REGULAR ARTICLE

blood advances

AML with complex karyotype: extreme genomic complexity revealed by combined long-read sequencing and Hi-C technology

Marius-Konstantin Klever,[1-3] 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,*] and Lars Bullinger[1,7,12,*]

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

Key Points

- Combination of long-read sequencing and Hi-C SV calling allows the characterization of genomic rearrangements at an unprecedented resolution.
- Integration of genomic SV calling data with transcriptomic data identifies novel oncogenic fusions and dysregulated candidate driver genes.

Acute myeloid leukemia with complex karyotype (CK-AML) is associated with poor prognosis, which is only in part explained by underlying *TP53* mutations. Especially in the presence of complex chromosomal rearrangements, such as chromothripsis, the outcome of CK-AML is dismal. However, this degree of complexity of genomic rearrangements contributes to the leukemogenic phenotype and treatment resistance of CK-AML remains largely unknown. Applying an integrative workflow for the detection of structural variants (SVs) based on Oxford Nanopore (ONT) genomic DNA long-read sequencing (gDNA-LRS) and high-throughput chromosome confirmation capture (Hi-C) in a well-defined cohort of CK-AML identified regions with an extreme density of SVs. These rearrangements consisted to a large degree of focal amplifications enriched in the proximity of mammalian-wide interspersed repeat elements, which often result in oncogenic fusion transcripts, such as *USP7::MVD*, or the deregulation of oncogenic driver genes as confirmed by RNA-seq and ONT direct complementary DNA sequencing. We termed this novel phenomenon chromocataclysm. Thus, our integrative SV detection workflow combing gDNA-LRS and Hi-C enables to unravel complex genomic rearrangements at a very high resolution in regions hard to analyze by conventional sequencing technology, thereby providing an important tool to identify novel important drivers underlying cancer with complex karyotypic changes.

Introduction

Acute myeloid leukemia (AML) is the most common acute leukemia in adults with an incidence of ~4 new cases annually per 100 000 inhabitants in the United States. Despite recent therapeutic advances, AML

Submitted 5 June 2023; accepted 30 July 2023; prepublished online on *Blood Advances* First Edition 15 August 2023. https://doi.org/10.1182/bloodadvances.2023010887.

*U.S.M., S.M., and L.B. contributed equally to this work.

Data are available on request from author, Marius-Konstantin Klever (marius.kevler@charite.de).

The full-text version of this article contains a data supplement.

still shows the shortest survival time of all leukemias.[1] The complex karyotype subtype of AML (CK-AML) is associated with an even poorer response to conventional therapy and has the worst outcome of all cases. A complex karyotype is apparent in about 10% to 14% of all patients with AML, and it is defined as the presence of ≥3 structural variants (SVs) in the absence of specific recurring translocations and/or inversions; that is, t(8;21), inv(16)/t(16;16), t(9;11), t(v;11)(v;q23.3), t(6;9), inv(3)/t(3;3), or AML with *BCR-ABL1*.[2-4] Inactivating *TP53* mutations, often associated with chromosomal instability, are present in the malignant cells of ~60% to 70% of all CK-AML cases.[5] Interestingly, frequently observed mutations in other subtypes of AML (eg, mutations in *FLT3, NPM1*, *KRAS*, *NRAS,* and *KIT*) are less common in CK-AML.[6,7] Consequently, it is likely that structural variants play an important role in the pathogenesis and contribute to the very poor prognosis of CK-AML.

The detection and functional interpretation of SVs are a challenging task in cancer research and poorly studied compared with the effects of single-nucleotide variants8[8]; even though SVs are thought to play a major role in cancer biology.[9] Several studies of CK-AML using low resolution genomic tools identified complex rearrangements associated with this disease. For instance, large copy number variation (CNVs) can be reliably detected with array-CGH but without positional information and precise breakpoint position. Moreover, smaller CNVs and balanced rearrangements cannot be detected by array-CGH (comparative genomic hybridization) or other low resolution genomic tools such as single nucleotide polymorphism (SNP) arrays and optical mapping.[10] Short-read genomic sequencing (GS) revolutionized the genomic field by identifying a plethora of variations in the human genome. In fact, several studies using short-read sequencing in cancer identified extremely complex genomic rearrangements like chromothripsis. Chromothripsis is a thought-to-be single catastrophic event, that is, currently thought to be rather rare in AML; however, it was recently shown to occur in a relatively high frequency (>50%) in several other cancers.[11,12] In AML, it is associated with an even poorer prognosis (average overall survival of only 2-4 months) than CK-AML with less complex rearrangements.[5,13] Although astonishing recent data based on short-read sequencing contribute to a great extent to our knowledge about structural variants in cancer,[12] different approaches integrating long- and short-read sequencing data bear great potential to resolve complex SVs at a very high resolution and confidence. The use of short-read whole GS is hampered by a low overlap of detected breakpoints with other technologies, particularly when it comes to the detection and reconstruction of complex events.[14,15] The results from different SV calling algorithms in short-read sequencing vary substantially and impair a comprehensive SV interpretation in health and disease.[16] Because of their long-read length, genomic DNA long-read sequencing (gDNA-LRS) was shown to be able to overcome some of the limitations of short-read sequencing, mostly regarding complex rearrangements and breakpoints in regions with repetitive elements.[17,18]

In this study, we provide a new workflow for precise SV characterization in tumors with very complex genomic rearrangements, including cases with chromothripsis. For this workflow, we integrated Oxford Nanopore (ONT) gDNA-LRS with short-read Hi-C (genomic analysis technique) in a cohort of well-defined CK-AML samples. The combination of both GS tools allowed us to unravel the complexity of the genomic rearrangements in CK-AML at an unprecedented resolution. We observed how catastrophic genomic events lead to extreme local breakpoint clustering, accompanied by focal amplifications. Combining genomic analyses with conventional RNA sequencing and ONT direct complementary DNA (cDNA) sequencing supported the potential pathogenic impact of rearranged driver genes by showing the effect of high confidence SVs on gene expression and the formation of fusion transcripts in CK-AML.

Methods

Participants and ethics approval

The study was performed with the approval of the Charité Ethics Committee, Berlin, Germany. Fresh peripheral blood or bone marrow biopsies of 11 patients with CK-AML and 5 healthy individuals were collected by the Hematology and Oncology departments of the Charité Universitätsmedizin Berlin and the University of Ulm, Germany. All samples were collected with informed and written consent from the patients and healthy individuals within the German and Austrian AML Study Group (AMLSG) BiO Registry study (NTC 01252485).

Karyotyping and mutational screening

For all patients with CK-AML, conventional karyotyping and genotyping for common AML mutations (*FLT3*, *CEBPA*, *KMTA2*, *NPM1*, *IDH1,* and *TP53*) were performed. Karyotype complexity was determined according to the European LeukemiaNet recommendations.[2] Basic clinical data as well as karyotyping and mutational screening findings are shown in supplemental Table 1. *TP53* pathogenic mutations were present in 5 of the 11 patients.

Samples collection and processing

Bone marrow or peripheral blood samples of all enrolled cases were collected and subsequently frozen in liquid nitrogen at an average density of 1×10^7 cells/mL, after Ficoll centrifugation to enrich for leukemic blasts (>90% of total cells). $CD34^+$ hematopoietic stem enriched cell fractions were obtained via peripheral blood apheresis of healthy hematopoietic stem cell donors after granulocyte colony-stimulating factor stimulation. CD34+ purity >95% of total cell count was confirmed by flow cytometry after purification. All cells used in this study were thawed in RPMI 1640 medium (Thermo Fisher Scientific), supplemented with 20% heat inactivated fetal bovine serum (FBS), DNAse I (Sigma-Aldrich), Heparin (Merck), and $MgCl_2$, and incubated for 1 hour at 37°C. Cells were then processed for Hi-C library preparation and DNA/RNA isolation.

Hi-C library preparation and data analysis

Hi-C libraries were prepared and data analysis was performed as described elsewhere.[19] To adjust the protocol to the input material of blood cells, fixation was performed at a final concentration of 1% formaldehyde in RPMI 1640 medium. A total of 5×10^5 to 1×10^6 cells per replicate were used as an input for our Hi-C sequencing pipeline and 2 to 4 Hi-C library replicates of each case were sequenced to an approximated mean sequencing depth of 320 million fragments. In addition, 2 automatized breakend (BND) detection tools, HiNT[15,20] and hic_breakfinder,[14] for Hi-C maps were run on all CK-AML cases, using the standard settings.

RNA and DNA extraction

RNA and DNA extraction were performed with the AllPrep DNA/RNA/Protein Mini Kit (Qiagen). RNA was quality checked on an

Agilent Technologies Tape Station (RNA ScreenTape) and used for downstream processing if an RNA integrity number value of ≥8.0 was reached.

ONT gDNA long-read sequencing library preparation, sequencing and analysis

Using the ligation sequencing kit, gDNA was prepared for ONT gDNA-LRS. DNA gDNA-LRS libraries were sequenced on a GridION on R9.4.1 flowcells. Sequencing of the gDNA libraries was performed until coverage of at least 10× for each patient was reached. For detection of SVs, the gDNA bamfiles were processed with the long-read SV caller NanoVar.[21]

ONT direct cDNA sequencing library preparation and analysis

ONT direct cDNA sequencing using messenger RNA (mRNA) was processed with the Dynabeads mRNA Purification Kit (Thermo Fisher Scientific) for total RNA isolation. The mRNA was reverse transcribed and prepared for ONT sequencing using the direct cDNA Sequencing Kit. ONT cDNA files were processed with the ONT version of the fusion caller JAFFA[22] with standard settings and alignment to human reference genome version 19 (hg19).

Illumina RNA sequencing and data analysis

RNA sequencing was performed on an Illumina NovaSeq 6000 in triplicates for all CK-AML samples but CK1-Mut and CK11-Wt. To benchmark gene expression in our cohort, we additionally performed RNA sequencing for 5 $CD34^+$ samples of healthy individuals in single replicates. Stranded mRNA was isolated by Poly-A selection and 100-bp paired-end sequencing was performed with 100 million reads per replicate. Trimmed reads were aligned to the hg19 using STAR,[23] and transcripts assembly was performed using stringtie with the GENCODE annotation (release 19).[24] Furthermore, Illumina RNA sequencing data were processed with the fusion caller JAFFA[22] using standard settings and alignment to hg19. Dysregulation of genes was assessed using DESeq2,[25] using protein-coding genes, long-noncoding RNAs, and pseudogenes extracted from the GENCODE annotation (release 19). Genes with an absolute log2 fold change of at least 1 and an adjusted P value of <.05 were determined differentially expressed. Overrepresentation analysis (ORA) of differentially expressed genes in gene ontology terms was carried out using the WebGestalt R package.[26]

RNA expression data sets

RNA sequencing data of the Beat AML data set were downloaded from GDC (genomic data commons) for 87 cases of AML, with myelodysplasia related changes, and $CD34^+$ cell samples of 21 healthy controls. Fragments per kilobase of transcript per million mapped reads (FPKM) values of this data set and our CK-AML RNA sequencing data set were used for the generation of z score expression data heatmaps using the pheatmap package in R. Microarray expression data for 30 CK-AML cases were previously generated by our laboratory.[27] This cohort also included data from $CD34^+$ cell samples of 3 healthy controls, which were not published to date.

Identification of BND signatures and genomic distribution

Categories and genomic location of repetitive elements were downloaded from Repbase.[28] The distribution of translocation and inversion BNDs in relation to genomic features was assessed as follows: Genomic features (exons, introns, UTR3' and UTR5') were downloaded from GENCODE (release 19). Promoter regions were defined as regions located 1.5 kb upstream and 0.5 kb downstream of the transcription start sites.

Functional evaluation of fusion genes

Functional evaluation of the USP7::MVD fusion transcripts was performed by cloning the cDNA in a pRSF91 retroviral vector. This construct was transfected in NIH3T3 cells that were seeded in high density with 1 to 2 µg/mL Polybrene (Sigma). Puromycin selection was started 24 hours after transduction at concentrations between 0.5 and 2 µg/mL. To monitor cellular proliferation, the fusion-gene transduced cells were seeded at 2×10^5 density per well and the number of viable cells was counted using trypan blue staining from days 1 to 5. The quantification was performed in triplicates.

Results

Cohort overview and structural variant detection using Hi-C and gDNA LRS

We applied a combination of innovative genomic and transcriptomic sequencing methods to comprehensively characterize a cohort of 11 patients with CK-AML plus 5 $CD34^+$ hematopoietic stem cell donor controls (Figure 1A; supplemental Table 1). Genomic technologies (Hi-C and gDNA-LRS) were used to develop a reliable SV detection workflow for complex rearrangements like chromothripsis, and by leveraging transcriptomic data (Illumina RNA and ONT direct cDNA sequencing), we parallelly identified fusion transcript and differential gene expression patterns to highlight potential functional consequences of the complex SVs detected in CK-AML (Figure 1A).

Our Hi-C SV detection workflow started with visual inspection of all Hi-C maps for SV BNDs of translocations and inversions (supplemental Note) (supplemental Figure 1), because automatized BND detection using HiNT[15] and hi_c breakfinder[14] showed a high rate of false-positive SVs (supplemental Figure 2). Next, all putative breakpoints identified by visual inspection in Hi-C maps were further examined for validation with gDNA-LRS calls. SVs occurring between different chromosomes (eg, translocations) or exchanging material from distant parts of a single chromosome (eg, insertions) create interaction patterns in Hi-C maps, making Hi-C a suitable tool for cross-validation.

To validate the Hi-C SV BNDs, we mapped onto the Hi-C maps all the BNDs detected by NanoVar (ONT gDNA-LRS caller) (supplemental Figure 1). Finally, we integrated the BND data set with copy number (CN) data from ACE, Absolute Copy number Estimation (gDNA-LRS) (supplemental Dataset 1) that we validated with the HiNT tools (Hi-C). The CN output of ACE showed a high correlation with those retrieved from the HiNT tool (supplemental Figure 3A-B), and fragments that were <20 kb were additionally analyzed by visual inspection of CN changes (supplemental Table 2). In summary, our filtering strategy yielded a high confident true-positive SV set supported by both Hi-C and gDNA-LRS and allowed us to understand the real landscape of the genomic complexity in CK-AML.

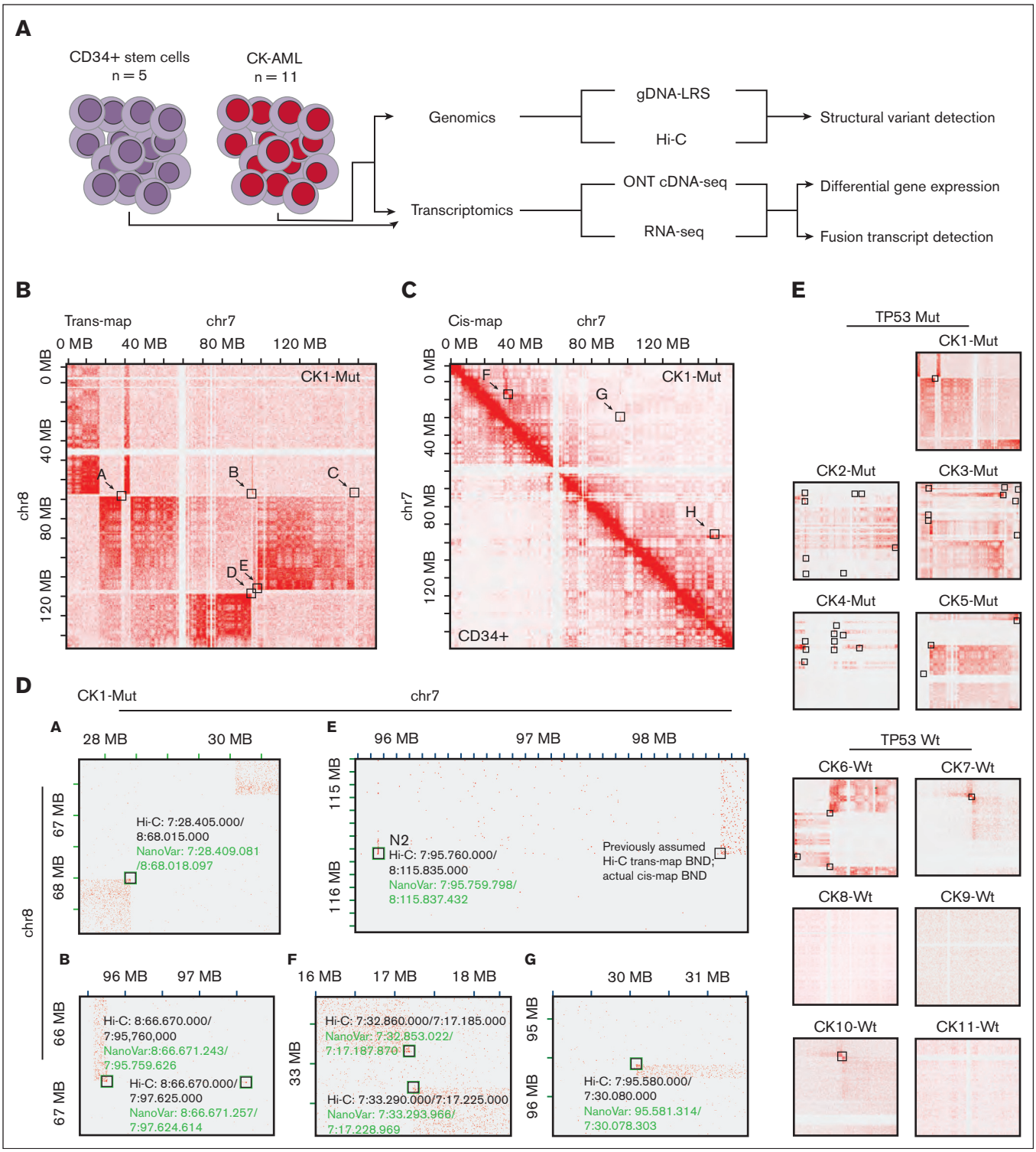

Figure 1. Cohort overview and complex genomic rearrangements detected by our SV detection workflow. (A) Samples from patients with CK-AML (n = 11) were subjected to genomic sequencing (Hi-C and ONT-GS) for SV detection. In addition, healthy CD34^{+} stem cell donors (n = 5) and the CK-AML samples were RNA-sequenced (Illumina RNA-Seq and ONT cDNA Seq) to study the functional consequences of the SVs. (B-C) Hi-C maps of patient CK1-Mut, chromosome 7/8 trans-map (B) and chromosome 7 cis-map (C). Hi-C breakend regions were inferred based on signal intensity at the breakends and are marked by black squares (named "a" to "h" for simplicity). Region d was shown to only harbor breakpoint-like patterns in the integrated analysis with NanoVar data. (D) Zoomed in detail of Hi-C maps showing breakends detected by both Hi-C and NanoVar (green squares with black squares inside). In these cases, the NanoVar SV calls were found to map in the same 10 kb range in which the BND were located estimated based on Hi-C. In region "e," we observed an indirect BND-like structure in the upper right corner (black square) without a corresponding NanoVar SV call. Interestingly, a NanoVar SV pointed out to a small fragment (<5 kb, named N2), also visible in Hi-C but missed in the primary visual inspection. This fragment represents the actual trans-map

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

In order to exemplify our SV analysis workflow and results, we selected 1 CK-AML case (CK1-Mut). The Hi-C map of this sample showed a complex rearrangement involving chromosome 7 (chr7) and chr8 (Figure 1B-C). Visual inspection of the Hi-C maps enabled us to identify 10 putative translocation and inversion BNDs (black squares), which were not observed in the healthy control sample ($CD34^+$; Figure 1C). Next, NanoVar SV calls matching to the Hi-C SV calls were projected onto the Hi-C maps for BND cross-validation (Figure 1D). Using this approach, 8 of 10 putative Hi-C breakpoints could be directly verified by NanoVar. The 2 remaining BNDs were shown to actually represent BND-like patterns but not factual BNDs, even if their visual appearance in Hi-C was very similar to the factual BNDs in Hi-C (Figure 1D); these BNDs lacked matching NanoVar SV calls and were in a detailed analysis linked to other fragments of the rearrangement (Supplemental Figure 4). With our combined approach, the entire rearrangement could be fully resolved (refer to supplemental Results; Supplemental Figure 4).

In the next step, we applied our integrative SV analyses to all CK-AML cases and observed a plethora of complex genomic rearrangements (Figure 1E). Interestingly, we observed that the rearrangements were much more pronounced in *TP53* mutated CK-AML cases (hereafter named CK1-Mut to CK5-Mut) than the cases that were *TP53* wildtype (CK6-Wt to CK11-Wt) (Figure 1E; supplemental Table 1). All *TP53*-mut cases showed chromothripsis, whereas all of the *TP53*-Wt cases, except 1(CK6-Wt), showed only simple rearrangements without chromothripsis (ie, larger CNVs or simple translocations). The difference in complexity of the *TP53* mutated and *TP53* wildtype cases could be observed by Hi-C alone (Figure 1E; supplemental Dataset 2). However, our high confidence SV workflow enabled us to examine the microstructure of the breakpoint regions of all cases in order to identify novel rearrangement patterns (supplemental Dataset 3).

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

We subsequently applied our SV detection workflow to all cases in this cohort (Figure 3A, Supplemental Figure 5). With this workflow, we could identify a novel phenomenon of extreme high clustering of SV BNDs in 3 (CK2-Mut, CK3-Mut, and CK-6Wt) of 6 chromothripsis cases (Supplemental Dataset 3). These showed a pattern of multiple aberrantly connected fragments with a size ranging from only a few hundred basepairs to a few kilobasepairs (Figure 2A; supplemental Figure 6A). Of all 122 fragments with a size of <20 kb that we found in our data set, 107 (87.7%) were present in the 3 cases showing chromocataclysm, further indicating the extreme local complexity of these cases in comparison with the other CK-AML cases. (Figure 3A-B). Most SV BNDs were associated with a CN change in close proximity or directly at the BND (CN change within <5 kb in 66% of all BNDs) (refer to supplemental Results; supplemental Figure 7A). The highest level of clustering was detected in 1 CK-AML (CK2-Mut), in which the genomic complexity reached up to 31 BNDs (inversions and translocations) distributed over a region of just 2.7 kb of size. Notably, we were able to detect a 297 bp fragment from chr16 as part of multiple subfragment connections, accompanied by CN changes inside the fragment (supplemental Figure 6B). The CN gain of this fragment against the surrounding regions was clearly visible in Hi-C as well as in the gDNA-LRS data (supplemental Figure 6A-B). Notably, this fragment was found to be connected with many other small fragments, which were often <20 kb in size and showed an elevated CN state, however, connections to larger chromosomal regions were also seen, leading to an extreme complex picture of genomic rearrangements in regions of only a few kb (Figure 2B). We suggest the name "chromocataclysm" for this phenomenon of extreme local breakpoint clustering that is accompanied by focal amplifications.

Chromothripsis with extreme local BND clustering showed breakpoint enrichment at promoters and MIR elements

Because chromocataclysm events have not been studied so far, we sought to investigate genomic aspects of these clustered BNDs. The relative occurrence of SV BNDs in relation to certain genomic features was investigated against 10 000 random breakpoints by means of a Mann-Whitney *U* test, with Benjamin-Hochberg correction (refer to supplemental Material). We found an overrepresentation of breakpoints inside gene promoters ($P < .001$), introns ($P < .001$), and an underrepresentation in intergenic regions ($P < .001$) (Figure 3C). MIR (mammalian-wide interspersed repeat) elements are short interspersed nuclear elements (SINEs) and are divided in 4 subcategories, namely, MIR, MIRb, MIRc and MIR3.[29] In our analysis, we observed a higher occurrence of breakpoints inside of the MIR subcategory ($P = .002$), with an additional enrichment in the vicinity of the MIR subcategory (average distance, 15 kbp; 95% confidence interval, 10.5-21.6; $P = .004$). This enrichment was most prominent in the chromocataclysm rearrangements (Figure 3D-E).

Chromothripsis and chromocataclysm result in novel fusion transcripts in CK-AML

Next, we wanted to better understand the potential impact of these complex genomic rearrangements in CK-AML. Thus, we performed Illumina RNA sequencing and ONT direct cDNA sequencing to detect candidate fusion genes as well as deregulated expression of genes that might emerge from the complex SVs. In 9 CK-AML cases, fusion transcripts were detected based on Illumina RNA sequencing (n = 6491) and the ONT direct cDNA sequencing (n = 271), which showed 115 identical fusion transcript calls (Figure 4A; supplemental Table 4**)** to be considered as fusion transcripts candidates if 2 corresponding genomic BNDs were present to each side of the

Figure 1 (continued) BND of chromosome 7/8 in breakpoint region "e" and is depicted also by a black square inside a green square (for Hi-C and NanoVar support) here. The previously assumed breakend in region "e" was shown to be connected to the N2 fragment in cis (data not shown). (E) Based on the Hi-C pattern, we identified 2 regimes of complexity in our cohort: all of the CK-AML cases that were TP53 mutated displayed chromothriptic rearrangements, whereas most cases that were TP53 wildtype showed far less complexity. Hi-C BND regions are highlighted by black squares.

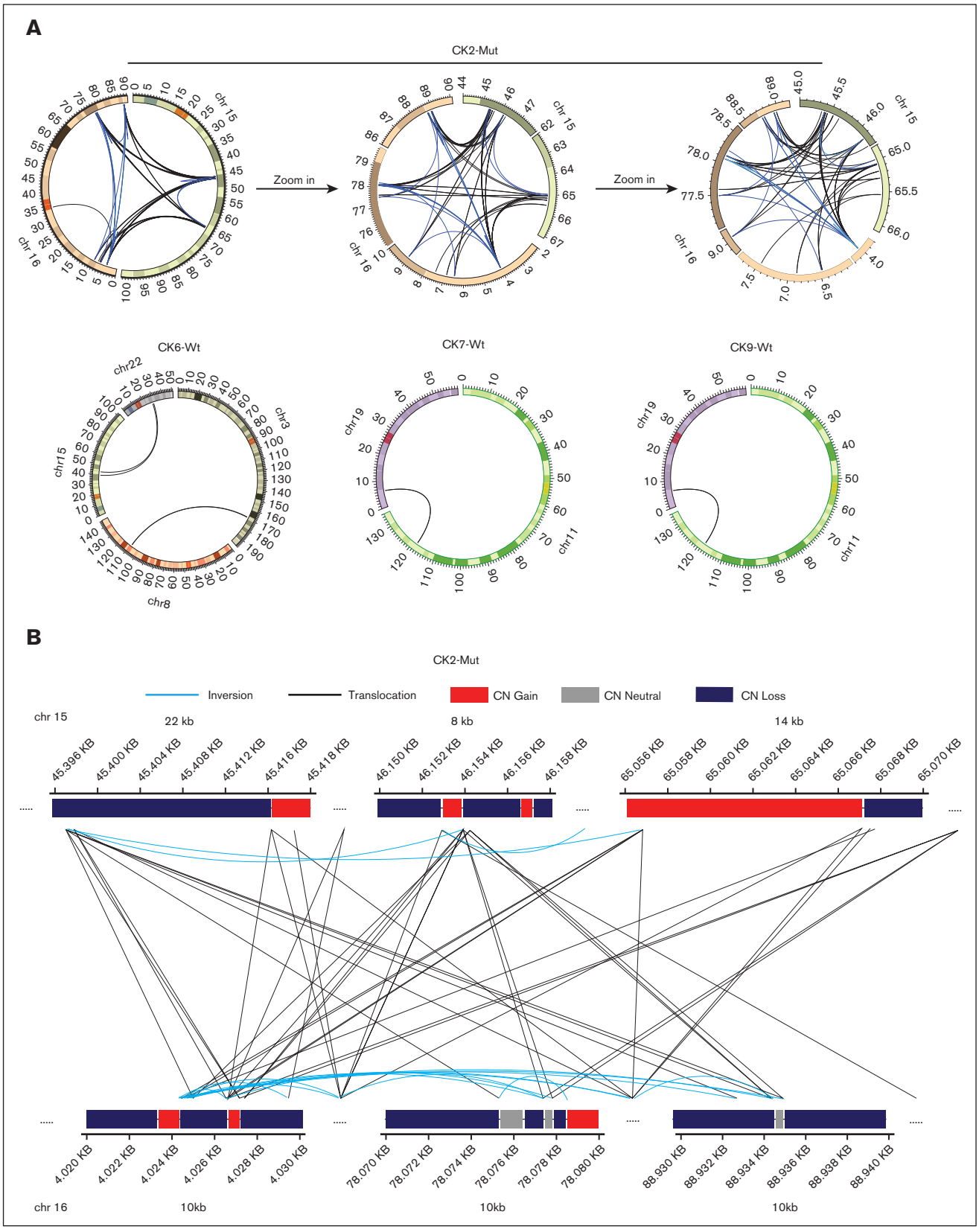

Figure 2. Chromocataclysm in CK-AML. (A) Circos plot of a chromocataclysm rearrangement in CK2-Mut and noncomplex rearrangements in CK6-Wt, CK7-Wt, and CK9-Wt. A clustering of breakends is preserved at all 3 stages of magnification shown here for case CK2-Mut. The clustering is here shown at full chromosome view on the left to increasing levels of magnifications in the middle and to the right indicating a chromocataclysm like pattern. Numbers indicate position on the chromosome in megabases. CK7-Wt

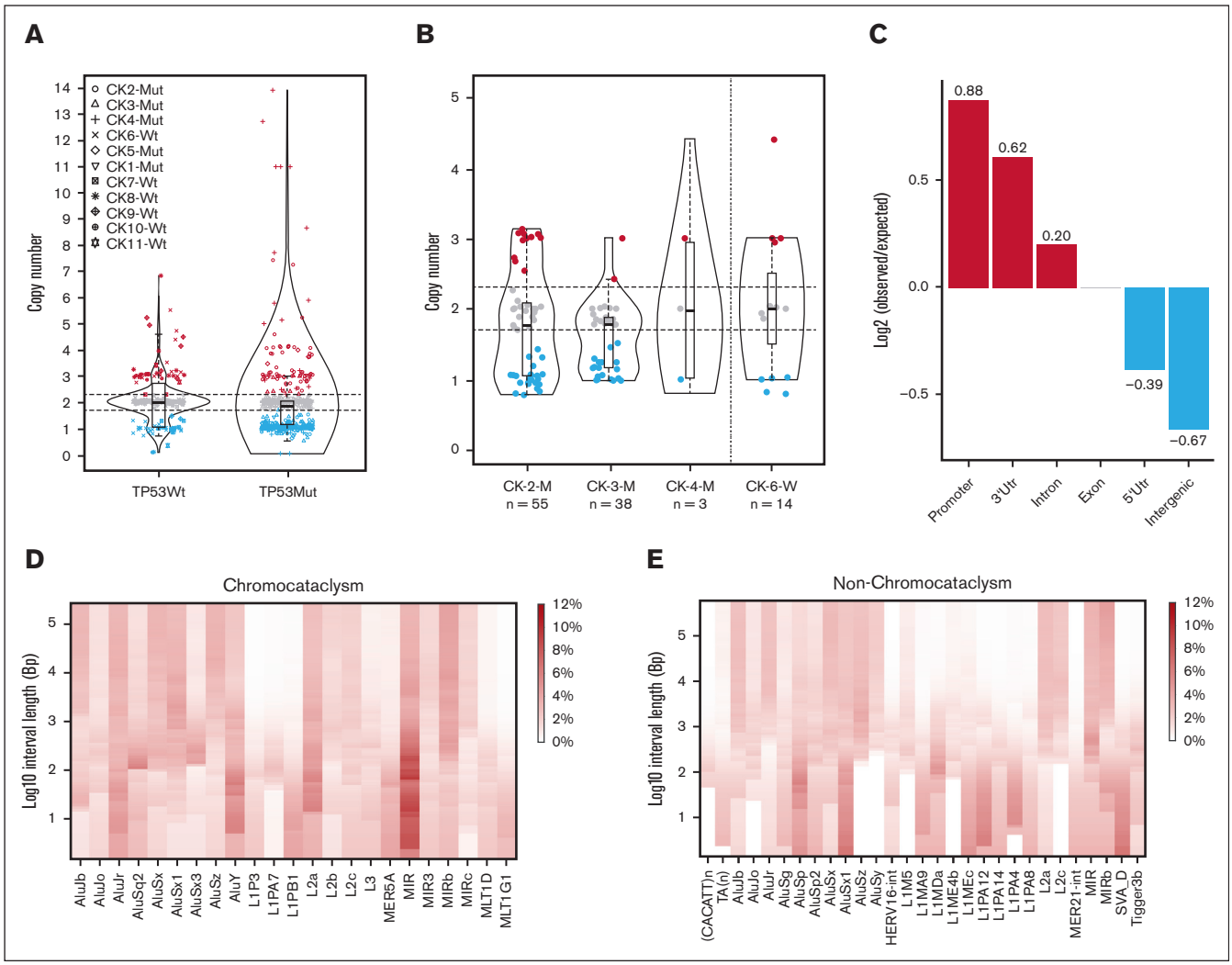

Figure 3. CN distribution and enrichment of breakends in the genome. (A) Violin plots of CN distribution in the final CNV data set of the TP53 mutated (n = 6) and TP53 wildtype (n = 5) cases. Each dot represents 1 fragment (distinct region on a genome of reference) and its respective CN. (B) Genomic fragments of <20 kb in size of the 4 cases with the highest complexity (total number of CN changes). The cases are ordered by rearrangement complexity. Blue: CN loss (CN < 1.7); red: CN gain (CN > 2.3). (C) Breakend enrichment analysis showed increased observed/expected ratio of breakends in gene promoters, 3'UTRs and introns; the opposite is observed for intergenic and 5'UTR regions. (D-E) Heatmap of the occurrence of BND in chromocataclysm cases (D) and chromothripsis cases without chromocataclysm (E) in the proximity of repetitive elements (repeat subcategories from RepeatMasker). The normalized relative occurrence was calculated for different intervals from the BNDs.

transcript (Figure 4A). This led to a high-curated data set of gene-fusion of known leukemia-associated transcripts, such as *KMT2A* (*MLL*) and *MLLT1* (CK7-Wt; Figure 4B), but also novel fusion events, such as *USP7::MVD* (CK2-Mut; Figure 4B), *ARGHA-P44::AC087294.2*, *ANKRD12::NUP88*, *PIP4K2B::ARHGAP23,* and *MTMR2::PRB1* (supplemental Table 5). Although some of the fusion partners have already been linked to oncogenesis in AML or other types of cancer, many fusion events have, to the best of our knowledge, not been reported yet (see supplemental Note).

Next, we sought to further evaluate the hypothesis that complex genomic rearrangements detected in CK-AML can lead to the activation of oncogenes. For instance, the *USP7::MVD* gene fusion (detected in CK2-Mut) resulted from the fusion of the first *USP7* exon, including the transcription start site, close to the transcription start site of *MVD* (Figure 4B). To test the potential oncogenic function of this fusion in vitro, we amplified the *USP7::MVD* fusion transcript by real-time polymerase chain reaction and cloned it into a pRSF91 retroviral vector. Transfected in NIH3T3 cells, the

Figure 2 (continued) and CK9-Wt have a similar BND connecting chromosomes 19 and 11. (B) Detailed view of some of the most complex regions that are involved in the chromocataclysm rearrangement of Chr15 and Chr16 in CK2-Mut, illustrating the extreme local complexity of CNVs and breakends. Bars show the local CN of the involved fragments. Blue: CN loss (CN < 1.7). Gray: CN stable (CN 1.7 ≤ x ≤ 2.3). Red: CN gain (CN > 2.3). Black lines show translocations (breakends on 2 different chromosomes), blue lines show inversions (breakends on the same chromosome). Dots connecting the displayed regions represent regions that are due to the complexity of the rearrangement not shown here. If breakends from the displayed regions projected to the nondisplayed regions, connections were still shown here by blue or black lines.

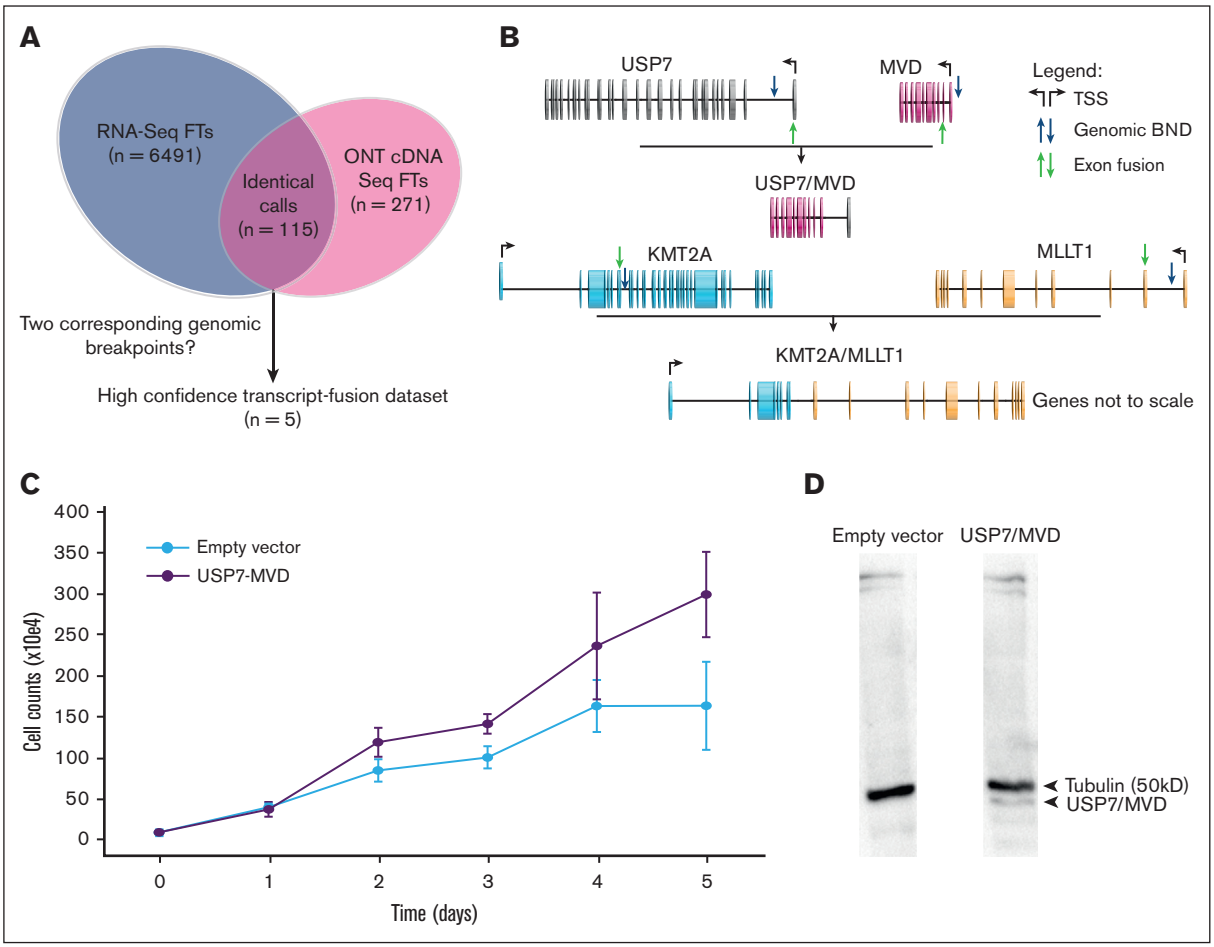

Figure 4. Identification of fusion transcripts. (A) Illustration of the fusion transcript detection pipeline, starting with integrating matching fusion transcript calls from JAFFA (Illumina RNA and ONTdirect cDNA data set) and filtering them by applying the criterion of 2 corresponding SV breakends to each identified fusion transcript. (B) Two fusion transcripts that were identified in RNA sequencing data set and also present matching genomic BNDs. The USP7/MVD fusion transcript was created by including the transcription start site (TSS) of USP7 next to the TSS of MVD, without disrupting MVD open reading frame. The fusion transcript was likely generated owing to a use of USP7 TSS and subsequent splicing-out of the first exon of MVD. TSSs are represented by black arrows; the genomic BNDs are marked by a blue arrow; and regions where the point of exon fusion was identified by JAFFA are marked with green arrows. (C) Cell culture growth of the NIH3T3 cell line transfected with a retroviral vector containing the USP7/MVD fusion transcript vs NIH3T3 cell line containing an empty vector. (D) Western blot results of the USP7/MVD fusion transcript compared with empty vector results.

USP7::*MVD* construct increased cell proliferation compared with empty vector transfected cells; thereby, pointing toward a potentially functional role of this fusion transcript and further supporting the relevance of these complex rearrangements (Figure 4C-D).

Gene expression analysis revealed a chromothripsis associated pattern of CN loss and down regulation of genes in the related genomic regions of CK-AML cases

Next, Illumina RNA sequencing data were integrated with CNV and BND information to identify dysregulation of genes in the respective regions. First, we conducted a differential expression analysis for individual patients with CK-AML against the combined $CD34^+$ controls. We then sought to identify differentially expressed genes that were potentially positively correlated with CN change (CN gain and gene expression upregulation; CN loss and gene expression downregulation). All genes that showed a CN gain and upregulation (supplemental Table 6) or a CN loss and downregulation (supplemental Table 7) in ≥3 cases, as well as genes that were disrupted because of SV BNDs occurring in the promoter or gene body and were downregulated (supplemental Table 8), were further analyzed This selection method was designed for selecting candidate genes that have a high likelihood of being relevant to CK-AML pathogenesis (Figure 5A). The CN gain and upregulation gene set included genes that were already shown to be amplified in cancer (*CEBPD* and *FBOX32*) and, in part, their overexpression has previously been linked to cancer (*CEBPD, RDH10, ASAP1, ARC,* and *TRIB1*) (supplemental Dataset 4). Among other cancer-related genes, the CN loss and downregulation gene set contained many genes with a confirmed tumor suppressor function (*CBFB, IRF5, ETV6, SMADA4, TNK1, SAMD9L, ZDHHC1, SIAH1, CUX1, ING3, RARRES2,* and *CPED1*). Interestingly, we found that IMMP2L, one of the candidate genes disrupted in a single case, was also affected by CN loss and gene downregulation events in 4 additional cases

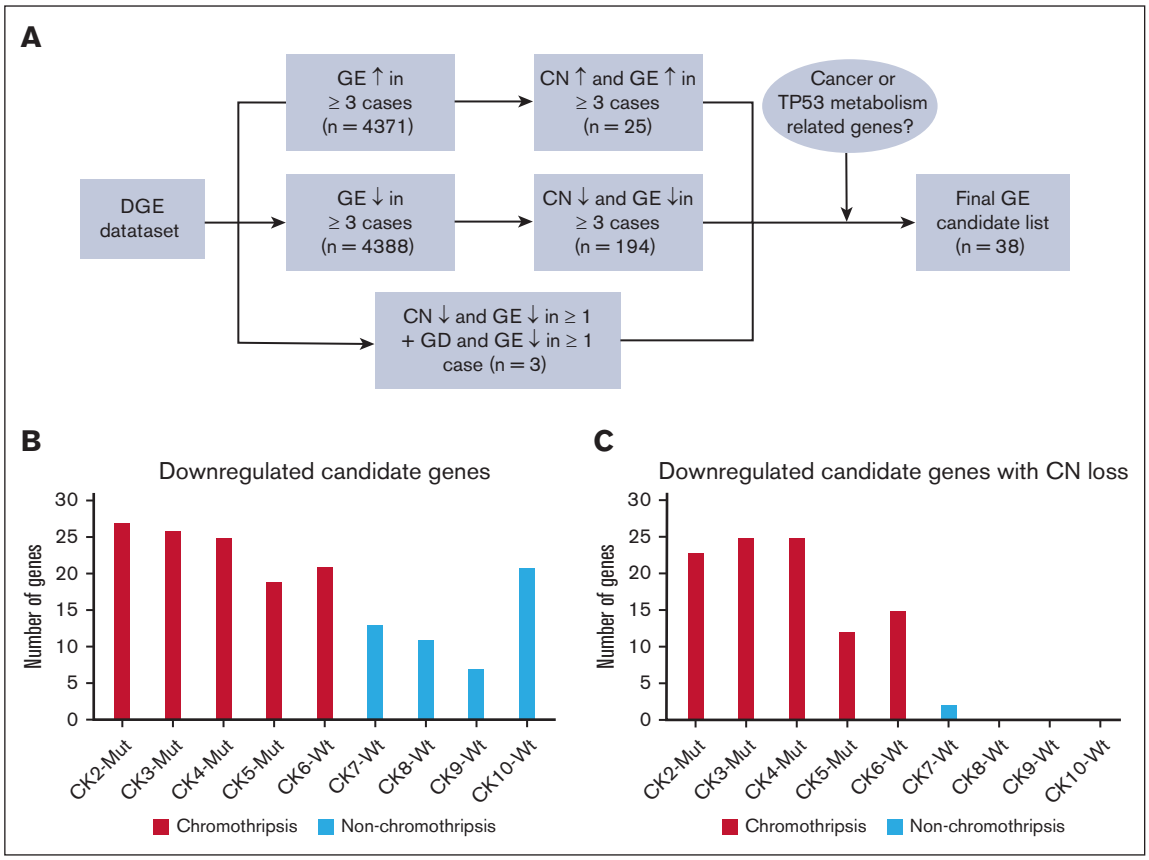

Figure 5. Gene dysregulation and influence of CN changes in CK-AML. (A) Schematic overview of our differentially expressed genes (DEG) analysis that integrates gene expression (GE) with CN information. CN ↓, CN loss; CN ↑, CN gain; GE ↓, gene downregulation; GE ↑, gene upregulation. (B) Number of genes from the 30 "downregulated candidate genes" that were downregulated in the respective case. (C) Number of genes from the 30 downregulated candidate genes that were downregulated and showed a CN loss at the gene locus.

(supplemental Dataset 4). Almost all of the 30 loss and downregulation candidate genes showed this pattern of CN loss and downregulation only in cases that we previously identified as chromothripsis. Only 2 of the candidate genes, ETV6 and LRP6, showed a CN loss and gene downregulation state in a case that was not classified as chromothripsis (CK7-Wt) (Figure 5B-C).

To further investigate a potential functional role of these genes in CK-AML, we compared our data set with AML CN data from The Cancer Genome Atlas (TCGA).[30] This analysis showed that 20 of 30 candidate genes with CN loss and downregulation were also linked to CN loss cases of the TCGA data set. Actually, these TCGA CN loss cases were to a high proportion CK-AML cases. On the other hand, 7 of 8 of the candidate genes with CN gain and upregulation were reported to be amplified in the TCGA data set (supplemental Note, supplemental Table 9).

Discussion

By combining Hi-C and ONT gDNA-LRS in patients with CK-AML, we were able to create a map of genomic aberrations with a previously unprecedented accuracy. By integrating our Hi-C SV calls with the data from NanoVar, detection of Hi-C fragments <1 kb in size has become possible. Such small fragments are extremely difficult to detect in Hi-C maps alone.[31] By applying our SV detection workflow to CK-AML data set, we found that ~50% (6/11) cases meet the criteria for chromothripsis.

Chromothripsis in cancer is regularly assessed as described by Korbel and Campbell. These criteria are composed of the following: (1) clustering of DNA BNDs; (2) oscillating CN patterns; (3) alternating pattern of retention and loss of heterozygosity; (4) occurrence on a single parental haplotype; (5) random rejoining of fragments on the derivative chromosome; and (6) alternation of BNDs between head and tail paired-end reads.[32] Using our combined approach, a large complexity of different features of chromothripsis could be revealed. Rearrangements differed significantly between cases, however, also within an individual CK-AML case, we observed strong variation regarding the level of BND clustering, CN states, presence of small local amplifications around the BNDs, as well as the relation to specific repetitive regions.

Even cases thought to harbor only a few aberrations based on karyotyping and primary visual inspection in Hi-C were shown to actually harbor many features of chromothripsis using our detailed SV analysis approach. One of the most striking differences to the original Korbel and Campbell criteria was the high presence of smaller amplifications clustering in cases with extreme local BND clustering (CK2-Mut, CK3-Mut, and CK6-Wt), but also pronouncedly in the form of larger sized amplifications (CK4-Mut),

which did not show such extreme BND clustering. Although chromothripsis is currently thought to be a single event, our data provide evidence that, in individual cases, chromothripsis events have possibly occurred independently. These rearrangements can show extreme BND clustering patterns, which we term as "chromocataclysm," whereas other rearrangements in the same case can show a much lower level of local BND clustering. Furthermore, these events were not physically linked to each other.

The huge variation of features in our cohort showed to some degree also the features of the recently introduced categories of chromoplexy and chromoanasynthesis.[10,11,33-35] Especially the presence of elevated CN state fragments and involvement of many chromosomes in some of our cases shows similarities to the concept of chromoanasynthesis that was yet thought to be a germ line-related phenomenon. However, the very complex and diverse features of the CK-AML–associated rearrangements discovered with our high resolution SV detection pipeline, leads to the hypothesis that the categories of chromothripsis, chromoplexy, and chromoanasynthesis are not fully delimited entities but rather represent a continuum of features of very complex rearrangements in cancer.

In the chromocataclysm cases, we found an enrichment of SV BNDs close to MIR elements. These elements were already shown to be associated with open chromatin and harbor enhancer functions in AML.[36,37] These regulatory elements that are closely related to the breakpoints in the chromocataclysm cases could cause a local dysregulation of gene expression around the breakpoints. Although the potentially functional role of MIR repeats in chromocataclysm remains to be elucidated in further functional studies, the association of BNDs to MIR repeats was interestingly the strongest in the most complex of our cases (CK2-Mut).

Combining our gDNA strategy with RNA-based approaches allowed us also to identify fusion transcripts with a very high accuracy, both known as well as novel fusion transcripts, such as *USP7::MVD. USP7* is known to play an important role in the *TP53/MDM2* network in many different ways, one of them is the stabilization of *TP53*.[38,39] The fusion of *USP7* to *MVD* leads to a functional deletion of 1 of the *USP7* alleles. In accordance, partial knockdown of *USP7* was shown to cause the destabilization of *TP53*,[39] and the *TP53* expression level were by far the lowest in CK2-Mut, which also showed mutation in 1 allele of TP53. Thus, we hypothesize that the fusion transcript USP7::MVD influences the *TP53/MDM2* network that contributes to the emergence of additional chromothripsis and/or chromocataclysm events in CK-AML.

By looking at gene expression changes that were associated with CN alterations or gene disruptions in our CK-AML cases, we could further delineate a list of potential CK-AML driver candidate genes. The potential pathogenic impact of these genes is further underlined by their exclusive location in regions that are known to be recurrently affected by CN changes in CK-AML. The almost exclusive association of CN loss and gene downregulation events in our candidate gene list suggests that these events may have an important role in CK-AML with chromothripsis. This can help to distinguish these cases from other CK-AML cases, which might help to further refine CK-AML management. In line with the enrichment of breakpoints in promoter regions, our results show how complex SVs can influence CK-AML pathogenesis by the disruption of specific tumor suppressor genes and activation of oncogenes in regions of BND and CNV clustering.

In line with a recent study in AML combining Hi-C and whole-genome sequencing,[40] our workflow integrating gDNA-LRS and Hi-C sequencing has the potential to provide an even more precise picture of SVs in tumors with complex genomic rearrangements, thereby enabling us to discover novel features of chromothripsis and SVs of potential functional impact. The main strength of our approach lies in the integration of 2 very different technologies that are not likely to suffer from the same bias, therefore, strongly reducing false-positive results. Another important advantage compared with approaches based on short-read sequencing is the possibility of long-read sequencing to span repetitive regions.[8,41,42] In our data set, 58% of all found breakpoints in the chromothripsis cases and 43% of all breakpoints in the cases with chromocataclysm were located in repetitive regions. The application of this workflow to various other cancers in the future could greatly enhance the understanding of the role of SVs in cancer and potentially lead to novel therapeutic options for patients in need.

Acknowledgments

The authors thank Thomas Risch for his valuable input about gene expression analysis strategies.

This study was supported in part by the Bundesministerium für Bildung und Forschung (ERA PerMed projects SYNtherapy 01KU1917 and MEET-AML 01KU2014 to L.B.). M.-K.K. was supported by an MD student research scholarship (Berlin Institute of Health) and a Peter-Scriba scholarship of the German Association for Internal Medicine. S.M. was supported by grant MU 880/16-1 from the Deutsche Forschungsgemeinschaft.

Authorship

Contribution: M.-K.K. designed and performed experiments, analyzed data, and wrote the manuscript; J.J., A.D. F.G.R., F.S., Z.X., and U.S.M. performed the experiments; E.S., S.H., and R.S. performed bioinformatic analyses; J.-F.S., O.B., J.W., K.D., and H.S. collected the data and provided essential samples; M.S., A.M., U.S.M., S.M., and L.B. supervised the project; and U.S.M., S.M., and L.B. conducted the overall project planning and revised the manuscript.

Conflict-of-interest disclosure: L.B. has advisory role in AbbVie, Amgen, Astellas, Bristol Myers Squibb, Celgene, Daiichi Sankyo, Gilead, Hexal, Janssen, Jazz Pharmaceuticals, Menarini, Novartis, Pfizer, Sanofi, and Seattle Genetics; and receives research funding from Bayer and Jazz Pharmaceuticals. The remaining authors declare no competing financial interests.

ORCID profiles: S.H., 0000-0002-4783-3814; M.S., 0000-0002-0583-4683; A.M., 0000-0001-8646-7469.

Correspondence: Lars Bullinger, Department of Hematology, Oncology and Tumorimmunology, Charité University Medicine, Augustenburger Platz 1, 13353 Berlin, Germany; email: lars.bullinger@charite.de; and Stefan Mundlos, Institute for Medical Genetics and Human Genetics, Charité University Medicine, Augustenburger Platz 1, 13353 Berlin, Germany; email: stefan.mundlos@charite.de.

References

1. Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: recent progress and enduring challenges. *Blood Rev*. 2019;36:70-87.
2. Döhner H, Wei AH, Appelbaum FR, et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood*. 2022;140(12):1345-1377.
3. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *Leukemia*. 2022;36(7):1703-1719.
4. Arber DA, Orazi A, Hasserjian RP, et al. International Consensus Classification of myeloid neoplasms and acute leukemias: integrating morphologic, clinical, and genomic data. *Blood*. 2022;140(11):1200-1228.
5. Rücker FG, Schlenk RF, Bullinger L, 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):2114-2121.
6. Bullinger L, Döhner K, Döhner H. Genomics of acute myeloid leukemia diagnosis and pathways. *J Clin Oncol*. 2017;35(9):934-946.
7. Mrózek K. Cytogenetic, molecular genetic, and clinical characteristics of acute myeloid leukemia with a complex karyotype. *Semin Oncol*. 2008;35(4):365-377.
8. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol*. 2019;20(1):246.
9. Macintyre G, Ylstra B, Brenton JD. Sequencing structural variants in cancer for precision therapeutics. *Trends Genet*. 2016;32(9):530-542.
10. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011;12(5):363-376.
11. Stephens PJ, Greenman CD, Fu B, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*. 2011;144(1):27-40.
12. Cortés-Ciriano I, Lee JJ, Xi R, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet*. 2020;52(3):331-341.
13. Fontana MC, Marconi G, Feenstra JDM, et al. Chromothripsis in acute myeloid leukemia: biological features and impact on survival. *Leukemia*. 2018;32(7):1609-1620.
14. Dixon JR, Xu J, Dileep V, et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet*. 2018;50(10):1388-1398.
15. Chaisson MJP, Sanders AD, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 2019;10(1):1784.
16. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*. 2019;10(1):3240.
17. Ramirez R, van Buuren N, Gamelin L, et al. Targeted long-read sequencing reveals comprehensive architecture, burden, and transcriptional signatures from hepatitis B virus-associated integrations and translocations in hepatocellular carcinoma cell lines. *J Virol*. 2021;95(19):e0029921.
18. Nattestad M, Goodwin S, Ng K, et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res*. 2018;28(8):1126-1135.
19. Melo US, Schöpflin R, Acuna-Hidalgo R, et al. Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases. *Am J Hum Genet*. 2020;106(6):872-884.
20. Wang S, Lee S, Chu C, et al. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol*. 2020;21(1):73.
21. Tham CY, Tirado-Magallanes R, Goh Y, et al. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biol*. 2020;21(1):56.
22. Davidson NM, Majewski IJ, Oshlack A. JAFFA: high sensitivity transcriptome-focused fusion gene detection. *Genome Med*. 2015;7(1):43.
23. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
24. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33(3):290-295.
25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
26. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199-W205.
27. Risueño A, Roson-Burgo B, Dolnik A, Hernandez-Rivas JM, Bullinger L, De Las Rivas J. A robust estimation of exon expression to identify alternative spliced genes applied to human tissues and cancer samples. *BMC Genomics*. 2014;15(1):879.
28. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11.
29. Carnevali D, Conti A, Pellegrini M, Dieci G. Whole-genome expression analysis of mammalian-wide interspersed repeat elements in human cell lines. *DNA Res*. 2017;24(1):59-69.

30. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120.
31. Harewood L, Kishore K, Eldridge MD, et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol*. 2017;18(1):125.
32. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell*. 2013;152(6):1226-1236.
33. Plaisancié J, Kleinfinger P, Cances C, et al. Constitutional chromoanasynthesis: description of a rare chromosomal event in a patient. *Eur J Med Genet*. 2014;57(10):567-570.
34. Pellestor F, Gatinois V. Chromoanasynthesis: another way for the formation of complex chromosomal abnormalities in human reproduction. *Hum Reprod*. 2018;33(8):1381-1387.
35. Baca SC, Prandi D, Lawrence MS, et al. Punctuated evolution of prostate cancer genomes. *Cell*. 2013;153(3):666-677.
36. Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak VV, Jordan IK. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob DNA*. 2014;5:14.
37. Zeng Y, Cao Y, Halevy RS, et al. Characterization of functional transposable element enhancers in acute myeloid leukemia. *Sci China Life Sci*. 2020;63(5):675-687.
38. Tavana O, Sun H, Gu W. Targeting HAUSP in both p53 wildtype and p53-mutant tumors. *Cell Cycle*. 2018;17(7):823-828.
39. Bhattacharya S, Chakraborty D, Basu M, Ghosh MK. Emerging insights into HAUSP (USP7) in physiology, cancer and other diseases. *Signal Transduct Target Ther*. 2018;3:17.
40. Xu J, Song F, Lyu H, et al. Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature*. 2022;611(7935):387-398.
41. Bolognini D, Magi A. Evaluation of germline structural variant calling methods for nanopore sequencing data. *Front Genet*. 2021;12:761791.
42. Sanchis-Juan A, Stephens J, French CE, et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med*. 2018;10(1):95.