Supplementary online data:

svpluscnv: analysis and visualization of complex structural variation data

Gonzalo Lopez[1,*], Laura E. Egolf[2], Federico M. Giorgi[3], Sharon J. Diskin[2,4] and Adam A. Margolin[1]

[1]Genetics and Genomics Sciences, Icahn School of Medicine, New York, USA, [2]Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, USA, [3]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [4]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

*To whom correspondence should be addressed.

Table of Contents

The *svpluscnv* R package is designed for integrative analyses of somatic DNA copy number variations (CNV) and structural variants calls (SVC) derived from WGS discordant alignments to the reference genome in tumor/normal paired studies. *svpluscnv* comprises multiple analytical and visualization tools that can be applied to large datasets from cancer patients as well as cell lines. The whole code is written in R and it takes advantage of the *algebra* of genomic ranges implemented in the *GenomicRanges* package (Lawrence, et al., 2013) as well as Circular plotting implemented in *Circlize* package (Gu, et al., 2014). The package also makes use of genomic annotations including cytogenetic bands and gene RefSeq annotations obtained from UCSC table browser (Karolchik, et al., 2004). Currently, two genome versions are supported: hg19/GRCh37 and hg38/GRCh38.

Code and data availability

The *svpluscnv* R package stable version source code is available in the GitHub repository (https://github.com/ccbiolab/svpluscnv). Accompanying *svpluscnv* code, we included a complete vignette describing all functionalities.

In addition, the code and data used for this manuscript has been uploaded to a GitHub repository: https://github.com/ccbiolab/svpluscnv_doc_code. The stand-alone code allows reproduction of all figures and analyses mentioned in the manuscript by cloning the repository to a local site.

Source of datasets used along this study

We used 3 genomics data sources throughout this manuscript, all of which are encoded in the hg19/GRCh37 coordinate system; the three datasets allow a variety of usage modes depending on data availability:

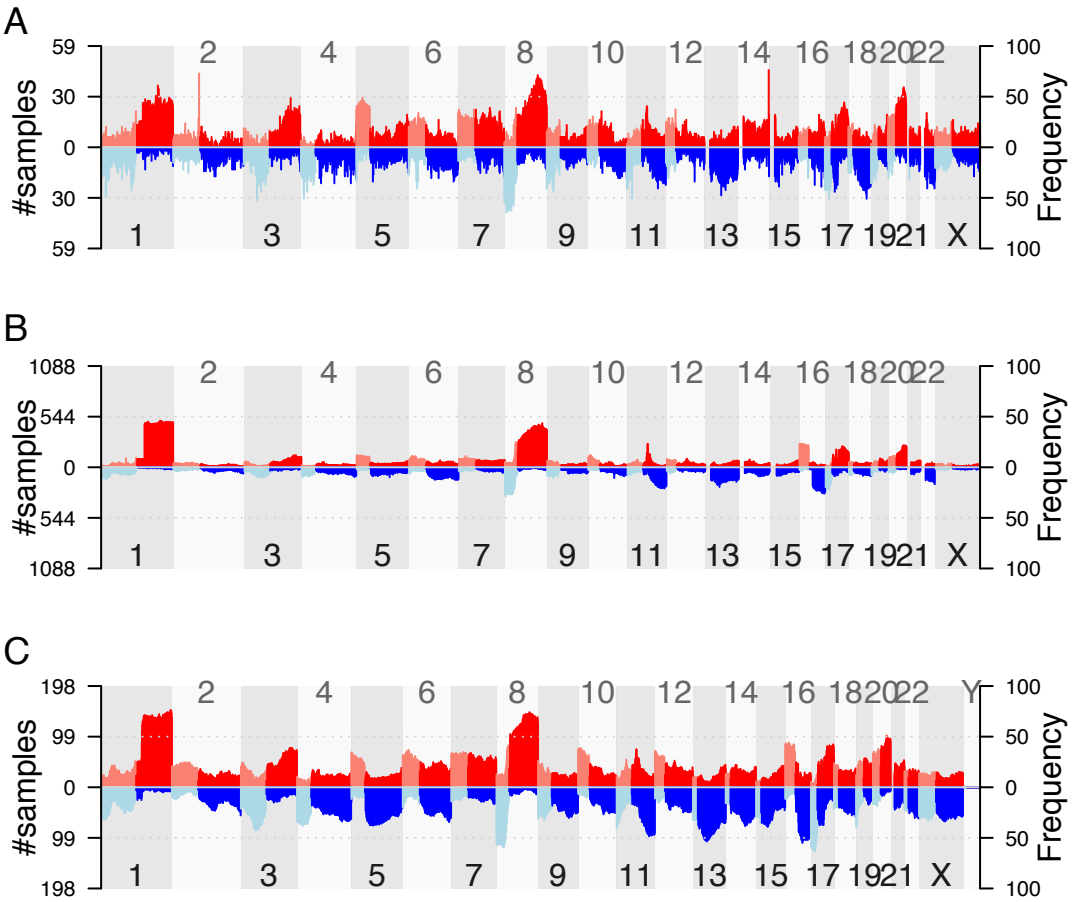
1) Breast Cancer Cell line genomic data from CCLE (Ghandi, et al., 2019) organized by DepMap repository (https://depmap.org/portal/download/) includes CNVs and SVC from orthogonal sources:
    a. SVC somatic SVs derived from 35 WGS sequenced cell lines were obtained from release 19Q2 (file name: CCLE_translocations_SvABA_20181221.csv.gz).
    b. CNV profiles from release CCLE_copynumber_2013-12-03; we used the SNP6.0 arrays from 59 cell lines from the legacy version since the most recent releases

included CNV profiles derived from different platforms that introduced considerable batch effect (data not shown).

2) TCGA data from Breast Cancer (brca) 1088 primary tumors were obtained from the hg19 GDC legacy archive (https://portal.gdc.cancer.gov/legacy-archive/). Profiles were derived from SNP6.0 arrays. This is the largest dataset, but only CNV data are available.
3) The PCAWG worldwide consortium for the analysis of cancer whole genomes (Consortium, 2020) comprises 198 Breast-Adenocarcinomas. These data derive from WGS and were downloaded from the consortium repository pages (https://dcc.icgc.org/releases/PCAWG). Although both CNV and SVC are available the source is not considered orthogonal:
    a. SVC data were obtained from the *consensus_sv* folder and reformatted as *svpluscnv* input format.
    b. CNV segmentation profiles were derived from the file consensus.20170119.somatic.cna.annotated.tar.gz. In order to recreate logR values the consensus '*total_cn'* CNV call was transformed: first, we added white noise ($\mu = 0$; $\sigma = 0.01$). Next, we divided the copy number by 2 (for autosomal chromosomes and the female X chromosome only; the male sex chromosomes were not divided) and applied log2 transformation.

Genome-wide visualization of CNV frequencies

To visualize CNV gain/loss frequencies across the genome, we used the *cnv.freq* function with arguments: *fc.pct = 0.3, ploidy = TRUE*. The function produces a 1Mb genomic binned table with gain/loss frequencies. The threshold *fc.pct* represents a percentage of the fold change in copy number dosage (e.g. 0.3 -> 30%). We applied the method to CCLE Breast (59 samples), TCGA brca (1088 samples) and PCAWG-Breast-AdenoCA (198 samples) (**Supplementary Figure S1**). Although the overall ploidy differs from among datasets, the overall shape is replicated.



Supplementary Figure S1. Visualization of CNV frequencies with *cnv.freq*
CNV genome map indicating the frequency of gains (red) and losses(blue) across breast cancer datasets: (**A**) 59 breast cancer cell lines with available CNV profiles, (**B**) 1088 primary breast cancers TCGA samples with available CNV and (**C**) 198 Breast adenocarcinomas from the PCAWG consortium with available CNVs derived from WGS.

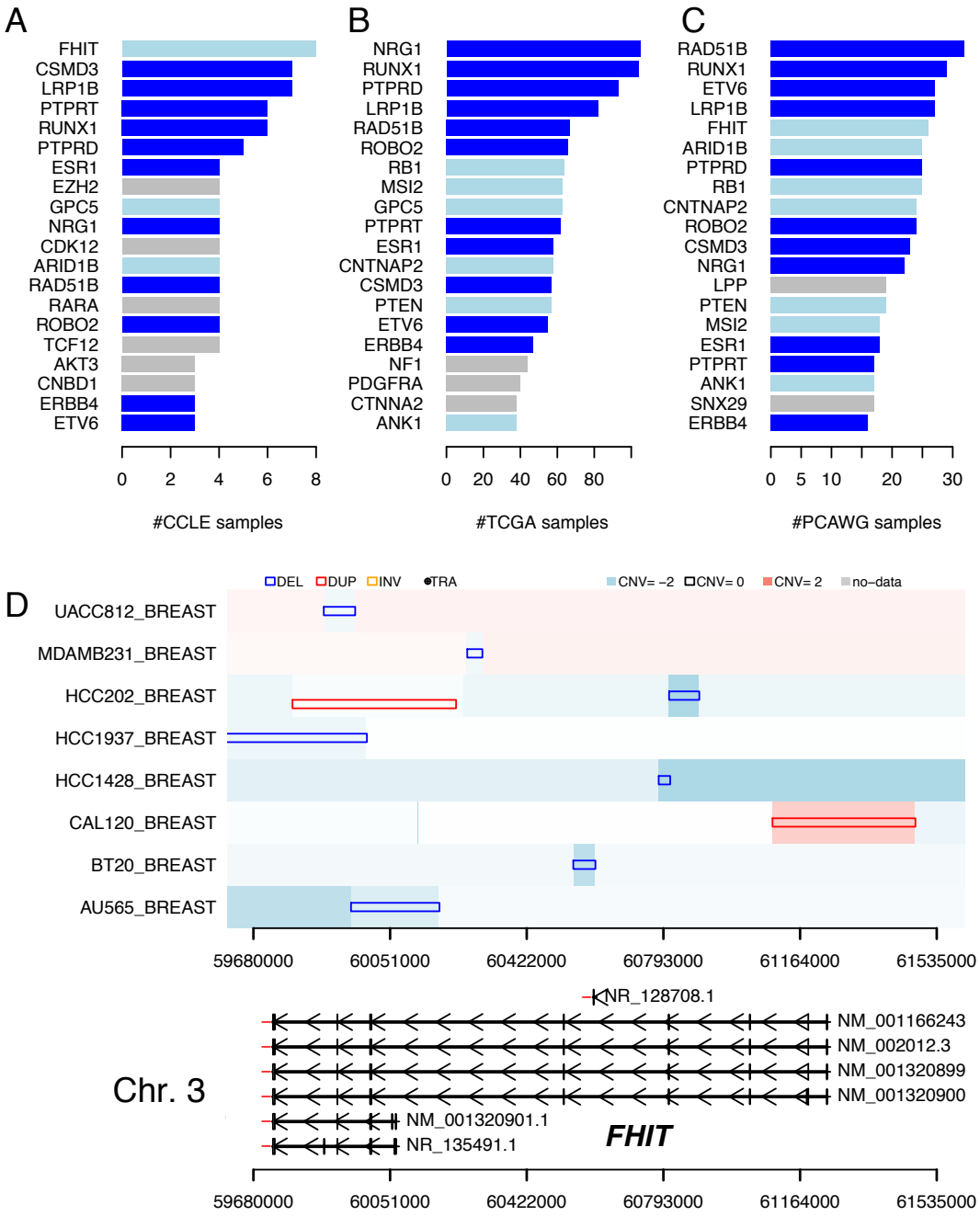
Identification and visualization of recurrently altered genes

Somatic pathogenic variants are characterized by presenting in recurrent patterns that differ across histotypes (Futreal, et al., 2004). Evaluating the recurrence of structural variations involves challenges as their interpretation is more complicated than other variant types (e.g. SNVs). Traditionally, CNVs have been used to detect dosage changes affecting whole genes (e.g. amplifications and deep-deletions) and their recurrence has been statistically evaluated (Mermel, et al., 2011). *svpluscnv* evaluates the recurrence of structural variants by focusing on the analysis of breakpoints overlapping with known genes, including sub-genic events (e.g. *ATRX* in frame fusion events (Kurihara, et al., 2014)) and their upstream and downstream regions that allow the identification of recurrent events altering the function of oncogenes and tumor suppressors (e.g. rearrangements near *TERT* (Peifer, et al., 2015; Valentijn, et al., 2015)).

The functions *cnv.break.annot* and *svc.break.annot* evaluate breakpoints obtained from CNV and SVC data, respectively. Both functions report a list of genes and their associated breakpoints and samples, which can be retrieved for further analyses.

We studied recurrently altered genes among the COSMIC cancer census (Futreal, et al., 2004) in the three breast datasets including CCLE (35 samples, SNP∩WGS), TCGA (1088 samples) and PCAWG (198). We considered a 'hit' whenever a given gene was altered by both *cnv.break.annot* and *svc.break.annot* in a given sample, then we aggregated the data across all samples and genes and obtained ranked lists of altered genes for each dataset.11 genes appear in the top 20 in all three datasets whereas another 9 are present in at least 2 datasets (**Supplementary Figure S2A-C**). In the case of TCGA, the results were obtained from CNV data only whereas CCLE and PCAWG results represent the intersection between both CNV and SVC breakpoint analyses.

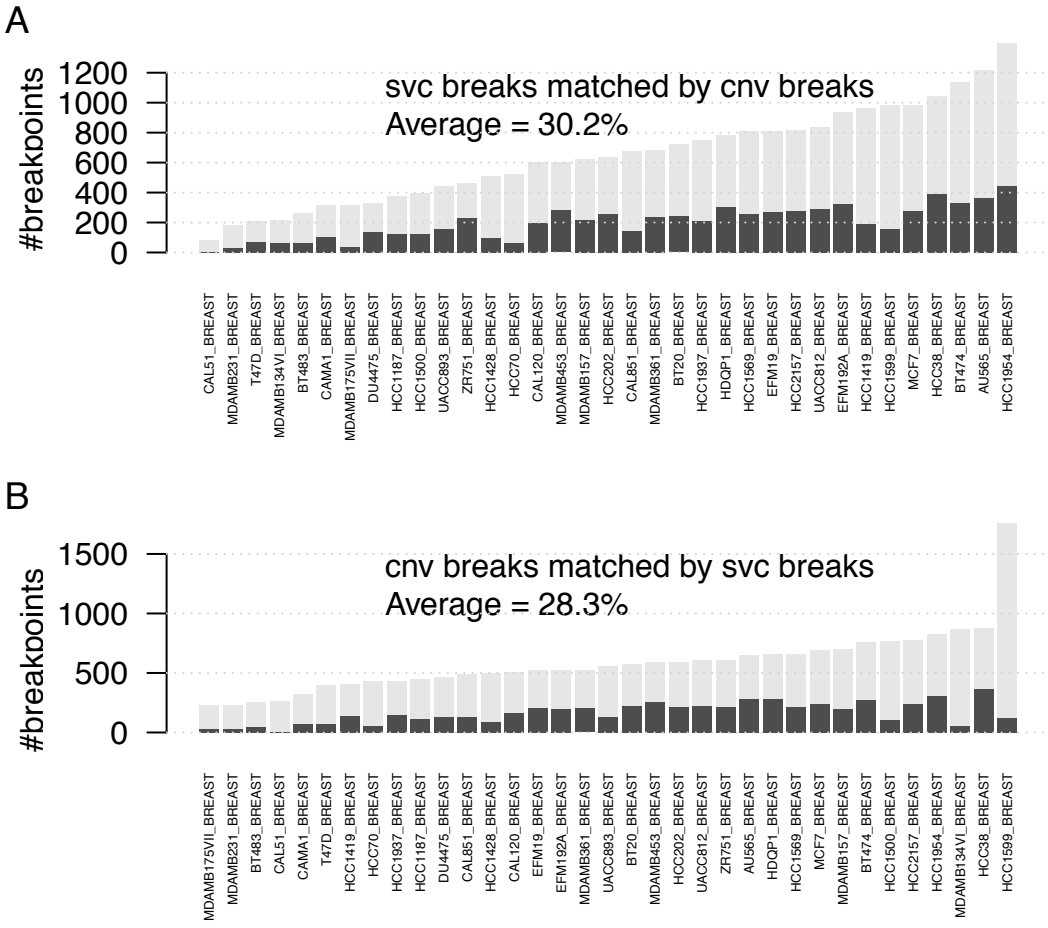
svcnvplus includes an integrated visualization tool *sv.model.view* that overlays data from CNV segmentation data and SVC calls. This function allows visualization of all variants affecting a specified genomic region (e.g. gene locus) at once. This functionality is complemented with a genomic track plot function (*gene.track.view*) that can be used to build layouts. To illustrate this functionality, we selected *FHIT* as the most frequently altered gene (n=8) in the subset of 35 CCLE breast cell lines with available CNV and SVC data. Observed breakpoints across CNV and SVC datatypes match correctly in all samples harboring SVs (**Supplementary Figure S2D**); this is remarkable since the data types were completely orthogonal in terms of data source and methodology.



Supplementary Figure S2. Recurrent SVs in Breast Cancer tumors and cell lines
(A-C) Recurrently altered genes based on breakpoints derived from CNV and SVC in 35 CCLE breast cell lines (A), CNV breakpoints in 1088 TCGA brca tumors (B) and CNV and SVC in 198 PCAWG breast adenocarcinomas (C); bars from genes present in the top 20 in two and three datasets colored in light-blue and blue respectively. (D) Visualization of orthogonal breakpoint data spanning *FHIT* genomic locus in 8 CCLE samples harboring variants; the top represents individual sample tracks where lines indicate SVC whereas the background color represents CNV. The bottom represents the known transcripts of *FHIT.*

Co-localization of breakpoints across orthogonal datasets

In order to integrate orthogonal approaches to study structural variants, we implemented tools to evaluate the overall concordance across datasets. To this end, breakpoints are obtained from each dataset using the retrieval functions *cnv.breaks* and *svc.breaks* for CNV and SVC data, respectively. Then, the function *match.breaks* is used to compare two sets of breakpoints. The function evaluates whether each breakpoint from source 'a' has a co-localizing breakpoint in source 'b' on a sample by sample basis given a maximum gap distance (*maxgap* = 10000 (bp)), thus reporting the overall validation rate across datasets. Here we compared the set of 35 CCLE Breast Cell lines with CNV and SVC data available; on average 30.2% SVC breakpoints co-localize with a CNVs and 28.3% CNV breakpoints co-localize with an SVC (**Supp Fig S3**).



Supplementary Figure S3. Validation of breakpoints from orthogonal sources
Per sample comparison of breakpoints; each bar plot represents the number of breakpoints in a given dataset that have colocalizing breakpoints in the orthogonal dataset. (**A**) Breakpoints from SVC (WGS) validated in CNV segmentation (SNP) data. (**B**) Breakpoints in CNV segmentation data validated in SVC.

Identification of shattered regions

Complex chromosomal rearrangements such as chromothripsis and chromoplexy are widespread events in many cancers and may have important pathogenic roles (Cortes-Ciriano, et al., 2020; Valentijn, et al., 2015; Zhang, et al., 2013). *svpluscnv* incorporates tools for the identification and visualization of shattered regions. The *shattered.regions* algorithm is highly parameterizable and executes the following steps:

1) Identification of genomic bins with high breakpoint density (HBD):
    a) The genome is binned into 10Mb windows (*window.size* = 10) and calculated every 2Mb (*slide.size* = 2).
    b) Breakpoints are defined using *cnv.breaks* (CNV), *svc.breaks* (SVC), and *match.breaks* (co-localizing CNV and SVC breakpoints) and then mapped into bins. Minimum thresholds are set for HBD using *num.cnv.breaks* = 5, *num.svc.breaks* = 5 and *num.common.breaks* = 3 respectively.
    c) To avoid biasing towards high chromosomal instability samples, the number of breakpoints in HBD bins are expected to be outliers within a given sample; therefore we set a threshold number of standard deviations above the average each HBDs must comply using the formula: $N = \mu + x\sigma$; where μ is the average number of breaks, σ the standard deviation and x is a parameter defined for each breakpoint type: *num.cnv.sd* = 5, *num.svc.sd* = 5 and *num.common.sd* = 0.
2) Identification of shattered regions:
    a) Overlapping HBDs that passed the thresholds are collapsed into shattered regions.
    b) To discard complex focal events such as circular amplifications or double minutes that tend to have breakpoints highly concentrated in small regions, the minimum interquartile average of the distances between breakpoints is set to *dist.iqm.cut* = 100000 (bp).
    c) Finally, shattered regions such as chromothripsis and chromoplexy produce interleaved SVs (non-consecutive break-ends). We set the minimum proportion of interleaved SVC *interleaved.cut* = 0.33 to discard regions with less than 33% interleaved variants.
    d) Shattered regions linked by translocations are identified.

Currently, available somatic CNV datasets are much larger compared to available somatic SV datasets obtained from WGS; for this reason, we implemented the algorithm *shattered.regions.cnv* that identifies catastrophic events using CNV segmentation data only. It

follows same steps as *shattered.regions*, using only parameters derived from CNV as described above in steps 1 a-c and 2a-b.

Comparison of different methods for the detection of shattered regions

In order to benchmark shattered.regions algorithm we obtained complex rearrangements from 2658 cancer whole genomes (Cortes-Ciriano, et al., 2020); the recently published study also included ShatterSeek, one of the two currently available methods for the detection of complex rearrangements in the literature; this study evaluates 2428 tumors after quality control and predictions were further manually curated. The other available method, ShatterProof (Govind, et al., 2014) was applied to the same dataset using default parameters. We also obtained predictions from the *shattered.regions* algorithm. In order to compare predictions, we considered as a positive, every chromosome with detected shattered regions (23 chromosomes x 2428 samples = 55844 total possible cases); thus, two methods agree if the same chromosome of a given sample is detected (regardless of the region boundaries). The overlaps across the three methods (Main Fig 1B) is highly significant when evaluated using a Fisher's exact test (p-value < $2.2e^{-16}$ in all pairwise comparisons).

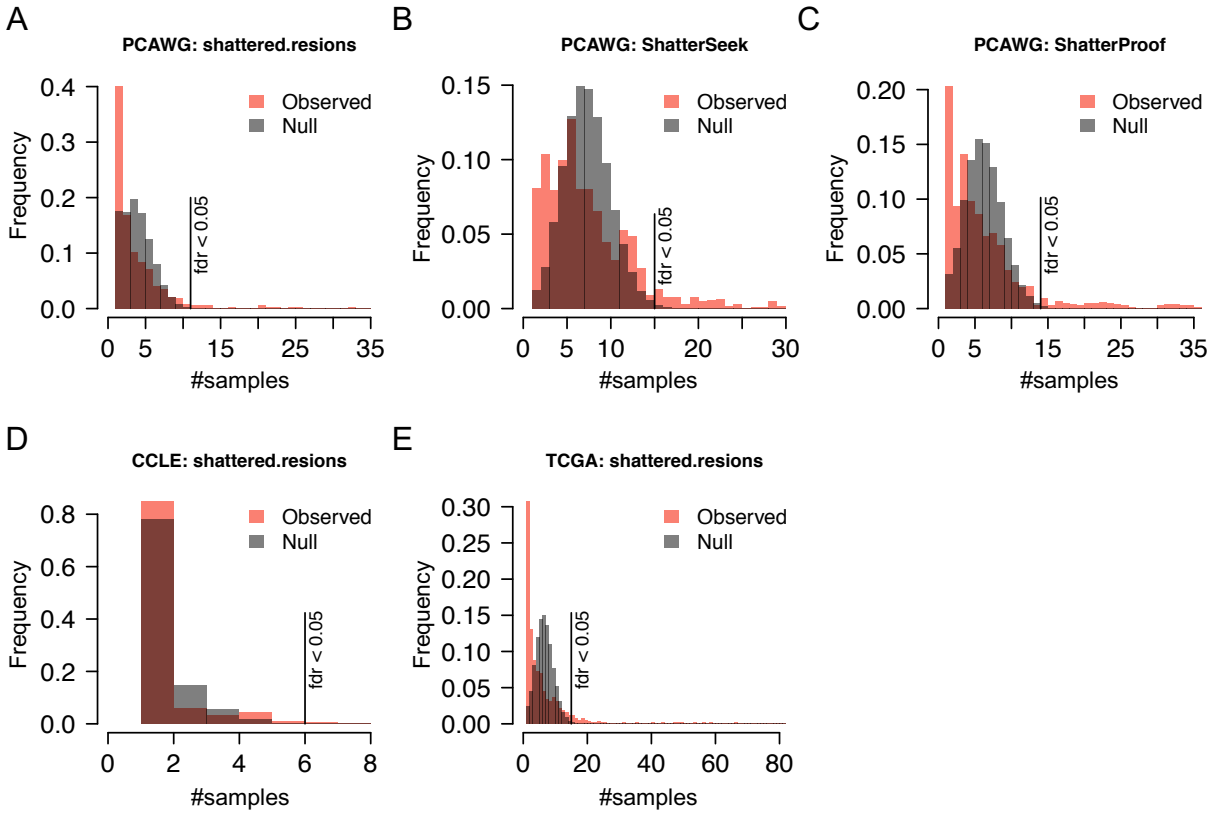
Localization of shattered region hot spots in breast cancer

To establish whether certain regions suffer chromosome shattering, we aim to discard the null hypothesis that shattered regions appear in the genome at random. To this end, we evaluate the frequency at which genomic bins are classified as HBDs. First, *shattered.regions* returns a binary matrix of $N_{samples}$ by N_{bins} dimensions (0 = normal; 1 = HBD); the sum of each bin represents the observed frequency distribution. Second, *freq.p.test* generates a null frequency distribution by permutating the bins in each sample with *I* (*iter = 100*) iterations (*null length = I x* N_{bins}). The observed distribution is compared with the null distribution to obtain empirical p-values and corrected for multiple testing using available methods implemented in *p.adjust* function (R stats package) (e.g. *method = "fdr"*). The corrected p-values deemed statistically significant under a specified cutoff (e.g. *p.cut = 0.05*) define genomic bins under selection pressure for chromosome shattering. This analysis is implemented in svpluscnv under the function name '*freq.p.test'*.

We tested this approach with shattered region predictions based on 198 PCAWG breast adenocarcinomas using *shattered.regions*, *ShatterSeek* and *ShatterProof* frequency maps (**Main Fig 1C**); svpluscnv incorporates functionalities to input defined predicted region (.bed format) into

the package framework using '*bed2chromo.reg*' function. We also evaluated alternative datasets (CCLE SNP∩WGS = 35, TCGA SNP = 1088 and) frequency maps (**Main Fig 1D**).

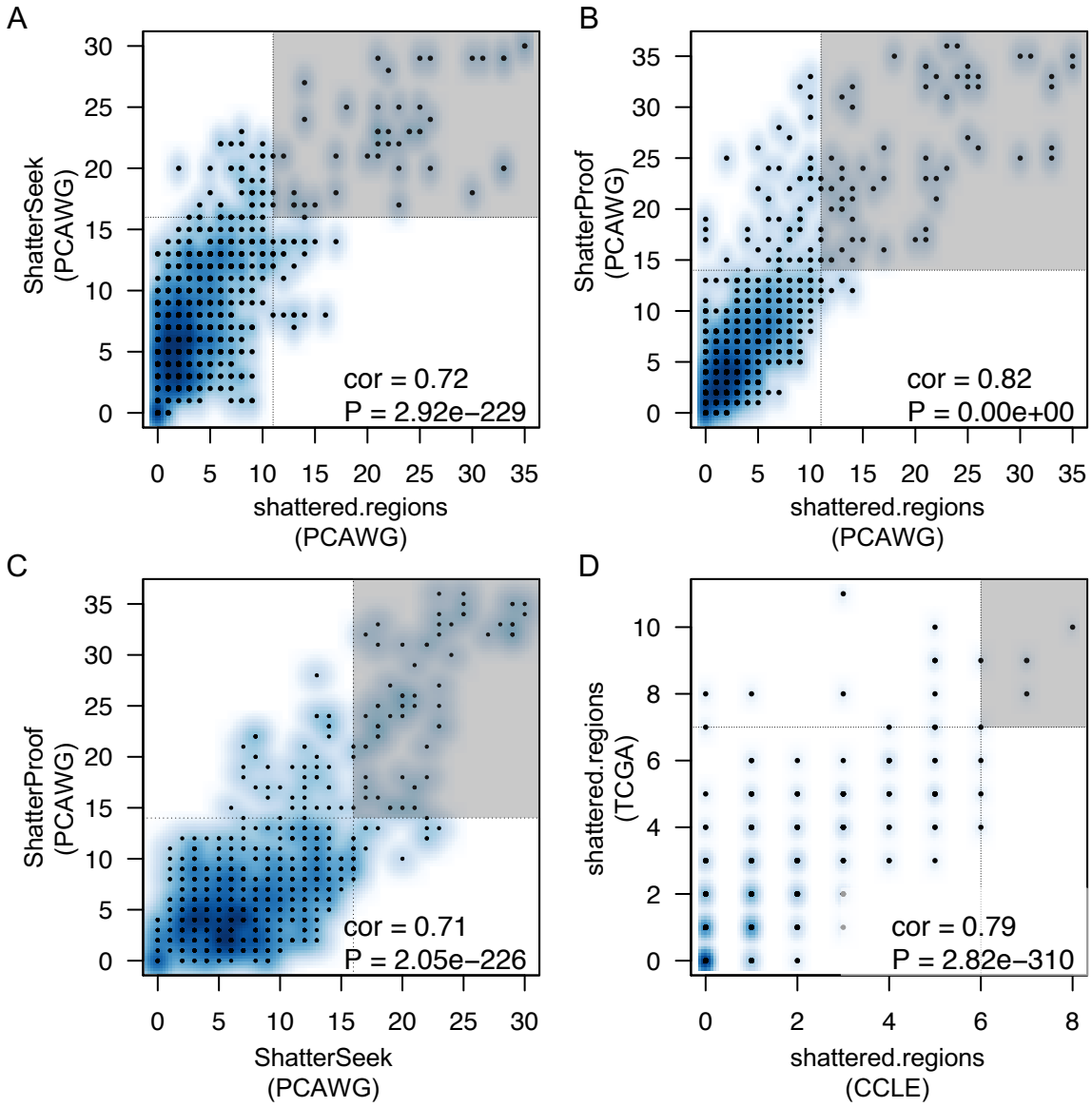
Frequency thresholds for significantly enriched regions in PCAWG were defined with FDR< 0.05 at 11, 15 and 14 in *shattered.regions*, *ShatterSeek* and *ShatterProof* respectively (**Supplementary Figure S4A-C**). Frequency thresholds for significantly enriched regions were defined with FDR< 0.05 at 6 for CCLE using shattered.regions and 15 for TCGA breast cancer dataset using *shattered.regions.cnv* (**Supplementary Figure S4D,E**). The five genome wide maps identify similar hot spots including: chr8.p11, chr8.q24, chr11q13, chr17q and chr20q. At least two of these regions encode well known oncogenes frequently amplified in cancer: *MYC* (chr8.q24) and *CCND1* (chr11q13) (Santarius, et al., 2010).



Supplementary Figure S4. Estimation of frequency thresholds for shattered region hot spots. Shattered frequency maps (related to Figure 1C,D) are evaluated using a permutation test (svpluscnv::freq.p.test) in order to estimate a significant cut off based on a false discovery rate < 0.05 using different methods and breast cancer and cell line datasets: (A-C) 198 Breast adenocarcinomas from the PCAWG consortium with available SVs and SNVs from WGS regions determined using (A) *shattered.regions* (svpluscnv), (B) ShatterSeek and (C) ShatterProof (based on Figure 1C frequency maps); (D) 35 breast cancer cell lines with available WGS (structural variant calls) and SNP arrays (CNV calls)

using *shattered.regions* (based on Figure 1D top frequency map); (E) 1088 primary breast cancers TCGA samples with available SNP6.0 arrays using *shattered.regions* (based on Figure 1D bottom frequency map);

To further evaluate the similarity across frequency maps we assessed the Pearson's correlation between pairs of frequency maps showing strong correlation (p-value < $2.2e^{-16}$) in all cases (**Supp Fig S5**).



Supplementary Figure S5. Similarity of shattered region frequencies.
Scatter plot and Pearson's correlation analysis of the HBD frequencies using pairs of genomic maps (A-C) derived from three methods against PCAWG 198 breast adenocarcinomas represented in **Fig 1C** and (D) breast cancer derived cell lines versus TCGA breast cancer tumors represented in **Fig 1D**. The grey shaded regions represent the area at which frequency values are significantly higher than expected by chance (FDR < 0.05) as defined in Supp Fig S4.

Data access and code availability

For further description of svpluscnv functions and code, visit the available vignette:

https://github.com/ccbiolab/svpluscnv

The code to reproduce the figures and analyses of this manuscript is available here:

https://github.com/ccbiolab/svpluscnv_doc_code

References

Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* 2020;578(7793):82-93.

Cortes-Ciriano, I.*, et al.* Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 2020.

Futreal, P.A.*, et al.* A census of human cancer genes. *Nat Rev Cancer* 2004;4(3):177-183.

Ghandi, M.*, et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 2019;569(7757):503-508.

Govind, S.K.*, et al.* ShatterProof: operational detection and quantification of chromothripsis. *BMC Bioinformatics* 2014;15:78.

Gu, Z.*, et al.* circlize Implements and enhances circular visualization in R. *Bioinformatics* 2014;30(19):2811-2812.

Karolchik, D.*, et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 2004;32(Database issue):D493-496.

Kurihara, S.*, et al.* Clinical features of ATRX or DAXX mutated neuroblastoma. *J Pediatr Surg* 2014;49(12):1835-1838.

Lawrence, M.*, et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;9(8):e1003118.

Mermel, C.H.*, et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;12(4):R41.

Peifer, M.*, et al.* Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature* 2015;526(7575):700-704.

Santarius, T.*, et al.* A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* 2010;10(1):59-64.

Valentijn, L.J.*, et al.* TERT rearrangements are frequent in neuroblastoma and identify aggressive tumors. *Nat Genet* 2015;47(12):1411-1414.

Zhang, C.Z., Leibowitz, M.L. and Pellman, D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes Dev* 2013;27(23):2513-2530.