doi: 10.1038/nature07385

Supplementary Information

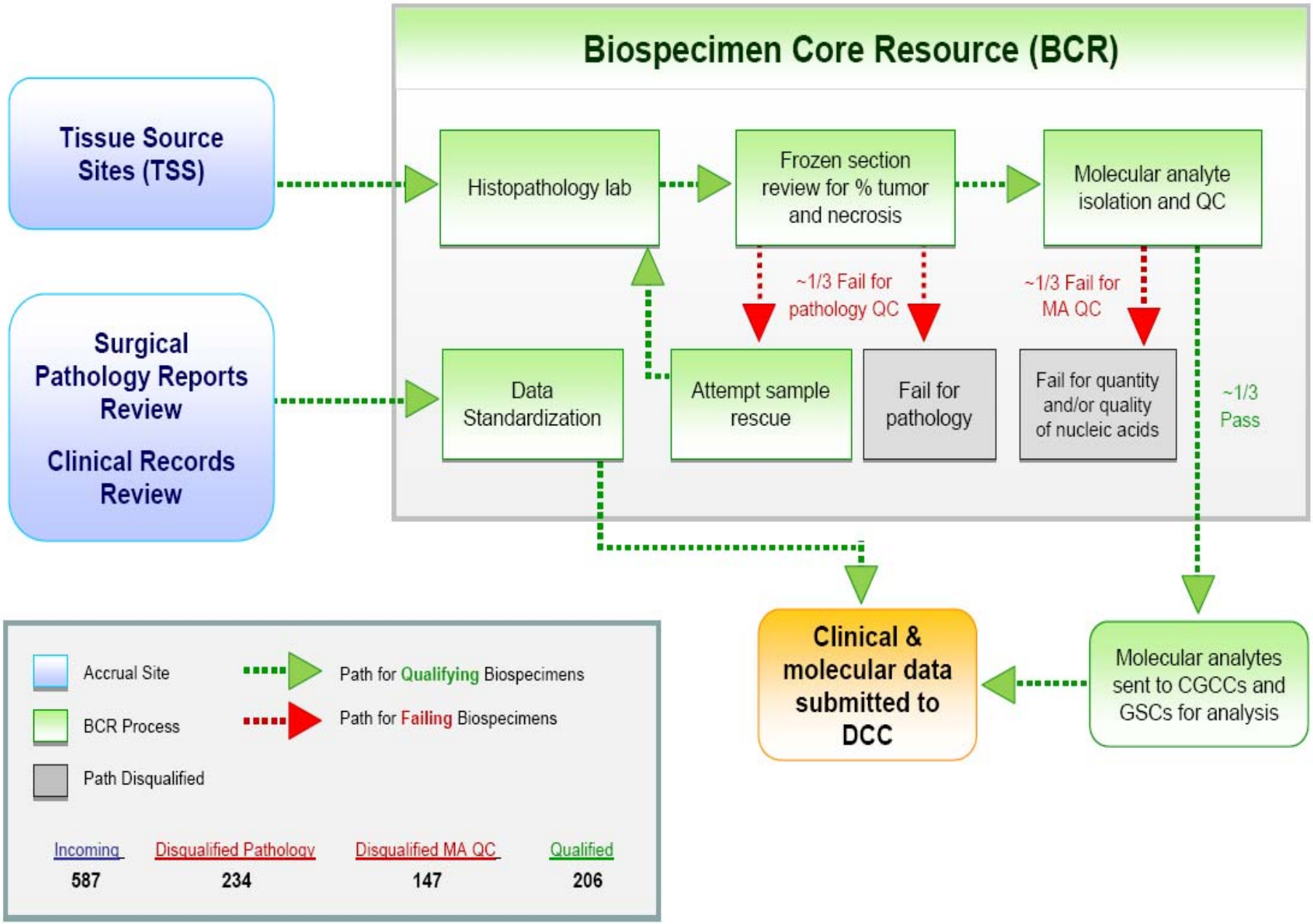
for

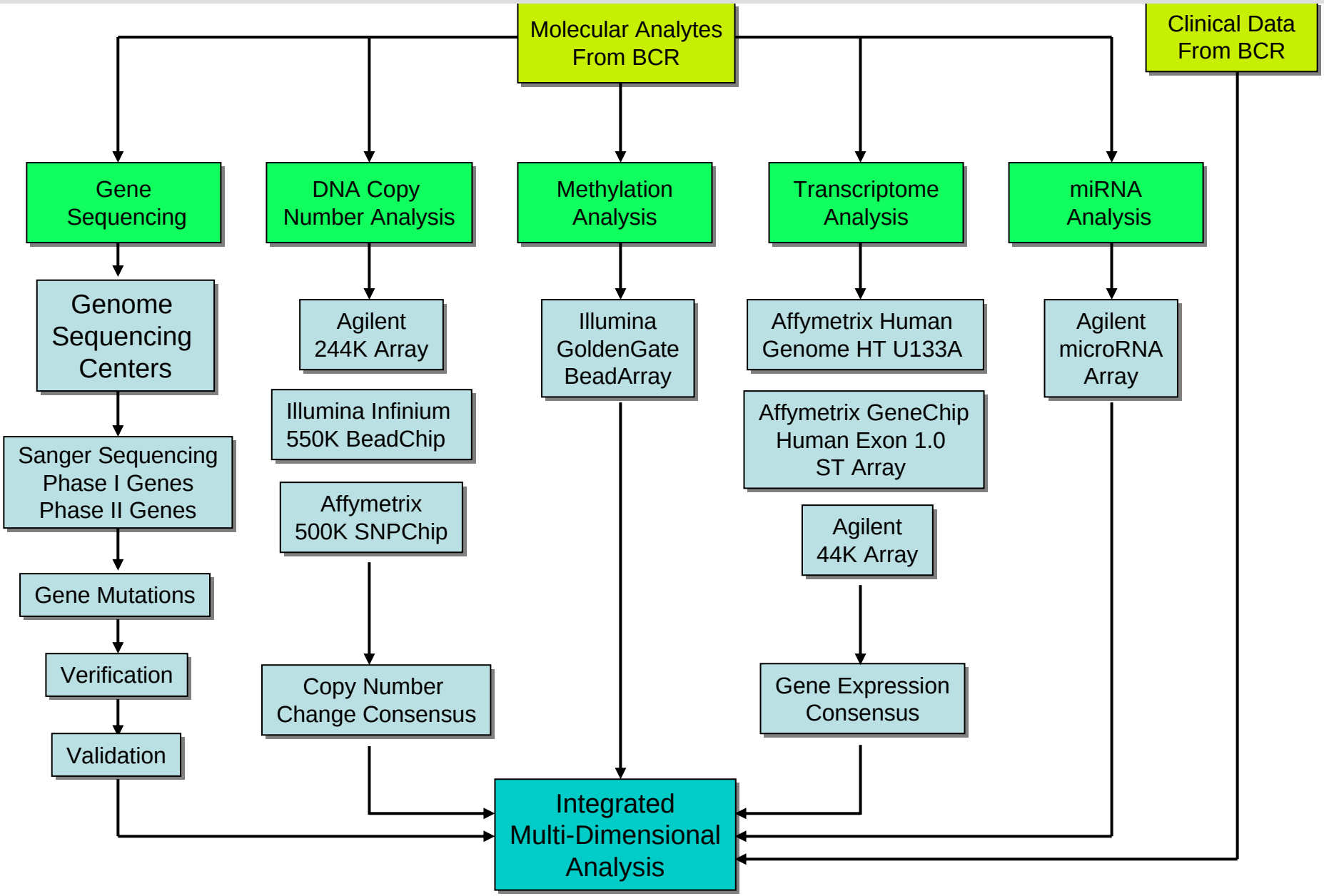
Comprehensive genomic characterization defines human glioblastoma genes and core pathways

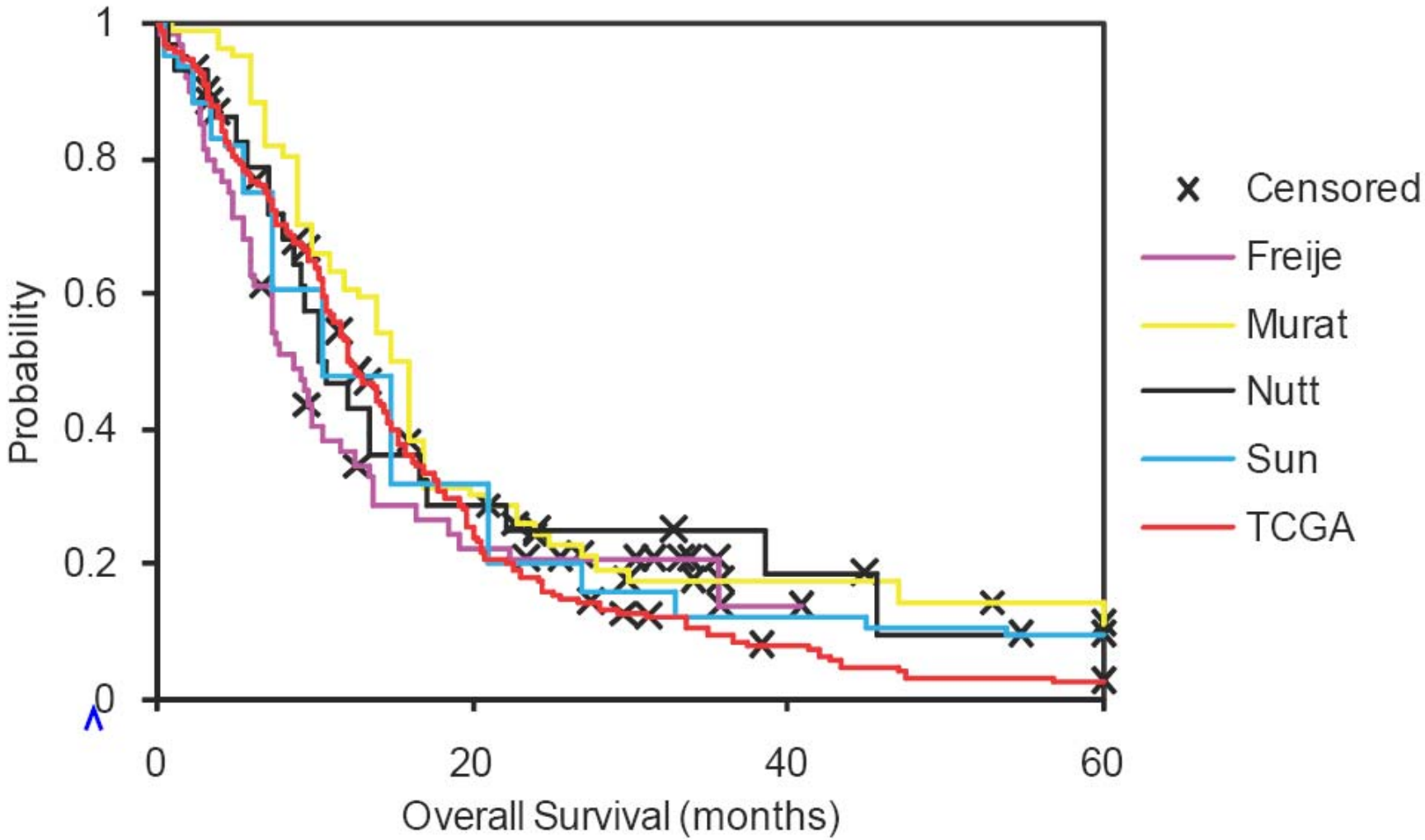
by

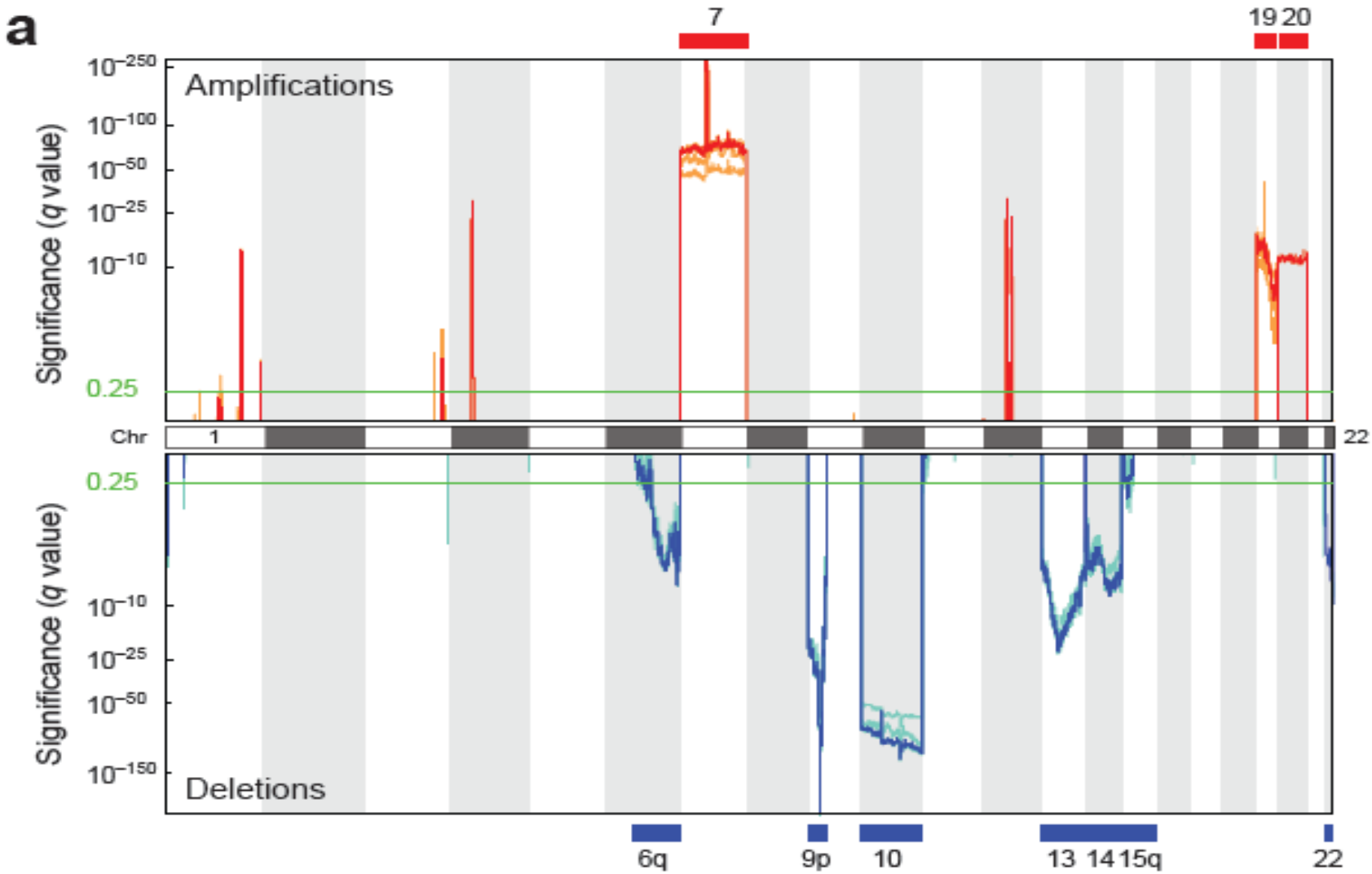
The Cancer Genome Atlas (TCGA) Research Network

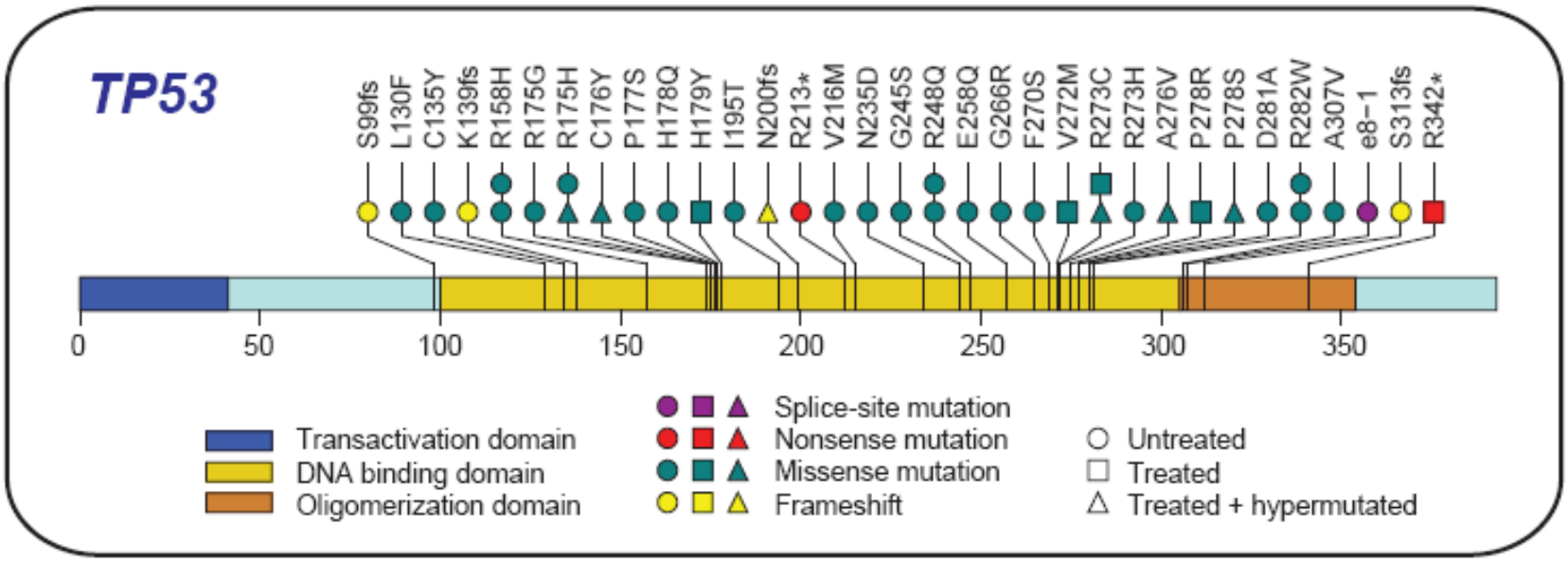
Table of Contents


Biospecimen Core Resource (BCR)
Tissue Source Sites (TSS)
Surgical Pathology Reports Review
Clinical Records Review
Histopathology lab
Frozen section review for % tumor and necrosis
Molecular analyte isolation and QC
~1/3 Fail for pathology QC
~1/3 Fail for MA QC
Data Standardization
Attempt sample rescue
Fail for pathology
Fail for quantity and/or quality of nucleic acids
~1/3 Pass
Clinical & molecular data submitted to DCC
Molecular analytes sent to CGCCs and GSCs for analysis
Accrual Site
BCR Process
Path Disqualified
Path for Qualifying Biospecimens
Path for Failing Biospecimens
Incoming 587
Disqualified Pathology 234
Disqualified MA QC 147
Qualified 206


Molecular Analytes From BCR
Clinical Data From BCR
Gene Sequencing
DNA Copy Number Analysis
Methylation Analysis
Transcriptome Analysis
miRNA Analysis
Genome Sequencing Centers
Agilent 244K Array
Illumina GoldenGate BeadArray
Affymetrix Human Genome HT U133A
Agilent microRNA Array
Illumina Infinium 550K BeadChip
Affymetrix GeneChip Human Exon 1.0 ST Array
Sanger Sequencing Phase I Genes Phase II Genes
Affymetrix 500K SNPChip
Agilent 44K Array
Gene Mutations
Verification
Copy Number Change Consensus
Gene Expression Consensus
Validation
Integrated Multi-Dimensional Analysis


1
0.8
0.6
0.4
0.2
0
Probability
0
20
40
60
Overall Survival (months)
Censored
Freije
Murat
Nutt
Sun
TCGA


a
7
19 20
Amplifications
Significance (q value)
10−250
10−100
10−50
10−25
10−10
0.25
Chr
1
22
0.25
Significance (q value)
10−10
10−25
10−50
10−150
Deletions
6q
9p
10
13 14 15q
22


TP53
S99fs
L130F
C135Y
K139fs
R158H
R175G
R175H
C176Y
P177S
H178Q
H179Y
I195T
N200fs
R213*
V216M
N235D
G245S
R248Q
E258Q
G266R
F270S
V272M
R273C
R273H
A276V
P278R
P278S
D281A
R282W
A307V
e8−1
S313fs
R342*
0
50
100
150
200
250
300
350
Transactivation domain
DNA binding domain
Oligomerization domain
Splice-site mutation
Nonsense mutation
Missense mutation
Frameshift
Untreated
Treated
Treated + hypermutated

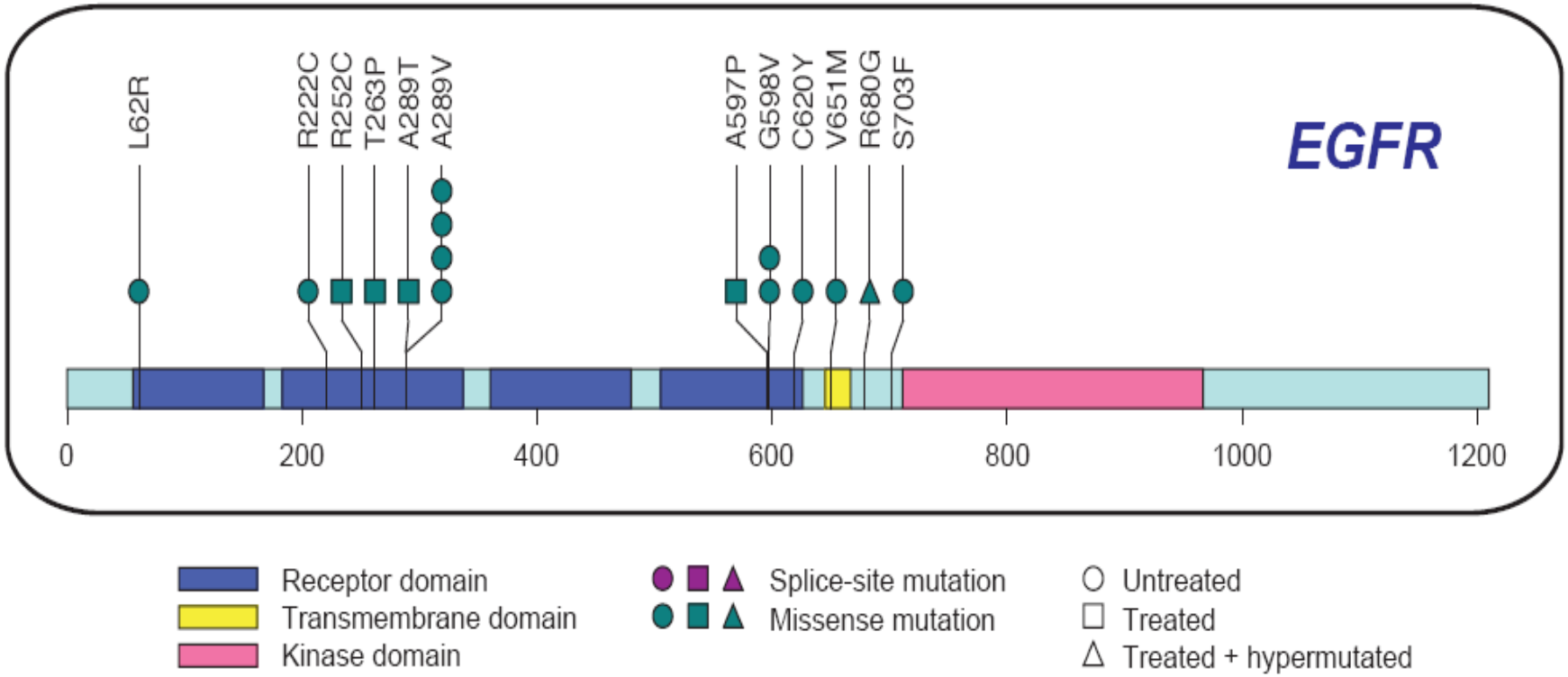

L62R
R222C
R252C
T263P
A289T
A289V
A597P
G598V
C620Y
V651M
R680G
S703F
EGFR
0
200
400
600
800
1000
1200
Receptor domain
Transmembrane domain
Kinase domain
Splice-site mutation
Missense mutation
Untreated
Treated
Treated + hypermutated

Figure S7. Signaling Pathway Alterations (DNA Copy Number, n=206)

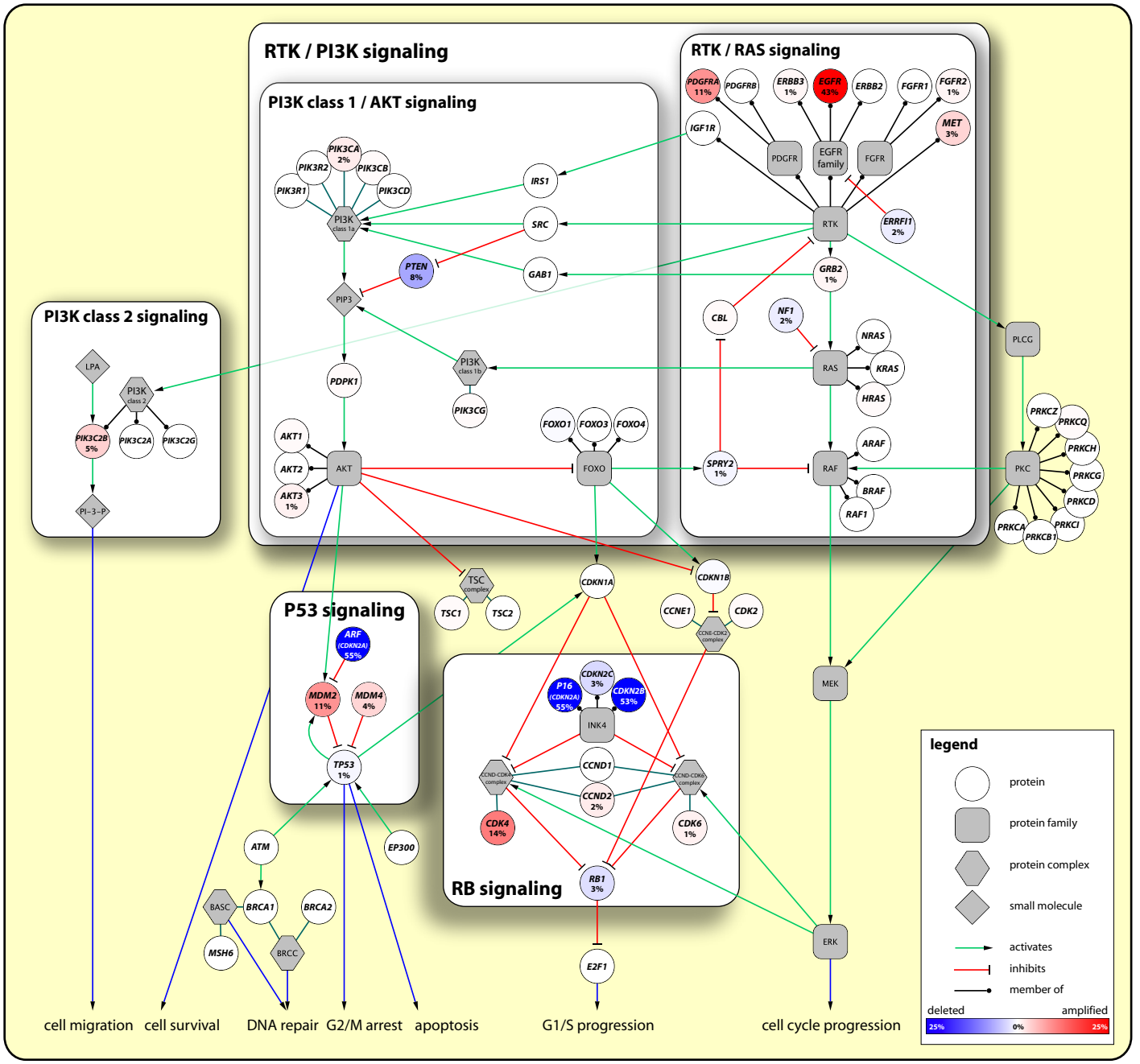
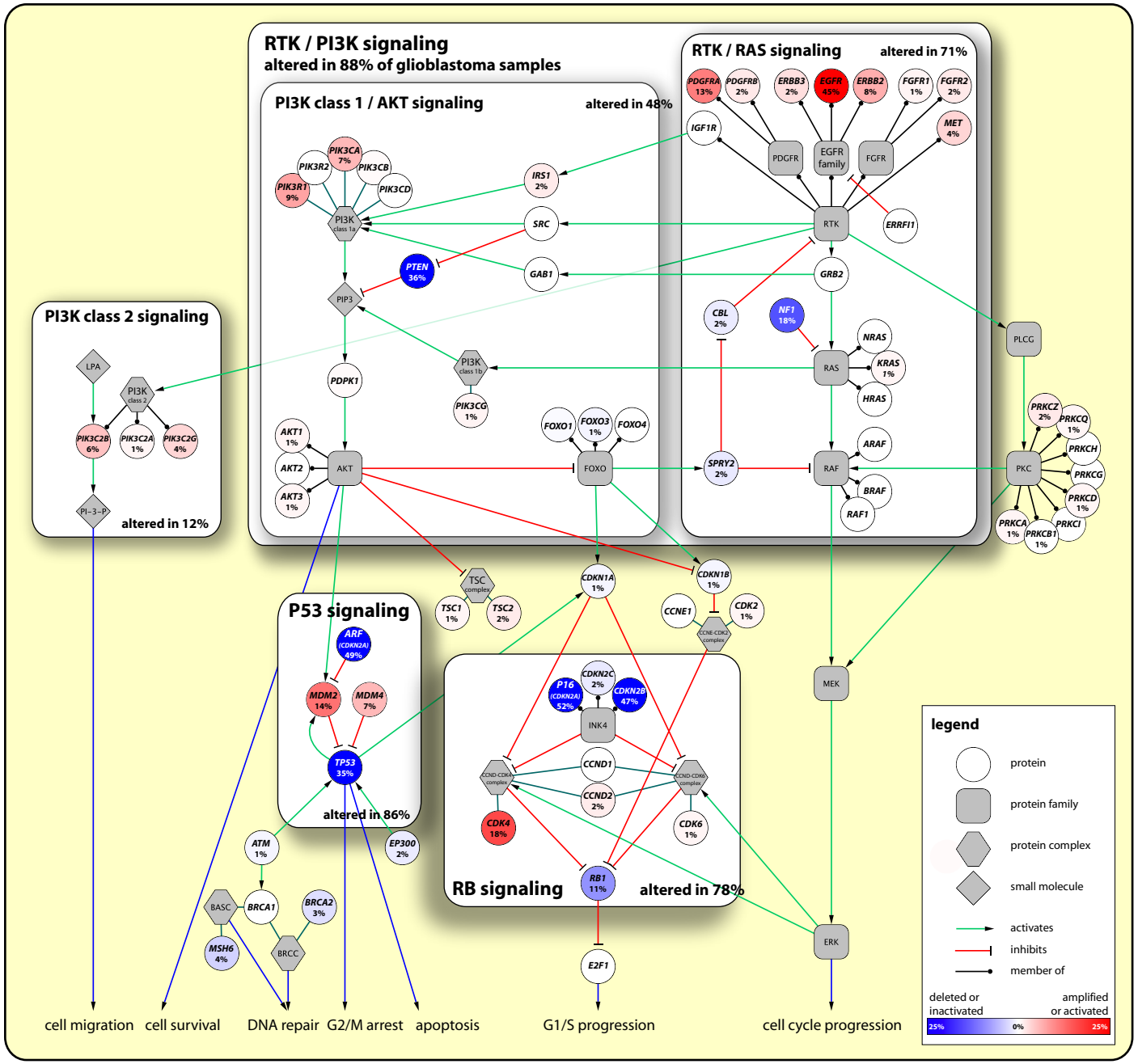


Figure S8. Signaling Pathway Alterations (DNA Copy Number and Mutations, n=91)



A. Supplementary Figure Legends

Figure S1. Biospecimen Processing and Quality Control. This figure summarizes the workflow for acquisition, quality control, and processing of clinically-annotated biospecimens into DNA and RNA for TCGA. Tumor tissues (and case-matched source of germline DNA) are shipped and maintained at -196^{o}C from the tissue source sites (TSSs). The BCR generates and reviews "top" and "bottom" sections from each portion of frozen tumor that is a candidate for analyte generation. These tissue in these sections must be ≥80% tumor nuclei and <50% necrotic, as subjectively determined by a pathologist reviewing a specific number of microscopic fields. Those that fail this test are reentered into a sample rescue pathway that either uses additional portions from the original specimen and/or attempts to physically trim the candidate portion to meet specifications. Approximately 1/3 of the cases failed at this step and were withdrawn. Samples that pass pathology QC enter the molecular analyte (MA) production pipeline for concomitant DNA and RNA isolation. MAs are checked for quality, quantity and the normal and tumor DNA is confirmed to be from the same individual (see Materials and Methods for details); those that pass are distributed to CGCCs and GSCs for characterization. About 1/3 of the total cases failed MA QC. Retrieval of clinical data is initiated once a case has passed MA QC. These data are entered into electronic systems, and formatted and re-coded to meet the NCI caBIG™ standards for Common Data Elements and approved terminologies. Clinical data (BCR box "Data Standardization") and molecular profiling/sequencing data (from the CGCC and GSC), are submitted to the Data Coordinating Center (DCC) from where they are accessible according to the project's data access policies.

Figure S2. Flow and distribution of biomolecules from BCR to TCGA centers for analyses. Flow chart showing the TCGA glioblastoma molecular analyses and their evaluation after provision of specimens and clinical data by the Biospecimen Core Resource (BCR) (see Figure S1). The data obtained from multiple analysis platforms is evaluated by multiple institutions and integrated to form a global representation of the genomic and transcriptomic alterations that occur in glioblastoma.

Figure S3. Kaplan-Meier survival plot for five GBM data sets. Survival data for TCGA samples (n=183) were similar to survival data from four GBM data sets: Freije et al (n=46), Murat et al (n=63), Nutt et al (n=23), and Sun et al (n=69). P=0.2

Figure S4. Significant copy-number gains and losses. Significance of copy number alterations, including low- and high-level events. Orange and cyan lines represent the significance level in data from each of the four genomic copy number platforms. Red and blue lines represent the second most significant value among the data sets – values below the significance threshold (q=0.25) represent consistent significant events. Red and blue bars on either side represent broad gains and losses.

Figure S5. *TP53* missense mutations identified in TCGA glioblastomas. Virtually all mutations (indicated by small boxes) occur in the central DNA binding domain. Frequent mutations were observed at classic p53 "hot spot" codons 175, 248, and 273.

Figure S6. *EGFR* somatic mutations in 91 glioblastoma tumors. Mutations cluster in the extracellular domain in both genes. Splice site mutation position is given in number of bases to the closest exon (e#); positive = 3' of exon.

Figure S7. Copy number alterations in glioblastoma. Frequencies of alterations by copy-number alteration (high-level amplifications and homozygous deletions) of genes altered in glioblastoma. Amplifications are shown in shades of red and homozygous deletions are shown in shades of blue. Based on consensus copy-number data for 206 samples (see Supplementary Methods).

Figure S8. Signaling pathway alterations in glioblastoma - based on mutations and copy number changes in 91 samples. Frequencies of alterations by mutation or copy-number alteration (high-level amplifications and homozygous deletions) of genes altered in glioblastoma. Amplifications and activating mutations are shown in shades of red and homozygous deletions and inactivating mutations are shown in shades of blue. Based on 91 samples with sequencing and copy-number data.

B. Supplementary Table Legends

Table S1. Summary (A) and Individual (B) listing of cases. S1A: Demographic summary the 206 cases generating samples used for this paper. S1B: Table of individual cases with associated demographic and basic clinical data, from which specimens were used to generate molecular profiles. A: **TCGA case ID**: a project unique reference and root identifier for all derived specimens and molecular aliquots. B: **Secondary or Recurrent GBM**: "No" indicates tumor was primary and previously untreated. C: **Gender**. D: **Vital Status**: vital status as of Q1, 2008. E: **Race**: using US Census Bureau standard. F: **Pathology Dx**: anatomic pathology diagnosis (GBM = astrocytoma grade IV). G: **Age at Procedure**: in years. H: **Age at Death**: in years. I: **Sample sequenced**: indicates whether or not this sample was characterized by sequencing, as well as genomic characterization. J: **Hypermutated**: Indicates whether or not this case's tumor DNA was identified as hypermutated by sequencing (see Discussion). K: **Neoadjuvant chemotherapy**: Indicates, for secondary or recurrent samples, whether or not the patient had received chemotherapy (and what agent, if known) prior to surgery yielding sample analyzed in this paper. L: **Neoadjuvant radiation therapy**: Indicates, for secondary or recurrent samples, whether or not the patient had been treated with radiation therapy prior to surgery yielding sample analyzed in this paper.

Table S2. GISTIC significant events and resident genes with copy number-correlated expression.

Table S3. Significant alterations identified by RAE including resident genes with correlated expression.

Table S4. GTS defined regions of informative CNAs.

Table S5. Phase I genes (n=601) selected for re-sequencing.

Table S6. Summary of validated somatic mutations identified in 91 samples.

Table S7. The genomic information for 1,498 DNA methylation reactions used for analyses of GBM tumors in a custom Illumina GoldenGate Methylation assay (OMA-003). The TargetID and ProbeID values are unique identifiers from Illumina during the initial probe design process. Each reaction, since it lies in the promoter/5' region of a gene, is described with accompanying GID, Accession Number, Gene Symbol, Gene ID, chromosomal location, gene synonym, gene annotation and gene product description. The genomic coordinates of the probed CpG dinucleotide, along with its relative distance to the transcription start site and the oligomer DNA sequences are provided. Finally, the CpG island status of the genomic locus containing each CpG dinucleotide is measured using a relaxed version of the Takai and Jones CpG Island criteria – although all genes were determined as being located in CpG islands, the Illumina probe design specifications placed some reactions on the edges of CpG islands or outside of the CpG island. An additional 1,505 reactions covering 807 genes was also tested using the commercially available Illumina DNA Methylation Cancer Panel I (OMA-002). The genomic information for these reactions are available at www.illumina.com. In total, 2,305 genes were assayed for DNA methylation on the Illumina GoldenGate platform.

Table S8. Copy number status of key components in p53, RB and RTK signaling pathways based on consensus copy number call (Supplementary methods) by 4 data sets on 206 samples.

Table S9. Copy number and mutation status of key components in p53, RB and RTK pathways based on data on 91 samples.

Table S10. Fisher's exact Odds ratios and p values on 91 samples. Top: Matrix of odds ratios showing relationships among gene alterations included in the pathway analysis. Odds ratios range between zero and infinity (INF). An odds ratio larger than 1 indicates co-occurrence of gene alterations beyond what would be expected by chance given the total number of samples altered for each of the two genes; an odds ratio less than 1 indicates a tendency toward mutual exclusivity of occurrence; an odds ratio of 1 indicates no association. Bottom: Matrix of p-values for positive and negative associations among gene alterations noted above. Bolded numbers are one-sided Fisher's exact test p-values for the null hypothesis that the true odds ratio is one or less (i.e., that an apparently positive association is just chance); non-bolded numbers are the same but for the null hypothesis that the true odds ratio is one or greater (i.e., that an apparently negative association is just chance). Orange and yellow highlight p-values <0.01 and <0.05, respectively, for positive association (i.e., co-occurrence); dark and light blue highlight p-values <0.01 and <0.05, respectively for negative association (i.e., mutual exclusivity).

Table S11. Fisher's exact Odds ratios and p values on 72 untreated samples. Same as above

C. Supplementary Methods

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas (TCGA) Research Network

SECTION I. BIOSPECIMEN COLLECTION AND PROCESSING

GBM biospecimen pathology quality control

Light microscopic evaluation was performed on a hematoxylin and eosin stained section of each frozen tumor specimen submitted to the Biospecimen Core Resource for assessment of percent tumor nuclei and percent necrosis in addition to other pathology annotations (Supplementary Figure S1). Each patient tumor frozen sample had top and bottom frozen sections evaluated for greater than or equal to 80% tumor nuclei and 50% or less necrosis. If both the top and bottom sections passed these quality metrics then the sample proceeded to biomolecule analyte extraction (Supplementary Methods Figure 1).

GBM biospecimen molecular analyte extraction

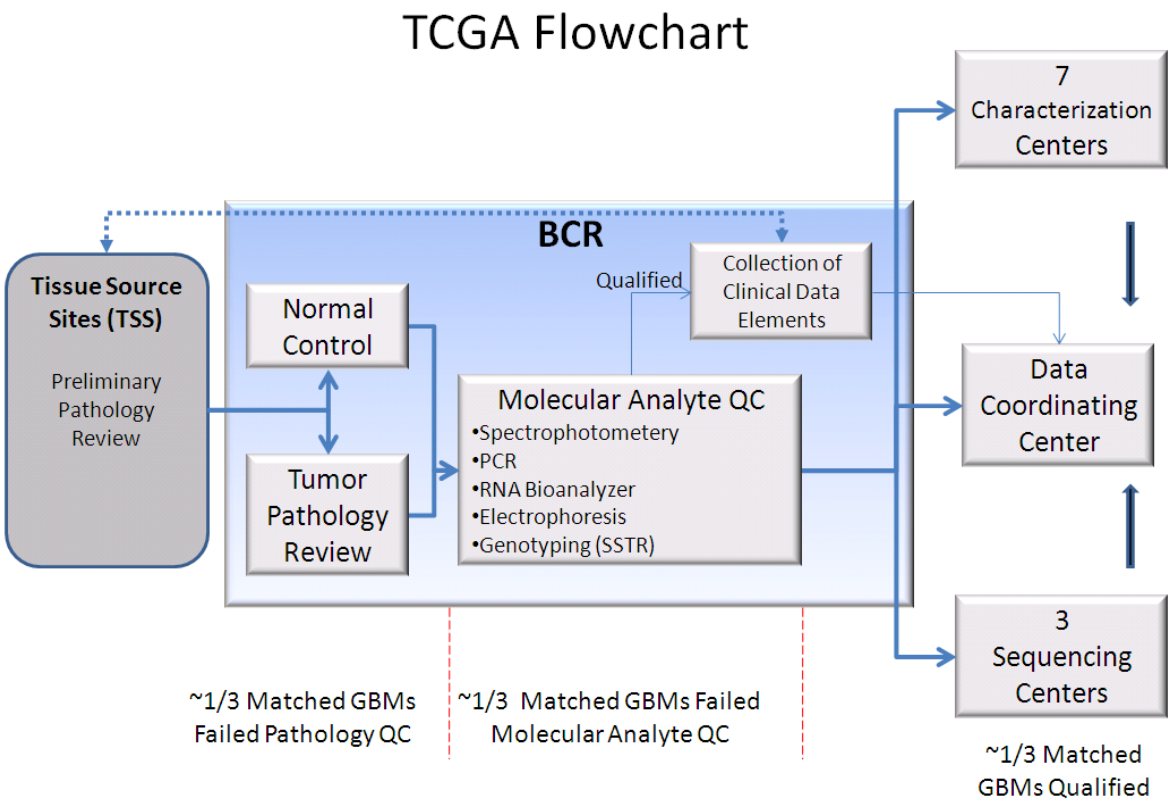
DNA and RNA fractions were isolated from the tissue using an AllPrep DNA/RNA mini kit (Qiagen) per the manufacturer's procedure. Approximately 120 mg of frozen GBM tissue was lysed in a buffer containing guanidine-isothiocyanate to inactivate DNases and RNases and to ensure isolation of intact DNA and RNA with a Covaris adaptive focused acoustics tissue disruptor. DNA was selectively recovered from the lysate by chromatography on a spin column. The column was washed and the bound DNA eluted in 0.1X TE buffer and then precipitated with 1/10 volume of 3M sodium acetate (pH 5.5) and 2.5 volumes of absolute ethanol. TRIzol was added to the flow-through from the DNA capture column, which contains RNA, and the solution is heated at 65°*C* for 5 minutes. After this step, chloroform was added and the phases were separated by high speed centrifugation. A 10% fraction of this total RNA fraction, containing micro RNA, was prepared by precipitation with 1/10 volume 3M sodium acetate (pH 5.5) and 2.5 volumes and absolute ethanol. Ethanol was added to the remaining 90% of the aqueous phase to provide appropriate binding conditions for RNA, and the sample was then applied to an RNeasy spin column, and treated with DNase I to remove residual contaminating DNA, then washed and eluted in 0.1X TE buffer. Residual salts were removed from the eluted RNA by diafiltration with water on a VivaSpin cartridge (VivaScience, VSO122).

Quality Control of Molecular Analytes

Matched normal patient DNA was extracted and purified from the blood or tissue using a QIAamp DNA Blood Midi Kit/QIAamp Mini Kit from QIAGEN. DNA and RNA from these purifications were quantitated by measuring optical density at A260 nm, A280 and A320 nm wavelengths. The purity was assessed by the A260 and A280 absorbance ratio. All DNA samples were further qualified by agarose gel electrophoresis to confirm molecular weight distributions. Suitability for use in sequencing was tested by PCR using primers that produce amplicons of the *GAPDH* gene having sizes of 435bp, 848 bp and 1960 bp. Amplification of at least the 2 smaller amplicons was verified on gel electrophoresis of the PCR products to meet the Quality Control metrics. To estimate the quality of the RNA, we used the RNA 6000 Nano assay on the Agilent Bioanalyzer. This analysis provides 2 estimates of the integrity of the 28S and 18S ribosomal RNA, RIN (RNA Integrity Number) and the 28S/18S ratio. Acceptable values are 28S/18S ratio ≥ 1 or RIN ≥ 7.

Genotyping Analysis

Matched normal (non-neoplastic) patient DNA was provided for each patient case qualified for TCGA. DNA was extracted from whole blood (or its derivatives) using a QIAamp DNA Blood Midi Kit (Qiagen). For a subset of patient cases, disease-free tissue collected from a site distant from the tumor was provided, and genomic DNA was extracted using the QIAamp Mini Kit (Qiagen). To permit DNA-fingerprint-based tracking of all of the derivative samples, and to confirm the matching between the tumor and control DNA samples from each patient, genotyping for highly polymorphic DNA markers was performed with the AmpFlSTR® Identifiler™, a short sequence-specific tandem repeat (SSTR) multiplex PCR assay (Promega, Madison, WI) which co-amplifies 15 SSTRs and the Amelogenin marker, the latter for gender identification.



Supplementary Methods Figure 1. Flowchart of sample flow through the Biospecimen Core Resource.

SECTION II DATA COORDINATING CENTER

Data Flow and Organization

The Biological Collection Resource (BCR) receives tissues and clinical metadata from Tissue Collection Centers (*e.g.* MD Anderson). The BCR provides each TCGA center with biospecimen analytes (DNA and RNA) and their corresponding sample identifiers (IDs). The BCR also transfers the clinical metadata and IDs to the TCGA Data Coordinating Center (DCC). Clinical data is represented using the BiospecimenCoreResource model. Those IDs persist with the results of each analyte.

Each center transfers their platform's data to the DCC in a compressed archive containing the experimental results of a set of assays conducted on a set of samples. Each archive represents many experimental assays from the same platform performed on many samples of the same tumor type. An *experiment* (*i.e.* complete study) for TCGA is defined as the sum of the results of assays for a particular platform from a particular center for all samples of a particular tumor type. That is, an experiment for a particular center is composed of all the assays of a particular platform for all the samples of a particular tumor type. *An experiment may be represented by many archives.*

Genomic Sequencing Centers (GSC) submit trace files to the NCBI Trace Archive. GSCs also transfer trace-to-ID relationship files and mutations to the DCC. Cancer Genomic Characterization Centers (CGCC) transfer experimental results for characterization assays (*e.g.* gene expression, copy number variation, and methylation). The MAGE-TAB specification is used to model and represent array-based data.

Transferred archives are distributed by the DCC to the TCGA public FTP site; data that are considered restricted are removed. Restricted data are distributed to the TCGA secure FTP (SFTP) site. In addition, restricted and unrestricted data are deposited into caBIG compatible repositories. The TCGA Data Portal provides user-friendly access to the FTP and SFTP sites. The DCC maintains relationships between all the data type and tracks metrics and ultimate locations of all data transferred.

As TCGA includes many centers using different platforms, TCGA organizes data by data type and data level. Each platform can potentially produce many kinds of data (data types) depending on the platform. For example, SNP-based platforms are the most complex in that the platform yields three data types: SNP, Copy Number Results, and Loss of Heterozygosity (LOH). The current TCGA platforms and the data types they produce are listed in Supplementary Methods Tables 1 through 3.

The concept of TCGA data level segregates raw data from derived data from higher-level analysis or interpreted results for each data type, platform, and center. Each center and platform may have a slightly different concept of data level depending on their data types, platforms, and the algorithms used for analysis. A normalized list of TCGA data levels for each data type is found below.

An in depth description of TCGA data enterprise including data classification and organization, how to access the data, and a description of how to aggregate TCGA data is presented in TCGA Data Primer (http://tcga-data.nci.nih.gov/docs/TCGA_Data_Primer.pdf).

Data Access

All data is submitted to and processed by the TCGA Data Coordination Center (DCC). The DCC distributes data as bulk downloads and provides access to that data *via* three methods:

1. Bulk downloads
    a. Open Access (ftp://ftp1.nci.nih.gov/tcga/)
    b. Controlled Access (sftp://caftps.nci.nih.gov)
2. TCGA Data Portal: Search-by-File and Archive (http://tcga-data.nci.nih.gov/tcga/findArchives.htm)
3. TCGA Data Portal: Data Access Matrix (http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm)

Bulk download sites have a particular directory structure that classifies distributed data files. The bulk download directory structure is depicted in Supplementary Methods Figure 2.That classification allows perusing for and location of particular datasets. Downloading of multiple archives or data sets is possible if a FTP/SFTP smart client is used. For example, all the characterization archives can be downloaded in one queue by downloading the CGCC directory. This classification facilitates programmatic download of data by using a consistent directory structure and naming process.

The TCGA Portal Search-by-File and Archive provides a user-friendly and searchable view of the FTP and SFTP sites. The Data Access Matrix (DAM), an application within the Portal, facilitates the download of specific data sets by cross-selecting a combination of center, platform, data type, data level, or batches of samples. The result of those selections is a subset of TCGA data files specific to those selections.

The TCGA Pilot Project produces large volumes of genomic information derived from human tumor specimens collected from patient populations, and grants access to significant amounts of clinical information associated with these specimens. The aggregated data generated is unique to each individual and, despite the lack of any direct identifying information within the data, there is a risk of individual re-identification by bioinformatic methods and/or third-party databases. Because patient privacy protection is paramount to NIH and TCGA, human subjects protection and data access policies are implemented to minimize the risk that the privacy of the donors and the confidentiality of their data will be compromised. As part of this effort, data generated from TCGA are available in two tiers.

The Open-Access Data tier is a publicly accessible tier of data that cannot be aggregated to generate a dataset unique to an individual. The open-access data tier does not require user certification for data access. The Controlled-Access Data tier is a controlled-access tier with clinical data and individually unique information. This tier requires user certification for data access.

For more information on these tiers see http://cancergenome.nih.gov/dataportal/data/access/. To learn how to gain access to the Controlled-Access data see http://cancergenome.nih.gov/dataportal/data/access/closed/.

Data Freezes

The DCC will provide a list of the archives that comprise each data freeze. That list represents all the most current new and revised data up to a certain date. Notification of a data freeze is posted to the public TCGA Data listserv (https://list.nih.gov/archives/tcga-data-l.html) and a TCGA Portal news item is also available. The Freeze lists are always published in the public FTP site under the "other" directory (*e.g.*ftp://ftp1.nci.nih.gov/tcga/other/TCGA_Data_Freeze_20080311.txt). The lists are always labeled by the date of the freeze. The contents of the Freeze file are tab-delimited and contain the following columns: archive_name, data_added (to the DCC Bulk Distribution site), and url (a direct URL to download the corresponding archive). Although data continues to be submitted and distributed after a data freeze, the freeze lists should be used as a reference for conducting analysis on common data sets. A freeze list may be referenced in publications using the date of the freeze (*e.g.* TCGA data freeze 03/11/2008).

Mapping of Characterization Platforms to a Common Genome

The vendors of TCGA array-based platforms provide platform-specific array design files that map probes to the genome or genetic elements. Each vendor may have used a different genomic build or have a different method of computing their mapping. It is advantageous for all probes from all TCGA platforms to be mapped using the same method and genomic build to facilitate integration the results of those platforms. The DCC was tasked to do that mapping.

The purpose of the DCC mapping was to:

1. Align all sequences targeted by TCGA platforms to the same build of the genome
2. Use the same methods and assumptions across all platforms for the alignment.
3. Provide a single format for alignment of all platforms.
4. Provide an explicit mapping of the relationships between features on arrays and composite measures such or genes.
5. Provide a "reasonable" estimate of the genomic locations of genes targeted by the composites described in #4

Reporters targeting genomic DNA were aligned to Genome Build 36.1 using BLAT. Reporters targeting RNA were aligned to a transcript database. The transcript database was aligned to Genome Build 36.1 to provide a mapping between transcript coordinates and chromosomal coordinates. The implementation of BLAT and transcript alignment was through the algorithm and software implementation SpliceMiner (http://discover.nci.nih.gov/spliceminer/) [51]. If no match was found, the reporters were aligned to Genome Build 36.1 as for genomic DNA targets. If a reporter aligned to a transcript, then the reporter was not aligned to genome directly. The transcript database being used is the SpliceMiner database composed of RefSeq 36.1 and GenBank 161 complete coding sequences. For genotyping arrays (*e.g.* Affymetrix SNP arrays and Illumina arrays) the target sequences of both alleles were aligned to the genome. The desired scenario is that only one (the common allele) will match build 36.1. In this case, the distinct reporters of both alleles will have the same genome coordinates. A useful alias is the dbSNP ID of these reporters.

Chromosome location are reported using the following syntax: GenomeBuild:Chromosome number:ChromosomeStart-ChromosomeEnd:strand (*e.g.*[36.1:chr3:1234-1890:+]). A comma is used to indicate gaps in a match, such as would be founds for reporters spanning exon-exon junctions (*e.g.*[36.1:chr3:1234-1890,2456-5432,9032-12300:+]).

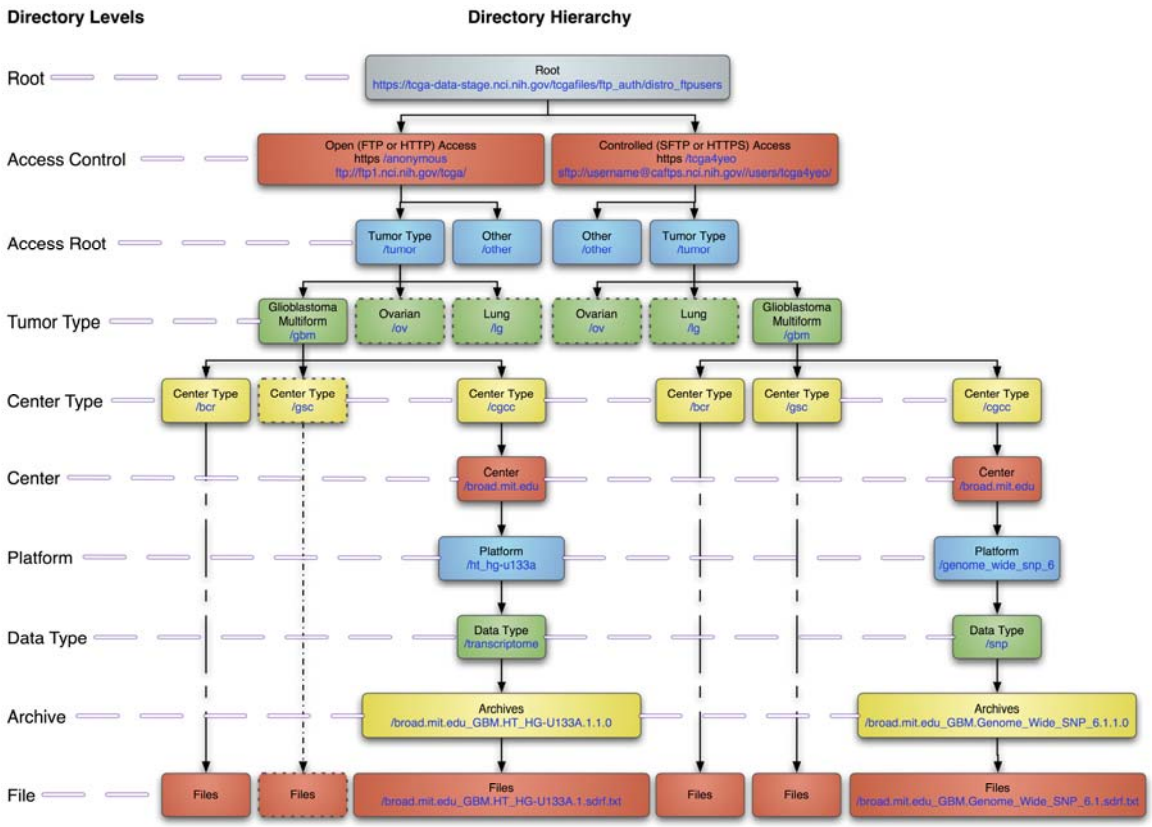
If a reporter does not have a perfect match in either build 36.1 of the genome or the transcript database (in the case of expression arrays) the chromosomal coordinates were reported as "NA" and the composite as "NOMATCH". If multiple perfect matches are found, all will be reported separated by commas.

To validate the SpliceMiner software and algorithm, a separate validation procedure was executed by members of TCGA cross platform genome alignment committee.To help the comparison between SpliceMiner and the validation alignment efforts an auxiliary file with the matched transcripts for each RNA reporter was generated.

TCGA array design files contain the following tab-delimited columns in the order specified:

- X_Block – metacolumn
- Y_Block – metarow
- X – x-coordinate on the array
- Y – y-coordinate on the array
- Feature_ID – Unique ID (one ID per reporter)
- Reporter_ID –ID of the sequence being measured
- Nucleic acid type – reporter is a transcript or genomic DNA
- Reporter_chr_coords – the reporter's chromosomal coordinates for all mappings on the genome
- Composite – the composites that the reporter matched (*e.g.* gene symbol)
- Composite_chr_coords – the union of all chromosomal coordinates of all sequences comprising the target of the composite
- Aliases – any original information that should be retained (*e.g.* Affymetrix probe set ID, dbSNP ID, vendor reported target)

The resulting TCGA ADFs are located at ftp://ftp1.nci.nih.gov/tcga/other/integration/.



Supplementary Methods Figure 2. Download directory structure and URL construction
Each rectangular object represents a directory of a particular type except "Files," which represent data files and the leaves of the hierarchy. Each level in the hierarchy represents a level in the directory structure. Colors are only meant to distinguish a level from its parent and children levels. Objects with dashed outlines represent planned directories. Arrowed lines represent the direction further down the hierarchy. Large dashed-arrowed lines indicate that directories for each level do exist but they are not shown to save space in the diagram. Small dashed-arrowed lines indicate that child directories for each level are planned. Wide-horizontal dashed lines indicate the directory level across objects. Blue text in objects represents the part of the directory path that should be concatenated onto the Root URL in the case of using the HTTP or HTTPS protocol or onto the Access Control URL in the case of using the FTP or SFTP protocols. For example, to download the HT_HG-U133A SDRF file listed at the File level, the following URLs would be appropriate: http://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/tumor/gbm/cgcc/broad.mit.edu/ht_hg-u133a/transcriptome/broad.mit.edu_GBM.HT_HG-U133A.1.2.0/broad.mit.edu_GBM.HT_HG-U133A.1.sdrf.txt or ftp://ftp1.nci.nih.gov/tcga/tumor/gbm/cgcc/broad.mit.edu/ht_hg-u133a/transcriptome/broad.mit.edu_GBM.HT_HG-U133A.1.2.0/broad.mit.edu_GBM.HT_HG-U133A.1.sdrf.txt.

Supplementary Methods Table 1 - TCGA Platforms and Data Types

Type	Center	Platform	Data Type (Base-Specific)
BCR	intgen.org	Biospecimen Metadata - Complete Set	Clinical-Complete Set
BCR	intgen.org	Biospecimen Metadata - Minimal Set	Clinical-Minimal Set
CGCC	broad.mit.edu	Affymetrix Genome-Wide Human SNP Array 6.0	SNP-Copy Number Results
CGCC	broad.mit.edu	Affymetrix Genome-Wide Human SNP Array 6.0	SNP-LOH
CGCC	broad.mit.edu	Affymetrix Genome-Wide Human SNP Array 6.0	SNP
CGCC	broad.mit.edu	Affymetrix HT Human Genome U133 Array Plate Set	Expression-Gene
CGCC	hms.harvard.edu	Agilent Human Genome CGH Microarray 244A	CGH-Copy Number Results
CGCC	jhu-usc.edu	Illumina DNA Methylation OMA002 Cancer Panel I	DNA Methylation
CGCC	jhu-usc.edu	Illumina DNA Methylation OMA003 Cancer Panel I	DNA Methylation
CGCC	lbl.gov	Affymetrix Human Exon 1.0 ST Array	Expression-Exon
CGCC	lbl.gov	Affymetrix Human Exon 1.0 ST Array	Expression-Gene
CGCC	mskcc.org	Agilent Human Genome CGH Microarray 244A	CGH-Copy Number Results
CGCC	stanford.edu	Illumina 550K Infinium HumanHap550 SNP Chip	SNP-Copy Number Results
CGCC	stanford.edu	Illumina 550K Infinium HumanHap550 SNP Chip	SNP-LOH
CGCC	stanford.edu	Illumina 550K Infinium HumanHap550 SNP Chip	SNP
CGCC	unc.edu	Agilent 244K Custom Gene Expression G4502A-07-1	Expression-Gene
CGCC	unc.edu	Agilent 244K Custom Gene Expression G4502A-07-2	Expression-Gene
CGCC	unc.edu	Agilent 8 x 15K Human miRNA-specific microarray	Expression-miRNA
GSC	broad.mit.edu	Applied Biosystems Sequence data	Mutations
GSC	broad.mit.edu	Applied Biosystems Sequence data	Trace-Gene-Sample Relationship
GSC	genome.wustl.edu	Applied Biosystems Sequence data	Mutations
GSC	genome.wustl.edu	Applied Biosystems Sequence data	Trace-Gene-Sample Relationship
GSC	hgsc.bcm.edu	Applied Biosystems Sequence data	Mutations
GSC	hgsc.bcm.edu	Applied Biosystems Sequence data	Trace-Gene-Sample Relationship

Supplementary Methods Table 2 – Description TCGA Data Levels pertaining to Data Types

Data Type (Base-Specific)	Level 1 (Raw)	Level 2 (Normalized/Processed)	Level 3 (Interpreted/Segmented)	Level 4 (Sumary Finding/ROI)
Clinical-Complete Set	Clinical data for 1 patient	NA	NA	NA
Clinical-Minimal Set	Clinical data for 1 patient	NA	NA	NA
CGH-Copy Number Results	Raw signals *per* probe	Normalized signals for copy number alterations of aggregated regions, *per* probe or probe set	Copy number alterations for aggregated/segmented regions, *per* sample	Regions with statistically significant copy number changes across samples
SNP-Copy Number Results	NA	Copy number alterations *per* probe or probe set	Copy number alterations for aggregated/segmented regions, *per* sample	Regions with statistically significant copy number changes across samples
SNP-LOH	NA	LOH calls *per* probe set	Aggregation of regions of LOH *per* sample	Statistically significant LOH across samples
SNP	Raw signals *per* probe	Normalized signals *per* probe or probe set and allele calls	NA	Statistically significant SNPs across samples
DNA Methylation	Raw signals *per* probe	Normalized signals *per* probe or probe set	Methylated sites/genes *per* sample	Statistically significant Methylated sites/genes across samples
Expression-Exon	Raw signals *per* probe	Normalized signals *per* probe or probe set	Expression calls for Exons/Variants *per* sample	Statistically significant exons/variants across samples
Expression-Gene	Raw signals *per* probe	Normalized signals *per* probe or probe set	Expression calls for Genes *per* sample	Statistically significant genes across samples
Expression-miRNA	Raw signals *per* probe	Normalized signals *per* probe or probe set	Expression calls for miRNAs *per* sample	Statistically significant miRNAs across samples
Trace-Gene-Sample Relationship	Trace file; Trace ID-sample relationship	NA	NA	NA
DNA Sequence Mutations	NA	Putative mutations	Validated somatic mutations	Statistically significant mutations across samples

Supplementary Methods Table 3 - Description of TCGA Data Levels

Level Number	Level Type	Description	Example
1	Raw	Low-level data for a single sample, not normalized across samples, and not interpreted for the presence or absence of specific molecular abnormalities.	Sequence trace file; Affymetrix .CEL file
2	Normalized/ Processed	Data for a single sample that has been normalized and interpreted for the presence or absence of specific molecular abnormalities.	Putative mutation call for a single sample; amplification/deletion/LOH signal for a probed locus in a sample; expression signal of a probe or probe set for a sample
3	Segmented/I nterpreted	Data for a single sample that has been further analyzed to aggregate individual probed loci into larger composite or contiguous regions.	Validated mutation call for a single sample; amplification/deletion/LOH signal of a region in the genome for a sample; expression signal of a gene for a sample
4	Summary Finding (ROI)	A quantified association, across classes of samples, among two or more specific molecular abnormalities, sample characteristics, or clinical variables.	A finding that a particular genomic region (a “region of interest”) is found to be amplified in 10% of TCGA glioma samples.

SECTION III. GENE RESEQUENCING

We targeted an initial set of 601 genes, comprised of 7932 coding exons (see Supplementary Table 5), for the initial phase of re-sequencing. These exons, plus 15 bases of adjacent sequence, were divided evenly among three sequencing centers. The individual samples used for the analysis were from batches 1 through 4, supplied as amplified whole genome DNA to the sequencing centers by the TCGA Biospecimen Core Resource. The de-identified individual IDs along with specimen IDs are provided in Supplementary Table ii.

Amplification primers were designed for all exons at the start of the project. The primer design process included a validation step with control DNA to ensure a high level of success and data quality. Likewise, individual DNA samples were also tested with control primers to ensure quality and quantity. Following PCR and exon re-sequencing using fluorescence-based Sanger chemistry on ABI 3730xl automated sequencers, the resulting data were screened for mutations through a series of automated and manual steps. A supplemental round of "confirmation" sequencing served to verify putative mutations, all of which were subsequently validated utilizing an orthogonal genotyping or sequencing method. All sequence traces were deposited in the NCBI Trace Archive [52] under randomized trace names to prevent de-identification of the patients [53]. Annotation provided with each trace is sufficient to link the traces for a single gene in a single individual, or to a 1 Mb segment, whichever is shorter. Annotation to permit linkage of all traces from a single individual is deposited in the TCGA Data Coordination Center (DCC) (http://cancergenome.nih.gov/dataportal/). Putative variants were identified using Polyphred 6.1 [54], Polyscan 3.0 [55], SNPdetector 3 [56,57], and SNPCompare [58]. SNPs and indels were screened against dbSNP for position/allele match.

Putative single nucleotide variants were validated by genotyping on either the Sequenom or Illumina Golden Gate platforms, using TaqMan or Biotage assays, and/or by 454-based re-sequencing. A subset of putative variants was also verified by second-pass sequencing. Verified and validated mutations were used for subsequent mutational analysis.

Putative indels were validated by a round of 454-based sequencing. Since the 3730 and 454 sequencing technologies utilize different chemistries and detection methods, the 454 provides anindependent platform for validating sequence variants. We mapped 454 reads to the human reference sequence (hg36) using BLAT with alignments refined by cross_match [59], and subsequent analysis identified gap positions, permitting validation of the indel positions initially predicted from 3730 data. The validated indels then were cross-referenced with the best BLAT alignments to determine the overall 454 sequence coverage in tumor and matched normal sample pairs. Indels detected by 3730-based sequencing were then aligned to their cognate 454 indels, by matching indels of the same type, of similar size (within 2 bp), and at a similar chromosomal position (within 2 bp). This step was performed separately for normal and tumor samples. When multiple 454-detected indels matched a target, the one with the highest number of supporting reads was retained. We tracked validation status by populating the list of targeted indels with the 454 read coverage, detected indels, and the number of indel-supporting reads. Validation was achieved when sufficient read coverage for both samples and a pre-determined threshold fraction of reads containing the indel were reached.

Background mutation rate estimation and significant gene test

Synonymous mutations identified in 601 genes were further evaluated to assess their somatic status as described above and were used in combination with codon usage of targeted regions to estimate the background mutation rate in GBM. The background mutation rate was then used in statistical calculations to identify significantly mutated genes (see below).

Indel annotation

Boundaries of insertion, deletion and complex rearrangements are annotated as follows (see also Reporting Mutations, "MAF file format", below)

Insertions:

- Start Position is the base before the insertion site
- End Position is the base after the insertion site
- The reference sequence is reported as -
- The inserted sequence is reported on the positive genomic strand
- When multiple alignments are possible the position is reported as the 3' most alignment on the annotated gene's strand.

Deletions:

- Start Position is the first base deleted
- End Position is the last base deleted
- The reference sequence is reported as the sequence from the Start Position to the End Position on the positive genomic strand
- The deleted sequence is reported as –
- When multiple alignments are possible the position is shifted to the 3' most alignment on the annotated gene's strand.

Complex Indels and Insertions and Deletions with multiple complex alignments:

- Start Position is the first base deleted or first base of the repeat
- End Position is the last base deleted or the last base of the repeat
- The reference sequence is reported as sequence deleted from the Start Position to End Position on the positive genomic strand
- The inserted sequence is reported as the sequence that replaces the reference sequence on the positive genomic strand

Reporting mutations (MAF file, data definitions and formats)

Somatic mutations and germline SNPs are deposited at the DCC in MAF files. The following data are reported in a candidate variation file prior to validation:

Somatic mutations:

- Missense and nonsense
- Splice site, defined as within 2 bp of the splice junction

- Silent mutations
- Indels that overlap the coding region or splice site of a gene or the targeted region of a genetic element of interest.

SNPs and indels:

- Germline, missense and nonsense (LOH will not be validated)
- Indels that overlap the coding region or splice site of a gene, or the targeted region of a genetic element of interest.

All candidate somatic missense, nonsense, splice site and indels are processed through "validation" by an independent (orthogonal) genotyping method as described above. Selected silent mutations may be processed through genotyping as well to help determine the background mutation rate. Verified and/orvalidated somatic mutations are reported in a separate MAF file deposited with the DCC. No germline (SNP or indel) candidates are processed through validation. However, if the validation process reveals a given candidate somatic variation event to be germline or loss of heterozygosity, those validated data are reported in the validation file.

MAF file format.
The MAF file has the following columns:

- Hugo_Symbol -- the HUGO symbol for the gene, e.g. EGFR
- Entrez_Gene_Id -- the entrez gene id, e.g. 1956
- GSC Center -- the genome sequencing center reporting the variant: BCM, Broad, or WUGSC
- NCBI_Build -- NCBI build number, currently build 36
- Chromosome -- chromosome number without prefix, e.g. X
- Start_position -- mutation start coordinate (1-based coordinate system)
- End_position -- mutation end coordinate (inclusive, 1-based coordinate system)
- Strand -- one of "+" or "-".
- Variant_Classification -- one of Missense_Mutation, Nonsense_Mutation, Silent, Splice_Site_SNP, Frame_Shift_Ins, Frame_Shift_Del, In_Frame_Del, In_Frame_Ins or Splice_Site_Indel
- Variant_Type -- one of SNP, Ins or Del
- Reference_Allele -- the plus strand reference allele at this position
- Tumor_Seq_Allele1 -- tumor sequencing (discovery) allele 1
- Tumor_Seq_Allele2 -- tumor sequencing (discovery) allele 2
- dbSNP_RS -- dbSNP id, e.g. rs12345
- dbSNP_Val_Status -- dbSNP validation status, e.g. by_frequency.
- Tumor_Sample_Barcode -- tumor sample identifier.
- Matched_Norm_Sample_Barcode -- normal sample identifier.
- Match_Norm_Seq_Allele1 -- matched normal sequencing allele 1
- Match_Norm_Seq_Allele2 -- matched normal sequencing allele 2
- Tumor_Validation_Allele1 -- tumor genotyping (validation) allele 1
- Tumor_Validation_Allele2 -- tumor genotyping (validation) allele 2
- Match_Norm_Validation_Allele1 -- matched normal genotyping (validation) allele 1
- Match_Norm_Validation_Allele2 -- matched normal genotyping (validation) allele 2
- Verification_Status -- one of Valid, Wildtype, Unknown

- Validation_Status -- one of Valid, Wildtype, Unknown.
- Mutation_Status -- one of Somatic, Germline, LOH, or Unknown

Mutation filtering

The sequencing results from each sequencing center were reported in a .maf file listing the details of each candidate mutation. The .maf files included the sample identifier, genomic coordinates ("site"), nucleotide change, and predicted functional consequences (missense, nonsense, synonymous, splice_site, etc.) of each candidate mutation, as well as the mutation status (germline, LOH, or somatic), and the validation or verification status of the mutation. The first step in analysis of the mutation data was to combine the .maf files from all centers into a .mut file containing at most one record for each site-sample pair. In the process of combining the files, care was taken to detect and resolve conflicts between multiple records for the same site-sample.

As part of our sequencing pipeline, non-synonymous mutations were subjected to an orthogonal validation or re-sequencing (verification) step to decrease the prevalence of false positives. In our analysis we considered only those mutations that were confirmed by validation or verification to be actual somatic mutations. Synonymous ("silent") mutations were subjected to manual review to confirm their status as actual somatic silent mutations. Lists of non-silent and silent somatic mutations can be found at http://tcga-data.nci.nih.gov/docs/somatic_mutations/tcga_mutations.htm.

We examined the distribution of mutation rates across samples in order to identify hypermutated tumors. We plotted the number of non-silent mutations and silent mutations for untreated and treated samples (Fig. 2a-2b). We identified 7 hypermutated samples among the treated samples using a standard outlier test. Samples with number of mutations > 75th peercentile + 2*IQR were considered outliers.

Background rate estimation

In order to identify genes that were mutated at a higher rate than would be expected from random background mutation, it was necessary to develop an estimate of the background mutation rate in the set of 72 GBMs carried forward in the sequencing analysis. To estimate the background mutation rate, we analyzed a subset of the filtered mutations that were highly likely to be passenger mutations. We looked at synonymous mutations, which are generally assumed to be functionally neutral [60]. In the set of 72 GBMs, there were 98 silent mutations, distributed over a total coverage of 75,710,450 sequenced bases. Dividing these two numbers yielded a silent mutation rate of $1.29 \pm 0.13 \times 10^{-6}$ mutations per total bases. To convert this background silent mutation rate to the background non-silent mutation rate (per total bases), it was necessary to estimate the expected ratio of silent to non-silent mutations, to correct for the fact that more bases are at risk for non-silent mutations than silent.

We weighted the set of all possible mutations according to the relative mutation rates in three different DNA contexts: (1) C's and G's in CpG dinucleotides; (2) other C's and G's; and (3) A's and T's. These three relative mutation rates were determined directly from the data. The

mutations used to compute the observed context-specific relative mutation rates were the set of all 98 silent mutations, plus a subset of the non-silent mutations: the 52 non-silent mutations that remained after the top 40 most highly mutated genes (based on their total number of mutations) were removed from the data set. These 52 mutations were considered to be predominantly passenger mutations, because they were in the least significantly mutated genes. They were included in the calculation in order to increase the robustness of the background rate estimation. The context-specific relative rates calculated from this set of 150 mutations were as follows: 6.08 ± 0.83 for CpG, 1.20 ± 0.13 for other C+G, and 0.20 ± 0.05 for A+T (all normalized to an overall rate of 1.0). Applying these as weights to the set of all possible mutations yielded an expected silent-to-non-silent ratio of 0.350 ± 0.041, the observed rate of silent mutations by this ratio yielded our estimate of the non-silent BMR (baseline mutation rate) of $3.70 \pm 0.57 \times 10^{-6}$. All data and parameters used for BMR calculation can be found in http://tcga-data.nci.nih.gov/docs/publications/gbm_2008/TCGA_GBM_Level4_Significant_Genes_by_Mutations_DataFreeze2.xls.

Note that using only the 98 silent mutations yield similar values: 4.60 ± 0.78 for CpG, 1.18 ± 0.16 for other C+G, and 0.20 ± 0.07 for A+T, expected silent-to-non-silent ratio of 0.342 ± 0.049 and BMR of $3.78 \pm 0.66 \times 10^{-6}$.

Identification of significantly mutated genes

We explored two methods of tallying mutations. The first method simply grouped all mutations together. For each gene we calculated the observed number of mutations and the total bases sequenced. We calculated a p-value for the gene based on these numbers and the BMR estimated above, using the binomial distribution. These p-values were then corrected for multiple hypotheses (601 genes) by the Benjamini and Hochberg FDR procedure [61]. Genes with an FDR (q) value of 0.1 or less were considered to be significantly mutated. Choosing this FDR cutoff insured that the expected fraction of false positives in our list of significantly mutated genes is not more than 10%.

Our second method took into account the DNA context of each mutation, in order to correct for the different context-specific mutation rates (for example, background mutation in CpG dinucleotides occur at >30-fold higher rate than in As or Ts). We considered four categories of mutations: the three categories listed above, plus a category for indels. The background mutation rate for indels was estimated from the indel mutations occurring in the non-top-40 genes (since there are no indel silent mutations). We combined the binomial distributions for the four mutation categories into a score for each gene, s_g, which was the sum of negative logarithms of the binomial distribution for each category [62]. Finally, to account for the multiple possible ways of achieving the observed score, we examined each possible permutation of mutations across the four categories and summed the probability of every permutation that yielded a score at least as high as the score for the observed permutation. This yielded a p value for the gene. After FDR correction, genes with a q value ≤ 0.1 were considered to be significantly mutated. Results and data used for these analyses can be found in http://tcga-data.nci.nih.gov/docs/publications/gbm_2008/TCGA_GBM_Level4_Significant_Genes_by_Mutations_DataFreeze2.xls.

Robustness of significant gene list

We tested how strongly the list of significant genes depended on the estimated BMR by repeating the analysis using the upper and lower ends of the 95% confidence interval calculated on our estimate for the BMR. These were 4.82 and 2.58 x 10^{-6}, respectively. We also examined the effect of changing which analysis method we used (simple or with-categories). We found that the list was extremely robust to these changes, with the same 8 genes being consistently identified as significant. We also used the BMR and relative rates as calculated based on the silent mutations alone (excluding the 52 non-silent mutations in the non-top 40 genes) and found the same 8 genes to be significant. Since there are no silent indel mutations it was unclear how to estimate the indel relative background mutation rate without using some non-silent mutation. Therefore we used the value from the foregoing analysis, which yielded the same results.

Effect of decreased sample size

We examined the effect of sample size by simulating experiments with a smaller number of samples. We analyzed 1000 random subsets of the 72 samples, in which the subsets contained 12, 24, or 48 randomly chosen samples. We tabulated the fraction of trials in which each gene was found to be significant in these smaller subsets. With a subset of 12 samples, only two of the 8 genes in Figure 2b (*PTEN* and *TP53*) had a greater than 50% chance of being discovered as significant. With 24 samples, only five genes (*PTEN, TP53, ERBB2, PIK3CA*, and *EGFR*) did. Using 48 samples and the context-based method for assessing significance (the second method described above) we were able to discover all of the 8 genes regardless of which of the three background mutations rates (low, mid and high) was used. The simpler method, however, yielded between 6 and 8 genes depending on the background mutation rate. This shows that the robustness of our gene list is due in part to the large sample size. Finally, it is important to note that there may be other genes beyond the identified 8 that have a genuine role in the development of GBM, but which would require even more samples than the 91 analyzed in order to be identified as statistically significant.

SECTION IV. COPY NUMBER ANALYSES

A. Generation of Level 1 and 2 Copy Number Data

A.1) Agilent 244K Arrays

Harvard Medical School / Dana-Farber Cancer Institute

Derivation of log2 ratios
Raw data were generated from scanned images using Agilent Feature Extraction Software (v9.5.11). The median background signal values of each channel are subtracted from the median signal values of the features (probes) of the corresponding channel to obtain the background corrected intensity values for each probe for both channels. Background corrected values from duplicated probes on an array are then merged by taking the median across the duplicating probes. The log2 ratios of background corrected values for the sample channel over the reference channel are then calculated.

Normalization
An in-house R package, aCGHNorm, is used for the normalization of the $\log_2$ ratio data. The normalization procedure involves the application of an invariant set LOWESS normalization algorithm to $\log_2$ ratio data. The algorithm assumes, in this case, that the majority of probe $\log_2$ ratios do not change and are independent of the background corrected intensities of the probes. To build the LOWESS model, the $\log_2$ ratios and the background corrected intensities of the sample and reference channels are used and a big window (21 probes) smoothing process is applied to $\log_2$ ratio after sorting by chromosome position. After mode-centering based on median-smoothed $\log_2$ ratio, unchanged probes (median-smoothed log2 ratio around zero) are then used to build the LOWESS model. The invariant set LOWESS normalization is applied iteratively to the log2 ratio data set until the sum of difference of LOWESS input and output $\log_2$ ratio is zero or stabilized. The artifact of the differences in probe GC content on log2 ratios is corrected by applying LOWESS using probe GC %, regional GC % (GC % of 20 KB of genome sequence containing the probe sequence), and $\log_2$ ratio. Data generated by the normalization process are then merged with in-house annotation data to form a data set containing probe name, chromosomal location, and normalized log2 ratio for each sample. Biological annotations are obtained by BLASTing the probe sequences against the genome.

Quality Control
A number of measurements are used to check for potential quality problem at various stages. i) Probes that are flagged out as non-uniform or saturated by Agilent feature extraction software are excluded; ii) Probes whose median signal values are lower than that of the background are considered faint and are also excluded; iii) The percentages of probes that are flagged during feature extraction or faint are calculated and arrays with over 5% of probes flagged out or being faint are considered as low quality; iv) The square root of the mean sum squares of variance in log2 ratios between consecutive probes arranged along chromosomes are calculated and used as another measurement of array quality. An array with a value over 0.30 is considered as low quality.

Memorial Sloan-Kettering Cancer Center

The raw data were obtained by the Feature Extraction program (v9.5.3.1) provided by Agilent. Although this program does normalize the two channels, this simple normalization misses an effect related to the local GC content within the probe region on the genome. These genomic-based artifacts can cause problems in the downstream analysis so a more complex normalization procedure was implemented to remove this effect. In addition, the usual intensity-dependent bias also needs to be normalized out. Therefore, a two step approach was used: first, a multi-dimensional LOWESS fit of log2 ratio to 3 vectors representing genomic GC percentage averaged over windows of 200bp, 2kbp, and 50kbp at each probe location; secondly, an intensity-dependent LOWESS fit on a set of invariant points. After this normalization, the log ratio (converted to the usual log base 2 convention) is generated for each sample, replicate probes are merged and the output is the normalized log ratio along with probe id and genomic coordinates (Level 2 data).

The following QC measures are computed on each sample: derivative noise, median background intensity in each channel and also the number of segments after segmentation (described below). In addition, the probe intensities and distribution of flagged probes are plotted to look for spatial artifacts or other potential defects in hybridization. Any samples that fail any of these tests are removed from further analysis. For the quantitative measures, a sample is flagged as failed if any measure deviates by more than three standard deviations from the mean of samples run through the MSKCC microarray facility.

A.2) Affymetrix SNP Array 6.0

SNP 6.0 data are processed from raw CEL files to segmented copy–number data using a GenePattern pipeline, which runs the following modules:

SNPFileCreator: this converts raw Affymetrix .CEL files to a single value for each probeset representing a SNP allele or a copy number probe. The module first performs brightness correction by scaling the probe-level values for each CEL file so that the sample-specific median value is 1000. Next, MBEI [63] is used to map probe-level values in each sample to a reference sample (chosen as the normal sample which has a total intensity closest to the median total intensity in the plate). Next, multiple probes are summarized using median polish across the samples in the plate (96 samples).

CopyNumberInference: this module converts summarized intensities, which are expressed in an arbitrary scale, to copy number values by estimating a probeset-specific linear calibration curve (background and scale). SNP probesets and copy-number probesets (CN) are handled separately. For CN probesets, the conversion is performed by using prior measurements of intensity in 5 cell lines with varying numbers (1 to 5) of X chromosomes and extrapolating to the entire genome [64]. For SNP probesets, the background and scale are estimated using the allele-specific cluster centers (i.e. mean intensities of the A and B probesets for the three possible genotypes; AA, AB, and BB) produced by the Birdseed algorithm [65]. Birdseed is applied only to normal samples within the analyzed plate that pass a quality control (FQC call rate ≥ 86%, Birdseed call rate > 90%).

RemoveCopyNumberOutliers: the Remove_CN_Outliers GP module is used to filter inconsistent estimated copy-number values. A value is considered to be an outlier if it satisfies several outlier criteria relative to neighboring values (treating CN probes and SNP probes identically) on the same sample, considered separately in the 5' and 3' directions:

Measure the median copy-number of the 5 nearest neighbors in the given direction. If the difference between the median and the value under consideration is greater than 0.3, and the difference between their log2 values is greater than $\log_2(6)$, the value is called an outlier. If the value is an outlier with respect to both its left neighbors and its right neighbors, it is replaced with the median of the three values centered on itself.

DivideByNormals: systematic bias in copy-number estimation is removed using 5-Nearest-Neighbor normalization [66]. For each tumor, the 5 most similar normal samples are identified among the entire TCGA normal samples (using Euclidean distance between $\log_2$-ratios measured on the entire genome except regions of known CNVs and the X and Y chromosomes). Next, the average log2-ratio of these normal samples are subtracted, at each position, from the tumor's $\log_2$-ratios.

Quality Control: we remove tumor samples that fail Birdseed quality control. In addition, samples are rejected if either their copy-number noise level (proportional to the median of pairwise absolute differences of $\log_2$-ratios of adjacent probes) or their number of segments as found by the segmentation is an outlier. We call a value an outlier if it falls k*IQRs (inter quartile range) above the third quartile. We use k=1 for the noise level and k=2 for number of segments. If several samples pass quality control for a single patient, the one with lowest noise is selected. In all, 169 tumors passed quality control for the SNP 6.0 platform.

A.3) Illumina 550K SNP Arrays

Raw data are given in the IDAT files, which are the binary data files produced by the Illumina scanner, one for each color channel of each sample, that contain the average intensity data for each SNP averaged over >20 beads. These files can be read by the Illumina BeadStudio analysis software to produce all the other data files. All the genotype calls for each sample, as well as a cluster file that defines the genotype cluster positions for each SNP, are provided in the DCC data portal. The raw intensity values and genotyping quality scores are exported from the Illumina BeadStudio software.

The Illumina Beadstudio software is used to generate normalized intensity values for each allele of every SNP, the logR (log of total intensity, summed over both alleles) and B allele frequency values for each SNP, as well as the differences in logR and B allele frequency between each pair of tumor and normal samples. Such pairwise differences are the basis of inferring copy number changes.

Additional normalized logR and B allele frequency data files were obtained from our custom normalization procedures. We developed these procedures to correct for additional sources of noise or bias, such as sample-specific, bead pool-specific, and SNP-specific effects in the intensity data that have not been adequately removed by Illumina's genotyping software.

Summary tables of CNVs discovered for each sample based on our segmentation software are provided at the DCC data portal.

B) Generation of Level 3 and 4 Copy Number Data: Segmentation and Identification of Regions of Interest

B.1) Segmentation

For each data set, segmentation of normalized $\log_2$-ratios was performed using the Circular Binary Segmentation (CBS) algorithm version 1.12.0 for Affymetrix or version 1.13.3 for Agilent data [67,68] with 10,000 permutations, an alpha value of 0.01, and undo splits (undo.sd=1). Post segmentation, we applied an additional level of normalization which centers the segment values around 0. This step was performed slightly differently between the centers; either setting the mode of the histogram of segment means, weighted by the number of probes per segment, to 0 or by subtracting the median segment value from all segments. Segmented data are available as Level 3 data from the DCC data portal site.

B.2) Genomic Identification of Significant Targets in Cancer (GISTIC)

We applied three variations of the GISTIC algorithm [66] on the selected segmented data for each center using the GISTIC GenePattern module (the latter two variations were developed as part of this study). The quality control steps for the Affymetrix SNP 6.0 arrays described above were applied to data from all platforms and yielded 4 data sets including 197 Agilent (HMS) samples, 195 Agilent (MSKCC) samples, 169 SNP6.0 samples and 194 Illumina 550K samples. These sets include at least one high quality sample representing each of the 206 characterization samples.

The three GISTIC variations are:

(i) *Standard* GISTIC: which uses a low-level cutoff (determined by estimating the noise in each platform) to find significant variation of all types; both broad low-level alterations and focal high level alterations. We used the output of this method to identify the broad regions which are significantly altered.

(ii) *Focal* GISTIC: uses sample-specific high-level thresholds (one for amplifications and one for deletions) in order to focus on focal gains or losses which are beyond the levels observed in whole chromosome arms in a given sample. We use this analysis to identify significant focal high-level events.

(iii) *Focal* GISTIC *with arm-peel-off*: This variant addresses the issue that some samples exhibit "choppy" gains or losses, which may cause GISTIC to identify individual significant regions for different parts of (what seems to be) the same chromosomal alteration. To avoid these "spurious" peaks we changed GISTIC's peel-off step to remove in each sample all segments in a chromosome arm which has an altered segment that covers the identified peak.

Standard GISTIC requires threshold parameters indicating the minimal copy-number variation sufficient to contribute to significance calculations. These parameters—one for amplification

and one for deletion--are determined by analyzing a histogram of segment copy-numbers and finding the first valleys to the left and right of the central peak at 0, representing the noise level. These threshold parameters are found independently for each center. In Focal GISTIC with and without arm-peel-off, thresholds are determined independently for each sample by identifying the maximum and minimum (for amplification and deletion, respectively) of medians observed for each chromosome arm, plus a small buffer (set to be the threshold values used for the Standard GISTIC). These aggressive thresholds result in all variation greater than half a chromosome arm being ignored, leaving only the focal high-level events.

All GISTIC runs were performed with cap-values (in $\log_2$-space) of -1.5 and 1.5 (0.7 copies and 5.65 copies) on each sample, i.e. any value above 1.5 was replaced by 1.5 and values below -1.5 were replaced with -1.5. These cap values were used to limit problems of hyper-segmentation that occur particularly in regions with extreme values due different attenuation curves of adjacent probes.

GISTIC reports regions of interest with an associated q-value, which were obtained by multiple hypotheses correction (Benjamini-Hochberg False Discovery Rate procedure) and represent an upper bound on the expected fraction of false positives in the resulting list. Regions with q-values below 0.25 are considered significant and are reported. GISTIC also outputs the genes and miRNAs contained within these regions.

To combine GISTIC results from several centers, we merged regions that overlap to any extent. We designate merged-regions that were formed by regions from at least two centers as 'validated' (see Table 2). The other regions include regions of interest that were detected only by a single center, possibly due to higher coverage in a particular genomic region or due to center-specific artifacts. As an example, using the SNP6.0 we were able to detect focal deletions in the *CSMD1* gene which were not observed by the other platforms most likely due to the higher-resolution of the SNP6.0 platform in that region.

Copy-Number Variants:

Before applying the GISTIC analysis we remove genomic regions which are associated with copy-number variations (CNVs). This step is necessary to avoid significant GISTIC peaks which are due to copy-number variations that appear in large enough fractions of samples. We compiled a list of genomic regions of CNVs by combining several sources. The combined list was used in all of the GISTIC runs. The sources for CNV regions are:

1) CNVs found in a SNP6.0 analysis of all HapMap normals [69].
2) CNVs identified in at least two independent publications listed in the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation, version 3): [70-79]
3) CNVs found in TCGA matched normals by an automated search (see B.3, below).
4) CNVs found in TCGA matched normals by manual investigation of preliminary GISTIC regions.

Gene-specific calls for genes in regions of interest

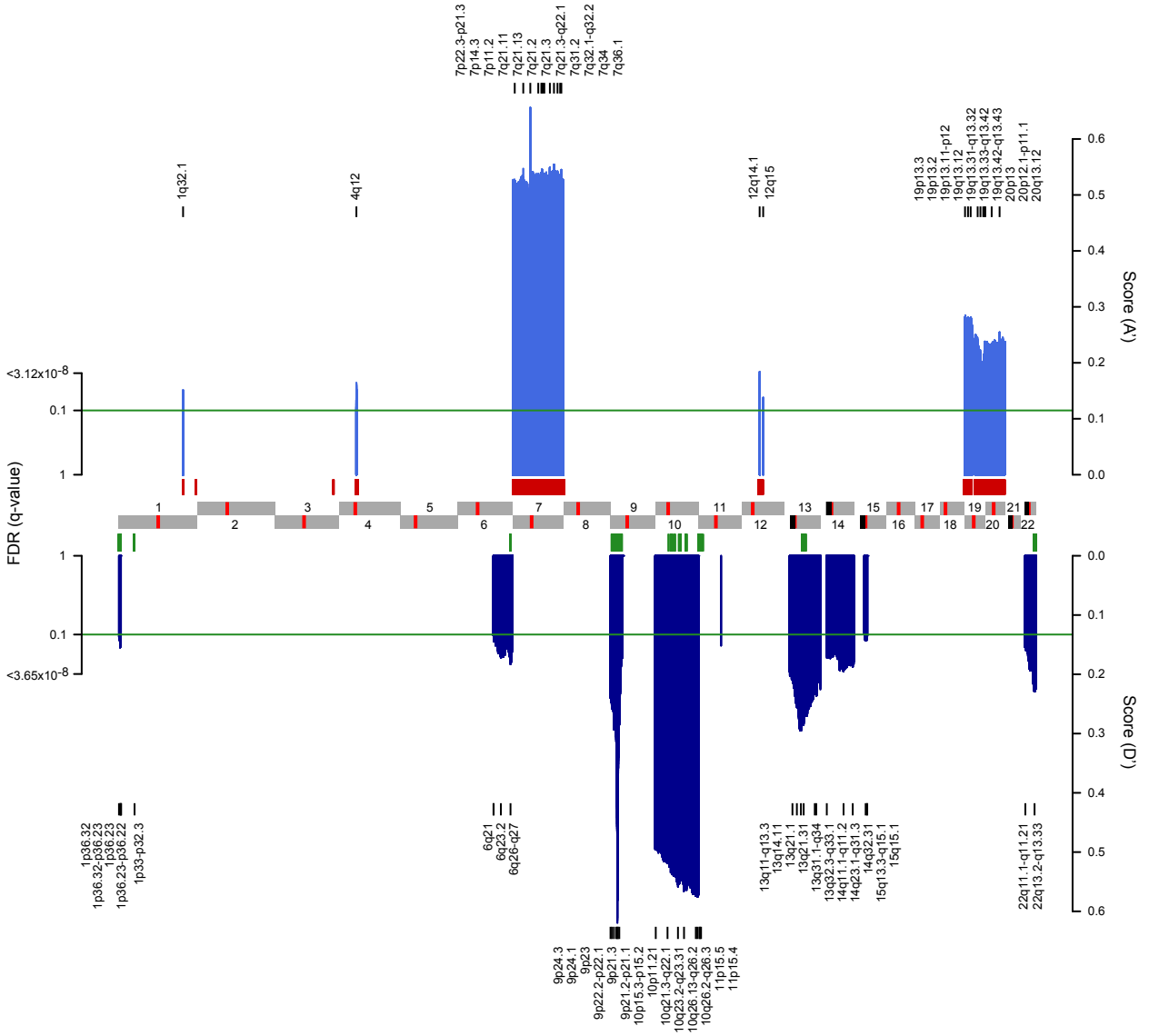
For each center, each patient, and each gene in the list of genes found within validated significant regions, a "call" was produced indicating the extent of copy-number alteration observed for that gene. Alteration magnitudes were gauged for each sample according to the extreme of alteration over the gene's region, compared to the Standard and Focal (sample-specific) copy-number thresholds as follows:

-∞ to Hemizygous Threshold:	Called Homozygous Deletion
Hemizygous to Standard Deletion Threshold:	Called Hemizygous Deletion
Standard Deletion to Standard Amplification:	Called Neutral
Standard Amplification to High-level Amplification:	Called Low-level Amplification
High-level Amplification to ∞:	Called High-level Amplification

From the table of center-specific gene calls, a combined call is determined for each patient and each gene. Using the same approach as for calling a region 'reliable', we combined the center-specific calls by taking the most extreme call which is supported by at least half the centers with available data (i.e. two centers if all centers had data). For example, if two centers observe No Alteration and two observe Low-level Amplification, the combined call is Low-level Amplification (ambiguous cases were not observed for this data set). There was a high level of agreement between centers (>85%) attesting to the high quality of the data and calling method.

B.3) RAE

In total, 216 glioblastoma tumors and 84 normal samples processed on the Agilent 244k platform at MSKCC were analyzed with RAE [80](RAE is available at http://cbio.mskcc.org/downloads/rae). Briefly, the RAE algorithm adapts to the noise characteristics of individual tumors producing sample-specific sigmoid-shaped discriminators of single-copy gain (A_0), amplification (A_1), hemizygous loss (D_0), and homozygous deletion (D_1). These are combined across samples in a common set of genomic regions derived from segmentation breakpoints of all tumors. A background model of random aberrations is produced through permutation of segmental DNA incorporating features of human recombination as a proxy for benign genetic turnover. Finally, RAE assesses the statistical significance of genomic gains and losses by comparing observed lesions across tumors to these random aberrations.



Supplementary Methods Figure 3: Genome-wide copy-number aberrations in GBM identified with RAE. The false discovery rate (q-value, left axis) and score (right axis) for statistically significant copy-number aberrations (light and dark blue respectively) in genomic coordinates (chromosomes indicated at center, centromere in red, acrocentric arms in black). The threshold for significance (FDR ≤ 10%) is indicated (horizontal green lines) as are identified regions of interest. Both independently significant amplification (A_1, red) and homozygous deletion (D_1, green) are also indicated (see Supplementary Methods text).

Input to RAE was level-3 individual-sample segmentation produced by the MSKCC CGCC pipeline (see section B.1). Each sample was normalized by first deriving the distribution of total autosomal segmentation at a width (bandwidth) based on each tumor's derivative noise, then identifying the diploid peak inside the bandwidth of this density distribution and centering the position of this peak to $\log_2$ of zero. These normalized profiles were then analyzed as previously described[80]. The HapMap reference normalization step, used for single-channel Affymetrix data, was excluded as the Agilent platform co-hybridizes a reference normal. A false discovery rate of <10% was the threshold for significance and from which regions of interest (ROI) were

identified with a two-stage algorithm incorporating the intrinsic error of segments spanning each loci of the genome (Supplementary Table 3). The assessment of total genomic loss was supplemented with regions of homozygous deletion significant at the same threshold in a model that excluded monoallelic losses. This is intended to detect uncommon biallelic events having little evidence of hemizygous loss (Supplementary Table 3).

Finally, a repository was created for the purpose of excluding known and presumed germline copy-number polymorphism from the GBM analysis. In total, 84 normal TCGA samples from the Agilent 244k platform were selected to profile copy-number polymorphism. Samples were segmented with CBS (see B.1) and transformed as previously described[80]. Any locus (segment) exceeding A_0 (single-copy gain) or D_0 (hemizygous loss) of 0.99 was considered altered and presumed polymorphic. These events were combined with variants identified by trusted studies obtained from the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation, version 3) [73] (see B.2), and from the analysis of the HapMap collection [69]. Consequently, regions identified by RAE from the tumor analysis, but appearing in two or more of these sources and having sequence coverage >50% were excluded as presumed polymorphism.

For each isoform of all autosomal genes in RefSeq (hg18), RAE additionally assigns a discrete copy-number alteration status in each tumor. These are determined from the values of the individual alteration detectors: A_0, A_1, D_0, and D_1. One of five classes are assigned to each gene/tumor: homozygous deletion (-2), hemizygous loss (-1), copy-neutral (0), single-copy gain (1), and multi-copy amplification (2). Formally, the region(s) of the unified breakpoint profile (UBP) derived by RAE that span the coding locus of a given isoform are identified. For genes spanned by a single region, single-copy gain is assigned to tumors with values of $A_0>0.9$ and $A_1<0.5$. Amplification requires the same of A_0 and $A_1\geq0.5$. A gene is hemizygous for values of $D_0>0.9$ and $D_1<0.9$, while homozygous deletion requires that both D_0 and D_1 exceed 0.9. In the event of discontinuous coverage of the coding locus by regions that harbor intragenic breakpoints in copy-number segmentation, the region of extreme value in either A_0 or D_0 determines the assignment per the aforementioned thresholds.

B.4) Genome Topography Scan (GTS)

The GTS algorithm was run on both SNP and aCGH datasets according to methodology previously described [81]. Briefly, GTS analyzes a set of copy number profiles to generate scores for each genomic position which summarize CNA recurrence, amplitude, and focality across samples. Focality is determined by a model of CNA formation which considers the potential joining of non-contiguous genomic regions during chromosomal rearrangement. GTS identifies genomic loci which appear to be focally targeted by CNA, even if seen rarely in the dataset. The algorithm is implemented in R (GTS R package, available at http://cbio.mskcc.org/brennan).

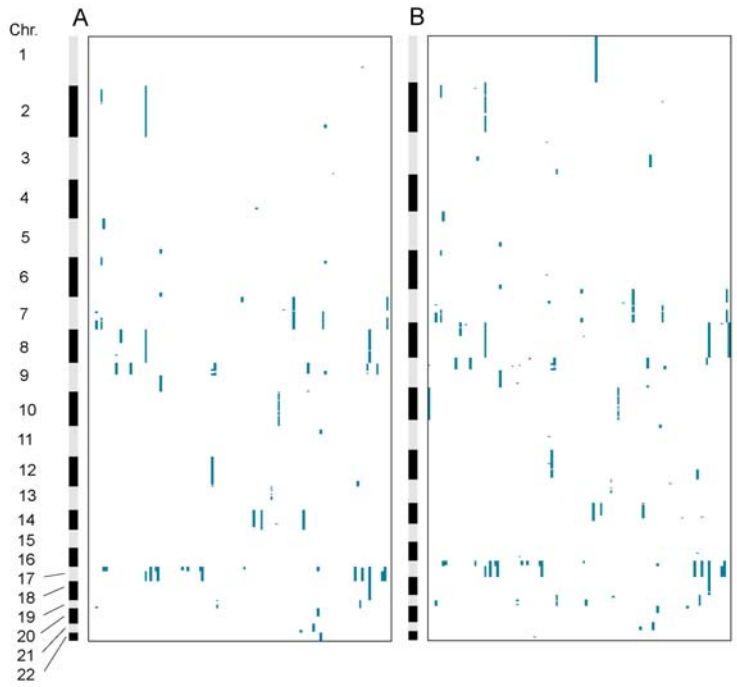
GTS cross-platform comparison: 139 unique tumor profiles were identified for which SNP data passed QC (Broad Institute, Affymetrix platform) and for which matching aCGH datasets were available (MSKCC, Agilent platform). To account for platform differences in signal saturation, segment means were normalized between platforms according to a polynomial regression. GTS was run with gene weighting and chromosomal linkage. Genes were ranked by GTS scores derived from each platform, excluding regions of known or suspected copy number variation

(CNV). The 200 genes with top-ranked GTS scores in the aCGH dataset were compared to the GTS rankings derived from the SNP dataset. For amplifications, 78% of the top 200 aCGH-identified genes were ranked within the top 400 SNP-identified genes. For deletions, the comparable overlap was 63%. The regions identified in common are given in Supplemental Table 4. In addition to the 139 paired samples in the comparative study, GTS was also run separately on the full set of 203 aCGH profiles (Agilent, MSKCC).

Enrichment for cancer-relevant genes at loci of focal CNA was assessed by Fisher's Exact test, considering the subset of genes from the Cancer Gene Census which did not reside in regions of known or suspected CNV. There is significant overrepresentation of cancer-relevant genes in these focally altered sets, as assessed by comparison with the Cancer Gene Census list [82], both for amplified regions (odds ratio 7.34, $p<0.00001$) and deletions (odds ratio 3.5, $p=0.01$).

C. Loss of Heterozygosity Analysis (LOH)

Only samples with paired normals were used for LOH analysis. In total, 137 samples from the Illumina platform and 123 from the Affymetrix 6.0 platform were used.



Supplementary Methods Figure 4. **Loss of Heterozygosity Analysis** The left and the right panel show the copy neutral LOH regions from the Affymetrix and Illumina arrays, respectively. The presence of copy-neutral LOH, which is most common on 17p, and which would impact the p53 gene, is shown for 123 identically sorted samples."

Affymetrix data

Allele-specific copy numbers and genotypes for normal samples were generated using the Birdseed algorithm and Affymetrix 6.0 array pipeline for all SNP probes [64,65]. SNP probes that were not heterozygous in a normal sample were discarded from its matched tumor sample. For each remaining SNP locus, each allele was then divided into either max or min channels, depending on which allele was greater than the other. The max and min channels were segmented separately using circular binary segmentation (CBS). The combined list of breakpoints from both channels was used to calculate the median copy number of each segment for each channel. Segments less than 10 SNPs in length were removed before further analysis.

Calls were made for copy neutral and copy loss LOH separately based on the copy number of the min and max allele. Copy neutral LOH was called when the min allele was <0.5 copies and the max allele was between 1.5 and 2.3 copies. Copy loss LOH was called when the min allele was less than 0.5 copies and the max allele was less than 1.5 copies.

Illumina data

The allelic ratio differences between each tumor and its normal control (delta B frequencies) were calculated at markers where the normal sample was genotyped as heterozygous. These values were subjected to circular binary segmentation (alpha=0.001, nperm=5000) to identify segments of LOH. The copy number data was then used to create a matrix, where 1 represents copy number change (post-segmentation log ratio means > 0.2 or < -0.2) and -1 represents no copy number change (means between -0.2 and 0.2). The LOH data was multiplied by this copy number matrix and cutoffs were applied to call LOH. Copy loss LOH was called when values were >0.25 and copy neutral LOH was called when values were < -0.25. Segments less than 10 SNPs were removed before further processing.

SECTION V. EXPRESSION PROFILING

A. Affymetrix Exon 1.0

Sample verification and RNA QC

Total RNA samples (n = 201) were received from Biospecimen Core Resource (BCR). Samples were normalized to approximately 100ng/ul concentration to perform the sample QC. Total RNA concentration, quality and protein contamination were determined by Nanodrop measurements. RNA integrity number (RIN) and 28s/18s ratio were determined by the Bioanalyzer (Agilent, Santa Clara, CA). To evaluate the possible DNA contamination in RNA, quantitative RT-PCR was performed using iScript one Step RT-PCR Kit SYBR Green assay, and delta CT values were computed against controls to check if the samples exceeded the genomic DNA contamination of 10ng/ul. All the quality values computed at LBNL CGCC were used to compare to the quality data provided by BCR. For each microarray experiment, with each batch of GBM samples, we included three universal RNA samples as controls for the experiment. We used Universal Human Reference RNA (Stratagene) Cat# 740000 (Stratagene, La Jolla, CA.), Human Universal Reference Total RNA Cat# 636538 (BD Clontech, Palo Alto, CA), and Brain total RNA Cat# R1234035-50 (Biochain Institute, Hayward, CA).

Whole transcript sense target labeling assay

2 µg of total RNA was subjected to ribosomal RNA removal procedure using Ribominus kit by Invitrogen Corporation (Carlsbad, CA). Double-stranded cDNA was synthesized from rRNA depleted RNA with random hexamers tagged with a T7 promoter sequence (T7-$(N)_6$ primer). The Double-stranded cDNA was then used as a template for T7 RNA polymerase producing cRNA. A second cycle of cDNA synthesis was performed using random hexamers to reverse transcribe the cRNA from the first cycle to produce single-stranded DNA (using dATP, dTTP, dGTP, and dUTP) in the sense orientation. cDNA was fragmented using DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE1). The fragmented DNA was then labeled with terminal deoxynucleotidyl transferase that conjugates biotinylated nucleotides. 5.5 µg of this biotin–labeled DNA was hybridized overnight with Affymetrix Human Exon1.0 ST microarrays and washed and scanned on Affymetrix GeneChip® Scanner 3000 7G scanner with an autoloader, according to the instructions from Affymetrix GeneChip Whole-TranscriptSense Target–Labeling Assay manual. Each scanned CEL image of the array was checked for any significant artifacts.

Data Processing

RMA was applied in combination with affymetrix.aroma to all 201 CEL files that met final quality control. This generated gene centric expression values, using a CDF file based on remapping of probes to the human genome 36.1 resulting in expression values for 18,632 genes.

B. Agilent 244K Whole Genome Expression Array

mRNA labeling

One to 2 ug of total RNA of sample and Stratagene Universal Human Reference were amplified and labeled using Agilent's Low RNA Input Linear Amplification Kit. The total yield of amplified RNA (aRNA) and Cy dye incorporation was measured by NanoDrop.

Array Hybridization and Imaging

Sample and reference (7-10 ug of each) were co-hybridized to a Custom Agilent 244K Gene Expression Microarray. Arrays were scanned on an Agilent Scanner and probe information was obtained with Agilent's Feature Extraction Software. Each scanned image is viewed for visible artifacts, and if multiple artifacts are present, the array is rejected. Agilent Feature Extraction software creates a QC report for each array that includes: (1) *Net Signal Statistics:* Signal range distributions for the red and green channels are presented and compared. Samples with large differences between the red and green channel for net signal are flagged as samples/arrays to be watched. (2) *Distribution of Outliers:* Samples with the % feature non-uniformity >1% are flagged as samples/arrays to be watched. (3) *MA plots:* Log of the Processed Signal is plotted versus Log of the Ratio (R/G) for each gene to help identify biases in intensity or dye. (4) *Reproducibility of SpikeIns (an internal hybridization control):* reproducibility of Agilent SpikeIns are measured by % coefficient of variation (<15) and SpikeIn linearity with R^2 values close to 1. If any array fails three of the QC criteria it is rejected. The 206 samples used in this study all passed quality control.

Data Processing

Data was lowess normalized and the ratio of the Cy5 channel (sample) and Cy3 channel (reference) was log2 transformed to create gene expression values for 18,624 genes.

C.Agilent 8x15K Human microRNA Microarray

miRNA Labeling

100-400ng of total RNA was labeled by ligation to cyanine 3-pCp molecules using the Agilent miRNA Micorarray labeling protocol (Agilent Technologies, Santa Clara, CA) using T4 ligase (NEB, Ipswich, MA).

Array Hybridization

Labeled miRNAs were hybridized to Agilent 8 x 15K Human miRNA-specific microarrays overnight. Arrays were scanned on an Agilent Technologies Scanner and probe information was obtained with Agilent's Feature Extraction Software. Each scanned image is viewed for visible artifacts. Agilent's Feature Extraction output reports four main microRNA-specific quality check criteria, including: (1) Additive Error Estimate, measure of the background. Samples with additive error between 5-12 counts/pixel are flagged as watched, samples with additive error greater than 12 counts/pixel are flagged as failed. (2) Percentage of Feature Population Outliers: samples with populations outlier between 7-10% are flagged as watched; samples with population outlier greater than 10% are flagged as failed. (3) Median Percent Coefficient of Variation (%CV) for replicate probes: measure of reproducibility. Samples with %CV between 8-15% are flagged as watched; samples with %CV greater than 15% are flagged as failed. (4)

75th Percentile of Total Gene Signal. Samples with 75th percentile total gene signal less than 35 are flagged as watched. Any sample that failed any of the first three criteria is repeated. All samples used in this study passed quality control.

Data Processing
Data was quantile normalized on the probe level. Signals from probes measuring the same microRNA are summed up to generate gene-centric total gene signal, followed by log2 transformation. Distance Weighted Discrimination (DWD) method is applied to data for batch-correction.

D. Affymetrix HT-HG-U133A

Sample Labeling
One μg of total RNA was converted to complementary RNA (cRNA) target using the Genechip® HT One-Cycle cDNA synthesis Kit (Affymetrix 900687) and the GeneChip® HT IVT Labeling Kit (Affymetrix 900688). Total RNA was first reverse transcribed using a T7-Oligo(dT) Promoter primer in the first strand cDNA synthesis reaction. Following RNAse H-mediated second strand cDNA synthesis, the double stranded cDNA was purified and served as a template for in an in vitro transcription (IVT) reaction. The IVT reaction was carried out in the presence of T7 RNA Polymerase and a biotinylated nucleotide analog / ribonucleotide mix for cRNA amplification and biotin labeling. The biotinlyated cRNA targets were then cleaned up and fragmented.

Array Hybridization
Samples were analyzed using Affymetrix HT-HG-U133A peg arrays (Affymetrix 900751). The hybridization and subsequent washing and staining were performed on the Affymetrix GeneChip® Array Station (GCAS) automation platform.

Data Processing
Of the 205 samples received by the Broad Institute, 204 Profiles of good quality were generated, where a very low percentage present indicated that hybridization failed for one sample. RMA (1) was applied in combination with affymetrix.aroma (2) in order to generate gene centric expression values, using a CDF file based on remapping of probes to the human genome 36.1. This resulted in expression values for 12,042 genes.

E. Creation of a unified Expression Dataset

For all gene expression analyses, a single gene expression data set was created. Each Affymetrix expression data sets were log transformed and the mean (or median) value was subtracted. As the Agilent platform generates log ratio data, only mean subtraction was applied. The resulting three expression data matrices were merged using the median value where three measurements were available (199 samples and 11,681 genes). The average value was used where two values were available, and the single value was used when a single value was available. Following this method, a data set consisting of 206 samples and 19,692 genes was generated.

SECTION VI. DNA METHYLATION PROFILING

Our approach for the TCGA is a two-tiered one which, first, involves pharmacological treatment of representative human GBM cell lines with DNA demethylating agents followed by transcriptome analysis. In the second tier, candidate genes identified from cell lines are used to generate a custom Illumina GoldenGate array with the capacity to monitor DNA methylation at a single CpG dinucleotide within 1,498 gene promoters. By including also the Illumina DNA Methylation Cancer Panel I platform, which queries methylation of 808 gene promoters, we analyzed more than 2,300 loci in the collection of TCGA GBM samples.

A. Identification of Candidate DNA Hypermethylated Genes

Cell treatment.
We performed drug treatment of four GBM cell lines (U87, T98 and D54MG; obtained from Dr. Greg Riggins, Johns Hopkins Oncology Center) and the human glioblastoma derived neurosphere cell line HSR-GBM1 (previously designated 20913; obtained from Dr. Angelo Vescovi, StemGen Inc.)[83] according to protocols developed for identification of the hypermethylome in colorectal cancer cell lines[84]. Briefly, log phase cells were cultured with 5 μM 5aza-2'-deoxycytidine (DAC; Sigma) for 96 hours, replacing media and DAC every 24 hours; or 300 nM Trichostatin A (Sigma) for 18 hours. Mock treatments were performed in parallel with PBS or ethanol instead of drugs.

Microarray analysis.
Total RNA was isolated, quantified, and checked for purity and integrity as previously described[84]. Sample amplification, labeling, and purification were carried out as previously described[84] using reagents and protocols from Agilent Technologies. Whole transcriptome screens were performed using Human 4x44K arrays from Agilent Technologies and the Agilent G2565BA scanner[84]. Mock and DAC samples were co-hybridized on a single array in parallel with mock and TSA samples and the complete collection of arrays have been deposited in the GEO Database (http://www.ncbi.nlm.nih.gov).

Data analysis and identification of hypermethylome candidate genes.
Raw data were processed and analyzed using the R statistical computing platform[85] and packages from Bioconductor bioinformatics software project[86]. The log ratio of red signal to green signal was calculated after LoEss normalization as implemented in the limma package from Bioconductor[87]. Mock/TSA changes (X axis) were plotted against mock/DAC (Y axis), and the characteristic spike of gene expression changes including DNA hypermethylated genes was visualized. Top tier (TT) probes were identified as having a greater than 2-fold change after DAC treatment and between 0.7 and 1.4 fold change after TSA treatment. Probes belonging to the next tier (NT) had identical parameters to the TT except that they fell within a zone of DAC responsiveness between 1.4 and 2.0 fold. In total, 3,703 genes were identified: 578 from TT, 1,070 from NT of U87; 573 from TT, 980 from NT U87 dye swap; 945 from TT, 1,225 from NT of T98; 272 from TT, 453 from NT of D54MG. It should be noted that the total number of genes (3,703) is lower than the sum of all TT and NT genes due to gene overlap.

Localizing probes to genes

The Agilent 4x44k human whole transcriptome array contains 1,639 control probes and 41,000 unique probes, containing 30,982 putative unique genes. Using the EnsEMBL Perl Application Programming Interface we mapped the exact genomic location of the 41,000 unique probes on EnsEMBL release 42 (NCBI assembly 36). Approximately 70% of the probes showed identity with an overlapping RNA species or transcript. We located another 5% of probes by extending the search for transcript matches 2 kb 5' or 3' of the probe location. Together, these approaches linked 30,854 probes (74%) with a transcript. The 3,703 genes from the GBM hypermethylome were mapped to this database, and 2,811 genes were identified.

Identifying CpG Islands

The 2,811 GBM candidate hypermethylome genes were first searched for a CpG island using the CpG island definition implemented in the newcpgreport package of EMBOSS[88,89] (http://emboss.bioinformatics.nl/) with a sliding window of 100 nucleotides, a minimal length of 200 nucleotides, a minimal GC% of 50% and a minimal observed/expected ratio higher than 60%. Application of these criteria generated 68% of genes with CpG islands. We further extended our search to sequences 300 nucleotides up and downstream of the transcription start site (TSS) and genes containing CpG islands identified in both searches were pooled[90]. Taken together, we identified CpG islands in 72% of the candidate hypermethylome genes.

Eliminating promoters with repeat sequences.

The presence of low complexity CpG-rich sequences within gene promoters may bias the identification of CpG island containing genes within the hypermethylome. We used the RepeatMasker (http://www.repeatmasker.org) to eliminate gene promoter regions containing LINE, SINE, ALU and other repetitive genomic elements. This analysis reduced the number of probes associated with a CpG island by 6%.

Selecting Autosomal Genes.

Non-autosomal genes may undergo allelic changes in expression due to DNA methylation dependent X-inactivation, perhaps resulting in spurious methylation values in human tumor samples. Thus, only autosomal genes were considered for the GoldenGate custom GBM platform.

B. Generation of a Custom GoldenGate Platform.

All autosomal transcript and probe IDs selected from the TT and NT probe lists were annotated as gene IDs that include the Ref Seq ID, NCBI Gene ID, HGNC ID and precise genomic coordinates. The Gene IDs for all candidate loci were submitted to Illumina for GoldenGate Methylation probe generation. Illumina first provided an *in silico* probe design for each gene using genomic coordinates between -500 and +200 nucleotides relative to the TSS. Oligomer sequences with known polymorphisms were removed from selection, since this could result in annealing mismatches during analysis of primary tumor tissue samples.

Each designed reaction is assigned a predictive performance score using a proprietary Illumina algorithm. This score is scaled from 0 to 1 and is based on performance experience at Illumina with prior designs and considers various oligonucleotide characteristics, such as G:C content, CpG density, self-complementarity, cross-hybridization to bisulfite-converted sequences

for other regions in the genome, as well as other parameters. We selected one reaction with the highest quality score for each gene locus and the reactions were ranked in order of decreasing quality score. The top 1,536 probes are retained for submission to Illumina based on quality score (a vast majority of which were above 0.8) and probes with the same quality score are resolved by transcript ranking. Illumina synthesized An Oligonucleotide Pool for Methylation Assays (OMA) specifically for use with TCGA GBM specimens. We refer to this panel as OMA-003.

In addition to the custom Illumina Goldengate gene panel, we also determined DNA methylation levels of 1,505 CpG dinucleotides spanning 808 gene regions on the commercially available Illumina DNA Methylation Cancer Panel I. This array is pre-selected to include genes previously shown to be methylated in human cancers, as well as candidate tumor suppressor genes, oncogenes, genes involved in DNA repair, apoptosis, cell cycle regulation, differentiation, and imprinting. We refer to this panel as OMA-002.

Illumina Bead Array Technology to generate methylation data.

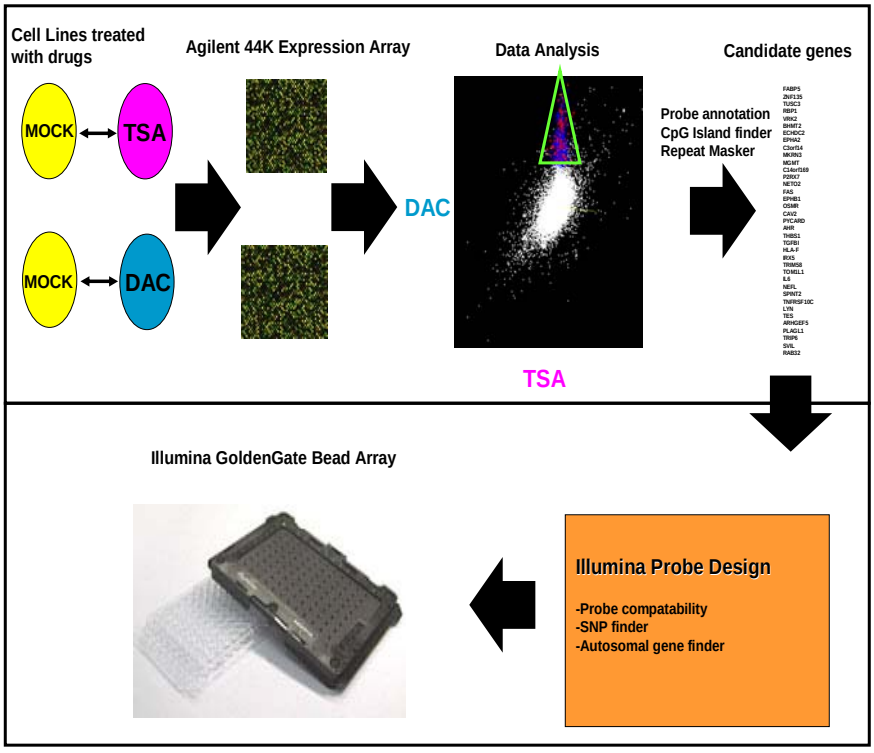
We processed 213 TCGA GBM samples from 195 patients. Each GBM tumor sample (1 µg genomic DNA) was converted with bisulfite using the EZ96 DNA Methylation Kit (Zymo Research, Orange, CA, cat #D5004) and eluted in an 18 µl volume. We retained 3 µl for use in post-bisulfite quality control experiments to determine the completeness and recovery of bisulfite conversion in the sample set. The Illumina GoldenGate DNA methylation assays were performed on all bisulfite-converted TCGA samples according to manufacturer's specifications for the custom GBM reaction set. A pair of probes are used for each assay, one designated M capturing methylated molecules, retains the complement to the unconverted cytosine, while the second, designated U complements the converted uracil. Intensities for each probe are an average of approximately 30 background subtracted, replicate measurements. The beta value, the calculated DNA methylation value for each locus is determined as: M / (U + M). We extracted U and M intensities and calculated beta values and detection p-values (the statistic after comparison of the intensities for each locus versus a panel of negative controls) for each locus and sample according to Illumina specifications. Beta value measurements with accompanying detection p-values > 0.05 were not significantly different from the panel of negative controls and were re-labeled as "N/A." After reducing the custom OMA-003 to practice on TCGA and control samples, we determined that 38 reactions performed poorly in more than 50% of TCGA samples, so these were masked from all analyses, leaving a total of 1,498 total reactions. For both OMA-002 and OMA-003 data sets, we also eliminated samples for which < 80% of the data points with detection p-value > 0.05 in order to ensure that data from high quality analyzed samples were released.

Selecting Cancer-Specific Hypermethylated Genes

Using Illumina OMA-002 or OMA-003 β-values obtained from the TCGA tumor samples, we first defined criteria to reduce the number of probes with high methylation in brain samples from non-cancer patients, and showing appropriate values for cell line controls, to generate a list of "cancer-specific" hypermethylated genes. Prior to filtration, DNA methylation profiles for patients with multiple samples were averaged. When multiple probes were available for the same gene, these were not averaged, but left in the dataset, in parallel. Steps of the algorithm include:

1. β values were scaled for each probe so that β' = (β - min(β))/((max(β) - min(β)). This scaling allows probes with a limited dynamic range of β values but which otherwise perform well available for downstream filtration.
2. Establish criteria, by default β' L = 0.25 and β' H = 0.75, to delineate nominally unmethylated, partially methylated and fully methylated sample groups for each probe.
3. The genes must satisfy these criteria:
   (a) unmethylated in negative control samples
   (b) methylated in positive control samples
   (c) unmethylated in all available normal samples
   (d) partially methylated (β' $\geq$ 0.25) in at least 5% of primary tumors
   (e) fully methylated (β' $\geq$ 0.75) in at least 2.5% of primary tumors

Because there are so few normal brain and so many GBM tumor samples, we expect that even if normals and tumors are equally (un)methylated, any outlying samples with high methylation values will tend to be tumors. In general, by requiring a few fully methylated samples, we hope to reduce the total number of partially and completely methylated samples required while still exercising some control over type I error. Under the assumption that a null correlation coefficient taken over N samples is approximately normally distributed with variance equal to 1/N, the threshold of -0.3 for the correlation coefficient would roughly correspond to a p-value of 0.005, but more importantly, we found this threshold to correspond to promising patterns of methylation which may be found in the Level 4 data on the DCC.



Supplementary Methods Figure 5. Schematic of the Methylation Screen. A two-tiered approach to identify and screen the GBM hypermethylome was developed. The first tier involves identification of the hypermethylated genes in GBM cell lines. GBM cell lines were treated with DAC or TSA, and gene expression monitored on 4x44K Agilent arrays. Mock/TSA changes (X axis) were plotted against mock/DAC (Y axis), and the characteristic spike of gene expression changes resulting from demethylation indicated by a green triangle. After data analysis candidate hypermethylated genes were filtered as shown and Illumina Goldengate probes identified. A custom Illumina Bead Array containing 1,498 probes allowing methylation analysis of an equal number of genes was generated.

SECTION VII. PATHWAY ANALYSIS

Genomic profiles

For the analysis of signaling pathways affected in glioblastoma, we included all validated somatic mutations reported in Supplemental Table 6 as well as gene-specific copy-number alterations (Section IV). Note that several types of alteration were excluded from the pathway analysis reported here: hemizygous deletions, alterations of microRNA genes and epigenetic alterations.

Copy number data from four different data acquisition platforms were analyzed using one or more of three different computational methods and reported as gene-specific amplification calls (gain or high-level amplification) and deletion calls (hemizygous or homozygous deletion). The following combinations of computational methods and experimental platforms were used:

- GISTIC on all four platforms (Affymetrix SNP6, Broad; Agilent, HMS; Agilent, MSKCC; Illumina, Stanford)
- RAE on Agilent, MSKCC
- GTS on Agilent, HMS

See Section IV for details on platforms and methods. Consensus calls from the different analyses were derived by requiring that a call was supported by at least two independent platforms and two independent methods.

Integration of mutation and copy-number data with pathway information

Initial inspection of validated somatic mutations and copy number variation using the Cancer Genome Workbench (CGWB, http://cgwb.nci.nih.gov) indicated that the majority of genomic alterations affected three well-known signaling pathways: the TP53 apoptotic pathway; the RB1 cell cycle arrest pathway; EGFR and other growth-factor receptors and PI-3 kinase / AKT signaling.

To further investigate the involvement of these pathways, we gathered detailed pathway information from the literature and pathway databases. We based our selection of genes and interactions on a pathway diagram published in a review article of alterations in glioblastoma [91], and then used pathway data from a number of publicly available databases via the Pathway Commons portal (http://www.pathwaycommons.org) to expand and modify the core pathway map at different levels of detail. Diagrams of the three individual sub-pathways are used in Fig. 5, and the more detailed global pathway figure, which shows connections between the sub-pathways, is used in Supplementary Figures 7 and 8. These pathway maps focus on glioblastoma-relevant genes with alterations in one or more samples in this study, i.e., genes or gene products without alterations are not shown.

We systematically tallied all mutations and all copy number alterations in the genes in these pathways. Using both mutation and copy-number data, alternative alterations of individual genes and sub-pathways in different samples emerged. These patterns are shown in a sample-by-sample fashion for mutations and copy-number alterations in 91 samples in Supplemental Table 9, and for copy-number alterations in 206 samples in Supplemental Table 8. The overall

frequencies of alterations in each gene are shown in Fig. 5 and Supplementary Figures 7 and 8. In these figures, high-level amplification events are indicated in shades of red, and homozygous deletion events are indicated in shades of blue. Mutations in genes known to be frequently deleted were considered as inactivating mutations and also indicated in shades of blue, while mutations in genes known to be frequently amplified were considered as activating mutations and indicated in shades of red. Saturated red indicates a frequency of activation of 25% or higher, and saturated blue indicates a frequency of inactivation of 25% or higher. We note that frequent alterations in glioblastoma within the RB1 pathway were previously reported based on lower-resolution technology [92].

The more comprehensive pathway figure (Supplementary Figures 7 and 8) was generated using the Mondrian software plugin (http://cbio.mskcc.org/mondrian/) in the Cytoscape network visualization software (http://cytoscape.org). Detailed data on all of the genetic alterations and expression profiles derived in this TCGA study can be viewed using Mondrian in any one or two of three dimensions: (1) by sample (2) by gene affected and (3) by data type. Details can be viewed either as heatmaps or as color levels in a GBM-specific pathway map, with interactive panning through the set of GBM samples. Software access, as well as access to the interactive data profiles, is available on the MSKCC Computational Biology Cancer Genomics website (http://cbio.mskcc.org/cancergenomics/gbm/).

Statistical Methods

On the basis of a statistical model of random assignment of alterations to genes, the probability that a given sample would have at least one aberration in each of the three pathways was calculated by identifying for each pathway the number of samples that did or did not have at least one aberration and calculating from those numbers the expected number of samples for which all three pathways would have at least one aberration. The p-values for mutual exclusivity of aberrations within pathways were calculated by comparing the expected (from the background model) and actual (as observed) numbers of samples among those with at least one aberration that had exactly one aberration. In the case of the Rb pathway, CDKN2B was omitted from the calculation because of its similarity to CDKN2A.

Fisher's exact odds ratios and one-tailed p-values, shown in Supplemental Tables 10 and 11, respectively, were calculated in the open-source R programming environment (R.app GUI 1.23 (4932), S.Urbanek & S.M.Iacus, R Foundation for Statistical Computing, 2008) using the "fisher.test" function. Each colored gene-sample pair in Supplemental Table 9 was coded as a "1"; all others were coded as "0" (no alteration). The resulting frequencies of 1's and 0's generated a two-by-two table for each gene. Positive and negative associations (conditioned on the marginals) were then assessed by Fisher's exact test. Interpretations of the odds ratios and one-tailed p-values are given in the respective Table legends.

D. REFERENCES

51. Kahn, A.B. et al. SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis. *BMC Bioinformatics* 8, 75 (2007).
52. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36, D13-21 (2008).
53. Ellis, M.J., Dixon, M., Dowsett, M., Nagarajan, R. & Mardis, E. A luminal breast cancer genome atlas: progress and barriers. *J Steroid Biochem Mol Biol* 106, 125-9 (2007).
54. Nickerson, D.A., Tobe, V.O. & Taylor, S.L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* 25, 2745-51 (1997).
55. Chen, K. et al. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res* 17, 659-66 (2007).
56. Zhang, J. et al. Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB). *Genome Res* 17, 1111-7 (2007).
57. Zhang, P. et al. Gene functional similarity search tool (GFSST). *BMC Bioinformatics* 7, 135 (2006).
58. Dutt, A. et al. Drug-sensitive FGFR2 mutations in endometrial carcinoma. *Proc Natl Acad Sci U S A* 105, 8713-7 (2008).
59. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-6 (2008).
60. Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R. & Easton, D.F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* 173, 2187-98 (2006).
61. Benjamini, Y. & Hochberg, Y. *J. Roy. Stat. Soc.* 57, 289-300 (1995).
62. Getz, G. et al. Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* 317, 1500 (2007).
63. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* 98, 31-6 (2001).
64. Monti, S. et al. DNA copy number inference pipeline for Affymetrix SNP6.0 arrays *http://www.broad.mit.edu/publications* (2008).
65. Korn, J. et al. Integrated genotype calling and assocaition analysis of SNPs, common copy number polymorphisms, and rare CNVs. *Nat Genet* in press(2008).
66. Beroukhim, R. et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* 104, 20007-12 (2007).
67. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557-72 (2004).
68. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23, 657-63 (2007).
69. McCarroll, S. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. . *Nature Genetics* in press(2008).
70. Conrad, B. Novel procedures for high-throughput analysis of a frequent insertion-deletion polymorphism in the human T-cell receptor beta locus. *Immunogenetics* 56, 220-4 (2004).

71. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 38, 75-81 (2006).
72. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 38, 82-5 (2006).
73. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet* 36, 949-51 (2004).
74. Locke, D.P. et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* 79, 275-90 (2006).
75. McCarroll, S.A. et al. Common deletion polymorphisms in the human genome. *Nat Genet* 38, 86-92 (2006).
76. Redon, R. et al. Global variation in copy number in the human genome. *Nature* 444, 444-54 (2006).
77. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* 305, 525-8 (2004).
78. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* 77, 78-88 (2005).
79. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* 37, 727-32 (2005).
80. Taylor, B.S. et al. Functional Copy-number Alterations in Cancer. *PLoS ONE* in press(2008).
81. Wiedemeyer, R. et al. Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. *Cancer Cell* 13, 355-64 (2008).
82. Futreal, P.A. et al. A census of human cancer genes. *Nat Rev Cancer* 4, 177-83 (2004).
83. Galli, R. et al. Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res* 64, 7011-21 (2004).
84. Schuebel, K.E. et al. Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. *PLoS Genet* 3, 1709-23 (2007).
85. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J Comput Graph Stat* 5, 299-314 (1996).
86. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, R80 (2004).
87. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* 31, 265-73 (2003).
88. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* 196, 261-82 (1987).
89. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16, 276-7 (2000).
90. Takai, D. & Jones , P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99, 3740-6 (2002).
91. Furnari, F.B. et al. Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev* 21, 2683-710 (2007).
92. Ichimura, K., Schmidt, E.E., Goike, H.M. & Collins, V.P. Human glioblastomas with no alterations of the CDKN2A (p16INK4A, MTS1) and CDK4 genes have frequent mutations of the retinoblastoma gene. *Oncogene* 13, 1065-72 (1996).