Extreme genomic complexity in acute myeloid leukemia revealed by combination of long-read sequencing technology and Hi-C

Marius-Konstantin Klever[1,2,3], Eric Sträng[1], Sara Hetzel[4], Julius Jungnitsch[3,5], Anna Dolnik[1], Robert Schöpflin[2,3,6], Jens-Florian Schrezenmeier[1], Felix Schick[1], Olga Blau[1,7], Jörg Westermann[1,7], Frank G. Rücker[8], Zuyao Xia[8], Konstanze Döhner[8], Hubert Schrezenmeier[9,10], Malte Spielmann[5,11], Alexander Meissner[4], Uirá Souto Melo[2,3]*, Stefan Mundlos[2,3,7,]* Lars Bullinger[1,7,12,]*

1. Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Medical Department, Division of Hematology, Oncology, and Cancer Immunology, 13353 Berlin, Germany
2. Max Planck Institute for Molecular Genetics, RG Development and Disease, 14195 Berlin, Germany
3. Institute for Medical Genetics and Human Genetics, Charité University Medicine Berlin, 13353 Berlin, Germany
4. Max Planck Institute for Molecular Genetics, Department of Genome Regulation, 14195 Berlin, Germany
5. Max Planck Institute for Molecular Genetics, Human Molecular Genomics Group, 14195 Berlin, Germany
6. Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, 14195 Berlin, Germany
7. Labor Berlin – Charité Vivantes GmbH, 13353 Berlin, Germany
8. Department of Internal Medicine III, University Hospital of Ulm, 89081 Ulm, Germany
9. Institute of Transfusion Medicine, University of Ulm, 89081 Ulm, Germany
10. Institute for Clinical Transfusion Medicine and Immunogenetics, German Red Cross Blood Transfusion Service Baden-Württemberg-Hessen and University Hospital Ulm, 89081, Ulm, Germany
11. Institut für Humangenetik Lübeck, Universität zu Lübeck, 23538 Lübeck, Germany
12. German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

*These authors contributed equally to this work

Corresponding authors:

Lars Bullinger, Dept. of Hematology, Oncology and Tumorimmunology, Charité University Medicine,

Augustenburger Platz 1, 13353 Berlin, Germany.

Phone: +49-30-450-553192, Fax: +49-30-450-553987,

E-mail: lars.bullinger@charite.de

Stefan Mundlos, Institute for Medical Genetics and Human Genetics, Charité University Medicine,

Augustenburger Platz 1, 13353 Berlin, Germany.

Phone: +49-30-450-569122, Fax: +49-30-450-569915,

E-mail: stefan.mundlos@charite.de

Supplemental Methods

Samples collection and processing

Supernumerary material, not needed for clinical allogenic hematopoietic stem cell transplantation, was used for this study. Stem cell preparation was performed by CD34-selection via CliniMACS-LScolumns (Miltenyi Biotech), and T-cell depletion using SAM-Beads (Miltenyi Biotec) and Orthoclone OKT3-antbodies (Janssen/Cilag).

Illumina RNA sequencing detailed bioinformatic analyses

RNA sequencing reads were processed by adapter and quality trimming with cutadapt[42], (version 2.4; parameters: --nextseq-trim 20 --overlap 5 --minimum-length 25 --adapter AGATCGGAAGAGC -A AGATCGGAAGAGC),[40] followed by poly-A trimming with cutadapt (parameters: --overlap 20 -minimum-length 25 --adapter "A[100]" --adapter "T[100]"). Trimmed reads were aligned to the human reference genome (hg19) using STAR[23], (version 2.7.5a; parameters: --runMode alignReads -chimSegmentMin 20 --outSAMstrandField intronMotif --quantMode GeneCounts) and transcripts assembly was performed using stringtie[24] (version 2.0.6; parameters: -e) with the GENCODE annotation (release 19). Furthermore, Illumina RNA sequencing data was processed with the fusion caller JAFFA,[22] using the direct mode of this software with standard settings and alignment to hg19. The identified fusion transcripts were compared to our translocation and inversion breakpoint dataset using bedtools closest[43] with a cutoff of less than 30kb distance to both corresponding genomic breakpoints. Per patient, genes with an average TPM of less than 0.5 across patient and control replicates were excluded from downstream analysis.

The GO terms resulting from our GO analysis were grouped by semantic similarity using the R package GOSemSim[44] with the semantic similarity measurement by Wang et al.[45] the following way: An undirected graph was constructed using GO terms as nodes. An edge was drawn between two GO terms if their semantic similarity score was above 0.7 and edges where weighted by the respective semantic similarity score. Groups of similar GO terms were determined by identifying the minimum spanning forest of the graph using Kruskal's algorithm.[46] For the resulting groups of similar GO terms a representative GO term was selected based on the minimum False-Discovery Rate (FDR) of all GO terms per group. Up- and down-regulated genes were tested separately per patient against the genomic background using the "Biological process" GO database.[47] The resulting GO terms were grouped by semantic similarity using the R package GOSemSim[44] with the semantic similarity measurement.[45] For the resulting groups of similar GO terms a representative GO term was picked based on the minimum False-Discovery Rate (FDR) of all GO terms per group.

Microarray expression data

The previously generated microarray data that we include in this study was generated as follows. All samples were processed as follows: Total RNA isolation was performed using Trizol reagent (Invitrogen, Carlsbad, CA). For targeted preparation according to the manufacturer's protocol, 0.2 microgram of RNA was used (GeneChip Whole Transcript Sense Target Labelling Assay manual, Affymetrix, Santa Clara, CA). The arrays were scanned with the Affymetrix GeneChip® Scanner 3000. Probe set summarization, background correction and normalization were performed by applying the robust multiarray average (RMA) algorithm. The exon array data were analyzed in R, oneChannelGUI 1.10.7.[48] A z-score expression data heatmap of this cohort was generated using Morpheus (Morpheus).

Oxford Nanopore Technology (ONT) DNA library preparation and sequencing

After extraction, genomic DNA (gDNA) was prepared for ONT WGS using the Ligation Sequencing Kit (ONT, SQK-LSK109). We also performed ONT direct cDNA sequencing using mRNA that was processed with the Dynabeads mRNA Purification Kit (Thermo Fisher Scientific, 61006) for mRNA isolation from total RNA. The mRNA was reverse transcribed and prepared for ONT sequencing using the direct cDNA Sequencing Kit (ONT, SQK-DCS109). All ONT libraries were sequenced on a GridION on R9.4.1 flowcells (ONT, FLO-MIN106D). ONT WGS of gDNA libraries was performed until a coverage of at least 10x for each patient was reached.

Oxford Nanopore Technology (ONT) DNA data analysis

The average N50 read length of all runs in our dataset was approx.7.1kb. Base calling was performed using ONT standard base caller Guppy. Fastq files of the gDNA runs were merged and aligned to the human reference genome with the NGMLR long read aligner[49] and processed with samtools to generate bamfiles.[50] The sorted and merged bamfiles of all sequencing runs for one patient were processed with the ACE tool[51] at 1 kb, 5 kb, 10 kb, 50 kb, and 100 kb binning size to search for CN alterations. For the final CN dataset, the 10 kb binning size ACE dataset was used, because of comparably high resolution but still very high accordance with the Hi-C coverage comparisons. This CN dataset was further refined by visual detection of CN changes in the Integrative Genomics Viewer (IGV) version 2.7.2.[52] The CN estimations that were derived from visual inspection were validated by comparing local genomic coverage with igvtools.[53] This enabled us to also identify the CN of fragments that were too small for the ACE CN analysis (< 10-20 kb fragments). For detection of SV breakpoints, the gDNA bamfiles were processed with the long-read SV caller NanoVar, Version 1.3.2.[21] NanoVar was primarily executed with standard filtering criteria, without filtering the SV calls by a confidence score (CS), to detect as many SVs as possible that were already detected by Hi-C sequencing. After comparing SV calls from Hi-C with the NanoVar SV calls, we were able to adjust the filtering criteria without excluding any true positives. For the final dataset, a CS of 0.4 was used for filtering. To also identify breakpoints of very small segments, which we did not identify by primary visual inspection of the Hi-C maps, all NanoVar translocation and inversion calls within 1 Mb at the Hi-C breakpoints were searched for in the Hi-C map.

Oxford Nanopore Technology (ONT) direct cDNA sequencing library preparation and analysis

The identified fusion transcripts were compared to our translocation and inversion SV call dataset using bedtools closest[43] with a cutoff of less than 30 kb distance to both corresponding genomic breakends. They were also compared to the fusion transcripts that were identified by running JAFFA on the Illumina RNA sequencing dataset. Fusion calls in the Illumina RNA and ONT direct cDNA datasets were identified as matching if they were located at the same exon-intron boundary.

Hi-C data analysis

Sequencing data of the Hi-C libraries was processed by the Juicer pipeline.[54] The output files of Juicer for all replicates were merged and read pairs with mapping quality (MAPQ) ≥ 30 were used to generate the final maps. We used Juicebox software for visual inspection of the Hi-C maps.[55] We analyzed the Hi-C data with HiNT-CNV, a coverage-based HI-C CN analysis method, that we executed using standard settings at 50kb binning size on the human reference genome (hg19). The high accordance of the HiNTCNV dataset with the ACE CN data is shown in supplemental Figure 2.

Functional evaluation of fusion genes

To evaluate potential biological effects of the USP7/MVD fusion transcript, RNA from case CK2-Mut was extracted and reversely transcribed to cDNA. The cDNA was then amplified and cloned in a pRSF91 retroviral vector. PCR primers for cloning in the pRSF91 retrovirus backbone were used as following: USP7 Ex1-Ex1 (AGAGACCGGTACCATGAACCACCAGCAGCAGCA (USP7_forward, Ex1) of USP7 and Ex2Ex6 (AGAGACGCGTGTCCCTGCTGAGGCAGTC (MVD_reverse, Ex6) of MVD. The retrovirus was produced by transfection of the pRSF91-USP7-MVD fusion transcript as well as VSV-G and packaging proteins using TRANS-LTI (Mirrus). The viral supernatant was collected after 36h and 48h. For viral infection, cells were seeded with 1 to 2 µg/ml Polybrene (Sigma). At concentrations of 0.5-2 µg/ml, the cells were puromycin selected 1 day after transduction. Further cultivation for assaying cell proliferation was conducted in RPMI 1640 (Mediatech) or DMEM (Mediatech) with 10-20% heat-inactivated FBS and 1% penicillin, streptomycin, and L-glutamine (Mediatech).

Identification of breakend signatures and genomic distribution

We calculated the distribution of breakends inside of repetitive elements and the distribution of breakends close to these breakends. The distribution shows the normalized occurrence of repetitive elements within a given distance to the breakend. Analysis of local breakend density was performed by kernel density estimation for each chromosome. The local density of translocation and inversion breakends was visualized using ggridges in R with kernel bandwidths 1 Mb, 1 kb and 0.5 kb. To calculate a genomic background rate to simulate a potential random distribution, the total size of these features was compared to the total size of the genome. This rate was used for the calculation of observed/expected values. All regions on sex chromosomes were excluded because none of these were affected by translocations or inversions in any of the cases reported here. Significance of enrichment of repetitive elements and other genomic features inside or in close proximity to breakends was investigated with a two-sided Mann-Whitney-U-test against 10000 random breakpoints (telomeres/centromeres and variable genomic regions excluded). Repeat classes from repeatmasker were investigated as well as 3'UTR, 5'UTR, introns , exons, intra- and intergenic-regions, excluding features with less than 1% prevalence in the human genome. These tests were performed for breakpoints inside the respective genomic regions and separately for increased occurrence of breakpoints around the respective genomic features. The findings were corrected by the Benjamini-Hochberg procedure. The distances and their 95% confidence intervals were estimated by boot-strap over 5000 iterations.

Supplemental Results

Cohort overview and structural variant detection using Hi-C and ONT DNA sequencing

Integrating Hi-C and ONT-GS analyses our workflow maximizes the potential of both technologies by thoroughly removing false-positives SV calls from the final dataset. For ONT GS data, this mainly concerns false-positives SV calls due to the high mismatch error of this technology. For Hi-C, exact identification of correct breakends and discrimination from breakend-like patterns of small fragments (<20 kb) is not possible with certainty.

Integrative SV analysis reveals genomic differences in TP53 mutated vs. TP53 wildtype cases

In order to identify the correct connections of these remaining fragments, we also projected additional NanoVar SV calls ranging from 100 kb up to 1 Mb around the putative BND on the Hi-C map. We observed that these two missing connections did not directly link to each other, but to two small fragments (N1, 1.3 kb and N2, 5.2 kb) (Figure 2c; supplemental Figure 3a). These fragments were located in between the presumed BND and joined them together (supplemental Figure 3a), and based on the combined ONT GS and Hi-C approach, we were able to likely reconstruct the derivate 7 and 8 chromosomes (supplemental Figure 3b,c).

Identification of "chromocataclysm" - extremely locally clustered chromothriptic rearrangements showing focal amplifications of kilobase and sub-kilobase regions of the genome

Most SV breakends in our dataset were associated with a CN change very close or directly at the breakend (supplemental Figure 7a). Due to this feature, which matches the definition of chromothripsis, we were able to display the CN state of fragments <1 kb, including their connections to other fragments. This can only partly be visualized due to the enormous complexity of some of the rearrangements (supplemental Figure 7b-d). Analysis of the CN state distribution of all fragments (including whole chromosomes but also very small fragments) showed that all CK-AML cases harbored fragments with CN gain as well as CN loss (Figure 3a). However, TP53 mutated cases showed higher genomic complexity than the TP53 wildtype cases and the distribution of the fragments varied substantially. Based on our analysis, many of the cases harbor amplified fragments with a CN of ≥ 5 (supplemental Figure 8a). Analysis of CN states of small fragments (<20 kb) in the four cases with the most total fragments showed that CK4-Mut harbors very few of these small fragments compared to the other cases (Figure 3b). We did not observe the chromocataclysm pattern in this case, while it was observable in the other three cases.

Overall, the distribution of BNDs in specific chromosomes and cytogenetic bands in all cases showed enrichment of translocation or inversion BNDs on chr7 (N=6), chr19 (N=5), chr3 and chr12 (N=5), and chr5, chr8, chr11 and chr17 (N=3). Furthermore, cytogenetic bands on chr7 and chr17 were recurrently affected by BNDs in three cases (7q11.21; 7q21.3; 7q22.1; 17p11.2; 17p12) (supplemental Table 3). This result led us to investigate if these frequently breakpoint-hit chromosomes in CK-AML show a similar pattern of gene expression among the cases reported here. Furthermore, we sought also to identify regions of the genome prone for being hit by a breakpoint in CK-AML cases.

Gene expression analysis revealed a chromothripsis associated pattern of CN losses and gene down regulations in CK-AML related genomic regions

Commonly up-regulated genes (up-regulated in at least 6 out of 9 patients) were enriched in gene ontology (GO) terms of leukocyte biology and function as well as processes of the immune system, while commonly down-regulated genes were enriched in GO terms of fatty acid biosynthetic process and blood circulation (supplemental Figure 9a,b).

The genomic distribution showed that all gain and upregulation candidate genes were located on 8q, a chromosome arm that is known to be repeatedly affected by CN gains in CK-AML. On the other hand, the loss and downregulation candidate genes were located on chromosome arms 7q, 16q, 12p, 17p and 18q, regions that are known to be repeatedly affected by CN losses in CK-AML (supplemental Figure 9c).[7,56] Interestingly, almost all of the 30 loss and downregulation candidate genes showed this pattern only in cases that we previously identified as chromothripsis. Only two of these genes, ETV6 and LRP6, showed a CN loss and gene downregulation state in a case that was not classified as chromothripsis (CK7-Wt).

While some of these genes could have a CNV alteration specific role in CK-AML, other genes seem to be also dysregulated in cases without CN alteration. In these cases, epigenetic mechanisms or the alteration of upstream transcription factor genes could lead to the resulting expression profile. To further investigate a CK-AML related role of our candidate genes, we compared our findings to our own microarray-based GE dataset previously generated in CK-AML cases (n=39) and CD34+ healthy controls (n=3),[27] and the Beat AML RNA expression dataset from AML with myelodysplasia related changes (n=87) and CD34+ healthy controls (n=21).[57] One of the CN gain and upregulation candidate genes was TRIB1, a pseudokinase that is also thought to function as an oncogene in several malignancies and is involved in TP53 signaling.[58] This gene showed high expression values (z-score transformed gene expression array signal) in our microarray GE dataset as well as in the Beat AML dataset, when compared to CD34+ healthy controls. Similarly, TNK1, a known tumor suppressor gene[59] and one of the candidate genes with CN loss and downregulation showed low expression values (zscore) in the AML cases of our two reference datasets (supplemental Dataset 5; supplemental Dataset 6). Taken together, our gene expression analyses revealed a common loss and downregulation/gain and upregulation pattern that seems to play a role in CK-AML in general, but further studies are warranted to elucidate the role of each of these genes in leukemia biology.

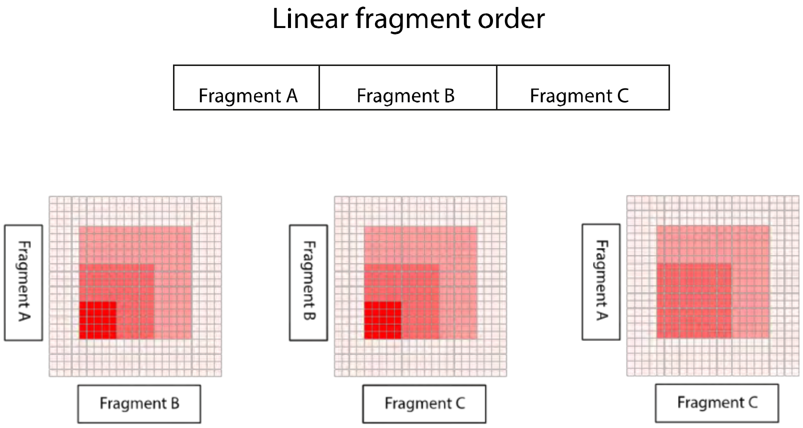
Supplemental Note

Breakend definition

We defined used the term breakend (BND) here to summarize interchromosomal and intrachromosomal breakpoints from our CK-AML cohort. These breakpoints were derived from trans (interchromosomal) and cis (intrachromosomal) Hi-C maps. In these maps, patterns of high visual signal intensity that were not present in the CD34+ stem cell control maps, were defined as potential breakpoints and integrated with potential breakends from the NanoVar tool (ONT WGS data). NanoVar SV calls from the BND and INV (inversion) categories were used for this integrative analysis.

Breakend signatures in Hi-C

In Hi-C, interaction of two genomic loci is represented by signal intensity in the Hi-C maps. The closer two genomic loci are in the 3D genome, the more interaction signal is visible in the Hi-C maps. Breakends of translocations and inversions are visible in Hi-C maps as a strong local signal, which is not visible in control maps (here we used CD34+ stem cell maps). In most cases, especially with simple rearrangements, the difference in signal intensity is very high and differences are easy visible. Translocations are visible in the trans-maps (interchromosomal) and inversions in the cis-maps (intrachromosomal). The point with the highest signal intensity is regularly the actual breakpoint, because the signal intensity decreases rapidly with genomic distance. An important but difficult to handle phenomenon that we observed here are breakend-like patterns, when dealing with very small fragments as reported in this study. If two loci are not directly linked via a breakpoint but show close proximity to each other due to for example a smaller fragment that is connected with both of them, the Hi-C signal of the connection of these fragments can appear as a breakend-like pattern. However, we saw in our dataset that these breakpoint like–patterns are well distinguishable from real breakpoints when the fragment size is in the range of about > 100kb, due to differences in local signal intensity at the putative breakpoint (supplemental Note Figure 1). For smaller fragments, a clear distinction of breakends and breakend like patterns would not be possible without integrating the data with the SV data from NanoVar.

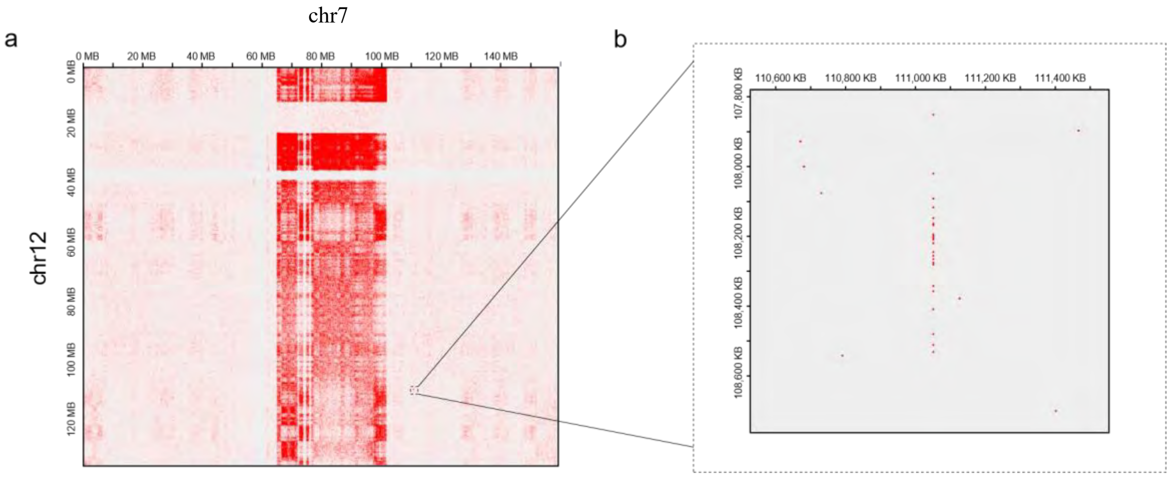


Structural variant detection pipeline

The pipeline that we report here analyses Hi-C and ONT WGS data in a novel integrative way, in order to maximize the potential of both technologies and thoroughly remove false-positives from out dataset. The SV detection for each patient started with a primary visual inspection of all Hi-C maps for SV breakpoints. The lists of all potential Hi-C breakpoints were then screened for directly corresponding breakpoints in the SV breakpoint call list that we generated based on the NanoVar SV caller which we executed on our ONT WGS data (see methods). Furthermore, all NanoVar SV calls in the range of 1 Mb around both coordinates of the putative Hi-C breakpoints were re-analyzed for additional Hi-C support. This was done in order to identify small fragments in clustered breakpoint regions, which were too small to be detected only by Hi-C visual inspection. If additional Hi-C support for a Nanovar SV call was detected, we started the pipeline again by executing the same procedure on these new Hi-C breakpoints that we performed on the Hi-C breakpoints from the primary visual inspection. This procedure was repeated until no additional NanoVar SV calls with Hi-C support were identified. This enabled us to generate genome wide CN estimations which we integrated with the SV breakpoint calls. In summary, we only integrated SV calls in our final SV call dataset that were supported by Hi-C sequencing as well as by NanoVar SV calls, except for a very small fraction of fragments (less than 3% of all found fragments) which were enclosed on both sides of the fragment by a NanoVar translocation or inversion call and were < 10 kb in size but lacked Hi-C support. This enabled us to thoroughly exclude false positives from our dataset.

Chromothripsis definition

For the assessment of rearrangements as chromothripsis, we used the definition of at least 10 copy number (CN) changes on a single chromosome that was already used by other studies previously.[13] In our final dataset all the five TP53 mutated and one of the TP53 wildtype CK-AML cases (CK6-Wt) fulfilled this definition. Case CK5-Mut primarily showed rather simple karyotyping results as well as a simple pattern of rearrangements at primary inspection of the Hi-C maps (supplemental Note Figure 2a). However, after detailed analysis, also this case fulfilled the criteria for chromothripsis and harbored complex rearrangements of smaller fragments (supplemental Note Figure 2). Two cases that we present here (CK2-Mut and CK3-Mut) were already examined by CN analysis using SNP microarray data. The same cutoff value for chromothripsis was used in this study but the data showed much fewer copy number losses and gains as seen by our integrative SV detection approach (supplemental Table 2).[27]



Chromothripsis with extreme local breakpoint clustering (Chromocataclysm)

We termed the phenomenon of extreme local breakpoint clustering that we that we found in small regions of some chromothriptic rearrangements as chromocataclysm. We defined chromocataclysm as the occurrence of 4 or more breakends in a region of 5 kb, in a single chromothriptic rearrangement. This requirement was fulfilled by the chromosome 15 and 16 rearrangement in CK2-Mut as well as a rearrangement including material from chromosomes 7, 9, 12, 17, 18, 19 and 20 in CK3-Mut and a rearrangement including chromosomes 5, 7, 10, 11, 12 and 19 in CK6-Wt. Interestingly, less complex rearrangements occurred in addition to these extreme complex rearrangements in all these three cases (supplemental Dataset 1). Most notably, CK2-Mut showed an additional chromothriptic rearrangement of chromosome 3, 7, 20 and 22, which was not connected to the chromosome 15 and 16 rearrangement but did show a much lower level of clustering thereby not meeting the criteria for a rearrangement with extreme local breakpoint clustering. This suggests that these rearrangements may occurred independently. These extreme complex rearrangements consisted largely of fragments that were much smaller than those detected with previously used methods and were amplified.[12,13,31]

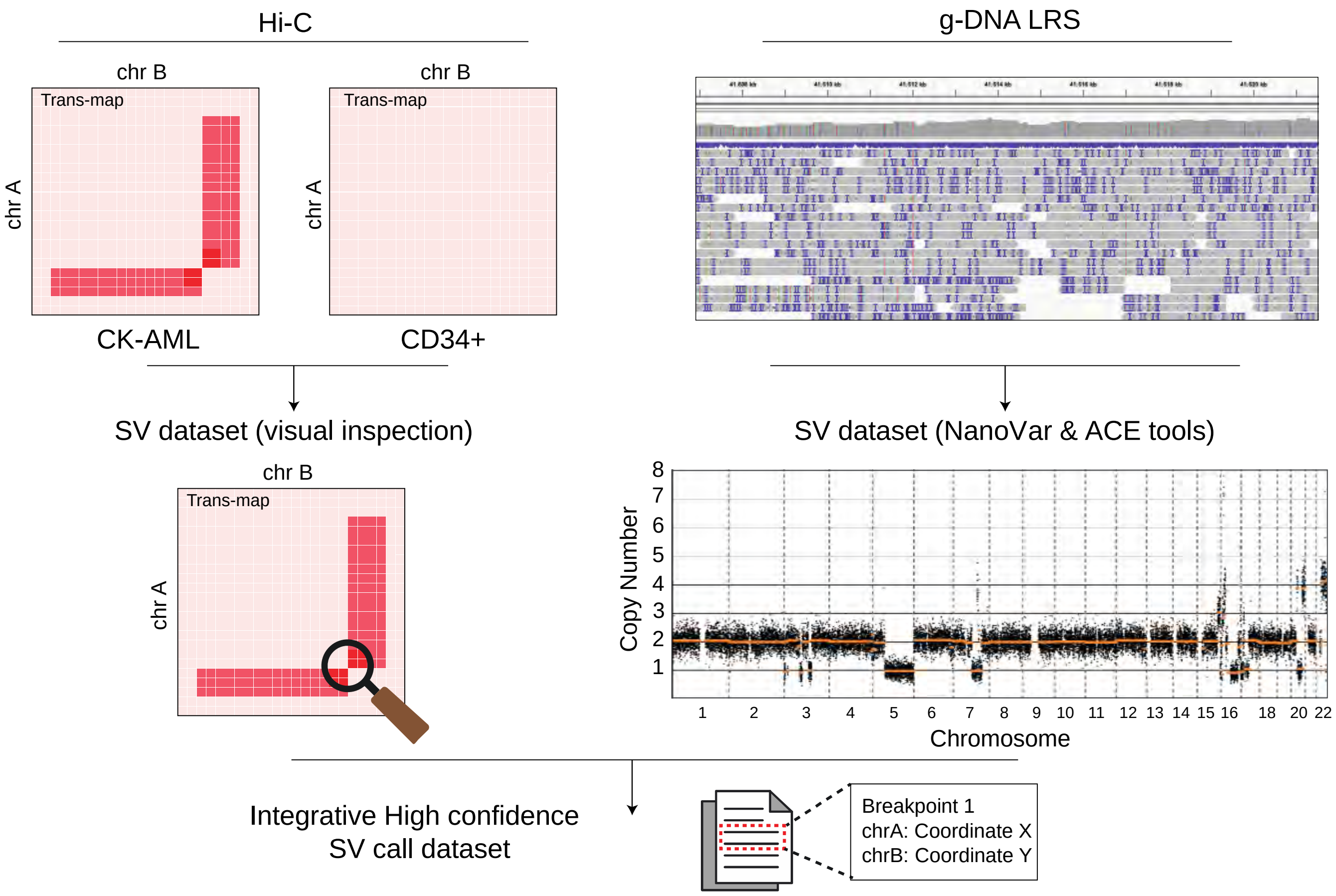
Fusion transcript detection pipeline

We developed an integrative pipeline for fusion transcript detection in our samples. For this pipeline, we wanted to be able to detect fusion transcripts with a high confidence. For this, we only integrated fusion transcripts that were supported by 2 transcriptome sequencing methods (ONT direct cDNA and Illumina RNA-Seq) and by corresponding genomic breakpoints supported by 2 genome sequencing methods (ONT WGS and Hi-C) in <30 kb distance from to both transcriptomic "breakpoints" of the fusion transcript. We set this cutoff value of <30 kb to also account for intronic breakpoints that are likely distant from the exon-exon fusions. The fusion transcripts that were discovered by this pipeline were matched with the fusion transcript information in the ChimerDB 4.0 database of known fusion transcripts.[60]

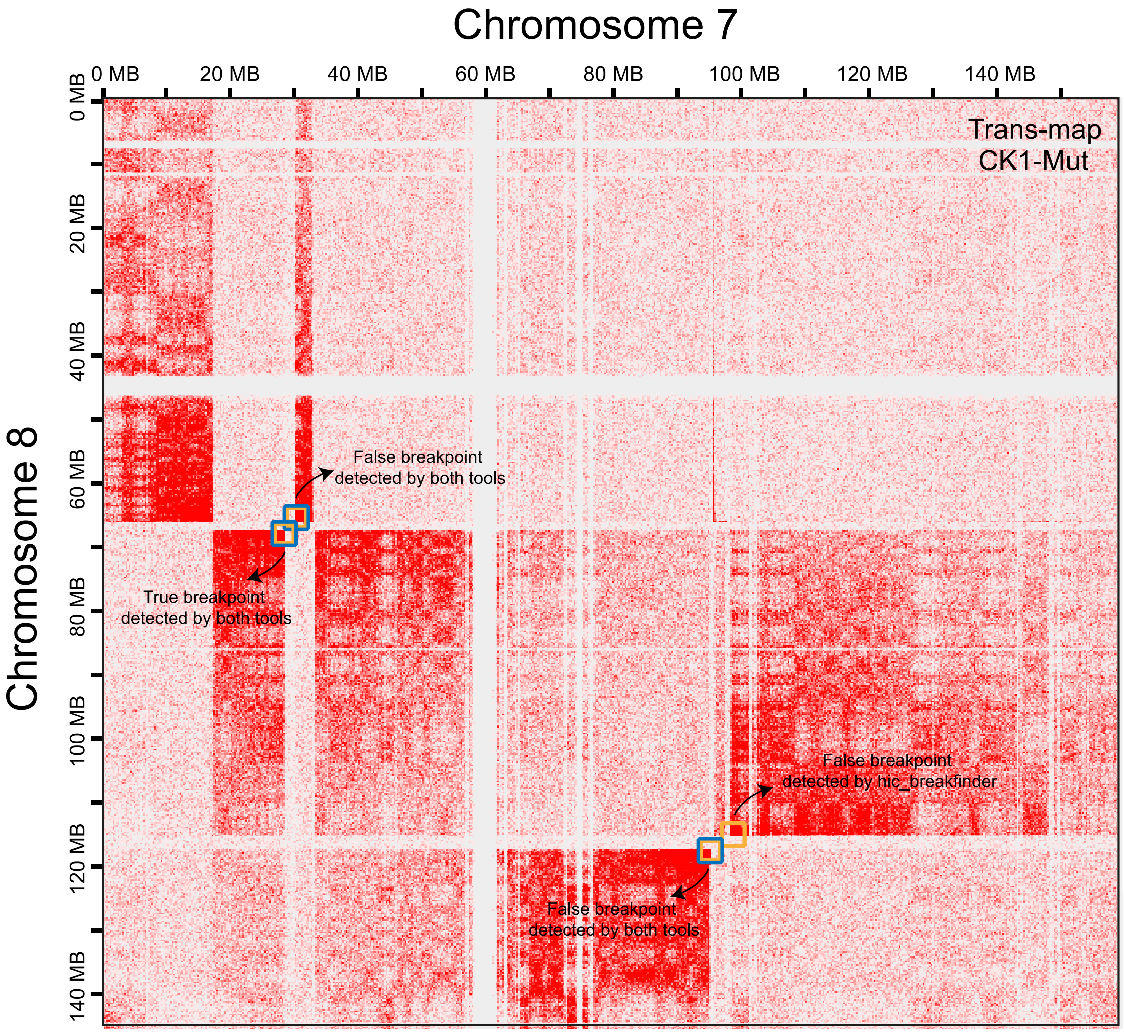
Function of genes that are part of the found fusion transcripts

USP7 is an ubiquitinase that plays a complex role in the interaction of MDM2 and TP53. Also, it is well known to play a role in several cancers including AML.[38,39,61] NUP88 is frequently overexpressed in cancer[62] and was also linked to NF-κB signaling in AML.[63] ANKRD12 is very similar paralog of ANKRD11, a putative tumor-suppressor gene and known co-activator of TP53 and is itself thought to activate TP53.[64,65] ARGHAP44 was recently described as a target of mutant TP53 in cancer.

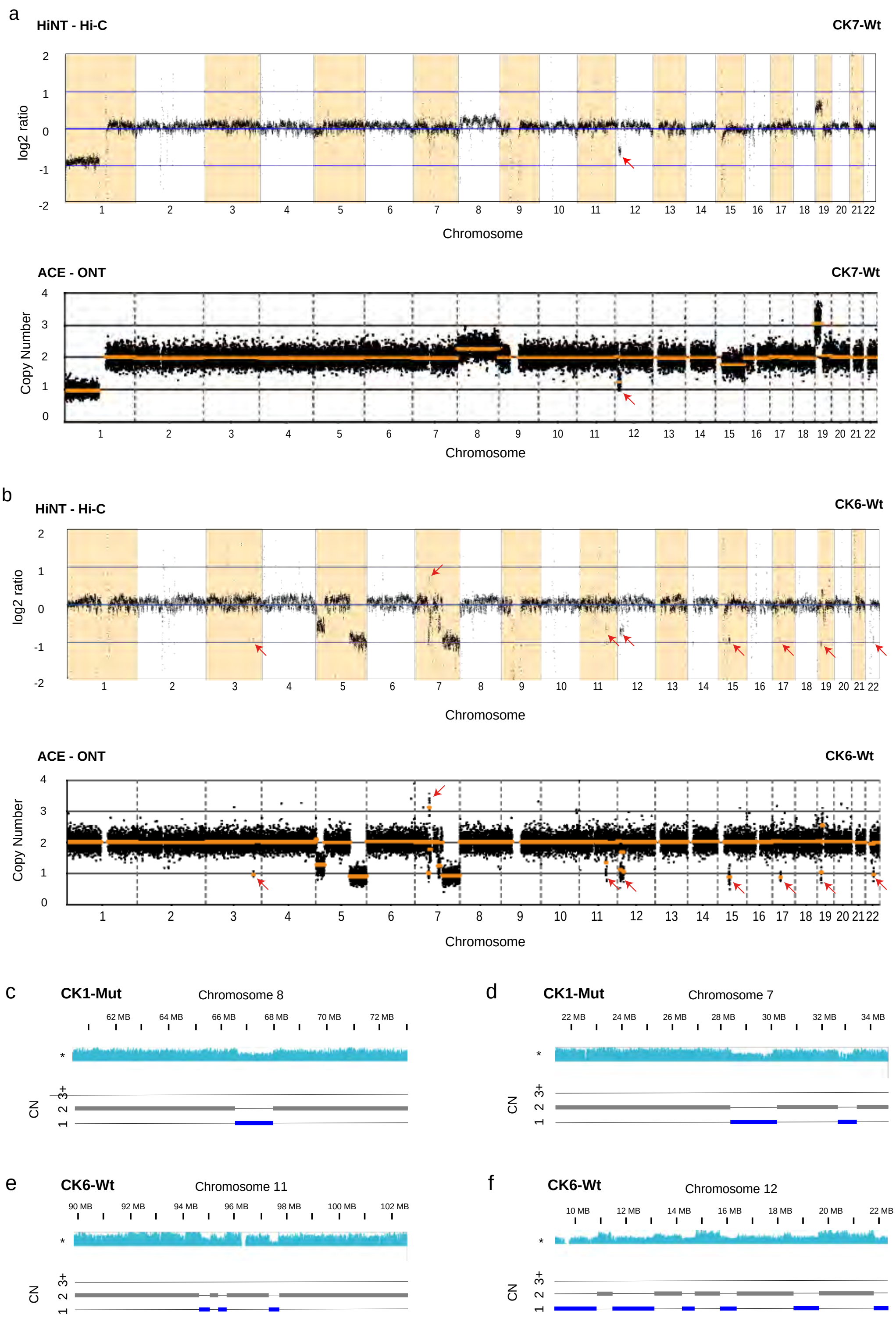
Supplemental Figures



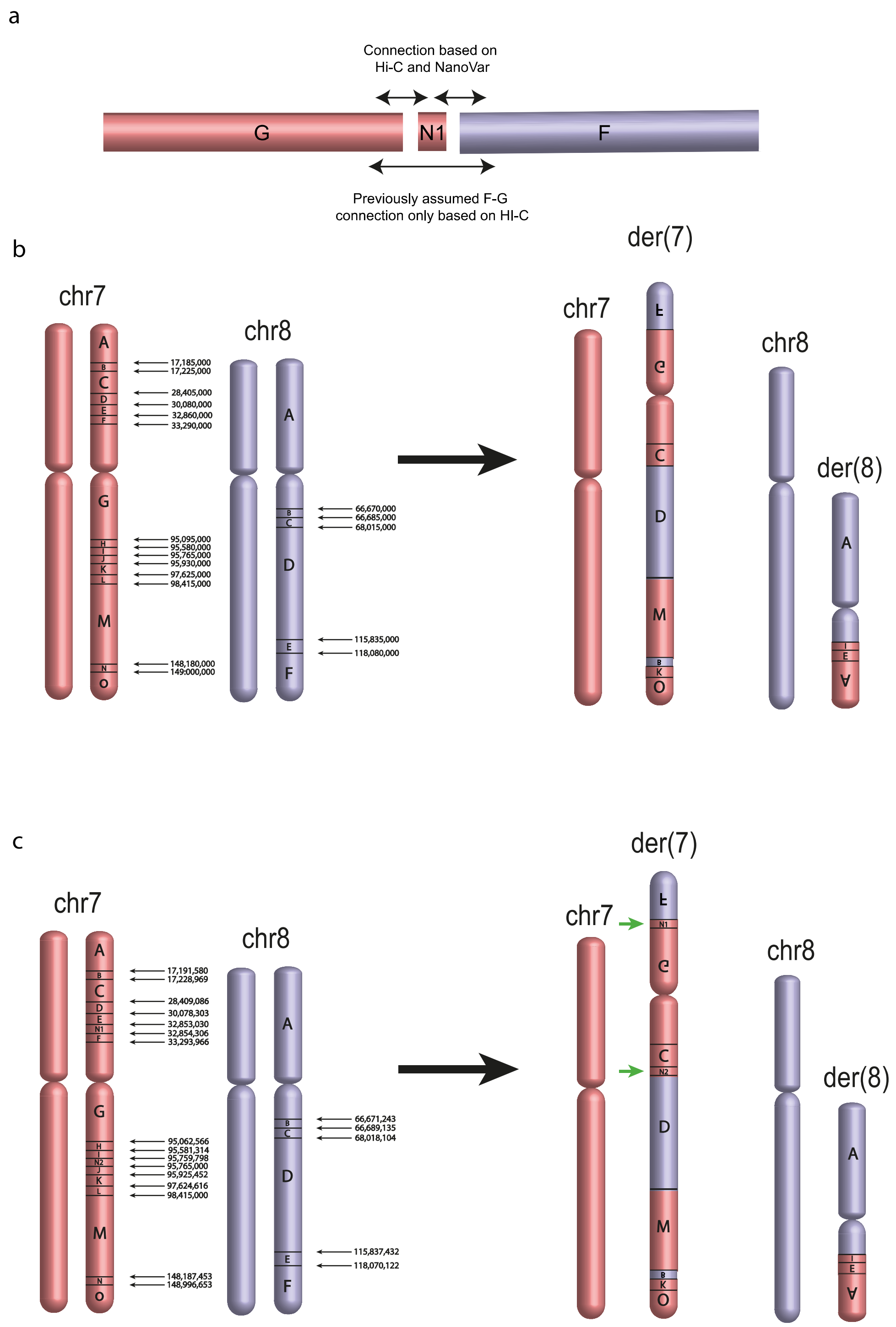
Supplemental Figure 1. Exemplary visualization of our SV detection workflow. Hi-C maps were analyzed based on visual inspection of the CK-AML maps compared to the CD34+ stem cell control maps. The SV calls that were observed in Hi-C were then integrated with ONT GS data that was analyzed with NanoVar and ACE tools. Based on these technologies, we obtained a final high confidence SV call dataset.



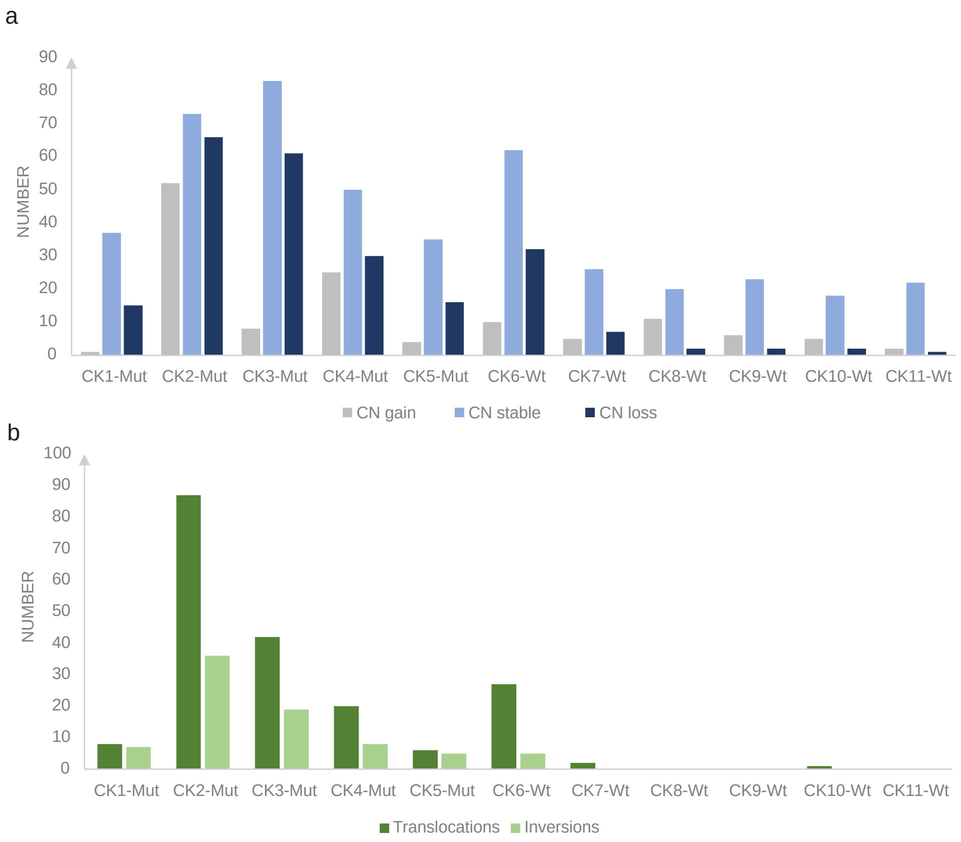
Supplemental Figure 2. Translocation calls from HiNT (black squares) and hic_breakfinder (yellow squares) projected on the Hi-C trans map (CK1-Mut) of chromosome 7 and 8 (**c**). For this rearrangement, both callers detected 1 out of of 8 Translocation calls that were present in our SV dataset. In addition, 2 false-positive calls were called by HiNT and 3 false-positive calls were called by hic_breakfinder. MB, genomic position in megabases.



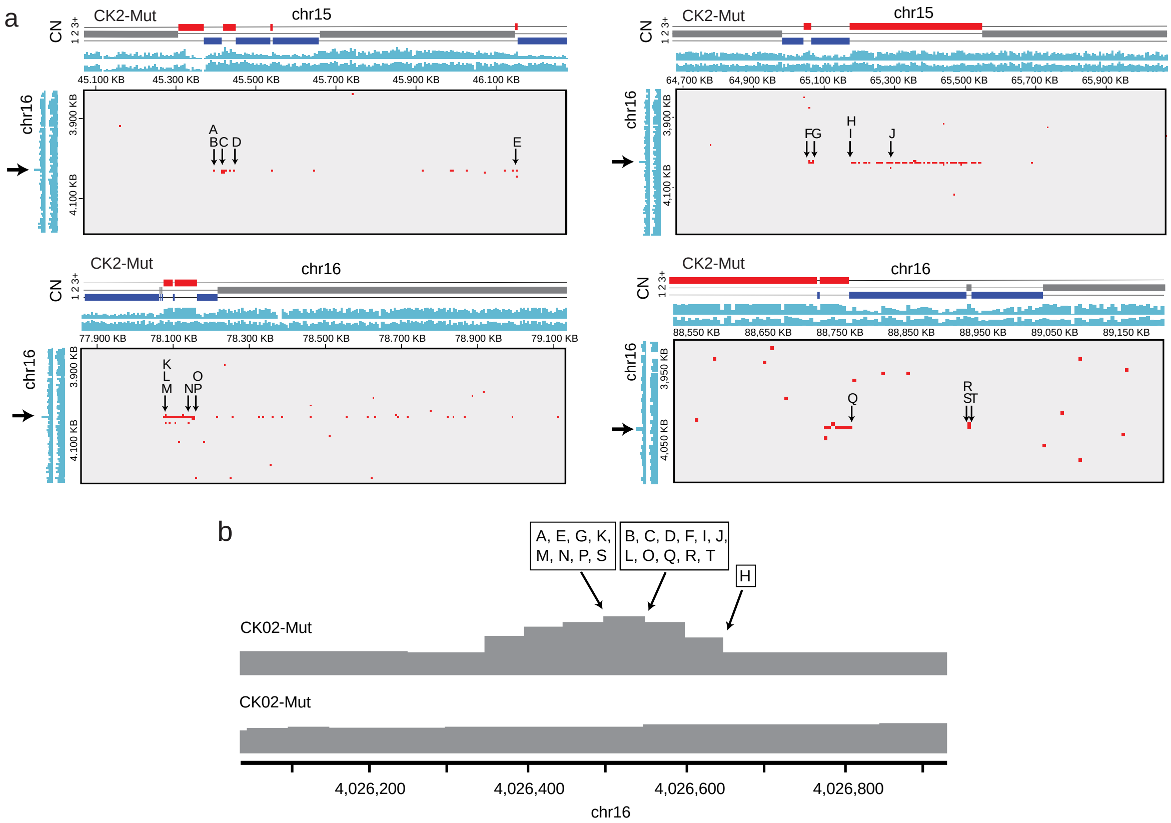
Supplemental Figure 3. HiNT and ACE copy number (CN) results. (**a**) and CK7-Wt and (**b**) CK6-Wt results from HiNT and ACE revealed a highly correlation between both tools. ACE CNV results at 100kb binning size (gDNA long read sequencing). Selected smaller fragments with a CN > 2,3 or < 1,7, which were present in the HiNT as well as in the ACE dataset are marked by red arrows (**a,b**). The accordance of the Hi-C coverage data with CN estimations from ACE was well preserved in all analyzed cases. (**c-f**) Hi-C coverage data is shown as turquoise coverage tracks above the CN data from ACE. ACE CN data is shown as blue bars in the size and position of the respective fragment. Selected regions with CN changes of CK1-Mt (**c,d**) and CK6-Wt (**e,f**). CN, Copy number; MB, genomic position in megabases.



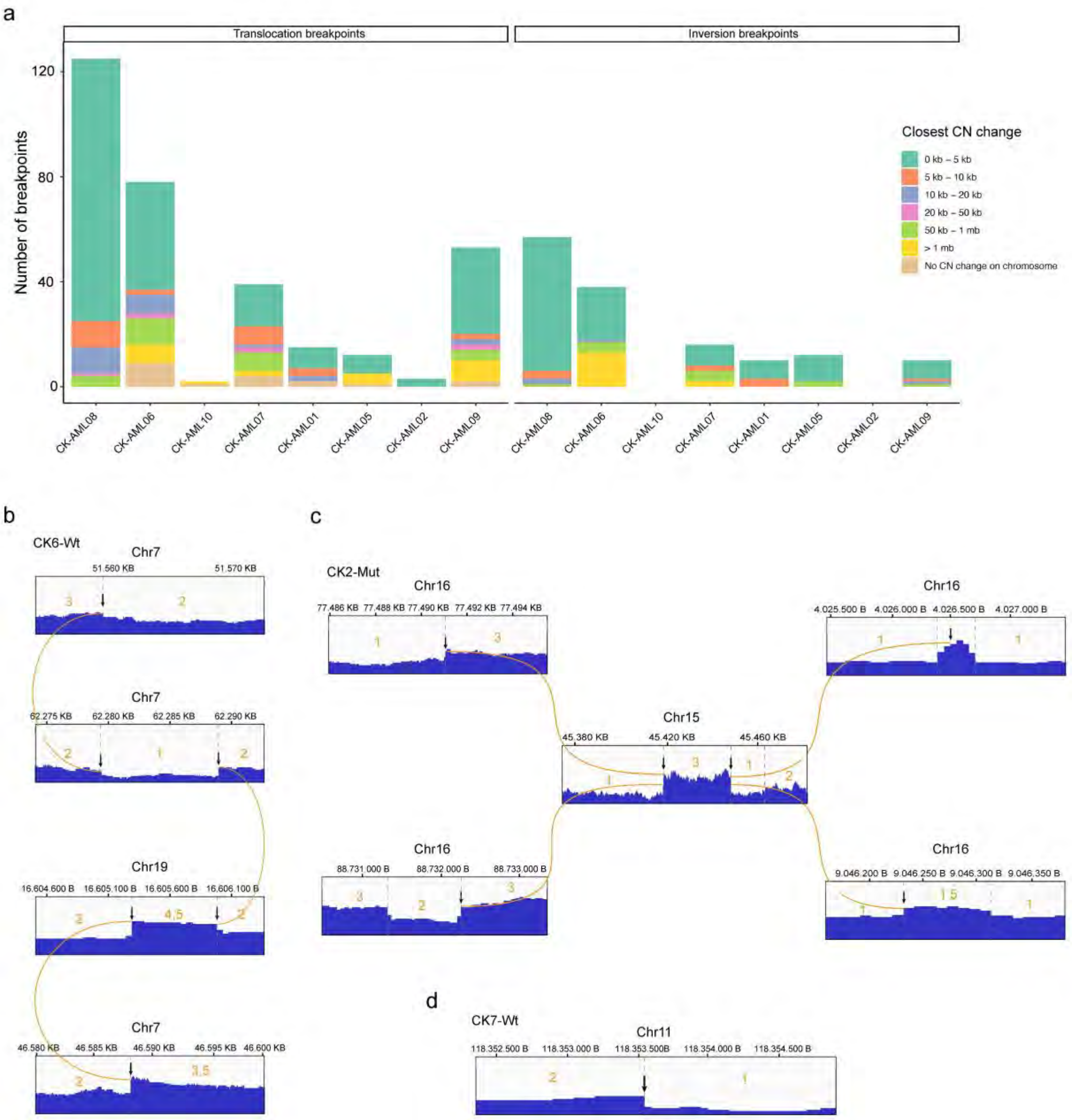
Supplemental Figure 4. Reconstruction of the derivate chromosome structure of a chromothriptic rearrangement (chromosome 7 and 8) in CK1-Mut based only on Hi-C sequencing. The position of Hi-C breakpoints is marked by small black arrows (**a**). Reconstruction of derivate chromosome structure based on Hi-C and NanoVar data (**b**). N1 and N2 mark fragments which were identified after integration of Hi-C data with NanoVar. The position of these fragments in the derivative chromosomes is marked by green arrows. The base pair position of the breakpoints after integration of Hi-C data with NanoVar in (**c**) are marked by small black arrows. Schematic view of the N1 fragments and the connections to parts F (chr7) and F (chr8) in the context of the previously assumed F-G connection. MB, genomic position in megabases; KB, genomic position in kilobases; der, derivative chromosome. All chromosome positions are based on the GRCh37/hg19 assembly.



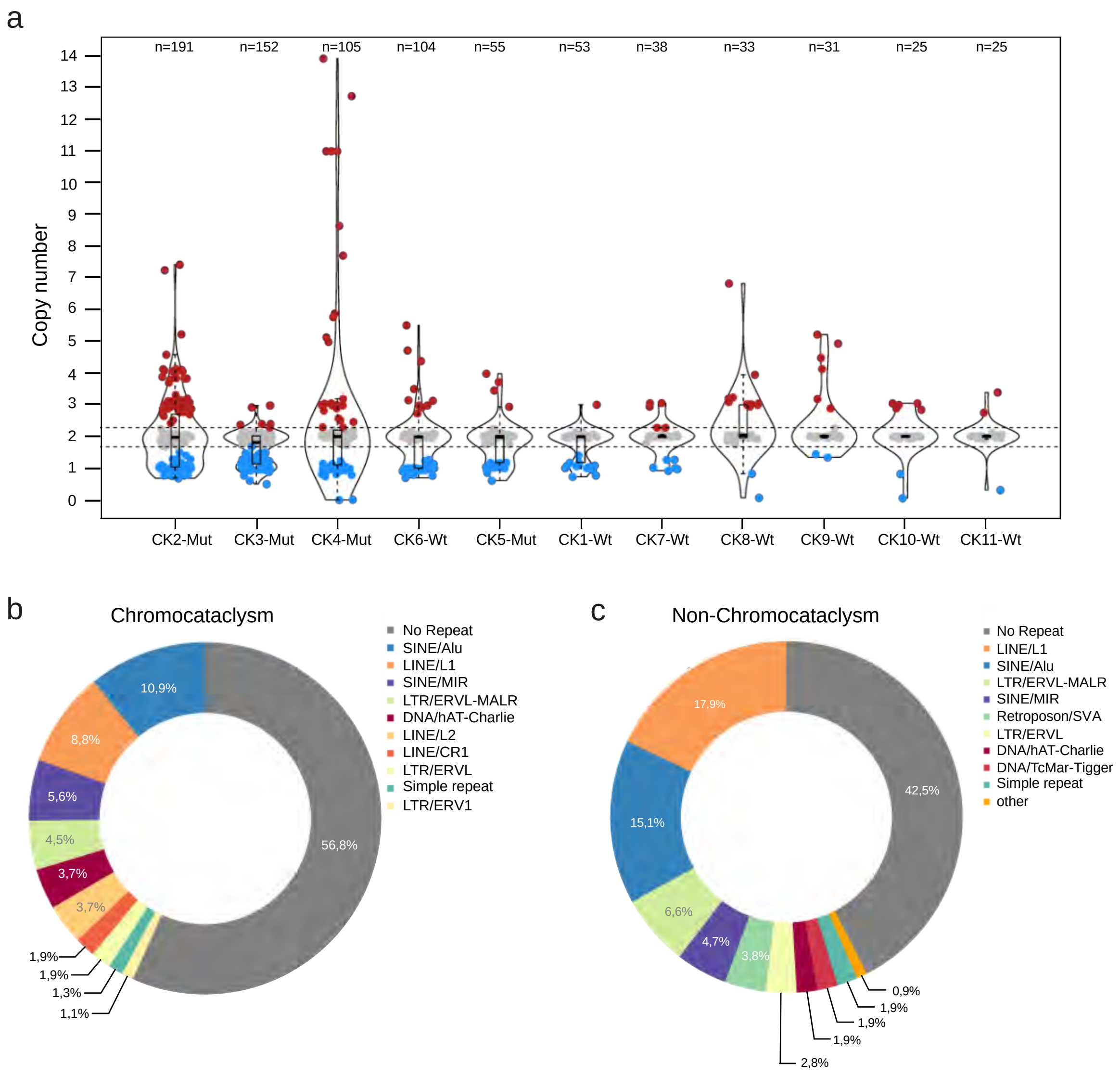
Supplemental Figure 5.: Overview of Structural Variants in the cohort presented in this work. Bar plot of the CN distribution for each case in our cohort. Cases on the x axis. Number of distinct fragments (all sizes) with a certain CN state on the y axis. CN gain (gray bars) is based on our CN detection approach defined as a CN > 2,3. CN stable (light blue) is defined as CN 1.7 ≤ x ≤ 2.3. CN. CN loss (dark blue) is defined as CN < 1,7 (**a**). Bar plot of the Translocation and Inversion distribution for each case in our cohort. Translocations (dark green) are here defined as interchromosomal breakends. Inversions (light green) as intrachromosomal breakends (**b**).



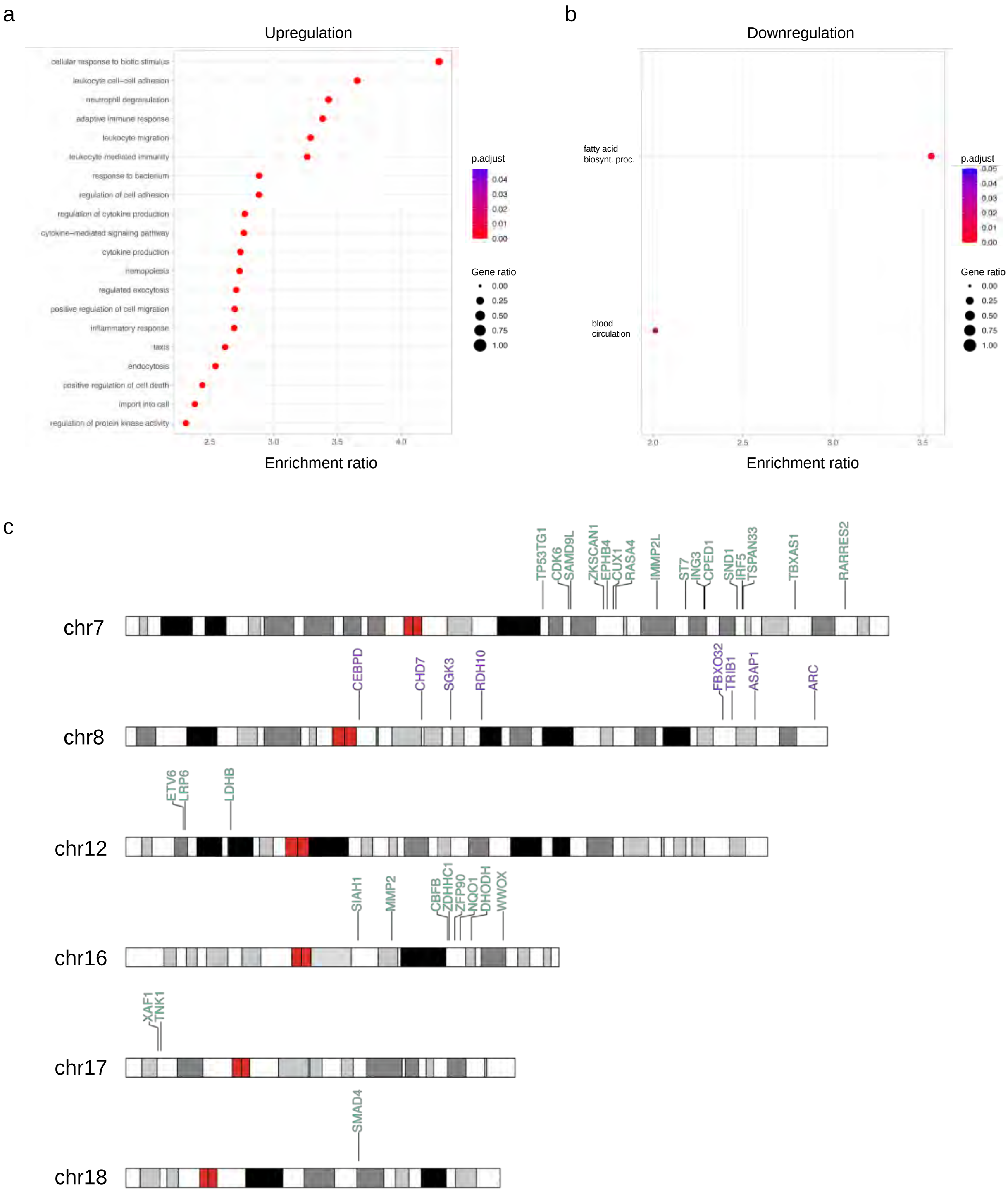
Supplemental Figure 6. Zoomed-in details of Hi-C trans-map of chromosome 15/16 and cis-map of chromosome16 (**a**). Chromosome positions are based on the GRCh37/hg19 assembly. The connections of a chromosome 16 small amplified region of 267 bp (position on chr 16: 4,026,350-4,026,647 bp) are visible as fragments in the Hi-C maps. Letters A-T mark Nanovar Translocation and Inversion calls of this region projected on the Hi-C map. Coverage estimations based on Hi-C are displayed around the Hi-C maps, the CN elevation of this region is very well visible in comparison with the coverage track of a sample in which this region is not affected by SVs (CK10-Wt). The segments around the amplified region shows a CN state of 1 based on ACE and Hi-C. The final CN of the segments connecting to the 297bp amplified region are shown above the coverage tracks and were generated based on our CNV algorithm. The 297 bp amplified region is marked on the CK-AML8 coverage track with a black arrowhead. * CK2-Mut coverage, ** CK10-Wt coverage. The 297 bp amplified region of CK2-Mut on chromosome 16 as it is represented in the .tdf files in IGV, consisting of multiple CN changes (CN state 3-5) and subfragments (**b**). The .tdf file track of CK2-Mut is shown below, here were no CN changes present in this region. The location of the Nanovar calls on the segment are marked by letters A-T. All chromosome positions are based on the GRCh37/hg19 assembly. KB, genomic position in kilobases; Numbers below (**b**), genomic position in bases.



Supplemental Figure 7. Number of breakpoints and their relative distance to the next point of CN change plotted for all samples that harbor any translocation or inversion breakpoint (**a**). Co-occurrence of CN changes with inversion or translocation breakpoints, shown for CK6-Wt (**b**), CK2-Mut (**c**) and CK7-Wt (**d**). Just small sections of the real complexity of these rearrangements can be visualized here. Translocation and/or inversion calls that are connecting the respective fragments here are marked by a black arrow. Connections of these fragments are shown as orange lines. The copy number of each fragment is written in orange numbers above the fragment. CN changes are marked by vertical dashed lines. KB, genomic position in kilobases; B, genomic position in bases.



Supplemental Figure 8. Violin plots of CN distribution in the final CNV dataset for all CK-AML cases (**a**). Each dot represents one fragment (distinct region on a chromosome without a CN change) and its respective CN. The cases are ordered by complexity, starting with the case with the highest number of CN changes to the left. Pie charts displaying the distribution of breakends of the chromocataclysm cases (**b**) and of chromothripsis cases that were not classified as chromocataclysm (**c**), regarding their occurrence inside of repetitive elements (repeat categories from repeatmasker) (**c**).

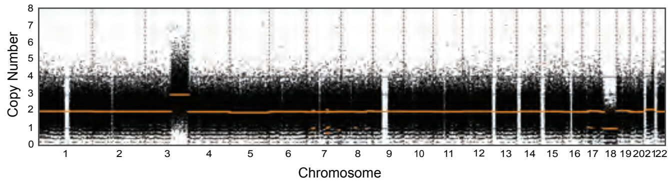


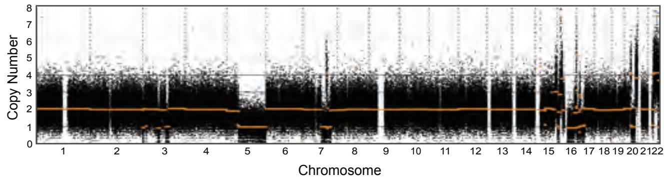
Supplemental Figure 9: Over-Representation Analysis (ORA) based on Gene Ontology (GO) terms. Overrepresentation of GO terms in the Illumina RNA-Seq data of the CK-AML cases vs. the CD34+ stem cell controls. Upregulated GO terms (**a**) downregulated GO terms (**b**). Genomic distribution of the candidate genes of our gene expression analysis (**c**), downregulation candidate genes in green letters, upregulation candidate genes in purple letters. p.adjust, False-Discovery Rate adjusted p values; Enrichment ratio, GO term overrepresentation compared to expected rate of occurrence; Gene ratio, fraction of GO term associated genes that are dysregulated. Chromosomal distribution of the candidate genes (**c**). CN ↑/GE ↑ candidate genes written in purple, CN ↓/GE ↓ candidate genes written in green.

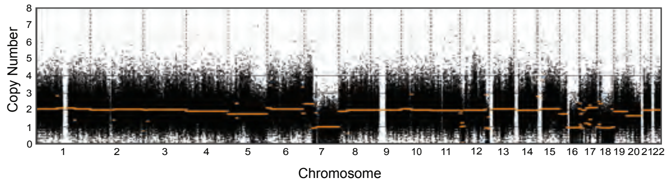
Supplemental Datasets

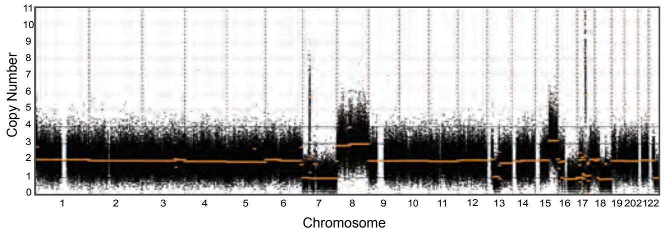
Supplemental Dataset 1: ACE tool ONT DNA long-read whole genome sequencing copy number data

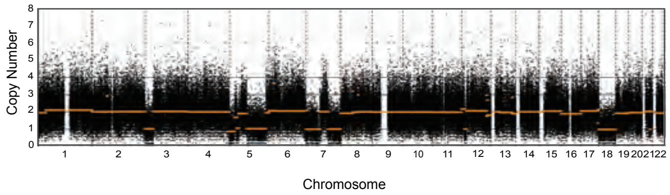
TP53 mut


CK1-Mut

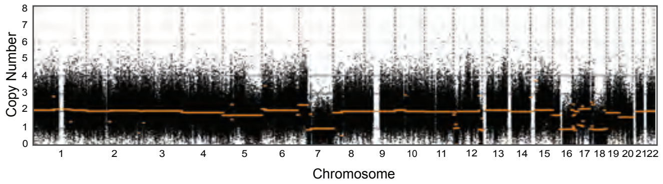

CK2-Mut

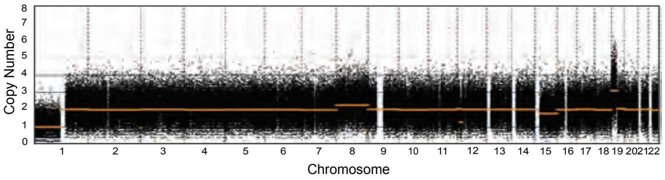

CK3-Mut

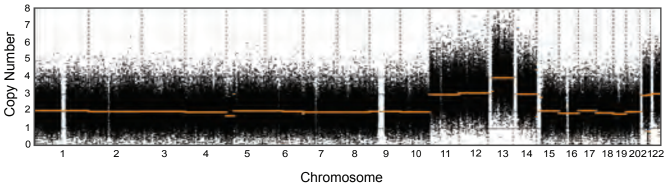

CK4-Mut

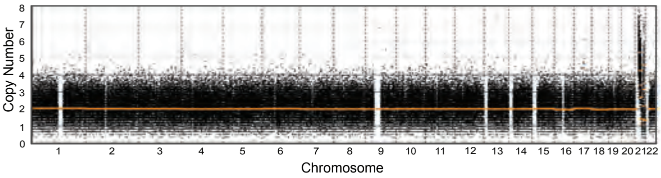

CK5-Wt

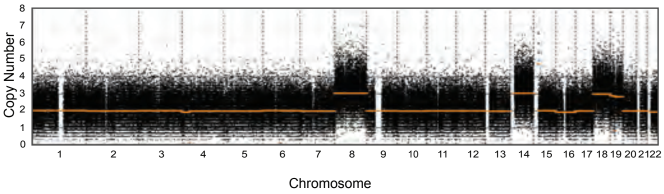
TP53 wt

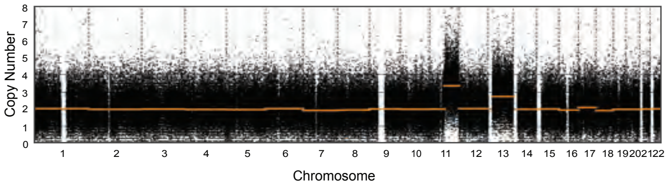

CK6-Wt


CK7-Wt


CK8-Wt

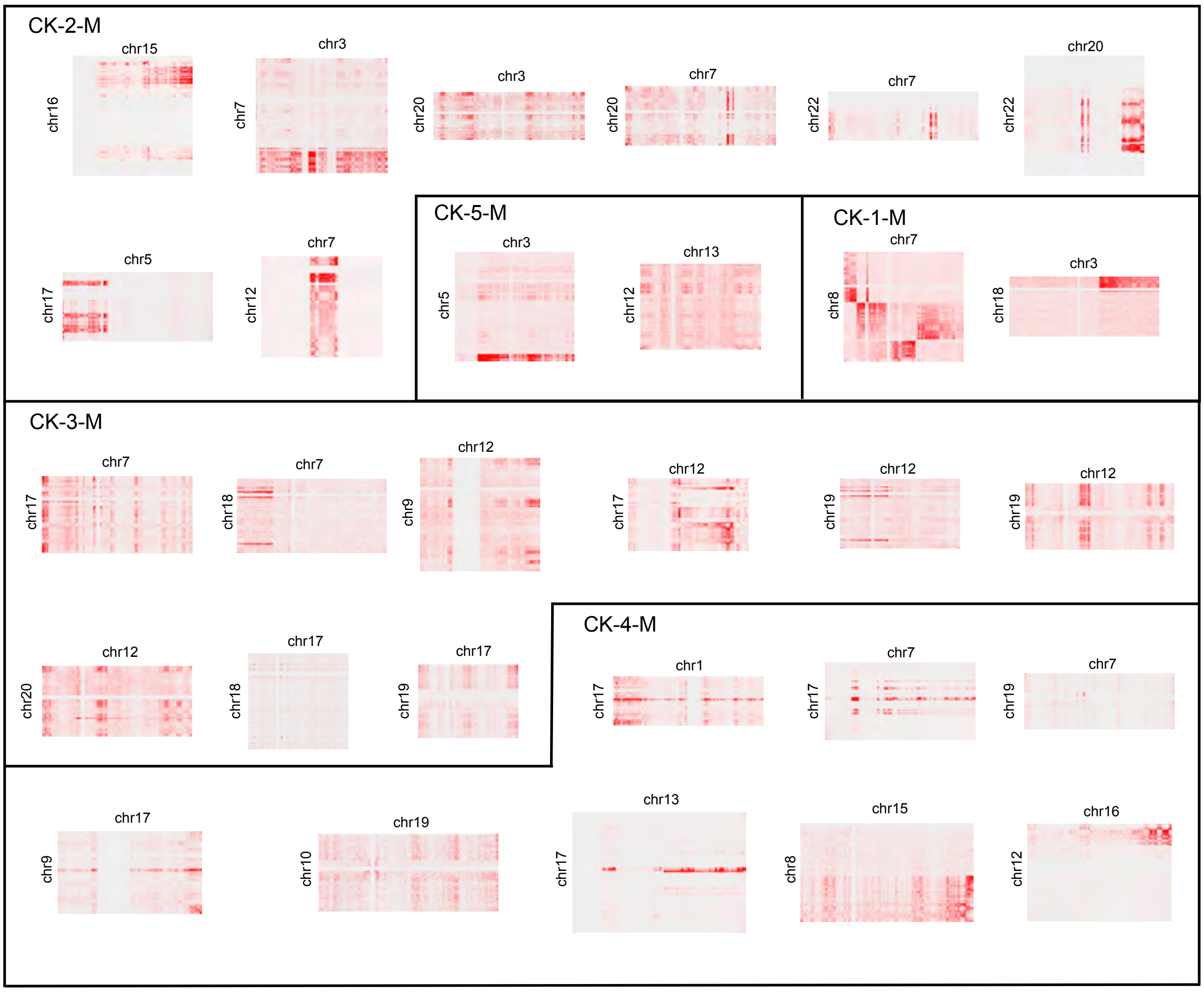

CK9-Wt


CK10-Wt

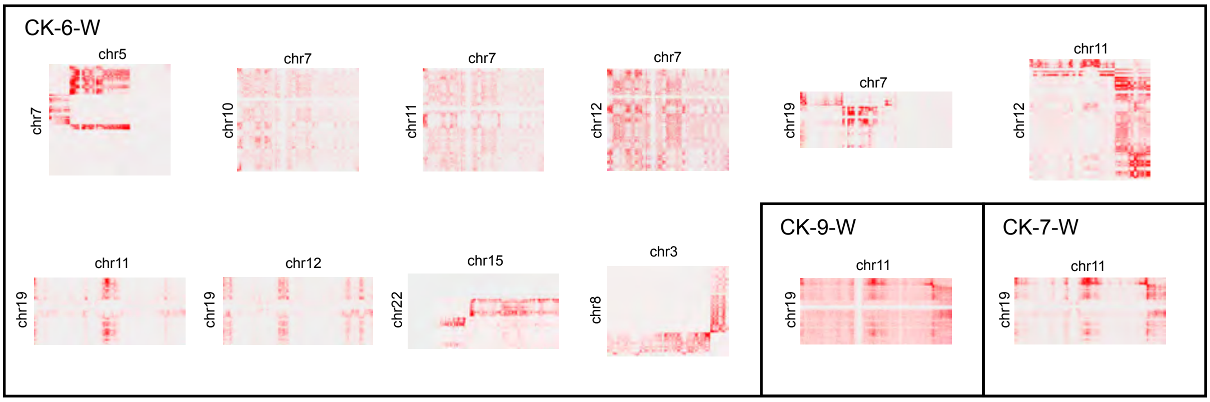

CK11-Wt

Supplemental Dataset 2: Hi-C maps for all rearrangements in this cohort

TP53 mut



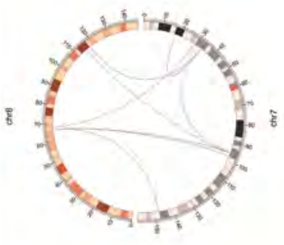
TP53 wt



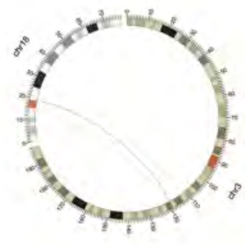
Supplemental Dataset 3: Circos plots for all Translocation and Inversion breakpoints in this cohort

TP53 mut

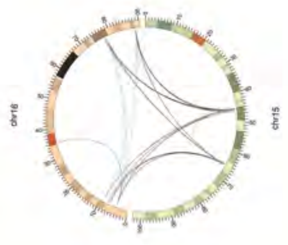
CK1-Mut
chr 7/8



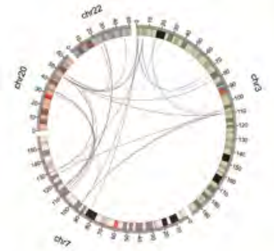
CK1-Mut
chr 3/18



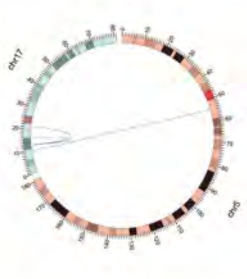
CK2-Mut
chr 15/16



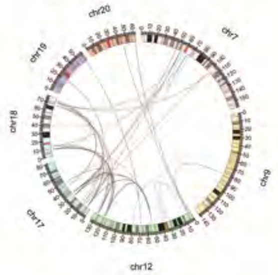
CK2-Mut
chr 3/7/20/22



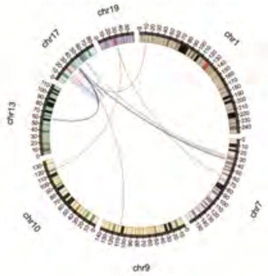
CK2-Mut
chr 5/17



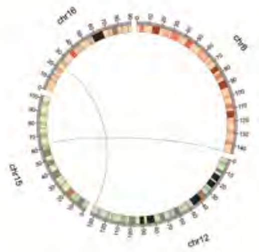
CK3-Mut
chr 7/9/12/17/18/19/20



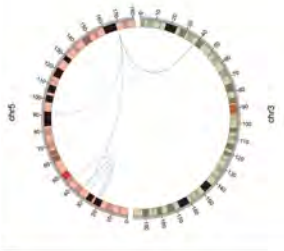
CK4-Mut
chr 1/7/9/10/13/17/19



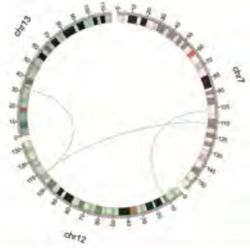
CK4-Mut
chr 8/12/15/16



CK5-Mut
chr 3/5

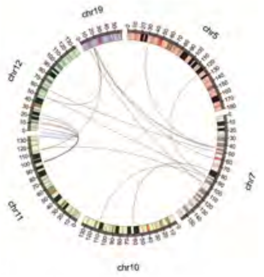


CK5-Mut
chr 7/12/13

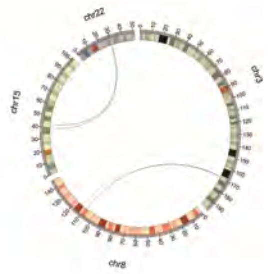


TP53 wt

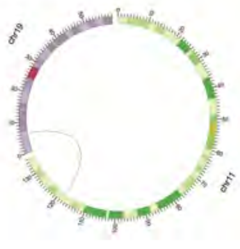
CK6-Wt
chr 5/7/10/11/12/19



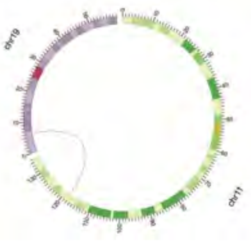
CK6-Wt
chr 3/8 and 15/22



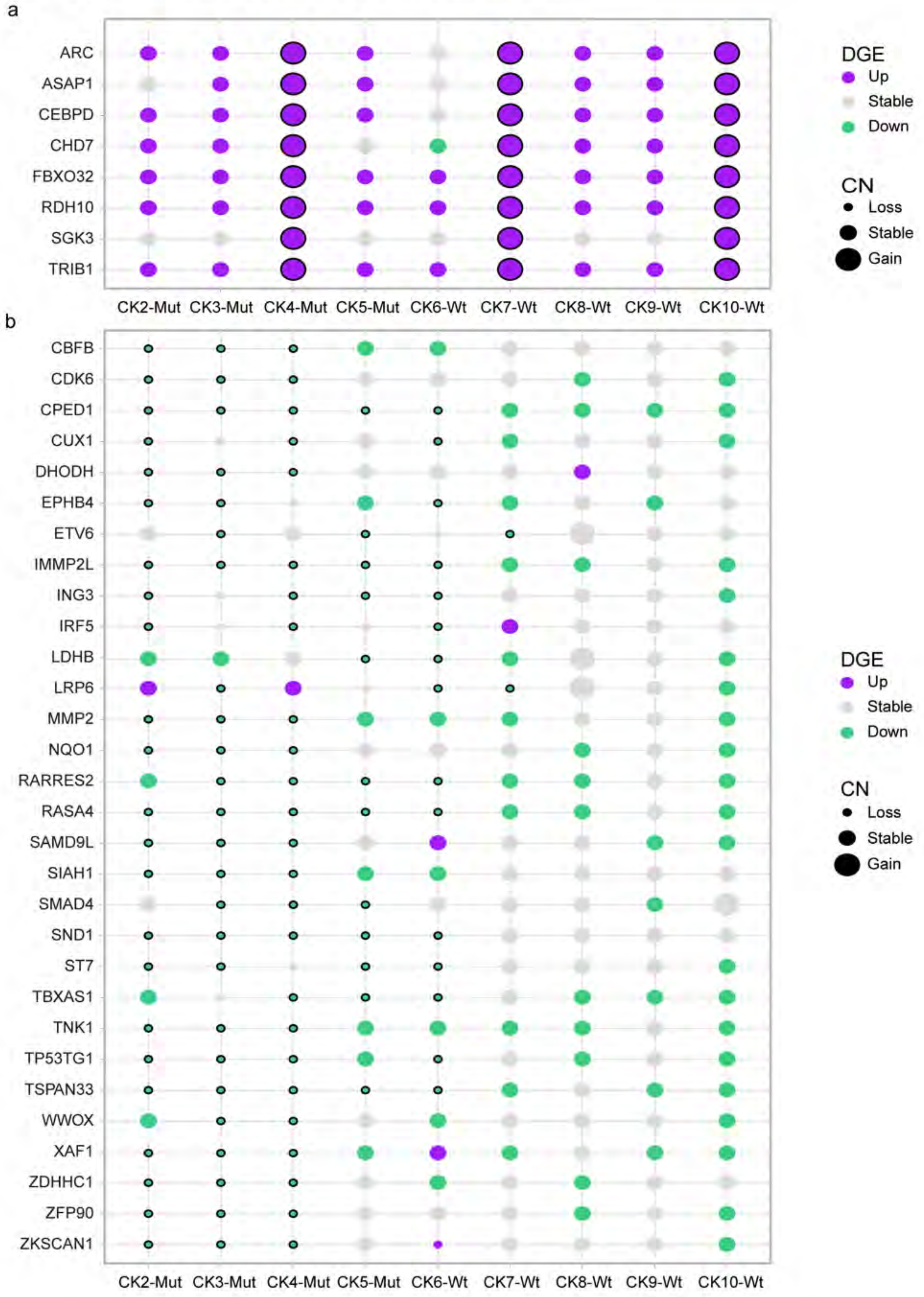
CK7-Wt
chr 11/19



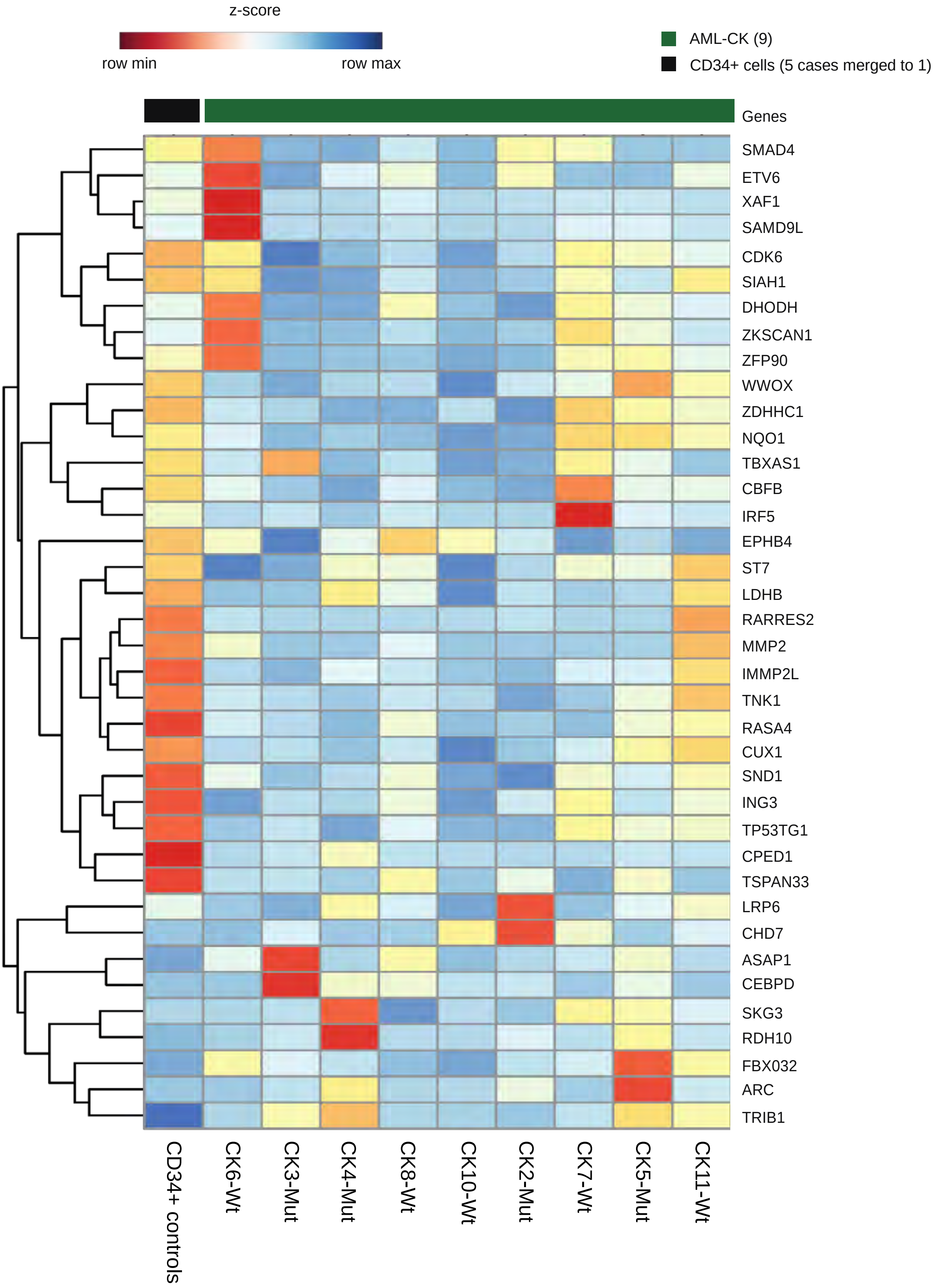
CK9-Wt
chr 11/19



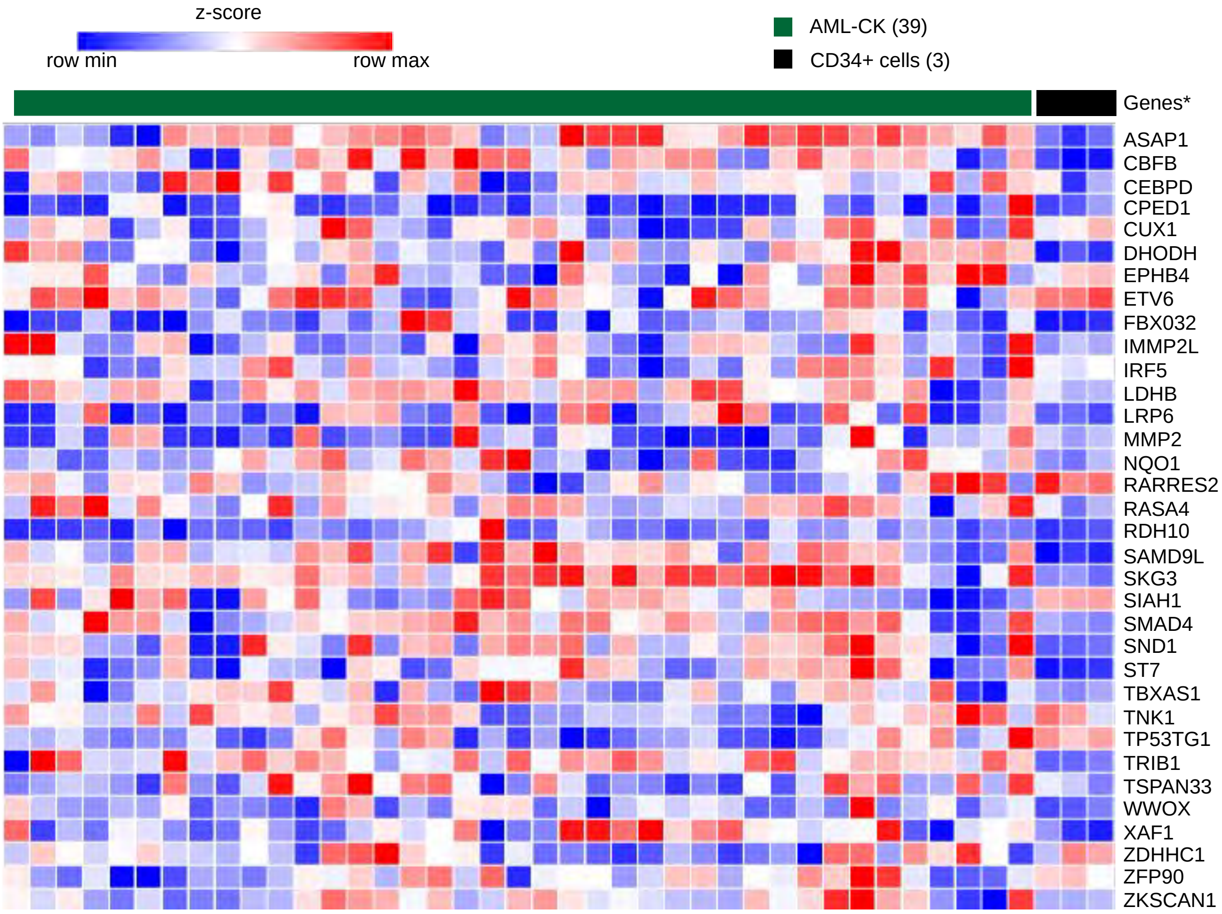
Supplemental Dataset 4: CK-AML copy number and gene expression dysregulation candidate genes



Supplemental Dataset 5: RNA sequencing data heatmap of the CK-AML and CD34+ hematopoetic stem cell control cohort reported in this monograph



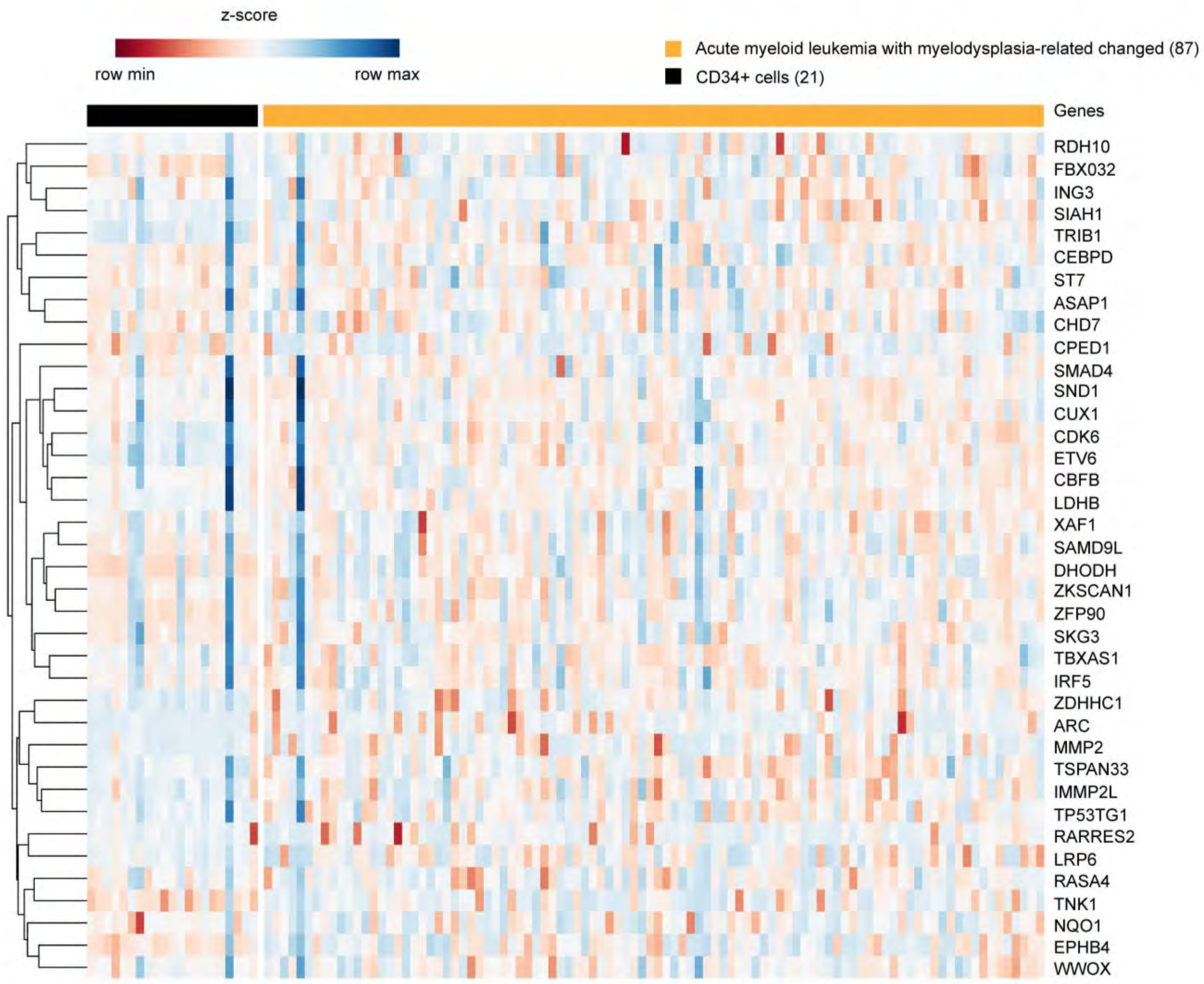
Supplemental Dataset 6: Microarray expression data heatmap of AML-CK cases and CD34+ hematopoetic stem cell samples of healthy controls



*CHD7, CKD6, ING3 and ARC not annotated

data derived from Risueño et al., 2014

Supplemental Dataset 7: RNA expression data heatmap of AML cases with myelodysplasia related changes and CD34+ hematopoetic stem cell samples of healthy controls



data derived from the Beat AML dataset, Tyner et al., 2018

Supplemental Reference

42. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011;17,10.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
44. Yu G. Gene Ontology Semantic Similarity Analysis Using GOSemSim. Methods Mol Biol. 2020;2117:207-215.
45. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274-81.
46. Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Amer. Math. Soc. 1956;7,48–48.
47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
48. Sanges R, Cordero F, Calogero RA. oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. Bioinformatics. 2007;23(24):3406-8.
49. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461-468.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
51. Poell JB, Mendeville M, Sie D, Brink A, Brakenhoff RH, Ylstra B. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. Bioinformatics. 2019;35(16):2847-2849.
52. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
53. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): highperformance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178-92.
54. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016;3(1):95-8.
55. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 2016;3(1):99-101.
56. Rücker FG, Schlenk RF, Bullinger L, Kayser S, Teleanu V, Kett H, et al. TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. Blood. 2012;119(9):2114-21.

57. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018;562(7728):526-531.
58. Ye Y, Wang G, Wang G, Zhuang J, He S, Song Y, et al. The Oncogenic Role of Tribbles 1 in Hepatocellular Carcinoma Is Mediated by a Feedback Loop Involving microRNA-23a and p53. Front Physiol. 2017;8:789.
59. May WS, Hoare K, Hoare S, Reinhard MK, Lee YJ, Oh SP. Tnk1/Kos1: a novel tumor suppressor. Trans Am Clin Climatol Assoc. 2010;121:281-92.
60. Jang YE, Jang I, Kim S, Cho S, Kim D, Kim K, et al. ChimerDB 4.0: an updated and expanded database of fusion genes. Nucleic Acids Res. 2020;48(D1):D817-D824.
61. Cartel M, Mouchel PL, Gotanègre M, David L, Bertoli S, Mansat-De Mas V, et al. Inhibition of ubiquitin-specific protease 7 sensitizes acute myeloid leukemia to chemotherapy. Leukemia. 2021;35(2):417-432.
62. Naylor RM, Jeganathan KB, Cao X, van Deursen JM. Nuclear pore protein NUP88 activates anaphase-promoting complex to promote aneuploidy. J Clin Invest. 2016;126(2):543-59.
63. Yi S, Chen Y, Wen L, Yang L, Cui G. Downregulation of nucleoporin 88 and 214 induced by oridonin may protect OCIM2 acute erythroleukemia cells from apoptosis through regulation of nucleocytoplasmic transport of NF-κB. Int J Mol Med. 2012;30(4):877-83.
64. Neilsen PM, Cheney KM, Li CW, Chen JD, Cawrse JE, Schulz RB, et al. Identification of ANKRD11 as a p53 coactivator. J Cell Sci. 2008;121(Pt 21):3541-52.
65. Lim SP, Wong NC, Suetani RJ, Ho K, Ng JL, Neilsen PM, et al. Specific-site methylation of tumour suppressor ANKRD11 in breast cancer. Eur J Cancer. 2012;48(17):3300-9.
66. Xu J, Jiao J, Xu W, Ji L, Jiang D, Xie S, et al. Mutant p53 promotes cell spreading and migration via ARHGAP44. Sci China Life Sci. 2017;60(9):1019-1029.