DISSERTATION

Genomische und transkriptomische Analyse komplexer struktureller Varianten und ihre Relevanz in der AML mit komplexem Karyotyp

Genomic and transcriptomic analysis of complex structural variants and their relevance in AML with a complex karyotype

zur Erlangung des akademischen Grades
Medical Doctor - Doctor of Philosophy (MD/PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von
Marius-Konstantin Klever

Erstbetreuung: Prof. Dr. Lars Bullinger

Datum der Promotion: 28.02.2025

Table of contents

List of figures

List of abbreviations

Abbreviation	Definition
AML	Acute myeloid leukemia
BND	Interchromosomal breakends
Bp	Basepair
cDNA	Complementary DNA
CK-AML	Acute myeloid leukemia with complex karyotype
CML	Chronic myeloid leukemia
CN	Copy number
CNV(s)	Copy-number variation(s)
gDNA	Genomic DNA
GO	Gene Ontology
Hi-C	High-throughput chromosome confirmation capture
INV	Intrachromosomal breakends
Kb	Kilobase
Mb	Megabase
MIR	Mammalian-wide interspersed repeat
mRNA	Messenger RNA
ONT	Oxford Nanopore Technologies
SNV(s)	Single-nucleotide variant(s)
SV(s)	Structural variant(s)
TCGA	The Cancer Genome Atlas
UTR	Untranslated region
WGS	Whole genome sequencing

Abstract

Structural variants (SVs) are of major importance for the field of cancer genomics as well human genetics. Their reliable detection and functional interpretation are still an important limitation to both fields. In this work, we applied SV detection methods based on gDNA long-read sequencing and high-throughput chromosome confirmation capture (Hi-C) to samples from patients with acute myeloid leukemia (AML) as well as patients with germline genetic disorders in order to develop a novel workflow for SV detection. This integrated workflow enabled us to characterize the SV landscape of the AML subtype of AML with complex karyotype (CK-AML), which is associated with a very poor prognosis and therapy resistance. The complexity of the structural variants that we identified exceeds the previous knowledge about complex genomic rearrangements in CK-AML and existing concepts like chromothripsis. The extreme local complexity in one CK-AML case reached up to an SV cluster of 31 breakends in a region only 2.7 kilobases in size, detectable with both Hi-C and gDNA long-read sequencing. These extremely complex SVs consisted in large part of focal amplifications and were enriched in the proximity of mammalian-wide interspersed repeat (MIR) elements. To get additional insight into the functional consequences of these structural variants, we further characterized the transcriptome of the cohort using RNA and long-read direct cDNA sequencing. This enabled us to identify novel fusion transcripts such as *USP7::MVD*. In the differential gene expression analysis, we identified new as well as previously described candidate genes associated with CK-AML. Our integrated structural variant detection workflow enabled us to identify complex genomic rearrangements with very high resolution and accurate elimination of likely false-positive SVs in the datasets. Our workflow can help to further elucidate the structure and consequences of complex genomic rearrangements in the future.

Zusammenfassung

Strukturelle Varianten (SVs) sind sowohl für die Krebsgenomik als auch für die Humangenetik von großer Bedeutung. Ihre zuverlässige Erkennung und funktionale Interpretation stellen für beide Bereiche immer noch eine wichtige Limitation dar. In dieser Arbeit wandten wir SV Detektionsmethoden basierend auf gDNA-long-read Sequenzierung und high-throughput chromosome confirmation capture (Hi-C) auf Proben von Patienten mit akuter myeloischer Leukämie (AML) sowie Patienten mit genetischen Erkrankungen der Keimbahn an, um eine neuartige Methode für den Nachweis von SVs zu entwickeln. Diese Methode ermöglichte es uns, die SV-Landschaft des Subtyps der AML mit komplexem Karyotyp (CK-AML) zu charakterisieren, welche mit einer sehr schlechten Prognose und Therapieresistenz assoziiert ist. Die Komplexität der von uns identifizierten SVs übersteigt das bisherige Wissen über komplexe genomische Varianten in der CK-AML und bekannte Phänomene wie Chromothripsis. Die extreme lokale Komplexität in einem CK-AML Fall erreichte in einem SV-Cluster 31 Bruchpunkte in einer Region mit einer Größe von nur 2,7 Kilobasen, welche sowohl mit Hi-C- als auch mit gDNA-Long-Read-Sequenzierung nachweisbar waren. Diese äußerst komplexen SVs bestanden zu einem großen Teil aus fokalen Amplifikationen und waren häufig in der Nähe von mammalian-wide interspersed repeat (MIR) Elementen angereichert. Um die funktionellen Konsequenzen dieser SVs zu untersuchen, charakterisierten wir ebenfalls das Transkriptom der Kohorte mithilfe von RNA und and long-read direct cDNA Sequenzierung. Dadurch konnten wir neuartige Fusionstranskripte wie *USP7::MVD* identifizieren. In der Genexpressionsanalyse konnten wir sowohl neue als auch bereits beschriebene, mit der CK-AML assoziierte. Kandidatengene identifizieren. Unsere integrative Detektionsmethodik zur Erkennung struktureller Varianten ermöglichte es uns, komplexe genomische Varianten mit sehr hoher Auflösung zu identifizieren und falsch positive SVs in den Datensätzen zu eliminieren. Unsere Methodik kann dazu beitragen, die Struktur und die Folgen komplexer genomischer Varianten in der Zukunft weiter aufzuklären.

1. Introduction

1.1 Structural variants

Structural variants (SVs) are a diverse class of genomic alterations, ranging from 50 basepair (bp) to megabase (mb) size alterations. SVs can be roughly divided into unbalanced copy-number variations (CNVs), including deletions, duplications and insertions as well as balanced SVs, including inversions and translocations. In complex genomic rearrangements, these subcategories are often combined in multiple ways and difficult to clearly distinguish [1]. SVs can be pathogenic due to various mechanisms. Unbalanced CNVs can lead to alteration in gene expression. A different number of copies of the respective gene can lead to an increase or decrease in gene expression of tumor suppressor genes or oncogenes. Balanced SVs like translocations and inversions but also forms of unbalanced translocation can lead to gene truncation and functional inactivation of the gene. On the other hand, such mechanisms can lead to the emergence of fusion genes with an aberrant function. In addition, SVs can lead to aberrant promoter enhancer interactions due to positional effects in the 3D genome, leading to local gene dysregulation.

SVs are known to be of great importance to the pathogenesis of multiple cancers [2]. The effect of structural variants to the genome is thought to be equivalent or even larger compared to the importance of single-nucleotide variants (SNVs) [3]. However, the pathogenetic role of SNVs were a large focus of cancer research in the past decades, while studying SVs was more complicated and limited by the ability to reliably detect SVs [4,5]. Great progress in the field of cancer genomics was made in identifying the pathogenetic role of simple rearrangements. One prominent is the Philadelphia chromosome, a chromosome 9 and 22 translocation leading to the emergence of the *BCR::ABL1* fusion gene. The *BCR::ABL1* gene is a functional tyrosine kinase and has an important pathogenic function in chronic myeloid leukemia (CML), as well as in a subset of acute leukemias. This discovery led to the development of tyrosine kinase inhibitors like Imatinib, revolutionizing the treatment of CML [6].

1.2 Acute myeloid leukemia with a complex karyotype

Acute myeloid leukemia (AML) is a malignant disease of the bone marrow. It is the most common acute leukemia in adults, accounting for about 80% of all cases, showing poor survival rates especially in the elderly [7]. AML with a complex karyotype (CK-AML) is a subtype of AML with a poor outcome and poor response to chemotherapy compared to most other subtypes of AML. Its prevalence increases with age and about 10-14% of all AML patients are diagnosed with CK-AML [8]. A complex karyotype in AML is defined as ≥3 SVs that occur without the presence of specific recurring SVs, i.e. 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* [9–11]. *TP53* mutations are present in about 70% of all CK-AML cases, while they are relatively rare in other AML subtypes. These mutations are associated with chromosomal instability and lead to a functional inactivation of this important tumor suppressor gene [12,13]. It is likely that the SVs present in CK-AML are of crucial importance for the pathogenesis, but the underlying pathomechanisms are not properly understood yet.

1.3 Complex genomic rearrangements and chromothripsis

In some CK-AML cases, very complex rearrangements like chromothripsis occur. Chromothripsis is a complex process of chromosome shattering, fragment deletion and complex rearrangement of the derivative chromosomes, occurring as an early event in cancer development [14,15]. In AML, chromothripsis is relatively rare, but occurs in about 35% of all CK-AML cases [16]. It also occurs in high frequencies in several other cancers. For example, in lung cancer, osteosarcoma, liposarcoma and melanoma, the frequency of chromothripsis is >50%, as reported by the PCAWG working group [17].

1.4 Challenges in the detection of structural variants in cancer

The detection of SVs based on short-read whole genome sequencing (WGS) is still hampered by false positives and often low recall rates [5]. Also, the results from different SV calling algorithms vary substantially [18]. A comparison of the translocation breakend calls from short-read WGS and high-throughput chromosome confirmation capture (Hi-C) data from leukemia cell lines revealed that the overlap of identical SV calls between these datasets, even in-between the two short read SV calling algorithms executed on the same dataset, was very low [19]. In another study, which compared the SV detection

capabilities of Hi-C, optical mapping, short-read WGS, karyotyping, fused transcript detection and paired-end tag sequencing, only 21% of all found inter-chromosomal translocations in a leukemia cell line were detected by at least 2 of these 6 methods, providing even more evidence for the current difficulties in correct SV identification [20]. Regarding the detection of unbalanced CNVs, microarray-based comparative genomic hybridization was used extensively to detect larger CNVs associated with cancer. However, the resolution is limited and precise breakend mapping is not possible with this technology [21].

1.5 Research question and choice of methods in this work

The main research question in this work was if the use of Hi-C and long-read sequencing could overcome some of the current difficulties of SV calling and enable us to understand more about the SVs in CK-AML. Therefore, we developed an integrated SV detection workflow based on Oxford Nanopore Technologies (ONT) genomic DNA (gDNA) long-read WGS and Hi-C and subsequently applied this workflow to a cohort of 11 CK-AML cases. We developed our workflow by first applying only Hi-C as a primary tool for breakend detection in the CK-AML cases as well as to additional cases with complex germline rearrangements, to assess the capabilities of Hi-C in solving such rearrangements. However, in Hi-C alone, exact identification of correct breakends and their discrimination from breakend-like patterns is not possible with certainty when it comes to the analysis especially of small fragments <5-10 kilobases (kb). Therefore, we sought to improve the results that we were able to obtain with Hi-C. This could be done by a method that overcomes some of the shortfalls of Hi-C and that could on the other hand also benefit from the advantages of Hi-C. Our overall goal was to integrate the technologies in a way that largely excludes false positives, which are known to heavily affect data analysis and hinder high-confidence SV calling.

A technology that has the potential to overcome many difficulties of SV detection is long-read WGS. The reads of short-read WGS, which is largely used in the field of SV detection, are not long enough to span repetitive elements of the genome. The reads of short-read WGS are usually only up to a few hundred basepairs long. Breakends in repetitive regions are thus likely to be missed by SV calling algorithms based on short read data [22]. Long-read WGS has the important advantage of spanning these repetitive regions with reads ranging from kilobases to even megabases in size, thereby potentially

enabling accurate SV detection also in repetitive regions, which make up about >50% of the genome [23,24]. In addition to the genomic SV dataset derived by our integrative workflow, we also characterized the transcriptome of the cohort, what enabled us to study the pathogenetic effect of the SVs in detail. To analyze the transcriptome, we used Illumina RNA-Seq as well as ONT direct complementary DNA (cDNA) sequencing. ONT direct cDNA sequencing is a novel transcriptome sequencing method, which provides the advantage of reads that can span full transcripts, therefor potentially enabling a better detection of transcript isoforms and fusion transcripts. We used Illumina RNA sequencing for differential gene expression analysis because of the still higher throughput compared to ONT direct cDNA sequencing and therefore more precise transcript quantification [25]. For fusion transcript detection, we used both transcriptome sequencing technologies. Integrating these methods with our SV dataset enabled us to identify known as well as novel fusions transcripts, which were not described to date, with high confidence. Furthermore, it enabled us to precisely study the effect of these SVs on gene expression.

2. Methods

2.1 Methods of the different projects

In the following, the methods of the main CK-AML project will be explained in detail. Further details of the methods used in this project can be found in the methods and supplementary methods of the respective publication [26]. The work that was done as part of the two other publications that are part of this dissertation involved mainly Hi-C experiments and analysis. These experiments were largely performed the same way that the Hi-C workflow of the CK-AML project was performed. Further information about the methods used in these projects can be found in the respective publications [27,28].

2.2 Ethics approval, screening for mutations and karyotyping

Approval of the Ethics committee of the Charité University Medicine Berlin was obtained for this study. Samples from bone marrow biopsies or fresh peripheral blood were collected at the respective Hematology and Oncology departments at the Charité and the University of Ulm, Germany. 11 CK-AML samples were collected from different patients with informed and written consent within the AMLSG BiO Registry study (NTC 01252485). Furthermore, CD34+ stem cells were collected from 5 healthy bone marrow donors with informed and written consent. The collection was performed via peripheral blood apheresis after G-CSF stimulation. For these 5 CD34+ stem cell samples, supernumerary material, not needed for clinical allogenic hematopoietic stem cell transplantation, was used. Karyotyping and screening for AML associated mutations (*FLT3*-ITD/TKD, *CEBPA*, *KMT2A* [previously *MLL*], *NPM1*, *IDH1* and *TP53*) was performed previously at collection of the respective samples. Only samples with a complex karyotype based on the European LeukemiaNet (ELN) classification [9] were included in this study.

2.3 Processing of the CK-AML and CD34+ stem cell samples

After collection of the samples of all CK-AML cases the cells were ficoll centrifuged to enrich for leukemic blasts (>90% of total cells) and frozen in liquid nitrogen at an average density of $1x10^7$ cells/ml. CD34+ stem cells were purified by CD34-selection via CliniMACS-LS-columns (Miltenyi Biotech, Bergisch-Gladbach, DE), and T-cell depletion with SAM-Beads (Miltenyi Biotec) and Orthoclone OKT3-antbodies (Janssen-Cilag,

Neuss, DE). Flow cytometry confirmed a CD34+ purity >95% of total cells after purification. The CD34+ stem cells were subsequently frozen in liquid nitrogen. The work listed above was carried out previously to this study by the respective departments. Before processing the samples with our sequencing pipelines, the cells were thawed for 1h in a water bath at 37°C. For this, a thawing medium, consisting of RPMI 1640 medium (Thermo Fisher Scientific, Waltham, MA, USA), supplemented with 20% heat inactivated fetal bovine serum (PAN Biotech, Aidenbach, Germany), Heparin (Merck KGaA, Darmstadt, DE), MgCl2 (Thermo Fisher Scientific) and DNAse I (Sigma-Aldrich, St. Louis, MO, USA) was used. After thawing, the cells were counted on an EVE Automated Cell Counter (Nano EnTek Inc, Seoul, KOR) after Trypan Blue staining (Nano EnTek Inc) and subsequently equally divided for Hi-C library preparation and DNA/RNA isolation.

2.4 Hi-C library preparation and sequencing data analysis

Hi-C libraries for sequencing were generated for 11 CK-AML cases and 1 CD34+ stem cell sample from a healthy donor, which served as a control Hi-C map. Here, a Hi-C protocol based on the in-situ Hi-C protocol developed by Rao et al., was used [29].

First, 500.000 to 1 million cells per replicate were crosslinked in RPMI 1640 medium supplemented with 1% formaldehyde (Sigma-Aldrich) and incubated for 10 min at room temperature. The crosslinking was quenched by adding glycine (Merck KGaA) and incubating on ice. After crosslinking, the cells were lysed in the lysis buffer. Restriction enzyme digestion of the DNA was performed using the *DpnII* enzyme (New England BioLabs, Ipswich, MA, USA). In the next step, the ends of the restriction enzyme cutting sited were filled up with biotinylated nucleotides (biotin-14-dATP, dCTP, dGTP, dTTP) (Thermo Fisher Scientific) and ligated using T4 DNA Ligase (New England BioLabs). In the following, the crosslinks of the biotinylated fragments were reversed by adding proteinase K (Thermo Fisher Scientific), SDS (Carl Roth GmbH + Co. KG, Karlsruhe, DE) and incubation at 55°C for 30 minutes with subsequent addition of 5M NaCl (Carl Roth GmbH + Co. KG) and incubation at 68°C for 2h. DNA precipitation was then performed by adding 3M Sodium Acetate (Merck KGaA) and Ethanol (Merck KGaA). In order to generate DNA fragments that are suitable for library sequencing (300-500 base pairs), DNA shearing with a S220 ultrasonicator (Covaris, Woburn, MA, USA**)** was used. After shearing, a biotin pulldown using Dynabeads MyOne Streptavidin T1 (Thermo Fisher Scientific) was done. Next, reparation of the fragment ends of the sheared DNA and biotin

removal was performed using T4 DNA polymerase (New England BioLabs) and DNA polymerase I, Large (Klenow) Fragment (New England BioLabs). The fragments were subsequently phosphorylated using T4 Polynucleotide Kinase (New England BioLabs).

Finally, the libraries were amplified using PCR and the DNA was purified. For the PCR, the NEBNext Multiplex Oligos for Illumina kit (New England BioLabs) was used, ligating Illumina adaptors to the samples and labeling them with Index and Universal primers (New England BioLabs). PCR was performed using the NEBNext Ultra II Q5 Master Mix (New England BioLabs) on a SimpliAmp Thermocycler (Thermo Fisher Scientific). The PCR amplified DNA was purified and size-selected using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA). 2-4 Hi-C library replicates of each case were sequenced with a target total sequencing depth of about 320 million fragments on a NovaSeq 6000 sequencing machine (Illumina, San Diego, CA, USA). The Hi-C sequencing data of the CK-AML project was processed here similar to the other published projects that are partially integrated in this monograph [27,28]. The replicates of the Hi-C library were processed, and quality checked by using the Juicer pipeline [30]. The final Hi-C files to create the maps were created with Juicer tools. For these final maps, all replicates of a single case were merged. The Hi-C maps were visually analyzed using the Juicebox desktop application [31]. To test the possibility of using automatized breakend detection tools in Hi-C instead of visual analysis, we used two breakend detection tools, hic_breakfinder [20] and automatized breakend detection with HiNT [19], that we executed with standard settings on the Hi-C files of all CK-AML cases.

2.5 DNA and RNA extraction and processing

DNA extraction for ONT genomic gDNA long read sequencing and RNA extraction for ONT direct cDNA long read sequencing as well as Illumina RNA Sequencing was done with the AllPrep DNA/RNA/Protein Mini Kit (Qiagen, Hilden, DE). The extracted RNA was quality checked using an Agilent Technologies Tape station on an RNA ScreenTape (Agilent Technologies, Santa Clara, CA, USA**)**. If the RNA Integrity Number was ≥ 8.0, the RNA was used for downstream processing and sequencing. For ONT long read direct cDNA sequencing, messenger RNA (mRNA) as input material was isolated from the total RNA using the Dynabeads mRNA Purification Kit (Thermo Fisher Scientific).

2.6 ONT long read library preparation and analysis (gDNA and cDNA)

After extraction, gDNA was prepared for gDNA long read sequencing using the Ligation Sequencing Kit (Oxford Nanopore Technologies, Oxford, UK, SQK-LSK109). The gDNA long read sequencing was performed until a genomic coverage of at least 10x for each patient was reached. The isolated mRNA for long read direct cDNA sequencing was reverse transcribed and prepared for sequencing using the ONT direct cDNA Sequencing Kit (Oxford Nanopore Technologies). All long read sequencing libraries were sequenced on a GridION on R9.4.1 flowcells (Oxford Nanopore Technologies). For processing of the raw sequencing data, the Oxford Nanopore Technologies software MinKNOW was used. Further details concerning the processing of the ONT long-read sequencing data, and the tools used for data analysis can be found in the methods and supplementary methods of the respective publication [26].

2.7 Fusion transcript analysis

To identify high confidence fusion transcripts based on our transcriptomic dataset, we analyzed the Illumina RNA sequencing data with the direct mode of the fusion caller JAFFA [32], using standard settings and alignment to the reference genome hg19. In addition, we also analyzed the ONT direct cDNA files with the long read version of JAFFA, using otherwise the same settings as for the short reads. Fusion calls in the Illumina RNA and ONT direct cDNA datasets were identified as matching if both transcriptomic “fusion break points” were located at the same exon-intron boundary.

2.8 Structural variant breakend signatures and genomic distribution

We further analyzed our SV call dataset: 1. for the occurrence of breakends inside or in the close proximity of repetitive elements and 2. for the distribution of breakends in different features of the genome. For the analysis of breakend distribution relative to repetitive elements, category and genomic location information from the Repbase repository was used [33]. The local density of breakends was calculated by kernel density estimation for the whole genome. This local density was visualized with the R package ggridges at kernel bandwidths 1 mb, 1 kb and 0.5 kb. For the calculation of the distribution of breakends relative to genomic features, genomic annotation data (transcription start sites, exons, introns, UTR3’ and UTR5’) from GENCODE (release 19) was used [34]. For

this calculation, promoter regions were defined as areas 1.5 kb upstream and 0.5 kb downstream of the respective transcription start site of the gene. 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.

To test the significance of enrichment of breakends relative to the genomic features and repetitive elements, a two-sided Mann-Whitney-U-test against 10000 random breakends was performed. For this analysis, low complexity genomic regions were excluded (telomeres, centromeres and variable genomic regions). Test correction was performed using the Benjamini-Hochberg method. Distances to the breakends as well as the respective confidence intervals were estimated by bootstrap over 5000 iterations. Regions on sex chromosomes were excluded from these analyzes, because sex chromosomes were not affected by any breakends in any of the cases reported here.

2.9 Illumina RNA sequencing and data processing

All cases but CK1-Mut and CK11-Wt were analyzed with Illumina RNA sequencing, based on material availability. We used 3 replicates for each of the cases that we sequenced. For the analysis of differential gene expression, we also sequenced single replicates of 5 healthy CD34+ hematopoietic stem cell donor controls. Poly-A selection was used to isolate stranded mRNA and 100 bp sequencing was performed with 100 million reads per replicate. Sequencing was done on a NovaSeq 6000 sequencing machine (Illumina). Further details concerning the processing of the RNA sequencing reads and the differential gene expression analysis as well as the comparison of the results of this analysis with existing datasets can be found in the methods and supplementary methods of the respective publication [26].

3. Results

3.1 Hi-C as a tool for SV detection in complex rearrangements

In Hi-C, interaction of two genomic loci is represented by signal intensity in the Hi-C maps. The closer the 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 at this genomic position. This represents increased interaction frequency of chromosome loci, that are normally not linked together and thus normally only interact to a low degree. In most cases, especially with simple rearrangements, the difference in signal intensity is high and differences are easily visible. Translocations are visible in trans-maps (interchromosomal) and inversions in cis-maps (intrachromosomal). The point with the highest signal intensity is regularly the actual breakend because the signal intensity decreases rapidly with genomic distance. Based on the direction of the fading signal, the orientation of the fragments can be inferred [27].

Good examples for visualization are simple large-scale translocations, as we observed in two cases with developmental delay (DD7 and DD8) (Figure 1) [27]. In the “all chromosomes view” of the visualization software Juicebox, it is already visible that normally only cis-maps (e.g. chromosome 2 and 2 cis-map) show high signal interaction frequencies (coded here in red), because different loci on a chromosome interact with each other. However, in the examples presented here, it is visible in the “all chromosomes view”, that also the chromosome 2 and 4 trans-maps in case DD7 as well as the chromosome 7 and 8 trans-maps in DD8 show high interaction signal intensities, which are not visible in the other trans-maps. This is a first hint that in these cases, translocations could have occurred. When looking at the detailed trans-map for the respective chromosomes, we see in each map two square like interaction patterns with a higher interaction frequency compared to the rest of the chromosome, indicating a balanced translocation. Each of these patterns represents a rearranged derivative chromosome (Figure 1) [27]. However, the complexity of rearrangements in Hi-C can be a lot higher. The Hi-C maps of two cases with a comparable phenotype (intellectual disability), which we analyzed in a different study [28], showed large differences when it comes to Hi-C map complexity. One of these cases (CP4) showed only low complexity rearrangements (here termed chromoplexy).

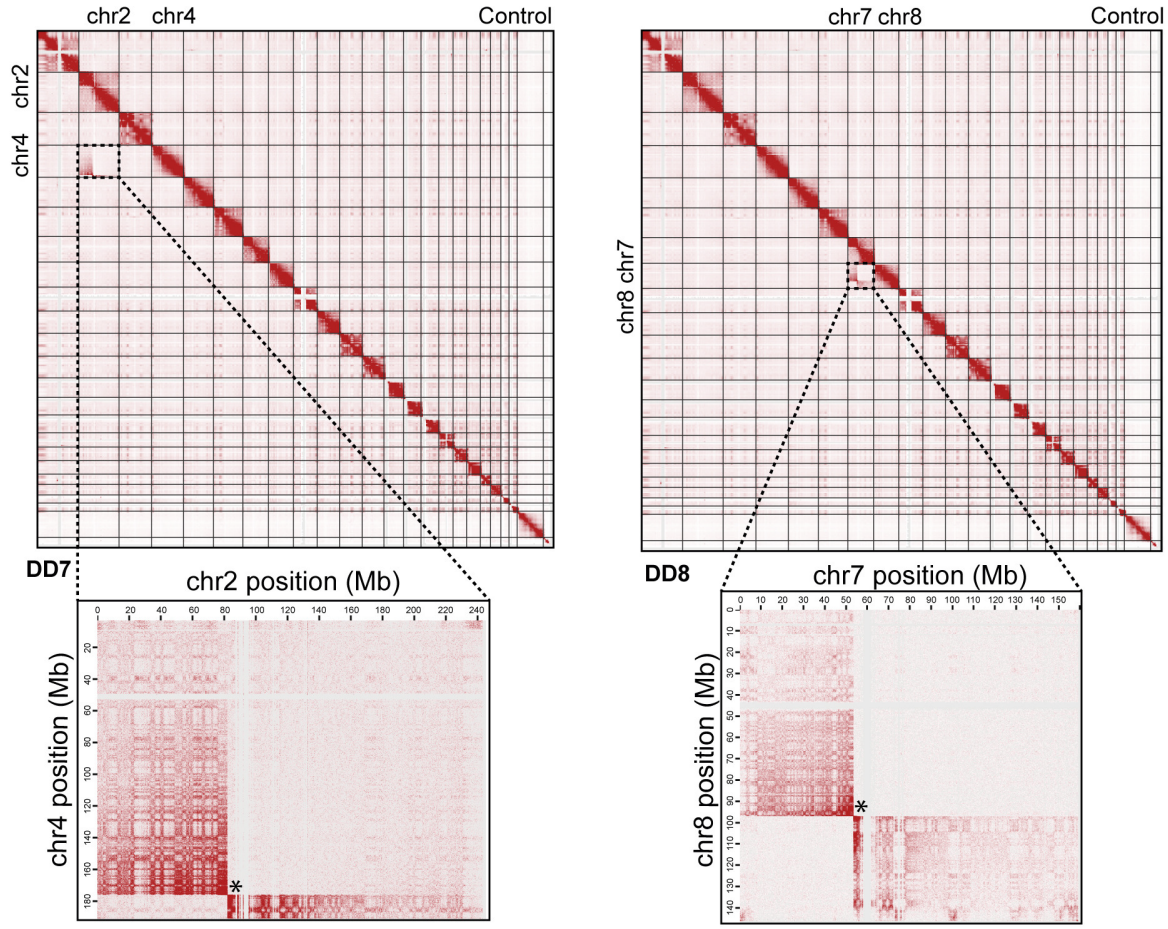

Figure 1: Detection of simple rearrangements in Hi-C. Shown here for two large scale germline translocations in individuals with developmental delay (DD7 and DD8). The translocations are in the upper part of the figure displayed in the “all chromosomes view” of Juicebox and below in detail in a trans-map, showing only the involved chromosomes. The breakend is the point of highest signal intensity, marked with an asterisk. Figure modified from a publication included in this work [27].

These included mostly translocation patterns in a chromosome 5 and 8 rearrangements (Figure 2a,c,e). The other case (CT3) showed a massive chromosome shattering as part of a chromothriptic rearrangement of chromosome 10 and 12 (Figure 2b). Many abnormal intensity patterns are seen in the chromosome 10/12 trans-map (Figure 2d) but also in the chromosome 12/12 cis-map (Figure 2f), representing multiple fragments of heavily fragmented and rearranged derivative chromosomes. The multiple interconnected fragments and the fragment sizes of sometimes <50 kb make it difficult to fully solve such rearrangements only based on Hi-C.

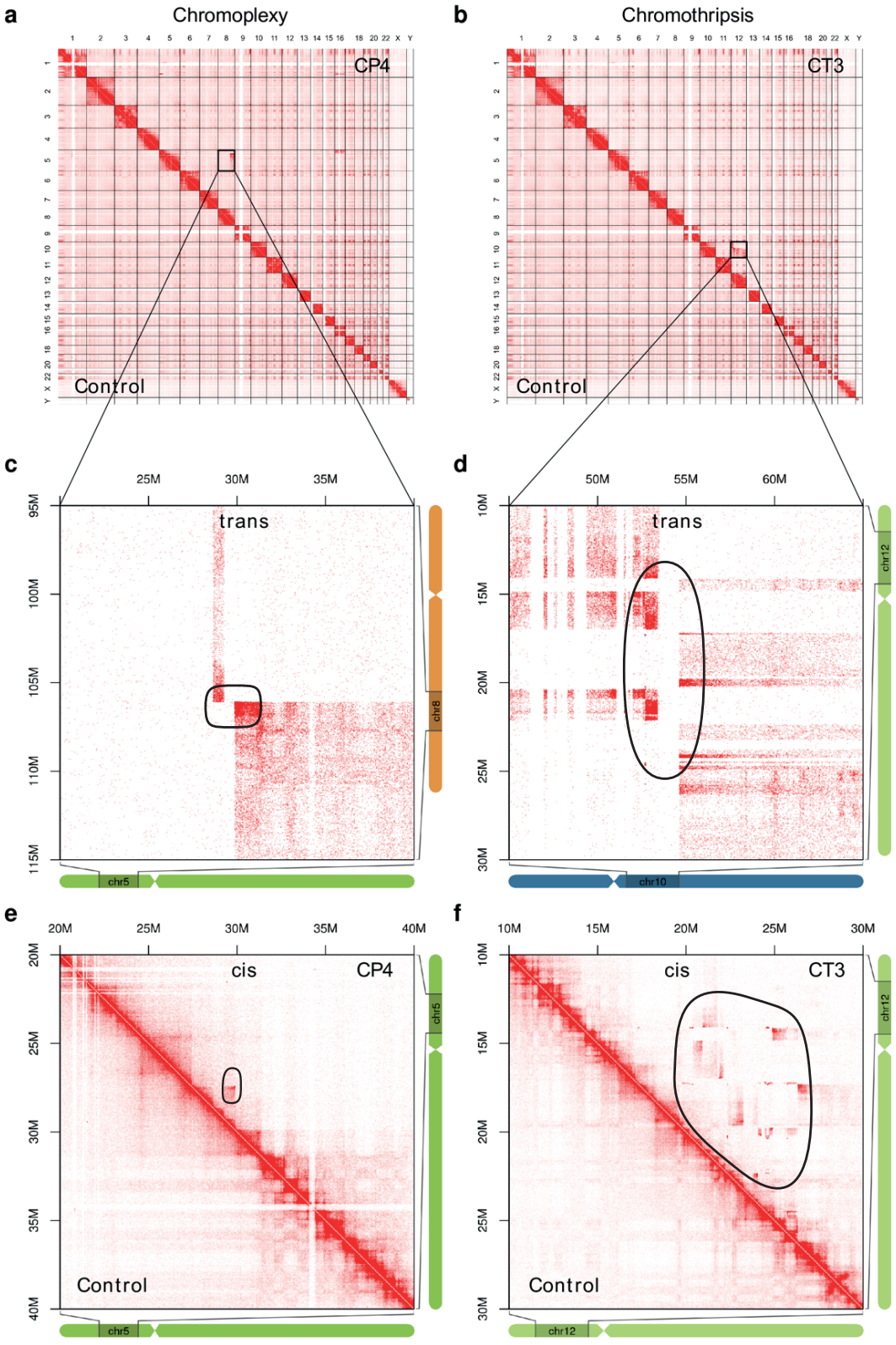

Figure 2: Different levels of complexity in genomic rearrangements analyzed with Hi-C. **a-f,** Low complexity (CP4) and chromothripsis cases (CT3) show different Hi-C patterns at different zoom levels. Genome-wide Hi-C maps (**a-b**) Zoom-in of trans Hi-C maps showing ectopic contacts between chr5 and chr8 (CP4), and between chr10 and chr12 (CT3), respectively (**c-d**). Zoom-in of cis Hi-C maps of chr5 (CP4) and chr12 (CT3). The chromothripsis case CT3 showing a chromothriptic rearrangement with breakends

occurring in clusters, leading to complex ectopic Hi-C patterns in cis (**f**) and trans (**d**). Breakend regions encircled in black. Figure and description modified from a publication included in this work [27].

Rearrangements showing comparable or even higher complexities could also be observed in Hi-C maps of the CK-AML cohort (Figure 3a, b) [26]. An important phenomenon that we observed when analyzing the Hi-C maps is the phenomenon of breakend like patterns. This phenomenon limits the analysis of Hi-C maps when you are looking at smaller fragments with a size of <100kb. If two loci are not directly linked via a breakend but show close proximity to each other due to e.g. a small fragment that is in the middle of these fragments and connected with both of them directly, the Hi-C signal of the connection of these fragments can appear as a breakend like pattern, even if these fragments are not connected. These patterns are difficult to distinguish from a real breakend, because the interaction frequency and thereby the Hi-C signal of two fragments that are only separated from each other by a few kb small in-between fragment, is almost the same as if they would be directly connected [26]. In our dataset, these breakend like–patterns were visually clearly different from real breakends, when the fragment size of the fragment that is located in-between was >100kb. For smaller fragments, integrating the data with further datasets is necessary in order to clearly distinguish breakends and breakend like patterns. We achieved this by integrating Hi-C with ONT long-read sequencing data [26].

3.2 Cohort overview of the CK-AML project

To characterize the genomic and transcriptomic features of 11 CK-AML cases included in this study, we used a combination of innovative technologies. Genomic sequencing data (Hi-C, ONT long-read WGS) was used to characterize the complex SVs in this cohort in detail. This data was then integrated with transcriptomic (ONT direct cDNA sequencing, Illumina RNA-Seq) sequencing data for fusion transcript detection and differential gene expression analysis to understand about the functional consequences of the complex SVs in CK-AML. To benchmark our findings, we performed Hi-C and Illumina RNA-seq in CD34+ hematopoietic stem cells of five healthy donors [26].

3.3 Structural variant detection by integrating Hi-C and long-read sequencing

As already discussed in the previous sections, we wanted to improve SV detection with Hi-C by integrating the results with SV data derived from long-read sequencing. For the detection of interchromosomal and intrachromosomal breakends, we integrated the Hi-C results with the ONT long read-sequencing SV calling dataset obtained with NanoVar [35]. To analyze CNVs, we developed our own workflow, which was based on the ACE tool [36], executed on our ONT long read-sequencing data. An overview of our whole SV calling workflow is shown in Figure 1a and Supplemental Figure 1a in the Klever et al. 2023 paper [26]. In the next sections, this workflow will be presented in detail.

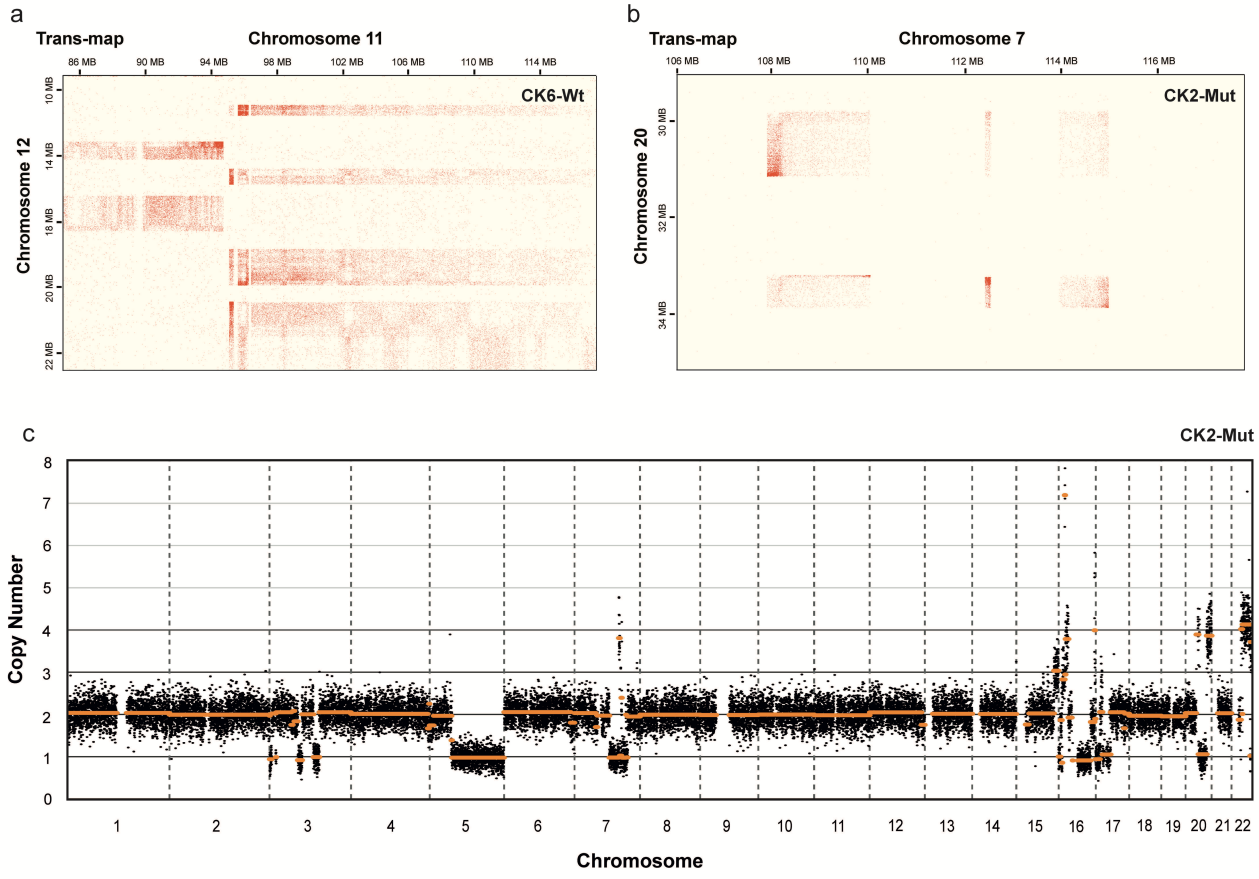

Figure 3: Extremely complex genomic rearrangements in Hi-C maps and copy number analysis with ACE. a-b, Hi-C trans- map of chromosome 11 and 12 showing a complex rearrangement in CK6-Wt (**a**), Hi-C trans-map of chromosome 7 and 20 showing a complex rearrangement in CK2-Mut (**b**). **c,** ACE results for CK2-Mut at 100kb binning size. CN state of the respective fragments is marked by orange bars. Figure created for this dissertation based on data from the Klever et al. 2023 paper [26].

3.4 Workflow for the detection of interchromosomal and intrachromosomal breakends

The first case, that we primarily analyzed with Hi-C and afterwards with our integrated Hi-C and long-read sequencing workflow, was CK1-Mut, which shall serve as an example to illustrate the resulting workflow. The Hi-C map of this sample showed a complex rearrangement involving chromosome 7 and chromosome 8. Visual inspection of trans- and cis-maps enabled us to identify 10 putative breakends in 8 breakend regions, which were not observed in the control sample (see Figure 1b,c in the Klever et al. 2023 paper) [26]. Some regions that only showed a breakend-like pattern could already be excluded based solely on Hi-C analysis, because they did not show a very intense signal pattern directly at the putative breakend and did not fit into a logical order of the fragments on the derivate chromosome. The application of Hi-C to this rearrangement enabled us to reconstruct the structure of the derivate chromosomes, including the breakend positions at kilobase resolution, which is the resolution limit of Hi-C analysis. When we started our analysis with first only applying Hi-C to the cohort, we already found a plethora of complex rearrangements, with fragments detectable at a much smaller size than what we expected.

In the next step, we analyzed the ONT long-read WGS data with NanoVar [35]. NanoVar was primarily executed with standard filtering criteria, without filtering the SV calls by a confidence score, to detect as many SVs as possible that we already detected by Hi-C. SV calls from the INV (intrachromosomal breakends) and BND (interchromosomal breakends) categories were extracted to integrate them with Hi-C. After mapping the NanoVar breakend calls on the Hi-C maps and comparing the results to the breakends detected with Hi-C, we were able to adjust the filtering criteria without excluding any true positives, in order to simplify the integration process. For the final dataset, a confidence score of 0.4 was used for filtering. The NanoVar breakends were identified as representing the same breakend as identified in Hi-C when both coordinates were less distant than 5 kb from the Hi-C breakend (resolution limit of Juicebox). To also identify breakends of small segments, which we may did not identify by visual inspection of the Hi-C maps alone, all NanoVar breakend calls within the range of 1 mb around all Hi-C breakends verified with Nanopore were re-analyzed in a secondary visual analysis. Here, we searched for additional Hi-C fragments that were too small to be detected in the

primary visual inspection but were detected by NanoVar. If an additional Hi-C fragment was detected in this secondary visual analysis based on a Nanovar breakend call, we started our workflow again, as if it would have been detected in the primary visual analysis. This procedure was repeated until no additional NanoVar SV calls with Hi-C support were identified. Finally, we only integrated breakend calls in our final SV call dataset, that were supported by Hi-C as well as by NanoVar breakend calls, except for a small fraction of fragments (less than 3%), which were enclosed on both sides of the fragment by fitting NanoVar breakend calls and were <10 kb in size but lacked Hi-C support. This enabled us to thoroughly exclude false positives from our dataset [26].

By applying this strategy to the case CK-1Mut, 8 out of 10 assumed Hi-C breakends could be verified by our integrative SV detection workflow. The two remaining ones were shown in the following to represent breakend-like patterns, even if their visual appearance in Hi-C was similar to the other identified Hi-C breakends. Searching the NanoVar breakend calls 1 mb around both breakends revealed connections of them to two previously unidentified small fragments, one of them (N1) 1.3 kb and the other one (N2) 5.2 kb in size (see Figure 1d in the Klever et al. 2023 paper) [26]. The fragment N2 was located in between the chromosome 7 fragment G and the chromosome 8 fragment F and linked these fragments, which were previously thought to be linked directly. Application of our SV detection workflow enabled us to completely reconstruct the derivate chromosome structure, with breakend detection at base pair resolution (see supplemental Figure 4 in the Klever et al. 2023 paper) [26].

3.5 Detection of Copy Number Variations with Hi-C and long-read sequencing

For the analysis of CNVs, we used the ACE tool executed on our ONT long-read WGS dataset [36]. When comparing the copy number (CN) estimations of the ACE tool with CNV data that could be inferred from Hi-C using the HiNT tool, these datasets showed a high visual accordance [19]. Even smaller fragments could be detected and were largely present in both datasets, demonstrating the robustness of these methods. However, likely due to the short Hi-C reads and alignment difficulties of these hybrid reads, the HiNT data appeared more noisy, especially around low complexity regions like the centromere (Figure 3c; see Supplemental Figure 3 in the Klever et al. 2023 paper) [26]. Therefore, we continued with using the ACE tool, which we could cross-validate with HiNT, for

generating our final CN dataset. To refine the CNV analysis by ACE, we compared the local TDF file genomic coverage information around the identified breakends and searched for smaller fragments <10-20 kb in size in IGV [37]. These fragments were too small for CNV analysis with ACE and were therefore manually identified. Their CN state was calculated by comparing the local coverage of these fragments to the coverage of known larger fragments in close proximity with a known copy number state based on ACE [26].

3.6 Overview of the structural variant detection workflow results

We subsequently applied our SV analysis workflow, which included the detection of interchromosomal and intrachromosomal breakends as well as the detection of CN gains and CN losses, to our whole cohort of 11 CK-AML cases. We were able to display the CN state of fragments of various sizes including small fragments <1 kb, and their connections to other fragments (see Supplemental Figure 6 and 7 in the Klever et al. 2023 paper) [26]. A detailed analysis of the of the CN state distribution showed that in all cases fragments with CN gain (CN >2.3) as well as CN loss (CN <1.7) were present. The distribution of the fragments varied substantially. In CK2-Mut, CK4-Mut, CK6-Wt, CK8-Wt and CK9-Wt, amplified fragments with a CN of ≥5 were detected. The four cases with the most detected fragments with a distinct CN state (CK2-Mut, CK3-Mut, CK4-Mut, CK6-Wt), i.e., the most complex cases in terms of CNV, were also the most complex cases in terms of the number of inter- and intrachromosomal breakends. The most complex of all cases was CK2-Mut, presenting with 191 fragments with a distinct CN state and 123 inter- and intrachromosomal breakends (Figure 4a,b; see Supplemental Figure 8a in the Klever et al. 2023 paper) [26].

CK2-Mut and CK3-Mut were examined in a previous study by CN analysis using microarray data [13]. This dataset showed much fewer copy number losses and gains in the respective cases, compared to our high-resolution approach (see Supplemental Table 2 in the Klever et al. 2023 paper) [26]. The distribution of inter- and intrachromosomal breakends in specific chromosomes and cytogenetic bands in the whole CK-AML cohort showed breakends recurrently hitting chromosome 7 (N=6), chromosome 19 (N=5), chromosome 12 (N=5) and chromosomes 5, 8, 11, 17 (N=3). Furthermore, specific cytogenetic bands on chr7 and chr17 were recurrently affected by breakends in three cases (7q11.21; 7q21.3; 7q22.1; 17p11.2; 17p12) [26].

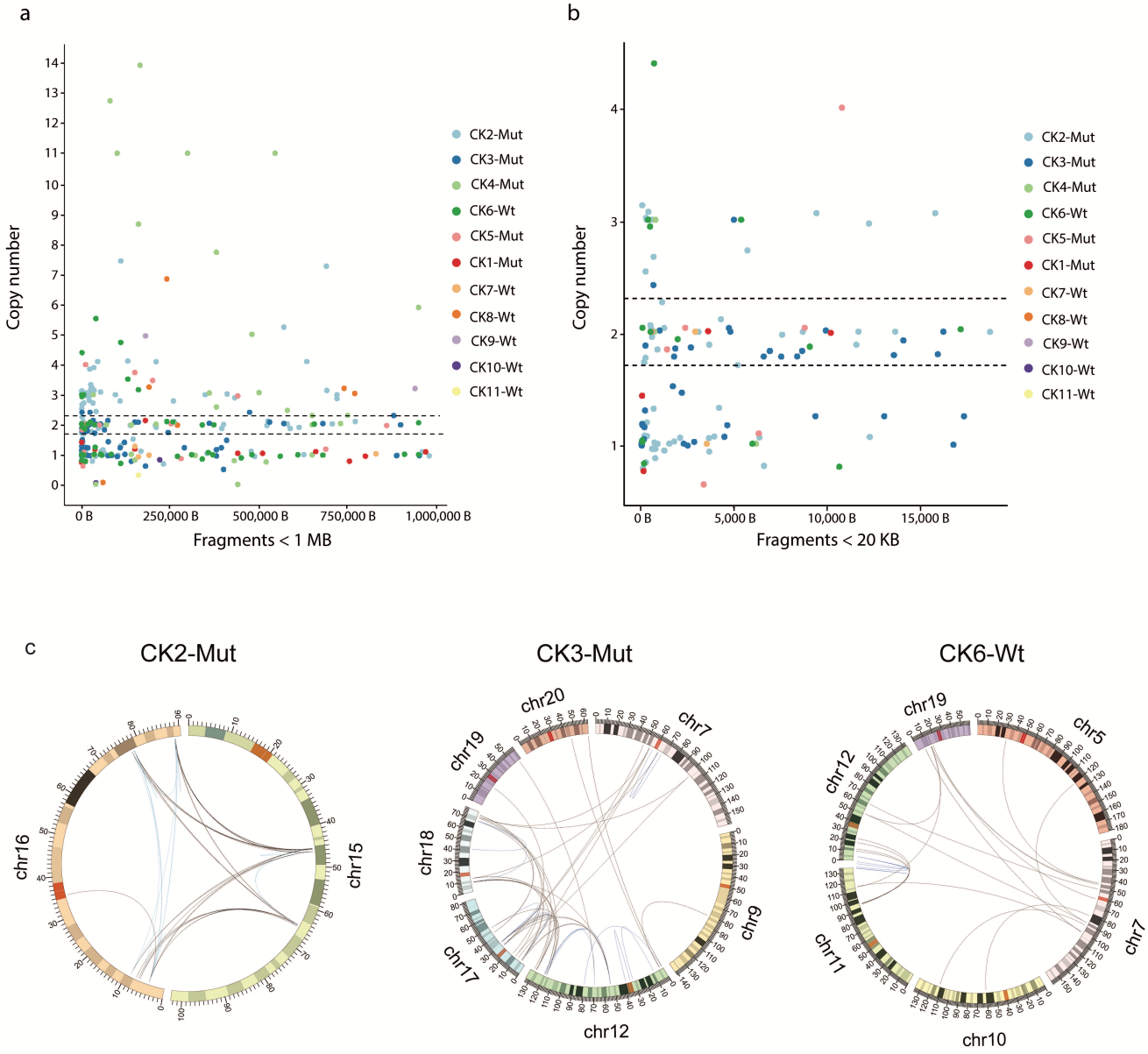

Figure 4: Size- and CN state distribution of fragments and chromocataclysm rearrangements in the CK-AML cohort. a-b, Scatterplots showing the distribution of fragment sizes of the different cases with the CN on the y axis and fragment size on the x axis, showing fragments < 1 mb in size (**a**) and fragments < 20 kb in size (**b**). Each dot represents one fragment (distinct region on a genome of reference) and its respective CN. Case ID coded with colored dots. Dashed black lines showing the limits of the CN stable fragment definition (1.7 ≤ CN fragment ≥ 2.3). **c,** Circos plot of all chromocataclysm rearrangements in this cohort (CK2-Mut, CK3-Mut and CK6-Wt). Intrachromosomal breakends are shown as blue lines and interchromosomal rearrangements as black lines. Numbers around the circos plots = genomic position in megabases. chr = chromosome. Figure created for this dissertation based on data from the Klever et al. 2023 paper [26].

3.7 Genomic differences in *TP53* mutated vs. *TP53* wildtype cases

One of the most striking features, that we already observed in the Hi-C analysis were the different levels of rearrangement complexity in *TP53* mutated CK-AML cases (CK1-Mut to CK5-Mut) compared to the cases that were *TP53* wildtype (CK6-Wt to CK11-Wt). All *TP53* mutated cases showed chromothripsis, which was here defined as the presence of 10 or more CN changes on a single chromosome, whereas all the *TP53* wildtype cases, but one (CK6-Wt, a case with chromothripsis), showed only simple rearrangements without chromothripsis. When looking at the distribution of CN fragments, the *TP53* mutated cases showed a larger number of fragments and a wider distribution of CN states, further underlining the much higher complexity of the *TP53* mutated cases (see Figure 3a and supplemental Figure 8a in the Klever et al. 2023 paper) [26].

3.8 Detection of a novel phenomenon in complex rearrangements - Chromocataclysm

As already discussed in the previous sections, the complexity of rearrangements showed large variations in cases that we present here. When looking at the rearrangements in detail, we detected a novel phenomenon of highly clustering breakends and accompanying amplifications. This pattern was present in three (CK2-Mut, CK3-Mut, CK-6Wt) out of six chromothripsis cases. It consisted mostly of multiple highly interconnected small fragments, kilobases and often only sub-kilobases in size, that were often amplified. These events occurred inside of larger chromothriptic rearrangements consisting of larger interconnected fragments in the range of multiple kilobases. Chromocataclysm was here defined as 4 or more breakends in a region of 5 kb. Chromocataclysm was detected in a chromosome 15 and 16 rearrangement in CK2-Mut as well as a rearrangement of chromosomes 7, 9, 12, 17, 18, 19 and 20 in CK3-Mut and a chromosome 5, 7, 10, 11, 12 and 19 rearrangement in CK6-Wt (Figure 4c). Interestingly, an additional rearrangement of chromosome 3, 7, 20 and 22 was detected in CK2-Mut, which was not connected to the chromosome 15 and 16 rearrangement and showed no chromocataclysm. These events could have occurred independently. 107 of all 122 fragments with a size of <20 kb in our cohort were present in the three chromocataclysm cases (87.7%). Interestingly, CK4-Mut, the chromothripsis case with the third highest number of total CN fragments, harbored only three fragments <20 kb in size. No chromocataclysm pattern was present

in this case [26]. Clustering of breakends in the chromocataclysm cases could not only be observed when looking at whole chromosomes, examining the fine structure of these breakends when looking at kilobase size regions of the chromosomes revealed extreme local clustering. In CK2-Mut, the complexity reached 31 breakends hitting a region only 2.7 kilobases in size (see Figure 2 in the Klever et al. 2023 paper) [26].

With our workflow, we were able to detect highly clustered fragments as well as their connections to other fragments, even when the fragment size was <1 kb. As an example, we detected one amplified fragment that was just 297 bp large and consisted of multiple sub fragments and accompanying CN changes inside the chromosome 15 and 16 chromocataclysm rearrangement in CK2-Mut. This extremely small fragment including the sub fragments was associated with 16 inter- and intrachromosomal breakends. The CN elevation of this fragment against the surrounding regions was clearly visible in Hi-C as well as in the ONT long-read WGS data. It was found to be connected to many other small fragments in highly clustered regions, which often showed an elevated CN state. Many of the connecting fragments were also located in regions of high local breakend clustering (see Supplemental Figure 6 in the Klever et al. 2023 paper) [26].

Identifying such small fragments would not have been possible by only using Hi-C based SV detection. These fragments are extremely hard to detect at primary visual analysis in Hi-C and it would be impossible to examine where the breakends are at basepair level. At this small size, breakend-like patterns are indistinguishable from the real ones. To exemplary validate the findings of this strong local clustering of breakends in the chromocataclysm case CK2-Mut, we calculated the local density of inter- and intrachromosomal breakends from NanoVar. We detected there the main highly clustered breakend regions on chromosome 15 of CK2-Mut around 45 mb and another region at around 65 mb by using our integrated SV detection workflow. These regions could be identified as peaks in the breakend density plots of chromosome 15. Interestingly, it is visible that breakend clustering is preserved when looking at the rearrangements in a detailed view of only kb to sub-kb sized regions (Figure 5a) [26].

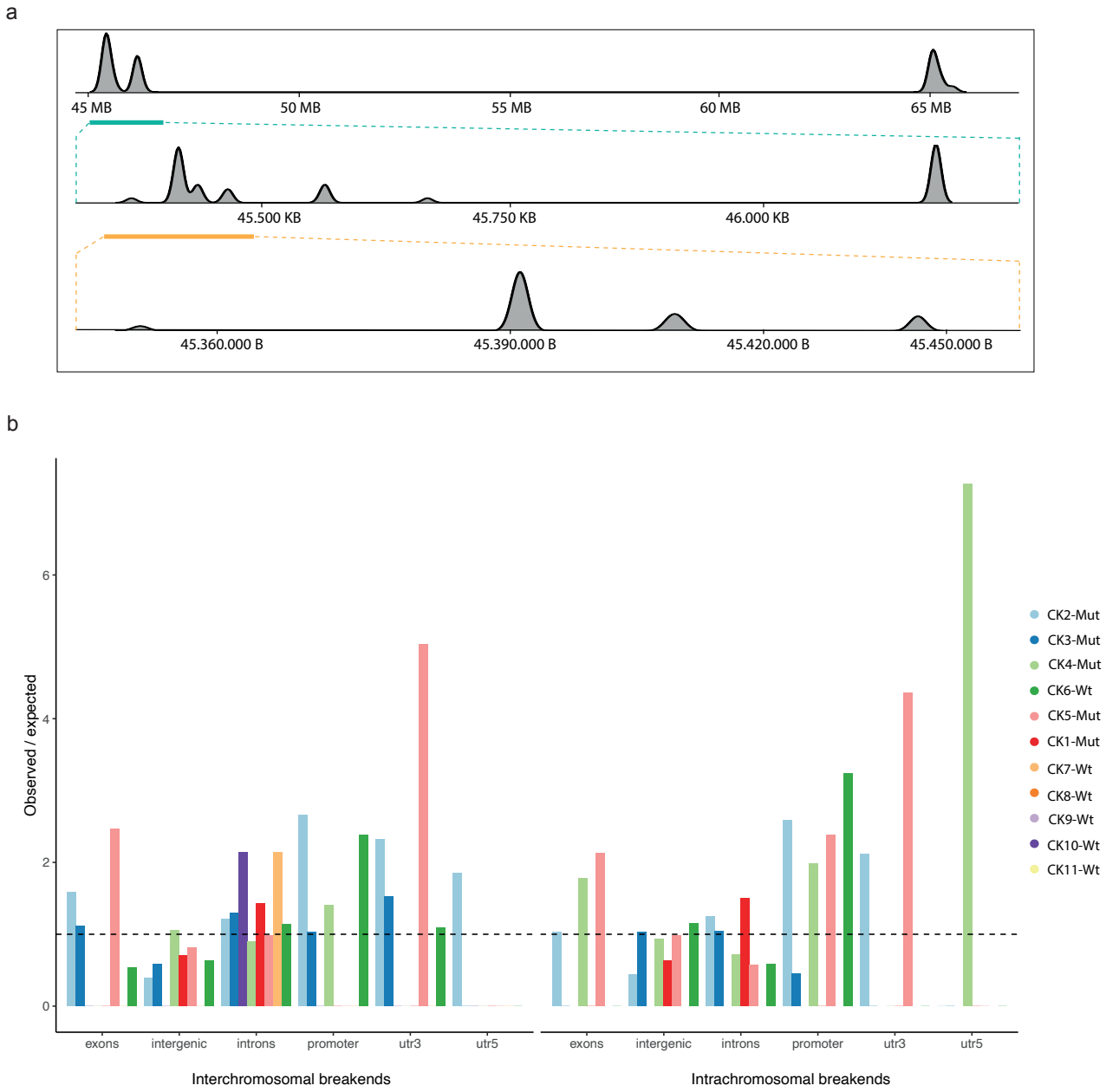

Figure 5: Breakend density plot in chromocataclysm and distribution in genomic features. a, Breakend density plot (inter- and intrachromosomal) of chromosome 15 in CK2-Mut, where this chromosome was affected by a chromocataclysm rearrangement. Regions with a high density of local breakend clustering are visible as peaks. The middle panel shows an enlarged breakend dense region of the top panel and the lowest panel an enlarged dense region of the middle panel. The enlarged regions are visualized by colored lines in green and orange. **b,** Bar plot showing the observed/expected ratios of breakends in different genomic features for all cases in the CK-AML cohort, interchromosomal breakends on the left and intrachromosomal breakends on the right. Utr3 = 3′ untranslated region. Utr5 = 5′ untranslated region. Figure created for this dissertation based on data from the Klever et al. 2023 paper [26].

3.9 Enrichment of breakends in repetitive regions and other genomic features

The phenomenon of chromocataclysm and the related high breakend clustering as well as the fact that some chromosomes and cytogenetic bands were hit recurrently by breakends, as presented previously, led us to investigate if some chromosome regions in CK-AML are may somehow prone for being hit by breakends and how the breakends are generally distributed in the genome. We primarily calculated the occurrence of breakends in relation to the main features structuring the genome (promoter regions, 3' untranslated regions (UTR), introns, exons, 5' UTR, intergenic regions). The relative occurrence was investigated against 10000 random breakends. Breakends were overrepresented inside gene promoters (p-value <0.001), introns (p-value <0.001), and an underrepresentation in intergenic regions (p-value <0.001) (Figure 5b; see Figure 3 in the Klever et al. 2023 paper) [26]. An overrepresentation in these very crucial features for gene function further suggests a functional role of the found breakends in our cohort of CK-AML cases.

We furthermore analyzed the distribution of inter- and intrachromosomal breakends from our dataset in relation to the genomic occurrence of repetitive elements. One prominent finding was that when looking at repeat categories, breakends occurred more often in MIR repeats, which represent about 2.5% of the genome, as expected by chance. 5.6% of all breakends of the three chromocataclysm cases occurred in MIR repeats, as well 4.7% of all breakends of the three chromothripsis cases that were not classified as chromocataclysm, indicating an overrepresentation by about 2x fold [26]. MIR elements are themselves a subgroup of short interspersed nuclear elements and can be further divided into 4 subcategories, MIR, MIRb, MIRc and MIR3 [38]. To further validate the overrepresentation of breakends inside of MIR repeats and in close proximity to MIR repeats, we calculated their distribution against 10000 random breakends. In this analysis we observed an overrepresentation of breakends directly inside of the MIR subcategory (p-value =0.002). Interestingly, we also found an overrepresentation in close proximity to the MIR subcategory elements (average distance 15 kbp, 95% confidence interval: 10.5 kbp -21.6 kbp, p-value =0.004). This enrichment in close proximity was most prominent for the chromocataclysm cases (see Figure 3d,e in the Klever et al. 2023 paper) [26]. It was strongest in case CK2-Mut, the most complex of all chromocataclysm and generally of all cases reported here (data not shown).

3.10 Functional studies based on SV data and transcriptome sequencing

After characterizing the whole CK-AML cohort with our integrated SV detection workflow, the comprehensive SV dataset enabled us to study the functional role of structural variants in CK-AML in detail. With this, we wanted to better understand the impact of the complex genomic rearrangements in CK-AML on gene expression and the emergence of fusion transcripts in detail. For all cases except CK1-Mut and CK11-Wt, enough sample material was available to perform ONT cDNA and Illumina RNA sequencing.

3.11 Complex genomic rearrangements result in novel fusion transcripts

We developed a workflow for fusion transcript detection that is based on multiple transcriptomic and genomic datasets, to detect fusion transcripts with high confidence and eliminate false-positive fusion calls. We used the ONT version of the fusion caller JAFFA [32] on our ONT direct cDNA sequencing data and the direct mode of JAFFA on our Illumina RNA sequencing data. In the nine CK-AML cases with available transcriptomic data, JAFFA detected a total of 271 candidate fusion transcripts in the ONT direct cDNA sequencing data and 6491 candidate fusion transcripts in the RNA sequencing data. Of these fusion transcripts, 115 were identical in both datasets. This dataset was compared to the genomic inter- and intrachromosomal breakends from the previous analyses. This was done to find fusions, which are supported by transcriptomic fusion transcripts present in two datasets as well as two matching genomic breakends, to further eliminate false positives. To also account for intronic breakends that are distant from the exon-exon fusions which are present in the fusion transcript datasets from JAFFA, we set a cutoff off <30 kb distance of both genomic SV breakends to both exonic “fusion points” of the fusion transcript. With this strategy, we identified a total of 5 high confidence fusion transcripts, which were all supported by two transcriptomic and one genomic dataset (see Figure 4a in the Klever et al. 2023 paper) [26]. Interestingly, all fusion transcripts that were in this distance range of <30 kb to genomic breakends, were anyways present in both transcriptomic datasets, the ONT direct cDNA as well as in the Illumina RNA sequencing fusion transcript dataset. One of the fusion transcripts that was supported by evidence from these three datasets was a known recurrent fusion event in leukemia, a chromosome 11 and 19 translocation leading to a fusion of *KMT2A* (*MLL*) and *MLLT1* in case CK8-Wt [39]. One of the novel fusion transcripts that we discovered

here was the fusion of *USP7* and *MVD* on chromosome 16 in the chromocataclysm case CK2-Mut. *USP7* is linked to *TP53* metabolism, it is well known to play a role in several cancers including AML (see Figure 4b and supplementary results in the Klever et al. 2023 paper) [26]. In this case, the first exon including the transcription start site of the truncated *USP7* was inserted close to the TSS of *MVD*. The fusion transcript, which contains a fusion of exon 1 of *USP7* and exon 2 of *MVD*, was likely generated due to a start of transcription at the TSS of *USP7*, and a subsequent splice-out of exon 1 of *MVD*. Other novel fusion transcripts that were supported by evidence from all datasets were a fusion of *ARGHAP44* to the antisense transcript *AC087294.2* in CK2-Mut, a fusion of *ANKRD12* to *NUP88* in CK3-Mut and a fusion of *MTMR2* and *PRB1* in CK6-Wt. Many of the involved genes were already linked to cancer biology (see supplementary results in the Klever et al. 2023 paper) [26].

3.12 Identification of CK-AML associated gene expression patterns

We analyzed the dysregulation of genes performing a differential gene expression analysis in the Illumina RNA sequencing dataset. This analysis was performed for individual CK-AML patients against the combined CD34+ controls. The results of this analysis were integrated with the CN data and the inter- and intrachromosomal breakend lists from our integrated SV detection dataset. We searched for gene expression signatures in regions affected by SVs that were shared by some of the CK-AML cases discussed here. The whole analysis was done in order to understand more about the pathophysiology of the complex structural variants. We started our analysis by conducting an Over-Representation Analysis for over- or underrepresented gene ontology (GO) terms from the “biological processes” category in our CK-AML cohort compared to the CD34+ hematopoietic stem cell controls. By conducting this analysis, we were able to detect if some of the differentially expressed genes may relate to the same biological processes and functions and indicating thereby different expression signatures. Commonly up-regulated genes (up-regulated in at least 6 out of 9 patients) were enriched in GO terms related to leukocyte biology, other processes of the immune system and general cell functions. Commonly down-regulated genes were only enriched in GO terms of fatty acid biosynthetic process and blood circulation (see supplemental Figure 9 in the Klever et al. 2023 paper) [26]. A large proportion of this result can likely be attributed to the different cell states of leukemic cells and hematopoietic stem cells. In the further

analysis, we analyzed genes that were affected by CN changes and gene dysregulation. We chose genes showing a CN gain and upregulation (N=25) or a CN loss and downregulation (N=194) in ≥4 cases as well as disrupted genes due to breakends in promoter or gene body that were downregulated for further analysis. Especially interesting for us were genes, that have already been linked to cancer or AML biology in the literature. Based on this strategy, we obtained a final list of cancer related candidate genes (see supplemental dataset 4 of the Klever et al. 2023 paper) [26].

Interestingly, genes that were affected by gene disruption and downregulation, were in some cases also affected by CN losses at the gene locus that were accompanied by gene downregulation in other cases. Most notably, *IMMP2L* was disrupted and downregulated in one case and showed CN loss and gene downregulation in 4 additional cases. *IMMP2L* is mitochondrial membrane associated gene that was shown to inhibit apoptosis when its signaling is switched off in a cellular assay [40]. Interestingly, all CN gain and upregulation candidate genes from the candidate gene list were located on the q arm of chromosome 8, a chromosome arm that is known to be repeatedly affected by CN gains in CK-AML. On the other hand, all candidate genes from the CN loss and downregulation category were on chromosome arms 7q, 12p, 16q, 17p and 18q, all representing known regions repeatedly affected by CN losses in CK-AML [41]. When examining all candidate genes, CN loss as well as gene downregulation occurred almost exclusively in cases that we previously identified as chromocataclysm and/or chromothripsis (see Figure 5, supplemental Figure 9 of the Klever et al. 2023 paper) [26]. In the following, we compared our CN loss and gene downregulation candidate genes to data from The Cancer Genome Atlas (TCGA) [42]. This analysis showed that 20 out of 30 candidate genes with CN loss and downregulation were also linked to CN loss in AML cases of the TCGA dataset. These TCGA AML cases, which showed CN losses, were to a high proportion actually CK-AML cases. 7 out of 8 CN gain and upregulation candidate were linked to CN gain in the TCGA dataset. However, all these amplifications were reported in the same patient, whose leukemia was not classified as CK-AML. 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 and is associated with chromothripsis, but further studies are warranted to elucidate the role of each of these genes in leukemia biology.

4. Discussion

4.1 SV detection result summary and interpretation

In this work, we analyzed samples of patients with germline genetic disorders as well as leukemia patients. The common interest was to understand more about complex genomic rearrangements occurring in these cases. We therefore integrated two different SV detection methods, Hi-C and ONT long-read WGS. We applied the developed workflow to a cohort of 11 CK-AML cases. Hereby, we were able to thoroughly eliminate false-positive SVs from both datasets and detect genomic aberrations in CK-AML with a high accuracy and resolution. With this workflow, we were able to detect extremely complex rearrangements including highly clustered breakends in very small genomic regions. The fragments inside these rearrangements were often a few kilobases and even sub-kilobases in size and were amplified, which is not a regular feature in classical chromothripsis [14,15]. Some cases showed a complexity of rearrangements and harbored features that lie beyond the current knowledge about complex genomic rearrangement in CK-AML in general and chromothripsis in particular. We named this new phenomenon, which seems to occur inside of comparably less complex chromothripsis events, chromocataclysm. The local clustering of the chromocataclysm rearrangements seemed to be much higher and the fragments that these rearrangements consist of much smaller than what could be detected so far. In a previous study about chromothripsis, a minimum CN segment size of 10 kb was used to detect chromothripsis [43].

One of the most striking differences to the original Korbel and Campbell criteria to assess chromothripsis [15], was the presence of small amplifications, most prominently in the chromocataclysm cases CK2-Mut and CK6-Wt, but also pronouncedly in the form of larger size amplifications in CK4-Mut, a case not classified as chromocataclysm. While chromothripsis is currently thought to be a single event that only occurs once in the development of a cell, we found in CK2-Mut some evidence for two chromothripsis events that were possibly not linked to each other. One of the rearrangements that was discussed here in detail (chromosome 15 and 16) showed the highest level of BND clustering and was one of the rearrangements that was classified here as chromocataclysm. Interestingly, another complex rearrangement in CK2-Mut, involving chromosome 3, 7, 20 and 22, showed a much lower level of local breakend clustering.

Furthermore, these two events were not physically linked to each other. The fact that small amplified fragments were not only connected to other complex fragments in chromocataclysm, but also showed connections to large genomic regions, providing evidence that such small amplified and rearranged regions exist not only as double minute-like structures, but can also be retained in the genome as part of highly clustered chromocataclysm events.

Some features of chromocataclysm seem to be similar to features of other categories of complex rearrangements e.g. chromoplexy and chromoanasynthesis [44]. Similar to chromocataclysm, in chromoanasynthesis, multiple chromosomes are involved and amplified fragments can be detected. In general, the extremely complex and distinctive features of the rearrangements that we discovered in this cohort with our high-resolution SV detection workflow, leads to the hypothesis that these categories may are not fully distinguishable from each other, but represent a continuum of features in complex rearrangements. Breakends of the chromocataclysm cases were enriched in the proximity of MIR elements. MIR elements are associated with open chromatin regions and could serve enhancer functions [45]. MIR elements functioning as enhancers in close spatial proximity to the breakends could cause a local gene expression dysregulation due to promoter-enhancer interactions. This could point towards a functional role of MIR elements in chromocataclysm.

4.2 Transcriptome analysis result summary and interpretation

Our fusion transcript detection workflow required support of ONT direct cDNA sequencing data, Illumina RNA-Seq data as well as corresponding genomic breakends. This enabled us to call fusion genes with a high confidence. One of the most prominent findings was the identification of a *USP7::MVD* fusion transcript in CK2-Mut. This fusion mechanism seems to be a combination of two fusion mechanisms. Firstly, the mechanism of generating fusion transcripts due to intronic SV breakends and secondly a read through mechanism, where a fusion transcript is formed because the transcription termination site of a gene is missed by the RNA polymerase and exon 1 of the next gene is spliced out. By conducting a differential gene expression analysis on our dataset and integrating the results with the SV dataset generated previously, we were able to identify 38 cancer associated candidate genes. Interestingly, these candidate genes were exclusively located in regions that are known to be affected by CN changes in CK-AML. 28 of 30

gene downregulation and CN loss candidate genes were dysregulated in the six chromothripsis cases. This suggests a chromothripsis related role of these genes and can may help to understand more about the role of chromothripsis in AML. Taken together with the breakend enrichment in promoters and examples of cancer related genes like *IMMP2L*, this points towards that a pathogenic mechanism of complex SVs is the functional inactivation of tumor suppressor genes in regions of breakend clustering.

4.3 Strengths and weaknesses of the work

By integrating two very different genome sequencing methods that are unlikely suffer from the same bias, we developed an SV detection workflow which detects CNVs and breakends at high resolution and enabled us to discover novel features of extremely complex genomic rearrangements. The usage of two different methods helped us to thoroughly eliminate false positives results from our dataset. Based on this dataset, we analyzed the distribution of breakends in the genome and performed fusion transcript detection and differential gene expression analyses. However, due to the low number of cases in this cohort, especially the results of the functional studies need to be validated and further mechanistical studies are needed in larger cohorts.

4.4 Implications for future research

The developed workflow is broadly applicable to solve complex rearrangements in cancer as well as other diseases. In cancer, complex SVs are often linked to a poor prognosis. To develop targeted therapies in the future, it is very important to understand the pathophysiological consequences of these SVs to identify druggable targets. An important but complicated task will be to partly automatize complex SV detection. This is important because the current high resolution analysis methods are still very laborious and not yet ready to be used in a standardized clinical setting. The workflow that we developed enables the detection of complex SVs at high resolution and with high confidence. However, further functional studies on the pathophysiological consequences of candidate gene dysregulation and the emergence of fusion transcripts are needed to understand their role in CK-AML pathogenesis.

Reference List

1. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020 Mar;21(3):171–89.

2. Wang WJ, Li LY, Cui JW. Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis. Epigenetics Chromatin. 2020 Dec;13(1):49.

3. Macintyre G, Ylstra B, Brenton JD. Sequencing Structural Variants in Cancer for Precision Therapeutics. Trends Genet. 2016 Sep;32(9):530–42.

4. Stadler ZK, Thom P, Robson ME, Weitzel JN, Kauff ND, Hurley KE, Devlin V, Gold B, Klein RJ, Offit K. Genome-Wide Association Studies of Cancer. J Clin Oncol. 2010 Sep 20;28(27):4255–67.

5. 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 Dec;20(1):246.

6. Kurzrock R, Kantarjian HM, Druker BJ, Talpaz M. Philadelphia ChromosomePositive Leukemias: From Basic Mechanisms to Molecular Therapeutics. Ann Intern Med. 2003 May 20;138(10):819.

7. Thein MS, Ershler WB, Jemal A, Yates JW, Baer MR. Outcome of older patients with acute myeloid leukemia: An Analysis of SEER Data Over 3 Decades. Cancer. 2013 Aug 1;119(15):2720–7.

8. Stölzel F, Mohr B, Kramer M, Oelschlägel U, Bochtler T, Berdel WE, Kaufmann M, Baldus CD, Schäfer-Eckart K, Stuhlmann R, Einsele H, Krause SW, Serve H, Hänel M, Herbst R, Neubauer A, Sohlbach K, Mayer J, Middeke JM, Platzbecker U, Schaich M, Krämer A, Röllig C, Schetelig J, Bornhäuser M, Ehninger G. Karyotype complexity and prognosis in acute myeloid leukemia. Blood Cancer J. 2016 Jan 15;6(1):e386–e386.

9. Döhner H, Wei AH, Appelbaum FR, Craddock C, DiNardo CD, Dombret H, Ebert BL, Fenaux P, Godley LA, Hasserjian RP, Larson RA, Levine RL, Miyazaki Y, Niederwieser D, Ossenkoppele G, Röllig C, Sierra J, Stein EM, Tallman MS, Tien HF,

Wang J, Wierzbowska A, Löwenberg B. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. Blood. 2022 Sep 22;140(12):1345–77.

10. Khoury JD, Solary E, Abla O, Akkari Y, Alaggio R, Apperley JF, Bejar R, Berti E, Busque L, Chan JKC, Chen W, Chen X, Chng WJ, Choi JK, Colmenero I, Coupland SE, Cross NCP, De Jong D, Elghetany MT, Takahashi E, Emile JF, Ferry J, Fogelstrand L, Fontenay M, Germing U, Gujral S, Haferlach T, Harrison C, Hodge JC, Hu S, Jansen JH, Kanagal-Shamanna R, Kantarjian HM, Kratz CP, Li XQ, Lim MS, Loeb K, Loghavi S, Marcogliese A, Meshinchi S, Michaels P, Naresh KN, Natkunam Y, Nejati R, Ott G, Padron E, Patel KP, Patkar N, Picarsic J, Platzbecker U, Roberts I, Schuh A, Sewell W, Siebert R, Tembhare P, Tyner J, Verstovsek S, Wang W, Wood B, Xiao W, Yeung C, Hochhaus A. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. Leukemia. 2022 Jul;36(7):1703–19.

11. Arber DA, Orazi A, Hasserjian RP, Borowitz MJ, Calvo KR, Kvasnicka HM, Wang SA, Bagg A, Barbui T, Branford S, Bueso-Ramos CE, Cortes JE, Dal Cin P, DiNardo CD, Dombret H, Duncavage EJ, Ebert BL, Estey EH, Facchetti F, Foucar K, Gangat N, Gianelli U, Godley LA, Gökbuget N, Gotlib J, Hellström-Lindberg E, Hobbs GS, Hoffman R, Jabbour EJ, Kiladjian JJ, Larson RA, Le Beau MM, Loh MLC, Löwenberg B, Macintyre E, Malcovati L, Mullighan CG, Niemeyer C, Odenike OM, Ogawa S, Orfao A, Papaemmanuil E, Passamonti F, Porkka K, Pui CH, Radich JP, Reiter A, Rozman M, Rudelius M, Savona MR, Schiffer CA, Schmitt-Graeff A, Shimamura A, Sierra J, Stock WA, Stone RM, Tallman MS, Thiele J, Tien HF, Tzankov A, Vannucchi AM, Vyas P, Wei AH, Weinberg OK, Wierzbowska A, Cazzola M, Döhner H, Tefferi A. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. Blood. 2022 Sep 15;140(11):1200–28.

12. Donehower LA, Soussi T, Korkut A, Liu Y, Schultz A, Cardenas M, Li X, Babur O, Hsu TK, Lichtarge O, Weinstein JN, Akbani R, Wheeler DA. Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. Cell Rep. 2019 Jul;28(5):1370-1384.e5.

13. Rücker FG, Schlenk RF, Bullinger L, Kayser S, Teleanu V, Kett H, Habdank M, Kugler CM, Holzmann K, Gaidzik VI, Paschka P, Held G, von Lilienfeld-Toal M, Lübbert M, Fröhling S, Zenz T, Krauter J, Schlegelberger B, Ganser A, Lichter P, Döhner K, Döhner H. TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. Blood. 2012 Mar 1;119(9):2114–21.

14. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, Futreal PA, Campbell PJ. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell. 2011 Jan;144(1):27–40.

15. Korbel JO, Campbell PJ. Criteria for Inference of Chromothripsis in Cancer Genomes. Cell. 2013 Mar;152(6):1226–36.

16. Rücker FG, Dolnik A, Blätte TJ, Teleanu V, Ernst A, Thol F, Heuser M, Ganser A, Döhner H, Döhner K, Bullinger L. Chromothripsis is linked to *TP53* alteration, cell cycle impairment, and dismal outcome in acute myeloid leukemia with complex karyotype. Haematologica. 2018 Jan;103(1):e17–20.

17. Cortés-Ciriano I, Lee JJK, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang CZ, Pellman DS, PCAWG Structural Variation Working Group, Akdemir KC, Alvarez EG, Baez-Ortega A, Beroukhim R, Boutros PC, Bowtell DDL, Brors B, Burns KH, Campbell PJ, Chan K, Chen K, Cortés-Ciriano I, Dueso-Barroso A, Dunford AJ, Edwards PA, Estivill X, Etemadmoghadam D, Feuerbach L, Fink JL, Frenkel-Morgenstern M, Garsed DW, Gerstein M, Gordenin DA, Haan D, Haber JE, Hess JM, Hutter B, Imielinski M, Jones DTW, Ju YS, Kazanov MD, Klimczak LJ, Koh Y, Korbel JO, Kumar K, Lee EA, Lee JJK, Li Y, Lynch AG, Macintyre G, Markowetz F, Martincorena I, Martinez-Fundichely A, Miyano S, Nakagawa H, Navarro FCP, Ossowski S, Park PJ, Pearson JV, Puiggròs M, Rippe K, Roberts ND, Roberts SA, Rodriguez-Martin B, Schumacher SE, Scully R, Shackleton M, Sidiropoulos N, Sieverling L, Stewart C, Torrents D, Tubio JMC, Villasante I, Waddell N, Wala JA,

Weischenfeldt J, Yang L, Yao X, Yoon SS, Zamora J, Zhang CZ, Park PJ, PCAWG Consortium. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet. 2020 Mar 2;52(3):331–41.

18. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019 Jul 19;10(1):3240.

19. Wang S, Lee S, Chu C, Jain D, Kerpedjiev P, Nelson GM, Walsh JM, Alver BH, Park PJ. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. Genome Biol. 2020 Dec;21(1):73.

20. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardımcı GG, Chakraborty A, Bann DV, Wang Y, Clark R, Zhang L, Yang H, Liu T, Iyyanki S, An L, Pool C, Sasaki T, Rivera-Mulia JC, Ozadam H, Lajoie BR, Kaul R, Buckley M, Lee K, Diegel M, Pezic D, Ernst C, Hadjur S, Odom DT, Stamatoyannopoulos JA, Broach JR, Hardison RC, Ay F, Noble WS, Dekker J, Gilbert DM, Yue F. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018 Oct;50(10):1388–98.

21. Davies JJ, Wilson IM, Lam WL. Array CGH technologies and their applications to cancer genomes. Chromosome Res. 2005 Apr;13(3):237–48.

22. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Mégy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E, Dolling H, Erwood M, Grozeva D, Stefanucci L, Arno G, Webster AR, Cole T, Austin T, Branco RG, Ouwehand WH, Raymond FL, Carss KJ. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. 2018 Dec;10(1):95.

23. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. Copenhaver GP, editor. PLoS Genet. 2011 Dec 1;7(12):e1002384.

24. Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. J Hum Genet. 2020 Jan;65(1):3–10.

25. Minervini CF, Cumbo C, Orsini P, Anelli L, Zagaria A, Specchia G, Albano F.

Nanopore Sequencing in Blood Diseases: A Wide Range of Opportunities. Front Genet. 2020 Feb 19;11:76.

26. Klever MK, Sträng E, Hetzel S, Jungnitsch J, Dolnik A, Schöpflin R, Schrezenmeier JFF, Schick F, Blau O, Westermann J, Rücker FG, Xia Z, Döhner K, Schrezenmeier H, Spielmann M, Meissner A, Melo US, Mundlos S, Bullinger L. AML with complex karyotype: extreme genomic complexity revealed by combined long-read sequencing and Hi-C technology. Blood Adv. 2023 Aug 15;bloodadvances.2023010887.

27. Melo US, Schöpflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, Klever MK, Türkmen S, Heinrich V, Pluym ID, Matoso E, Bernardo de Sousa S, Louro P, Hülsemann W, Cohen M, Dufke A, Latos-Bieleńska A, Vingron M, Kalscheuer V, Quintero-Rivera F, Spielmann M, Mundlos S. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. Am J Hum Genet. 2020 Jun;106(6):872–84.

28. Schöpflin R, Melo US, Moeinzadeh H, Heller D, Laupert V, Hertzberg J, Holtgrewe M, Alavi N, Klever MK, Jungnitsch J, Comak E, Türkmen S, Horn D, Duffourd Y, Faivre L, Callier P, Sanlaville D, Zuffardi O, Tenconi R, Kurtas NE, Giglio S, Prager B, Latos-Bielenska A, Vogel I, Bugge M, Tommerup N, Spielmann M, Vitobello A, Kalscheuer VM, Vingron M, Mundlos S. Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. Nat Commun. 2022 Oct 29;13(1):6470.

29. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014 Dec;159(7):1665–80.

30. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016 Jul;3(1):95–8.

31. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 2016 Jul;3(1):99–101.

32. Davidson NM, Majewski IJ, Oshlack A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. Genome Med. 2015 Dec;7(1):43.

33. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015 Dec;6(1):11.

34. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019 Jan 8;47(D1):D766–73.

35. Tham CY, Tirado-Magallanes R, Goh Y, Fullwood MJ, Koh BTH, Wang W, Ng CH, Chng WJ, Thiery A, Tenen DG, Benoukraf T. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol. 2020 Dec;21(1):56.

36. Poell JB, Mendeville M, Sie D, Brink A, Brakenhoff RH, Ylstra B. ACE: absolute copy number estimation from low-coverage whole-genome sequencing data. Birol I, editor. Bioinformatics. 2019 Aug 15;35(16):2847–9.

37. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar 1;14(2):178–92.

38. Carnevali D, Conti A, Pellegrini M, Dieci G. Whole-genome expression analysis of mammalian-wide interspersed repeat elements in human cell lines. DNA Res. 2016 Dec 27;dsw048.

39. Bhatnagar B, Blachly JS, Kohlschmidt J, Eisfeld AK, Volinia S, Nicolet D, Carroll AJ, Block AW, Kolitz JE, Stone RM, Mrózek K, Byrd JC, Bloomfield CD. Clinical features and gene- and microRNA-expression patterns in adult acute leukemia patients with t(11;19)(q23;p13.1) and t(11;19)(q23;p13.3). Leukemia. 2016 Jul;30(7):1586–9.

40. Yuan L, Zhai L, Qian L, Huang D, Ding Y, Xiang H, Liu X, Thompson JW, Liu J, He YH, Chen XQ, Hu J, Kong QP, Tan M, Wang XF. Switching off IMMP2L signaling drives senescence via simultaneous metabolic alteration and blockage of cell death. Cell Res. 2018 Jun;28(6):625–43.

41. Mrózek K. Cytogenetic, Molecular Genetic, and Clinical Characteristics of Acute Myeloid Leukemia With a Complex Karyotype. Semin Oncol. 2008 Aug;35(4):365–77.

42. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.

43. Fontana MC, Marconi G, Feenstra JDM, Fonzi E, Papayannidis C, Ghelli Luserna di Rorá A, Padella A, Solli V, Franchini E, Ottaviani E, Ferrari A, Baldazzi C, Testoni N, Iacobucci I, Soverini S, Haferlach T, Guadagnuolo V, Semerad L, Doubek M, Steurer M, Racil Z, Paolini S, Manfrini M, Cavo M, Simonetti G, Kralovics R, Martinelli G. Chromothripsis in acute myeloid leukemia: biological features and impact on survival. Leukemia. 2018 Jul;32(7):1609–20.

44. Zhang CZ, Leibowitz ML, Pellman D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. Genes Dev. 2013 Dec 1;27(23):2513–30.

45. 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 Dec;5(1):14.

Statutory Declaration

"I, Marius-Konstantin Klever, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic: Genomic and transcriptomic analysis of complex structural variants and their relevance in AML with a complex karyotype (Genomische und transkriptomische Analyse komplexer struktureller Varianten und ihre Relevanz in der AML mit komplexem Karyotyp), independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me."

_____	_____
Date	Signature

Declaration of your own contribution to the publications

Marius-Konstantin Klever contributed the following to the below listed publications:

Publication 1: **Klever MK**, Sträng E, Hetzel S, Jungnitsch J, Dolnik A, Schöpflin R, Schrezenmeier JF, Schick F, Blau O, Westermann J, Rücker FG, Xia Z, Döhner K, Schrezenmeier H, Spielmann M, Meissner A, Melo US, Mundlos S, Bullinger L. AML with complex karyotype: extreme genomic complexity revealed by combined long-read sequencing and Hi-C technology. Blood Advances. 2023 Nov 14;7(21):6520–31.

Contribution: My work focused primarily on the implementation of the laboratory experiments on which this work is based, as well as on the bioinformatic analyses. The analysis of the Hi-C and Nanopore DNA data as well as the development of the detection method for structural variants (SVs) can be attributed almost exclusively to my work.

The majority of the work was carried out by me in the following areas:

1. Laboratory chemical generation of the Hi-C library for the Hi-C sequencing (n=12) (This was the most laborious part in terms of laboratory chemistry, since the preparation of the library for 2 cases takes about 1-2 full working weeks) (support by J. Jungnitsch and U. Melo)
2. Nanopore gDNA library preparation and sequencing on the gridION (n=12)
3. RNA isolation and library preparation for Illumina RNA sequencing (n=14)
4. Nanopore direct cDNA library preparation and sequencing on the gridION (n=9)
5. Administration and processing of the already existing clinical and genetic data of the patients (support by A. Dolnik, J. Schrezenmeier)
6. Analysis of Hi-C maps and Hi-C SV detection
7. Bioinformatic analysis of the Nanopore gDNA data using NanoVar and development of own filter strategies (n=11) (support by E. Sträng)
8. Development of an SV detection methodology for the integrative analysis of genomic Nanopore and Hi-C data (support by E. Sträng)
9. Analysis of the Nanopore cDNA and Illumina RNA data using software for identifying fusion transcripts (JAFFA) and integration of these two datasets in the context of the known genomic datasets (support by E. Sträng)
10. Writing of the first manuscript draft and creating all figures and tables (support by U. Melo, E. Sträng, J. Jungnitsch)

I worked in the following areas, the contribution of co-authors of the manuscript had a workload comparable or larger work than mine:

11. Bioinformatic generation of Hi-C maps (support by R. Schöpflin)
12. Automated Hi-C Map Analysis (support by R. Schöpflin)
13. Participation in the acquisition of patient samples and stem cells from healthy individuals (n=14) (support by J. Schrezenmeier, H. Schrezenmeier, A. Dolnik, L. Bullinger, J. Westermann)
14. Analysis of the distribution of breakends in genomic features (support by E. Sträng, S. Hetzel))
15. Bioinformatic analysis of Illumina RNA gene expression data (support by S. Hetzel)
16. General project planning and revision of the manuscript (support by U. Melo, S. Mundlos, L. Bullinger)

Publication 2: Schöpflin R, Melo US, Moeinzadeh H, Heller D, Laupert V, Hertzberg J, Holtgrewe M, Alavi N, **Klever MK,** Jungnitsch J, Comak E, Türkmen S, Horn D, Duffourd Y, Faivre L, Callier P, Sanlaville D, Zuffardi O, Tenconi R, Kurtas NE, Giglio S, Prager B, Latos-Bielenska A, Vogel I, Bugge M, Tommerup N, Spielmann M, Vitobello A, Kalscheuer VM, Vingron M, Mundlos S. Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. Nature Communications. 2022 Oct 29;13(1):6470.

Contribution: Together with U. Melo and J. Jungnitsch I performed the library preparations for several of the cases. In the data analysis, I worked on the manual curation of SV calls based on Hi-C maps (together with U.Melo and J. Jungnitsch). Furthermore, I was involved in general project planning. I especially contributed to the strategies of integrating long-read sequencing with Hi-C SV calls. These cases of germline chromothripsis showed different patterns and lower local complexity compared to the cases of chromothripsis in AML that I analyzed for my main project. Therefore, different strategies of SV calling and data integration were found to be more suitable here (see methods). Figure 1 and 2 (especially 2c) as well as Supplementary Figure 2, which is included in this dissertation, are partly based on my SV analyses. I also revised the final manuscript and figures.

Publication 3: Melo US, Schöpflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, **Klever MK**, Türkmen S, Heinrich V, Pluym ID, Matoso E, Bernardo de Sousa S, Louro P, Hülsemann W, Cohen M, Dufke A, Latos-Bieleńska A, Vingron M, Kalscheuer V, Quintero-Rivera F, Spielmann M, Mundlos S. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. American Journal of Human Genetics. 2020 Jun;106(6):872–84.

Contribution: With this project I started my work on complex genomic rearrangements. I performed the library preparations for several of the cases together with the first author (U. Melo). I analyzed the HI-C maps of the complex case DD3 and reconstructed the rearranged linear genomic structure. Hereby, I contributed the underlying data to Figure 2. I also contributed to the analysis of several other cases, especially to the resulting Figure 3, which is included in this dissertation, and partially also to Figure 4. Furthermore, I worked on revising the manuscript and figures. In this work, I developed a lot of the techniques that I used for analyzing complex Hi-C maps in my subsequent work.

_____

Signature, date and stamp of first supervising university professor / lecturer

_____

Signature of doctoral candidate

Printing copy of Publication 1: AML with complex karyotype: extreme genomic complexity revealed by combined long-read sequencing and Hi-C technology.

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.

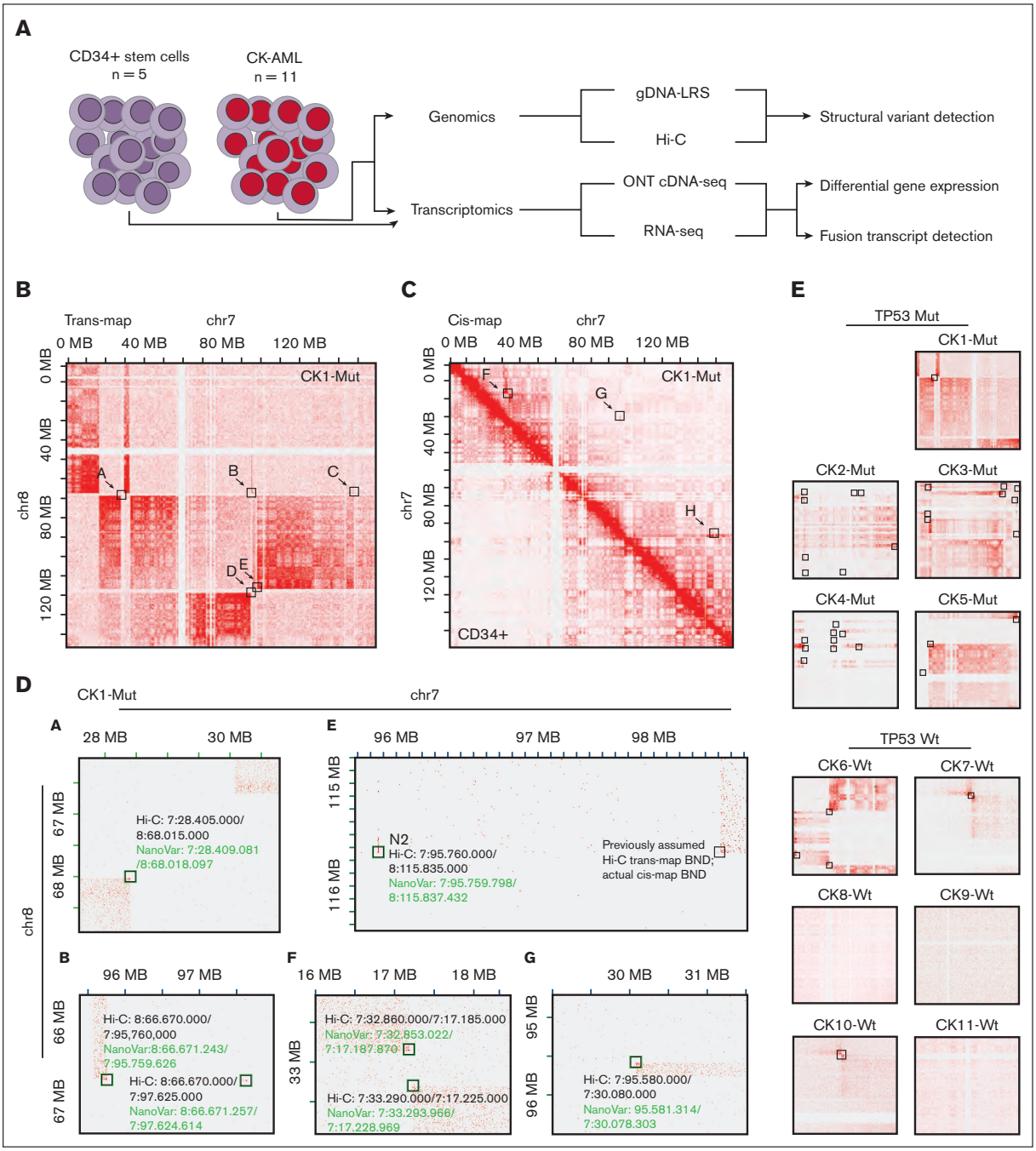

Figure 1. Cohort overview and complex genomic rearrangements detected by our SV detection workflow. (A) Samples from patients with CK-AML (n = 11) were subjected to genomic sequencing (Hi-C and ONT-GS) for SV detection. In addition, healthy CD34^{+} stem cell donors (n = 5) and the CK-AML samples were 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.

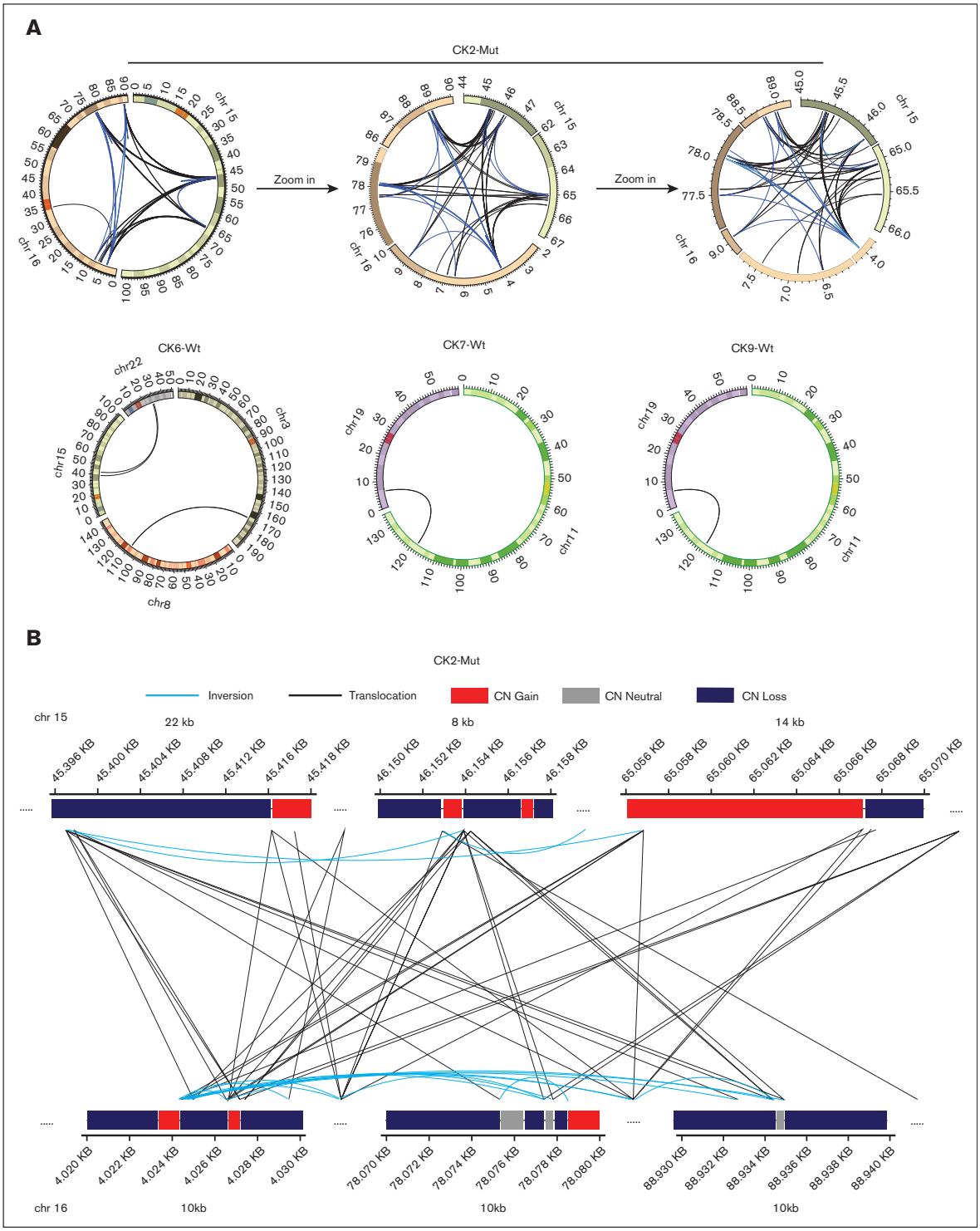

Figure 2. Chromocataclysm in CK-AML. (A) Circos plot of a chromocataclysm rearrangement in CK2-Mut and noncomplex rearrangements in CK6-Wt, CK7-Wt, and CK9-Wt. A clustering of breakends is preserved at all 3 stages of magnification shown here for case CK2-Mut. The clustering is here shown at full chromosome 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

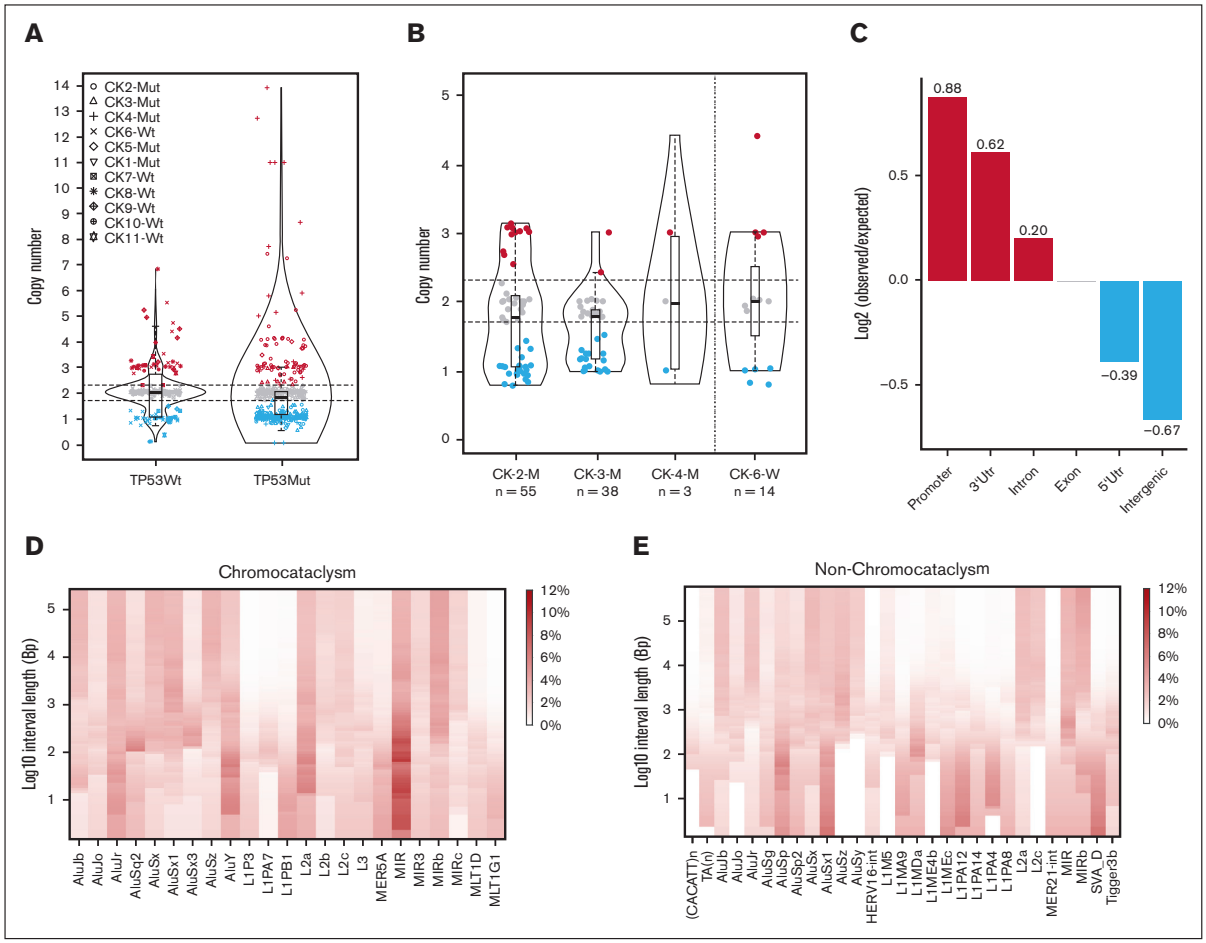

Figure 3. CN distribution and enrichment of breakends in the genome. (A) Violin plots of CN distribution in the final CNV data set of the TP53 mutated (n = 6) and TP53 wildtype (n = 5) cases. Each dot represents 1 fragment (distinct region on a genome of reference) and its respective CN. (B) Genomic 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.

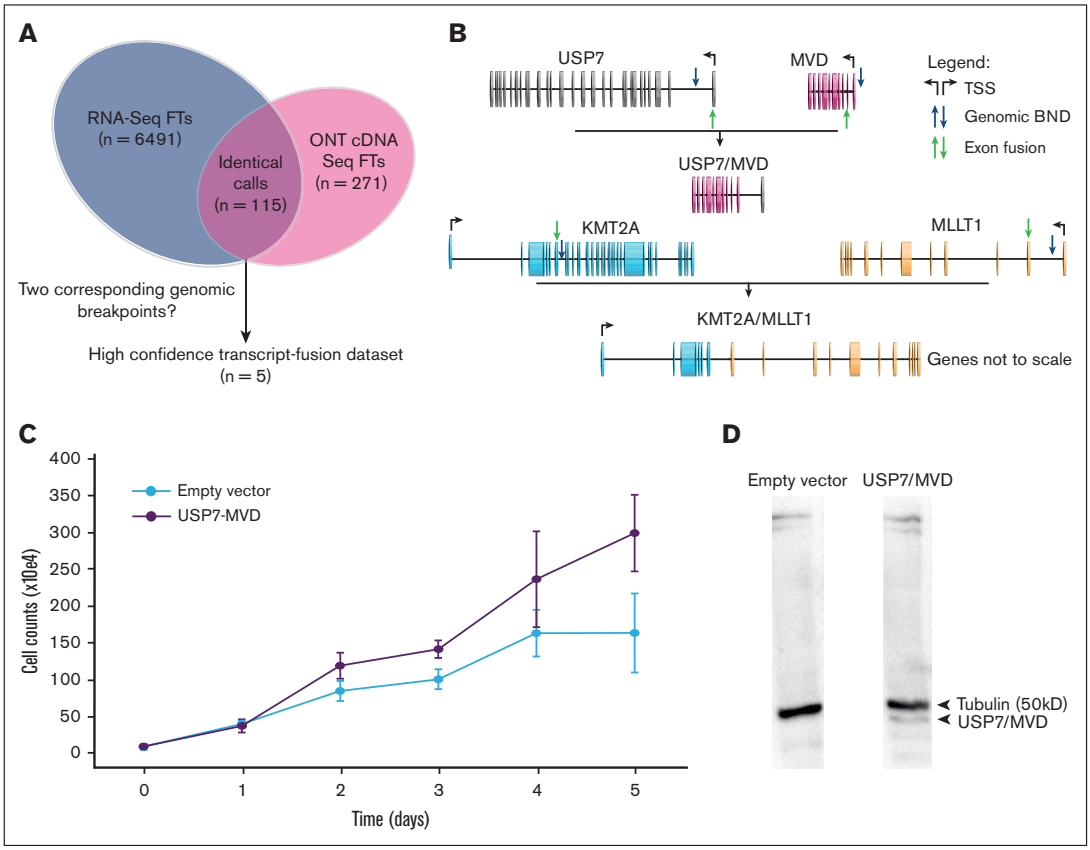

Figure 4. Identification of fusion transcripts. (A) Illustration of the fusion transcript detection pipeline, starting with integrating matching fusion transcript calls from JAFFA (Illumina RNA and ONTdirect cDNA data set) and filtering them by applying the criterion of 2 corresponding SV breakends to 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

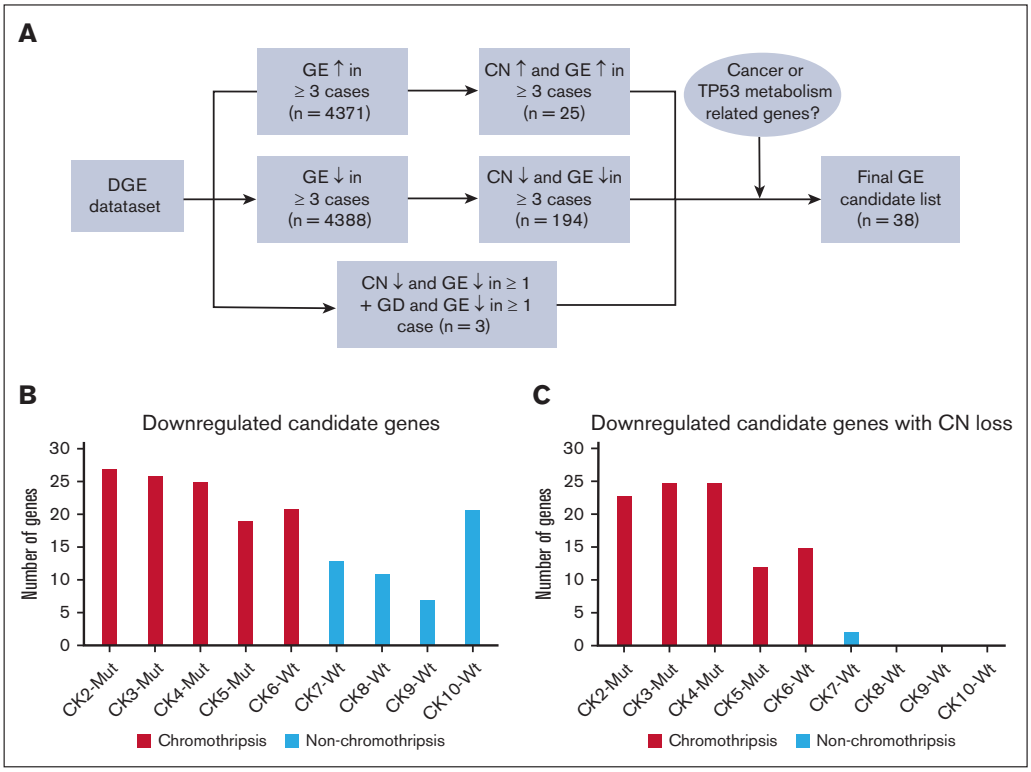

Figure 5. Gene dysregulation and influence of CN changes in CK-AML. (A) Schematic overview of our differentially expressed genes (DEG) analysis that integrates gene expression (GE) with CN information. CN ↓, CN loss; CN ↑, CN gain; GE ↓, gene downregulation; GE ↑, gene upregulation. (B) Number of genes 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.

Printing copy of Publication 2: Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes

nature communications

Article

https://doi.org/10.1038/s41467-022-34053-7

Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes

Received: 17 October 2021

Accepted: 7 October 2022

Published online: 29 October 2022

Check for updates

Robert Schöpflin[1,2,3,21], Uirá Souto Melo[1,2,21], Hossein Moeinzadeh[3], David Heller[3], Verena Laupert[3], Jakob Hertzberg[1,2,3], Manuel Holtgrewe[4,5], Nico Alavi[3], Marius-Konstantin Klever[1,2], Julius Jungnitsch[1,2], Emel Comak[3], Seval Türkmen[2,6], Denise Horn[2], Yannis Duffourd[7,8], Laurence Faivre[7,9], Patrick Callier[7,8], Damien Sanlaville[10], Orsetta Zuffardi[11], Romano Tenconi[12], Nehir Edibe Kurtas[13], Sabrina Giglio[14], Bettina Prager[15], Anna Latos-Bielenska[16], Ida Vogel[17], Merete Bugge[18], Niels Tommerup[18], Malte Spielmann[1,19,20], Antonio Vitobello[7,8], Vera M. Kalscheuer[1], Martin Vingron[3] ✉ & Stefan Mundlos[1,2] ✉

Structural variants are a common cause of disease and contribute to a large extent to inter-individual variability, but their detection and interpretation remain a challenge. Here, we investigate 11 individuals with complex genomic rearrangements including germline chromothripsis by combining short- and long-read genome sequencing (GS) with Hi-C. Large-scale genomic rearrangements are identified in Hi-C interaction maps, allowing for an independent assessment of breakpoint calls derived from the GS methods, resulting in >300 genomic junctions. Based on a comprehensive breakpoint detection and Hi-C, we achieve a reconstruction of whole rearranged chromosomes. Integrating information on the three-dimensional organization of chromatin, we observe that breakpoints occur more frequently than expected in lamina-associated domains (LADs) and that a majority reshuffle topologically associating domains (TADs). By applying phased RNA-seq, we observe an enrichment of genes showing allelic imbalanced expression (AIG) within 100 kb around the breakpoints. Interestingly, the AIGs hit by a breakpoint (19/22) display both up- and downregulation, thereby suggesting different mechanisms at play, such as gene disruption and rearrangements of regulatory information. However, the majority of interpretable genes located 200 kb around a breakpoint do not show significant expression changes. Thus, there is an overall robustness in the genome towards large-scale chromosome rearrangements.

Genomic rearrangements, also called structural variants (SVs), contribute to a large extent to genomic variability and are a common cause of genetic disease. Despite advances in genome sequencing (GS) technologies, their detection remains a challenge, particularly in complex cases with many nested rearrangements. Furthermore, long-read and short-read SV pipelines show different sensitivity and specificity, depending on the type and the size of the SV[1,2]. The interpretation of SVs with respect to pathogenicity has been the subject of many

A full list of affiliations appears at the end of the paper. ✉e-mail: vingron@molgen.mpg.de; stefan.mundlos@charite.de

studies, and multiple algorithmic attempts have been made to account for the many effects SVs can have. While changes in copy number may exert their effects via a change in gene dosage, others disrupt genes or result in gene fusions. Furthermore, large-scale rearrangements can alter the three-dimensional chromatin architecture, e.g., by disrupting or reshuffling topologically associating domains (TADs), thereby rewiring gene regulatory landscapes[3]. Even though this mechanism has been observed in congenital malformation disorders[4,5], as well as for cancer[6], it is unclear how generalizable the effect of TAD disruptions on gene expression is and how they might be related to an individuals' phenotype. Conversely, genome-wide depletion of cohesin and CTCF, both important for the formation of TAD boundaries, did not lead to large-scale changes in gene expression[7,8], and the disruption of a TAD boundary alone does not necessarily lead to a change in gene expression[9]. Furthermore, only a small fraction of genes become misregulated upon disruption of TADs in shattered balancer chromosomes in *Drosophila melanogaster*[10].

We approached these open questions by studying extreme cases of chromosomal rearrangements, as they are observed in congenital chromoanagenesis. Chromoanagenesis is an umbrella term for complex large-scale genomic rearrangements, which can further be divided into chromoplexy (Fig. 1a), chromothripsis (Fig. 1a), and chromoanasynthesis, depending on the complexity of the events and the gain/loss of genetic material[11]. Complex genomic rearrangements (CGR) are frequent in many cancer genomes[12,13] but are rare in congenital disorders[11,14]. Despite the massive rearrangements, some affected individuals show only mild clinical symptoms. The complex nature of these rearrangements with many breakpoints and their highly nested form makes it exceedingly difficult to resolve them. They can thus be considered a perfect test bed for technologies to detect and interpret genomic rearrangements.

Here, we investigate the genomes of 11 individuals with complex constitutional chromosome rearrangements, including chromothripsis. We perform an extensive characterization of breakpoints using a combination of short-read (Illumina) and long-read (PacBio CLR) genome sequencing, as well as Hi-C (Fig. 1b). We use these three technologies jointly to detect breakpoints, remove likely false-positive calls, reconstruct shattered chromosomes, characterize the 3D chromatin landscape and investigate novel adjacencies created upon the genomic rearrangements. The analysis of single nucleotide polymorphisms (SNPs) in conjunction with PacBio and Hi-C reads *is* further used for an allele-specific quantification of RNA-seq data from patient cells. The results indicate that the fraction of genes with altered expression is associated with the distance of the gene to a breakpoint.

Results

Combining genome sequencing and Hi-C reveals the complexity of genomic rearrangements in germline chromoanagenesis

Ten out of eleven individuals included in this study presented with intellectual disability (ID), while one did not present any pathogenic feature. Their diagnostic workup included karyotyping and microarray-based comparative genomic hybridization (array CGH) (Supplementary Table 1). One case was tested with multicolor FISH, confirming the presence of several translocations (Fig. 1c). Lymphoblastoid cell lines

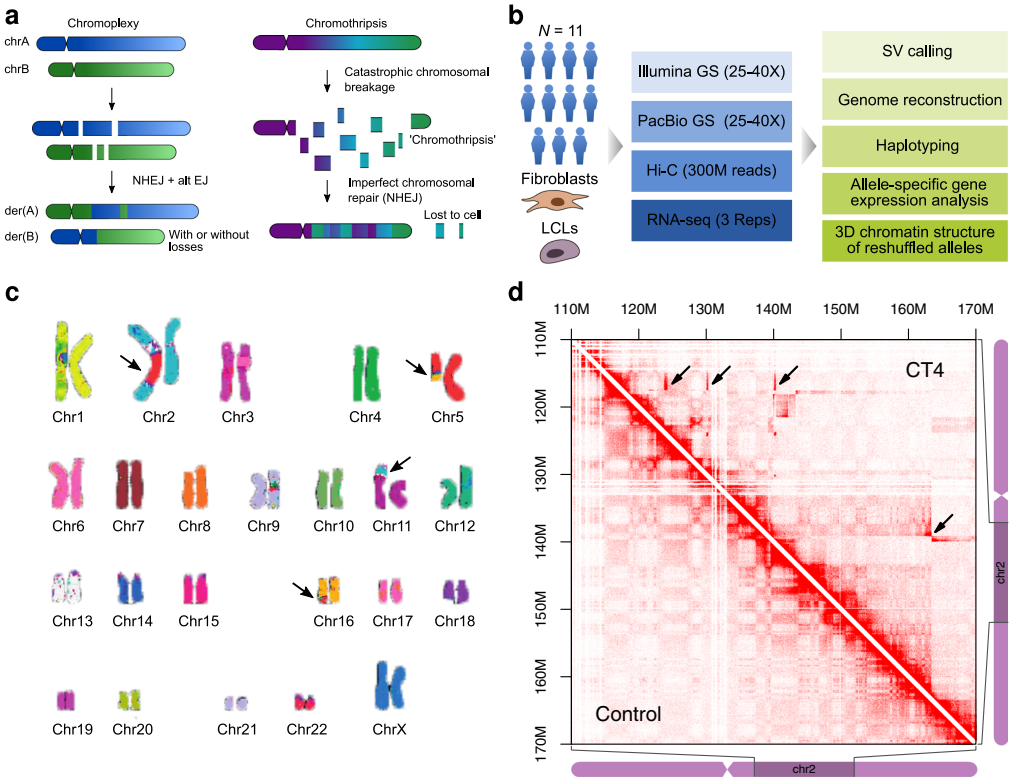

Fig. 1 | Complex genomic rearrangements investigated in this study. a Forms of complex genomic rearrangements: Chromoplexy is characterized by the exchange of larger fragments between chromosomes. Chromothripsis is characterized by a shattering of one or several chromosomal fragments followed by an imperfect repair. Schematic is based on[11,57]. **b** Outline of the study: cohort, clinical sample available, sequencing technologies, and analyses. Schematics of cells were created with BioRender.com. **c** Multicolor FISH of sample CT4 indicating several translocations between chr2, chr5, chr11, and chr16. **d** Hi-C map of chr2 of CT4 showing several large-scale rearrangements. Ectopic interactions are only visible in the CT4 sample (examples indicated by arrows, upper triangular matrix), but not the in a control (lower triangular matrix).

https://doi.org/10.1038/s41467-022-34053-7

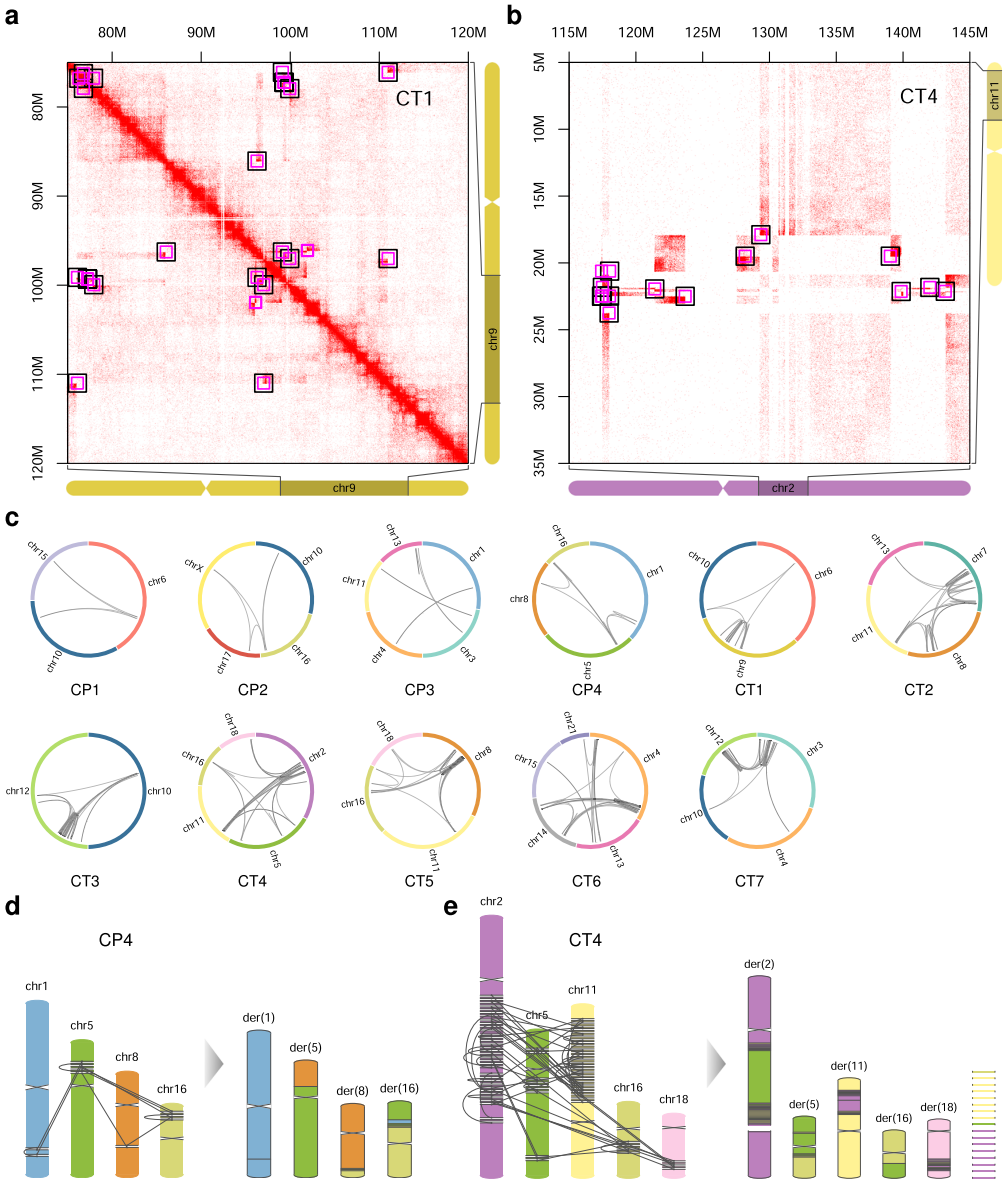

Fig. 2 | Reconstruction of shattered chromosomes. a Overlay of curated SV calls from Illumina GS (pink squares) and PacBio GS (black squares) for a cis Hi-C map as well as for **b** a trans Hi-C map. **c** Circos plots for 11 samples showing all curated large-scale novel adjacencies. **d** Reconstruction graph and derivative chromosomes for CP4. The chromosome reconstruction strategy comprises several steps starting with (i) the placement of all curated breakpoints to generate chromosomal fragments, (ii) connecting chromosomal fragments according to SV calls, (iii) tracing all possible paths in the fragment graph to obtain derivative chromosomes. **e** Reconstruction graph and derivative chromosomes for CT4. The last column shows leftover singletons. Note, small fragments and telomeric ends are shown enlarged.

(LCLs) were available for 10 cases, and fibroblasts for one case. To further characterize the rearrangements in detail, we applied Illumina short-read as well as PacBio long-read GS (CLR) (Fig. 1b).

We first applied the structural variant caller SVIM[15] to PacBio long-read GS data and detected, on average 243 large-scale novel adjacencies (i.e., >100 kb between the fused positions or trans) per sample when filtering for a quality-value of at least 5 (i.e., 5% of the mean alignment coverage). Large-scale rearrangements create novel adjacencies that cause ectopic interaction patterns in Hi-C maps[16,17] (Fig. 1d). A novel adjacency becomes especially prominent in the Hi-C map when the genomic distance between the fused positions is large and when the fused chromosomal fragments are large. True novel adjacencies between large fragments colocalize with an ectopic chromatin interaction pattern in the Hi-C map, whereas those that show no evidence of ectopic interaction are likely false positives. We, therefore, focused on novel adjacencies of >100 kb between the fused positions and projected them onto the Hi-C maps (Fig. 2a, b).

After the manual curation of the initial SV calls (See Methods), we revised our SV filtering strategy for PacBio to reduce the call set further and excluded SV calls with at least one of the criteria described in the Methods section. Additionally, we projected Illumina GS large-scale SV calls (>100 kb or trans) from the Delly tool[18] onto the Hi-C map and

curated the SV calls from PacBio and Illumina jointly. Besides the coordinates, the strand information of the novel adjacency can also be informative, because it indicates the direction (Supplementary Fig. 1a), in which the ectopic Hi-C signal is expected[19]. The joint curation of PacBio and Illumina-based calls yielded between 4 and 73 large-scale novel adjacencies per case, most of them detected by both methods (Supplementary Fig. 1b). Only a small subset of calls was unique to one of the methods. We did not consider small-scale novel adjacencies (<100 kb) or novel adjacencies not related to complex rearrangements. However, small-scale novel adjacencies (1–100 kb between anchor points) were considered later in the reconstruction of derivative chromosomes, when their coordinates matched the anchor points of curated large-scale novel adjacencies (See Methods).

Genome sequencing analysis reveals two regimes of genomic complexity

The analysis of novel adjacencies revealed large differences between the investigated cases in the number of breakpoints, as well as in their distribution throughout the genome. Four cases showed patterns of chromoplexy (hereafter named CP1–CP4), and seven cases exhibited chromothripsis-like structures (CT1–CT7) (Fig. 2c). The chromoplexy cases were characterized by a lower number of large-scale novel adjacencies (4–11) (Fig. 2c) and less complex ectopic Hi-C patterns (Supplementary Fig. 2c, e). Often two or more chromosomes were involved in the rearrangement, with no or only a few cis junctions (Supplementary Fig. 2e). The resulting chromosomal fragments were usually large (Supplementary Fig. 3A), and no copy number gain of fragments was observed. One case (CP3) showed a loss of 6 Mb (Supplementary Fig. 3b). The chromothripsis cases, in contrast, had a higher number of large-scale novel adjacencies (16–73) (Fig. 2c, Supplementary Fig. 1b) and the resulting chromosomal fragments were often smaller (Supplementary Fig. 3a), due to a clustering of breakpoints, resulting in regions of shattering (Fig. 2c). The rearrangements occurred in nested conformation, leading to complex contact patterns in the corresponding Hi-C maps (Supplementary Fig. 2d, f). In some cases, the exchange of genetic material between the chromothriptic patches of several chromosomes was observed as an ectopic signal in the trans Hi-C maps (Supplementary Fig. 2d). The number of affected chromosomes ranged between two and five. Losses of chromosomal fragments were frequent (between 1 and 46 deleted fragments) (Supplementary Fig. 3b), and copy number gains of small fragments (<1 kb) were observed in two cases (Supplementary Fig. 3b). We compared the curated SVs to the database gnomAD-SV[20] and did not find SVs with matching coordinates (tolerance of the breakpoints ±1 kb and allele frequency >0.01).

To identify SV-breakpoints that are potentially disease-causing, we searched for genes associated with intellectual disability (ID), the most prevalent phenotype in our cohort, that was hit by a curated small- or large-scale SV (Supplementary Table 2). Six out of eleven cases harbor breakpoints within genes associated with ID (Supplementary Table 2). For instance, we detected a breakpoint in GRIN2B (CT3), which was previously identified in this case and described as causative[21]. Moreover, further breakpoints affecting SOX5 were identified, which might contribute to the phenotypic spectrum in this case. Furthermore, we identified a breakpoint in USP7, which is likely to explain the CP4 phenotype. All other genes associated with ID in our cohort were inherited in an autosomal recessive fashion and were therefore discarded as candidates.

Novel genomic adjacencies are enriched in chromatin at the nuclear periphery

We investigated the occurrence of novel adjacencies with respect to known features of the genome and chromatin organization, such as TADs, A/B compartments, and lamina-associated domains (LADs)[22]. To evaluate the enrichment of breakpoints with respect to genomic features, we computed *P*-values based on empirical background models, such as novel adjacencies with random coordinates (See methods and Supplementary Fig. 18). Novel adjacencies are potentially TAD disrupting, as they may fuse the regulatory content of different TADs (Fig. 3a), potentially causing gene misregulation[3]. According to this, evolutionary rearrangement breakpoints of different vertebrate species were found to be enriched at TAD boundaries. Breaks within TADS are thus being avoided by negative selection[23]. We thus tested if TAD boundaries are enriched for breakpoints (Fig. 3b), but did not observe an enrichment (*P*-value = 0.819).

An important layer of chromatin organization is the segregation of chromatin into A and B compartments. Compartments are defined based on Hi-C maps[24,25], showing preferential long-range interactions between regions of the same compartment type. A and B compartments were observed to largely overlap with the transcriptional state and epigenetic features of euchromatin and heterochromatin, respectively. Interestingly, we observed an enrichment of breakpoints in the B-compartment (*P*-value = 0.037) (Fig. 3c). Besides the fusion of two B-compartment locations, we also observed the fusion of an A-compartment locus to a B-compartment locus (Fig. 3d). In line with this observation, we investigated the abundance of breakpoints in lamina-associated domains (LADs). LADs are chromatin domains interacting with the nuclear envelop and can be detected by DamID-seq[26,27]. They are generally transcriptionally less active than non-LAD regions of the genome. Similar to the A/B-compartment analysis, we observed an enrichment of breakpoints in LADs (*P*-value = 0.008) (Fig. 3e). LAD-LAD fusions are an abundant type of novel adjacencies (Fig. 3f). Additionally, also a fusion between regions that are LAD and non-LAD regions in a WT genome occurred (Fig. 3f).

An analysis of repeats at breakpoints did not show a general preference of breakpoints to occur in repeats (*P*-value = 0.879) (Supplementary Fig. 4a). Cases CP2, CP4, and CT1 show values above the expected value, but have relatively few breakpoints in general. An analysis of repeat classes overlapping with breakpoints did not reveal a striking, recurring pattern that would be shared by several cases (Supplementary Fig. 4b).

An in-depth analysis of the flanking sequences around the breakpoints (See Methods) showed that novel adjacencies are often associated with small losses or gains of genetic material indicating imperfect fusion of chromosomal fragments (Supplementary Fig. 5). Additionally, we searched for microhomology, i.e., short regions of DNA sequence homology[28], around the breakpoints of novel adjacencies. The analysis revealed only low degrees of microhomology (Supplementary Fig. 5b), indicating that sequence homology is a minor aspect of the fusion events.

In summary, we observed an enrichment of breakpoints in the B-compartment and LADs (regions with low gene expression), while we found no striking association with TAD boundaries and repeat regions.

Reconstruction of whole derivative chromosomes

Next, we used the high-confidence set of novel-adjacency calls to build a reference-based reconstruction graph of the derivative chromosomes. In the first step, we used the curated breakpoint positions to split the original WT chromosomes into the corresponding chromosomal fragments (Fig. 2d). In a reconstruction graph, each chromosomal fragment defines two nodes, i.e., at the 5′ and at the 3′ end of each fragment. The 5′ and 3′ nodes of the same fragment are implicitly connected by an edge. Fragments with a telomere have only one node on the non-telomeric side. All high-confidence novel adjacencies were included as additional edges to the reconstruction graph. Afterward, the reconstruction graph was traversed, and the order and orientation of chromosomal fragments along the paths revealed the layout of the derivative chromosome.

In chromoplexy cases, which are depleted of cis novel adjacencies, the reconstruction yielded whole derivative chromosomes

https://doi.org/10.1038/s41467-022-34053-7

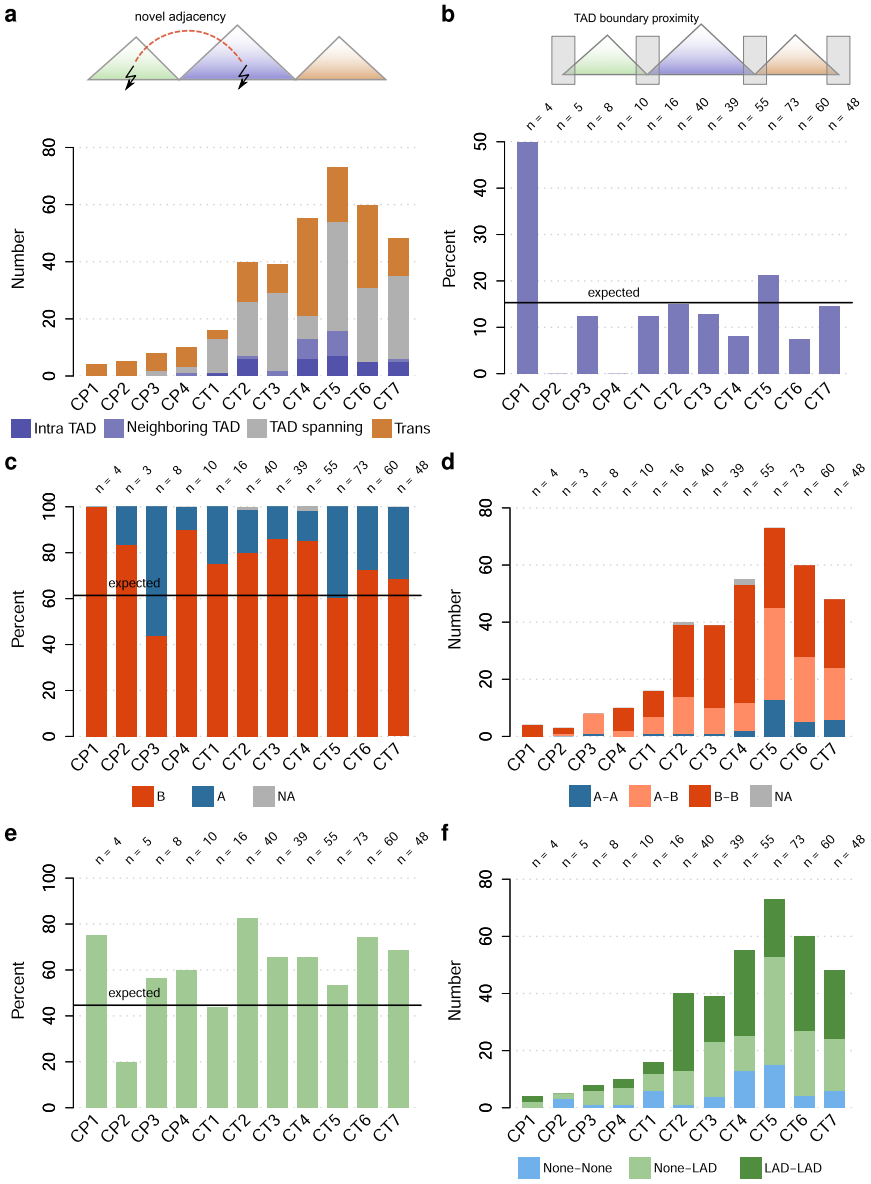

Fig. 3 | Analysis of the 3D chromatin structure around breakpoints. For selected genomic features, a horizontal line indicates the expected value, which is derived from the genome-wide fraction of the corresponding feature. **a** Number of novel adjacencies with breakpoints located in the same TAD, in neighboring TADs, spanning at least one TAD and on different chromosomes, respectively. **b** Percentage of breakpoints locating in TAD boundary regions (±50 kb). The black line indicates the percentage expected by chance. N is the number of novel adjacencies evaluated. **c** Fraction of compartment type of the breakpoint location for individual samples. The expected line shows here the genomic fraction of the B-compartment. **d** Number and type of compartment fusions induced by the large-scale novel adjacencies. **e** Fraction of breakpoints located in LADs. The black line indicates the percentage expected by chance. **f** Number and type of LAD/Non-LAD fusions.

(Fig. 2d). For most of the complex chromothripsis cases (CT2, CT4–CT7), the reconstruction graph was not complete because not all novel adjacencies could be identified leading to missing junctions. In these cases, the reconstruction stops at the stage of incomplete chromosomal scaffolds (Fig. 2e). To overcome this problem, we included Hi-C information to complement the reconstruction analysis. In the Hi-C derived interaction matrix, we used a 2D grid defined by the breakpoints and searched within this grid for ectopic Hi-C contacts shared between scaffolds. In case of ectopic interactions, the corresponding scaffolds can be further grouped together, assuming their origin from the same derivative chromosome (See Methods, Supplementary Fig. 6). The principle also works in reverse; for checking reconstructions, the pattern of the derivative chromosome can be projected back onto the Hi-C map to determine if it is compatible with the ectopic Hi-C patterns. The grid representation of breakpoints has another advantage because it indicates that breakpoints are missing when sharp edges of ectopic Hi-C patterns are not bordered by a breakpoint line. These indicators provide valuable information about derivate chromosome reconstruction.

After grouping incomplete scaffolds to putative derivative chromosomes (Supplementary Fig. 6), Hi-C was instrumental to identify the

https://doi.org/10.1038/s41467-022-34053-7

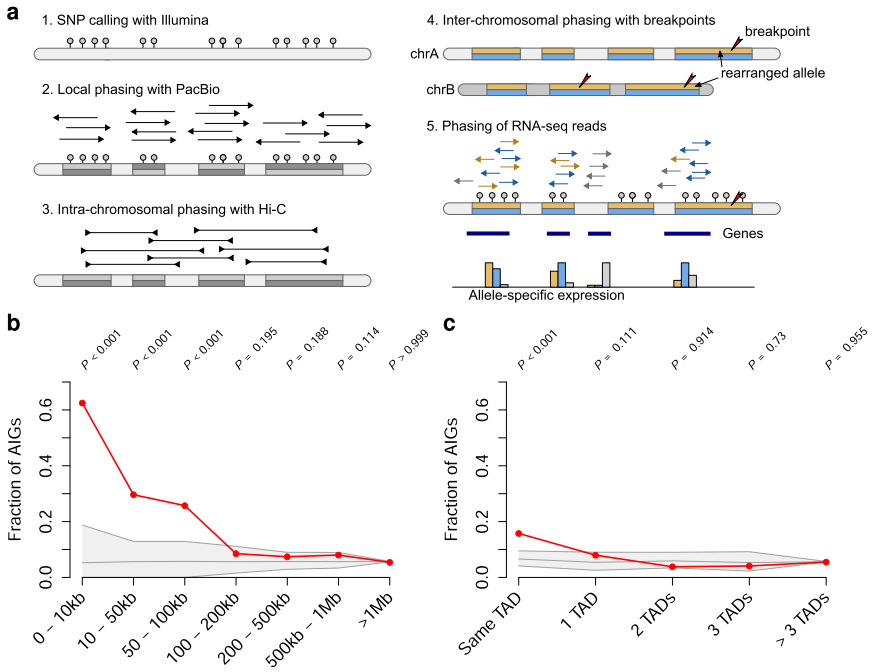

Fig. 4 | Haplotyping and allele-specific analysis of RNA-seq data. a Haplotyping: Based on SNPs from Illumina sequencing larger haplotype blocks are created using PacBio long-reads. Large haplotype blocks are connected with the help of Hi-C. The breakpoints of rearrangements are also phased and used to label WT allele and affected allele. The haplotype information is used to phase RNA-seq data at positions with informative SNPs and to derive allele-specific gene expression values. **b** Differential gene expression analysis around breakpoints: Fraction of allelic imbalance genes (AIGs) with respect to the distance between an expressed gene and the closest breakpoint (red line). As a control, simulations of random breakpoints were performed. The light gray area with the gray line indicates the 5th percentile, median and 95th percentile, respectively, expected by chance. *P*-values were computed by comparing each observed value against values obtained from an empirical background model. The tests were performed right-sided, adjustments for multiple testing were not performed. **c** Same as **b**, but the distance of a gene to the next breakpoint is measured in TAD units (Same TAD, 1 TAD, etc.). *P*-values were computed as described in **b**.

order and orientation of the components. For the majority of cases, the derivate chromosomes could be readily resolved with only a few large components remaining (Supplementary Fig. 7). For a small number of components, all possible solutions can be enumerated in a permutation approach (See Methods, Supplementary Fig. 8) and each of the solutions can either be inspected visually for its plausibility, or it can be evaluated with a scoring function to identify the best solution.

The reconstruction graph also contains leftover fragments which have no connecting edge (Fig. 2e, Supplementary Fig. 7). These singletons are candidates for a loss of genetic material. When checking the coverage of Illumina GS, indeed, many of these fragments appear as deletions (Supplementary Fig. 2, Supplementary Fig. 10). The remaining fragments, usually very small (<5 kb), seem to be still present in the genome, but their location remained unknown. Rarely, fragments of the reconstructed derivative chromosomes have low coverage, suggesting that they are rather deleted than part of the derivative chromosome (Supplementary Fig. 10). For case CT7, it was not possible to place the telomeric part of the q-arm of chr12 based on Hi-C, even though the fragment does not appear to be deleted.

Chromosomal rearrangements are associated with changes in gene expression based on genomic distance

To quantify the effect of chromosomal rearrangements on gene expression, we haplotyped the genomes using single nucleotide polymorphisms (SNPs) from Illumina GS, SV calls from PacBio together with PacBio long-reads and Hi-C reads, in order to produce allele-specific expression data (Fig. 4a). After phasing the RNA-seq data, DESeq2[29] was used to compare the expression of the rearranged allele vs. the WT allele (See Methods). Thus, for each individual, the expression was analyzed within the sample, instead of compared across individuals. Only genes which are expressed on at least one allele and for which the phasing of RNA-seq data was possible are informative (Supplementary Fig. 9b). Overall, we found 70 genes with the transcription start site (TSS) within 100 kb to the nearest breakpoint and phased RNA-seq signal. Out of these, 22 genes showed allelic imbalance expression (abs(Log2FC) > 1 and padj < 0.05). Simulations with random breakpoints showed that the effect is statistically significant up to a distance of 100 kb from the breakpoint and then decreases quickly (Fig. 4b). As an alternative approach, we performed simulations with permuted gene expression tables instead of random breakpoints and observed similar statistical trends (Supplementary Fig. 20). For distances up to 1 Mb, the fraction of allelic imbalanced gene expression appeared elevated, but was not statistically significant. In short, we observed an association between AIG and breakpoint proximity, and we further investigated the cause of this pattern.

Gene disruption is a likely explanation for reduced gene expression (Supplementary Fig. 9a, Supplementary Fig. 11a), and indeed, we found 39 informative genes with an intragenic breakpoint. However, not all of them showed downregulation: We also observed unchanged expression or even upregulation (Supplementary Fig. 11a). For genes hit by a breakpoint, we observed, that not necessarily all transcript variants were truncated by the breakpoint (Supplementary Fig. 12). Upregulation of gene expression could be caused by TAD fusions and/or enhancer–promoter rewiring. Alternatively, fusion transcripts, for which we found evidence on RNA-seq level (Supplementary Data 1), could also result in upregulation. These features can be overlapping; e.g., for seven genes, we found allelic imbalanced expression, an

https://doi.org/10.1038/s41467-022-34053-7

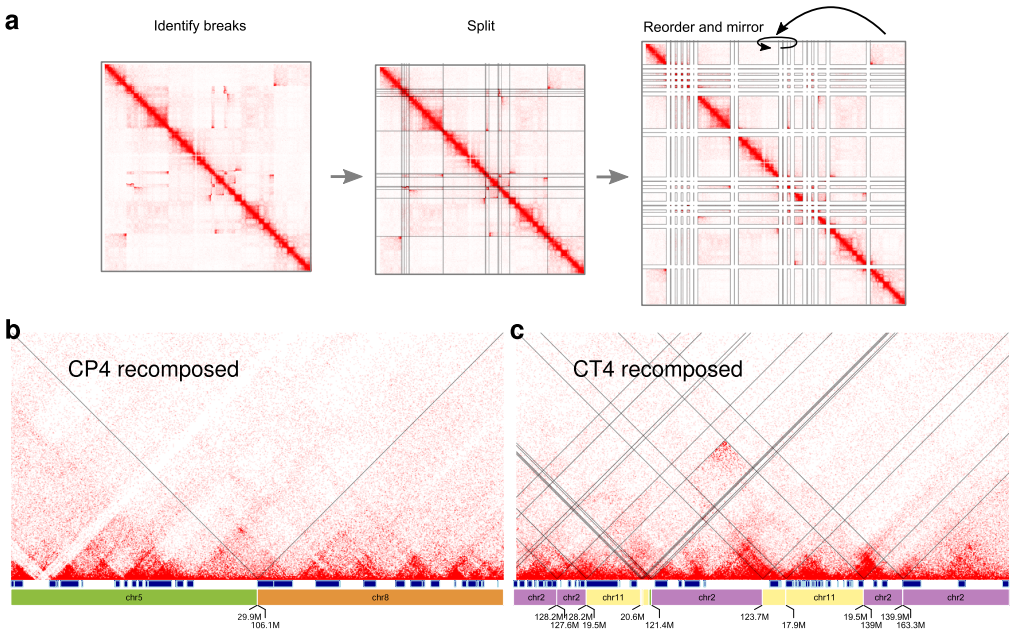

Fig. 5 | Analysis of the chromatin structure of reshuffled chromosomes. **a** Recomposing of Hi-C maps by cutting a Hi-C map at all breakpoint positions and reordering and reorienting all rows and columns according to the reconstruction scheme. **b** Recomposed Hi-C map for CP4 (**c**) and CT4. Labels for small chromosomal fragments are not shown to improve clarity. Note, the subtraction of a Hi-C map from a control sample could not remove Hi-C patterns of the WT allele entirely, visible in the map for CT4.

intragenic breakpoint, and a predicted fusion transcript at the same time (Supplementary Fig. 13).

In the group of upregulated genes, the fraction of housekeeping genes appears to be lower than expected by the genome-wide fraction of housekeeping genes (Supplementary Fig. 11b).

When grouping genes not by genomic distance to the closest breakpoint but by measuring the distance in the number of separating TADs, the enrichment of allelic imbalance genes (AIGs) was less pronounced (Fig. 4c), supporting the concept that genomic distance has a stronger effect on gene expression alterations than separating TADs.

Lastly, the majority of the breakpoints (69%) did not have an informative gene with its TSS in the region ±200 kb around the breakpoint (Supplementary Fig. 14). However, out of the 31% of breakpoints with at least one informative gene in ±200 kb proximity, 45% have at least one AIG (Supplementary Fig. 14).

Reshuffled Hi-C maps of derivative chromosomes reveal abundant TAD fusion events

Next, we sought to evaluate the rearranged 3D chromatin structure pattern in *chromoanagenesis*. The reconstruction of shattered chromosomes provides the linear layout of the derivative chromosomes, i.e., the location of functional elements such as genes, known regulatory elements, and insulators. The reconstructed derivative chromosomes can also be used to rearrange the Hi-C map accordingly (Fig. 5a). This recomposition brings the Hi-C signal from rearranged chromosomes into order, resolving the Hi-C patterns from rearrangements, which have been visible before. Interestingly, the derivative Hi-C maps reveal chromatin structures that emerged upon the rearrangements, such as TAD fusions and loops (Fig. 5b, c). We checked the rearranged Hi-C maps for TAD fusion events, focusing on candidate locations that have sufficient Hi-C signal and are interpretable. 86% of these locations in the chromoplexy cases and 80% in the chromothripsis cases showed also in the Hi-C map evidence for a TAD fusion. Taken together, massive chromosomal rearrangements in germline chromoanagenesis disturb gene expression. However, this occurs only to a certain degree, with most of the observable genes located in close proximity to a breakpoint not changing their expression. Thus, in our cohort, we observed overall robustness of the genome towards constitutional large-scale chromosome rearrangements.

Evaluation of the accuracy of reconstructed derivative chromosomes

We next evaluated the accuracy of the reconstructed derivative chromosomes using different approaches *because* a comprehensive ground truth is missing. On a coarse scale, karyotyping confirmed the majority of chromosomes we identified as being affected. For CT1 we found no involvement of chrX reported by karyotyping (Supplementary Table 1). Additionally, our approach detected the involvement of chr10 (CP2), chr11 (CP3), chr1 (CP4), and chr18 (CT5) that was not reported by karyotyping.

A previous study on case CT2 using mate-pair sequencing and Sanger sequencing[30] reported 41 novel adjacencies, the same number we curated. 37 novel adjacencies are common with our set, while 4 are unique to each of the sets when requiring a distance <1 kb between breakpoints and the same strand orientation at the junction to define common novel adjacencies.

As a next evaluation approach, we mapped the PacBio long-read data to custom genomes generated based on our reconstructions. Next, we checked, if PacBio long-reads span breakpoint junctions predicted by our curated novel-adjacency set. For 365/376 junctions, one or several supporting PacBio reads could be found (Supplementary Fig. 19b). For 9/11 cases, all tested junctions are covered by junction-spanning reads (Supplementary Fig. 19b). For case CT2, a single junction has no supporting PacBio read, but the fraction of junction-spanning reads is, for unknown reasons, very low for this sample in general, which also leads to large deviations in the SV calling when comparing Illumina-based and PacBio-based SV calling (Supplementary Fig. 1b). For the most complex case CT5, 10 junctions had no support by spanning PacBio reads. This indicates that, on a local level, the reconstruction still contains several errors. Reconstruction

errors might be attributed to different reasons, such as (i) false-positive novel adjacencies in the curated novel-adjacency set used for the reconstruction, (ii) missed complexity by undetected SVs or not considered SVs, (iii) in rare cases, the breakpoint position indicated by SVIM had an offset to the real breakpoint. This can also lead to deviations in the reconstruction process, especially when the fragments are very small, as in the case of CT5. Additionally, also the analysis of the Illumina GS coverage (Supplementary Fig. 10) indicates missing breakpoints because a few fragments which appear deleted based on coverage are still present in the reconstructed derivative chromosomes.

The last evaluation we performed is the reprocessing of the Hi-C data with custom genomes. This approach is similar to the recomposition of the Hi-C map described above but starts already at the level of the genomic sequence. We did not observe larger misjoins in cis parts of the custom Hi-C maps, which would usually be accompanied by a discontinuity of the Hi-C signal (Supplementary Data 2).

Discussion

The first step towards a better understanding of the implications of complex rearrangements on chromatin structure, gene regulation, and phenotype is the comprehensive detection of all breakpoints and reconstruction of shattered chromosomes. In previous studies of germline CGRs, breakpoints were detected using mate-pair sequencing, followed by filtering for sample-specific SVs and subsequent validation of breakpoints using PCR and Sanger sequencing[30,31]. Here, we combined Illumina and PacBio GS with Hi-C to identify breakpoints and resolve chromosomal rearrangements in 11 cases with CGRs. The majority of novel adjacencies with large distances between the breakpoints were detected by the Illumina-based caller, as well as the PacBio-based caller.

In general, GS methods suffer from a high false-positive rate when using the initial call set of common tools such as Delly[18] or SVIM[15]. In contrast to GS, Hi-C probes the 3D interactions of the genome and projects them onto a two-dimensional map[24]. We used this property of Hi-C for independent validation of large-scale SV-breakpoints[16] detected by Illumina and/or PacBio GS. Additionally, Hi-C is a powerful tool to order and orient scaffolds[32–34]. The disentangling of genomic rearrangements based on Hi-C was suggested for developmental diseases[17] and demonstrated in an automated manner in cancer cell lines[35]. The use of Hi-C for a manual reconstruction of derivative chromosomes was described recently for somatic chromothripsis induced in a cell line[19]. Here, we combined Hi-C with SV calls from short and long-read sequencing to reconstruct germline shattered chromosomes, which was possible for whole chromosomes in many instances. In the remaining cases, a reconstruction was possible to the level of large chromosomal blocks, which were grouped further to derivative chromosomes. By applying a permutation approach, a chromosome-wide reconstruction was achieved. However, this came with more uncertainty, because no direct evidence for the junction between chromosomal fragments was available, and the details of the junction remain unclear.

The approach presented here has some limitations. We analyzed only large novel adjacencies >100 kb or translocations and added afterward a few smaller novel adjacencies (1–100 kb) to the reconstruction. However, we did not consider the complexity of rearrangements from novel adjacencies with <1 kb between the fused positions. The presence of small rearranged fragments can be overlooked by Hi-C and make an evaluation difficult. For example, the local shattering of CT5 had such a high complexity that the reconstruction approach reached its limits. In some cases, Illumina- and PacBio-derived breakpoint positions disagreed by more than 50 bp. Especially when breakpoints are very close to each other, this can create ambiguity and alter the reconstruction. To match breakpoints in the reconstruction, we allowed a tolerance of <50 bp. With this tolerance, we obtained sometimes more than one junction per fragment end. We resolved these conflicts by removing the novel adjacency, which had the lowest evidence, i.e., was found by one technology only or had poor support in Hi-C. As the judgment of novel adjacencies based on Hi-C is difficult for small fragments, different reconstructions are possible, depending on which novel adjacencies are removed in case of ambiguity. The presence of some remaining reconstruction issues and missed novel adjacencies are indicated by mapping PacBio long-reads to custom genomes (Supplementary Fig. 18 and section 'Evaluation of the accuracy'), by the reconstructed Hi-C maps, as well as by fragments, which appear deleted based on coverage, but are still in the reconstruction (Supplementary Fig. 10). Additionally, copy number gains can create ambiguity in the reconstruction and imply chromosomal fragments with overlap, which would not be properly handled here. As Hi-C is based on short-read mapping, regions with low mappability, such as centromeres, have poor or missing Hi-C signals hampering the evaluation of rearrangements in these regions. The reconstruction of derivative chromosomes, as well as the assignment of haplotypes for phasing RNA-seq data, is based on the assumption that all large-scale rearrangements occurred in the same allele and were manifested in the germline. A useful extension of our approach could be the testing for polymorphisms. During the recomposition, the different pieces of the Hi-C map are assembled, but no additional normalization for the individual pieces, such as proposed in ref. 35 was implemented. The recomposition of the Hi-C map, as well as the permutation of possible reconstructions, was limited to fragments that cover a complete Hi-C bin (25 and 100 kb, respectively); smaller fragments were removed for these steps. An alternative approach without the issues of small fragments is the mapping of Hi-C reads to a custom rearranged genome as performed for the evaluation of our reconstructions (Supplementary Data 2).

Our curated set of novel adjacencies provided the basis for an in-depth analysis of the rearranged chromosomes, the distribution of the breakpoints, and their effect on gene expression. The breakpoints showed no preference for TAD boundaries, in contrast to what was reported for rearrangements which manifested over the course of evolution[23]. However, we did observe an enrichment of breakpoints for the B-compartment as well as for LADs. These transcriptionally inactive regions of the genome are located closer to the nuclear periphery and may thus be prone to damage from defective isolation of the nucleoplasm or being more tolerant towards rearrangements. It is important to note, that we are looking here only at a small number of samples, limiting the possibility of drawing general conclusions on enrichments of breakpoints with respect to genome organization.

The combination of Illumina GS, PacBio GS, and Hi-C allowed haplotyping of the patients' genomes. By phasing RNA-seq data, we were able to investigate the difference in RNA-expression profiles between the intact WT alleles and the shattered alleles. The majority of genes did not show a significant difference in the allelic balance, suggesting that many large-scale rearrangements had no effect on the investigated genes and cell type. However, we observed an enrichment of regulated genes within a region of 100 kb around breakpoints. At larger distances up to 1 Mb, the level of AIGs appeared still elevated, although not statistically significant. Most cases of altered gene regulation within 100 kb breakpoint distance were associated with breakpoints within the gene. While gene disruption is a likely scenario for downregulated genes, we observed upregulation for 50% of genes (11/22). For these cases, misregulation by e.g., enhancer adoption is a possible scenario, modulating the expression of intact transcript isoforms, as well as fusion transcripts in some cases as indicated by fusion transcript analysis of RNA-seq data.

The observation that a rather small fraction of genes close to a breakpoint showed allelic imbalanced expression has to be also seen in the context of the experimental setup of this study. The tested lymphoblastoid cell lines represent only a small fraction of potentially

regulated genes, and the number of observable genes is further reduced because phasing RNA-seq could only be done at SNP positions. During development and in all tissues/organs, this will be very different, thus dramatically increasing the number of potentially regulated genes. Nevertheless, the results are in line with a study in *Drosophila*[10] in which only a minority of genes showed deviations in expression in the presence of large-scale rearrangements in shattered balancer chromosomes. However, as in the *Drosophila* study, it has to be considered that the individuals studied here are survivors, i.e., this is a negative selection against more severe effects. In addition, we acknowledge that the cell lines studied do not reflect the complexities of an embryonic environment in which gene regulation happens at a very different level.

Methods

Subjects

From our in-house cohort of individuals with chromosomal rearrangements detected by karyotyping, we selected seven cases with >3 chromosomal rearrangements to be enrolled in this study. Through collaborative efforts, we obtained an additional four cases. In total, 11 constitutional CGRs were selected for this study. Informed consent to publish genomic and clinical data were obtained from all patients (or their legal guardian). The study adhered to the Declaration of Helsinki standards, and was approved by the internal Ethics Committee of the Institute for Human Genetics of Charité—Universitätsmedizin Berlin, Berlin, Germany. Peripheral blood lymphocytes were used for establishing lymphoblastoid cell lines by EBV transformation. Molecular cytogenetics investigations were performed before for CP2[36] and CT3[21]. For CT3, previous serial FISH mapping found one of the breakpoints to disrupt *GRIN2B*[21]. For CT2 a fibroblast cell line was established and characterized by molecular cytogenetic and mate-pair sequencing in previous studies[30,37].

Cell culture

LCLs were cultured in RPMI medium with 15% fetal bovine serum (FBS) and 1% pen-strep. Fibroblasts were cultured in DMEM with 10% FBS, 1% L-glutamine, and 1% pen-strep.

Illumina genome sequencing

Short-read Illumina whole-genome sequencing (GS; 30× coverage) was performed on DNA isolated from the cell lines. Sequencing was performed by Macrogen (South Korea) on Illumina HiSeq X machines with Illumina TruSeq PCR-free chemistry. After quality control (QC), reads were aligned to the GRCh37 reference genome with BWA-MEM (version 0.7.17)[38] duplicates were masked using SAMBLASTER (version 0.1.24)[39], and the reads were sorted and converted to BAM files using Samtools (version 1.9)[40]. SVs were detected using Delly (version 0.8.1)[18]. Coordinates of novel adjacencies were derived from the SV calls considering the classes DUP, DEL, INV, and BND. Bcftools (version 1.10.2)[40] was used for the processing of VCF files.

PacBio genome sequencing

We cultured 4×10^7 cells (LCLs and fibroblasts) for PacBio CLR GS, and high molecular weight (HMW) DNA (for >30 kb SMRTbell Libraries) was extracted using a smart DNA prep kit (Analytik Jena). Quality control step was performed using the DNF-467 Genomic DNA 50 kb Analysis Kit on a 5200 Fragment Analyzer system (Agilent). Library preparation: briefly, all samples were sonicated using the Megaruptor 3 shearing kit on the Megaruptor 3 instrument (Diagenode; parameters 20 μg HMW-DNA; Speed: 3). Purification step was performed with AMPure PB Beads (Ratio 0,46×). Library preparation QC was performed using the DNF-464 High Sensitivity Large Fragment 50 Kb kit. We used the kit SMRTbell Express Template Prep Kit 2.0 (100-938-900) and performed size selection using BluePippin Size-Selection System (Sage Science). Range selection mode "BPstart" 30,000 bp, "BPEnd" 80,000 bp with library input of 3–5 μg. Library sequencing was performed on Sequel II system (Pacific Biosciences).

The eleven PacBio CLR datasets were generated on a PacBio Sequel II machine. One sequencing run with a single SMRT cell was performed per patient with the exception of CT3 (2 SMRT cells). We aligned all datasets to the GRCh37 human reference genome using pbmm2 (version 1.3.0, parameters: -preset "SUBREAD", -median-filter), yielding alignment coverages between ~20× and 40× (Supplementary Fig. 15). Median read lengths varied between 8 kb for patient CT3 and 29 kb for patient CP4.

PacBio novel-adjacency calling and filtering

Novel-adjacency calls were produced with the SV caller SVIM[15] (version 1.4.1, parameters: -all_bnds -max_sv_size 5,000,000 -segment_gap_tolerance 300 -segment_overlap_tolerance 100 -zmws). In the all_bnds mode, SVIM collects all novel adjacencies indicated by PacBio read alignments considering adjacencies from translocations, deletions, inversions, interspersed, and tandem duplications. Simple insertions were not considered because the insertion of bases does not create any novel adjacencies between reference loci. Novel-adjacency calls from different reads were clustered using a hierarchical clustering approach based on the sum of distances between breakend positions.

To reduce the rate of false positives calls, we removed novel-adjacency calls matching at least one of the following criteria: (i) low number of supporting reads (below 5% of the mean alignment coverage); (ii) artificially high read coverage; (iii) at gaps in the reference genome; (iv) in proximity to segmental duplications; and (v) SV calls occurring in more than one sample.

In case, we detected a novel adjacency by both technologies at the same location in the Hi-C map with matching strand orientations, we selected the Illumina-based call for further downstream analysis when labeled as 'precise' by Delly. Otherwise, we used the PacBio-based call.

The PacBio-based novel adjacencies calls <100 kb are without filtering step (v).

Score-based filtering. To further reduce the false-positive rate, adjacencies supported by a low number of reads were removed. We applied sample-specific thresholds to accommodate for the large variance in alignment coverage across samples. To retain high sensitivity, a lenient threshold of 5% of the genome-wide average alignment coverage was used for each sample.

Coverage-based filtering. We computed the average alignment coverage in nonoverlapping genomic windows of 10 kb. Windows with an average coverage higher than three times the genome-wide average coverage were annotated as high-coverage regions, and novel adjacencies found in such regions were filtered out.

Gap-based filtering. To remove spurious novel adjacencies caused by the presence of gaps in the reference sequence, novel adjacencies with a distance of less than 10 kb to a reference gap were filtered out. Gap locations were detected using seqtk (version 1.3, https://github.com/lh3/seqtk, parameters: cutN -n 1000).

Duplication-based filtering. Due to their length and similarity, segmental duplication regions can confuse the read alignment algorithm leading to erroneous alignments. To remove unreliable novel-adjacency calls between related segmental duplication regions, novel adjacencies overlapping annotated segmental duplication regions (source: http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab) were filtered out.

Cohort-based filtering. Finally, the remaining novel adjacencies from all patients were merged, and similar adjacencies were clustered using

a breakend distance cutoff of 1 kb. Non-unique adjacencies, i.e., those present in more than one sample, were particularly common in the repetitive genomic regions close to the centromeres and telomeres and were enriched in false positives and population polymorphisms. To retain only genomic rearrangements unique to each patient, adjacencies present in more than one sample were filtered out.

It is noted that the filtering steps could potentially also remove true positive SV calls, e.g., genomic variation can also be close to segmental duplications[41].

Preparation of Hi-C libraries

Hi-C libraries were processed as described in the previously published in situ protocol and with minor modification using our in-house modified version[17]. Briefly, ~1 million cells were fixed in 2% formaldehyde, lysed, and digested overnight with DpnII enzyme (New England BioLabs, M0202). PCR amplification (4–8 cycles) using the NEBNext Ultra II Q5 Master Mix (New England BioLabs, M0544). PCR purification and size selection were carried out using Agencourt AMPure XP beads (Beckman Coulter, A63881). Libraries were deep sequenced (~360 million fragments) in 75 bp, or 100 bp paired-end runs on a NovaSeq6000 (Illumina). For each individual, the Hi-C library was created by pooling between two and four technical replicates generated from two different cell cultures, to ensure higher complexity of the sequencing library.

Hi-C bioinformatics analysis

Paired-end sequencing data were processed using the Juicer pipeline v1.5.6, CPU version[42] with BWA-*MEM* (version 0.7.17)[38] for aligning short reads to reference genome hg19. Alternative haplotypes were removed from hg19, and the sequence of the Epstein-Barr virus (NC_007605.1) was added. Replicates were merged by combining filtered and deduplicated read-pairs output from the Juicer pipeline. Juicer tools (version *1.7.5*)[42] was used to create hic-files for visualization and downstream analysis. Juicebox (Desktop version 1.8.8)[43] was used for the inspection of raw count maps, as well as maps normalized with Knight and Ruiz (KR) normalization[25,42,44] at different bin sizes. For the generation of Hi-C maps, we used read-pairs with mapping quality (MAPQ) ≥ 30. However, spotting genomic rearrangements, it can be helpful additions to generate and inspect Hi-C maps with lower, more permissive MAPQ thresholds.

For the display of Hi-C maps in figures, simple raw count maps were used to show ectopic Hi-C patterns. For visualization as heatmaps with a linear scale, high values were truncated to improve visualization. Novel-adjacency calls were overlayed with the Hi-C map as 2D annotation in Juicebox for visual inspection.

RNA-seq library preparation

RNA extraction was performed in all 11 samples using the RNeasy mini kit (Qiagen, Hilden, Germany). Poly(A) mRNA capture was performed using the KAPA mRNA HyperPrep Kit (KR1352—v5.17), and the RNA-seq was performed on a HiSeq4000 (Illumina) in three technical replicates (PE75, 50 million fragments per sample), except for CT5 for which two technical replicates were available.

DESeq2 analysis

Raw reads were mapped to the human genome build hs37d5 using STAR (version 020201)[45] and further filtered for a minimum mapping quality of MAPQ = 5.

Alignments were then further used to distinguish the wild-type (WT) allele and the rearranged allele (See Methods: Halpotyping and Phasing). This resulted in a table with read counts per gene and per allele for each replicate. DESeq2 (version 1.26.0)[29] was then applied for an adapted differential gene expression analysis by contrasting the read counts from the rearranged allele against the read counts from the WT allele taking all three replicates into account. We used an adjusted *P*-value < 0.05 and an abs(Log2FC) >1 to define allelic imbalance genes.

Sample CT2 was excluded from the analysis of the allelic imbalance genes due to a high base level of allelic imbalance genes (Supplementary Fig. 9C). Additionally, genes on the sex chromosomes were excluded from the analysis of the allelic imbalance. After the DESeq2 analysis, genes for which no padj-value could be computed were not considered further in the analyses.

For computing, the distance between a gene and a breakpoint, the distance between the TSS and the closest curated breakpoint was computed. In case a gene had alternative TSS, the most 5′ TSS was used for the distance computation.

Fusion transcript detection

We used the software Arriba (version 2.1.0)[46] with default parameters for the detection of fusion transcripts in RNA-seq. Each RNA-seq replicate was analyzed individually. Afterward, we projected the coordinates of the predicted breakpoints onto the Hi-C map and manually selected fusion transcript candidates, which were located at ectopic Hi-C interaction patterns associated with novel adjacencies or close by. For the downstream analysis, we kept only candidate genes, which were labeled as 'high confidence' by the Arriba tool in at least one replicate.

Haplotyping and phasing

Preparation of Hi-C reads for haplotyping. In order to use Hi-C reads for haplotyping, we separately mapped the Hi-C reads of different replicates to the human reference genome hg19. For this, we followed the practice suggested by HapCUT2[47]. We used BWA-MEM (version 0.7.17)[38] and Samtools (version 1.10))[40] for mapping single reads of Hi-C pairs and for sorting the aligned files, respectively. We used the Hi-C_repair tool developed in HapCUT2[47] and Samtools to combine single-read alignment files. We finally used the Picard tools (version 2.20.8-0) (https://broadinstitute.github.io/picard/) for marking duplicates. Finally, we merged the alignment files of different technical replicates into one Hi-C BAM file for each sample.

Haplotyping with small variants. PacBio, Illumina, and Hi-C data were integrated to compute the haplotypes. PacBio and Illumina reads were mapped to *the* GRCh37 reference genome (See sections:'PacBio genome sequencing' and 'Illumina genome sequencing', respectively). We used Illumina reads for variant calling. We parallelized the jobs over the chromosomes to speed up the processes. The following steps were performed on individual chromosomes. First, we called small variants (variants < 50 bp) using freebayes (version 1.2.0)[48]. At this step, we filtered out homozygous variants and those heterozygous variants with *a* calling quality of <30. We then extracted haplotype informative reads, the reads with at least two variants, from PacBio and Hi-C data using extractHAIRS module of HapCUT2. Afterward, Hi-C and PacBio informative reads were merged and fed to HapCUT2 for haplotyping. The HapCUT2 results were converted to the VCF format using Whatshap (version 0.18)[49]. In the next step, *we* tagged the PacBio reads with haplotypes and haplotype blocks using Whatshap. We then merged the variants and tagged PacBio reads of each sample into one phased VCF and one BAM file, respectively. After these steps, the variants are grouped into so-called phase sets.

Investigating the feasibility of integrating breakpoints into haplotyping. We clustered long-reads on each breakpoint into three groups, *namely* haplotype one (H1), haplotype two (H2), and unassigned reads (NoHapInfo). For this, we used PacBio reads haplotag, i.e., the reads tagged by the haplotypes. Supplementary Fig. 16 shows a screenshot of the genomics viewer IGV[50] at a breakpoint (sample: CT4, locus: chr2:140, 217, 329–140, 217, 537). The reads in H1, H2, and NoHapInfo groups are shown in brown, purple, and gray color, respectively. In addition, we analyzed the breakpoints using split-mapped PacBio

reads. We clustered the reads into rearranged (in short CGR), wild-type (in short WT), and undecided (NoBpInfo) groups (see Supplementary Fig. 16). We assigned a read into a WT group if the mapping quality ≥20 and the read spans ±90 bp of a breakpoint (dashed red lines in the example shown in Supplementary Fig. 16). We assign a read into a CGR group if *the* mapping quality ≥20 and the read-only pass one side of a breakpoint. In addition to that, a CGR read needs to be mapped to another region in the genome as supplementary alignment. We excluded the reads ending ±90 bp of breakpoints and assigned them to the NoBpInfo group.

There are nine possibilities for tagging the reads on the breakpoints of a phase set. Note that the phase sets may span several breakpoints. Each read can be tagged with H1 or H2, or NoHapInfo. Also, each read can also be assigned to CGR, WT, or NoBpInfo classes. Among the nine groups obtained by combining the two tags, the four groups of WT/H1, WT/H2, CGR/H1, and CGR/H2 are informative for the investigation of the feasibility of haplotyping using breakpoints. We observed that most of the reads cluster mainly into two of these four groups. They either cluster to WT/H1-CGR/H2 or WT/H2-CGR/H1 (See Supplementary Fig. 17 for an example shown for *sample* CT4). This supports that the two independent procedures of labeling the reads, i.e., haplotyping via small variants and the breakpoint analysis with split reads, are agreeing. Thus, we leveraged haplotyping using breakpoints.

Now, for each phase set, it has to be determined, which haplotype (H1 or H2) represents the rearranged allele. In many cases, the majority of reads votes for one scenario (e.g., WT/H1-CGR/H2). However, few reads may provide evidence for the opposite scenario (e.g., WT/H2-CGR/H1), leading to a conflicting situation. Conflicting reads might occur due to errors in haplotyping, mapping, or deviations from the assumptions that all breakpoints originate from the same haplotype.

In case of conflicting reads, the scenario with the majority of votes was selected. After these steps, all phase sets were labeled CGR and WT, respectively.

Expanding haplotyping to a set of chromosomes. The assumption here is that the chromosomal rearrangements originate exclusively from one allele, i.e., maternal or paternal. To expand the haplotyping across chromosomes, we combine all the CGR phase sets and WT phase sets, respectively, across the affected chromosomes.

Phasing RNA-seq reads using phased small variants. We used phased small variants for haplotyping of RNA-seq reads. To phase the RNA-seq, we used the variants located in gene bodies. We tagged the RNA-seq reads sampled from CGR or WT chromosomes. This step is implemented to be parallelized on samples and chromosomes. For that, we implemented a tool that provides *a* BAM file with tagged RNA-seq, a table providing the variants carried by genes and the read count for CGR or WT variants. We filtered *out* genes with less than 16 reads coverage on phased heterozygous variants.

Analysis of breakpoint signature

For each breakpoint, Illumina reads with a minimum of 10 soft-clipped bases and a mapping quality ≥20 that are located ≤10 bp away from the called location were fetched using the python package pysam (version 0.15.2). The reads were separated into left (L) and right (R) depending on the distance of their alignment to the breakpoint location. If no supporting clipped reads were found in one or both directions, the window was iteratively extended up to 10 kb. The genomic distance between the median clipped-positions of the L- and R-read-groups was computed, representing the amount of gained or lost material i.e., the InDel size at hg19 aligned breakpoints. Only breakpoints supported by 3 or more reads clipped at the same reference position for each side were included in the InDel analysis. All others are specified as NA. To compute the junction homology, the 50 bp consensus sequence of L/R-reads around the respective median clipped position was aligned to the 25 bp aligned consensus sequences of the L/R-reads at the target breakpoint. The alignment was performed using the pairwise2.align.localms function of the python package biopython (version 1.73), assigning a +2 for a match, −1 for mismatch or opening a new gap, and −0.1 for extending an existing gap. These sequences were chosen with respect to the individual breakpoint's strand annotation. For example: Considering a breakpoint with a "++"-annotation, the 50 bp consensus sequence of all L-reads around the median clipped position (i.e., 25 bp aligned sequence and 25 bp clipped) was aligned to the 25 bp aligned consensus sequence of all R-reads at the target breakpoint. A perfect alignment (score = 100) with start position 26 i.e., the start of the clipped sequence, was regarded as a blunt end breakpoint with no homology. Any perfect alignment with a shifted start position indicated a stretch of homology-based on the difference between the expected and observed alignment start position. Imperfect alignments were investigated for template-independent insertions by aligning 50 bp clipped consensus sequence of the initial breakpoint's L/R-reads to 25 bp aligned consensus sequence of the target breakpoint. The final homology was defined as the total length of junction homology and template-independent insertions. Finally, each breakpoint was manually investigated to confirm the computed homology, discarding breakpoints with questionable L- or R-support.

Statistical testing with empirical background models

The overall fractions of genomic features, shown as horizontal lines (expected value) in the Fig. 3 and Supplementary Fig. 4, were computed as the fraction of the genomic feature with respect to the length of all chromosomes, excluding gaps of the reference genome in the computation.

We implemented empirical background models to evaluate the association of breakpoints with features of the chromatin organization, repeats, as well as the enrichment of allelic imbalance genes proximal to breakpoints. For each type of background model, 1000 random configurations were generated. The *P*-value for enrichment was computed as the fraction of random configurations, which achieved a value greater or equal to than obtained by the original data. *P*-values were derived by aggregating the information from all cases, except for the analysis of allelic imbalance genes, where CT2 was excluded due to the high overall level of allelic imbalance genes. In the following, the different background models are explained in more detail.

(1) Novel adjacencies with random coordinates: For each set of novel adjacencies, random sets were generated by shifting the coordinates. All coordinates on the same chromosome were shifted by the same random offset. In case a coordinate was shifted outside of the chromosome, it was inserted again at the other side of the chromosome. Thus, the relative distances between breakpoints were maintained, except for coordinates, which had to be inserted again. Configurations, in which at least one coordinate overlapped a region marked as the gap in the reference genome, were not considered. This type of random model was used to evaluate the enrichment of breakpoints in TAD boundaries, A/B compartments, LADs, repeat regions, and the analysis of allelic imbalance genes in the proximity of breakpoints.
(2) Novel adjacencies with random connections: For the set of novel adjacencies from all cases, the connection between anchor points were permuted. Thus, the coordinates stayed the same, but the connection between breakpoints was altered. This permutation was used to evaluate A/B-compartment fusions and LAD/None-LAD fusions.
(3) As an alternative to (1) for the analysis of allelic imbalance genes in the proximity of breakpoints, we implemented a permutation of

the gene expression values and permuted the assignment of expression values to genes while the coordinates of the breakpoints were not changed.

Manual curation of novel adjacencies

In a manual curation step, novel adjacencies candidates from PacBio and Illumina-based SV calls were checked for Hi-C patterns of rearrangements in Juicebox[43]. SV calls that appeared isolated and unrelated to the complex rearrangements were not considered. In case PacBio and Illumina-based novel-adjacency candidates were overlapping or very close to each other, the Illumina-based novel adjacency was used for downstream analysis when it was labeled as 'precise' by Delly. Otherwise, the PacBio-based call was selected. Besides the coordinates, the strand information was also considered in the curation process, when possible.

Sometimes, a novel-adjacency call was also selected, without having clear evidence in Hi-C, when it was fitting to novel adjacencies in the proximity that had Hi-C support. This was especially the case in complex regions, such as occurring in CT4 and CT5, showing a high degree of shattering and local clustering of novel adjacencies. Here, novel-adjacency calls were also selected without having prominent Hi-C support, when they were localized in a shattered region. Also, in case the strand information was not matching the Hi-C pattern, it was, in some cases, selected with reservation. However, novel adjacencies were removed again, when they led to ambiguous edges in the reconstruction graph or to connected chromosomal components, which were not compatible with Hi-C later. For the removal of conflicting novel adjacencies, those were preferred, which had low support in Hi-C.

For case CT2, one chromosomal fragment was broken up manually in the reconstruction, because it led to the connection of chromosomal components, which were not compatible with the Hi-C grid (See Section reconstruction). The position of the breakpoint was approximated from Hi-C alone.

Generating scaffolds and derivative chromosomes

Manually curated large-scale novel adjacencies were used for the reconstruction of derivative chromosomes. Additionally, we added *a* few small-scale novel-adjacency calls (1–100 kb between the anchor points), when they were matching the anchor points of the curated large-scale novel adjacencies. We added the following numbers of small-scale novel adjacencies per case (CP2: 1, CT2: 1, CT3: 1, CT4: 3, CT5: 8, CT6: 2, CT7: 4) for the reconstruction.

In the first step, the breakpoints in the combined set were simplified, such that complementary breakpoints from different novel adjacencies, with almost identical coordinates, were adjusted to fit exactly. This shifting of breakpoints was done for distances below 50 bp and reduced the number of resulting fragments to avoid very small fragments. The resulting fragments can be connected at the start and at the end, except for telomeric fragments, which only have one connection possibility in our model. All novel adjacencies represent edges that connect fragments. Traversing the different paths of the graph yields the order and orientation of the fragments of the derivative chromosome. However, this requires that each fragment side has a maximum of one connecting edge, i.e. all paths are non-overlapping. In case, more than one connecting edge was found, the ambiguity was resolved manually by removing novel adjacencies from the initial call set. This was necessary for the following cases (CT2: 3 removed, CT4: 3 removed, CT5: 4 removed, CT6: 4 removed). In case, two novel adjacencies were compatible with the same fragment end, they were prioritized based on how well they were supported by Hi-C, if they were found by PacBio and Illumina-based callers and how well the strandedness agreed with the Hi-C pattern. Often, this information ruled out novel adjacencies, which initially only have been taken into the set with reservation.

In case connections are missing, only incomplete derivative chromosomes can be reconstructed. These scaffolds can be grouped further based on shared Hi-C signal.

Grouping scaffolds to derivative chromosomes

The reconstruction graph is sometimes incomplete for individual derivative chromosomes of complex cases (Fig. 2d), such that traversing the graph does not result in complete reconstructions. However, often still larger components, i.e., scaffolds can be derived. Afterward, these scaffolds can be grouped together by Hi-C based on the ectopic interactions between different scaffolds. For this task, we span a two-dimensional grid across the Hi-C map. The grid lines are defined by the breakpoint coordinates. The grid cells can be checked for ectopic interactions. The assumption is that fragments pairs with ectopic interactions are located in the same derivative chromosome. However, the other way around the assumption is not valid, because chromosomal fragments, which do not share any ectopic contacts, are not necessarily in different derivative chromosomes. The absence of ectopic interactions could as well result from large genomic distances between the fragments or the occurrence of different arms in the derivative chromosome. Nevertheless, the first rule is powerful enough to group almost all of the non-completed components into derivative chromosomes. Even though the order in the group is not necessarily clear, they still belong to the same derivative chromosome. The described approach can only work when the same chromosomal fragment is not part of different derivate chromosomes, which could happen in case of duplications or other copy number gains. In this case, the assumption that groups from different derivative chromosomes have mutually exclusive ectopic interaction patterns would be violated.

Recomposing Hi-C maps of reconstructed derivative chromosomes

The reconstruction of Hi-C maps was done based on the original Hi-C maps created for reference genome hg19. According to the order of the chromosomal fragments in the derivative chromosomes, the corresponding parts of the Hi-C map were extracted using the straw-library[43] and composed. In case the chromosomal fragment appeared in an inverted orientation, the extracted part of the original Hi-C map was inverted as well. In contrast to breakpoint coordinates, which are in base-pair resolution, the Hi-C map consists of larger bins. Therefore, the starts of chromosomal fragments were rounded up to the start of the next bin, and the end of the fragments was rounded down to the closest end of the bin. If the resulting piece of the Hi-C map was smaller than one bin, the chromosomal fragment was ignored in the reconstruction, thus small fragments could not be considered in the reconstruction.

The Hi-C map is an overlay of the signal from the WT allele and the rearranged allele. By the process of recomposing the Hi-C map, the Hi-C patterns originating from the rearranged allele are brought into order. However, at the same time, the Hi-C patterns from the WT allele become reshuffled, resulting in rearrangements patterns as observed before for the mutated allele. To mitigate artifacts from the WT allele caused by the recomposition, we subtracted from the recomposed map a control WT map, which was recomposed in the same manner. As the WT allele is expected to contribute roughly half of the signal in the Hi-C map, we scaled the control map to have 50% percent of the overall signal of the sample map. The main diagonal was excluded for the computation of the scaling factor. The aim of the subtraction is to enrich the signal from the rearranged allele, even it is not working perfectly, and artifacts may remain the in the map.

Permutation of reconstructed scaffolds to complete reconstructions

For some cases, it was only possible to reconstruct larger scaffolds that were not complete yet. However, by the use of Hi-C, the scaffolds could

be grouped and assigned to the derivative chromosome they belong to. Within these groups, the order and orientation of scaffold is unknown, because no direct sequencing-based support for a connection between the scaffolds is available. In this case, we applied a permutation-based approach to propose a draft of the derivative chromosome.

In case two scaffolds were grouped together, and both of them have a telomeric end, the scaffold was connected in the only possible manner. In case three or more scaffolds were grouped into a derivative chromosome, the two scaffolds with a telomere defined the ends of the derivative chromosomes, and for the remaining scaffolds, all possible orders and orientations were enumerated. By this, $(N-2)! * 2^{(N-2)}$ different combinations had to be tried out, where N is the number of scaffolds. This strategy was applied for groups of scaffolds with exactly two telomeric ends and group sizes >2. We limited so far the reconstruction to five components, i.e., two telomeric fragments and three middle parts. It is noted, that some scenarios, such as appending fragments to a telomeric end or the presence of duplicated genomic regions[19], were not considered in the current implementation.

In order to assess the plausibility of each solution, we visualized the corresponding recomposed Hi-C maps (See section 'Recomposing Hi-C maps') and evaluated the visualizations manually. A recomposed Hi-C map, which does not show any rearrangements, is assumed to indicate a proper reconstruction, and only one of the permutations should represent the correct or at least the best solution. An exception is recomposed Hi-C maps with very small scaffolds that contain only very small pieces of the Hi-C map, in the most extreme case, only one bin. Here, the differences between different solutions can be absent (1 bin) or very small, making the distinction of solutions difficult or impossible.

Additional to the manual inspection, we computed a simple score for each of the permutations. The associated recomposed Hi-C map can be separated into tiles (See Fig. 5), according to the underlying chromosomal fragments. We computed the score as the sum over all tiles defined by all non-redundant pairwise combinations of fragments.

The area of a tile in the Hi-C map is defined by the two fragments, start s and end e, describing the start bin and end bin of the corresponding fragment in the recomposed Hi-C map m. The subscore of a tile is computed by weighting the Hi-C signal at position ij with the distance of the pixel to the main diagonal $|i-j|$. These weighted Hi-C values are summed up in the four corners of the tile:

$$\begin{aligned}\mathrm{subscore}(s_1,e_1,s_2,e_2) = &\sum_{i=s_1}^{s_1+w-1}\sum_{j=s_2}^{s_2+w-1}|i-j|\cdot m_{ij} + \sum_{i=s_1}^{s_1+w-1}\sum_{j=e_2-w+1}^{e_2}|i-j| \\ &\cdot m_{ij} + \sum_{i=e_1-w+1}^{e_1}\sum_{j=s_2}^{s_2+w-1}|i-j|\cdot m_{ij} + \sum_{i=e_1-w+1}^{e_2}\sum_{j=e_2-w+1}^{e_2}|i-j|\cdot m_{ij}\end{aligned} \quad (1)$$

The window size w, was set to 5, or to $(e-s+1)$ in case the corresponding dimension of the tile was smaller than 5.

The rational of the score is that the parts of the Hi-C map with the largest values should be at the main diagonal or close by. In case of an erroneous reconstruction, high-intensity values will occur further away from the main diagonal, because reconstruction error should appear as rearrangement. Thus, the working assumption is that errors in the reconstruction lead to an increase in the score. For our cases, the solution with the lowest score was identical to the solution selected by the manual inspection.

Evaluation of the reconstructions of derivative chromosomes

We performed several evaluations of our reconstruction. We generated custom genomes as Fasta files for each case according to the respective reconstruction of derivative chromosomes. In the next step, we aligned PacBio long-reads to the corresponding custom genome and checked, how many reads spanned the junctions in a window of 100 bp around the junction Supplementary Fig. 19b. We also mapped the Hi-C data to custom genomes and inspected the Hi-C maps. The Hi-C data is more difficult to evaluate because the map is an overlay of the rearranged allele, which is now ideally in order, but the WT allele produces patterns of rearrangements. We tried to mitigate the effect of the WT allele, by subtracting a control sample which was scaled to 50% of the overall intensity on autosomes (neglecting the diagonal and first subdiagonal). This operation reduced the intensity, but some patterns remained Supplementary Data 2.

Additional software tools

BioRender (https://biorender.com/) was used to create schematics of cells in Fig. 1b. The Circlize package[51] was used to create circus plots. Color palettes were created with https://colorbrewer2.org and the Paletton online tool (https://paletton.com). The read coverage of Illumina GS reads for chromosomal fragments was computed using Bioconductor package bamsignals (https://bioconductor.org/packages/release/bioc/html/bamsignals.html). Bioconductor package GenomicRanges[52] was used to compute distances and overlaps between genomic intervals.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support this study are available from the corresponding authors upon reasonable request. Informed consents from patients do not cover the deposition of sequencing data from the patient samples. These data are available only upon request from S.M. (stefan.mundlos@charite.de). Data can be shared for research purposes with permission of the patient or his/her legal guardian. The gene annotation from Gencode (v19) was used. TAD annotations for hg19 were downloaded from the website of the 3D genome browser for cell line GM12878[53]. A/B-compartment annotation was taken from previously published work[25] (GEO GSE63525), the subcompartments were collapsed to A and B compartments. LAD annotation for T cells was taken from previously published work[26] (GEO GSE94971). A list of housekeeping genes was taken from previously published work[54]. Repeats were downloaded from UCSC genome browser[55] for hg19 via the table browser[56] (group: Repeats, track: RepeatMasker, table: rmsk).

Code availability

Code for the reconstruction approach: https://github.molgen.mpg.de/schoepfl/chromosome_reconstruction. Code for haplotyping and RNA-seq phasing: https://github.com/moeinzadeh/Chromothripsis_haplotyping.

References

1. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, 6537 (2021).
2. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
3. Spielmann, M., Lupianez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
4. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
5. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
6. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
7. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944.e922 (2017).

8. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320.e324 (2017).
9. Despang, A. et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
10. Ghavi-Helm, Y. et al. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282 (2019).
11. Zepeda-Mendoza, C. J. & Morton, C. C. The iceberg under water: unexplored complexity of chromoanagenesis in congenital disorders. *Am. J. Hum. Genet.* **104**, 565–577 (2019).
12. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
13. Cortes-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).
14. Chatron, N. et al. The enrichment of breakpoints in late-replicating chromatin provides novel insights into chromoanagenesis mechanisms. *bioRxiv* https://doi.org/10.1101/2020.07.17.206771 (2020).
15. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
16. Dixon, J. R. et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
17. Melo, U. S. et al. Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases. *Am. J. Hum. Genet.* **106**, 872–884 (2020).
18. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
19. Sidiropoulos, N. et al. Somatic structural variant formation is guided by and influences genome architecture. *bioRxiv* https://doi.org/10.1101/2021.1105.1118.444682 (2021).
20. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
21. Endele, S. et al. Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–1026 (2010).
22. Robson, M. I., Ringel, A. R. & Mundlos, S. Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. *Mol. Cell* **74**, 1110–1122 (2019).
23. Krefting, J., Andrade-Navarro, M. A. & Ibn-Salem, J. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol.* **16**, 87 (2018).
24. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
25. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
26. Robson, M. I. et al. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Res.* **27**, 1126–1138 (2017).
27. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
28. Ottaviani, D., LeCain, M. & Sheer, D. The role of microhomology in genomic structural variation. *Trends Genet.* **30**, 85–94 (2014).
29. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
30. Nazaryan-Petersen, L. et al. Multigenic truncation of the semaphorin-plexin pathway by a germline chromothriptic rearrangement associated with Moebius syndrome. *Hum. Mutat.* **40**, 1057–1062 (2019).
31. Kloosterman, W. P. et al. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep.* **1**, 648–655 (2012).
32. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
33. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
34. Marie-Nelly, H. et al. High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).
35. Wang, X. et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
36. Seidel, J. et al. A multiple translocation event in a patient with hexadactyly, facial dysmorphism, mental retardation and behaviour disorder characterised comprehensively by molecular cytogenetics. Case report and review of the literature. *Eur. J. Pediatr.* **162**, 582–588 (2003).
37. Borck, G. et al. Molecular cytogenetic characterisation of a complex 46,XY,t(7;8;11;13) chromosome rearrangement in a patient with Moebius syndrome. *J. Med. Genet.* **38**, 117–121 (2001).
38. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
39. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
40. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 2 (2021).
41. Mohajeri, K. et al. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016).
42. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
43. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
44. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2012).
45. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
46. Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
47. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
48. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv*, 1207.3907 [q-bio.GN] 2012 (2012).
49. Patterson, M. et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
50. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
51. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
52. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
53. Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 151 (2018).
54. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).

55. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
56. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
57. Shen, M. M. Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer Cell* **23**, 567–569 (2013).

Acknowledgements

We would like to thank the individuals and families for their collaboration and contribution to this project. We thank Volkmar Beensen and Gotthold Barbi for the collaboration on two published cases. We thank Michael Robson, Lila Allou, and Alexandra Despang for their comments on the manuscript. This work was supported by grants MU 880/16-1 from the Deutsche Forschungsgemeinschaft (DFG) to S.M and the Bundesministerium für Bildung und Forschung (BMBF) FKZ 031L0169B to S.M. and FKZ 031L0169A to M.V. M.S. is a DZHK principal investigator and supported by grants from the DFG (SP1532/3-1, SP1532/4-1, and SP1532/5-1) and the Deutsches Zentrum für Luft- und Raumfahrt (DLR 01GM1925).

Author contributions

R.S., U.S.M., M.V., and S.M. conceived the study. R.S. and U.S.M. performed the manual curation of novel adjacencies and downstream data analysis. R.S. processed Hi-C data, created chromosome reconstructions, and the derived Hi-C maps. U.S.M. performed cell culture experiments, DNA- and RNA-library preparations. U.S.M. and J.J. performed Hi-C experiments. R.S., U.S.M., M.V., and S.M. wrote the manuscript with contributions from other authors. Da.He. performed PacBio GS read alignment and SV calling and filtering. M.H. performed Illumina GS read alignment and SV calling. J.H. performed InDel and microhomology analysis of breakpoints, PacBio GS read alignment to custom genomes, and analysis of junction-spanning reads. E.C., M.K., and J.J. analyzed chromosomal rearrangements in individual Hi-C maps. H.M. and N.A. performed haplotyping of patient genomes and phasing of RNA-seq data. V.L. performed differential gene expression analysis. S.T. and provided multicolor FISH. De.Ho.,Y.D., L.F., P.C., D.S., A.V., O.Z., R.T., N.K., S.G., B.P., A.L.B., I.V., M.B., and N.T. contributed with sample collection and clinical evaluation. M.S. and V.K. contributed with samples and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34053-7.

Correspondence and requests for materials should be addressed to Martin Vingron or Stefan Mundlos.

Peer review information *Nature Communications* thanks Peter Park, Adam Ameur, Jan Korbel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Max Planck Institute for Molecular Genetics, RG Development & Disease, Berlin, Germany. [2]Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, Berlin, Germany. [3]Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Berlin, Germany. [4]CUBI—Core Unit Bioinformatics, Berlin Institute of Health, Berlin, Germany. [5]Charité—University Medicine Berlin, Berlin, Germany. [6]Laboratoire national de santé, Dudelange, Luxembourg. [7]UFR Des Sciences de Santé, INSERM-Université de Bourgogne UMR1231 GAD « Génétique des Anomalies du Développement », FHU-TRANSLAD, Dijon, France. [8]Unité Fonctionnelle d'Innovation diagnostique des maladies rares, FHU-TRANSLAD, CHU Dijon Bourgogne, Dijon, France. [9]Department of Genetics and Centres of Reference for rare disorders, developmental abnormalities and intellectual disabilities, FHU TRANSLAD and GIMI Institute, University Hospital Dijon, Dijon, France. [10]Department of Medical Genetics, University Hospital of Lyon, 69007 Lyon, France. [11]Medical Genetics, Department of Molecular Medicine, University of Pavia, Pavia, Italy. [12]Genetica Clinica, Dipartimento di Pediatria, Università di Padova, Padova, Italy. [13]Medical Genetics Unit, Meyer Children's University Hospital, Florence, Italy. [14]Medical Genetics Unit, University of Cagliari, Cagliari, Italy. [15]Praxis für Humangenetik, Kinderzentrum Dresden-Friedrichstadt, Dresden, Germany. [16]Department of Medical Genetics, University of Medical Sciences in Poznan, Poznan, Poland. [17]Department for Clinical Medicine, Aarhus University, Aarhus, Denmark. [18]Wilhelm Johannsen Center for Functional Genome Research, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark. [19]Institute of Human Genetics, University Hospitals Schleswig-Holstein, University of Lübeck and Kiel University, 23562 Lübeck, 24105 Kiel, Germany. [20]DZHK (German Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, 23562 Lübeck, Germany. [21]These authors contributed equally: Robert Schöpflin, Uirá Souto Melo. ✉e-mail: vingron@molgen.mpg.de; stefan.mundlos@charite.de

Printing copy of Publication 3: Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases.

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

ARTICLE

Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases

Uirá Souto Melo,[1,2,18] Robert Schöpflin,[1,2,18] Rocio Acuna-Hidalgo,[1,2,18] Martin Atta Mensah,[2] Björn Fischer-Zirnsak,[1,2] Manuel Holtgrewe,[2,3] Marius-Konstantin Klever,[1,2] Seval Türkmen,[2] Verena Heinrich,[4] Ilina Datkhaeva Pluym,[5] Eunice Matoso,[6,7] Sérgio Bernardo de Sousa,[6] Pedro Louro,[6,8,9] Wiebke Hülsemann,[10] Monika Cohen,[11] Andreas Dufke,[12] Anna Latos-Bieleńska,[13,14] Martin Vingron,[4] Vera Kalscheuer,[1] Fabiola Quintero-Rivera,[15] Malte Spielmann,[16,17,*] and Stefan Mundlos[1,2,*]

Genome-wide analysis methods, such as array comparative genomic hybridization (CGH) and whole-genome sequencing (WGS), have greatly advanced the identification of structural variants (SVs) in the human genome. However, even with standard high-throughput sequencing techniques, complex rearrangements with multiple breakpoints are often difficult to resolve, and predicting their effects on gene expression and phenotype remains a challenge. Here, we address these problems by using high-throughput chromosome conformation capture (Hi-C) generated from cultured cells of nine individuals with developmental disorders (DDs). Three individuals had previously been identified as harboring duplications at the SOX9 locus and six had been identified with translocations. Hi-C resolved the positions of the duplications and was instructive in interpreting their distinct pathogenic effects, including the formation of new topologically associating domains (neo-TADs). Hi-C was very sensitive in detecting translocations, and it revealed previously unrecognized complex rearrangements at the breakpoints. In several cases, we observed the formation of fused-TADs promoting ectopic enhancer-promoter interactions that were likely to be involved in the disease pathology. In summary, we show that Hi-C is a sensible method for the detection of complex SVs in a clinical setting. The results help interpret the possible pathogenic effects of the SVs in individuals with DDs.

Introduction

Over the last decade, the development and refinement of genomic technologies has paved the way for a major transformation regarding genotype-phenotype correlations in the field of human genetics.[1] Genome-wide methods such as microarrays and high-throughput sequencing techniques have improved the identification of genetic alterations involved in disease, increasing our understanding of genetic disorders.[2–5] Genetic variation occurs at various levels ranging from single-nucleotide variations (SNVs) and small indels to larger structural variants (SVs). SVs are a large class of genome alterations with a size >50 bp; this class includes deletions, duplications, inversions, insertions, and translocations.[6] SVs can be detected by various genetic and/or genomic screening tools that each hold certain advantages and disadvantages in regard to detection sensitivity and specificity. For instance, while array comparative genomic hybridization (CGH) can reliably detect large copy number changes, it does not reveal their precise genomic position and fails to detect copy-number-neutral variants, such as inversions and translocations. Short-read-based whole-genome sequencing (WGS) can in principle detect all SV types genome-wide, but its resolution and specificity is limited in repetitive regions of the genome. Furthermore, the detection of balanced SVs relies on the presence of chimeric reads spanning the breakpoint. Thus, detecting and resolving the layout of complex genomic rearrangements (CGRs) with multiple breakpoints can be challenging with WGS.[7,8]

Aside from the detection of SVs, their clinical interpretation is a largely unsolved problem, especially when they do not disrupt protein coding parts of the genome. Recent work has shown that SVs can disrupt the complex three-dimensional (3D) architecture of the human genome, causing position effects and thereby contributing to developmental disorders (DDs).[9] The development of high-throughput chromosome conformation capture (Hi-C)[10,11]

[1]Max Planck Institute for Molecular Genetics, RG Development and Disease, 13353 Berlin, Germany; [2]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, 13353 Berlin, Germany; [3]Berlin Institute of Health (BIH), Core Unit Bioinformatics, 10117 Berlin, Germany; [4]Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, 13353 Berlin, Germany; [5]Department of Obstetrics and Gynecology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA; [6]Medical Genetics Unit, Centro Hospitalar e Universitário de Coimbra, 3000-075 Coimbra, Portugal; [7]Center of Investigation on Environment Genetics and Oncobiology (iCBR-CIMAGO), Faculty of Medicine, University of Coimbra, 3000-548 Coimbra, Portugal; [8]Familial Risk Clinic, Instituto Português de Oncologia de Lisboa Francisco Gentil, 1099-023 Lisboa, Portugal; [9]Faculty of Health Sciences, Universidade da Beira Interior, 6201-001 Covilhã, Portugal; [10]Handchirurgie Kinderkrankenhaus Wilhelmstift, 22149 Hamburg, Germany; [11]kbo-Kinderzentrum München, 81377 München, Germany; [12]Institut für Medizinische Genetik und Angewandte Genomik, 72076 Tübingen, Germany; [13]Department of Medical Genetics, University of Medical Sciences in Poznan, 60-806 Poznan, Poland; [14]Centers for Medical Genetics GENESIS, Grudzieniec st, 60-601 Poznan, Poland; [15]Department of Pathology and Laboratory Medicine, UCLA Clinical Genomics Center, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA; [16]Max Planck Institute for Molecular Genetics, Human Molecular Genomics Group, 13353 Berlin, Germany; [17]Institut für Humangenetik Lübeck, Universität zu Lübeck, 23538 Lübeck, Germany
[18]These authors contributed equally to this work
*Correspondence: spielman@molgen.mpg.de (M.S.), mundlos@molgen.mpg.de (S.M.)
https://doi.org/10.1016/j.ajhg.2020.04.016.

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

revealed that chromatin interactions occur preferentially within defined and stable regions of the genome; these regions are known as topologically associating domains (TADs).[12,13] TADs are domains in the genome which are separated by insulating proteins (e.g., CTCF) and which build a framework for contacts of regulatory elements and genes. SVs can disrupt the TAD architecture and rewire the regulatory landscape of genes, which can lead to gene misregulation and cause DDs.[14–16] Therefore, genomic alterations leading to TAD disruption, abnormal chromatin interactions, and subsequent misregulation of gene expression are emerging as largely unexplored mechanisms involved in genetic disorders.[17]

Here we applied Hi-C to samples from nine individuals with DDs. For all of these cases, a candidate SV related to each individual had previously been detected via karyotyping and/or array CGH. Hi-C can detect and help resolve complex SVs in clinically available tissues and, in some cases, a plethora of CGRs was detected which had been overlooked with conventional molecular tools. Furthermore, we evaluated the changes of the TAD landscapes associated with the SVs in order to predict their disease-causing potential. Based on the changes of the 3D chromatin structure, we suggest putative disease-associated genes for several cases, and we derive hypotheses of the disease mechanism on a genetic level. This demonstrates that Hi-C can be a powerful tool for the identification and clinical interpretation of SVs in the context of DDs, helping improve the understanding of the mechanisms through which SVs can perturb development.

Material and Methods

Subjects and Ethics Approval

The study was performed with the approval of the Charité Ethics Committee, Berlin, Germany. All subjects, their parents, or their legal guardians provided informed and written consent for participation in this study. The clinical evaluation included medical history interviews, a physical examination, and review of medical records. Blood samples were obtained from each participating individual except for one fetus, and DNA was extracted using standard procedures. Additionally, skin biopsies were obtained from two individuals and amniocytes from one fetus.

Cytogenetic and Genomic Screening

Samples from individuals with DD (n = 7) were submitted for chromosome banding using the trypsin-Giemsa technique (Table 1). The detected chromosomal translocations were further confirmed through the use of fluorescence *in situ* hybridization (FISH) on metaphases. Copy number analysis on individuals with DD (n = 4) was carried out via array CGH using a whole-genome 1 M oligonucleotide array (Agilent; Table 1 and Table S1), or 2.6M Affymetrix CytoScan HD (Affymetrix). 1 M arrays were analyzed through the use of Feature Extraction v9.5.3.1 and CGH Analytics v3.4.40 or Cytogenomics v2.5.8.11 software, respectively (Agilent). The analysis settings were as follows: aberration algorithm—ADM-2; threshold—6.0; window size—0.2 Mb; filter—five probes, log2 ratio = 0.29. The 2.6M array data was analyzed using Chromosome Analysis Suite 4.0 (ChAS) version 3.3.0.139 (r10838). Copy number variants (CNVs) detected via array CGH were confirmed using qPCR.

Chromosomal analysis was performed on metaphases obtained from 72 h phytohemagglutinin (PHA) stimulated peripheral blood lymphocyte cultures according to standard procedures. Analysis of GTG-banded chromosomes was done at a resolution of 700 bands per haploid genome according to the International System for Human Cytogenomic Nomenclature (ISCN) 2016. Confirmatory FISH was performed on metaphases obtained from 72 h PHA stimulated peripheral blood lymphocyte cultures according to standard procedures.

WGS was also performed in blood from individuals DD1, DD3, and DD4; in lymphoblastoid cell lines (LCLs) from DD5–DD9; and in fetal amniocytes from individual DD2 in order to validate the SVs' breakpoints (Table 1 and Tables S1 and S2). Sequencing was performed using Macrogen on Illumina HiSeq X machines with Illumina TruSeq PCR-free chemistry. After quality control, reads were aligned to the GRCh37 sequence (hs37d5.fa) with BWA-MEM,[18] duplicates were masked using SAMBLASTER,[19] and the reads were sorted and converted to BAM files using Samtools.[20] SVs were detected using Delly,[21] PopDel,[22] and ERDS.[23] Phenotype based exome analysis was performed using Exomiser[24] and MutationDistiller.[25] All rare (MAF < 0.01) heterozygous loss-of-function variants are shown in Table S3.

Cell Culture

Fibroblast cell lines were established from skin biopsies of two individuals with DD and one healthy control by following a standard procedure. Fibroblasts were cultured in Dulbecco's Modified Eagle Medium (DMEM; Thermo Fisher Scientific) supplemented with 15% fetal bovine serum (FBS; Thermo Fisher Scientific), 1% L-glutamine (Thermo Fisher Scientific), and 1% penicillin-streptomycin (Thermo Fisher Scientific). Amniocytes from one individual with DD were cultured in AmnioMAX II Complete Medium (Thermo Fisher Scientific). LCLs were established by Epstein-Barr virus (EBV) transformation of leucocytes from peripheral blood samples of individuals with DD (n = 6) and of one control. LCLs were cultured in Roswell Park Memorial Institute (RPMI) medium (Thermo Fisher Scientific) with 15% fetal calf serum (FCS) and 1% penicillin-streptomycin. Fibroblasts were grown to confluence prior to fixation for the preparation of the Hi-C libraries, whereas LCLs and amniocytes were cultured until the plateau phase of growth was reached.

Preparation of Hi-C Libraries

Hi-C libraries were processed as described in the previously published *in situ* protocol.[11] In brief, ~1 million cells were fixed in 2% formaldehyde, lysed, and digested overnight with *DpnII* enzyme (New England BioLabs, R0543). Digested DNA ends were marked with biotin-14-dATP (Thermo Fisher Scientific, 19524016) and ligated overnight using T4 DNA ligase (New England BioLabs, M0202). Formaldehyde crosslinking was reversed by incubation in 5 M NaCl for 2 h at 68°C, followed by ethanol precipitation. Covaris (S-Series 220) was used to shear the DNA to fragments of 300–600 bp for library preparation, and biotin-filled DNA fragments were pulled down using Dynabeads MyOne Streptavidin T1 beads (Thermo Fisher Scientific, 65602). The DNA ends were subsequently repaired using T4 DNA polymerase and the Klenow fragment of DNA polymerase I (New England BioLabs, M0203 and M0210) and phosphorylated with T4 Polynucleotide Kinase NK (New England BioLabs, M0201). The DNA was further

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

Table 1. Overview of Our Cohort with Nine Individuals Presenting with DDs and Their Respective SVs

Cytogenetics/Cytogenomics Results Prior to This Study									WGS		Hi-C	
Subjects	Detection Tool	Structural Variant				ClinGen CNVPathogenicity Calculator		Case SolvedPrior to Hi-C?	-		-	
-	-	ISCN 2016 Nomenclature	Type	Size	Inheritance	Prediction	Score	-	Detectedthe SV?*	Cell Type	SV Is Visible?	Case Solved Using Hi-C?
DD1	array CGH	arr[GRCh37] 17q24.3 (67956481_ 70087077)x3	duplication	2.1 Mb	*de novo*	Pathogenic	2.35	yes	yes	fibroblasts	yes	NA
DD2	array CGH	arr[GRCh37] 17q24.3 (67442273_ 70559193)x3	duplication	3.1 Mb	*de novo*	likely pathogenic	0.9	no	no	amniocytes	yes	no
DD3	array CGH	arr[GRCh37] 17q24.3q25.1 (68620187_ 71083594)x3	duplication	2.4 Mb	NR	VUS	−0.3	no	no	fibroblasts	yes	yes
DD4	karyotyping	46,XX,t(5;18) (q31.1;q12.3)	reciprocal translocation	NA	*de novo*	NA	NA	yes	yes	LCLs	yes	NA
DD5	karyotyping	46,XY,t(14;20) (q12;q13.2)	reciprocal translocation	NA	*de novo*	NA	NA	no	no	LCLs	yes	yes
DD6	karyotyping	46,XX,t(3;5) (q24;q14.3)	reciprocal translocation	NA	NR	NA	NA	no	yes**	LCLs	yes	yes
DD7	karyotyping	46,XY,t(2;4) (p12;q34)	reciprocal translocation	NA	*de novo*	NA	NA	no	yes	LCLs	yes	no
DD8	karyotyping	46,XY,t(7;8) (p12;q22.1)	reciprocal translocation	NA	NR	NA	NA	no	yes	LCLs	yes	no
DD9	karyotyping	46,XX,t(2;7) (q33.1;p21)	reciprocal translocation	NA	*de novo*	NA	NA	no	yes	LCLs	yes	no

Legend: DDs—developmental disorders; SV—structural variant; CNV—copy number variant; Hi-C—high-throughput chromosome conformation capture; CGH—comparative genomic hybridization; ISCN—International System for Human Cytogenomic Nomenclature; LCLs—lymphoblastoid cell line; WGS—whole-genome sequencing; NR—not reported; NA—Not applicable; *SV callers; **One or more breakpoint(s) missed by callers.

prepared for sequencing by ligating adaptors to the DNA fragments, using the NEBNext Multiplex Oligos for Illumina kit (New England BioLabs, E7335 and E7500). Indexes were added via PCR amplification (4–8 cycles) using the NEBNext Ultra II Q5 Master Mix (New England BioLabs, M0544). PCR purification and size selection were carried out using Agencourt AMPure XP beads (Beckman Coulter, A63881). Libraries were deep sequenced (~370 million fragments) in a 75 bp paired-end run on a HiSeq4000 (Illumina). For each individual, the Hi-C library was created by pooling a total of four technical replicates generated from two different cell cultures in order to ensure higher complexity of the sequencing library.

Hi-C Bioinformatics Analysis

Paired-end sequencing data were processed using the Juicer pipeline v1.5.6, CPU version.[26] The pipeline was set up with BWA v0.7.17[27] for aligning short reads with BWA-MEM to the reference genome hg19. Alternative haplotypes were removed from hg19, and the sequence of the Epstein-Barr virus (NC_007605.1) was added. Replicates were merged by combining the output files of the Juicer pipeline containing the information from filtered and deduplicated read-pairs. Juicer Tools[26] was used to create raw count maps, as well as maps normalized with Knight and Ruiz (KR) matrix balancing;[11,26,28] we inspected both. In addition to generating Hi-C maps binned to a regular grid, we also generated restriction-fragment-based Hi-C maps for further inspection. For the generation of Hi-C maps, we used read-pairs with mapping quality (MAPQ) ≥ 30. However, for spotting genomic rearrangements, it can be helpful to also generate and inspect Hi-C maps with lower, more permissive MAPQ thresholds. In some cases, such as copy number variations, the KR normalization can be disadvantageous because the algorithm assumes equal visibility of all loci,[11] and that is not the case when the number of copies differs locally. In case of duplications, the algorithm would tend to reduce the signal of the duplicated region to compensate for the additional allele. Due to this, we directly used raw count maps for the display of duplication events. This has the disadvantage that locus-specific biases, which are addressed by the KR normalization, are still contained in the map, but has the advantage that the copy number variation becomes more clearly visible (see also Supplemental Notes).

In order to generate Hi-C maps with reduced sequencing depth, we extracted the same number of reads for each replicate from the FASTQ files, such that the number of extracted read-pairs summed to 1 M, 10 M, 50 M, and 100 M, respectively. These subsampled FASTQ files were processed as described above.

Genome-wide Hi-C maps, inter-chromosomal Hi-C maps, and symmetric intra-chromosomal Hi-C maps were visualized using

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

Juicebox (Desktop version 1.8.8).[29] For specific loci of interest, intra-chromosomal Hi-C maps were visualized as a heatmap of the upper triangle matrix rotated by 45 degrees. For this type of visualization, very high values were truncated to improve the display of smaller count values. We used ChIP-seq ENCODE data for CTCF (AG09309; human toe fibroblast from an apparently healthy adult), H3K27ac, H3K4me1, and H3K4me3.[30]

Criteria for Selecting Candidate Genes Related to the Individual's Phenotype

Chromosomal alterations can move genes or regulatory elements into new regulatory landscapes. When an SV disrupts the integrity of a TAD, new hybrid domains may emerge between the TAD boundaries adjacent to the breakpoints, fusing together parts from different regulatory landscapes. Depending on the nature of the SVs, these hybrid domains emerge as neo-TADs (duplication), fused-TADs (deletion), or shuffled-TADs (inversion/translocation).[9] By these events, new enhancer-promoter interactions (EPIs) as well as loss of EPIs are possible. Based on the gain or loss of EPIs, we applied the following two criteria to select candidate genes putatively linked to the disease, depending on the individual's phenotype: (A) the gene was already associated with a neurodevelopmental phenotype or limb phenotype, described in Online Mendelian Inheritance in Man (OMIM) or (B) a phenotype similar to the individuals' symptoms was reported from an animal model (*Mus musculus*) as retrieved from specific databases (Mouse Genomic Informatics and Mouse Phenotype). We selected candidate genes related to the individual's phenotype when fulfilling both criteria listed above.

We applied the new ClinGen CNV Pathogenicity Calculator to evaluate the pathogenicity of all duplication cases (individuals DD1–DD3). The scores and criteria were the following:

- For DD1:
  - 1A. Contains protein-coding or other known functionally important elements (assigned points: 0)
  - 2A. Complete overlap; the TS gene or minimal critical region is fully contained within the observed copy number gain (assigned points: 1)
  - 3A. 0–34 genes (assigned points: 0)
  - 4A. … the reported phenotype is highly specific and relatively unique to the gene or genomic region (assigned points: 0.9)
  - 5A. Use appropriate category from *de novo* scoring section in Section 4 (assigned points: 0.45)
  - Prediction: pathogenic (total score: 2.35)
- For DD2:
  - 1A. Contains protein-coding or other known functionally important elements (assigned points: 0)
  - 2H. HI gene fully contained within observed copy number gain (assigned points: 0)
  - 3A. 0–34 genes (assigned points: 0)
  - 4A. … the reported phenotype is highly specific and relatively unique to the gene or genomic region. Confirmed *de novo*: 0.45 points (assigned points: 0.45)
  - Observed copy number gain is *de novo*
  - 5A. Use appropriate category from *de novo* scoring section in Section 4. (assigned points: 0.45)
  - Prediction: likely pathogenic (total score: 0.9).
- For DD3:
  - 1A. Contains protein-coding or other known functionally important elements (assigned points: 0)
  - 2H. HI gene fully contained within observed copy number gain
  - 3A. 0–34 genes (assigned points: 0)
  - 4D. … the reported phenotype is *not* consistent with the gene/genomic region or not consistent in general (assigned points: −3)
  - 5A. Use appropriate category from *de novo* scoring section in Section 4
  - Prediction: variant of uncertain significance (total score: −0.3).

Real-Time Quantitative (qPCR and RT-qPCR) Analyses

qPCR was performed to validate small duplications detected by using Hi-C in DNA from individual DD3. We designed two primer pairs as calibrators outside the duplicated region and three located within each duplication. RT-qPCR was performed to check *MEF2C* expression in DD6. RNA was extracted from LCLs (n = 5) by using RNeasy Mini Kit (QIAGEN). Total RNA (1 μg/μL) was reverse-transcribed into cDNA with oligo(dT) primers by using SuperScript IV First-strand Synthesis System (Thermo Fisher). qPCR and RT-qPCR were performed using the PowerUp SYBR® Green Master Mix (Thermo Fisher) and subjected to QuantStudio 6 Flex Real-Time PCR System, 384-well (Applied Biosystems). Copy number and gene expression were calculated using the $2^{-\Delta\Delta CT}$ method.[31] Each experiment was performed once with three technical triplicates per sample.

Results

Cohort of Individuals with DD and Detection of SVs

We generated genomic data from a cohort of nine individuals with DD. Based on initial diagnostic array CGH or karyotype results, at least one candidate SV had been detected in each individual (Table 1). The detected SVs were mostly *de novo* events with high impact on genomic structure and were therefore likely pathogenic, although the molecular mechanism of disease in most of the cases was unclear (seven out of nine). Clinical aspects and their respective SVs are described in detail in Table S1. Three out of nine affected individuals had microduplications (> 2 Mb) involving 17q24, and six had reciprocal chromosomal translocations. The SV breakpoints of all cases were validated using WGS. We performed Hi-C in all nine samples using the available clinical cells for each case (e.g., fibroblasts, amniocytes, or LCLs). Samples were sequenced for, on average, 370 million reads; and ∼160-250 (∼55%) million reads passed the quality control within the Juicer pipeline per sample (Figure S1).

Hi-C Reveals the Complex Nature of Duplications at the SOX9 Locus

We investigated three individuals (DD1, DD2, and DD3) with microduplications at the SOX9 locus on chromosome 17 (17q24.3). DD1 was an individual with Cooks syndrome (MIM: 106995), a condition previously described by us as being caused by microduplications containing enhancers upstream of the SOX9 locus but not the *SOX9* gene itself, and the neighboring *KCNJ2* gene (Table 1).[15,32] The microduplication in DD2 (3.1 Mb) encompassed *SOX9*, *KCNJ2*, and *KCNJ16*, as well as parts of *MAP2K6*. This

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

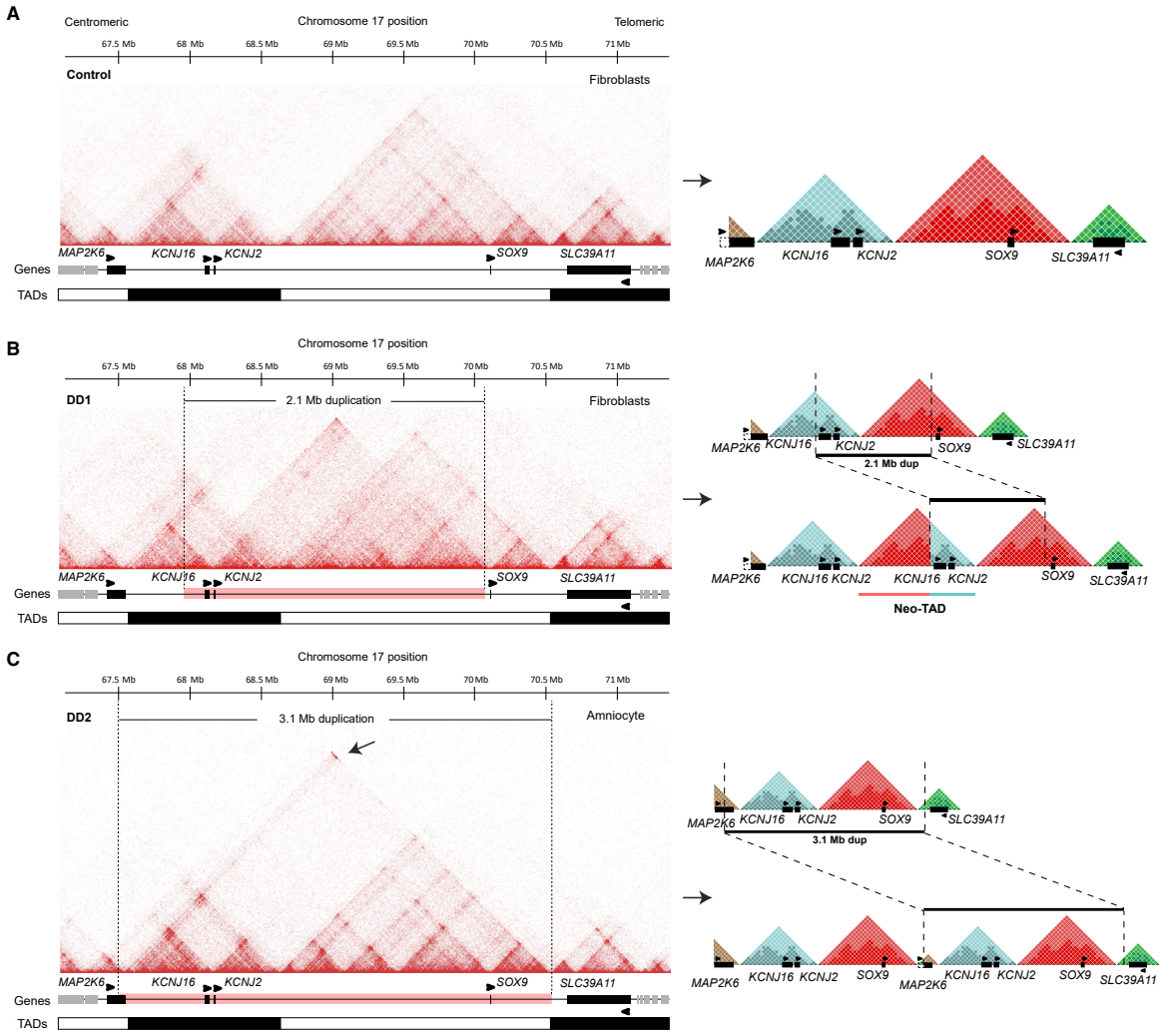

Figure 1. Hi-C Reveals Disruption of 3D Chromatin Folding Resulting from SVs at the 17q24 Region
(A) Hi-C map of a control sample (skin fibroblast; 10 kb resolution; raw count map) from an unaffected individual, showing the 3D landscape of the 17q24 region. TADs are represented by white and black bars on the track below. On the right: schematic representation of TAD structure on the 17q24 region (*MAP2K6* TAD in brown; *KCNJs* TAD in light blue; *SOX9* TAD in red; *SLC39A11* TAD in green).
(B) Hi-C map from fibroblasts of DD1 with a 2.1 Mb tandem duplication (chr17:67,958,880–70,085,143; GRCh37/hg19; highlighted in pink in the gene track below the respective map). Note that *SOX9* is not included in the duplication. The duplication induces the formation of a neo-TAD containing *KCNJ2* and *KCNJ16* and known regulatory elements of the *SOX9* locus, leading to ectopic expression of *KCNJ2*.
(C) Hi-C map from DD2 (amniocytes) showing a 3.1 Mb tandem duplication (highlighted in pink) on 17q24.3 (chr17:67,441,705–70,564,888), probably associated with the individual's phenotype. Note that the TSS of *MAP2K6* and *SLC39A11* are not included in the duplication. The underlying SV enables contacts between genomic regions at the margins of the duplication chr17:67,440,000–67,560,000 and chr17:70,545,000–70,565,000 (highlighted by an arrow), neither region containing ORFs, while the TADs of *SOX9* and *KCNJ* genes are copied as a whole.

individual had a severe heart defect, skeletal anomalies, and hydrops fetalis. The duplication (2.4 Mb) in individual DD3 is shifted to the telomeric side, including *SOX9* and the next flanking gene *SLC39A11*, but not *KCNJ2*. Individual DD3 had psychomotor developmental delay, microcephaly, and mild intellectual disability. Thus, in spite of their partially overlapping duplications, the outcome was very different, resulting in distinct phenotypes. We employed Hi-C in these individuals to investigate the effects of the rearrangements on chromatin structure in order to better understand their very different effects on development.

First, we performed Hi-C in fibroblasts from a healthy individual to serve as a control (Control; Figure 1A). At the locus of the *SOX9* gene, several well-defined TADs are visible in the Hi-C map. The TAD containing *SOX9* is

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

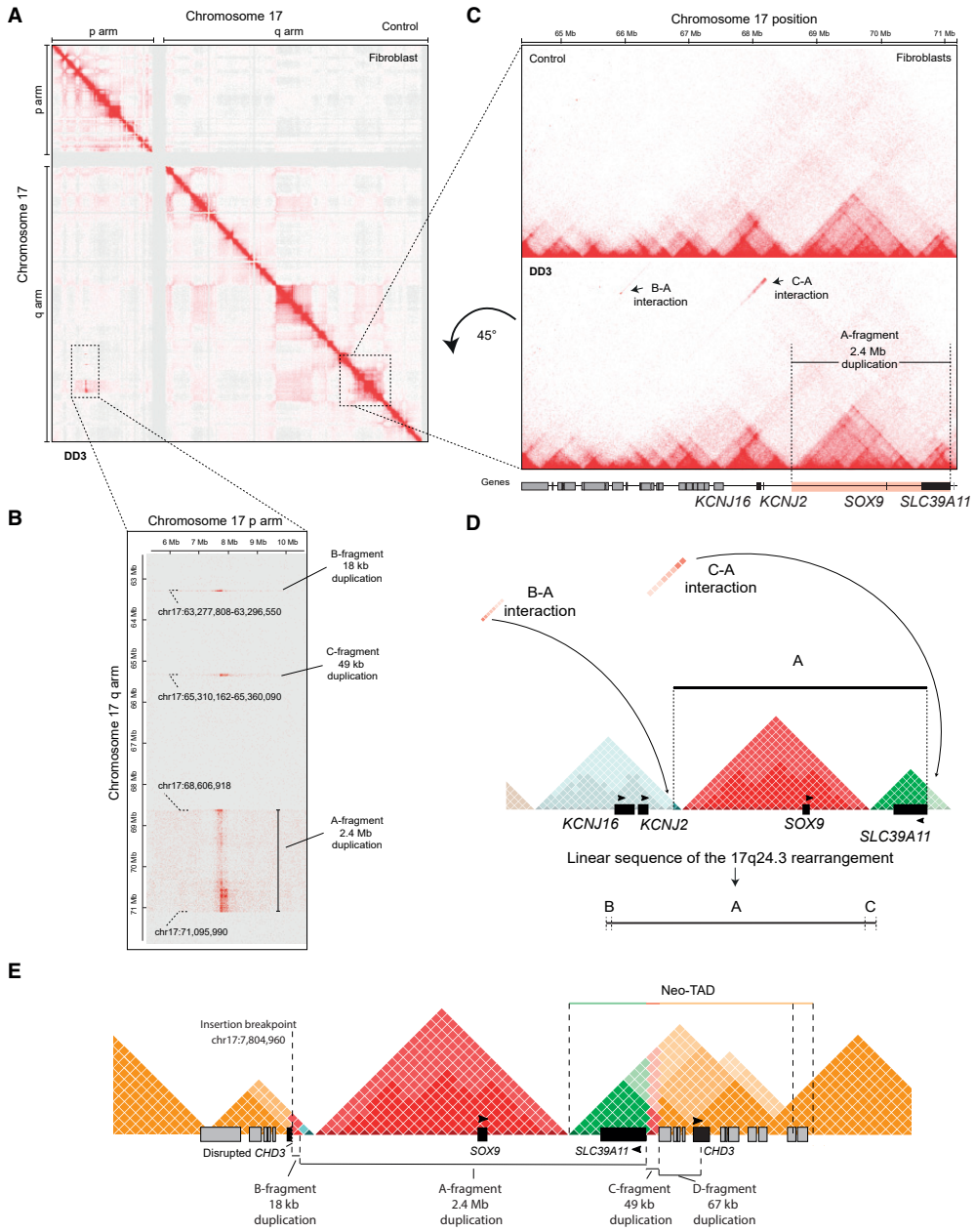

Figure 2. Hi-C Reveals a Complex Rearrangement Composed of Several SVs in an Individual with DD

(A) Hi-C map from DD3 fibroblasts reveals a complex rearrangement involving p- and q-arms of chromosome 17 (500 kb resolution; raw count map).

(B) Zoom-in on the ectopic signal of chr17 showing the large 2.4 Mb duplication (A-fragment), plus two other duplications (18 kb, B-fragment; and 49 kb, C-fragment). Note: Based on the orientation of the gradient of the ectopic Hi-C signal (fades away in horizontal direction, along the p-arm), we could infer that the three duplications from the q-arm are inserted on the p-arm (insertion breakpoint: chr17:7,804,960; for details see Figure S3).

(C) Hi-C cis-map of the 17q24 region (control and DD3) showing the interaction of the small duplications (B- and C-fragments) with the large one (A-fragment). Note that the *SLC39A11* ORF is part of the duplication (25 kb resolution; raw count map).

(D) Schematic view of the 17q24 rearrangement containing the B-A-C rearranged fragments. The 18 kb duplication is inserted upstream of the 2.4 Mb duplication, in front of the SOX9 TAD boundary (B-A interaction), and the 49 kb duplication (C-fragment) is inserted downstream of the large duplication, within the SLC39A11 neo-TAD (C-A interaction).

(legend continued on next page)

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

delimited by boundaries insulating it from the neighboring TADs and contains no further genes. The TAD flanking the *SOX9* TAD on the left (centromeric direction) contains the potassium channel genes *KCNJ2* and *KCNJ16*, while the TAD flanking the *SOX9* TAD on the right (telomeric direction) contains the gene *SLC39A11*.

The individual with Cooks syndrome (DD1) presented with hypoplasia and aplasia of nails and distal phalanges, similar to the previously described cases,[28] and a 2.1 Mb heterozygous duplication at the 17q24.3 locus had been detected initially using array CGH (Figure S1C and Table 1). Hi-C performed in fibroblasts showed a gain of new chromatin interactions at the SOX9 locus not comprising *SOX9* but including *KCNJ2* and *KCNJ16* (Figure 1B). Hi-C showed a strong interaction between the beginning and end of the duplicated region; this indicates that the 2.1 Mb duplication was in tandem (Figure 1B). The data show the emergence of a new chromatin domain (neo-TAD), which is visible as a new chromatin interaction superimposed on the wild-type SOX9 and KCNJs TADs (boxed region in Figure 1B). The newly formed TAD contains copies of *KCNJ2* and *KCNJ16*, as well as regulatory elements of the SOX9 TAD, which can now ectopically interact with each other (Figure 1B, right panel). As shown previously, the *SOX9* enhancers in the neo-TAD are capable of driving *KCNJ2* misexpression in a SOX9-like fashion in the developing limb, leading to hypoplasia and/or aplasia of nails and distal phalanges.[15]

Individual DD2 was a fetus with severe heart defects, skeletal anomalies, and hydrops fetalis (Table 1 and Table S1). A diagnostic array CGH revealed an ~3.1 Mb duplication (Figure S1D and Table 1), which included the complete SOX9 TAD and the neighboring KCNJ2/16 TADs, as well as parts of *MAP2K6* but not the transcription start site (TSS). Hi-C generated from amniocytes revealed that the duplication was in tandem (Figure 1C). A new chromatin interaction (neo-TAD) occurred between the centromeric margin of the duplication (estimated coordinates, chr17:67,440,000–67,560,000; ~120 kb) and the telomeric margin (estimated coordinates, chr17:70,545,000–70,565,000; 20 kb), but neither region contains genes, except for the remaining truncated region of *MAP2K6* (arrow on Figure 1C). The SOX9 and KCNJs TADs genes were copied as a whole including their boundaries, thus preserving their integrity. The effect of this 3.1 Mb microduplication is thus increased gene dosage without disruption of any regulatory landscape. Based on a gene dosage sensitivity tool (ClinGen), pure dosage effects of this gene are not expected to result in significant phenotypes. Inward-rectifier potassium channels (IRK) such as *KCNJ2* are important regulators of resting membrane potential and cell excitability. Dominant-negative mutations in *KCNJ2* result in Andersen syndrome (MIM: 170390), a condition characterized by periodic paralysis, cardiac arrhythmias, and dysmorphic features.[33] In addition, autosomal dominant atrial fibrillation (MIM: 613980) has been associated with activating mutations in *KCNJ2*.[34] Duplications have not been reported. It is possible that an increased gene dosage of *KCNJ2* leads to fetal cardiac arrhythmias and thus to hydrops. However, this is unlikely given the absence of arrhythmias in all the ultrasounds; this also would not explain the heart malformation. We conclude that while Hi-C provided more information on the 3D configuration of the genome of individual DD2, it still does not fully explain the clinical findings observed in this case.

In the third case with duplication at the SOX9 locus (DD3), array CGH detected an ~2.4 Mb duplication (Figure S2A) that included *SOX9* and the flanking *SLC39A11* gene, a presumed cellular zinc transporter. G-banded karyotype detected additional material on the short arm of chromosome 17 (46,XX,add(17)(p13.1); Figure S2B), which was later confirmed via FISH to be an insertion of the 2.4 Mb duplication (Figure S2C). The resulting rearrangement is described as 46,XX,add(17)(p13.1).ish der(17)ins(17)(p13.1) (RP11-84E24+).arr[GRCh37] 17q24.3q25.1(68620187_71083594) x3. Hi-C generated from DD3 fibroblasts revealed a gain of a new chromatin contact on the cis-map of chromosome 17 (Figure 2B) resulting from physical interaction between the 2.4 Mb fragment from the q-arm (here named A-fragment) with the p-arm of the same chromosome (Figure 2B). Hi-C also detected two small duplications from the 17q24.3 locus of 18 kb (named B-fragment) and 49 kb (named C-fragment), all validated through the use of WGS and qPCR (Figure 2C; Figure S2D). The 2.4 Mb duplication was poorly visible in the DD3 Hi-C map (Figure 2D), but it became more apparent after we subtracted a map from a reference sample (Figure S2E). Hi-C revealed that the 18 and 49 kb duplications were inserted up- (B-A interaction) and downstream (C-A interaction) of the 2.4 Mb duplication, and this was further validated using WGS (Figure S3). The linear sequence of the rearranged fragment is B-A-C (Figure 2C and 2D).

Hi-C indicated that the B-A-C fragment was inserted in 17p13 (Figure 2B). The region on 17p13 is gene dense with small chromatin domains without clear boundaries (Figures S4). The inserted fragment from the long arm disrupted the *CHD3* open reading frame (ORF) (Figure S3A). However, further analysis revealed a third duplication of 67 kb (D-fragment; validated by qPCR, Figure S4), containing four genes including a complete copy of *CHD3* (Figure S5). To unravel the complex 3D landscape of the nested rearrangement, we produced a schematic linear map of the region (Figure 2E). This reconstructed map suggests a TAD-fusion of the remaining SLC39A11 TAD (green) with the duplicated C-fragment (red) plus

(E) Schematic representation of the 3D chromatin landscape on the der(17p13) region. Hi-C maps helped to unveil the derivative 3D landscape and the order of fragments in a linear sequence, which consists of a neo-TAD containing *SLC39A11*, *CHD3*, and several other genes.

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

the 17p13 region (including the D-fragment; yellow), permitting ectopic interaction of several genes and enhancers (Figure S6). Lastly, Hi-C also revealed that the duplicated *CHD3* loses contact with its wild-type downstream region, which may contain regulatory elements essential for its spatio-temporal expression (Figure S4B). *CHD3* is associated with a known DD (Snijders Blok-Campeau syndrome; MIM: 618205),[35] and this might explain part of the individual's symptoms. However, due to the complex nature of these CGRs, other alterations to gene regulation are likely to have an effect on the DD3 phenotype.

In summary, Hi-C of duplications at the SOX9 locus detected by array-CGH revealed a higher degree of rearrangement complexity than was previously thought. Hi-C maps help reconstruct the new genomic architecture and interpret the molecular mechanism involved in the pathogenesis of these phenotypes.

Hi-C Is Highly Effective in Translocation Detection and the Reconstruction of Complex Genomic Breakpoints

Inversions and translocations are copy-number-neutral SVs that can disrupt chromatin domains by rearranging enhancer elements with respect to their cognate target genes, potentially resulting in a regulatory loss and/or gain of function. Here, we sought to investigate whether Hi-C could identify complex SVs in clinical samples harboring chromosomal translocations previously detected through the use of karyotyping. Hi-C maps were generated from LCLs of six individuals with DD carrying reciprocal chromosomal translocations (Table 1). Hi-C readily shows the translocations in all cases (Figure 3). We also asked if the detected SVs (from DD1 to DD9) could be observed in the Hi-C maps when using lower sequencing depth. After reducing the number of initial paired-end reads for all maps (from 100 M down to 1 M sequenced fragments), we could visualize the tandem duplications with 50 M, the intra-chromosomal insertions with 10 M, and, surprisingly, the reciprocal translocations with only 1 M sequenced fragments (Figure S7). We also compared the breakpoints estimated from binned Hi-C data and the breakpoints provided by WGS, yielding on average a distance of 3.5 kb (143 bp to 8.6 kb) between both approaches (Table S2). When switching from binned Hi-C maps to restriction fragment based Hi-C maps, the estimation of the breakpoint may be further improved (Table S2). Interestingly, the WGS SV callers Delly, PopDel, and ERDS failed to detect the 3.1 Mb microduplication from DD2, the 2.4 Mb and the 49 kb duplications from DD3, one reciprocal translocation (DD5), and one breakpoint in the CGR of DD6 (Table 1).

We selected two cases in the series of translocations to illustrate the Hi-C results. An individual presenting with dysmorphic features and severe intellectual disability (DD6) harbored a reciprocal translocation 46,XX,t(3;5)(q24;q14.3) previously detected through the use of karyotyping. Hi-C from DD6 LCLs confirmed the translocation (Figure 4A and B) but revealed a much more complex structure of the derivative chromosomes 3 and 5. For simplification, we named the four fragments on chromosome 5 "A to D" based on their breakpoint location. The Hi-C trans-map indicated a direct contact of fragment B with chr3. The most likely explanation is that this fragment is inserted in reverse orientation after fragment C (Figure 4C). At the region around the breakpoint on chr5 (Figure 4D), we observed the CGRs shown in the trans-map (Figure 4B). The estimated breakpoint chr5:88,045 kb disrupts the *MEF2C* ORF. The fragment B (light gray), containing *MEF2C* and several known enhancers, was inserted downstream of fragment C, however, it was inserted in an inverted orientation (arrow on Figure 4D and 4E). Thus, the shuffled-TAD contains only a disrupted *MEF2C*. Indeed, RT-qPCR of DD6 LCLs revealed reduced expression (50%) of *MEF2C* compared to controls (Figure 4F). Haploinsufficiency of *MEF2C* causes severe ID, epilepsy, and/or cerebral malformations (MIM: 613443), and this makes it a likely candidate for the patient's phenotype.[36]

In a second case, the molecular mechanism of DD5 was solved with the help of Hi-C. In brief, the translocation in DD5 disrupts the FOXG1 TAD on 14q12, separating this gene from several of its known brain enhancers (Figure S8).[37,38] *FOXG1* haploinsufficiency is associated with the congenital variant of Rett syndrome (MIM: 613454), and several non-coding SVs that disrupt *FOXG1* downstream enhancers have been identified in individuals with Rett syndrome-like phenotypes.[37,38] Therefore, loss of interactions between *FOXG1* and its regulatory elements caused by the reciprocal translocation putatively causes a regulatory loss of function, thus explaining the phenotype of DD5.

The remaining three unsolved translocation cases, individuals with intellectual disability and development delay, are described in Figures S9 and S10. Although the exact molecular mechanisms were not completely solved in these cases, the translocation breakpoints are located inside TADs, potentially disrupting wild-type EPI (i.e., loss-of-function) or/and creating enhancer adoption by TAD reshuffling (i.e., gain of function). In two cases (DD7 and DD9), genes involved in neurological disorders and highly expressed in the brain (e.g., *SATB2* and *CTNNA2*) are likely to be involved in the patients' phenotypes. In summary, Hi-C effectively maps reciprocal translocations and helps identify likely pathogenic mechanisms associated with the phenotypes in two to four (out of five) unsolved translocation cases.

Discussion

High throughput technologies are revolutionizing the field of human genetics, and the introduction of WGS in the clinic as a one-test-for-all promises to replace all other molecular technologies in the near future.[1–5] Although WGS has been shown to outperform exome sequencing for SNV detection,[39] substantial challenges remain for the

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

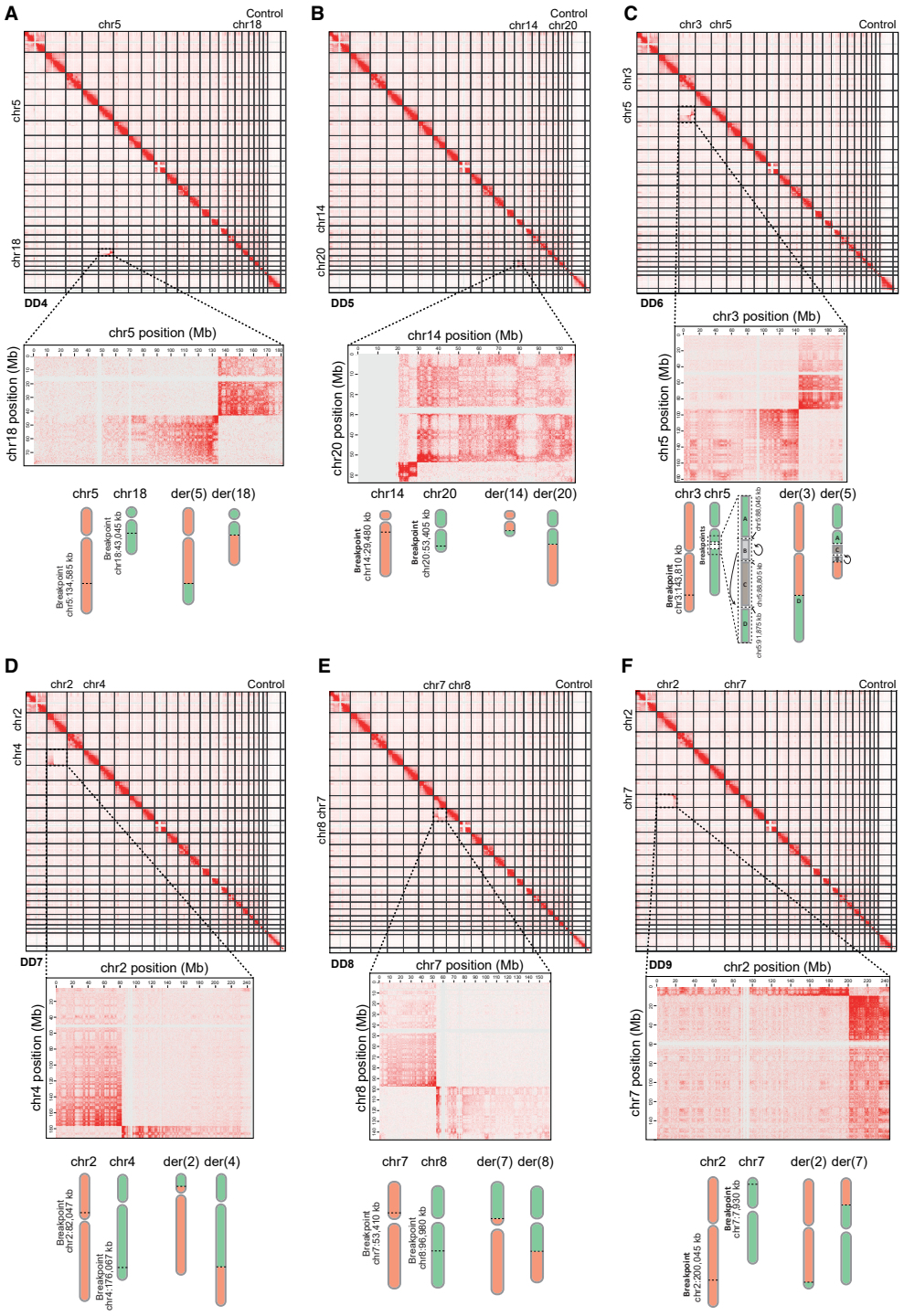

Figure 3. Reciprocal Translocations Are Readily Detected by Hi-C
(A–F) First panel: Hi-C maps of LCLs from individuals DD4–DD9 detect the reciprocal translocations in each case. Second panel: Reciprocal translocations are visible as bow-tie-like patterns in the Hi-C trans-maps (250 kb resolution). Third panel: Schematic representation of the derivative chromosomes based on the pattern of the signal intensity from the trans-map.

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

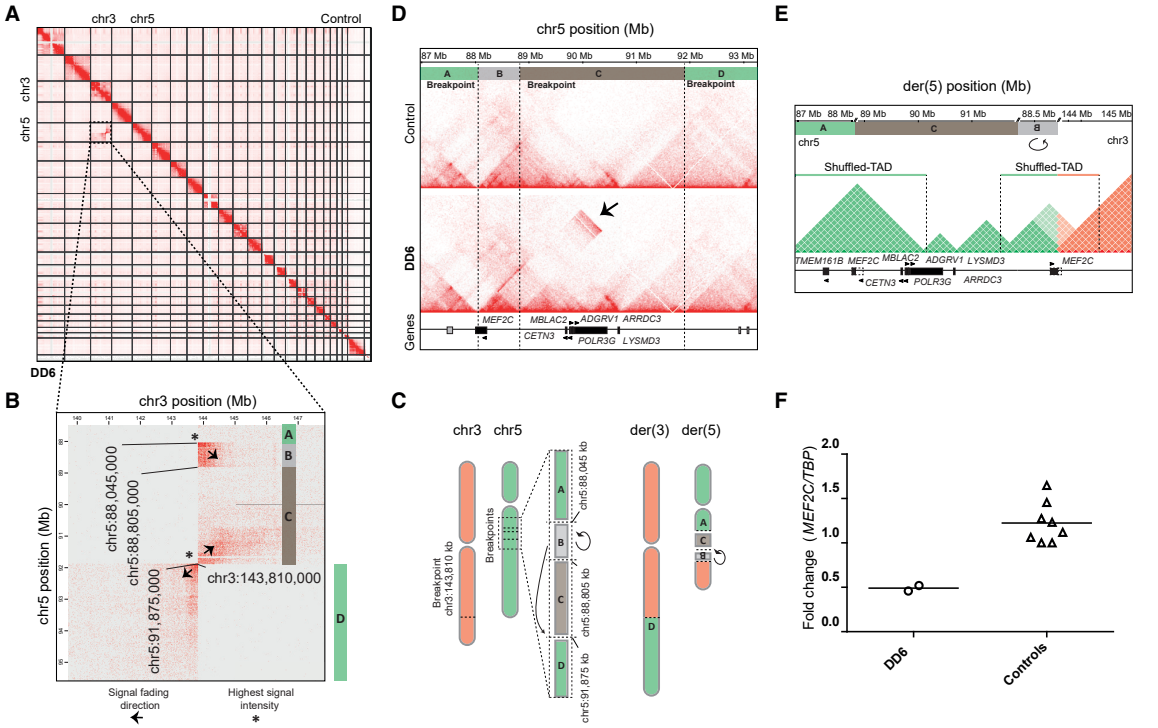

Figure 4. Hi-C Maps Showing a Reciprocal Translocation t(3;5) with a CGR Changing the 3D Landscape of the Derivative Chromosomes

(A) Hi-C map from LCLs of individual DD6 confirmed the t(3;5).

(B) Hi-C detected one breakpoint on chr3 and three on chr5 (chr5:88,045 kb; chr5:88,805 kb; and chr5:91,875 kb). Based on the gradient of the Hi-C signal observed in the trans-map, we could infer that fragment B is inserted downstream of fragment C, also inversely oriented and fused to chr3 (250 kb resolution; raw count map).

(C) Schematics of the derivative chromosomes based on the breakpoints and fragment orientations inferred from Hi-C bin signal from Figure 4B. The der(5) harbors a CGR composed of an insertion and inversion.

(D) The cis-map of 5q14.3 (25 kb resolution, KR-normalized). Three breakpoints on 5q14.3 can be observed and are represented by four fragments on chr5 (red—A, light gray—B, brown—C, and red—D). Note the inversion of the fragment B on the cis-map of DD6 (arrow).

(E) Schematic representation of der(5) showing two novel shuffled-TADs, one containing several genes within the same domain and the other devoid of an ORF.

(F) RT-qPCR showing half the expression of *MEF2C* in DD6 compared to controls.

detection and interpretation of complex SVs with multiple breakpoints. These SVs, including duplications, deletions, translocations, insertions, and inversions, have the potential to disrupt higher-order chromatin organization, thereby rewiring the complex 3D chromatin organization of a locus.[9] This may lead to the repositioning of TAD boundaries and/or the relocation of enhancer elements into other regulatory landscapes, causing gene misexpression and in some cases a deleterious phenotype.[14,15,40,41] Here, we set out to use Hi-C for the identification and interpretation of SVs in clinical samples. As proof of concept, we therefore performed Hi-C in nine samples from individuals with DD, samples with known large SVs detected by commonly used diagnostic cytogenetic and/or cytogenomic tools, and we interrogated whether Hi-C would be useful to detect, resolve, and interpret these variants. Indeed, Hi-C could show all previously known SVs plus a number of additional chromosomal rearrangements, revealing a high degree of complexity in some cases that had remained undetected with conventional methods. Due to these structural changes, the Hi-C maps enabled us to reconstruct the altered 3D genomic architecture. Based on this, we were able to identify regions in which the formation of aberrant regulatory interactions presumably contributed to the phenotypes of the screened individuals (five out of nine).

Our data show that Hi-C is a powerful tool for detecting SVs in a clinical setting. Compared to clinical array CGH, Hi-C has better resolution (3–5 kb in Hi-C compared to around 30–50 kb in array CGH), plus Hi-C can also identify copy number neutral variants, such as inversions and translocations, and add positional information to CNVs.[11,17] Here, for instance, we identified duplications in tandem (DD1 and DD2), duplications and an insertion (DD3), and a translocation with an inversion (DD6). Hi-C also emerges as a versatile tool that can detect large SVs through the use of lower sequencing resources, whereby we were able to spot reciprocal translocations in

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

reduced sequencing-depth maps down to 1 M fragments (Figure S7). It is noteworthy that the WGS SV callers used in this study failed to detect three out of nine main SVs in our cohort. Although Hi-C has been used before to identify large SVs in primary brain tumor samples[42] and B cells lymphomas[43] and in combination with optical mapping and WGS in other cancer cell lines,[17] the potential of Hi-C for the investigation of the genetic causes of DDs has not been evaluated yet. Our data indicate that Hi-C has a large diagnostic potential. Low-coverage Hi-C could be used in patient cells as an alternative to array CGH or as a secondary tool for excluding balanced SVs instead of classical karyotyping. Hi-C could also be extremely useful in the field of cancer cytogenetics[17,42,43] and in precision medicine.[44]

Our study also has several limitations. First, the limited resolution in our binned Hi-C maps of around 5 kb, the high noise level in Hi-C, and the effects of binning short reads from ligated restriction fragments to a regular grid allow for only a rough estimate of breakpoints and do not show specific enhancer promoter contacts in the newly formed TADs. Improvements in algorithms and tools for data analysis may solve these limitations,[8] e.g., analyzing restriction-fragment-based Hi-C maps could improve the estimation of the breakpoint (Table S2). Second, the computational analysis for Hi-C is still not as user-friendly as with current array CGH and WGS tools. Third, better algorithms should be developed and tested to automatize the detection, in a genome-wide perspective, of all SVs and ectopic Hi-C contacts in samples from clinical individuals.[8,17,45,46]

In summary, our study illustrates the importance of considering not only the linear sequence of the genome, but also the 3D chromatin structure when interpreting the impact of SVs. We show that Hi-C is a powerful tool useful for detecting and resolving the structure of complex chromosomal rearrangements in DDs. If developed further, Hi-C may become a method of choice for efficient SV detection in cases with suspected genetic causes. To gain an even deeper understanding of the complex nature of SVs, we propose combining Hi-C with WGS, long read sequencing, and/or optical mapping.[17] The combination of these powerful tools will allow for a more precise determination of cryptic rearrangements and enables more accurate predictions of phenotypic consequences.

Data and Code Availability

The accession numbers for the CNVs coordinates reported in this manuscript are in DECIPHER: 412419, 412420, 412421.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.04.016.

Acknowledgments

We would like to thank the individuals and families for their collaboration and contribution to this project. Technical assistance: We thank Susanne Rothe, Vanessa Suckow, and Celina São-José for their excellent work. We thank Angela Maria Vianna-Morgante for assistance in reviewing the cytogenetics nomenclature. Funding: M.S. and S.M. are supported by grants from the Deutsche Forschungsgemeinschaft (DFG) (SP1532/3-1, SP1532/4-1, SP1532/5-1, and MU 880/16-1) and the Max Planck Foundation. U.S.M. is a fellow of the Capes-Alexander von Humboldt Foundation.

Declaration of Interests

Rocio Acuna-Hidalgo is a founder, shareholder, and full-time employee of Nostos Genomics.

Conflict of Interest

The other authors declare no competing interests.

Received: October 8, 2019
Accepted: April 29, 2020
Published: May 28, 2020

Web Resources

ClinGen CNV Pathogenicity Calculator, http://cnvcalc.clinicalgenome.org/cnvcalc/cnv-gain
DECIPHER, https://decipher.sanger.ac.uk/
Mouse Genome Informatics, http://www.informatics.jax.org/
Mouse Phenotype, https://www.mousephenotype.org/
Online Mendelian Inheritance in Man (OMIM), https://www.omim.org/

References

1. Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.
2. Vissers, L.E.L.M., van Ravenswaaij, C.M.A., Admiraal, R., Hurst, J.A., de Vries, B.B.A., Janssen, I.M., van der Vliet, W.A., Huys, E.H.L.P.G., de Jong, P.J., Hamel, B.C.J., et al. (2004). Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nat. Genet. *36*, 955–957.
3. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.
4. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.
5. Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al.; Centers for Mendelian Genomics (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet. *97*, 199–215.

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

6. Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. *7*, 85–97.
7. Sanchis-Juan, A., Stephens, J., French, C.E., Gleadall, N., Mégy, K., Penkett, C., Shamardina, O., Stirrups, K., Delon, I., Dewhurst, E., et al. (2018). Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. Genome Med. *10*, 95.
8. Wang, S., Lee, S., Chu, C., Jain, D., Nelson, G., Walsh, J.M., et al. (2020). HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. Genome Biology *21* ((1):73), In press. https://doi.org/10.1186/s13059-020-01986-5.
9. Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467.
10. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.
11. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.
12. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.
13. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.
14. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.
15. Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.
16. Flöttmann, R., Kragesteen, B.K., Geuer, S., Socha, M., Allou, L., Sowińska-Seidler, A., Bosquillon de Jarcy, L., Wagner, J., Jamsheer, A., Oehl-Jaschkowitz, B., et al. (2018). Noncoding copy-number variations are associated with congenital limb malformation. Genet. Med. *20*, 599–607.
17. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardımcı, G.G., Chakraborty, A., Bann, D.V., Wang, Y., et al. (2018). Integrative detection and analysis of structural variation in cancer genomes. Nat. Genet. *50*, 1388–1398.
18. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv:1303.3997.
19. Faust, G.G., and Hall, I.M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503–2505.
20. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
21. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.
22. Roskosch, S., Jónsson, H., Björnsson, E., Beyter, D., Eggertsson, H.P., Sulem, P., Stefánsson, K., Halldórsson, B.V., and Kehr, B. (2019). PopDel identifies medium-size deletions jointly in tens of thousands of genomes. bioRxiv. https://doi.org/10.1101/740225.
23. Zhu, M., Need, A.C., Han, Y., Ge, D., Maia, J.M., Zhu, Q., Heinzen, E.L., Cirulli, E.T., Pelak, K., He, M., et al. (2012). Using ERDS to infer copy-number variants in high-coverage genomes. Am. J. Hum. Genet. *91*, 408–421.
24. Smedley, D., Jacobsen, J.O., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat. Protoc. *10*, 2,004–2,015.
25. Hombach, D., Schuelke, M., Knierim, E., Ehmke, N., Schwarz, J.M., Fischer-Zirnsak, B., and Seelow, D. (2019). MutationDistiller: user-driven identification of pathogenic DNA variants. Nucleic Acids Res. *47* (W1), W114–W120.
26. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016a). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. *3*, 95–98.
27. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
28. Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J. Numer. Anal. *33*, 1029–1047.
29. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016b). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. *3*, 99–101.
30. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
31. Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. Nat. Protoc. *3*, 1101–1108.
32. Kurth, I., Klopocki, E., Stricker, S., van Oosterwijk, J., Vanek, S., Altmann, J., Santos, H.G., van Harssel, J.J., de Ravel, T., Wilkie, A.O., et al. (2009). Duplications of noncoding elements 5′ of SOX9 are associated with brachydactyly-anonychia. Nat. Genet. *41*, 862–863.
33. Plaster, N.M., Tawil, R., Tristani-Firouzi, M., Canún, S., Bendahhou, S., Tsunoda, A., Donaldson, M.R., Iannaccone, S.T., Brunt, E., Barohn, R., et al. (2001). Mutations in Kir2.1 cause the developmental and episodic electrical phenotypes of Andersen's syndrome. Cell *105*, 511–519.
34. Xia, M., Jin, Q., Bendahhou, S., He, Y., Larroque, M.-M., Chen, Y., Zhou, Q., Yang, Y., Liu, Y., Liu, B., et al. (2005). A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. Biochem. Biophys. Res. Commun. *332*, 1012–1019.
35. Snijders Blok, L., Rousseau, J., Twist, J., Ehresmann, S., Takaku, M., Venselaar, H., Rodan, L.H., Nowak, C.B., Douglas, J., Swoboda, K.J., et al.; DDD study (2018). CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language. Nat. Commun. *9*, 4619.

Please cite this article in press as: Melo et al., Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases, The American Journal of Human Genetics (2020), https://doi.org/10.1016/j.ajhg.2020.04.016

36. Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. *49*, 36–45.
37. Allou, L., Lambert, L., Amsallem, D., Bieth, E., Edery, P., Destrée, A., Rivier, F., Amor, D., Thompson, E., Nicholl, J., et al. (2012). 14q12 and severe Rett-like phenotypes: new clinical insights and physical mapping of FOXG1-regulatory elements. Eur. J. Hum. Genet. *20*, 1216–1223.
38. Ibn-Salem, J., Köhler, S., Love, M.I., Chung, H.R., Huang, N., Hurles, M.E., Haendel, M., Washington, N.L., Smedley, D., Mungall, C.J., et al. (2014). Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. *15*, 423.
39. Lelieveld, S.H., Spielmann, M., Mundlos, S., Veltman, J.A., and Gilissen, C. (2015). Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. Hum. Mutat. *36*, 815–822.
40. Kragesteen, B.K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A.M., Jerković, I., et al. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet. *50*, 1463–1473.
41. Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schöpflin, R., Markowski, J., Ibrahim, D.M., Acuna-Hidalgo, R., Despang, A., Andrey, G., et al. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. Nat. Cell Biol. *21*, 305–310.
42. Harewood, L., Kishore, K., Eldridge, M.D., Wingett, S., Pearson, D., Schoenfelder, S., Collins, V.P., and Fraser, P. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biol. *18*, 125.
43. Díaz, N., Kruse, K., Erdmann, T., Staiger, A.M., Ott, G., Lenz, G., and Vaquerizas, J.M. (2018). Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. Nat. Commun. *9*, 4938.
44. Li, Y., Tao, T., Du, L., and Zhu, X. (2020). Three-dimensional genome: developmental technologies and applications in precision medicine. J. Hum. Genet. *65*, 497–511.
45. Chakraborty, A., and Ay, F. (2018). Identification of copy number variations and translocations in cancer cells from Hi-C data. Bioinformatics *34*, 338–345.
46. Stansfield, J.C., Cresswell, K.G., Vladimirov, V.I., and Dozmorov, M.G. (2018). HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. BMC Bioinformatics *19*, 279.

List of Publications

1. Klever, M. K., Sträng, E., Hetzel, S., Jungnitsch, J., *et al.* AML with complex karyotype: extreme genomic complexity revealed by combined long-read sequencing and Hi-C technology. *Blood Advances.* 2023 Nov 14;7(21):6520–31.

2. Melo, U., Jatzlau, J., Prada, C., Flex, E., Hartmann, S., Ali, S., Schoepflin, R., Bernardini, L., Ciolfi, A., Moeinzadeh, MH, **Klever, M. K.,** *et al*. Enhancer hijacking at the *ARHGAP36* locus is associated with connective tissue to bone transformation. *Nature Communications* 11;14(1):2034 (2023).

3. Schöpflin, R., Melo, U.S., Moeinzadeh, H., Moeinzadeh, H., Heller, D., Laupert, V., Hertzberg, J., Holtgrewe, M., Alavi, N., **Klever, M. K.,** *et al.* Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. *Nature Communications* 13, 6470 (2022).

4. Melo, U. S., Piard, J., Fischer-Zirnsak, B., **Klever, M. K.,** *et al.* Complete lung agenesis caused by complex genomic rearrangements with neo-TAD formation at the SHH locus. *Human Genetics* 140, 1459–1469 (2021).

5. Melo, U. S., Schöpflin, R., Acuna-Hidalgo, R., Mensah, M. A., Fischer-Zirnsak, B., Holtgrewe, M., **Klever, M. K.,** *et al.* Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *American journal of human genetics 106*(6), 872–884 (2020).

Curriculum Vitae

My curriculum vitae does not appear in the electronic version of my paper for reasons of data protection

Acknowledgements

I would like to thank everyone who was involved in the projects and the associated publications, on which this dissertation is based, for their support. Here I would like to mention Dr. Eric Sträng, Sara Hetzel, Dr. Anna Dolnik and Dr. Robert Schöpflin and my supervisors, Dr. Uirá Souto Melo, Prof. Dr. Stefan Mundlos and Prof. Dr. Lars Bullinger. I would also like to thank the Berlin Institute for Health Research (BIH) and the German Society for Internal Medicine (DGIM), which supported me with doctoral scholarships for the duration of my doctoral thesis.

I would also like to mention my friend and colleague Mr. Julius Jungnitsch, who was a doctoral student at the same time at the Max Planck Institute for Molecular Genetics, and with whom I share many positive memories of this time.

Furthermore, I would like to especially thank my fiancée, Ms. Nicola Junghaus. She was a huge support to me during the not always easy times of my doctoral thesis, for which I am deeply grateful to her.

My family was also a great support for me in these times. I would especially like to thank my mother Barbara Klever and my father André Klever, on whose tireless support I could rely on my whole life.