Published in final edited form as:

ACS Chem Neurosci. 2018 January 17; 9(1): 73–79. doi:10.1021/acscchemneuro.7b00362.

Large Scale Identification of Variant Proteins in Glioma Stem Cells

Ekaterina Mostovenko¹, Ákos Végvári², Melinda Rezeli², Cheryl F. Lichti^{1,3}, David Fenyő⁴, Qianghu Wang^{5,7}, Frederick F. Lang⁶, Erik P. Sulman^{5,7,8}, K. Barbara Sahlin², György Marko-Varga², and Carol L. Nilsson^{9,*}

¹Department of Anatomy and Neurobiology, Virginia Commonwealth University School of Medicine, 1217 E. Marshall St., Richmond, VA 23284

²Clinical Protein Science & Imaging, Biomedical Center, Department of Biomedical Engineering, Lund University, SE-221 84 Lund, Sweden

³Department of Pathology and Immunology, Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, Missouri, 63110

⁴Department of Biochemistry and Molecular Pharmacology and Institute for Systems Genetics, New York University School of Medicine, New York City, New York 10016, United States

⁵Department of Genomic Medicine, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, United States

⁶Department of Neurosurgery, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, United States

⁷Department of Radiation Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, United States

⁸Translational Molecular Pathology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, United States

⁹Center of Excellence in Biological and Medical Mass Spectrometry, Lund University, Klinikgatan 32, Lund, SE-221 84 Sweden

Abstract

Glioblastoma (GBM), the most malignant of primary brain tumors, is a devastating and deadly disease, with a median survival of 14 months from diagnosis, despite standard regimens of radical brain tumor surgery, maximal safe radiation, and concomitant chemotherapy. GBM tumors nearly

Corresponding Author. Email: carol.nilsson@med.lu.se, **ORCID:** 0000-0002-2838-8751.

Supporting Information

Non-redundant list and sequences of variant peptides validated by transcript matching, list and sequences of 86 heavy labeled peptides used to quantify variant peptides, and list and sequences of 211 heavy labeled peptides for parallel reaction monitoring of variant peptides (PDF).

Author Contributions. EM, CFL, DF, KBS, and QW critically analyzed 'omics datasets. FFL and EPS supplied the samples and provided insight into neuro-oncological aspects related to the findings. EM, MR, CFL, GMV and CLN designed the experiments. AV created the custom SAV database. All co-authors contributed to the writing of this report.

The authors declare no competing financial interest.

always re-emerge after initial treatment and frequently display resistance to current treatments. One theory that may explain GBM re-emergence is the existence of glioma stem-like cells (GSCs). We sought to identify variant protein features expressed in low passage GSCs derived from patient tumors. To this end, we developed a proteomic database that reflected variant and non-variant sequences in the human proteome, and applied a novel retrograde proteomic workflow, to identify and validate the expression of 126 protein variants in 33 glioma stem cell strains. These newly identified proteins may harbor a subset of novel protein targets for future development of GBM therapy.

TOC Graphic

Keywords

glioblastoma; GBM; precision medicine; protein single amino acid variants; proteomics; transcriptomics; bioinformatics; targeted mass spectrometry; protein quantification; parallel reaction monitoring

Introduction

Brain cancer, especially glioblastoma (GBM), is among the deadliest of cancers. Despite aggressive therapy with maximal safe surgical resection, chemotherapy, and radiation, survival averages just over one year^{1, 2}. Recurrence of the primary tumor is nearly inevitable, and recurrent tumors are often more aggressive than the primary tumor. One hypothesis for the recurrence of GBM is the existence of GSCs^{3–7} in areas of the brain that are surgically inaccessible or resistant to standard-of-care treatments. Recently, a comparative study of primary and recurrent tumor samples from 70 patients treated with radio- and chemotherapy showed a >10-fold enrichment in tumor stem-like cells in the recurrent tumors⁸. Those treatments also alter the phenotype of the GSCs, turning them into a more proliferative cell type. The underlying events that lead to the phenotypic change involve selective pressures upon the GSC populations, resulting in alterations in metabolic and signaling pathways that maintain cell proliferation and resistance to cytotoxic stimuli. Thus, the dismal clinical outlook of GBM is at least partly due to the ineffectiveness of standard of care treatments on GSCs.

Most GBM tumors are sporadic, although there is an increased risk of malignant brain tumors in neurofibromatosis⁹. Familial glioma in which heritable single gene mutations occur comprises another small fraction of GBM^{10–12}. GBM tumors display characteristic somatic genetic mutations in TP53, NF1, PTEN, and EGFR^{13, 14} while low grade gliomas (LGG) frequently harbor IDH1/2 mutations¹⁵. Tumors may be classified and subclassified

by their genetic, epigenetic¹⁵, or even metabolic characteristics¹⁶. Deep datasets that describe molecular features can be studied to provide knowledge that may guide the treatment of tumors. Despite these efforts, the outlook and median survival of GBM patients remains at just over one year following diagnosis.

GBM therapeutics that target disease-related proteins may be useful as adjuvant therapies. We have previously studied GSC protein expression in the context of the chromosome-centric human proteome project (C-HPP)^{17–20}, a global research consortium. Our initial findings in these early reports were not as deep as we hoped. After much reflection on the subject of the failure of proteomics to provide new targets for GBM, we devised a novel retrograde proteomic workflow to identify protein variant features in low passage GSC lines derived from patient GBM tumors and then examined each variant for correlation with the risk of decreased survival in glioma²¹.

Proteomics is the comprehensive study of protein abundances, localization, function and interactions. Because the proteome is related to the genome, proteomics may be viewed as a gene annotation tool. Recent advances in mass spectrometry allow great depth and routine measurement of thousands of species^{22, 23}. However, the bottleneck in protein identification is the task to assign all of the acquired spectra correctly. Unlike DNA and RNA sequencing, proteomic data analysis workflows rely on the quality of the tandem mass spectrometric data and the degree of completeness of the protein sequence database. Currently, the human proteome coverage in ProteomicDB is approximately 80%. At least 16% of these proteins entries (neXtProt 2017-04-12, www.nextprot.org) still lack evidence at the protein level, let alone the annotation of their isoforms, the so-called “missing proteins”^{24, 25}. These proteoforms may have altered biological functions compared to the canonical/consensus proteins, and could therefore be associated with altered qualities in a disease state of an organism, for example, BRCA1 and BRCA2 in human breast cancer²⁶.

The latest version of neXtProt contains 5,317,610 protein variants, including natural and disease mutations. As part of the C-HPP consortium, we previously annotated single amino acid variants (SAVs) in GSCs derived from chromosome 19¹⁹. In that pilot study²⁷, we identified a metabolic enzyme variant now in development as a potential novel therapeutic target in GSCs²⁸. As investigators of resistance mechanisms in GSCs, we sought to identify all SAVs derived from all chromosomes, in order to study their potential link to pathogenesis in future studies. Here, we report for the first time, the proteome-wide annotation of SAVs derived from the full complement of the genome, in GSCs.

Results and Discussion

To identify variant peptides, we created a custom database containing SAVs from dbSNP, 2,267,533 in total. We generated short peptide sequences for each variant. Because the cell proteomes were digested with trypsin, we allowed one hypothetical tryptic missed cleavage on each side of the variant sequence and assigned a unique identifier reflecting the original NextProt accession number, reference amino acid position of the SAV in the protein, and the substitute amino acid (i.e. NX_Q9H7N4-T-5-P denotes the substitution of proline for threonine at amino acid position 5 in Splicing factor, arginine/serine-rich 19 also known as

SCAF1). Each peptide containing a SAV could be easily discriminated from the rest of the database. Furthermore, the risk of identifying non-variant, non-unique peptides corresponding to the reference protein, was eliminated in a filtering step. The creation of a database containing only short sequences provided a different challenge caused by decreased complexity. We addressed this shortcoming by appending a UniProt canonical to the variant database as described in detail in Lichti et al.¹⁹.

We compared several search engines and found that PEAKS searches resulted in the most accurate identification of variant peptides¹⁹. Following the database search was a step to filter out all canonical sequences and non-unique peptides. In a conventional database search, the next step would be to filter out erroneous identifications based on the false discovery rate (FDR). However, in our opinion, the global FDR value is not indicative of the quality of the identification, due to the nature of our custom database. FDR estimation is based on the target-decoy approach, in which the decoy database is usually automatically generated by the search engine or provided by the researcher, by use of a scrambled or reverse target database. Despite the apparent simplicity of the task, it remains challenging and there is no consensus on the best strategy to generate a decoy database²⁹.

Variant peptide sequences were matched to the RNAseq data from corresponding cell lines, provided by M.D. Anderson Cancer Center. The non-variant peptide sequence was matched to the reference genome (GRCh37/hg19) using PGx (25) (Fig. 1), the subsequent output file contained the chromosomal location of the target peptide, which was applied to parse RNA-seq raw data (SAM tools). The positioning of the variant in the peptide served as an anchor to predict open reading frames. Each nucleotide sequence of the predicted area was then translated back to amino acids and compared with the peptide of interest (Fig. 1) resulting in a distribution of variant and wild-type read counts across all cell lines.

We identified 1,022 variant peptides initially. We compared Protein BLAST (<https://blast.ncbi.nlm.nih.gov/>) against the canonical human database. All peptides were required to match canonical sequence 100% with maximum 1 missed match. The accession code of the query was compared with the subject and only unique matches were kept. We then assigned new IDs to the peptide, containing information regarding the peptide sequence location of the variant. Additionally, position and reference (derived from the unique identifier) were required to be correct. Many structural proteins, such as myosin, have numerous homologues. An SAV in one protein often matched perfectly the canonical sequence in a different human myosin isoform; therefore, all homologous matches were removed. Furthermore, we found matching sequences to SAVs in proteins other than myosin. Our final list contained ~400 unique peptides (Suppl. Table 1).

Not all identified variant peptides were appropriate for mass spectrometric quantification by parallel reaction monitoring (PRM): many were longer than 20 amino acids, had challenging retention properties on the LC column, or were difficult to synthesize as heavy peptides. Despite the issues encountered, we were able to confirm the presence of 126 variant peptides by use of PRM. For instance, cytoplasmic C-1-tetrahydrofolate synthase (MTHFD1) R653Q was found in fifteen of the GSC lines (Fig. 2). 43 variant peptides were selected for further investigation to determine their concentration together with the corresponding reference

peptides in 33 GSCs. Altogether, 58 peptides were quantified in at least one cell line, including 29 SAV-containing peptides (Suppl. Table 2).

During the experimental phase, each GSC culture plate was split for sample analysis by proteomics and transcriptomics. Therefore, all peptides were validated against the corresponding, culture-matched RNA-seq data. For each variant peptide, its reference sequence was used to identify a match against a reference genome by use of PGx³⁰, providing the exact chromosomal location for a corresponding nucleotide sequence, as well as its orientation in DNA (3' to 5' or 5' to 3' strand). Thus, we were able to match the peptide sequences to the RNA-seq data. RNA-seq provides short reads (75-mers or shorter) which contain information on their start and finish as well as the nature of the match: exact match (to reference), deletion (when a number of nucleotides are missing compared to the reference), insertion (when a number of nucleotides are inserted compared to the reference), and an intron area. In order to match with RNA-seq data, a corresponding peptide sequence was required to be at least partially located within the read and nucleotides corresponding to the variant amino acid present in full. RNA sequences can be read in six different ways, based on the open reading frame; however, based on the exact location of the variant and the direction of transcription, we were able to calculate the open reading frame and locate the nucleotide sequence corresponding to the peptide. To confirm the match, we translated the nucleotide sequence to amino acids and compared the peptide sequence to the reference and variant peptides. Each match was accounted for, resulting in a distribution of reference vs variant peptides across all cell lines (Fig. 3).

We investigated the population genomics information from the 1000 Genomes study³¹. Based on the plots of RNA-seq data, we identified four types of distributions (Fig. 3). A fraction of the variant RNAs were not present in any of the cell lines and were grouped together with the variants strongly dominated by the wild type (Fig. 3A). In other cases, the variants demonstrated a seemingly random distribution across cell lines (Fig. 3B). A third group of the variants that were confirmed in almost all of the cell lines, with no wild type present (Fig. 3C). The last subset was characterized by overexpression of variant peptides relative to the reference (Fig. 3D). Interestingly, a few percent of the variants reported in the study indicate that the “minor” allele is widely present in certain populations, making that designation questionable. For example, A variant of double-stranded RNA-specific adenosine deaminase (ADAR) K384R was validated in ~70% of the cell lines. This variant naturally occurs in 76% of the American population (1000 Genomes, rs2229857), clouding the significance of its finding in the GSC data.

Conclusions

Protein variants that are present in a majority of cell lines could be generally characteristic to cancers but are understudied. Many variants have a link to human disease, for instance, a germline variant of protein farnesyltransferase type-1 subunit alpha (FNTA) Gln184Glu was confirmed in >90% of the cell lines. It is an enzyme involved in the apoptotic cascade and germline variants reported to date lead to a decreased enzymatic activity. Novel large-scale protein variant detection may provide new insights and novel therapeutic targets for GBM, a devastating and deadly disease. Our novel approach yielded approximately 400 SAV

peptides in GSCs (0.0035% of all known germline SAVs in the human genome). While this is a small number compared to all known variants, we did not expect every finding to have relevance to GBM, or even cancer as a disease. In a recent proteogenomic analysis of Jurkat cells in which a custom database was created, over 1,000 SAV peptides and 463 SAV-containing phosphopeptides were identified, some with an altered kinase recognition motif³².

We performed an epidemiological analysis of public cancer datasets using our SAV data and identified ten protein variants in our data associated with the risk of dying of glioma at P-values <0.05 ²¹. Molecular epidemiology strategies, developed and applied to genomic data, will be informative to proteomic datasets as well^{33, 34}. The variant proteins with the lowest P-values were associated with GBM: PREX1 S1559T (0.005) and DNAAF5 V632A (0.016) and with LGG: NUP210 R786Q (0.006) and PCMT1 V143I (0.016). In the future we plan to investigate the mechanistic connection of the proteins by Crispr/Cas9 manipulation of GSC genomes followed by phenotypic studies in cell culture and xenograft implantation in nude mice³⁵ to either encourage or negate the variant proteins' applicability as adjuvant targets in glioma.

Methods

Study Materials

Isolation of GSCs from patient tumors at The University of Texas M.D. Anderson Cancer Center, Department of Neurosurgery, was performed as previously described⁴ in accordance with approval from institutional review board. GSCs were cultured according to a published method^{4, 36}. All cell lines were certified free of *Mycoplasma*. Identical cell culture batches were divided for transcriptomic and proteomic studies, eliminating batch variance in the comparative assays. The detailed method descriptions for extraction of proteins and RNA from GSCs, nano-LC/MS/MS of digested proteins, and data analysis have been previously published¹⁸.

Database design

The proprietary protein database of 2,226,998 SAV sequences (ProteoFinder v4, release 2014-06-10) was created in fasta file format³⁷ and combined with UniProtKB/SwissProt³⁸ human database (version 2014_07) and 58 common Repository of Adventitious Proteins (cRAP) contaminant sequences (<http://www.thegpm.org/crap/index.html>). The SAV sequences were generated from 40,191 protein entries of the entire human proteome (excluding titin) by altering single amino acids at known positions according to neXtProt (release 2014-05-27)³⁹. Duplicates originating from alternative splicing variants of the same protein entries were removed always keeping the first listed protein variants. The resulting SAV sequences were then shortened to reduce redundancy, keeping only that part of the protein sequence where the amino acid exchange took place surrounded by one additional tryptic peptide at both N- and C-termini. These entries were distinguished from the UniProtKB and cRAP sections of the database in the entry name by use of the ">NX_access code" syntax.

Database Searches and Peptide Assignment

Label-free MS/MS data from 36 GSC lines (ProteomeXchange identifier PXD000563) was obtained in an Orbitrap Elite (Thermo Fisher) in triplicate according to a published proteomics workflow described in Lichti et al¹⁸. The data were searched using PEAKS (version 7, Bioinformatics Solutions, Waterloo, ON⁴⁰) against the custom protein database (2,267,533 entries). Searches were performed with a parent ion tolerance of 10 ppm, fragment ion tolerance of 0.02 Da, fixed carbamidomethyl cysteine, and variable modifications of oxidation (M). Trypsin was specified as the enzyme, allowing for two missed cleavages and a maximum of three PTMs per peptide. FDR estimation was enabled.

The resulting peptide identification list was filtered based on the peptide-spectrum match score ($-10\log P > 30$), uniqueness and source (only peptides that matched variant database were kept). To ensure, that identified peptides were variant, they were analyzed by blastp (similarity matrix BLOSUM90) against the human proteome. Only peptides that matched their corresponding protein with one mismatch at the appropriate position were retained. Peptides that matched multiple homologues proteins, or redundant variants, were discarded.

Variant verification using RNAseq data

In order to verify peptide sequences, we matched them to the RNAseq data corresponding to each cell line, provided by M.D. Anderson Cancer Center. The non-variant peptide sequence was used to match to the reference genome (GRCh37/hg19) using PGx³⁰ (Fig. 1). The output file contained the chromosomal location of the target peptide; this information was applied to parse RNA-seq raw data (SAM tools) and count matching reads. The positioning of the variant in the peptide served as an anchor to calculate open reading frames. Each nucleotide sequence of the predicted area was then translated back to amino acids and compared with the peptide of interest (Fig. 3) resulting in a distribution of variant and wild-type read counts across all cell lines.

Parallel reaction monitoring validation

We ordered 211 heavy labeled synthetic peptides (PEPotec SRM library Grade 2) containing SAVs identified in the proteomic data (Thermo Fisher Scientific, Ulm, Germany, Suppl. Table 3). For optimization of the assay, peptides were mixed in equal volumes (10 μ l) and diluted with 0.1% formic acid (FA) in water to the final overall concentration of 50 fmol/ μ L. A small fraction of peptides (~30) were of poor quality and had to be analyzed at a higher concentration (500 fmol/ μ L). MS analysis was run on a Q Exactive mass spectrometer (Thermo Scientific, Bremen, Germany) equipped with a nano-ESI interface (Dream spray, AMR Inc., Tokyo) and connected to an Easy n-LC 1000 pump (Thermo Scientific, Waltham, MA). Peptides were injected onto an Acclaim PepMap 100 precolumn (100 μ m $\times$ 2 cm, Thermo Scientific, Waltham, MA), and then were separated on a reverse phase column (Zaplous α Pep-C18, 100 μ m $\times$ 20 cm, AMR Inc., Tokyo). Separation was performed in a 90-min linear gradient (5 to 35% acetonitrile containing 0.1% formic acid) at 500 nL/min. The PRM analysis was conducted with the following parameters: spray voltage 1.4 kV, transfer capillary temperature 300 $^{\circ}$ C, resolution 17,500 @ 200 m/z, AGC target value 1×10^6 , maximum injection time 50/100 ms, isolation window 2 Th, and normalized collision energy 27. GSC protein digests (33 samples, 100 μ g) were dried down and then resuspended

in 100 μ L of 0.1 % FA in water. Protein digests were spiked with the heavy peptide mix and diluted with 0.1 % FA in water to a final concentration 0.25 μ g/ μ L (digest) and 25 fmol/ μ L (peptide mix). PRM analysis was performed with the same parameters as described above using a scheduled acquisition method (7 min window) with < 30 consecutive precursor ions. The inclusion list was created using Skyline v3.1 (MacCoss Lab)⁴¹.

Quantification with selected reaction monitoring (SRM)

In total, 86 heavy labeled synthetic peptides with a C-terminal cleavable tag (SpikeTides) were ordered (JPT Peptide Technologies, Berlin, Germany, Suppl. Table 2) containing SAVs and their reference sequences. The peptides were mixed and digested with trypsin to remove the cleavable tag, and this peptide mixture (with and without background matrix) was used for optimization of the assay parameters. Sample analysis was performed on a TSQ Quantiva mass spectrometer (Thermo Scientific) equipped with an electrospray ion source (EASY-Spray NG) and connected to an EASY n-LC 1000 pump (Thermo Scientific, Waltham, MA). Peptides were loaded onto an Acclaim PepMap 100 precolumn (100 μ m $\times$ 2cm, Thermo Scientific, Waltham, MA), and then separated on an EASY-Spray column (15 cm $\times$ 75 μ m ID, PepMap C18 3 μ m, 100 Å) with the flow rate set to 300 nL/min and the column temperature to 35 °C. Solvent A (0.1% formic acid) and solvent B (0.1% formic acid in acetonitrile) were used to create a nonlinear gradient to elute the peptides. Separations were performed with a gradient of 5% to 22% B in 50 minutes, followed by 22% to 32% B in 10 minutes, and 32% to 85% B in 5 minutes and finished by holding 85% B for 8 minutes. SRM transitions were acquired in Q1 and Q3 operated at unit resolution (0.7 FWHM); the collision gas pressure in Q2 was set to 1.5 mTorr. The cycle time was 2 s, and calibrated RF and S-lens values were used. GSC protein digests were spiked with heavy peptide mixture and analyzed with the above described parameters using 5-min detection windows in scheduled SRM mode.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding. D.F. was supported by the National Cancer Institute (NCI) CPTAC award U24 CA210972, contract 13XS068 from Leidos Biomedical Research, Inc., and by a grant from the Shifrin-Myers Breast Cancer Discovery Fund. F.F.L. and E.P.S. were supported by MD Anderson Brain SPORE P50-CA127001, E.P.S. was supported by R01-CA1902, and E.P.S. and Q. W. were supported by the National Brain Tumor Society Defeat GBM project. C.L.N. acknowledges the support of the University of Texas Medical Branch, The University of Texas M. D. Anderson Cancer Center, The Cancer Prevention Research Institute of Texas (RML1122), the Center of Excellence in Biological and Medical Mass Spectrometry, and Lund University.

References

1. Hegi ME, Diserens A-C, Gorlia T, Hamou M-F, de Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L, Bromberg JEC, Hau P, Mirimanoff RO, Cairncross JG, Janzer RC, Stupp R. MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *New England Journal of Medicine*. 2005; 352:997–1003. [PubMed: 15758010]
2. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJB, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff R-

- O. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The Lancet Oncology*. 2009; 10:459–466. [PubMed: 19269895]
3. Fidoamore A, Cristiano L, Antonosante A, d'Angelo M, Di Giacomo E, Astarita C, Giordano A, Ippoliti R, Benedetti E, Cimini A. Glioblastoma Stem Cells Microenvironment: The Paracrine Roles of the Niche in Drug and Radioresistance. *Stem Cells International*. 2016; 2016:6809105. [PubMed: 26880981]
 4. Galli R, Binda E, Orfanelli U, Cipelletti B, Gritti A, De Vitis S, Fiocco R, Foroni C, Dimeco F, Vescovi A. Isolation and Characterization of Tumorigenic, Stem-like Neural Precursors from Human Glioblastoma. *Cancer Research*. 2004; 64:7011–7021. [PubMed: 15466194]
 5. He H, Nilsson CL, Emmett MR, Marshall AG, Kroes RA, Moskal JR, Ji Y, Colman H, Priebe W, Lang FF, Conrad CA. Glycomic and transcriptomic response of GSC11 glioblastoma stem cells to STAT3 phosphorylation inhibition and serum-induced differentiation. *J Proteome Res*. 2010; 9:2098–2108. [PubMed: 20199106]
 6. Nilsson CL, Dillon R, Devakumar A, Rogers JC, Krastins B, Rosenblatt MM, Majo M, Kaboord BJ, Sarracino D, Rezai T, Prakash A, Lopez M, Ji Y, Priebe W, Colman H, Lang FF, Conrad CA. Quantitative phosphoproteomic analysis of STAT3/IL-6/HIF1 α signaling network: An initial study in GSC11 glioblastoma stem cells. *J. Proteome Res*. 2010; 9:430–443. [PubMed: 19899826]
 7. Sharma A, Shiras A. Cancer stem cell-vascular endothelial cell interactions in glioblastoma. *Biochemical and Biophysical Research Communications*. 2016; 473:688–692. [PubMed: 26692486]
 8. Tamura K, Aoyagi M, Ando N, Ogishima T, Wakimoto H, Yamamoto M, Ohno K. Expansion of CD133-positive glioma cells in recurrent de novo glioblastomas after radiotherapy and chemotherapy. *Journal of Neurosurgery*. 2013; 119:1145–1155. [PubMed: 23991844]
 9. Monroe CL, Dahiya S, Gutmann DH. Dissecting Clinical Heterogeneity in Neurofibromatosis Type 1. *Annual Review of Pathology: Mechanisms of Disease*. 2017; 12:53–74.
 10. Andersson U, Wibom C, Cederquist K, Aradottir S, Borg Å, Armstrong GN, Shete S, Lau CC, Bainbridge MN, Claus EB, Barnholtz-Sloan J, Lai R, Il'yasova D, Houlston RS, Schildkraut J, Bernstein JL, Olson SH, Jenkins RB, Lachance DH, Wrensch M, Davis FG, Merrell R, Johansen C, Sadetzki S, Bondy ML, Melin BS, Adatto P, Morice F, Payen S, McQuinn L, McGaha R, Guerra S, Paith L, Roth K, Zeng D, Zhang H, Yung A, Aldape K, Gilbert M, Weinberger J, Colman H, Conrad C, de Groot J, Forman A, Groves M, Levin V, Loghin M, Puduvalli V, Sawaya R, Heimberger A, Lang F, Levine N, Tolentino L, Saunders K, Thach T-T, Iacono DD, Sloan A, Gerson S, Selman W, Bambakidis N, Hart D, Miller J, Hoffer A, Cohen M, Rogers L, Nock CJ, Wolinsky Y, Devine K, Fulop J, Barrett W, Shimmel K, Ostrom Q, Barnett G, Rosenfeld S, Vogelbaum M, Weil R, Ahluwalia M, Peereboom D, Staugaitis S, Schilero C, Brewer C, Smolenski K, McGraw M, Naska T, Rosenfeld S, Ram Z, Blumenthal DT, Bokstein F, Umansky F, Zaaroor M, a Cohen A, Tzuk-Shina T, Voldby B, Laursen R, Andersen C, Brennum J, Henriksen MB, Marzouk M, Davis ME, Boland E, Smith M, Eze O, Way M, Lada P, Miedzianowski N, Frechette M, Paleologos N, Bystrom G, Svedberg E, Huggert S, Kimdal M, Sandström M, Brännström N, Hayat A, Tihan T, Zheng S, Berger M, Butowski N, Chang S, Clarke J, Prados M, Rice T, Sison J, Kivett V, Duo X, Hansen H, Hsuang G, Lamela R, Ramos C, Patoka J, Wagenman K, Zhou M, Klein A, McGee N, Pfefferle J, Wilson C, Morris P, Hughes M, Britt-Williams M, Foft J, Madsen J, Polony C, McCarthy B, Zahora C, Villano J, Engelhard H, Borg A, Chanock SK, Collins P, Elston R, Kleihues P, Kruchko C, Petersen G, Plon S, Thompson P, Johansen C, Sadetzki S, Melin B, Bondy ML, Lau CC, Scheurer ME, Armstrong GN, Liu Y, Shete S, Yu RK, Aldape KD, Gilbert MR, Weinberg J, Houlston RS, Hosking FJ, Robertson L, Papaemmanuil E, Claus EB, Claus EB, Barnholtz-Sloan J, Sloan AE, Barnett G, Devine K, Wolinsky Y, Lai R, McKean-Cowdin R, Il'yasova D, Schildkraut J, Sadetzki S, Yechezkel GH, Bruchim RB-S, Aslanov L, Sadetzki S, Johansen C, Kosteljanetz M, Broholm H, Bernstein JL, Olson SH, Schubert E, DeAngelis L, Jenkins RB, Yang P, Ryneerson A, Andersson U, Wibom C, Henriksson R, Melin BS, Cederquist K, Aradottir S, Borg Å, Merrell R, Lada P, Wrensch M, Wiencke J, Wiemels J, McCoy L, McCarthy BJ, Davis FG. Germline rearrangements in families with strong family history of glioma and malignant melanoma, colon, and breast cancer. *Neuro-Oncology*. 2014; 16:1333–1340. [PubMed: 24723567]
 11. Bainbridge MN, Armstrong GN, Gramatges MM, Bertuch AA, Jhangiani SN, Doddapaneni H, Lewis L, Tombrello J, Tsavachidis S, Liu Y, Jalali A, Plon SE, Lau CC, Parsons DW, Claus EB,

- Barnholtz-Sloan J, Il'yasova D, Schildkraut J, Ali-Osman F, Sadetzki S, Johansen C, Houlston RS, Jenkins RB, Lachance D, Olson SH, Bernstein JL, Merrell RT, Wensch MR, Walsh KM, Davis FG, Lai R, Shete S, Aldape K, Amos CI, Thompson PA, Muzny DM, Gibbs RA, Melin BS, Bondy ML. Germline Mutations in Shelterin Complex Genes Are Associated With Familial Glioma. *JNCI: Journal of the National Cancer Institute*. 2015; 107:dju384–dju384. [PubMed: 25482530]
12. Jalali A, Amirian ES, Bainbridge MN, Armstrong GN, Liu Y, Tsavachidis S, Jhangiani SN, Plon SE, Lau CC, Claus EB, Barnholtz-Sloan JS, Il'yasova D, Schildkraut J, Ali-Osman F, Sadetzki S, Johansen C, Houlston RS, Jenkins RB, Lachance D, Olson SH, Bernstein JL, Merrell RT, Wensch MR, Davis FG, Lai R, Shete S, Aldape K, Amos CI, Muzny DM, Gibbs RA, Melin BS, Bondy ML. Targeted Sequencing in Chromosome 17q Linkage Region Identifies Familial Glioma Candidates in the Gliogene Consortium. *Scientific Reports*. 2015; 5:8278. [PubMed: 25652157]
 13. Maire CL, Ligon KL. Molecular pathologic diagnosis of epidermal growth factor receptor. *Neuro-Oncology*. 2014; 16:viii1–viii6. [PubMed: 25342599]
 14. Endersby R, Baker SJ. PTEN signaling in brain: neuropathology and tumorigenesis. *Oncogene*. 2008; 27:5416. [PubMed: 18794877]
 15. Gusyatiner O, Hegi ME. Glioma epigenetics: From subclassification to novel treatment options. *Seminars in Cancer Biology*. 2017
 16. Masui K, Cavenee WK, Mischel PS. Cancer metabolism as a central driving force of glioma pathogenesis. *Brain Tumor Pathology*. 2016; 33:161–168. [PubMed: 27295313]
 17. Paik Y-K, Jeong S-K, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee H-J, Na K, Choi E-Y, Yan F, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon J-Y, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee E-Y, Hochstrasser D, Legrain P, Hancock WS. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotech*. 2012; 30:221–223.
 18. Lichti CF, Liu H, Shavkunov AS, Mostovenko E, Sulman EP, Ezhilarasan R, Wang Q, Kroes RA, Moskal JC, Fenyö D, Oksuz BA, Conrad CA, Lang FF, Berven FS, Végvári Á, Rezeli M, Marko-Varga G, Hober S, Nilsson CL. Integrated Chromosome 19 Transcriptomic and Proteomic Data Sets Derived from Glioma Cancer Stem-Cell Lines. *Journal of Proteome Research*. 2014; 13:191–199. [PubMed: 24266786]
 19. Lichti CF, Mostovenko E, Wadsworth PA, Lynch GC, Pettitt BM, Sulman EP, Wang Q, Lang FF, Rezeli M, Marko-Varga G, Végvári Á, Nilsson CL. Systematic Identification of Single Amino Acid Variants in Glioma Stem-Cell-Derived Chromosome 19 Proteins. *Journal of Proteome Research*. 2015; 14:778–786. [PubMed: 25399873]
 20. Nilsson CL, Berven F, Selheim F, Liu H, Moskal JR, Kroes RA, Sulman EP, Conrad CA, Lang FF, Andrén PE, Nilsson A, Carlsohn E, Lilja H, Malm J, Fenyö D, Subramaniam D, Wang X, Gonzales-Gonzales M, Dasilva N, Diez P, Fuentes M, Végvári Á, Sjödin K, Welinder C, Laurell T, Fehniger TE, Lindberg H, Rezeli M, Edula G, Hober S, Marko-Varga G. Chromosome 19 Annotations with Disease Speciation: A First Report from the Global Research Consortium. *Journal of Proteome Research*. 2012; 12:135–150. [PubMed: 23249167]
 21. Mostovenko E, Liu Y, Amirian ES, Tsavachidis S, Armstrong GN, Bondy ML, Nilsson CL. Combined Proteomic-Molecular Epidemiology Approach to Identify Precision Targets in Brain Cancer. *ACS Chemical Neuroscience*. 2017
 22. Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, Fenyö D, Zhang B, Mani DR. Methods, Tools and Current Perspectives in Proteogenomics. *Molecular & Cellular Proteomics*. 2017; 16:959–981. [PubMed: 28456751]
 23. Scifo E, Calza G, Fuhrmann M, Soliymani R, Baumann M, Lalowski M. Recent advances in applying mass spectrometry and systems biology to determine brain dynamics. *Expert Review of Proteomics*. 2017; 14:545–559. [PubMed: 28539064]
 24. Lane L, Bairoch A, Beavis RC, Deutsch EW, Gaudet P, Lundberg E, Omenn GS. Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. *Journal of Proteome Research*. 2013; 13:15–20. [PubMed: 24364385]
 25. Nilsson CL, Mostovenko E, Lichti CF, Ruggles K, Fenyö D, Rosenbloom KR, Hancock WS, Paik Y-K, Omenn GS, LaBaer J, Kroes RA, Uhlén M, Hober S, Végvári Á, Andrén PE, Sulman EP, Lang FF, Fuentes M, Carlsohn E, Emmett MR, Moskal JR, Berven FS, Fehniger TE, Marko-Varga

- G. Use of ENCODE Resources to Characterize Novel Proteoforms and Missing Proteins in the Human Proteome. *Journal of Proteome Research*. 2015; 14:603–608. [PubMed: 25369122]
26. Valencia OM, Samuel SE, Viscusi RK, Riall TS, Neumayer LA, Aziz H. The role of genetic testing in patients with breast cancer: A review. *JAMA Surgery*. 2017; 152:589–594. [PubMed: 28423155]
27. Anderson LC, Hakansson M, Walse B, Nilsson CL. Combined Front-End ETD and CID Tandem Mass Spectrometric Analysis at 21 Tesla and X-Ray Crystallography Define Structural Differences in Single Amino Acid Variants of Human Mitochondrial Branched-Chain Amino Acid Aminotransferase 2 (BCAT2). *J. Am. Soc. Mass Spectrom.* 2017
28. Anderson LC, Håkansson M, Walse B, Nilsson CL. Intact Protein Analysis at 21 Tesla and X-Ray Crystallography Define Structural Differences in Single Amino Acid Variants of Human Mitochondrial Branched-Chain Amino Acid Aminotransferase 2 (BCAT2). *Journal of The American Society for Mass Spectrometry*. 2017
29. Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. *BMC Bioinformatics*. 2012; 13:S2.
30. Askenazi M, Ruggles KV, Fenyö D. PGx: Putting Peptides to BED. *Journal of Proteome Research*. 2016; 15:795–799. [PubMed: 26638927]
31. The Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015; 526:68–74. [PubMed: 26432245]
32. Ma S, Menon R, Poulos RC, Wong JWH. Proteogenomic analysis prioritises functional single nucleotide variants in cancer samples. *Oncotarget*. 2017; 8:95841–95852. [PubMed: 29221171]
33. Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, Lena PD, Casadio R, Edwards M, Gifford D, Jones DT, Sundaram L, Bhat R, Li X, Pal LR, Kundu K, Yin Y, Moulton J, Jiang Y, Pejaver V, Pagel KA, Li B, Mooney SD, Radivojac P, Shah S, Carraro M, Gasparini A, Leonardi E, Giollo M, Ferrari C, Tosatto SCE, Bachar E, Azaria JR, Ofra Y, Unger R, Niroula A, Vihinen M, Chang B, Wang MH, Franke A, Petersen B-S, Pirooznia M, Zandi P, McCombie R, Potash JB, Altman RB, Klein TE, Hoskins RA, Repo S, Brenner SE, Morgan AA. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*. n/a-n/a.
34. Schaafsma GCP, Vihinen M. Large differences in proportions of harmful and benign amino acid substitutions between proteins and diseases. *Human Mutation*. 2017; 38:839–848. [PubMed: 28444810]
35. Sumeer, Lal, Lacroix, Michel, Tofilon, Philip, Fuller, Gregory N., Sawaya, Raymond, Lang, Frederick F. An implantable guide-screw system for brain tumor studies in small animals. *Journal of Neurosurgery*. 2000; 92:326–333. [PubMed: 10659021]
36. Jiang H, Gomez-Manzano C, Aoki H, Kondo S, McCormick F, Xu J, Kondo Y, Bekele BN, Colman H, Lang FF, Fucyo J. Examination of the therapeutic potential of Delta-24-RGD in brain tumor stem cells: role of autophagic cell death. *J. Natl. Cancer Inst.* 2007; 99:1410–1414. [PubMed: 17848677]
37. Vegvari, A. Mutant Proteogenomics. In: Vegvari, A., editor. *Proteogenomics*. Springer Verlag; 2016. p. 77-91.
38. UniProt: a hub for protein information. *Nucleic Acids Research*. 2015; 43:D204–D212. [PubMed: 25348405]
39. Gaudet P, Michel P-A, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, Duek PD, Gateau A, Gleizes A, Hinard V, Rech de Laval V, Lin J, Nikitin F, Schaeffer M, Teixeira D, Lane L, Bairoch A. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Research*. 2017; 45:D177–D182. [PubMed: 27899619]
40. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics*. 2012; 11
41. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics (Oxford, England)*. 2010; 26:966–968.

Figure 1.

Schematic representation of the variant validation workflow, in which variant peptide sequences were identified by MS/MS and searches of a custom database, then verified by matching to RNAseq data. Comparison of each nucleotide sequence corresponding to a matched peptide sequence resulted in a distribution of variant and wild-type read counts across all cell lines.

Figure 2.

The presence of SAV- containing peptides were validated by PRM. Results for cell samples containing variant cytoplasmic C-1-tetrahydrofolate synthase (MTHFD1) R653Q, peptide sequence IA HGNSSIIADQIALK, are shown below. Top: PRM traces showing retention time (X-axis, minutes) and signal intensity (Y-axis, arbitrary units). Bottom: Quantitative ratio of native peptide to spiked in heavy standard peptide.

Figure 3.

An analysis of RNA-seq count distributions of variant and non-variant transcripts across all cell lines revealed four patterns: Strong expression of the wild-type (A), a mix of non-variant and variant (B), variant expression exclusively (C), or relatively higher expression of the variant compared to the non-variant transcript. The X-axis lists the cell line and the Y-axis shows relative amounts of transcript.