Neuro-Oncology
19(7), 940–950, 2017 | doi:10.1093/neuonc/now288 | Advance Access date 22 February 2017

Genotype-based gene signature of glioma risk

Yen-Tsung Huang, Yi Zhang, Zhijin Wu, Dominique S. Michaud

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan (Y-T.H); Department of Epidemiology (Y-T.H, D.S.M.); Department of Biostatistics, Brown University, Providence, Rhode Island (Y-T.H, Y.Z., Z.W.); Department of Public Health and Community Medicine, Tufts University, Boston, Massachusetts (D.S.M.)

Corresponding Author: Yen-Tsung Huang, MD, ScD, Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan (ythuang@stat.sinica.edu.tw).

Abstract

Background. Glioma accounts for 80% of malignant brain tumors, but its etiologic determinants remain elusive. Despite genetic susceptibility loci identified by genome-wide association study (GWAS), the agnostic approach leaves open the possibility that other susceptibility genes remain to be discovered. Here we conduct a gene-centric integrative GWAS (iGWAS) of glioma risk that combines transcriptomics and genetics.

Methods. We synthesized a brain transcriptomics dataset ($n = 354$), a GWAS dataset ($n = 4203$), and an advanced glioma tumor transcriptomic dataset ($n = 483$) to conduct an iGWAS. Using the expression quantitative trait loci (eQTL) dataset, we built models to predict gene expression for the GWAS data, based on eQTL genotypes. With the predicted gene expression, iGWAS analyses were performed using a novel statistical method. Gene signature risk score was constructed using a penalized logistic regression model.

Results. A total of 30 527 transcripts were analyzed using the iGWAS approach. Four novel glioma susceptibility genes were identified with internal and external validation, including *DRD5* ($P = 3.0 \times 10^{-79}$), *WDR1* ($P = 8.4 \times 10^{-77}$), *NOMO1* ($P = 1.3 \times 10^{-25}$), and *PDXDC1* ($P = 8.3 \times 10^{-24}$). The genotype-predicted transcription pattern between cases and controls is consistent with that between tumor and its matched normal tissue. The genotype-based 4-gene signature improved the classification between glioma cases and controls based on age, gender, and population stratification, with area under the receiver operating characteristic curve increasing from 0.77 to 0.85 ($P = 8.1 \times 10^{-23}$).

Conclusion. A new genotype-based gene signature of glioma was identified using a novel iGWAS approach, which integrates multiplatform genomic data as well as different genetic association studies.

Key words

expression quantitative trait loci | gene-signature | genome-wide association study | glioma | integrative genomics

Gliomas account for 32% of all brain tumors and 80% of all malignant brain tumors.[1,2] Mortality is high in advanced gliomas (60% of gliomas), and patients with glioblastoma multiforme (GBM) have survival rates of 4.7% at 5 years.[2] Other than ionizing radiation, there are no established environmental risk factors for gliomas. Various environmental factors have been implicated in the etiology of gliomas, but most findings have been inconsistent, which suggests the importance of genetics and genomics in the susceptibility of this devastating disease. Several rare genetic syndromes, such as neurofibromatosis type I, have been associated with the glioma risk,[3] and a positive family history is associated with a 2-fold elevated risk of glioma,[4,5] supporting the role for a genetic component to glioma. A number of genetic regions have been identified in genome-wide association studies (GWAS).[6–11] While recent findings from GWAS revealed important regions that are associated with the glioma risk, the agnostic approach inherent to GWAS and the strict adjustment for multiple comparisons leave open the possibility that other susceptibility regions remain to be discovered. Given the lack of well-established causes for glioma, understanding genetic susceptibility will provide new insights and opportunities for progress in unraveling the biological mechanisms behind this fatal cancer.

In GWAS, a large number of single nucleotide polymorphism (SNP) markers are tested across the genome. As multiple comparison adjustments are needed in GWAS, there has been a

Importance of the study

We present a genomic study that utilized a novel approach to discover new susceptibility genes of glioma and to construct a gene signature for this devastating disease. With external reference data of transcriptomics and genetics in brain tissue, we are able to predict the cerebral tissue-specific transcriptomic profile for the subjects based on their genotypes. We applied this approach to synthesize a multiplatform genomic study where we jointly analyzed a GWAS dataset, a brain transcriptomics dataset, and a tumor genomic dataset. Our integrative approach identified 4 susceptibility genes of glioma: *DRD5*, *WDR1*, *NOMO1*, and *PDXDC1,* which were validated in multiple data and were used to construct a 4-gene signature for the glioma risk.

substantial interest in improving the statistical power of testing SNP effects by borrowing additional biological information. A major criticism of GWAS lies in its agnostic style[12]: no biological knowledge is encoded in the standard GWAS analyses. To address such limitations, SNP-set analyses have been advocated to integrate biological information into statistical analyses and to decrease the number of tests.[13,14] Analyses using SNP sets grouped by physical locations have shown a better performance than the standard single SNP analyses in re-analyzing the breast cancer GWAS dataset.[14] SNPs can also be grouped into a set according to biological functions and have been utilized in studying skin cancer,[15] bladder cancer,[16] and lung cancer.[17] By decreasing the number of tests and incorporating biological knowledge in the analysis, the SNP-set approach has provided a biologically relevant alternative to pursue genetic association analyses. However, this approach has not yet been applied to glioma.

Expression quantitative trait loci (eQTL) are the SNPs that are associated with gene expression. Several studies have incorporated eQTL data into GWAS using different approaches. Some studies used eQTL to prefilter or prioritize the SNPs: a GWAS of basal cell carcinoma that focused on eQTL SNPs was reported[15]; an osteoporosis GWAS used eQTL to re-prioritize the ranking from the result of genome-wide scans.[18] Other studies focused on the overlap between GWAS and eQTL analyses: an asthma GWAS showed that the susceptibility loci at 17q21 were also eQTL for the *ORMDL3* gene.[19] Still others found that eQTL were enriched in the trait-associated SNPs; eQTL SNPs were also found enriched among common susceptibility loci of type 2 diabetes,[20] bipolar disorder,[21] lymphocyte count,[22] and the published GWAS SNPs in the online database collected by the National Human Genome Research Institute.[23] These studies have shown the promise of integrating eQTL into GWAS analyses.

Currently the existing approaches aim at the overlap of SNP–disease (GWAS) and SNP–expression (eQTL) associations using different strategies, but whether the association of SNPs with expression really contributes to the disease risk has not been directly examined in previous studies. Statistical methods have been developed to jointly analyze SNP and gene expression data provided that both data types are collected from the same individuals.[24–26] However, how to analytically integrate the genetic and transcriptomic data when the GWAS and the eQTL study are conducted in different subjects remains a challenge. Here we introduce a new analytic framework that utilizes a novel statistical methodology[27] to jointly analyze SNP and gene expression data by integrating a GWAS with an eQTL study.

Methods

Study Population

Three population datasets were included in this analysis: a brain eQTL dataset, a glioma GWAS dataset, and a GBM dataset from The Cancer Genome Atlas (TCGA). The brain eQTL study was conducted at the National Institute on Aging and consisted of genomic data obtained from fresh frozen tissue samples from the frontal lobe of cerebral cortex in 354 neurologically normal Caucasian subjects. Genome-wide genotyping was performed using Illumina Infinium HumanHap 550K, 610Q, or 660W BeadChips, and mRNA expression profiling was measured using Illumina HumanRef-8 Expression Beadchips. Additional details on this dataset can be found in Gibbs et al.[28]

The glioma GWAS dataset was archived in the Database of Genotypes and Phenotypes (study accession phs000652.v1.p1). The genome-wide genotype data were collected on 556 glioma cases and 3647 controls using Illumina 550K, 610Q, or 660W BeadChip. We randomly divided the GWAS data into a Discovery Set consisting of 370 cases and 2430 controls and Validation Set consisting of 186 cases and 1217 controls. Missing genotypes were imputed using IMPUTE2 software (see Supplement Section I).

To validate the differentially expressed genes estimated for cases and controls in the GWAS data, we collected the array-based transcriptomic data from 473 GBM tumors or stage IV astrocytomas and 10 organ-specific normal control tissues from individuals without cancer who donated tissue for other reasons, both archived in TCGA.[29] The transcriptomic data were preprocessed level 3 data measured with University of North Carolina AgilentG4502A_07 array. The study is based on de-identified data that have been made publicly available, and thus does not involve human subjects. The demographics of the 3 datasets are summarized in Supplementary Table 1.

eQTL Analyses and Estimation of Gene Expression for GWAS Data

Three hundred fifty-four subjects of the eQTL dataset were analyzed to identify eQTL. We first defined *cis*-eQTL of a gene as SNPs locating within 0.5 Mb of the gene. We then constructed univariate eQTL models using linear regression models for each transcript expression value, regressing log2-transformed expression level on an SNP genotype under additive mode (0, 1, 2 as the number of the minor allele), adjusting for age and

gender. For the *cis*-eQTL SNPs with *P*-value smaller than .05, we considered them as potential eQTL and built a multivariate eQTL model based on these eQTL SNPs. The distribution for the number of potential eQTL for 30527 transcripts is shown in Supplementary Figure 1 (median = 9).

We assumed a multiple linear regression model that log2-transformed expression level of each transcript G was determined by covariates $\boldsymbol{X}$ (1 [for the intercept], age, and gender), and SNPs $\boldsymbol{S} = (S_1,\ldots,S_p)^T$ from its corresponding eQTL set with size of p SNPs:

$$G_i = \boldsymbol{X}_i^T\alpha_X + \boldsymbol{S}_i^T\alpha_S + \epsilon_{Gi} \text{ and } \epsilon_G \sim N(0,\sigma_G^2). \quad (1)$$

Due to the large number of eQTL SNPs, we employed a ridge estimator to estimate $\hat{\alpha}_X$ and $\hat{\alpha}_S$ that minimized $\sum_i (G_i - \boldsymbol{X}_i^T\alpha_X - \boldsymbol{S}_i^T\alpha_S) + \lambda \| \alpha_S \|^2$; the tuning parameter λ was chosen from 10-fold cross-validations. We then predicted the unmeasured expression levels of probes for all 4349 samples in the GWAS study using the estimated model coefficients of SNPs, $\mu_{Gi} = \boldsymbol{S}_i^T\hat{\alpha}_S$. We obtained the predicted expression values of 30527 unique probes.

Integrative Genome-wide Association Study

We have developed a novel statistical method to integrate an eQTL study into a GWAS where the 2 studies were conducted in different study subjects.[27] We built eQTL models in (1) to estimate the expression value for each subject in the GWAS data and then combined the *cis*-eQTL SNPs mapped to the gene (or transcript) and its estimated expression level to assess their joint effect on the glioma risk by the following efficient testing procedure.

We first assumed a disease risk model, one transcript at a time: for subject i ($i = 1,\ldots,n$; n is the sample size of GWAS data), the glioma outcome Y_i ($Y_i = 1$ and 0 for case and control, respectively) is determined by p eQTL SNPs $\boldsymbol{S}_i = (S_{1i},\ldots,S_{pi})$ identified from the above, one mRNA expression of a transcript/gene G_i, their possible cross-product interactions as well as covariates $\boldsymbol{X}_i$: 1 for the intercept, age, gender and 4 principal components for population stratification, through a logistic model:

$$\text{logit } P(Y_i = 1 | \boldsymbol{S}_i, G_i, \boldsymbol{X}_i) = \boldsymbol{X}_i^T\beta_X + \boldsymbol{S}_i^T\beta_S + G_i\beta_G + G_i\boldsymbol{S}_i^T\beta_{SG}.$$

We then obtain the marginal model for the glioma risk under the rare disease assumption that only depends on the eQTL SNPs $\boldsymbol{S}_i$ and the covariates $\boldsymbol{X}_i$ by taking an integral with respect to gene expression G:

$$\text{logit } P(Y_i = 1 | \boldsymbol{S}_i, \boldsymbol{X}_i) \approx \log \int e^{\boldsymbol{X}_i^T\beta_X + \boldsymbol{S}_i^T\beta_S + (\beta_G + \boldsymbol{S}_i^T\beta_{SG})G_i} dF_G(g).$$

The null hypothesis of no genetic effect is specified as:

$$H_0 : \Delta \equiv \text{logit } P(Y = 1 | \boldsymbol{X}, \boldsymbol{S} = \boldsymbol{s}_1) - \text{logit } P(Y = 1 | \boldsymbol{X}, \boldsymbol{S} = \boldsymbol{s}_0) = 0.$$

It has been shown that $H_0 : \Delta = 0 \leftrightarrow H_0 : \beta = (\beta_S^T, \beta_G, \beta_{SG}^T)^T = 0$, provided $\boldsymbol{S}$ are eQTL SNPs.[27] Because of the low power in conventional multivariate tests, we further assumed β_S, β_G, and β_{SG} followed arbitrary working distributions with mean zeros and variances τ_S, τ_G, and τ_{SG}, respectively. Thus the null hypothesis $H_0 : \Delta = 0$ became equivalent to:

$$H_0 : \tau_S = \tau_G = \tau_{SG} = 0. \quad (2)$$

We constructed the test statistic Q as a weighted sum of noncentered scores U_{τ_S}, U_{τ_G}, and $U_{\tau_{SG}}$ for τ_S, τ_G, and τ_{SG} under the null (2):

$$Q = w_1U_{\tau_S} + w_2U_{\tau_G} + w_3U_{\tau_{SG}} = (\boldsymbol{Y} - \mu_0)^T \boldsymbol{K} (\boldsymbol{Y} - \mu_0),$$

where $\mu_0 = (\mu_{01},\ldots,\mu_{0n})^T$, $\boldsymbol{K} = w_1\boldsymbol{S}\boldsymbol{S}^T + w_2\mu_G\mu_G^T + w_3\boldsymbol{C}_{SG}\boldsymbol{C}_{SG}^T$, $\boldsymbol{S} = (\boldsymbol{S}_1,\ldots,\boldsymbol{S}_n)^T$, $\mu_G = (\mu_{G1},\ldots,\mu_{Gn})^T$, $\boldsymbol{C}_{SG} = (\boldsymbol{C}_{SG1},\ldots,\boldsymbol{C}_{SGn})^T$, and $\boldsymbol{C}_{SGi} = \mu_{Gi}\boldsymbol{S}_i$. We chose the weights w_1, w_2, and w_3 to be the square root of the information for U_{τ_S}, U_{τ_G}, and $U_{\tau_{SG}}$, respectively. μ_{0i}, the glioma risk for subject i under the null (2), was estimated by $e^{\hat{\beta}_X^T\boldsymbol{x}_i}\left(1 + e^{\hat{\beta}_X^T\boldsymbol{x}_i}\right)^{-1}$, where $\hat{\beta}_X$ was estimated from the logistic model under the null. We calculated *P*-value for the test statistic Q by comparing with its underlying distribution, a mixture of chi-square distributions.[27]

We consider the following 3 hypotheses: (1) the SNPs-only model, $H_0 : \tau_S = 0$; (2) the model with SNPs and gene expression effects but no SNPs-by-expression interactions, $H_0 : \tau_S = \tau_G = 0$; and (3) the model with SNPs and gene expression effects as well as their interactions, $H_0 : \tau_S = \tau_G = \tau_{SG} = 0$. Note the models (1) and (2) are parsimonious models nested within (3). In order to synthesize the information from the 3 candidate models, we conducted an omnibus test. Specifically, we calculated *P*-values for the 3 models p_S, p_{SG}, and p_{SGC} and used the smallest *P*-value $p_{min} = \min(p_s, p_{SG}, p_{SGC})$ as the test statistic of the omnibus test (see Supplement Section II for further discussion). We were able to obtain the realization of the approximated distribution for the minimum *P*-value using a perturbation procedure, $\{p_{min}^{(b)}\} = \{p_{min}^{(1)},\ldots,p_{min}^{(B)}\}$, where B is the number of perturbation.[27] By comparing p_{min} with $\{p_{min}^{(b)}\}$, we calculated the tail probability as the omnibus *P*-value. As the proposed method protects type I error for each transcript,[27] standard methods of adjusting multiple comparisons on *P*-values can be used (ie, Bonferroni's correction, false discovery rate, etc). We used Bonferroni's correction in the analysis of glioma risk. Sensitivity analyses by adjusting for population stratification in eQTL analyses are present in Supplement Section III (Supplementary Tables 4 and 5; Supplementary Figures 2 and 3).

Gene Signature

The transcripts identified in the iGWAS analysis were further combined to construct a risk prediction model that used transcript signature to predict glioma risk with the 2-stage discovery-validation process. We fit the following logistic regression model using the Discovery Set:

$$\text{logit } P(Y_i = 1 | X_i, \mu_{G1i},\ldots\mu_{Gmi}) = \boldsymbol{X}_i^T\gamma_X + \gamma_1\mu_{G1i} + \ldots + \gamma_m\mu_{Gmi},$$

where m is the number of genes or transcripts. Due to the large degrees of freedom and the correlation among

$\mu_{G1}, \ldots, \mu_{Gm}$, we introduced an $L2$ penalty into the maximum likelihood estimation to stabilize the estimator of $\hat{\gamma}_{train} = \left(\hat{\gamma}_1, \ldots, \hat{\gamma}_m\right)^T$ and $\hat{\gamma}_{X,train}$. The tuning parameter for the penalty was chosen to minimize mean squared error from 10-fold cross-validation. The risk score of the gene signature was derived as $\psi_i = \mu_{GSi}^T \hat{\gamma}_{train}$, where $\mu_{GSi} = \left(\mu_{G1i}, \ldots, \mu_{Gmi}\right)^T$. We compared the classification performance between the model with covariates and the m-transcript/gene risk score ($\boldsymbol{X}_i^T \hat{\gamma}_{X,train} + \psi_i$) and that with only covariates ($\boldsymbol{X}_i^T \hat{\gamma}_{X,train}$), using receiver operating characteristic (ROC) curve. To avoid overfitting in ROC, we calculated ψ_i using the expression values μ_{GSi} and the covariates $\boldsymbol{X}_i$ in the Validation Set and $\hat{\gamma}_{train}$ and $\hat{\gamma}_{X,train}$ estimated from the Discovery Set (Figure 4A and C). We also swapped the two sets, ie, using the Validation Set to estimate $\hat{\gamma}_{train}$ and $\hat{\gamma}_{X,train}$ and calculate ψ_i with the expression values μ_{GSi} and the covariates $\boldsymbol{X}_i$ in the Discovery Set (Figure 4B and D). Note that the risk score of the gene signature is the weighted sum of the genotype:

$$\psi_i(\text{genotype based risk score}) = \sum_{j=1}^{m} \mu_{Gji} \hat{\gamma}_j = \sum_{j=1}^{m} \left(\boldsymbol{S}_{ji}^T \hat{\alpha}_{Sj}\right) \hat{\gamma}_j,$$

where S_{ji} is the genotype of eQTL for gene (or transcript) j and $\hat{\alpha}_{Sj}$ is the association of the eQTL SNPs with the gene (or transcript) expression.

Results

iGWAS

The analysis strategy is depicted in Figure 1. The glioma GWAS dataset was randomly divided into Discovery and Validation Sets. We conducted iGWAS to analyze 30 527 transcripts using the Discovery Set. The distribution of the 30 527 omnibus *P*-values is very close to a uniform distribution with a spike at the low end (Supplementary Figure 4). The quantile-quantile plot is shown in Supplementary Figure 5 with a genomic inflation factor of 1.061. Note that the 30 527 transcripts may not be independent due to the fact that eQTL could be mapped to multiple transcripts and that multiple transcripts could be derived from the same gene. iGWAS omnibus *P*-values of the 30 527 transcripts in the Discovery Set were presented in a Manhattan plot (Figure 2). Note that each dot represents a statistical significance level of a transcript where we jointly analyzed its expression value and the genotypes of its eQTL. The black dots are the 55 transcripts with $P < .001$ in both Discovery and Validation Sets.

Among the 30 527 transcripts, 371 (1.30%), 397 (1.39%), 400 (1.40%), and 467 (1.63%) transcripts were significant at discovery $P < .01$ in tests for the SNP-only model, tests for SNP and gene expression main effect model, tests for interaction model, and omnibus tests, respectively; 98 (0.34%), 96 (0.34%), 101 (0.35%) and 107 (0.37%) transcripts were significant at discovery $P < .001$. It suggests that better statistical power was achieved by incorporating genotype-estimated gene expression, which is consistent with the numerical studies.[27]

The 55 validated transcripts are summarized in Table 1 and Supplementary Table 2. Fifty-four of them had discovery *P*-values lower than the Bonferroni-adjusted genome-wide significance level, and all 55 combined *P*-values reached the genome-wide significance. Results of the SNP-only model, the SNP and gene expression main effect model, and the interaction model are summarized in Supplementary Table 3. The estimated gene expression provided advantage in detecting statistical significance in transcripts of *OR7E85P*, *SLC2A9*, *DRD5*, *WDR1*, *VNN3*, *SNORD100*, *SNORA33*, *NOMO1*, and *PDXD1*, but not the 40 transcripts of the ubiquitin specific peptidase 17-like family, *NTAN1*, *STMN3*, and *LIME1*, which had higher statistical significance in the SNP-only model. The most significant genes included *DRD5* (dopamine receptor D5; $P = 3.0 \times 10^{-79}$), *WDR1* (WD repeat domain 1; $P = 8.4 \times 10^{-77}$), *SLC2A9* (solute carrier family 2 [facilitated glucose transporter] member 9; $P = 1.4 \times 10^{-27}$), *NOMO1* (NODAL modulator 1; $P = 1.3 \times 10^{-25}$), and *PDXDC1* (pyridoxal-dependent decarboxylase domain containing 1; $P = 8.3 \times 10^{-24}$).

In addition to the iGWAS approach, we investigated candidate regions and previous GWAS loci using univariate genetic association analyses. For the 20 previously reported glioma (or GBM) susceptibility loci that can be mapped in our data, we confirmed 10 SNPs: rs2736100 (5p15.33; *TERT*), rs2853676 (5p15.33; *TERT*), rs2252586 (7p11.2; *EGFR*), rs4977756 (9p21.3; *CDKN2BAS*), rs1412829 (9p21.3; *CDKN2BAS*), rs1063192 (9p21.3; *CDKN2A/B*), rs2157719 (9p21.3; *CDKN2A/B*), rs3851634 (12q23.3; *POLR3B*), rs2297440 (20q13.33; *RTEL1*), rs6010620 (20q13.33; *RTEL1*, *TNFRSF6B*) (Supplementary Table 6). Based on our eQTL analyses, rs2297440 and rs6010620 were associated with the expression of *STMN3* and *LIME3*.

As shown in the Manhattan plot, there are 4 hot spots of glioma susceptibility, including 4p16.1, 6q23.2, 16p33.11, and 20q13.33 (Figure 2). In 4p16.1, rs6824806 had highly significant genetic association with glioma risk, and the SNP also had very strong association with transcription levels of *DRD5*, *WDR1*, *SLC2A9*, *OR7E85P*, and ubiquitin specific pepetidase 17-like family. In 6q23.2, rs4458717 and r12200377 had significant association with *SNORD100*, *SNORA33*, and *VNN3* expressions as well as the glioma risk. In 16p33.11, rs11075260 was associated with *PDXDC1*, *NOMO1*, *NTAN1* expressions, and the glioma risk. In 20q13.33, rs2297440, rs6010620, rs6089953 had strong association with the glioma risk and moderate association with *STMN3* and *LIME3* transcription level. We note that 20q13.33 is also the region that harbors significant SNPs (rs2297440 and rs6010620) in previous GWAS.[6,9]

Genotype-based Gene Signature

Hierarchical clustering was performed based on Euclidean distance among the estimated expression values of the 55 validated transcripts (Figure 3). Glioma cases are enriched in the red cluster, compared with the other clusters (20.1% vs 8.1%, $P = 1.2 \times 10^{-29}$). With the 55 transcripts, we developed a gene signature with use of $L2$ penalized logistic regression. Compared with the model with only covariates: age, gender, and 4 principal components of population stratification, the model with the additional 55 transcripts had better discrimination between glioma cases and controls with area under the ROC curve (AUC) increasing from

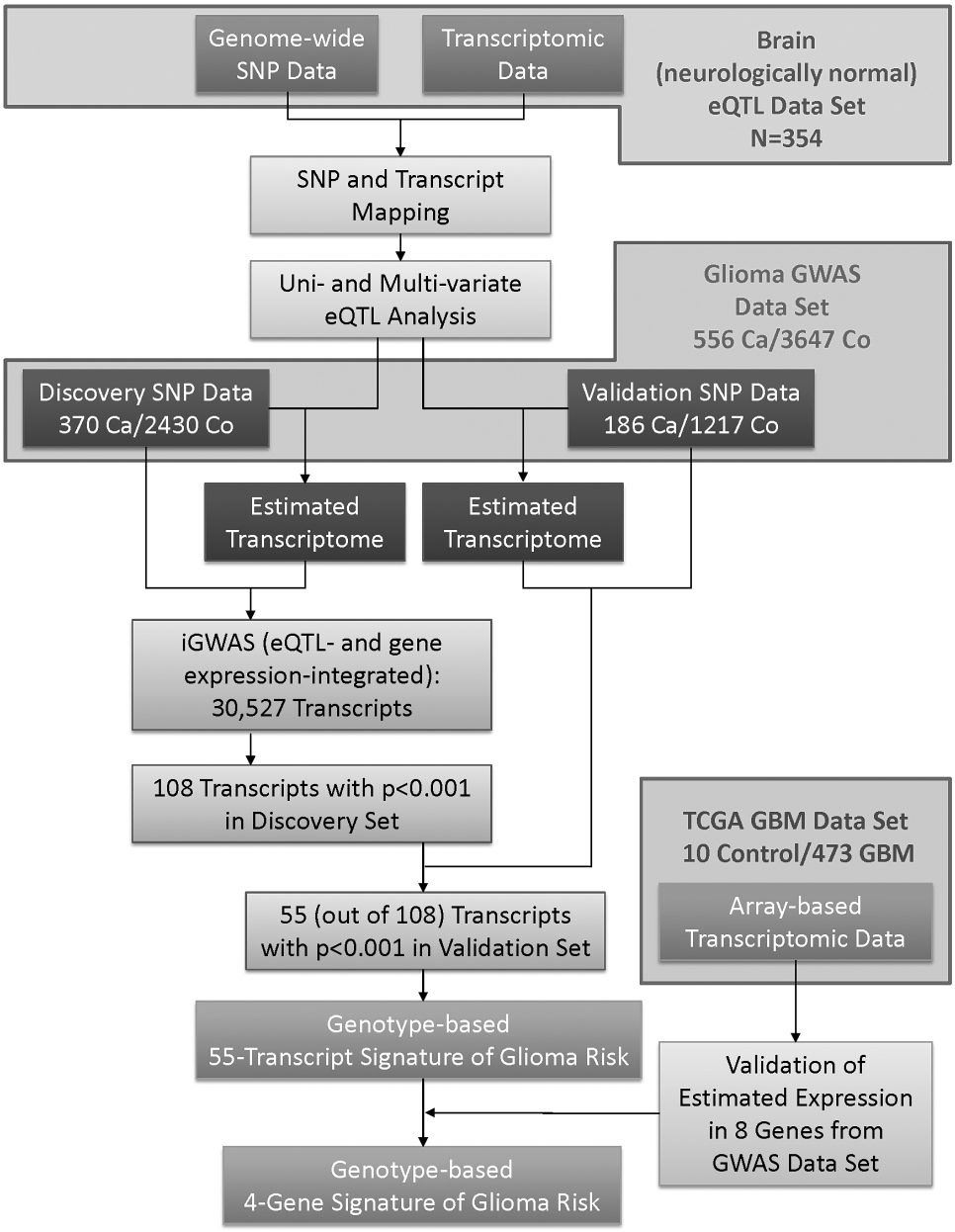

Fig. 1 The schematic of design and analysis of the integrative genome-wide association study (iGWAS).

0.76 to 0.83 ($P = 1.3 \times 10^{-5}$) in the Validation Set with models built by the Discovery Set (Figure 4A), from 0.77 to 0.88 ($P = 1.1 \times 10^{-35}$) in the Discovery Set with models built by the Validation Set (Figure 4B), and from 0.77 to 0.88 ($P = 1.0 \times 10^{-35}$) with model building and testing using both sets combined. ROC curves for a single transcript of all 55 transcripts are shown in Supplementary Figure 6.

Among the 55 transcripts, 8 (*DRD5*, *WDR1*, *SLC2A9*, *VNN3*, *NOMO1*, *PDXDC1*, *STMN3*, and *LIME1*) can be identified in TCGA gene expression data of GBM. We investigated the estimated gene expression by comparing differential expression patterns between cases and controls with those inTCGA GBM tumors versus organ-specific control tissues. Four genes (*DRD5*, *WDR1*, *NOMO1*, and *PDXDC1*) were validated: expression of *DRD5*, *NOMO1*, and *PDXDC1* was significantly lower and that of *WDR1* was significantly higher in GBM tumor tissue than normal brain tissue, which is consistent with findings based on the estimated gene expression for the glioma GWAS data (Figure 5). With the 4 genes further validated by TCGA data, we developed a parsimonious 4-gene signature. Its performance is similar to the 55-transcript signature, with AUC increasing from 0.76 (the model with only covariates) to 0.82 (the model with covariates and the 4 genes) ($P = 2.3 \times 10^{-4}$) in the Validation Set (Figure 4C), from 0.77 to 0.86 ($P =$

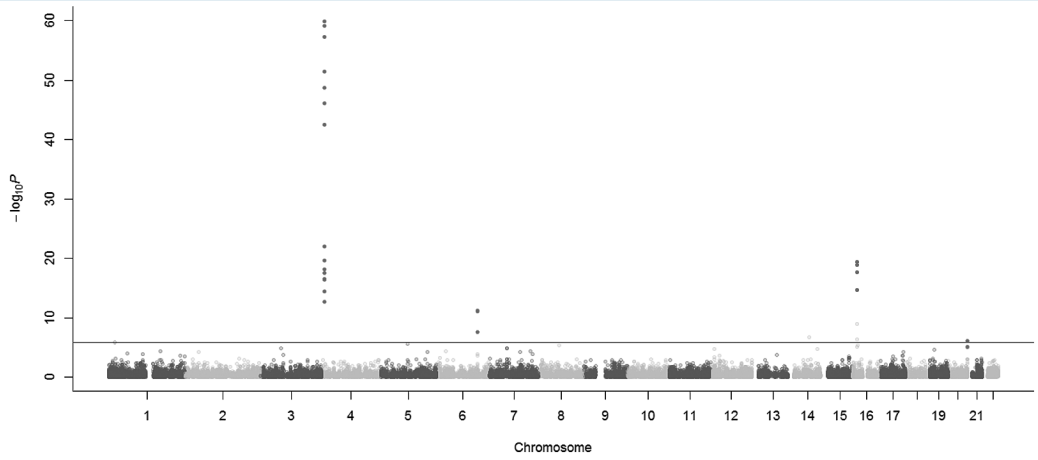

Fig. 2 Manhattan plot of eQTL- and gene expression–integrated genome-wide association studies of 30 527 transcripts with glioma risk. The black dots are $-\log_{10}$p of the 55 transcripts confirmed by the Validation Set. The horizontal broken line indicates the Bonferroni genome-wide significance level.

Table 1 The 16 out of 55 transcripts discovered at the significance level of omnibus $P < .001$ and validated at the level of omnibus $P < .001$

Illumina ID	Ensembl Gene ID	HGNC Symbol	Chromosome	Start Position	End Position	No. of SNPs	Omnibus *P*-value		
							Discovery	Validation	Combined
ILMN_1684499	ENSG00000251694	*USP17L9P*	4	9360109	9361701	27	7.17E-47	3.27E-14	1.13E-59
ILMN_2079225	ENSG00000250884	*OR7E85P*	4	9485365	9486349	61	2.98E-52	3.15E-16	7.17E-63
ILMN_1738406	ENSG00000109667	*SLC2A9*	4	9772777	10056560	186	2.20E-20	1.78E-05	1.40E-27
ILMN_1723803	ENSG00000109667	*SLC2A9*	4	9772777	10056560	186	6.78E-60	7.36E-18	3.09E-71
ILMN_1689043	ENSG00000169676	*DRD5*	4	9783258	9785632	126	5.59E-58	8.42E-16	2.95E-79
ILMN_1780036	ENSG00000071127	*WDR1*	4	10075963	10118573	188	2.89E-43	2.02E-13	8.27E-58
ILMN_1675844	ENSG00000071127	*WDR1*	4	10075963	10118573	188	1.27E-60	5.83E-19	8.39E-77
ILMN_1804935	ENSG00000093134	*VNN3*	6	133043926	133055904	226	2.90E-08	4.60E-05	1.89E-12
ILMN_2096747	ENSG00000221500	*SNORD100*	6	133137941	133138016	231	5.58E-12	0.000821	9.84E-16
ILMN_2096747	ENSG00000200534	*SNORA33*	6	133138358	133138487	231	8.75E-12	0.000967	1.92E-15
ILMN_2126957	ENSG00000103512	*NOMO1*	16	14927538	14990017	43	4.15E-20	9.43E-08	1.28E-25
ILMN_1702114	ENSG00000103512	*NOMO1*	16	14927538	14990017	43	2.34E-18	1.16E-05	1.67E-22
ILMN_1703969	ENSG00000179889	*PDXDC1*	16	15068448	15233196	46	1.35E-19	1.97E-06	8.32E-24
ILMN_1815552	ENSG00000157045	*NTAN1*	16	15131710	15149921	35	2.37E-15	0.000821	6.78E-17
ILMN_1728645	ENSG00000197457	*STMN3*	20	62271061	62284780	136	8.16E-07	0.000322	5.42E-11
ILMN_2344079	ENSG00000203896	*LIME1*	20	62366815	62370456	145	7.72E-06	0.000682	1.14E-09

HGNC = Human Genome Organisation (HUGO) Gene Nomenclature Committee. The omnibus *P*-value characterizes the statistical significance incorporating SNP-only models, SNP and gene expression main effect only models, and interaction models. Only 1 of the 40 ubiquitin specific peptidase 17-like family members (*USP17L9P*) is presented, and the remaining 39 transcripts are presented in Supplementary Table 1.

1.4×10^{-27}) in the Discovery Set (Figure 4C), and from 0.77 to 0.85 ($P = 8.1 \times 10^{-23}$) in both sets combined. The other 4 genes that were not validated by TCGA data were *SLC2A9*, *VNN3*, *STMN3*, and *LIME1*. Interestingly, the signature constructed by these 4 genes had only modest increase in AUC (from 0.76 to 0.81 in the Discovery Set; from 0.76 to 0.79 in the Validation Set) (Supplementary Figure 7). The genotype-estimated expressions of the 55 transcripts between glioma cases and controls are shown in Supplementary Figure 8.

Discussion

Here we present a new approach to conduct a genome-wide association study, which identifies novel susceptibility genes of glioma risk: *DRD5*, *NOMO1*, *PDXDC1*, and *WDR1*. Gene-centric GWAS analysis decreases the number of tests and has become popular. Genotype-based methods such as the SNP-set Sequence Kernel Association Test (SKAT)

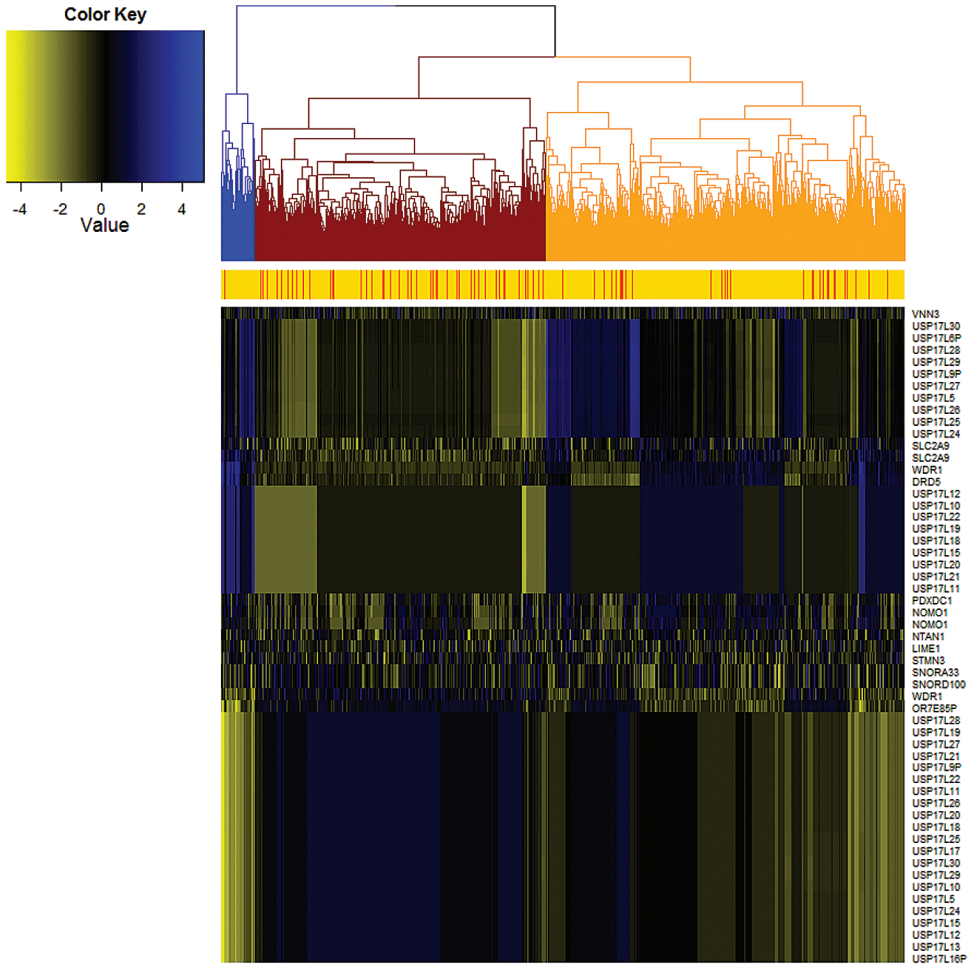

Fig. 3 Hierarchical clustering of the 55 validated transcripts based on the Euclidean distance of estimated expression level. The red stripe indicates the glioma case and the orange indicates the control.

analyze all genetic variants within a gene and have shown better power than single-SNP analyses.[14,30] A special case of our method focusing on an SNP-only model can be construed as SKAT analyses with the linear kernel. Similar to our approach, an eQTL-based method PrediXcan proposed to utilize reference data to impute transcriptome in GWAS.[31] The difference is that PrediXcan focused only on the association of imputed gene expression with phenotype, and we jointly analyze the genotype data, the imputed gene expression data, as well as the genotype-by-transcription interaction using a score-type variance component test.[27] It is critical to include the genetic effect, which captures the *trans*-eQTL effect through other genes or effects of other transcriptional regulation and biological pathways not via the imputed transcription. It has also been shown that analysis with single platform loses statistical power compared with the proposed joint analyses.[27] By integrating both genetic and transcriptomic data, we present a unified framework where SKAT and PrediXcan are 2 special cases, and our approach further provides omnibus tests to synthesize the optimal information.

A challenge of jointly analyzing SNP and gene expression data in genome-wide association studies is its difficulty in gaining access to the target tissue. In most GWAS, investigators collect peripheral blood samples and genotype DNA extracted from blood cells. While mRNA can also be extracted and profiled from blood cells to study immune-related diseases,[19] it is subject to serious limitations in interpretation when generalizing to non-immune related diseases such as glioma, since brain tissue is the target tissue, not blood cells. For glioma, it is not impossible to obtain target tissue from cases; however, it is often difficult and unethical if not impossible to obtain target tissue from controls. To integrate genomic data collected from different genetic association studies, we have developed a statistical method[27] under the framework of causal mediation modeling.[32,33] Utilizing the newly developed algorithm, our proposed iGWAS approach provides an analytic paradigm to exploit the information from an external eQTL study to infer the missing transcriptomic profile for GWAS subjects. We note that our eQTL analyses focused on the expression profile from only the cerebral tissue that is only limited to the frontal lobe, which may attenuate the signals or even diminish the likelihood of identifying signals. The information of anatomical sites can be easily incorporated in our analytic framework by

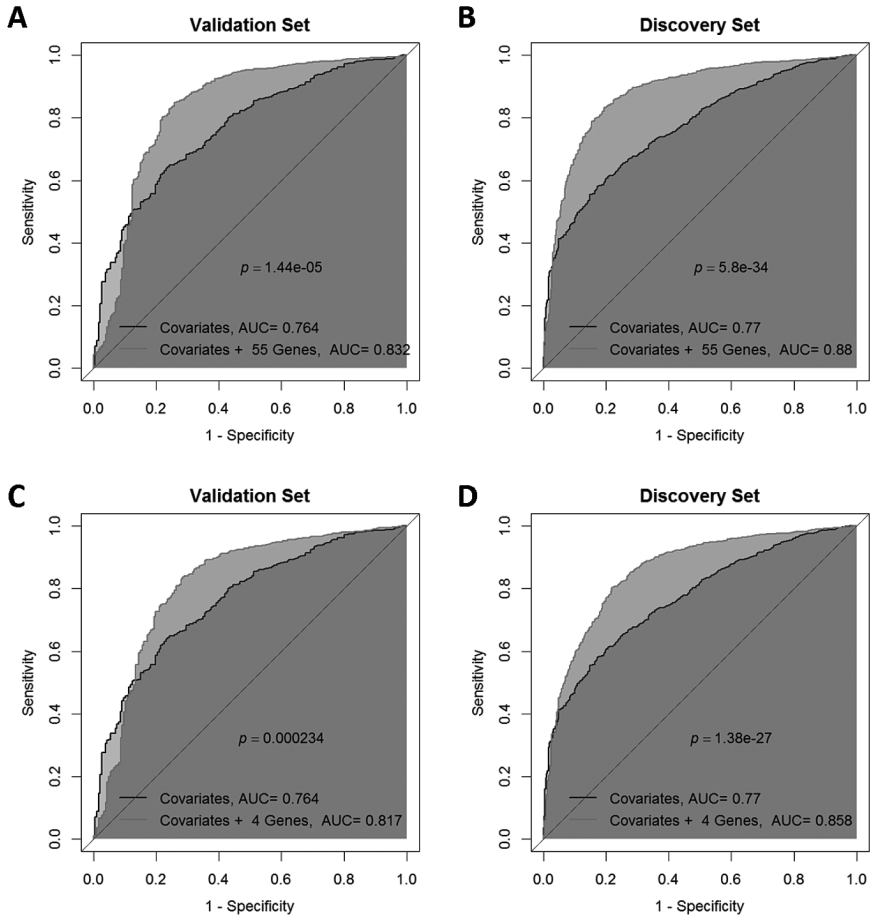

Fig. 4 The gene signature of glioma risk. (A, B) The ROC curve and its area under the curve (AUC) of the model with only covariates (dark gray curve) and that with covariates and the 55 transcripts (light gray curve) in Validation Set (A) and Discovery Set (B). (C, D) The ROC curve and its AUC of the model with only covariates (dark gray curve) and that with covariates and the 4 genes: *DRD5*, *WDR1*, *NOMO1*, and *PDXDC1* (light gray curve) in Validation Set (C) and Discovery Set (D).

developing lobe-specific eQTL models for prediction, which may increase the power of detecting susceptibility genes.

In the example illustrated here, we identified 4 new genes that have not been linked to glioma before: *DRD5*, *NOMO1*, *PDXDC1*, and *WDR1*. *DRD5* encodes the D5 subtype of the dopamine receptor, a G-protein coupled receptor which stimulates adenylyl cyclase.[34] Consistent with our findings that low expression was associated with glioma risk (Figure 5A and B), the Human Protein Atlas shows that DRD5 protein has high expression in cerebral cortex but is not detectable in glioma tumor tissue; the expression in cerebral cortex is mostly observed in neuropil but not in glial cells, endothelial cells, or neuronal cells.[35,36] *DRD5* has been associated with neurologic disorders such as Parkinson's disease,[37] multiple sclerosis,[38] schizophrenia,[39] and attention-deficit hyperactivity disorder,[40] and is also an FDA-approved drug target.[41] However, the literature is very limited on its role in carcinogenesis, which deserves more research.

NOMO1, also known as *PM5*, is one of the 3 highly similar genes located in a duplication region on p arm of chromosome 16[42]; the 3 genes encode proteins that may have the same function, and one protein has been identified as part of a complex involved in the nodal signaling pathway during development.[34,43] NOMO1 protein is highly expressed in normal cerebral cortex, particularly in glial cells and neuronal cells, but not in endothelial cells or neuropil; in contrast, its expression in glioma tissue ranges from medium to nondetectable.[35,36] The protein expression pattern is consistent with our findings in the GBM data from TCGA (Figure 5F) as well as in the eQTL-estimated expression (Figure 5E). *NOMO1* was found overexpressed in the cutaneous T-cell lymphoma cell line compared with normal peripheral blood monocytes,[44] but little is known about its role in glioma. Discussion of *WDR1* and *PDXDC1* is provided in the Supplementary material.

We note that the results from TCGA should be interpreted with caution because the set from TCGA was derived from tumor DNA, and GWAS and eQTL sets were derived from normal DNA. Because gene expression is profiled after tumors have developed, such differential expression could be from either causal genes that induce carcinogenesis or reactive genes with their expression altered secondary to cancer development.[45] The gene expression of the tumor consists of 2 components: the

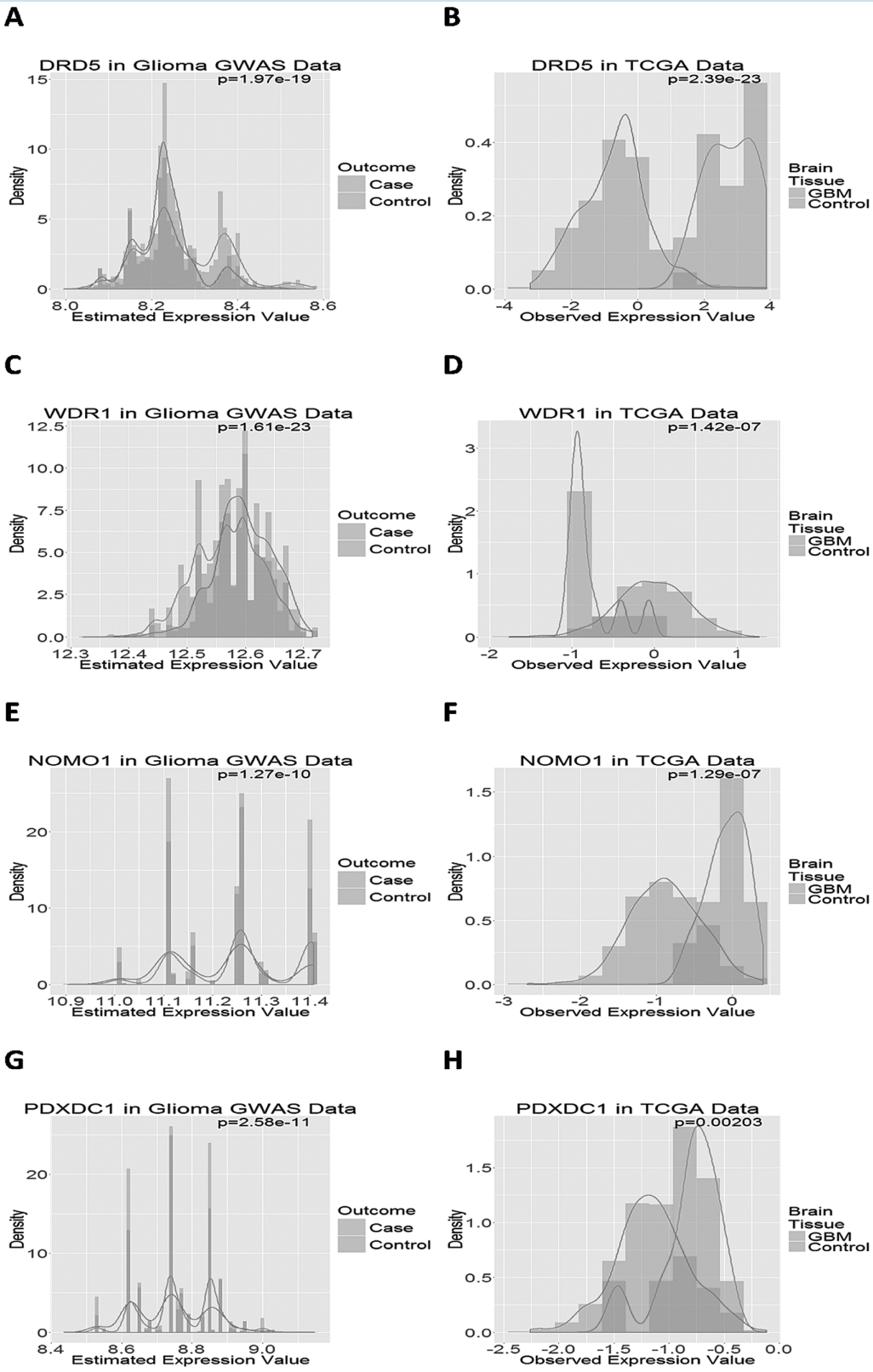

Fig. 5 Gene expression in glioma GWAS data and TCGA GBM data. (A) The gene expression level of DRD5 estimated for glioma cases and controls in our glioma GWAS data: (A) *DRD5*, (C) *WDR1*, (E) *NOMO1*, (G) *PDXDC1*. The observed gene expression level of brain tissue from GBM patients and organ-specific controls: (B) *DRD5*, (D) *WDR1*, (F) *NOMO1*, (H) *PDXDC1*.

causal gene expression and reactive gene expression. Since both GWAS and eQTL sets were derived from the nontumor DNA, we hypothesize that the predicted transcriptome is more likely to represent the causal gene expression than the reactive one. The 4 genes, *SLC2A9*, *LIME1*, *STMN3*, and *VNN3* with estimated expression not validated by TCGA data could be due to the difference of causal (in iGWAS data) and reactive (in TCGA data) expression, or simply false positives. Moreover, the validity of the estimated transcriptomic data for the GWAS subjects by the eQTL data relies on the assumption that the eQTL subjects are representative of the GWAS subjects, conditioning on age, gender, and ethnicity/population stratification. Exploratory analyses using low-grade glioma tumors were also conducted to validate *DRD5* and *WDR1* (Supplement Section IV). We expect that the set from TCGA tends to provide a more conservative validation, ie, we may miss causal signals due to its being intertwined with the reactive genes.

In conclusion, we identified 4 novel susceptibility genes of glioma: *DRD5*, *NOMO1*, *PDXDC1*, and *WDR1*, and constructed a genotype-based gene signature of glioma risk using integrative genomic analytics, iGWAS. The iGWAS approach integrates multiplatform genomic data, that is, SNP and gene expression data as well as different genetic association studies (ie, an eQTL study and a GWAS). The iGWAS has advantages of identifying biologically plausible results and improving statistical power, which make it a promising strategy to revisit existing GWAS data identifying new disease susceptibility genes that are transcriptionally functional.

Supplementary Material

Supplementary material is available at *Neuro-Oncology* online.

Funding

National Cancer Institute grant 5R03CA182937-02 (Integrative Modeling of Gene Expression into GWAS of Glioma, to Y-T.H., D.S.M.)

Acknowledgments

We thank the US National Institutes of Health/National Cancer Institute grant 5R03CA182937-02 (Integrative Modeling of Gene Expression into GWAS of Glioma, to Y-T.H., D.S.M.) for the support of the analysis project we present in this paper, and thank Dr J. Raphael Gibbs for sharing the brain eQTL data and providing the data preprocessing information.

Conflict of interest statement. The authors report they have none to declare.

References

1. Porter KR, McCarthy BJ, Freels S, et al. Prevalence estimates for primary brain tumors in the United States by age, gender, behavior, and histology. *Neuro Oncol*. 2010;12(6):520–527.
2. Ostrom QT, Gittleman H, Fulop J, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008–2012. *Neuro Oncol*. 2015;17(Suppl 4):iv1–iv62.
3. Farrell CJ, Plotkin SR. Genetic causes of brain tumors: neurofibromatosis, tuberous sclerosis, von Hippel-Lindau, and other syndromes. *Neurol Clin*. 2007;25(4):925–46, viii.
4. Malmer B, Henriksson R, Grönberg H. Familial brain tumours-genetics or environment? A nationwide cohort study of cancer risk in spouses and first-degree relatives of brain tumour patients. *Int J Cancer*. 2003;106(2):260–263.
5. Hemminki K, Tretli S, Sundquist J, et al. Familial risks in nervous-system tumours: a histology-specific analysis from Sweden and Norway. *Lancet Oncol*. 2009;10(5):481–488.
6. 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.
7. Wrensch M, Jenkins RB, Chang JS, et al. Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet*. 2009;41(8):905–908.
8. Sanson M, Hosking FJ, Shete S, et al. Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Hum Mol Genet*. 2011;20(14):2897–2904.
9. Rajaraman P, Melin BS, Wang Z, et al. Genome-wide association study of glioma and meta-analysis. *Hum Genet*. 2012;131(12):1877–1888.
10. Kinnersley B, Labussière M, Holroyd A, et al. Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat Commun*. 2015;6:8559.
11. Walsh KM, Codd V, Smirnov IV, et al.; ENGAGE Consortium Telomere Group. Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk. *Nat Genet*. 2014;46(7):731–735.
12. Hunter DJ, Chanock SJ. Genome-wide association studies and "the art of the soluble". *J Natl Cancer Inst*. 2010;102(12):836–837.
13. Kwee LC, Liu D, Lin X, et al. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet*. 2008;82(2):386–397.
14. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010;86(6):929–942.
15. Zhang M, Liang L, Morar N, et al. Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. *Hum Genet*. 2012;131(4):615–623.
16. Menashe I, Figueroa JD, Garcia-Closas M, et al. Large-scale pathway-based analysis of bladder cancer genome-wide association data from five studies of European background. *PLoS One*. 2012;7(1):e29396.
17. Spitz MR, Gorlov IP, Dong Q, et al. Multistage analysis of variants in the inflammation pathway and lung cancer risk in smokers. *Cancer Epidemiol Biomarkers Prev*. 2012;21(7):1213–1221.
18. Hsu YH, Zillikens MC, Wilson SG, et al. An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits. *PLoS Genet*. 2010;6(6):e1000977.
19. Moffatt MF, Kabesch M, Liang L, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*. 2007;448(7152):470–473.
20. Zhong H, Beaulaurier J, Lum PY, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet*. 2010;6(5):e1000932.

21. Gamazon ER, Badner JA, Cheng L, et al. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry*. 2013;18(3):340–346.
22. Cusanovich DA, Billstrand C, Zhou X, et al. The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum Mol Genet*. 2012;21(9):2111–2123.
23. Nicolae DL, Gamazon E, Zhang W, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010;6(4):e1000888.
24. Huang YT, Vanderweele TJ, Lin X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann Appl Stat*. 2014;8(1):352–376.
25. Zhao SD, Cai TT, Li H. More powerful genetic association testing via a new statistical framework for integrative genomics. *Biometrics*. 2014;70(4):881–890.
26. Huang YT, Liang L, Moffatt MF, et al. iGWAS: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genet Epidemiol*. 2015;39(5):347–356.
27. Huang YT. Integrative modeling of multiple genomic data from different types of genetic association studies. *Biostatistics*. 2014;15(4):587–602.
28. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*. 2010;6(5):e1000952.
29. Hudson TJ, Anderson W, Artez A, et al.; International Cancer Genome Consortium. International network of cancer genome projects. *Nature*. 2010;464(7291):993–998.
30. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82–93.
31. Gamazon ER, Wheeler HE, Shah KP, et al.; GTEx Consortium. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091–1098.
32. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143–155.
33. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol*. 2010;172(12):1339–1348.
34. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2011;39(Database issue):D52–D57.
35. Uhlen M, Oksvold P, Fagerberg L, et al. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*. 2010;28(12):1248–1250.
36. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419.
37. Wang J, Liu ZL, Chen B. Dopamine D5 receptor gene polymorphism and the risk of levodopa-induced motor fluctuations in patients with Parkinson's disease. *Neurosci Lett*. 2001;308(1):21–24.
38. Cosentino M, Zaffaroni M, Marino F. Levels of mRNA for dopaminergic receptor D_5 in circulating lymphocytes may be associated with subsequent response to interferon-β in patients with multiple sclerosis. *J Neuroimmunol*. 2014;277(1-2):193–196.
39. Zhao Y, Ding M, Pang H, et al. Relationship between genetic polymorphisms in the DRD5 gene and paranoid schizophrenia in northern Han Chinese. *Genet Mol Res*. 2014;13(1):1609–1618.
40. Mill J, Curran S, Richards S, et al. Polymorphisms in the dopamine D5 receptor (DRD5) gene and ADHD. *Am J Med Genet B Neuropsychiatr Genet*. 2004;125B(1):38–42.
41. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*. 2014;42(Database issue):D1091–D1097.
42. Martin J, Han C, Gordon LA, et al. The sequence and analysis of duplication-rich human chromosome 16. *Nature*. 2004;432(7020):988–994.
43. Haffner C, Frauli M, Topp S, et al. Nicalin and its binding partner Nomo are novel nodal signaling antagonists. *EMBO J*. 2004;23(15):3041–3050.
44. Lange A, Kistler C, Jutzi TB, et al. Detergent fractionation with subsequent subtractive suppression hybridization as a tool for identifying genes coding for plasma membrane proteins. *Exp Dermatol*. 2009;18(6):527–535.
45. Schadt EE, Lamb J, Yang X, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. 2005;37(7):710–717.