Title

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

Journal

Neuro-Oncology, 23(4)

ISSN

1522-8517

Authors

Manjunath, Mohith
Yan, Jialu
Youn, Yeoan
et al.

Publication Date

2021-04-12

DOI

10.1093/neuonc/noaa248

Peer reviewed

Neuro-Oncology
23(4), 638–649, 2021 | doi:10.1093/neuonc/noaa248 | Advance Access date 1 November 2020

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

Mohith Manjunath,[†] Jialu Yan,[†] Yeoan Youn, Kristen L. Drucker, Thomas M. Kollmeyer, Andrew M. McKinney, Valter Zazubovich, Yi Zhang, Joseph F. Costello, Jeanette Eckel-Passow, Paul R. Selvin, Robert B. Jenkins, and Jun S. Song

Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA (M.M., J.Y., P.R.S., J.S.S.); Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA (M.M., J.Y., Y.Z., J.S.S.); Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA (Y.Y., P.R.S.); Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA (K.L.D., T.M.K., R.B.J.); Department of Neurological Surgery, University of California San Francisco, San Francisco, California, USA (A.M.M., J.F.C.); Department of Physics, Concordia University, Montreal, Québec, Canada (V.Z.); Department of Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA (Y.Z.); Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA (J.E-P)*

**Current affiliation: Department of Data Sciences, Dana-Farber Cancer Institute, Boston, Massachusetts, USA (Y.Z.)*

[†]Mohith Manjunath and Jialu Yan contributed equally to this work.

Corresponding Author: Prof. Jun S. Song, Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (songj@illinois.edu).

Abstract

Background. Large-scale genome-wide association studies (GWAS) have implicated thousands of germline genetic variants in modulating individuals' risk to various diseases, including cancer. At least 25 risk loci have been identified for low-grade gliomas (LGGs), but their molecular functions remain largely unknown.

Methods. We hypothesized that GWAS loci contain causal single nucleotide polymorphisms (SNPs) that reside in accessible open chromatin regions and modulate the expression of target genes by perturbing the binding affinity of transcription factors (TFs). We performed an integrative analysis of genomic and epigenomic data from The Cancer Genome Atlas and other public repositories to identify candidate causal SNPs within linkage disequilibrium blocks of LGG GWAS loci. We assessed their potential regulatory role via in silico TF binding sequence perturbations, convolutional neural network trained on TF binding data, and simulated annealing–based interpretation methods.

Results. We built an interactive website (http://education.knoweng.org/alg3/) summarizing the functional footprinting of 280 variants in 25 LGG GWAS regions, providing rich information for further computational and experimental scrutiny. We identified as case studies *PHLDB1* and *SLC25A26* as candidate target genes of rs12803321 and rs11706832, respectively, and predicted the GWAS variant rs648044 to be the causal SNP modulating *ZBTB16*, a known tumor suppressor in multiple cancers. We showed that rs648044 likely perturbed the binding affinity of the TF MAFF, as supported by RNA interference and in vitro MAFF binding experiments.

Conclusions. The identified candidate (causal SNP, target gene, TF) triplets and the accompanying resource will help accelerate our understanding of the molecular mechanisms underlying genetic risk factors for gliomas.

Key Points

1. Analysis of 25 low-grade glioma-associated genetic loci reveals candidate functional mechanisms.
2. The variant rs648044 likely modulates *ZBTB16* expression through perturbation of MAFF binding.
3. *PHLDB1* and *SLC25A26* are candidate target genes of rs12803321 and rs11706832, respectively.

Importance of the Study

Recent large-scale GWAS have implicated at least 25 genetic loci in modulating LGG susceptibility, but their molecular pathways remain elusive. To better understand the molecular functions of germline variants in modulating LGG risk, we developed an integrative framework utilizing genomic, epigenomic, and transcriptomic data to identify candidate (causal SNP, target gene, transcription factor) triplets. For the GWAS locus harboring the SNP rs648044, our framework revealed that this SNP likely modulates the expression of the target gene *ZBTB16* through perturbing the binding affinity of MAFF. We provide evidence that ZBTB16 directly regulates *CIC* (capicua transcriptional repressor), a tumor suppressor frequently mutated in isocitrate dehydrogenase-mutant oligodendrogliomas. We also developed an interactive web resource to summarize the functional annotation of 280 germline variants in 25 LGG GWAS regions. The results of our study will help accelerate the discovery of molecular mechanisms underlying genetic risk factors for gliomas and guide the design of new therapeutic preventions and interventions.

Gliomas are tumors originating in the glial cells of the brain. According to the 2016 World Health Organization (WHO) classification of tumors of the central nervous system, low-grade glioma (LGG) mainly includes diffuse astrocytic and oligodendroglial tumors.[1] The 2016 WHO classification further incorporated molecular features such as the mutations in either isocitrate dehydrogenase 1 (*IDH1*) or *IDH2* (collectively referred to as IDH^{mut}) and codeletion of the chromosome arms 1p and 19q (1p/19q codeletion). By including the status of telomerase reverse transcriptase (*TERT*) promoter mutations, gliomas can be further classified into 5 main molecular groups based on the presence or absence of the 3 molecular alterations.[2] The 5 molecular groups are:

1. "*TERT* promoter mutation only,"
2. "IDH^{mut} only,"
3. "*TERT* promoter and IDH^{mut},"
4. triple-positive (IDH^{mut}, *TERT* promoter mutant, 1p/19q codeleted), and
5. triple-negative (*IDH* wild-type, *TERT* wild-type, 1p/19q non-codeleted).

The triple-positive and "IDH^{mut} only" groups compose the majority of LGGs, while "*TERT* promoter mutation only" is prevalent in glioblastoma multiforme[2] (GBM). This study considers LGGs only, excluding GBM, with a focus on the triple-positive and "IDH^{mut} only" groups, which are usually oligodendrogliomas and astrocytomas, respectively, in terms of the 2016 WHO classification.

Genome-wide association studies (GWAS) have identified several single nucleotide polymorphisms (SNPs) associated with LGG susceptibility,[3–6] but only a few studies have hitherto discovered the corresponding genes directly regulated by these SNPs.[7,8] Most of the LGG GWAS SNPs reside in noncoding regions of the human genome, posing severe challenges to studying their molecular function and identifying susceptibility genes that may inform preventive and therapeutic measures. An integrative and systematic analysis of the LGG GWAS loci is thus needed to identify molecular mechanisms of tumorigenesis and help accelerate neuro-oncology research.

Our main hypothesis is that the GWAS loci contain causal SNPs that reside in functional regulatory regions of the human genome and modulate the expression of target genes by directly perturbing the binding affinity of transcription factors (TFs). In this study, we utilized large-scale heterogeneous datasets from The Cancer Genome Atlas (TCGA), Encyclopedia of DNA Elements[9] (ENCODE), and Roadmap Epigenomics Mapping Consortium[10] (REMC) databases for a comprehensive analysis of LGG germline GWAS variants. To provide easy access to all our findings, we integrated the results into an interactive web database, Analysis of Low-Grade Glioma GWAS (ALG^3), accessible at http://education.knoweng.org/alg3/.

Materials and Methods

LGG GWAS SNPs and SNPs in High Linkage Disequilibrium

We obtained a list of GWAS SNPs from Melin et al,[6] passing the combined meta-analysis (8 studies) *P*-value

cutoff of 5×10^{-8} for non-glioblastoma gliomas, yielding 25 GWAS SNPs significantly associated with LGG (Supplementary Table 1). Out of these 25, eight SNPs were also found to be significant ($P < 5 \times 10^{-8}$) in glioblastoma. The median odds ratio for the 25 GWAS SNPs was 1.2, where 23 of the 25 SNPs had odds ratio less than 1.5, typical of low-penetrance genetic variants.[6] We then used LDlink[11] to obtain all SNPs in high linkage disequilibrium (LD; $r^2 \geq 0.8$, 1000 Genomes Phase 3, EUR population) with the 25 GWAS SNPs and analyzed the functional footprinting of 280 SNPs in total. We obtained the glioma molecular subgroup information of the GWAS SNPs from Eckel-Passow et al.[12]

TCGA LGG Data

We utilized 5 types of TCGA LGG datasets[13]: germline genotype data of 513 patients, primary tumor copy number segmentation data of 513 patients, tumor RNA-seq aligned bam files of 516 patients, processed gene-level RSEM (RNA-Seq by Expectation Maximization) expression data of 516 patients, and clinical data of 515 patients. Out of 508 patients with all 5 data types, 427 patients' molecular subtype information was available.[14,15] Assigning these 427 patients to the 5 molecular subgroups yielded 204 patients in the "*IDH*^{mut} only" subgroup and 137 patients in the triple-positive subgroup.

Phased Allele-Specific Expression Analysis

From the expression quantitative trait loci (eQTL) analysis, genes with false discovery rate[16] (FDR) adjusted $p_i \leq 0.2$, where p_i is the *P*-value of the genotype linear regression coefficient, were selected as candidate target genes. For each candidate gene, we performed a phased allele-specific expression (ASE) analysis to test the differential transcription between the 2 chromosomes harboring different alleles of a given GWAS SNP.[17] We first obtained a subset of patients having heterozygous genotypes both at the GWAS SNP and at exonic SNPs of the candidate gene. We then extracted the imputed haplotype (Supplementary Methods) to determine the phase between the GWAS SNP and the exonic SNPs. Allele-specific coverage of the exonic SNPs by RNA-seq reads (*MAPQ* ≥ 20) was obtained, and Wilcoxon signed-rank sum test (for sample size $n \geq 5$) was used to examine the transcription imbalance between the 2 copies of chromosomes at a *P*-value threshold of 0.05.

Convolutional Neural Network and Simulated Annealing Methods

We trained a convolutional neural network (CNN) model on TF chromatin immunoprecipitation followed by sequencing (ChIP-seq) and DNase I hypersensitive sites sequencing (DNase-seq) data in available cell types from ENCODE and used the model to predict the allele-specific binding pattern of the same TF in the human fetal brain tissue (Supplementary Methods). We then used a Markov Chain Monte Carlo sampling[18,19] method to perform probabilistic optimization of the CNN-learned motif over the set of input sequences (Supplementary Methods).

Results

Integrative Analysis Identifies Candidate (Causal SNP, Target Gene, TF) Triplets

We developed an integrative analysis framework incorporating heterogeneous genomic, epigenomic, and transcriptomic datasets to understand the functional impact of GWAS variants (Figure 1). In the genomic context, we started with a list of 25 GWAS loci associated with increased risk for LGG (Methods); each locus contained a GWAS SNP showing the best association with LGG in the population, but the reported SNP might not necessarily be the functionally causal SNP, and nearby SNPs in high LD could act as true molecular effectors. We therefore examined all 280 SNPs in high LD with the GWAS SNPs ($r^2 \geq 0.8$, 1000 Genomes Phase 3, EUR population) (Methods). Genotypes for TCGA LGG cohort were imputed to obtain high-confidence genotypes for the high LD SNPs (Methods, Supplementary Methods). Epigenomic information contained histone modification and open chromatin signals from ChIP-seq, assay for transposase-accessible chromatin sequencing (ATAC-seq) and DNase-seq, as well as chromatin interactions from proximity ligation-assisted ChIP-seq (PLAC-seq) (Supplementary Methods). Using these datasets, we identified candidate causal SNPs residing within accessible regulatory DNA elements in the human brain and performed motif perturbation analyses to obtain TFs whose binding affinity might be modulated by the SNPs (Supplementary Methods). To further assess the impact of SNPs, we trained a CNN model on TF ChIP-seq data to predict allele-specific TF binding and deployed a simulated annealing method[18] to extract the optimal TF motif learned by the CNN (Methods, Supplementary Methods). In the transcriptomic context, we performed eQTL and phased ASE analyses using TCGA gene expression profiles to obtain a set of credible target genes (Methods, Supplementary Methods). We further filtered candidate TFs based on TF-target gene expression correlation analysis (Supplementary Methods). Our framework thus revealed candidate (causal SNP, target gene, TF) triplets, which could be prioritized for experimental validation. As case studies of detailed analysis and interpretation, we focused on 3 loci that had (1) one of the lowest GWAS *P*-values (*PHLDB1* locus), (2) a target gene with known tumor suppressor functions in other cancers (*ZBTB16* locus), and (3) no convincing eQTL candidate gene in a previous study[6] (*LRIG1* locus), respectively.

ZBTB16 Locus: 11q23.2 GWAS SNP rs648044

The lead SNP rs648044 modulates the expression of ZBTB16 through chromatin looping

The lead GWAS SNP rs648044 (Methods) contained no other SNP in high LD within its haplotype block (Methods) and was thus our candidate causal variant. As functional variants often interact with their target genes through active regulatory elements, we examined the epigenetic landscape surrounding the SNP in brain-related tissues and cell lines. Independent ATAC-seq[20,21] and DNase-seq

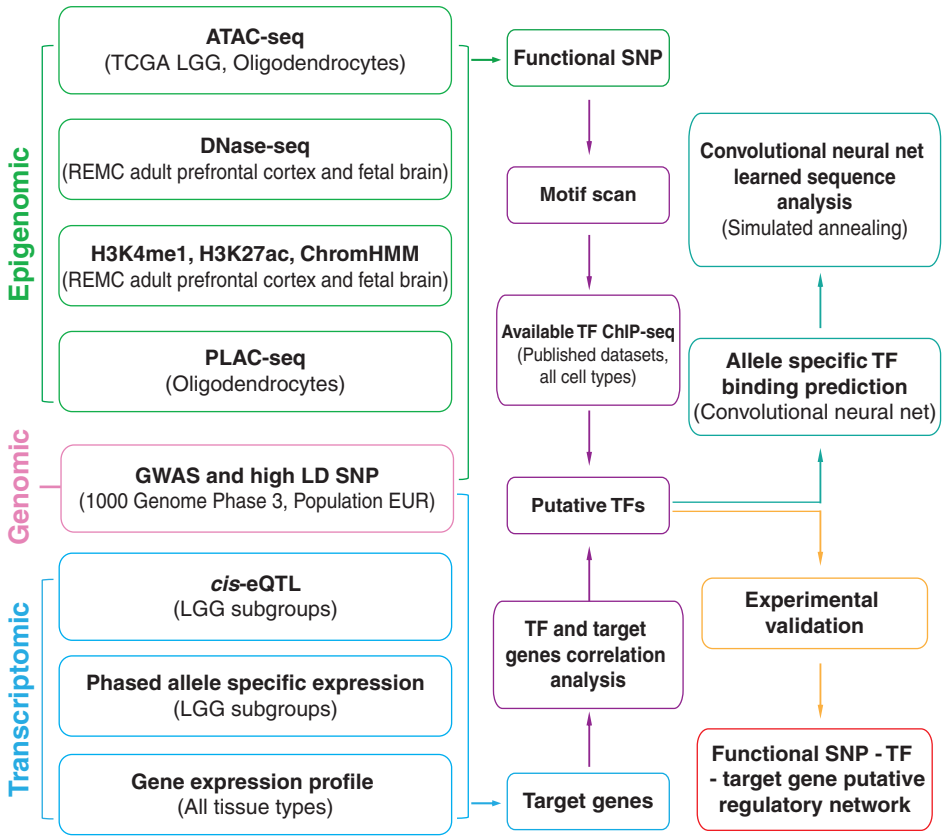

Fig. 1 Integrated framework for functional analysis of LGG GWAS SNPs. Green: epigenomic data; pink: genomic information; blue: transcriptomic data and analysis; purple: motif and TF-gene expression correlation analyses; ocean blue: deep learning approaches for TF binding prediction; yellow: experimental validation; red: candidate triplets.

datasets confirmed the SNP to be located within an open chromatin region in TCGA LGG samples, oligodendrocytes, and fetal brain tissue samples (Figure 2A) (Supplementary Methods). Histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac) ChIP-seq in fetal brain and dorsolateral prefrontal cortex tissues showed an active enhancer activity at the location (Figure 2A), as also annotated by REMC (Supplementary Figure 1A).

The eQTL analysis using the TCGA LGG genotype and gene expression data suggested *NCAM1* and *ZBTB16* to be the top candidate target genes (Supplementary Figure 1B, Figure 2B, *NCAM1* $P = 0.0054$ in the combined IDH^{mut} only and triple-positive group, Supplementary Methods). *NCAM1* is located ~1.1 Mb away from *ZBTB16*. H3K4me3 PLAC-seq confirmed a physical looping interaction only between the active *ZBTB16* promoter and the enhancer harboring rs648044 in oligodendrocytes[21] (Figure 3A; Supplementary Methods). We thus prioritized *ZBTB16* for further analysis. Correlation analysis between *ZBTB16* normalized expression values and genotype status at rs648044 in different molecular groups found a significant association in the combined group of "IDH^{mut} only" and triple-positive ($P = 0.0118$, FDR = 0.124; Supplementary Figure 1B). The expression level of *ZBTB16* was suppressed by the rs648044-A risk allele, indicating that *ZBTB16* might act as a tumor suppressor. Consistent with this hypothesis, *ZBTB16* encodes a zinc-finger TF[22] implicated in inhibiting proliferation, metastasis, or epithelial-mesenchymal transition in multiple cancers and is genetically lost in metastatic castration-resistant prostate cancer,[23] supporting its tumor suppressor role.[24–26]

rs648044 likely perturbs the binding affinity of MAFF

We next sought to identify the TF whose binding affinity might be perturbed by rs648044. We first utilized known TF binding motifs to perform in silico TF binding affinity perturbation analysis based on a sequence permutation test (Supplementary Methods). For each candidate TF, we then computed molecular group-wise Pearson correlation coefficient between the TF and *ZBTB16* expression levels stratified into 3 genotype groups of rs648044 (Supplementary Methods). Based on the eQTL finding that *ZBTB16* expression was lower in the risk group (AA genotype), we expected a candidate repressor TF to have higher binding affinity toward the risk allele A and show a

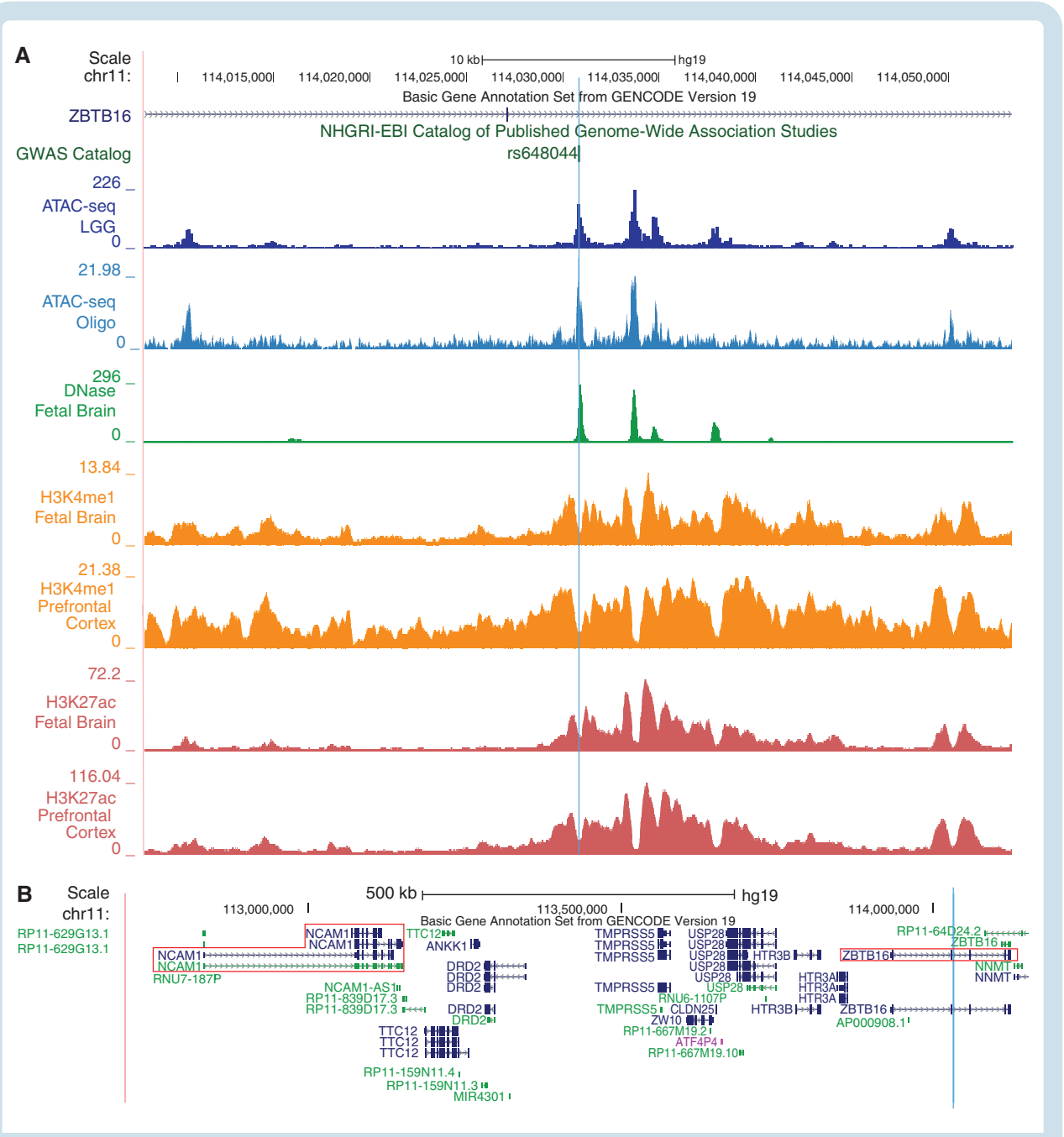

Fig. 2 *ZBTB16* intronic enhancer harboring rs648044 modulates mRNA expression of nearby genes. (A) A snapshot of the *ZBTB16* locus where the GWAS SNP rs648044 is denoted by a blue vertical line. The shown epigenomic tracks are: TCGA-LGG ATAC-seq,[20] oligodendrocytes ATAC-seq,[21] and REMC data in fetal brain and prefrontal cortex. (B) A zoomed-out view of the *ZBTB16* locus encompassing the eQTL target genes, *NCAM1* and *ZBTB16*, indicated by red boxes.

greater negative correlation with *ZBTB16* in the risk group compared with the GG genotype group; conversely, we expected a candidate activator TF to have lower binding affinity towards the risk allele and show a weaker positive correlation with *ZBTB16* in the risk group. ATAC-seq data in TCGA LGG samples showed a significant skew toward the rs648044-A risk allele, indicating that the TF might act as a repressor ($P = 0.010$, Fisher's method for combining binomial test P values; Supplementary Table 2; Supplementary Methods). These criteria together identified MAFF as the top candidate TF for further experimental validation. MAFF is a member of the small Maf basic leucine zipper TFs that can homodimerize and repress target genes. Its motif[27] clearly preferred the risk allele A (Figure 3B; permutation test $P = 0.0029$, Supplementary Methods), and the structure of expression correlation showed attenuation of the negative correlation between *MAFF* and *ZBTB16* in the AG and GG

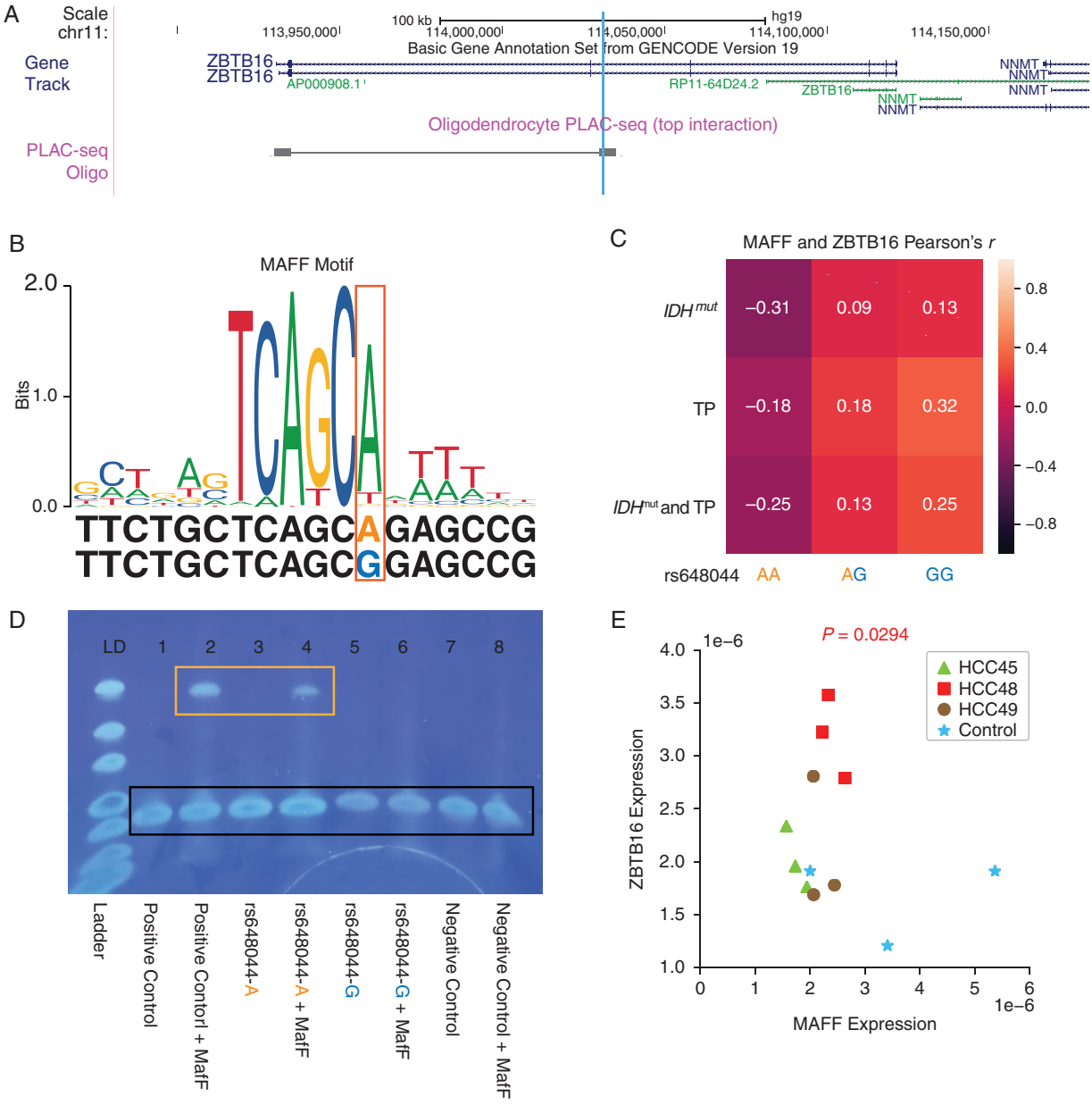

Fig. 3 The GWAS SNP rs648044 likely perturbs the binding affinity of MAFF that represses *ZBTB16*. (A) Oligodendrocyte PLAC-seq track[21] showing the rs648044 locus (blue vertical line) interacting with the *ZBTB16* promoter, about 100 kb away. (B) JASPAR[27] motif logo of the predicted TF MAFF and 2 variants of the flanking sequence harboring rs648044-A and rs648044-G alleles. Throughout the text, the risk and non-risk alleles of a SNP are colored orange and blue, respectively. (C) Pearson's correlation coefficient between *ZBTB16* and MAFF in the combined "*IDH*^{mut} only" and TP group, "*IDH*^{mut} only" subgroup and TP subgroup. (D) Gel picture from the EMSA experiment showing a ladder ("LD") and eight lanes using the mixture of the recombinant MafF protein and 4 different DNA sequences (Supplementary Methods): 81 bp positive control ("PC") sequence, 81 bp sequence flanking rs648044-A, 81 bp sequence flanking rs648044-G and negative control ("NC") sequence. The lower molecular weight bands in black box correspond to free DNA. Orange box highlights the bands of MafF-bound DNA, corresponding to the results of "positive control DNA + MafF" and "rs648044-A flanking sequence + MafF." (E) MAFF RNA interference knockdown experiment results, showing a significant increase in *ZBTB16* mRNA expression after MAFF knockdown. One-sided *t*-test *P*-value between the control group and the combined group of 3 independent short hairpin RNA clones is shown.

genotypes that were predicted to weaken the affinity of MAFF to DNA (Figure 3C).

To confirm that MAFF preferentially binds the rs648044-A allele, we performed an electrophoretic mobility shift assay (EMSA) (Figure 3D; Supplementary Methods). We detected binding of MAFF on positive control DNA (from a top consensus ChIP-seq peak region in HepG2, K562, and HelaS3, Supplementary Methods; lane 2) and the sequence containing the risk A allele (lane 4), but not on the sequence containing the alternative G allele (lane 6) and negative control DNA (a permuted sequence with no MAFF core binding motif, Supplementary Methods; lane 8). Knockdown of MAFF using short hairpin RNA in a cell line—derived from an *IDH1*^{R132H} mutant, *TERT* promoter-mutant, 1p/19q-codeleted (triple positive) oligodendroglioma patient and heterozygous at rs648044 (Supplementary Methods) – led to a significant increase in *ZBTB16* mRNA expression compared with non-target controls (Figure 3E; $P = 0.0294$, two-group one-sided *t*-test), but not in *NCAM1* mRNA expression (Supplementary Figure 2; $P = 0.37$, two-group one-sided *t*-test). These results support our prediction that MAFF preferentially binds the risk allele rs648044-A and represses the putative tumor suppressor *ZBTB16*. We further analyzed the prevalence of capicua transcriptional repressor (*CIC*) mutations in the context of rs648044 genotypes, as CIC is an important tumor suppressor frequently mutated in *IDH*^{mut} gliomas. *CIC* inactivating mutations tended to occur more frequently in the homozygous non-risk GG genotype than the combined AA and AG genotypes in TCGA triple-positive gliomas (odds ratio 2.0, Fisher's exact test $P = 0.076$; Supplementary Table 3), although statistical significance could not be reached, potentially due to small sample size. This finding suggested that the predicted suppression of *ZBTB16* by the risk rs648044-A allele could be an alternate mechanism for LGG tumorigenesis in *CIC* wild-type gliomas.

PHLDB1 Locus: 11q23.3 GWAS SNP rs12803321

eQTL and ASE analyses implicate PHLDB1 as a candidate target gene

We next applied our computational framework to the locus containing rs12803321 (reference allele: G (risk), alternative allele: C), one of the most significant LGG GWAS SNPs. The SNP rs12803321, located in the first intron of Pleckstrin Homology Like Domain Family B Member 1 (*PHLDB1*) (Supplementary Figure 3), was reported to be significantly associated with the "*IDH*^{mut} only" subgroup.[12,28] An eQTL analysis of 71 genes within 4 Mb of rs12803321 in *IDH*^{mut} only subgroup (Supplementary Methods) identified *PHLDB1* and Trehalase (*TREH*) as the top candidate target genes (*PHLDB1* $P = 2.5 \times 10^{-9}$, FDR $= 1.82 \times 10^{-7}$; *TREH* $P = 8 \times 10^{-5}$, FDR $= 2.84 \times 10^{-3}$; Figure 4A, Supplementary Figure 4A, B). The number of risk alleles was anticorrelated with the expression level of *PHLDB1* and *TREH* adjusted for covariates (Supplementary Methods). Since *TREH* expression was low (zero RSEM in 68 patients out of total 193), we prioritized *PHLDB1* for further analysis. We analyzed the allele-specific transcription pattern of *PHLDB1* using TCGA RNA-seq raw reads and the exonic SNPs' phased haplotype information (Methods). There were 20 exonic SNPs with more than 5 cases in the "*IDH*^{mut} only" group having a heterozygous genotype at both rs12803321 and the exonic SNP. Wilcoxon signed-rank sum test on the RNA-seq read counts from the 2 chromosomes[17] detected a statistically significant skew at 9 exonic SNPs out of 20 ($P < 0.05$). All these 9 SNPs showed higher transcription emanating from the rs12803321-C haplotype (Supplementary Figure 5). These results together demonstrated that the risk allele rs12803321-G was associated with decreased expression of *PHLDB1* in "*IDH*^{mut} only" group.

Candidate causal SNP rs12225399 perturbs the binding affinity of SP1/SP2

We next prioritized candidate functional SNPs using epigenomic data. There were 3 SNPs in high LD with rs12803321 (Methods): rs67307131 ($r^2 = 0.98$), rs12225399 ($r^2 = 0.97$) and rs7125115 ($r^2 = 0.90$). The GWAS SNP and all 3 high LD SNPs were located in open chromatin and active enhancer regions, as assessed by the fetal brain DNase-seq, TCGA LGG ATAC-seq,[20] oligodendrocyte ATAC-seq,[21] and prefrontal cortex histone modification (H3K4me1, H3K27ac) ChIP-seq data (Supplementary Figure 3). Motif analysis using FIMO[29] yielded candidate TFs whose binding affinity might be perturbed by any of the above four SNPs (Supplementary Table 4). Further filtering the TF list through TF-target gene expression correlation analysis (Supplementary Methods), we determined rs12225399 to be the best candidate causal SNP, and SP1/SP2 the top candidate TFs: first, rs12225399 was located near a local peak center in TCGA LGG and oligodendrocyte ATAC-seq (Supplementary Figure 3, Supplementary Figure 4C); second, sequence perturbation analyses demonstrated that the rs12225399-C allele, in phase with the rs12803321-C allele, created a high-scoring SP1/SP2 binding motif, whereas the rs12225399-G allele significantly perturbed the motif (FIMO SP1 $P = 4.25 \times 10^{-5}$, Figure 4B; FIMO SP2 $P = 5.53 \times 10^{-5}$, Supplementary Figure 6A; permutation test SP1 $P = 0.015$, SP2 $P = 0.0023$; Supplementary Methods); third, Pearson correlation coefficient between SP2 and *PHLDB1* in "*IDH*^{mut} only" group was highest in the rs12225399-CC genotype ($r = 0.40$) and decreased in rs12225399-GC ($r = 0.26$) and rs12225399-GG genotypes ($r = 0.23$) (Supplementary Figure 6B). The correlation between SP1 and *PHLDB1* did not show the same trend as SP2 and *PHLDB1* (Supplementary Figure 6C); however, since SP1 and SP2 recognize similar sequences (Figure 4B, Supplementary Figure 6A), we could not rule out SP1 as not being functional at the SNP. The high LD SNP rs7125115 was not selected as a candidate causal SNP, because our analysis did not yield a good candidate TF (Supplementary Table 4). These results together implied that the rs12225399-C allele likely increased the binding affinity of SP1/SP2, functioning as transcription activators to enhance the expression of *PHLDB1*.

Because of the lack of SP1/SP2 ChIP-seq data in brain cell types, we could not verify directly whether SP1/SP2 actually bound the predicted causal SNP. We thus applied a deep learning method to predict TF

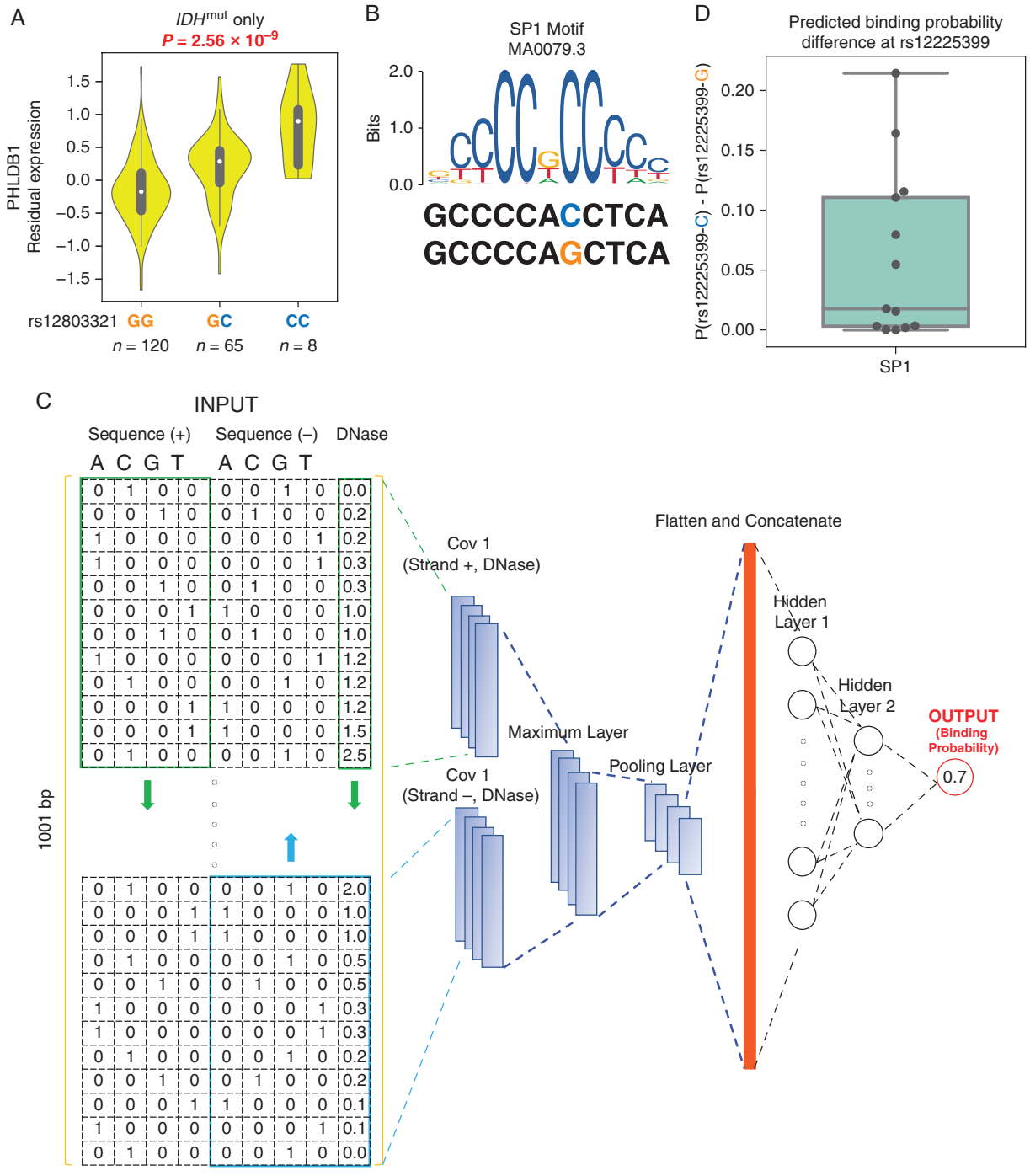

Fig. 4 The high LD SNP rs12225399 likely modulates *PHLDB1* expression by perturbing the binding affinity of SP1/SP2. (A) eQTL result for rs12803321 and *PHLDB1* in the TCGA-LGG "*IDH*^{mut} only" subgroup. (B) SP1 motif logo MA0079.3 (JASPAR[27]) and two variants of the flanking sequence harboring rs12225399-C and rs12225399-G alleles. (C) CNN for predicting the binding pattern of SP1 based on DNA sequence and open chromatin information. From left to right: 1001 bp × 9 input matrix incorporating sequence information and quantile-normalized DNase-seq signal at each base; convolutional layer using filters of length 12 bp; maximum layer, extracting the maximum of the convolutional layer output from the positive and negative strands; maximum pooling layer; flatten and concatenate layer; fully connected layer with 80 neurons; fully connected layer with 10 neurons; output. (D) The difference of SP1 binding probability between the two alleles of rs12225399, predicted by the CNN model based on 13 REMC fetal brain DNase-seq datasets from 10 donors.

binding affinity in fetal brain samples (Supplementary Methods). Although SP2 was a better candidate, SP2 ChIP-seq data were available in only one ENCODE cell line (Supplementary Methods), while SP1 ChIP-seq data were available in seven cell lines (H1-hESC, HEK293T, HepG2, Liver, K562, MCF-7, and A549; Supplementary Table 5). We thus trained a CNN for SP1 only, using sequence information and cell type-matched DNase-seq to predict the SP1 ChIP-seq signals (Figure 4C). A549 dataset was used as a test set, and the CNN was trained on the remaining six datasets (Supplementary Methods). The receiver operating characteristic area under the curve was 0.95 for the test set (Supplementary Figure 7A, Supplementary Methods). Moreover, we confirmed that the optimal CNN-learned motif, extracted via a simulated annealing method, closely resembled the known SP1 motif[27] (Figure 4B, Supplementary Figure 7B). The trained CNN was then used to evaluate the impact of rs12225399 on SP1 binding in the brain, taking the allele information and DNase-seq profiles in 13 REMC fetal brain samples as input. Our model predicted differential binding of SP1 at the two alleles of rs12225399, showing higher predicted probability of binding at the C allele than the G allele across all 13 REMC samples (Figure 4D).

LRIG1 Locus: 3p14.1 GWAS SNP rs11706832

Functional analysis of rs11706832 locus identifies the (rs11706832, SLC25A26, LEF1) triplet

The LGG GWAS SNP rs11706832 (reference allele: A, alternative allele: C (risk)), located in an intron of Leucine rich repeats and immunoglobulin like domains 1 (*LRIG1*) (Supplementary Figure 8), was reported to be associated with "*IDH*^{mut} only" and triple-positive glioma subgroups.[12] Although highly expressed in the brain, *LRIG1* did not show a significant eQTL association with rs11706832 in TCGA LGG data ($P = 0.52$ and 0.34 for "*IDH*^{mut} only" and triple-positive, respectively), in agreement with a previous report.[6] By contrast, we found that Solute carrier family 25 member 26 (*SLC25A26*), a gene 432 kb away from *LRIG1*, was significantly associated with rs11706832 in eQTL and phased ASE analyses: the number of rs11706832 risk allele C was positively correlated with the expression level of *SLC25A26* (Figure 5A-C; genotype $P = 2.9 \times 10^{-3}$, "*IDH*^{mut} only"; 4.11×10^{-2}, triple-positive; 2.11×10^{-4}, "*IDH*^{mut} only" and triple-positive combined; FDR = 1.48×10^{-3}, "*IDH*^{mut} only" and triple-positive combined). Phased ASE analysis identified seven exonic SNPs with a Wilcoxon signed-rank sum test $P < 0.05$ ("*IDH*^{mut} only" and triple-positive combined group, case number > 5). 5 of these 7 exonic SNPs showed a significant transcriptional skew toward the rs11706832:C allele (Supplementary Figure 9), in agreement with the eQTL result, while the other two showed an opposite trend. These results suggested that a functional consequence of the GWAS risk allele rs11706832:C was to increase the expression of *SLC25A26*.

Of all 3 SNPs in high LD with rs11706832 (Methods), rs4402869 ($r^2 = 0.87$) and the GWAS SNP rs11706832 resided in open chromatin and active enhancer regions (Supplementary Figure 8). Motif analysis and gene-TF expression correlation analysis for rs11706832 and rs4402869 identified rs11706832-LEF1 to be the best candidate SNP-TF pair (Supplementary Table 6), with the rs11706832-A allele potentially creating a LEF1 binding motif (FIMO $P = 9.4 \times 10^{-4}$, Supplementary Figure 10A) and the A-to-C conversion significantly perturbing the binding motif (permutation test $P = 0.012$). The correlation structure between *LEF1* and *SLC25A26* expression was inconclusive in the combined "*IDH*^{mut} only" and triple-positive group, but the anticorrelation was clearly strongest in the AA genotype when all LGG samples were used (Supplementary Figure 10B, C). These results together suggested that the

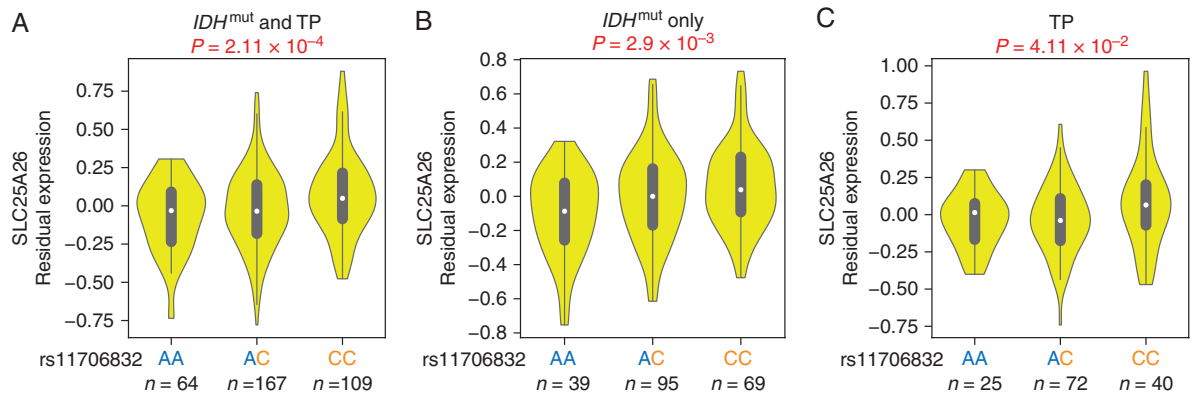

Fig. 5 The GWAS SNP rs11706832 is associated with an increased expression of *SLC25A26*. (A) eQTL result for rs11706832 and *SLC25A26* in the combined TCGA-LGG "*IDH*^{mut} only" and TP group. (B) Similar to (A), but for the "*IDH*^{mut} only" subgroup. (C) Similar to (A), but for the TP subgroup.

rs11706832-A allele might create a binding site of LEF1, a known transcriptional repressor,[30] thereby suppressing the expression of *SLC25A26*.

Analysis of Low-Grade Glioma GWAS (ALG³): An Interactive Web Resource

We have developed the web portal ALG³ (http://education.knoweng.org/alg3/) to provide an interactive visual summary of the functional footprinting of LGG GWAS loci, facilitating additional analysis or experimental validation. ALG³ includes an embedded genome browser,[31] copy number information, eQTL results, relevant ENCODE ChIP-seq information, motif analysis and expression correlation analysis (Supplementary Methods). The processed ATAC-seq and PLAC-seq data in oligodendrocytes[21] are also linked to the University of California Santa Cruz genome browser.

Discussion

We have shown that the 11q23.2 GWAS SNP rs648044 may modulate the expression of *ZBTB16* by perturbing the binding affinity of MAFF. Although ENCODE ChIP-seq data show a MAFF peak (q-value = 3.1×10^{-4}) covering the SNP rs648044 in K562 cells, as well as a similar MAFK peak (q-value = 1.6×10^{-4}) in HepG2 cells, further studies are needed to confirm the allele-specific binding of MAFF at rs648044 in glioma cells, as predicted by our computational analysis and in vitro data. ZBTB16 has been shown to regulate self-renewal and differentiation of hematopoietic stem cells, mainly acting as a transcriptional activator and antagonized by a noncanonical function of the histone methyltransferase EZH2[32]. It also acts as a tumor suppressor in prostate cancer, melanoma, gallbladder cancer, and leukemia.[25,26,33,34] Although no ZBTB16 ChIP-seq data are currently available in oligodendrocytes, ChIP-seq data in human mesenchymal stem cells,[35] endometrial stromal cells,[36] and acute myelogenous leukemia cells[32] show ZBTB16 binding the *CIC* promoter in these cell types (Supplementary Figure 11). The mRNA expression level of ZBTB16 is also highly correlated with that of CIC in prefrontal cortex (Spearman's $\rho = 0.65$, GTEx v8), supporting that *CIC* is likely a direct transcriptional target of ZBTB16. Importantly, *CIC* is one of the most commonly mutated genes in IDH^{mut} oligodendrogliomas and located on chromosome 19q, which is often codeleted with chromosome 1p in oligodendrogliomas. These observations thus suggest a potentially important interaction network involving the regulation of *CIC* by ZBTB16 and disruption of this interaction by rs648044 in the tumorigenesis of LGG. The fact that *CIC* mutation shows a trend of being more frequent in the homozygous non-risk GG genotype of rs648044, where the expression level of ZBTB16 is elevated, is consistent with this potential interaction between the two tumor suppressors. However, the sample size of patients in our study may be too small to understand the genetic interactions accurately; furthermore, some patients having the non-risk GG genotype of rs648044 may have mutations in other genes or harbor other risk SNPs, leading to alternate mechanisms of LGG pathogenesis.[6,14]

We have proposed *PHLDB1* to be a candidate target gene repressed in the risk genotype of rs12803321. Our identified causal SNP rs12225399 also appears as one of top candidate causal SNPs in a previous study implicating *PHLDB1* for a different GWAS SNP.[8] Knockdown of *PHLDB1* has been shown to increase cell death and reduce neurosphere formation in the U87MG glioma cell line,[8] but its molecular function remains poorly understood. We have developed a deep learning approach for predicting the binding pattern of TFs when their ChIP-seq data are not available in the human brain. Most previous machine learning approaches have been using only sequence information for predicting protein binding patterns,[37] and some recent studies have begun to utilize other genomic and epigenomic information.[38] Our deep learning model integrates DNase-seq signal with sequence information into one convolutional filter. Using the CNN trained on non-brain cell data to evaluate sequence and open chromatin information in brain tissues has allowed us to predict allelic preference of SP1 binding. A similar approach may benefit future functional genomics studies in the brain, where TF ChIP-seq data are not readily available.

At the rs11706832 locus, we have shown *SLC25A26* expression to be elevated in the risk group. This gene belongs to the mitochondrial carrier family and encodes a protein involved in transporting S-adenosylmethionine into the mitochondria.[39] It has been shown that overexpression of *SLC25A26* in CaSki cells contributes to mitochondrial DNA (mtDNA) hypermethylation[40] and that mtDNA methylation level tends to decrease during glioblastoma progression.[41] Future studies may reveal how potential mtDNA methylation changes attributable to *SLC25A26* modulation by rs11706832 contribute to LGG tumorigenesis.

Analysis of pan-cancer TCGA ATAC-seq data shows that rs648044, rs12225399 and rs11706832 also reside in open chromatin regions of several cancer types. We cannot thus conclude at this point that the proposed regulatory functions of these SNPs are specific to LGG; however, the effects of their regulatory functions on modulating cancer risk seem specific to LGG, as the GWAS SNPs were associated only with non-glioblastoma gliomas.[42] Even though our eQTL analysis modeled copy number alterations in TCGA LGG expression data, it is possible that other uncharacterized somatic mutations that could alter transcription levels or mRNA stability might have strongly perturbed the mRNA abundance in tumor samples and complicated the target gene identification. This study has focused on assessing the molecular function of genetic variants in altering the binding affinity of TFs. Other molecular functions might include DNA methylation changes and protein modifications, although the effects of differential methylation themselves could be mediated through differential binding of TFs or other chromatin binding proteins.[43] To facilitate the rapid identification of candidate (causal SNP, target gene, TF) triplets, we have summarized our results into an interactive user-friendly web database, ALG³.

Supplementary Material

Supplementary data are available online at *Neuro-Oncology* (http://neuro-oncology.oxfordjournals.org/).

Keywords

functional genomics | genetic variants | GWAS | low-grade glioma

Funding

This project was supported by an Oligodendroglioma Research Award from the National Brain Tumor Society (R.B.J. and J.S.S.), a generous gift from the Dabbiere family (J.F.C., R.B.J., J.S.S.), National Institutes of Health (NIH) R01CA163336 (J.S.S.), NIH R01NS100019 (P.R.S.), NIH R01CA230712 (R.B.J.), and NCI fellowship F31CA243187 (A.M.M.).

Acknowledgments

This study utilizes the data generated by the TCGA Research Network (http://cancergenome.nih.gov/) and the GTEx Project (dbGaP accession number phs000424.v8.p2 on 09/05/2019) supported by the Common Fund of the Office of the Director of the NIH and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Conflict of interest statement. The authors declare no conflicts of interest.

Authorship statement. Analysis design: RBJ, JSS. Experimental design: JFC, PRS, RBJ, JSS. Implementation: MM, JY, YY, KLD, TMK, AMM, VZ, YZ. Analysis and interpretation of the data: all authors. All authors were involved in the writing of the manuscript and have read and approved the final version.

References

1. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 2016;131(6):803–820.
2. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *N Engl J Med.* 2015;372(26):2499–2508.
3. Shete S, Hosking FJ, Robertson LB, et al. Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet.* 2009;41(8):899–904.
4. Sanson M, Hosking FJ, Shete S, et al. Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Hum Mol Genet.* 2011;20(14):2897–2904.
5. Kinnersley B, Labussière M, Holroyd A, et al. Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat Commun.* 2015;6:8559.
6. Melin BS, Barnholtz-Sloan JS, Wrensch MR, et al; GliomaScan Consortium. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet.* 2017;49(5):789–794.
7. Atkins I, Kinnersley B, Ostrom QT, et al. Transcriptome-wide association study identifies new candidate susceptibility genes for glioma. *Cancer Res.* 2019;79(8):2065–2071.
8. Baskin R, Woods NT, Mendoza-Fandiño G, Forsyth P, Egan KM, Monteiro AN. Functional analysis of the 11q23.3 glioma susceptibility locus implicates PHLDB1 and DDX6 in glioma susceptibility. *Sci Rep.* 2015;5:17367.
9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
10. Kundaje A, Meuleman W, Ernst J, et al; Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317–330.
11. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics.* 2015;31(21):3555–3557.
12. Eckel-Passow JE, Decker PA, Kosel ML, et al. Using germline variants to estimate glioma and subtype risks. *Neuro Oncol.* 2019;21(4):451–461.
13. 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.
14. Brat DJ, Verhaak RG, Aldapeet KD, et al; Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med.* 2015;372(26):2481–2498.
15. Ceccarelli M, Barthel FP, Malta TM, et al; TCGA Research Network. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell.* 2016;164(3):550–563.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995;57(1):289–300.
17. 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.
18. 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.
19. Finnegan A, Song JS. Maximum entropy methods for extracting the learned features of deep neural networks. *PLoS Comput Biol.* 2017;13(10):e1005836.
20. Corces MR, Granja JM, Shams S, et al. The chromatin accessibility landscape of primary human cancers. *Science.* 2018;362(6413):eaav1898.
21. 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.
22. Lin DY, Huang CC, Hsieh YT, et al. Analysis of the interaction between zinc finger protein 179 (Znf179) and promyelocytic leukemia zinc finger (Plzf). *J Biomed Sci.* 2013;20:98.
23. Hsieh CL, Botta G, Gao S, et al. PLZF, a tumor suppressor genetically lost in metastatic castration-resistant prostate cancer, is a mediator of resistance to androgen deprivation therapy. *Cancer Res.* 2015;75(10):1944–1948.
24. Wang JB, Jin Y, Wu P, et al. Tumor suppressor PLZF regulated by lncRNA ANRIL suppresses proliferation and epithelial mesenchymal transformation of gastric cancer cells. *Oncol Rep.* 2019;41(2):1007–1018.

25. Shen H, Zhan M, Zhang Y, et al. PLZF inhibits proliferation and metastasis of gallbladder cancer by regulating IFIT2. *Cell Death Dis.* 2018;9(2):71.
26. Jin Y, Nenseth HZ, Saatcioglu F. Role of PLZF as a tumor suppressor in prostate cancer. *Oncotarget.* 2017;8(41):71317–71324.
27. 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.
28. Labreche K, Kinnersley B, Berzero G, et al. Diffuse gliomas classified by 1p/19q co-deletion, TERT promoter and IDH mutation status are associated with specific genetic risk loci. *Acta Neuropathol.* 2018;135(5):743–755.
29. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics.* 2011;27(7):1017–1018.
30. Nawshad A, Medici D, Liu CC, Hay ED. TGFbeta3 inhibits E-cadherin gene expression in palate medial-edge epithelial cells through a Smad2-Smad4-LEF1 transcription complex. *J Cell Sci.* 2007;120(Pt 9):1646–1653.
31. WashU. EpiGenome gateway—WashU EpiGenome browser. 2019. https://github.com/epgg/eg. Accessed March 19, 2019.
32. 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.
33. Felicetti F, Bottero L, Felli N, et al. Role of PLZF in melanoma progression. *Oncogene.* 2004;23(26):4567–4576.
34. Hobbs RM, Pandolfi PP. Shape-shifting and tumor suppression by PLZF. *Oncotarget.* 2010;1(1):3–5.
35. 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:e40364.
36. Kommagani R, Szwarc MM, Vasquez YM, et al. The promyelocytic leukemia zinc finger transcription factor is critical for human endometrial stromal cell decidualization. *PLoS Genet.* 2016;12(4):e1005937.
37. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol.* 2015;33(8):831–838.
38. Schmidt F, Gasparoni N, Gasparoni G, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* 2017;45(1):54–66.
39. Agrimi G, Di Noia MA, Marobbio CM, Fiermonte G, Lasorsa FM, Palmieri F. Identification of the human mitochondrial S-adenosylmethionine transporter: bacterial expression, reconstitution, functional characterization and tissue distribution. *Biochem J.* 2004;379(Pt 1):183–190.
40. Menga A, Palmieri EM, Cianciulli A, et al. SLC25A26 overexpression impairs cell function via mtDNA hypermethylation and rewiring of methyl metabolism. *FEBS J.* 2017;284(6):967–984.
41. Sun X, Vaghjiani V, Jayasekara WSN, Cain JE, St John JC. The degree of mitochondrial DNA methylation in tumor models of glioblastoma and osteosarcoma. *Clin Epigenetics.* 2018;10(1):157.
42. 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.
43. Bonder MJ, Luijk R, Zhernakova DV, et al; BIOS Consortium. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet.* 2017;49(1):131–138.