Supplementary Materials

Supplementary Table 1: Sequencing statistics

MCF7 96-BAC Pools	Fosmid Clones	Mapped Sanger Fosmid Ends	Fosmids Spanning Breakpoints	Raw Pyroseqs	Mapped Pyroseqs
plate1	10,450	2,532 (**24.2%**)	546	300,862	197,815 (**65.8%**)
plate2	12,635	2,796 (**22.1%**)	440		
plate3	10,070	2,510 (**24.9%**)	442	508,381	367,648 (**72.3%**)
plate4	9,690	2,587 (**33.1%**)	623		
plate5	8,550	2,901 (**38.7%**)	405	457,940	319,801 (**69.8%**)
plate6	8,026	20 (**0.3%**)	3		

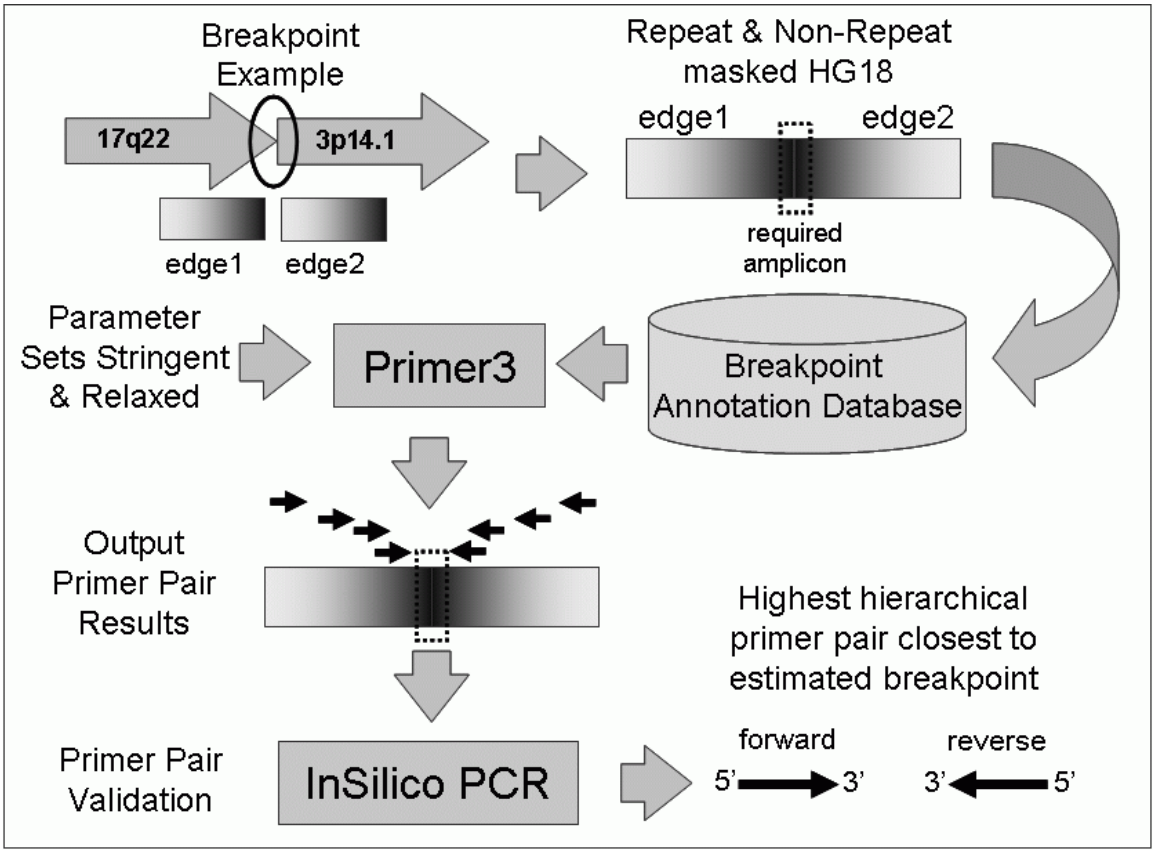
BLAT parameters used for mapping

Fast sequence search command line tool BLAT v. 23 was used to map Sanger-derived fosmid end sequences (FES) and the BAC pool derived 454 pyrosequences. The gfServer parameters for the Sanger-derived sequences are: tileSize=11, minMatch=2, maxGap=2; and the gfServer parameters for the 454-derived pyrosequences are: tileSize=8, minMatch=1, repMatch=65536. The gfClient parameters for the Sanger-derived sequences are: minScore=20, minIdentity=0, maxIntron=50; and the gfClient parameters for the 454-derived pyrosequences are: minScore=10, minIdentity=0, maxIntron=20.

Sanger-derived FES BLAT mappings are filtered such that only the highest 3% of scoring hits are retained where that set does not exceed seven members. The 100bp average 454-derived pyrosequence BLAT mappings are filtered more stringently such that only the highest 2% of scoring hits are retained where that set does not exceed five members.

It is important to note that these sequence filters do not eliminate false positive rearrangement detection due to mapping to repetitive DNA elements; this need is fulfilled by the criteria to unambiguously call breakpoints given multiple bridging FESs. The fosmid clone coverage of the originating BAC pool is approximately 24X, thus allowing multiple FESs to span any given breakpoint. In order to report a breakpoint, there must exist, within the population of ESPs that bridge an aberrant join, a uniquely and maximally mapped FES or BES pair.

Primer design pipeline



Supplemental Figure 1: Primer Design Pipeline

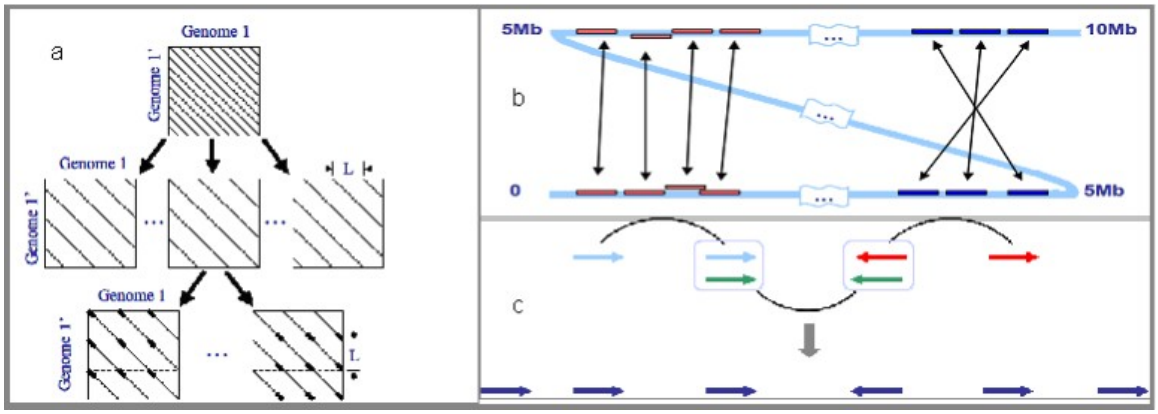
Stringent Primer3 Parameters
PRIMER_MIN_SIZE=20
PRIMER_OPT_SIZE=25
PRIMER_MAX_SIZE=27
PRIMER_MIN_TM=60.0
PRIMER_OPT_TM=60.0
PRIMER_MAX_TM=65.0
PRIMER_GC_CLAMP=1
PRIMER_MIN_GC=45.0
PRIMER_MAX_GC=60.0
PRIMER_MAX_DIFF_TM=5.0

Relaxed Primer3 Parameters
PRIMER_MIN_SIZE=18
PRIMER_OPT_SIZE=22
PRIMER_MAX_SIZE=26
PRIMER_MIN_TM=60.0
PRIMER_OPT_TM=64.0
PRIMER_MAX_TM=68.0
PRIMER_GC_CLAMP=0
PRIMER_MIN_GC=30.0
PRIMER_MAX_GC=70.0
PRIMER_MAX_DIFF_TM=10

MCF-7 aberrant join annotations are assembled such that sufficient upstream and downstream sequence straddles the breakpoint for multiple primers designs to be considered. Primer3 was employed to design PCR primers across novel aberrant join annotations. Resulting breakpoint spanning primer designs are validated against InSilico PCR such that no primer set is allowed to produce an amplicon less than 40Kb when applied to the reference human genome.

Identification of Low Copy Repeats

The method of identifying Low Copy repeats (LCRs) is illustrated in **Supplemental Figure 2**. In the following we describe the specific steps in more detail.



Supplemental Figure 2: Identification of Low Copy repeats**:** a(Step 1): PASH; b(Step 2): Reciprocal filtering and merging; c(Step 3): Linking and clustering

Step 1: PASH (**Supplementary Figure 2.a**)
In order to predict LCRs in the human genome both inter- and intra-chromosomally, we start identifying similarity pieces in the genome by comparing the current version human genome sequence (build 36, Mar. 2006) against itself using Pash. Pash is a computer program for efficient, parallel, all-against-all comparison of very long DNA sequences which has proven its efficiency in the application of mammalian genomes comparison and whole-genome shotgun sequencing reads mapping. Pash implements Positional Hashing, a parallelizable method for sequence comparison based on k-mer representation of sequences. As Figure 2.a. illustrates, this method divides the problem of whole genome comparison into groups of comparison diagonals, all of which can be processed in parallel. Pash does not require high-copy repeat (HCR) masking and is therefore well suited to detect LCRs that are in fact frequently composed to a large degree from ancient HCRs. Rather that macking repeats, Pash uses k-mer frequency information to ignore k-mers that are overrepresented because of their presence in the high copy repeats.

Step 2: Merging and clustering (**Supplementary Figure 2.b**)
The similarities detected by Pash are post-processed by applying a “reciprocal best match” filter. This filter ensures that for each pairwise similarity reported, each of the two sequences must appear on the other’s list of top matches. The filter is adjusted to keep the best match for each fragment, but allows mapping multiple fragments to the same genomic location, thus allowing for multiple duplication events. This filter implicitly uses

the full set of sequences as positive controls to increase the specificity of anchoring and to reduce the number of false positive matches.

The filtered list of matching sequences then go through a merge step. In this step, multiple fragments close to each other in genomic location are aggregated into one chunk if their matching partners are also located within certain range, and if the Pash similarity score density (=score/chunk length) exceeds certain threshold. The score density is used here because the score of the merged chunk is strictly the sum of its members' scores, and there could be thousands of extremely low-score members or just several members with very high scores, which would otherwise result in the same merged-score. As **Supplementary Figure 2.b**. shows, positive ordering (direct) mapping and negative ordering (reversed) mapping were both allowed.

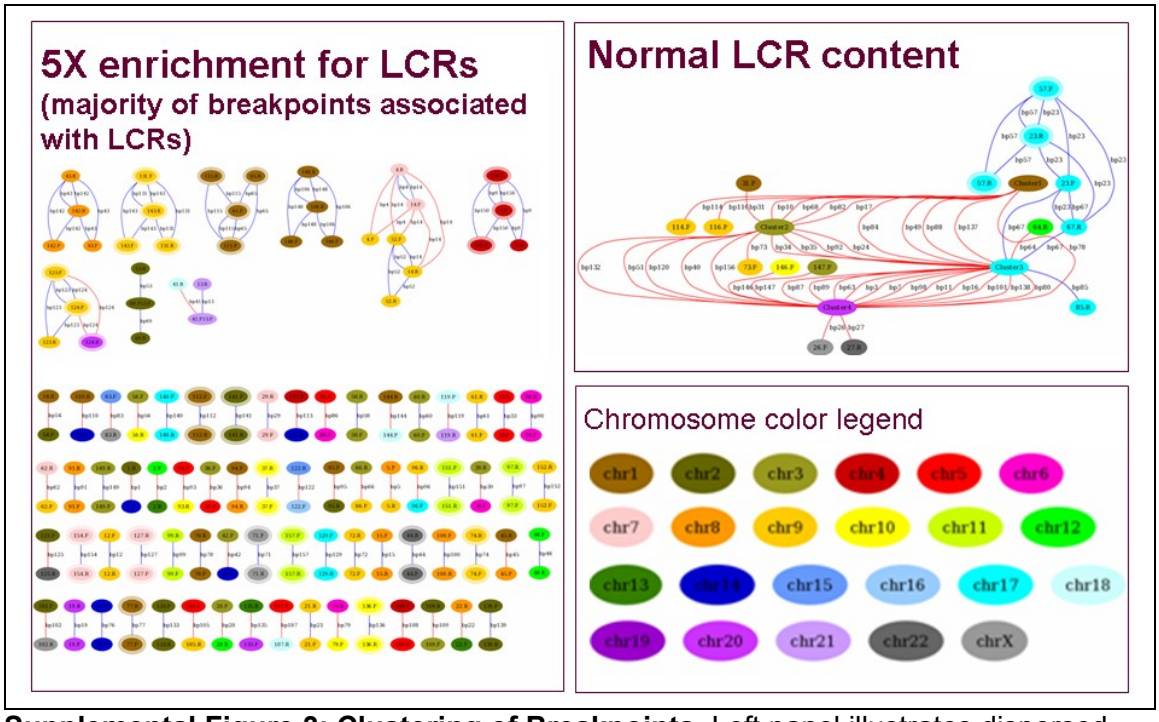
Step 3: Linking and Clustering: (**Supplementary Figure 2.c**)

Up to this point, we have identified pairwise similarities across the genome. We now need to cluster these pairwise LCRs into homologous groups according to their k-mer features. The clustering is based on two criteria: first, k-mer content similarity, which is measured by

$$[\text{No. of kmerDiff} + \log(1 + \text{sizeDiff})] / [(\text{No. of kmersInBothSets})]$$

secondly, there must be positional overlapping between members from different pairs. This step is applied recursively to all the paired up segments until all of them have been compared and grouped into a cluster of related LCRs..

Breakpoint Cluster Analysis



Supplemental Figure 3: Clustering of Breakpoints. Left panel illustrates dispersed breakpoints, which are enriched for LCRs and the top right panel illustrates clustered

breakpoints which are not. The nodes indicate chromosomal loci and lines connecting the nodes indicate fosmid clones bridging the loci. Chromosome color legend is in the bottom panel on the right.

Calculating Recurrent Copy Number at 157 Somatic Breakpoints

Copy number data was gathered as described below, then regions of amplification were extracted and intersected with all breakpoints, using a radius of 10,000 base pairs. Copy number variation at clustered breakpoints was compared to that at dispersed breakpoints.

Supplementary Table 2: Breakpoint Table

See file “SuppTable2-BreakpointTable”.

Calculating Recurrent Copy Number and Expression Change for 79 Breakpoint-Associated Genes

Normalized Affymetrix HG-U133A microarray data was obtained from Neve, et al. and fold change was calculated with respect to the HBL100 normal breast tissue cell line. If a gene's fold change was more than one standard deviation from the mean, it was considered differentially expressed.

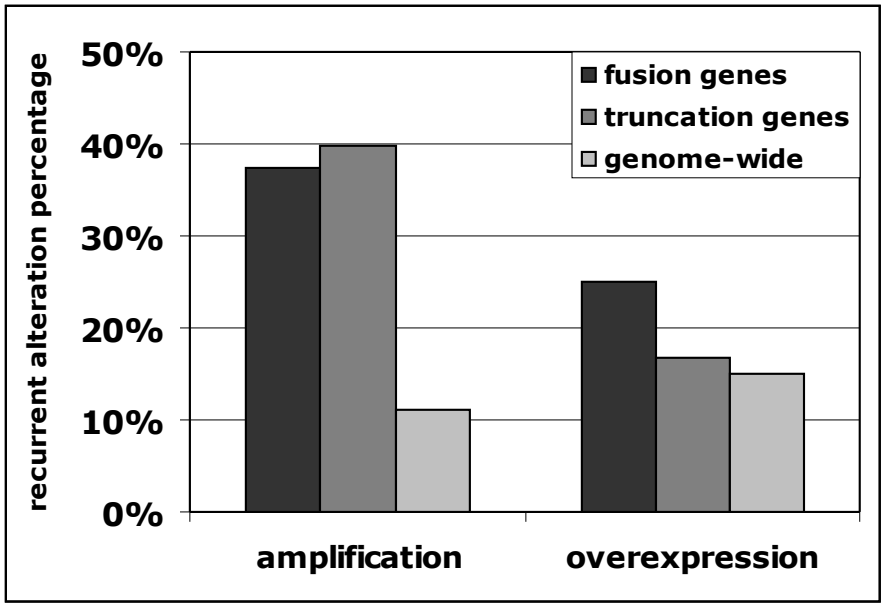
Copy number data from 56 cancer cell lines and 145 breast tumors was obtained from four separate papers. All data sets were from BAC arrayCGH and copy number change was analyzed on a per-BAC basis, with the boundaries of each BAC extended to halfway between it and its neighboring BACs. In edge cases where there was no neighboring BAC, the segment was extended to the beginning or end of the chromosome. Thresholds for calling copy number were set at a log ratio of 0.3 for all data sets except Shadeo, where the threshold was set at 0.222 in order to see comparable results on overlapping cell line data.

Expression and Copy Number data was integrated into Supplementary Table 3, and a recurrence score was assigned to each gene in our breakpoint set based on the number of cell lines in which it was differentially expressed or showed copy number change. We define recurrence as a gain or loss that appears in at least 20% of samples. Copy-number enrichment was calculated by comparing the percentage of all genes in the RefSeq and Known Gene tracks that show recurrent changes to the percentage of breakpoint-associated genes with recurrent changes. Significance and p-values were calculated using Fisher's exact test.

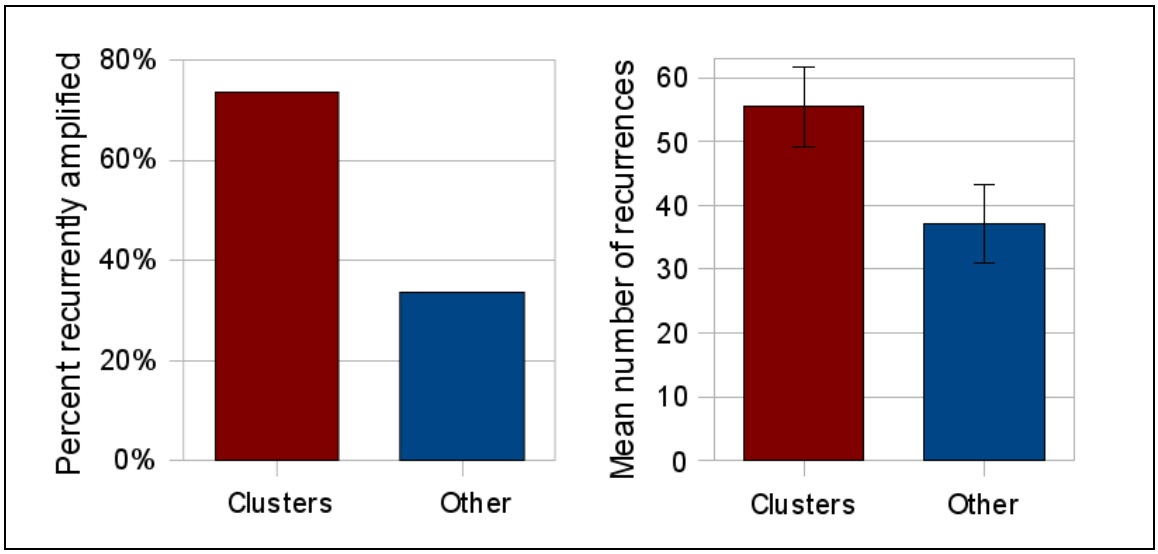
In addition, Affymetrix 100k SNP array copy number data, originally made available by Affymetrix, was downloaded from the Pevsner Laboratory website at (http://pevsnerlab.kennedykrieger.org/snpscan_05_sampledata.htm). It was segmented using the Circular Binary Segmentation algorithm (Venkatraman and Olshen 2007) and regions encompassing at least three probes and having a mean copy number of +/-0.5 were called aberrant.

Based on an integrated analysis of copy number and expression change reported in other studies involving 145 breast tumors and 56 breast cancer cell lines (Chin, DeVries et al. 2006; Neve, Chin et al. 2006; Shadeo and Lam 2006; Jonsson, Staaf et al. 2007),

we identified genes that we find to be both disrupted in MCF7 and also recurrently reported as altered in other studies. Among our breakpoint genes, we find over two-fold enrichment for recurrent upregulation.



Supplemental Figure 4: Breakpoint-associated genes in MCF-7 show enrichment for recurrent amplification and overexpression in other breast cancer cell lines and tumors. Left: Sets of genes involved in fusion and truncation events each show over 3-fold enrichment for copy number amplifications (Fisher's exact test: $p=5.524e^{-3}$ and $p=1.105e^{-9}$, respectively) Right: Recurrent overexpression is also observed, but fails to pass a high significance threshold.



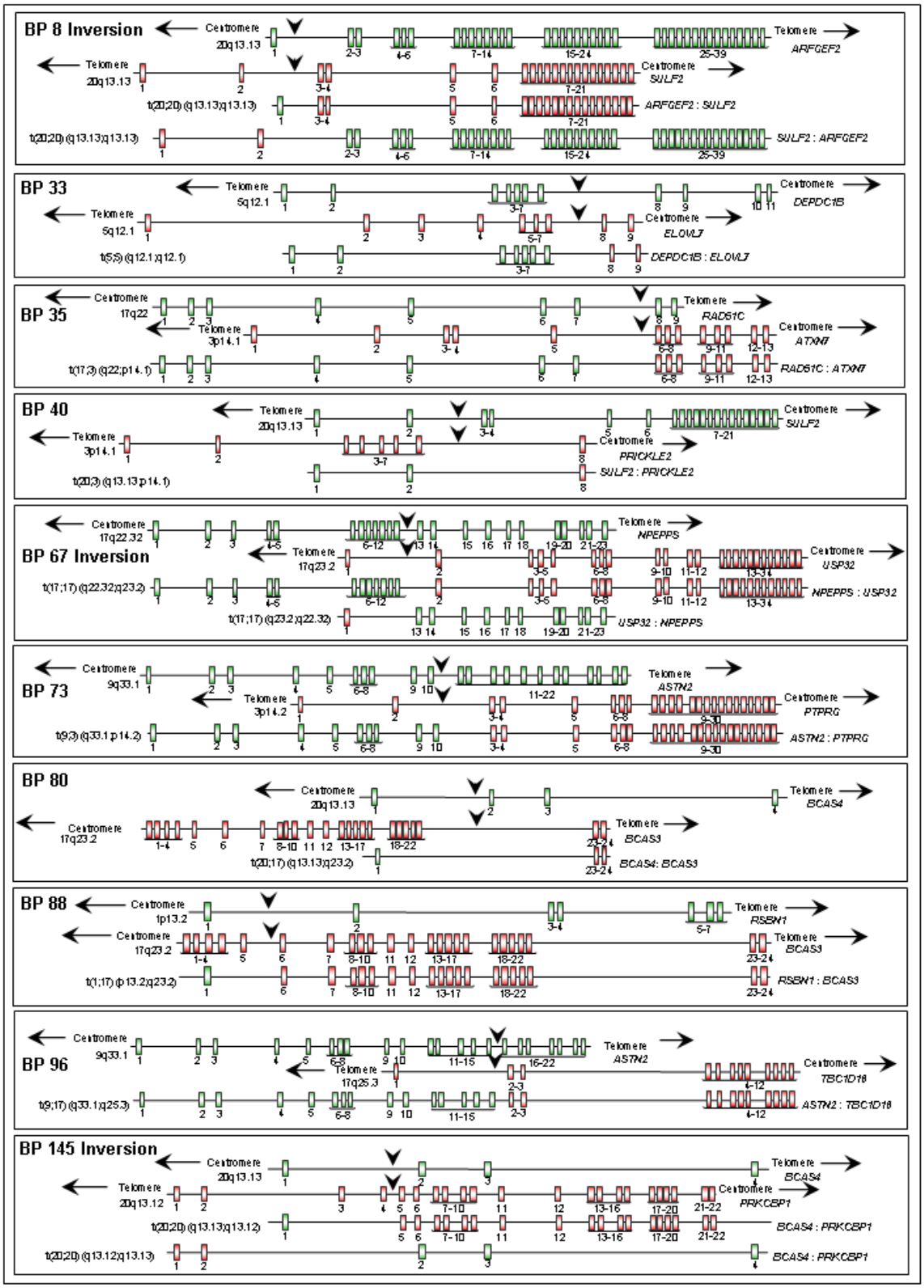
Supplementary Figure 5: Left panel: 73% of breakpoints found in the four clusters are in recurrently amplified regions in other breast cancer tumors and cell lines. This is more than a two-fold enrichment over dispersed breakpoints (Fisher's exact test – p-value =9.9e-7). Right panel: The mean number of amplifications seen in other cell lines and tumors is also significantly higher for clustered breakpoints (Student's t-test - p=2.3e-5).

Supplementary Table 3: Gene Table Including Expression and Copy Number Variation from Other Studies

See file “SuppTable3-GeneTable”:

Supplementary Table 4: Breakpoint Table Including Copy Number Variation from Other Studies

See file “SuppTable4-BreakPointGainLoss”:

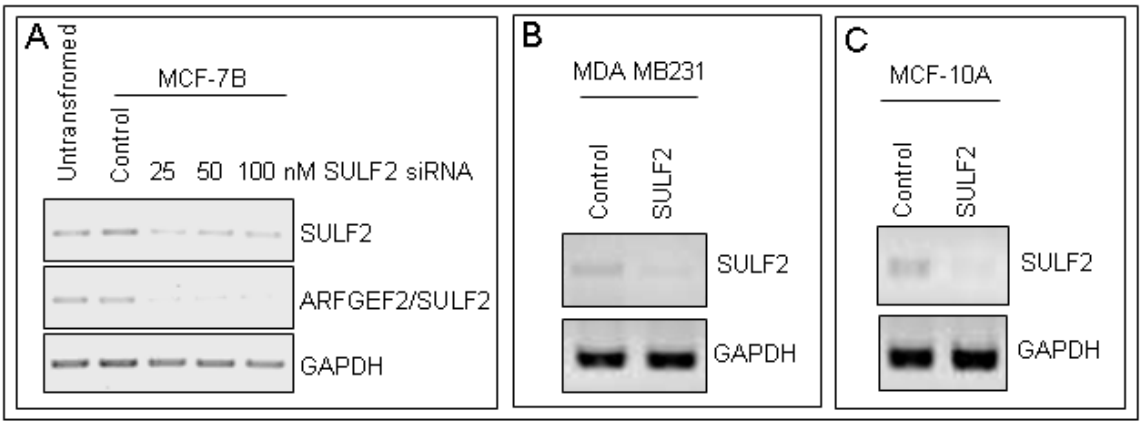


Supplementary Figure 5: Ten Gene Fusions

Supplemental Table 5: RT-PCR primers for amplification of predicted fusion transcripts

Breakpoint	Forward Primer	Reverse Primer	Nested Forward	Nested Reverse
BP 8	caggagagccagaccaagag	acttgccagtgaggatggag		
BP 33	ggcccaactgttctgatttg	cccaatgcagaaagtccatag		
BP 35	agttggggacatgctgctac	ggatgcgctaagaacacctc		
BP 40	gagcgagagtgtgtcgagtg	tttcctgctgtgcaactgtc	cccagctctgcgttcact	ggaaatagccctcgttgtca
BP 67	ggttgctgctaaaaccttgc	cctccagatagaagccatcg	tgcagactttgcagctggt	cgtggagcattctttccatt
BP 73	ttgctgtgagttcctgcatc	ccctcggactgacttgaaaa	ctggacatctccgactggtt	aaattgcggtttgaaagctg
BP 88	tgtgtttgtgggtgaaatgg	tgtcaggccacttttcaatg	gaggggaaggagaaacctca	tggacatggatagcagcttg
BP 96	accactccaaaggcattgac	ccaccaggatgtcctcatct	cacttcgagcaccatcttca	gctctcgggtgtgatgtagc
BP 145	cctcctgatgctgctcgt	cagcacaggcttgctcatc	ctcgcgctcttcctgacc	gcttgaccttctcctgcttg

To give insight into the function of the *ARFGEF2-SULF2* fusion, *SULF2* mRNA was knocked down using siRNA specifically targeting *SULF2* in MCF-7B, MDA MB231 and MCF-10A cells (**Supplementary Figure 6**).



Supplementary Figure 6: A) Different amounts of SULF2 siRNA were tested on MCF-7B cells. By RT-PCR is shown that both SULF2 and ARFGEF2/SULF2 fusion expression is reduced with the transfection of SULF2 siRNA. **B,C)** 50uM of control or SULF2 siRNA was transfected in MDA MD231 (b) and MCF-10A (c) cells. RT-PCR for SULF2 mRNA shows reduced levels of SULF2 mRNA after treatment with SULF2 siRNA. GAPDH is used as the loading control.

Supplementary Material References

Chin, K., S. DeVries, et al. (2006). "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies." Cancer Cell **10**(6): 529-41.

Jonsson, G., J. Staaf, et al. (2007). "High-resolution genomic profiles of breast cancer cell lines assessed by tiling BAC array comparative genomic hybridization." Genes Chromosomes Cancer **46**(6): 543-58.

Neve, R. M., K. Chin, et al. (2006). "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes." Cancer Cell **10**(6): 515-27.

Shadeo, A. and W. L. Lam (2006). "Comprehensive copy number profiles of breast cancer cell model genomes." Breast Cancer Res **8**(1): R9.

Venkatraman, E. S. and A. B. Olshen (2007). "A faster circular binary segmentation algorithm for the analysis of array CGH data." Bioinformatics **23**(6): 657-63.