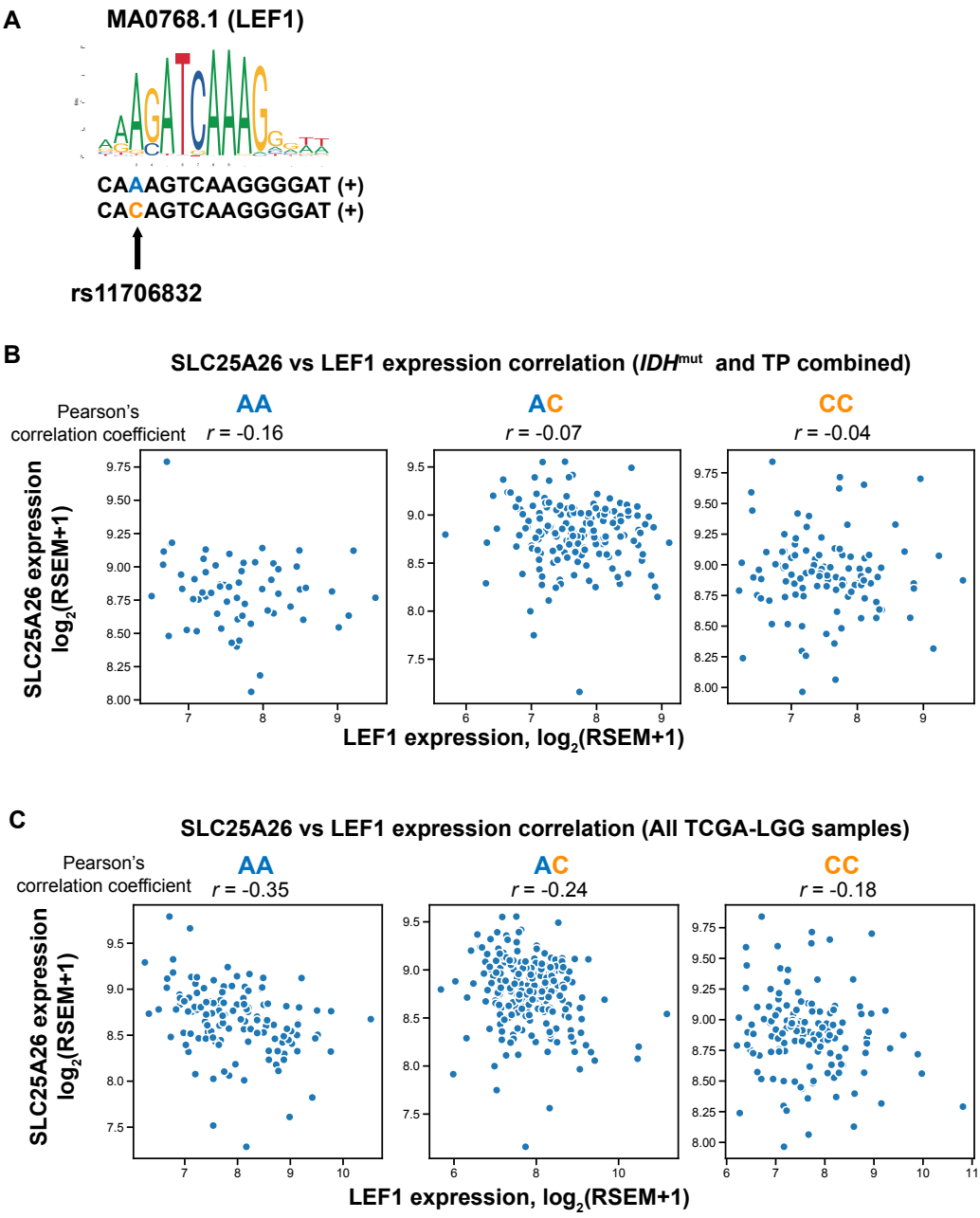

**Supplementary Fig. 7:** The receiver operating characteristic (ROC) curve assessing the performance of the CNN model and the CNN learned motif from simulated annealing interpretation. (A) Receiver operating characteristic (ROC) curve assessing the performance of the CNN model trained on six ENCODE SP1 ChIP-seq data sets (H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7). Test set was chr1 data in the A549 cell line. Area under the curve (AUC) = 0.95. (B) CNN-learned motif visualized through a motif logo obtained from WebLogo[25] 3. The core motif inside the red box resembles the core motif of SP1 MA0079.3 (Fig. 4B).



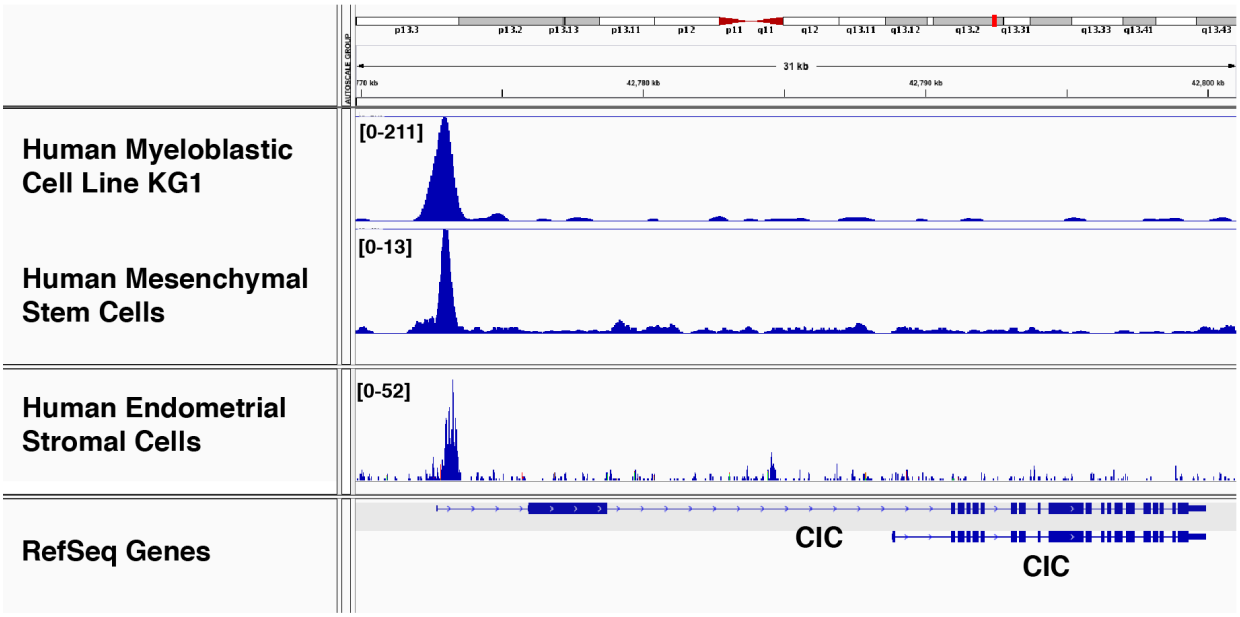

**Supplementary Fig. 8:** The GWAS SNP rs11706832 resides in a regulatory element within an *LRIG1* intron. A snapshot of the *LRIG1* locus where the GWAS SNP rs11706832 is denoted by a blue vertical line. The top three tracks are: basic gene annotation set from GENCODE version 19, LGG GWAS SNPs from GWAS catalog[30] and SNPs in high LD with rs11706832 (dbSNP[31] 151). The lower epigenomic tracks are: TCGA-LGG ATAC-seq, oligodendrocytes ATAC-seq[7], and REMC data in fetal brain and prefrontal cortex.




**Supplementary Fig. 9:** Phased allele-specific transcription pattern of *SLC25A26* at 7 exonic SNPs in the combined TCGA-LGG "*IDH*$^{mut}$ only" and triple-positive group. Five of these 7 exonic SNPs show a significant transcriptional skew toward the rs11706832-C allele (marked by magenta lines), in agreement with the eQTL result, while the other two show an opposite trend (marked by brown lines). For each exonic SNP, RNA-seq read count differences between the two chromosomes harboring the rs11706832-A allele and the rs11706832-C allele are sorted across patients and shown as bar plots. The *P*-value from the Wilcoxon signed-rank sum test is shown at the top of each bar plot. The genomic locations of these 7 exonic SNPs are shown in the GENCODE Version 19 track.




**Supplementary Fig. 10.** The GWAS SNP rs11706832 likely modulates *SLC25A26* expression through perturbing the binding affinity of LEF1. (A) LEF1 motif MA0768.1 (JASPAR[10]) and two versions of the flanking sequence harboring the rs11706832-A and rs11706832-C alleles. (B) Scatter plots of *LEF1* vs. *SLC25A26* expression in the three genotypes of rs11706832 in the combined TCGA-LGG "*IDH*$^{mut}$ only" and triple-positive group. (C) Scatter plots of *LEF1* vs. *SLC25A26* expression in the three genotypes of rs11706832 in all TCGA-LGG samples.




**Supplementary Fig. 11.** ZBTB16 binds the *CIC* promoter in multiple cell types. A snapshot at the *CIC* locus from the Integrative Genomics Viewer[32] showing ZBTB16 ChIP-seq data in human acute myelogenous leukemia cell line KG1[33], mesenchymal stem cells[34] and endometrial stromal cells[35].

## SUPPLEMENTARY REFERENCES


**1.** Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med.* 2016; 375(12):1109-1112.

**2.** Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016; 48(10):1284-1287.

**3.** Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet.* 2016; 48(11):1443-1448.

**4.** Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet.* 2016; 48(7):811-816.

**5.** Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol.* 2010; 28(10):1045-1048.

**6.** Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science.* 2018; 362(6413).

**7.** Nott A, Holtman IR, Coufal NG, et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science.* 2019; 366(6469):1134-1139.

**8.** Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011; 27(21):2987-2993.

**9.** Dewey M. metap: meta-analysis of significance values. *R package version 1.3.* 2020.

**10.** Fornes O, Castro-Mondragon JA, Khan A, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2020; 48(D1):D87-D92.

**11.** Kulakovskiy IV, Vorontsov IE, Yevshin IS, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 2018; 46(D1):D252-D259.

**12.** Jolma A, Yan J, Whitington T, et al. DNA-binding specificities of human transcription factors. *Cell.* 2013; 152(1-2):327-339.

**13.** Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform.* 2008; 9(4):326-332.

**14.** Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics.* 2011; 27(7):1017-1018.

**15.** Zhang Y, Manjunath M, Zhang S, Chasman D, Roy S, Song JS. Integrative Genomic Analysis Predicts Causative Cis-Regulatory Mechanisms of the Breast Cancer-Associated Genetic Variant rs4415084. *Cancer Res.* 2018; 78(7):1579-1591.

**16.** Fan Y, Xi L, Hughes DS, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* 2016; 17(1):178.

**17.** Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv.* 2019.

**18.** Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics.* 2012; 28(3):311-317.

**19.** Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012; 22(3):568-576.

**20.** Gleize V, Alentorn A, Connen de Kerillis L, et al. CIC inactivating mutations identify aggressive subset of 1p19q codeleted gliomas. *Ann Neurol.* 2015; 78(3):355-374.

**21.** Chollet Fao. Keras. *https://keras.io*. 2015.

**22.** Jia Y, Abadi M, Agarwal A, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems2015.

**23.** Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv.* 2014; 1412.6980(cs.LG).

**24.** Finnegan AI, Kim S, Jin H, et al. Epigenetic engineering of yeast reveals dynamic molecular adaptation to methylation stress and genetic modulators of specific DNMT3 family members. *Nucleic Acids Res.* 2020; 48(8):4081-4099.

**25.** Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res.* 2004; 14(6):1188-1190.

**26.** WashU. EpiGenome Gateway - WashU EpiGenome Browser. *https://github.com/epgg/eg*. 2019; Accessed 19 Mar 2019.

**27.** Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* 2019; 47(W1):W158-W165.

**28.** Cheneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 2018; 46(D1):D267-D275.

**29.** Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012; 9(3):215-216.

**30.** MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 2017; 45(D1):D896-D901.

**31.** Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001; 29(1):308-311.

**32.** Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011; 29(1):24-26.

**33.** Koubi M, Poplineau M, Vernerey J, et al. Regulation of the positive transcriptional effect of PLZF through a non-canonical EZH2 activity. *Nucleic Acids Res.* 2018; 46(7):3339-3350.

**34.** Agrawal Singh S, Lerdrup M, Gomes AR, et al. PLZF targets developmental enhancers for activation during osteogenic differentiation of human mesenchymal stem cells. *eLife.* 2019; 8.

**35.** Kommagani R, Szwarc MM, Vasquez YM, et al. The Promyelocytic Leukemia Zinc Finger Transcription Factor Is Critical for Human Endometrial Stromal Cell Decidualization. *PLoS genetics.* 2016; 12(4):e1005937.