## Table of Contents

**Supplementary file 1 – Power for detection of somatic point mutations in coding exons of TARGET genes.**
**Supplementary file 2 – Detailed plots of SCNA calls in TARGET genes.**
**Supplementary file 3 – Detailed plots of the data used for phylogenetic inference in each case.**

## Sample Selection and Preparation

The study was reviewed and approved by the human subjects institutional review boards (IRBs) of the Dana-Farber Cancer Institute, Brigham and Women's Hospital, Broad Institute of Harvard and MIT, Massachusetts General Hospital, Seoul National University College of Medicine and Vall d'Hebron University Hospital. Written informed consent was obtained from all participants. Histologic diagnosis was re-confirmed on all samples by a board certified neuropathologist (S.S., A.S.R. and D.N.L) and representative fresh frozen or paraffin embedded blocks with estimated purity of ≥ 40% were selected. DNA was extracted from tissue shavings of frozen tissue or 3-5 1 mm core punch biopsies (Miltex, cat# 33-31AA-P/25) from FFPE tissue and from buffy coat preparations of paired blood using standard techniques (QIAGEN, Valencia CA). The DNA was then quantified using PicoGreen® dye (InVitrogen, Carlsbad CA). Mass spectrometric genotyping with a well-established 48-SNP panel was used to confirm the identity of tumor-normal pairs (Sequenom, San Diego CA).(1)

## Whole Exome Sequencing

Whole exome sequencing was performed using the platforms at the Broad Institute and the Center for Cancer Genome Discovery (CCGD)(2) as previously described. In brief, at CCGD, DNA was fragmented by sonication (Covaris Inc, Woburn, MA) to 150bp and further purified using Agencourt AMPure XP beads. 50 ng size selected DNA was then ligated to specific adaptors during library preparation (Illumina TruSeq, Illumina Inc, San Diego, CA). Each library was made with sample specific barcodes, quantified by QPCR (Kapa Biosystems, Inc, Woburn, MA) and 2 libraries were pooled to a total of 500 ng for Exome enrichment using the Agilent SureSelect hybrid capture kit (Whole Exome_v2 Agilent / Agilent 1.5, Santa Clara, CA). Several captures were pooled further and sequenced in one or more lanes to a final equivalent of 2 exomes per lane on a Hiseq 2500 (Illumina Inc, San Diego, CA). At the Broad Institute, libraries for whole exome (WE) sequencing were constructed and sequenced on either an Illumina HiSeq 2000 or Illumina GA-IIX using 76 bp paired-end reads. Details of whole exome library construction have been detailed elsewhere.(3, 4)

Standard quality control metrics, including error rates, percentage of passing filter reads, and total Gb produced, were used to characterize process performance before downstream analysis. The Illumina pipeline generates data files (BAM files) that contain the reads together with quality

parameters. Output from Illumina software was processed by the "Picard" data processing pipeline to yield BAM files containing aligned reads with well-calibrated quality scores.(3)

All data have been deposited in dbGaP: accession number **phs000730.v1.p1**

## Targeted validation sequencing

DNA from sample 0024-P and 0218-P was fragmented by sonication (Covaris Inc., Woburn, MA) to 250 bp, and DNA from sample 0013-P and 0244-P was fragmented to 150bp. Fragmented DNA was further purified using Agencourt AMPure XP beads. Size selected DNA was then ligated to specific adaptors during library preparation (Illumina TruSeq, Illumina Inc, San Diego, CA). Each library was made with sample specific barcodes, quantified by qPCR (Kapa Biosystems, Inc, Woburn, MA) and 2 libraries (0024-P and 0218-P; 0135-P and 0244-P) were pooled to a total of 500 ng for OncoPanel_v2 enrichment using the Agilent SureSelect hybrid capture kit (Agilent Technologies, Santa Clara, CA). Two captures were pooled further and sequenced in 2 lanes on a Hiseq 2500 in Rapid Run Mode (Illumina Inc, San Diego, CA).

## Immunohistochemistry

Immunohistochemical studies were performed on five-micrometer-thick sections of formalin-fixed, paraffin-embedded tissue in a Bond 3 automated immunostainer (Leica Microsystems, Bannockburn, IL, USA), and primary antibody against Her2/neu (4B5, 1:15, Ventana Medical System Inc.). Overexpression of Her2 was defined as positive membranous staining in more than 30% of the neoplastic cells. Partial and faint staining in less than 10% of tumor cells, weak to moderate complete membrane staining in more than 10% of the tumor cells, or intense or thick circumferential membrane staining in more than 30% of the tumor cells were scored as 1+ (negative), 2+ (equivocal), or 3+ (positive), respectively.

## Processing of raw whole-exome sequencing data

Read pairs were aligned to the hg19 reference sequence using the Burrows-Wheeler Aligner(5) and sample reads were de-multiplexed using Picard tools (http://picard.sourceforge.net). Data were sorted and duplicate-marked using Samtools (http://samtools.sourceforge.net) and Picard. Bias in base quality score assignments due to flowcell, lane, dinucleotide context, and machine cycle were analyzed and recalibrated, and local realignment around insertions/deletions was achieved using the Genome Analysis Toolkit (GATK) (http://www.broadinstitute.org/gatk/).(3, 6) All sample pairs passed the Firehose pipeline including a QC pipeline to test for any tumor/normal and inter-individual mix-ups by comparing insert-size distribution, copy-number profile, and SNP fingerprinting as described previously.(4) In addition, samples were assessed for DNA contamination having occurred during sample handling or library preparation using ContEst.(4) Somatic single-nulcleotide variants, insertions, and deletions were annotated using Oncotator (www.oncotator.org) which uses

information from publicly available databases, including the UCSC Genome Browser's UCSC Genes track(7), dbSNP build 132(8, 9), UCSC Genome Browser's ORegAnno track(9), UniProt release 2011_03(10), and COSMIC v51(11).

## Somatic mutation calling

Somatic point-mutations (single nucleotide variants and short insertions/deletions) were initially called using MuTect(12) and Strelka,(13) respectively, to compare each individual primary tumor or metastatic sample to the matched germline sample. Spurious calls caused by mismapping and other previously identified systematic errors were removed using an established list of known problematic sites.(2, 12, 14-16) A list of mutated sites seen in two or more of the normal samples was created and all somatic mutation calls at these sites were removed from analysis.

An additional filter was applied to exclude artifact mutations introduced by the processing of formalin-fixed, paraffin-embedded (FFPE) specimens, which are characterized by artificially induced C>T mutations in the context of a preceding C base (similar to the OxoG filter(17)). These artifact mutations are distinct from real CpG>TpG mutations, which dominate in most cancers. In brief, the FFPE filter consists of two steps. First, the filter estimates the component of total sequencing error rate due to FFPE artifacts in C>T by scanning all reference C (or G) sites counting sites sequenced as T (or A) in the two possible read pair orientations. Second, the orientation of each C>T (or G>A) mutation is compared to a model of balanced read pair orientation (binomial with p=0.5: no artifact) and a biased orientation characteristic of FFPE artifacts (binomial with p=0.96). The filter removes mutations consistent with the FFPE orientation bias to the degree where less than 1% of the surviving mutations in a given sample are consistent with FFPE artifacts.

To analyze somatic mutations across multiple sequenced tissue-samples from single patients, we utilized a tool designed to increase detection sensitivity for point mutations (single nucleotide variants and small insertions or deletions) called originally in one of the samples constituting the case. This method, termed 'forced calling', counted the number of reads supporting the reference or alternate allele at each called site in the BAM file for each sample. Reads were discarded if their base quality at the called site was < 20 or if their read-mapping quality was < 5, or if they were part of a duplicate read-pair overlapping the site. We then provisionally accepted mutation calls at this site if > 0 reads supporting the originally called allele were observed. This step ensured maximal sensitivity for mutations. In order not to increase false positives, further consideration of mutation call validity was made jointly in the phylogenetic inference, as described below.

## Calculation of total and allelic copy-numbers from whole exome sequencing data

Somatic total copy-ratios for each captured exon were calculated by comparison of the exon-average coverage with those obtained in a panel of normal samples. Read-depth at informative capture targets in tumor samples was calibrated to estimate copy-ratio using depths observed in the normal (non-cancer) diploid genomes. The resulting copy-ratio profiles were then segmented using the circular binary segmentation (CBS) algorithm(18) (**Fig. S8A,D**). Allelic copy-number analysis was

then performed by examination of alternate and reference read counts at heterozygous SNP positions (as determined by analysis of the matched normal sample). These counts were used to infer the fractional contribution of the two homologous chromosomes to the observed copy-ratio in each segment (**Fig. S8B,E**). Further analysis of change-points in these allelic-ratios was performed using PSCBS(19), refining the segmentation. These data were then input to ABSOLUTE(20), which was used to jointly estimate the fraction of cancer nuclei, average cancer genome ploidy, and absolute allelic copy-numbers (**Fig. S8F,G, Fig. S2**). An updated version of ABSOLUTE (v1.2) was used for all analyses. The major differences in the updated version are (i) improved identification of segmental regions with non-integer copy number (i.e., with different copy numbers in various cancer-cell populations contributing to the sequenced tissue sample); and (ii) improved estimation of CCF for point mutations occurring in regions of non-integer copy number (described below).

## Somatic copy number alteration calling

Somatic copy-number alterations (SCNAs) were called using ABSOLUTE(20) to correct observed copy-ratios for variations in sample purity and ploidy, yielding *rescaled copy-numbers* for each genomic segment (i.e. linear transformations of the observed copy-ratio to units of absolute copy-number) . We classified segments into four categories: (i) no alteration; (ii) homozygous deletion (referred to as deletion hereafter); (iii) amplification; and (iv) high-level amplification. To assist in this classification, we calculated, for each genomic segment, its *focality* – the fraction of that sample's genome with lower copy number (for amplified regions) or higher copy number (for deletions). Segmental SCNAs were then projected onto genes by intersecting each segment with gene footprints defined by GENCODE (v19)

SCNAs with greater focality had more evidence specifically nominating them as driver events in a given sample. Segments were considered *deleted* if their rescaled copy-number (*rCN*) was < 0.25 and their focality was > 0.995 (**Fig. S8**). To call amplifications, we used a threshold that was a linear function of both the *rCN* and focality, so that events with higher focality required lower rCN values to be called (**Fig. S20**). Segments were called *amplified* if their focality was > $0.98 - 0.2 \times \log_2 (rCN / 5)$ (**Fig. S9, S12-18**)

. In order to be more conservative in calling amplifications in a subset of related cancer-tissue samples, we lowered the thresholds for calling amplification in samples for which any related sample was called using the above criterion. In these cases, we required focality > $0.88 - 0.2 \times \log_2 (rCN / 5)$. This was done to ensure that we did not designate an amplification as unique to a given sample when it was present in a related sample at a slightly lower level. Calling of *high-level amplification* events was performed in a similar manner, requiring focality > $0.98 – (1/7) \times \log_2 (rCN / 7)$.

All potentially clinically informative (TARGET) SCNA calls were manually reviewed (**Supplementary file 2**). A small number of calls were revised to correct segmentation errors, most of which resulted in false-negative SCNA calls in a subset of related cancer samples (these tended to occur when the number of affected exome-capture targets was very small or when the raw copy-ratio signal was very noisy). The SCNA calls which were manually altered were as follows: in patient 308 a deletion of *STK11* was changed to shared; in patient 405, a deletion of *CDKN2A* was changed to

shared; in patient 199, a deletion of *SMARCA4* was changed to shared; in patient 26, a deletion of *RB1* was changed to shared; in patient 98, amplifications of *KIT* and *MYC* were changed to shared; in patient 137 an amplification of *MET* was changed to shared; in patient 296 an amplification of *EGFR* was changed from shared to primary only.

## Calculation of statistical power for somatic mutation detection

To calculate power, we considered, for each targeted region, the expected variant allele fraction (VAF) of a point mutation with CCF=1 and multiplicity=1, given the sample purity and local copy number. Detection power was calculated as the probability to observe at least the number of reads supporting an alternate allele required by MuTect(12) in order to call the mutation. This was done given the expected VAF and observed coverage depth and assuming the estimates of the sequencing error rate and desired FDR used by MuTect(12), as previously described(20). These parameters typically imply that 3 reads supporting a given non-reference allele are required to call a mutation (e.g. given typical sequencing depth at that site). These calculations were used to calculate *unpaired* power (disregarding information from any other sequenced cancer-tissue samples available from the same patient; **Fig. S19B,E**). Unpaired detection power in each targeted exon of all TARGET genes nominated by mutation is shown for the primary and metastasis samples of each case (as in **Fig. S19E**; **Supplementary file 1.**)

For mutations detected in only a subset of the tissue-samples comprising a given case, we calculated the *paired*-detection power. This calculation assumed a multiplicity of 1 and a CCF=1 in all samples harboring the mutation (the latter as implied by the branched-sibling model; **Fig. S1**), and took into account the forced-calling procedure used for related tissue-samples, whereby candidate mutations were called with a single supporting read matching the called allele. For clinically informative (TARGET) point-mutations (**Fig. 2**) detected exclusively in either a primary tumor or matched brain-metastasis sample, we assumed that the mutations were present in both samples if the mutant site had paired-detection power less than 0.99 in the sample in which it was not detected (**Fig. S11**). This affected 4 of 19 total TARGET mutations detected exclusively in one sample (**Fig. 2, S10**).

## Calculation of point-mutation CCF distributions

For each somatic mutation, we estimated the fraction of cancer cells that harbors the mutation, i.e. its *cancer cell fraction* (or CCF), represented as a distribution over the possible CCF values, between 0 and 1(20, 21). A CCF value of 1 implies mutations are present in 100% of cancer cells of that biopsy. A CCF value of <1 implies that subclonal mutations are present in only a subset of the cancer cells that were sampled. Probability distributions over CCF were computed for each by correcting mutant and reference read fractions for sample purity and local copy-number(20), as previously described(21).

Additional modifications to this procedure were made in order to better account for cases where the underlying local copy-number was non-integer (ABSOLUTE v1.2). This was accomplished

by first estimating, for each such region, ancestral and derived copy-numbers and the corresponding CCF of the SCNA. This was performed in a manner similar to that previously described(21), except that the procedure for choosing ancestral and derived copy-numbers was modified to take into account genome doubling, as follows: SCNAs below the modal integer copy-number in the given sample were assumed to have derived copy numbers of one less than the ancestral copy-number, and vice-versa for SCNAs above the modal copy-number. We then considered four possible scenarios relating the mutation to the ancestral and derived copy-numbers (regarding which cell-populations harbored them and whether they involved the same DNA molecules or not): (i) The mutation occurred in an ancestral population with the derived SCNA in *trans*; (ii) The mutation occurred in an ancestral population with the derived SCNA in *cis*; (iii) Both the mutation and SCNA occurred in derived cell-populations, with the population harboring the mutation potentially nested inside that harboring the SCNA; and (iv) the mutation and SCNA occurred in sibling cell-populations. We integrated out the model parameters corresponding to each of these four scenarios and performed Bayesian model averaging by computing the average of CCF distributions implied by each model, weighted by the model evidence. We note that this led to multi-modal CCF distributions for some mutations.

To reduce uncertainty in the CCF distribution of individual mutations, we assumed that each cancer-tissue sample contained a small (but unknown) number of distinct cancer-cell populations, within which all mutations share the same CCF (enabling more accurate estimation of the population's CCF). We therefore applied a Bayesian clustering method (which jointly estimated the CCF values and the number of populations) to the set of CCF distributions corresponding to mutations detected in each individual sample. This was accomplished by sampling from a mixture of Dirichlet processes(22) using a Markov chain Monte Carlo (MCMC) sampler, as previously described(21, 23). We used 250 MCMC iterations with the 125 initial ones discarded as 'burn-in'. A prior over the number of mutation clusters in a given sample was specified using a negative binomial distribution ($r$=10, $mu$=3); these values favored 1-5 clusters (**Fig. S3D**). A hard partition of mutations was obtained by counting the number of times each pair of mutations were assigned to the same cluster over the MCMC iterations (after convergence). A distance metric was created by taking the inverse of each pair-count, and hierarchical clustering was performed using complete linkage. The resulting tree was then cut into $k$ clusters, where $k$ was chosen as the lowest number of sampled clusters during the MCMC simulation (**Fig. S3C,D**).

## Supplementary figure legends

### Figure S1. Branched evolution leads to tissue-sampling bias in primary tumor samples.

**A**. Plot of cancer-cell fraction (CCF) values for somatic mutations in two samples; a primary tumor sample ($x$-axis) vs. a patient-matched brain metastasis sample ($y$-axis). CCF values denote the fraction of cancer cells bearing a mutation in a cancer-tissue sample. Colored circles denote clusters of mutations with similar (x,y) CCF values with the number of mutations indicated. For example, the gray circle in the upper right represents 261 mutations present in all of the cancer cells from the

primary and the metastatic sample (CCF = 1.0 in both samples). The blue cluster of 18 mutations is present in 35-50% of cancer cells sampled in the primary tumor sample, but none from the brain metastasis sample. Cells bearing these mutations must be descended from cells bearing the 14 light-blue mutations present in all cancer cells sampled from the primary tumor but none from the brain metastasis (lower right). Similar reasoning applies to the mutations detected exclusively in the brain metastasis (red and dark-red clusters).
**B**. Phylogenetic tree constructed from the data in **A**. Colored circles represent cell populations, with colored arrows (branches) indicating the inheritance of mutations from ancestral populations (gray circle reflects the inferred common ancestor). Branch lengths are proportional to the number of mutations assigned to that branch, indicated by the numbers. Branch thickness corresponds to the CCF of mutations on each branch.

**C**. Representation of tumor evolution and tissue-sampling scenario consistent with all related pairs of primary and metastatic cancers analyzed. Vertical lines indicate the sampled cancer tissue from the primary tumor and metastasis. The *y*-axis represents cell number. The light blue section denotes a subclone of the primary tumor with a nested subclone shown in dark blue; all of the cancer cells of the primary-tumor sample are derived from these subclones. The purple section denotes a subclone of the primary tumor containing an ancestor leading to (red arrow) the brain metastasis (red section), possibly via multiple intermediate sites (not shown). This ancestor subclone was not represented in the primary-tumor sample analyzed. The dark-red section denotes a subclone of the brain metastasis. This scenario implies that genetic characterization of primary tumor samples will result in tissue-sampling bias, whereby mutations in the brain metastasis are not present in the primary tumor sample, and vice-versa.

### Figure S2: Results of ABSOLUTE on samples from patient 418

**A**. Allelic segmental copy-ratios (*y*-axis) are shown for each pair of homologous chromosomes across the entire genome (*x*-axis). The height of each segment indicates the 95% confidence interval for the copy-ratio. Dotted horizontal lines correspond to modeled absolute allelic copy-numbers, as indicated on the right-hand y-axis of the adjacent summary histogram (at right). Segments are colored as blue if the copy-ratios are consistent with all cancer cells contributing to the sample having the same copy-number in that region (CCF=1), or pink if they appear to derive from a heterogeneous mixture of copy-numbers (CCF<1).
**B**. The summation of probability distributions (*y*-axis) over the fraction of alternate reads (VAF; *x*-axis) for the somatic mutations detected in the sample is shown as filled transparent curves for mutations with CCF=1.0 (green) and with CCF < 1.0 (pink). The dotted vertical line indicates the inferred sample purity divided by 2. For samples with fewer than 500 mutations detected, the distributions for individual mutations are plotted using solid lines.
**C**. Distributions over mutation VAF are rescaled using ABSOLUTE to correct for estimated sample purity and local copy-number (displayed as in **B**). This yields distributions over *multiplicity* (*x*-axis), the expected number of mutant alleles present per cell in the sampled cancer-cell population. Mutations with multiplicity < 1 are restricted to a subset of this population (CCF < 1.0). Mutations with multiplicity = 2 likely occurred by whole genome doubling, or by chromosomal gain including the mutant allele.

Similar plots for all 86 cases are available in **Supplementary file 3**.

### Figure S3: 2D Bayesian clustering analysis of point-mutation CCF distributions in case 418

Two-dimensional cancer-cell fraction (CCF) probability distributions for each unique pair of samples from patient 418 are shown for mutations detected in each pair. The plots on the lower-left (lower diagonal matrix) show the result of the 2D Bayesian clustering algorithm applied to the input CCF distributions, shown in the upper-triangular matrix (each plot in the lower-triangle corresponds to its transposed coordinates in the upper-triangle; note that tissue-sample labels refer only to plots in the lower-triangle.) The lower-left plots show the result of the 2D Bayesian clustering algorithm applied to the input CCF distributions in the upper-right. CCF distributions for each mutation are displayed as transparent filled-contours corresponding to density equal to scales of 0.99, 0.95, 0.8, 0.5, and 0.1 times each density at the mode. The degree of transparency is inversely proportional to each density at the mode (i.e., mutations with more uncertain CCF have greater transparency). Mutations are colored according to the expected value of their post-clustering CCF density, as follows: grey – CCF ≥0.8 in both samples; green – CCF ≥ 0.8 in one sample and uncalled in the other; pink – CCF < 0.8 in one sample and uncalled in the other; orange: CCF < 0.8 one sample and called in the other. For clarity, only mutations covered by ≥ 20 sequencing reads are displayed. We note that, for the purposes of this plot, small numbers of reads supporting the alternate allele for each mutation were discarded if they were insufficient to result in a mutation call using MuTect (i.e., mutations were not force-called).

Similar plots for all 86 cases are available in **Supplementary file 3**.

### Figure S4: Bayesian clustering of private point-mutation CCF distributions in all sequenced tissue-samples from case 418

**A.** Pre-clustering probability distributions over CCF (from ABSOLUTE) are plotted for each mutation as transparent filled curves with color determined by hard-cluster assignment (shown in **C**).
**B.** Post-clustering probability distributions over CCF are shown for each mutation.
**C.** Posterior densities of the hard-clusters.
**D**. Prior and posterior distributions over *k* (cluster number) are shown, as well as the histogram of sampled *k* values from the MCMC run (after discarding the burn-in samples).

Similar plots for all 86 cases are available in **Supplementary file 3**.

### Figure S5: Genetic alterations supporting phylogeny construction in case 418

**A.** The matrix of all point mutations (SSNVs and indels; columns) detected in any of the tissue samples from this case (rows) is shown. The mutations are colored according to the legend in **C**, with the exception of undetected mutations, which are colored according to a linear color-scale indicating paired-detection power ranging from black (power=0) to white (power > 0.99). The color bar on the top of the matrix indicates the assignment of mutations to branches of the phylogenetic tree (shown in **Fig. S10**). A single mutation on the far right (under light-grey color bar) was not assigned to a branch of the tree.
**B.** The data from **A** is shown after removing mutations rejected by the automatic phylogeny inference procedure (i.e., mutations present in > 1 sample with any CCF value < 1.0 and fewer than 3 supporting reads).
**C.** Stacked bar-plot of the data in **A**, summarized over each of the tissue samples from this case.

**D.** The matrix of all SCNAs, called at the gene level, (columns) detected in any of the tissue samples from this case (rows) is shown. The color bar at the top indicates phylogenetic branch assignment, as in **A**.

Similar plots for all 86 cases are available in **Supplementary file 3**.

### Figure S6: Phylogenetic tree for case 418

The phylogenetic tree obtained by maximum parsimony on the matrix of binary mutation presence / absence data (**Fig. S5A,B**) is shown. The thickness of each line corresponds to the CCF of the mutations on that branch. For each tissue sample, the longest sub-branch is labeled with the name of the tissue sample.

Similar plots for all 86 cases are available in **Supplementary file 3**.

### Figure S7. Evolutionary relationships between primary tumor-samples and brain metastasis samples

Evolutionary and tissue sampling scenarios are shown (right), corresponding to various configurations of observed mutation CCF clusters (left).
**A**. Branched-sibling, **B**. Clonally unrelated. **C**. Ancestor.

### Figure S8. Detection of homozygous deletion in *CDKN2A* in the brain metastasis of case 24

Plots on the left refer to the primary-tumor sample; plots on the right refer to the matched brain-metastasis sample. **A-C** A 100kb genomic region (*x*-axis) around *CDKN2A* is shown.
**A.** Copy-ratios (*y*-axis) for all targeted exons (points) in the region are shown with color alternating between light and dark grey for each segment. Dashed horizontal lines indicate the segmental copy-ratios. The right axis indicates the absolute copy-numbers, which are calculated by adjusting for the sample's purity and ploidy using ABSOLUTE. Text on the top of the plot gives the rescaled copy-number and focality score (fraction of the genome with lower copy-number) for each segment.
**B.** Tumor variant allelic fractions (VAF; *y*-axis) are shown for each SNP called heterozygous in the paired normal-sample's exome. For each SNP, the phase is indicated by color (red: major-copy homologous chromosome, blue: minor-copy homologous chromosome, purple: allelic balance, no phase information). Vertical lines through each point correspond to 95% confidence intervals around the VAF of each SNP. Grey horizontal lines indicate the inferred fraction of the minor-copy homologous chromosome in the cancer sample (bottom line), and of that of the major-copy (top line). Points plotted as black open diamonds correspond to outliers.
**C.** Transcript models from refgene are shown. Vertical blue lines indicate exons.
**D**., **E**. Data are shown as in **A** and **B**, except that the *x*-axis spans the entire chromosome. *CDKN2A* was called as homozygously deleted in the brain metastasis (right); genomic markers in the gene footprint are colored in blue.
**F**. Total segmental copy-ratios (*y*-axis) are shown across the entire genome (*x*-axis). Dotted horizontal lines correspond to modeled absolute copy-numbers, as indicated on the right-hand axis of the adjacent summary histogram (right).
**G**. Allelic segmental copy-ratios (*y*-axis) are shown for each pair of homologous chromosomes across the entire genome (*x*-axis). The height of each segment indicates the 95% confidence interval for the copy-ratio. Dotted horizontal lines correspond to modeled absolute allelic copy-numbers, as indicated on the right-hand axis of the adjacent summary histogram (right). Segments are colored as blue if the copy-ratios are consistent with all tumor cells contributing to the sample having the same copy-number in that region, or pink if they appear to derive from a heterogeneous mixture of copy-numbers.

Similar plots for all TARGET genes nominated by amplification or deletion where at least one sample was called are available as **Supplementary file 2**.

### Figure S9. Amplification of *FGFR1* and *MYC* detected in the brain metastasis of case 331

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 8 in all sequenced samples from case 331. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S10. Additional alterations under investigation for association with various targeted therapies

**A-H**. Additional alterations in TARGET genes under investigation for association with the indicated therapies. Mutations are plotted as in **Fig. 2**.
Alterations which may be associated with sensitivity to
**A.** Ephrin inhibitors(24, 25)
**B**. Epigenetic therapy(26) including histone deacetylase inhibitors
**C**. Notch inhibitors(24, 27)
**D**. WNT inhibitors(24)
**E**. AURKA inhibitors
**F**. Multitargeted TKIs including sorafenib, sunitinib and dasatinib
**G**. MDM inhibitors
**H**. PARP inhibitors
**I**. Alterations in TARGET which may be diagnostic / prognostic in multiple tumor types(5, 20, 28, 29)

### Figure S11. Power for paired-detection of somatic mutations

For somatic point-mutations (single-nucleotide variants or small indels), detected in one sample of a pair, the cumulative fraction (*y*-axis) of paired-detection power in the other sample (*x*-axis) is shown. For mutations detected in both samples (grey), the lower of the paired-detection power values is shown. The dashed vertical line corresponds to power = 0.99. The left-hand plot refers to all mutations detected, whereas the right-hand plot refers to only actionable mutations in TARGET genes. Only four such mutations were detected exclusively in one sample (the brain metastasis) and underpowered at the 0.99 level in their paired primary-tumor sample. These mutations were therefore considered to be shared in both samples and their lack of detection power in the primary sample is indicated in **Fig. 2**. The four mutations were: (i) *KRAS* p.G12C, case 114, power = 0.90 (we note that in this case the samples were clonally unrelated and the primary sample harbored a different activating mutation in *KRAS*); (ii) *RB1* p.D701E, case 201, power = 0.77; (iii) *STK11* p.GP279fs, case 321, power = 0.84; and (iv) *SMAD4* p.G352E, case 132, power = 0.96.

### Figure S12. Amplification of *CCNE1* detected in the brain metastasis of case 314.

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 19 in all sequenced samples from case 314. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S13. Amplification of *EGFR* detected in the brain metastasis of case 314

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 7 in all sequenced samples from case 314. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S14. Amplification of *MYC* detected in the brain metastasis of case 308

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 8 in all sequenced samples from case 308. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S15. Amplification of *MYC* detected in the brain metastasis of case 138

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 8 in all sequenced samples from case 138. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S16. Amplifications of *CDK6* and *MET* detected in the brain metastasis of case 138

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 7 in all sequenced samples from case 138. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S17. Amplifications of *CCNE1* and *AKT2* detected in the brain metastasis of case 138

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 19 in all sequenced samples from case 138. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S18. Amplification of *EGFR* detected in a regional lymph node from case 296

Raw total copy-ratio data and segmentation results are displayed as in **Fig. S8A** for chromosome 7 in all sequenced samples from case 296. Genes called amplified are labeled in red, with the exon-capture targets intersecting the gene footprint colored in red.

### Figure S19. Power for somatic mutation detection in 86 matched primary-tumor and brain-metastasis samples.

**A**. Stacked bar-chart showing the somatic mutation density in each case, colored by phylogenetic branch (grey: shared, blue: primary, red: metastasis, blue: primary CCF < 1, dark red: metastasis CCF < 1.
**B**. Bar chart showing the fraction of targeted bases in the exome for which the unpaired detection-power was at least 0.8. The primary tumor and metastasis samples are denoted with transparent blue and red, respectively. Transparent bar-charts for the primary and metastatic samples are plotted on top of one another.
**C**. Bar chart showing the average mean-coverage over targeted regions for primary and metastatic samples, with colors as in **B.**
**D**. Bar chart showing the inferred sample purity (fraction of cancer nuclei) for primary and metastatic samples, with colors as in **B**. Purity was estimated using ABSOLUTE.
**E**. Detection power in each targeted exon of *PIK3CA* (rows) in the primary and metastatic samples of each case (columns). Each cell in the matrix shows the (unpaired) detection power in the primary sample (upper triangle) and metastatic sample (lower triangle). Power is shown using a linear color scale from white (power = 0) to blue (power > 0.99) for the primary samples and red for the metastatic

samples. Case numbers are shown on the bottom of the matrix, and refer to Panels A-E. Similar plots for all TARGET genes nominated by mutation are available in **Supplementary file 1**.

**Figure S20. Calling of amplifications in primary-tumor samples and paired brain metastases.**
Each plot shows the log2 rescaled copy-number vs. focality for the indicated gene in each paired primary and metastatic sample, which are connected by lines. Rescaled copy-numbers are calculated by a linear transformation of copy ratio adjusting for sample purity and ploidy as determined by ABSOLUTE. Focality refers to the fraction of a sample's genome which is at a lower copy-number than the given gene. Diagonal dotted lines correspond to thresholds for calling genes amplified (center line) or high-level amplified (top-right line). For a given pair of samples, if one sample falls above either of these lines, then the paired sample will receive the same call if it is above the adjacent line to the lower-left. The color of each line indicates the amplification call on that sample pair (as indicated in the legend). Colored circles at the endpoints of lines indicate the type of amplification call made in that sample (as indicated in the legend).

## References

1. Demichelis F, Greulich H, Macoska JA, Beroukhim R, Sellers WR, Garraway L, et al. SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines. Nucleic acids research. 2008;36:2446-56.
2. Brastianos PK, Horowitz PM, Santagata S, Jones RT, McKenna A, Getz G, et al. Genomic sequencing of meningiomas identifies oncogenic SMO and AKT1 mutations. Nature genetics. 2013;45:285-9.
3. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011;27:2601-2.
4. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011;470:214-20.
5. Blackford A, Serrano OK, Wolfgang CL, Parmigiani G, Jones S, Zhang X, et al. SMAD4 gene mutations are associated with poor prognosis in pancreatic cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2009;15:4674-9.
6. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297-303.
7. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, et al. The UCSC genome browser database: update 2011. Nucleic acids research. 2010:gkq963.
8. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29:308-11.
9. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, et al. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic acids research. 2008;36:D107-D13.
10. Consortium U. Ongoing and future developments at the Universal Protein Resource. Nucleic acids research. 2011;39:D214-D9.
11. Shepherd R, Forbes SA, Beare D, Bamford S, Cole CG, Ward S, et al. Data mining using the catalogue of somatic mutations in cancer BioMart. Database. 2011;2011:bar018.

12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013;31:213-9.
13. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811-7.
14. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, et al. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011;471:467-72.
15. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2012;109:3879-84.
16. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011;333:1157-60.
17. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic acids research. 2013:gks1443.
18. Venkatraman E, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007;23:657-63.
19. Olshen AB, Bengtsson H, Neuvial P, Spellman PT, Olshen RA, Seshan VE. Parent-specific copy number in paired tumor–normal studies using circular binary segmentation. Bioinformatics. 2011;27:2038-46.
20. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology. 2012.
21. Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. Leukemia. 2013.
22. Escobar MD, West M. Bayesian density estimation and inference using mixtures. Journal of the american statistical association. 1995;90:577-88.
23. Lohr JG, Stojanov P, Carter SL, Cruz-Gordillo P, Lawrence MS, Auclair D, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. Cancer cell. 2014;25:91-101.
24. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nature medicine. 2014;20:682-8.
25. Day BW, Stringer BW, Al-Ejeh F, Ting MJ, Wilson J, Ensbey KS, et al. EphA3 maintains tumorigenicity and is a therapeutic target in glioblastoma multiforme. Cancer cell. 2013;23:238-48.
26. Kikuchi J, Takashina T, Kinoshita I, Kikuchi E, Shimizu Y, Sakakibara-Konishi J, et al. Epigenetic therapy with 3-deazaneplanocin A, an inhibitor of the histone methyltransferase EZH2, inhibits growth of non-small cell lung cancer cells. Lung Cancer. 2012;78:138-43.
27. Stoeck A, Lejnine S, Truong A, Pan L, Wang H, Zang C, et al. Discovery of biomarkers predictive of GSI response in triple-negative breast cancer and adenoid cystic carcinoma. Cancer discovery. 2014;4:1154-67.
28. Powell B, Soong R, Iacopetta B, Seshadri R, Smith DR. Prognostic significance of mutations to different structural and functional regions of the p53 gene in breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2000;6:443-51.
29. Johnson BE, Ihde DC, Makuch RW, Gazdar AF, Carney DN, Oie H, et al. myc family oncogene amplification in tumor cell lines established from small cell lung cancer patients and its relationship to clinical status and course. The Journal of clinical investigation. 1987;79:1629-34.

Figure 1

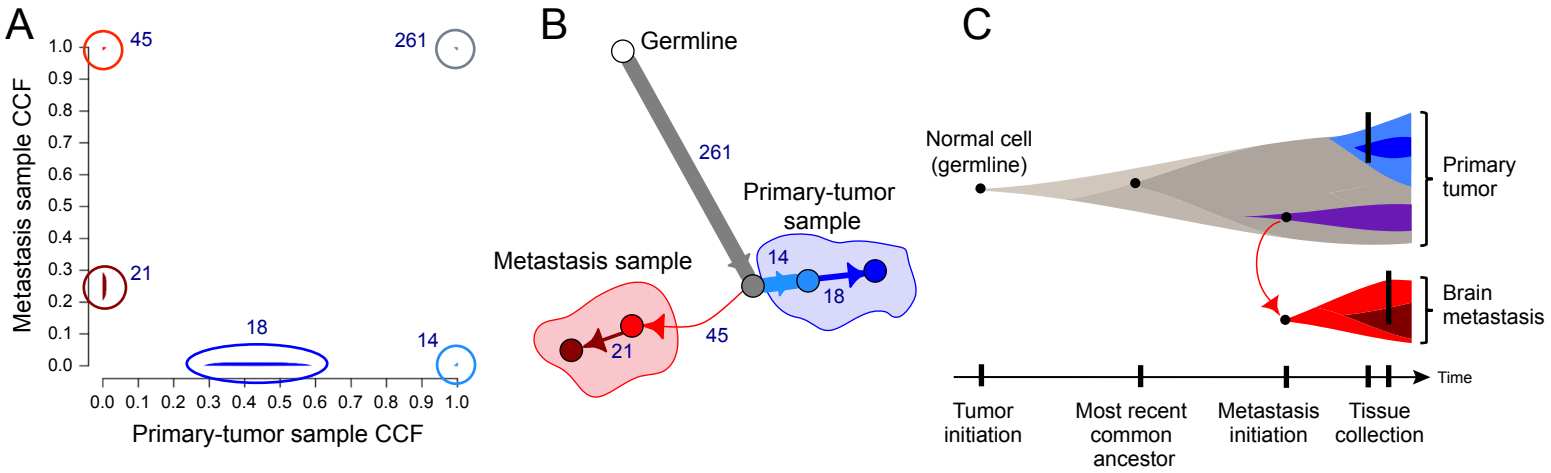



**Figure S1. Branched evolution leads to tissue-sampling bias in primary tumor samples**

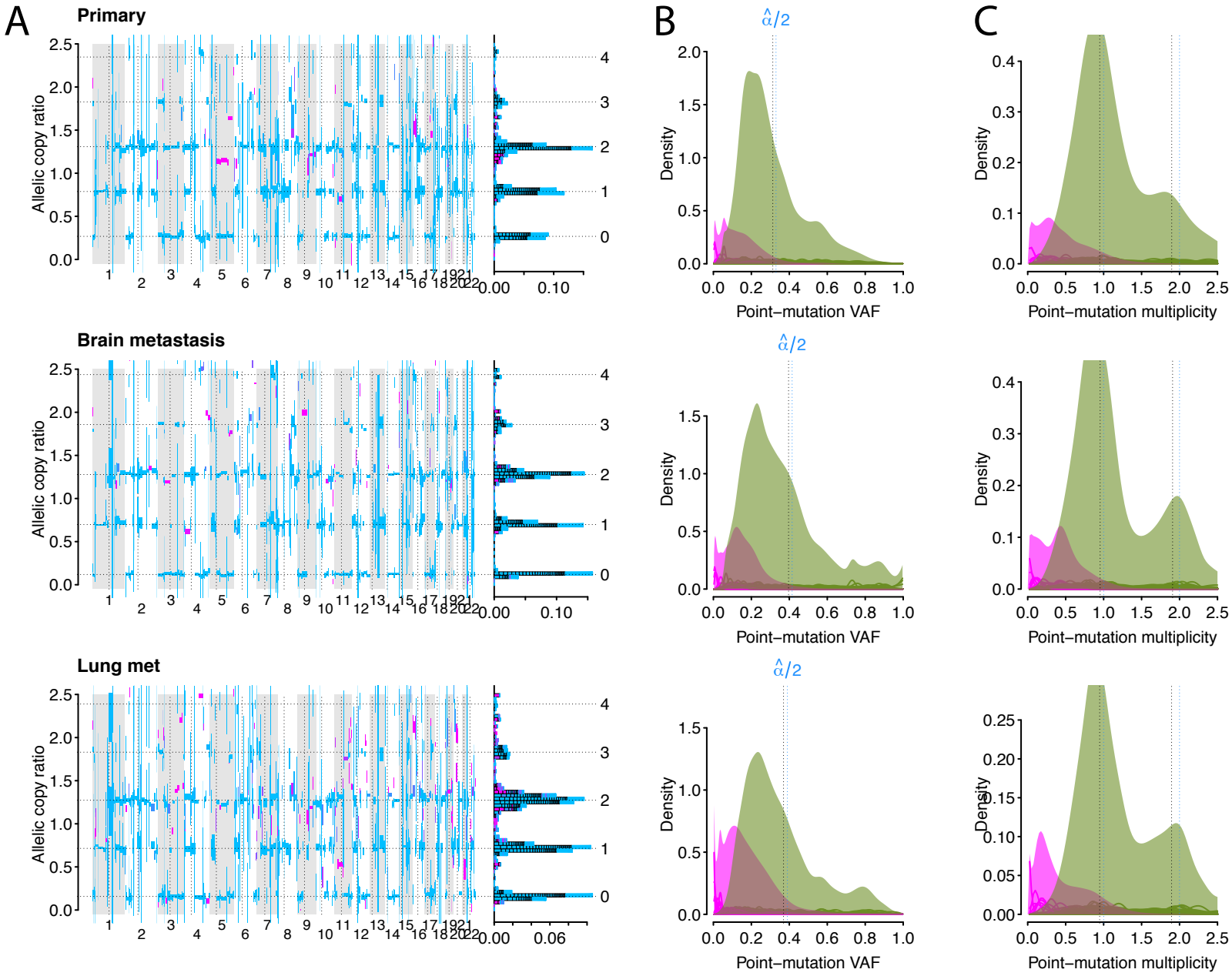



**Figure S2. Results of ABSOLUTE on samples from patient 418**

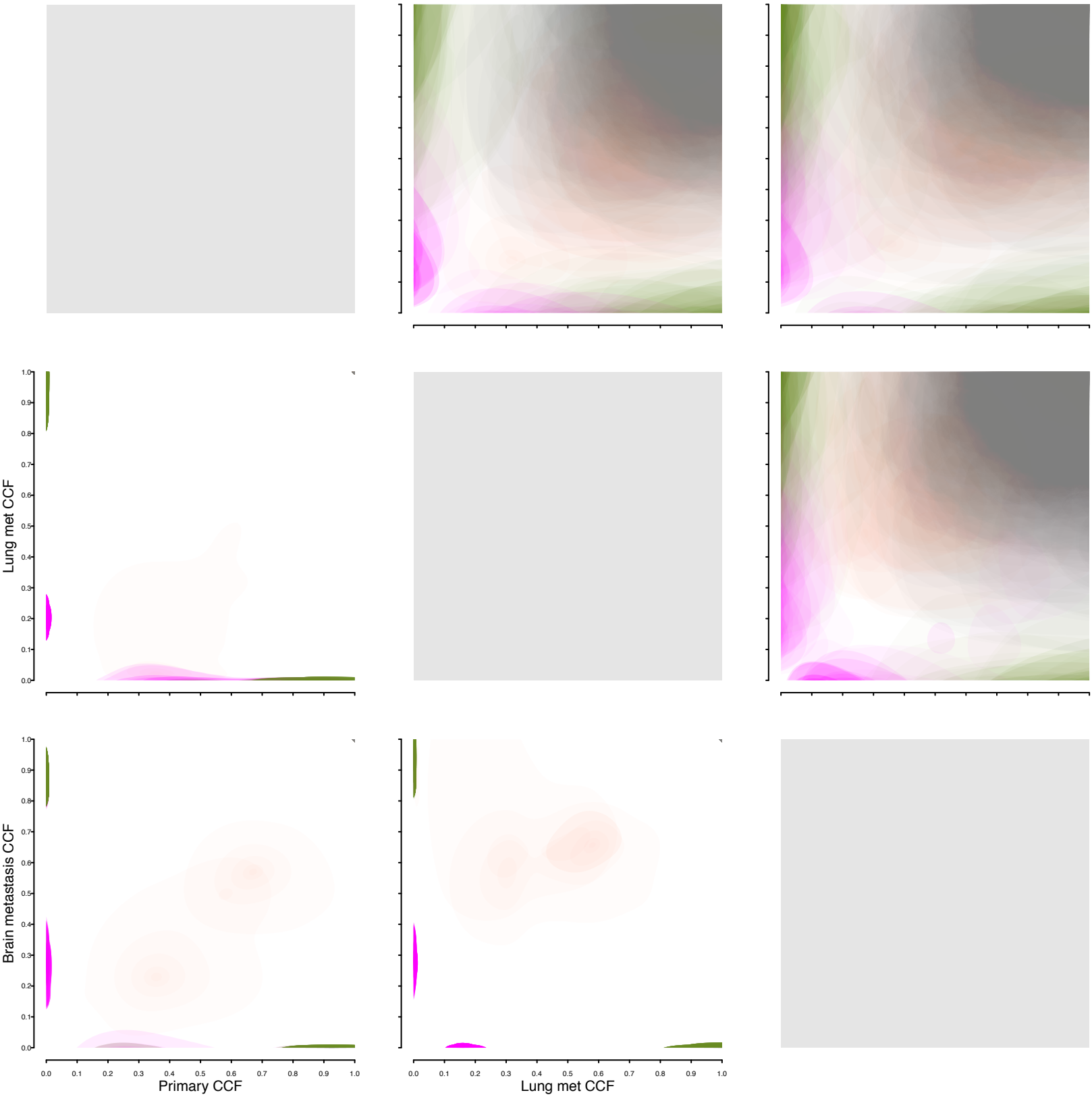



**Figure S3. 2D Bayesian clustering analysis of point-mutation CCF distributions in case 418**

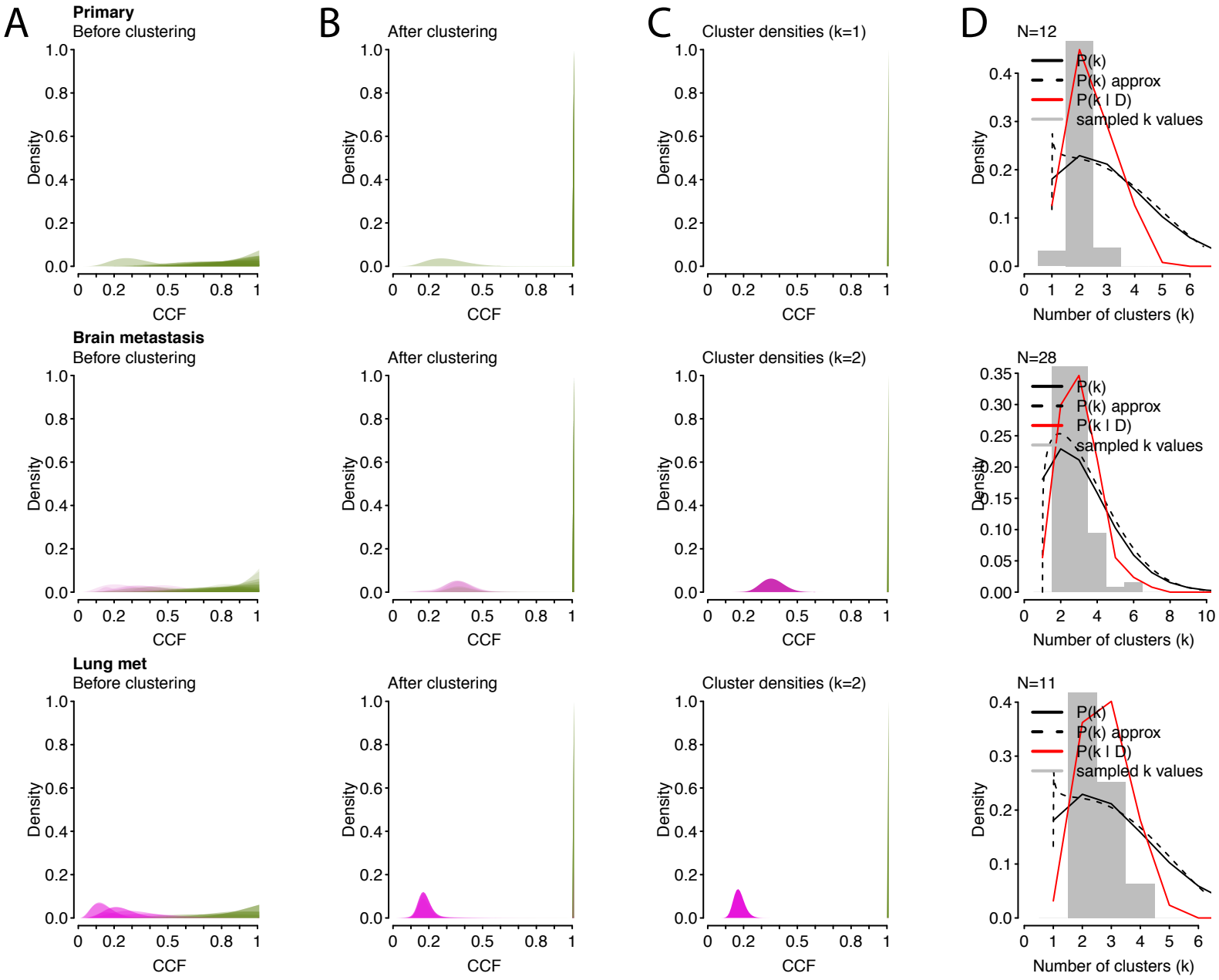



**Figure S4. Bayesian clustering of private point-mutation CCF distributions in all sequenced tissue- samples from case 418**

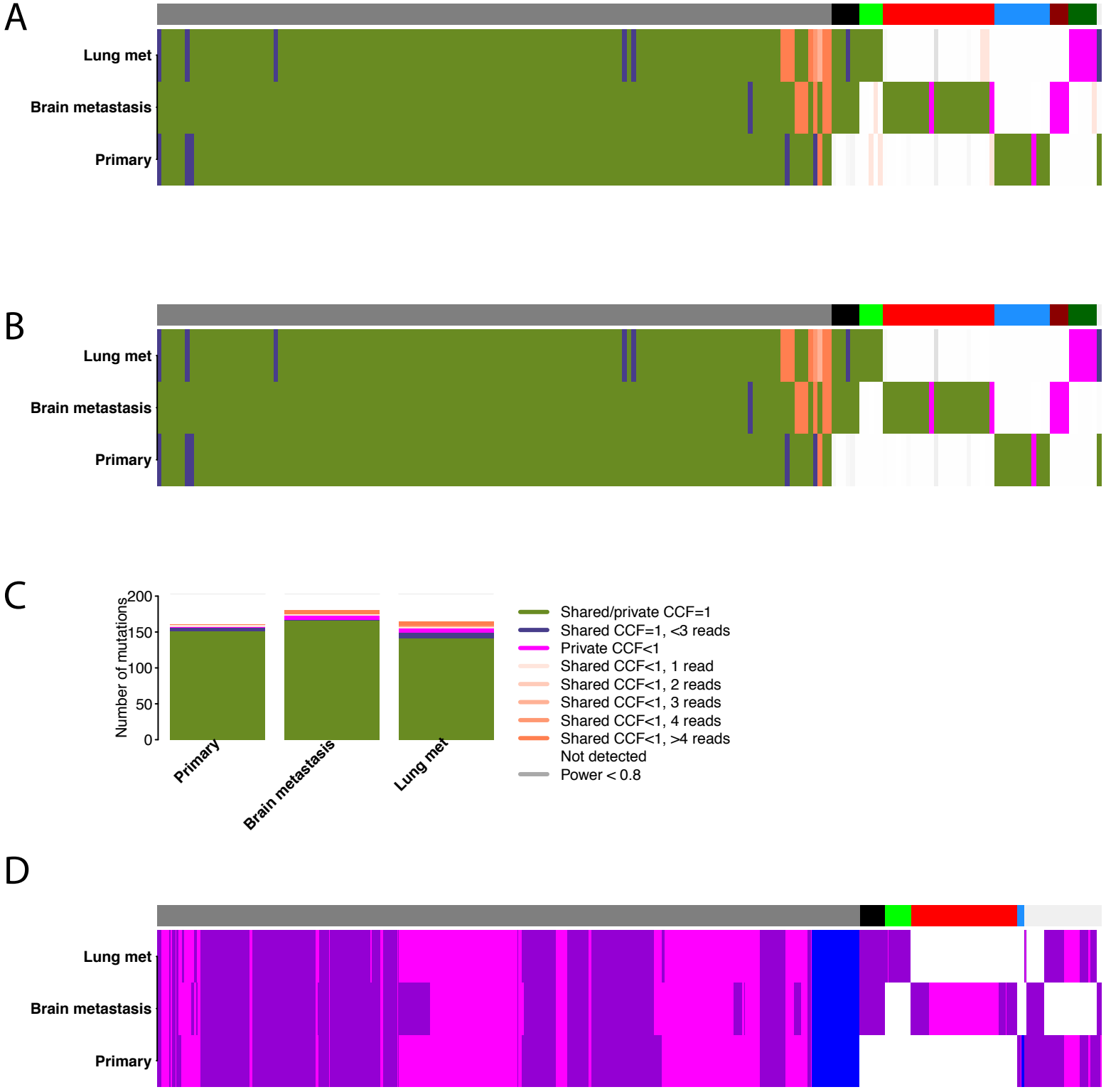



**Figure S5. Genetic alterations supporting phylogeny construction in case 418**

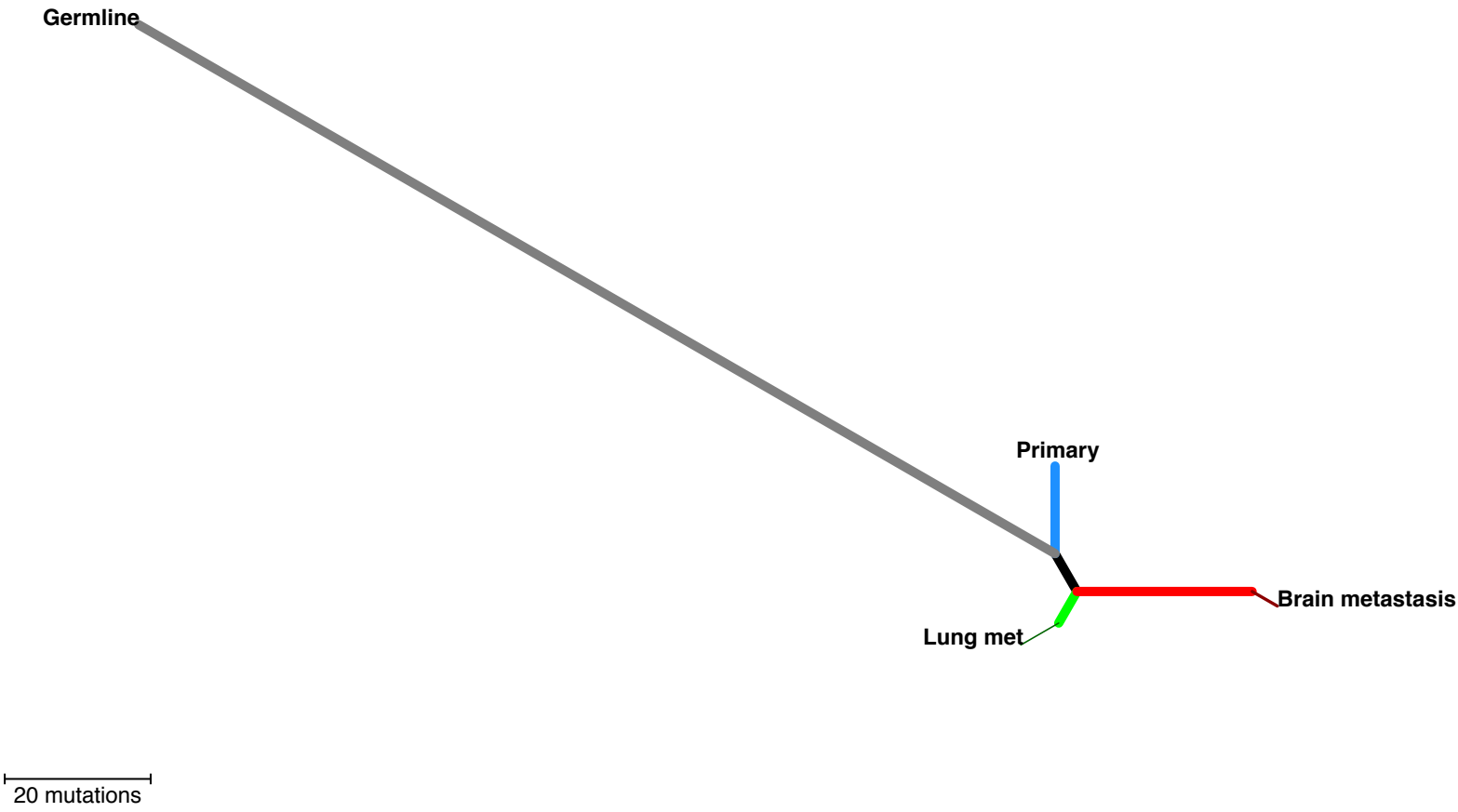



**Figure S6. Phylogenetic tree for case 418**

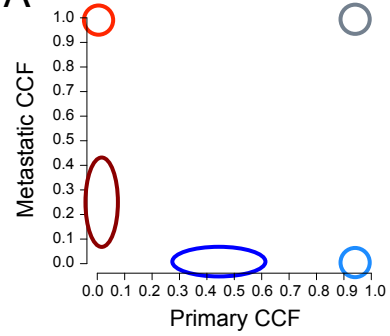
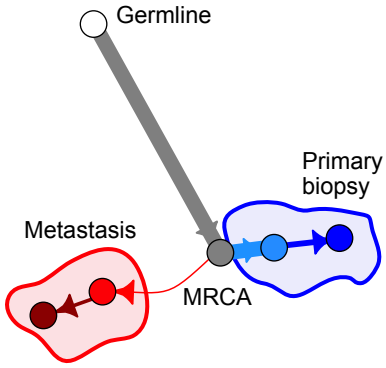
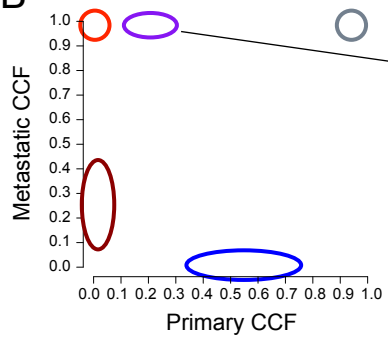
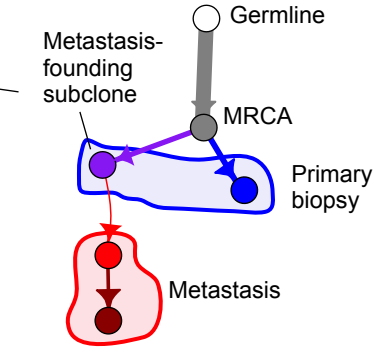
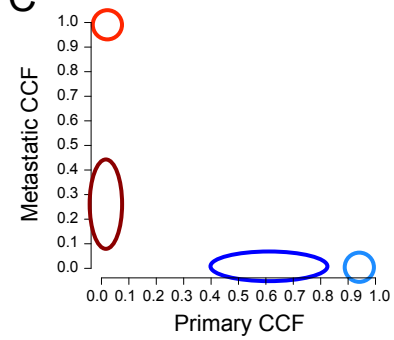
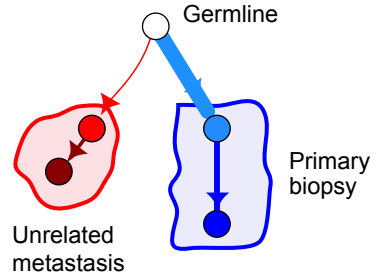
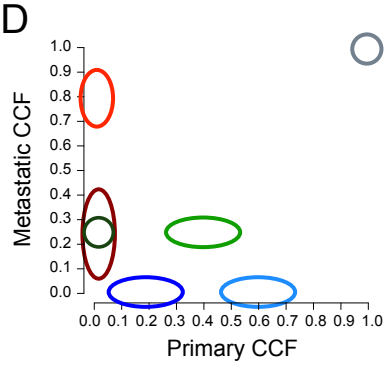
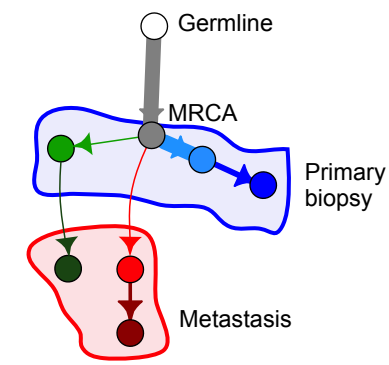
| Modeled mutation data | Inferred phylogeny and tissue sampling | Description |
|---|---|---|
| A   |   | • Branched (sibling) relationship.<br>• The primary and metastasis share a most-recent common ancestor (MRCA), but have continued to evolve independently.<br>• 82 of 86 cases. |
| B   |   | • Metastasis-founding subclone detected in the primary biopsy.<br>• 0 of 86 cases. |
| C   |   | • Unrelated primary and metastasis samples.<br>• 4 of 86 cases. |
| D   |   | • Polyclonal seeding of metastasis, with one of the ancestor subclones detected in the primary tumor.<br>• 0 of 86 cases. |

**Figure S7. Evolutionary relationships between primary tumor-samples and brain metastasis samples**

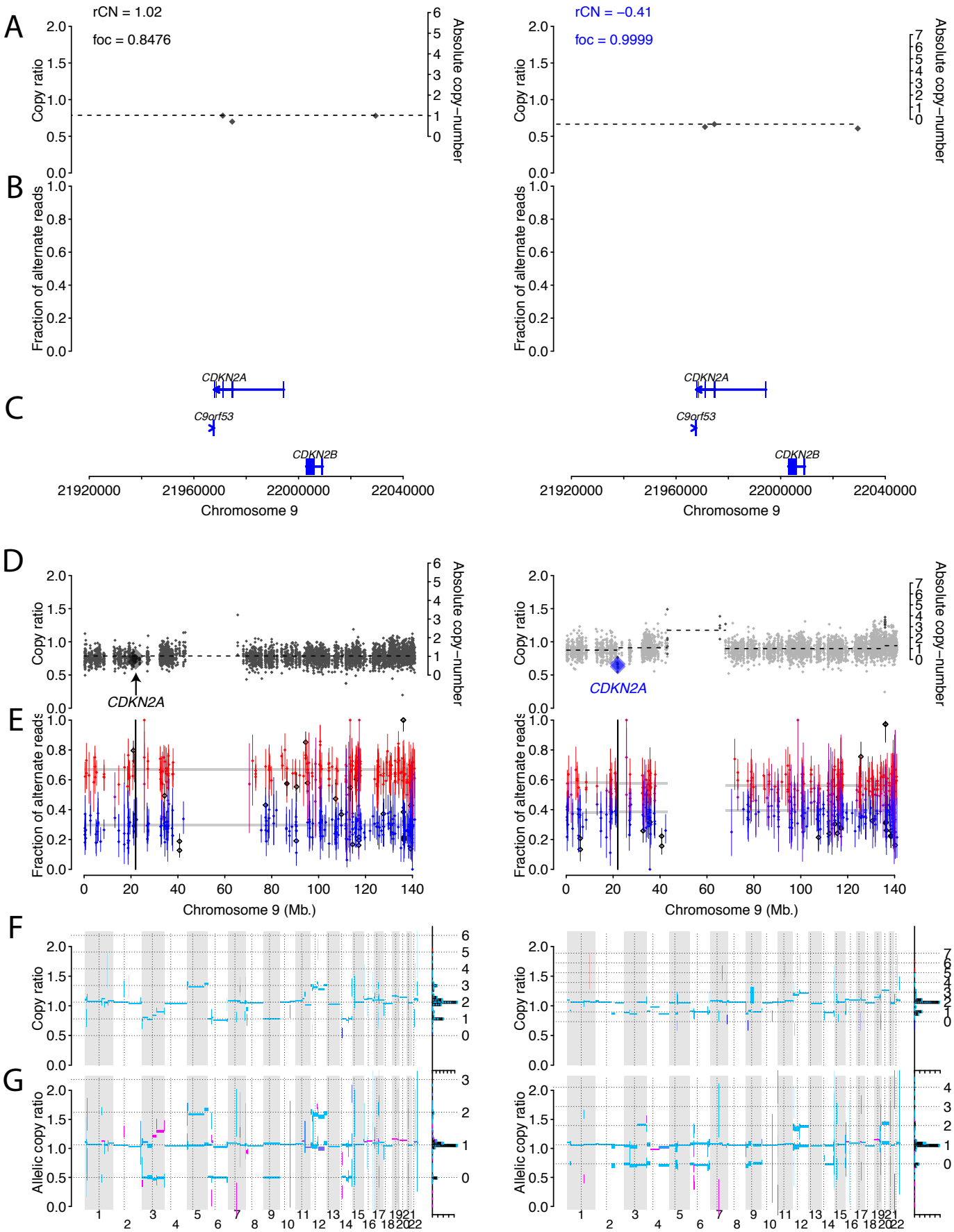



**Figure S8. Detection of homozygous deletion in *CDKN2A* in the brain metastasis of case 24**

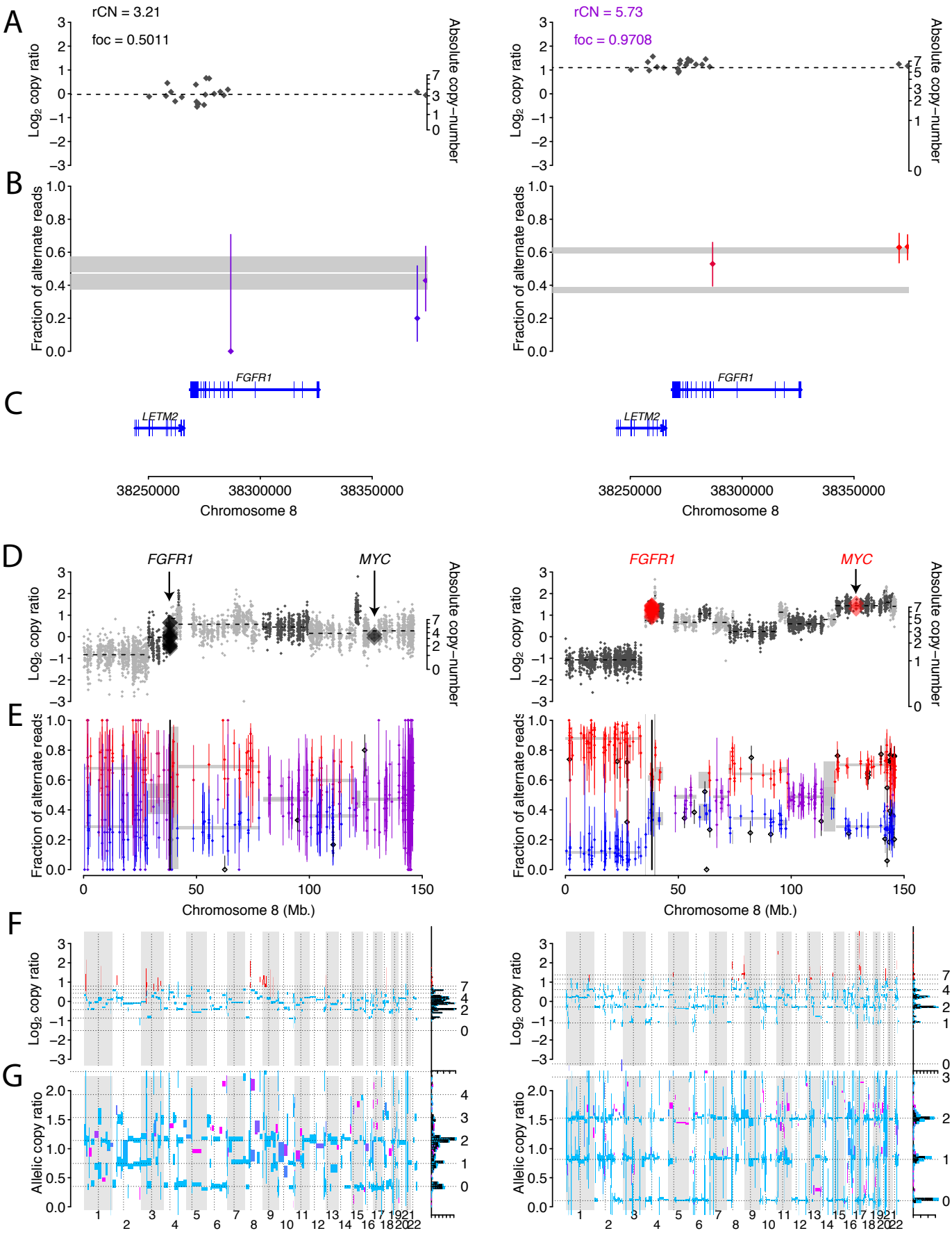



**Figure S9. Amplification of *FGFR1* and *MYC* detected in the brain metastasis of case 331**

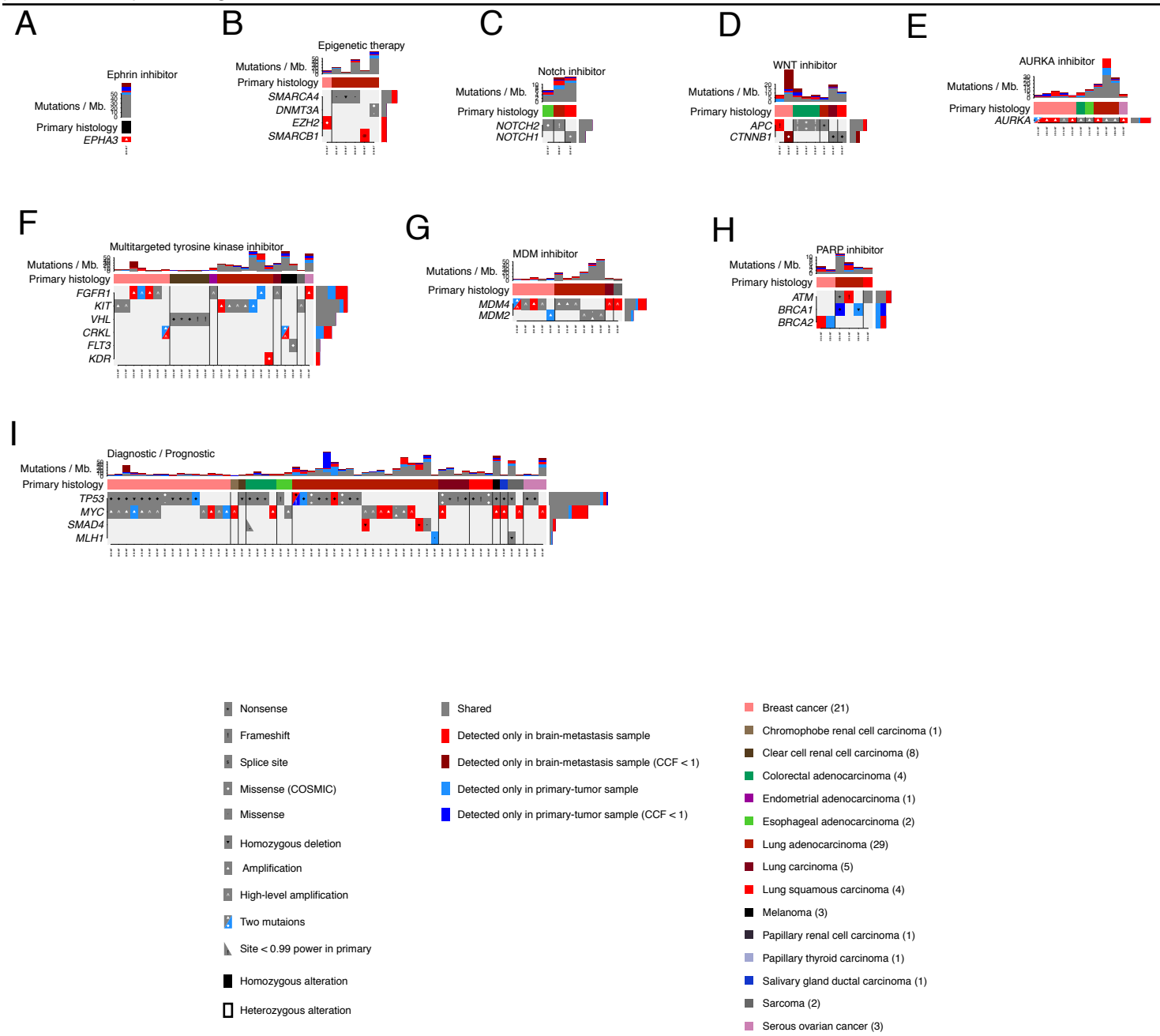



**Figure S10. Additional alterations under investigation for association with various targeted therapies**

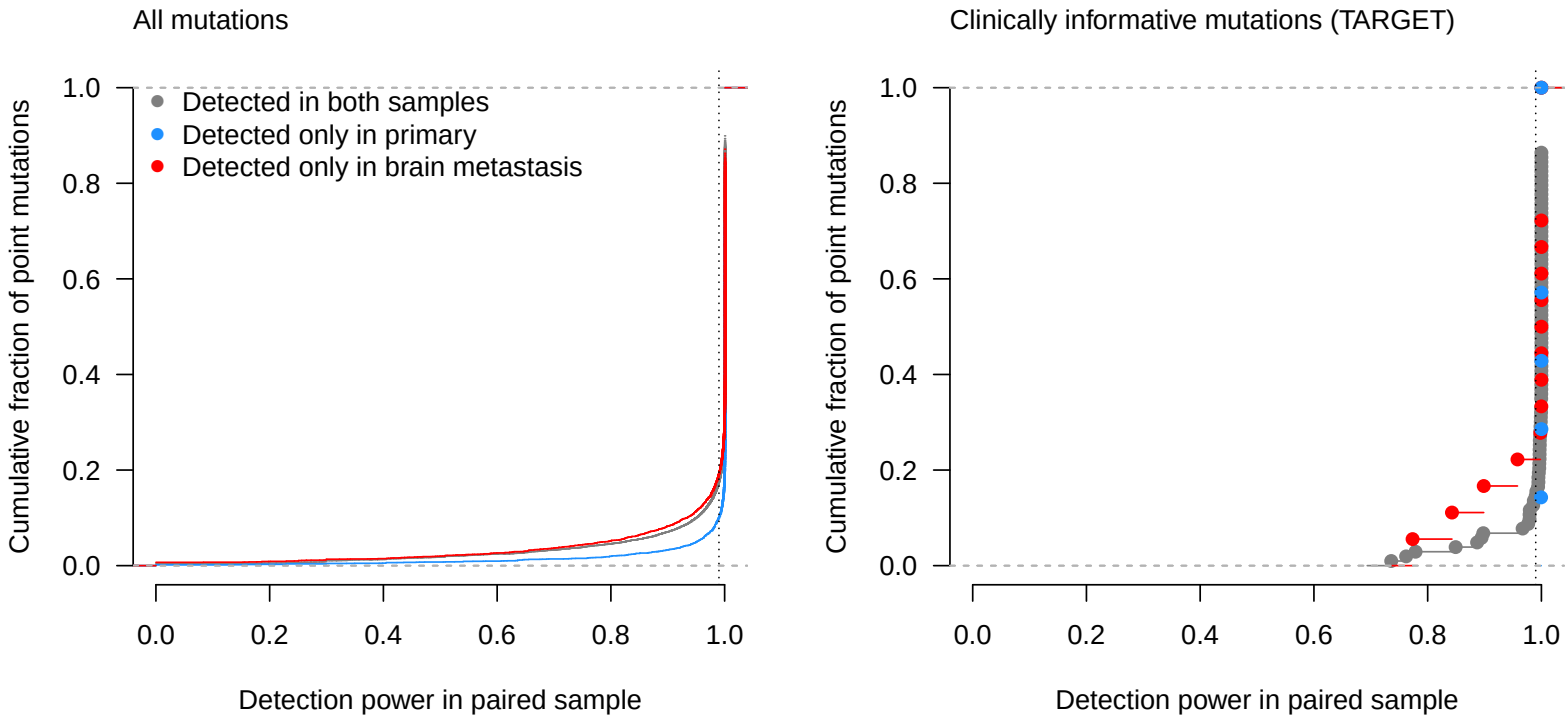



**Figure S11. Power for paired-detection of somatic mutations**

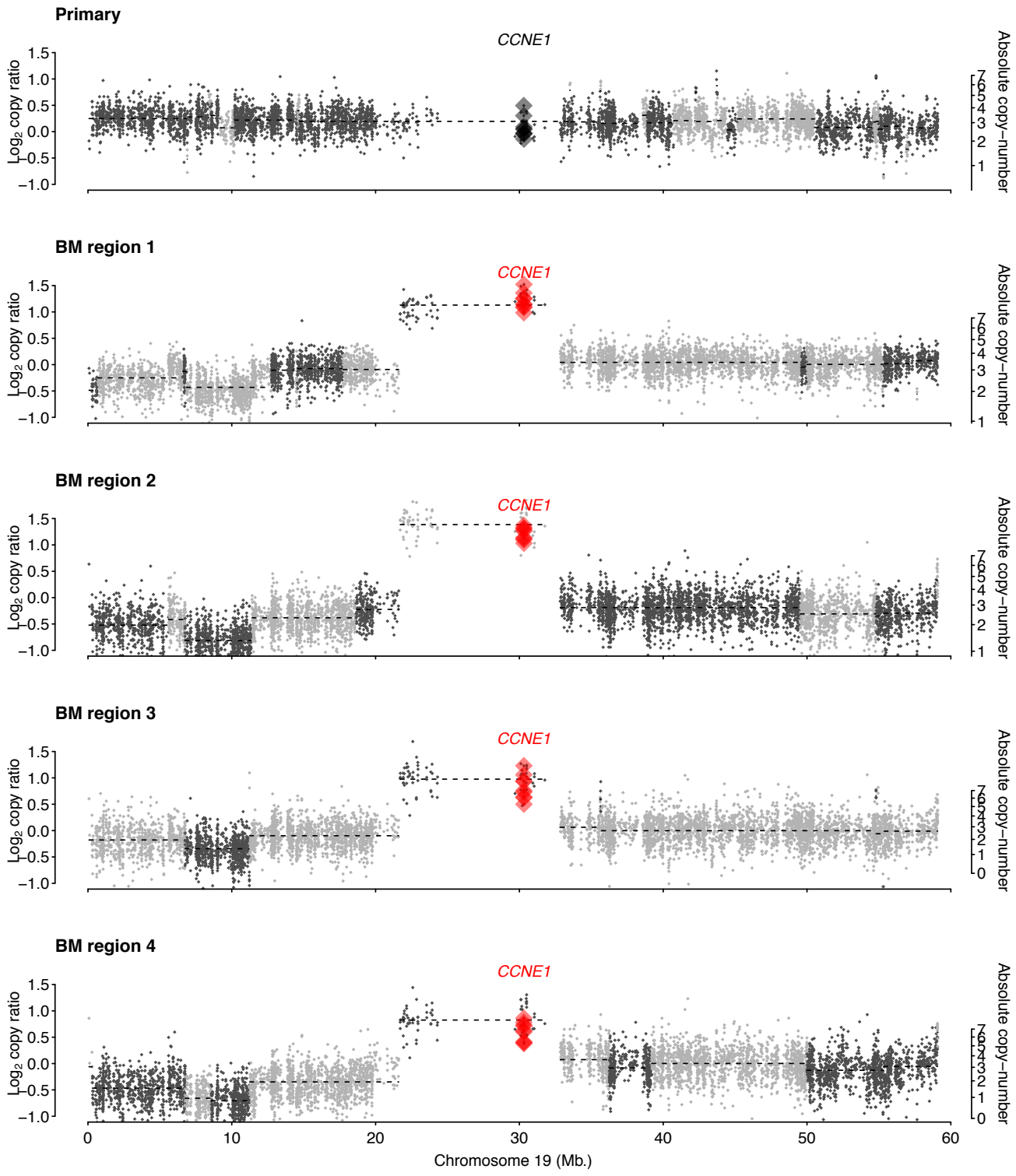



**Figure S12. Amplification of *CCNE1* detected in the brain metastasis of case 314**

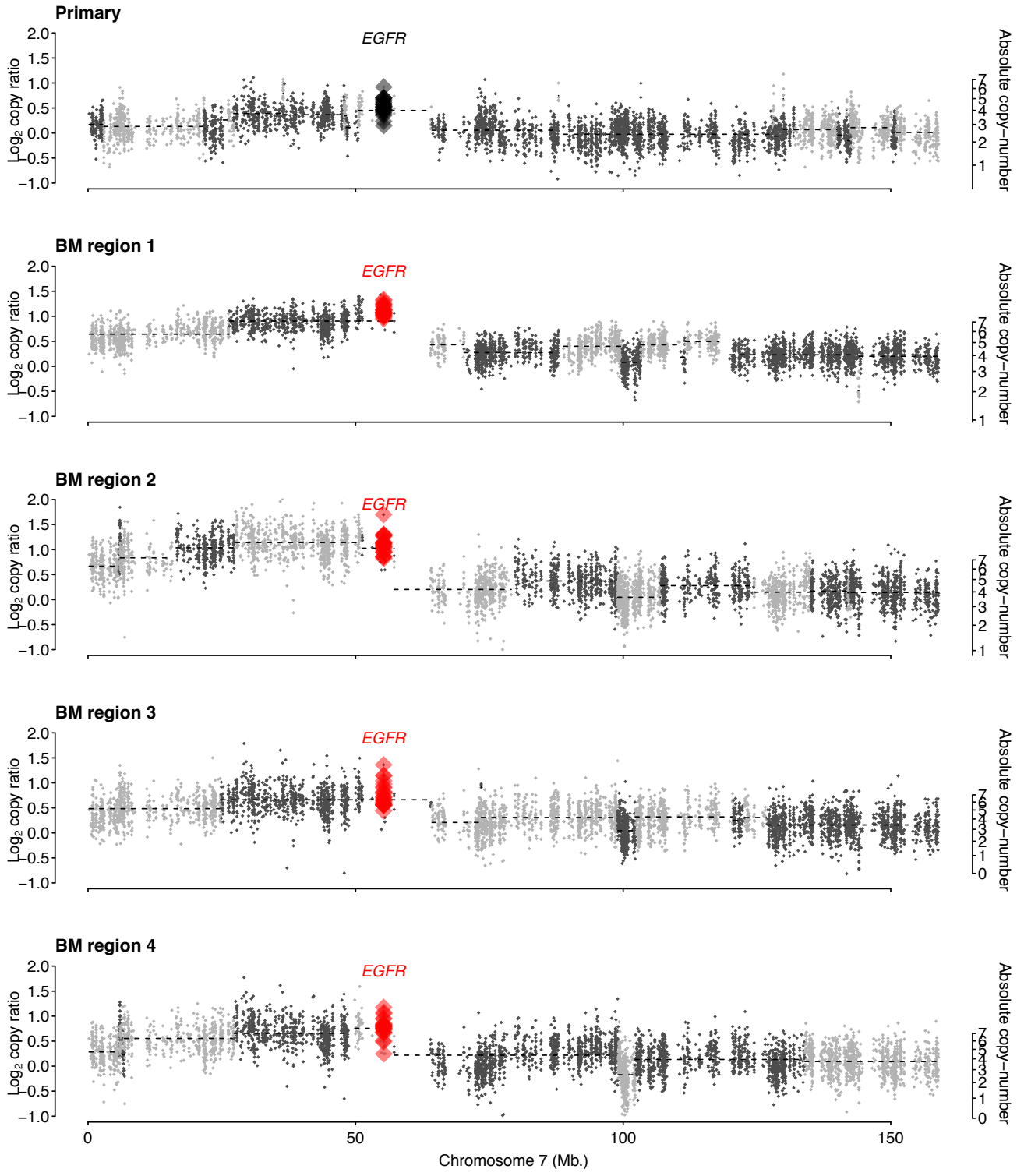



**Figure S13. Amplification of *EGFR* detected in the brain metastasis of case 314**

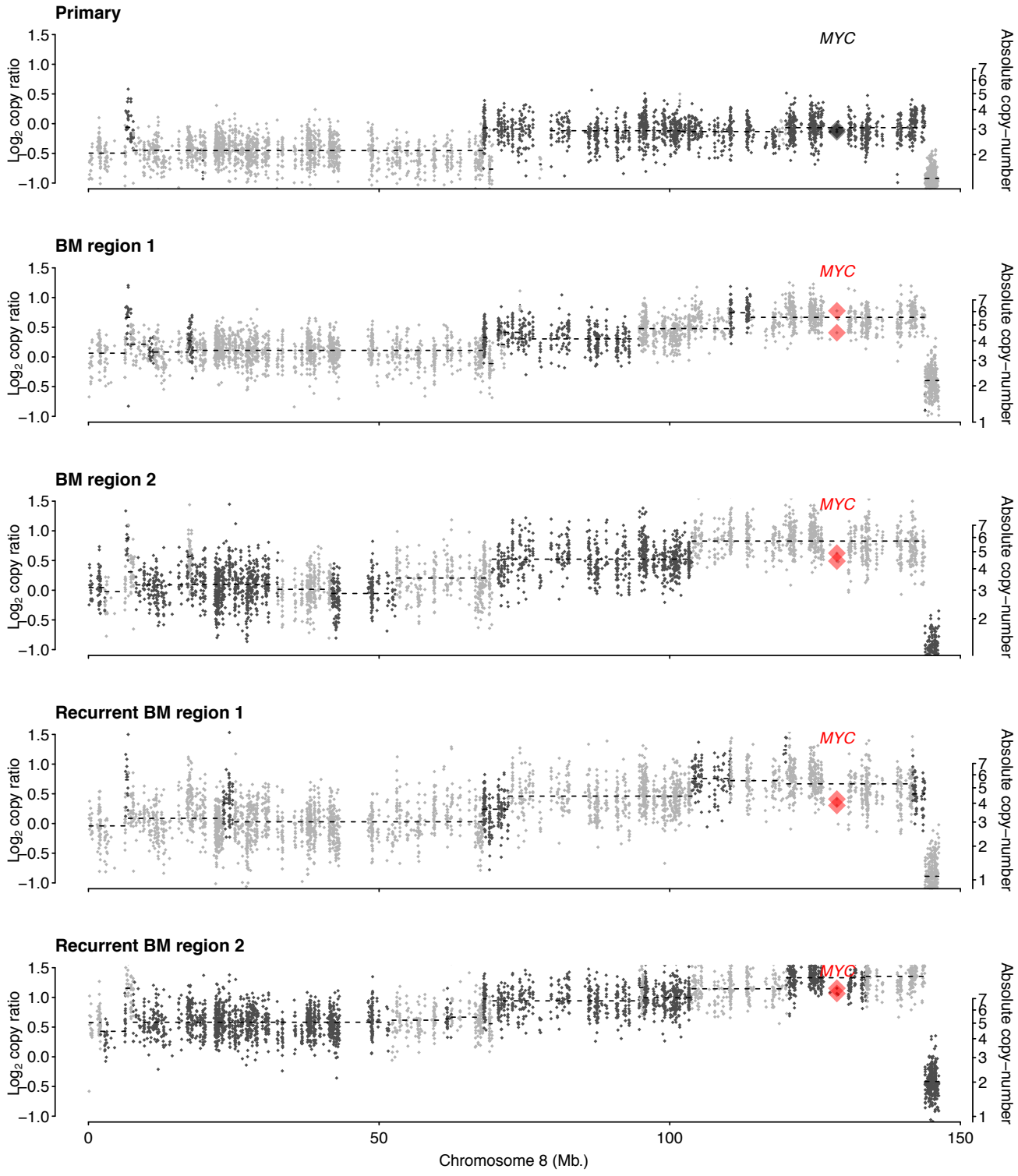



**Figure S14. Amplification of *MYC* detected in the brain metastasis of case 308**

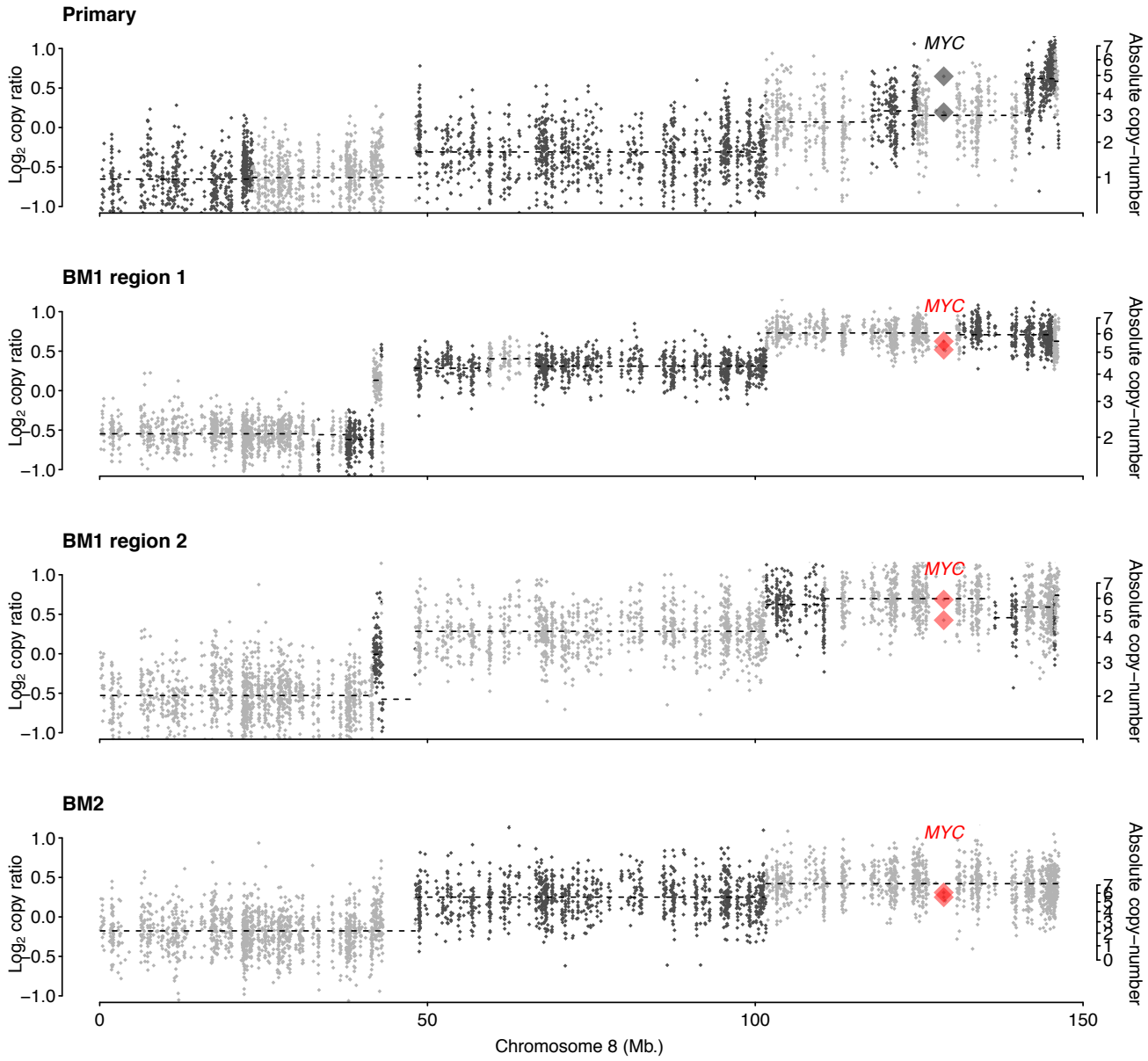



**Figure S15. Amplification of *MYC* detected in the brain metastasis of case 138**

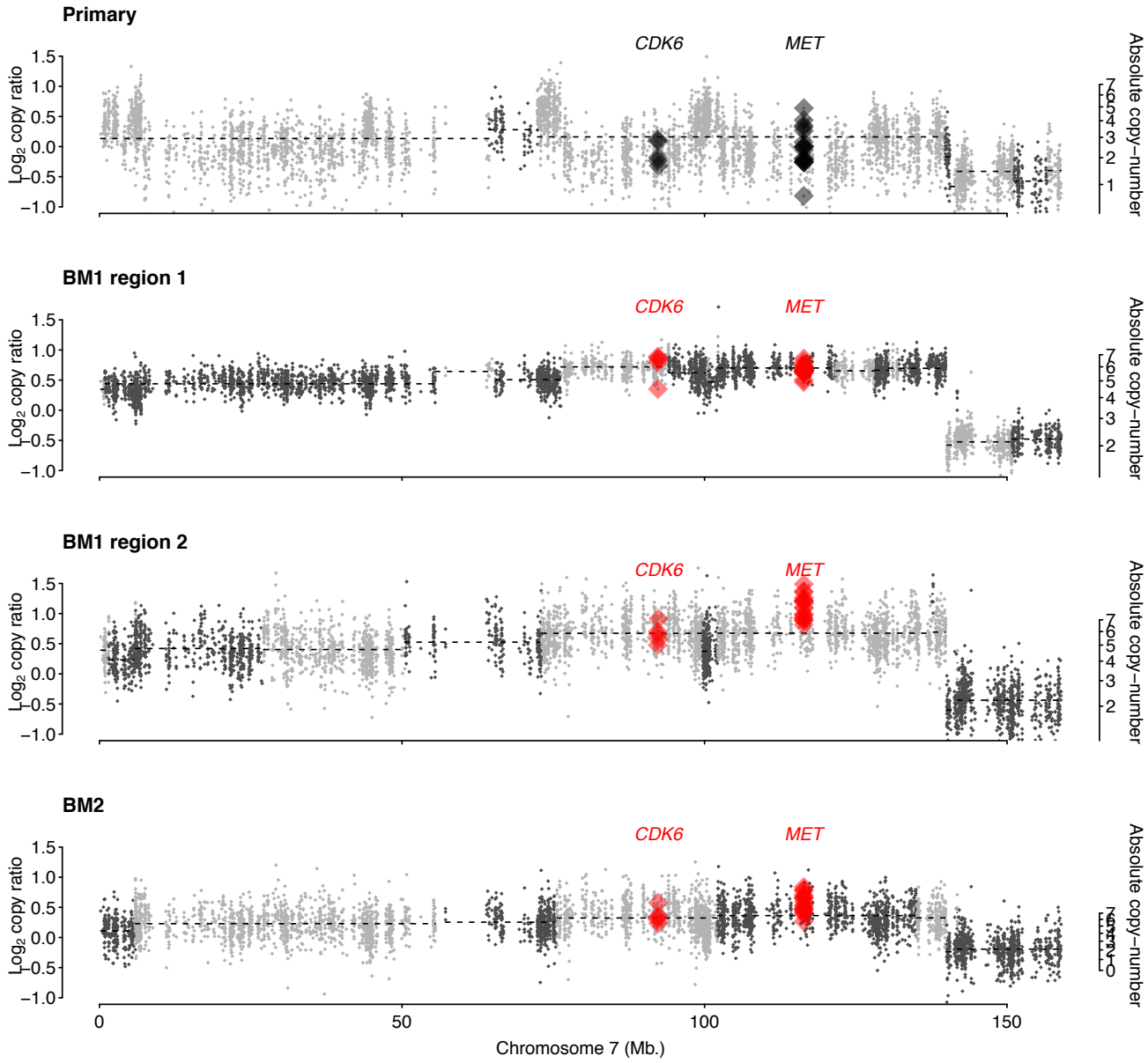



**Figure S16. Amplifications of *CDK6* and *MET* detected in the brain metastasis of case 138**

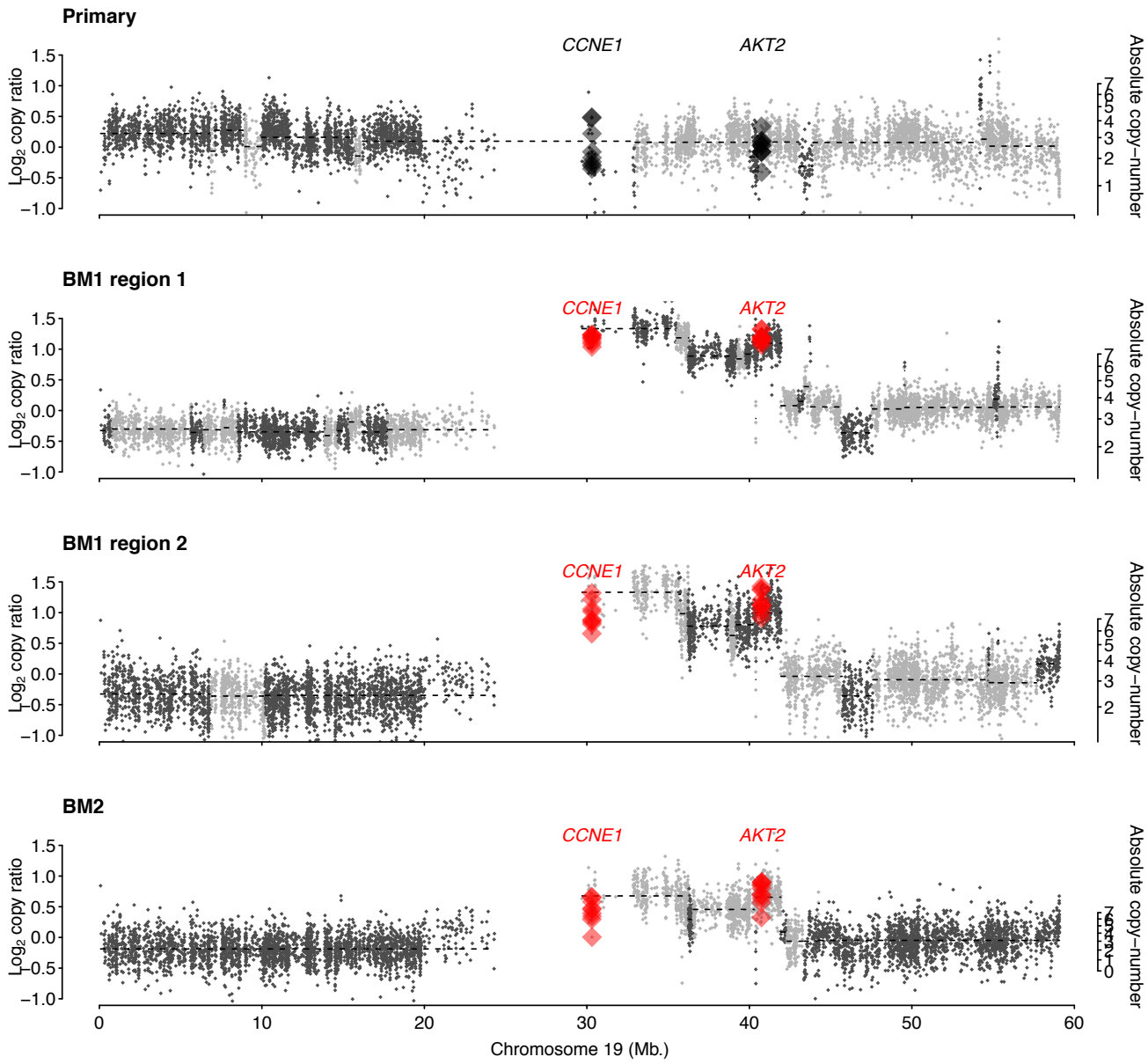



**Figure S17. Amplifications of *CCNE1* and *AKT2* detected in the brain metastasis of case 138**

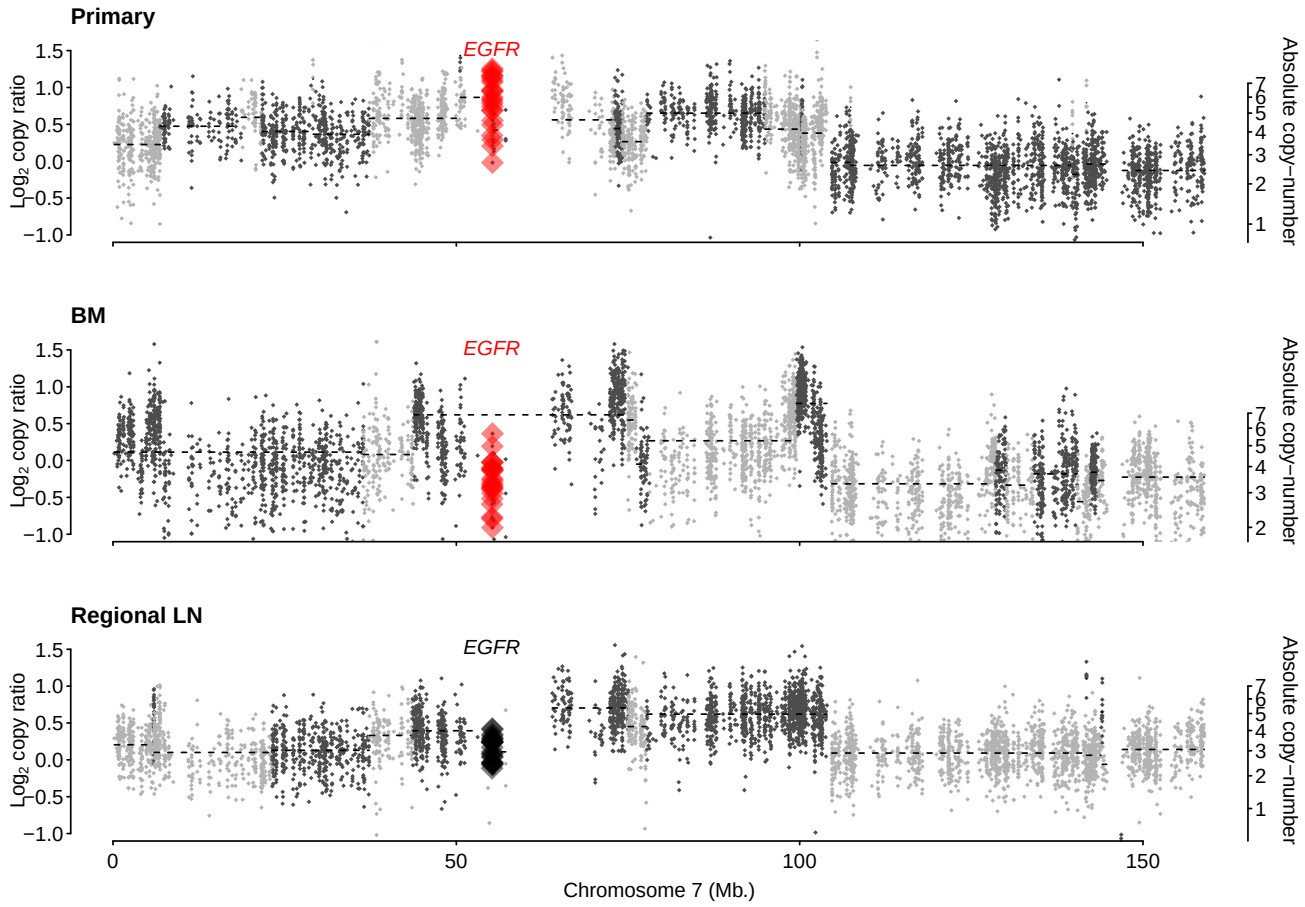



**Figure S18. Amplification of *EGFR* detected in the primary-tumor sample from case 296**

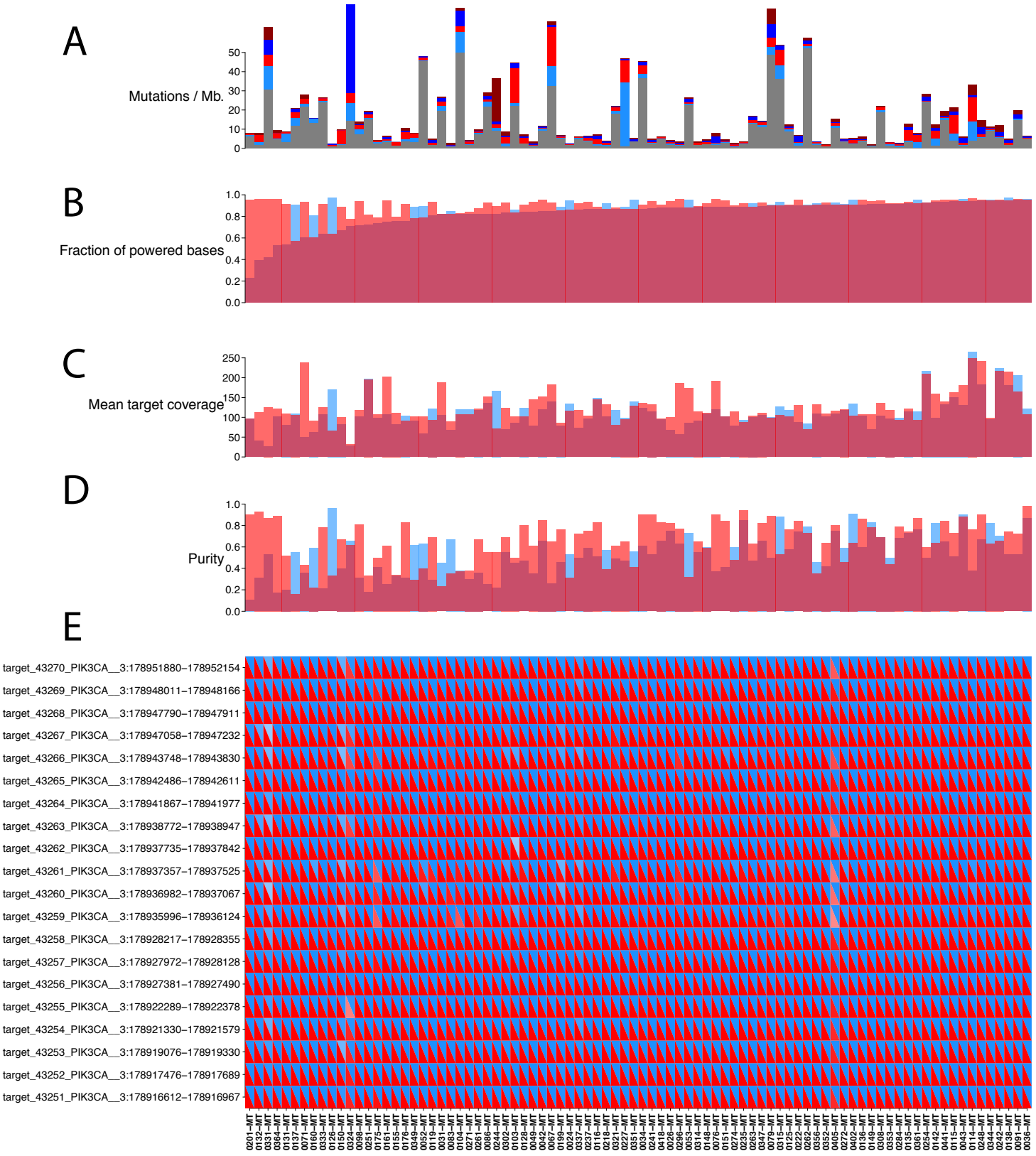



**Figure S19. Power for somatic mutation detection in 86 matched primary-tumor and brain-metastasis samples**

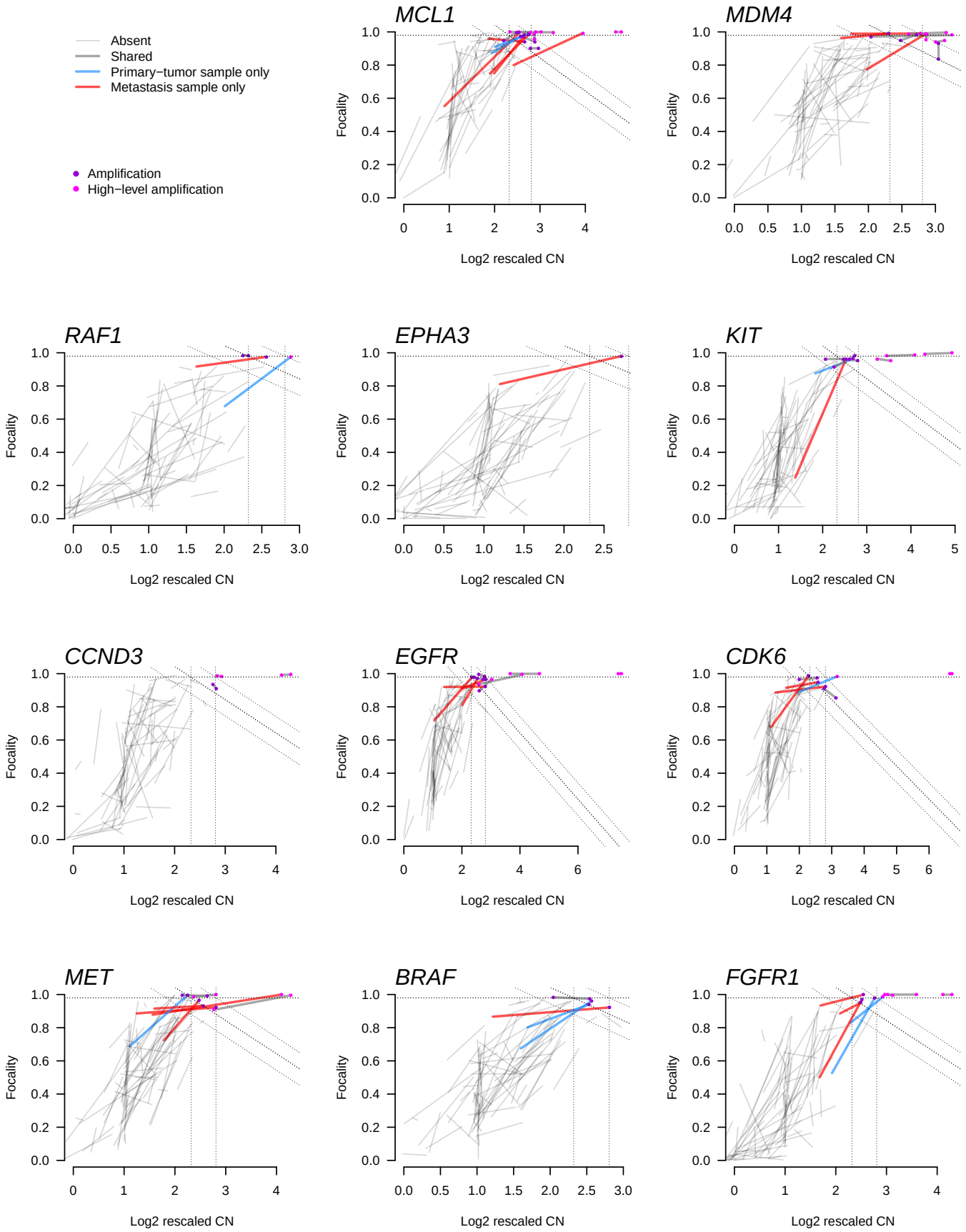



**Figure S20. Calling of amplifications in primary-tumor samples and paired brain metastases**

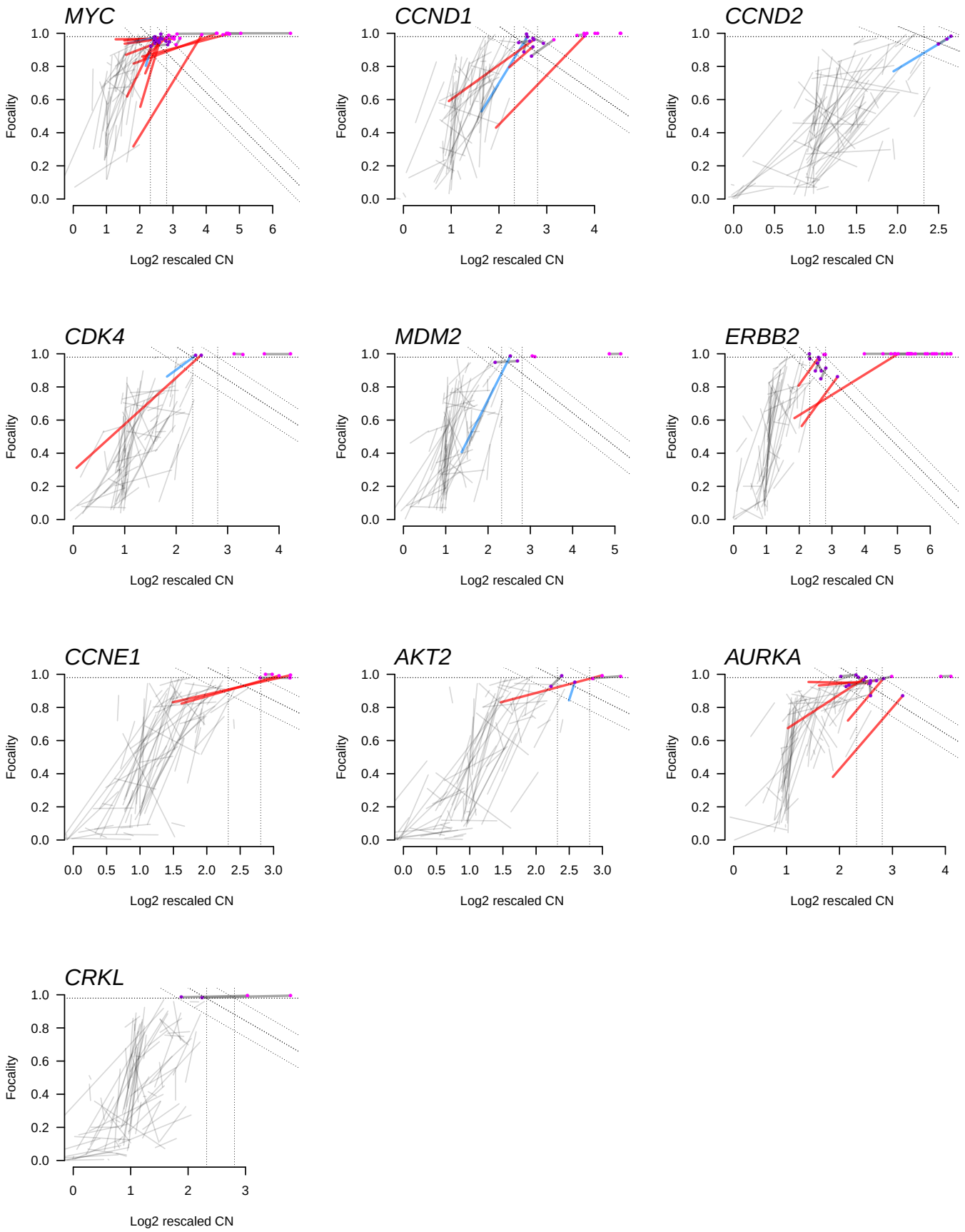



**Figure S20. Calling of amplifications in primary-tumor samples and paired brain metastases**