Targeted Sequencing in Chromosome 17q Linkage Region Identifies Familial Glioma Candidates in the Gliogene Consortium

Ali Jalali1*, E. Susan Amirian2*, Matthew N. Bainbridge3,4, Georgina N. Armstrong2, Yanhong Liu2, Spyros Tsavachidis2, Shalini N. Jhangiani4, Sharon E. Plon2, Ching C. Lau2, Elizabeth B. Claus5,6, Jill S. Barnholtz-Sloan7, Dora Illyasova8,9, Joellen Schildkraut9, Francis Ali-Osman10, Siegad Sadetzki11,12, Christoffer Johansen13,14, Richard S. Houlston15, Robert B. Jenkins16, Daniel Lachance16, Sara H. Olson17, Jonine L. Bernstein17, Ryan T. Merrell18, Margaret R. Wrench19, Faith G. Davis20, Rose Lai21, Sanjay Shete22, Kenneth Aldape23, Christopher I. Amos24, Donna M. Muzny4, Richard A. Gibbs4, Beatrice S. Melin25 & Melissa L. Bondy2

Gliomas comprise approximately 75% of all malignant primary brain tumors (PBTs) and account for an estimated 4% of cancer deaths in the United States1–4. Glioblastoma multiforme (GBM) is the most common type of glioma, constituting nearly 65% of cases, with an incidence rate of 2–3 per 100,000 in the United States.
United States and Europe. The 5-year survival for individuals with GBM is only about 10%, and median survival time is an estimated 12–14 months.

Despite decades of research, there are few established risk factors for glioma. A number of candidate gene and genome wide association studies have been conducted, and have, thus far, revealed seven low-penetrant susceptibility loci associated with sporadic glioma formation. With regard to familial glioma, known single gene disorders, such as neurofibromatosis, tuberous sclerosis, and Li-Fraumeni and Turcot’s syndromes, predispose patients to glioma formation, but cannot explain more than a minute proportion of cases. Therefore, the factors responsible for first-degree relatives of glioma patients having approximately twice the risk of glioma formation compared to unrelated individuals remains unclear.

Because the genetic basis of familial glioma remains enigmatic, the “Genetic Epidemiology of Glioma International Consortium” (Gliogene Consortium) was formed in 2006 to recruit families affected by glioma in 14 different institutions across five countries. This consortium has provided an unprecedented opportunity to further our understanding of the heritability of this rare, though highly fatal, condition, with the ultimate goal of uncovering enough information about glioma susceptibility to allow for the screening and genetic counseling of high-risk individuals and families. Our previous studies, which involved non-parametric and parametric linkage analyses, both yielded significant linkage peaks only on chromosome 17q (parametric linkage score: 3.1, nonparametric linkage score: 3.39). Because of the concordance between these prior findings, we conducted targeted sequencing focused on this region of chromosome 17q, with the aim of identifying the variant(s) or gene(s) that could explain linkage across this region to familial glioma. We, additionally, characterized deleterious rare variants within this chromosomal region among these glioma families.

### Results

The linkage region was defined as the 1.7 LOD drop region from linkage peaks, which spans bases 34,355,567-52,135,011 and 54,612,056-61,596,548 (GRCh37 coordinates) on chromosome 17 (solid bars in Fig. 1).

A total of 203 individuals from 23 families were successfully sequenced (33 affected, 170 unaffected) to an average depth of coverage of 95 fold in the target regions. Of these families, 20 had at least one affected individual sequenced (probands of three families failed sequencing and were subsequently excluded from additional analysis). Information on the demographics of affected individuals, whether sequencing was completed, and the relationship between affected individuals in each family is summarized in Table 1. Affected individuals were not sequenced if they were deceased prior to study initiation, or if a specimen was otherwise unobtainable.

After alignment and variant calling, a total of 4,830 variants were annotated. After removing common variants, variants mapping outside our selected target regions, variants only present in unaffected individuals, and variants not segregating among affected individuals in the same family, there were 539 remaining variants, which were subsequently submitted for Sanger sequencing verification for the affected individuals (Fig. 2A). All of the variants in our final list (Table 2) were located within the linkage region.

**Sanger sequencing.** Out of 539 variants interrogated by Sanger sequencing, some failed initial design and quality checks, including several attempts at successful primer design, in part due to the location of the variants within high-repeat or low-complexity regions or within duplicated regions. A total of 278 variants (51.6%) were validated in at least in a subset of affected individuals (Fig. 2B). Reads not meeting the aforementioned quality control criteria were removed, leaving 186 variants (66.9%) that were fully verified by Sanger sequencing. Our final filtering criteria were then applied (Fig. 2B). A final list of 21 candidate variants, all missense mutations, was obtained (Table 2).

**Candidate variants.** The 21 candidate variants on our final list are each private to individual families (Table 2). There were 15 candidate variants present in more than one affected individual per family (“2/2” affected ratio in Table 2), and the remaining six variants were in families with only one sequenced affected individual. Of the 15 shared variants, three variants are novel, never having been reported in a publicly available variant database (dbSNP, ESP, 1000 Genome Project). These include two missense mutations in RSADI and MYO19 in both affected and 3 out of 8 unaffected
| Relationship | Histology            | Age Dx | Sex    | Sequenced |
|--------------|----------------------|--------|--------|-----------|
| **Family A** | **Proband**          | 48     | Female | Yes       |
|              | **Sibling**          | 33     | Male   | Yes       |
|              | **P 2nd Cousin**     | 20     | Female | No        |
| **Family B** | **Proband**          | 40     | Female | Yes       |
|              | **P 1st Cousin**     | 30     | Female | Yes       |
| **Family C** | **Proband**          | 40     | Male   | Yes       |
|              | **Sibling**          | 46     | Male   | Yes       |
|              | **Sibling**          | 16     | Female | No        |
| **Family D** | **Proband**          | 54     | Female | Yes       |
|              | **Father**           | 31     | Male   | No        |
|              | **P Aunt**           | 55     | Female | No        |
| **Family E** | **Proband**          | 39     | Male   | Yes       |
|              | **Sibling**          | 27     | Female | Yes       |
| **Family F** | **Proband**          | 61     | Male   | Yes       |
|              | **Sibling**          | 52     | Female | Yes       |
| **Family G** | **Proband**          | 28     | Male   | Yes       |
|              | **Father**           | 29     | Male   | Yes       |
|              | **P Grandfather**    | 50     | Male   | No        |
| **Family H** | **Proband**          | 59     | Male   | Yes       |
|              | **Sibling**          | 42     | Male   | No        |
| **Family I** | **Proband**          | 58     | Female | Yes       |
|              | **Nephew**           | 32     | Male   | Yes       |
| **Family J** | **Proband**          | 27     | Female | Yes       |
|              | **Sibling**          | 38     | Male   | Yes       |
| **Family K** | **Proband**          | 65     | Male   | No        |
|              | **M Aunt**           | 82     | Female | Yes       |
|              | **M Great Aunt**     | 79     | Female | No        |
|              | **M Great Aunt**     | 81     | Female | No        |
|              | **M 1st Cousin Once Removed** | 63  | Female | No        |
|              | **M 1st Cousin Once Removed** | 61  | Male   | No        |
| **Family L** | **Proband**          | 39     | Female | Yes       |
|              | **M Aunt**           | 66     | Female | Yes       |
| **Family M** | **Proband**          | 66     | Female | Yes       |
|              | **M Aunt**           | 59     | Female | No        |
| **Family N** | **Proband**          | 66     | Male   | Yes       |
|              | **Sibling**          | 53     | Male   | No        |
| **Family O** | **Proband**          | 43     | Female | Yes       |
|              | **Mother**           | 27     | Female | No        |
|              | **Sibling**          | 45     | Male   | No        |
| **Family P** | **Proband**          | 28     | Male   | Yes       |
|              | **Identical Twin**   | 18     | Male   | Yes       |
| **Family Q** | **Proband**          | 10     | Female | Yes       |
|              | **M Grandmother**    | 65     | Female | No        |
| **Family R** | **Proband**          | 44     | Female | Yes       |
|              | **Sibling**          | 22     | Male   | No        |
|              | **Niece**            | 24     | Female | Yes       |
|              | **Niece**            | 4      | Female | No        |
|              | **M Aunt**           | 11     | Female | No        |
|              | **M 1st Cousin**     | 38     | Female | Yes       |
|              | **M 2nd Cousin Once Removed** | 49  | Female | No        |
members of Family J, and a missense mutation in G6PC in both affected and 0 out of 5 unaffected members of Family A.

Although each specific variant on our final list was private to an individual family, there were two genes for which more than one family had variants that made the final list. MYO19 was implicated in Families J and O, and KIF18B was implicated in Families G and B (Table 2).

Of the 21 candidate variants identified by our filtering schema, 19 had scaled C-scores above 10, indicating that these variants are predicted to be in the top 10% of the most deleterious possible substitutions in the human genome. The highest scaled C-scores belonged to SPAG9 rs143491486 (C-score: 32.0), RUNDC1 rs61995866 (C-score: 28.2), and the variant in TTC25 (chr17:40091564 C > G; C-score: 25.9). In Family L, both sequenced affected individuals and 3 out of 4 sequenced unaffected individuals had the SPAG9 rs143491486 variant. However, 2 of the 3 unaffected individuals who had this SPAG9 variant were children (Fig. 3a). The abbreviated CADD output is provided as supplementary information.

Discussion
In this study, we attempted to identify variants on targeted regions of chromosome 17q that could explain linkage across familial glioma cases. We also sought to describe the distribution of potentially deleterious rare variants within participating glioma families. As in any NGS study, we obtained a large list of variants and had to devise a filtering and verification strategy to refine our list for inclusion of only the most relevant variants (Fig. 2). Although we did not detect a variant (or gene) that could explain linkage across all participating families, we did identify several very rare or novel missense variants (many of which are predicted to be in the top 10% of the most deleterious substitutions possible in the human genome) that segregated in sequenced affected members of individual families. In complex diseases such as glioma, the importance of identifying deleterious rare variants (e.g., private mutations) should not be underestimated, as such variants may pinpoint genes that could be important to disease pathogenesis. Furthermore, despite being infre-
quent in the general population, the identification of these rare variants can highlight genes or genomic regions in which a series of less deleterious and more common mutations may interact to cumulatively increase disease risk in a larger population.

Two interesting candidate genes identified in this study are \textit{MYO19} and \textit{KIF18B}, which were the only genes that were implicated in more than one family. \textit{MYO19} codes for a myosin that is involved in mitochondrial motility\textsuperscript{16}. The \textit{MYO19} gene has not been well-studied in relation to carcinogenesis; however, one study has implied that it may act as a fusion gene in breast cancer tissues\textsuperscript{17}.

\textit{KIF18B} encodes a protein product that regulates microtubule dynamics (i.e., microtubule length) and thus, plays an important role in cell division through its involvement in mitotic spindle assembly\textsuperscript{18–20}. \textit{KIF18B} has been shown to be overexpressed in hepatocellular carcinoma, and some evidence suggests that the expression of this gene may also be deregulated in several other types of tumor tissues\textsuperscript{20,21}. Both of these genes warrant additional examination in future etiologic studies.

A promising candidate variant identified in our study, \textit{SPAG9} rs143491486 (p.Ser269Leu), is predicted to be among the top 0.1% of the most deleterious variants possible in the human genome (Table 2, Family L), based on its Combined Annotation-Dependent Depletion\textsuperscript{22} scaled C-score of 32. The product of this

**Table 2 | List of 21 very rare or novel missense variants within the linkage locus that segregate among all affected members of individual families.** Listed in order of the Combined Annotation-Dependent Depletion scaled C-scores\textsuperscript{22}, each variant is private to one family. Families not listed had no variants that survived filtration. The affected (or unaffected) ratio denotes the number of sequenced affected (or unaffected) individuals carrying the variant over total number of sequenced affected (or unaffected) individuals within the family.

| Family | Chr. 17 Position | Ref. Allele | Var. Allele | Gene Symbol | ESP6500 Freq. | 1000G Freq. | dbSNP ID | Affected Ratio | Unaffected Ratio | Scaled C-Score |
|--------|------------------|------------|------------|-------------|---------------|-------------|----------|---------------|----------------|---------------|
| L      | 49098662         | G A        | SPAG9      | 0.069%      | rs143491486   | 2/2         | 3/4      | 32.0          |                |               |
| C      | 41134047         | G C        | RUNC1      | 0.069%      | rs61995866    | 2/2         | 2/4      | 28.2          |                |               |
| M      | 40091564         | C G        | TTC25      | 1/1         |               | 5/8         | 25.9     |               |                |               |
| O      | 34883425         | G A        | MYO19      | 1/1         |               | 4/5        | 19.2     |               |                |               |
| G      | 39503458         | C T        | KRT33A     | 0.015%      | rs142400197   | 2/2         | 1/4      | 19.2          |                |               |
| B      | 43006370         | C T        | KIF18B     | 0.008%      | rs201865018   | 2/2         | 1/6      | 18.4          |                |               |
| A      | 43859014         | C T        | MYO19      | 0.03%       | rs149568450   | 2/2         | 2/4      | 18.1          |                |               |
| R      | 42750898         | A T        | C17orf104  | 0.05%       | rs192757598   | 2/2         | 2/4      | 16.2          |                |               |
| A      | 41063291         | G A        | G6PC       | 2/2         | 0/5          | 14.5      |          |               |                |               |
| H      | 37785802         | C A        | PPP1R1B    | 2/2         | 2/4          | 14.1      |          |               |                |               |
| C      | 38932391         | G A        | KRT27      | 0.077%      | rs148928902   | 1/1         | 2/6      | 14.0          |                |               |
| J      | 45904542         | C T        | MRP10      | 0.046%      | rs149631185   | 2/2         | 3/4      | 14.0          |                |               |
| J      | 48561815         | T C        | RASD1      | 0.042%      | rs200709037   | 2/2         | 3/6      | 12.2          |                |               |
| S      | 41338453         | G C        | NBR1       | 0.028%      | rs147094229   | 2/2         | 4/8      | 7.2           |                |               |
| B      | 48141461         | G A        | ITGA3      | 0.05%       | rs14210478    | 2/2         | 1/6      | 10.7          |                |               |
| K      | 39036435         | C T        | KRT20      | 0.008%      | rs147094229   | 1/1         | 0/24     | 10.5          |                |               |
| J      | 39620632         | G A        | KRT32      | 0.008%      | rs147094229   | 2/2         | 4/8      | 7.2           |                |               |
| K      | 55194252         | G A        | AKAP1      | 0.062%      | rs149147838   | 1/1         | 1/24     | 6.7           |                |               |
| N      | 45419305         | C G        | EFCAB13    | 0.038%      | rs138179179   | 1/1         | 4/7      | 5.0           |                |               |

**Figure 3 |** (a) \textit{SPAG9(L)} and (b) \textit{RUNC1(C)} mutations in familial glioma pedigrees. Individuals with glioma are shown as filled. Individuals with other cancers are shown as half-filled. Disease and age in years (y) at first diagnosis is given underneath the symbol, current age or age at death (+) above it. Glioma type is shown (AA, anaplastic astrocytoma; Astro, astrocytoma; GBM, glioblastoma multiforme). Other cancers in the pedigree are shown (OvC, ovarian cancer; Mel, melanoma; TesC, testicular cancer; Lym, lymphoma). Carriers of \textit{SPAG9} (a) and \textit{RUNC1} (b) mutations are shown with their specific mutation, whereas individuals who tested negative for the mutation in the specific pedigree are depicted as wild-type (WT).
gene, sperm-associated antigen 9 (SPAG9), is a member of a scaffolding protein family that helps MAP-kinases bind with their transcription factors to activate specific signal transduction pathways. SPAG9 is actually a putative oncoprotein that has been implicated in prostate, breast, hepatocellular, thyroid, cervical, lung, bladder, and endometrial cancers. Recently, Yi et al. showed that SPAG9 was differentially overexpressed in human astrocytomas, compared to normal astrocytes, and that SPAG9 expression levels were positively correlated with tumor grade (ρ < 0.001). Interestingly, in our study, one of the two sequenced affected individuals carrying the SPAG9 variant in Family L had a Grade II astrocytoma. Although the exact mechanism through which this particular SPAG9 variant may be involved in gliomagenesis is unknown, it has been suggested that the SPAG9 protein may influence cell invasion by upregulating the expression of MMP9 (matrix metalloproteinase 9, a.k.a. gelatinase-B), which, in turn, has been shown to be involved in the neovascularization of malignant gliomas.

The non-synonymous SPAG9 variant identified in our study contributes to existing evidence that this gene may play an important role in glioma susceptibility. This variant is predicted to be protein damaging by SIFT (SIFT score: 0; http://sift.jcvi.org/) and PolyPhen (Polymorphism Phenotyping score: 0.997), and is at a location that is highly evolutionarily conserved. Ascertainment of whether currently unaffected members of Family L who carry this variant develop glioma in the future will be extremely important, as the unaffected individuals carrying this variant were under the age of 20 and may develop glioma later in life.

There is little known about the variant with the next highest C-score, RUNDC1 rs61995866 (p.Glu386Gln). This variant is predicted to be in the 1% most deleterious of all possible substitutions in the human genome. The RUNDC1 protein has been shown by a high-throughput RNA interference-based screening study to be a p53 regulator. Interference with this gene resulted in increased p53 transcription, thus indicating that the RUNDC1 protein may be a p53 antagonist. However, this variant has not previously been studied in relation to familial glioma. Both sequences affected individuals in Family C carried this mutation, as did two of four sequenced unaffected individuals (implying incomplete penetrance) [Fig. 3b]. However, one of the two unaffected individuals carrying the RUNDC1 variant was only in her early 20s at last contact.

Even the variants on our final list that have lower C-scores may provide interesting candidates for further investigation. For example, the candidate gene with the highest number of somatic non-silent mutations in The Cancer Genome Atlas (TCGA) low-grade glioma database is G6PC with three tumor samples containing missense somatic variants, followed by one tumor sample for each of MYO19, GHD迦 and KRT27. Looking at the combined low-grade glioma and glioblastoma multiforme TCGA data, G6PC remains the gene with the highest number of somatic non-silent mutations (four tumor samples) followed by MYO19 (three tumor samples) and KIF1BB (two tumor samples). This was based on an analysis of 279 low-grade glioma and 268 glioblastoma multiforme tumor samples from the TCGA project as reported by the International Cancer Genome Consortium (https://dcc.icgc.org/).

Targeted deletion of G6PC in mice leads to hepatic tumorigenesis, and this gene is expressed at comparable levels in the central nervous system according to the Cancer Genome Anatomy Project. Along with the finding that both affected (with low grade glioma) and no unaffected members of Family A carried the G6PC missense variant in their germline (scaled C-score: 14.5), this information makes the G6PC variant another interesting candidate for future study.

It is also noteworthy that four different keratin genes were implicated in our final list (Table 2, Families C, H, K, and J). Overexpression of cytokeratins has been reported in some cancers, but little is known about the potential role of these genes in gliomagenesis. There is a cluster of keratin-related genes or pseudo-genes on chromosome 17q, thereby obscuring whether these findings may truly be biologically relevant or simply represent an artifact of our loci of interest.

Glioma is a complex disease where multiple loci are likely to be important for disease development within a family. In order to maximize resources, the regions targeted for sequencing in our study were chosen strategically to ensure comprehensive coverage of what were likely to be the most pertinent linkage regions. However, within our sequencing targets, there are numerous regions that are notoriously difficult to sequence via the NGS platform and Sanger sequencing, possibly due to genomic duplications or high repeat content. Additionally, we acknowledge that other loci, outside of the regions covered by this study, have previously been linked to familial glioma (although no causative variants in these regions have been definitively established). For example, in our other analyses, we have found that two Gliogene families (D and R) had TP53 mutations that may explain their increased glioma risk. Interestingly, neither of these families had chromosome 17q variants that met our final filtering criteria in the current study.

One of the challenging aspects of studying diseases with high mortality, such as glioma, is that there are very few families for which we can obtain biologic samples from more than one or two affected family members, which limits the possibility of performing segregation analyses of rare variants. Only about 5% of glioma patients have a family history of glioma, and in 83% of glioma families, there are only two glioma patients in the family, indicating a complex disease with low penetrance of causal variants. Unfortunately, we were unable to detect a variant that could explain linkage across all participating families in this study. Because there are no established filtering strategies for studies such as this one, we had to develop the filtering schema presented here to maximize the potential of finding such variants, while simultaneously excluding variants common among healthy populations.

Despite these limitations, the findings of this study may lend a substantive amount of insight into genes involved in gliomagenesis, and future studies should evaluate the candidate genes/regions identified here for the presence of more common/less deleterious variants that may cumulatively impact familial glioma risk in a larger population. Overall, our study complements the previous research conducted by the Gliogene Consortium on the enigmatic factors contributing to familial glioma risk. The importance of the chromosome 17q region is becoming increasingly clear, though several additional studies are necessary before definitively conclusions can be drawn. We are currently in the process of obtaining copy number variation (CNV) data for a subset of the Gliogene families. The CNV data will be examined in the context of what is currently known about factors that predict familial glioma susceptibility in the hopes that enough information can be amassed from the Gliogene Consortium’s series of studies to be able to differentiate and provide genetic counseling to high-risk families and individuals.

Methods

Study Population. Families recruited for participation in the Gliogene Consortium had at least two biologically related members (83% of families with two or 17% with three or more affected family members) who were diagnosed with a histologically-confirmed glioma. The study population and recruitment scheme have been described in detail elsewhere. Briefly, for the analyses presented here, the glioma families (n = 23) contributing most to the linkage peaks identified in Shete et al. (2011) and Sun et al. (2012) were selected for genomic sequencing.

Sample collection. Blood or saliva samples from glioma family members were obtained under written informed consent at each Gliogene Consortium recruitment center. DNA samples were prepared as previously described. This study was approved by the Institutional Review Board of each Gliogene institution, including Baylor College of Medicine, and was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants or their guardians.
Density of Sequencing

Successful enrichment of the capture libraries was estimated to range from a 6 to 9 of 7500. The efficiency of the capture was evaluated by performing a qPCR-based quality quantification and size distribution of the pre-capture LM-PCR product was determined using the LabChip GX electrophoresis system (PerkinElmer).

Custom capture design. The DNA capture reagent was designed to target all coding exons of all genes, the UTRs of cancer related genes, miRNA binding sites in all 3 UTRs, transcription factor binding sites, miRNA and small nucleolar RNA within the chromosome 17 linkage peak region.

The pre-capture libraries were pooled as a 46-plex (totaling 1 ug per pool) and hybridized in solution to the custom Gliogene capture reagent (1.6 Mb, NimbleGen) according to the manufacturer’s protocol (NimbleGen SeqCap EZ Exome Library SR User’s Guide) with minor revisions. Human Cot1 DNA was added into the hybridization to block repetitive genomic sequences, followed by post-capture LM-PCR amplification using the 2X SOLID Library High Fidelity Amplification Mix (a custom product manufactured by Invitrogen). Purification was performed with Agencourt AMPure XP beads after enzymatic reactions, and following the final purification, quantity and size distribution of the pre-capture LM-PCR product was determined using the LabChip GX electrophoresis system (PerkinElmer).

Sequencing. Library templates were prepared for sequencing using Illumina’s cBot cluster generation system with TruSeq PE Cluster Generation Kits. Briefly, these libraries were denatured and diluted in hybridization buffer in order to achieve a load density of ~800 K clusters/mm². Each library pool was loaded in a single lane of a HiSeq flow cell, with 2% phiX control library spiked into the lane for run quality control. The sample libraries underwent a ligation bridge amplification to form clonal clusters, followed by hybridization with the sequencing primer. The sequencing run was performed in paired-end mode using the Illumina HiSeq 2000 platform. Using the TruSeq SBS Kits, sequencing-by-synthesis reactions were extended for 101 cycles from each end, with an additional 7 cycles for the index read. With the sequencing runs yielding an average of ~838 Mb per sample, samples achieved an average of 86% of the targeted exome bases covered to a depth of 20X or greater.

Initial sequence analysis was performed using the HGSC Mercury analysis pipeline (https://www.hgsc.bcm.edu/content/mercury). In summary, the .bcf files produced on-instrument were first transferred into the HGSC analysis infrastructure by the HiSeq Real-time Analysis module. Mercury then ran the vendor’s primary analysis software, which multiplexed pooled samples and generated sequence reads and base-call confidence values (qualities), followed by the mapping of reads to the GRCh37 Human reference genome (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grchuman/) using the Burrows-Wheeler aligner (BWA, http://bio-bwa.sourceforge.net/), resulting BAM (binary alignment/map) file undergoent quality recalibration using GATK (http://www.broadinstitute.org/gatk/), and where necessary the merging of separate sequence-event BAMs into a single sample-level BAM. BAM sorting, duplicate read marking, and realignment to improve insertion/deletion discovery all occur at this step.

Sanger validation. Mutation validation was performed with bidirectional sequencing of the selected sample sites. PCR reactions were prepared using 5 ng of genomic DNA, 0.4 μM oligonucleotide primers, and 0.7X Qiagen Multiplex Master Mix (Cat. no. 200145) containing HotStar Taq, buffer, and polymerase. Reactions were performed with a 2X SOLiD PCR and final sequencing PCR and final sequence performance. Touchdown PCR was performed with the following parameters: 98°C for 5 min. and 10 cycles of 98°C for 30 sec., 72°C for 30 sec. (decreasing by 1°C per cycle), and 72°C for 1 min. The reaction then continued with 30 cycles of 98°C for 30 sec., 63°C for 30 sec., and 72°C for 1 min, followed by a final extension at 72°C for 5 min. The PCR products were purified with a 1:15 dilution of Exo-SAP, diluted by 0.6X, and then cycle-sequenced for 25 cycles using a 1/64th dilution of BigDye Terminator v3.1 reaction mix (Applied Biosystems, Cat. No. 4337456). Finally, reactions were precipitated with ethanol, resuspended in 0.1 mM EDTA, and analyzed on an ABI 3730xl sequencing instruments using the Rapid36 run module and 3x base-calling. SNPs were identified using SNP Detector software and then visually validated with Consed.

Variant annotation and filtering. Variants were annotated for functional effect on protein and presence in variant databases (i.e., dbSNP), as well as known minor allele frequency using AnnoVar [21]. We developed a filtering schema (described below) to evaluate these findings and reduce it to a subset of variants more likely to be etiologically relevant (i.e., very rare protein-altering candidate variants within the linkage region). We implemented a filtering strategies using scripts written in R (version 2.15.3).

Variants that met initial filtering criteria were verified by Sanger sequencing among affected individuals, and an additional post-Sanger filtering strategy was then applied. Filtering was carried out in a stepwise fashion with verification of the excluded variants at each step to ensure that biologically significant variants were not being eliminated.

Initial Filtering Strategy. The borders of the linkage region were defined based on a 1.7 LOD drop region (~ 99% linkage confidence interval) from the linkage peaks identified in our previous two analyses [13,14], which encompassed two adjacent regions on chromosome 17q. Variants located outside this linkage region were excluded (“Linked locus” filter), unless the variants outside the region were in genes found in the Cancer Gene Census (CGC) [22]. The CGC, which is maintained by Welcome Trust Sanger Institute, is a catalogue of variants that have previously been implicated to have a causal role in oncogenesis. At the time of this analysis, the following 23 genes on chromosome 17 were included in the CGC: TAF15, SUZ12, LASPi, CLTC, CDKR2, NP1, RARA, BCR1, DDx5, ETv4, MLLt6, Erbb2, Col1A1, Brrip1, Hlf, CD79B, Dysbindin, Ino, Mcr1, Srsf2, AspSCR1, Canti, Srsf2, and Alop.

Among non-CGC genes, variants that were not within 200 bases of the sequencing target region were excluded (“Targeted Region” filter). Next, common variants, defined as those that had >5% frequency in the Exome Sequencing Project (ESP5400 release) [48] or 1000 Genome (1000G, 2010_07 release) [15] CEU population, were also excluded (“Rare Variant” filter). Finally, we removed any variant that was not shared by all sequenced affected individuals of at least one family, prior to Sanger verification (“Family Segregation” filter).

Post-Sanger Filtering Strategy. Variants were considered verified if Sanger reads had full concordance with NGS reads in all affected individuals, and both the forward and reverse reads agreed. Final filtering involved excluding the following: variants outside of the linkage region (“Linked locus” filter); variants with a minor allele frequency >0.1% in the ESP5400 release or 1000G 2010_04 release (“Very Rare Variant” filter); variants not within the exonic sequencing target regions (“Targeted Exome” filter); and variants that did not alter the amino acid sequence of the product protein (“Protein Alteration” filter). The aim of this filtering strategy was to narrow down the list of variants to a list of very rare protein-altering candidate variants within our linkage locus. This was, however, carried out in a step-wise fashion with verification of the excluded variants at each step for potential biological function to ensure we are not excluding any biologically significant variants.

Combined Annotation-Dependent Depletion C-Score. After implementing the above filtering schema, we used scaled C-scores from the Combined Annotation-Dependent Depletion (CADD) method to rank the variants remaining on our final list potential to deleteriousness [22]. The CADD C-score provides an integration of 63 different annotations (i.e., alledic diversity, functionality, putative pathogenicity, and evolutionary conservation) into one score per variant. For example, the C-score integrates scores from PolyPhen2 [25] and SIFT [26] predicted impact of the variant on protein function, GERP++ [27] and PhyloP [28] for evolutionary conservation, and a number of other annotations for prioritization of causal variants [22]. A scaled C-score ≥10 indicates variants predicted to be in the top 10% of the most deleterious possible substitutions in the human genome, a score of ≥20 indicates variants predicted to be in the top 1%, and a score of ≥30 indicates variants predicted to be in the top 0.1%. More details on the CADD C-score can be found elsewhere [22].

Ethics. Written informed consent was obtained from each subject or from his or her guardian. Approval from local institutional review boards was received at each Gliogene participating institution. This study was conducted in accordance with the Declaration of Helsinki.

1. Adamson, C. et al. Glioblastoma multiforme: a review of what we have been and where we are going. Expert Opin Investig Drugs 18, 1061–1089 (2009).
2. Ogbuki, H. & Kleihues, P. Epidemiology and etiology of gliomas. Acta Neuropathol 109, 93–108 (2005).
3. Schwartzbaum, J. A., Fisher, J. I., Aldehyde, D. K. & Wrensch, M. Epidemiology and molecular pathology of glioma. Nat Clin Pract Neurol 2, 494–503; quiz 1 p following 516 (2006).
4. Ostrom, Q. T. et al. CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006–2010. Neuro Oncol 15, Suppl 2, i–56 (2013).
5. Ostrom, Q. T. et al. The epidemiology of glioma in adults: a ‘state of the science’ review. Neuro Oncol 16, 896–913 (2014).
6. Liu, Y. et al. Genetic advances in glioma: susceptibility genes and networks. Curr Opin Genet Dev 20, 238–244 (2010).
7. Melin, B. & Jenkins, R. Genetics in glioma: lessons learned from genome-wide association studies. Curr Opin Neurol 26, 688–692 (2013).
8. Rajaraman, P. et al. Genome-wide association study of glioma and meta-analysis. Hum Genet 131, 1877–1882 (2012).
9. Melin, B. et al. Known glioma risk loci are associated with glioma with a family history of brain tumours -- a case-control gene association study. Neuropathol Appl Neurobiol 39, 999–1012 (2013).
10. Shete, S. et al. Genome-wide association study identifies five susceptibility loci for glioma. Nat Genet 41, 899–904 (2009).
