INSTRUCTIONS ON THE ANNOTATION OF PDF FILES

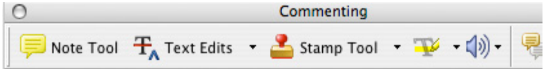
To view, print and annotate your article you will need Adobe Reader version 9 (or higher). This program is freely available for a whole series of platforms that include PC, Mac, and UNIX and can be downloaded from http://get.adobe.com/reader/. The exact system requirements are given at the Adobe site: http://www.adobe.com/products/reader/tech-specs.html.

Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file.

PDF ANNOTATIONS

Adobe Reader version 9

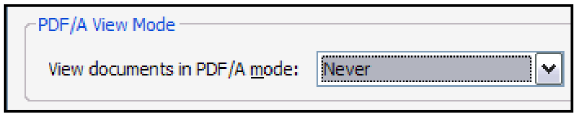
When you open the PDF file using Adobe Reader, the Commenting tool bar should be displayed automatically; if not, click on 'Tools', select 'Comment & Markup', then click on 'Show Comment & Markup tool bar' (or 'Show Commenting bar' on the Mac). If these options are not available in your Adobe Reader menus then it is possible that your Adobe Acrobat version is lower than 9 or the PDF has not been prepared properly.



(Mac)

PDF ANNOTATIONS (Adobe Reader version 9)

The default for the Commenting tool bar is set to 'off' in version 9. To change this setting select 'Edit | Preferences', then 'Documents' (at left under 'Categories'), then select the option 'Never' for 'PDF/A View Mode'.



PDF/A View Mode

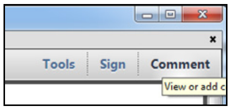
View documents in PDF/A mode: Never

(Changing the default setting, Adobe version 9)

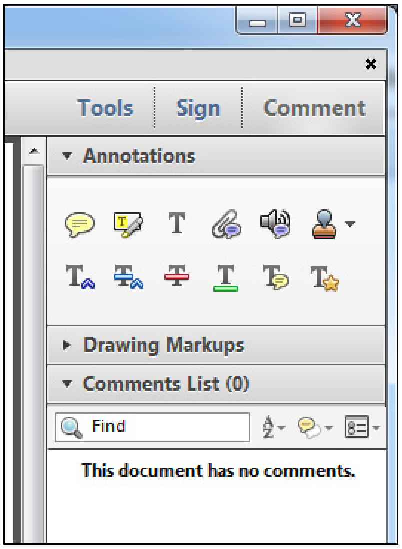
Adobe Reader version X and XI

To make annotations in the PDF file, open the PDF file using Adobe Reader XI, click on 'Comment'.

If this option is not available in your Adobe Reader menus then it is possible that your Adobe Acrobat version is lower than XI or the PDF has not been prepared properly.



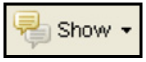
This opens a task pane and, below that, a list of all Comments in the text. These comments initially show all the changes made by our copyeditor to your file.



This document has no comments.

HOW TO...

Action	Adobe Reader version 9	Adobe Reader version X and XI
Insert text	Click the 'Text Edits' button Text Edits on the Commenting tool bar. Click to set the cursor location in the text and simply start typing. The text will appear in a commenting box. You may also cut-and-paste text from another file into the commenting box. Close the box by clicking on 'x' in the top right-hand corner.	Click the 'Insert Text' icon on the Comment tool bar. Click to set the cursor location in the text and simply start typing. The text will appear in a commenting box. You may also cut-and-paste text from another file into the commenting box. Close the box by clicking on '_' in the top right-hand corner.
Replace text	Click the 'Text Edits' button Text Edits on the Commenting tool bar. To highlight the text to be replaced, click and drag the cursor over the text. Then simply type in the replacement text. The replacement text will appear in a commenting box. You may also cut-and-paste text from another file into this box. To replace formatted text (an equation for example) please Attach a file (see below).	Click the 'Replace (Ins)' icon on the Comment tool bar. To highlight the text to be replaced, click and drag the cursor over the text. Then simply type in the replacement text. The replacement text will appear in a commenting box. You may also cut-and-paste text from another file into this box. To replace formatted text (an equation for example) please Attach a file (see below).
Remove text	Click the 'Text Edits' button Text Edits on the Commenting tool bar. Click and drag over the text to be deleted. Then press the delete button on your keyboard. The text to be deleted will then be struck through.	Click the 'Strikethrough (Del)' icon on the Comment tool bar. Click and drag over the text to be deleted. Then press the delete button on your keyboard. The text to be deleted will then be struck through.
Highlight text/ make a comment	Click on the 'Highlight' button on the Commenting tool bar. Click and drag over the text. To make a comment, double click on the highlighted text and simply start typing.	Click on the 'Highlight Text' icon on the Comment tool bar. Click and drag over the text. To make a comment, double click on the highlighted text and simply start typing.
Attach a file	Click on the 'Attach a File' button on the Commenting tool bar. Click on the figure, table or formatted text to be replaced. A window will automatically open allowing you to attach the file. To make a comment, go to 'General' in the 'Properties' window, and then 'Description'. A graphic will appear in the PDF file indicating the insertion of a file.	Click on the 'Attach File' icon on the Comment tool bar. Click on the figure, table or formatted text to be replaced. A window will automatically open allowing you to attach the file. A graphic will appear indicating the insertion of a file.
Leave a note/ comment	Click on the 'Note Tool' button Note Tool on the Commenting tool bar. Click to set the location of the note on the document and simply start typing. Do not use this feature to make text edits.	Click on the 'Add Sticky Note' icon on the Comment tool bar. Click to set the location of the note on the document and simply start typing. Do not use this feature to make text edits.

HOW TO...		
Action	**Adobe Reader version 9**	**Adobe Reader version X and XI**
Review	To review your changes, click on the 'Show' button  on the Commenting tool bar. Choose 'Show Comments List'. Navigate by clicking on a correction in the list. Alternatively, double click on any mark-up to open the commenting box.	Your changes will appear automatically in a list below the Comment tool bar. Navigate by clicking on a correction in the list. Alternatively, double click on any mark-up to open the commenting box.
Undo/delete change	To undo any changes made, use the right click button on your mouse (for PCs, Ctrl-Click for the Mac). Alternatively click on 'Edit' in the main Adobe menu and then 'Undo'. You can also delete edits using the right click (Ctrl-click on the Mac) and selecting 'Delete'.	To undo any changes made, use the right click button on your mouse (for PCs, Ctrl-Click for the Mac). Alternatively click on 'Edit' in the main Adobe menu and then 'Undo'. You can also delete edits using the right click (Ctrl-click on the Mac) and selecting 'Delete'.

SEND YOUR ANNOTATED PDF FILE BACK TO ELSEVIER

Save the annotations to your file and return as instructed by Elsevier. Before returning, please ensure you have answered any questions raised on the Query Form and that you have inserted all corrections: later inclusion of any subsequent corrections cannot be guaranteed.

FURTHER POINTS

- Any (grey) halftones (photographs, micrographs, etc.) are best viewed on screen, for which they are optimized, and your local printer may not be able to output the greys correctly.
- If the PDF files contain colour images, and if you do have a local colour printer available, then it will be likely that you will not be able to correctly reproduce the colours on it, as local variations can occur.
- If you print the PDF file attached, and notice some 'non-standard' output, please check if the problem is also present on screen. If the correct printer driver for your printer is not installed on your PC, the printed output will be distorted.

Please cite this article in press as: Yang et al., Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes, Cell (2013), http://dx.doi.org/10.1016/j.cell.2013.04.010

Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes

Lixing Yang,[1] Lovelace J. Luquette,[1] Nils Gehlenborg,[1,3] Ruibin Xi,[1,4] Psalm S. Haseley,[1,5] Chih-Heng Hsieh,[6] Chengsheng Zhang,[6] Xiaojia Ren,[5] Alexei Protopopov,[7] Lynda Chin,[7] Raju Kucherlapati,[2,5] Charles Lee,[6,8] and Peter J. Park[1,5,9,*]

[1]Center for Biomedical Informatics
[2]Department of Genetics
Harvard Medical School, Boston, MA 02115, USA
[3]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA 02115, USA
[4]School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China
[5]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA
[6]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[7]Department of Genomic Medicine and Institute for Applied Cancer Science, MD Anderson Cancer Center, Houston, TX 77054, USA
[8]Department of Medicine, Seoul National University School of Medicine, Seoul 110-799, South Korea
[9]Informatics Program, Children's Hospital, Boston, MA 02115, USA
*Correspondence: peter_park@harvard.edu

http://dx.doi.org/10.1016/j.cell.2013.04.010

SUMMARY

Identification of somatic rearrangements in cancer genomes has accelerated through analysis of high-throughput sequencing data. However, characterization of complex structural alterations and their underlying mechanisms remains inadequate. Here, applying an algorithm to predict structural variations from short reads, we report a comprehensive catalog of somatic structural variations and the mechanisms generating them, using high-coverage whole-genome sequencing data from 140 patients across ten tumor types. We characterize the relative contributions of different types of rearrangements and their mutational mechanisms, find that ~20% of the somatic deletions are complex deletions formed by replication errors, and describe the differences between the mutational mechanisms in somatic and germline alterations. Importantly, we provide detailed reconstructions of the events responsible for loss of *CDKN2A/B* and gain of *EGFR* in glioblastoma, revealing that these alterations can result from multiple mechanisms even in a single genome and that both DNA double-strand breaks and replication errors drive somatic rearrangements.

INTRODUCTION

Cancer is a disease driven by genetic alterations, which include single nucleotide variations (SNVs), structural variations (SVs), and aneuploidy. The spectrum of somatic SNVs studied suggests that mismatch repair deficiency and specific mutagenic exposure such as smoking, UV light, and chemotherapy can be inferred from the mutational signatures for specific tumor types (Greenman et al., 2007; Lee et al., 2010; Pleasance et al., 2010a, 2010b). Larger scale SVs including deletions, insertions, inversions, tandem duplications, translocations, and more complex rearrangements constitute another frequent type of alterations that could alter normal gene function in tumors. Somatic SVs have been characterized in several previous studies (Bass et al., 2011; Berger et al., 2011; Campbell et al., 2010; Hillmer et al., 2011; Stephens et al., 2009); however, which driving forces exist for SV formation is still unclear.

Three main types of mechanisms known to cause SVs are homologous recombination, nonreplicative nonhomologous repair, and replication-based mechanisms (Gu et al., 2008; Hastings et al., 2009b). Homologous recombination is the most common DNA repair mechanism and is generally accurate, except when the pairing is between incorrect homologous regions, as in nonallelic homologous recombination (NAHR). Deficiency in homologous recombination is believed to be a major source of cancer genome instability (Hoeijmakers, 2001). For example, *BRCA1* and *BRCA2* are required for the homology-directed repair of chromosomal breaks, and loss of these two genes often results in genome instability and cancer (Venkitaraman, 2002). Nonhomologous end joining (NHEJ) is a nonreplicative nonhomologous repair mechanism that requires no homology and sometimes can generate very short microhomology or small insertions at the breakpoint (Mahaney et al., 2009). In addition, alternative end joining (alt-EJ) mechanism, also called microhomology-mediated end joining (MMEJ), can generate blunt ends, small insertions, and, more frequently, microhomology at the deletion breakpoints (Bennardo et al., 2008; McVey and Lee, 2008). Which factors are involved in alt-EJ is much less clear than it is for NHEJ, and alt-EJ appears to be independent of the canonical NHEJ factors such as Ku70 and XRCC4 (Arlt et al., 2012; Bennardo et al., 2008). NHEJ has often been implicated in tumor genomes, based on the very few overlapping sequences at breakpoints (Stephens et al., 2009, 2011). For complex rearrangements, a replicative mechanism called fork stalling and template switching (FoSTeS) has been described (Lee et al., 2007), which later was generalized to microhomology-mediated break-induced repair (MMBIR). It has been proposed that the replication at the fork can stall and the polymerase can shift template via microhomology to any nearby single-stranded DNA, resulting in inversion, tandem duplication, translocation, or more complex rearrangements (Hastings et al., 2009a; Zhang et al., 2009). A recent assessment of SV formation (mostly focusing on deletions) in a normal human population by the 1000 Genomes Project (Mills et al., 2011) did not address the role of replication-based mechanisms. In cancer, a recent study hypothesized that complex genomic rearrangements can arise from a single catastrophic event (Stephens et al., 2011) driven by NHEJ; similar complex rearrangements have also been observed in pathogenic germline alterations (Chiang et al., 2012). Another study suggested that multiple amplifications in developmental delay and cognitive anomalous patients are generated by a replication-based mechanism (Liu et al., 2011) because microhomology is frequently observed at the breakpoints. However, there is no comprehensive study of the mechanisms that underlie somatic SVs in cancer genomes, and many aspects are still poorly characterized, including the forces that drive SV formation, relative contribution of different mechanisms across tumor types, and whether additional mechanisms play a role.

In our analysis of somatic SVs across ten tumor types, nonhomology-based and microhomology-based mechanisms are consistently the dominant mutational mechanisms responsible for genomic rearrangements, driven by DNA double-strand breaks and replication errors. Importantly, multiple mechanisms sometimes act on a single gene in a genome, and two driving forces can act together on a different part of a genome to create mutations and promote tumorigenesis.

RESULTS

Identification of Germline Complex Deletion Events Using Meerkat

To characterize the mutational spectrum of somatic SVs in cancer, it is important to identify both simple (e.g., deletion, insertion, and inversion) and complex SVs at base-pair resolution. The most common type of complex SV is a deletion with an insertion or inversion at the breakpoint (Conrad et al., 2010; Kidd et al., 2010) generated by FoSTeS/MMBIR. Previously, the identification of such events involved capturing and sequencing of the segments adjacent to the deletion breakpoints (Conrad et al., 2010; Kidd et al., 2010). Here, we predict both germline and somatic SVs directly from short read data, focusing on complex events such as those generated by FoSTeS/MMBIR (an example shown in Figures 1A–1D). This is made possible by a new algorithm called Meerkat (see Experimental Procedures; see also the Extended Experimental Procedures available online). With the base-pair resolution of the breakpoints identified by our method, the mechanisms forming SVs are inferred based on sequence homology at the breakpoints (Kidd et al., 2010; Lam et al., 2010; Mills et al., 2011) (Figure 1E; see also Discussion). Identification of somatic SVs from short read data is challenging due to several factors, including sequencing errors, GC content, and other biases in sequencing, ambiguous alignments due to repetitive sequences, a large number of germline SVs, chimeric molecules generated during library construction, normal cell contamination in tumor samples, and heterogeneity within tumor cell populations. The distinguishing feature of our method is that it considers the configuration of *multiple* clusters of discordant read pairs (read pairs in which the mapped reads are at unexpected distance or orientation) to recognize complex events with high accuracy, in addition to efficiently utilizing split, clipped, and multiple-aligned reads (Extended Experimental Procedures).

To verify the accuracy of our method, we applied Meerkat to two HapMap genomes (NA18507 and NA12878) that have been sequenced at high coverage on the Illumina platform and for which complex deletions have been previously reported (Kidd et al., 2010) based on the sequencing of fosmid library with 40 kb inserts. We identified a total of 3,508 and 2,327 SVs, respectively (Table S1). Of our simple deletion predictions from NA18507 and NA12878, 91.4% (2,102/2,301) and 93.7% (1,304/1,391) were reported in the Database of Genomic Variants (DGV10) (Iafrate et al., 2004) or the 1000 Genomes Project (Mills et al., 2011), respectively, suggesting that the vast majority of our predictions were true events (Figure S1). To further validate the events we detected, we randomly selected 49 events across different types in NA18507 including 24 complex deletion events. We were able to validate 48 events by PCR, including all complex deletions (Table S1). We identified a total of 379 and 253 complex deletions in NA18507 and NA12878, demonstrating that our method is far more sensitive than the previous effort, which reported 2 and 17 complex deletions, respectively (Kidd et al., 2010). Therefore, with the Meerkat algorithm, we can provide a more comprehensive spectrum of mechanisms of SVs in a genome. An example for a complex deletion in NA18507 identified by Meerkat is shown in Figures 1A–1D (the same event was reported by Kidd et al. (2010) but in a different individual NA18956). Comparing our predictions to the simple and complex deletions reported in Kidd et al. (2010), most of the events we failed to identify occur in repetitive regions of the genome (Figure S1; Table S2). This is expected because events reported by Kidd et al. (2010) were based on Sanger sequencing. The examination of repetitive elements with short reads is a challenging problem that we have addressed in a separate study (Lee et al., 2012).

Somatic Structural Variations across Tumor Types

We analyzed high-coverage whole-genome sequencing data from 140 individuals across ten tumor types, including 14 colorectal adenocarcinoma (Bass et al., 2011; Lee et al., 2012), seven multiple myeloma (MM) (Chapman et al., 2011), seven prostate adenocarcinoma (Berger et al., 2011), nine ovarian serous cystadenocarcinoma (OV) (Lee et al., 2012), 16 glioblastoma multiforme (GBM) (Lee et al., 2012), 19 hepatocellular carcinoma

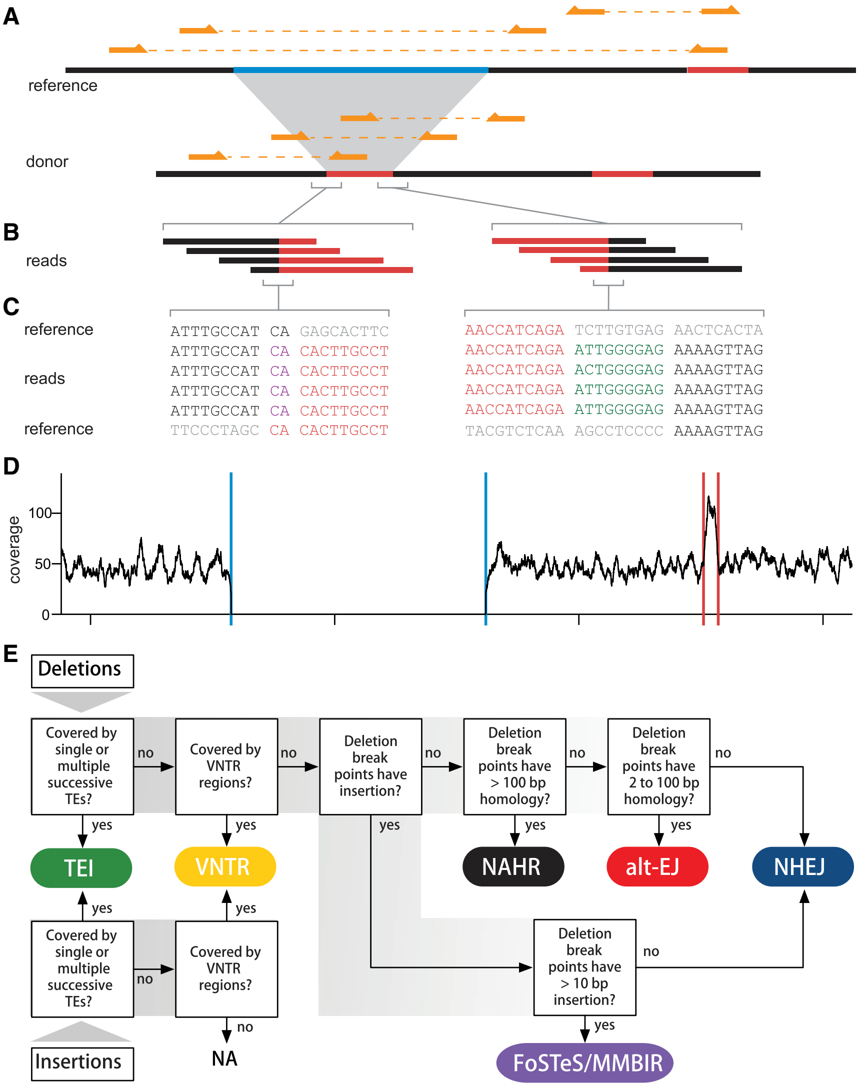

A
reference
donor
B
reads
C
reference
reads
reference
ATTTGCCAT CA GAGCACTTC
ATTTGCCAT CA CACTTGCCT
ATTTGCCAT CA CACTTGCCT
ATTTGCCAT CA CACTTGCCT
ATTTGCCAT CA CACTTGCCT
TTCCCTAGC CA CACTTGCCT
AACCATCAGA TCTTGTGAG AACTCACTA
AACCATCAGA ATTGGGGAG AAAAGTTAG
AACCATCAGA ACTGGGGAG AAAAGTTAG
AACCATCAGA ATTGGGGAG AAAAGTTAG
AACCATCAGA ATTGGGGAG AAAAGTTAG
TACGTCTCAA AGCCTCCCC AAAAGTTAG
D
coverage
100
50
0
E
Deletions
Covered by single or multiple successive TEs?
no
Covered by VNTR regions?
no
Deletion break points have insertion?
no
Deletion break points have > 100 bp homology?
no
Deletion break points have 2 to 100 bp homology?
no
yes
yes
yes
yes
yes
TEI
VNTR
NAHR
alt-EJ
NHEJ
yes
yes
Covered by single or multiple successive TEs?
no
Covered by VNTR regions?
no
NA
Deletion break points have > 10 bp insertion?
no
yes
FoSTeS/MMBIR
Insertions

Figure 1. Example of a Complex Deletion Generated by FoSTeS/MMBIR and a Pipeline for Predicting SV Mechanisms

(A) A complex deletion is predicted by three discordant clusters. The sequence in light blue on the reference is deleted; the sequence in red on the reference is duplicated and inserted into the deletion breakpoints. Three read pairs from the donor are shown above the donor sequence. Three discordant read pairs mapped to the reference are shown above the reference sequence.

(B) Reads covering the breakpoints of insertion. The breakpoints are covered by 27 and 11 reads, respectively (only four are shown for each). Reads matching different parts of the reference genome are shown in the corresponding colors.

(C) Nucleotide sequences of the reads covering the breakpoints of insertion. Black and red colors indicate the reads and the reference sequences that match each other and the gray sequences indicate unmatched references. There are a 2 bp microhomology (shown in purple) at the breakpoint on the left and a 9 bp insertion of unknown source (shown in dark green) at the breakpoint on the right.

(D) Sequencing depth. Blue and red lines denote the predicted deletion and the predicted insertion donor sites, respectively, showing that the copy number is consistent with the SV call.

(E) This flowchart shows the breakpoint features for determining the mechanism that is likely to have generated the observed SV. The classification criteria are mainly adapted from Kidd et al. (2010). Six types of mechanisms are assigned: transposable element insertion (TEI), variable number of tandem repeats (VNTR), nonhomologous end joining (NHEJ), alternative end joining (alt-EJ), nonallelic homologous recombination (NAHR), and fork stalling and template switching/microhomology-mediated break induced repair (FoSTeS/MMBIR).

See also Figure S1 and Tables S1 and S2.

(Sung et al., 2012), 35 breast invasive carcinoma (BRCA) (Cancer Genome Atlas Research Network, 2012a), 19 lung squamous cell carcinoma (LUSC) (Cancer Genome Atlas Research Network, 2012b), ten uterine corpus endometrioid carcinoma (UCEC), and four kidney renal clear cell carcinoma (KIRC) patients. With data from both tumor and germline samples for each patient to distinguish germline and somatic variants, a total of 140 pairs of genomes consisting of about half trillion (458 billion, ~35× coverage per genome on average) paired-end reads (75–101 bp) were analyzed.

A total of 25,874 high-confidence somatic SVs (Table S3) were identified from the 140 cancer genomes (Table S4), ranging from 0 to 3,160 per genome with an average of 185 (Figure 2A). To assess the accuracy of somatic SV predictions, we randomly selected 78 out of 138 SVs in an ovarian tumor (OV0725), including two complex events with two breakpoints each, and examined them in both tumor and normal tissues. By PCR, we were able to validate 73 out of 80 (91%) breakpoints (Table S3) as somatic events.

The frequency of different types of somatic SVs in each sample is shown in Figure 2A. We first note the remarkable variation in the number of SVs among individuals within and across tumor types (the x axis for each tumor type is scaled differently in Figure 2A). Some genomes contain no SV (e.g., LUSC1078 and KIRC4856), whereas others show thousands of SVs (e.g., BRCAA0J6); in a single tumor type, the number of SVs can vary by an order of magnitude between individuals. Among the tumor types, breast tumors and lung squamous cell tumors have significantly more SVs than any other tumors (p = 4.70e-23 and 6.68e-5, respectively, ANOVA tests using a negative binomial model). The number of SVs identified in breast cancer patients here (16,125 in 35 patients, ~461 per patient on average) is much larger than those in previous studies (2,166 in 24 patients, ~90 per patient on average (Stephens et al., 2009) and 2,476 in 22 patients, ~113 per patient on average (Banerji et al., 2012), due to increased sequencing coverage, sensitivity in the detection method, and sample variation. Kidney cancer patients have significantly less SVs than other tumor types

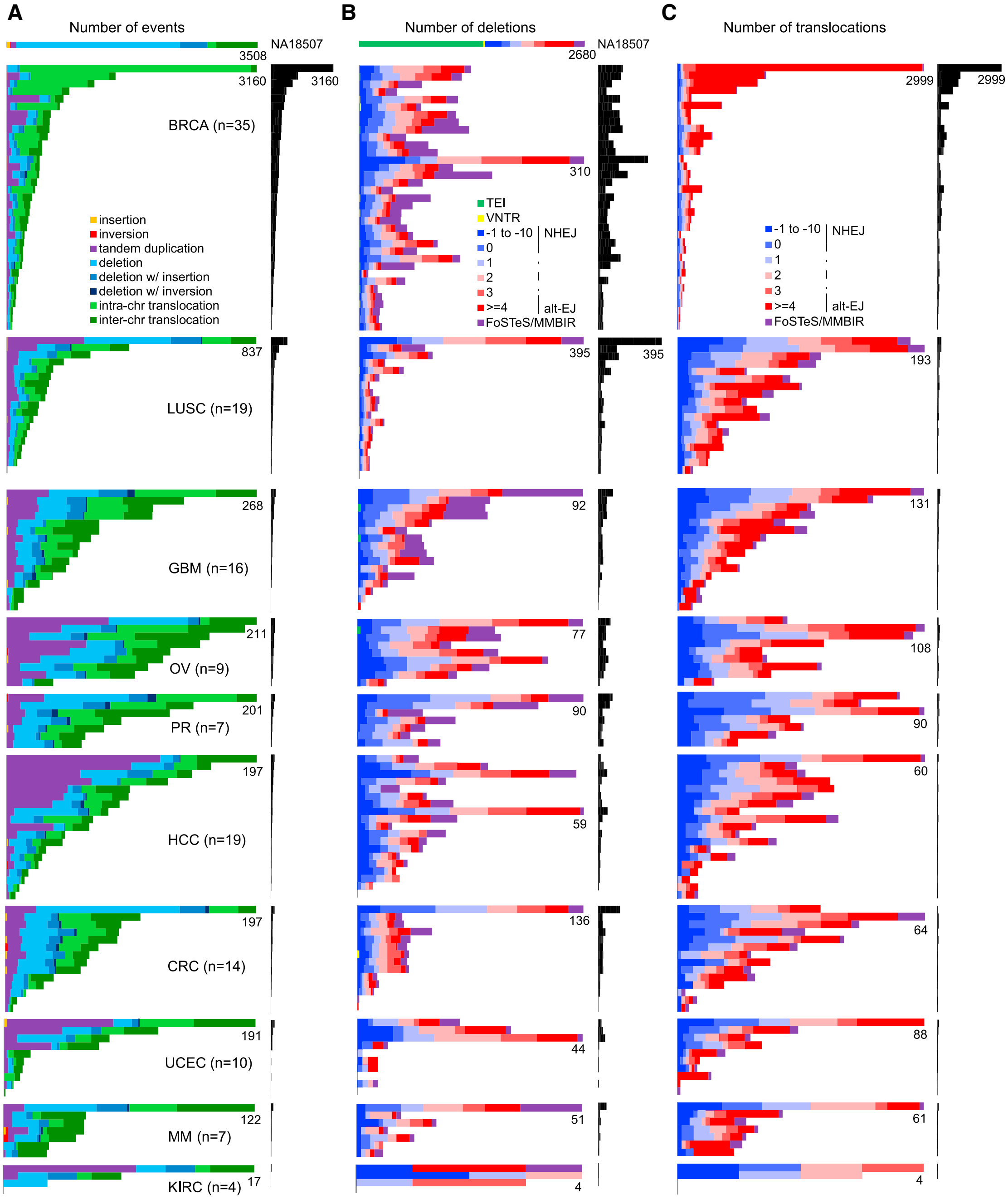


Figure 2. Spectrum of Somatic SV Types and Mechanisms

(A) Frequencies of types of somatic SVs identified in each patient. Each horizontal bar displays the number of SVs for one sample. The colored bar charts on the left show the number of events scaled by the maximum number of events (as noted) in each tumor type. The black bar charts on the right show the number of events for all patients on the same scale. A HapMap genome (NA18507) is shown at the top as an example of germline events; see Figure S2 for germline events for all patients. Most (59%) of the translocations in NA18507 are TE insertions, as described previously (Lee et al., 2012), 18% are repeat-related events including TE insertions not identified by Lee et al. (2012), and the remaining ones might be events too complex to be identified by Meerkat.

(B) Frequencies of somatic deletion mechanisms. The order of the samples is the same as in (A).

(C) Frequencies of somatic translocation mechanisms. The order of the samples is the same as in (A).

See also Figure S2 and Tables S3, S4, and S5.

(p = 1.74e-5, ANOVA test using a negative binomial model). In terms of event types, translocations (57%) are the most abundant SV type, whereas deletions and tandem duplications make up 25% and 17%, respectively. The proportions of different types of SVs across different tumor types are highly variable (Figure 2A). For instance, the breast tumors have significantly more intrachromosomal translocations than other tumors types do (p = 1.22e-53, ANOVA test using a negative binomial model). There are also considerable differences between individual genomes. For example, all rearrangements in a kidney sample (KIRC5010) are deletions, whereas there are no deletions in a liver cancer (HCC13). In nontumor samples, each individual has about 3,000 germline SVs with deletions always being the most abundant (~60%) (Figure S2A), similar to what we find in the HapMap individual NA18507 (Figure 2A).

By pairing multiple clusters of discordant reads to predict complex events, we achieved a better description of the nature of SVs than previously obtained. For example, in the PR0581 genome, a "close chain" pattern had been described to form the *TMPRSS2-ERG* fusion gene, involving *C21orf45* (Berger et al., 2011). We identified two related events in this genome. The first event is a 3 Mb deletion that causes the *TMPRSS2-ERG* fusion. The second event is a 74 bp deletion in the first intron of *C21orf45* at which the 3 Mb deletion from the first event was inserted. The copy numbers of the aforementioned regions were unchanged, supporting the two events we predicted. Detailed descriptions of the events involving *CDKN2A/B, EGFR*, and *CDK4* are provided later.

Certain pathways, such as DNA replication, DNA repair, and cell-cycle pathways, are likely to malfunction in order for the cell to generate and maintain the genomic rearrangements. To investigate this, we identified mutations in genes that in above pathways caused by SVs as well as single nucleotide variants (Bass et al., 2011; Berger et al., 2011; Chapman et al., 2011; Cancer Genome Atlas Research Network, 2012a, 2012b). As expected, almost all patients have at least one gene altered in at least one of these pathways (Table S5); nearly half of the mutations are caused by SVs.

Mutational Mechanisms for Somatic SVs

The number of deletions per genome ranges from 0 to 395, with an average of 46 (Figure 2B). Deletions are usually a result of DNA double-strand break repair. The mechanisms of deletion formation are predicted as shown in Figure 1E (see the Figure 1E legend for information on how mechanistic categories were assigned). In the cancer genomes we studied, NHEJ (39%) and alt-EJ (41%) are the dominant mechanisms. This is in contrast to the mechanisms in the HapMap genome NA18507 (Figure 2B) and nontumor genomes (Figure S2B), in which transposable element insertion (TEI) is always the dominant mechanism, and the frequencies of NHEJ and alt-EJ in germline deletions are ~15% and ~22%, respectively. The increased ratio of NHEJ to alt-EJ in somatic deletions compared to that in germline is statistically significant (p = 1.41e-12, Wilcoxon's paired rank sum test). We also find that about 20% of the somatic deletions are complex deletions formed by FoSTeS/MMBIR (Figure 2B), in contrast to ~5% for the germline deletions (Figures 2B and S2B). The mechanisms of somatic deletions are also variable across different tumor types and between samples. For instance, some genomes have a notable portion of FoSTeS/MMBIR in somatic deletions, whereas others have none (Figure 2B).

We identified between 0 and 2,999 inter- and intrachromosomal translocations in the cancer genomes with an average of 106 translocations per genome (Figure 2C). Again, NHEJ and alt-EJ are the dominant mechanisms with alt-EJ being more abundant in most cases. A small number of translocations are formed by FoSTeS/MMBIR with variable frequencies across genomes. Breast cancer patients have significantly more translocations formed by alt-EJ than other tumor types (p = 4.21e-19, ANOVA test using a negative binomial model). We note that the proportion of translocations formed by NHEJ and alt-EJ are comparable in most tumor types, but alt-EJ is much more prominent in breast tumors that have a large number of translocations (>500). The reason for this difference in translocation formation in those samples is not clear, but it may be due to an alteration in a specific pathway that induces translocations.

Tandem duplications are known to result from unequal crossing over (Edlund and Normark, 1981) or by FoSTeS/MMBIR (Hastings et al., 2009a). Short sequence homologies are required for FoSTeS/MMBIR—the microhomology can be as short as 2 bp to allow new DNA synthesis to start (Zhang et al., 2009)—whereas a larger degree of homology is required for unequal crossing over. Complex rearrangements, especially ones involving dosage gains, are often driven by FoSTeS/MMBIR (Hastings et al., 2009a; Liu et al., 2011) as evidenced by the microhomology frequently observed at the breakpoints. Events with no more than 10 bp insertions at breakpoints were classified as NHEJ because NHEJ is known to generate small insertions at breakpoints (Haviv-Chesner et al., 2007). In HapMap (Figure 3) and other nontumor samples (Figures S3A and S3B), the majority of the breakpoints of tandem duplications (73%) and complex deletions (71%) have microhomology that support the MMBIR models. In contrast, the fraction of breakpoints with microhomology is significantly less in somatic tandem duplications (46%) and complex deletions (52%) (p = 6.78e-19 and p = 1.09e-13, respectively, using Wilcoxon's paired rank sum test; Figures 3, S3C, and S3D). Although most of the germline tandem duplications and complex deletions were generated by FoSTeS/MMBIR (Figures S3A and S3C), a small number have no homology at the breakpoints. In somatic tandem duplications and complex deletions, we do not observe homology at many breakpoints (Figure 3). Thus, we suspect that a template-switching mechanism that does not require microhomology or another non-homology-based mechanism is often utilized in somatic cells to form tandem duplications and complex deletions.

Reconstruction of Complex Rearrangements in GBM Patients

We are particularly interested in GBM genomes because several recurrent copy number alterations we found are known to play an important role in tumorigenesis (Cancer Genome Atlas Research Network, 2008). In our 16 GBM whole-genome data sets, 15 genomes have loss of heterozygosity of chromosome 10, 12 have homozygous deletions of *CDKN2A/B*, 14 have *EGFR* amplifications, and five have *CDK4* amplifications (Table 1; see also Extended Experimental Procedures; Figure S4). We tested 26 SVs involving loss or gain of *CDKN2A/B*, *EGFR*, and *CDK4* and validated 25 as somatic SVs by PCR (Table S3).

Although the copy number changes in these regions have been documented previously based on array data, the exact configuration of the rearrangements and the mechanisms underlying those events are largely unknown. Using Meerkat, we not

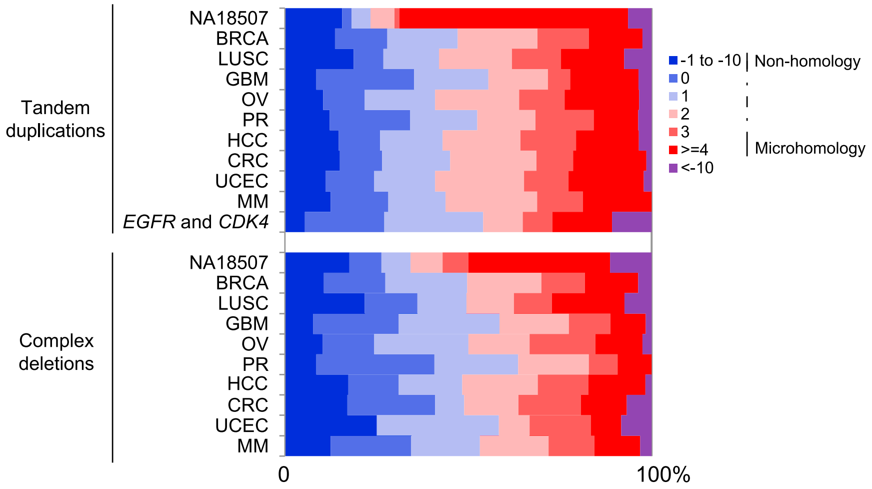


Figure 3. Proportion of Homologies at the Breakpoints of Somatic Tandem Duplications and Complex Deletions Compared with NA18507

Homologies in base pairs are shown for each breakpoint as a positive number. A blunt end has a homology of 0 bp. Small insertions with unknown source are shown as negative numbers. Somatic tandem duplications and complex tandem duplications that are responsible for *EGFR* and *CDK4* amplifications in GBM patients are shown in a separate category.

See also Figure S3.

only ascertained the types of events that generated the observed configuration, but also gained insights into the mechanisms by analyzing sequence homology at the breakpoints. It is interesting to note that in both *CDKN2A/B* loss and *EGFR* gain, most tumor genomes have both arm-level and focal loss/gain (Table 1). Out of 12 patients harboring *CDKN2A/B* loss, six have arm-level loss and focal deletions (Figure 4A), two have two independent focal deletions (Figure 4B), and four have complex rearrangements (Figure 4C). SVs responsible for *CDKN2A/B* loss in other patients are displayed in Figure S5. Most (11 of 13) of the focal deletions were generated by NHEJ, which suggests these alterations are mostly formed through erroneous repair of DNA double-strand breaks.

In the 14 GBM genomes with *EGFR* amplification, most have more than one event contributing to the copy gain: nine with a chromosome arm gain, eight with tandem duplication(s), one with a complex tandem duplication, and eight with complex events. For GBM0155 (Figure 5A), three tandem duplications (1.6 Mb, 983 kb, and 28 kb) involving *EGFR* were identified. In GBM0145 (Figure 5B), *EGFR* is amplified by a 789 kb tandem duplication, but a complex deletion was also found (deletion of a 417 kb fragment with insertion of a 50 kb fragment in the breakpoint). From the copy ratios, it appears that this deletion only affects a subset of the tandem-duplicated copies, suggesting that it happened during the tandem duplication. The complex deletion may not have been generated by FoSTeS/MMBIR because no microhomology was found at the breakpoints. Similarly, GBM0214 (Figure 5C) contains a 59 kb tandem duplication and multiple subsequent rearrangements of various types in the *EGFR* region that are exceedingly difficult to disentangle. SVs responsible for *EGFR* amplifications in other patients are displayed in Figure S6.

In GBM0152, a 923 kb fragment covering *EGFR* (Figure 6A) is merged with two fragments from chromosome 12 (a 5,620 bp fragment and a 286 bp fragment in inverted orientation) and then tandem-duplicated (Figure 6B). Moreover, nearly 40 regions on chromosome 12 (including *CDK4*) are coalesced in an elaborate complex series of events with the copy ratios of various fragments at approximately 40-fold, 75-fold, and 110-fold gain (Figure 6C). In this case, the three prominent copy ratios and all the amplified segments being connected by discordant read pair clusters make it possible to disentangle the underlying events (Figure 6D). Based on the pattern of segments connected by discordant read pairs and the corresponding copy ratio for each segment (Figure 6E), we present one possible coamplified unit (Figure 6F) that is consistent with all the observed copy ratios and discordant read pairs, whereas other compatible configurations are also possible. This single unit (Figure 6F), composed of dozens of fragments, was tandem-duplicated to reach a copy ratio of about 40.

While the complexity of the rearrangements in Figure 6C is reminiscent of chromothripsis (Stephens et al., 2011), it is unlikely to be the case here; instead, it is likely to have been generated by a replication-based mechanism. Chromosomes that have undergone chromothripsis have copy numbers that oscillate between two levels. The complex tandem duplications in our example have several distinct copy numbers, indicating that they are more likely to be a result of a series of replication-based template-switching events. A single unit of amplification contains multiple instances of the segment (junctions 2–4, 12–14, 14–20, and 18–29 in Figure 6F); it is unlikely that at least two copies of chromosome 12 were shattered at the same place and joined with the same segments at the exact breakpoints after a ‘‘one-off’’ catastrophic event. Therefore, we suspect that certain junctions (such as 14–20 in Figure 6F) were formed first by a template switching event, and then the resulting fragment served as an additional template in subsequent switching events to form more rearranged fragments.

Most of the GBM patients examined in this study have both *EGFR* gain, most likely through a replication-based mechanism, and *CDKN2A/B* loss, mostly by NHEJ repair of DNA double-strand breaks. Furthermore, in all tumor types, a significant portion of the focal deletions was generated by FoSTeS/MMBIR in addition to the dominant NHEJ and alt-EJ mechanisms. This suggests that cancer genomes are likely to have more than one driving force (e.g., replication error and erroneous repair of DNA double-strand breaks) acting together in the same individual to initiate different types of rearrangements in different parts of the genome and provide advantageous mutations for cancer progression.

Table 1. *CDKN2A/B* Loss, Chromosome 10 Loss, and *EGFR*, *CDK4* Amplification in GBM Samples

	CDKN2A/B Event Types				*EGFR*								*CDK4*				
		Del/Del_ins			Event Types						Allelic	Copy	Event Types			Allelic	Copy
ID	AL	NHEJ	MMEJ	CP	AG	Del	Del_ins	Dup	CDup	CP	Amplification	Ratio	AG	CDup	CP	Amplification	Ratio
GBM0145	1	1			1		1	1		1	bi	66.9		1		mono	6.0
GBM0185	1		1		1			1		1	bi	24.8					1.0
GBM0188	1	1			1						mono	1.3					1.2
GBM0208	1	1						1		1	mono	19.1					1.1
GBM0214	1			1	1			1		1	mono	8.9			1	mono	1.5[c]
GBM0152									1		mono	61.8		1		mono	38.8
GBM0155		1		1	1			3			bi	5.7					0.9
GBM0648		1		1	1						mono	1.5[c]					1.2
GBM0786	1	1			1					1	bi	70.0					1.1
GBM0877		1	1					1		1	mono	17.9	1			mono	1.3
GBM0881[a]								1			mono	4.2					1.0
GBM1086		2										1.2					1.3
GBM1401		1[b]			1					1	bi	25.5					1.5[c]
GBM1438					1						bi	1.4			1	mono	11.6
GBM1454												1.1[c]					0.8
GBM1459		1		1		1		1		1	mono	33.3					0.8

Samples with IDs in bold are the ones in which experimental validation of *CDKN2A/B* loss, *EGFR* and *CDK4* amplifications has been performed. Copy loss/gain were predicted jointly from copy ratios and allele ratios of germline heterozygous SNPs. AL, arm-level loss; Del, deletion; Del_ins, deletion with insertion in the breakpoints; CP, complex events; AG, arm-level gain; Dup, tandem duplication, CDup, complex tandem duplication. See also Extended Experimental Procedures and Figure S4.

[a] The only sample without chromosome 10 loss.

[b] Deletion that is also involved in a copy-neutral loss of heterozygosity event which caused the loss of both copies of *CDKN2A/B*.

[c] Inconsistent copy ratio estimates between the read-depth and Affymetrix SNP Array 6.0 (http://tcga-data.nci.nih.gov/tcga/) data.

Double Minute Chromosomes and Complex Rearrangements

Double minute chromosomes (DMs) are extra circular chromosomal DNA with neither centromeres nor telomeres that can duplicate autonomously. They have been found in a variety of solid tumors as well as in leukemia (Thomas et al., 2004). *EGFR* has been shown to be amplified by DMs in glioma and glioblastoma (Vogt et al., 2004). All of the *EGFR* amplifications we identified above are likely to be DMs because the amplified fragments in DM loop structures would be predicted as tandem duplications. With paired-end sequencing, we are not able to determine if the amplifications were tandem duplications on the same chromosome or circularized as double minute chromosomes. In one patient, we identified a deletion whose breakpoints matched the tandem duplication (Figure S6G), suggesting an excision of the DNA fragment followed by circularization of that fragment, similar to the excisions of amplified DM fragments reported based on FISH (Storlazzi et al., 2010; Van Roy et al., 2006).

It was previously shown that most DMs that amplified *EGFR* in gliomas are a single fragment circularized by a microhomology-based mechanism (Vogt et al., 2004), likely FoSTeS/MMBIR (microhomology was detected at six out of seven breakpoints), and subsequently amplified by recombination to join multiple fragments into one larger circular DNA or by rolling circle replication. The copy number of the amplified region we observed is the average across many tumor cells; each cell or a subpopulation of cells could have a different number of the amplified unit. Most of the initial circularizations of DMs in neuroblastoma and small cell lung carcinoma (Storlazzi et al., 2010) were also generated by FoSTeS/MMBIR, with 23 out of 32 breakpoints showing either microhomologies or large insertions. Similar to *EGFR* amplifications in glioma, most of the *EGFR* and *CDK4* amplifications in GBM patients reported here can be explained by an initial circularization of a single DNA fragment resulting from replication error; others can be explained by the circularization of multiple DNA fragments. However, we observed more breakpoints without homology than with microhomology (Figure 3), suggesting that some of these initial circularizations were generated by FoSTeS/MMBIR but more were formed by non-homology-based replicative mechanisms.

DISCUSSION

We have reported a comprehensive catalog of somatic rearrangements in cancer, revealing the diversity in the types of somatic SVs and the mechanisms that generate them across different tumor types and individuals. Given the disruptive nature of some genomic rearrangements and their role in promoting cancer progression, precise characterization of the rearrangements and their mechanisms is crucial. While much of the work on structural variations so far has focused on their impact on genes and the mutations occurring in intergenic regions have often been considered "passenger" events, recent work by the

Please cite this article in press as: Yang et al., Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes, Cell (2013), http://dx.doi.org/10.1016/j.cell.2013.04.010

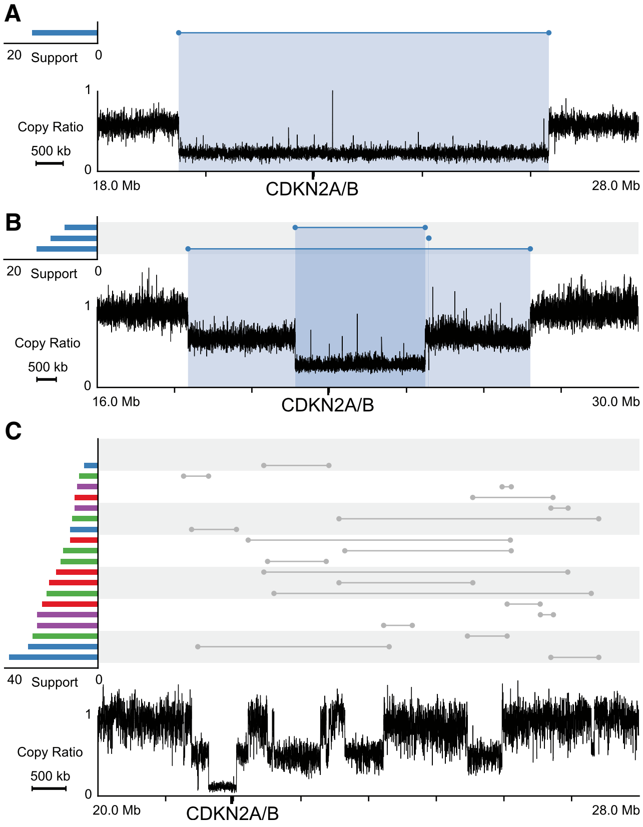


Figure 4. *CDKN2A/B* Losses in GBM Patients

(A–C) Profiles in the lower part of the plots show copy ratios (tumor versus matched normal). Above the copy ratio profiles, predicted somatic SVs are represented by lines with the breakpoints indicated by dots. SVs corresponding to a notable copy number change are colored, with the color indicating the orientation of the breakpoints. A red cluster typically suggests a tandem duplication; a blue cluster typically suggests a deletion. The number of supporting discordant read pairs for each SV is shown on the left using the same color coding. The copy-loss regions are highlighted with blue shades.

(A) GBM0208, an arm-level loss and a focal deletion.

(B) GBM1086, two focal deletions.

(C) GBM0648, complex rearrangements.

See also Figure S5.

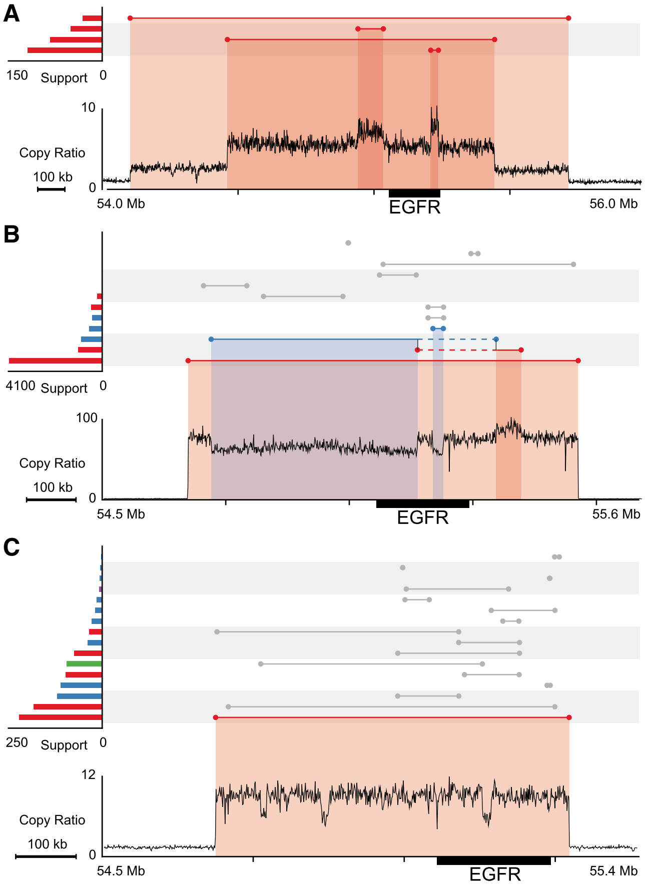


Figure 5. *EGFR* Amplifications in GBM Patients

(A–C) SVs and copy ratios are displayed as described in Figure 4. The copy-loss and gain regions are highlighted with blue and red shades, respectively.

(A) GBM0155, three tandem duplications.

(B) GBM0145, one tandem duplication and a deletion with insertion at the breakpoints. Two vertical black lines connecting two single events denote a complex deletion, which was predicted by combining two discordant read pair clusters. The solid blue and red lines represent segments that have been deleted and duplicated. The dashed lines denote a region of no copy number change.

(C) GBM0214, one tandem duplication and complex rearrangements.

See also Figure S6.

ENCODE consortium (Dunham et al., 2012) has shown that the fraction of the noncoding genome that plays a role in gene regulation is much larger than previously thought. This suggests that, in addition to a direct impact on protein structure (e.g., by fusion transcripts), other, perhaps more subtle types of misregulations may result from rearrangements that involve noncoding regions (e.g., disruption of enhancer activity or binding of a noncoding RNA). Thus, it is advantageous to know not simply whether a genomic region is amplified or not but also where the amplified segments are located. At some point in the future, improved DNA sequencing technology will accommodate much longer reads (on the order of kilobases or longer) to make reconstruction of structural alterations easier; in the meantime, innovative approaches such as the one we report here are needed to dissect the evolution of the cancer genome based on short-read data.

It is important to note that assigning mechanisms to events based on sequence features at the breakpoints is an inexact process. For instance, we found that NHEJ and alt-EJ contribute the most to focal deletions and translocations. The events generated by alt-EJ tend to have more microhomology than by NHEJ (Bennardo et al., 2008; McVey and Lee, 2008), but there is no consensus cutoff for distinguishing NHEJ and alt-EJ (Arlt et al., 2012), as both mechanisms can generate rearrangements with blunt ends, microhomology, or small insertions at the breakpoints. In addition, events generated by FoSTeS/MMBIR frequently have microhomology at the breakpoints. These ambiguities in the thresholds, however, are unlikely to materially affect our comparisons of germline and somatic events or comparisons across tumor types because we apply the same criteria to all events. We also note that SVs generated by NAHR typically

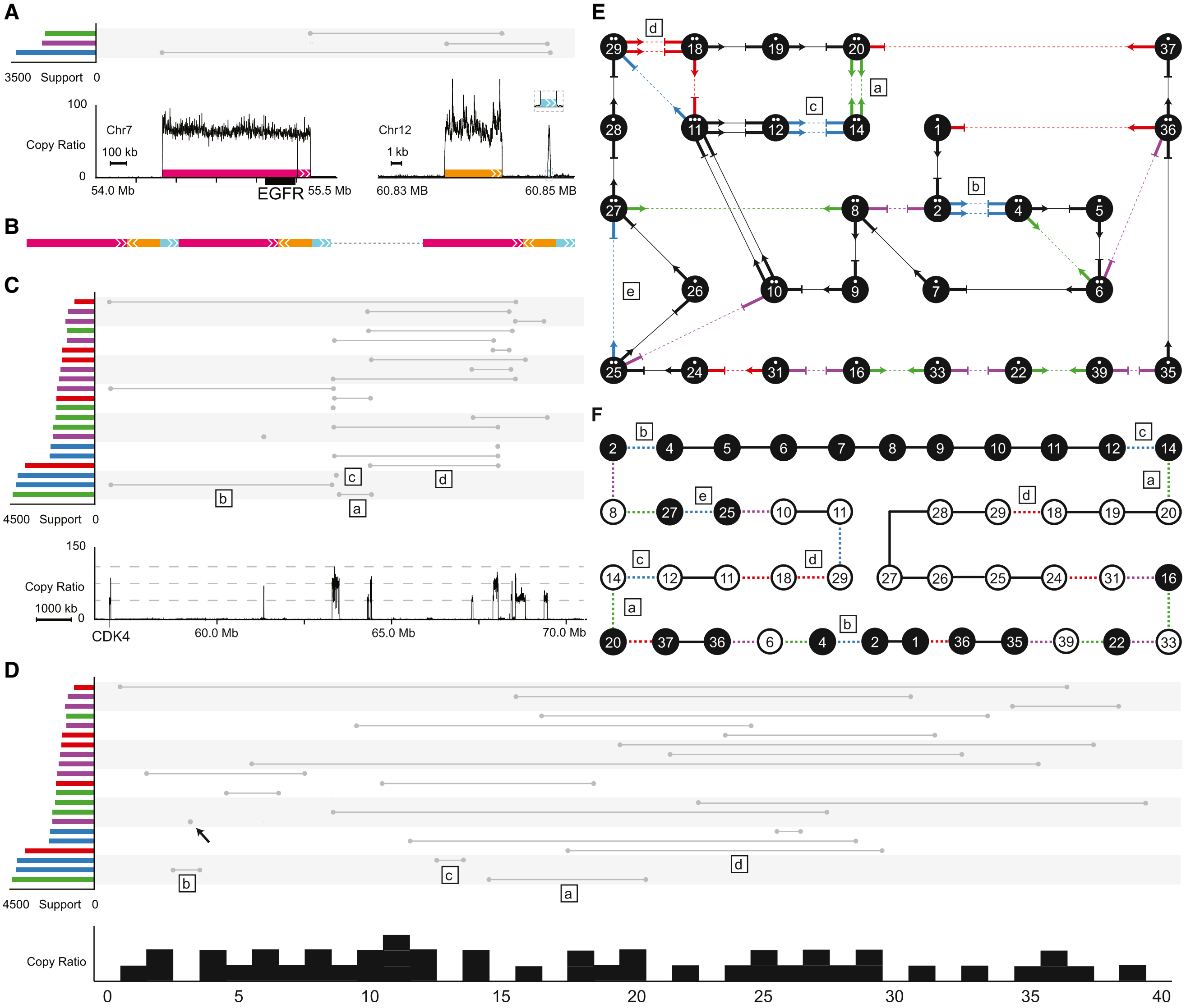


Figure 6. Amplifications of *EGFR* and Chromosome 12 in GBM0152

(A) Copy ratio and rearrangements involving *EGFR*. Colored boxes with arrows denote the amplified regions and their orientations.

(B) Diagram of the resulting rearrangements. Three segments of DNA from chromosome 7 and chromosome 12 are merged into one and tandem-duplicated.

(C) Copy ratio and somatic rearrangements on chromosome 12. The three gray dashed lines in copy ratio panel (bottom of this figure) denote copy ratios of 40, 75, and 110. The rearrangements marked by “a,” “b,” “c,” and “d” have approximately twice as many supporting discordant read pairs as other rearrangements. These rearrangements are also marked in (D)–(F).

(D) The 14 Mb region of chromosome 12 shown in (C) was segmented according to copy ratios. Each segment was rescaled and assigned an identifier from 0 to 40. The rearrangement marked with a black arrow is not involved in the amplifications of other segments on chromosome 12, but is involved in the amplification of *EGFR* on chromosome 7 as displayed in (A).

(E) Each segment in (D) is shown as a numbered node connected by arrows and lines. Black arrows connected by lines denote concordant connections. Ratios of segments are denoted by the number of dots above the segment IDs inside each node. Nonamplified segments are not shown. The connection marked with “e” (also marked in F) is a germline deletion.

(F) This diagram shows one possible solution on how segments are connected. Segments with a white background are in an inverted orientation. Colored dashed lines denote discordant connections, whereas black lines denote concordant connections.

require at least 100 bp of homology (Liskay et al., 1987; Waldman and Liskay, 1988); therefore, we could not identify these SVs generate by NAHR based on the short (≤100 bp) reads we have.

We found more microhomology-based mechanisms (alt-EJ and FoSTeS/MMBIR) for germline SVs (e.g., deletions, tandem duplications and complex events) than for somatic SVs, suggesting that those mechanisms may be suppressed in cancer cells. It is also possible that DNA breakage and replication fork stalling are more frequent in cancer cells, and a non-homology-based mechanism is the easiest way to repair. A similar trend was observed in pathogenic germline rearrangements (Chiang et al., 2012), with less microhomology at the breakpoints of pathogenic balanced translocations and inversions.

The driving forces behind large-scale genomic rearrangements have been less well characterized than those for single nucleotide alterations. In addition to the chromosome arm-level alterations (induced, e.g., by mutations in genes that maintain genome stability; Solomon et al., 2011), focal losses and gains of *CDKN2A/B* and *EGFR* in GBM patients involved distinct mechanisms acting together on the same locus in the same genome. Why multiple mechanisms act on certain regions of the genome repeatedly remains unclear. The rearrangements we observe are a snapshot of the combined effect of bias in formation and selection of the alternations in cancer genomes. It is possible that specific regions are biased toward the formation of genomic rearrangement, driven by their genomic and epigenetic features as well as their regulatory function (De and Michor, 2011a, 2011b; Fudenberg et al., 2011). For example, the recruitment of specific proteins induced by androgen can trigger DNA double-strand breaks that result in *TMPRSS2-ERG* gene fusion in prostate cancer (Haffner et al., 2010). It is also possible that alterations occur randomly for the most part and the fitness of the cell increases with certain alterations. Further studies are needed to better understand the relationships between the driving forces and their targets and how each step of the alteration confers growth advantage to the selected clones.

EXPERIMENTAL PROCEDURES

Identification of SVs Using the Meerkat Algorithm

In short, we predict SVs based on discordant read pairs and refine the precise breakpoints by looking for the reads that cover the SV breakpoint junctions. Mutational mechanisms are predicted based on homology and sequencing features at the breakpoints (Figure 1E), which is adapted from Kidd et al. (2010). See Extended Experimental Procedures for more details. The Meerkat package is available online at http://compbio.med.harvard.edu/Meerkat/.

Experimental Validation of SVs

A set of SVs predicted by Meerkat was validated by PCR. PCR primers were designed using Primer3 (http://frodo.wi.mit.edu/primer3/) to amplify the predicted SV breakpoints. The primer pairs were designed to produce a product under 10 kb for SVs of NA18507 or about 200 bp long for somatic SVs predicted in cancer samples. For NA18507, PCRs were run on genomic DNA. For somatic SVs, PCRs were run on whole-genome amplified DNA of both tumor and matched normal samples to ensure that the somatic SVs were found only in tumor but not in the matched normal sample. Whole-genome DNA amplification was performed with Sigma/Rubicon's WGA kit per manufacturer's instructions. For NA18507, five to ten SVs were randomly selected from each event type and an SV was considered validated if the predicted product across breakpoints was detectable in genomic DNA. For deletions with a large insertion in the breakpoints, large insertions, deletions with an inversion in the breakpoints, three PCRs were performed. Two PCRs were aimed at amplifying across the two breakpoints, and the third PCR was targeted to amplify the entire insertion or inversion event if the insertion or inversion was <10 kb. An event was considered validated if all three PCRs yielded products of expected sizes. If a PCR validation was not successful, two more pairs of primers were attempted.

ACCESSION NUMBERS

All primary sequences (BRCA, KIRC, and UCEC) newly reported in this paper were generated by The Cancer Genome Atlas (TCGA) Research Network and are available at CGhub (https://cghub.ucsc.edu/).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.04.010.

ACKNOWLEDGMENTS

This work was supported by grants from NIH (RC1HG005482 to P.J.P., U24CA144025 to R.K., and U24CA143845 to L.C.). We thank The Cancer Genome Atlas project for providing the whole-genome data, the Information Technology Department at Harvard Medical School for providing computational resources, and Michael Talkowski and members of the Park laboratory for a careful reading of the manuscript. L.C. is founder and/or consultant for AVEO, Metamark, Karyopharm, and Eden. R.K. is on the board of directors of AVEO, Metamark, Enlight, and KEW Group.

Received: October 4, 2012
Revised: February 28, 2013
Accepted: March 29, 2013
Published: May 9, 2013

REFERENCES

Arlt, M.F., Rajendran, S., Birkeland, S.R., Wilson, T.E., and Glover, T.W. (2012). De novo CNV formation in mouse embryonic stem cells occurs in the absence of Xrcc4-dependent nonhomologous end joining. PLoS Genet. *8*, e1002981.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., et al. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. Nature *486*, 405–409.

Bass, A.J., Lawrence, M.S., Brace, L.E., Ramos, A.H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G., Sivachenko, A., et al. (2011). Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat. Genet. *43*, 964–968.

Bennardo, N., Cheng, A., Huang, N., and Stark, J.M. (2008). Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. PLoS Genet. *4*, e1000110.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C., et al. (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214–220.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.-L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Cancer Genome Atlas Research Network. (2012a). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Cancer Genome Atlas Research Network. (2012b). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.-P., Ahmann, G.J., Adli, M., et al. (2011). Initial genome sequencing and analysis of multiple myeloma. Nature *471*, 467–472.

Chiang, C., Jacobsen, J.C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R.E., Kirby, A., Lindgren, A.M., Rudiger, S.R., et al. (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat. Genet. *44*, 390–397, S1.

Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J., and Hurles, M.E. (2010). Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat. Genet. *42*, 385–391.

De, S., and Michor, F. (2011a). DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat. Biotechnol. *29*, 1103–1108.

De, S., and Michor, F. (2011b). DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat. Struct. Mol. Biol. *18*, 950–955.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Edlund, T., and Normark, S. (1981). Recombination between short DNA homologies causes tandem duplication. Nature *292*, 269–271.

Fudenberg, G., Getz, G., Meyerson, M., and Mirny, L.A. (2011). High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat. Biotechnol. *29*, 1109–1113.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. Pathogenetics *1*, 4.

Haffner, M.C., Aryee, M.J., Toubaji, A., Esopi, D.M., Albadine, R., Gurel, B., Isaacs, W.B., Bova, G.S., Liu, W., Xu, J., et al. (2010). Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. Nat. Genet. *42*, 668–675.

Hastings, P.J., Ira, G., and Lupski, J.R. (2009a). A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. *5*, e1000327.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009b). Mechanisms of change in gene copy number. Nat. Rev. Genet. *10*, 551–564.

Haviv-Chesner, A., Kobayashi, Y., Gabriel, A., and Kupiec, M. (2007). Capture of linear fragments at a double-strand break in yeast. Nucleic Acids Res. *35*, 5192–5202.

Hillmer, A.M., Yao, F., Inaki, K., Lee, W.H., Ariyaratne, P.N., Teo, A.S.M., Woo, X.Y., Zhang, Z., Zhao, H., Ukil, L., et al. (2011). Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res. *21*, 665–675.

Hoeijmakers, J.H.J. (2001). Genome maintenance mechanisms for preventing cancer. Nature *411*, 366–374.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837–847.

Lam, H.Y.K., Mu, X.J., Stütz, A.M., Tanzer, A., Cayting, P.D., Snyder, M., Kim, P.M., Korbel, J.O., and Gerstein, M.B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat. Biotechnol. *28*, 47–55.

Lee, J.A., Carvalho, C.M., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235–1247.

Lee, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K.P., Bhatt, D., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature *465*, 473–477.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., et al.; Cancer Genome Atlas Research Network. (2012). Landscape of somatic retrotransposition in human cancers. Science *337*, 967–971.

Liskay, R.M., Letsou, A., and Stachelek, J.L. (1987). Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics *115*, 161–167.

Liu, P., Erez, A., Nagamani, S.C., Dhar, S.U., Kołodziejska, K.E., Dharmadhikari, A.V., Cooper, M.L., Wiszniewska, J., Zhang, F., Withers, M.A., et al. (2011). Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell *146*, 889–903.

Mahaney, B.L., Meek, K., and Lees-Miller, S.P. (2009). Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining. Biochem. J. *417*, 639–650.

McVey, M., and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. *24*, 529–538.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al.; 1000 Genomes Project. (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K.W., Greenman, C., et al. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184–190.

Solomon, D.A., Kim, T., Diaz-Martinez, L.A., Fair, J., Elkahloun, A.G., Harris, B.T., Toretsky, J.A., Rosenberg, S.A., Shukla, N., Ladanyi, M., et al. (2011). Mutational inactivation of STAG2 causes aneuploidy in human cancer. Science *333*, 1039–1043.

Stephens, P.J., McBride, D.J., Lin, M.-L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., et al. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005–1010.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

Storlazzi, C.T., Lonoce, A., Guastadisegni, M.C., Trombetta, D., D'Addabbo, P., Daniele, G., L'Abbate, A., Macchia, G., Surace, C., Kok, K., et al. (2010). Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. Genome Res. *20*, 1198–1206.

Sung, W.K., Zheng, H., Li, S., Chen, R., Liu, X., Li, Y., Lee, N.P., Lee, W.H., Ariyaratne, P.N., Tennakoon, C., et al. (2012). Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat. Genet. *44*, 765–769.

Thomas, L., Stamberg, J., Gojo, I., Ning, Y., and Rapoport, A.P. (2004). Double minute chromosomes in monoblastic (M5) and myeloblastic (M2) acute myeloid leukemia: two case reports and a review of literature. Am. J. Hematol. *77*, 55–61.

Van Roy, N., Vandesompele, J., Menten, B., Nilsson, H., De Smet, E., Rocchi, M., De Paepe, A., Påhlman, S., and Speleman, F. (2006). Translocation-excision-deletion-amplification mechanism leading to nonsyntenic coamplification of MYC and ATBF1. Genes Chromosomes Cancer *45*, 107–117.

Venkitaraman, A.R. (2002). Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell *108*, 171–182.

Vogt, N., Lefèvre, S.H., Apiou, F., Dutrillaux, A.M., Cör, A., Leuraud, P., Poupon, M.F., Dutrillaux, B., Debatisse, M., and Malfoy, B. (2004). Molecular structure of double-minute chromosomes bearing amplified copies of the epidermal growth factor receptor gene in gliomas. Proc. Natl. Acad. Sci. USA *101*, 11368–11373.

Waldman, A.S., and Liskay, R.M. (1988). Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol. Cell. Biol. *8*, 5350–5357.

Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat. Genet. *41*, 849–853.

EXTENDED EXPERIMENTAL PROCEDURES

Identification of Structural Variations Using the Meerkat Algorithm

Identification of Discordant Read Pairs

The BAM files are first scanned to calculate the median insert size and standard deviation for each read group (designated in the BAM files). Read groups with less than 30% uniquely mapped reads are discarded. For increased sensitivity, if one read in a pair is mapped and the other read is unmapped, the unmapped read is cleaved into two parts (35 bp from both end), each part is paired with the mapped read, and those pairs are re-mapped to the reference genome. If one read in a pair is mapped and the other read is soft-clipped (partially mapped), the clipped part is paired with the mapped read and re-mapped to the reference as a pair. Then, BAM files are scanned again to identify discordant read pairs. We call a read pair discordant when the reads are (1) mapped to different chromosomes, (2) mapped to the same chromosome but in incompatible orientations, or (3) mapped in compatible orientations but at unexpected distance (>3 × standard deviation + median insert size). Duplicate read pairs are discarded. Discordant read pairs are identified in re-mapped pairs and merged with the ones identified in the original BAM files. Four types of discordant pairs ("+−," "−+," "++," "−") are defined based on the mapping orientation of the reads. The "+−" pair is defined when reads mapped to the smaller coordinate are on the forward strand and reads mapped to the larger coordinate are on the reverse strand. Other types are defined in a similar manner.

To identify the confidence interval (defined loosely here) for a breakpoint, related read pairs must first be combined into a cluster. Two read pairs that span a breakpoint are merged into a cluster if the two reads (one from each pair) on each side of the breakpoint map to the same chromosome with the same orientation and the distance between the ends are within 3 × σ^* from μ^*, where μ^* is the difference in the median insert sizes of the read groups to which the two read pairs belong and σ^* is the standard deviation estimated by the pooled standard deviation of insert sizes for the two read groups. With the read pairs combined into clusters, the confidence interval for a discordant read pair is estimated by the mapping position of a read toward the breakpoint and $\mu + 3\,\sigma$, where μ and σ are the median and the standard deviation of insert size in the corresponding read group. Finally, the confidence interval for a discordant cluster is defined as the intersection of the confidence intervals for all read pairs in that cluster.

Typical read pair-based SV prediction algorithms depend on SV breakpoints located between a pair of reads to identify discordant read pair. If reads cover an SV breakpoint, they will be soft-clipped or unmapped; these discordant read pairs would not be identified unless handled separately. The whole-genome sequencing data from The Cancer Genome Atlas were mostly generated with libraries of insert size less than 200 bp (with the rest of the libraries with ~400 bp inserts) and read length of 75 bp or 100 bp. Thus, in most cases, SV breakpoints would be within reads rather than between them. Re-mapping unmapped or clipped portions of soft-clipped reads allows us to obtain additional discordant read pairs in order to improve sensitivity.

Then, for all unmapped or soft-clipped reads, the base pairs with quality score lower than 15 from the end of a read are trimmed. Twenty base pairs from both ends of each read are extracted to form split read pairs. Split read pairs are aligned to sequences of confidence intervals of SV breakpoints by BWA (Li and Durbin, 2009). Each SV breakpoint should be covered by at least one split read pair with one end mapped to the left and the other end mapped to the right; otherwise, the predicted SVs are discarded. For each breakpoint with split read support, the entire read corresponding to the split reads are locally aligned against the breakpoint sequence by BLAST. Hence, the precise SV breakpoint positions, the level of homology at breakpoints, and the number of unmatched base pairs can be found.

Handling of Nonuniquely Mapped Reads

Nonuniquely mapped reads (reads that can be mapped to multiple positions equally well) pose computational challenges. The approach used to handle these reads in Meerkat is similar to those in VariationHunter (Hormozdiari et al., 2009) and HYDRA (Quinlan et al., 2010). The key difference, however, is that the position of a nonuniquely mapped read is determined not only at the cluster level but also at the event level. A read pair in which both reads are nonuniquely mapped is discarded. If one read in a pair is uniquely mapped and the other read is nonuniquely mapped, up to 100 possible mapping positions are extracted and paired with the uniquely mapped read to construct pseudo-discordant read pairs. Each pair is assigned a weight of 1/*N*, where *N* is the number of possible mapping positions, for the use in the clustering step (pairs with both reads uniquely mapped are assigned the weight of 1). Reads with > 100 possible mapping positions are discarded due to a heavy computational burden. All pseudo-discordant pairs are then merged with discordant pairs in which both ends are uniquely mapped and are clustered as described above. To obtain the most parsimonious solution, an attempt is made to select the configuration with the least number of events and smallest event sizes. To produce the least number of clusters, discordant clusters are sorted by weight and the cluster with the largest weight is selected in each iteration. After selecting a cluster, all pseudo-discordant pairs belonging to that cluster are eliminated from the remaining clusters. All clusters of equal weight are maintained to be decided in the next step.

Analysis of Multiple Clusters of Discordant Reads

To the extent possible, all discordant clusters are paired with other clusters (Figure S1) to predict SV events. When a pairing does not occur, the event is simple to call: unpaired "+−" clusters are predicted as deletions (Figure S1A), unpaired "−+" clusters are predicted as tandem duplications (Figure S1F), and other unpaired clusters are predicted as translocations. Large insertions generate two discordant read pair clusters that span both insertion breakpoints. Therefore, such events can be identified by pairing two clusters (Hillmer et al., 2011) (Figures S1G, S1H, and S1J). Meerkat also attempts to pair discordant clusters to predict complex deletions—e.g., a deletion with insertion at the breakpoint (Figures S1B, S1C, and S1D). This event is typically a result of DNA double-strand break (DSB) repair by the FoSTeS/MMBIR mechanism (Hastings et al., 2009) and is widely observed in normal genomes (Conrad et al., 2010; Kidd et al., 2010; Zhang et al., 2009). Specifically, in Figure S1B, a "+−" cluster paired with a "−+" cluster is predicted as a deletion with an insertion at the breakpoint, where the insertion is from the same chromosome and with the same orientation. If a cluster can pair with multiple clusters, the event with the smallest size is selected. The event with the smallest size is also selected among clusters with equal weights. For equally-weighted clusters that cannot be paired with other clusters, intrachromosomal events with the smallest size or interchromosomal events with the least number of mismatches are selected. Meerkat also attempts to pair three discordant clusters to predict complex deletions (a detailed example illustrated with DNA sequences is shown in Figure 1). Without pairing multiple clusters, the "+−" cluster alone would be predicted as a deletion. Such a deletion would have one side of the breakpoint predicted correctly but the other side would be wrong. The "−+" cluster alone would be predicted as some other event (intrachromosomal translocation, divergent event, fold-back inversion, tandem duplication, etc.) depending on the algorithm. Similarly, an event shown in Figure S1C might be incorrectly predicted as two inversions if they are not considered together; likewise, an event shown in Figure S1D might be predicted as two interchromosomal translocations. Hence, pairing of multiple clusters is essential for characterizing the nature of SVs more accurately. A deletion with an insertion that comes from the deleted region in an inverted orientation has been previously reported (Perry et al., 2008). Such events can be called in a similar way as inversions (Figure S1E).

Analysis of Germline SVs

We downloaded Illumina paired-end sequencing data for the HapMap individual NA18507 (accession number SRA010896) and NA12878 (accession number ERA089523) from the National Center for Biotechnology Information Sequence Read Achive (NCBI SRA). The coverages were both ∼46X and the read lengths were 100 bp and 101 bp. Reads were mapped to hg18 by BWA. Meerkat was used to predict SVs with the criteria that each was supported by at least 3 discordant read pairs and at least 1 split read. The following cases were filtered out: SVs < 100 bp (small events are not in the scope of this study, they can be identified better by gapped alignment rather than from discordant read pairs) or > 1 Mb (very large germline events are rare (Mills et al., 2011) and not the main focus of our study), SVs where both breakpoints from the same discordant cluster are mapped to satellite or simple repeats (likely to be alignment artifacts), SVs with homology (or unmatched nucleotides) of > 40 bp at the breakpoints (the vast majority of such SVs are alignment artifacts if the breakpoints have extensive homology), complex events predicted from single clusters, and SVs predicted from nonuniquely mapped reads but not the smallest in size. We also used BreakDancer (Chen et al., 2009) and CREST (Wang et al., 2011) (3 supporting pairs or 3 supporting reads required) to identify deletions in NA18507 and compared to the ones identified by Meerkat from discordant read pairs without requiring split-read support. We used DGV10 (Iafrate et al., 2004) and Mills et al. (2011) as the gold standard to assess the sensitivity and specificity of the three algorithms.

Analysis of Somatic SVs in Cancer Samples

SVs were predicted by Meerkat in the cancer and the matched normal genomes in the same way as in NA18507. SVs predicted from the normal genomes were filtered in the same way as in NA18507 to generate germline variants. Candidate SVs predicted from the cancer genomes were removed if similar discordant pairs were found in any of the normal genomes. A series of filters were applied (similar to the germline case but with additional criteria). An event was removed if: the SV is smaller than 100 bp; both breakpoints from the same discordant cluster are in satellite or simple repeats; the event featured more than 40bp of homology (or unmatched nucleotides) at breakpoints; the homology (or unmatched nucleotides) at breakpoints for interchromosomal translocation is >20 bp; complex events that are predicted from only a single cluster; the event contains too many discordant pairs (>3) around the predicted breakpoints in the matched normal genome; the event contains too many nonuniquely mapped reads (>25%) around the predicted breakpoints in the matched normal genome; and if the event contains too many soft clipped reads (>3) around the predicted breakpoints in the matched normal genome. Furthermore, to increase specificity, every discordant cluster must be supported by at least 1 read pair in which both mates are uniquely mapped. If the number of supporting discordant pairs and supporting split reads combined were $\geq$ 6, such variants were considered high confidence. Manual inspection was also performed in order to reconstruct certain complex rearrangements.

Allele Ratio Test for Copy Number Changes in Cancer Samples

The identification of copy-neutral regions and copy loss/gain events can be supplemented by analysis of SNP allele ratios (Gardina et al., 2008). We first identified germline heterozygous SNPs in normal genomes using the SAMtools package (Li et al., 2009). Then, the allele ratio (major allele/minor allele) for each germline SNP was calculated from the matched tumor genome. If there is no copy number change, these ratios should be close to 1. If one copy is amplified one or more times, the increase in allele ratio will depend on the copy number and tumor purity. If both copies are amplified equal number of times, the ratio should remain close to 1. If there is one copy loss, the increase in allele ratio should also depend on tumor purity. Therefore, the copy neutral regions can be determined from examination of the baseline of allele ratio. For a given region, we tested whether there was a copy change by assessing the allele ratio of germline heterozygous SNPs in the cancer genome. For a copy-gain region, we also determined whether the amplification is monoallelic or biallelic and to what copy number it was amplified. We have developed an algorithm to jointly determine tumor purity, copy loss/gain, loss of heterozygosity, and subclones in tumor population (manuscript in preparation) based on the above rationale (see examples in Figure S4).

SUPPLEMENTAL REFERENCES

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Gardina, P.J., Lo, K.C., Lee, W., Cowell, J.K., and Turpaz, Y. (2008). Ploidy status and copy number aberrations in primary glioblastomas defined by integrated analysis of allelic ratios, signal ratios and loss of heterozygosity using 500K SNP Mapping Arrays. BMC Genomics *9*, 489.

Hormozdiari, F., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. *19*, 1270–1278.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A., et al. (2008). The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet. *82*, 685–695.

Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E., Mell, J.C., and Hall, I.M. (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Res. *20*, 623–635.

Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L., Ma, J., Rusch, M.C., Chen, K., Harris, C.C., Ding, L., et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat. Methods *8*, 652–654.

Xi, R., Hadjipanayis, A.G., Luquette, L.J., Kim, T.M., Lee, E., Zhang, J., Johnson, M.D., Muzny, D.M., Wheeler, D.A., Gibbs, R.A., et al. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc. Natl. Acad. Sci. USA *108*, E1128–E1136.

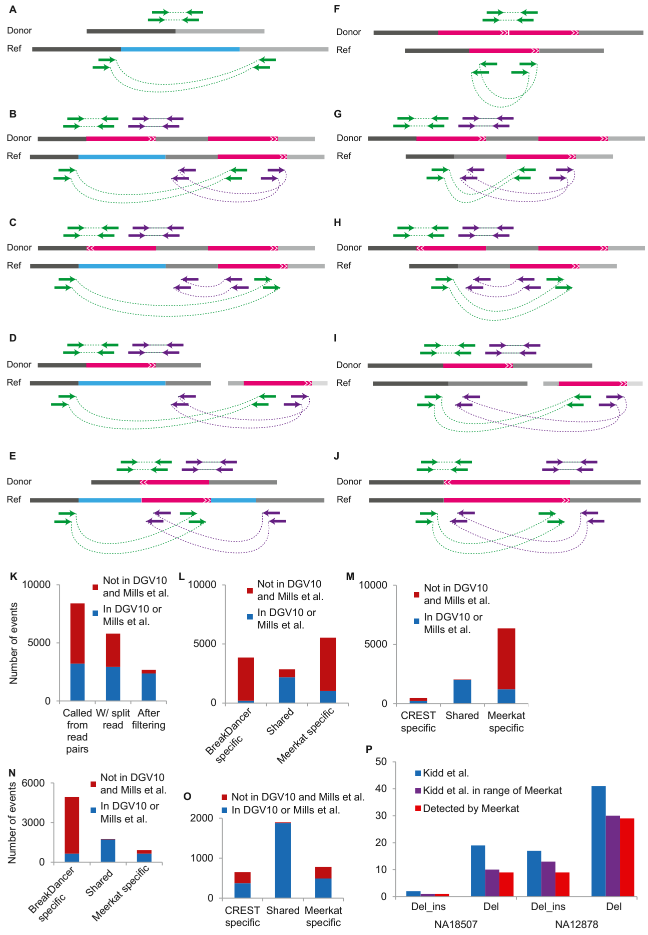


Figure S1. Analysis of Discordant Read Pair Configurations and Comparison of Predicted SVs in the HapMap Individual NA18507 to Those Reported in DGV10 and Mills et al. (2011), Related to Figure 1

(A–J) Blue lines on the reference genome represent deleted segments in the donor genome; red lines represent inserted or inverted segments. Orientations of the sequences are denoted by arrows within the red lines. Green and purple arrows connected by dashed lines represent read pairs. We use “+” and “−” below to denote reads in the positive and negative orientations, respectively.

(A) Deletion: “+−” cluster.

(B) Deletion with insertion from the same chromosome in the same orientation: “+−” paired with “−+” cluster.

(C) Deletion with insertion from the same chromosome in the opposite orientation: “++” paired with “−” cluster.

(D) Deletion with insertion from a different chromosome in the same orientation (in a different orientation not shown): “+−” paired with “−+” cluster (or “++” paired with “−” cluster for different orientation).

(E) Deletion with inversion: “++” paired with “−” cluster.

(F) Tandem duplication: “−+” cluster.

(G) Insertion from the same chromosome in the same orientation: “+−” paired with “−+” cluster.

(H) Insertion from the same chromosome in the opposite orientation: “++” paired with “−” cluster.

(I) Insertion from a different chromosome in the same orientation (in a different orientation not shown): “+−” paired with “−+” cluster (or “++” paired with “−” cluster for a different orientation).

(J) Inversion: “++” paired with “−” cluster.

(K–O) After identifying reads cover the breakpoint junctions (split read support) and applying additional filters, the sensitivity and specificity of Meerkat are both better than BreakDancer and CREST if events reported in DGV10 and Mills et al. (2011) are taken as ground truth. Blue bars indicate number of events reported in DGV10 or Mills et al. (2011), whereas red bars indicate number of events not reported in DGV or Mills et al. (2011).

(K) Meerkat calls.

(L) Meerkat calls from read pairs without split read support compared to BreakDancer calls.

(M) Meerkat calls from read pairs without split read support compared to CREST calls.

(N) Meerkat calls with split read support and additional filters compared to BreakDancer calls.

(O) Meerkat calls with split read support and additional filters compared to CREST calls.

(P) Comparisons of complex and simple deletions to Kidd et al. (2010). For complex deletions (del_ins), Meerkat can only detect ones with insertion > 100 bp or ≤ 40 bp. For simple deletions (del), Meerkat can only detect ones with homology at breakpoint ≤ 40 bp.

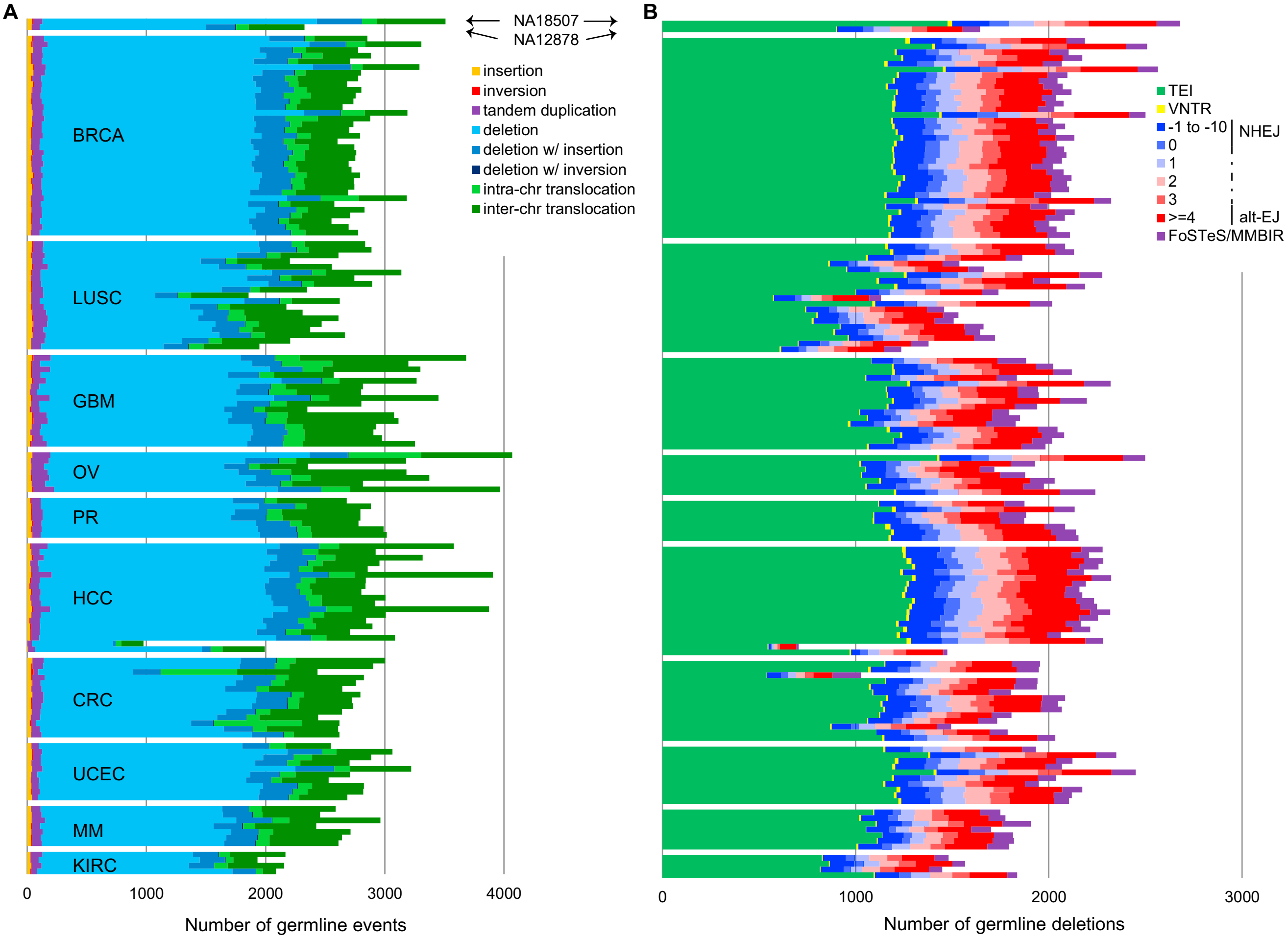


Figure S2. Spectrum of Germline SV Types and Mechanisms, Related to Figure 2

(A and B) Two HapMap genomes NA18507 and NA12878 are shown on top. Patients are in the same order as in Figure 2A.

(A) Frequencies of germline SV types.

(B) Frequencies of germline deletion mechanisms.

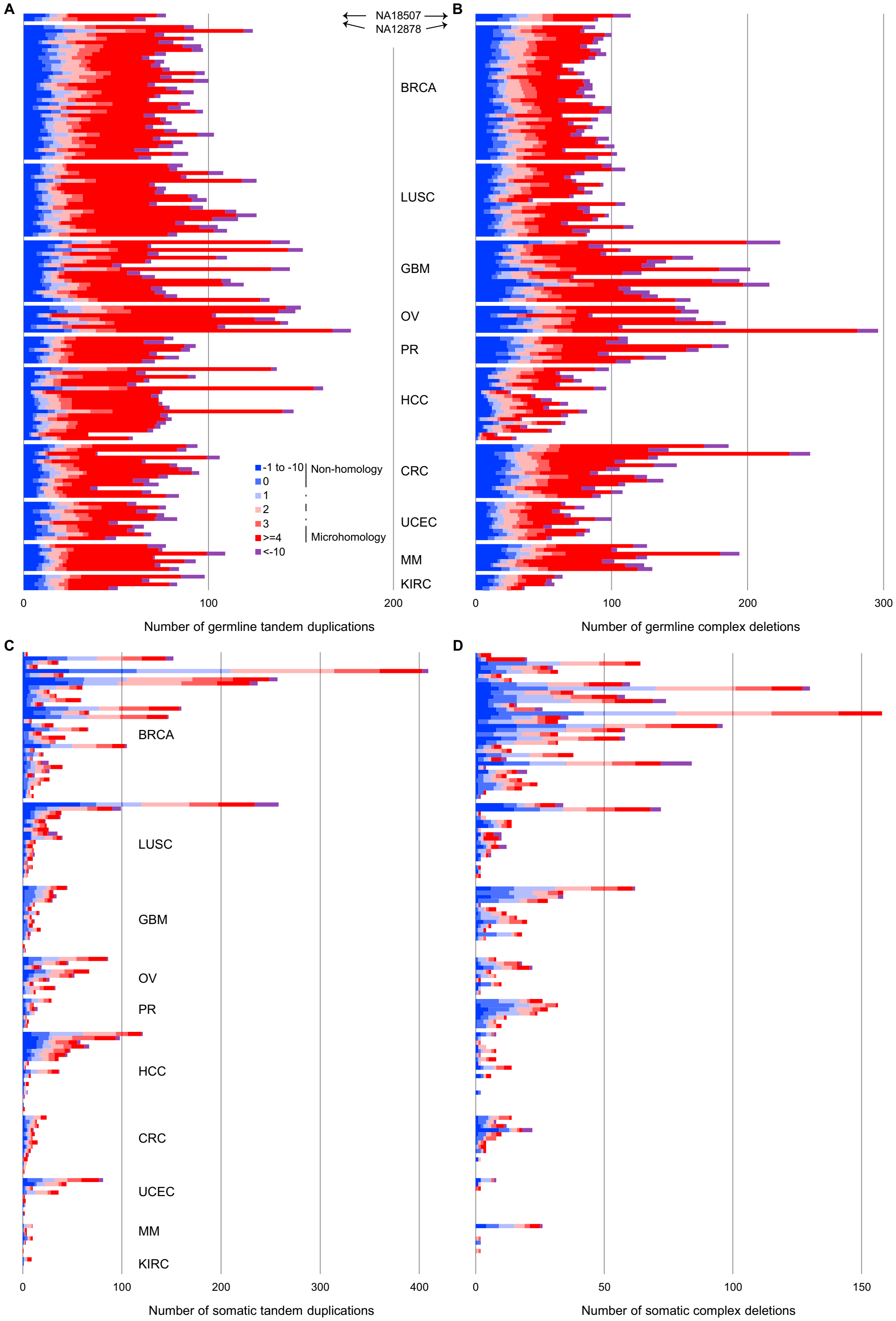


Figure S3. Homology of Breakpoints for Germline and Somatic Tandem Duplications and Complex Deletions, Related to Figure 3

(A–D) Two HapMap genomes NA18507 and NA12878 are shown on top for germline events. Patients are in the same order as in Figure 2A. Sequence homologies in base pairs are shown for each breakpoint as a positive number. A blunt end has a homology of 0 bp. Small insertions with unknown source are shown as negative numbers. (A) germline tandem duplications, (B) germline complex deletions, (C) somatic tandem duplications and (D) somatic complex deletions.

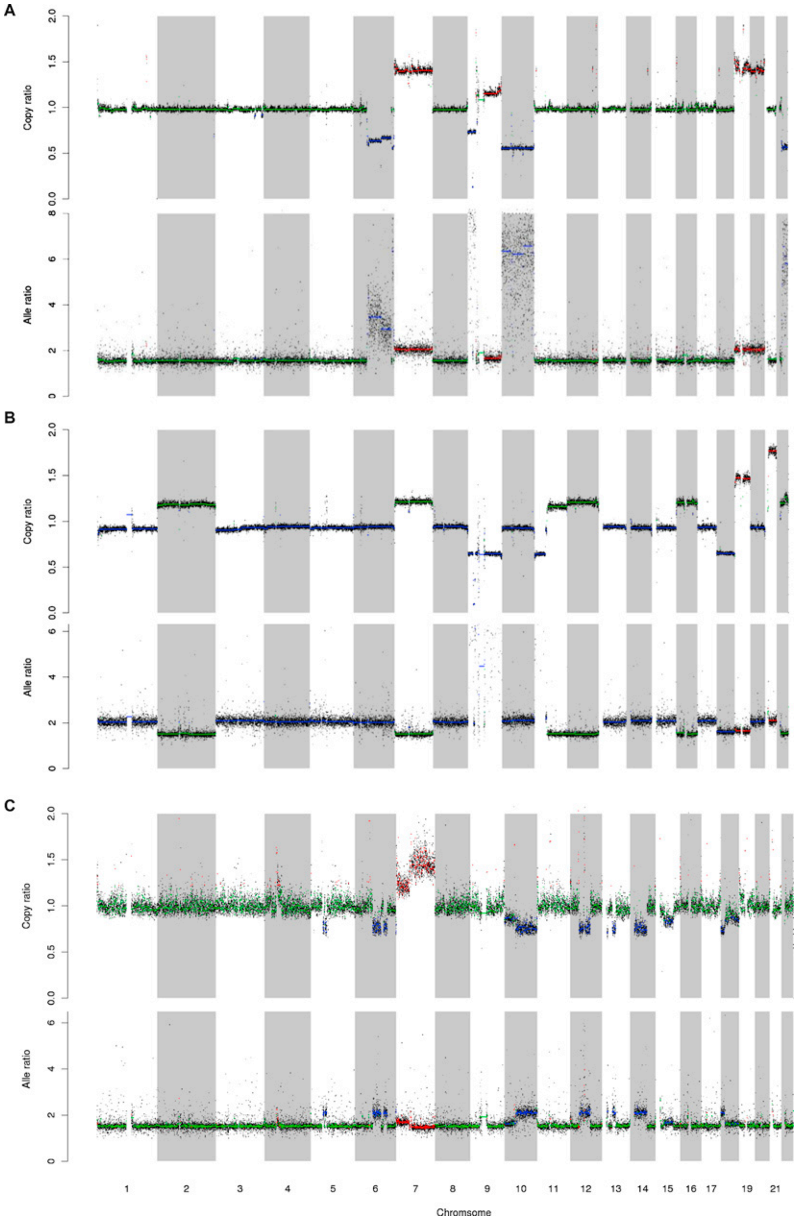


Figure S4. Predicting Copy Number Alterations by Copy Ratio and Allele Ratio, Major Allele/Minor Allele, Related to Table 1

(A–C) Black dots denote the average copy ratio on upper panel and average allele ratio in lower panel per 100 kb bin. Colored lines denote segments that are predicted to have the same copy ratio by BIC-seq (Xi et al., 2011), which merges neighboring bins that have similar copy ratios based on a statistical criterion. Green: copy neutral regions, blue: copy loss regions, red: copy gain regions. The copy-neutral regions should have allele ratios of exactly 1. However, due to the variation of data, the ratio of major allele to minor allele is greater than 1.

(A) GBM0185. The copy-neutral regions have allele ratios of about 1.5; Chr7 has one copy gain and the allele ratios in the region are about 2. Chr10 has one-copy loss and the allele ratios are between 6 and 7, which are affected by the proportion of cells harboring the copy loss (86% of cell in this case).

(B) GBM1086. The copy-neutral regions have a copy ratio of 1.15. About 90% of the cells have one-copy loss on Chr9, which has allele ratios larger than 10. About 40% of the cells have one-copy loss at Chr1, Chr3, Chr4, etc, which have allele ratios of about 2. Although those chromosomes have copy ratios about 0.93, which are typically not considered as copy loss, the allele ratios are clearly distinct from copy-neutral chromosomes, thus allowing us to identify them as copy loss. The copy ratio of copy-neutral regions is not centered at 1 but at 1.15 and only a small portion (~40%) of cells have lost Chr1, Chr3 and Chr4, therefore, those regions could generate copy ratios very close to 1. About 40% of the cells have a two-copy loss at Chr18 whose allele ratios are almost the same as copy-neutral chromosomes.

(C) GBM1438. Most of the short arm of Chr7 has one copy gain which gives allele ratios significantly different from copy-neutral chromosomes. The long arm and a small portion of short arm of Chr7 have a two-copy gain and both parental copies are amplified. Therefore, the allele ratios are the same as copy neutral chromosomes.

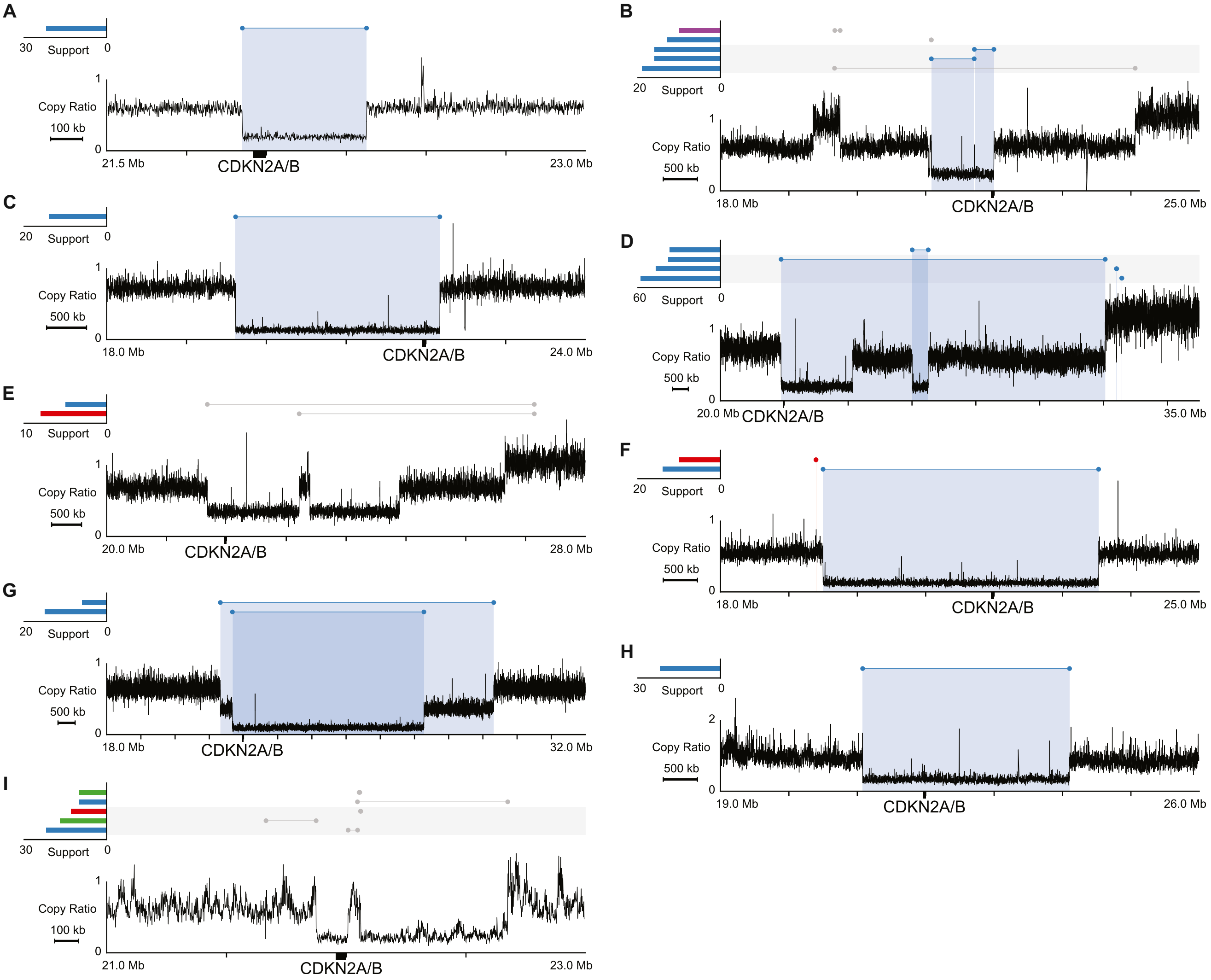


Figure S5. *CDKN2A/B* Loss in Other GBM Patients, Related to Figure 4

(A–I) Above the copy ratio profiles, predicted somatic SVs are represented by lines with the breakpoints noted by dots. SVs corresponding to a notable copy number change are colored, with the color indicating the orientation of the breakpoints. A red cluster typically suggests a tandem duplication event. A blue cluster typically suggests a deletion event. The number of supporting discordant read pairs for each SV is shown on the left using the same color-coding. (A) GBM0145, (B) GBM0155, (C) GBM0185, (D) GBM0188, (E) GBM0214, (F) GBM0786, (G) GBM1086, (H) GBM1401 and (I) GBM1459.

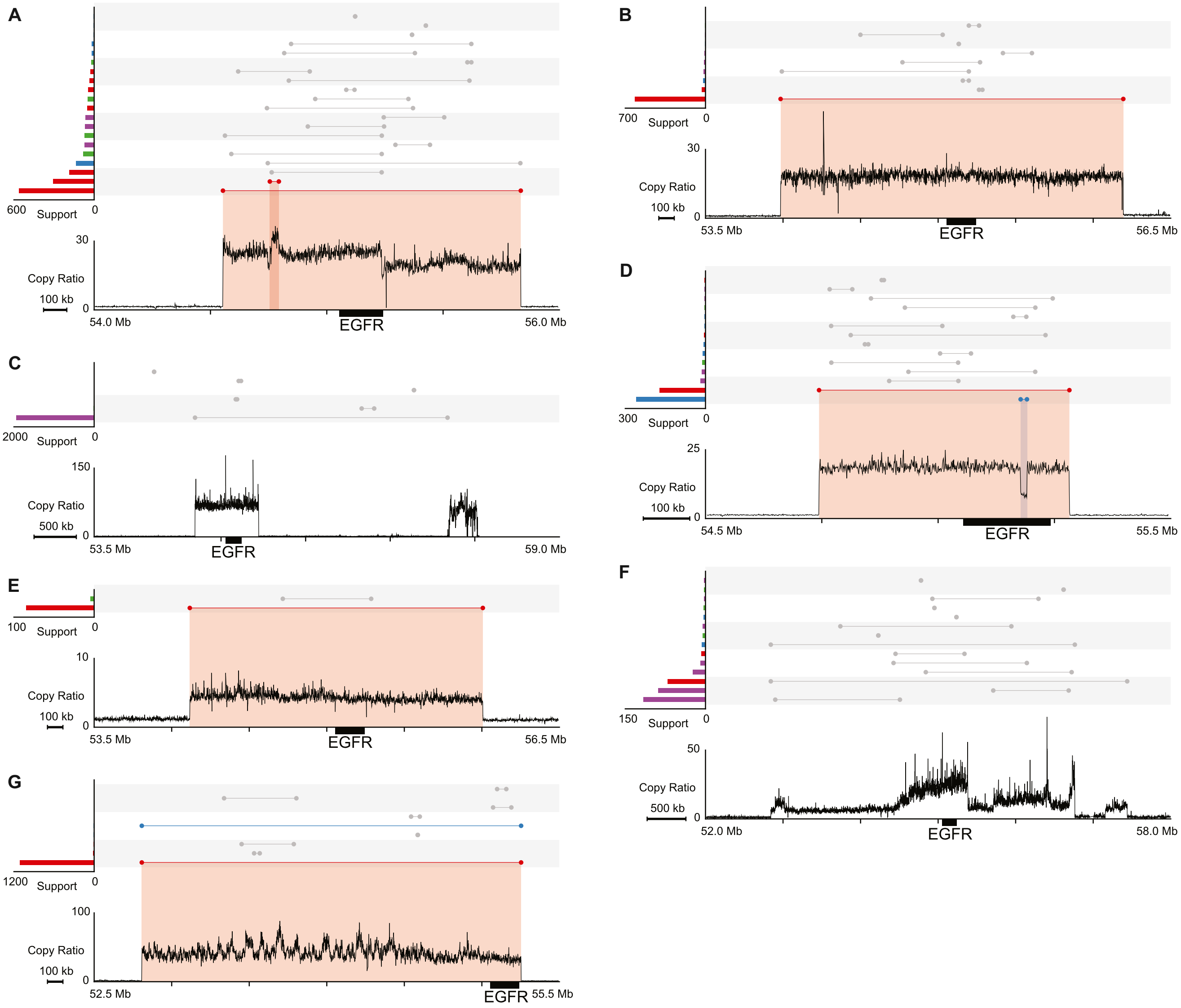


Figure S6. *EGFR* Amplifications in Other GBM Patients, Related to Figure 5

(A–G) See Figure S5 for description. (A) GBM0185, (B) GBM0208, (C) GBM0786, (D) GBM0877, (E) GBM0881, (F) GBM1401 and (G) GBM1459.