Title

Double Minute Chromosomes in Glioblastoma Multiforme Are Revealed by Precise Reconstruction of Oncogenic Amplicons

Journal

Cancer Research, 73(19)

ISSN

0008-5472

Authors

Sanborn, J Zachary
Salama, Sofie R
Grifford, Mia
et al.

Publication Date

2013-10-01

DOI

10.1158/0008-5472.can-13-0186

Peer reviewed

Published in final edited form as:
Cancer Res. 2013 October 1; 73(19): 6036–6045. doi:10.1158/0008-5472.CAN-13-0186.

Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons

J. Zachary Sanborn[1,2], **Sofie R. Salama**[2,3,*], **Mia Grifford**[2], **Cameron W. Brennan**[4], **Tom Mikkelsen**[5], **Suresh Jhanwar**[6], **Sol Katzman**[2], **Lynda Chin**[7], and **David Haussler**[2,3]

[1]Five3 Genomics, LLC, Santa Cruz, CA, 9506

[2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA, 95064

[3]Howard Hughes Medical Institute, Santa Cruz, CA, 95064

[4]Human Oncology & Pathogenesis Program and Department of Neurosurgery, Memorial Sloan-Kettering Cancer Center, New York, New York 10065

[5]Depts. Neurology & Neurosurgery, Henry Ford Hospital Detroit, MI 48202

[6]Cytogenetics Laboratory, Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065

[7]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract

DNA sequencing offers a powerful tool in oncology based on the precise definition of structural rearrangements, copy number in tumor genomes. Here we describe the development of methods to compute copy number and detect structural variants with data synthesis to locally reconstruct highly rearranged regions of the tumor genome with high precision from standard short read, paired-end sequencing datasets. We find that circular assemblies are the most parsimonious explanation for a set of highly amplified tumor regions in a subset of glioblastoma multiforme (GBM) samples sequenced by The Cancer Genome Atlas (TCGA) consortium, revealing evidence for double minute chromosomes (DM) in these tumors. Further, we find that some samples harbor multiple circular amplicons and in some cases further rearrangements occurred after the initial amplicon-generating event. Fluorescence in situ hybridization (FISH) analysis offered an initial confirmation of the presence of DMs. Gene content in these assemblies helps identify likely driver oncogenes for these amplicons. RNA-seq data available for one DM offered additional support for our local tumor genome assemblies, identifying the birth of a novel exon made possible through rearranged sequences present in the DM. Consistent with previous estimates, our method was also useful for analysis of a larger set of GBM tumors for which exome sequencing data is available, finding evidence for oncogenic DMs in over 20% of clinical specimens examined.

Keywords

glioblastoma; genomic rearrangements; double minute chromosomes; tumor sequencing algorithms; oncogenic amplicons

[*]To whom correspondence should be addressed: Sofie R. Salama, Mail Stop CBSE/ITI, 1156 High St., Santa Cruz, CA, 95064, ph: 831-459-2814, fax: 831-459-1809, ssalama@soe.ucsc.edu.

Conflict of interest: J. Zachary Sanborn is a co-founder, equity holder and Chief Technology Officer, and David Haussler is a co-founder, equity holder and member of the Scientific Advisory Board of Five3 Genomics, LLC.

Introduction

High throughput methods for whole-genome sequencing have provided researchers with an unprecedented ability to measure the complex state of genomic rearrangements characteristic of most cancers. Numerous methods for inferring structural variation from paired-end sequencing data have been developed (1-3), but the structural variants called by such methods are often considered only in isolation, and used primarily to identify potential fusion genes. The difficulty in discovering all true structural variants and filtering out false positives makes it hard to use the output of such methods to reassemble large regions of the tumor genome. However, such tumor genome assemblies can reveal the complex structure of the tumor genome and can be used to infer the mechanism by which somatic alterations critical to cancer progression occur, such as amplifications of oncogenes and deletions of tumor suppressors. Ideally these tumor assemblies will also reveal unique features of the cancer that can be used as diagnostic features to monitor disease progression in the patient.

It is well documented that one mechanism by which genes become highly amplified in tumors is via the circularization of double-stranded DNA into what are known as double minute chromosomes (4). Double minutes (DMs) have been shown to confer resistance to certain drugs, as well as pass along this resistance non-uniformly to daughter cells. They have been observed up to a few megabases in size, and contain chromatin similar to actual chromosomes, but lack the centromere or telomeres found in normal chromosomes. Since DMs lack centromeres, they are, like mitochondria, randomly distributed to daughter cells during cell division (5). They are generally lost in future generations unless there is some selective pressure to maintain them. For example, when they confer selective growth advantage to tumor cells, they are readily retained at high copy number. In particular, DMs containing oncogenes may serve to amplify these genes to hundreds of copies per cell. DMs are common in several types of cancer, including brain cancers, with an estimated 10% of glioblastoma multiforme (GBM) tumors bearing DMs (6).

Homogeneously staining regions (HSRs) represent another common mode of extreme gene amplification in cancer, observed in both solid and hematological cancers (7-10). An HSR arises from high copy number tandem duplication of a genomic segment such that it expands the affected chromosome. Because they are embedded within larger chromosomes, fluorescent in situ hybridization (FISH) probes specific to genomic sequences within the HSR give a broad band of staining in a specific position within the larger chromosome, distinct from the focal staining usually observed with a locus specific FISH probe. It is believed that DMs and HSRs are related, in that DMs can derive from HSRs as well as create HSRs via chromosomal reinsertion (7,11,12). FISH is used to distinguish whether HSRs, DMs, or both are present in a given tumor sample. Whatever their form, these highly amplified oncogenes are often key drivers of their respective tumors, and may be vital in their detection, diagnosis, and treatment.

Here we present methods that analyze high-throughput whole genome sequencing data from tumor and matched-normal samples to detect these extremely amplified and rearranged regions of the tumor genome. We show that these low level results can be synthesized to construct accurate local genome assemblies that are circular in nature, suggesting the existence of DM chromosomes and/or HSRs in the tumor. We use FISH analysis to independently identify DMs and HSRs in two tumor samples. In addition, we show that RNA-seq data further supports our circular assemblies, in one case identifying a novel isoform of the gene CPM that co-opts intergenic DNA to create a novel exon. Further analysis of the amplicons is done to distinguish likely driver genes, such as MDM2, from likely passenger genes, such as those disrupted by the amplicon structure. We show that this

is made possible by the configuration of the rearranged regions represented by the predicted circular assembly. Finally, our analysis of a much larger set of samples with whole exome sequencing data suggests that 20% of GBM samples harbor highly amplified oncogenic amplicons.

Materials and Methods

Tumor and normal genome sequencing data

Tumor and matched-normal whole genome BAM files were downloaded from CGHub (cghub.ucsc.edu) under the following sample identifiers: TCGA-06-0648-01A-01D-0507-08, TCGA-06-0648-10A-01D-0507-08, TCGA-06-0145-01A-01D-0507-08, and TCGA-06-0145-10A-01D-0507-08. . In addition, 264 exome BAM files downloaded from CGHub, with sample identifiers given in Supplemental Table S1. Whole genome and exome BAM files were indexed by samtools (13) and processed by BamBam (See Supplemental Methods) to determine relative copy number, allele fraction, and structural variants. Supplemental Tables S1 and S2 list the copy number and rearrangement breakpoints detected by BamBam for exome and whole genome sequencing data, respectively.

Determining breakpoints related to highly amplified regions

The read support (number of overlapping reads) for a given breakpoint is directly proportional to the copy number of the regions it connects. Thus, by requiring breakpoints to have a high level of read support, we can filter out breakpoints that are part of a copy-neutral rearrangement, and breakpoints that led to low-copy amplifications and deletions. We can then focus on the breakpoints that are part of highly amplified regions in the tumor. The particular read support threshold is chosen such that breakpoints that have read support expected of copy-neutral regions of the tumor genome are removed.

Reconstructing amplicons by walking a breakpoint graph

Similar to a recently published method (14), we construct a breakpoint graph by defining a node for each side of an amplified segment of the tumor genome, connecting these two nodes by a segment edge that represents the amplified segment, and defining a set of bond edges, each of which represents a pair of segment sides that are adjacent in the tumor genome. Which amplified segments of the tumor genome are included as segment edges in the breakpoint graph is determined by relative copy number. The bond edges are the highly supported breakpoints found in the manner described above. If an amplified segment is interrupted by a breakpoint, then that segment will be split into two segment edges.

We visualize the graph with the segments laid out according to genomic position. Assuming all segments have the same relative copy number and each segment side has exactly one bond edge attached to it, we determine the arrangement of the amplified segments in the tumor genome by starting at the node for the left side of the first segment and traversing the segment edge to the node at the other side. The path then follows a bond edge attached to that node over to the segment side to which it is connected, then traverses the new segment edge, continuing in this manner of alternatively traversing segment and bond edges until we have returned to the starting node. If there are segment edges remaining that we have not traversed, then we choose a new starting node among them and repeat this procedure, until all segment edges are accounted for. Each cyclic path we determine defines a separate circular chromosome via the concatenation of its segments in the direction of traversal. This interpretation of the graph is unambiguous, as the overall direction of traversal of any circular chromosome is immaterial, as is the order in which we discover them. When the number of bond edges per segment side (i.e. node) in the graph is not uniformly 1, or not all

segments have the same relative copy number, some segments may require more than one traversal to account for the extra copies and some walks may terminate at segment sides that have no bond. Whenever multiple walks can be made through the set of bond edges and segments, we enumerate the different solutions, including potentially circular solutions that cannot close the loop due to one or more missing bond edges. A single solution that features multiple independent cyclic paths occurs when multiple circular walks are required in the procedure described above. Here distinct sets of bond edges and segments are connected in independent closed loops that lack any bond edges that could fuse the independent loops together into a single, larger closed loop. A toy example breakpoint graph and its solution are described in Supplemental Figure 1.

When there are multiple solutions, the optimal path(s) through the graph are taken to be those that most closely agree with the observed relative copy number. The number of times a solution traverses a given segment produces an estimate of that segment's copy number. The root mean square deviation (RMSD) of the segment traversal counts to the observed relative copy number for each solution is calculated, and then the solution(s) with the smallest RMSD value are labeled as optimal.

RNA-Seq analysis

RNA-Seq reads were mapped using BWA(15) and coverage was calculated using BEDtools (16) . We calculated the total coverage per transcript per million uniquely mapped read pairs for each gene within the TCGA-06-0648 DM. For CPM the coverage over a truncated version of the gene was used as described in the text. These normalized expression metrics of TCGA-06-0648 were compared with a set of 9 other TCGA GBM samples that comprised the original RNA-Seq cohort for GBM (Table 1).

FISH analysis

FISH was performed for EGFR/Cep7 and MDM2/Cep12 using pre-labeled probes (Vysis, Abbott Molecular). Charged slides with 4 micron paraffin sections were dewaxed, rinsed, microwaved in 10mM sodium citrate buffer, then digested in pepsin-HCl (40 μg/mL, 10 min at 37 °C), rinsed, and dehydrated. Probe and slides were codenatured using a HYBrite automated hybridizer at 80 °C for 8 min then hybridized for 2 to 3 d at 37 °C.

Exome analysis to estimate prevalence of DMs

Additional samples that potentially bear DMs were identified in TCGA GBM exome datasets as follows. Tumor and matched normal exomes were processed by BamBam to compute relative coverage and identify somatic rearrangements. High copy number peaks were defined as regions with 5-fold increased relative coverage versus their matched normal, a threshold chosen to conservatively filter out peaks caused by low level amplifications and noise (see Supplemental Methods for details). GBM tumors with multiple such high copy number peaks were manually analyzed to discover any oncogenes within peaks, associate peaks with nearby somatic rearrangements, and determine if a sample exhibits multiple peaks with similar copy number levels. Samples were scored as having a possible DM if they either contain multiple distinct high copy number peaks with similar copy number levels and at least one peak contained an oncogene, or a single distinct peak containing an oncogene, spanning approximately 1Mb, and having an associated rearrangement.

Association of DM/HSR samples with other GBM tumor features

Samples containing likely chromothripsis events were identified from the set of 26 samples with whole genome sequencing data determined to have a possible DM by choosing the subset that had multiple (>3) high copy number peaks (18 samples), and were compared to

samples containing no high copy number peaks as determined by whole exome data (112 samples). T-tests were performed in R comparing several features between the two groups, including molecular subtype, survival, mutation in PTEN, TP53, KEL, IDH1, PIK3R1, PIK3CA, POTEB, NF1, RB1, and EGFR, amplification of PRDM2, MET, MDM2, EGFR, CDK4, CCNE1, and CCND2, and deletion of RB1, PTEN, PRDM2, MET, and CCND2. Bonferroni-adjusted p-values were reported. TP53 mutations within the two groups were further evaluated using a two-tailed Fisher's exact test.

Results

BamBam: a robust method for identifying tumor-specific variation

Considering that a BAM file storing a single patient's whole genome sequence at high coverage (>30×) can be hundreds of gigabytes in compressed form, a serial analysis of two sequencing datasets requires researchers to store intermediate results that must be merged to perform a comparative analysis, such as identifying mutations found only in the tumor sample. To overcome this problem, we developed BamBam, a tool that performs a comparative analysis of a patient's tumor genome versus his/her germline by simultaneously processing the tumor and matched-normal short-read alignments stored in SAM/BAM-formatted files (13). Simultaneous processing of both BAM files enables BamBam to efficiently calculate tumor relative coverage and allele fraction, discover somatic mutations and germline SNPs, and infer regions of structural variation. The relative coverage and allele fraction estimates made by BamBam can be used to estimate tumor copy number and normal contamination in the sequenced tumor sample (See Supplemental Methods for details).

Reconstruction of candidate double minute chromosomes using whole genome sequencing

We focused on three samples from the initial set of 19 glioblastoma multiforme (GBM) samples from TCGA subjected to whole genome sequence analysis where we detected highly amplified segments overlapping oncogenes, suggestive of DMs. We applied these methods to samples designated TCGA-06-0152, TCGA-06-0648, and TCGA-06-0145. In each case, whole genome sequencing of a tumor biopsy was available separately from whole genome sequencing of a blood sample (matched normal tissue sample). For two of these samples (TCGA-06-0152 and TCGA-06-0648), multiple segments had similar levels of amplification whereas the third sample (TCGA-06-0145) had one large amplified region with further rearrangements internal to the region.

Sample TCGA-06-0648 had a striking pattern of genome amplification in which 16 distinct segments (15 from chromosome 12 and a small fragment from chromosome 9) had similarly high levels of amplification (>10 copies) and also appeared to be linked to each other by high confidence rearrangement events identified by BamBam (Figure 1a). One of the chromosome 12 segments contains the MDM2 oncogene. Out of the total of 701 putative somatic breakpoints called, 97 breakpoints met the filtering criteria specified in Supplemental Methods. Only 16 of these 97 breakpoints further met or surpassed a chosen minimum read support threshold of 100 to identify breakpoints likely associated with highly amplified regions, including two breakpoints that did not have split read evidence. All of these highly supported breakpoints are proximate to the boundaries of the highly amplified segments clustered on chromosome 12 suggesting that the highly supported breakpoints and the amplifications are directly associated and may represent the rearranged configuration of one (or multiple) amplicons in the tumor's genome.

Figure 1b shows a schematic of the amplified segments of TCGA-06-0648 and their associated breakpoints. This diagram predicts a circular path that completely accounts for all observed breakpoints and amplified segments, resulting in a single 891kb circular amplicon containing a single copy of MDM2. Since all but two breakpoints were refined by split read solutions, the estimated size of this amplicon is accurate to within ~100bp. This reconstructed circle is consistent with either an array within a larger chromosome of precise tandemly duplicated copies of an initial circular amplicon formed from these 16 genomic segments (assuming single-copy breakpoints joining this tandem array to non-repeated DNA were not sufficiently covered to be observed) or an extra-chromosomal circular DNA (DM) (17) with average copy number of ~84 in the tumor sample. Since we can assume neither a perfectly clonal tumor nor a stable number of DM copies in every generation of tumor cells, the average copy number of 84 should be considered the number of DM copies in the average tumor cell sequenced. We specifically searched for breakpoints with lower read support within the amplicon region connecting it to other single-copy genomic locations, as these could provide evidence for an insertion site of a tandem array, but found none. Thus, the presence of a DM is the more parsimonious explanation of these data, as it does not require us to postulate the existence of one or more pairs of unobserved breakpoints where one or more HSRs each containing multiple exact copies of this 891kb region are inserted into larger chromosomes.

DMs have been identified in many tumor types where they often contain oncogenes important for that cancer, such as EGFR in the case of GBM (18). The TCGA-06-0648 DM contains several protein coding genes from chromosome 12, including intact copies of the MDM2 oncogene and CAND1, which encodes an inhibitor of cullin ring-ubiquitin ligase complexes (19). It also includes a truncated allele of carboxypeptidase M (CPM), a membrane-bound and secreted protease that cleaves the C-terminal residue of epidermal growth factor (20). The amplified allele of CPM lacks the last exon, which should not affect the catalytic or major structural domains of the protein, but removes the amino acids necessary for GPI anchoring of CPM to the plasma membrane and would be expected to result in an exclusively secreted form of CPM (21). A partial allele of the ras-family protein RAP1B is also present, but as it lacks the promoter and first exon, it is unlikely that this allele is expressed. It seems likely that MDM2 drove the high copy maintenance of this DM in TCGA-06-0648, but CPM and CAND1 could contribute as well.

This region of the tumor genome has all of the hallmarks of a chromothripsis event, suggesting that the DM was created by connecting shattered fragments of chromosome 12 and a small region of chromosome 9 into a single circular episome (22). The observation that multiple regions with uniformly high read depths are connected by a set of structural variants with similarly high read support suggests that these alterations likely occurred together during a single event, such as chromothripsis, instead of a series of independent focal rearrangements. Furthermore, all breakpoints lack homology or exhibit 2-6 bp microhomologies at their junctions, indicating that non-homologous end-joining (NHEJ) and microhomology-mediated end joining (MMEJ) are the primary DNA double stranded break repair mechanisms responsible for constructing the double minute (23).

We applied these same methods to sample TCGA-06-0152, which also had several highly amplified segments, including segments containing the MDM2, CDK4 and EGFR oncogenes (TCGA, manuscript in preparation). This analysis predicted two amplicons, in which each amplicon harbored at least one oncogene (Supp. Figure 2).

The final case, TCGA-06-0145, exhibits an extreme level of amplification (>50-fold) of a single ~800kb genomic segment including *EGFR* that could indicate the presence of an EGFR-DM or HSR (Figure 2). In contrast with the other samples, the amplified region of

TCGA-06-0145 contains significant variations in the major and minor allele frequency as well as deletion and duplication events with lower read support, which are more compatible with an HSR interpretation. However, again, we were unable to find evidence of breakpoints that would link this amplicon to another genomic region, which argues against an HSR.

The solution to the breakpoint graph of TCGA-06-0145 shows the possibility of three distinct paths that incorporate all breakpoints and explain the observed copy number, and each path predicts a different form of EGFR. EGFR is intact in the dominant path (7 of every 9 copies). The remaining two paths, each present in 1 of 9 copies, feature breakpoints that are internal to the EGFR gene, with one path producing a non-functional form of EGFR and the other deleting exons 2-7 of EGFR. This form of EGFR is known as EGFRvIII, a highly oncogenic, constitutively active form of EGFR that is expressed in multiple tumor types (24). This is interesting since it suggests two scenarios: (1) EGFRvIII emerges after wildtype EGFR is significantly amplified or (2) EGFRvIII is created early but cells with more copies of wildtype EGFR have a selective advantage in the tumor population. The former scenario seems most plausible, as the increased number of copies subsequently improves the chance that the EGFRvIII mutant will arise. Regardless of the true scenario, the ratio of EGFRvIII to wildtype EGFR suggests that high copy number of oncogenic EGFRvIII may not be necessary to provide significant advantage over the wildtype amplification to the growing tumor cell.

Transcriptome data reveals a novel DM-associated fusion protein

For one of the three tumor samples examined in this study (TCGA-06-0648), RNA sequencing was also performed by TCGA. We analyzed these data along with that from the nine other samples in the initial GBM RNA-Seq batch to examine the expression of alleles associated with the TCGA-06-0648 DM. As expected from the absence of the promoter and first exon, RAP1B expression was half that of the other GBM samples that do not have amplifications in this region suggesting that only the intact copy of RAP1B on chromosome 12 is expressed and the DM allele is not expressed. In contrast, MDM2, CAND1 and the first eight exons of CPM were expressed at >15-fold higher levels than was observed in the GBM samples lacking amplification of these genes (Table 1).

For CPM, we observed that many reads originating in exon 8 terminated in a region 1.47 Mb away from it in the normal version of chromosome 12, but only 13.5 kb away in the DM. Closer analysis of this region revealed that the 5′ end of these reads are just downstream of a canonical splice site acceptor sequence that generates a new exon encoding a novel 30 amino acid carboxy terminus for the DM-derived CPM allele (Fig. 3). This region is not part of any known transcript and the resulting protein sequence has no strong homology to any other proteins. This sequence is unlikely to provide a GPI anchor site so we anticipate that the DM-derived CPM protein would be secreted. It is not clear what the functional effect of expressing this altered CPM gene would be, although it may affect EGF metabolism as both membrane-bound and secreted forms of CPM are known to cleave the carboxy-terminal arginine of EGF (25).

In summary, from the point of view of gene expression, the DM results in overexpression of MDM2, CAND1 and a novel form of CPM in this tumor sample.

TCGA-06-0648 and TCGA-06-0145 amplicons exist as DMs

The ability to distinguish HSRs from DMs from short read sequencing data is limited. To independently assess the nature of the amplification events in these tumors, we performed FISH analysis on paraffin sections derived from tumors TCGA-06-0648 and TCGA-06-0145 using probes to MDM2 (amplified in TCGA-06-0648) and EGFR

(amplified in TCGA-06-0145) (Figure 4). Material was unavailable for TCGA-06-0152. In 0648, the MDM2 probe gave a punctate pattern throughout the nucleus indicating many non-chromosomal sites of MDM2, typical of DMs. In contrast, the EGFR probe gave a broad pattern of staining in addition to punctate spots for TCGA-06-0145, consistent with a combination HSR and DM.

Prevalence of putative DMs / HSRs in exome sequencing data

Exome sequencing is cheaper than whole-genome sequencing, and samples of tumors that have been subjected to exome sequencing are more plentiful. Evidence for DMs can be obtained from exome-sequencing by searching for patterns of multiple regions of high-level amplification overlapping at least one oncogene, a pattern common to the DM-containing samples we analyzed. In contrast, broad or chromosome arm level amplification events as well as focal events with a modest level of amplification (e.g. duplications) are not expected to be part of a DM. In order to estimate the prevalence of DMs/HSRs in glioblastoma multiforme, we searched a set of 264 TCGA samples with tumor and matched-normal exome sequencing data for signatures of DMs. As detailed in Methods, we first performed a computational survey of samples to identify samples exhibiting focal extreme amplification(s), prioritizing those samples with multiple distinct peaks on one or more chromosomes. A careful manual review of these samples was performed to assess the likelihood that the sample contains a DM, looking for the quality of relative copy number calls, any evidence of structural variation associated with the amplified peaks, and the presence of potential oncogenes.

As described in Table 2 and Supplemental Table S1, 61 samples (23%) have features suggesting the presence of a DM. A total of 121 oncogenes were amplified across these 61 samples, with at least one oncogene identified in every putative DM. *EGFR* was the oncogene most frequently associated with these high level amplicons, followed by *CDK4* and *MDM2*. *MYCN*, which has been identified in a number of GBM DMs (6), was associated with amplicons in two samples. As a group, the putative DM/HSR-containing samples showed similar survival to the cohort of samples of analyzed by exome sequencing (data not shown). Overall these results suggest that almost a quarter of GBM samples have oncogenic amplicons present at high copy number.

Validation of exome sequencing-based prediction of DM / HSR containing samples

Recently, the TCGA project performed whole genome sequencing on an additional 25 tumor/normal pairs from the GBM cohort. This new data set contains 23 samples that we predicted to harbor a DM/HSR based on our curation of the exome data for these samples. While complete analysis of this data is being pursued by the Analysis Working Group, we performed a preliminary analysis using BamBam and the methods described above to identify circular amplicons in these samples (summarized in Supplemental Table S3). For 16/23 samples, we were able to reconstruct at least one circular amplicon. For the remaining 7/23 samples, multiple highly amplified peaks were identified with breakpoints connecting many, but not all of the peaks. We could not reconstruct circular amplicons for these samples, although the possibility remains that rearrangement breakpoints allowing for a circular solution were not detected for technical reasons. For the 2/25 samples where we did not predict a DM/HSR based on the exome sequencing data, we also did not detect highly amplified genomic regions with associated rearrangements in the whole genome sequencing data.

This larger set of samples with circular amplicons allowed us to look for common features associated with these samples compared to TCGA GBM samples where exome data suggests no amplicons. Previous studies identifying medullablastomas with chromothripsis-

associated DMs and other complex genetic rearrangements noted that such samples harbor p53 mutations (26,27). However, there was no enrichment of p53 mutations in the samples with circular amplicons in this sample cohort. The only significant association we observed was with PTEN deletion (Bonferroni-adjusted p-value=0.0052), a common event in GBM tumors.

Focusing on the sixteen samples where we successfully reconstructed circular amplicons with at least one oncogene, a mixture of simple and complex amplicons was observed. One sample harbored two simple circular amplicons each containing one genomic segment with an oncogene and the remaining five had a single oncogenic amplicon containing one genomic segment (Supplemental Table S3). Ten samples are complex, harboring multiple highly amplified segments from one or more chromosomes. Of these ten, six samples have at least two circular amplicons. Together these results strengthen our estimates of the prevalence DM/HSR amplicons in GBMs and suggest that samples containing such amplicons can often harbor multiple independent amplicons.

Discussion

The ability to integrate relative copy number with breakpoints enables us to understand the topology of vital parts of the cancer genome. By examining the overall pattern of amplification events in the TCGA GBM whole genome sequencing data, we found that some samples had multiple highly amplified segments of similar copy number. Surprisingly, for many of these highly amplified tumor regions, we are able to completely explain both the observed copy number and highly supported breakpoints surrounding them by solving a simple breakpoint graph, which describes the order and orientation of the highly amplified segments in the tumor genome. For the three GBM samples analyzed in detail here, the optimal solutions to the breakpoint graphs of amplified segments are circular amplicons. These circular solutions suggest that the observed amplified regions are DMs or HSRs.

Sequence coverage of 30X could be limiting our ability to detect breakpoints associated with the chromosomal integration site of an array of these amplicons, as the copy number of the breakpoint at the integration site is much lower than that of the amplicon. Thus, at this coverage we cannot reliably distinguish between a DM and an HSR with our bioinformatic analysis. The availability of tissue sections derived from these same tumor samples did allow us to directly address this issue for two samples. FISH analysis of one sample, TCGA-06-0648, is consistent with a DM whereas another, TCGA-06-0145, gives a pattern suggestive of a combination of DM and HSR.

Much can be learned through precise knowledge of the amplicon structure. Genes whose coding or promoter regions are disrupted by the amplicon structure are obvious passenger gene candidates, provided that transcriptional machinery is unable to create a novel transcript like the one observed with the birth of a new exon for the CPM gene. Highly expressed genes such as MDM2 that remain intact may drive tumor development and/or play a role in tumorigenesis, utilizing the DM's ability to rapidly reproduce to significantly increase the oncogenic capacity of these cells. The highly amplified state of these oncogene-harboring amplicons indicates that they have strong oncogenic potential, and thus confer a selective advantage to the tumor cell. Their formation was likely a key event in the tumorigenesis of these GBM tumors and they are likely to persist over time.

Examining the larger set of over 250 TCGA GBM samples for which there is exome data suggests that oncogenic amplicons are quite prevalent as they are found in almost a quarter of the samples. Using recently generated whole genome sequencing data for an additional 23 samples from the set of samples where we had predicted DM/HSRs from exome sequencing

data, we were able to confirm all of these samples had highly amplified genomic segments with rearrangements connecting at least some of the amplified segments. Further, we were able to reconstruct circular amplicons for 16 of these samples and discovered that 6 had multiple amplicons as we had observed for TCGA-06-0152, suggesting that the presence of multiple amplicons is common, and that they persist by conferring a combined selective advantage to the tumor cells. These results bolster the notion that chromothripsis-type events occur with reasonable frequency in GBM and through amplification of oncogene expression contribute to tumorogenesis.

The prevalence of these amplicons suggests that tumor-specific breakpoints associated with DM amplicons may be a potential diagnostic for the presence of tumor cells in a significant fraction of GBM patients, especially if DM-derived DNA is present in the blood. Several recent observations suggest the possibility of finding such tumor DNA in the blood of glioma patients. Skog et al. reported microvesicles that bud off from GBM tumor cells with lipid bilayers intact, carrying cytoplasmic contents of the tumor cell such as mRNA, miRNA, and angiogenic proteins (28). These vesicles can deliver their contents to other cells or blood. More recently, Balaj et al. have isolated microvesicles containing single-stranded DNA (ssDNA) with amplified oncogenic sequences, in particular c-Myc (29). Tumor nucleic acids leaked into the blood in this manner or via macrophages that engulf necrotic or apoptotic cells have been proposed as possible disease biomarkers in several cancers including glioma (30,31). Indeed, a recent study robustly detected tumor-associated rearrangements from breast and colorectal cancer patients via high throughput sequencing of the cell-free, plasma fraction of blood (32). The methods described in this study could readily be applied to such sequencing data and thus potentially provide a non-invasive method to characterize and monitor glioma patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the members of the TCGA GBM Analysis Working Group and the Haussler and Josh Stuart lab cancer genomics group for helpful discussions of our results. We also thank Melissa A. Wynne for technical assistance with the FISH assay. The results published here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Grant Support: This work was supported by the following grants to DH: NCI 5U24ACA143858, NHGRI, Stand Up To Cancer SU2C-AACR-DT0409, NHGRI 5U01ES017154 and the following grants to LC:NCI 5U24CA143845, NCI P01CA095616. DH is an investigator of the Howard Hughes Medical Institute, SK is supported by funds from the California Institute for Quantitative Biosciences and CWB is supported by the Leon Levy Foundation.

Grant Support: Howard Hughes Medical Institute (DH, SRS), NCI 5U24ACA143858 (DH, JZS, MG), NCI 5U24CA143845 (LC), NCI P01CA095616 (LC), NHGRI (DH, MG), Stand Up To Cancer SU2C-AACR-DT0409 (DH, MG), NHGRI 5U01ES017154 (DH, MG), California Institute for Quantitative Biosciences (SK), Leon Levy Foundation (CWB)

References

1. Sindi S, Helman E, Bashir A, Raphael BJ. A geometric approach for classification and comparison of structural variants. Bioinformatics. 2009; 25:i222–30. [PubMed: 19477992]
2. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Meth. 2009; 6:677–81.

3. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nature genetics. 2011; 43:964–8. [PubMed: 21892161]
4. Barker PE. Double minutes in human tumor cells. Cancer Genetics and Cytogenetics. 1982; 5:81–94. [PubMed: 6175392]
5. Lundberg G, Rosengren AH, Hakanson U, Stewenius H, Jin Y, Stewenius Y, et al. Binomial mitotic segregation of MYCN-carrying double minutes in neuroblastoma illustrates the role of randomness in oncogene amplification. PLoS ONE. 2008; 3:e3099. [PubMed: 18769732]
6. Bigner SH, Mark J, Bigner DD. Cytogenetics of human brain tumors. Cancer Genet Cytogenet. 1990; 47:141–54. [PubMed: 2192793]
7. Balaban-Malenbaum G, Gilbert F. Double minute chromosomes and the homogeneously staining regions in chromosomes of a human neuroblastoma cell line. Science. 1977; 198:739–41. [PubMed: 71759]
8. Yoshimoto M, Caminada De Toledo SR, Monteiro Caran EM, de Seixas MT, de Martino Lee ML, de Campos Vieira Abib S, et al. MYCN gene amplification. Identification of cell populations containing double minutes and homogeneously staining regions in neuroblastoma tumors. The American Journal of Pathology. 1999; 155:1439–43. [PubMed: 10550298]
9. Streubel B, Valent P, Jager U, Edelhauser M, Wandt H, Wagner T, et al. Amplification of the MLL gene on double minutes, a homogeneously staining region, and ring chromosomes in five patients with acute myeloid leukemia or myelodysplastic syndrome. Genes Chromosomes Cancer. 2000; 27:380–6. [PubMed: 10719368]
10. Storlazzi, CT.; Lonoce, A.; Guastadisegni, MC.; Trombetta, D.; D'Addabbo, P.; Daniele, G., et al. Genome Research. Vol. 20. Cold Spring Harbor Laboratory Press; 2010. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure; p. 1198-206.
11. Brookwell R, Hunt FA. Formation of double minutes by breakdown of a homogeneously staining region in a refractory anemia with excess blasts. Cancer Genet Cytogenet. 1988; 34:47–52. [PubMed: 3395993]
12. Reddy KS. Double minutes (dmin) and homogeneously staining regions (hsr) in myeloid disorders: a new case suggesting that dmin form hsr in vivo. Cytogenet Genome Res. 2007; 119:53–9. [PubMed: 18160782]
13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
14. Greenman CD, Pleasance ED, Newman S, Yang F, Fu B, Nik-Zainal S, et al. Estimation of rearrangement phylogeny for cancer genomes. Genome Research. 2012; 22:346–61. [PubMed: 21994251]
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. [PubMed: 19451168]
16. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–2. [PubMed: 20110278]
17. Kuttler F, Mai S. Formation of non-random extrachromosomal elements during development, differentiation and oncogenesis. Semin Cancer Biol. 2007; 17:56–64. [PubMed: 17116402]
18. Vogt N, Lefèvre S-H, Apiou F, Dutrillaux A-M, Cör A, Leuraud P, et al. Molecular structure of double-minute chromosomes bearing amplified copies of the epidermal growth factor receptor gene in gliomas. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:11368–73. [PubMed: 15269346]
19. Duda DM, Scott DC, Calabrese MF, Zimmerman ES, Zheng N, Schulman BA. Structural regulation of cullin-RING ubiquitin ligase complexes. Curr Opin Struct Biol. 2011; 21:257–64. [PubMed: 21288713]
20. Deiteren K, Hendriks D, Scharpe S, Lambeir AM. Carboxypeptidase M: Multiple alliances and unknown partners. Clin Chim Acta. 2009; 399:24–39. [PubMed: 18957287]
21. Reverter D, Maskos K, Tan F, Skidgel RA, Bode W. Crystal structure of human carboxypeptidase M, a membrane-bound enzyme that regulates peptide hormone activity. Journal of Molecular Biology. 2004; 338:257–69. [PubMed: 15066430]

22. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell. 2011; 144:27–40. [PubMed: 21215367]

23. McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 2008; 24:529–38. [PubMed: 18809224]

24. Moscatello DK, Holgado-Madruga M, Godwin AK, Ramirez G, Gunn G, Zoltick PW, et al. Frequent expression of a mutant epidermal growth factor receptor in multiple human tumors. Cancer Research. 1995; 55:5536–9. [PubMed: 7585629]

25. McGwire GB, Skidgel RA. Extracellular conversion of epidermal growth factor (EGF) to des-Arg53-EGF by carboxypeptidase M. J Biol Chem. 1995; 270:17154–8. [PubMed: 7615511]

26. Rausch T, Jones DTW, Zapatka M, Stutz AM, Zichner T, Weischenfeldt J, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012; 148:59–71. [PubMed: 22265402]

27. Northcott PA, Shih DJH, Peacock J, Garzia L, Morrissy AS, Zichner T, et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature. 2012; 488:49–56. [PubMed: 22832581]

28. Skog J, Würdinger T, van Rijn S, Meijer DH, Gainche L, Curry WT, et al. Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. Nature. 2008; 10:1470–6.

29. Balaj L, Lessard R, Dai L, Cho Y-J, Pomeroy SL, Breakefield XO, et al. Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nat Comms NIH Public Access. 2011; 2:180.

30. Schwarzenbach H, Hoon DSB, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nature. 2011; 11:426–37.

31. Lavon I, Refael M, Zelikovitch B, Shalom E, Siegal T. Serum DNA can define tumor-specific genetic and epigenetic markers in gliomas of various grades. Neuro-Oncology. 2010; 12:173–80. [PubMed: 20150384]

32. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Craig D, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Science Translational Medicine. 2012; 4:162ra154.

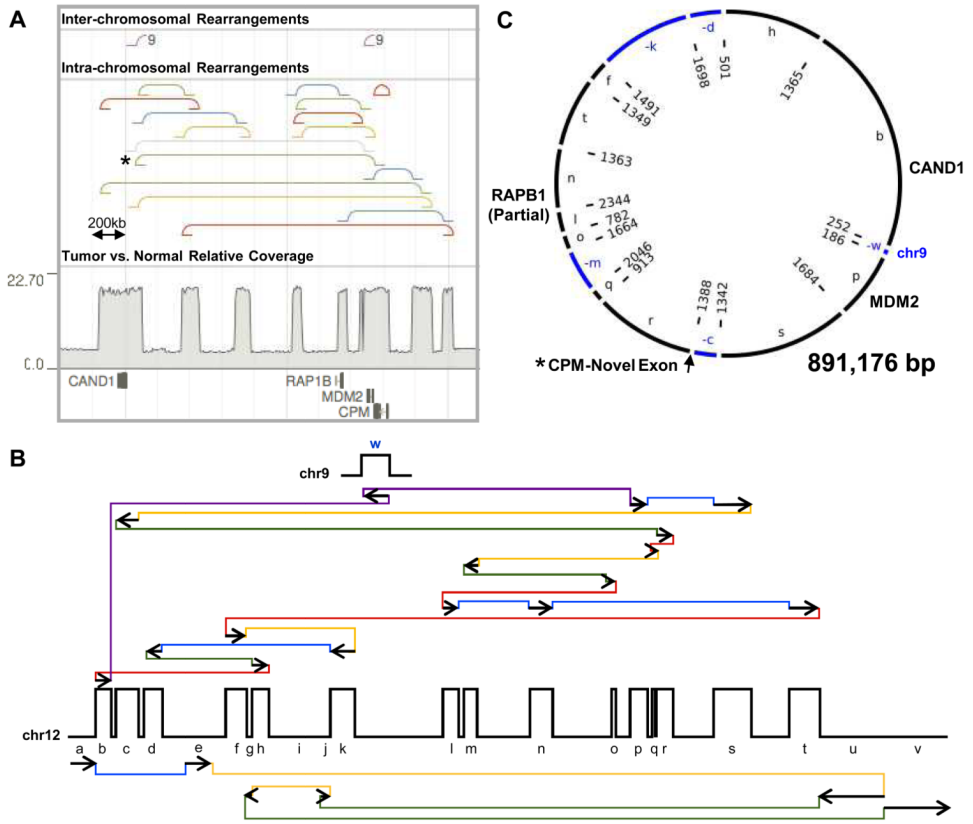

Figure 1. Reconstruction of TCGA-06-0648 double minutes

(a) Tumor browser view of a region of chromosome 12 from TCGA-06-0648. The bottom track shows protein-coding genes overlapping the amplified segments. The track above shows relative copy number of the tumor DNA compared to the normal, showing distinct blocks of elevated copy number with similar total copy number between blocks. The next two tracks show intra- and inter-chromosomal rearrangement breakpoints. All rearrangements shown are supported by at least 100 discordant reads. The type of rearrangement is indicated by the color of the line: duplication (red), deletion (blue), inversion (yellow and green), inter-chromosomal rearrangement (purple). (b) A diagram of the amplified segments and structural variants identified on chromosome 12 and 9 showing. Walking through this diagram results in a circular solution suggesting the DM diagrammed in (c) where segments inverted relative to their orientation in the reference genome are indicated by (–) and colored blue. The letters inside the circle correspond to the segments in (b). The numbers inside the circle indicate the number of sequencing reads supporting each breakpoint.

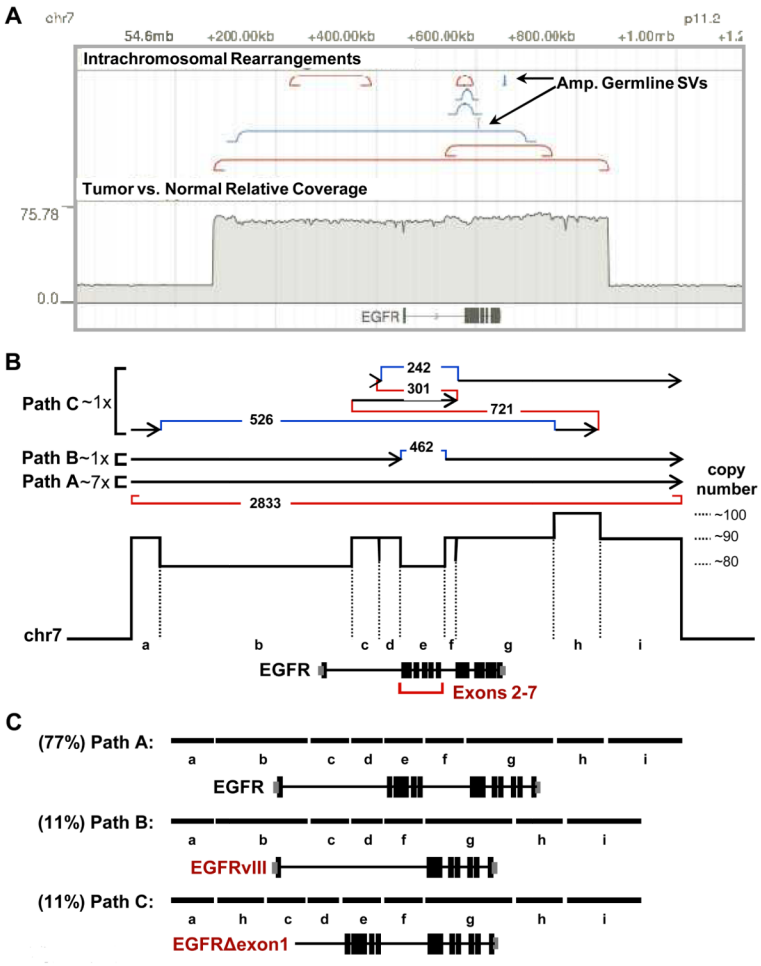

Figure 2. Reconstruction of 06-0145 amplicons

(a) Browser shot of the amplicon on chromosome 7. Tracks as described for figure 1. The two “Amp. Germline SVs” are known germline breakpoints present on the amplicon. (b) Diagram of the three paths that together form a potential solution of the breakpoint graph for sample TCGA-06-0145 that accounts for all observed, highly- supported breakpoints. The location of the gene EGFR is noted. Note that all paths, A, B and C, can create circular solutions by using the encompassing tandem duplication shown in red to connect the end of each path to the beginning. The tandem duplication may also connect path A to path B, path C to path B, etc. (c) The effects of each path on the EGFR gene, showing that path B creates an oncogenic form of EGFR, EGFRvIII, through specific deletion of exons 2-7.

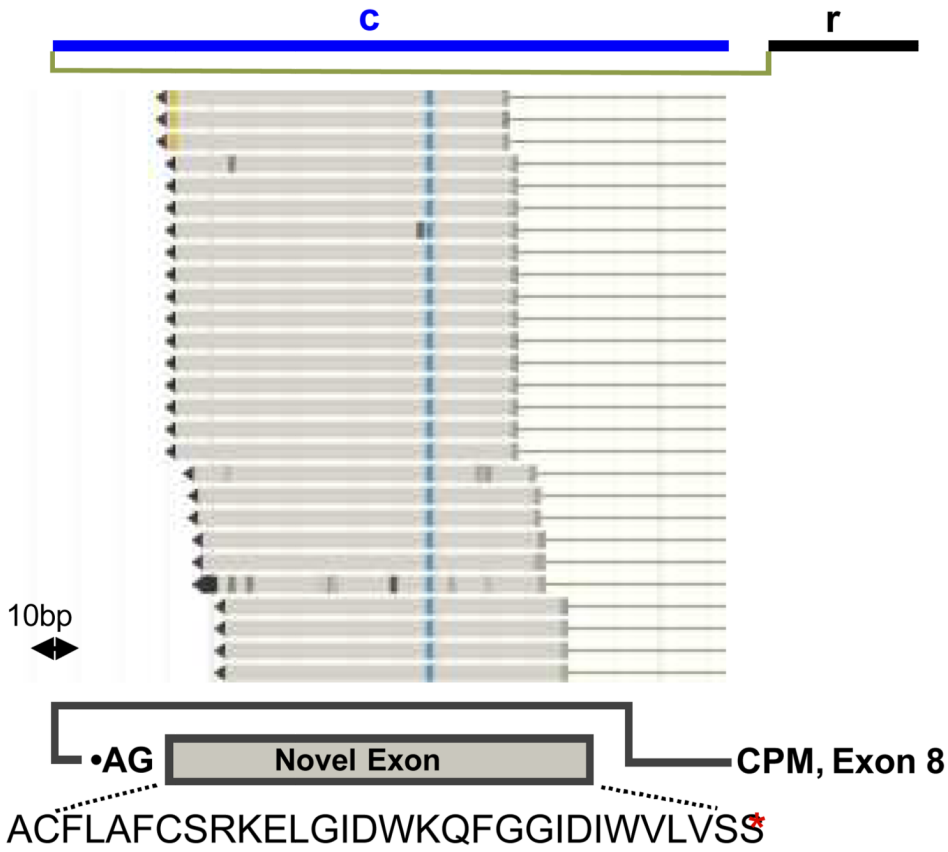

Figure 3. Novel CPM C-terminal exon expressed from the TCGA-06-0648 DM

Top line shows the appropriate region from (Fig. 1c) with sequencing reads where the other pair maps to exon 8 of CPM below. The bottom panel shows the orientation of the new exon relative to CPM and its amino acid sequence.

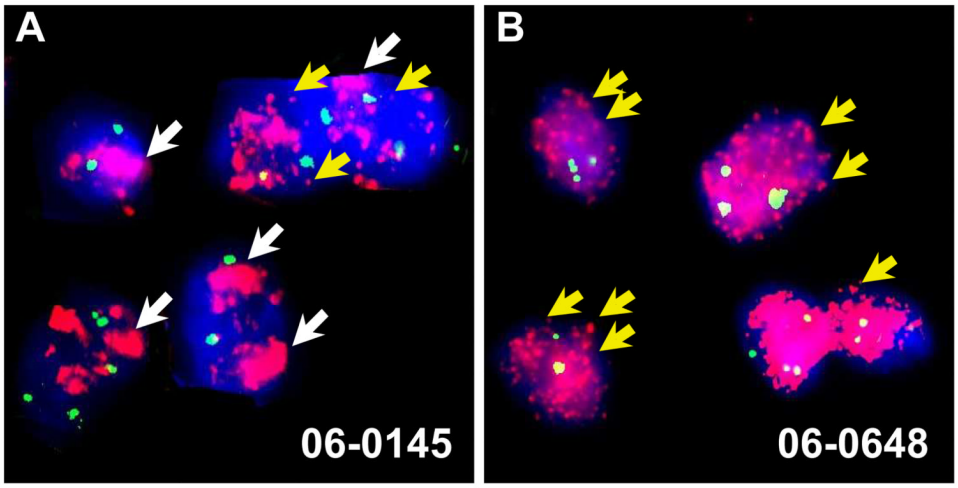

Figure 4. Visualization of GBM tumor oncogenic amplicons

FISH analysis of formalin-fixed paraffin embedded (FFPE) blocks matching samples 06-0145 (a) and 06-0648 (b). Probes for the centromeric region of chromosome 7 (a) and chromosome 12 (b) shown in green gave 2-4 foci per cell. Probes for EGFR (a) and MDM2 (b) are shown in pink. White arrows point to broad, HSR-like staining patterns, and yellow arrows point to discrete extrachromosomal spots.

Table 1

RNA-Seq-based expression estimates of 06-0648 DM-associated genes

	Expression[1]			
Sample	**MDM2**	**CPM (exons 1-8)**	**CAND1**	**RAP1B**
TCGA-GBM-02-0033	2849	1180	18717	22098
TCGA-GBM-06-0124	3504	662	15256	14654
TCGA-GBM-06-0126	3247	906	14808	10403
TCGA-GBM-06-0155	1955	911	17741	32690
TCGA-GBM-06-0214	5382	2783	15000	15429
TCGA-GBM-06-0216	3050	143	20966	14595
TCGA-GBM-06-0648	57126	9888	480414	8493
TCGA-GBM-06-0879	1235	498	12833	14488
TCGA-GBM-12-0692	909	224	14716	11022
TCGA-GBM-14-0786	2305	514	17484	13911
Avg[2]	2382	630	16565	16733
0648 Fold Increase	23.98	15.70	29.00	0.51

[1] Expression reported as normalized transcript coverage (coverage per transcript per million uniquely mapped paired-end reads)

[2] Excludes 06-0214 and 06-0648, which have MDM2 amplification

Table 2
Summary of amplicons found in TCGA-GBM exomes

Total Samples	264
Total Potential DM	61
Potential DM with *EGFR*	46
Potential DM with *CDK4*	18
Potential DM with *MDM2*	14