Journal of proteome research

Article

pubs.acs.org/jpr

Systematic Identification of Single Amino Acid Variants in Glioma Stem-Cell-Derived Chromosome 19 Proteins

Cheryl F. Lichti,[†,▽] Ekaterina Mostovenko,[†,▽] Paul A. Wadsworth,[†] Gillian C. Lynch,[‡] B. Montgomery Pettitt,[‡] Erik P. Sulman,[§] Qianghu Wang,[∥] Frederick F. Lang,[⊥] Melinda Rezeli,[¶] György Marko-Varga,[¶] Ákos Végvári,*[,¶] and Carol L. Nilsson*[,†]

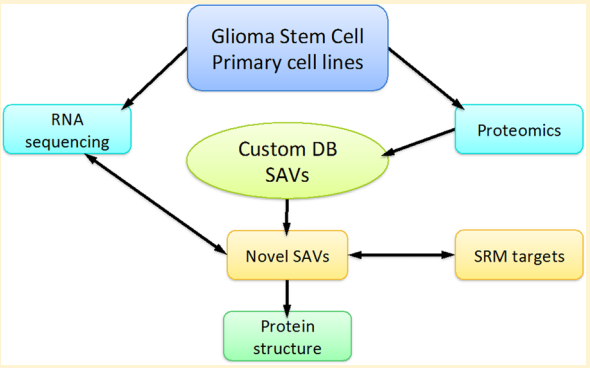
[†]Department of Pharmacology and Toxicology and [‡]Biochemistry and Molecular Biology, UTMB Cancer Center, University of Texas Medical Branch, Galveston, Texas 77555, United States

[§]Department of Radiation Oncology, [∥]Department of Bioinformatics, and [⊥]Department of Neurosurgery, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, United States

[¶]Clinical Protein Science & Imaging, Biomedical Center, Department of Biomedical Engineering, Lund University, 221 84 Lund, Sweden

Supporting Information

ABSTRACT: Novel proteoforms with single amino acid variations represent proteins that often have altered biological functions but are less explored in the human proteome. We have developed an approach, searching high quality shotgun proteomic data against an extended protein database, to identify expressed mutant proteoforms in glioma stem cell (GSC) lines. The systematic search of MS/MS spectra using PEAKS 7.0 as the search engine has recognized 17 chromosome 19 proteins in GSCs with altered amino acid sequences. The results were further verified by manual spectral examination, validating 19 proteoforms. One of the novel findings, a mutant form of branched-chain aminotransferase 2 (*p.Thr186Arg*), was verified at the transcript level and by targeted proteomics in several glioma stem cell lines. The structure of this proteoform was examined by molecular modeling in order to estimate conformational changes due to mutation that might lead to functional modifications potentially linked to glioma. Based on our initial findings, we believe that our approach presented could contribute to construct a more complete map of the human functional proteome.



KEYWORDS: *Chromosome-Centric Human Proteome Project, chromosome 19, mass spectrometry, missense single nucleotide variants, single amino acid variants, glioma stem cells, cancer proteomics, bioinformatics, SRM assay, BCAT2 p.Thr186Arg*

1. INTRODUCTION

One important goal of the Chromosome-Centric Human Proteome Project (C-HPP), an international consortium, is to completely map the human proteome by identification of proteins in selected tissues and cells.[1,2] As an articulated goal, the C-HPP is also determined to capture biological features of gene variation, gene regulation, and protein expression mapped by chromosome localization. Consequently, all C-HPP projects are generating and reporting protein data in a format that is aligned with the DNA sequence of individual chromosomes and with the output of transcriptome data (RNA sequencing). In addition to sequential data derived from the most frequent proteoforms (consensus or canonical sequences), it is desirable to characterize additional major proteoforms, such as alternative splicing transcript (AST), single amino acid variants (SAV), and post-translational modifications (PTMs). A complementary HPP activity, the Biology/Disease-driven HPP (B/D-HPP), is focused on generation of knowledge from studies of cellular mechanisms and biochemical processes, analyzing proteomes associated with human diseases.[3] The results are expected to facilitate routine determinations of processes and disease relevant proteins in life science research.

At present, human protein sequence data is collected at a rapid pace, predominantly due to the technological advances in mass spectrometry (MS). However, roughly 10% is still missing, due to the lack of quality observations of certain proteins, incorrect gene annotation, very low abundance or absence of expression in most tissues, or unfavorable structure (or cleavage sites) for bottom-up MS studies. On the other hand, two recent publications have set new milestones in providing the most complete draft of the human proteome to date.[4,5] According to the Proteomics DB (http://www.proteomicsdb.org), the current state of the chromosome 19 is

Received: August 1, 2014
Published: November 17, 2014

at 96.4% completion (1352 genes and 1304 proteins) and includes details about a high number of ASTs ("isoforms").[5] Nevertheless, if all proteoforms with different sequences that have biological functions are considered, then the exact size of the human proteome is still unknown today and may reach extremely high numbers, up to several million.[2,6] Because the HPP's directive is to identify at least one AST and one SAV of each consensus proteins as well as three major PTMs (phosphorylation, glycosylation and acetylation),[7] the number of proteoform entries in the complete human proteome is estimated in the range of 100 000 to 1 000 000.

The latest downloadable version of the neXtProt database (http://www.nextprot.org) includes 20 055 protein entries (released on 19 September 2014). There are 1430 genes and 1426 protein entries reported for chromosome 19, which are identified at the protein level (1129 entries), transcript level (248 entries), uncertain (36 entries), homology (10 entries), and predicted (7 entries). Although this database fasta file does not include any information about SAVs (nor do the UniProt databases), references to mutant proteoforms are listed on the neXtProt Web site when certain proteins are selected. However, the level of identification for these SAVs is not indicated otherwise. Genomic databases are typically more detailed, providing a list of missense single nucleotide variations (missense SNVs) of each human gene. Databases such as 1000 Genomes (http://www.1000genomes.org) provide population-based frequency information as well. Other genomic databases such as the Catalogue of Somatic Mutations in Cancer (COSMIC at http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) have accumulated gene expression variants and revealed disease relation.

Genome investigations have identified a large number of gene mutations, including missense SNVs[8] that are currently used for risk analysis of cancer, for example, *BRCA1* and *BRCA2*. However, it is important to underline that not the genes but the proteins have biological function and are the working units of cell machinery. Consequently, their expression patterns in both qualitative and quantitative aspects should be characterized when the altered biology of diseases are in the focus of study. Despite this fact, the majority of protein identifications are based upon MS/MS data from shotgun proteomics experiments searched against protein databases such as UniProtKB and neXtProt, which are typically restricted to proteoforms with consensus and AST sequences only. Thus, direct searches of SAVs cannot be performed with tandem spectra and such data is seldom reported, despite the fact that SAVs may play important role in disease biology.[9]

Some search engines are constructed by use of various approaches and algorithms in order to provide tools to recognize SAVs. As an example, PEAKS (Bioinformatics Solutions Inc., Waterloo, ON, Canada) offers an algorithm (SPIDER)[10] that can utilize the match of peptide spectrum matches (PSMs) with replaced amino acids. Others, like the freely available X!Tandem (http://www.thegpm.org) employs another strategy that systematically changes each residue in a peptide for all other possible amino acids and score the mutated peptide (and all potential modifications) against all of the available MS/MS spectra. Therefore, bioinformatic tools do exist that aid in the process of identifying SAVs in shotgun proteomics data, and these tools can prove to be valuable in the characterization of SAVs associated with a particular disease.

Glioblastomas are among the deadliest of cancers; fewer than 10% of patients survive five years after diagnosis. Even with standard-of-care treatments, tumors nearly always recur. Recurrence is at least in part due the existence of glioma stem cells (GSCs), which are resistant to radiation and chemotherapy. We have acquired comprehensive, quantitative data sets from thirty-six GSC lines, including gene activity and protein expression. We have previously examined patterns of global as well as chromosome-19-specific protein and gene expression in a subset of the 36 GSC lines.[11,12] Our hypothesis is that GSCs, as a rare type of diseased cell type, harbor protein SAVs due to germline or somatic mutations. Those proteins may normally have cell protective properties, but proteins with amino acid substitutions have the potential to be transformed into promoters of genetic instability or invasivity, as well as ineffective regulators of epigenetic or metabolic control. Through the use of a customized database and bioinformatic tools, we provide the first comparative data of chromosome 19 SNP products on 36 glioma stem cell lines and compelling evidence for the expression of a SAV in branched-chain aminotransferase 2 (BCAT2), a protein encoded by chromosome 19, in six of the GSC lines.

We have devised a systematic workflow to address the second level of protein expression complexity (represented by the SAVs) and have applied this workflow in the identification of missense SNV products at the protein level in glioma samples based on high resolution MS/MS data. We used a custom protein database to widen the search window for mutant proteoforms. We have identified and further verified novel mutant proteoforms that might be strongly associated with glioma.

2. EXPERIMENTAL SECTION

2.1. Samples and LC−MS/MS

Isolation of GSCs from patient tumors was performed as previously described[13] in accordance with the institutional review board of The University of Texas M.D. Anderson Cancer Center and are named in the order that they were acquired. GSCs were cultured according to a published method.[13,14] All cell lines were tested to exclude the presence of *Mycoplasma* infection. Downstream proteomic analyses were performed on identical cell culture batches in order to reduce the influence of batch variance in the comparative assays.

Cell lysates from 2×10^6 GSCs were reduced, alkylated, and analyzed in triplicate by LC−MS/MS on an Orbitrap Elite equipped with an Easy nanoLC 1000 pump (Thermo Fisher Scientific, Waltham, MA) as described in a previous publication.[11] Briefly, peptide mixtures were separated on a C18 column (ProteoPep II, New Objective, 10 cm × 75 μm) using a 240 min gradient (Solvent A, 0.1% formic acid in water; Solvent B, 0.1% formic acid in acetonitrile). Data were acquired using high-resolution data-driven analysis (DDA), with the survey scan (MS) acquired in the Orbitrap at 60 000 resolution (at m/z 400) in profile mode. The survey scan was followed by ten HCD MS/MS spectra, acquired in centroid mode at 15 000 resolution in the Orbitrap.

2.2. Protein Identification Strategy

For database searching the technical replicates were combined and searched against a combination of the UniProtKB/SwissProt-Human database (2014_06 version, 40 548 protein entries) with all known chromosome 19 SAV sequences (132 264) together with 115 sequences of the common Repository of Adventitious Proteins (cRAP) contaminant database (http://www.thegpm.org/crap/index.html). Searches

dx.doi.org/10.1021/pr500810g | J. Proteome Res. 2015, 14, 778−786

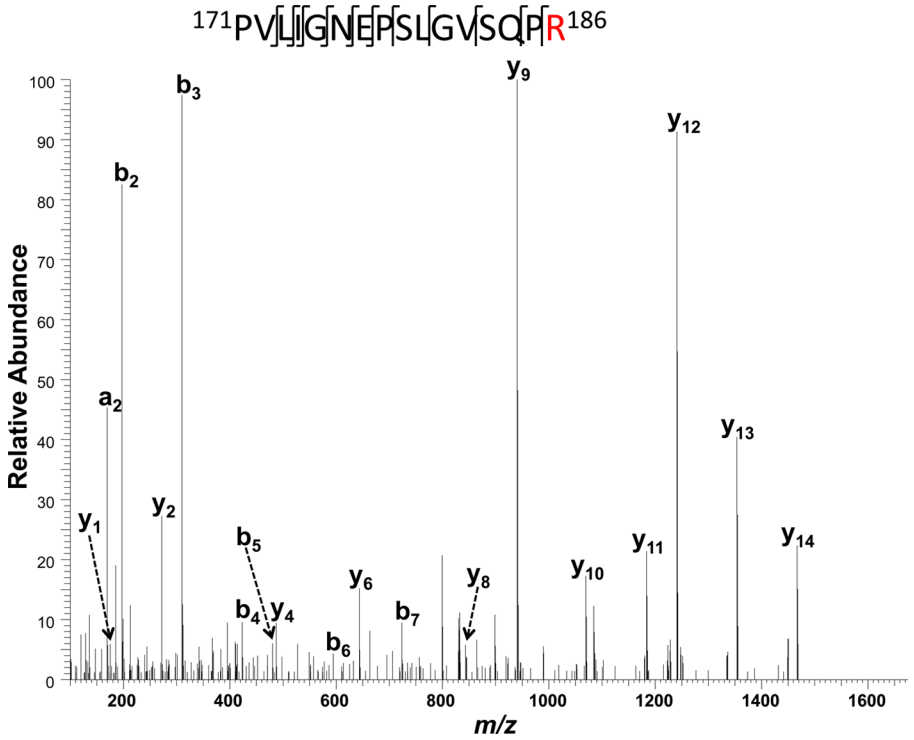


Figure 1. HCD-MS/MS spectrum of ^{178}PVLIGNEPSLGVSQP**R**^{186}, supporting assignment of the *p.Thr186Arg* proteoform of BCAT2. A complete list of theoretical and observed *m*/*z* values for ions observed in this spectrum can be found in Supporting Information Table 4.

were performed using PEAKS 7.0 (Bioinformatics Solutions) with 10 ppm parent mass error tolerance and 0.025 fragment mass tolerance, allowing for a maximum of two missed cleavages and one nonspecific cleavage. Carbamidomethyl cysteine was set as a fixed modification, and oxidation (M) and phosphorylation (STY) were set as variable modifications. FDR estimation was enabled.

Peptides assigned as SAVs with a −10log *P* score >30 were selected for further validation. Homology searching was performed using the BLAST utility (www.uniprot.org) against the UniProt-Human database in order to confirm that each peptide sequence was unique. For those peptides that passed these initial filters, manual verification of MS/MS spectral assignments of SAV peptides was performed by comparing *m*/*z* values for ions observed in the MS/MS spectrum with a theoretical *m*/*z* list generated using the MS-Product utility on the Protein Prospector Web site (http://prospector.ucsf.edu).

2.3. Selected Reaction Monitoring Verification Assay

For optimization of the assay, isotopically labeled peptides were mixed and diluted with 5% acetonitrile (ACN) at a concentration of 3−25 pmol/μL for each synthetic peptide. The mixture was analyzed by nanoLC−MS/MS using a TSQ Vantage triple quadrupole mass spectrometer equipped with an Easy n-LC II pump (Thermo Scientific, Waltham, MA). Samples (2 μL) were injected onto an Easy C18-A1 precolumn (Thermo Scientific, Waltham, MA), and following online desalting and trapping at a pressure of 280 bar, the peptides were separated on a 75 μm × 150 mm fused silica column packed with ReproSil C18 (3 μm, 120 Å from Dr. Maisch GmbH, Germany). Separations were performed in a 55 min linear gradient from 5 to 40% ACN containing 0.1% formic acid at the flow rate of 300 nL/min. The MS analysis was conducted in positive ion mode at 1750 V applied spray voltage. The transfer capillary temperature was 270 °C and tuned S-lens value was used. Selected reaction monitoring (SRM) transitions were acquired in Q1 and Q3 operated at unit resolution (0.7 fwhm), the collision gas pressure in Q2 was set to 1.2 mTorr. Scheduled method was used for data acquisition with 4 min time windows and the cycle time was set to 1.5 s, whereas the maximum number of consecutive transitions was 50.

The SRM assay optimization was done with the aid of Skyline v2.5.0.6079 software (MacCoss Lab). Primarily, high numbers of transitions, including b- and y-ion series, were chosen for each peptide at both 2+ and 3+ charge states. The five transitions, which produced the most abundant signals without observed interferences in the glioma samples, were selected for further analysis and utilized for unambiguous identification.

For the sample analysis, the same chromatographic conditions were used as described above for the assay development. Identical SRM parameters were used for the heavy and native forms of each peptide. All raw data generated on the triple quadrupole mass spectrometer were imported to Skyline for data analysis. The peak integration was done automatically using Savitzky-Golay smoothing. All data were manually inspected to confirm the correct peak detection.

2.4. Transcriptome Analysis

Paired-end whole transcriptome sequencing was performed on the Illumina HiSeq platform after random priming and rRNA reduction. Transcript reads were aligned to human reference transcriptome (ENSEMBL version 64) using PRADA.[15] Downstream data analyses including variant calling were performed using Burroughs-Wheeler alignment, Samtools, and Genome Analysis Toolkit. More details on transcriptomic data acquisition and analysis are available in Lichti et al.[11]

2.5. Structure Preparation

The BCAT2 SAV was modeled using a combination of two crystal structures: the substrate, L-isoleucine covalently bound to the cofactor pyridoxamine phosphate (PMP), bound form of the enzyme (PDB code: 1KT8), and the PMP bound form (PDB code: 1KTA).[16] Both were downloaded from www.pdb.org and prepared using the Visual Molecular Dynamics (VMD) software.[17] As 1KT8 is missing residues near the mutation site, it was used as a reference structure onto which the coordinates of 1KTA were superimposed in order to include the coordinates of residues 175 to 191. Crystallographic waters, glycerol, acetic acid, and the substrate were removed. Residue 186 (please, note that our residue numbering scheme in this paper follows the recommendations of Human Genome Variation Society and thus may differ by six from other authors) was then mutated from threonine to arginine, and hydrogens were added.

The resulting coordinates were stored and from them an additional starting structure was generated in which the Arg186 backbone ψ dihedral was rotated from +132° to −60° by rotating parts of both Arg186 and Pro187, while φ was held constant at −80. After rotation, residues 185−188 were minimized with the remainder of the protein fixed. Both these final coordinates and the original unrotated structure were then minimized, holding all atoms of the protein fixed except for those in residues 175−191. The final coordinates of each were stored and used as the basis for side chain rotamer generation.

2.6. Side Chain Rotamer Generation and Minimization

For both the original and rotated ψ backbone structures, the side chain angles χ_1 and χ_2 of Arg186 were subsequently rotated in six increments of 60°, yielding a total of 72 structures. From the original structure, the side chain was rotated about the χ_1 bond without altering the original χ_2 torsion, generating an additional six structures.

Two independent minimizations were performed on each of the 78 structures: one in which the entire system was minimized, and one in which only residues 177−188 were minimized while all else was held fixed. The final coordinates of each minimization were used for electrostatic free energy calculations and structure analysis.

2.7. Electrostatic Free Energy Calculation

Electrostatic free energy calculations were conducted on the first monomer (chain A) of each of the 156 BCAT2 structures. The PDB 2PQR software[18] was used to generate PQR files for each structure. Here, the CHARMM27 parameter set was used to obtain partial atomic charges and atomic radii, and a pH of 7.0 was used in assigning protonation states for titratable residues. The total electrostatic free energy of the system was calculated using the APBS software package.[19] The structures of each minimization set were ranked by their corresponding energies and the best ten structures of both sets were analyzed using VMD.

3. RESULTS AND DISCUSSION

3.1. Identification of SAVs in Glioma Stem Cell Lines

We addressed the determination of mutant proteoforms in biological samples, a high priority within the C-HPP. A novel systematic approach was developed that is based on searching shotgun proteome tandem spectra, collected via data dependent analysis, against a customized protein sequence database containing chromosome 19 SAVs. The new database used for identification of mutant proteoforms was compiled with all human consensus proteins together with a new set of missense SNV-derived sequences of chromosome 19. To improve search quality, the frequently used cRAP contaminant protein database was also included, resulting in a database of 172 927 protein sequences in total.

When search space is confined to short sequences derived from only one chromosome, a problem is expected that is similar to one observed when a search is performed against a database for a partially sequenced organism. In a standard target-decoy approach, a number of high quality spectra will be wrongly matched to decoy sequences simply due to the absence of correct peptide sequences in the target database. This makes adequate significance estimation, often represented in proteomics as global false discovery rate, rather challenging. Because many proteins exist in multiple isoforms, for example, structural proteins such as myosin or tubulin, there is a high risk that a spectrum may be also erroneously assigned to a SAV-containing peptide. Those issues can be addressed by creating a custom decoy database,[20] a task that is not trivial in the case of short, low complexity sequences. Alternatively, protein sequences from the other homologous species, or in our case the rest of the canonical proteome from the same species, could be appended to a database of interest, adding more possibilities for the correct peptide-spectrum matches in a target data set and, therefore, enabling target-decoy approach implemented by standard search engines. Therefore, we used the modified decoy database generated by PEAKS.[21]

Because protein databases for shotgun proteomic searches do not commonly include SAV sequences, and different search engines may perform differently with the same database; three search engines (SEQUEST, PEAKS, and Mascot) were employed and the corresponding search results were compared. Triplicate data files for each cell line were used as a single input for database searching, initially performed with nearly identical search parameters.

An initial search and validation on a subset of the GSCs revealed several SAVs. Manual verification of spectra containing SAVs led to confirmation of a T186R substitution of branched-chain aminotransferase 2 (BCAT2) (see Figure 1), a known natural variant of the protein,[22,23] in GSC28. This finding was further validated at the transcript level and by SRM (see sections 3.2 and 3.3, respectively).

Due to the high confidence in this finding, identification of the SAV-containing peptide of BCAT2 (^{178}PVLIGNEPSLGVSQP**R**^{186}) in GSC 28 was used as a positive control for all database searches. According to the SNAP database of GPMdb, this tryptic peptide of BCAT2 is less frequently observed compared with its longer, miscleaved forms (http://snap.thegpm.org/%7E/dblist_protein_mut/label=ENSP00000322991). However, this may be a biased result of search algorithms that consider such peptides as false.[24] Interestingly, PEAKS DB was the only search engine that successfully identified the peptide containing BCAT2 *p.Thr186Arg*, with a −10log *P* score of 74. Neither SEQUEST nor Mascot identified this peptide, despite the high confidence of the assignment in PEAKS. Our initial thought was that both failed to identify the peptide due to trypsin cleavage specificity assigned in the search parameters, as the N-terminus of the SAV-containing peptide arises from cleavage between arginine and proline. Therefore, we adjusted the trypsin specificity in Mascot to allow for cleavage N-terminal to proline and

dx.doi.org/10.1021/pr500810g | J. Proteome Res. 2015, 14, 778−786

Table 1. Chromosome 19 SAVs Identified from Custom Database Searches of Untargeted Proteomic Data from 30 Cell Lines[a]

accession	gene symbol	protein name	peptide	−10log *P*	ppm	*m/z*	dbSNP reference SNP number	HGVS notation	COSMIC reference SNP number
NX_Q9ULX6 *p.His458Gln*	AKAP8L	A-kinase anchor protein 8-like	TVEDLDGLIQQIYR	65.25	1.2	831.9395	rs2058322[b]	ENSP00000380557:p.Q458H[b]	n.a.
NX_O15382 *p.Thr186Arg*	BCAT2	branched-chain-amino-acid aminotransferase, mitochondrial	PVLIGNEPSLGVSQPR	74.31	0.5	831.9627	rs11548193	ENSP00000322991:p.T196R	n.a.
NX_Q13011 *p.Glu41Ala*	ECH1	delta(3,5)-delta(2,4)-dienoyl-CoA isomerase, mitochondrial	LTGSSAQEAASGVALGEAPDHSYESLR	68	3.8	901.7701	rs9419	ENSP00000221418:p.E41A	n.a.
NX_P06744 *p.Ile208Thr*	GPI	glucose-6-phosphate isomerase	TLAQLNPESSLFITASK	72.4	0.5	910.4941	rs8191371	ENSP00000348877:p.I208.T	n.a.
NX_Q9BQ67 *p.Arg319Gln*	GRWD1	glutamate-rich WD repeat-containing protein 1	QEPFLLSGGDDGALK	50.95	−0.4	773.8907	rs2302951	ENSP00000253237:p.R319Q	n.a.
NX_Q92945 *p.Ala92Ser*	KHSRP	far upstream element-binding protein 2	IGGDSATTVNNSTPDFGFGGQK	70.56	−0.7	1085.5051	rs61751242	ENSP00000381216:p.A92S	n.a.
NX_O00754 *p.Leu278Val*	MAN2B1	Lysosomal alpha-mannosidase	NLcWDVLcVDQPVVEDPR	76.03	1.4	1107.521	rs1054486	ENSP00000221363:p.L278 V ENSP00000395473:p.L278 V	n.a.
NX_Q66K74 *p.Cys440Tyr*	MAP1S	Microtubule-associated protein 1S	VLFPGcTPPAYLLDGLVR	68.39	−0.5	994.5368	rs12983721	ENSP00000325313:p.C440Y	n.a.
NX_P37198 *p.Gly3Trp*	NUP62	nuclear pore glycoprotein p62	WFNFGGTGAPTGGFTFGTAK	66.9	−4	1010.9722	n.a.	ENSP00000305503:p.G3W	567508
NX_P12955 *p.Glu170Val*	PEPD	Xaa-Pro dipeptidase	FVVNNTILHPEIVEcR	51.33	1.4	647.341	rs61748998	ENSP00000244137:p.E170 V ENSP00000380226:p.E170 V	n.a.
nxlNX_O60664 *p.Val275Ala*	PLIN3	perilipin-3	AQEALLQLSQALSLMETVK	91.87	0.5	1037.0671	rs9973235	ENSP00000221957:p.V275A ENSP00000465596:p.V275A	n.a.
NX_O60664−1: *p.Gly171Ser*	PLIN3	perilipin-3	SVVTSGVQSVMGSR	65.07	3.1	697.361	rs144123988	ENSP00000221957:p.G171S ENSP00000465596:p.G171S	n.a.
NX_P14314 *p.Ile453Val*	PRKCSH	glucosidase 2 subunit beta	LGGSPTSLGTWGSWVGPDHDK	70.85	−0.7	1077.5154	rs34351170	ENSP00000252455:p.I453 V ENSP00000465461:p.I453 V	n.a.
NX_P14314 *p.Ala291Thr*	PRKCSH	glucosidase 2 subunit beta	SEALPTDLPTPSAPDLTEPK	84.56	2.4	1040.0306	rs11557488	ENSP00000252455:p.A291T ENSP00000395616:p.A291T ENSP00000465489:p.A291T ENSP00000466134:p.A291T ENSP00000466012:p.A291T ENSP00000465461:p.A291T	n.a.
NX_Q8IY67 *p.Glu273Asp*	RAVER1	ribonucleoprotein PTB-binding 1	GFAVLEYETADmAEEAQQQADGLSLGGSHLR	51.84	1.4	1103.8506	rs12610701	ENSP00000482277:p.E273D ENSP00000479520:p.E273D	n.a.
NX_P40429 *p.Ala154Asp*	RPL13A	60S ribosomal protein L13a	YQAVTDTLEEK	68.2	0.4	648.8198	rs150697570	ENSP00000375730:p.A154D	n.a.

dx.doi.org/10.1021/pr500810g | J. Proteome Res. 2015, 14, 778−786

Table 1. continued

accession	gene symbol	protein name	peptide	−10log *P*	ppm	*m/z*	dbSNP reference SNP number	HGVS notation	COSMIC reference SNP number
NX_Q9UBE0 *p.Glu298Asp*	SAE1	SUMO-activating enzyme subunit 1	YcFSDmAPVcAVVGGILAQEIVK	39.65	3	848.4188	n.a.	ENSP00000270225:p.E298D	566726
NX_Q9H7N4 *p.Thr420Pro*	SCAF1	splicing factor, arginine/serine-rich 19	AARPPPAASATPTAQPLPQPPAPR	70.73	0.8	787.4332	rs7251334	ENSP00000353769:p.T420P	n.a.
NX_Q15758 *p.Val512Leu*	SLC1A5	neutral amino acid transporter B(0)	SELPLDPLPLPTEEGNPLLK	73.31	1.7	1086.5953	rs3027961	ENSP00000444408:p.V512L	n.a.

[a]Annotated HCD−MS/MS spectra and the corresponding ion tables can be found in Supporting Information Figure 1. [b]Conflict in dbSNP and Ensembl databases, which indicates this mutation Q → H.

repeated the search using GSC 28 as a positive control. Because Mascot again failed to identify the peptide containing BCAT2 *p.Thr186Arg*, we focused on PEAKS DB to identify SAVs in the remaining cell lines (see Table 1 for a verified list and Supporting Information Table 1 for a complete list of SAV peptides identified by PEAKS).

For those peptides that were confirmed to be truly unique by homology searching, manual verification of spectra was performed in order to ensure that the assigned SAVs were correct. Because our data were acquired using high resolution MS in an Orbitrap, mass error for the precursor ions was used as an initial filter of quality. Because the mass errors for peptide assignments in each LC−MS/MS file fall into a normal distribution, and 95% of correct assignments fall within two standard deviations of the mean error,[25] we removed peptides whose mass error was dramatically higher than for other peptides within the same analytical run. In a similar manner, high mass accuracy MS/MS data facilitates spectral verification. However, high mass accuracy cannot always be used to confirm SAVs. Leucine and isoleucine are indistinguishable from one another by exact mass and chemical deamidation can convert glutamine and asparagine to glutamic acid and aspartic acid, respectively. Furthermore, the carbamidomethyl group (57.02146) has the same exact mass as a glycine residue. Therefore, species formed due to overalkylation (addition to nucleophilic amino acid side chains or the peptide N-terminus) can be assigned incorrectly as being due to SAVs. Therefore, it is often critical to have orthogonal validation of SAVs at the transcript level. A list of verified SAV-containing peptides identified by LC−MS/MS is presented in Table 1. A further list of SAV-containing peptides that require additional transcript level validation is provided in Supporting Information Table 2.

3.2. Validation of BCAT2 *p.Thr186Arg* at RNA Level

Each identified peptide containing a SAV was verified manually as described above. For the orthogonal validation, we evaluated the individual reads from the transcriptome sequencing to quantitate the frequency of C (ACG coding Thr) and G (AGG coding Arg) alleles. The C → G variant occurs at position 613 in the reference mRNA sequence of BCAT2 NM_001190 (rs11548193 in dbSNP).[26] A total of 8 GSC lines contained reads of >50% with C613G variant, including GSCs 28, 274, 7−11, 275, 17, 280, 289, and 293 (data not shown).

3.3. SRM Verification of Selected SAVs

The successful identification of SAV proteoforms in GSCs was further validated by SRM. Selected unique peptide sequences with a single mutation were synthesized and SRM assays were developed using multiple transitions for each peptide (Supporting Information Table 3). A subset of ten GSC samples (GSCs 28, 6−27, 20, 112, 129, 5−22, 275, 7−11, 274, and 7−2) was tested for four newly identified mutant proteoforms. The primary target of SRM confirmation was the BCAT2 *p.Thr186Arg* mutation that was unambiguously verified in samples GSC 28, GSC 7−11, and GSC 275 by perfectly matching transitions and retention times of heavy labeled internal standard and native peptides (see Figure 2A).

The specificity of SRM assays was utilized in investigation of Leu → Ile and Ile → Leu mutations that were initially identified in database search despite of the isobaric nature of these SAVs. The difference in retention times observed in reversed phase chromatography due to the lower hydrophobicity of peptides containing isoleucine provided a utility. A SAV of elongation factor 2 (P13639-1 *p.Ile665Leu*) was expected with longer

dx.doi.org/10.1021/pr500810g | J. Proteome Res. 2015, 14, 778−786

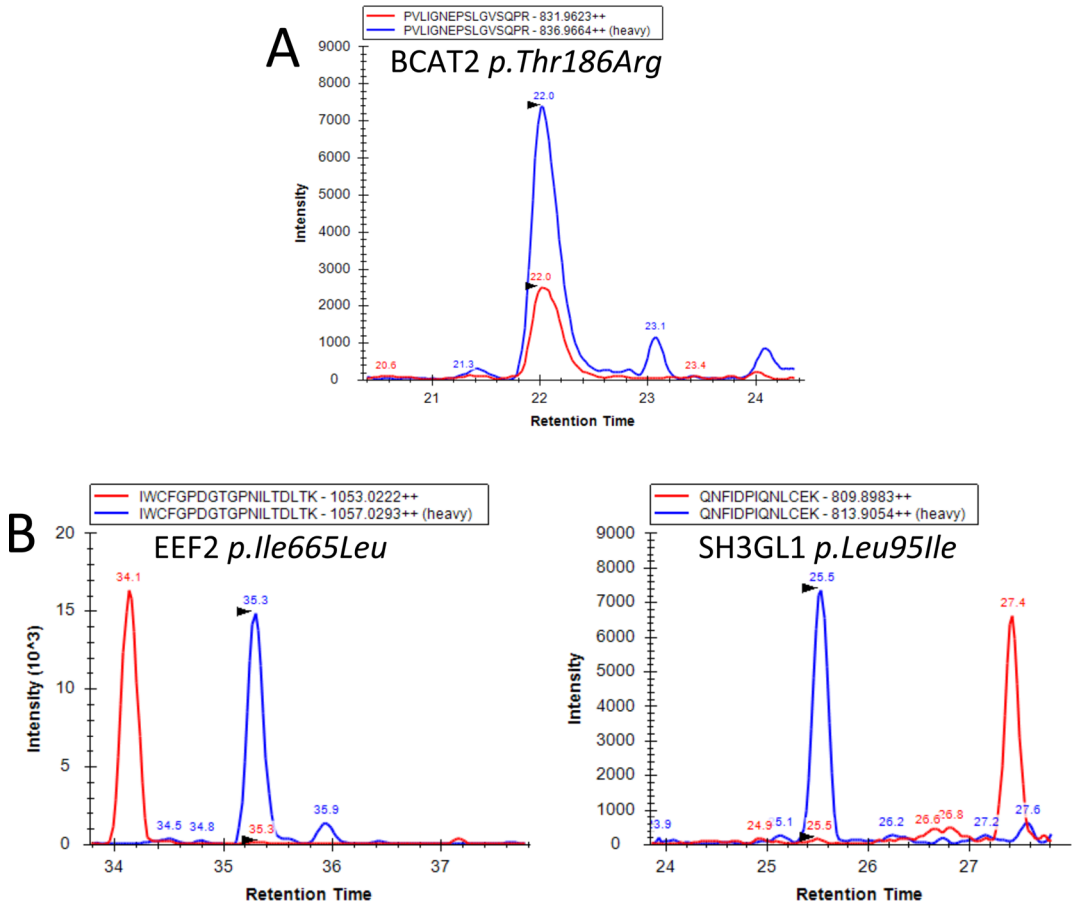


Figure 2. Validation of SAVs by SRM assay. (A) BCAT2 *p.Thr186Arg* was confirmed by matching heavy-isotope-labeled synthetic peptide (blue) with the endogenous PVLIGNEPSLGVSQPR (red), whereas (B) the endogenous signals of elongation factor 2 *p.Ile665Leu* and endophilin-A2 *p.Leu95Ile* could not agree with the internal standards.

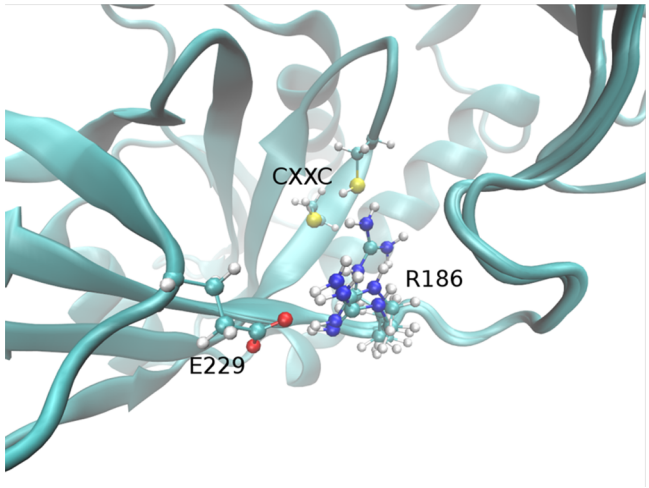


Figure 3. Best four BCAT2 *p.Thr186Arg* structures are superimposed with Arg186, Glu229, and the CXXC (Cys321 and Cys324) highlighted. Among rotamers with residues 177−188 minimized, the four overlaid here were hundreds of kcal/m lower than the average of all rotamers. The arginine side chain is found in a salt bridge with Glu229 and toward the inside of the flexible arm, where the arginine side chain has the most stable hydrogen bonding. These lower energy positions of the arginine side chain all greatly change the electrostatic environment as compared with the Thr186 wild type, and this may influence the CXXC oxidation reaction that regulates enzyme activity.

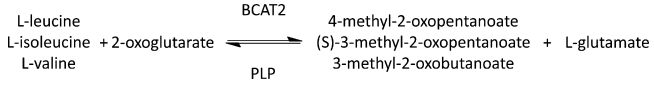


Figure 4. Reaction mechanism for BCAT2, which catalyzes the transfer of an amino group from leucine, isoleucine, and valine to α-ketoglutarate to form glutamate. PLP is a required cofactor for this reaction.

retention time but the endogenous signal with identical transitions eluted earlier than the synthetic isotope labeled standard. A reversed case was also observed when an endophilin-A2 mutation (Q99961-2 *p.Leu95Ile*) was investigated (see Figure 2B).

Besides their validation power, SRM assays can be further utilized in quantitative analysis of mutant allele expressions in biological samples. This information is necessary in the case of heterozygous samples, in particular when the expression of the mutant allele has a strong concordance with biological activity of cells.

3.4. Structural/Functional Implications of the BCAT2 *p.Thr186Arg* Mutation

The *p.Thr186Arg* mutation on BCAT2, confirmed at the RNA level to be in eight of the 36 GSC lines, is located along a flexible loop in close proximity to the CXXC center (Figure 3). To study the possible effects of this mutation, we generated 78 *p.Thr186Arg* BCAT2 variant structures by rotating the arginine side chain about the χ_1 and χ_2 and scanning the ψ backbond.

dx.doi.org/10.1021/pr500810g | J. Proteome Res. 2015, 14, 778−786

Following minimization, electrostatic free energy calculations yielded a set of Arg186 rotamers. The structures that where highest in electrostatic free energy or contained severe violations of the chemical constrains were removed from the data set. Of the structures lowest in electrostatic free energy, which satisfied the packing and stereochemical constraints in minimization, four were grouped within a few kilocalories per meters, whereas the rest were separated by over 100 kcal/m in electrostatic free energy and so were not considered for further analysis.

As seen by the four lowest energy structures (Figure 3), the Arg186 side chain gravitates toward the carboxylate group of Glu229, as well as inward toward the CXXC site, both of which provide stability through hydrogen bonding pairs. Compared with the Thr186 wild type, the arginine side chain significantly changes the electrostatic environment near the CXXC, potentially shifting the pK_a of the cysteine residues. The electrostatically induced pK_a shift could influence the cysteine protonation state and therefore the equilibrium of the oxidation reaction (Figure 4), consequently altering enzyme activity. Although direct interaction, for example, through hydrogen bonding, is not sterically forbidden, it is also not required. The electrostatic shift of the intrinsic pK_a of the cysteine would be sufficient to explain a change in activity. It has been demonstrated that the CXXC center (Cys321 and Cys324) contributes to enzyme activity through reversible disulfide bond formation that prevents substrate from correctly orienting with the pyridoxal phosphate (PLP) cofactor.[27]

4. CONCLUSIONS

We demonstrated the utility of our custom protein database in the identification of chromosome 19 SAVs in GSCs. Through a combination of homology searching and manual verification of spectra identified by PEAKS DB, we identified 19 SAV-containing peptides. These peptides represent 19 SAVs in 17 chromosome 19 proteins. One of these SAVs, BCAT2 *p.Thr186Arg*, was validated orthogonally in multiple cell lines by SRM and by RNA-Seq Future studies will be directed toward verification of the remainder of these SAVs and a study of the biological implications of these SAVs. The identification and characterization of SAVs in glioma has already yielded the identity proteins that will be of further interest to study as potential modulators of gliomagenesis, contributors to glioma pathology, and resistance to standard of care treatments. We expect that our methods are readily transferrable to other cancer cells and tissues as well.

To improve visibility of SAVs, we strongly suggest extending the registry of protein sequences in neXtProt with information about the mutant proteoforms pointing to identification levels as well as expression sites.

■ ASSOCIATED CONTENT

Supporting Information

Complete list of theoretical and observed m/z values for ions observed in HCD-MS/MS spectrum of ^{178}PVLIGNEPSLGVSQP**R**^{186}, list of SAV-containing peptides that require additional transcript level validation, elected unique peptide sequences with a single mutation, protein accession numbers and sequencing, and ion m/z information. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Á.Végvári. E-mail: akos.vegvari@bme.lth.se. Phone: +46-46-222-3721. Address: Clinical Protein Science & Imaging, Department of Biomedical Engineering, Lund University, BMC D13, SE-221 84 Lund, Sweden.

*C. L. Nilsson. E-mail: carol.nilsson@utmb.edu. Phone: +1-409-747-1840. Address: Department of Pharmacology and Toxicology, UTMB Cancer Center, University of Texas Medical Branch, Galveston, Texas 77555-1048, United States.

Author Contributions

$^{\nabla}$These authors contributed equally to this work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Drs. Huiling Liu and Alexander Shavkunov at UTMB are acknowledged for sample preparation and acquisition of MS/MS data. Financial support from the Cancer Prevention Research Institute of Texas (R1122, C.L.N.), the University of Texas Medical Branch (C.L.N., B.M.P.), the National Institutes of Health (RO1 GM037657, B.M.P.), and the Crafoord Foundation (20130635, A.V.) are gratefully acknowledged.

■ ABBREVIATIONS

C-HPP, Chromosome-Centric Human Proteome Project; LC−MS/MS, liquid chromatography tandem mass spectrometry

■ REFERENCES

(1) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F. F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−223.

(2) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111.009993.

(3) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23−7.

(4) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.;

Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575−81.

(5) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−7.

(6) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013−2014 and Strategies for Finding Missing Proteins. *J. Proteome Res.* **2014**, *13* (1), 15−20.

(7) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11* (4), 2005−13.

(8) Wood, L. D.; Parsons, D. W.; Jones, S.; Lin, J.; Sjoblom, T.; Leary, R. J.; Shen, D.; Boca, S. M.; Barber, T.; Ptak, J.; Silliman, N.; Szabo, S.; Dezso, Z.; Ustyanksky, V.; Nikolskaya, T.; Nikolsky, Y.; Karchin, R.; Wilson, P. A.; Kaminker, J. S.; Zhang, Z.; Croshaw, R.; Willis, J.; Dawson, D.; Shipitsin, M.; Willson, J. K.; Sukumar, S.; Polyak, K.; Park, B. H.; Pethiyagoda, C. L.; Pant, P. V.; Ballinger, D. G.; Sparks, A. B.; Hartigan, J.; Smith, D. R.; Suh, E.; Papadopoulos, N.; Buckhaults, P.; Markowitz, S. D.; Parmigiani, G.; Kinzler, K. W.; Velculescu, V. E.; Vogelstein, B. The genomic landscapes of human breast and colorectal cancers. *Science* **2007**, *318* (5853), 1108−13.

(9) Mathivanan, S.; Ji, H.; Tauro, B. J.; Chen, Y. S.; Simpson, R. J. Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *J. Proteomics* **2012**, *76* (Spec No.), 141−9.

(10) Han, Y.; Ma, B.; Zhang, K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J. Bioinf. Comput. Biol.* **2005**, *3* (3), 697−716.

(11) Lichti, C. F.; Liu, H.; Shavkunov, A. S.; Mostovenko, E.; Sulman, E. P.; Ezhilarasan, R.; Wang, Q.; Kroes, R. A.; Moskal, J. C.; Fenyö, D.; Oksuz, B. A.; Conrad, C. A.; Lang, F. F.; Berven, F. S.; Végvári, Á.; Rezeli, M.; Marko-Varga, G.; Hober, S.; Nilsson, C. L. Integrated Chromosome 19 Transcriptomic and Proteomic Data Sets Derived from Glioma Cancer Stem-Cell Lines. *J. Proteome Res.* **2014**, *13* (1), 191−199.

(12) Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andrén, P. E.; Nilsson, A.; Carlsohn, E.; Lilja, H.; Malm, J.; Fenyö, D.; Subramaniyam, D.; Wang, X.; Gonzales-Gonzales, M.; Dasilva, N.; Diez, P.; Fuentes, M.; Végvári, Á.; Sjödin, K.; Welinder, C.; Laurell, T.; Fehniger, T. E.; 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. *J. Proteome Res.* **2012**, *12* (1), 135−150.

(13) 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 Res.* **2004**, *64* (19), 7011−7021.

(14) Jiang, H.; Gomez-Manzano, C.; Aoki, H.; Alonso, M. M.; Kondo, S.; McCormick, F.; Xu, J.; Kondo, Y.; Bekele, B. N.; Colman, H.; Lang, F. F.; Fueyo, 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* (18), 1410−1414.

(15) Torres-Garcia, W.; Zheng, S.; Sivachenko, A.; Vegesna, R.; Wang, Q.; Yao, R.; Berger, M. F.; Weinstein, J. N.; Getz, G.; Verhaak, R. G. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* **2014**, *30* (15), 2224−6.

(16) Yennawar, N. H.; Conway, M. E.; Yennawar, H. P.; Farber, G. K.; Hutson, S. M. Crystal structures of human mitochondrial branched chain aminotransferase reaction intermediates: ketimine and pyridoxamine phosphate forms. *Biochemistry* **2002**, *41* (39), 11592−601.

(17) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33−8 27−8..

(18) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32* (Web Server issue), W665−7.

(19) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (18), 10037−10041.

(20) Victor, B.; Gabriel, S.; Kanobana, K.; Mostovenko, E.; Polman, K.; Dorny, P.; Deelder, A. M.; Palmblad, M. Partially sequenced organisms, decoy searches and false discovery rates. *J. Proteome Res.* **2012**, *11* (3), 1991−5.

(21) Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics* **2012**, *11* (4), M111 010587.

(22) Bledsoe, R. K.; Dawson, P. A.; Hutson, S. M. Cloning of the rat and human mitochondrial branched chain aminotransferases (BCATm). *Biochim. Biophys. Acta* **1997**, *1339* (1), 9−13.

(23) Eden, A.; Simchen, G.; Benvenisty, N. Two yeast homologs of ECA39, a target for c-Myc regulation, code for cytosolic and mitochondrial branched-chain amino acid aminotransferases. *J. Biol. Chem.* **1996**, *271* (34), 20242−5.

(24) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does Trypsin Cut Before Proline. *J. Proteome Res.* **2008**, *7*, 300−305.

(25) Zubarev, R.; Mann, M. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics* **2007**, *6* (3), 377−81.

(26) Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29* (1), 308−11.

(27) Yennawar, N. H.; Islam, M. M.; Conway, M.; Wallin, R.; Hutson, S. M. Human mitochondrial branched chain aminotransferase isozyme: structural role of the CXXC center in catalysis. *J. Biol. Chem.* **2006**, *281* (51), 39660−71.