Supplementary Figure Legends

Supplementary Figure 1. Histograms showing the inter-breakpoint distance at different variant classes. For each breakpoint, the inter-breakpoint distance is defined as the distance to the nearest somatic breakpoint from the same tumor sample. From top, the left panels show the entire set of somatic breakpoints, simple breakpoints, and breakpoints found in clusters. The right panel shows different classes of breakpoint clusters including mild complex genomic rearrangements (CGRs) with <10 breakpoints, extreme CGRs with >9 breakpoints, and stepwise rearrangements. Beneath the label in the top right corner of each plot we state the median inter-breakpoint distance for each class. Note that each bar on the x-axis increases an order of magnitude, and that the rightmost bar corresponds to breakpoints that do not have a nearest breakpoint on the same chromosome.

Supplementary Figure 2. Expanded version of Figure 2B showing additional controls for breakpoint cluster identification. *(A)* From left, we show the results clustering of tumor-specific and normal-specific breakpoints, as in Figure 2B. However, we also include a third set of "individual-specific" variants shown at right, which are germline variants that were found in a single tumor-normal pair. Note that some *bona-fide* individual-specific breakpoint clusters are expected to exist given the existence of complex germline SV. *(B)* From left, we show the results of the Monte-Carlo simulation entailing random shuffling of tumor specific breakpoints, as well as the random sampling of 1000 Genomes deletion breakpoints, as in Figure 2B. Here, we also include results from randomly sampling from the set of validated germline breakpoints and the set of "rare" individual-specific breakpoints unique to a single tumor-normal pair. In contrast to the rightmost section of panel A, where we simply use the actual set of individual-specific breakpoints identified in each tumor-normal pair, in part B we randomly sample from the entire set of individual-specific breakpoints so as to match the number of tumor-specific somatic breakpoints observed in each tumor.

Supplementary Figure 3. Association between breakpoints and genome annotations. We assessed whether CGR breakpoints are enriched in SV hotspot regions such as segmental duplications or fragile sites, or at repetitive elements known to cause read-mapping artifacts. Each row is a breakpoint class, as shown at left, and each column is a genome annotation. For each comparison we counted the observed number of overlaps between breakpoints and the above genome annotations and calculated

an enrichment score relative to a Monte Carlo simulation, as defined at the top of the figure. When the observed number of intersections is what is expected by chance, the enrichment score will be near zero. When the number of intersections is higher than expected by chance, the score will be greater than zero. Scores less than zero reflect cases where less intersections were observed than expected by chance.

Supplementary Figure 4. Correlation of CGRs with copy number alterations (CNAs). *(A)* Plot showing the number of SV breakpoints detected by HYDRA-MULTI (x-axis) versus the number of CNA change-points detected by read-depth analysis for each breakpoint cluster. In the upper left corner the correlation coefficient is shown, as calculated using the MATLAB *corrcoef* function. Note that each duplication or deletion detected by both methods will generate one HYDRA-MULTI breakpoint but two CNA change-points. *(B)* A zoomed-in version of A. *(C)* Table showing the overlap between CNAs and different classes of breakpoint clusters. From left, shown are the number of breakpoint clusters, the number of HYDRA-MULTI breakpoint calls, the number of CNAs found within 50kb of a cluster, the number of clusters with at least one CNA, the percentage of clusters found with at least one CNA, and fold enrichment calculated relative to the mean overlap detected in a Monte Carlo simulation in which breakpoint clusters were randomly shuffled 100 times.

Supplementary Figure 5. CIRCOS plots of mild one-off CGRs composed of 3-9 breakpoints and predicted to result from a single complex mutation. Only the chromosome(s) and breakpoints involved in the rearrangement are shown. Chromosome coordinates increase in the clockwise direction. The chromosome name is indicated outside the circle. The outermost track is the cytogenetic band, with the centromeres colored red. Moving inward, the second track is COSMIC cancer genes. Next is a plot showing the copy number profile obtained from read-depth analysis. This profile includes germline CNVs and somatic CNAs. Blue dots are the normalized read-depth, represented as a Z-score. The red line plotted on top of the blue dots is segmented read-depth data. The Y-axis limits correspond to the median Z-score plus or minus 7.5 median absolute deviations. The lighter gray track inside of the read-depth track corresponds to somatic CNA change-points. Rearrangements are depicted as lines connecting points on the circular chromosome(s). Deletion breakpoints are shown in red, duplications in green, and inversions in blue. Note that these breakpoint classes are defined by the relative orientation of the joined genomic segments, and may not actually involve deletion or duplication of

sequence.

Supplementary Figure 6. CIRCOS plots of extreme one-off CGRs composed of at least 10 breakpoints, following the conventions outlined in the legend for Supplementary Fig. 5.

Supplementary Figure 7. CIRCOS plots of stepwise breakpoint clusters following the conventions outlined in the legend for Supplementary Fig. 5.

Supplementary Figure 8. CNA state clustering results for mild one-off CGRs. Each red dot is the predicted copy number value from one of the two CNA segment pairs that make up a CNA change-point. Change-point values are shown in sorted order. Blue lines plotted on top of change-point values indicate which values were clustered together. At left the Y-axis is determined by the minimum and maximum change-point value. At right the Y-axis is shown from predicted copy number of 0 to 10.

Supplementary Figure 9. CNA state clustering results for extreme one-off CGRs (chromothripsis) following the conventions outlined in the legend for Supplementary Fig. 8.

Supplementary Figure 10. CNA state clustering results for stepwise breakpoint clusters following the conventions outlined in the legend for Supplementary Fig. 8.

Supplementary Figure 11. CIRCOS plots for the 16 breakpoint clusters involving multiple amplifications, following the conventions outlined in the legend for Supplementary Fig. 5.

Supplementary Figure 12. CIRCOS plots of the entire genome for each of the 64 tumors, following the conventions outlined in the legend for Supplementary Fig. 5.

Supplementary Table 1. Summary statistics for each dataset analyzed in this study. A key describing the columns is included as the first sheet in the excel file. All dataset statistics were empirically determined using 50 million reads extracted from the position sorted BAM file, excluding the first 30 million reads.

Supplementary Table 2. The high confidence tumor-specific breakpoints. A key describing the columns is included as the first sheet in the excel file. Note that only breakpoints that were successfully assembled will have the data for the columns describing the validating contig, and only those that were successfully genotyped by alignment of reads to breakpoints will have data in the allele frequency field. Empty fields are indicated with "NA". All coordinates are from NCBI Build 37 (1000 Genomes Version) of the human genome.

Supplementary Table 3. The germline control breakpoints. A key describing the columns is included as the first sheet in the excel file. Note that only breakpoints that were successfully assembled will have the data for the columns describing the validating contig, only those that were successfully genotyped by alignment of reads to breakpoints will have data in the allele frequency field. Empty fields are indicated with "NA". All coordinates are from NCBI Build 37 (1000 Genomes Version) of the human genome.

Supplementary Table 4. All high confidence breakpoints. A key describing the columns is included as the first sheet in the excel file. Note that only breakpoints that were successfully assembled will have the data for the columns describing the validating contig, only those that were successfully genotyped by alignment of reads to breakpoints will have data in the allele frequency field. Empty fields are indicated with "NA". All coordinates are from NCBI Build 37 (1000 Genomes Version) of the human genome.

Supplementary Table 5. A summary of the breakpoint clusters identified in this study. A key describing the columns is included as the first sheet in the excel file.

Supplementary Table 6. The somatic CNA change-points identified in this study. A key describing the columns is included as the first sheet in the excel file.

Supplementary Table 7. Association of CNA change-points with HYDRA-MULTI breakpoint calls, broken down by breakpoint class and size. Overlap between change-points and breakpoints calls is defined as within 10kb. There are six sheets in this excel file representing all somatic breakpoints, simple SV breakpoints, all breakpoints found in clusters , both mild (<10 breaks) and extreme (>9 breaks) CGR

breakpoints judged to be due to complex one-off mutations, and breakpoints at stepwise clusters judged to be due to progressive mutation. Each sheet has exactly the same format. From top, deletion, tandem duplications, inversions and inter-chromosomal rearrangement breakpoints are shown separately and broken down into 5 size classes, as labeled in the left-most column. Note that the deletions, duplications and inversions larger than 1mb correspond to the intra-chromosomal rearrangement class shown in other figures and tables. From left the columns show the total number of breakpoints (Total), the number that are associated with a CNA (CNA Assoc.), the percentage that are associated with a CNA (% CNA Assoc.), the mean number of breakpoints that are associated with a CNA a Monte-Carlo simulation shuffling breakpoint coordinates 100 times within uniquely mappable genomic regions (RandomMean) (see **Supplementary Methods**), the standard deviation from random shuffling (RandomStd), the number of standard deviations that the observed CNA association differs from the mean expected value determined by random shuffling (Z-Score), and the number of observed CNA associations divided by the expected (FoldEnrichment).

Supplementary Table 8. Assembly-based validation results broken down by dataset and breakpoint class. The tumor sample names are shown at the far left. For each breakpoint class there are three columns corresponding to the number of validated HYDRA-MULTI breakpoint calls, the total number of calls, and the raw validation rate. Unlike the validation rate shown in Figure 1, this validation rate is not corrected for the efficiency of *de novo* assembly.

Supplementary Table 9. Assembly-based validation results broken down by breakpoint class (columns) and size (rows). Note that the deletions, tandem duplications and inversions larger than 1mb correspond to the intra-chromosomal rearrangement class used in other figures and tables. *(A)* The raw assembly-based validation rate, calculated as validated breakpoints divided by total. *(B)* The validation rate corrected for the efficiency of assembly-based validation using 1000 Genomes deletions (76.8%). *(C)* The false discovery rate (FDR) corrected as in part B *(D)* The number of breakpoints in each class.