Supplemental Material

Somatic structural variant formation is guided by and influences genome architecture

Nikos Sidiropoulos[1,2,9], Balca R. Mardin[3,4,9], F. Germán Rodríguez-González[1,2], Shilpa Garg[5], Adrian M. Stütz[3], Jan O. Korbel[3], Erez Lieberman Aiden[6,7], Joachim Weischenfeldt[1,2,8,*]

[1] Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, 2200, Denmark

[2] The Finsen Laboratory, Rigshospitalet, Copenhagen, 2200, Denmark

[3] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, 69117, Germany

[4] BioMed X Institute, Heidelberg, 69120, Germany

[5] Department of Biology, University of Copenhagen, Denmark

[6] The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA

[7] Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USA

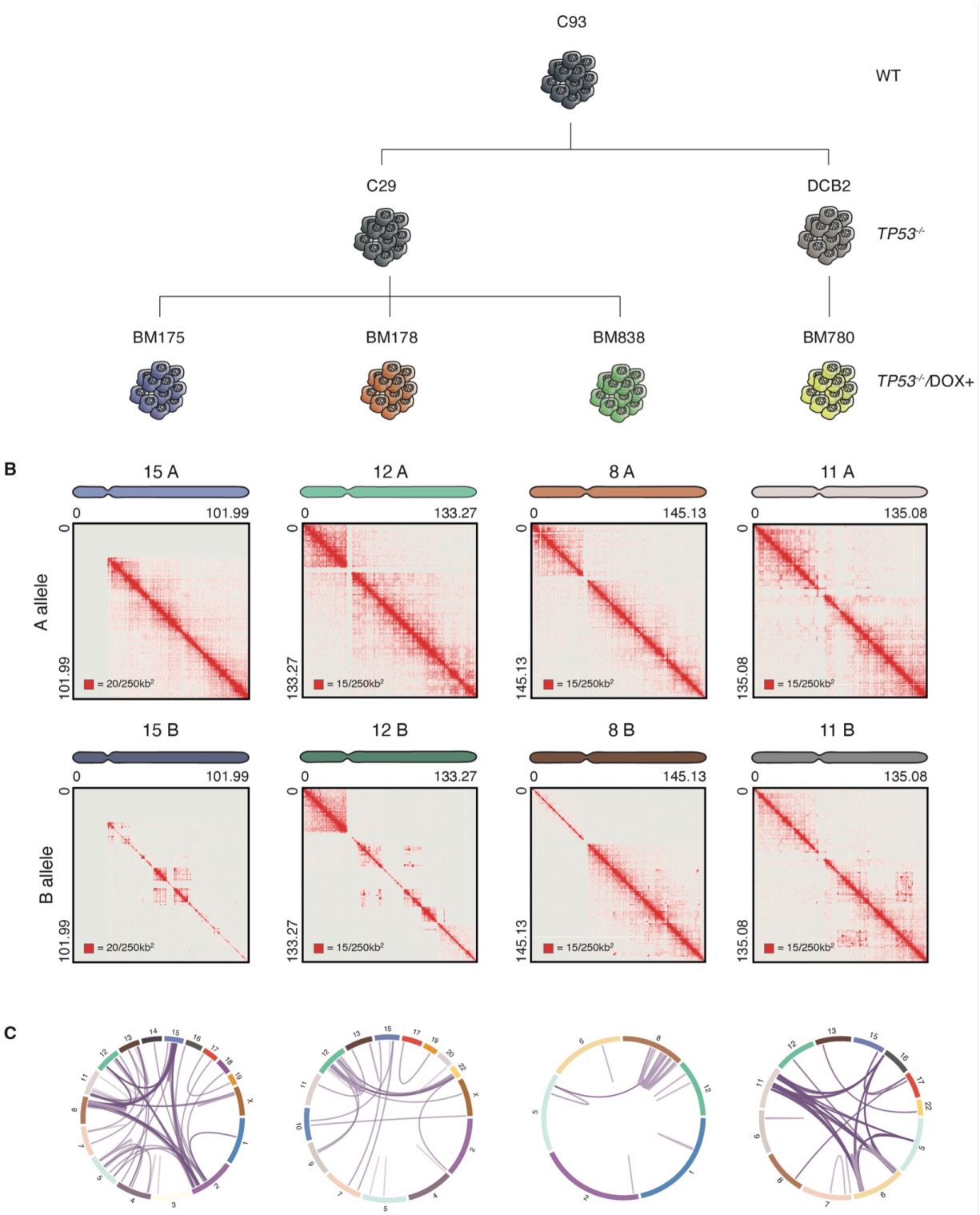
[8] Department of Urology, Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany

[9] These authors contributed equally

[*] email: joachim.weischenfeldt@bric.ku.dk

Table of Contents

SUPPLEMENTAL FIGURES



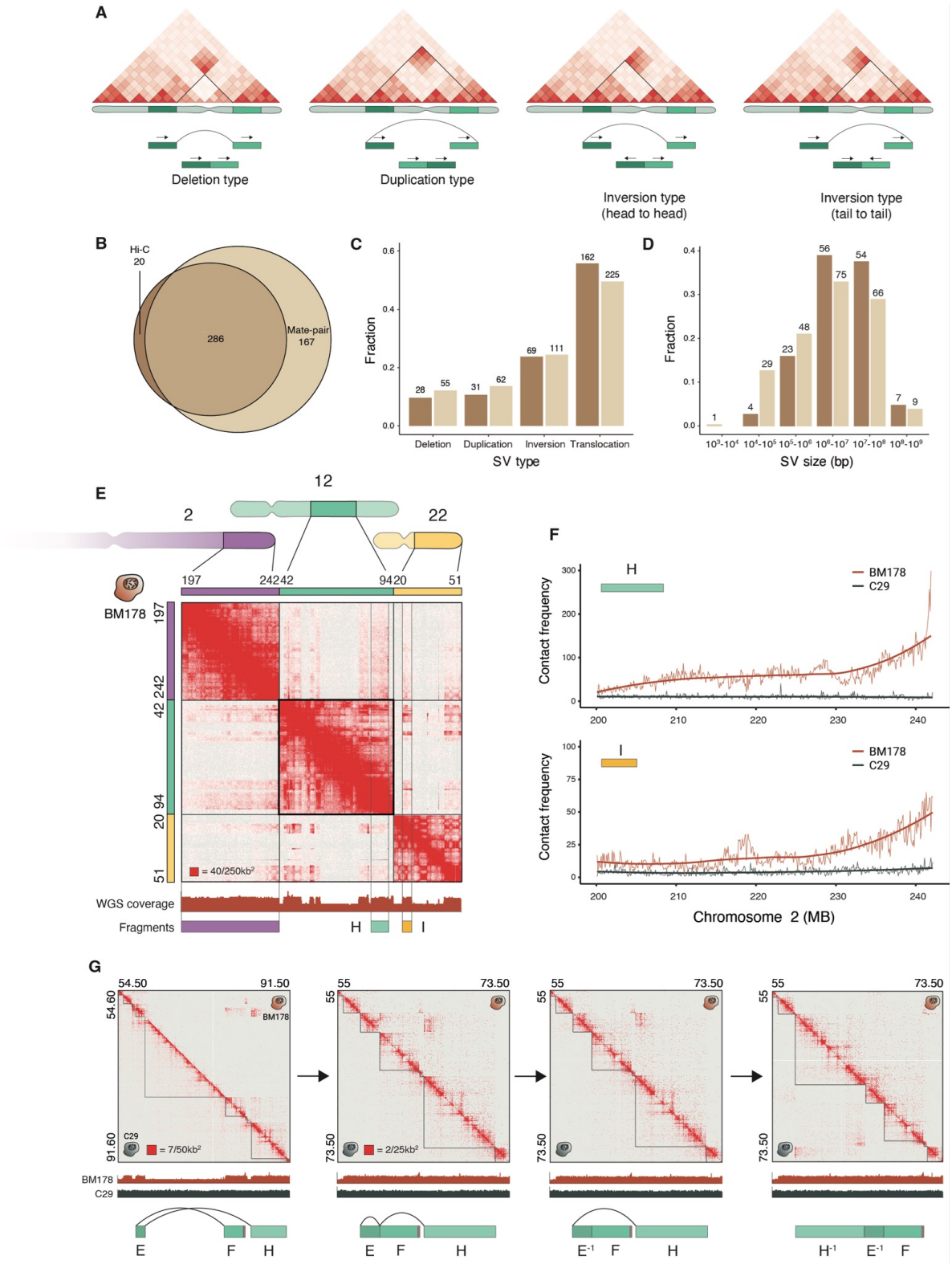
Supplemental Figure S1. Model system overview

A) A 3-tier model system to study SVs. C93 cells (WT) were used to generate two populations of *TP53*-deficient cells, C29 and DCB2 (*TP53*$^{-/-}$). Upon DSB induction with doxorubicin on *TP53*-

deficient cells, we generated four distinct populations carrying complex genomic rearrangements (*TP53*$^{-/-}$DOX+), BM175, BM178 and BM838 derived from C29 and BM780 from DCB2.

B) Allele-specific Hi-C maps of whole chromosomes carrying complex rearrangements, corresponding to the *TP53*$^{-/-}$DOX+ samples in the panel above.

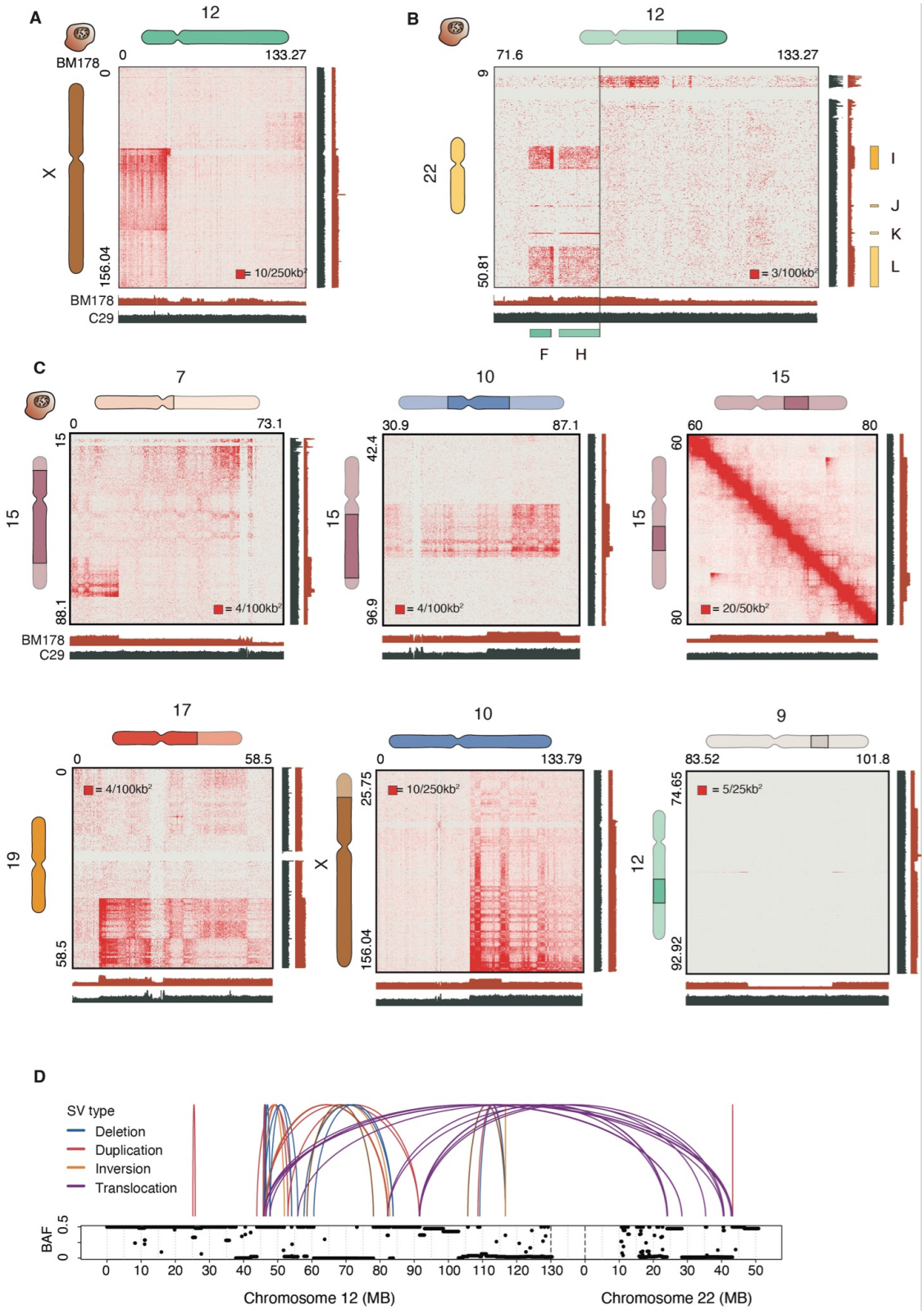
C) Genome-wide SVs of the *TP53*$^{-/-}$DOX+ samples, BM175 (*n* = 138), BM178 (*n* = 44), BM838 (n = 25) and BM780 (*n* = 99). The set of SVs used is a union of SV calls from whole-genome sequencing data and Hi-C.



Supplemental Figure S2. Hi-C based SVs

A) Footprints of deletion-, duplication- and inversion-type SVs on a Hi-C map.

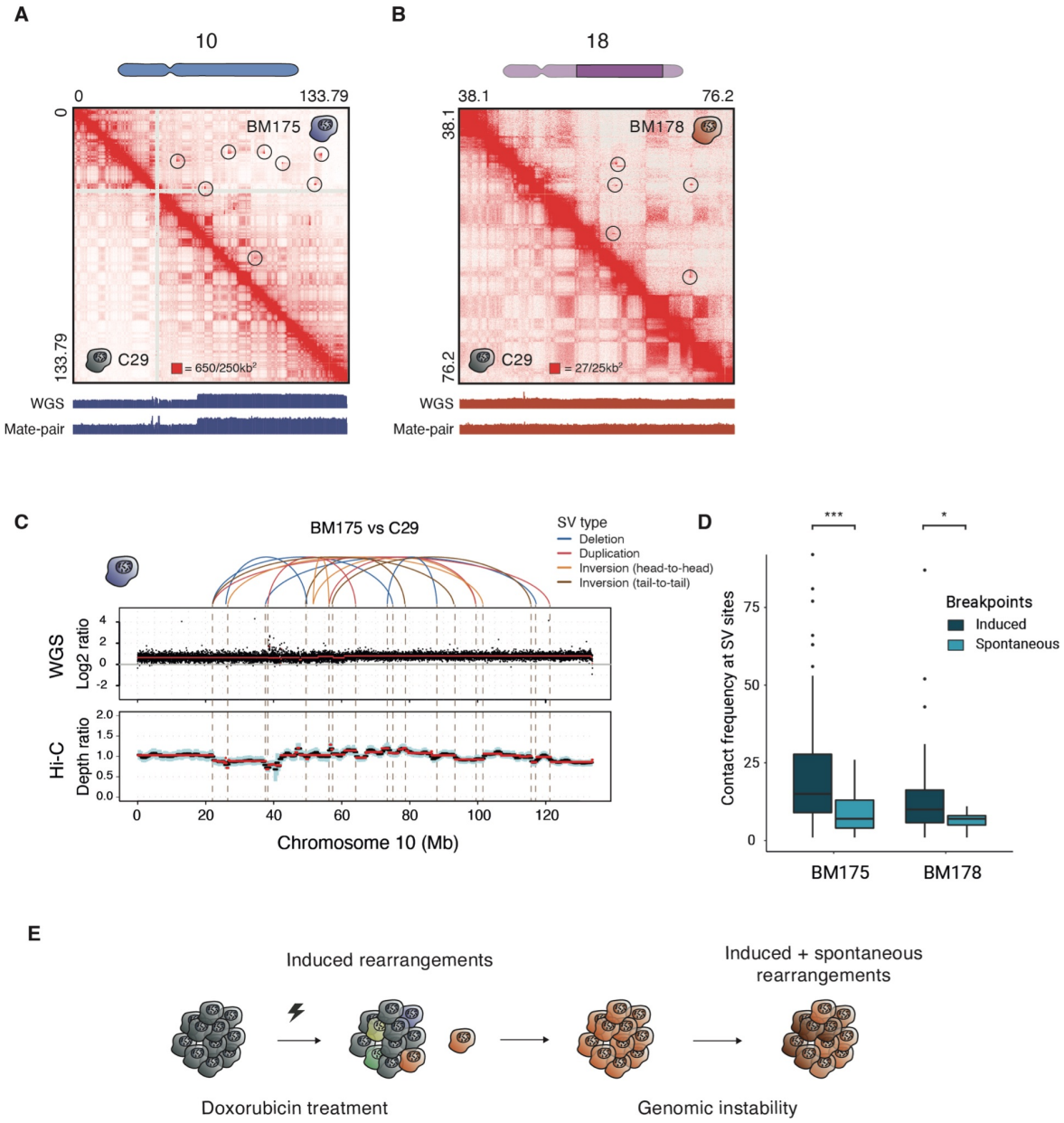
B) Overlap of Hi-C and Mate-pair SV calls.
C-D) Fraction of each rearrangement type (C) and size distribution (D) in Hi-C (brown) and Mate-pair (beige) SV calls.
E) Partial map of Chromosomes 2, 12 and 22. A faint signal on the telomere of Chromosome 2 provides evidence of direct contact with rearranged fragments from Chromosome 12 and 22, annotated H and I respectively. The genomic regions depicted on the Hi-C maps correspond to the highlighted regions on the chromosome ideograms.
F) Average contact frequency of fragments H and I with the q arm telomere of Chromosome 2 for BM178 and the maternal C29 cell line.
G) Assembly of fragments E, F and H using Juicebox Assembly Tools (JBAT). All plots display a partial view of the B allele of Chromosome 12 with BM178 on the upper and C29 on the lower triangle. The illustrations below the plots follow the fragment position in the respective assembly. In a stepwise procedure, we split the chromosome on the predicted sites (left) and remove the deleted regions (left-center). We invert fragment E (right-center) to resolve the first rearrangement, followed by inversion and repositioning of fragment H (right) to resolve the second and last rearrangement. Ectopic domains form between the fragments in the derivative assembly.



Supplemental Figure S3. Complex SV formation

A) Centromere joining of 12p and Xq

B) Evidence of the remaining 12q residing along the 22p arm. The enriched contacts downstream of fragment H align with changes in copy-number in WGS data. Due to the low complexity of 22p, we were unable to identify interactions at higher resolutions.
C) Additional, large-scale rearrangements between several chromosomes of BM178.
D) B-Allele-Frequency (BAF) as estimated by Sequenza on WGS data. Arches and colors represent types of SVs within and between Chromosome 12 (left) and 22 (right).
The genomic regions depicted on the Hi-C maps correspond to the highlighted regions on the chromosome ideograms.



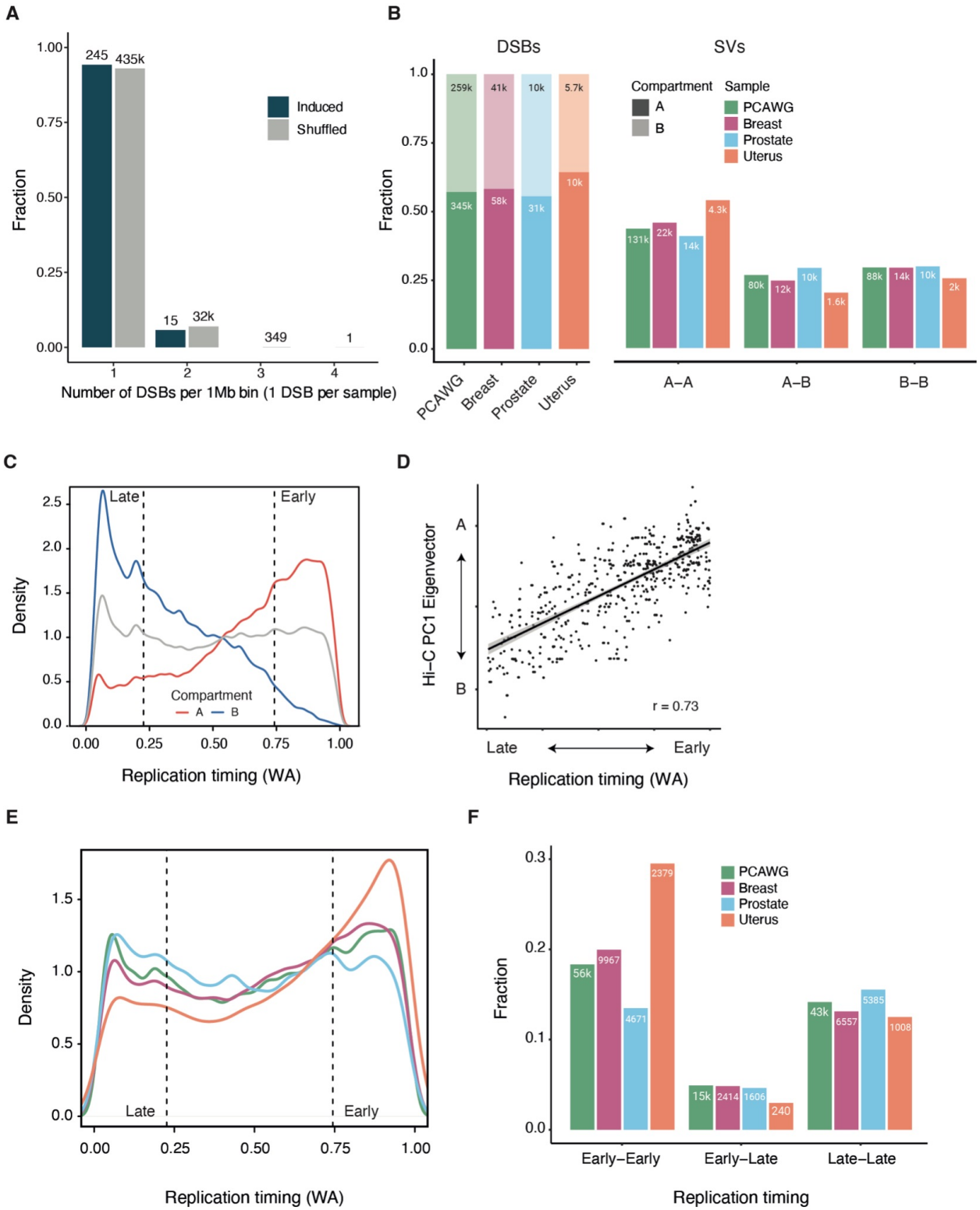
Supplemental Figure S4. Induced and spontaneous SVs in BM175 and BM178

A-B) Structural variants in BM175 Chromosome 10 (A, *n* = 13) and BM178 Chromosome 18 (B, *n* = 9) detected in Hi-C data with no evidence on WGS or mate-pair support. The genomic regions depicted on the Hi-C maps correspond to the highlighted regions on the chromosome ideograms. C) (top) Spontaneous SVs calls on Chromosome 10 of BM175. (middle) WGS coverage of BM175, normalized by the coverage of the maternal C29 cell-line. (bottom) Normalized Hi-C coverage of BM175. Black dots correspond to 10kb genomic regions. Horizontal red lines show

the segmentation output of Sequenza. Vertical dashed gray lines trace the genomic position of the spontaneous SVs.

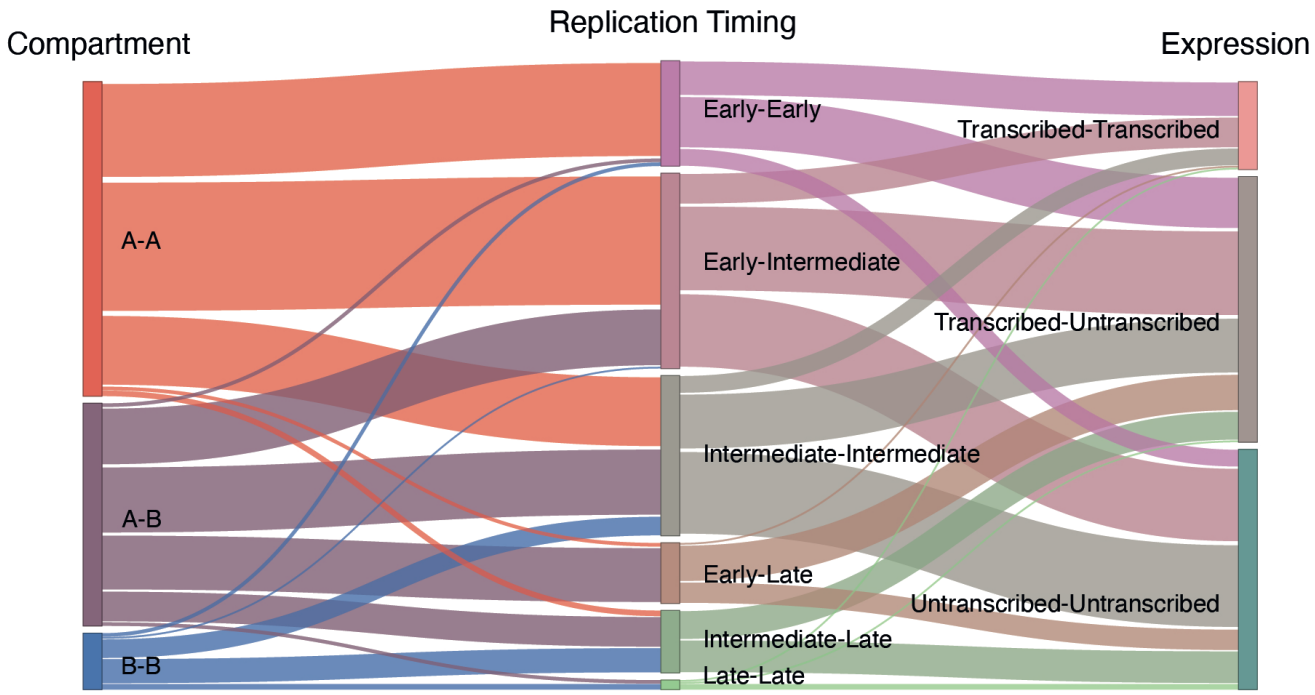
D) Comparison of contact frequency at juxtaposed loci between induced and spontaneous rearrangements. Frequencies are computed at 5kb resolution within a 25×25kb window around the juxtaposed locus. Mann-Whitney *U* one-sided tests between the two distributions reveal significant contact depletion in the spontaneous events (p-value = 0.003 and 0.03 in BM175 and BM178 respectively). Box plots show the median, first and third quartiles, and outliers are shown if outside the 1.5× interquartile range.

E) Formation of DSBs because of doxorubicin treatment (induced) and genomic instability (spontaneous).



Supplemental Figure S5. Association of somatic rearrangements in 2,700 tumor samples with chromosomal compartments and replication timing

A) Induced vs Shuffled DSB recurrence per 1MB genomic bin across the 4 chromothriptic daughter clones (BM175, BM178, BM780, BM838). Only 1 DSB per clone per genomic bin was considered. The shuffled set comprises 1,000 permutations of the original Induced DSB set (see Methods).
B) Distribution of DSBs (left) and SVs (right) on A and B compartments in pan-, breast, prostate, and uterine cancer samples. Uterine tumors show an enrichment in A-to-A juxtapositions.
C) Distribution of replication timing weighted averages (WA) scores in the active A (red) and inactive B (blue) chromosomal compartments. Grey line shows the average WA score genome-wide. Values below the 25% and above the 75% quantile (dashed vertical lines) are assigned to early and late replication respectively.
D) A/B compartment scores are highly correlated with replication timing (Pearson's correlation = 0.77, p-value < 2.2×10-16). Compartment scores are represented by the first principal component eigenvector of the observed/expected Hi-C map.
E, F) Distribution of DSBs (E) and SVs (F) with regards to replication timing. Uterine samples show a striking abundance of Early-to-Early SVs, suggesting an effect of replication timing in SV formation. Early-to-Late SVs are depleted overall.



Supplemental Figure S6. Association between chromatin compartments, replication timing and gene expression on SV formation

In the gene expression column, 'Transcribed-Transcribed' refers to both loci of the SV (*n* = 304) being associated with transcribed genes, and vice versa.

SUPPLEMENTAL METHODS

Cell lines

hTERT RPE-1 cells were grown in DMEM/F-12 (Thermo Fischer Scientific, 11320074) medium supplemented with 10% FBS (Thermo Fischer Scientific, 10500064) and Antibiotic-Antimycotic (Thermo Fischer Scientific, 15240062). All cell lines were maintained in appropriate densities and were incubated in a humidity-controlled environment (37°C, 5% CO2). All cell lines tested negative for mycoplasma contamination.

To generate the Doxorubicin clones, cells were treated with sublethal concentrations of doxorubicin (0.15 μM) for 16h, released into fresh media, and incubated for 3 days before they were sorted for single cells in 96-well plates. From this moment the cells were expanded as single clones and were subjected to WGS (Mardin et al. 2015). The process in total takes roughly 10 days for the cells to grow from single clones.

Sequenced reads from WGS and mate-pair were produced in the original study of induced chromothripsis in RPE-1 cells (Mardin et al. 2015). RNA-seq assay for DCB2 was also produced in the original study. For a detailed description, refer to the Materials and Methods section of that paper.

Preparation of RNA-seq libraries and analysis

Total RNA was purified in duplicate from 1 million frozen late-passage cells using the Direct-zol RNA MiniPrep kit from Zymo Research (#R2050), following the manufacturer instructions. RNA purity, A260/A280 and A260/A230 were assessed with NanoDrop 2000. RNA integrity was evaluated by electrophoresis on a denaturing agarose gel using 350 ng RNA. Clear 28S and 28S bands were observed indicating no RNA degradation. Before strand specific RNA-seq library preparation in duplicate, using the Directional RNA Library Prep kit for Illumina (NEB#7420), in combination with NEBNext Multiplex Oligos for Illumina Set 1 (NEB#E7335), ribosomal RNA depletion was performed using 1 μg RNA using NEBNext rRNA Depletion kit (NEB#6310). Concentration of final libraries was measured using Qubit 2.0 Fluorometer in combination with Qubit dsDNA HS Assay Kit (Invitrogen). Size distribution and fragment length were assessed using Agilent Bioanalyzer HS-DNA Chips. The molarity of individual libraries was calculated considering the concentrations and average fragment size. Pair-end sequencing of 1.6 pM pooled libraries were sequenced with Illumina NextSeq 500 Sequencing System using NextSeq 500/550 Mid Output v2 kit (300 cycles). PhiX Control (Illumina) was added at 1% to the pool as an internal control before sequencing.

RNA-seq reads were aligned using STAR v2.5.3 with --chimSegmentMin 20 and the rest parameters at default. Read coverage was computed at 50bp bins and normalized by Bins Per Million mapped reads with bamCoverage v3.2.0 from the deepTools suite (Ramírez et al. 2016). Transcript expression levels were quantified as Transcripts Per Million (TPM) for all samples and collapsed to gene-level expression by salmon v0.14.0 (Patro et al. 2017) with default parameters and quasi-mapping-based mode. For every gene, we calculated the mean TPM across all available replicates per sample and classified genes with mean TPM < 1 as inactive. Furthermore, we classified active genes as high or low based on the quantile of the TPM distribution of the active genes; high ≥50%, low < 50%.

Differential gene expression between technical replicates of BM175 and C29 was performed by the R package DESeq2 v1.26.0 (Love et al. 2014). To minimize effects of copy number alterations, we adjusted the gene read counts of BM175 based on the respective gene copy number state as estimated by *sequenza*.

Distance effect of DSBs to gene expression was measured as a function of $\log_2$ fold change of the gene with the highest $\log_2$ fold change within non-overlapping 10kb bins from a DSB. To avoid biases introduced by genomic loci with multiple breakpoints, we clustered DSBs that resided within a 50kb distance from each other with BEDTools v2.26.0 (Quinlan and Hall 2010).

Allele specific gene expression was evaluated as in (Ghavi-Helm et al. 2019). Briefly, RNA-seq reads that overlapped one or more informative RPE-1 SNPs were assigned to the respective haplotype. Reads with ambiguous allele assignment were discarded.

Preparation of in situ Hi-C libraries

Five million late-passage RPE cells were crosslinked with 1% formaldehyde for 10 min at room temperature. The crosslinking reaction was quenched with Glycine, then the cells were lysed with Hi-C lysis buffer (10mM Tris-HCl pH8.0 10mM NaCl, 0.2% NP-40) with 1× of protease inhibitors (Sigma-Aldrich complete protease inhibitor cocktail) and the nuclei were recovered. DNA was digested with 100 units of MboI, and the ends of restriction fragments were labeled using biotinylated nucleotides and ligated. After reversal of crosslinks, ligated DNA was purified and sheared to an average length of 400bp using a Covaris S2 instrument (LGC Genomics). DNA

was recovered, and the ligation junctions were pulled down with streptavidin beads and prepped for Illumina sequencing using NEBUltra kit.

Mate-pair based SV calling

Sequenced reads were aligned to the hg19 genome reference with BWA-MEM v0.7.15 and with -T 0 -M parameters. Structural variants were detected by DELLY2 v0.7.7 (Rausch et al. 2012b), using the BAM files of the daughter clones as tumor and the matching maternal clones as normal. A high stringency filter was applied to remove SVs detected in ≥ 1% a set of germline SV calls from 1105 healthy individuals from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). Furthermore, we required at least 4 supporting reads with MAPQ ≥ 20. High-quality SVs coordinates were converted to the hg38 genome assembly by liftOver (Hinrichs 2006). We have previously shown that this filtering approach has a false-positive rate lower than 10% (Rausch et al. 2012a). Complex SVs were identified as described in (Li et al. 2020) and chromothripsis was identified using Shatterseek with default parameters (Cortés-Ciriano et al. 2020). See Supplemental Table 2 for comparative sequencing metrics.

Preparation of ChIP-seq libraries

ChIP-seq libraries were prepared using the CHIP-IT protocol from Active Motif according to manufacturer's instructions. H3K4me3 antibodies were obtained from Active Motif. Libraries were prepared and sequenced on the Illumina HiSeq 2000 platform with 51 cycles according to the manufacturer's instructions.

Repli-seq

20 million cells were grown to an exponential phase and pulsed with BrdU to label the newly replicating DNA. Cells were collected and the cell pellets were resuspended in DAPI staining

solution (1 ×PBS with 0.1% Triton X-100 and 2μg/mL DAPI). Nuclei were then sorted in 6 fractions (G1, S1, S2, S3, S4 and G2) based on the DNA amount. Sorting was done on a FACS Aria Fusion instrument (BD Biosciences) using a 355 nm laser (450/50 BP). Cells were lysed by the addition of SDS-PK buffer (dH_2O with 50mM Tris-HCl, pH8.0/ 10mM EDTA/ 1M NaCl/ 0.5%SDS) containing 0.2mg/mL Proteinase K and 0.05mg/mL glycogen for every 100,000 cells collected. The DNA from these cells is isolated by phenol-chloroform extractions and then sonicated into fragments to an average size of ~0.7-0.8 kb using a Covaris S2 instrument (LGC Genomics). Immunoprecipitation was performed using an anti-BrdU antibody (BD Biosciences Pharmingen, Cat.#555627). Precipitated DNA was recovered by another round of phenol-chloroform extraction and then a second strand synthesis was done by the random hexamers and exo- Klenow enzyme (NEB). Amount of final DNA was measured by Qubit High sensitivity kit and sequencing libraries were prepared by NEBnext ultra kit (NEB) according to the manufacturer's instructions.

SUPPLEMENTAL TEXT

Evidence for bona fide Chromothripsis

To test whether the observed SVs in our samples were compatible with chromothripsis, we followed several lines of evidence. First, we used our Hi-C data to phase the SVs (Tourdot et al. 2021), which confirmed that all SVs were associated with the same haplotype (Supplemental Fig. S1B-C), further supporting that the SVs are induced in a single cell cycle and represent a bona fide example of chromothripsis (Korbel and Campbell 2013). As expected, our analysis of four different daughter cells revealed no breakpoint recurrence and random orientation of SVs (Supplemental Fig. S1C, S5A), supporting that these events are associated with a more general source of genomic instability and not disruption of specific genomic sites or selection of specific SVs. Mutations in key DNA repair enzymes can give rise to specific patterns of complex SVs in cancer genomes (Li et al. 2020; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). We next tested whether specific types of complex SVs were detectable in the genomes besides chromothripsis. We recently developed computational approaches to classify complex SVs into different categories including chromoplexy, templated insertions and breakage-fusion bridge cycle, as well as chromothripsis (Li et al. 2020; Cortés-Ciriano et al. 2020). Applying the methodologies to our highly rearranged samples only revealed chromothripsis as the mechanism of formation, further supporting a random breakpoint formation process and that no other faulty DNA repair-based processes are active in these genomes.

Supportive evidence of a derivative chromosome following chromothripsis

The derivative chromosome assemblies allowed us to gain insight into the complex constituents of a highly rearranged genome in an allele-specific manner. We found that the derivative 2-12-22 chromosome involved the q-arm of Chromosome 22 and fragments of 12q. The Chromosome 22p arm was ligated to the remaining q arm of 12, while 12p acquired the q-arm of the X-Chromosome (Supplemental Fig. S3A-B). We additionally identified another highly rearranged derivative 10-15-7 chromosome as well as 7-15, 11-13, 17-19 and 10-X (Supplemental Fig. S3C). The 10-X fusion was the only large-scale rearrangement inherited from WT RPE-1 cells (Janssen et al. 2011).

To validate the assembly, we used Juicebox Assembly Tools (JBAT) (Dudchenko et al. 2018) on allele-specific Hi-C maps to reconstruct a part of the predicted derivative chromosome, which confirmed the formation of ectopic domains between fragments E, F and H (Supplemental Fig. 2G).

Identifying clonal and subclonal SVs

Dox-induced SVs are expected to be clonal, present in all cells and evident in both WGS and Hi-C assays while spontaneous SVs are expected to be more subclonal and may only be supported by Hi-C. We detected 22 spontaneous events (Supplemental Table 5) on two daughter clones, 13 on Chromosome 10 of BM175 (Supplemental Fig. S4A) and 9 on Chromosomes 11, 12 and 18 of BM178 (Supplemental Fig. S4B) that were not associated with copy-number switches detected from the WGS data generated 20 generations earlier. We further verified that the Hi-C based SVs were indeed associated with copy number changes, when assessed using the Hi-C data (Supplemental Fig. S4C). We also detected 14 pre-existing SVs in C29 which were excluded from the spontaneous SV analysis. We computed the contact frequency in a 25××25kb window around induced and spontaneous rearrangements to create a clonality metric for all SVs. This clonality analysis revealed significantly lower contact frequencies for spontaneous SVs, characterizing them as subclonal events (Supplemental Fig. S4D). The two groups of SVs, induced and spontaneous, gave us a unique opportunity to identify the mechanistic basis for SVs associated with inhibition of TOP2 and SVs associated with subsequent genomic instability in the daughter cell genomes (Supplemental Fig. S4E).

REFERENCES FOR SUPPLEMENTAL MATERIAL

Cortés-Ciriano I, Lee JJ-K, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang C-Z, Pellman DS, et al. 2020. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* **52**: 331–341.

Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, Glenn St Hilaire B, Yao W, Stamenova E, et al. 2018. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *Genomics* 18019.

Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. 2019. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat Genet* **51**: 1272–1282.

Janssen A, van der Burg M, Szuhai K, Kops GJPL, Medema RH. 2011. Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science* **333**: 1895–1898.

Korbel JO, Campbell PJ. 2013. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**: 1226–1236.

Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, et al. 2020. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**: 112–121.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Mardin BR, Drainas AP, Waszak SM, Weischenfeldt J, Isokane M, Stütz AM, Raeder B, Efthymiopoulos T, Buccitelli C, Segura-Wang M, et al. 2015. A cell-based model system links chromothripsis with hyperploidy. *Mol Syst Biol* **11**: 828.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**: 417–419.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–5.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. Nature 578: 82–93.

Tourdot RW, Brunette GJ, Pinto RA, Zhang C-Z. 2021. Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol* **22**: 139.