Supplementary Information

Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma

with a description of the method

Genomic Identification of Significant Targets in Cancer (GISTIC)

Rameen Beroukhim*, Gad Getz*, Leia Nghiemphu, Jordi Barretina, Teli Hsueh, David Linhart, Igor Vivanco, Jeffrey C. Lee, Julie H. Huang, Sethu Alexander, Jinyan Du, Tweeny Kau, Roman K. Thomas, Kinjal Shah, Horacio Soto, Sven Perner, John Prensner, Ralph M. Debiasi, Francesca Demichelis, Charlie Hatton, Mark A. Rubin, Levi A. Garraway, Stan F. Nelson, Linda Liau, Paul Mischel, Tim F. Cloughesy, Matthew Meyerson, Todd A. Golub, Eric S. Lander, Ingo K. Mellinghoff & William R. Sellers

*Contributed equally.

A Frequently Asked Questions (FAQ) document is available at http://www.broad.mit.edu/cancer/pub/GISTIC/

Contents

Supplementary Figure Legends

SI Figure 1. Fewer significant regions are identified in an analysis restricted to primary GBMs. The results of the original GISTIC analysis of glioma (displayed as in Fig. 2b) are presented alongside a similar analysis of only the primary GBMs in the dataset. All of the regions that are significant among primary GBMs are also significant in the larger dataset including secondary GBMs and lower-grade gliomas. Some events, such as 8q gain and 19q loss, are significant in the larger dataset but not among only primary GBMs. This loss of significance may be either due to a decreased prevalence of these events among primary GBMs, or decreased power to detect low-prevalence aberrations in a smaller tumor set.

SI Figure 2. GISTIC applied to different glioma datasets generates nearly identical results. The results of the original GISTIC analysis (displayed as in Fig. 2b) are presented alongside similar analyses of 178 tumors on 100K SNP arrays (1) and 37 tumors on 16K CGH arrays (2). Only minor differences in results are seen; these are due to differences in the distribution of glioma subtypes within each dataset, (a high proportion of grade III gliomas among the 178 tumors and of secondary GBMs in the CGH analysis) and to stochastic fluctuation. As expected, significant aberrations tend to reach higher levels of significance (lower q-values) as the number of samples increases.

SI Figure 3. Comparison between GISTIC analyses of glioma and lung cancer reveals distinct profiles. The results of the original GISTIC analysis of glioma (displayed as in Fig. 2b) are presented alongside a similar analysis of 81 lung cancer samples using 100K SNP arrays (3). The overall pattern is strikingly different, although both tumor types exhibit similar amplifications of chr7 (including *EGFR*)

and deletions of chr9p (*CDKN2A/B*) and chr13 (*RB1*). A more detailed analysis of the lung cancer genome using GISTIC is the subject of a forthcoming manuscript (4).

SI Figure 4. Broad amplification of chromosome 7 vs. focal amplification of *EGFR*. (a) A histogram of the copy numbers (displayed as $\log_2$ ratios) across samples at the *EGFR* locus shows tumors divide into 3 classes: $\log_2$ ratios less than 0.1 (unamplified, associated with 7^{norm}), between 0.1 and 0.9 (low-level amplifications, associated with 7^{gain}), and exceeding 0.9 (high-level amplifications, or $7^{gain}EGFR^{amp}$). No samples had log2 ratios between 0.7 and 1.3, suggesting a qualitative difference between 7^{gain} and $7^{gain}EGFR^{amp}$. Note that the values 0.1 and 0.9 coincide with θ^{amp} and θ^{hi_amp} (see SI Methods). (b) Copy-number profiles (displayed as $\log_2$ ratios, blue line) across chr7 (Mb coordinates on left) are displayed for representative samples with 7^{norm}, 7^{gain}, and $7^{gain}EGFR^{amp}$. The presence of low-level amplification at the *EGFR* locus does not imply a focal amplification. In fact, 42 out of 44 cases exhibited low-level amplification across most of the chromosome. The high level of *EGFR* amplification seen in $7^{gain}EGFR^{amp}$, however, is always focal and never extends over most of the chromosome.

SI Figure 5. MET/HGF$^+$ cell lines activate MET and AKT in an HGF-dependent manner but do not activate EGFR. (a) Treatment with SU11274 reduces MET and AKT activation in MET/HGF$^+$ cells. Whole-cell lysates from MET/HGF$^+$ (see Fig. 3) Hs683 and LN18 cells were obtained after 24-hour serum starvation in the presence of the indicated concentrations of SU11274. (b) MET and AKT activation in MET/HGF$^+$ cells are HGF-dependent. Whole-cell lysates were obtained after 24-hour serum starvation in the presence of (as indicated) anti-HGF antibodies (5 μg/ml) or SU11274 (2.5 μM), or with HGF (50 μg/ml) added for the final 10 minutes. (c) Neither the presence of 7^{gain} nor high levels of *EGFR* expression are associated with activation of EGFR. Cell lines were characterized as having

high *EGFR* expression if their median-normalized, median absolute deviation-scaled RNA expression levels (using only concordant *EGFR* probesets on Affymetrix U133 arrays) were greater than zero. None of these cell lines have focal amplification of *EGFR*. Immunoblots to the indicated epitopes were performed on whole-cell lysates prepared after 24-hour serum starvation. EGFR-dependent lung cancer cells (H3255) (5) were included as positive controls. (d) These cells also do not exhibit decreased viability when treated with the EGFR inhibitor erlotinib. Viability was measured using WST dye after exposure to inhibitor at the indicated concentrations for 96 hours.

SI Figure 6. Flow chart representing the components of the 4 stages of the GISTIC algorithm. Each step in GISTIC is represented by a block or bullet and is described in a subsection of Supplementary Methods.

SI Figure 7. Evolution of the data as it progresses through GISTIC. (a) Raw signal intensities ($\log_2$ scaled) are displayed across the genome (y axis) for 141 tumors and 33 normal samples (y axis; sample characteristics indicated on top). Copy-number aberrations are difficult to distinguish at this step, prior to division by normal controls. (b) Raw genotyping data are displayed for these same samples. Large regions of homozygosity (seen as stripes lacking the usual frequency of yellow heterozygous markers) likely represent LOH events. (c) Batch effect correction removes artifactual copy-number changes associated with date of data generation. (Top panel) Inspection the dates on which data was obtained (batches, indicated by color bar at the top) shows that high-level changes in signal intensity are restricted to a single batch in some markers (arrows). If not corrected, these signal intensity changes will be seen as recurrent copy-number changes. (Bottom panel) After batch correction, these artifacts are removed. (d) Normalized signal intensities displayed for the tumor samples only reveal copy-number aberrations,

including losses (blue) and amplifications (red). Although systematic errors have been minimized by batch correction and selection of appropriate normalization controls (see Supplementary Methods), substantial random errors persist. The last row of sample characteristics (top) indicates samples with low tumor purity (in green; see Supplementary Methods), which are removed from subsequent steps. (e) Histograms of normalized data in the quality control step enable identification of datasets in which contamination with normal cells obscures the signal contributed by tumor cells. (Top panel) Histograms depicting normalized signal intensity distributions that would be expected from the indicated tumor purities. A pure tumor would be expected to display separate peaks corresponding to the different copy-number levels in the tumor. The width of each peak will vary according to the level of noise in the array, and the distance between peaks represents the amount of signal contributing to copy-number estimates. As the proportion of tumor cells decreases, so does this signal, leading to a smaller distance between peaks. At low proportions of tumor, peaks associated with different copy-number levels become indistinguishable, indicating the signal is obscured by noise. (Bottom panel) Actual histograms (grey) and smoothed versions (dark lines) representative of the patterns seen among the 141 gliomas analyzed. These roughly correspond to the expected distributions seen in the top panel. (f) Segmented signal intensity data for the 105 glioma samples with high tumor purity reveal the copy number aberrations with much lower levels of random noise. (g) Segments with $\log_2$ signal intensity ratios greater than 0.1 are considered amplified and displayed here. (h) We identified loss and retention of heterozygosity events (blue and yellow, respectively) among the 105 tumors with high tumor purity by comparing the observed frequency of heterozygous SNP markers to the expected frequency in each region of the genome (6). (i) The frequency of amplification and average level of amplification across the genome are displayed in panels to the right of the amplification data (from panel [g]). High scores for either one of these indicates a high likelihood that amplifications in that region of the genome are not

solely chance events. GISTIC uses a *G* score (far right panel) that integrates both of these measures to identify aberrations that are associated with cancer. (j) Comparison of observed *G* scores to similar scores generated after permuting the marker labels allows us to determine the statistical significance of aberrations in each region (displayed on the right as FDR q-values to account for multiple hypotheses; see Supplementary Methods). Regions with q-values less than 0.25 (green line) are considered significantly aberrant. (k) "Peel-off" method identifies independent peaks within a statistically significant region. The top left panel displays a chromosome from an idealized set of tumors, with amplified regions in orange; q values associated with these regions are shown in the top right panel. For every chromosome in which some of these q values attain statistical significance (the significance threshold is denoted by the green line), the "peel-off" algorithm identifies the region with minimal q value (red line) as the primary peak. All aberrations involving the primary peak (marked by red stars) are then removed (faded orange, bottom left panel), and *G* scores and FDR q values are recalculated (bottom right) using only the remaining aberrations. If these reach significance, the region with minimal q value is selected as the secondary peak. The process iterates until no statistically significant regions remain.(l) In the case of chr7 amplfication, this "peel-off" algorithm enables us to identify separate peaks associated with *EGFR* and *MET*. The original amplification data for chr7 is displayed in the top panel along with the associated *G* scores. The entire chromosome is associated with *G* scores that are greater than the significance threshold, but a clear peak is observed at the *EGFR* locus. When we remove all amplicons that cover this peak region, we find a second peak that crosses the significance threshold at the MET locus (middle panel). When amplicons covering this second peak are removed, the remaining amplicons do not reach statistical significance (bottom panel).

SI Figure 8. Copy-number changes tend either to be focal or near the size of a chromosome arm. The distribution of sizes of all amplifications and deletions in the dataset is displayed. The majority of events fall into one of 2 peaks: either focal events covering less than 10% of a chromosome arm, or broad events covering more than 90% of a chromosome arm.

SI Figure 9. LOH is usually, but not always, associated with deletions. (a) The statistical significance of deletions (blue) and LOH (purple) are displayed as in Fig. 2b. All significant regions of LOH are also significant regions of deletion, with two exceptions: (i) a small region containing *EGFR* gives the appearance of LOH in highly amplified samples due to allelic imbalance, and (ii) chromosome 17p, containing *TP53*. (b) *TP53* primarily undergoes copy-neutral LOH. The top panel displays loss of heterozygosity (LOH, blue) and retention of heterozygosity (yellow) along chromosome 17 for 8 gliomas (labeled A-H) with LOH at the *TP53* locus. The bottom panel displays signal intensities (red = high, blue = low, white = neutral) and copy-number calls (red bar = amplified, blue bar = deleted) for those gliomas. LOH at *TP53* is associated with neutral copy numbers in gliomas A-G. Across our dataset, copy loss (as in glioma H) is seen in only 3 of 23 gliomas observed to have LOH at the *TP53* locus.

Supplementary Methods

Introduction: Genomic Identification of Significant Targets in Cancer (GISTIC)

We describe a general method for Genomic Identification of Significant Targets in Cancer (GISTIC). GISTIC can be divided into four stages (**SI Fig. 6**):

(1) Characterization of chromosomal aberrations on a per-tumor basis

(2) Aggregation of data from different tumors to differentiate between driver and passenger aberrations.

(3) Identification of peak regions most likely to contain the oncogene and tumor suppressor gene (TSG) targets.

(4) Classification of tumors on the basis of their driver aberrations.

Stage 2 contains the 2 central features of the algorithm: that it scores each genomic marker according to an integrated measure of the prevalence and amplitude of copy-number changes (and only prevalence in the case of LOH), and that it assesses the statistical significance of each score by comparison to the results expected from the background aberration rate alone.

In the following section we provide an overview of the motivations and methods behind each of the 4 stages. Detailed descriptions of each stage, to allow reproduction of the results, are included in the following 4 sections, with subsections dedicated for each block or bullet in **SI Fig. 6**. For clarity, the first time a parameter is described it is marked with a boldface font and the value that we used appears in parentheses. The evolution of the data as it progresses through the algorithm is visualized in **SI Fig. 7**.

Overview of the method

With data describing chromosomal aberrations in large tumor sets, the aberrations that drive tumorigenesis and the oncogenes and TSGs they most likely target can be identified if the following 4 issues are addressed. (1) The aberrations in each of the tumors must be accurately mapped. (2) Driver aberrations that rise above the background rate of random passenger aberrations must be identified. (3) For each driver aberration, the loci most likely to contain the targeted oncogenes or TSGs must be identified. (4) Tumors must be classified as to whether they are aberrant at the predicted driver loci, so

that the effects of those aberrations can be studied. GISTIC represents an example of such an approach, in which these 4 issues are addressed in the 4 stages of the algorithm.

Stage 1

In this stage, chromosomal aberrations are mapped in each tumor. Here, the goal is to maximize the accuracy with which these aberrations are identified, by (1) minimizing systematic error, (2) minimizing random error, and (3) discarding poor-quality datasets. Chromosomal regions with high signal intensities are designated as amplified, regions with low signal intensities are designated as deleted, and regions with an excess of homozygous SNP markers are designated as having lost heterozygosity.

Systematic errors arise when datasets from different samples are generated under slightly different experimental conditions. A primary example is *batch effect*, in which data generated on different days varies slightly. We limit batch effects on our copy-number assessments by using a batch effect correction module, in which we identify and correct markers that show consistent signal within batches but large variations between batches. Other experimental variables, such as day of manufacture of the array, or slight variations in PCR conditions, can also lead to systematic errors even between samples within a batch. Many array comparative genomic hybridization platforms minimize these by using 2-color systems in which tumor and control DNA are hybridized simultaneously to the same array. In single-color systems such as the Affymetrix SNP arrays that we use prominently in this study, these systematic errors can be minimized by selecting appropriate controls for each tumor, such that the controls share similar variations in their noise profiles across the genome.

Several methods exist to reduce the effects of random noise in copy-number datasets, most often by identifying regions of copy-number change and averaging the signal intensities for all markers within them (7). Examples include segmentation algorithms such as Circular Binary Segmentation (8) and Gain and Loss Analysis of DNA (GLAD) (9), Hidden Markov Model-based approaches (10, 11), and clustering methods (12). Each has advantages and disadvantages that may vary with the noise characteristics of the dataset. We used GLAD due to its high sensitivity for identifying copy-number changes (7). However, this high level of sensitivity occasionally leads GLAD to report non-existent copy-number changes in very small segments (fewer than 4 markers). We therefore filter these out.

In poor-quality datasets, the signal intensity variations due to copy-number changes are obscured by noise. We therefore identify high-quality datasets as having separate peaks, corresponding to different copy numbers, in histograms of the signal intensity data. Poor-quality samples, particularly those with extensive contamination with normal DNA, generate insufficient signal to distinguish separate peaks, and are discarded. Likewise, duplicate samples from the same individual, identified by similar SNP genotypes, are eliminated.

Stage 2

This stage contains the two core features of GISTIC (Figure 1). First, we score each genomic marker for the sources of evidence that it is in a region affected by driver aberrations (the *G*-score). Here, we treat amplifications, deletions, and LOH events separately—allowing for the possibility that a region could be significantly amplified and deleted simultaneously (for instance if an oncogene and TSG neighbor each other, with some samples amplified and others deleted). In the cases of amplifications and deletions, we assume that both the prevalence and average amplitude of these events independently indicate the likelihood with which a region is affected by such driver aberrations. Therefore, we use a simple integrated score of the prevalence of the copy-number change times the average ($\log_2$-transformed) amplitude. In the case of LOH, amplitudes do not apply and we therefore score each marker only by the prevalence of events.

Second, we compare these *G*-scores to the distribution of scores expected if only random aberrations were observed. This distribution can be determined by rescoring the genome after permuting marker locations within each sample; we instead derive a semi-exact estimate. The comparison of actual scores to those generated by our null model of random aberrations allows us to calculate the statistical significance of each *G*-score (represented by False Discovery Rate q-values(13)), representing the likelihood that the observed data could have been generated by chance alone.

Regions of the genome that are too frequently or highly aberrant to be explained by chance alone are selected as likely to harbor driver aberrations.

Stage 3

In this stage, GISTIC identifies the most likely locations of the oncogene or TSG targets of the driver aberrations identified in stage 2. This stage is designed with 4 considerations in mind: (1) these gene

targets are most likely to lie in the regions most frequently aberrant to the highest degree (similar to the minimal common region of aberration, with high-amplitude aberrations are given greater weight); (2) occasional random aberrations may occur near, but not overlapping, real oncogenes or TSGs, distracting us from their true locations; (3) a single region may contain more than one independently targeted gene; and (4) some aberrations may exert their effects through broad-based changes across much of the length of the aberration. This latter consideration is suggested (but not proven) by the high prevalence of broad aberrations that consistently affect large regions of the genome (near the size of a chromosome arm) (**SI Fig. 8**).

Given (1), for each region found in stage 2 to contain likely driver aberrations, we select the "peak" regions with maximal *G*-scores and (an equivalent statement) minimal q-values as most likely to contain the oncogene or TSG targets. In each case, we allow for (2) (the possibility that random aberrations are skewing the location of the peak) by leaving each sample out in turn, and recalculating the peak boundaries—only the widest boundaries are taken. We also allow for (3) (that a single region may contain two or more independent gene target) by applying a "peel-off" method designed to identify aberrations that overlap but are independently statistically significant. Finally, we allow for (4) by determining, for each peak region, whether the aberrations at this locus are primarily focal or broad, or whether both focal and broad aberrations are independently significant.

Stage 4

To determine the effects of driver aberrations identified in stage 2, we must classify tumors as to whether they have these aberrations. Because the peak regions are most likely to contain the oncogene or TSG targets of these aberrations, GISTIC first classify each tumor according to its copy-number status at the peak regions. For broad aberrations, which may be specifically disrupting a large region of the genome, GISTIC also classifies each tumor as to whether it is aberrant across most of the length of the region.

Detailed description

Required inputs

The inputs to GISTIC are the following files (details regarding software availability and exact file formats can be found at http://www.broad.mit.edu/cancer/pub/GISTIC)

(a) A **.snp** file that represents either the signal intensities or $\log_2$ ratio for each of the genomic markers (in our case, SNPs, although non-polymorphic loci interrogated by comparative genomic hybridization methods may also be used) across a set of samples.

(b) A **.loh** file representing the inferred loss of heterozygosity (LOH) status, either as discrete calls or probabilities (similar format to .snp files).

(c) A **sample info** file which denotes for each array its array name, sample name, tumor type, ploidy, paired normal, batch, gender and platform. Additional information for each sample can be supplied for visualization and correlation purposes.

(d) A **genome info** file with the location of each marker.

(e) A **cytogenetic info** file with cytoband locations.

(f) A **copy number variation** file with the locations of germline copy-number polymorphisms.

(g) A **transcript** database with gene locations.

(h) An optional list of **known target** gene symbols for visualization purposes.

(i) An optional list of **general cancer** gene symbols for reporting purposes.

(j) A **parameter file** with values for the various parameters used in the algorithm.

Stage 1: Characterization of chromosomal aberrations on a per-tumor basis

In this stage, we systematically characterize on a genome-wide basis the amplifications, deletions, and loss-of-heterozygosity (LOH) events affecting each tumor. We aim to reduce inaccuracies in these determinations due to systematic artifacts, random error, and poor-quality data. In the case of copy-number determinations, systematic errors are controlled by correcting for batch effect, selecting appropriate germline datasets for normalization, and controlling for germline copy-number polymorphisms. The effects of random noise are minimized by use of a segmentation algorithm and

application of a threshold for calling amplification or deletion that is rarely attained by fluctuations in segmented copy-number values in normal samples. Duplicate samples from one individual and samples with poor-quality data (i.e. copy-number changes were not reliably distinguishable) are eliminated.

The initial steps are aimed at controlling for systematic errors that can lead to false amplifications and deletions at a single genomic location across tumors. Even when these artifacts occur at a very low frequency, when we consider the hundreds of thousands of markers that may be present in the dataset, we are likely to encounter a few artifacts whose consistency across multiple tumors will lead them to appear even more significant than real changes associated with tumorigenesis. Therefore, controlling for these systematic errors is an essential step in a high-resolution genome-wide approach. We and others (14-16) have found that slight variations in experimental conditions between successive arrays can lead to these systematic changes in signal intensities. We therefore control for these experimental variations in two steps: (1) correcting for variations due to batch effect, which are defined by the date and core-facility in which the data was generated; and (2) selecting for normalization a set of normal samples that are most similar to the tumor sample according to their baseline signal intensity variations across the genome.

Source data

GISTIC can be applied to any dataset representing copy-number or LOH data measured across the genome. As an example of its application, we used 100K SNP array data from 141 gliomas, along with a set of normal controls. Here, probe-level signal intensity data were normalized to a baseline array with median intensity, using the invariant set normalization method (17). The signal intensity of each SNP was then obtained using a model-based (PM/MM) method (18). Genotyping calls were made by Affymetrix Genotyping Tools Version 2.0.

SI Fig. 7a-b show the raw signal and genotyping calls as heatmaps.

Data preprocessing

The noise in signal intensities is dominated by a multiplicative component. Hence, to make the noise constant across signal intensity, we $\log_2$-transform the data using a floor value of 1 to avoid small or

negative numbers. Next, we bring the samples to the same scale by subtracting the median value across all markers for each sample.

Batch effect correction

In this step, we assume that signal intensity variations due solely to batch effect are likely to be marked by their consistency within a batch, and variance from other batches. We therefore compare, for each marker independently, the distribution of signal intensities from all tumor and normal samples in a given batch to that of the tumor and normal samples from all other batches, using a variance-thresholded t-test with minimal variance of σ^2_{min} (=0.16 in our case), which represents the typical level of noise per marker and can be estimated using replicate datasets. For markers and batches where the t-test yields an asymptotic p-value less than $P_{batch\_effect\_cutoff}$ (= 0.001), a constant is added to the signal intensities of that marker in all samples in each variant batch, to yield the same mean signal intensity as all non-variant batches. Batches with fewer than N_{min} (=5) samples are not modified in this manner.

Among our data, 4.9% +/- SD 9.4% of loci were modified in each batch in this manner. The locations of these loci varied widely between batches, such that 63% of loci were corrected in at least one of the 14 batches (**SI Fig. 7c**). However, in a majority of cases these corrections were small, with the signal intensity difference averaging 4.2% +/- SD 2.5% of the unperturbed signal intensity. Thus the benefit from batch correction largely derives from correcting the small number of markers with more pronounced batch effects.

Selection of germline samples for normalization

In order to obtain copy number estimates for a sample we first calculate $\log_2$ tumor-to-normal copy-number ratios at each marker. These are calculated by subtracting the average of $\log_2$-transformed signal intensities from a set of normal controls from the $\log_2$-transformed signal intensity of the tumor. When examining the normal controls, we observed that subsets of samples exhibit systematic deviations in signal intensity across megabases of the genome. Replicate datasets representing the same sample often display different patterns of systematic deviation (data not shown). This suggests that some of the variation in signal intensities between samples is due to experimental factors that enter during data generation. Some of these experimental factors have been previously modeled (14, 15); it is possible that

many have not. We chose to correct the systematic effects due to these factors, including those that have not been modeled, simply by selecting the set of normal controls that share similar noise profiles to the sample being normalized. To this end, we identify the N_{close}(=5) normal samples that are closest to the sample being normalized measured by Euclidean distance between the $\log_2$-transformed signal profiles (again ensuring these profiles are at the same scale between samples by subtracting the median value across all markers for each sample), and use these for normalization.

SI Fig. 7d shows the normalized data as a heatmap.

Merging platforms

In some cases each sample is profiled using more than one platform, interrogating different sets of loci. For example, the 100K SNP arrays we used consist of independent 50K Xba and Hind array platforms. In this step, we merge the data from these platforms, interlacing markers according to position on the genome. In principle, this step can be used to merge data from different technologies such as array CGH or BAC arrays. However, care must be taken if the different platforms have highly variant dynamic rage or noise characteristics.

A more general problem is that of merging datasets in which each sample was assayed using a separate set of platforms. We do not address this issue here.

Quality control

In this step, samples with poor-quality data are removed. Copy-number profiles can suffer from either of two features that will make them non-informative. First, extreme levels of noise can lead to the inability to distinguish copy-number changes; second, high levels of contamination with normal cells (even in samples that appear highly enriched for tumor) can dampen the signal intensity differences between copy-number changes to the extent that copy-number changes are not robustly resolved. These two features work in tandem: a larger amount of contaminating normal DNA may be tolerated if the signal-to-noise ratio is high and small changes in signal can be robustly detected.

We assessed both noise level and normal contamination simultaneously by generating histograms of the log2 ratios collected at each autosomal marker locus (**SI Fig. 7e**). We first smooth the $\log_2$ ratios by taking the mean value across a running window of $\boldsymbol{H}_{\mathbf{window}}$ (=501) markers. The histogram is generated using a bin size $\boldsymbol{H}_{\mathbf{bin}}$ (=0.01), and smoothed by convoluting with a Gaussian distribution that has a standard deviation of $\boldsymbol{H}_{\mathbf{sigma}}$ (=0.05).

For a dataset from a homogenous tumor sample, we expect to identify separate peaks in this histogram corresponding to the separate copy numbers in the tumor's genome (**SI Fig. 7e**). As more contaminating normal DNA is mixed with the tumor's, however, the observed signal in all cases will approach normal levels and the peaks will tend to coalesce. Conversely, as the noise level increases, so will the width of each peak, resulting in a single broad hump from which separate peaks cannot be resolved. Therefore, each tumor whose smoothed histogram has only a single peak is marked as having failed quality control.

Some datasets may not display separate peaks despite high-quality data from highly enriched tumors because copy-number changes are not extensive enough in the tumor to be visible as separate peaks in the histogram. This is likely to be true particularly among tumor types with predominantly diploid genomes. In the case of glioma, however, almost all tumors appear to suffer significant levels of aneuploidy. Eight samples were analyzed after tumor purity was assured by obtaining DNA after needle dissection. In all eight, separate peaks indicative of abnormal copy numbers could be observed. However, in four of these cases where DNA was obtained without needle dissection, separate peaks were not observed. This suggests that the majority of tumor samples will have detectable copy-number alterations by histogram analysis if sufficiently pure.

Based on these analyses, measurable copy number differences were resolved in the histograms of 105 samples; all of these were included in further analysis. The clinical characteristics of these included samples are similar to the overall tumor set (**SI Table 1**), suggesting the selection process is not biased. The copy-number profiles of these 105 samples (segregated on the left in **Fig. 2a**) are similar to those that were removed (on the right), but with greater amplitude of variation. Nevertheless, we removed the samples on the right because the low amplitudes with which their copy numbers vary makes their classification (aberrant vs. not aberrant) at any locus less reliable.

Removal of duplicates

Due to inaccuracies in sample tracking, large tumor sets can contain duplicate samples from the same individual. This can bias downstream analyses of the frequencies and common boundaries of chromosomal aberrations. The use of SNP arrays for genome analysis also provides genotype data that allows for the elimination of duplicates by identifying samples with similar genotypes. Here, we score by genotype each of the m_{tot} SNPs that are assayed in each sample: A=1, AB =2, and B = 3. For every pair of tumor samples, we calculate the Euclidean distance between all m_{inf} SNPs that are informative (i.e. not "No Call") in both samples, and divide by $m_{inf} \cdot \sqrt{m_{tot}}$. Pairs for which this normalized Euclidean distance is less than a threshold θ^{dup} (= 0.4 on the basis of experience with known replicates) are identified as coming from the same individual. For such duplicates, only the tumor with higher-quality data (represented by more distinct peaks in its quality control histogram, above) is retained.

Copy-number assessment

Copy-number determinations are most reliable when data from neighboring markers with the same underlying copy number are combined to reduce the effects of noise. Several methods for such noise reduction have been reported (7-10, 12). We chose to use the segmentation method Gain and Loss Analysis of DNA (GLAD) (9). The input to the segmentation algorithm are the $\log_2$ ratios r_{ij} for each marker i and sample j. We denote the segmented (and smoothed) data by c_{ij}. (N.B. we do not utilize GLAD postprocessing clustering steps, but only utilize the initial steps aimed at segmenting and smoothing the data). GLAD tends to misidentify outliers as separate segments (7). To correct this, we join any segment with fewer than N_{short}(=5) markers to the neighboring segment with the closest c_{ij}, and assign the new segment a new c_{ij} reflecting the median r_{ij} across all markers in the combined segment. This step is performed recursively until no segments with fewer than N_{short} markers are left.

SI Fig. 7f shows the segmented data for the 105 samples that passed quality control.

Copy number variation (CNV) control

To eliminate copy-number variations derived from polymorphic germline events, markers from regions with known germline copy number variations as listed in http://projects.tcag.ca/variation/ (19-22) are removed.

Identification of copy-number aberrations

In order to call regions amplified or deleted we first need to set $\log_2$ ratio thresholds: $\boldsymbol{\theta}^{\mathbf{amp}}$ and $\boldsymbol{\theta}^{\mathbf{del}}$. Markers with $c_{ij}>\theta^{amp}$ are called amplified and ones with $c_{ij}<\theta^{del}$ are called deleted. We set these thresholds based on the empirical distribution of values in normal samples. First, all samples are brought to the same baseline signal intensity by subtracting the median c_{ij} value across all markers from each sample, to generate new c_{ij} values. The thresholds are then set to be such that only $\boldsymbol{F}_{\mathbf{normal}}$(=0.5%) of markers on autosomes pass each threshold among normal samples. In the case of the glioma dataset, this yielded θ^{amp}=0.10 and θ^{del} = -0.10.

To test the reliability of these calls, in the case of 13 tumors we obtained DNA from separate aliquots of the same tumor, where both aliquots produced data with measurable copy number differences on histogram analysis (see above). A median of 90.2% of copy-number calls were identical between the separate aliquots of each tumor. This is a conservative estimate of the reliability of these calls, as some of the difference between aliquots reflects real differences due to tumor heterogeneity.

SI Fig. 7g displays the $\log_2$ signal intensities in the regions of the genome for which $c_{ij}>\theta^{amp}$.

LOH assessment

When using paired tumor and normal genotyping data, the LOH status at all loci (including non-informative loci) is inferred by applying an HMM that takes into account the LOH calls at informative loci (6). When using genotyping data from unpaired tumors (as is our case), the LOH status at all loci is inferred on the basis of extent of regional homozygosity, taking into account the haplotype structure of the genome (6).

SI Fig. 7h displays the LOH identified among the 105 tumors with high tumor purity.

At this point, each sample has been assessed for amplifications, deletions, and LOH, and in Stage 2 we will distinguish between likely driver and passenger aberrations.

Stage 2: Aggregation of data from different tumors to differentiate between driver and passenger aberrations

To determine which of the aberrations identified in Stage 1 are likely to represent driver events, we aggregate the data from all tumors used in the analysis to generate summary scores for amplifications, deletions, and LOH. The statistical significance of each score is determined by comparison to the distribution of scores obtained by all permutations of the data (using a semi-exact approximation), with correction for multiple hypothesis testing.

Scoring the copy-number genome

We assume there are two sources of evidence that a copy-number aberration is not a chance event: *f*, the frequency of the aberration across the sample set and $\overline{c}$, its average amplitude. Therefore, the scores we generate for amplification and deletion events reflect both sources of evidence:

$$G_i^{\text{amp}} = f_i^{\text{amp}} \times \overline{c_i^{\text{amp}}}, \text{ and}$$

$$G_i^{\text{del}} = -f_i^{\text{del}} \times \overline{c_i^{\text{del}}}. \qquad (1)$$

We wanted the score to represent the negative log of the likelihood of observing the contributing aberrations by chance alone. We found that $\log_2$ ratios approximate these negative log likelihoods, both for amplifications and deletions. as estimated by the overall frequency of aberrations, as a function of amplitude, across our glioma dataset (data not shown).

Note that the scores in equation (1) can also be represented as a sum across the *n* samples in the set:

$$G_i^{\text{amp}} = f_i^{\text{amp}} \times \overline{c_i^{\text{amp}}} = \frac{1}{n} \sum_{j|c_{ij}>\theta^{\text{amp}}} c_{ij}, \text{ and}$$

$$G_i^{\text{del}} = -f_i^{\text{del}} \times \overline{c_i^{\text{del}}} = -\frac{1}{n} \sum_{j|c_{ij}>\theta^{\text{del}}} c_{ij}. \qquad (2)$$

Using the sum representations we can calculate the G_i^{amp} scores, for example, by first replacing $c_{ij} \leq \theta^{\text{amp}}$ with 0 and then summing over *j*.

As LOH is not associated with an amplitude, the LOH score represents only the frequency of LOH across the sample set which can also be rewritten as a sum:

$$G_i^{\mathrm{LOH}} = f_i^{\mathrm{LOH}} = \frac{1}{n} \sum_{j|\mathrm{LOH}_{ij}=1} 1 . \qquad (3)$$

SI Fig. 7i displays the *G* scores associated with amplifications in our tumor set, along with the frequency and average amplitude components of these scores and the per-tumor amplification data (after replacing values $\leq\theta^{\mathrm{amp}}$ with 0) that gave rise to them.

Null hypothesis generation: an analytic derivation of the null distribution

To assess which of these peaks are statistically significant, we identify those *G* scores which rise above the null distribution of values one would expect to obtain from random passenger aberrations alone. Since passenger aberrations could occur anywhere in the genome, one may model this null distribution by recalculating the *G* scores across all combinations of permutations of the marker labels within each sample. Note that by assuming, in these permutations, that all observed aberrations (including driver aberrations) are passengers, we generate a conservative, high estimate of the background aberration rate.

Although one can simulate the null distribution by performing each of these permutations in turn, we in fact derive a semi-exact estimate of this null distribution. For amplifications and deletions separately, we replace the $\log_2$ ratios in each marker not called aberrant with zero:

$$\tilde{c}_{ij}^{\mathrm{amp}} = c_{ij} \times I\left(c_{ij} > \theta^{\mathrm{amp}}\right) \quad \text{and}$$

$$\tilde{c}_{ij}^{\mathrm{del}} = c_{ij} \times I\left(c_{ij} < \theta^{\mathrm{del}}\right) \qquad (4)$$

As noted above, the *G* scores for each marker can be calculated by summing the corresponding $\tilde{c}$ across all samples. Under the null hypothesis, the arrangement of $\tilde{c}$ values is independent between samples and therefore the distribution of the sum of $\tilde{c}_{ij}$ across the samples is the same for all markers and equals the convolution of the distributions of $\tilde{c}$ values in each sample. We approximate these distributions by

generating histograms for each sample: $h_j^{\text{amp}}\left(\tilde{c}^{\text{amp}}\right)$ and $h_j^{\text{del}}\left(\tilde{c}^{\text{del}}\right)$ using a bin size of $\boldsymbol{C_{\text{bin}}}$ (=0.001). Note that as C_{bin} approaches zero the approximation becomes exact. For LOH the histograms have two values: the fraction of markers that do not have LOH and the fraction that do. The final distribution for G^{amp} is given by $h_1^{\text{amp}} \otimes h_2^{\text{amp}} \cdots \otimes h_n^{\text{amp}}$ and similarly for G^{del} and G^{LOH}.

Significance testing

We next assign statistical significance to the observed *G* scores using the null distribution calculated in the previous step. The p-value for an observed *G* score is simply the sum of the tail of the null distribution from the observed score and above. Next, in order to correct for multiple hypothesis testing we apply the Benjamini-Hochberg FDR procedure (13) to obtain q-values. These corrected probabilities are an upper bound for the expected fraction of false positives. Note that these q values are conservative since we treat all markers as independent hypotheses when in fact close markers are highly positively correlated.

Regions with q values of less than 0.25 are marked as significantly aberrant (**Fig. 2**).

SI Fig. 7j displays the q-values associated with amplifications in our tumor set, along with the *G* scores they are associated with and the per-tumor amplification profiles on which these scores are based.

Stage 3: Identification of peak regions most likely to be contain the oncogene and tumor suppressor gene targets

In this stage, we consider for each significantly aberrant region which are the most likely oncogene and TSG targets. We consider the possibility that the region may encompass two or more independently aberrant genes. We also consider the possibility that a random "passenger" mutation occurring in a single sample near, but not overlapping, the oncogene or TSG genes will distract us from those genes.

Identification of minimal targeted loci

If the driver aberrations within a region are selected due to their effects on a single gene, we would expect that gene to lie in the region where the largest number of tumors are aberrant to the highest degree. This locus equates with the locus with the minimal q value (and maximal *G* score). Therefore, within each region found to have a q value less than 0.25, we identify the peak region with minimal q value as the primary target. This peak region might contain many markers, as long as they have exactly the same q value. Usually these are neighboring markers that lie on the same copy-number or LOH segment in every sample in the dataset.

Identification of independent peak regions—"Peel-off" algorithm

It is possible that two or more peaks within a significant region are independently aberrant, but due to overlap between some aberrations associated with each peak, the entire region appears statistically significant. To recapture all of these independent peak regions, we implement an iterative "peel-off" algorithm (**SI Fig. 7k**). Here, for each chromosome that has a region with a q value less than 0.25, we remove from the data all aberrations overlapping the region with minimal q-value on the chromosome (the primary peak). We then recalculate *G* scores and q values taking a conservative approach, where we calculate p and q values based on the original null distribution including all aberrations. We remove aberrations by setting all consecutive markers that exceed the θ threshold to zero. If any part of the chromosome continues to have a q value less than 0.25, we reiterate the procedure by identifying the region with the minimal q value as a separate peak and "peel-off" aberrations it overlaps. These iterations continue until no q values less than 0.25 are obtained in the chromosome. Note that this method greedily assigns an aberration that overlaps two or more peaks to the most significant locus.

In the glioma dataset, the "peel-off" algorithm identified 2 peak regions (corresponding to *EGFR* and *MET*) independent amplified within chromosome 7, although all of chromosome 7 constitutes a single region of significant amplification (**Table 1; SI Fig. 7l**).

Determination of boundaries for each peak region

For each independent peak, the boundaries of the region of minimal q value encompass the region with the greatest evidence for containing the oncogenes or TSGs, as that region is most aberrant in the largest

number of samples. These particular boundaries, however, may be shifted from the oncogenes or TSGs due to the presence of a nearby random passenger mutation or by errors in the boundaries determined by the segmentation analysis in a single sample. Therefore, to ensure robustness of the boundaries that we identify, GISTIC recalculates the boundaries of each peak region after leaving out each sample in turn, and takes the maximum upper and minimum lower boundary of the peak of the score among all iterations. Note that this procedure uses only the data which corresponds to the "peeled-off" segments that are associated with the analyzed peak. All genes that lie wholly or partially within these boundaries are considered candidate oncogenes or TSGs. If no gene is within these boundaries, the nearest gene is considered the likeliest candidate.

Broad vs. focal aberrations

Examination of the glioma genome (**Fig. 2**) reveals broad regions undergoing significant amplification or deletion in addition to focal events. The finding that some significant focal events lie within significant broad regions, whereas others do not, suggests the possibility that overlapping broad and focal aberrations may target different genes (see **Main Text**). Therefore, for each peak region we determine whether it is subject to significant broad or focal aberrations or both.

Any region that is statistically significant over more than half a chromosome arm harbors significant broad aberrations. Also, for each peak, the G score required to attain significance (G_{sig}) is subtracted from the maximal G score, and the width of the region attaining this score is assessed. If this region does not cover more than half a chromosome arm the peak harbors significant focal aberrations. For peaks that rise to G scores less than twice G_{sig}, the width of the region at half the maximal G score is used to determine whether the peak is due primarily to broad or focal events. The result in glioma is the identification of 16 significant broad events and 16 significant focal events (**Table 1**).

Here, we use a cutoff of half a chromosome arm to define broad aberrations due to the finding that most copy-number aberrations in the glioma dataset were either substantially larger than this (generally the size of a chromosome arm or greater), or substantially smaller.

Stage 4: Classification of tumors on the basis of their driver aberrations

To study the effects of driver aberrations, tumors must be classified according to whether they have them. For each tumor we determine whether it is aberrant at each peak region and, in the case of copy-number aberrations, whether it has a high- or low-level copy-number change. In the case of statistically significant broad regions, we classify tumors as to whether they are aberrant across most of the region.

Tumor classification per peak regions

Samples are classified according to whether they have the appropriate aberration at each peak region. For instance, for each peak region of amplification, samples that were called amplified in Stage 1 are classified as aberrant; likewise for peak regions of deletion and LOH. In cases where these peaks comprise more than one marker, any sample that was called aberrant in the majority of these markers is classified as aberrant. In most cases, these calls are identical between markers within the peak region of minimal q value, as changes in any one sample will lead to changes in the *G* score and therefore the q value.

For peak regions of copy-number change, samples are also classified as to the amplitude of that change at each locus. The signal intensity distribution at *EGFR* (**SI Fig. 4a**) suggests a qualitative difference between samples with low-level amplification ($\theta^{amp} < c_{ij} < 0.9$) and samples with high-level amplification ($c_{ij} > 0.9$, corresponding to at least 3.7 copies in a diploid cell). Therefore, we classify each tumor according to whether it has a low- or high-level amplification at each peak region of amplification, using cutoffs of θ^{amp} and $\boldsymbol{\theta^{hi\_amp}}$ (=0.9). To similarly distinguish between low-level (e.g. hemizygous) and high-level deletions, we applied cutoffs of θ^{del} and $\boldsymbol{\theta^{lo\_del}}$ (= -1.3, corresponding to less than 0.9 copies in a diploid cell).

Tumor classification per broad regions

Samples are also classified as to whether they have each of the broad aberrations identified in Stage 3, using the boundaries of the broad region as determined in Stage 3. Any sample that in Stage 1 is called with the appropriate aberration (e.g. amplified in a significantly amplified region) in more than half of the markers within this broad region is classified as having a broad aberration in the region.

Having classified every tumor as to its status at every targeted locus and broad region of aberrancy, the GISTIC algorithm is complete.

Output from GISTIC

The results of the algorithm are contained in the following files:

(a) **Display** files in .pdf, .eps and .fig formats showing the variation in *G* scores and associated q values for all markers along the genome.

(b) An **all lesions** file that describes all of the significant aberrations and peak regions, and the status of each sample at each focal and broad region.

(c) A **segmented_data** file that represents the c_{ij} values after batch correction, normalization, segmentation analysis, and removal of copy-number polymorphisms.

(d) A **gene table** which lists the genes that overlap with each of the peak regions. Genes that are listed as known targets or generally related to cancer are highlighted (if such lists are provided).

(e) A **histograms file** (.pdf) with a histogram plot for each sample and a mark indicating whether the sample has passed the histogram quality control step.

Supplementary Notes

Supplementary Note 1: LOH analysis

The *G* scores and corresponding significance levels for LOH (**SI Fig. 9**) yield a similar pattern to deletions, with 2 exceptions: (i) High-level amplifications of EGFR on chr7 are scored as LOH because they give rise to an allelic imbalance that obscures the minor allele; and (ii) chr17p (containing the TSG *TP53*) appears to primarily undergo copy-neutral LOH, with multiple samples exhibiting regional homozygosity despite retaining two copies of the chromosome (**SI Fig. 9**). Other than these cases, the similar pattern between LOH and deletions indicates that the reduction to homozygosity that represents LOH is usually due to hemizygous deletion of one allele. However, the ability to map deletions is superior to LOH, due to 2 factors: (i) LOH is obscured by low levels of contaminating normal DNA that are tolerated by deletion mapping, and (ii) the resolution of LOH analysis is poorer than for deletions. This latter factor is true when paired normal samples are used to map LOH (because most SNP markers are homozygous in the normal sample and therefore uninformative as to LOH status of the tumor) or when paired normal samples are not used (given the necessary reduction in resolution this implies) (6). For these reasons, we placed more emphasis on the results for deletions except in the primarily copy-neutral case of LOH at chr17p.

Supplementary Note 2: Minimal common region analysis of 141 gliomas

As a comparison to the GISTIC method, we performed an analysis of the minimal common regions of copy-number variation in our 100K SNP array data from 141 gliomas. Here, GLAD (9) was used to segment the raw $\log_2$ ratios generated from the signal intensity (after brightness correction (17) and model-based expression(18)) of the tumor divided by the mean signal intensity of all normal controls at each SNP locus. Segments for which the median $\log_2$ ratio across all SNPs was greater than 0.1 or less than -0.1 were called amplified or deleted, respectively. For each region found to be amplified or deleted in over 5% of samples, the minimal common regions of amplification or deletion were identified as potentially harboring oncogenes or tumor suppressor genes. This approach yielded 144 minimal common regions of amplification or deletion, harboring 5 of the known oncogenes and tumor suppressor

genes in glioma. These results are similar to prior analyses of the glioma genome in terms of the number of regions selected and sensitivity to known oncogenes and tumor suppressor genes (**Table 2**).

The GISTIC analysis of the same dataset appears to provide superior specificity (identifying only 27 peak regions for copy-number aberrations) and sensitivity (identifying 9 of the known oncogenes and tumor suppressor genes in glioma) (see **Main Text**, **Table 2**). Three factors may contribute to the high level of specificity of GISTIC: (1) When using very high-resolution datasets, even systematic errors occurring in a small fraction of markers and tumors can give rise to large numbers of artifactual aberrations across the dataset. GISTIC minimizes these in multiple preprocessing steps. (2) Without controlling for the background aberration rate, random events may be identified as interesting candidates. GISTIC uses a statistical test to eliminate these. (3) Within a region that is frequently aberrant, multiple loci often share the same, maximal frequency of aberration—leading them all to be considered minimal common regions of amplification or deletion. GISTIC prioritizes those loci with the highest average amplitude of change.

Supplementary Note 3: Comparative outlier analysis

To identify genes responsible for the functional effects of 7^{gain}, we applied a 'comparative outlier analysis' in which we identified genes on the chromosome that show extreme outliers among at least 10% of the tumors among tumors with 7^{gain} compared to 7^{normal} (**Table 3**). Specifically, for each probeset 'PRBST' matching a gene on chr7, primary GBMs were classified according to their copy-number status at the gene locus (defined as the mean of segmented values across the minimal set of SNP markers that contain the gene) as 7^{norm} (if $\theta^{del} < c_{ij} < \theta^{amp}$), 7^{gain} (if $\theta^{amp} < c_{ij} < 0.9$), or $7^{gain}PRBST^{amp}$ (if $c_{ij} > 0.9$). All expression values were normalized by subtracting the median and scaling by the median absolute deviation of 7^{norm} samples. The outlier score represents the top 10^{th} percentile of these transformed expression values among the 7^{gain} samples.

The assumption behind this analysis is that broad aberrations, because they affect large numbers of genes, may have (1) polygenic effects, and (2) heterogeneous effects across tumors (sometimes affecting one set of genes and other times affecting a different set). Therefore, we did not look for genes that are

consistently upregulated in 7^{gain}, but rather genes that are overexpressed in some samples with 7^{gain}, compared to the distribution expected from 7^{norm}.

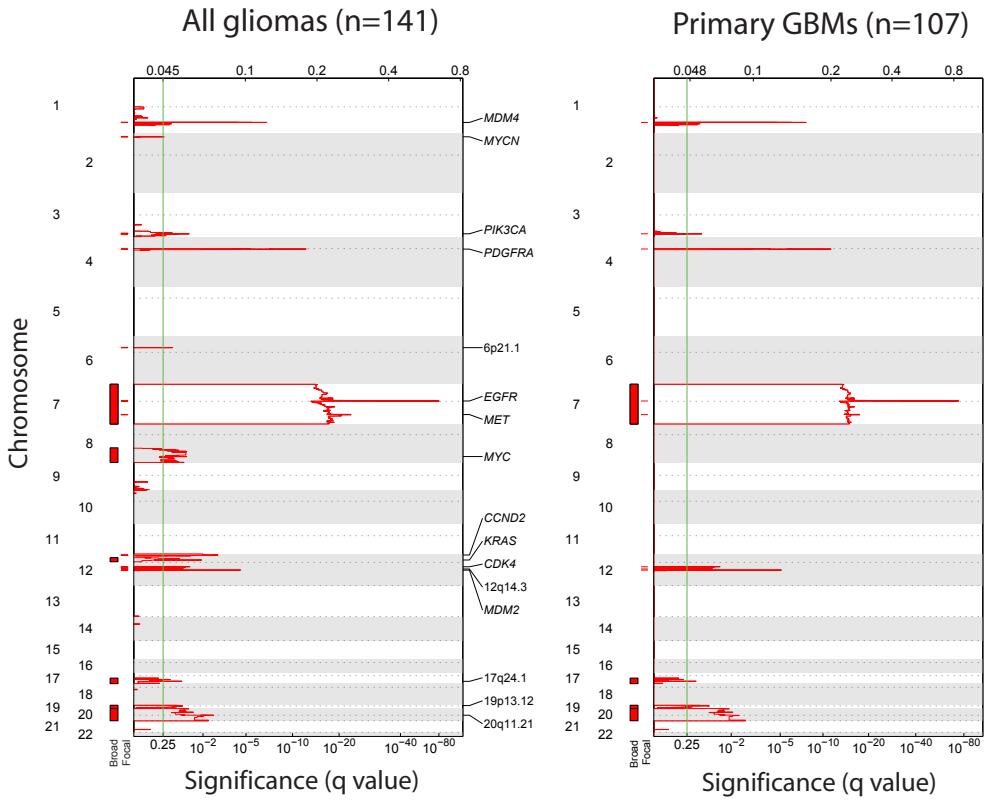
The results are striking. Although this outlier analysis was not restricted to potential oncogenes, the four top-scoring genes (out of 568 mapping to the chromosome; **Table 3**) are all likely candidates: *MET* (a known glioma oncogene) and its ligand *HGF* (see **Main Text**), *PDAP1*, an enhancer of the glioma oncogene *PDGFRA* (23), and *HOXA9*, an oncogene in acute myelogenous leukemia (24, 25). All of these candidates merit follow-up studies.

References:

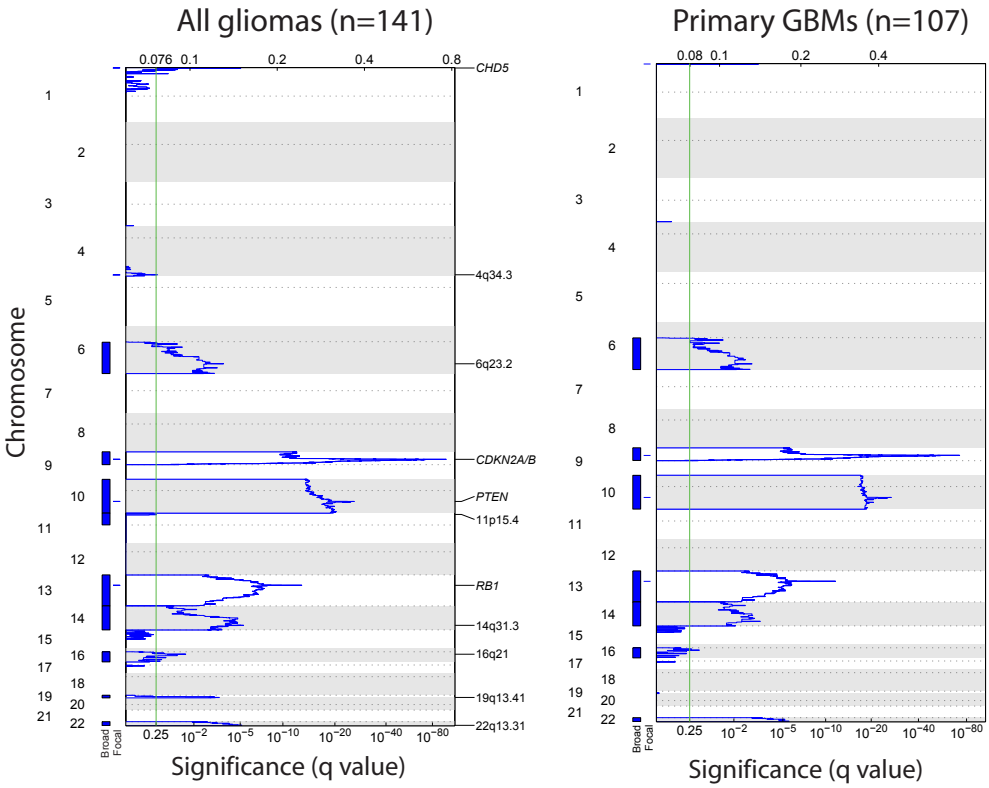
1. Kotliarov, Y., Steed, M. E., Christopher, N., Walling, J., Su, Q., Center, A., Heiss, J., Rosenblum, M., Mikkelsen, T., Zenklusen, J. C. & Fine, H. A. (2006) *Cancer Res* **66,** 9428-36.
2. Maher, E. A., Brennan, C., Wen, P. Y., Durso, L., Ligon, K. L., Richardson, A., Khatry, D., Feng, B., Sinha, R., Louis, D. N., et al (2006) *Cancer Res* **66,** 11502-13.
3. Zhao, X., Weir, B. A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J. C., Naoki, K., Richards, W. G., et al (2005) *Cancer Res* **65,** 5561-5570.
4. Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W. M., Province, M. A., Kraja, A., Johnson, L. A., et al (in press) *Nature*.
5. Tracy, S., Mukohara, T., Hansen, M., Meyerson, M., Johnson, B. E. & Janne, P. A. (2004) *Cancer Res* **64,** 7241-4.
6. Beroukhim, R., Lin, M., Park, Y., Hao, K., Zhao, X., Garraway, L. A., Fox, E. A., Hochberg, E. P., Mellinghoff, I. K., Hofer, M. D., et al (2006) *PLoS Comput Biol* **2,** e41.
7. Lai, W. R., Johnson, M. D., Kucherlapati, R. & Park, P. J. (2005) *Bioinformatics***,** bti611.
8. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. (2004) *Biostat* **5,** 557-572.
9. Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F. & Barillot, E. (2004) *Bioinformatics* **20,** 3413-3422.
10. Fridlyand, J., et al. (2004) *J. Multivariate Anal.* **90,** 132-153.
11. Zhao, X., Li, C., Paez, J. G., Chin, K., Janne, P. A., Chen, T.-H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J. W., Sellers, W. R. & Meyerson, M. (2004) *Cancer Res* **64,** 3060-3071.
12. Wang, P., Kim, Y., Pollack, J., Narasimhan, B. & Tibshirani, R. (2005) *Biostatistics* **6,** 45-58.
13. Benjamini, Y. & Hochberg, Y. (1995) *J Roy Stat Soc, Ser B* **57,** 289-300.
14. Ishikawa, S., Komura, D., Tsuji, S., Nishimura, K., Yamamoto, S., Panda, B., Huang, J., Fukayama, M., Jones, K. W. & Aburatani, H. (2005) **333,** 1309.
15. Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C. & Ogawa, S. (2005) *Cancer Res* **65,** 6071-6079.
16. Carvalho, B., Bengtsson, H., Speed, T. P. & Irizarry, R. A. (2006) *Biostatistics*.
17. Li, C. & Hung Wong, W. (2001) *Genome Biology* **2,** research0032.1 - research0032.11.
18. Li, C. & Wong, W. H. (2001) *PNAS* **98,** 31-36.
19. Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. & Lee, C. (2004) *Nat Genet* **36,** 949-51.
20. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. (2006) *Nat Genet* **38,** 75-81.
21. Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. (2006) *Nat Genet* **38,** 82.
22. McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., Dallaire, S., Gabriel, S. B., Lee, C., Daly, M. J., et al (2006) *Nat Genet* **38,** 86.
23. Fischer, W. H. & Schubert, D. (1996) *J Neurochem* **66,** 2213-6.
24. Borrow, J., Shearman, A. M., Stanton, V. P., Jr., Becher, R., Collins, T., Williams, A. J., Dube, I., Katz, F., Kwong, Y. L., Morris, C., et al (1996) *Nat Genet* **12,** 159-67.
25. Nakamura, T., Largaespada, D. A., Lee, M. P., Johnson, L. A., Ohyashiki, K., Toyama, K., Chen, S. J., Willman, C. L., Chen, I. M., Feinberg, A. P., et al (1996) *Nat Genet* **12,** 154-8.

Supplementary Figure 1

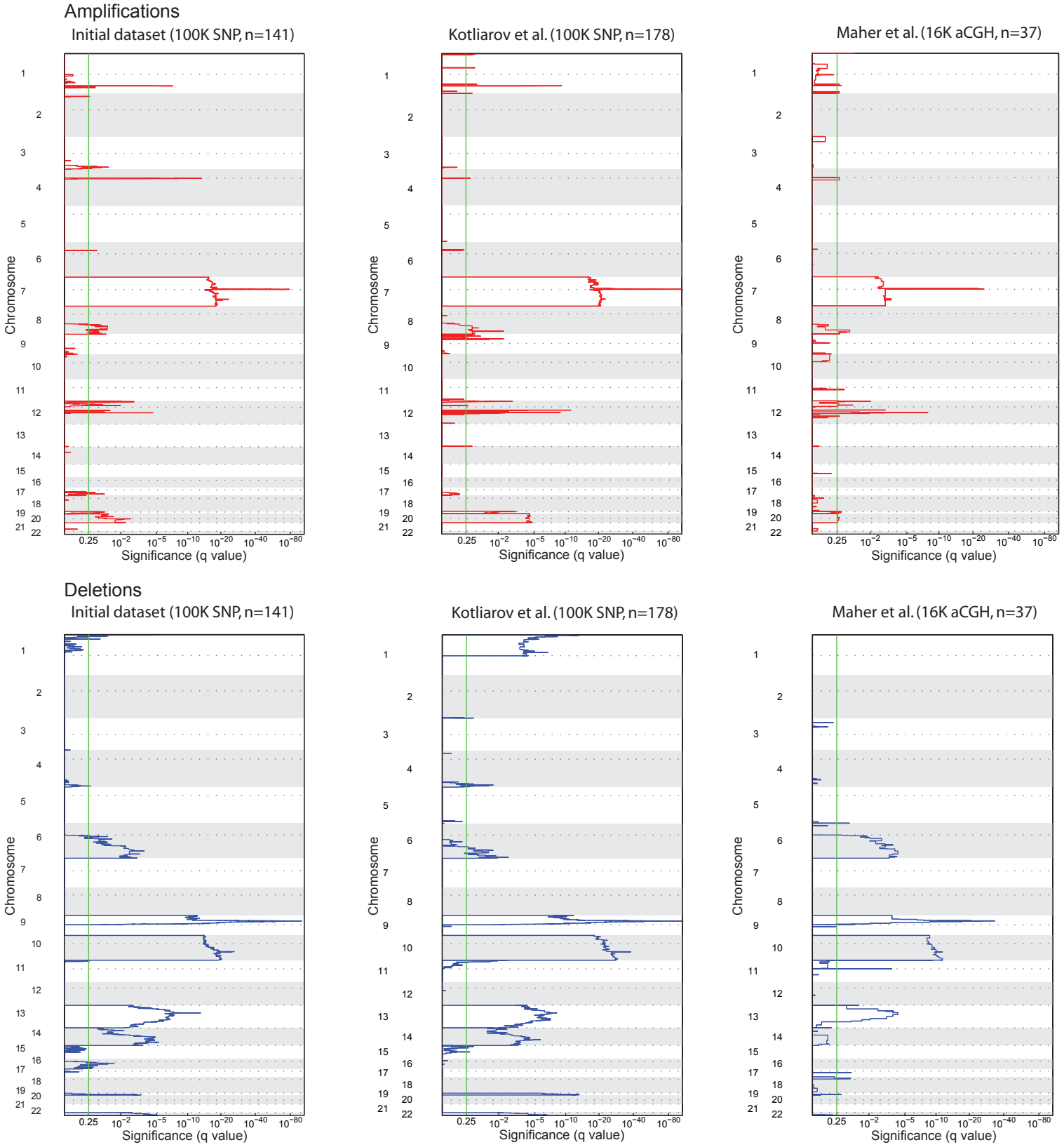
Amplifications



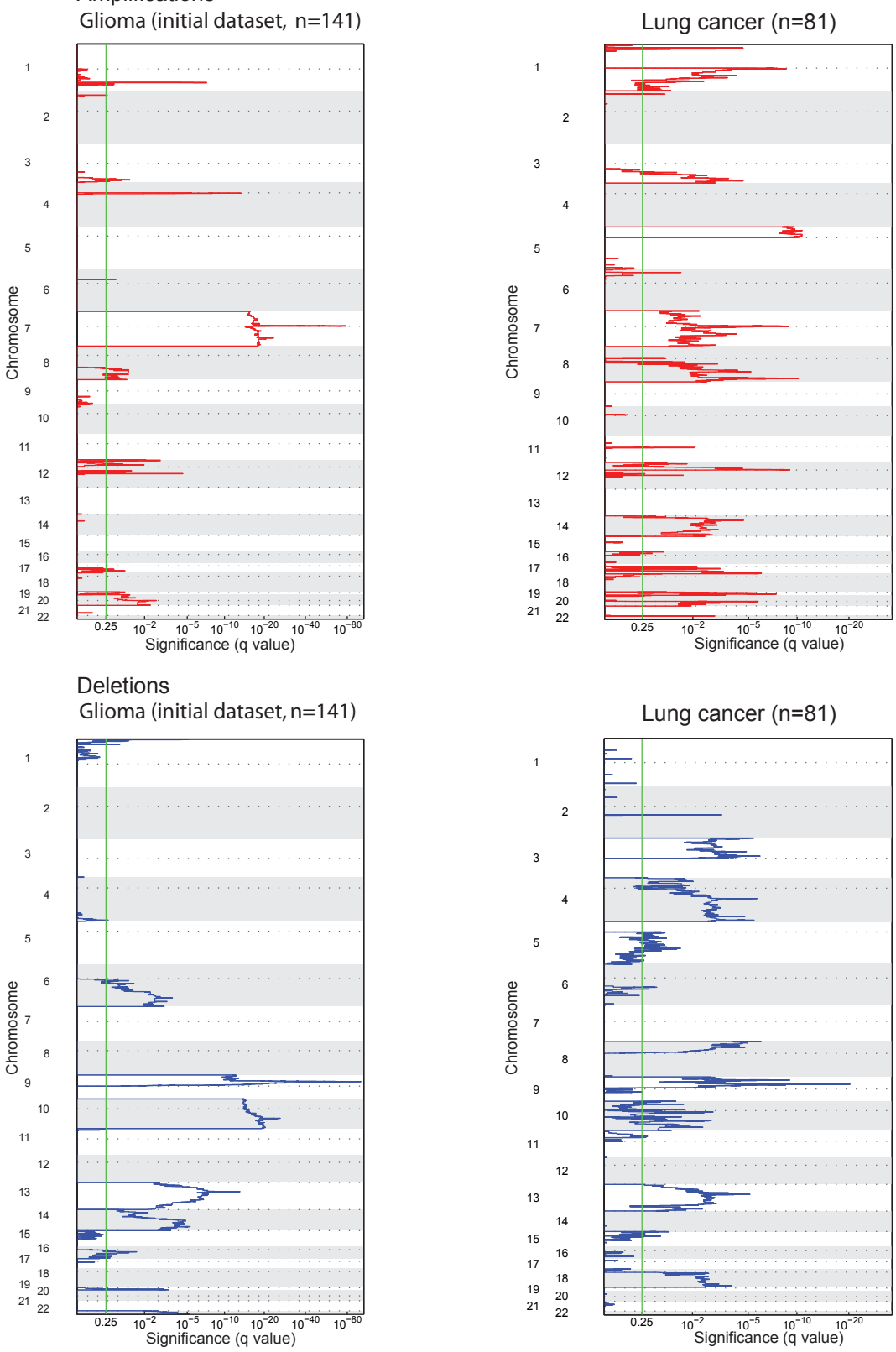
Deletions



Supplementary Figure 2

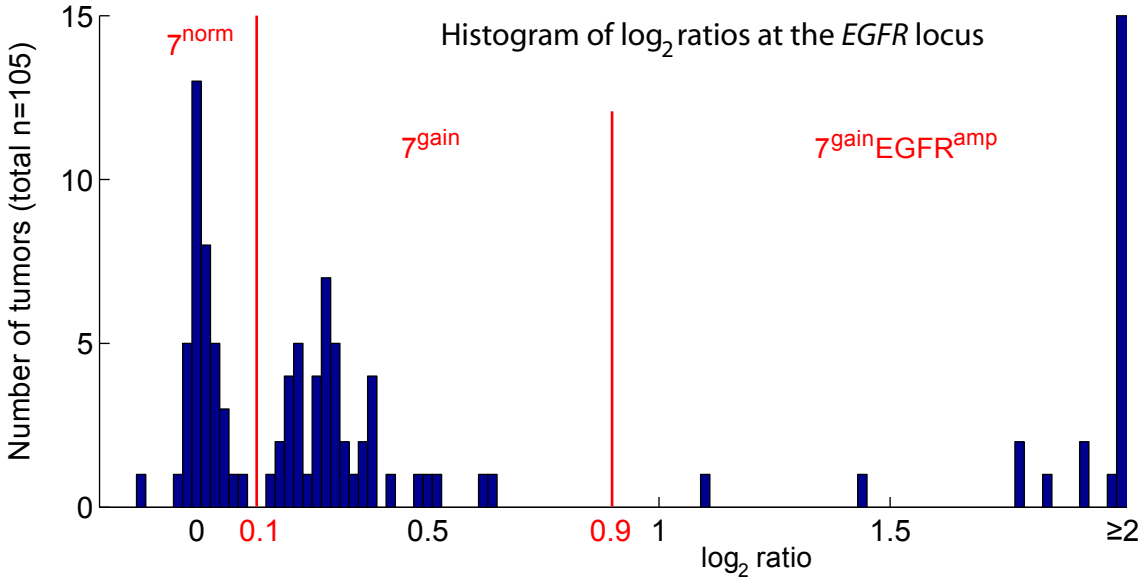


Supplementary Figure 3

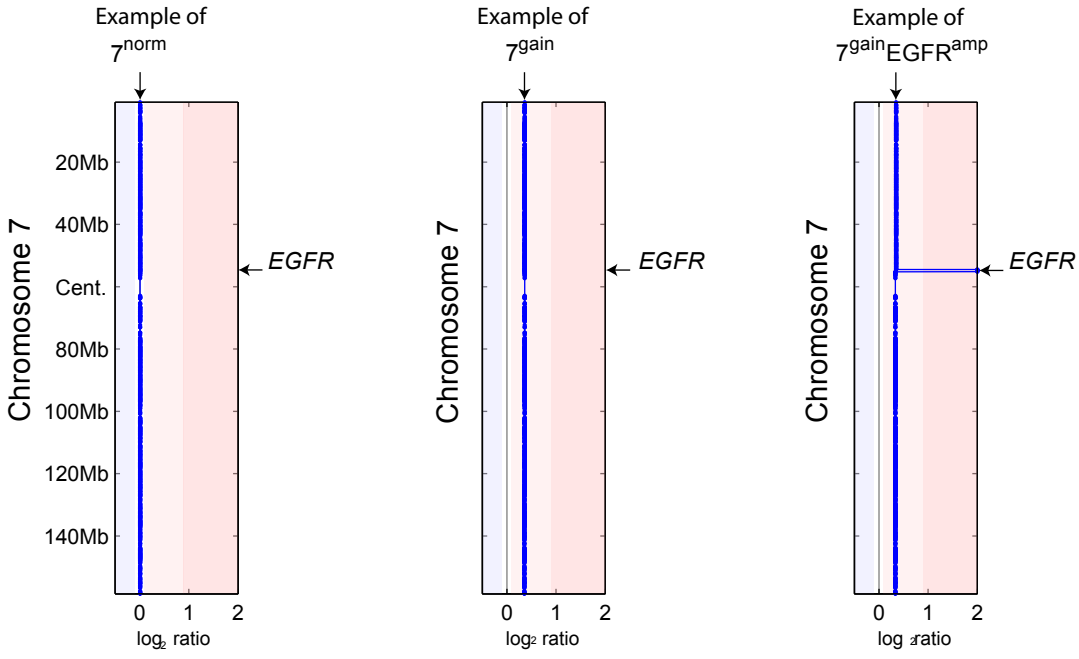


Supplementary Figure 4

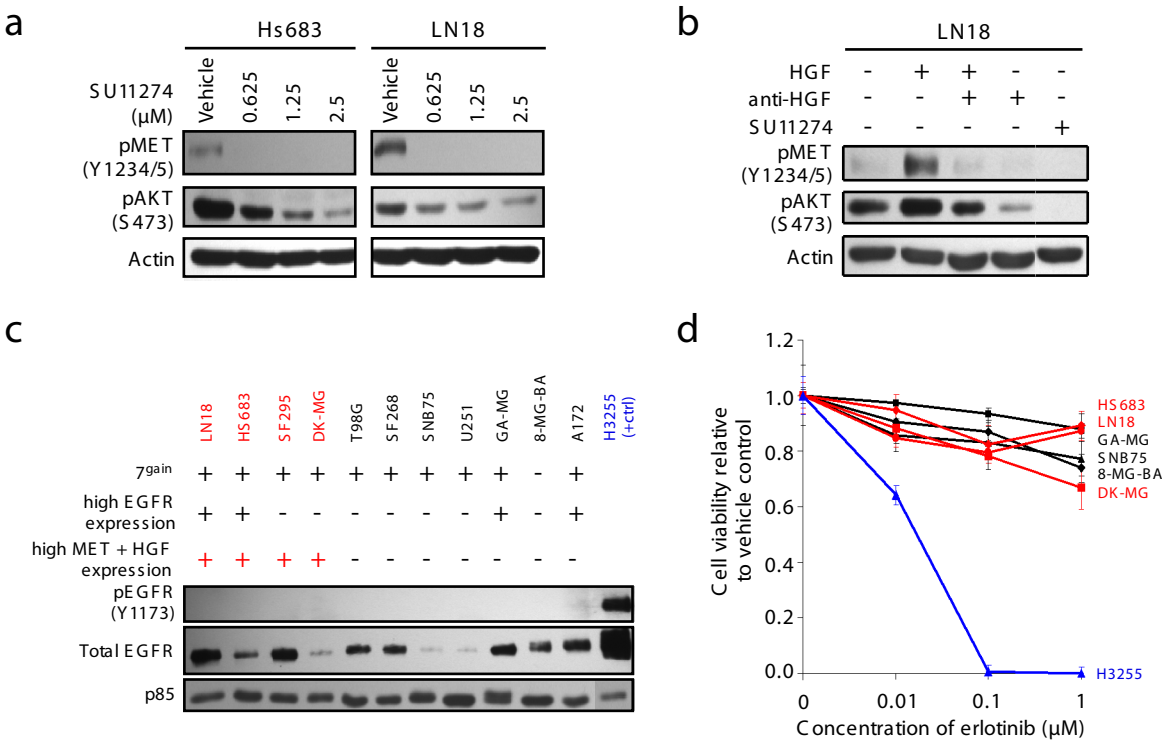
a



b

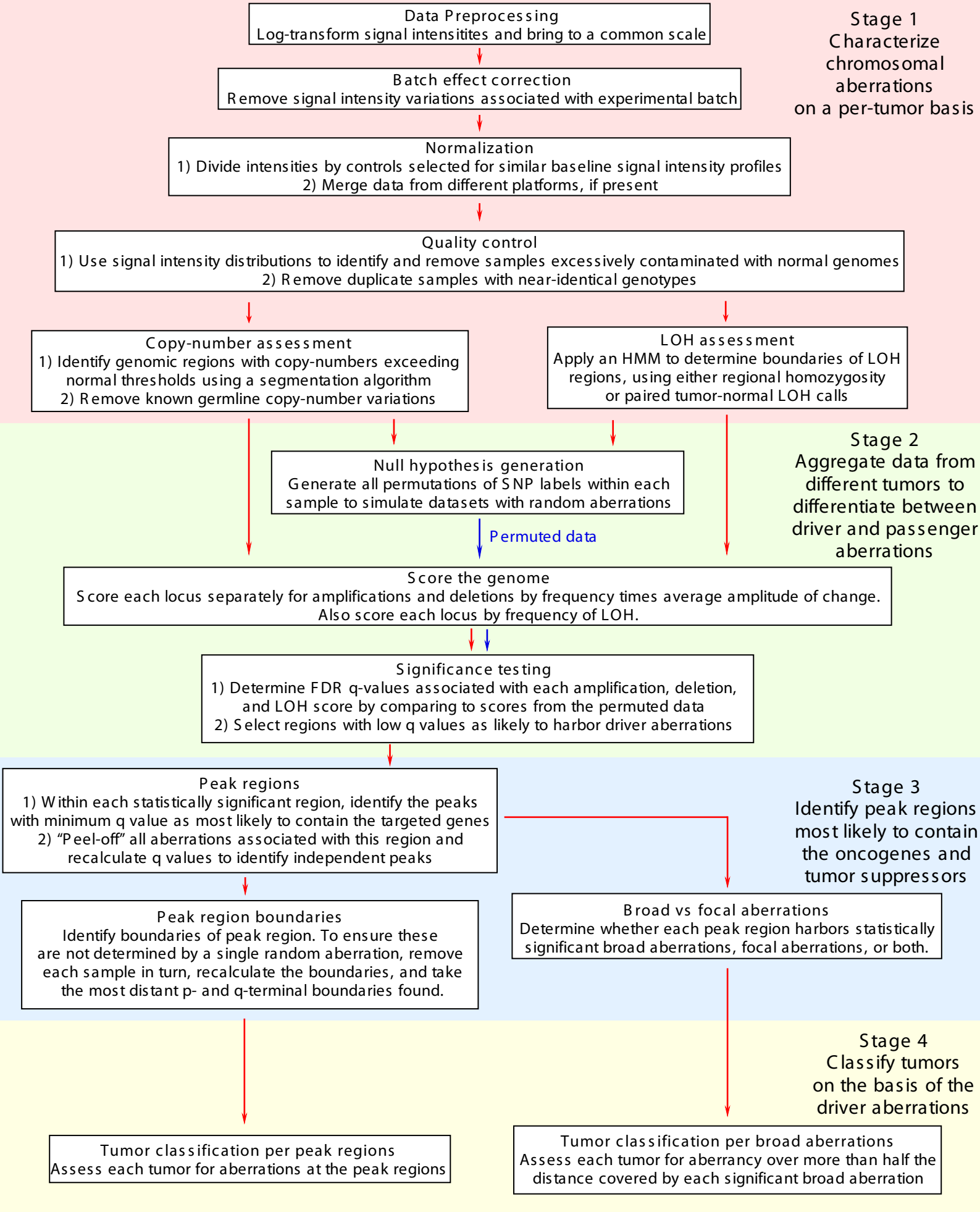


Supplementary Figure 5



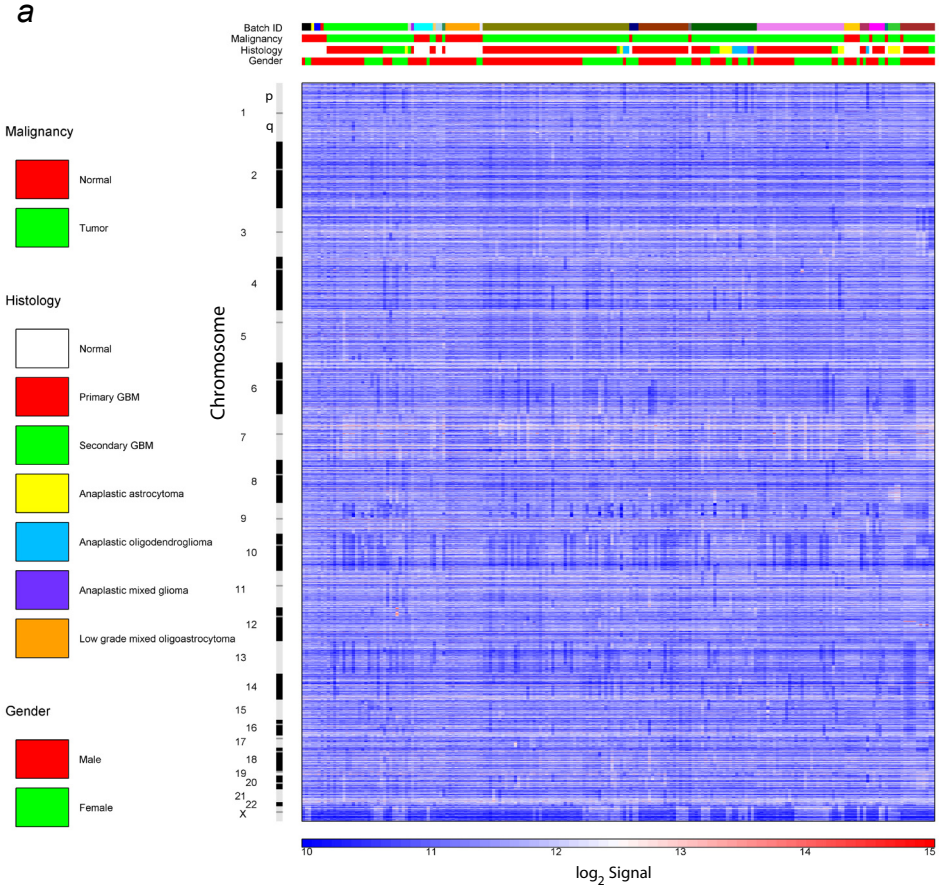
Supplementary Figure 6

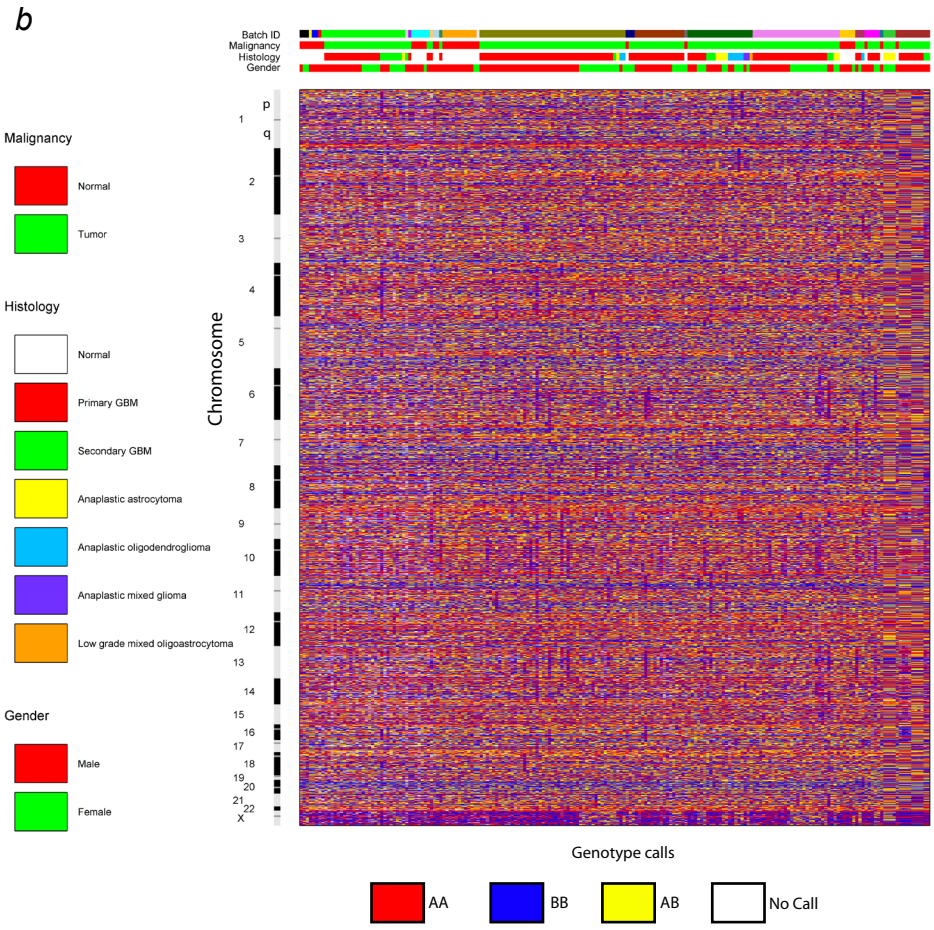
GISTIC Flow Chart



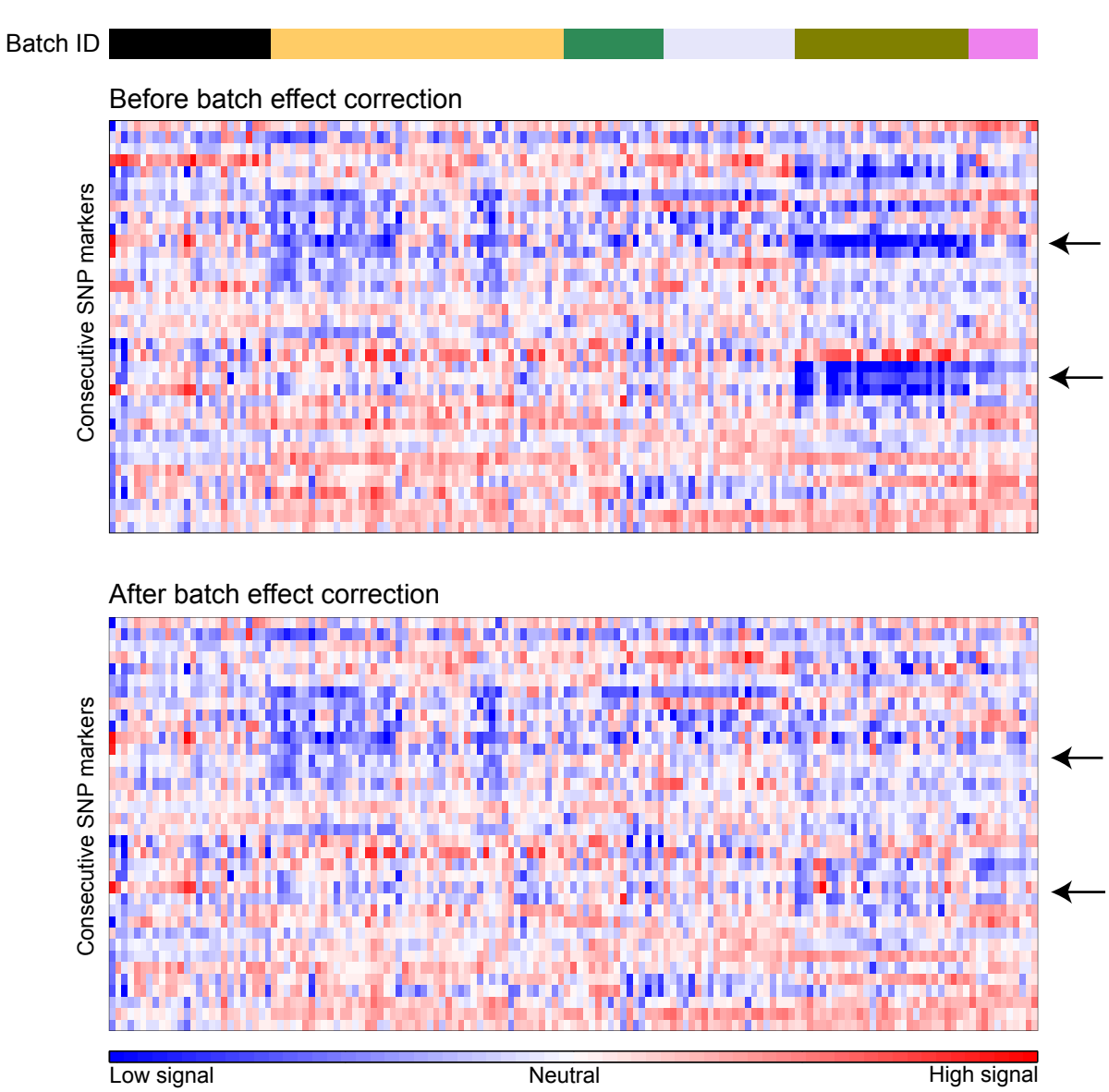
Supplementary Figure 7

a

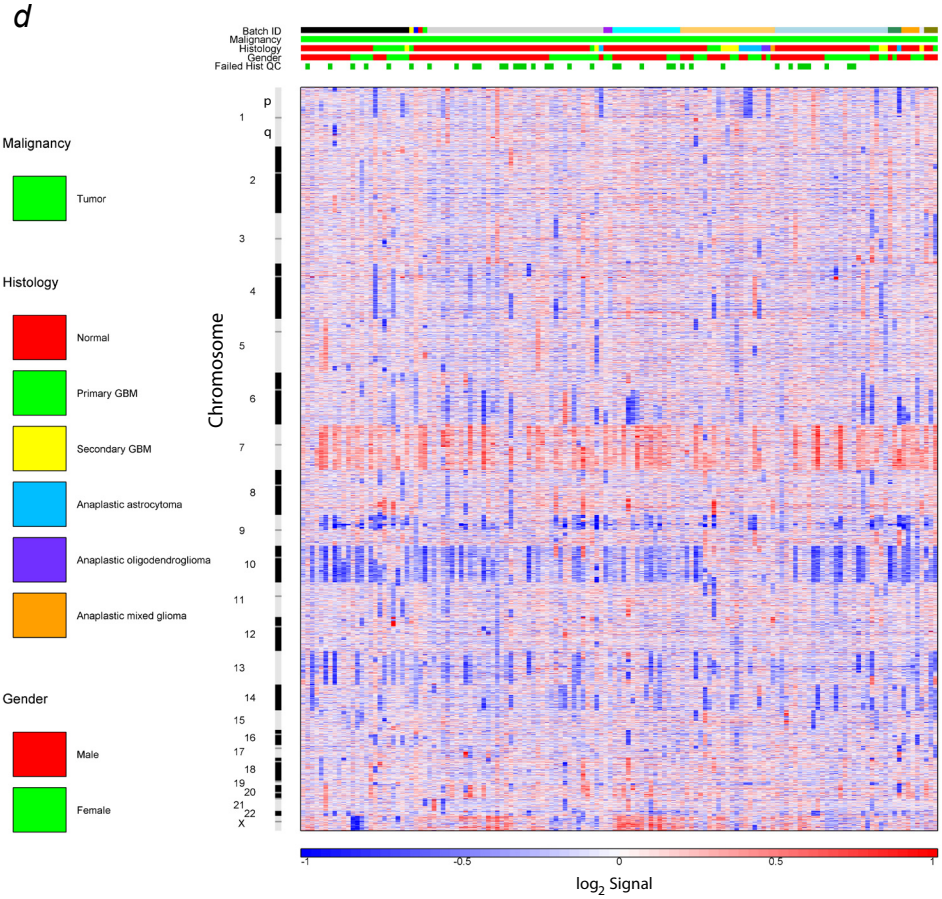



b
Batch ID
Malignancy
Histology
Gender
Malignancy
Normal
Tumor
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Low grade mixed oligoastrocytoma
Gender
Male
Female
Chromosome
p
q
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Genotype calls
AA
BB
AB
No Call

C

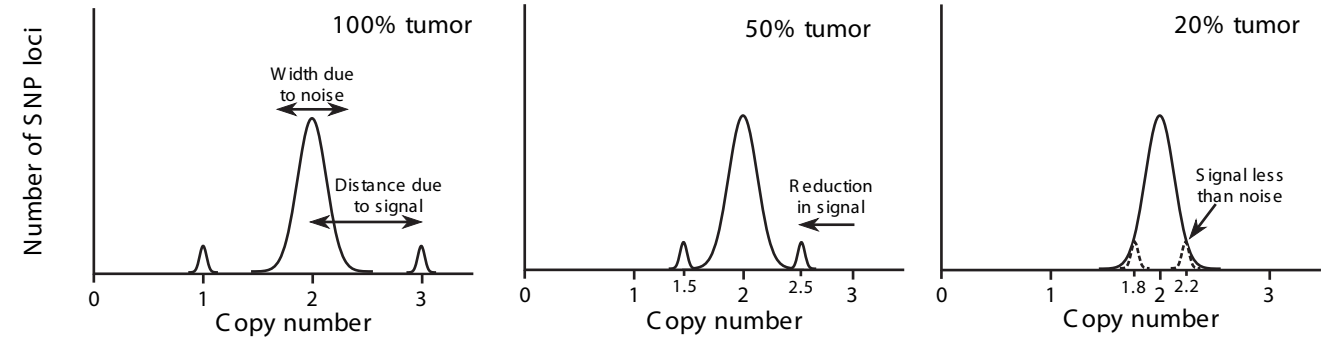

Batch ID
Before batch effect correction
Consecutive SNP markers
After batch effect correction
Consecutive SNP markers
Low signal
Neutral
High signal

d

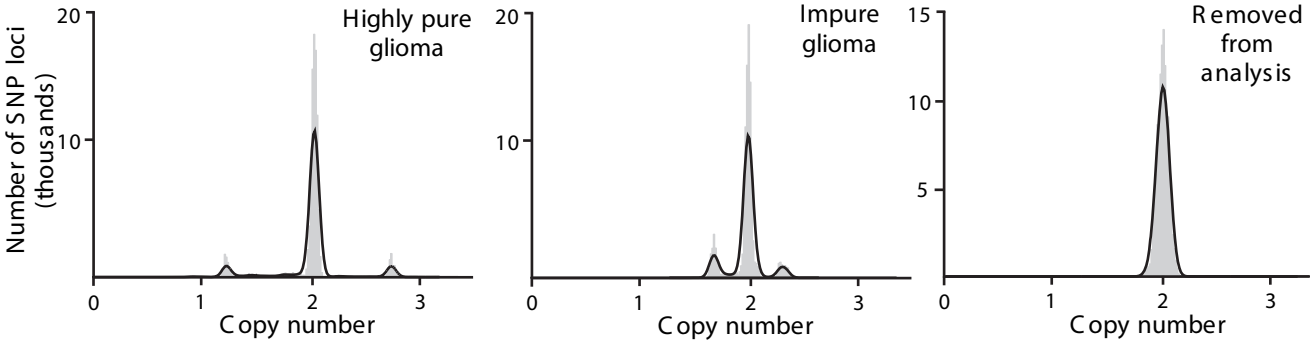

Batch ID
Malignancy
Histology
Gender
Failed Hist QC
Malignancy
Tumor
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
Chromosome
p
q
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
-1
-0.5
0
0.5
1
log2 Signal

e

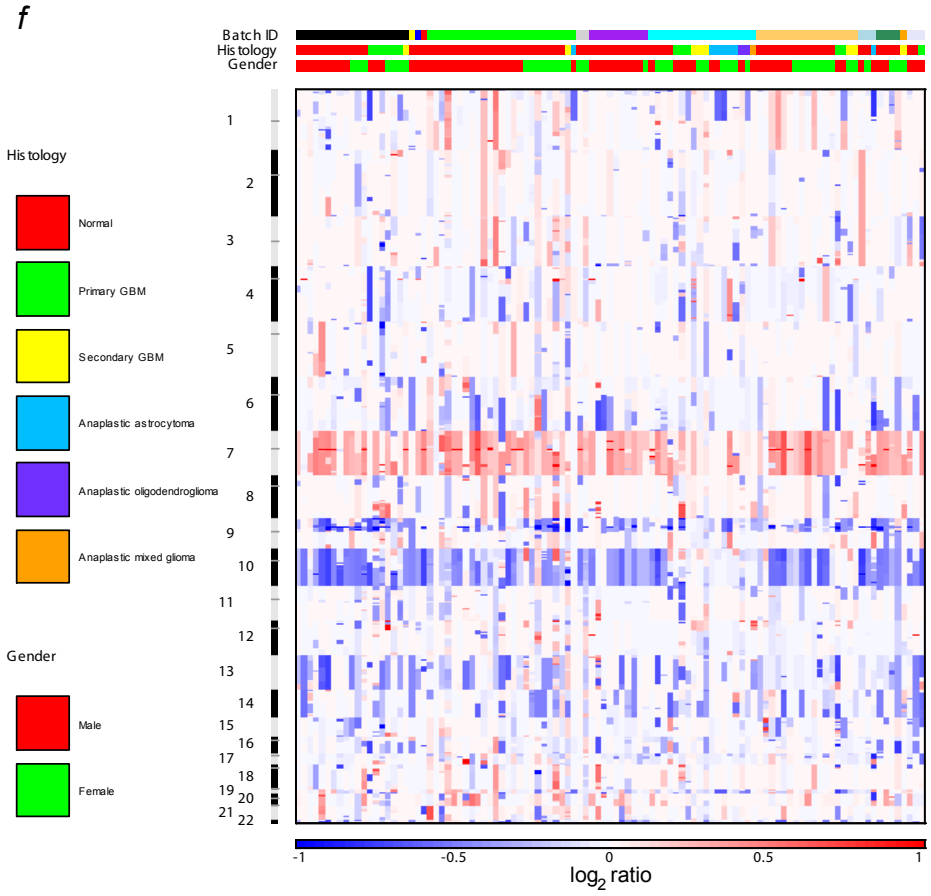
Diagrams of theory

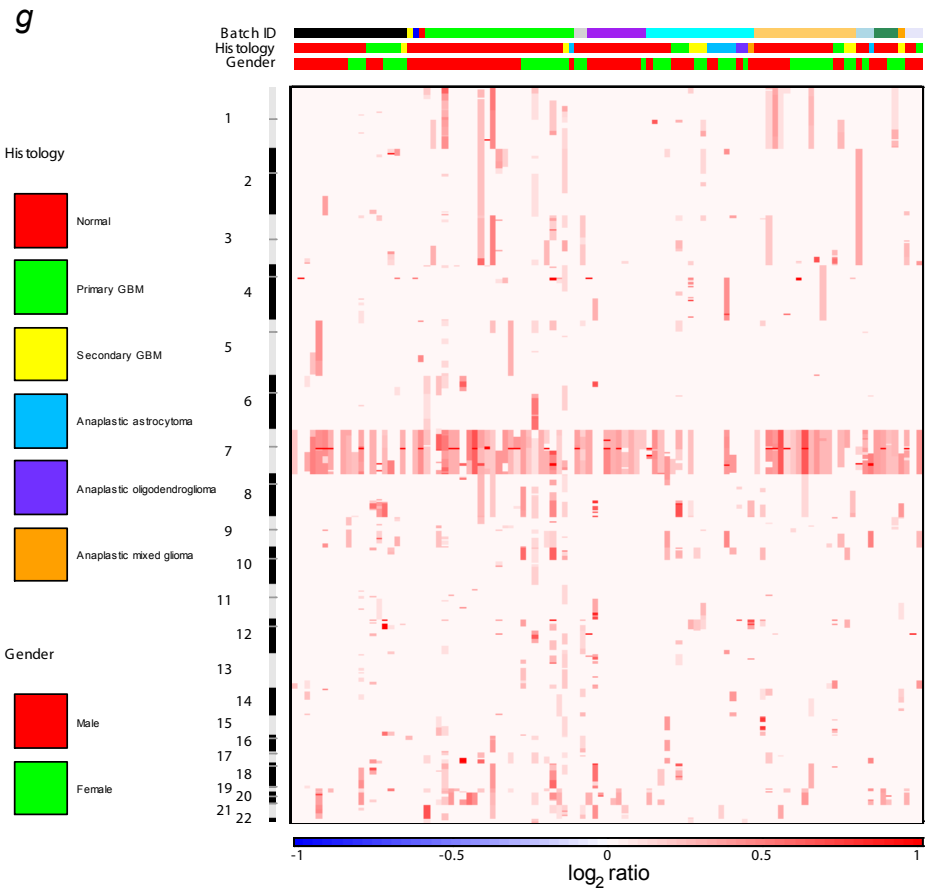


Example datasets

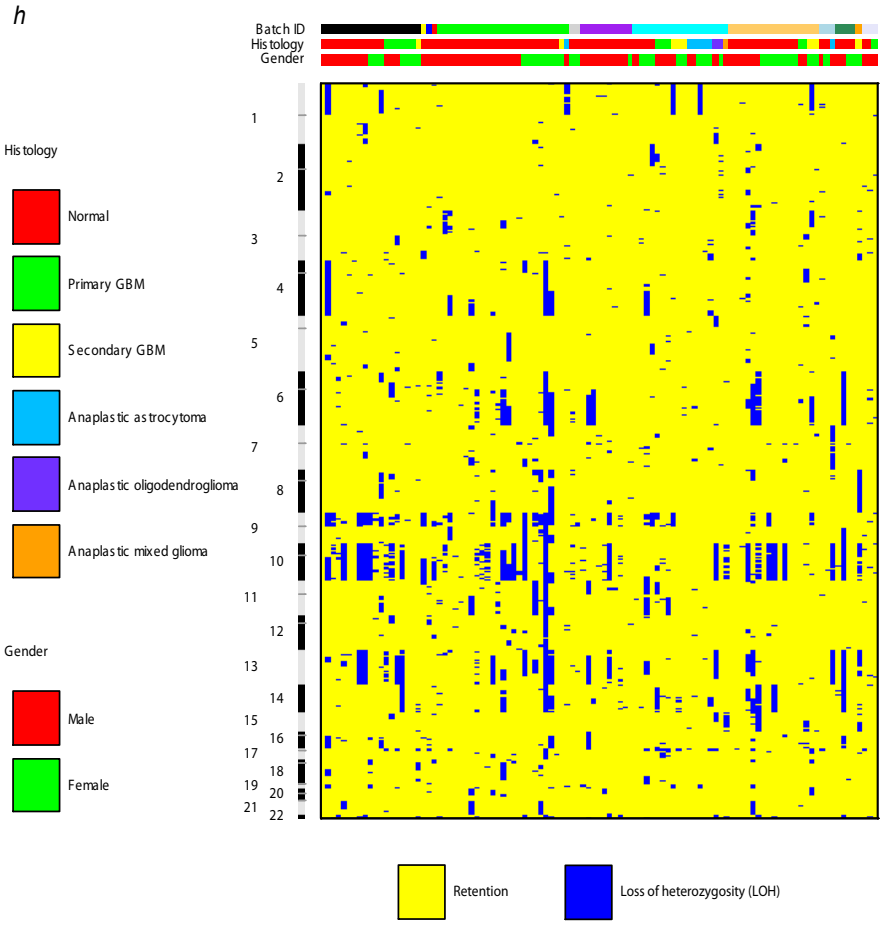


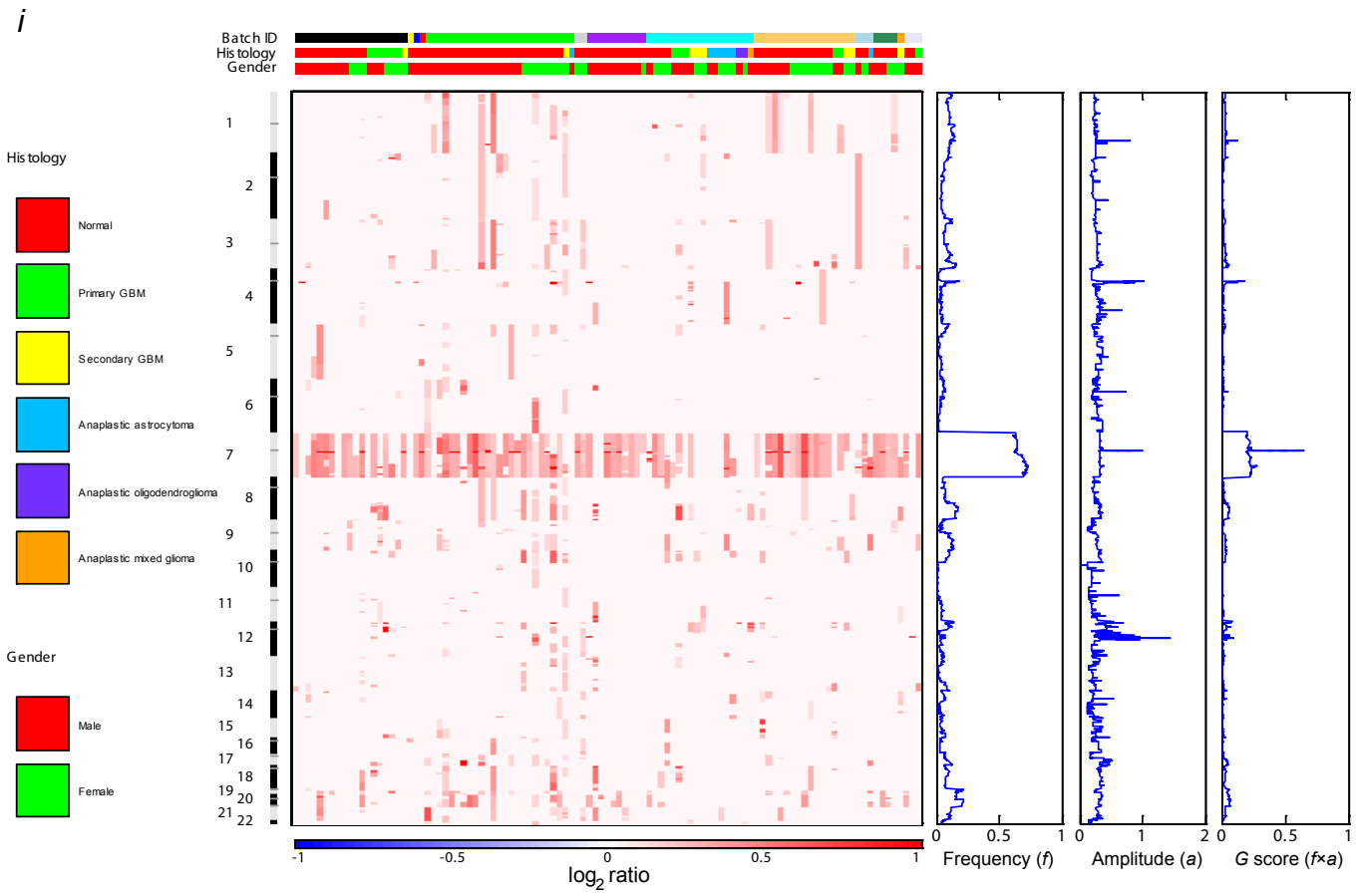
f


Batch ID
Histology
Gender
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
-1
-0.5
0
0.5
1
log2 ratio

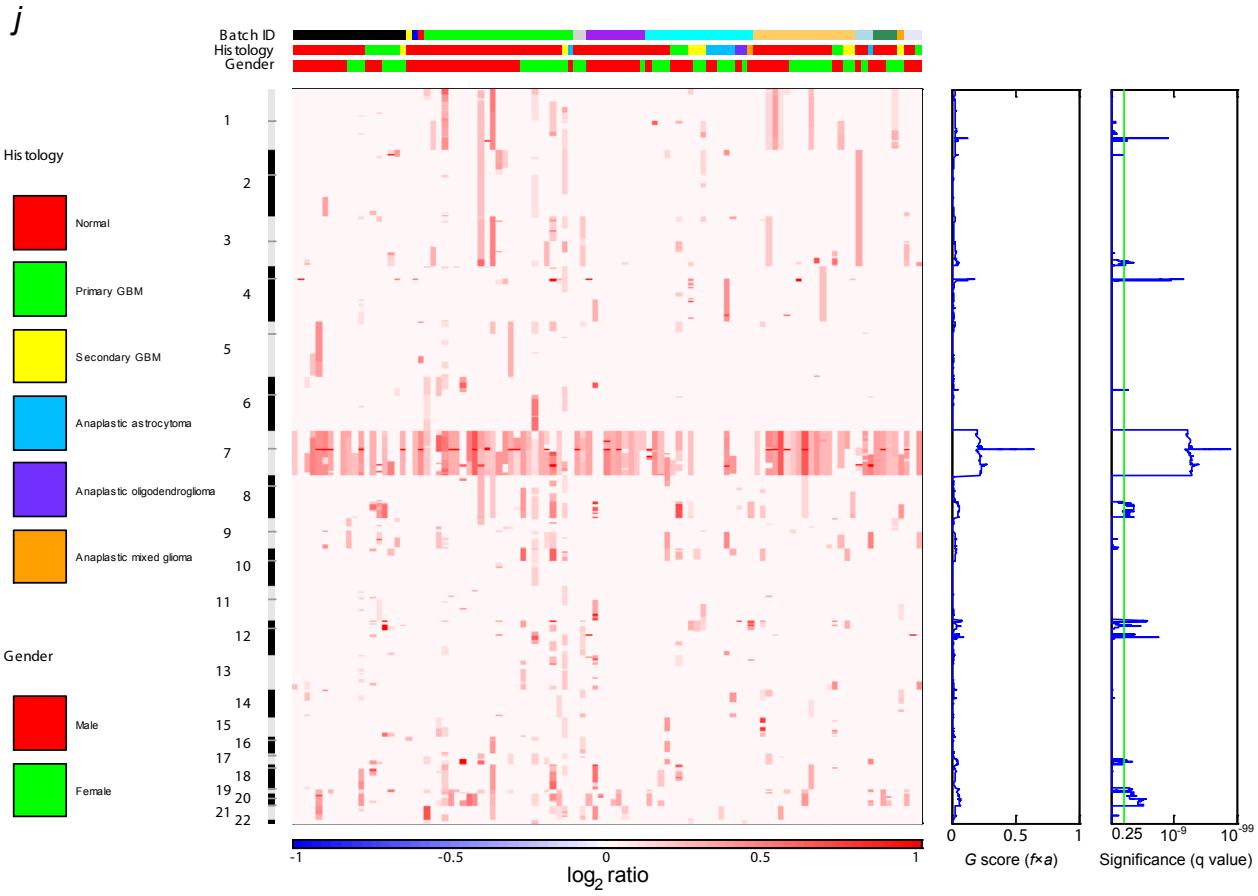

g
Batch ID
Histology
Gender
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
-1
-0.5
0
0.5
1
log2 ratio

h

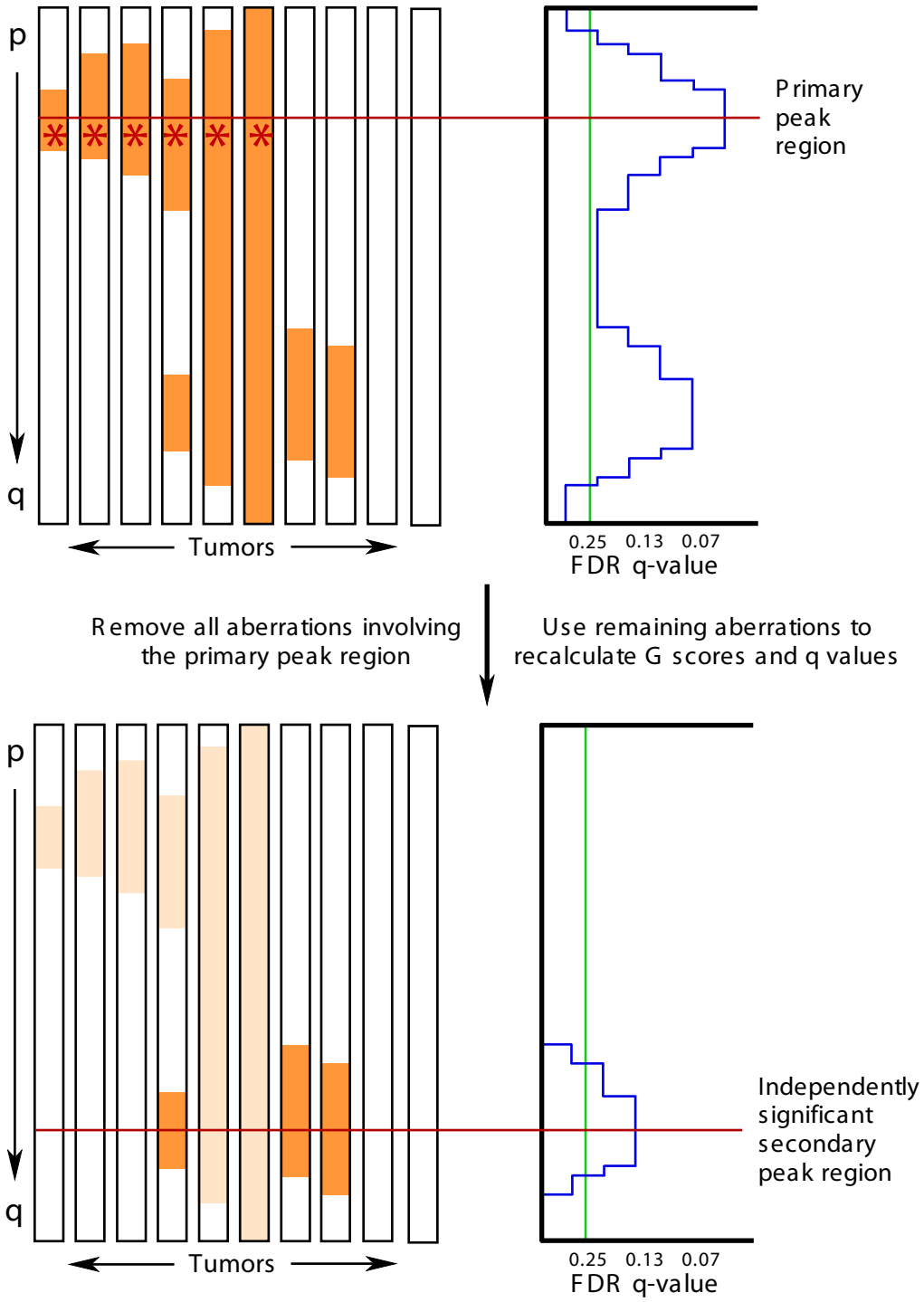

Batch ID
Histology
Gender
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Retention
Loss of heterozygosity (LOH)

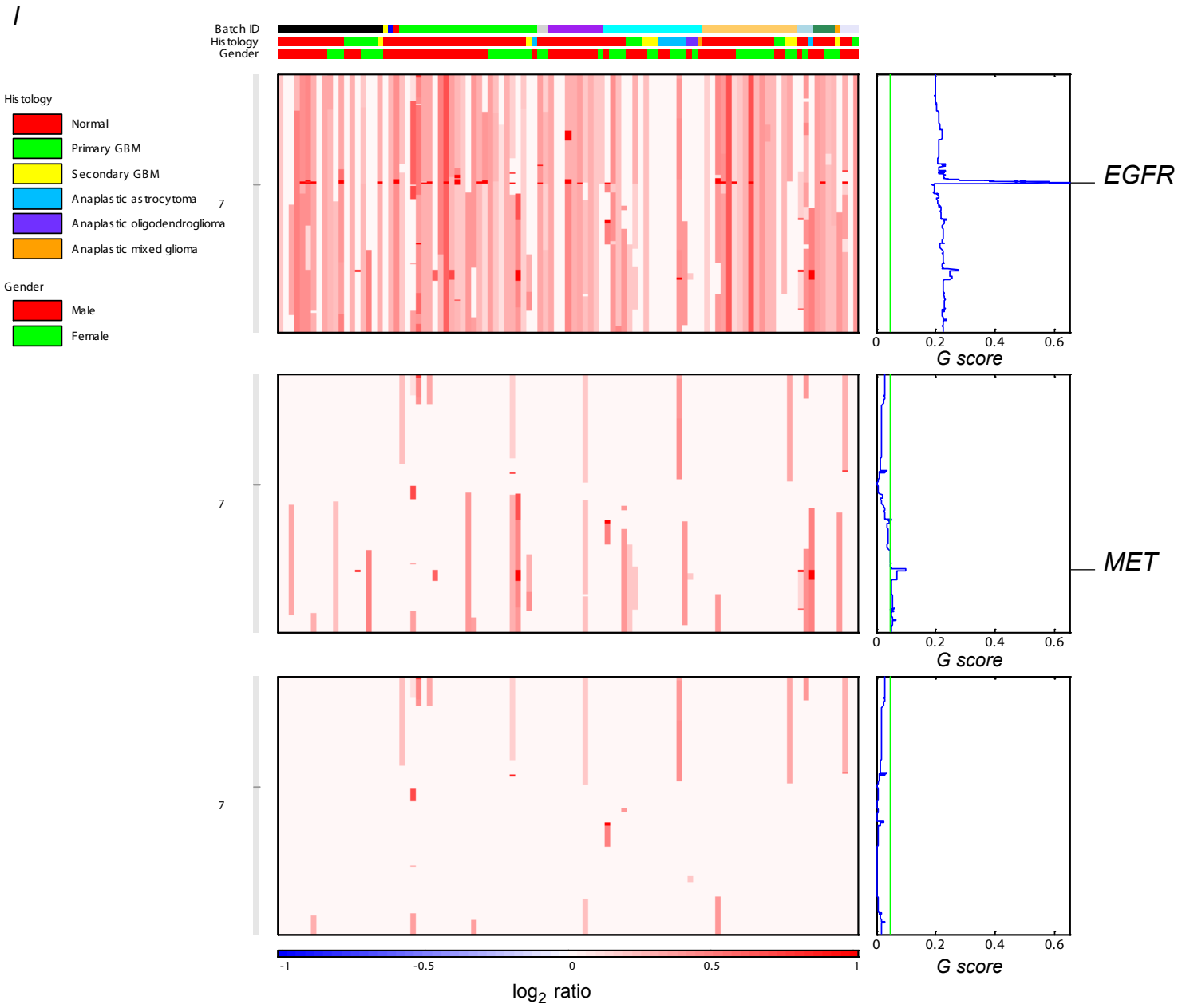

i
Batch ID
Histology
Gender
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
-1
-0.5
0
0.5
1
log2 ratio
0
0.5
1
Frequency (f)
0
1
2
Amplitude (a)
0
0.5
1
G score (f×a)

j

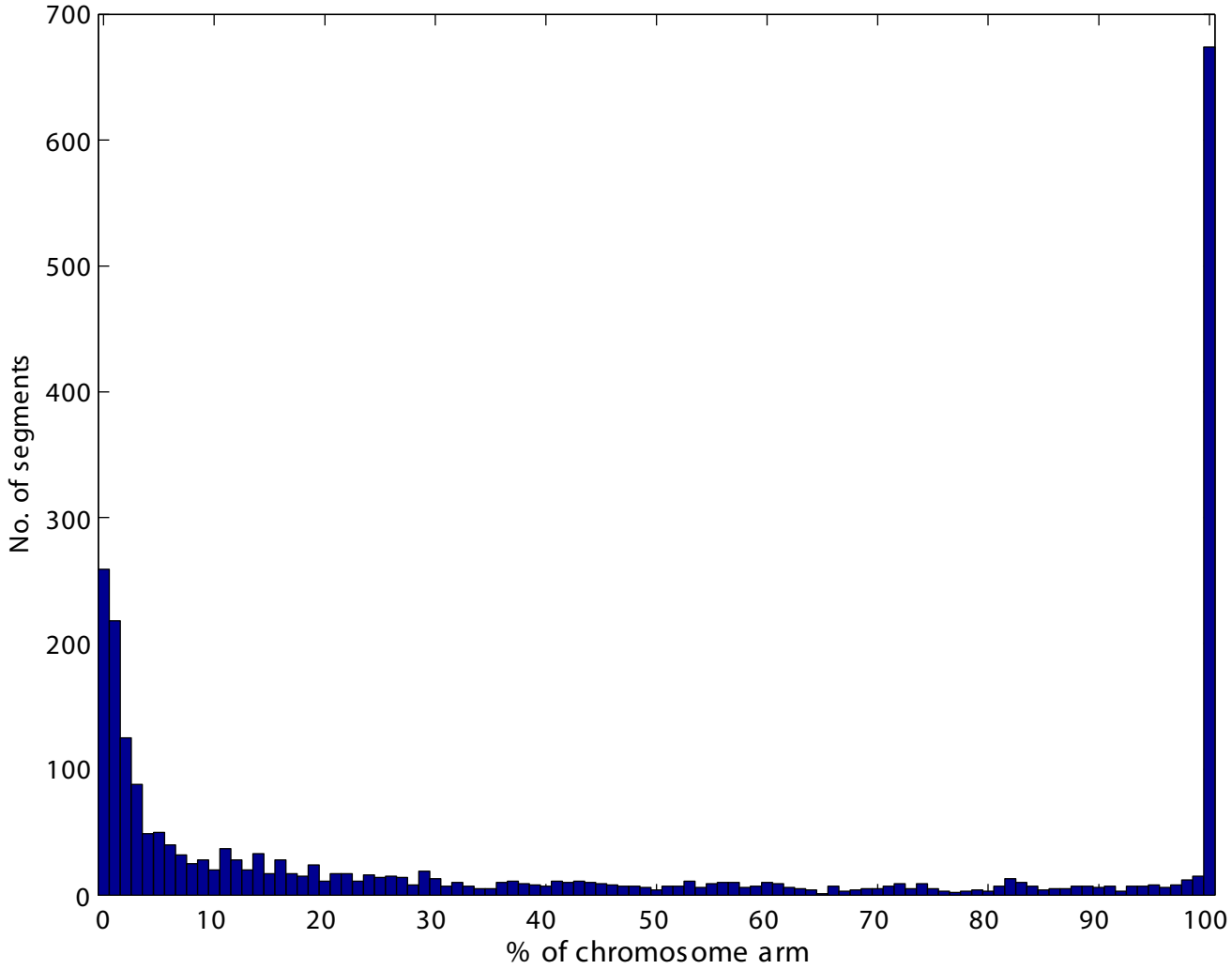

Batch ID
Histology
Gender
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
-1
-0.5
0
0.5
1
log2 ratio
0
0.5
1
G score (f×a)
0.25
10-9
10-99
Significance (q value)

k

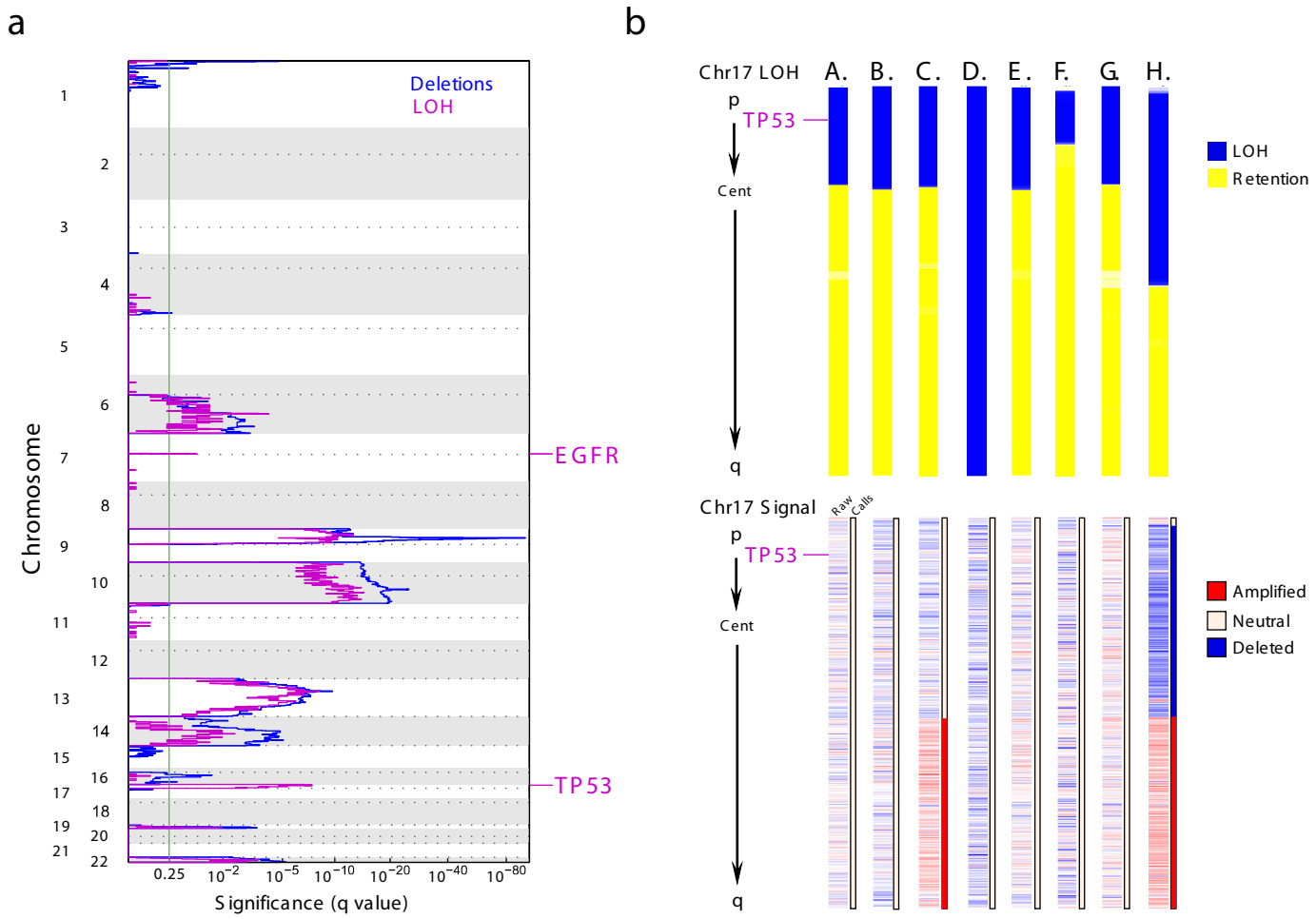

p
q
Tumors
Primary peak region
0.25 0.13 0.07
FDR q-value
Remove all aberrations involving the primary peak region
Use remaining aberrations to recalculate G scores and q values
p
q
Tumors
Independently significant secondary peak region
0.25 0.13 0.07
FDR q-value


I
Batch ID
Histology
Gender
Histology
Normal
Primary GBM
Secondary GBM
Anaplastic astrocytoma
Anaplastic oligodendroglioma
Anaplastic mixed glioma
Gender
Male
Female
7
7
7
EGFR
MET
0
0.2
0.4
0.6
G score
-1
-0.5
0
0.5
1
log2 ratio

Supplementary Figure 8



Supplementary Figure 9



Supplementary Table 1 Patient characteristics.

	Overall (n=141)	Passed quality control (n=105)	p value†
Primary GBM	107 (76%)	75 (71%)	0.46
Secondary GBM	15 (11%)	12 (11%)	0.84
Anaplastic Astrocytoma	9 (6%)	7 (7%)	1
Anaplastic oligodendroglioma	7 (5%)	8 (8%)	0.43
Anaplastic mixed glioma	2 (1%)	2 (2%)	1
Low grade mixed oligoastrocytoma	1 (1%)	1 (1%)	1
Median survival* (range)	538 (37-5026)	552 (63-5026)	0.62
Median age (range)	52 (20-82)	50 (20-82)	0.62
% male	64	62	0.89

* In days

† Calculated using Fisher's exact statistic (histologies, gender) or *t*-test (age, survival)

Supplementary Table 2 Peak regions of amplification, deletion, and non-overlapping LOH

			Initial dataset (141 gliomas)				Combined dataset (319 gliomas)				Oncogene or tumor suppressor gene in region
Amplification	Broad or focal*	Cytoband*	Center of peak**	q value	Frequency†	# of genes in peak§	Center of peak**	q value	Frequency†	# of genes in peak§	
1	both	7p11.2	54.7	1e-79	65% (26% focal)	1	54.7	1e-237	63% (31% focal)	1	*EGFR*
2	both	7q31.2	116.4	1e-24	73% (7% focal)	7	116.4	1e-49	66% (5% focal)	7	*MET*
3	focal	4q12	54.9	1e-14	18%	4	54.9	1e-13	11%	4	*PDGFRA*
4	focal	1q32.1	201.5	1e-7	15%	12	201.7	1e-18	15%	5	*MDM4*
5	focal	12q15	67.6	1e-5	7%	4	67.5	1e-16	7%	4	*MDM2*
6	broad	20q11.21	44.7	1e-3	22%	348	61.0	1e-9	24%	107	
7	focal	12p13.32	4.1	1e-3	14%	21	4.2	1e-7	12%	13	*CCND2*
8	broad	12p12.1	27.2	0.01	11%	28	Merged with region 7‡				*KRAS*
9	focal	12q14.1	56.9	0.04	7%	33	56.3	1e-14	9%	32	*CDK4*
10	focal	3q26.33	180.8	0.04	16%	24	183.6	1e-3	12%	8	*PIK3CA*
11	broad	8q24.12	122.0	0.05	15%	1	129.2	1e-5	15%	2	*MYC*
12	broad	19p13.12	11.9	0.07	22%	548	15.3	1e-6	23%	253	
13	focal	12q14.3	65.2	0.07	6%	1	66.1	1e-4	5%	1	
14	broad	17q24.1	58.8	0.07	11%	100	56.1	0.01	10%	61	
15	focal	6p21.1	43.1	0.13	7%	25	not significant				
16	focal	2p24.3	16.2	0.23	11%	4	16.2	0.09	7%	2	*MYCN*
17	focal	12q15	Merged with region 5‡				68.8	1e-4	5%	1	
18	focal	13q33.3	not significant				107.8	0.01	6%	5	
19	focal	9p22.1	not significant				19.1	0.02	5%	9	
20	broad	9q34.11	not significant				124.8	0.04	12%	184	
Deletion											
1	both	9p21.3	21.9	1e-100	69% (49% focal)	3	22.0	1e-242	61% (43% focal)	2	*CDKN2A/B*
2	both	10q23.31	89.4	1e-26	71% (7% focal)	2	89.5	1e-73	63% (7% focal)	1	*PTEN*
3	both	13q14.2	47.2	1e-13	47% (9% focal)	8	47.1	1e-25	39% (8% focal)	6	*RB1*
4	focal	1p36.31	5.2	1e-5	35%	11	7.8	1e-20	35%	8	*CHD5*
5	broad	22q13.31	44.4	1e-5	37%	120	45.3	1e-15	37%	21	
6	broad	14q31.3	82.6	1e-5	36%	1	72.0	1e-14	32%	25	
7	broad	19q13.41	55.2	1e-4	31%	151	55.3	1e-19	36%	204	
8	broad	6q23.2	140.1	1e-4	30%	57	163.7	1e-7	22%	2	
9	broad	16q21	58.2	0.02	23%	18	61.4	1e-3	16%	18	
10	focal	4q34.3	183.9	0.22	17%	1	180.0	1e-3	18%	1	
11	broad	11p15.4	6.0	0.24	24%	230	2.6	1e-5	23%	169	
12	broad	15q13.3	not significant				31.0	1e-3	15%	1	
13	focal	2q37.3	not significant				241.0	0.09	14%	68	
14	focal	5q34	not significant				166.1	0.20	12%	1	
Copy-neutral LOH											
1	broad	17p13.2	4.3	1e-8	22%	202	5.7	1e-12	19%	97	*TP53*

* Based upon the original dataset of 141 gliomas, except for regions detected only in the combined dataset

** Mb coordinates using hg16 build.

§ Number of known or predicted RefSeq genes within peak.

† The frequency of focal events is reported against all samples. The frequency of 17p LOH is understated because a number of events were obscured by contaminating normal DNA.

‡ Although this locus remains in a significantly amplified region, the “peel-off” algorithm (Supplementary Methods) no longer identifies it as an independent event.

Supplementary Table 3 List of gene probes with highest outlier scores in comparison of 7gain to 7norm among 528 genes on chromosome 7

Rank	Gene Probe	Genes	Outlier Score
1	217624_at	*PDAP1*	52.8922
2	211599_x_at	***MET***	43.6612
3	213816_s_at	***MET***	36.1742
4	214651_s_at	*HOXA9*	35.6697
5	210755_at	***HGF***	31.5658
6	217599_s_at	*MDFIC*	23.9145
7	210111_s_at	*KIAA0265*	21.6106
8	207561_s_at	*ACCN3*	20.4218
9	203291_at	*CNOT4*	19.7519
10	207060_at	*EN2*	18.868
11	202904_s_at	*LSM5*	18.5608
12	215198_s_at	*CALD1*	17.3731
13	200756_x_at	*CALU*	17.0493
14	204148_s_at	*ZP3* and *POMZP3*	16.9143
15	213360_s_at	*POM121* and *LOC340318*	15.9566
16	220618_s_at	*ZCWPW1*	15.9184
17	213807_x_at	***MET***	15.7769
18	210997_at	***HGF***	15.6395
19	219758_at	*FLJ12571*	15.4237
20	203630_s_at	*COG5*	15.365