Supplementary Materials for

An anatomic transcriptional atlas of human glioblastoma

Authors: Ralph B. Puchalski*†, Nameeta Shah*†, Jeremy Miller, Rachel Dalley, Steve R. Nomura, Jae-Guen Yoon, Kimberly A. Smith, Michael Lankerovich, Darren Bertagnolli, Kris Bickley, Andrew F. Boe, Krissy Brouner, Stephanie Butler, Shiella Caldejon, Mike Chapin, Suvro Datta, Nick Dee, Tsega Desta, Tim Dolbeare, Nadezhda Dotson, Amanda Ebbert, David Feng, Xu Feng, Michael Fisher, Garrett Gee, Jeff Goldy, Lindsey Gorley, Benjamin W. Gregor, Guangyu Gu, Nika Hejazinia, John Hohmann, Parvinder Hothi, Robert Howard, Kevin Joines, Ali Kriedberg , Leonard Kuan, Chris Lau, Felix Lee, Hwahyung Lee, Tracy Lemon, Fuhui Long, Naveed Mastan, Erika Mott, Chantal Murthy, Kiet Ngo, Eric Olson, Melissa Reding, Zack Riley, David Rosen, David Sandman, Nadiya Shapovalova, Clifford R. Slaughterbeck, Andrew Sodt, Graham Stockdale, Aaron Szafer, Wayne Wakeman, Paul E. Wohnoutka, Steven J. White, Don Marsh, Robert C. Rostomily, Lydia Ng, Chinh Dang, Allan Jones, Bart Keogh, Haley R. Gittleman, Jill S. Barnholtz-Sloan, Patrick J. Cimino, Megha S. Uppin, C. Dirk Keene, Farrokh R. Farrokhi, Justin D. Lathia, Michael E. Berens, Antonio Iavarone, Amy Bernard, Ed Lein, John W. Phillips, Steven W. Rostad, Charles Cobbs, Michael J. Hawrylycz†, and Greg D. Foltz

*These authors contributed equally to this work.
†Corresponding authors. E-mail: MikeH@alleninstitute.org (M.J.H); rbpuchalski@gmail.com (R.B.P); nameeta.shah@gmail.com (N.S.)

This PDF file includes:

Materials and Methods
Tables S2, S5-S8
Figures S1 to S8
Captions for separate files of data tables S1, S3, S4, S9-S16

Other Supplementary Materials for this manuscript include:

Files of data tables S1, S3, S4, S9-S16

MATERIALS AND METHODS

Scope

Overall, the Ivy GAP cohort consisted of 41 patients (table S1), and their 42 tumors were used to generate ~440 tissue blocks, 270 transcriptomes, ~11,500 machine learning (ML)-annotated H&E images registered to ~23,000 ISH images , ~400 MRI scans, tumor-derived cell lines and xenografts, and supporting longitudinal clinical information.

Tumor acquisition and clinical data collection

This study was reviewed and approved by Western IRB (IRB20091429) in compliance with the ethical principles as set forth in the report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research entitled "Ethical Principles and Guidelines for the Protection of Human Subjects of Research (Belmont Report)". The research protocol was also approved by the Swedish Neuroscience Institute research steering committee. All participants provided written informed consent according to IRB guidelines prior to participation in this study. MRI data sets were collected at clinically determined intervals for each patient. Intra-operative photographs and MRI Stealth images were obtained before and after *en bloc* resection when possible. Patients received chemotherapy and radiation treatment after surgery. Clinical data sets that were collected included age, gender, molecular subtype, *MGMT* methylation, EGFR amplification, *EGFRvIII* deletion, *PTEN* deletion, *IDH1* point mutation, initial KPS, neurosurgery (resection number), hemisphere, chemotherapy, radiation therapy, recurrence by 6 months, and multifocality (table S1). Patient samples used in this study were diagnosed as WHO grade IV glioblastoma.

Cell line generation

Glioma stem cell cultures were established from freshly resected tumor tissues and maintained in NeuroCult® NSA medium (Stem Cell Technologies) with B-27 serum-free supplement (Invitrogen), 20 ng/mL epidermal growth factor (EGF) and 20 ng/mL fibroblast growth factor (FGF-2) as described (*34*). In brief, tissue samples were minced into 1 mm^3 fragments and digested with Accutase (Sigma) at 37 °C for 15-20 minutes.

NSA medium was added to quench Accutase activity and cell suspensions were passed through 70 μm nylon mesh. The suspensions were centrifuged at 1000 rpm for 5 minutes, resuspended in fresh NSA, and plated into T75 flasks pre-coated with laminin (1:100 in PBS; Sigma). To evaluate neurosphere formation ability of glioma stem cells, single cell suspensions were plated directly into ultra-low attachment dishes (Corning) and maintained in NSA medium.

Affymetrix genome-wide SNP arrays

Genomic DNA was extracted from tumor tissues by using ChargeSwitch gDNA mini tissue kit (Life Tech, CA) per company protocol. Measurement of quality and quantity of genomic DNA was carried out by using Nanodrop ND-1000. Samples were then sent for profiling on Affymetrix Genome-Wide SNP 6.0 arrays to Genome Technology Access Center, Washington University School of Medicine, St. Louis, MO.

Tissue acquisition, subdivision, and freezing

Immediately after *en bloc* resection, each tumor was placed on a surgical towel, and subdivided into 9 x 7.5 x 18 mm high (tumors W1-W12) or 9 x 7.5 x 9 mm high (tumors W13-W55) tissue blocks so that tissue sections also measured ~9 x 7.5 mm on glass slides. Each block was supported in a custom-fabricated Teflon-coated metal L-bar assembly set on aluminum foil over wet ice, and frozen with Flash Freeze Rapid Freezing Spray containing 1,1,1,2 Tetrafluoroethane (Decon Laboratories). An OCT base formed with a disposable cryomold was frozen to the 1-2 mm of tissue block that protruded from the top of the L-bars with the freezing spray to facilitate mounting of the block to a cryostat specimen disk. The 18 mm high blocks were divided into two 9 mm pieces on a refrigerated dissection table (-15°C) using a custom fabricated block chopper equipped with a standard razor blade. The bottom (block .1) contained the original OCT base, whereas the top (block .2) was embedded in OCT at the chopped interface. Blocks were stored at -80°C before processing.

Tissue processing, staining, and ISH

Tissue processing (cryosectioning, quality control, fixation, dehydration, and acetylation), staining (H&E staining, Toluidine blue staining, Feulgen-HP yellow counterstaining after colorimetric *in situ* hybridization (ISH)), as well as ISH and probes (colorimetric *in situ* hybridization, PCR and IVT reactions for ISH probes, ISH controls), are described in detail at http://glioblastoma.alleninstitute.org/. The core infrastructure developed for generating the Allen Mouse Brain Atlas (*35*) data (http://www.brain-map.org) was used for processing glioblastoma tissue for the Ivy GAP.

Fixation, dehydration, acetylation

The procedures described for the Allen Mouse Brain Atlas were used without modification. Slides destined for ISH were fixed, dehydrated, and acetylated, whereas slides destined for H&E staining were fixed and dehydrated. All slides were stored in plastic boxes at room temperature until processed by ISH within 3 weeks or stained with H&E within one week.

Hematoxylin and eosin (H&E) staining

Quality control slides were stained with H&E to facilitate assessment of tissue integrity and to review neuropathology. Slides were processed on a Leica Autostainer XL with a regressive H&E staining protocol stain (*36*). Sections were treated in Formula 83 (10 min), rehydrated in a graded series of 100% (2x), 95%, and 70% ethanol ending with de-ionized water (each 1 min), stained with non-mercuric Harris Hematoxylin (commercially prepared and filtered before every use, 13 min), rinsed with de-ionized water (1 min), differentiated in 1% HCl in 70% ethanol (45 sec), rinsed with de-ionized water (1 min), blued with 1% lithium carbonate (1 min), rinsed with de-ionized water (1 min), and stained in 1% eosin Y in 1% aqueous calcium chloride (6 min). Sections were then dehydrated in a graded series of 50%, 70%, 95% and 100% (3x) ethanol, cleared in Formula 83, and coverslipped with DPX.

Toluidine blue staining

After each tissue block was sectioned onto slides, one section was collected for manual staining and review of tissue integrity and neuropathology prior to sectioning the same

block onto the next set of slides. Just after collection, the section was stained with 2% Toluidine Blue (Fisher Scientific; 1 min), rinsed in tap water twice (5 dips over 10 sec each), dehydrated in 100% ethanol twice (1 min each), air dried (2 min), then mounted with Vecta Mount, coverslipped and reviewed immediately.

Tissue block, neuropathology, and RNA quality control

Blocks passed tissue integrity criteria if 2-4 Hematoxylin and Eosin (H&E)-stained sections collected from the block face had less than, 10% freezing artifact and 75% necrosis per unit area. Tissue blocks with excessive necrosis were failed and therefore were not included in the studies, and as a result, some tumors did not have a sufficient number of passing blocks to be included in the atlas. Blocks passed neuropathology criteria if the H&E-stained sections had at least 25% viable cells of core tumor at a high density of cells (50-100 cells/100μm^2). The remaining 75% typically consisted of viable cells in core tumor at low to high density, tissue at various stages of necrosis, tissue affected by hemorrhage, absence of cells or tissue, leading edge (tumor cell-free margin), or infiltrating tumor. For blocks that passed tissue integrity and neuropathology criteria, RNA quality was assessed. RNA Integrity Numbers (RINs) ranged from 5.6 to 8.7, with most measurements between 7 and 8.

Cryosectioning for ISH surveys

Fresh frozen tissue blocks were removed from -80°C, equilibrated at -15°C in cryostats, mounted on chilled chucks, and sectioned at 20μm with object temperature of -10°C or -11°C to reduce chatter through the necrotic areas and folds on the leading edge that contacted the blade first. Each of 8 blocks of tissue (1.x – 8.x) was sectioned onto sequential slides for ISH and HP-Yellow counterstain (slides 1, 3, 4… 32 or 45) or H&E-stained (slides 2, 5…44).

Cryosectioning for laser microdissection

Fresh frozen tissue blocks were removed from -80°C, equilibrated for sectioning at -15°C, and sectioned at 14μm with the cryostat object temperature at -10°C or -11°C to reduce chatter through the necrotic areas and folds on the leading edge that contacted the

blade first. Tissue sections were mounted onto glass slides with polyethylene naphthalate (PEN) membranes (Leica Microsystems, Inc., Bannockburn, IL). For the Anatomic Structures RNA-Seq Study, 1 H&E slide with 2 sections was included as a reference for every 3 PEN slides with 4 sections each. For the Cancer Stem Cells RNA-Seq Study, 1 H&E slide with 2 sections was included for every 3 PEN slides with 4 sections and 2 to 8 ISH reference slides with 2 sections. After drying for 30 minutes at room temperature, PEN slides were frozen at -80°C. Slides destined for ISH were fixed, dehydrated, and acetylated, whereas slides destined for H&E staining were fixed and dehydrated. The slides were stored in plastic boxes at room temperature until processed by ISH within 3 weeks or stained with H&E within one week.

Laser microdissection

In preparation for laser microdissection, PEN slides were removed from -80°C and quickly processed through cresyl violet and Eosin to lightly stain the tissue. Sections were fixed in ice-cold 70% ethanol for 30 seconds, washed for 15 seconds in nuclease-free water, stained with 0.7% cresyl violet in 0.05% NaOAc, pH 3.4 for 4 minutes, rinsed in nuclease-free water for 10 seconds, 15 seconds in 70% ethanol, followed by 2 dips in 0.25% Eosin, and 20 seconds each in 95%, 100%, and 100% ethanol rinses. Slides were air-dried for 2 minutes and desiccated by vacuum for 1 hour at room temperature, then frozen at -80°C until microdissection. For both RNA-Seq studies, cresyl violet/Eosin-stained sections mounted on PEN membranes were microdissected while visually referring to H&E-stained sections that had been curated to identify matched target regions. For the Cancer Stem Cells RNA-Seq study, ISH reference gene expression patterns informed the curation of the H&E-stained sections. A Leica LMD6000 (Leica Microsystems, Inc., Bannockburn, IL) was used for microdissection, and the system included an upright research microscope fitted with a diode laser and a CCD camera to acquire live images of slides. The scope and laser were controlled via a dedicated computer running Leica LMD software (v.6.6.2.3552).

ISH probe design, synthesis, and testing

For labeling target mRNA in tissue sections using ISH, digoxigenin-labeled riboprobes were designed and synthesized according to specific criteria. Briefly, using sequences obtained from RefSeq and a semi-automated process based on Primer3 software (*37*), probes were designed to be between 400-1000 bases in length (optimally > 600 bases) and to contain no more than 200 bp with > 90% homology to non-target transcripts. To facilitate cross-platform comparisons with existing Affymetrix microarray data from glioblastoma tissue, probes for the Ivy GAP were designed to completely overlap their corresponding Affymetrix array probe when possible. Secondarily, the probes were then designed to have >50% overlap with the existing Allen Mouse Brain Atlas probe when the mouse and human genes were orthologous. Riboprobes were synthesized using standard in vitro transcription (IVT) reactions based on PCR templates prepared from human cDNA clones (NIH Mammalian Gene Collection, Open Biosystems, Huntsville, AL) or pooled cDNA synthesized from human brain total RNA. cDNA was prepared from human brain RNA from prefrontal, temporal, parietal, occipital, and frontal cortical areas as well as medulla and cerebellum (Ambion, Austin, TX) using Superscript III RTS First-Strand cDNA Synthesis Kit (Invitrogen, Carlsbad, CA), then pooled in equal amounts to provide templates for PCR. PCR primers were obtained from Integrated DNA Technologies (Coralville, IA) at a final concentration of 10μM, and designed with GC content between 42% – 62% and an optimal size of 22nt with lower and upper limits of 18nt and 26nt, respectively. For cDNA clones, the clone sequence was compared with RefSeq sequences, and consensus sequences with >98% homology across 80% of the total length were used to develop probes. When a clone was used as a template, a single PCR was used requiring only a forward and reverse primer with an additional SP6 RNA polymerase binding sequence (GCGATTTAGGTGACACTATAG). When using brain cDNA as a template, probes were generated against sequences within a region 3000 bp from the 3' end using 3 primers: forward, reverse, and a nested reverse primer containing the SP6 RNA polymerase binding sequence. cDNA primers underwent a BLAST analysis to verify amplification of only target sequence. All cDNA reactions were run on the Bioanalyzer for quality control.

PCR and IVT reactions for ISH probes

Standard conditions for PCR and IVT reactions were as described (*35*). IVT reactions were diluted to working stocks of 30ng/µl with THE (0.1mM Sodium Citrate pH 6.4, Ambion). Aliquots were stored in low volumes to minimize freeze/thaw cycles. IVT dilutions were stored at -80°C. For hybridization, the probe was diluted 1:100 (to 300ng/ml) or 1:50 (to 600ng/ml) into in situ hybridization buffer (Ambion) in 96-well ISH Probe Plates. A probe was hybridized at 600ng/ml if its gene expression in The Cancer Genome Atlas (TCGA) dataset fell in the lowest or third quartile rank. A list of probes and the concentrations used for each study appear in the Gene List in Documentation. MECOM was always hybridized at 300ng/ml. Each well provides probe for one ISH slide. Probe plates were stored at -20°C until used in an ISH run. All PCR and IVT products were run on the bioanalyzer for size and morphology quality control. Specifically, PCR products that were not of the correct size (+/- 100bp) or that showed multiple products were not used to generate riboprobes. IVT products that were shorter than their predicted size were not used. It is common to see IVT products that run slightly larger than their predicted molecular weight, or as multiple peaks, due to secondary structure of the RNA. IVT products with multiple bands were not used for ISH unless the additional bands were determined to result from secondary structure. For the Anatomic Structures ISH Survey and the Cancer Stem Cells ISH Survey, probes were tested on the ISH platform optimized for glioblastoma tissue. About 90-95% of the probes passed and was used in the atlas. Probes were failed if they caused excessive ISH artifacts such as streaks, background and signal gradients, and spots on multiple tissue sections. For the Anatomic Structures ISH for Enriched Genes Study and Cancer Stem Cells ISH for Enriched Genes Study, newly designed and synthesized probes were not tested prior to being used in the studies.

Colorimetric ISH hybridization and Feulgen-HP yellow counterstain

The ISH procedures developed for generating the Allen Mouse Brain Atlas (*35*) data (http://www.brain-map.org) were used for glioblastoma tissue, except that the proteinase K concentration used was one cycle of 5 min at 0.0007U/mL, and TSA+ (2.0µl/slide) was used instead of TSA for all studies other than the first (Anatomic Structures ISH Survey). The chemistry was optimized to detect low signal levels and therefore the ISH reaction

product is saturated in many cases. Colorimetric ISH is semi-quantitative at best and is not an absolute measurement of gene expression levels; it is relative to mRNA levels in the range of non-saturating conditions. Feulgen-HP yellow DNA counterstain is a nuclear stain that adds definition to tissue with low ISH background to facilitate automated focus processes during image acquisition and to support gene expression analysis, whether done manually or informatically. After ISH, slides were removed and treated with acid alcohol (70% ethanol adjusted to pH 2.1) to reduce background, 5N hydrochloric acid to prepare the tissue for the counterstain, HP yellow counterstain (Catalog #869, Anatech Ltd) to stain the nuclei, and two final acid alcohol washes to remove non-covalently bound HP yellow stain, all on a Leica autostainer. Slides were coverslipped with Hydro-Matrix Mounting Medium on a Leica CV5030 coverslipper. Coverslipped slides were incubated overnight at 37°C to promote solidification of the mounting media. Prior to scanning, slides were cleaned to remove excess mounting media and debris.

ISH controls

The experimental variability of the automated ISH process was monitored with ISH run controls and ISH tissue controls. The run control probes Drd1a and Calb1 were hybridized to mouse brain tissue sections cut at 25μm thick, one slide per probe and 4 brain sections per slide. If the hybridization pattern of intensity and density of signal, or the background, was considered by personnel dedicated to the quality control of the process to be outside the normal variability, then the experiment was failed and repeated. The tissue control probe MECOM was hybridized to glioblastoma tissue sections cut at 20μm thick for the ISH studies, which contained 8 sections per slide, and 14 μm thick for the ISH reference slides associated with laser microdissections used for the Cancer Stem Cells RNA-Seq Study, which contained 2 sections per slide. MECOM is expressed at low levels predominately in vascular tissue, and was used to monitor ISH signal on all specimen sub-blocks on all ISH runs. Slides containing the mouse or glioblastoma tissue were also processed through hybridization without probe, which was used to monitor non-specific hybridization.

Image acquisition and processing

Whole slides were scanned directly to SVS file format at a resolution of 0.5µm/pixel without down sampling on ScanScope® scanners (Aperio Technologies, Inc; Vista, CA) equipped with a 20x objective and Spectrum software. The raw image files of ~5 GB per image were archived after they were converted to JPEG 2000 file format. The preprocessed images were flipped along the horizontal axis, white balanced, and compressed at a rate of 0.8 to ~400 MB per image. During post-processing (Informatics Data Processing), colorized expression values or heat masks showing ISH signal intensity were generated and the closest H&E stained image of the same specimen was determined for each ISH section.

Image and annotation quality control

During review of images, the automated bounding box overlay was manually adjusted if necessary so that each of 8 bounding boxes per slide was placed over the corresponding tissue section, and images of slides with focus or image tile stich misalignments were re-scanned. Images were failed if data analysis was compromised by artifacts (e.g. mechanical damage, mounting medium bubbles, hybridization bubbles, and NBT/BCIP precipitated aggregates) associated with the corresponding tissue section.

Image processing pipeline

An informatics processing pipeline was developed to support the quantification of gene expression in tumor features, which were labeled in the nearest H&E image with the semi-automated annotation application, and for which the nuclear density was counted in each 45 x 45 pixel grid.

Image signal detection

Images of ISH sections counterstained with HP-Yellow were first down-sampled by a factor of 2. The underlying method for ISH detection algorithm used is in principal similar to the one for the Allen Mouse Brain Atlas (http://www.brain-map.org), which is based on adaptive filtering in combination of mathematic morphology operations. However, the image spectral information used for detection is mainly from the Red

channel in order to avoid the impact of the HP-Yellow counterstain, though the contrast of the expressors in Red are lower than in Green channel but still deemed sufficient. Additional process was included to separate and remove the expressor-look-alike dark brown objects artifacts. There are also morphometric limitations set for expressor objects to exclude various other types of artifacts such as fragments of coverslip edges.

Image signal unionization and search

Classification results in feature labels and nuclear coverage fractional area for each 45x45 pixel block for each H&E image. Automated ISH annotation was achieved by registering to the ISH to the nearest H&E image and warping the feature labels to fit the ISH image. Small features were dilated to compensate for potential misalignment. An *expression energy* value is computed for each feature by summing expressing pixel intensity for each block labeled for that feature normalized by the sum of nuclear coverage fractional area. A search service was developed to allow users search over the whole dataset for user defined specimen and/or expression profiles. Tumor feature search: find all tumor blocks containing a specific feature sorted by the normalized area occupied by the feature. Expression search: find all ISH sections containing a specific feature sorted by expression energy within the feature. Differential search: find all ISH section with expression energy with higher expression in the target feature compared to a contrast feature.

Image registration

A multi-resolution elastic registration algorithm was developed to register the ISH image to the closet H&E image to enable the transfer the anatomical region annotation onto the expression data. Both the H&E and ISH images were first down-sampled and split into RGB channels. Image pyramids were then constructed on the red channel of H&E and the blue channel of ISH to enable multi-resolution registration. An elastic registration algorithm (*38*) was then applied to each level of the H&E and ISH image pyramids. It minimized the energy function: $E = w_i E_{img} + w_s E_{smooth}$, where $E_{img}$is the intensity dissimilarity between the target (H&E) and the warped subject image (ISH); $E_{smooth}$is the smoothness of the deformation field of the warped subject image defined by divergence and curl; w_i, $w_s$are weights of E_{img}, and E_{smooth}, respectively. After

registration, the deformation field of the warped ISH was reversely-mapped. The deformation field of the H&E was generated, which indicated where each pixel in the H&E needed to move in order to match a pixel in the corresponding ISH image. The H&E annotation was further mapped onto the unwarped ISH images.

Machine learning algorithm and processing/annotation of H&E images

A ML application called "Mill" was created by White Marsh Forests, Inc., and was used by the Ivy GAP to identify and label the anatomic features in ~12,000 histological images based on the statistical ML algorithm called Decision Forests (*39-41*). The features labeled included leading edge (LE), infiltrating tumor (IT), cellular tumor (CT), early necrosis (EN), necrosis (NE), pseudopalisading cells around necrosis (PAN), hyperplastic blood vessels (HBV), and microvascular proliferation (MVP). The ML application was trained to identify each feature in each tissue sub-block by manually labeling the features in about 1 of 4 to 6 high resolution JPG images derived from a series of scanned H&E-stained tissue sections. The training was dynamic in that a small subset of images for a sub-block was labeled, predicted, and reviewed by the expert. In one to two iterations, the expert revised the training using positive and negative (i.e. labeling the region as "not X") training, which improved the ML labeling accuracy as determined by the expert as well as by the ML using holdout data. The alternative, training on a large set of images over several sub-blocks, blocks, or tumors, was much less accurate. The original images of ~15,000 x 18,000 pixels were down sampled by powers of two into image pyramids and then sampled by an array of image samplers. Once manually trained on a sub-block, Mill was used to automatically predict the features in all H&E images of that sub-block at a label density of one prediction point per 140 x 140 pixel area of the original image. The results were reviewed manually, and if the results were deemed unsatisfactory by the expert reviewer, the sub-block was tagged for manual editing. The prediction spacing was then reduced to a 45 x 45 pixel matrix, which required four times as long as for the processing of the preceding prediction step. A mode sampling filter was applied to these predictions, which reduces noise in the predictor by taking into consideration the predictions of each pixel's 8 neighbors. Then the remaining images for the sub-block were batched processed overnight. The errors in the tagged images were then manually

corrected, and the comma delimited label files were exported. A board-certified neuropathologist was consulted on a regular basis throughout the semi-automated annotation effort.

The parameter settings used within the Decision Forest algorithm include:

1. 1 forest per batch
2. 400 trees per forest
3. 400 random trials per split node to optimize information gain in training
4. 30% of each training set withheld from training each tree generated for testing (the test confusion matrix)
5. Image Sampling to generate train data for forest
    - Image samples randomly sampled within 800 pixels of the point being labeled
    - Random down-sampling using a Multi-resolution image pyramid
    - RGB and HSV color spaces sampled
    - Individual single samples and differences in value between two random samples used

Post-processing of csv annotation files

The final set of anatomic annotations for each H&E image was prepared in a series of post-processing steps, and then the CSV file was imported into the Allen Institute Laboratory Information Management System (LIMS). The steps were: (1) merged the CSV file of annotations with the file for nuclear count, (2) converted pseudopalisading cells around necrosis (PAN) to pseudopalisading cells no necrosis (PNN) if there was no necrosis (NE) within 900 pixels of PAN, (3) provided hierarchy for blood vessels (BV), hyperplastic blood vessels (HBV), and microvascular proliferation (MVP), (4) merged early necrosis (EN) with necrosis (NE) to make revised and final NE, (5) converted cellular tumor (CT) to cellular tumor perinecrotic zone (PNZ) if there was necrosis (NE) within 180 pixels, (6) made all labels consistent with Ivy GAP ontology, (7) replaced early training labels hemorrhage (HE), tissue fold (FOLD), ice damage (ICE), and space or no tissue (SPA) with 0, (8) replaced all predictions with 0 on selected images (<10) that were distorted due to sectioning artifacts and that could not be accurately registered to ISH, and (9) sorted files first by y coordinate and then by x coordinate.

Quantitation of nuclei in H&E tissue section images

The rule-set based application called Developer (Definiens, Inc.) was used to count the nuclei in the original JPG image of each tissue section, using the same 45 x 45 pixel matrix used by Mill after rescaling. Two layers, representing Hematoxylin and Eosin stained populations, were separated from the JPG image, and nuclei were counted using the watershed algorithm on the Hematoxylin layer. Nuclei fraction area was calculated for each grid as area covered by nuclei divided by total area. The nuclei count was saved in a CSV file, later merged with the post-processed CSV file of automated annotations from Mill, and used to normalize ISH signal intensity in each anatomic feature for the final gene expression calculations in the Ivy GAP web application.

Neuropathology concordance analyses

Three analyses consisted of (1) measuring variability between neuropathologists, (2) assessing the accuracy of manual annotations used for guiding laser microdissection, and (3) assessing the accuracy of the ML annotations used to label the anatomic features in ~12,000 H&E images of the atlas. The inter-neuropathologist variability was measured with a randomly organized set of 94 test images of the anatomic features curated to reflect the definitions (fig. S1) used throughout the construction of the atlas and training of the ML Decision Forest with the guidance of a neuropathologist (S.W.R.). It was critical for each neuropathologist involved in the analyses to adopt these definitions, even though they deviate from WHO definitions (e.g. HBV vs. MVP). If the definitions used for ML had not been adopted, assessing the accuracy of the ML annotations would not have been possible, because lack of concordance could reflect disagreement between neuropathologists, disagreement relative to a standard set of definitions, or disagreement with the ML labels. Our goal was to assess the accuracy of the manual and ML annotations so that the user of the atlas would know the reliability of the data. For each analysis, three neuropathologists (P.J.C., C.D.K., and M.U.) were presented with images of H&E-stained tissue sections and identified the main feature or agreement with the manual or ML label using controlled vocabulary. Online forms for assessing the manual and ML annotations guided neuropathologists to the images at http://glioblastoma.alleninstitute.org/, and recorded survey responses automatically with a timestamp. Details of the methods supporting the three analyses appear in the table

legends. Cohen's kappa **(κ)** (*42*) was calculated across all features for each neuropathologist as a measure of concordance, and the combined accuracy was calculated across all neuropathologists for each feature, an approach consistent with other neuropathology concordance analyses (*43*). Interpreting Cohen's kappa depends on how challenging the measurements are, but in general, a kappa of 0.66-0.96 is considered good to near perfect (https://stats.stackexchange.com/questions/82162/cohens-kappa-in-plain-english).

Analysis of the data collected from the neuropathologists' assessment of the 10 ML annotations for table S8 proceeded in four steps. The 10 ML annotations refer to LE, IT, CT, PAN, PNZ, HBV, MVP, NE, HBV/MVP, and LE/IT. All Features refers to LE, IT, CT, PAN, PNZ, HBV, MVP and NE. All Features Merged refers to CT, PAN, PNZ, NE, HBV/MVP, and LE/IT. Cohen's kappa (binary) is based on Question 1 input, and Cohen's kappa (Agreement ≥75%) is based on combining input from Question 2 and Question 3.

Step 1: Calculate sensitivity, specificity and Cohen's kappa.

Neuropathologists' input from Question 1 was collected, and true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) were calculated using the binary values as shown below.

	ML Identified (% area ≥ 1%)	ML Not Identified (% area < 1%)
Neuropathologist Identified	TP	FN
Neuropathologist Not Identified	FP	TN

Sensitivity = TP / (TP+FN)

Specificity = TN / (TN+FP)

These values were then used to calculate sensitivity, specificity and Cohen's kappa (binary) for All Features (LE, IT, CT, PAN, PNZ, HBV, MVP and NE) and All Features Merged (CT, PAN, PNZ, NE, HBV/MVP, LE/IT).

Step 2: Process online form data collected from neuropathologists

I = variable for value entered by neuropathologist for Question 2, "% of annotation labeling correct feature"

N = variable for value entered by neuropathologist for Question 3, "% of feature not labeled by correct annotation"

		ML	
		Identified (I)	***Not Identified (N)***
Neuropathologist	***Identified (I)***	I=V, N=V	I=NA, N=100%
	Not identified (N)	I=0%, N=NA	I=NA, N=NA
	V = value entered by user		

The data was processed using above table to appropriately use NA, 0% and 100% irrespective of neuropathologists' input.

Step 3: Generate neuropathologist %Agreement data

I = value entered by neuropathologist for Question 2, "% of annotation labeling correct feature"

N = value entered by neuropathologist for Question 3, "% of feature not labeled by correct annotation"

True positive (TP) = **I** = the amount of feature correctly identified by ML

False positive (FP) = **100 - I** = the amount of feature incorrectly identified by ML

True negative (TN) = **100 - N** = the amount of feature correctly identified as other feature

False negative (FN) = **N** = the amount of feature incorrectly identified as other feature

Accuracy (%Agreement) = (TP+TN) / (TP+TN+FP+FN) = (**I** + 100 - **N**) / (**I**+1-**N**+100-I+**N**) = (**I** + 100 - **N**)/2

Sensitivity = TP / (TP+FN) = I / (**I**+**N**)

Specificity = TN / (TN+FP) = (100-**N**) / (100 - **N** + 100 - **I**)

Example 1. **I** = 76-100%, **N** = 1-25%

worst-case

Accuracy = (76+75)/2

best-case

Accuracy = (100+99)/2

%Agreement range = 76-100%

Example 2. **I** = 51-75%, **N** = 26-50%

worst-case

Accuracy = (51+50)/2

best-case

Accuracy = (75+74)/2

%Agreement range = 50-75%

Step 4: Generate accuracy using multiple cutoffs

We used %Agreement data for calculating accuracy and Cohen's kappa (Agreement≥75%) using stringent, moderate and relaxed cutoffs (as shown below) for All Features (LE, IT, CT, PAN, PNZ, HBV, MVP and NE) and All Features Merged (CT, PAN, PNZ, NE, HBV/MVP, LE/IT). We also calculated accuracy, but not Cohen's kappa for Question 2 and Question 3 separately.

Q2 (% of annotation labeling correct feature)	Q3 (% of feature not labeled by correct annotation)	%Agreement	Cutoff
76-100%	0%	88-100%	Stringent

76-100%	1-25%	75.5-99.5%	Stringent
51-75%	0%	75.5-87.5%	Stringent
76-100%	26-50%	63-87%	Moderate
51-75%	1-25%	63-87%	Moderate
26-50%	0%	63-75%	Moderate
76-100%	51-75%	50.5-74.5%	Relaxed
51-75%	26-50%	50.5-74.5%	Relaxed
26-50%	1-25%	50.5-74.5%	Relaxed
1-25%	0%	50.5-62.5%	Relaxed
76-100%	76-100%	38-62%	
51-75%	51-75%	38-62%	
26-50%	26-50%	38-62%	
1-25%	1-25%	38-62%	
0%	1-25%	37.5-49.5%	
51-75%	76-100%	25.5-49.5%	
26-50%	51-75%	25.5-49.5%	
1-25%	26-50%	25.5-49.5%	
0%	26-50%	25-37%	
26-50%	76-100%	13-37%	
1-25%	51-75%	13-37%	
0%	51-75%	12.5-24.5%	
1-25%	76-100%	0.5-24.5%	
NA	76-100%	0%	
0%	NA	0%	
0%	76-100%	0-12%	

RNA isolation and RNA sequencing

Microdissected tissue was collected directly into RLT buffer from the RNeasy Micro PLUS kit (Qiagen Inc., Valencia, CA) with 1:100 β-mercaptoethanol diluation, per manufacturer's instructions. Samples were volume-adjusted with water to 75µl, vortexed, centrifuged, and frozen at -80°C. RNA samples, after following the manufacturer's

directions, were eluted in 14µl, and 1µl was run on the Agilent 2100 Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA) using the Pico assay. Due to low sample volume and incompatibility of the eluent with the Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE), samples were quantitated using the Bioanalyzer concentration output. This was done by running a 1ng/µl RNA standard on the same Pico chip and then dividing the sample concentration output by the output of the standard concentration. The average RNA Integrity Number (RIN) of all passed samples was 7.1. Samples were failed when the Bioanalyzer traces showed degraded 18S and 28S bands, with RINs typically lower than 4.5 failing. In most cases, 5ng of total RNA was used as the input amount for the library prep. 5ng total RNA was used as input into ClonTech SMARTer Ultra Low Input RNA Kit for Illumina Sequencing-HV (# 634820). 12 PCR cycles were used for amplification as suggested in the manufacturer's instructions (ClonTech SMARTer Kit Manual 120213). The Modified Nextera DNA sample preparation was used after step V.B of the ClonTech SMARTer kit, instead of Covaris shearing and instead of step VI in the ClonTech SMARTer kit. RNA Sequencing was done on Illumina HiSeq 2000, producing approximately 30M 50bp paired-end clusters per sample. In most cases, 5 samples per lane were run. RNA-Seq libraries were assessed for quality by yield and visual inspection of the presence, quality, and size of cDNA product on a Bioanalyzer. Initially, 11 samples failed (7 for no product and 4 for majority of product <500bp) the quality control criteria. However, upon a second attempt at synthesis, all 275 samples passed. One of the 275 samples was failed for low inter-array-correlation (IAC) and was excluded from the data set. The average concentration of samples that passed the criteria was 1582 pM.

RNA-Seq data alignment

The data generation, collection, alignment, and normalization is described in detail online at http://glioblastoma.alleninstitute.org/ in the Documentation tab. Raw read (fastq) files were aligned to the hg19 human genome sequence (*44*) with the RefSeq transcriptome version 54 (downloaded 8/25/2012 and updated by removing duplicate gene entries from the gtf reference file for consistency with the LIMS). For alignment, Illumina sequencing adapters were clipped from the reads using the fastqMCF program (*45*). After clipping,

the paired-end reads were mapped using RNA-Seq by Expectation-Maximization (RSEM) (*46*) using default settings except for two mismatch parameters: bowtie-e (set to 500) and bowtie-m (set to 100). RSEM aligns reads to known isoforms and then calculates gene expression as the sum of isoform expression for a given gene, assigning ambiguous reads to multiple isoforms using a maximum likelihood statistical model. Reads that did not map to the transcriptome were then aligned to the hg19 genome sequence using Bowtie with default settings (*47*). Reads that mapped to neither the transcriptome with RSEM nor to the genome with Bowtie were mapped against the ERCC sequences (in this project as a negative control). The final results files included quantification of the mapped reads (raw read counts, FPKM, and TPM values for the transcriptome-mapped reads, chromosome-wide counts for the genomic-mapped reads), BAM files including both transcriptome and genome-mapped reads, and fastq files for the unmapped reads. Anonymized BAM files (where sequence-level information has been removed) and gene-level quantification (TPM, FPKM, and number of reads) are available as part of the resource (see Download tab). Resulting FPKM values (normalized for gene length and sequencing depth) used for the analyses of this paper were further adjusted for the total transcript count using TbT normalization as described below.

RNA-Seq data normalization

In the Allen Human Brain Atlas, analysis of the RNA-Seq data showed minimal process batch effects but improvements in variability after normalization could be made (*48*), and therefore a comparable post-hoc data normalization strategy was used for this project. Gene expression values were summarized as transcripts per million (TPM) and fragments per kilobase per million (FPKM), as described above, both of which normalize read counts by gene length and for the total number of reads in slightly different ways. In this manuscript (as well as for the heatmaps shown on the website), the FPKM data matrix was further adjusted for the total transcript count using TbT normalization (*49*), which scales each sample based on the summed expression of all genes that are not differentially expressed. FPKM values were TbT normalized in linear space, with the differential expression vector defined as TRUE if a sample was from cellular tumor and

FALSE otherwise. Sample data was then scaled such that the total log2(FPKM) across the entire data set remained unchanged after normalization. The result of this step was that expression levels for all genes in a particular sample were multiplied by a scalar value close to 1 (in most cases between 0.7-1.3).

Differential expression between anatomic structures

Differentially expressed genes were identified for each pair of anatomic structures (except IT) using the 'edgeR' library (version 3.12.1) in R (version 3.2.5) (*50*), in combination with a fold-change threshold. First, genes with low or no expression (< 98 total counts, one per anatomic structure sample) were excluded from the data. Next, each pair of anatomic structures was compared with edgeR using the standard method and all default settings. Resulting p-values were corrected for multiple comparisons based on local false discovery rate using the "lfdr" function in the 'qvalue' library (version 2.2.2) (*51*) in R. Genes with FDR<0.01 and at least 2-fold average enrichment were considered differentially expressed between one anatomic structure and another. Markers genes for each structure were defined as genes with significant enrichment in that structure for all three other structures tested.

Admixture analysis

The proportions of CT, LE, MVP, and PAN in each tumor sample were estimated using "cell-type identification by estimating relative subsets of RNA transcripts" (CIBERSORT; https://cibersort.stanford.edu/) (*52*). This program estimates the proportions of different populations in a mixture using a novel application of linear support vector regression (SVR), which is robust to noise. The expression profile of a mixed sample is modeled as a linear combination of the expression profile of each population comprising that sample, in this case CT, LE, MVP, and PAN.

CIBERSORT takes as input a reference sample file with gene expression from pure cell populations, and a phenotype classes file indicating which sample corresponds to which population. Using these data, a signature gene file is derived, which contains average gene expression of marker genes for each population, and this signature is used to extract estimated proportions from any uploaded mixture file. Here, gene expression

signatures for each feature were calculated using the 98 samples dissected from distinct anatomic features (excluding the 24/122 from IT). After calculating these proportions, 15 samples with <75% purity for the histologically defined feature were excluded, and the above strategy was repeated with the remaining 83 samples. This iterative approach was taken because, although our anatomical structure data set was designed to specifically target these areas using laser microdissection, we do not have reference samples of known 100% purity. The success of targeting is quite good based on the dense staining of ISH markers for targeted regions as well as expression levels for marker genes of each anatomic feature (Fig. 2). Laser microdissection highly enriches for the desired cell populations, as 85% of samples have at least 75% purity for a particular anatomic structure.

These final gene signatures (table S4) were then applied to all expression profiles of LMD samples from this resource, as well as to bulk tumor tissues from TCGA. Resulting estimates of the proportions of each population in each heterogeneous sample are provided as table S15.

Multi-dimensional scaling

We used multi-dimensional scaling, a relative of principal components analysis, to define two-dimensional representations of samples collected for the anatomic features component of the study and for all samples using all genes. This was done by first converting gene expression data into logarithmic space and calculating the Pearson correlation across all genes for each pair of samples. A scaled correlation distance was defined as $\sqrt{(1-COR^2)}$, where COR is the correlation matrix described above. Finally, the first two principal coordinates and corresponding percent variances explained were calculated using the "cmdscale" function in R, with k=2.

Gene set enrichment analysis

ToppFun tool (*53*) was used to detect functional enrichment of all gene lists based on Ontologies (GO, Pathway). The tool was used with default settings and a p-value cutoff of 0.05.

Anatomic feature fractional area calculation

Anatomic feature area for each of the anatomic features for a given tumor was calculated by aggregating the grids labelled as the given anatomic feature by Mill for all tissue sections from that tumor and dividing it by total number of labelled grids for that tumor.

Anatomic feature specific expression of a gene

Anatomic feature specific expression of a gene for each of the features was calculated as median gene expression value for the subset of samples assigned as a given anatomic feature.

Molecular subtype calculation

The molecular subtype was calculated for each study using all samples. The set of 840 transcripts from the RNA-Seq data as per Verhaak et al., 2010 (6) was used to calculate four components corresponding to each subtype: 1. Neural, 2. Proneural, 3. Classical, and 4. Mesenchymal. The 840 genes were divided in to eight sets, four sets positively correlated with each subtype and four sets negatively correlated with each subtype. The following was done for each of the three datasets: 1. Ivy GAP – LMD, 2. Ivy – Bulk, and 3. TCGA – bulk. For each gene we obtained the average Δct (μ) and standard deviation (σ). For each tissue sample we calculate standard scores (z) for 840 genes as follows:

$$z_g = \Delta ct_g - \mu / \sigma_g \text{, where } g \in 840 \text{ genes}$$

Four components: 1. Neural, 2. Proneural, 3. Classical, and 4. Mesenchymal are calculated by taking the average of z scores of all genes belonging to the corresponding subtype gene set.

Neural =

$\mu(z_g) - \mu(z_h)$, where $g \in$ positively correlated with Neural, $h \in$ negatively correlated with Neural

Proneural =

$\mu(z_g) - \mu(z_h)$, where $g \in$ positively correlated with Proneural, $h \in$ negatively correlated with Proneural

Classical =

$\mu(z_g) - \mu(z_h)$, where g $\in$ positively correlated with Classical, h $\in$ negatively correlated with Classical

Mesenchymal =

$\mu(z_g) - \mu(z_h)$, where g $\in$ positively correlated with Mesenchymal, h $\in$ negatively correlated with Mesenchymal

Copy number data analysis

CRMA v2 vignette from http://www.aroma-project.org/vignettes/CRMAv2/ was used for calculation of raw copy numbers. This vignette involves the following steps that are described in detail at the website.

Step 1 - Calibration for crosstalk between allele probe pairs
Step 2 - Normalization for nucleotide-position probe sequence effects
Step 3 - Probe summarization
Step 4 - Normalization for PCR fragment-length effects
Step 5 - Calculation of raw copy numbers: For reference dataset we used the Affymetrix SNP 6.0 data from PBMCs of TCGA glioblastoma patient cohort.

We then used the non-paired Circular Binary Segmentation (CBS) (*54*) method for partitioning the copy number signals data set into segments of equal underlying copy number levels for segmentation. Details described at http://www.aroma-project.org/vignettes/NonPairedCBS/.

Integrated copy number and gene expression analysis

The Ivy GAP - LMD gene expression data set was log transformed and z-score normalized. For each copy number segment for a given sample S with values > 0.2 or < -0.2, the gene expression of all genes located in that segment was recalculated as an average of all z-scores of genes in that segment for sample S to obtain RNA-Seq derived copy number segment.

Anatomic feature study samples from the Ivy GAP – LMD data were chosen for this analysis. To calculate gene expression changes coordinated with copy number changes the following measure was used:

$\Sigma R_s / \Sigma C_g$, where

R is RNA-seq derived copy number for genes in set s,

C is Copy number value for genes in set g,

s ∈ all genes with (R > 0.1 and C ≥ 0.2) and (R < -0.1 and C ≤ -0.2),

g ∈ all genes located within the copy number segments for a given sample S with values > 0.2 or < -0.2

The same measure was also calculated after shuffling the copy number data tumor labels for 100 iterations.

SNV analysis

SNVs were calculated by using mpileup command from SAMtools (version 1.3.1) suite. The following parameters were used:

samtools mpileup -Aug -t DP,AD,SP -Q 10 -q 10 -d 100000 –f hg19.fa -r

-A, --count-orphans do not discard anomalous read pairs

-u, --uncompressed generate uncompressed VCF/BCF output

-g, --BCF generate genotype likelihoods in BCF format

-t -t, --output-tags LIST optional tags to output:

DP,AD,ADF,ADR,SP,INFO/AD,INFO/ADF,INFO/ADR

-f, --fasta-ref FILE faidx indexed reference sequence file (hg19 was used)

-q, --min-MQ INT skip alignments with mapQ smaller than INT

-Q, --min-BQ INT skip bases with baseQ/BAQ smaller than INT

-r, --region REG region in which pileup is generated

The output was then parsed for variants-only data using Bcftools, version 1.3.1 (https://samtools.github.io/bcftools/) with following command.

bcftools call -Amv

call SNP/indel calling

-A, --keep-alts keep all possible alternate alleles at variant sites

-m, --multiallelic-caller alternative model for multiallelic and rare-variant calling (conflicts with -c)

-v, --variants-only output variant sites only

To assess the accuracy of using RNA-seq data for variant calling we used TCGA dataset of 144 glioblastoma patient samples where both DNA-seq and RNA-seq data was

available. We limited our analysis to eight most frequently mutated genes in glioblastoma, *TP53*, *PTEN*, *EGFR*, *NF1*, *IDH1*, *PIK3CA*, *PIK3R1*, and *ATRX*. Mutations with at least five supporting reads were included. We found that 93% of the mutations identified in RNA-seq data were also present in DNA-seq data. Out of nine mutations not detected in DNA-seq data, eight of those were mutations in the *EGFR* gene and one in *TP53* gene.

	DNA-seq present	**DNA-seq absent**
All	93% (118)	7% (9)
without EGFR	99% (89)	1% (1)

On the other hand only 60% of mutations identified by DNA-seq were present in RNA-seq data. When we looked at which mutations were not detected by RNA-seq we found that most of them were frameshift and nonsense mutations as well as non-synonymous mutations in *EGFR*. Frameshift and nonsense mutations are likely to be subjected to nonsense-mediated mRNA decay and that might explain that 76% of such mutations are not detected in RNA-seq data as compared to only 24% of undetected non-synonymous single amino acid change mutations. Other potential reasons for lower sensitivity could be subclonal mutations (likely in case of *EGFR* mutations) and insufficient sequencing depth.

	RNA-seq present	**RNA-seq absent**
All	60% (127)	40% (85)
non-synonymous SNV	76% (111)	24% (35)
non-synonymous SNVs without EGFR	92% (81)	8% (7)

indel, frameshift, nonsense SNV	24% (16)	76% (50)

SIFT tool (http://sift.jcvi.org/) was used to identify somatic mutations vs. germline variations.

TCGA RNA-seq data is 75 bp paired-end reads generated using Illumina HiSeq with sequencing depths ranging from 27 to 126 million reads per sample which is more than double the Ivy GAP-LMD RNA-seq data with about 30M 50bp paired-end reads per sample. Also, the bulk tumor samples have better RNA quality compared to laser-captured micro-dissected samples. So we used an additional criterion to detect SNVs. Only SNVs that were present in at least one of the samples with ≥ 5 reads and MSE > 0.5 were considered. Two SNVs that were present in 13 and 26 tumors respectively were considered sequencing artifacts as the SNV region was low complexity region.

We recognize important intrinsic limitations in calling SNVs from RNA-Seq data. SNVs in genes that are expressed at low levels not covered by sufficient sequencing depth will be missed. This also results in missing out mutations that cause loss of expression. Potential problems calling variants from RNA-Seq data were mitigated by (i) restricting analysis to the frequently mutated genes *TP53*, *PTEN*, *EGFR*, *NF1*, *IDH1*, *PIK3CA*, *PIK3R1*, and *ATRX*, (ii) visualizing and validating each SNV using IGV viewer, (iii) demonstrating excellent concordance with DNA-seq variant calls and RNA-seq variant calls using our method on TCGA data, (iv) showing good exome coverage from the FPKM table, (v) considering only those SNVs that were present in at least one of the samples with ≥ 5 reads and MSE > 0.5 were considered, and (vi) comparing somatic mutations to germline mutations in the same set of genes across anatomic features. In total 93.8% of the 25,873 quantified genes show detectible expression (FPKM>0) in at least one of the 270 samples, and 79.8% of the genes show reliable expression (FPKM>1) in at least one of the 270 samples, while 58.8% of genes show reliable expression (average FPKM>1) in at least one anatomic feature. The number of reads per SNV call ranged from 2 to 1591, with an average of 110 and a median of 23 (table S13).

Single sample gene set enrichment analysis

Single-sample GSEA (ssGSEA), an extension of Gene Set Enrichment Analysis (GSEA) (*55*), calculates separate enrichment scores for each pairing of a sample and gene set. Each ssGSEA enrichment score represents the degree to which the genes in a particular gene set are coordinately up- or down-regulated within a sample. We used the web-based service hosted at the Broad Institute (http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/ssGSEAProjection/4) to calculate ssGSEA enrichment scores with default parameters and gene sets as provided in supplemental table S16. Scores for each gene set was z-score normalized.

Validation of RNA-Seq by anatomic feature ISH for enriched genes

ISH validation of 31 genes identified by RNA-Seq as enriched in CT, PAN, or MVP was performed using automated analysis of images from the Anatomic Features ISH for Enriched Genes Study. Of the 37 genes in the study, 6 were excluded from analysis: 5 "LE/IT" genes were chosen to be a part of the ISH study for reasons other than enrichment in LE or IT RNA-Seq samples, and 1 gene was missing from some of the tables. Expected RNA-Seq enrichment of 37 genes in the anatomic features is listed in Gene_lists_Table2_anat_fea_vali tab (also in Table 2 of Ivy Gene Lists white paper on the resource's website), and of the 31 analyzed genes in the ISH_gene_expression_summarizati tab, of table S3. Expression levels for ISH in each structure were quantified using expression energy (defined above in "Image signal unionization and search"), and are available on the "Download" page of the website. Specifically, the "Gene expression energy value for each anatomic structure in each tumor sub-block" file was joined with normalized values from the "Fractional area of anatomic structures in each tumor sub-block" file and presented in the followupISH tab of table S3. For each gene, average expression energy was calculated in each anatomic structure across all available ISH, considering only sections with at least 1% area covered by CT, PAN, and MVP (columns E-G of ISH_gene_expression_summarizati tab of table S3). In total 754 images were analyzed, or an average of 24 per gene. The structure (or structures) with the highest expression levels are then recorded for comparison with RNA-Seq in column "Enrichment (ISH)" of the ISH_gene_expression_summarizati tab.

Genes are defined as having “good agreement” between RNA-Seq and ISH when all of the structures with the highest expression levels match between methods (value 1), while “partial agreement” is when some (but not all) regions match (value 0.5), and no agreement is when no regions match (value 0), as listed in column “Agree?” of the ISH_gene_expression_summarizati tab.

SUPPLEMENTARY TABLES

Feature	Neuropathologist 1		Neuropathologist 2		Neuropathologist 3		Combined Accuracy (%)
	Agree	Disagree	Agree	Disagree	Agree	Disagree	
CT	30	0	29	1	30	0	99%
IT	23	0	22	1	23	0	99%
LE	19	0	19	0	19	0	100%
MVP	22	3	16	9	25	0	84%
PAN	24	0	21	3	24	0	96%
TOTAL:	118	3	107	14	121	0	95%
Combined Accuracy (%)	98%		88%		100%		
Concordance (κ)	0.97		0.86		1.00		

Table S2. Neuropathologist-determined accuracy and concordance of manual annotations supporting laser-microdissected anatomic features in atlas. Three neuropathologists (P.J.C., C.D.K., and M.U.) were presented with all the images of the H&E-stained tissue sections that had been manually annotated with laser microdissection (LMD) guide lines under the supervision of a fourth neuropathologist (S.W.R.). An online survey form guided neuropathologists to the 47 sub-blocks from the 10 tumors with the LMD guidelines for the 122 samples isolated for the Anatomic Structures RNA-Seq study at http://glioblastoma.alleninstitute.org/, and recorded each response with a timestamp. Multiple tissue sections of each sub-block contained guide lines drawn on 1-5 anatomic features, which were easily-identified classic examples of the features. Each neuropathologist independently answered two questions for each sample's guide lines across multiple images of a sub-block. In reply to "What % of a particular annotation labels the correct feature?" the neuropathologist selected 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100%. Selecting 70% meant that about 70% of the annotated area for a particular feature (e.g. PAN) labeled the feature correctly throughout the images of the sub-block, and 30% of the annotation labeled another feature incorrectly. In reply to "What is the quality of the sub-block?" the neuropathologist selected High, Medium, Low, or

Unsatisfactory technically (not possible to score accuracy of annotations). The replies for the 10% scale were converted to a binary output, Agree/Disagree, using a cutoff of 80%, with the result that 80%, 90%, and 100% replies were converted to Agree (i.e. the feature had the correct LMD guide line), and all replies less than 80% were converted to Disagree (i.e. the feature had the incorrect LMD guide line). Low quality of the sub-block was rare, and was not correlated with Agree/Disagree outcome**.** Cohen's kappa **(*κ*)** (*42*) was calculated across all features for each neuropathologist as a measure of concordance, and the combined accuracy was calculated across all neuropathologists for each feature, an approach consistent with other neuropathology concordance analyses (*43*). Neuropathologists' data and confusion matrix for Cohen's kappa calculations are available in table S9.

	LE	CT	IT	NE	PAN	MVP	HBV	ED	ICE	FOLD	SPA	EN	Total	Recall
LE	101336	40	797	158	0	0	7	0	9	13	2	65	102427	98.93%
CT	205	306309	1496	1419	2095	806	1815	0	27	47	2	1724	315945	96.95%
IT	1269	437	152562	148	18	3	55	0	36	17	0	263	154808	98.55%
NE	318	2589	557	215087	732	54	877	0	9	93	64	3312	223697	96.15%
PAN	2	1843	29	212	133975	229	725	0	16	57	0	240	137328	97.56%
MVP	2	754	10	23	223	56567	259	0	0	19	4	139	58000	97.53%
HBV	11	2080	64	193	701	203	136295	0	12	18	0	190	139767	97.52%
ED	0	0	0	0	0	0	0	750	0	0	0	0	750	100.00%
ICE	0	7	8	19	7	0	4	0	1455	0	0	0	1500	97.00%
FOLD	7	87	9	46	31	6	15	0	0	304423	0	7	304631	99.93%
SPA	2	3	0	31	0	0	0	0	1	0	310443	14	310494	99.98%
EN	168	1310	445	1776	300	87	195	0	0	5	28	178379	182691	97.64%
Total	103320	315459	155977	219112	138082	57955	140247	750	1565	304697	310541	184333	1932038	
Precision	98.08%	97.10%	97.81%	98.16%	97.03%	97.61%	97.18%	100.00%	92.97%	99.91%	99.97%	96.77%		
													ACCURACY	98.22%

Table S5. Machine learning-determined accuracy and precision of machine learning annotations of anatomic features in atlas. About 12,000 images corresponding to 300 tissue blocks from 42 tumors were labeled by ML with expert supervision. Accuracy and precision of the labeling process by the Decision Forest algorithm (*40*) are represented in confusion matrix table. The Decision Forest trees were not trained on the samples that were tested so this kind of 'holdout' testing is a reasonably robust measure of the accuracy of the algorithm. In the matrix the rows represent the points tested and the columns the test results. For example, the entry in the second row of the leftmost column indicates 205 samples of tissue expertly labeled LE, were labeled incorrectly as CT by the algorithm. Precision summaries (percentage of results correctly labeled) and Recall summaries (percentage of source samples correctly labeled) as well as the overall accuracy (percentage of all results correct) are shown. It should be noted that the metrics represented in the matrix were used as a performance guide for the algorithm only, and final labeling was subject to expert examination in consultation with a board-certified neuropathologist (S.W.R.), and manual correction. All abbreviations: LE (leading edge), CT (cellular tumor), IT (infiltrating tumor), NE (necrosis), PAN (pseudopalisading cells around necrosis), MVP (microvascular proliferation, HBV (hyperplastic blood vessel), ED (edema), ICE (freezing artifact), FOLD (fold in tissue), SPA (space without tissue), and EN (early necrosis).

Feature: none merged	Neuropathologist 1		Neuropathologist 2		Neuropathologist 3		Combined Accuracy (%)
	Agree	Disagree	Agree	Disagree	Agree	Disagree	
CT	8	5	13	0	12	1	85%
HBV	4	8	9	3	2	10	42%
IT	8	3	11	0	10	1	88%
LE	7	6	10	3	9	4	67%
MVP	11	1	12	0	11	1	94%
NE	9	1	6	4	8	2	77%
PAN	8	3	7	4	8	3	70%
PNZ	8	4	8	4	5	7	58%
TOTAL:	63	31	76	18	65	29	72%
Combined Accuracy (%)	67%		81%		69%		
Concordance (*κ*)	0.62		0.78		0.65		

Table S6. Assessment of inter-neuropathologist variability without merging LE and IT, and HBV and MVP. Three neuropathologists (P.J.C., C.D.K., and M.U.) were presented with 94 test images representing various examples of each H&E-stained feature, and each independently identified the main feature in each image using the definitions developed for training the ML algorithm (fig. S1). Each neuropathologist's label for a particular image was compared to the ground truth for that image, and assigned Agree if they were the same and Disagree if they were different. Ground truth was established by the investigator who created a reference set of images with a fourth neuropathologist (S.W.R.) who guided the training of the Decision Forest for the ML algorithm. Distinguishing LE from IT, and HBV from MVP, were especially

challenging, and were analyzed independently here, or as merged features in Table S8. Cohen's kappa **(κ)** (*42*) was calculated across all features for each neuropathologist as a measure of concordance, and the combined accuracy was calculated across all neuropathologists for each feature, an approach consistent with other neuropathology concordance analyses (*43*). Neuropathologists' data and confusion matrix for Cohen's kappa calculations are available in table S10.

Feature: some merged	Neuropathologist 1		Neuropathologist 2		Neuropathologist 3		Combined Accuracy (%)
	Agree	Disagree	Agree	Disagree	Agree	Disagree	
CT	8	5	13	0	12	1	85%
HBV/MVP	22	2	24	0	19	5	90%
LE/IT	22	2	23	1	22	2	93%
NE	9	1	6	4	8	2	77%
PAN	8	3	7	4	8	3	70%
PNZ	8	4	8	4	5	7	58%
TOTAL:	77	17	81	13	74	20	82%
Combined Accuracy (%)	82%		86%		79%		
Concordance (*κ*)	0.78		0.83		0.74		

Table S7. Assessment of inter-neuropathologist variability with merging of LE and IT, and HBV and MVP. Same data set as table S6, but for analysis the neuropathologists' labels of LE or IT images were merged into one measure, LE/IT, and labels of HBV or MVP images were merged into one measure, HBV/MVP, to demonstrate the degree to which concordance increases as a result of merging these features that are difficult to readily distinguish from one another. Neuropathologists' data and confusion matrix for Cohen's kappa calculations are available in table S10.

Feature		Neuropathologist 1					Neuropathologist 2					Neuropathologist 3					Ave Agreement (≥ 75 %)
		Sensitivity	Specificity	Accuracy			Sensitivity	Specificity	Accuracy			Sensitivity	Specificity	Accuracy			
				Q1 (> 75 %)	Q2 (< 26 %)	Agreement (≥ 75 %)			Q1 (> 75 %)	Q2 (< 26 %)	Agreement (≥ 75 %)			Q1 (> 75 %)	Q2 (< 26 %)	Agreement (≥ 75 %)	
All Features		0.85	0.98	0.76	0.78	0.72	0.94	0.98	0.78	0.90	0.81	0.95	0.95	0.91	0.98	0.94	0.82
All Features Merged		0.93	0.96	0.85	0.86	0.80	0.97	0.96	0.78	0.94	0.84	0.96	0.91	0.93	0.98	0.95	0.86
LE		0.81	0.98	0.71	0.69	0.60	0.91	0.98	0.90	0.88	0.84	0.93	0.97	0.96	0.97	0.93	0.79
IT		0.92	1.00	0.86	0.88	0.84	0.92	0.96	0.98	0.92	0.90	1.00	1.00	0.80	0.96	0.80	0.85
CT		1.00	1.00	0.97	0.95	0.92	1.00	1.00	0.95	0.97	0.95	0.99	0.67	0.92	0.98	0.96	0.94
PAN		0.79	0.90	0.75	0.66	0.63	0.90	0.93	0.72	0.88	0.73	0.86	0.91	0.95	1.00	0.95	0.77
PNZ		0.96	0.96	0.81	0.92	0.76	0.97	0.96	0.78	0.94	0.90	1.00	0.84	0.95	1.00	0.97	0.88
HBV		0.66	0.98	0.48	0.58	0.38	0.78	1.00	0.46	0.74	0.50	0.85	0.95	0.94	0.94	0.94	0.61
MVP		0.39	1.00	0.33	0.37	0.34	0.94	1.00	0.44	0.76	0.53	0.88	0.99	0.93	0.94	0.93	0.60
NE		0.97	1.00	0.86	0.90	0.82	0.97	1.00	0.66	0.95	0.79	0.96	0.95	0.88	1.00	0.96	0.86
LE-IT		0.98	0.98	0.98	0.94	0.94	0.96	0.93	0.92	0.96	0.94	1.00	0.98	0.92	0.96	0.92	0.93
HBV-MVP		0.81	0.97	0.70	0.69	0.65	0.96	1.00	0.55	0.91	0.68	0.94	0.91	0.94	0.94	0.94	0.75
κ (binary)	All Features	0.90					0.94					0.93					
	All Features Merged	0.91					0.96					0.94					
κ (Agreeme	All Features	0.66					0.81					0.92					

nt ≥75 %)	All Features Merged	0.76	0.87	0.93	

Table S8. Neuropathologist-determined accuracy and concordance of machine learning annotations of anatomic features in atlas. The 10 ML annotations reviewed include LE, IT, CT, PAN, PNZ, HBV, MVP, NE, HBV/MVP, and LE/IT. All Features refers to LE, IT, CT, PAN, PNZ, HBV, MVP and NE. All Features Merged refers to CT, PAN, PNZ, NE, HBV/MVP, and LE/IT. Cohen's kappa (binary) is based on Question 1 input, and Cohen's kappa (Agreement ≥75%) is based on combining input from Question 2 and Question 3. The neuropathology concordance analyses showed overall combined accuracy was greater than 70% for all neuropathologists, while accuracy for individual features varied from 34% for MVP to as high as 96% for CT. Neuropathologists' input data and confusion matrices for each validation are available in tables S9-S11.

Three neuropathologists (P.J.C., C.D.K., and M.U.) were presented with a random sampling of the ~12,000 ML-annotated images of all H&E-stained tissue sections of the atlas. Of the 946 sub-blocks from 42 tumors used to create the atlas, 95 that represented the full range of complexity or heterogeneity in the atlas were randomly chosen such that 10% of the sub-blocks at each level of complexity with 1, 2, 3,… or 8 anatomic features (LE, IT, CT, NE, PNZ, PAN, HBV, and MVP) were included in the review. The ML annotations were far more complex than the manual annotations of classic examples of features because the algorithm required labeling of every pixel in every image as one of the anatomic features, even when it was extremely challenging for an expert to distinguish for example LE from IT, or HBV from MVP. Each neuropathologist reviewed the 3rd image of the 10-18 H&E images of each sub-block, the 3rd likely being the least accurately annotated (worst case scenario) given the physical distance from the images used for training the ML Decision Forest for each sub-block, i.e. the 1st, 6th, 12th, and 15th images.

Guided by an online survey form that was equipped to record each response for the 95 sub-blocks presented from http://glioblastoma.alleninstitute.org/, each neuropathologist independently answered four questions for each of 10 features (LE, IT, CT, NE, PNZ, PAN, HBV, MVP, HBV/MVP, and LE/IT). The survey form included for each sub-block a bar graph showing the fractional area labeled for each feature calculated on the basis of ML annotations. In reply to Question 1, "Is the feature identified?" the neuropathologist selected Identified or Not Identified. In reply to Question 2, "What percentage of a particular annotation labels the correct feature?" the neuropathologist selected NA, 0%, 1-25%, 26-50%, 51-75%, or 76-100%. Selecting 26-50% meant that about 26-50% of the annotated area for a particular feature (e.g. PAN) labeled the feature correctly throughout the image, and therefore 50-74% of the annotation labeled another feature incorrectly. In reply to Question 3, "What percentage of the feature is not labeled by the correct annotation?" the neuropathologist selected NA, 0%, 1-25%, 26-50%, 51-75%, or 76-100%. Selecting 26-50% meant that about 26-50% of the feature (e.g. PAN) was not labeled by the correct (e.g. PAN) annotation, and therefore this fraction of the feature was mistakenly labeled as another feature, or not labeled at all. The neuropathologist selected NA for a particular feature when the ML scored it 1-100%, but the neuropathologist scored it Not Identified, or when the ML scored it 0%, and the neuropathologist scored it Identified or Not Identified. In reply to Question 4, "What is the quality of the sub-block?" the neuropathologist selected High, Medium, Low, or Unsatisfactory technically (not possible to score accuracy of annotations).

Input from the neuropathologists was processed and analyzed, and the replies for Question 1 - Question 3 were treated independently (Binary) or combined (Agreement, Materials and Methods). For each neuropathologist and anatomic feature, including the merged features LE/IT and HBV/MVP, sensitivity, specificity, and accuracy were calculated. To calculate accuracy for Question 2, the replies for the quartile scale were converted to a binary output, Agree/Disagree, with a cutoff of 75%, with the result that a reply of 76-100% was converted to Agree, meaning that the neuropathologist agreed that the ML annotation labeled the correct feature, and all replies less or equal to 75% were converted to Disagree. To calculate accuracy for Question 3, the replies for the quartile scale were converted to a binary output, Agree/Disagree, with a cutoff of 26%, with a

result that a reply of 0% or 1-25% was converted to Agree, meaning that the neuropathologist agreed that the ML labeled most of that particular feature correctly, and all replies greater or equal to 26% were converted to Disagree. To calculate accuracy for the combined results for Question 2 and Question 3, output from the Agreement formula (Materials and Methods) greater or equal to 75% was converted to Agree, meaning that the neuropathologist agreed that ML annotations were accurate, and values less than 75% were converted to Disagree, meaning that the ML annotations were not accurate. Overall combined accuracy was greater than 70% for all neuropathologists with accuracy for individual features varying from 34% for MVP to as high as 96% for CT. Low quality of the sub-block was rare, and was not correlated with Agree/Disagree outcome**.** Cohen's kappa **(*κ*)** (*42*) was calculated across all features for each neuropathologist as a measure of concordance as done for other neuropathology concordance analyses (*43*) using the binary data from Question 1 or the combined data (Agreement$\geq$75%) from Question 2 and Question 3. Neuropathologists' data and confusion matrix for Cohen's kappa calculations are available in table S11.

SUPPLEMENTARY FIGURES

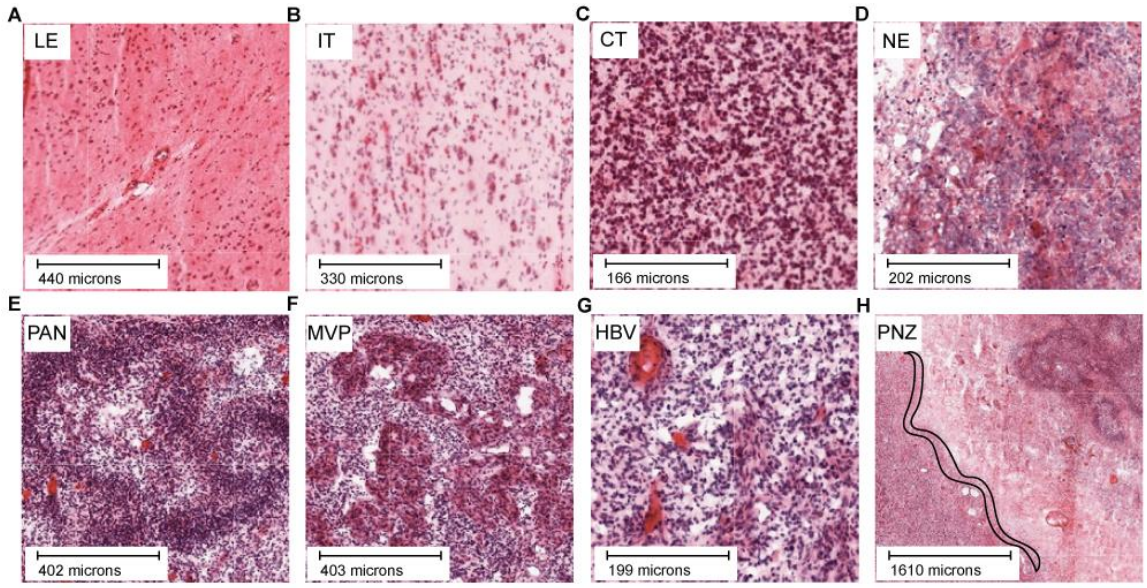


Fig. S1. Histologically-distinct anatomic features. All sections were stained with H&E. (**A**) Leading Edge (LE) is the outermost boundary of the tumor, where the ratio of tumor to normal cells is about 1-3/100, and the laminar architecture of the cortical layers is frequently evident. (**B**) Infiltrating Tumor (IT) is the intermediate zone between the Leading Edge (LE) and Cellular Tumor (CT), and is frequently marked by perineuronal satellitosis. (**C**) Cellular Tumor (CT) constitutes the major part of core, where the ratio of tumor cells to normal cells is about 100/1 to 500/1. (**D**) Necrosis (NE) is dead or dying tissue, marked by presence of karyorrhectic or cellular debris and absence of crisp cytological architecture. (**E**) Pseudopalisading Cells around Necrosis (PAN) is the narrow boundary of cells arranged like pseudopalisades along the perimeter of NE in the core. The extent of these constituents varies from little NE and high density of pseudopalisading cells in immature PAN to extensive NE and low density or faint ridges of pseudopalisading cells in mature PAN. (**F**) Microvascular Proliferation (MVP) refers to two or more blood vessels sharing a common vessel wall of endothelial and smooth muscle cells typically in the core, and arranged in the shape of a glomerulus or garland of multiple interconnected blood vessels. (**G**) Hyperplastic Blood Vessels (HBV) are blood vessels with thickened walls found anywhere in a tumor. (**H**) Perinecrotic zone (PNZ) refers to a boundary of tumor cells in the tumor core along the edge of necrosis that lacks a clear demarcation of PAN (PNZ depicted in black outline). These definitions were

developed with neuropathologists (P.J.C., C.D.K., M.U., and S.W.R.) for establishing a common framework for concordance analyses and for training of Decision Forest to support ML, and may deviate from WHO definitions.

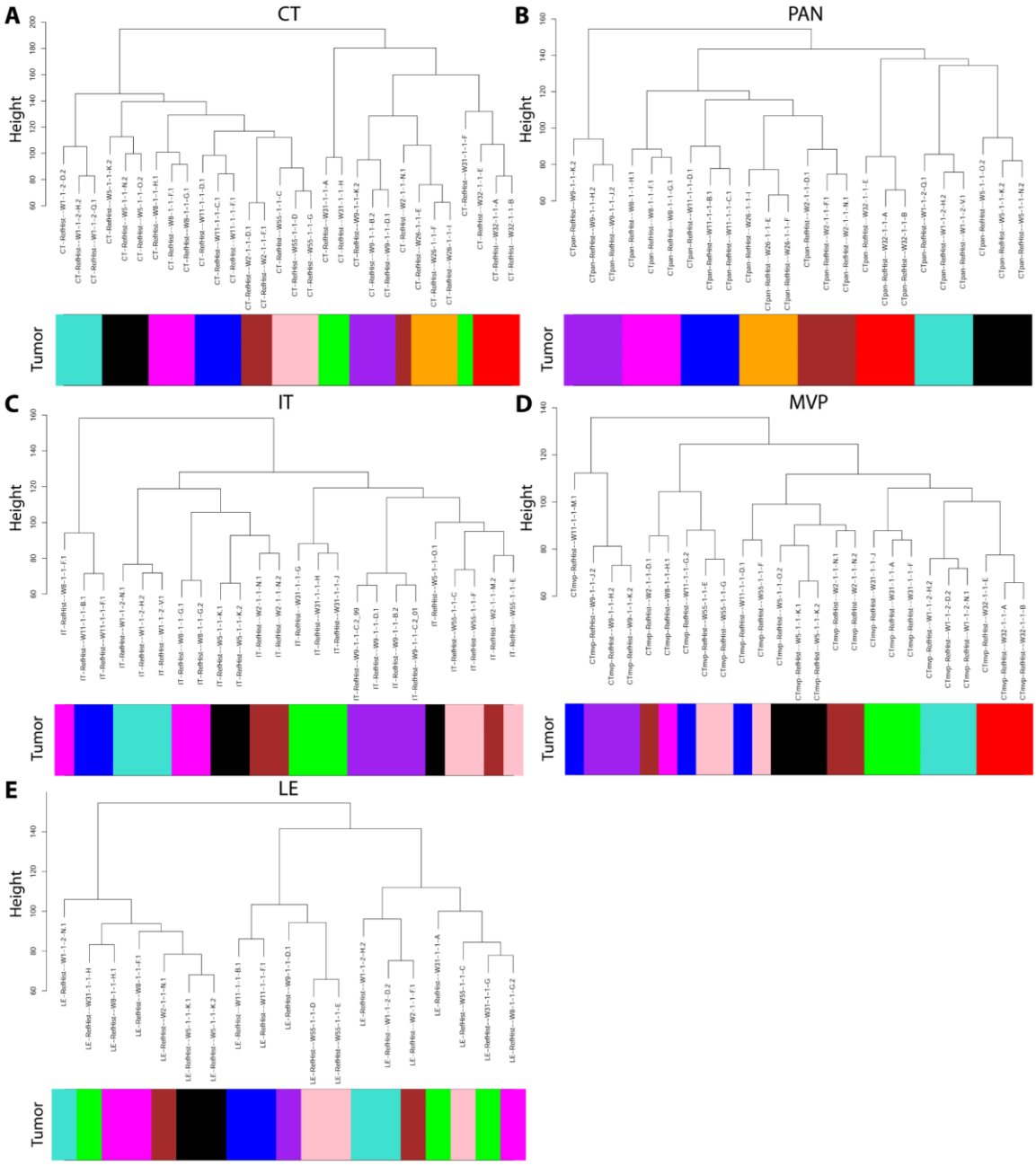


Fig. S2 Hierarchical clustering of anatomic feature RNA-Seq samples. Hierarchical clustering of all 122 anatomic feature RNA-Seq for CT (**A**), PAN (**B**), IT (**C**), MVP (**D**), and LE (**E**) using Euclidean distance between samples based on all genes as the distance measure. Grouping of samples by tumor (color bar) indicates intratumor variability is less than intertumor variability, particularly for CT and PAN.

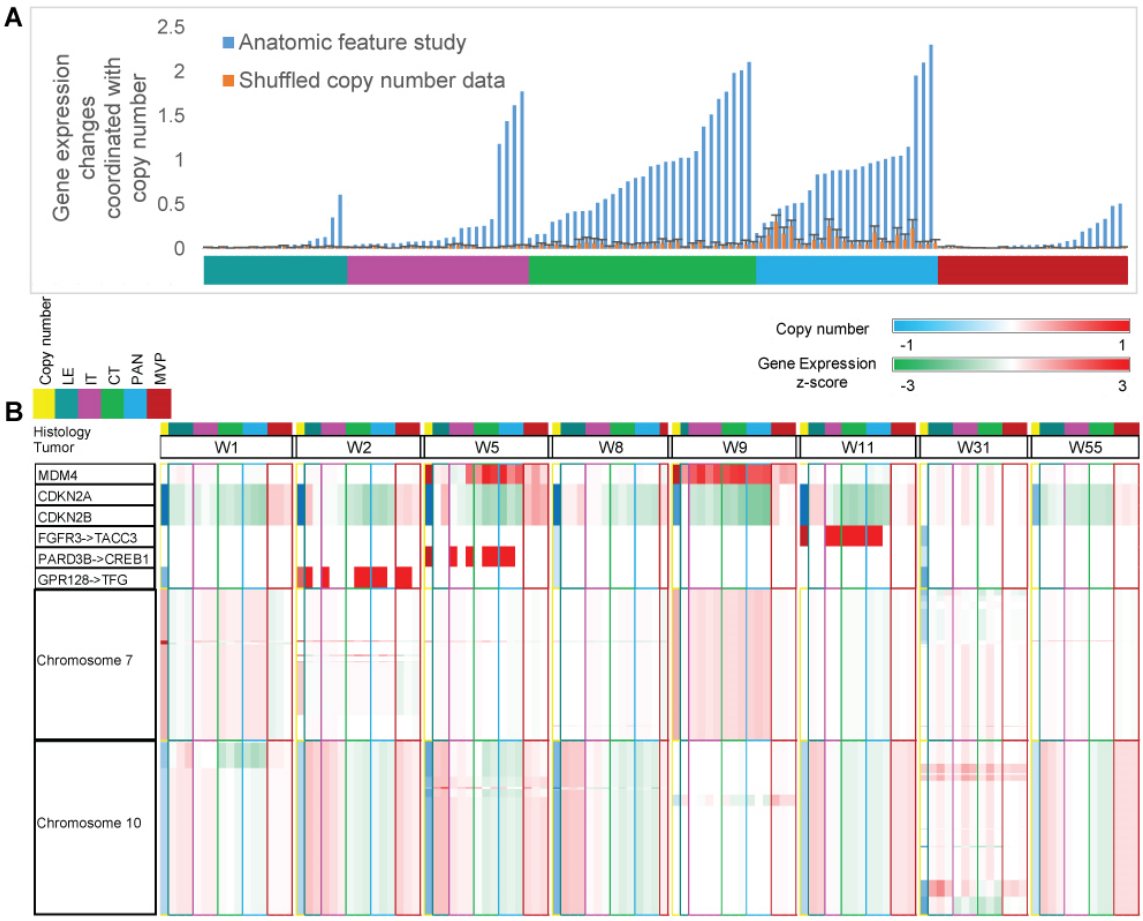


Fig. S3. Effect of copy number changes on gene expression across anatomic features. (**A**) Gene expression changes for each sample from the anatomic feature study coordinated and normalized to copy number changes from the matched tumor. For 39 of 42 tumors, genomic DNA was isolated from bulk tissue for generating copy number data using Affymetrix SNP 6.0 array, amplification and deletion segments at all loci with detectable alterations were identified (table S12), and the expression of genes located within a copy number segment was averaged. If the average expression was positive for amplification segments and negative for deletion segments then it was counted as gene expression change coordinated with copy number change. LE samples showed minimal gene expression changes coordinated with copy number changes, IT samples showed increased changes, CT and PAN samples showed significantly higher changes, and MVP samples showed changes slightly higher than LE samples but considerably lower than CT and PAN samples. For assessing statistical significance, a shuffled dataset was generated

by shuffling the tumor labels for the copy number dataset, and gene expression changes were calculated, which was done for 100 iterations. (**B**) Examples of frequently amplified *MDM4* gene, frequently deleted *CDKN2A/2B* gene, and *FGFR3→TACC3* gene fusion, frequently gained chromosome 7, and frequently lost chromosome 10. Consistent with (**A**), CT samples showed gene expression changes coordinated with copy number changes, and LE samples showed gene expression not impacted by copy number changes. PAN samples exhibited changes similar to CT samples, while MVP samples exhibited changes in between CT and LE samples, which might reflect a mixture of normal cells and tumor-derived cells, e.g. tumor derived vascular pericytes (*56*), or from inaccuracies inherent in large-scale LMD resulting in inclusion of nearby tumor cells. The *MDM4* gene was amplified in tumor W5 (top row), which is reflected by overexpression of the genes in that region in CT and PAN. MVP samples also showed moderate overexpression, but LE samples showed no differential gene expression in that region of W5. A similar trend was evident for the fusion gene (*PARD3B→CREB1*) resulting from tandem duplication on chromosome 2, fusion gene (*FGFR3→TACC3*) on chromosome 4, *EGFR* amplification on chromosome 7, and the *CDKN2A/2B* locus on chromosome 9. Assessments of other tumors showed similar effects of tumor-associated copy number changes on anatomic features. In contrast, a germline fusion *GPR128→TFG* on chromosome 3 (*57*) showed overexpression in all anatomic features of W2.

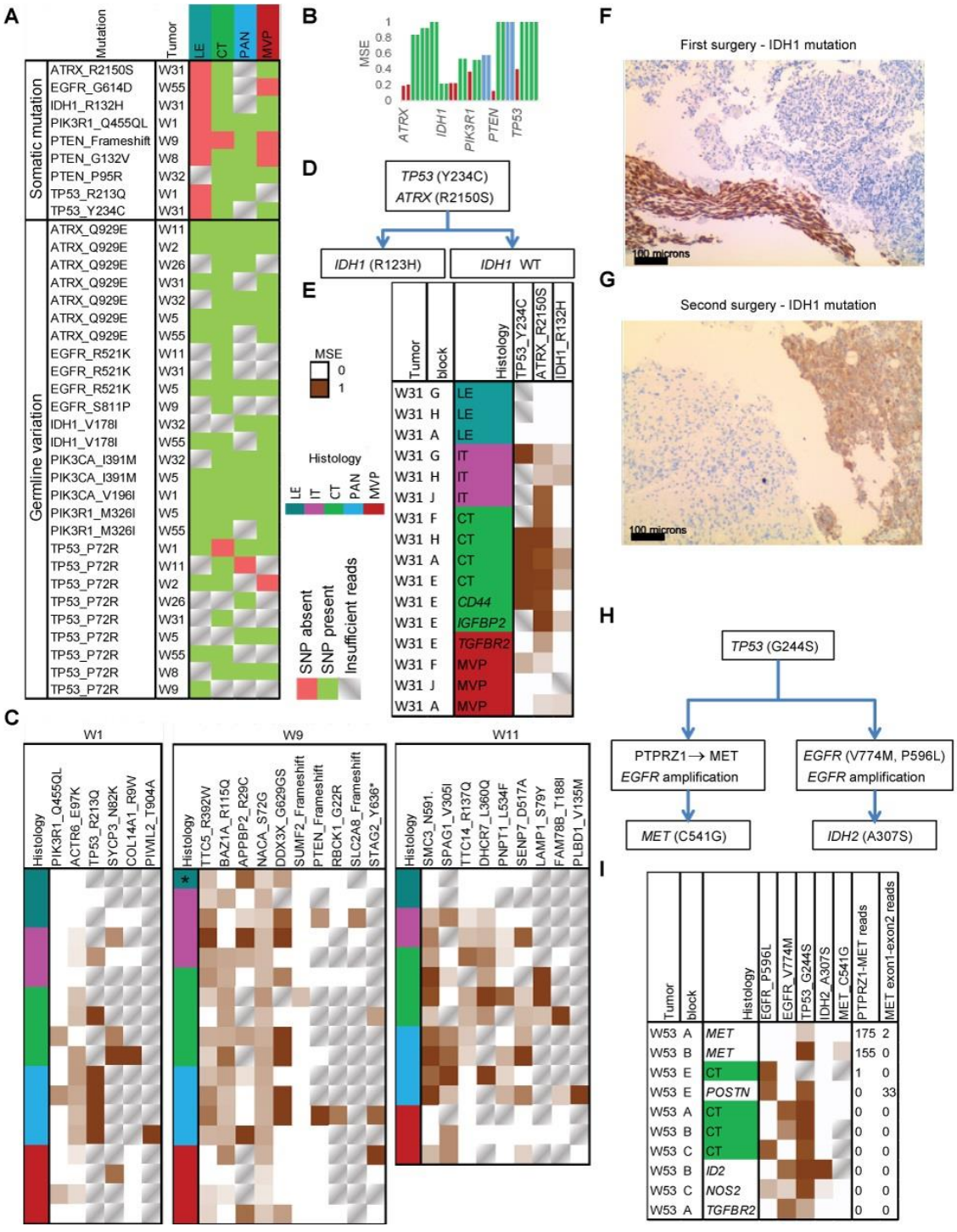
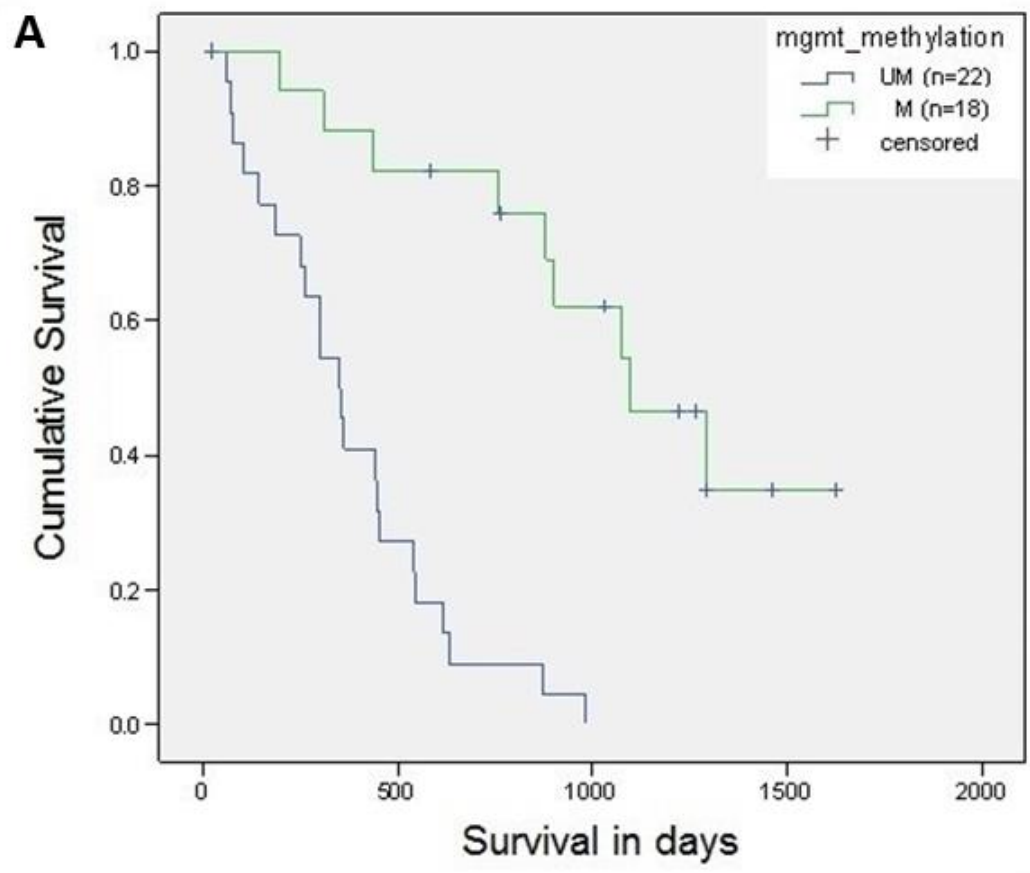


Fig. S4. SNV profile across anatomic features and exemplary cases of intratumor heterogeneity of clinically relevant mutations. Mutated allele specific expression (#SNV reads/ #total reads). We recognize important intrinsic limitations in calling SNVs from RNA-Seq data. SNVs in genes that are expressed at low levels not covered by

sufficient sequencing depth will be missed. This also results in missing out mutations that cause loss of expression (Supplementary Materials, Methods). (**A**) Of eight genes (*TP53, PTEN, EGFR, ATRX, IDH1, NF1, PIK3R1, PIK3CA*) that are often recurrently mutated in glioblastoma in the TCGA glioblastoma data set (*58*), 45 unique non-synonymous SNVs in 28 of the tumors (table S13) were detected in the LMD/RNA-Seq dataset. The number of high-confidence SNVs (>3 reads/sample) varied by anatomic feature. Reflecting the heterogeneity in copy number and gene expression data (fig. S3), we detected somatic mutations in CT, PAN and MVP samples but not in LE samples. In contrast germline mutations in the same set of genes were detected in all anatomic features. (**B**) The mutated allele specific expression for SNVs found in MVP samples was lower compared to the CT samples from the same tumor. (**C**) Three of the tumors from Ivy cohort with samples from all anatomic features were also part of the TCGA cohort with DNA-seq data available. Mutations identified in both TCGA DNA-seq and Ivy GAP RNA-seq data are shown. The mutated allele specific expression varied widely across samples from the same tumor for a majority of the SNVs, suggesting that many of the SNVs are sub-clonal, a result consistent with a previous report (9). It also shows the trend described for panels **A** and **B**. Some of the somatic mutations for tumor W9 were detected in the LE sample from W9 denoted by an asterisk. This same sample also has the highest amount of copy number coordinated with gene expression changes, and the CIBERSORT algorithm predicts this sample to have 40% CT fraction. (**D**) As an example of a tumor with a sub-clonal mutation, W31 has SNVs in three genes, *TP53, ATRX* and *IDH1*, which are more frequently mutated in lower grade gliomas (*59*) than glioblastoma and are associated with better survival. However, a fraction of the CT samples with SNVs in *TP53* and *ATRX* have wildtype (WT) *IDH1*, which is associated with poorer survival rates. (**E**) In tumor W31, the mutated allele specific expression for *TP53* and *ATRX* are highest for CT samples, which are enriched for the greatest number of tumor cells, and lowest for LE, which corresponds to near normal tissue. (For block F, >95% of cells in CT samples stain positively for TP53). (**F, G**) As further evidence for the sub-clonal mutation of IDH1 in W31, immunostaining for IDH1 mutant protein shows a mixture of labeling ranging from negative to strong diffuse staining, and this heterogeneous staining pattern persisted at recurrence, which represents a good example

of heterogeneity that could alter the patient's eligibility in clinical trials depending on the staining outcome at the biopsy site. In contrast, tissue sections from the first and second surgery show robust, uniformly-diffuse, labeling of TP53 mutant protein, with ~95% or 99% of the cells staining positively, respectively (not shown). (**H**) W53, has SNVs in *TP53, EGFR, IDH2,* and *MET*, as well as evidence of *EGFR* amplification and the previously described *PTPRZ1→MET* fusion (*60*). The possible clonal evolution for this tumor is presented, starting with *TP53* SNV (G244S) and then diverging into two sub-populations; one with *PTPRZ1→MET* fusion and another with *EGFR* V774M, P596L SNV. One of the samples with a *PTPRZ1→MET* fusion has a SNV in the *MET* gene (C541G), whereas another sample with the *EGFR* V774M SNV acquires a mutation in *IDH2* (A307S) gene. (**I**) W53 sub-clonal SNVs that support the schematic presented in **G**. In total 93.8% of the 25,873 quantified genes show detectible expression (FPKM>0) in at least one of the 270 samples, and 79.8% of the genes show reliable expression (FPKM>1) in at least one of the 270 samples, while 58.8% of genes show reliable expression (average FPKM>1) in at least one anatomic feature. The number of reads per SNV call ranged from 2 to 1591, with an average of 110 and a median of 23 (table S13).



B

MGMT methylation	Median			
	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
UM	351.000	48.311	256.310	445.690
M	1096.000	190.142	723.321	1468.679
Overall	540.000	106.747	330.776	749.224

Log Rank (Mantel-Cox) p < 0.0001

Fig. S5. *MGMT* methylation status and survival. (**A**) Kaplan-Meier estimation of survival according to O-6-methylguanine-DNA methyltransferase (*MGMT*) promoter methylation status determined using quantitative methylation-specific PCR (MSP) or methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA) tests. Methylation status was positive (M) if either MSP or MS-MLPA test was positive as described in (*61*) and status was negative (UM) if both tests were negative. (**B**) The median survival for the methylated group was superior at 1096 days vs. 351 days for the unmethylated group (P<0.0001).

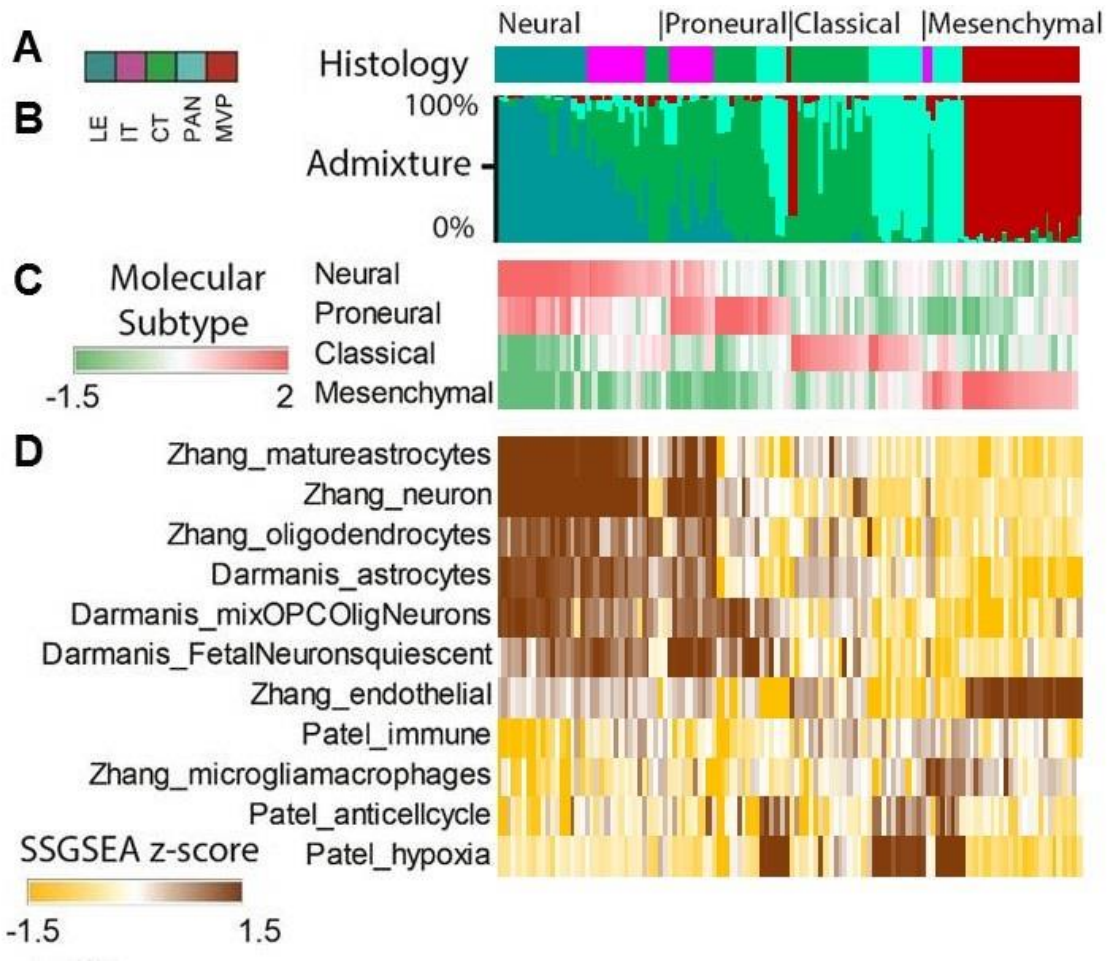


Fig. S6. Classification of Ivy GAP anatomic feature samples by histology, admixture, subtype, and cell types. Classification of 122 Ivy GAP anatomic feature LMD/RNA-Seq samples by (**A**) histology, (**B**) admixture analysis showing homogeneity of samples with respect to anatomy, (**C**) molecular subtype, and (**D**) published cell type and/or phenotype gene sets determined from single cells or cell types (*12*, *62*, *63*). Samples in the figure were sorted by subtype (neural, proneural, classical, mesenchymal), histology, and then subtype value. List of gene set signatures, detailed names, and references is provided in table S15. Fractional composition determined with CIBERSORT (*52*). Gene enrichment determined with single sample Gene Set Enrichment Analysis (ssGSEA). Classifications, and correlations between anatomic feature gene sets and molecular subtypes, calculated in table S16.

Each of the 122 Ivy LMD samples was computationally reduced to its anatomical makeup by admixture analysis, and validated by histology. The admixture model correctly matched the majority proportion to the corresponding histological label of LE, CT, PAN and MVP for 97% of the samples. We calculated the subtype of the Ivy LMD and TCGA bulk samples using the previously identified 840 gene set (6). This analysis revealed a striking relationship between anatomic feature and molecular subtype: the mesenchymal component predominated in HBV, MVP, and to a lesser extent in PAN and PNZ samples; the neural component predominated in the LE and IT samples; and the classical and proneural components predominated in CT and PAN samples.

Endothelial cell gene sets (i.e., Zhang_endothelial) were highly enriched in MVP and HBV of the mesenchymal subtype. Immune signatures, including gene sets representing microglia and macrophages, were highly enriched in CT, PAN, PNZ, and HBV of the mesenchymal subtype. Hypoxia and anti-cell cycle gene sets (e.g. Patel_hypoxia) were highly enriched in all PNZ and PAN samples, including some CT samples with minority fractions of PNZ and PAN of the mesenchymal, classical, proneural subtypes. Gene sets for neurons, astrocytes, oligodendrocytes, and quiescent fetal neurons at lower levels were enriched in LE and IT samples of the neural and proneural subtypes.

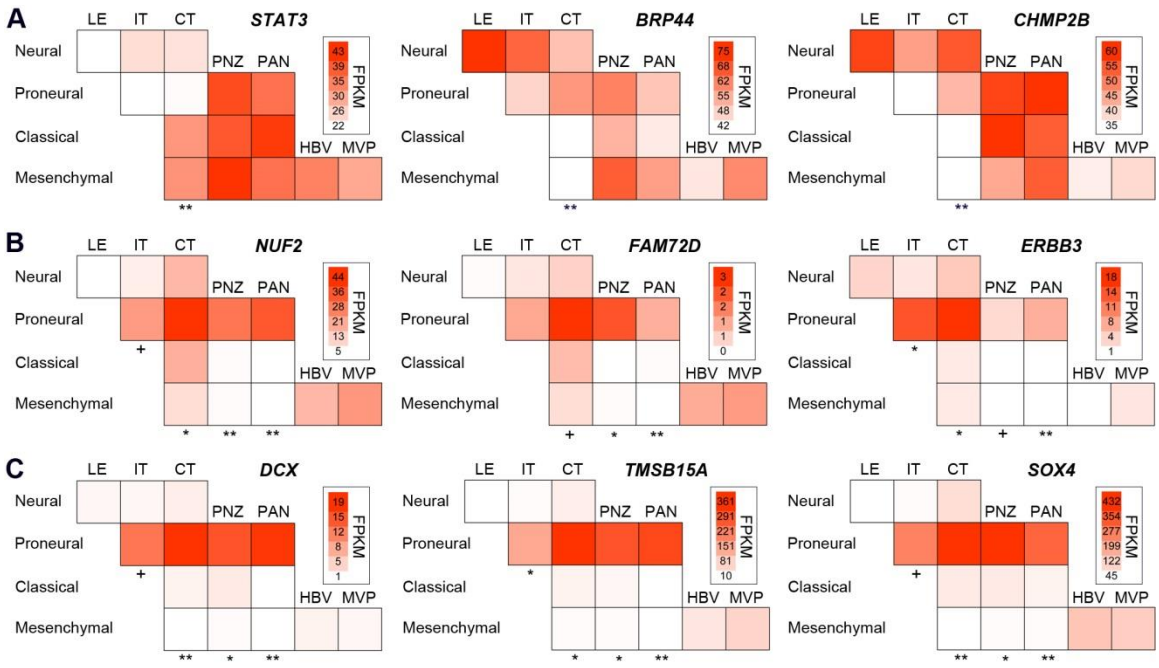


Fig. S7. Subtype selective gene expression across anatomic features. Examples of genes showing significant differential expression by RNA sequencing across subtypes within each anatomic feature. (**A**) Genes with significant difference in expression across subtype in CT but not in PAN or PNZ. (**B**) Genes with significant difference in expression across subtypes in CT, PAN, and PNZ. (**C**) Genes with difference in expression across subtypes in IT, CT, PAN, and PNZ. ** FDR-corrected $p<0.01$, ANOVA; * $p<0.05$; + $p<0.1$. Note that LE, HBV, and MVP are each represented almost exclusively by a single subtype, and therefore differences in subtype could not be assessed. Each box corresponds to at least 5 samples.

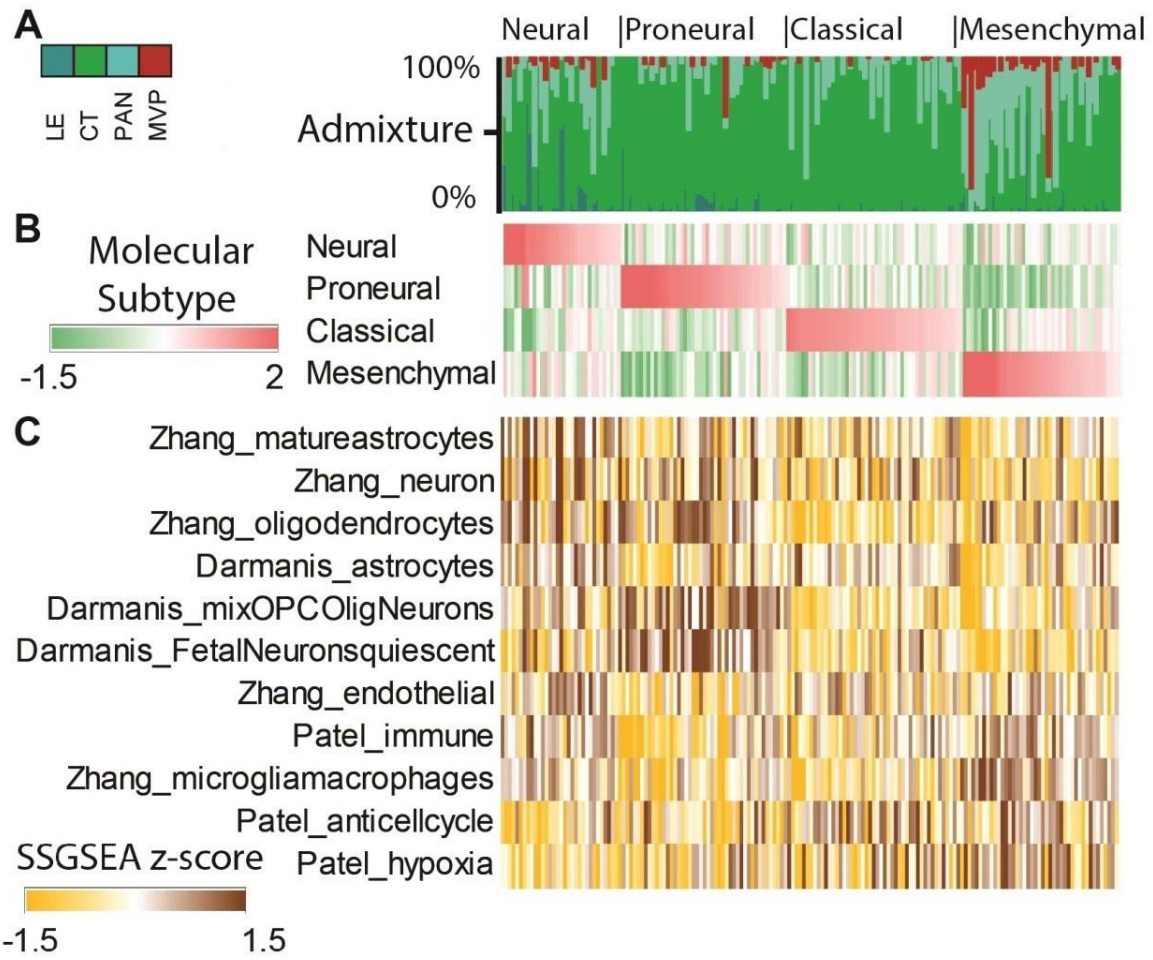


Fig. S8. Classification of TCGA bulk tumor samples by admixture, subtype, and cell types. Classification of 167 TCGA bulk tumor RNA-Seq samples by (**A**) admixture, (**B**) molecular subtype, and (**C**) published cell type and/or phenotype gene sets determined from single cells or cell types (*12*, *62*, *63*). Samples in the figure were sorted first by subtype and then by subtype value with some manual ordering. List of gene set signatures, detailed names, and references is provided in table S15. Fractional composition determined with CIBERSORT (*52*). Gene enrichment determined with ssGSEA. Classifications, and correlations between anatomic feature gene sets and molecular subtypes, calculated in table S16.

We calculated the subtype of the Ivy LMD and TCGA bulk samples using the previously identified 840 gene set (*6*). Enrichment of the gene sets in the TCGA samples confirmed the findings of the Ivy LMD samples to the extent possible with its limited representation of LE, HBV, and MVP components (*5*).

Other Supplementary Materials for this manuscript include the following as separate Excel files (available at www.sciencemag.org/content/_____:

Table S1: Patient data – 1p19q codeletion and *EGFR* amplification (**EGFR* amplified in a small sub-population) determined by FISH. *IDH1* mutation determined by immunohistochemistry or genotyping with PCR(*64*) (wildtype; WT), *EGFRvIII* detected by PCR(*65*) and *MGMT* methylation determined by quantitative methylation specific PCR(*61*) (unmethylated, UM; methylated, M), *EGFR*, *PDGRA*, and *PTEN* copy number data, *PTEN*, *TP53*, *IDH1*, *IDH2*, *ATRX*, *EGFR*, *PDGFRA*, *PIK3CA*, and *PIK3R1* RNA-Seq derived mutations, *FGFR3-TACC3* and *PTPRZ1-MET* gene fusions, molecular subtype as provided in the web resource. Additional information from TCGA for samples from six patients from Ivy cohort that were provided to TCGA is also included.

Table S3: Anatomic feature differentially expressed gene sets – All genes showing at least 2-fold enrichment and a maximum false discovery rate of 0.01 between LE (1998), CT (114), PAN (389), and MVP (1126) and each of the other anatomic features by edgeR analysis. These are the genes along the diagonal of Fig 2a.

Table S4: Validation of RNA-Seq by anatomic feature ISH for enriched genes– Comparison between RNA-Seq and ISH assessments of enrichment in PAN, CT, or MVP. LE and IT could not be compared because LE/IT genes enriched in RNA-Seq samples were not selected for assessment in the Anatomic Feature ISH for Enriched Genes study.

Table S9: **Validation of manual annotations** – Confusion matrix and input data for validation by neuropathology concordance analysis of manual annotations to support LMD guide lines around anatomic features.

Table S10: **Validation of inter-neuropathologists agreement** – Confusion matrix and input data for validation of neuropathologists' identification of anatomic features.

Table S11: **Validation of machine learning annotations** – Confusion matrix and input data for validation by neuropathology concordance analysis of ML annotations of anatomic features in atlas.

Table S12: Copy number changes – Copy number segments showing amplifications and deletions in bulk tissue from 39 tumors at all loci with detectable genomic alterations.

Table S13: Single Nucleotide Variants (SNVs) – Anatomic feature RNA-Seq derived SNVs identified in 10 genes that are often mutated in glioblastoma.

Table S14: Anatomic feature gene sets for admixture analysis – Sets of 293 genes highly specific for individual anatomic features that were generated by CIBERSORT for admixture analysis in figures S6B and 8A, and Table S16.

Table S15: Gene sets for published brain cell types and/or phenotypes – Published gene sets used as input for single sample gene set enrichment analysis of Ivy and TCGA samples.

Table S16: Ivy GAP and TCGA gene set enrichment heat map and correlations – For 122 Ivy LMD/RNA-Seq samples: sample, tumor, histology, admixture, subtype, published cell type signature enrichments and correlations. For 167 TCGA bulk samples: sample, tumor, admixture, subtype, cell type signatures enrichments and correlations.