LUND UNIVERSITY

Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia

Moura-Castro, Larissa Helena

2024

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Moura-Castro, L. H. (2024). *Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia*. [Doctoral Thesis (compilation), Department of Laboratory Medicine]. Lund University, Faculty of Medicine.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

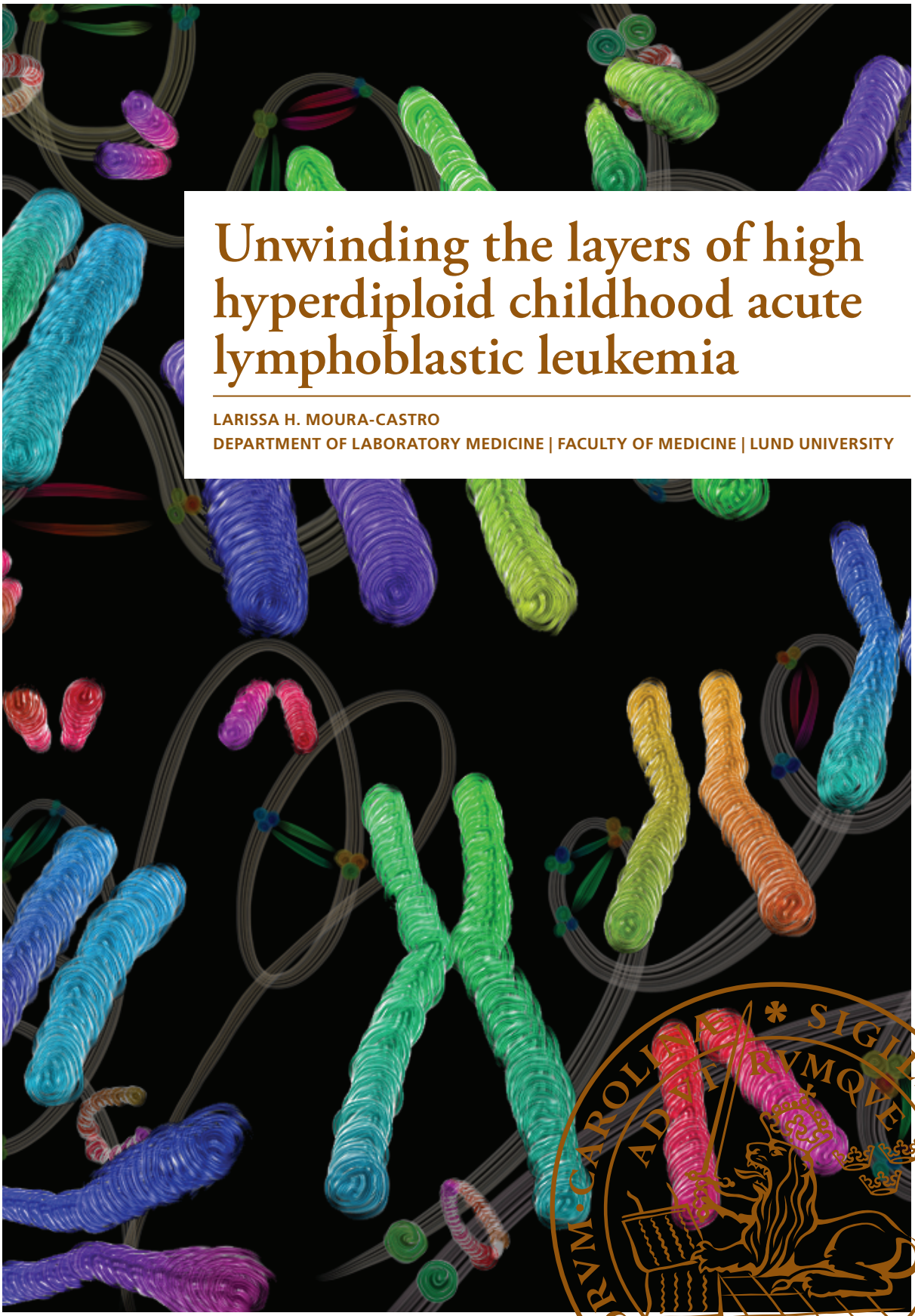
Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 02. May. 2024



Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia

LARISSA H. MOURA-CASTRO
DEPARTMENT OF LABORATORY MEDICINE | FACULTY OF MEDICINE | LUND UNIVERSITY

Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia

Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia

Larissa H. Moura-Castro

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Medicine at Lund University to be publicly defended on 18th of April 2024 at 09.00 at Belfragesalen, BMC D15, Lund

Faculty opponent
Professor Nick Cross
University of Southampton, Southampton, United Kingdom

Organization: LUND UNIVERSITY

Document name: Doctoral Dissertation **Date of issue** 18 April 2024

Author(s): Larissa Helena Moura Castro **Sponsoring organization:**

Title and subtitle: Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia

Abstract:

High hyperdiploid (HeH) acute lymphoblastic leukemia (ALL) is one of the most common pediatric cancer types. It is more frequent in children between 2-4 years old and usually has a favorable prognosis. The modal number in the leukemic cells from this subtype ranges from 51 to 67 chromosomes, marked by non-random chromosome gains. In this thesis, I investigated the leukemogenesis of HeH ALL, focusing on the origins and effects of aneuploidy (whole chromosome gains), as well as genome organization and transcription regulation.

In ***Article I*** we found that HeH ALL primary samples, compared to *ETV6::RUNX1* ALL, harbor genome-wide transcription dysregulation associated with the chromosome gains. We further found that HeH cases display low levels of CTCF and the cohesin complex, as well as aberrant interphase and metaphase chromatin architecture. In ***Article II***, we showed that the HeH ALL subtype displays recurrent sister chromatid cohesion defects. We further observed low gene expression of cohesin subunit *RAD21* and condensin subunit *NCAPG*, as well as chromosome copy number variation associated with cohesion defects. In ***Article III***, we investigated the 3D chromatin landscape of nine genetic subtypes of B-cell precursor ALL, identifying different clusters based on their chromatin signature, and revealing major chromosome disorganization in aneuploid samples affecting gene regulation, including the leukemia-related genes *FLT3* and *IKZF1*. In ***Article IV***, we studied clonal evolution in primary HeH ALL samples at the single-cell level, revealing that this subtype displays stable aneuploid cells with little chromosome number variation. Our results from *in silico* modeling and analysis of 577 HeH samples suggest that HeH cells originate from a punctuated event caused by a diploid cell undergoing tripolar division.

Ultimately, this thesis contributes to our understanding of the origins and development of HeH ALL. I characterize its chromosome architecture at the interphase and metaphase level and show how these features associate with topological gene regulation.

Key words: acute lymphoblastic leukemia, high hyperdiploidy, chromatin organization, transcriptional regulation, CTCF, cohesin, clonal heterogeneity

Classification system and/or index terms (if any) Supplementary bibliographical information

Language English **ISSN and key title**: 1652-8220

ISBN: 978-91-8021-536-7

Recipient's notes **Number of pages**:81

Price Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature  Date 2024-03-06

Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia

Larissa H. Moura-Castro

LUND
UNIVERSITY

Coverphoto by Larissa Helena Moura-Castro
Copyright pp 1-81 Larissa Helena Moura-Castro

Paper 1 © Springer Nature Limited
Paper 2 © John Wiley & Sons, Inc.
Paper 3 © by the Authors (Manuscript unpublished)
Paper 4 © Springer Nature Limited

Lund University, Faculty of Medicine
Department of Laboratory Medicine

ISBN 978-91-8021-536-7
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2024

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se
MADE IN SWEDEN

To my family.

Table of Contents

Original Papers

Article I

Yang M, Vesterlund M, Siavelis I, Moura-Castro LH, Castor A, Fioretos T, Jafari R, Lilljebjörn H, Odom DT, Olsson L, Ravi N, Woodward EL, Harewood L, Lehtiö J and Paulsson K. Proteogenomics and hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Commun*. 2019;**10(1)**:1519.

Article II

Moura-Castro LH, Peña-Martínez P, Castor A, Galeev R, Larsson J, Järås M, Yang M and Paulsson K. Sister chromatid cohesion defects are associated with chromosomal copy number heterogeneity in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2021;**60**:410–417.

Article III

Moura-Castro LH, Yang M, Dushime GT, Woodward EL, Aydin E, Castor A, Olsson-Arvidsson L, Fioretos T, Johansson B, Järås M, Hagström-Andersson AK and Paulsson K. The 3D genome of pediatric B-cell precursor acute lymphoblastic leukemia. *Manuscript*.

Article IV

Woodward EL, Yang M, Moura-Castro LH, van den Bos H, Gunnarsson R, Olsson-Arvidsson L, Spierings DCJ, Castor A, Duployez N, Zaliova M, Zuna J, Johansson B, Foijer F and Paulsson K. Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukaemia. *Nat Commun*. 2023;**14**:1658.

Abbreviations

3C	Chromatin conformation capture
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
APC/C	Anaphase promoting complex or cyclosome
ATAC-seq	Assay for transposase accessible chromatin sequencing
B/ATRI	Before/After trisomy
BCP	B-cell precursor
BSA	Bovine serum albumin
BTRI	Before trisomy
cDNA	Copy DNA
CDKs	Cyclin-dependent kinases
CIN	Chromosomal instability
CLL	Chronic lymphocytic leukemia
CLP	Common lymphoid progenitor
CML	Chronic myeloid leukemia
CMP	Common myeloid progenitor
CMS	Chromosome morphology score
CNA	Copy number alteration
C-NHEJ	Canonical non-homologous end-joining
CNS	Central nervous system
DC	Dendritic cell
der()	Derivative chromosome
DGE	Differential gene expression
DSB	Double-strand break
E-P	Enhancer – promoter
FACS	Fluorescence-activating cell sorting
FISH	Fluorescence *in situ* hybridization
GMP	Granulocyte/macrophage progenitor
GWAS	Genome-wide association study
HeH	High hyperdiploid
HSC	Hematopoietic stem cell
iAMP21	Intrachromosomal amplification of chromosome 21
ISH	*In situ* hybridization
LMPP	Lymphoid-primed multipotent progenitor
MEP	Megakaryocyte/erythrocyte progenitor

mDC	Myeloid dendritic cells
MNC	Modal number of chromosomes
mRNA	Messenger RNA
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MTX	Methotrexate
NEB	Nuclear envelope breakdown
NGS	Next-generation sequencing
NK	Natural killer
NuMA	Nuclear mitotic apparatus
PCA	Principal component analysis
PCR	Polymerase chain reaction
PCG	Primary constriction gap
pDC	Plasmacytoid dendritic cells
Pol II	RNA polymerase II
P-P	Promoter – promoter
qPCR	Quantitative polymerase chain reaction
RAD21-KD	*RAD21* knockdown
RNAi	RNA interference
RNA-seq	RNA sequencing
SAC	Spindle assembly checkpoint
scWGS	Single-cell whole genome sequencing
SMC	Structural maintenance of chromosome
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural variant
TAD	Topologically associating domain
t()	Translocation
UPID	Uniparental isodisomy
WCP	Whole-chromosome paint
WES	Whole exome sequencing
WGBS	Whole genome bisulfite sequencing
WGS	Whole genome sequencing
wUPID	Whole chromosome uniparental isodisomy

Overview

Unwinding the layers of high hyperdiploid childhood acute lymphoblastic leukemia seeks to explore the molecular pathogenesis of this subtype and increase our knowledge of its characteristics, development, and origin. In order to understand the biological meaning behind many genetic features discussed in this thesis, the introduction will cover the basis of several mechanisms and fundaments that are relevant to our research.

Firstly, the processes of cell division, chromosome formation, genome organization and cancer genetics are briefly presented. The second part of the introduction focuses on formation of blood cells, leukemogenesis and, finally, the clinical and genetic aspects of B-cell precursor childhood acute lymphoblastic leukemia (ALL), with emphasis on the high hyperdiploid (HeH) subtype. The specific aims of this thesis are then presented, and I discuss the methods that have been performed, followed by a description of the main results of each included article. The four articles are discussed together and divided into two main themes: ***What drives leukemogenesis in HeH ALL?*** and ***Stable or unstable? Origins and clonal evolution of HeH ALL***, where I put our findings in the context of the present literature, closing the thesis summary with concluding remarks.

Introduction – Part I

The cell cycle

Cell division and the first stages of cell cycle

The concept of cell division was first proposed in the 1850s by the embryologist Robert Remak while he studied frog embryos, describing the cell creation as a continuous division starting at the nucleus. Still, it took 3 more decades to understand mitosis, when Walther Flemming in 1882 described that the longitudinal half of each chromosome went to each daughter cell **(1)**. After over 150 years of research from the cell theory to cell cycle division, we now have a profound knowledge of the stages a cell undergoes to generate daughter cells **(2)**.

A cell in a state of quiescence is in the G0 phase. From the moment it starts the cell division cycle, it enters the interphase, which comprises G1, S and G2. Hereafter the cell enters the M phase, including firstly mitosis, which is subdivided into prophase, prometaphase, metaphase, anaphase and telophase, and secondly cytokinesis, when the cytoplasm of the cell divides to form two daughter cells. Mediation of cell cycle is ensured by regulatory cyclin subunits and cyclin-dependent kinases (CDKs), including CDK1 (G2 and M), CDK2 (G1 and S) and CDK4/6 (G1), and its progression is tightly controlled by checkpoints that secure engagement of essential steps before moving from one phase to the next one. The main checkpoints are the G1/S restriction checkpoint, the G2/M DNA damage checkpoint, and the spindle assembly checkpoint (SAC) **(2)**.

The G1 is a growth phase, where the cells synthesize a large number of proteins, and by reaching a certain size the cell decides whether to enter S phase by CDK activation, differentiate or undergo cell death. Next, the cell enters the DNA synthesis, or S phase, when DNA replication occurs and doubles the content of each chromosome into two sister chromatids. The G2 phase follows DNA synthesis, and is marked by several preparations for mitosis, most remarkably DNA double-strand breaks repair, which is essential to pass the G2/M checkpoint. Lastly, the nucleus of

the cell divides during mitosis, and the cytoplasm divides to form the daughter cells **(2)**.

Prophase, prometaphase and mitotic spindle pole formation

Mitotic progression starts in prophase when topoisomerase II and the condensin complex initiate chromatin condensation and begin to shape the very recognizable individual mitotic chromosomes. Another important event is the separation of centrosomes and subsequent migration to the opposite poles to form the bipolar mitotic spindle apparatus. In late prophase, the cell becomes committed to mitosis, i.e. the process is no longer reversible, as the nuclear envelope breakdown (NEB) occurs. At this point, centrosomes are already generating dynamic microtubule arrays, that after NEB can reach for the chromosomes **(2-4)**.

During prometaphase, chromosomes are fully condensed, and each sister chromatid must attach to opposing microtubules to form the spindle poles. Specifically, it is the disk-shape proteins located in the centromeric region of each chromosome, known as kinetochores, that interact with the microtubules **(2-4)**.

Metaphase, SAC and transition to anaphase

The mitotic chromosomes, now captured by microtubules, are moved towards the cell equator, and aligned in the so-called metaphase plate, as the cell prepares for the SAC. Here, if a single chromosome is not attached to microtubules, anaphase onset is delayed. This delay is ensured by the production of inhibitory complexes, catalyzed by unattached kinetochores, that blocks the destruction of securin and cyclin B regulatory subunit CDK1, preventing sister chromatid separation and mitotic exit. Meanwhile, aurora-B kinase plays the main role in correcting errors in spindle pole attachment **(3)**.

One of the key events in metaphase to anaphase transition is activation of the anaphase promoting complex or cyclosome (APC/C) by forming a protein complex with CDC20. Early anaphase then initiates as APC/C-CDC20 degrades securin in chromosome centromeres, enabling separase to cleave cohesin and allowing sister chromatid segregation towards opposing poles. Simultaneously, the same protein complex degrades cyclin B1, terminating CDK1 activity and enabling mitotic exit. PP1γ dephosphorylates histone H3 bound to chromatin, which will allow later chromosome decompaction. Among other biochemical and mechanical changes,

aurora B kinase is dephosphorylated and moves from the chromatin to the center of the spindle pole **(4)**.

Late anaphase, telophase and cytokinesis

As microtubules have been continuously shortening, in late anaphase the spindle poles finally move apart. The phosphorylation gradient caused by aurora B kinase localizing in the equator of the spindle pole is essential to keep the two forming chromatin masses from invading the center zone, reinforcing the division that is being created. The nuclear mitotic apparatus (NuMA), which had been removed from chromatin, is phosphorylated, and allowed to return to its position for further chromatin decondensation. A new nuclear envelope begins to be formed in each side, together with nuclear pore complexes, and start separating the nuclear DNA from the cytoplasm **(2, 4)**.

Entering telophase, spastin promotes microtubule disassembly from the chromatin while ESCRT closes the gaps in the nuclear envelope. Actin cytoskeleton forms an actomyosin-ring in the midbody of the dividing cell. Furrow ingression via cytokinesis then takes part by contraction of the actomyosin-ring, separating the cytoplasm of the two emerging daughter cells. Lastly, cytoskeleton structures are removed from the intercellular bridge, and further contraction of the cell cortex seals the plasma membrane, releasing the two daughter cells **(2, 4)**.

The cohesin complex

Cohesin is a ring-like, multiprotein complex, also known as a structural maintenance of chromosome (SMC) protein complex **(5, 6) (Figure 1)**. In somatic cells, the complex is composed of two SMC heterodimer proteins, SMC1/3, the kleisin subunit RAD21 and either a STAG1 or STAG2 subunit, while different versions of these proteins – except SMC3 – occur in meiosis-related cohesins **(5, 7)**. Cohesin entraps chromatin in a topological manner, without direct binding, making the integrity of the ring essential for its proper function **(7)**. Association of cohesin to DNA starts in the G1 phase of the cell cycle, and requires the loading factors NIPBL and MAU2, a process antagonized by the releasing complex, WAPL and PDS5A/B **(7, 8)**.

Cohesin first known function was to embrace sister chromatids and maintain cohesion from DNA replication to cell division and thus assist proper chromosome segregation **(5, 8)**. In G1, cohesin associates with single DNA strands, entrapping

both sister chromatids after the DNA replication fork passes. Cohesion is then maintained by acetylation of SMC3 and recruitment of Sororin to PDS5, which displaces WAPL and prevents the ring complex from unloading **(7, 8)**. Here, cohesin facilitates double-stranded DNA repair by homologous recombination and assists the replication fork to restart if it stalls. When cells enter mitosis, most cohesins are unloaded by activation of CDK1, Aurora kinase B and PLK1, which phosphorylate Sororin and STAG1/2 along the chromosomes and allow hyperactivity of the releasing complex. In the centromeric regions, however, Shugoshin 1 and protein phosphatase 2A (PP2A) protect Sororin from phosphorylation and cohesion is maintained, allowing chromosome alignment in the metaphase plate and ensuring correct orientation of sister kinetochores. At the beginning of anaphase, the APC/C cleaves Securin, releasing Separase to dissolve the remaining cohesins and promoting separation of the sister chromatids **(7-9)**.

Formation of mitotic chromosomes

At the onset of mitotic progression, interphase chromatin must be highly compacted and individualized into cylindrical chromosomes to ensure the integrity of the replicated genome into each daughter cell. Mitotic chromosomes are formed by consecutive DNA loops, anchored by an axis and organized in radial arrays **(10)**. It is believed that the condensin complexes are responsible for the topological organization of mitotic chromosomes via loop extrusion, that result in the shortening of the chromatin **(10, 11)**.

Condensins I and II, similarly to cohesin, are also ring-shaped SMC protein complexes, composed of SMC2, SMC4, the CAPH/H2 kleisin and the subunits CAPD2/D3 and CAPG/G2 **(6, 12) (Figure 1)**. Condensin II is responsible for the first layer of chromatin loops during prophase, followed by condensin I loop nesting for further condensation after NEB. By prometaphase, the mitotic chromosome loop formation has acquired a helical shape, being the base for the chromosome bodies. Another vital function of condensin is to provide rigidity for mitotic chromosomes to resist the pulling forces suffered when microtubules attach to the kinetochores **(10)**.

Although essential for mitotic chromosome organization, condensin is not the main player in global compaction of mitotic chromatin. Chromatin fiber compaction is mostly regulated by histones that suffers mitosis-specific post-translational modifications **(10)**. In particular, de-acetylation of H2B, H3 and H4 during mitosis

is very prominent, suggesting that electrostatic interactions of histone tails trigger chromatin compaction **(13)**.

Resolution of sister chromatids begin soon after DNA replication, and in G2 the replicated genome is already separated. Condensin, cohesin and topoisomerase II have been suggested to take part in chromosome resolution, although the full mechanism behind this process remains uncertain **(10)**. Recently, it was found that Ki-67, a protein located in the mitotic chromosome periphery, prevents mitotic chromosomes from crumpling into a single body after nuclear envelope disassembly. During late prophase, Ki-67 assembles a layer of proteins and RNAs in the surroundings of mitotic chromosomes and forms a brush-like steric and electrostatic barrier. The charge barrier disperses particles on the surface of mitotic chromosomes, promoting their individualization **(10, 14)**.

In early anaphase, the DNA cross-bridging protein BAF is dephosphorylated and binds to chromatin, forming a stiff cross-bridged chromatin layer around each set of anaphase chromosomes. The BAF-induced chromatin network enables the assembly of new nuclear envelopes during mitotic exit **(10, 15)**.

Chromatin folding and organization

Hierarchy of chromatin folding

Over 100 years ago Carl Rabl, and then Theodor Boveri, proposed that chromosomes occupy distinct territories during interphase **(11)**. Later in 1974, a chromatin structure of 8 histone molecules wrapping around 200 base pairs (bp) of DNA was described by Roger Kornberg and named as nucleosome – the smallest folding unit in chromatin organization **(11, 16)**. From the largest to the smallest, we now understand much of how chromatin is structured thanks to the constant emergence of technologies to study genome folding.

The process of chromatin folding is hierarchical and highly conserved in eukaryotes (**Figure 2**). Below the distinct chromosome territories, there are sub-chromosomal compartments separating transcriptionally active and inactive genomic regions, known as A and B compartments respectively **(17, 18)**. Furthermore, areas of the same chromatin state, A or B, show a preference for interacting with each other, not only within the same chromosome but genome-wide **(18)**. At the submegabase-scale, stretches of 1 to 5 genes are contained in topologically associating domains

(TADs), which favor internal interactions while insulating communication between different TADs **(19-21)**.

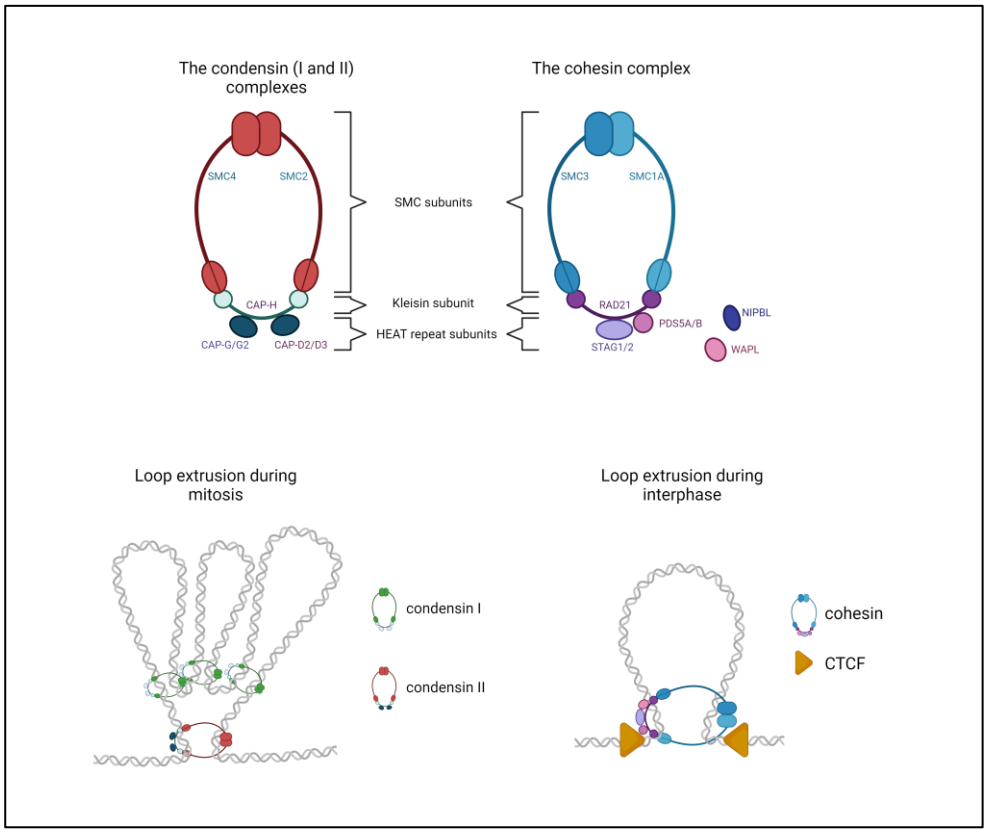


Figure 1 | **The cohesin and condensin complexes.** Both the cohesin and the condensin complexes have a ring-like shape and consist of two SMC subunits, a kleisin subunit and associating HEAT repeat subunits. During mitosis, condensin II binds to DNA and form loops to fold the chromatin, while condensin I further compacts the chromatin by pumping inner loops. Similarly, cohesin binds to CTCF during interphase and pumps DNA inwards to form loops that will become topologically associating domains (TADs). Created using Biorender.

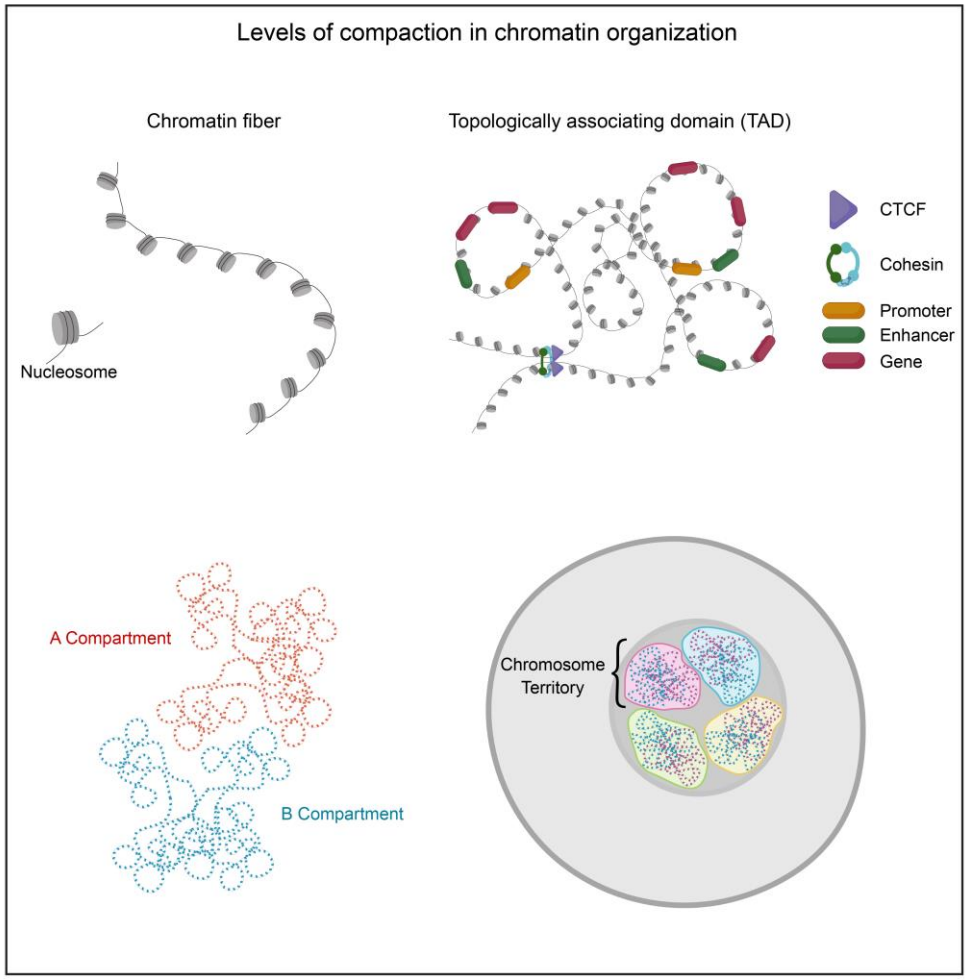


Figure 2 | **Hierarchy of chromatin folding.** The first level of chromatin compaction is the chromatin fiber, formed by nucleosomes that wind approximately 200 bp of DNA. Cohesin binds to CTCF in specific binding sites and forms TADs by loop extrusion, insulating 1-5 genes per domain. Chromatin with the same transcriptional activity level clusters together, forming A (more active) and B (less active) compartments. Finally, each chromosome during interphase occupies a specific area of the nucleus, known as chromosome territory. Created using Biorender.

Entering the fine-scale area of organization, we encounter chromatin loops. In 2015 it was shown that 99% of loops were anchored by the insulator protein CTCF, and 80% co-localized with cohesin complex binding sites **(11, 22)**. In this same level, there are also cis-regulatory elements controlling gene expression in long-range interactions, known as enhancer-promoter (E-P) and promoter-promoter (P-P) links, usually 5-200 kb long **(23-25)**. Transcription co-activators such as Mediator and YY1 and chromatin remodelers like BRG1 mediate E-P links together with the

transcription machinery of RNA Polymerase II (Pol II). Interestingly, a recent study showed that E-P and P-P links nest stretches of chromatin, and form loops with a different set of boundary markers, including transcription factors and histones **(25)**. E-P and P-P loops are formed without CTCF and cohesin binding sites, and acute depletion of such proteins have little effect on this type of loop formation **(26)**. This knowledge together with the fact that gene expression has a better correlation to E-P and P-P loops than to cohesin loops suggest that the cohesin complex might have a greater role in chromatin structure than in topological gene regulation **(25, 26)**.

CTCF, Cohesin and Condensin

CTCF is an 11-zinc-finger, sequence-specific DNA binding protein, and the only known insulator protein in vertebrates. This protein is, consequently, indispensable for genome folding, and not surprisingly enriched at transitions between different chromatin states and involved in interactions with the nuclear lamina. CTCF is sensitive to DNA methylation, implying that epigenetic changes interfere with its binding and affect gene expression. Lastly, CTCF shares a functional relationship with the cohesin complex, as mentioned before, co-occupying binding sites throughout the genome and defining anchor points of chromatin folding **(5)**.

Cohesin plays an important role in topological organization of the genome as CTCF and cohesin share a large amount of co-occupied chromatin sites, where STAG1/2 interacts with the C-terminus of CTCF **(5)**. Following this discovery, the loop-extrusion model (**Figure 1**) was suggested and heavily evidenced to explain TAD formation, where the cohesin complex embraces the chromatin, pumping it inwards until blocked by two convergent CTCF proteins bound to the DNA **(25, 27, 28)**. While cohesin is responsible for loop and TAD formation, recent data showed that condensin II and transcription factor III C (TFIIIC) co-localize at TAD boundaries that interact to form compartments, revealing a new role for this SMC complex **(6)**.

Cancer genetics

The hallmarks of cancer

Hanahan and Weinberg first proposed in 2000 that tumorigenesis is a multistep process that relies in the acquirement of six biological abilities necessary to overt cancer. These included self-sufficient proliferative signaling, evading growth suppressors, tissue invasion and metastasis in solid tumors, replicative immortality,

inducing or accessing vasculature and resisting apoptosis **(29)**. As our understanding of cancer continues to increase, the hallmarks of cancer have been through two revisions **(30, 31)**, adding new distinguished characteristics of cancer cells, such as genome instability and mutation, unlocking phenotypic plasticity and disrupting differentiation. The number of accumulated mutations necessary in the process of cancer formation can be extensive, and this progressive transformation is frequently long termed **(29-31)**.

Defects during mitosis and mitotic slippage

Mitosis is a fundamental process that requires meticulous regulation to ensure fidelity of the genetic material that will be passed on to the daughter cells. Any abnormalities related to microtubule attachment, sister chromatid cohesion or centrosome function activates the SAC, and leads to mitotic arrest until the errors are corrected. Failure in satisfying SAC causes cell death, either by mitotic catastrophe or in the next G1 phase. However, there are cases where cells manage to evade cell death after extended periods in mitotic arrest, known as mitotic slippage **(3, 4, 32)**.

During the S phase of the cell cycle the centrosome is duplicated and should then follow on to disjunction and separation in G2, and migration to form bipolar spindles in M phase. A major centrosome dysfunction is its amplification during the S phase, generating several instead of two units, and culminating in multipolar mitosis formation **(33)**. Conversely, at the onset of spindle assembly, sister kinetochores might become attached to microtubules from the same pole (syntelic attachment), or a single kinetochore might be captured by microtubules from both poles (merotelic attachment) **(3)**. Likewise, errors might occur during chromosome segregation, including lagging chromosomes and chromosome bridges and defects in sister chromatid cohesion **(4, 34)**. Such errors play key roles in triggering two of the most distinguished characteristics of cancer cells: chromosomal instability and aneuploidy **(4)**.

Chromosomal instability

Chromosomal instability (CIN) is a hallmark of cancer and refers to an increased rate of chromosome mis-segregation during mitosis **(35, 36)**. A frequent consequence of CIN is aneuploidy (see section below), although they are not synonyms and aneuploid cells can become stable shortly after being originated **(35-37)**. CIN can also promote chromosome breaks that result in chromosome structural rearrangements, as well as chromothripsis – a catastrophic event where a

chromosome shatters and gives rise to hundreds of rearrangements in one or several chromosome regions **(35)**. Moreover, chromosomal instability plays an important role in tumor evolution, formation of subclones and drug resistance **(35, 37)**.

Aneuploidy

Aneuploidy is classically defined as gains and losses of whole chromosomes, a description that has been recently expanded to numerical changes in whole chromosome arms as well **(36, 37)**. Although the vast majority of cancer types display some level of aneuploidy – ~90% in solid and ~60% in non-solid cancers **(37, 38)** – the consequences of these anomalies can be very distinct in different tumor types. Depending on the genomic context and chromosome gains, aneuploidy can both promote or suppress tumor development. It also has high prognostic value, since methods for detecting chromosomal gains and losses can be fast and straightforward, and recurrent aneuploidies are often used for risk stratification **(36, 37)**. An example of how diverse the impact of chromosome gains and losses can be is seen in the clinical outcome of tumors when classifying the degree of aneuploidy. High degree of aneuploidy is associated with poor prognosis in solid tumors, while hypodiploidy has the same effect in hematological malignancies. Hyperdiploidy is favorable in leukemias and lymphomas, while single aneuploidies have different outcomes depending on the involved chromosome and cancer type **(37)**. Aneuploidy has been well-established as an early event that plays a role in tumorigenesis, highlighting its importance and the need for further understanding how it arises and its effects **(37, 39, 40)**.

Chromosome rearrangements

Large rearrangements in chromosome structure are recurrent features in cancer cells. DNA damage caused by mitotic errors and exogenous or endogenous factors may lead to balanced and unbalanced chromosome translocations, formation of marker chromosomes and ring chromosomes, dicentric and isochromosomes, deletions, duplications and inversions of chromosome regions **(41, 42)**.

Mitotic errors such as lagging chromosomes trapped during furrow ingression and formation of micronulei from mis-segregated chromosomes pave the way to double-strand breaks (DSB) in the DNA, and often lead to chromosome translocations and chromothripsis, respectively **(42)**. DNA damage can also be introduced by exogenous sources, such as radiation or chemical exposure, and by endogenous events, among others oncogene activation, oxidative stress and DNA replication stress **(41, 42)**.

Although there are several mechanisms for DNA repair, these can be prone to errors and thus compromise genome integrity, especially in the presence of DSB. Methods such as canonical non-homologous end-joining (C-NHEJ) for DSB repair very often result in chromosome rearrangements, since it ligates any DSB regardless of specificity. Single-strand annealing utilizes long homologous sequences shared by both ends of DSB for repairing, and leads to deletions of repeated regions. Homologous recombination by break-induced replication promotes repair between two DSB that display one homologous side with the template, a mechanism that is considerably prone to error and can lead to loss of heterozygosity, duplications and unbalanced translocations **(42)**.

Single nucleotide variants

Every DNA molecule has a sequence composed of nucleotide bases, being two purine bases – adenine (A) and guanine (G), and two pyrimidine bases – cytosine (C) and thymine (T). Mutations where one nucleotide substitutes another are known as single nucleotide variants (SNVs), and their majority are harmless, representing natural variability of the human genome. Nevertheless, the position of an SNV and type of base substitution can have dire consequences and affect amino acid coding **(43)**.

Synonymous mutations are substitutions that do not affect the encoded amino acid due to codon bias. Nonsynonymous mutations, which have greater detrimental effects, are further divided into missense mutations when it leads to change in amino acids, or nonsense mutations when it inserts a premature stop codon, resulting in truncated proteins **(43, 44)**. Similar to aneuploidies and chromosomal rearrangements, point mutations have the potential to affect oncogenes and tumor suppressors – genes that are particularly important in cancer development. Well-known examples of cancer-related genes often targeted by point mutations are *TP53* and *RAS* **(42, 45, 46)**.

Introduction – Part II

Hematopoiesis

The history of hematopoiesis

In 1949, a study involving total body X-radiation of mice observed, for the first time, blood cells capable of shielding the subjects from lethal doses of irradiation **(47)**. This was just the beginning that instigated investigations of these remarkable blood forming cells, highly proliferative and capable of regeneration. During the 70's, a series of studies noted that the same blood cell type had the ability to generate myeloid, erythroid, megakaryocytic, and lymphoid cells **(48-50)**. It was then proposed that the bone marrow must contain a self-renewing, multilineage progenitor cell type, which was then named hematopoietic stem cell (HSC), also giving rise to the concept of the stem cell itself **(49, 51)**. Today it is known that HSCs are responsible for the production of over 10 different cell types in mammalians, although there is much yet not fully understood in the complex formation of blood cells. Research in genetics has, nonetheless, helped clarify how HSCs take different paths in this differentiation process known as hematopoiesis **(52)**.

HSCs differentiation

In adults, maintenance of HSCs is carefully regulated by transcription factor genes responsible for production, survival, and self-renewal of these cells, including the well-known *RUNX1*, *KMT2A* (former *MLL*) and *GATA2* **(53)**. While a small population of HSCs keep their abilities of self-renewal, others lose this capacity as they go under differentiation, the first step being the formation of a multipotent progenitor (MPP). Downstream progenitor cells slowly become more restricted, as MPPs form the oligopotent progenitors – capable of generating several but not all blood lineages. Traditionally, the common lymphoid progenitor (CLP) will generate B-cells, T-cells, and natural killers (NKs), where *IKZF1* (Ikaros) is essential for the

development of all lymphoid cells **(52, 53)**. The common myeloid progenitor (CMP) differentiates into two paths: the megakaryocyte/erythrocyte progenitor (MEP), where *GATA1* is a key gene to determine megakaryocytes (platelets); and the granulocyte/macrophage progenitor (GMP), having *SFPI1* (PU.1) as a master regulator for both granulocyte and monocyte formation **(52, 53)**. Recent studies have, however, resulted in changes in the hematopoietic lineage commitment tree, as dendritic cells (DCs) were found to be formed both by the CLP path, as plamacytoid dendritic cells (pDC), and the CMP path, as myeloid dendritic cells (mDC) **(54)**. Likewise, another oligopotent progenitor known as lymphoid-primed multipotent progenitor (LMPP) was discovered, having a combined potential for B-cells, T-cells and granulocytes/macrophages, but no megakaryocyte/erythrocyte potential **(55)**.

Lymphopoiesis and transcription factor requirements

Through the differentiation process known as lymphopoiesis, a series of steps defining cell fate leads HSCs to give rise to mature T-cells, B-cells, and natural killers (NKs) **(56)**. Among common lymphoid transcription factors, *IKZF1* is essential for lymphoid cell development, recruiting chromatin remodeling complexes to DNA regulatory elements, and regulating both self-renewal in HSCs and lineage-specific gene expression **(57, 58)**. Other factors that are common between B and T lymphopoiesis include GFI1, which inhibits PU.1 expression in MPP and thus GMP commitment, C-MYB, that regulates IL-7 receptors, and TCF3, necessary for lymphoid-lineage priming in MPPs **(57)**.

While B-cells mature in the bone marrow, when multipotent progenitor cells enter the T-cell commitment pathway they migrate to the thymus according to thymic microenvironmental stimulus **(57, 59)**. In contact with ligands in the thymic epithelial cells, the Notch pathway is activated, and signaling will continuously influence T-cell development until they begin to express TCRβ or TCRγδ – which will define αβ T-cells and γδ T-cells. Other transcription factors that are specific to T-cells are BCl11B, which regulates αβ T-lineage, GATA-3, taking part in initial specification and TCRαβ-dependent positive selection, and TCF-1, a signal-dependent transducer involved in the Wnt pathway **(59)**.

Early B-lineage fate restriction is highly determined by IKZF1, as its extended list of functions also includes promoting B-cell identity and modulating B-cell specific genes **(58)**. The transcription factor TCF3 is required in later stages of development, beyond pre- and pro-B cells, and also regulates expression of the B-lineage specific

transcription factors EBF1 and PAX5. These two transcription factors are fundamental for maintaining B-cell identity as they have collaborative but distinct roles, where EBF1 is required for B-lineage genes expression, and PAX5 has a dose dependent effect on inhibiting NK and T-lineage genes to allow B-cell commitment **(57, 58)**.

Leukemogenesis

Similar to what has been described for tumorigenesis in general, leukemogenesis takes a series of genetic alterations, accumulated over time, to transform normal hematopoietic cells into leukemic cells, and both disrupted differentiation and non-exhaustive proliferation are likely the early basis of leukemia **(60)**. The term pre-leukemic cells emerged as a definition of the intermediate stage where early genetic events have begun to transform normal hematopoietic cells, however secondary events are yet to come to overt leukemia **(60, 61)**.

From hematopoietic stem cells to pre-leukemic cells

The pathogenesis of leukemias have been studied for decades, and as early as the late 60's, aneuploidies – gain and loss of whole chromosomes – and partial chromosomal aberrations were identified as causatives of these malignancies **(61, 62)**. In acute myeloid leukemia (AML), loss of the long arm of chromosome 5 (5q), including the overlapping region 5q31 was described from several patient samples as an early event, and present in pre-leukemic cells **(63)**. The study from 1989 also investigated genes located in the region, only to find that several growth factors involved in regulating hematopoiesis were mapped to 5q31 and adjacent areas **(61, 63)**. Among others, these included *CSF2* and *IL3*, which sustain viability of early HSCs, *CD14* – a surface marker of monocytes and macrophages, and the tumor-suppressor gene *EGR1* **(61)**.

A more recent study has proposed a model for pre-leukemic cells, discussing evidence that the early events target HSCs. As leukemogenesis is believed to be a long-term process with a prolonged pre-leukemic stage, an early leukemogenic alteration should occur in a self-renewing cell with developmental plasticity or grant self-renewal to a more differentiated cell. In this context, the pre-leukemic HSC would be able to, while still retaining multi-lineage potency, proliferate without exhaustion and give time for additional mutations to accumulate. Ultimately, a secondary driver change emerges, originating leukemic cells that are reprogrammed

to be lineage-restricted as they lose normal functions **(60)**. Evidence of functional pre-leukemic cells harboring early genetic events was described in patients with chronic myeloid leukemia (CML), where the canonical *BRC::ABL1* fusion gene was detected not only in differentiated lymphoid cells but also in HSCs, a fact also corroborated by observations in chronic lymphocytic leukemia (CLL) and acute lymphoblastic leukemia (ALL) **(60, 64-72)**.

Pre-leukemic HSCs require secondary events to become full leukemic cells **(60, 61, 73)**. Early or first events usually involve genes that affect transcriptional and epigenetic regulators capable of modifying lineage options, including *ABL1, BCR, CEBPA, ETV6, PAX5, RARA* and *RUNX1*, among others. The secondary group of mutations frequently offer proliferative or survival advantages by activating signal-transduction pathways, such as alterations in *FLT3, RAS* or *KIT* **(73)**. Interestingly, studies suggest that pre-leukemic HSCs are therapy-resistant, possibly taking part in relapse cases. In agreement with this statement, early leukemogenic events were reported to have high concordance between diagnosis and relapse, while second-hit mutations are frequently gained or lost in relapse **(60)**. Infections and immune system modulation have also been suggested to trigger pre-leukemic cells into leukemia, a concept known as delayed infection hypothesis. Briefly, lack of natural infection exposures in infants may hinder modulation or priming of their immune system and may possibly lead to highly dysregulated immune responses when they are exposed to common infections in their second or third year of life **(74)**.

The origins of childhood leukemia

Accounting for 30% of all childhood cancers, leukemia is the most common neoplastic disease in children **(75)**. The origin of infant and childhood leukemia has relentlessly puzzled researchers. In fact, many wondered whether pre-leukemic cells emerged before birth. Several studies have shown indisputable evidence that most cases in fact arise in-utero **(74, 76, 77)**. Studies involving monozygotic twins with leukemia were the first indication of in-utero origins, where the siblings were found to share the exact same breakpoints in canonical leukemic fusion genes, suggesting that the leukemia cell of origin arose in one twin fetus and spread to the other twin **(74, 78)**. Further evidence was gathered from analysis of archived neonatal blood spots, collected routinely, showing that most cases of pediatric ALL and AML could be detected by fusion genes and other markers in pre-leukemic cells right after birth **(74, 77, 79, 80)**. Lastly, screening of 567 normal blood cord samples in 2002 observed the *ETV6::RUNX1* fusion in 1% of cases, approximately 100 times the incidence of childhood *ETV6::RUNX1*-positive ALL, indicating that functional pre-

leukemic cells arise more often prenatally, but secondary events to overt leukemia are rarer **(74, 81)**.

Clinical aspects of childhood acute lymphoblastic leukemia

Acute lymphoblastic leukemia is a heterogeneous entity that includes several subtypes, being the most common pediatric malignancy. Apart from the *KMT2A*-rearranged subtype, which occurs most commonly in infants (< 12 months old), the age peak of ALL is between 2 and 5 years old **(82)**. The overall survival rate in high-income countries is over 90%, while low- and middle-income countries show a lower range of success (22-79%) **(83)**. Besides stratification based on genetic subtypes, higher white blood cell count (WBC $\geq 50 \times 10^9$/L) at diagnosis, age below 12 months or above 10 years old, T-cell immunophenotype and central nervous system (CNS) involvement are considered poor prognostic factors and used for risk assignment **(82-84)**. Treatment of ALL consists of three stages with a duration of 2 to 2.5 years in total. Briefly, the first stage is remission-induction therapy to eradicate the initial leukemic cell burden with a duration of 4-6 weeks, where glucocorticoid (usually prednisone), vincristine and asparaginase are administered. This is followed by intensification (consolidation) therapy that aims to eradicate leukemic residual cells, including administration of cyclophosphamide, cytarabine, and mercaptopurine combined with high dosage of methotrexate (MTX). Lastly, maintenance therapy is carried out for ≥ 1 year, consisting of daily mercaptopurine and weekly methotrexate with or without vincristine and steroid pulses **(82, 84)**. The current protocol for treating ALL in children and adolescents in Sweden is the ALLTogether protocol. In 2019 several European study groups, including the Scandinavian program NOPHO, have started this collaborative protocol to enhance ALL treatment in children and adolescents, with a focus on decreasing over-treatment and improving the survival of high-risk cases **(85)**.

B-cell precursor childhood acute lymphoblastic leukemia

B-cell precursor (BCP) childhood ALL is characterized by proliferation of B-lymphoid progenitor cells. It is the most common type of childhood cancer, with a five-year overall survival rate of 90% **(86-88)**. Approximately 75% of BCP-ALL cases display aneuploidy or recurrent chromosomal rearrangements that occur as early or first events in leukemogenesis and are used as the basis for classification of

genetic subtypes **(89)**. Genetic aberrations in BCP-ALL commonly disrupt genes related to lymphoid development, such as *ETV6* and *RUNX1*, or activates oncogenes and tyrosine kinases, e.g. *ABL1* **(82)**. Additional alterations in B-cell transcription factor genes – e.g. *PAX5*, *IKZF1* and *EBF1*, causing developmental arrest in pro- and pre-B-cells, also play an important role in the pathogenesis of the disease, together with mutations in cell cycle and tumor suppressors – *CDKN2A/B* and *TP53*, lymphoid signaling – *BTLA*, *TOX*, *CD200*, and regulators of hematopoiesis, such as *FLT3* **(40, 82, 90)**. Other frequent targets are genes involved in the RTK-RAS signaling pathway, especially *NRAS*, *KRAS* and *PTPN11*, and histone modifiers such as *CREBBP* **(40, 82, 90, 91)**. Recent efforts to investigate cases that are not classified within any of the canonical genetic alterations revealed several emerging molecular subtypes in BCP-ALL, such as *PAX5*-driven alterations, *MEF2D* rearrangements *ZNF384* rearrangements, *ETV6::RUNX1*-like and *BCR::ABL1*-like **(88)**. Since these emerging subtypes are not completely settled, the focus of this section will be the well-established genetic subtypes.

BCR::ABL1-positive

The *BCR::ABL1* fusion gene arises by the rearrangement t(9;22)(q34;q11), where the der(22) is called Philadelphia chromosome. Comprising 3-5% of pediatric ALL cases, *BCR::ABL1* is associated with older age (median of 7.9 years), higher leukocyte count and poor prognosis **(89, 92)**. This fusion gene increases cell proliferation and dysregulate differentiation, while the most prominent additional alterations are deletions in *IKZF1*, which occur in more than 80% of *BCR::ABL1* ALL cases and has been associated with treatment resistance in this subgroup. *IKZF1* deletions are not present in *BCR::ABL1* CML, suggesting that such alterations are important in the development of *BCR::ABL1* ALL **(87, 89)**.

DUX4-rearranged

The recently described *DUX4*-rearranged BCP ALL is a subtype that accounts for 4-7% of pediatric cases, and is associated with a good prognosis **(84, 88)**. The most common fusion gene is *IGH::DUX4*, while *ERG::DUX4* is rarer. *DUX4* is overexpressed in this subtype, although not expressed in normal B-cells due to its specific role in embryonic development regulation **(88, 93)**. Whereas *IKZF1* deletions are often (40-50% of the events) associated with poor prognosis in other BCP ALL subtypes, deletions targeting this gene are favorable in *DUX4*-rearranged cases. Moreover, deletions in *ERG* also associates with better outcome **(84, 88)**.

Both *IGH::DUX4* and *ERG::DUX4* result in a truncated DUX4 protein which varies in length from patient to patient, but retains its DNA-binding properties **(93)**.

ETV6::RUNX1-positive

The most common fusion gene in pediatric BCP-ALL is *ETV6::RUNX1*, caused by the rearrangement t(12;21)(p13;q22). Accounting for 25% of childhood ALL cases, this subtype has superior molecular response to treatment, thus being associated with favorable prognosis. Nevertheless, late relapses – several years after treatment cessation – occur in up to 20% of patients, while very early relapses are rare **(94)**. The *ETV6::RUNX1* fusion gene is a primary event in leukemogenesis, and may convert *RUNX1* to a transcriptional repressor, as well as activate JAK-STAT signaling downstream. Recurrent second events include deletions in *PAX5*, *EBF1*, and the second copy of *ETV6*, and additional mutations are also seen in *BTLA*, *TOX* and *BTG1*, among others **(89)**.

Intrachromosomal amplification of chromosome 21 (iAMP21)

Intrachromosomal amplification of chromosome 21 (iAMP21) has an incidence of up to 2% in BCP ALL cases, with a median age of 9 to 10 years at diagnosis, low white blood cell count, a 5-year event-free survival rate of ~30% and overall survival rate of ~70% **(89, 95, 96)**. The subtype is characterized by a highly complex region on chromosome 21, mainly located between 32.8 and 37.9 Mb, with various regions of amplification, inversions and deletions in a phenomenon known as breakage-fusion-bridge, usually followed by chromothripsis, and which are heterogeneous among patients. The hallmark that defines a case as iAMP21 is to have three or more extra copies of *RUNX1* **(89, 95)**.

KMT2A-rearranged

Rearrangements involving *KMT2A* correspond to 3-4% of childhood BCP ALL, occurring in 70-75% of infants, and there is strong evidence that this subtype has prenatal origin **(84, 88, 97)**. It is considered a high-risk subtype with poor outcome, where the long-term event-free survival rate is less than 60%. Approximately 130 genes transcripts from *KMT2A* rearranged with over 90 different partner genes have been described, where the most common translocations are t(4;11)(q21;q23) – *KMT2A::AFF1*, t(11;19)(q23;p13.3) – *KMT2A::MLLT1*, and t(9;11)(p21;q23) –

KMT2A::MLLT3 **(88, 97)**. While *KMT2A* maintains homeotic gene expression in hematopoiesis, additional genetic alterations are uncommon in this subtype **(89)**.

Low hypodiploid and near-haploid

Low hypodiploid BCP ALL includes aneuploid cases with 31-39 chromosomes, with a frequency of 1% in pediatric cases **(84)** and poor prognosis. It harbors deletions in *IKZF2*, and high frequency of *TP53* mutations, frequently constitutional, which are otherwise rare in BCP ALL **(84, 89)**. Additionally, low hypodiploid cells often double and form near-triploid clones **(98)**.

Near-haploidy (24-30 chromosomes) is also a rare aneuploid subtype of BCP ALL, with a frequency of 2% in children and poor prognosis **(84, 99)**. Mutations in the RAS signaling pathway are very frequent, as well as *IKZF3* deletions **(84, 89)**. This subtype can often be mis-diagnosed due to the presence of duplicated hyperdiploid clones, which are characterized by disomies and tetrasomies, and frequent gains of chromosomes X, 14, 18 and 21 **(62, 99)**.

TCF3::PBX1-positive

TCF3::PBX1 ALL – caused by t(1;19)(q23;p13) – comprises 6% of childhood BCP ALL cases, with a median age of 7, and displays more frequent CNS relapse. Historically, the prognosis for the subtype was poor, which has changed with more recent treatment regimens **(84, 89)**. This translocation is often unbalanced and, as a consequence of this rearrangement, the subtype often displays gains of 1q. *TCF3* plays an important role in lymphoid development, and the *TCF3::PBX1* fusion gene disrupts HOX-regulated gene expression, which affects hematopoietic differentiation **(89)**.

High hyperdiploid (HeH)

High hyperdiploid (HeH) is the largest subtype of childhood BCP ALL (~30%), and accounts for hyperdiploid cases with 51-67 chromosomes that display a non-random gain of chromosomes X, 4, 6, 10, 14, 17, 18 and 21. With a median age of ~4 years, this subtype is characterized by low white blood cell count, usually below 10^9/l, rare instances of extramedullary leukemia (less than 5%), and mostly lymphoblasts with L1 morphology **(62, 100)**.

Trisomy or tetrasomy 21 occur in all HeH ALL cases, followed by gains of chromosome X with an incidence of 90-95%, and, in order of likelihood, gains of chromosomes 14, 6, 18, 4, 17, and 10. Notably, class-defining genes of this subtype are located predominantly in chromosomes 21 and X **(62, 100-102)**. Cases from this subtype display low frequency of subclonality, usually involving gains or losses of 1 to 3 whole chromosomes, or additional structural rearrangements **(62)**. Evidence suggests that the aneuploidy itself is the main driver change in HeH ALL, causing dosage effects – increased expression of genes located in the gained chromosomes, but that additional mutations are required to overt leukemia. Studies have showed that high hyperdiploidy has pre-natal origins in most, if not all, cases **(78, 103)**. Furthermore, analysis of mutational patterns among homologues in trisomies showed that the majority of mutations arise after chromosome gains, reinforcing that aneuploidy is the first hit in HeH ALL leukemogenesis **(40, 62)**.

Other key characteristics of classical HeH includes 2:2 allelic ratios of tetrasomies in virtually all cases and common (~30%) loss of heterozygosity through uniparental isodisomy in whole chromosomes (wUPID) – both chromosomes of a disomy come from the same parent, where wUPID 9 occurs more often **(40, 62, 102)**. Structural rearrangements are observed in ~50% of HeH ALL cases, where the most frequent unbalanced events – rearrangements that result in gain or loss of genetic material – are gains of 1q, deletions in 6q, and isochromosomes of 7q and 17q **(62)**.

The most targeted metabolic pathway in HeH ALL is the RAS pathway **(40, 62, 100)**. The RAS-RAF-MEK-ERK pathway is responsible for extracellular signaling to intracellular substrates, playing a role in regulation of cell proliferation and differentiation **(104)**. The commonly targeted genes of the Ras pathway in HeH cells are *KRAS*, *NRAS*, *FLT3* and *PTPN11*, being present in 50% of the cases. Histone modifiers and chromatin-remodeling genes are also frequently targeted in this subtype (20% of the cases), including *CREBBP*, *WHSC1*, *SUV420H1*, *SETD2* and *EZH2* **(40, 100)**. In regard to germline mutations and susceptibility, the most relevant affected gene in HeH ALL is *ARID5B*, an epigenetic activator of gene expression and regulator of cell cycle that is essential for normal lymphocyte development **(100, 102)**.

Since HeH ALL is associated with a favorable prognosis, early identification of higher-risk cases is important to prevent treatment failures and relapses, while providing a chance for deintensification of the remaining cases with better prognosis. Higher chromosome number, more than 53/55 chromosomes, is related to superior prognosis. Likewise, the presence of the so-called triple trisomies (+4,

+10 and +17), as well as concurrent +17 and +18 or +17/+18 in the absence of +5 and +20, leads to better outcome **(62, 100, 105)**. Conversely, the intronic risk allele rs7090445-C in *ARID5B* is associated with drug resistance and increased relapse, and mutations in *KRAS* and *CREBBP* often co-exist in relapsed clones **(91, 100)**. Recently, a study has showed that HeH cases with 56-67 chromosomes and with triple trisomies are highly sensitive to asparaginase, cytarabine, mercaptopurine and thioguanine, while cases with +7 and +9 were resistant to asparaginase, reinforcing the importance of investigations for case-specific treatments **(106, 107)**.

The present investigation

Aims

The overall aim of this thesis was to investigate the molecular pathogenesis of HeH childhood ALL, focusing on the global effects of hyperdiploidy, its nature and origin, and ultimately to shed light to its role in leukemogenesis. More specifically, the individual aims of the included articles were:

Article I

To determine the effects of aneuploidy on gene expression, protein levels and chromatin architecture in HeH ALL.

Article II

To investigate the incidence of sister chromatid cohesion defects and chromosome copy number variation in HeH ALL.

Article III

To explore the 3D genomic landscape of different subtypes of BCP ALL, its relationship with genomic features and its effects on the transcriptome.

Article IV

To investigate the clonal heterogeneity of HeH ALL at the single cell level, and to elucidate the origin of high hyperdiploidy.

Methods

Cytogenetics and cell biology techniques

Classic cytogenetics

In 1956, Tjio and Levan published their work of identifying the diploid number of human chromosomes at Lund University **(108).** The following years expanded our knowledge of the 46 chromosomes, and soon the term cytogenetics was invented to define the study of chromosomes. Classic cytogenetics has long been used to investigate changes in chromosomes, as genetic disorders and tumors began to be correlated with chromosome gains and losses as well as structural chromosome changes. Before the discovery of banding techniques, human chromosomes were divided into groups from A to G based on their size and centromere position, however it was difficult to infer more about the individual chromosomes within groups **(109)**. Then in 1968, Caspersson developed the first chromosome banding technique using quinacrine-based fluorescent dye, which was named Q-banding. The Q-banding revealed that each chromosome has a unique banding pattern, which led to the identification of all human chromosomes and rapid development of other banding techniques, as well as the publication of the first banded human karyogram in 1970. Currently, G-banding is the most commonly used banding technique due to its permanence and clarity when compared to Q-banding fluorescent patterns **(110)**.

Since 1960, the International System for Human Cytogenetic Nomenclature (ISCN) has been used by cytogeneticists world-wide to describe human chromosomes in a concise way. It includes idiograms for every chromosome, identifiable bands in different banding resolutions, definition and abbreviation of chromosome abnormalities, and how a karyotype should be written down **(111)**. Chromosomes are divided at the centromere into short arm (p) and long arm (q). Based on G-banding patterns, chromosome regions are divided into bands and consecutive subbands, and given a numerical designation according to their distance from the centromere **(112)**. The cytogenetic techniques used in this thesis will be described below.

Cytogenetic slide preparations

In order to analyze mitotic chromosomes using microscopy, it is necessary to harvest cells and prepare them before producing slides with metaphase spreads. There is variation in protocols between different laboratories in terms of reagents, incubation time and concentrations, and based on the cell type being used. However, all protocols follow the same principles, and the main steps remain similar. Briefly, cultured cells are incubated with Colcemid to arrest them in metaphase and increase the number of metaphase chromosomes for analysis. The next step is to make cells swollen so they burst and chromosomes spread more easily on the slide, a procedure that involves incubation with a hypotonic solution, usually 0.075 M KCl for lymphoblasts. Afterwards, cells are washed 2-3 times with fixative solution, consisting of 3:1 methanol:acetic acid. Fixed cells, commonly referred to as fixatives, can be stored at -20°C for decades, or directly used for slide preparation by dripping a few drops on a glass slide and letting it air dry. Slides with metaphase spreads can also be stored at -20°C, or used directly for staining. If not stored, slides must be aged before staining to ensure that metaphase chromosomes will not degrade, and aging can be achieved by air drying the slides for a couple of days, microwaving them for 1 to 2 minutes or placing them on a 60°C slide oven overnight.

G-banding

The G-banding technique was described in 1971 by Drets and Shaw, as they noticed that treating metaphase spreads with NaOH followed by sodium chloride-trisodium citrate and staining with Giemsa resulted in chromosome banding patterns **(113)**. Such a technique created dark bands where chromosomes had lower G-C content, alternated by lighter interbands. G-banding formed banding patterns equal but opposite to Q-banding, meaning that G-banding lighter interbands were the stained darker bands in Q-banding. G-banding became preferably used rather than Q-banding since it was fluorescence-free and could be analyzed in a classic optical microscope **(112, 113)**.

Fluorescence in situ hybridization (FISH)

The first technique of *in situ* hybridization (ISH) for labeling specific DNA regions on metaphase chromosomes was described in 1969, using radioactive labeling as the principle **(114)**. In the 1980's and early 1990's, the field of molecular cytogenetics spread as researchers discovered that DNA could be labeled with biotin, and fluorochromes – fluorescent dyes – were developed for a safer, nonradioactive DNA-labeling, known as fluorescence *in situ* hybridization (FISH)

(112). Suitable for both metaphase chromosomes and interphase cells, most FISH protocols consist of pretreating the cytogenetic slides with SSC buffer followed by pepsin to enhance permeability. Fluorescent probes are dripped on the slide, which is shortly incubated at high temperatures (~72°C) for DNA denaturation followed by a long incubation time at 37°C for hybridization. After washing, slides are commonly counter-stained with DAPI and can then be analyzed on a fluorescence microscope. Commercially available probes vary in color, which allow multiple regions to be inquired simultaneously, and can be sequence-specific or whole chromosome paint (WCP).

Cohesion defects assay

As described in the introduction of this thesis, the normal progression of the M phase in cell division includes the alignment of mitotic chromosomes in the metaphase plate, where kinetochores of each sister chromatid are attached to microtubules from opposing poles **(3)**. The cohesin complex counteracts the pulling forces of microtubules by keeping sister chromatids united or cohesed, a mechanism that is essential to pass the SAC checkpoint and ensure proper chromosome segregation. Premature separation of sister chromatids, still in metaphase, may cause chromosome mis-segregation and lead to aneuploidy **(8)**. Sister chromatid cohesion defects began to be quantified in a systematic way by Barber in 2008 **(34)** to measure the effects of cohesion-related genes in colorectal cancer cell lines, which was similarly done by Sajesh in 2013 using Hodgkin lymphoma cell lines **(115)**. In both studies, the term primary constriction gap (PCG) is used to describe gaps between the centromeres, a clear sign of premature separation of sister chromatids during metaphase (**Figure 3**). Although mitotic chromosomes are either cohesed or not, cohesion defects as a whole can exist at different levels. We can measure the severity of cohesion defects by counting the number of chromosomes displaying PCG in a single cell, and cells within a case can have different scores, meaning that a patient case may have from mild to very severe defects, or perhaps zero to severe defects. One can also picture the overall incidence of cohesion defects by taking the average of cells with any level of PCG and presenting the result as total percentage of PCGs.

In this thesis, we measured both total percentage of PCGs and severity of defects, based on the PCG classification used by Barber and Sajesh **(34, 115)**, with modifications we found pertinent when analyzing high hyperdiploid cells. Using both FISH slides stained with DAPI and G-banded slides for analysis, we classified the severity of cohesion defects per cells as follows: PCG-I (mild), 1-4 chromosomes with PCGs; PCG-II (moderate), 5-19 chromosomes are affected; PCG-III (severe), 20 or more, but not all, chromosomes are affected; PCG-IV (very severe), complete loss of sister chromatid cohesion.

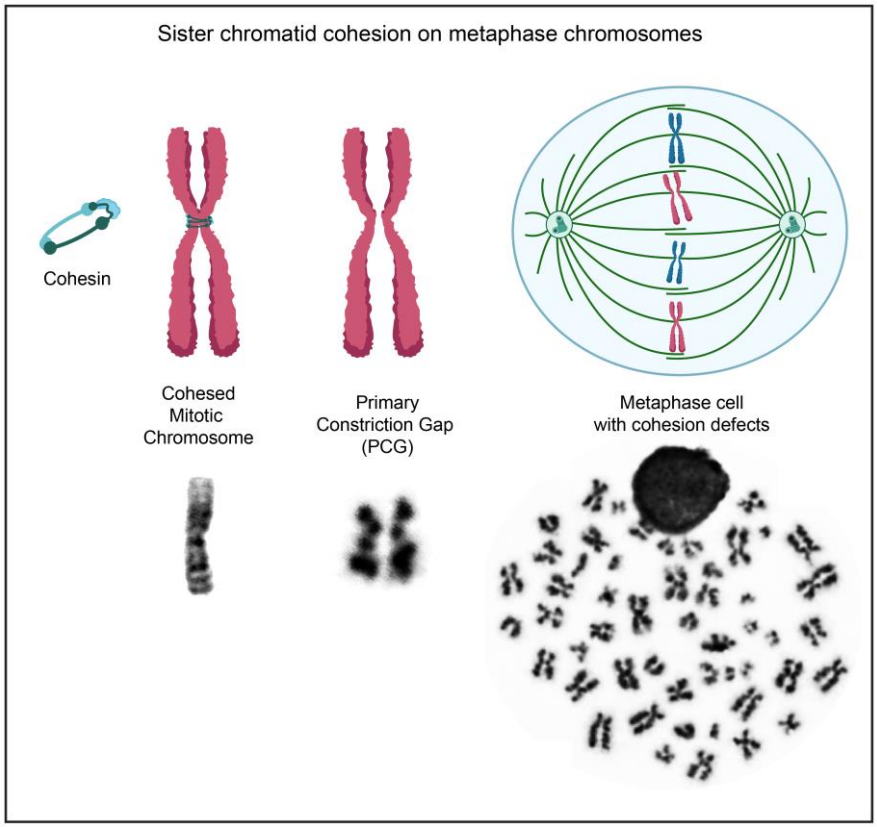


Figure 3 | **Detection of sister chromatid cohesion defects.** Illustrations (above) and microscopy pictures (below) of how we performed the cohesion defects assay to detect premature separation of sister chromatids during metaphase in leukemic samples. A completely cohesed mitotic chromosome shows no gaps between their duplicated (sister) chromatids. When cohesin fails to keep sister chromatids together at the onset of metaphase, we can observe a space or primary constriction gap (PCG) in the middle of the chromosome, indicative of defective cohesion. On the same cell under metaphase, there can be cohesed chromosomes and chromosomes with PCGs. The number of chromosomes displaying PCGs can be used to measure the severity of cohesion defects on a cell. Created using Biorender and pictures from our own patient cohort.

Chromosome morphology study

A common opinion among cytogeneticists is that leukemic cells have poor metaphase chromosome morphology when compared to normal cells, and HeH ALL cells have particularly bad chromosome morphology. Therefore, we developed a method to score metaphase chromosome morphology in order to compare average scores between different BCP ALL subtypes in a quantitative way. As classic cytogenetics may be subjected to the eyes of the analyst, we established features that should be accounted for when classifying chromosome morphology, which

includes chromosome condensation, band resolution, and overall shape and appearance.

The chromosome morphology score (CMS) is preferably assessed in G-banded slides, and each cell is scored as follows (**Figure 4**): CMS 1 – poor morphology, where chromosomes are too condensed to display any substantial banding pattern, are difficult to identify, and have poorly-defined shape, usually with a "fuzzy" appearance; CMS 2 – fair morphology, where chromosome resolution is approximately 200-300 bands, chromosomes are less constricted and have a sharper appearance, facilitating their identification; CMS 3 – good morphology, where chromosomes are long and display banding patterns above 350 bands, with well-defined shape allowing easy karyotyping.

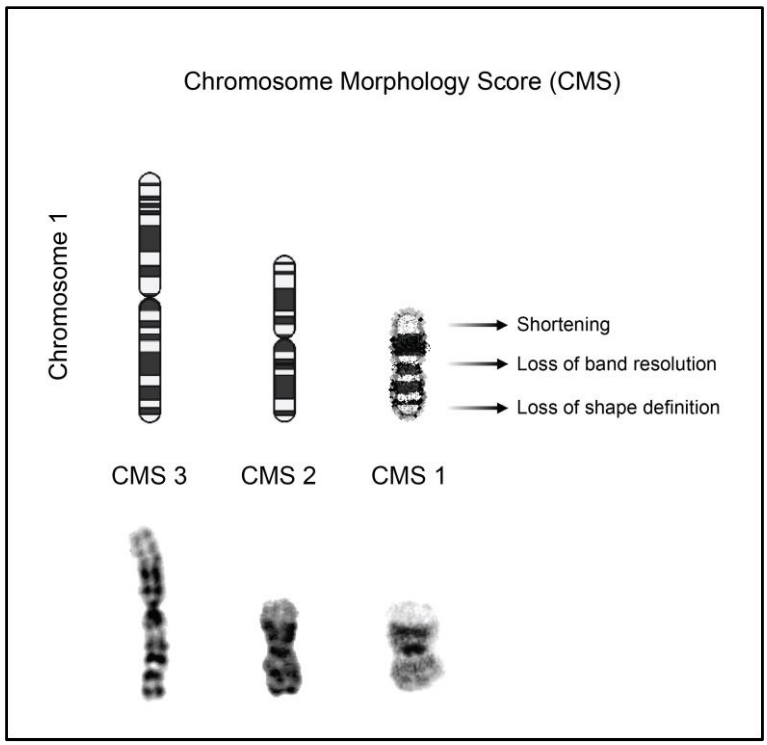


Figure 4 | **Chromosome morphology score.** Illustrations/Idiograms (above) and microscopy pictures (below), using chromosome 1 as an example, of how we analyze mitotic chromosome morphology (Chromosome morphology score, CMS). Each cell receives an average score based on all of its chromosomes, ranging from 1 (poor) to 3 (good). Band resolution, level of compaction and overall shape are considered when scoring chromosome morphology. Created using Biorender and pictures from our own patient cohort.

Immunofluorescence

Immunofluorescence can be applied in a wide variety of tissues and cell types, and target many different components using antibodies tagged with fluorochromes, which can then be visualized by fluorescence microscopy **(116)**. In this thesis, I utilize immunofluorescence to visualize microtubules and centrosomes during mitosis and investigate the incidence of mitotic defects in leukemic cell lines with low cohesin expression.

Briefly, a poly-L-lysine coated glass slide should be used for preparation of non-adherent cells, such as lymphoblasts. The cell suspension is incubated on the coated slide, followed by incubation with paraformaldehyde, a less aggressive fixative agent that ensures cells remain intact. The slides are then treated with a blocking solution based on bovine serum albumin (BSA) to minimize unspecific antibody absorption. Afterwards, slides are incubated with primary antibodies, which in our studies were targeting microtubuli (anti-tubulin alpha) and centrosomes (anti-tubulin gamma). Lastly, we perform incubation with secondary antibodies compatible with the primary ones (i.e. same donor species) and that are conjugated with fluorochromes, and slides are mounted with DAPI before microscope analysis.

Gene knockdown using short hairpin RNA (shRNA)

Many functional studies wish to investigate the effect on live cells of lowering the expression (knockdown) or deactivating (knockout) a target gene, and the use of RNA interference (RNAi) has become a very popular and powerful tool for such purposes. The procedure consists of delivering double-stranded RNA – identical to the target sequence – to the cell, leading to degradation of the host messenger RNA (mRNA) and thus affecting the gene expression of the target. Among other options for shRNA vector production and delivery, lentiviral-mediated transduction is a popular choice since it is straight-forward and less toxic for the targeted cells **(117)**. After vector production, the lentivirus is added to cultured cells for the transduction, including a 60-minute centrifugation step and overnight incubation at 37°C **(118)**. The transduced cells – which now contain a fluorescently labelled vector – are sorted by fluorescence-activating cell sorting (FACS) after 48 hours, cultured for one week, and quantitative PCR (qPCR) is performed to inquire the expression level of the targeted gene, or knockdown efficiency.

Proteomic techniques

Mass spectrometry

Mass spectrometry (MS) is an indispensable tool in proteomic studies, among other fields, due to its ability of both quantifying and identifying proteins, which can be used to investigate protein expression levels **(119)**. The sample preparation for MS usually involves as main steps protein precipitation and filtration, phase extraction and affinity enrichment. When samples are loaded into the mass spectrometer, they go through a first stage of ionization, followed by separation of ions based on their mass-to-charge ratio. Finally, ions are measured and can be visualized in a mass spectrum chart **(120)**. Single-stage MS is commonly used when the focus of a study is to measure the molecular mass of a polypeptide. However, to retrieve information about posttranslational modifications or amino acid sequences, tandem mass spectrometry (MS/MS) is performed instead. In MS/MS, selected ions are fragmented through collision after mass determination, and analysis of the masses of such fragments will generate information about additional structural features of the targeted peptides **(119)**.

Genotyping and sequencing-based techniques

Single-nucleotide polymorphism (SNP) array

Single-nucleotide polymorphism (SNP) array analysis, which was first developed for SNP typing to enable genome-wide association studies (GWAS), is a genome-wide genotyping method that uses a large number of allele-specific probes synthetized on microarrays and is often used to detect copy number alterations (CNAs) and allelic imbalances. SNP array can only detect alterations that cause copy number changes, missing balanced chromosome rearrangements. Another limitation of this technique is that one cannot infer the exact position of an alteration, but rather in which SNP range it is located. Nevertheless, this method is very cost-efficient, allows rapid and high-throughput information on CNAs and it can detect loss of heterozygosity caused by UPIDs **(121)**.

Next-generation sequencing

The first-generation sequencing method known as Sanger sequencing was described in 1977, using chain-terminating labelled dideoxynucleotides, fragmentation and size separation to sequence DNA strands based on a DNA template **(122, 123)**. As this method was time-consuming, expensive, and limited in terms of depth, there were constant efforts to develop new technologies. In the early 2000's, next generation sequencing (NGS) emerged, allowing high-throughput parallel sequencing of single DNA molecules. Briefly, NGS is based on short reads, where samples are fragmented, DNA ends are repaired and ligated with adapters, followed by surface attachment and *in situ* amplification, in a way that millions of sequencing reactions occur simultaneously. Third-generation sequencing refers to long-read sequencing approaches, capable of reading fragments of up to 10 Kb, and such technologies are superior to short-read NGS in terms of *de novo* assembly, handling genome-wide repeats and structural variants detection **(123)**.

Usually, considering that both forward and reverse reads of a DNA template in a library have equal probability of being sequenced, NGS methods sequence only one end of the DNA templates **(124)**. However, forward and reverse reads can be paired to map both ends of DNA fragments. In mate-pair sequencing, library preparation includes circularizing and biotin-labelling the DNA ends that were brought together, followed by further fragmentation, and allowing mapping of long-distance genomic regions, which is advantageous for interrogating whole genomes with less sequencing coverage **(124, 125)**. In paired-end sequencing, DNA is fragmented, repaired and ligated with adapters, then both ends of a linear DNA fragment can be mapped. Since the distance between each paired read is known, paired-end sequencing allows higher coverage and better mapping over repetitive regions **(124, 126)**.

Whole-genome sequencing (WGS)

Whole-genome sequencing (WGS) is a powerful tool for genome-wide investigations and has been paramount to study cancer biology. This method can be used for sensitive detection of small variants, copy number variations, loss of heterozygosity and structural variants. WGS can be used to inquire functional predictions, mutation signatures, pathway integration and therapeutic targets, making it one of the most complete approaches for in-depth cancer research. Although efficient, WGS is very expensive compared to targeted deep sequencing

due to the sequencing depth and amount of data required to study the whole genome **(127)**.

Whole-exome sequencing (WES)

Whole-exome sequencing (WES) has become the most popular targeted enrichment approach among sequencing methods. It consists of sequencing all gene-coding regions, or exons, of the genome, being far more feasible since it requires approximately 2% of sequencing load compared to WGS. Although genomic information from non-coding regions is not accounted for, 85% of disease-related mutations occur within exons, and since the whole exome is sequenced, there is no need for selection of candidate genes, a requirement of other targeted methods **(128)**.

Single-cell whole-genome sequencing (scWGS)

Single-cell sequencing methods have emerged in the early 2010's as a powerful tool to study subtle differences in the genome using NGS at the single cell level. The approach requires single-cell isolation, usually by FACS, DNA extraction and amplification, and library preparation, followed by high-throughput sequencing. In terms of application, scWGS is frequently used for investigating tumor heterogeneity, clonal evolution, and chromosomal instability in cancer. Moreover, single-cell technology can also be coupled with RNA sequencing and epigenetic analysis for in-depth transcriptome and methylome studies **(129)**.

RNA sequencing (RNA-seq)

Not long after the emergence of NGS, RNA sequencing was developed as a way to study gene expression, translation, and RNA structure. Sample preparation includes RNA extraction, mRNA enrichment and cDNA synthesis, followed by the usual NGS library preparation. The sequencing depth required, even for high-throughput studies, is considerably lower when compared to WGS, ranging from 10 to 30 million reads per sample. RNA-seq is commonly used in cancer research for differential gene expression (DGE) studies and for fusion gene detection, since detecting structural rearrangements at the genome level does not necessarily mean that a fusion gene will be translated **(130)**.

High-throughput chromosome conformation capture (Hi-C/Micro-C)

Chromatin architecture at the interphase level can be investigated through chromosome conformation capture (3C) methods. In the early 2000's, genome-wide 3C methods began to be developed, using proximity ligation as its basis **(21)**. Briefly, intact nuclei are submitted to covalent crosslinking to bind genomic loci that are physically close in the 3D setting of the cell. The chromatin is then fragmented using restriction enzymes (used in Hi-C) or micrococcal nuclease (MNase) (used in Micro-C), and usual NGS library preparation is carried out for paired-end sequencing (**Figure 5**). The greatest difference between Hi-C and Micro-C that impacts resolution is the enzyme digestion. While restriction enzymes cleave DNA only in their specific recognition sites, resulting in uneven coverage of the genome, MNase recognizes nucleosomes and cleaves DNA into mono-, di- and trinucleosomes, which guarantees a better coverage and nucleosome-level resolution **(131)**. Both Hi-C and Micro-C require deep sequencing to allow visualization of fine-scale chromatin structures. While 600 million reads suffice for WGS studies, genome-wide 3C methods often surpass 1 billion reads.

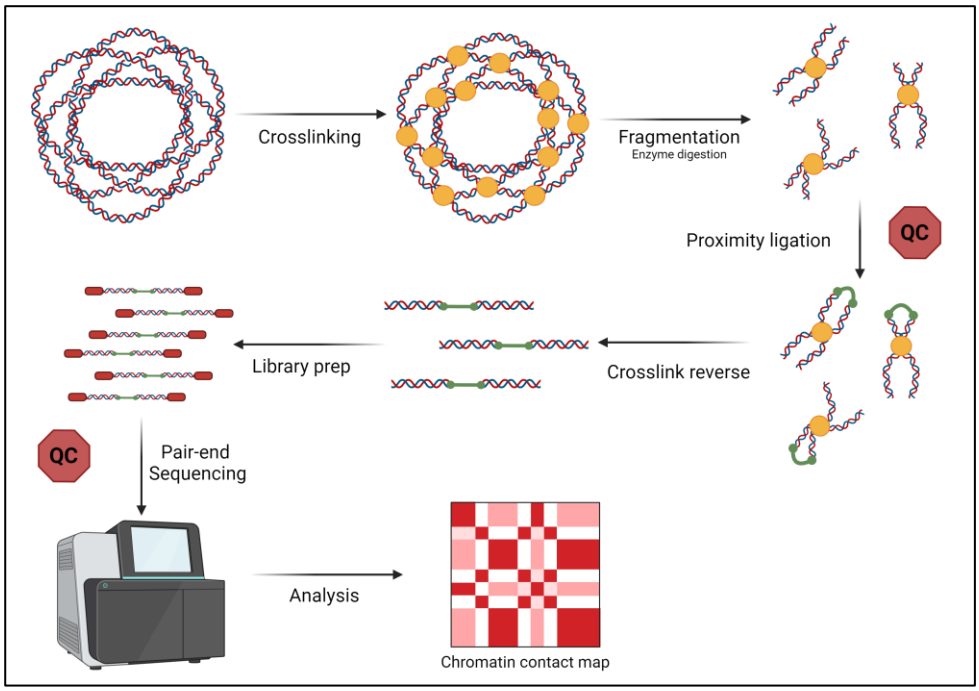


Figure 5 | **Micro-C protocol for generating 3D genomic data.** Proximity ligation-based Micro-C protocol, that begins with fixing live cells to crosslink genetic regions that are in contact in the 3D setting of the nucleus. MNase is used to fragment the crosslinked DNA, followed by quality control check to ensure that fragments have sizes corresponding to mono-, di- and trinucleosomes. Afterwards, bridge ligation is performed followed by crosslink reversal, resulting in hybrid, linear fragments in which library preparation can be normally carried out for pair-end sequencing. A shallow pair-end sequencing is performed first to ensure the quality of the libraries, followed by high-throughput sequencing. Finally, bioinformatic tools are used for identifying chromatin structures and generating contact maps that will be used for analysis.

Results

Article I

Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia

In ***Article I***, we performed analyses on matched datasets of WGS/WES/SNP array and RNA-seq including 48 HeH and 41 *ETV6::RUNX1* ALL primary samples, and proteomic analysis of 18 HeH and 9 *ETV6::RUNX1* cases using liquid chromatography MS/MS. We detected 8222 proteins that were ascertained by RNA-seq, and expression levels were positively correlated for 75% of the mRNA-protein pairs. By comparing the mean RNA and protein expression according to the copy number of HeH cases, we observed a clear dosage effect – genes and proteins with higher copy number displayed higher expression, although a negative correlation between gained chromosomes and expression was seen in 16% of the genes and 25% of the proteins. Moreover, RNA expression of 83 cases from a previous study was used to investigate *cis* – genes within a region – and *trans* – genes in other genomic regions – dosage effects of the copy number gains in HeH ALL, which showed that chromosome gains have a genome-wide impact on gene and protein expression, but with no correlation with known cancer driver genes.

The proteomic landscape of HeH and *ETV6::RUNX1* ALL are different from each other, as shown by hierarchical cluster and principal component analysis (PCA) in our study. The HeH subtype displayed upregulation of 2423 genes and 1286 proteins, and downregulation of 2222 genes and 1127 proteins compared to *ETV6::RUNX1* cases. Top differentially expressed proteins included known players in ALL, including previously reported high expression of CD44 and FLT3 in HeH ALL, and IGF2BP1, CLIC5, RAG1 and RAG2 in *ETV6::RUNX1* ALL. Investigation of pathway dysregulation by gene set enrichment analysis showed that pathways related to translation and ribosomes, innate immunity, cell adhesion, cytokines and activated signaling, protein folding and proteolysis, and the endosome are enriched in HeH ALL. Meanwhile, enriched pathways in *ETV6::RUNX1* ALL

included chromatin organization and modification, G2/M checkpoint and mitochondria.

Our proteogenomic analysis further revealed that CTCF and members of the cohesin complex display low levels in HeH compared to *ETV6::RUNX1* ALL and normal B-cells. We could also show that the DEG between HeH and *ETV6::RUNX1* were strongly correlated to CTCF binding sites, indicating that CTCF and cohesin levels have genome-wide effects in HeH ALL. We further explored this matter by using publicly available data to compare the expression of gene pairs that belong to the same TAD versus genes separated by TAD boundaries, as genes within a TAD should be similarly regulated. All datasets, which included childhood ALL cases (n=201), AML cases (n=151) and papillary renal-cell carcinoma (n=270), showed higher correlation between gene expression of gene pairs from the same TAD. When performing the same type of analysis in our HeH and *ETV6::RUNX1* ALL cases, the HeH samples showed no difference in expression of inter- and intra-TAD gene pairs, and such results did not change when analyzing only commonly gained or non-gained chromosomes.

In light of our findings, we next performed Hi-C in four HeH and two *ETV6::RUNX1* ALL primary samples from our cohort. At the A/B compartment level, HeH cases were 90% similar to publicly available Hi-C data from the GM12878 cell line, and 98% similar to *ETV6::RUNX1* cases. At the TAD level of resolution, three out of the four HeH ALL samples displayed longer TAD structures and fewer boundaries compared to *ETV6::RUNX1* ALL, a result from partial TAD fusions. Weakened or absent TAD boundaries were frequently seen in at least two HeH ALL cases (131 boundaries in total, 97% overlap with CTCF binding sites), while *ETV6::RUNX1* samples were missing only 14 TAD boundaries. Additionally, a total of 134 (45%) genes and 65 (31%) proteins encoded within 1 Mb of boundary losses were differentially expressed between HeH and *ETV6::RUNX1* samples, and were more often down-regulated in HeH ALL. Analysis of boundary strength by directionality index and insulation scores showed that TAD boundary positions that remained unchanged between the two ALL subtypes lost strength in some of the HeH samples (2 out of 4), further confirming the relationship between abnormal chromatin architecture and transcriptional dysregulation in at least some HeH ALL cases. Lastly, we investigated whether chromosome architecture at the metaphase level was also disorganized in HeH ALL by scoring chromosome morphology in a total of 37 HeH and 33 *ETV6::RUNX1* samples. Comparison of mean CMS revealed that mitotic chromosomes display significantly poorer morphology in HeH ALL than *ETV6::RUNX1*, in line with the Hi-C and RNA-seq data.

Article II

Sister chromatid cohesion defects are associated with chromosomal copy number heterogeneity in high hyperdiploid childhood acute lymphoblastic leukemia

In ***Article II***, we performed knockdown of the cohesin member *RAD21* in REH cells, an *ETV6::RUNX1*-positive ALL cell line, and analyzed the incidence of cohesion defects. The *RAD21*-knockdown (RAD21-KD) cells displayed a higher percentage of PCGs compared to controls (33-36% versus 4% in control cells) as well as higher severity, ranging from PCG I-III (only PCG I seen in control cells). Next, we carried out the cohesion defects assay on primary samples of HeH (n=45) and *ETV6::RUNX1* (n=37) ALL, where PCGs were detected in 86% and 49% of the cases, respectively. The incidence and severity of cohesion defects in the HeH ALL subtype ranged from 0-85% of total PCGs and from mild defects to complete loss of cohesion (PCG I-IV), although moderate defects or PCG II were the most common (40% incidence). *ETV6::RUNX1* ALL cases displayed from 0-18% of total PCGs and predominantly mild (PCG I) defects.

Afterwards, we performed interphase FISH on HeH samples classified as high PCG levels versus low PCG levels to examine whether cohesion defects were associated with chromosome copy number heterogeneity. We compared copy number changes of chromosomes X, 2, 3, 6, 10 and 21 on cases that shared the same ploidy for the given chromosome, where HeH ALL samples with high PCG levels displayed significantly higher copy number heterogeneity for chromosomes 3, 6, 10 and 21. Similar analysis on the RAD21-KD cells versus control showed increased copy number variation for chromosome 21 in knocked down cells, but not for chromosomes X, 2 and 3. Immunofluorescence was also performed in the RAD21-KD and control cells, and while no disturbances were found in control cells, the RAD21-KD cells displayed 6.5% of mitotic aberrations, including lagging chromosomes/chromatin bridges, and mono-, tri- and tetrapolar mitoses.

Finally, we combined our cytogenetic data with RNA-seq data (HeH n=36; *ETV6::RUNX1* n=32) from previous studies to infer whether sister chromatid cohesion defects were associated with levels of cohesin, condensin, or both. By classifying cases into low or high mRNA expression per gene, we compared the PCG percentage of cases from each group and found that low expression of *RAD21* is associated with higher PCG levels. Likewise, low expression of the condensin subunit *NCAPG* correlated with incidence of cohesion defects.

Article III

The 3D genome of pediatric B-cell precursor acute lymphoblastic leukemia

In this study we used the genome-wide 3C method Micro-C to investigate the chromatin architecture of 33 BCP ALL primary patient samples with matched RNA-seq and cytogenetic analyses, including the genetic subtypes HeH (n=14), *ETV6::RUNX1* (n=8), *TCF3::PBX1* (n=4), *DUX4*-rearranged (n=2), *BCR::ABL1* (n=1), iAMP21 (n=1), *KMT2A*-rearranged (n=1), near-haploid (n=1) and near-triploid (n=1). We profiled the A/B compartments of BCP ALL cases and compared them to publicly available Hi-C data from peripheral blood mononuclear cells and CD34+ HSC samples (referred to as normal cells). We observed that BCP ALL cases have significantly less open chromatin than our normal cells control, and among the BCP ALL genetic subtypes, *ETV6::RUNX1* cases displayed the least amount of open chromatin. We also observed a negative correlation between levels of open chromatin and both TAD boundary strength and total number of TADs. We further examined A/B compartments shifts between the HeH, *ETV6::RUNX1* and *TCF3::PBX1* cases, identifying 235 genes located in such regions, and discovered that the first two subtypes were more similar than the latter, and compartment shifts were accompanied by changes in gene expression.

In regard to TAD organization, we identified 10318 TAD boundaries at 25 kb resolution. PCA based on TAD boundary strength revealed that the three main genetic subtypes of our cohort form separate clusters. Cluster 1 comprised all aneuploid cases except the near-triploid, i.e. HeH cases, a near-haploid and a *BCR::ABL1* case with 51 chromosomes. Cluster 2 included all *ETV6::RUNX1* cases together with the iAMP21 case and one *DUX4*-rearranged case with an *ETV6* deletion, and cluster 3 consisted of the *TCF3::PBX1* cases and the near-triploid case, while the remaining BCP ALL cases were not in any cluster. Cases in cluster 1 displayed a considerable reduction in total number of TAD boundaries, lower TAD strength scores and increased TAD length when compared to clusters 2 and 3. Agreeing with our observations in ***Article I***, TAD boundary loss in cluster 1 was accompanied by significantly lower expression of CTCF, and, interestingly, gained TADs in this cluster showed a reduction in CTCF binding sites, indicating a possible CTCF-independent mechanism for new TAD formation in aneuploid samples.

Next, we merged Micro-C data from cases belonging to the same genetic subtype to generate high-resolution maps and perform loop calling at 5 kb. Analysis focused on leukemia-related genes evidenced that the high expression of *FLT3* in HeH ALL was associated with increased chromatin interactions between the *FLT3* promoter and an enhancer in *PAN3* intron 8, in a CTCF-independent manner. We also observed lower expression of *IKZF1* in cluster 1 correlating with reduced interactions of CTCF-dependent E-P loops. Additionally, histone modifier genes *EP300*, *EZH2* and *SETD1B* showed loss of E-P loop strength and gene expression alterations in cluster 1. Transcription dysregulation associated with chromatin changes were also identified in cluster 2 for the known cancer-related genes *BRAF*, *MYC* and *NOTCH1*.

This study also included investigating the chromatin architecture of mitotic chromosomes, using the cohesion defects assay and chromosome morphology score combined with RNA-seq and Micro-C data. Consistent with ***Article II***, HeH ALL cases displayed considerably higher levels of cohesion defects and one of the worst median CMS, while *ETV6::RUNX1* and *DUX4*-rearranged cases consistently showed good mitotic chromosome structuring. Both cohesion defects and chromosome morphology significantly correlated with total number of TADs, as fewer TADs were associated with high PCG levels and poorer CMS. Moreover, we found that low expression of cohesion subunit *SMC1A* correlated with good chromosome morphology, while condensin subunits *NCAPG2* and *SMC4* low expression coincided with higher PCG levels.

As Micro-C data can be used for structural variant (SV) calling, we investigated the incidence of these events in 31 cases of our cohort with supporting data from WGS (n=19), SNP array (n=31) or scWGS (n=8). We detected 116 SVs, which mostly were intrachromosomal rearrangements, and validated 86.2% of them with supporting data. Moreover, we identified the formation of 24 neo-TADs and 278 neoloops caused by somatic SVs. We showed that neoloops driven by *PAX5* deletions increased the expression of truncated *PAX5* transcripts in four *ETV6::RUNX1* cases, and a t(1;19) derived neoloop potentially results in upregulation of *UHRF1* by enhancer-hijacking in a HeH ALL case.

Article IV

Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukemia

Here, we explored the clonal origin and heterogeneity in HeH ALL by using scWGS of nine primary patient samples, bulk copy number data from WGS, WES and SNP array of 577 samples, and *in silico* modelling. Copy number analysis of single cells of nine HeH samples revealed very homogeneous genomes, where 5 out of 9 cases had the same chromosomal content in more than 99% of the cells, and 4 of the cases displaying 3 to 5 subclones with whole chromosome changes. Heterogeneity scores were calculated for each sample, showing that cases with higher scores also had higher subclonality, which was associated with high levels of cohesion defects for one sample but not for the others. Analysis of the phylogenetic trees showed early chromosome gains and late structural changes, suggestive of punctuated evolution. Moreover, many cases displayed copy number changes of the same chromosome in different clonal events, revealing strong selective pressure for gains of chromosome 21 and 17, and losses of chromosome 9.

We further explored selective pressures in primary HeH samples by analyzing bulk copy number data of 577 cases. A very strong selection for extra copies of chromosomes 21 (100%), X (97%), 14 (95%), 6 (89%), 18 (83%), 4 (82%), 17 (78%) and 10 (74%) were observed, followed by additional common gains of chromosomes 8 (38%), 5 (23%), 9 (19%), 11 (14%), 12 (14%), and 22 (11%). While the remaining chromosomes seem to have a neutral stance, chromosomes 13 and 20 may possibly be selected against as they were recurrently monosomic. UPIDs were seen in 36% of the cases in a ratio of 0-5% of UPID/all disomies, with the exception of UPID 9 which was present in a ratio of 17%. In total, 72% of the cases in bulk analysis did not have detectable subclones, in line with the scWGS findings, and chromosomes 9 (8.7%), X (5.1%), 8 (4.5%) and 21 (3.6%) were the most commonly involved in subclonality. Agreeing with our observations, investigation of matched diagnostic and relapsed samples showed a positive selection for chromosome 8 and negative for chromosome 9. Additionally, we also noticed that cases with modal chromosome number (MNC) 51-61 (n=545) and 62-67 (n=32) were presented with different ratios of trisomy/tetrasomy, which was further explored in our simulations.

In this study we have also explored the possible origins of high hyperdiploidy in BCP ALL by *in silico* modelling. We simulated 50000 cells over several generations

taking into consideration the chromosome gains with strong positive selection and the average UPID ratio (~ 2.5%) seen in primary HeH ALL, exploring five possible routes to aneuploidy: (1) diploid cell with sequential gains; (2) initial tetraploidy with sequential chromosomal losses; (3) diploid cell with a tripolar division; (4) initial tetraploid cell with tripolar division; (5) mitotic catastrophe by complete loss of sister chromatid cohesion. While simulations were stopped when UPID levels reached 2.5%, routes (4) and (5) were removed from further testing due to inconsistently higher UPID frequencies (initial values of 18.2% and 9.9%, respectively). Considering HeH cells with MNC 51-61, route (2) was also discarded as tetraploidy with sequential losses resulted in very few cells with such MNC, and we continued the investigations by comparing the pattern of trisomies and tetrasomies between primary samples and routes (1) and (3). Although the chromosome pattern of both remaining routes fit well with the patient data, the allelic ratio of 3:1 on tetrasomies had a much higher frequency in route (1) than in route (3) and patient data, supporting a diploid/tripolar division origin. Regarding HeH with MNC 62-67, both routes (2) and (3) agreed with patient data, and as such cases also displayed higher levels of subclonality, it is possible that HeH ALL with MNC 51-61 and MNC 62-67 have different mechanisms of development, but the great majority of HeH ALL cases likely originate from a tripolar division on an initial diploid cell as a punctuated evolution event, followed by low-level clonal evolution. In line with our hypothesis, frequency assessment of mutations that occurred before or after trisomy (BTRI or B/ATRI) in primary samples indicated that trisomies of chromosomes with strong positive selection harbor more B/ATRI mutations and thus were more recent than the ones with neutral or negative selection.

Lastly, we also observed that chromosome rearrangements and somatic mutations occur later than the bulk chromosomal gains. Most structural rearrangements were subclonal, including duplication of 1q, deletion of 6q and gains of 17q, while isochromosome 7q was usually present in the major clone. Between 10 and 40% of the primary samples harbored subclones with deletions in *IKZF1, CDKN2A, PAX5, ETV6, CREBBP*, and *TCF3*. Analysis of 338 driver mutations in 218 samples with either WES or WGS available also revealed that 44% of the events were subclonal, and mutational age estimation in trisomies and tetrasomies showed that 92% of driver mutations were B/ATRI, and only five – including a *IKZF1* mutation – were B/ATRI events.

Discussion

What drives leukemogenesis in HeH ALL?

Taking into consideration that hyperdiploidy is an early event and a driver change in leukemogenesis, a long-lasting question that hovers around HeH ALL is what the effects of aneuploidy in this subtype are. A recurrent hypothesis is that genes located in the commonly gained chromosomes would be overly expressed, a phenomenon known as dosage effects. Indeed, studies have shown higher RNA expression of genes located in gained chromosomes **(40, 132, 133)**. In ***Article I*** we addressed the impact of copy number alterations in HeH compared with *ETV6::RUNX1* ALL in a comprehensive way, combining RNA-seq, MS/MS, WGS, WES, SNP array and to some extent Hi-C.

Although RNA and protein levels are directly correlated, posttranslational regulation mechanisms may interfere in protein expression, and a given gene that is highly expressed in the transcriptome might not translate into high protein levels **(134-136)**. This reinforces the importance of incorporating both transcriptome and proteome in cancer research. Here, we compared the mean RNA and protein expression according to chromosome copy number and showed a clear dosage effect, as expected. Gains of X, 14 and 21 displayed stronger dosage effects, an interesting discovery since chromosome 21 is acquired in all HeH cases and chromosomes X and 14 are the second and third most gained ones **(62, 100, 101)**. Intriguingly, 16% of the genes and 25% of the inquired proteins showed instead a lower expression on gained chromosomes, meaning that not all genes are implicated by dosage effect. We further investigated the consequences of aneuploidy by inquiring *trans* effect, or the effect of chromosome gains in other genomic regions, and we found that hyperdiploidy has a wide-spread influence on the transcriptome and proteome.

The proteogenomic study in ***Article I*** also culminated in the discovery that CTCF and members of the cohesin complex display low expression in HeH ALL compared to *ETV6::RUNX1* and normal pre-B cells, which was further explored in ***Articles II***

and ***III***. CTCF and cohesin are essential for topological regulation of gene expression and are recurrently involved in several cancer types **(5, 11, 34, 115, 137)**. Disruption of TAD boundaries allow chromatin contacts that were otherwise insulated, leading to abnormal gene expression **(11)**, and may cause oncogene activation and/or tumor-suppressor inactivation **(11, 137)**. This led us to investigate in ***Article I*** if the wide-spread transcription dysregulation seen in hyperdiploid cells were related to the low levels of these proteins. In fact, CTCF binding sites were considerably enriched in genes that were differentially expressed between HeH and *ETV6::RUNX1* ALL samples. Analysis of publicly available databases also showed that genes located within the same TAD were similarly expressed compared to genes from different TADs, an observation that held true for several cancer types, except for HeH ALL. Therefore, we performed Hi-C in four HeH and two *ETV6::RUNX1* ALL primary samples, which revealed clear abnormal chromatin architecture in HeH cells, even with the limited size of the cohort. Briefly, partial fusions of several TADs into one resulted in HeH samples displaying fewer number of TADs, although larger, compared to *ETV6::RUNX1* samples and GM12878 publicly available data. Moreover, 131 of the apparently intact TADs – those where size and location did not change – had weakened or absent boundaries, strongly associated with CTCF-cohesin binding sites. This permissive state of topological boundaries is likely one of the phenomena underlying the genome-wide dysregulation of gene expression in HeH ALL and could be an important feature that distinguishes this subtype. A recurrent event in the context of chromatin architecture aberrations in cancer is oncogene activation by enhancer-hijacking, where the lack of specific TAD boundaries leads to a TAD fusion, and an enhancer that was once isolated from a given proto-oncogene can now activate it **(137)**. In fact, our group has described in 2020 that deletions in 13q12.2, present in ~2% of BCP ALL cases, led to a TAD fusion in a HeH case that allowed *FLT3* upregulation through enhancer hijacking **(138)**.

Based on such discoveries, ***Article III*** was designed to thoroughly investigate the 3D chromatin architecture of a larger cohort of primary patient samples of pediatric BCP ALL using Micro-C, including all main genetic subtypes. PCA of TAD boundaries showed that the three most representative BCP ALL subtypes of our cohort, HeH (n=14), *ETV6::RUNX1* (n=8) and *TCF3::PBX1* (n=4), cluster separately, in line with previous studies where PCA on T-ALL primary samples also pointed that different genetic subtypes display distinct 3D genomic signatures **(139, 140)**. A very interesting discovery was that all aneuploid cases, except the near-triploid sample, clustered together, and this included HeH samples, one near-haploid

and one *BCR::ABL1* case with hyperdiploidy. Since the chromosome gains of the near-haploid (27/54 chromosomes) and BA_1 (51 chromosomes) included chromosomes X, 14 and 21, the question remains of whether they cluster with HeH cases due to similar chromosome pattern or simply because they share similar chromosome number. Consistent with our findings in ***Article I***, in ***Article III*** we observed overall weakened TAD boundaries, as well as fewer number of TADs, in the aneuploid group, and again affected TADs were strongly correlated to CTCF binding sites. Moreover, gained TADs in the aneuploid samples are not enriched for CTCF binding sites, suggestive of alternative mechanisms for TAD formation. It remains to be investigated whether these alternative mechanisms involve other recently described players in chromatin conformation, such as RNA pol II, YY1 or others **(25)**.

Both cohesin and CTCF are frequently mutated in different cancer types, including in acute myeloid leukemia **(141-143)**. Yet, such mutations are very rare in HeH ALL, and it remains to be answered whether other unidentified regulators/mechanisms are involved in the low expression of cohesin and CTCF in these samples or if the hyperdiploidy itself causes it. Studies on the effect of aneuploidy in chromosome territories have tried to clarify if the sole existence of extra material in the nucleus could culminate in genome-wide spatial changes **(144, 145)**. One study in particular showed that a cell line in which the only acquired somatic event was trisomy 7 displayed several *trans* effects, such as global gene dysregulation, large A/B compartment shifts in chromosome 14 and loss of TAD boundaries in chromosome 4 **(144)**. Single-chromosome gains also have an impact in the position of their own and other chromosome territories in the 3D setting of the nucleus **(144, 145)**. Hence, it would be interesting to investigate the impact of hyperdiploidy in chromosome territories in future studies, and perhaps this could bring us another piece of the puzzle that HeH is.

In ***Article I***, we inquired whether DGE associated with chromosome gains *per se* in HeH ALL was targeting oncogenes or tumor-suppressor genes but found no correlation. Regarding the abnormal chromatin architecture seen in HeH cells, in ***Article III*** we observed events associated with well-known leukemia-related genes. The high-expression of *FLT3*, a hallmark of HeH ALL, was found to be caused by increased chromatin interactions between the gene and a strong upstream enhancer, a specific feature of HeH cases that is unrelated to somatic mutations – HeH cases with deletions in 13q12.2 display even higher *FLT3* expression **(138)**. Furthermore, *IKZF1* low expression without the presence of deletions was also clarified by Micro-C, as this gene shows decreased E-P interactions, suggesting weakening of

chromatin loops. Lastly, we observed the effects of structural events in chromatin architecture as deletions in *PAX5* caused the formation of a neoloop that upregulates a truncated *PAX5* transcript. Previous studies have shown that neoloop formation enables enhancer hijacking that affects *TLX3*, *TAL2*, *HOXA* and *MYC* in T-ALL **(139, 140)**, and *HSF4*, *MYC* and *CBL* in AML **(146)**. In sum, many transcriptional features in leukemia can be explained by topological events in the 3D genomic landscape, and in ***Article III*** we bring a comprehensive study of the 3D genome of nine genetic subtypes of childhood BCP ALL.

Considering the functional roles of the cohesin complex during mitosis, we have also investigated metaphase chromosome architecture in HeH ALL and other BCP ALL subtypes in ***Articles I-III***. Firstly, we interrogated whether low expression of CTCF and cohesin affected the overall morphology of metaphase chromosomes in the same way as in interphase chromosome organization. Consistent with our hypothesis, we show in ***Article I*** that HeH ALL cases display poor chromosome morphology compared to *ETV6::RUNX1* ALL. In ***Article III*** we observed that CMS varies between genetic subtypes, and HeH samples continue to display low scores, but most importantly we saw a correlation between number of TAD boundaries and metaphase chromosome morphology, where fewer TADs were associated with low CMS. In ***Article II*** we scrutinized the incidence and extent of sister chromatid cohesion defects in HeH compared to *ETV6::RUNX1* ALL, an assay that was repeated for the cases included in ***Articles III*** and ***IV***. Although heterogeneous and not present in all cases, the incidence and severity of cohesion defects in HeH ALL is higher than in any other BCP ALL genetic subtype included in this thesis. In ***Article III*** we could also show a negative correlation between number of TAD boundaries and total PCG percentage in BCP ALL, in line with the known functions of the cohesin complex.

By combining RNA-seq data with the cytogenetic assays, we showed, in fact, a negative correlation between expression of the cohesin subunit *RAD21* and percentage of PCG in ***Article II***, as well as expression of the *SMC1A* subunit and chromosome morphology in ***Article III***. Another study has put into context impaired condensin complex and mislocated aurora B in HeH ALL as causative of cohesion defects, although the cohesin complex was not investigated (**141**). Likewise, we show in ***Article II*** a correlation between low expression of condensin subunit *NCAPG* and cohesion defects in a larger cohort of primary HeH samples, and again in ***Article III*** for the *NCAPG2* and *SMC4* subunits.

Taken altogether, ***Articles I-III*** paint a chaotic scenery in chromosome organization, both at the interphase and metaphase level, of HeH ALL primary samples. The abnormal interphase chromatin architecture in HeH samples has an evident impact on dysregulation of gene expression. Concurrently, the combination of low levels of CTCF and cohesin (***Article I*****),** and impairment of condensin and aurora B **(147)** may explain the low proliferative rates described in HeH ALL, as sister chromatid cohesion defects and mitotic chromosome organization defects delay the mitotic process, and very likely culminate in mitotic slippage **(62, 100, 147)**. We have yet to discover what causes low expression of CTCF and cohesin in HeH ALL, since no mutations are involved and the dosage effect seen in the subtype does not affect members of the cohesin complex, which are virtually all located in commonly gained chromosomes. As to how aneuploidy affects HeH ALL, the first clear answers are dosage effect and genome-wide *trans* effects in gene expression. Moreover, and most strikingly, chromosome gains seem to interfere in 3D genomic organization, as we observed that the whole aneuploid cluster in ***Article III*** displayed weakened and fewer TADs, even for a *BCR::ABL1*-positive case with hyperdiploidy, indicating that chromosome gains from different origins may have similar implications in TAD organization. In the future, functional studies of CTCF, cohesin and condensin knockdowns in aneuploid cell lines followed by Hi-C or Micro-C analysis could reinforce and help clarify the scenery seen in ***Articles I*** and ***III***, as well as investigating primary samples of other aneuploid cancer types. Additionally, investigating the spatial organization of chromosome territories in HeH and other aneuploid BCP ALL subtypes could bring valuable information of the effect of extra chromosomes in leukemic cells.

Stable or unstable? Origins and clonal evolution of HeH ALL

The non-random pattern of chromosome gains in HeH ALL has always puzzled researchers in the field. One of the continuously asked questions is whether aneuploidy in HeH cells is accompanied by chromosomal instability. As reviewed in the introduction of this thesis, CIN and aneuploidy are not synonyms, and aneuploid cells do not necessarily display CIN, as they can become stable shortly after arising **(35-37)**. Cytogenetically, the karyotypes of HeH cells appear to be quite stable, very few subclones are seen, and subclonal variation is of relatively low complexity when compared with other cancer types **(62)**. Studies using interphase FISH have reported that HeH samples display cell-to-cell copy number variation of its commonly gained chromosomes **(148-150)**. Interphase FISH is, however, a problematic method of inquiring chromosome copy number heterogeneity depending on the control samples one uses and considering the higher incidence of

technical artifacts of cytogenetics compared to molecular biology techniques. Furthermore, one needs to account for the quality of the slide and the number of times it has been re-hybridized to target different chromosomes, since this increases unspecific signals from the fluorescent probes. As regards the choice of control samples, it is important to have the same baseline number of the inquired chromosomes between target and controls. That is because trisomies and tetrasomies require a higher cut-off level for the probes than disomies to avoid false-negative signals. In ***Article II***, we attempted to circumvent such issues by comparing specific chromosomes from HeH cases that had the same copy number.

Sister chromatid cohesion defects are a known cause of CIN. However, we found it puzzling that there was such a great variation of incidence and severity of PCGs in HeH ALL samples from our cohort in ***Article II***. We then decided to perform interphase FISH to investigate whether cases with high levels of PCG would display increased chromosome copy number variation compared to cases with low or no PCGs. Indeed, we observed an increased variation in four out of the six investigated chromosomes associated with higher levels of cohesion defects, namely chromosomes 3, 6, 10 and 21. Moreover, knockdown of *RAD21* in an *ETV6::RUNX1*-positive cell line resulted in cohesion defects, increased formation of multipolar mitoses, and increased copy number variation of chromosome 21. Based on our findings as detailed above and previous knowledge on how chromosomes are gained in HeH ALL **(40, 62)**, cohesion defects are not likely to be the cause of aneuploidy in this subtype, although it might play a role in clonal evolution through increased chromosome heterogeneity. We continued our investigations in ***Article IV***, where we performed scWGS in nine HeH samples and one normal bone marrow to analyze whole chromosome copy number variation from 257 to 348 individual cells per case. Here, we showed that HeH ALL displays a generally stable genome, as 5 out of 9 cases had the same chromosome number in more than 99% of cells, and the remaining of the cases displayed 3 to 5 subclones with numerical changes each. Combining these findings with bulk WGS screens, no correlation was seen between mutations in genes related to genomic instability and increased number of subclones. The HeH case with the second highest level of heterogeneity also displayed 85% of cohesion defects, however the other cases with several subclones displayed mild levels – below 21% of PCGs. Unfortunately, the cohort used for scWGS included only one case with high levels of cohesion defects. Complementary single-cell analysis of such cases would help clarify if there is a stronger relationship between clonal heterogeneity and cohesion defects in HeH ALL, or if cohesion defects are an independent phenotype with a distinct underlying mechanism, perhaps a reflection of abnormal chromatin structuring rather than a cause of missegregation.

As mentioned before, many causes of aneuploidy have been described in the literature. Defects in the SAC, sister chromatid cohesion defects, merotelic

attachments and multipolar mitosis are known pathways leading to chromosome copy number alterations **(3, 4, 33, 34, 151)**. The routes that might lead to a stable hyperdiploid karyotype, however, are still a matter of discussion. Namely, four of these hypotheses are: (I) sequential chromosome gains in consecutive cell divisions due to sister chromatid nondisjunction; (II) an initial near-haploid cell that suffers duplication of its chromosomes; (III) an initial tetraploid cell with subsequent chromosome losses; or (IV) a punctuated event where a single abnormal mitosis leads to massive chromosome gains **(62, 152-154)**. Among the possible mechanisms suggested in literature, it is more likely that high hyperdiploid cells arise either by a tetraploid pathway (III) or by simultaneous gains (IV). Initial tetraploidization with subsequent chromosome losses would explain 2:2 allelic ratios in tetrasomies, and one third of the disomies resulted from loss of tetrasomy would be expected to display wUPID. Simultaneous chromosome gains in a single mitotic catastrophe would also explain the tetrasomy pattern in HeH and would result in no wUPID – which agrees with the majority of the cases **(40, 62, 154, 155)**. In ***Article IV***, we analyzed the chromosome pattern and copy number changes in subclones of 577 HeH ALL samples, unveiling strong positive selection for chromosomes X, 4, 6, 10, 14, 17, 18 and 21, weaker positive selection for chromosomes 5, 8, 11, 12 and 22, and neutral or negative selection for chromosomes Y, 1,2,3,7,9,13,15,16,19 and 20. We applied the acquired knowledge of positively selected chromosomes, as well as observed UPID rates (~2.5%), to an *in silico* model to simulate five possible scenarios that could give rise to high hyperdiploidy: (1) initial diploidy with sequential gains; (2) initial tetraploidy with sequential losses; (3) initial diploidy with a tripolar division; (4) initial tetraploidy with a tripolar division; (5) mitotic catastrophe due to complete loss of cohesion. Simulations were run starting with 50 000 cells, following through multiple generations, and stopped when the frequency of UPID reached 2.5%. Then, comparison between simulations and copy number data from the 577 HeH samples showed that a tripolar division from an initial diploid cell best satisfied the required patterns to resemble HeH ALL primary samples. The tripolar origin was further validated by assessing the age of trisomies, based on whether somatic SNVs were present in one (before and non-duplicated or after trisomy origin, B/ATRI) or two (before trisomy, BTRI) homologues. Here, we observed that chromosome gains from the positive selection group were newer as they more often displayed B/ATRI mutations, reinforcing that the bulk of aneuploidies arises at once, followed by low-level clonal evolution with acquirement of strongly selected chromosome gains.

The genetic features of HeH ALL are well-fitted for a tumorigenesis model known as punctuated evolution **(156)**. This model, which somewhat defies the perception of tumor evolution as a long and multi-stepped accumulation of genetic aberrations **(29-31)**, consists of a short burst of massive genetic alterations that occur very early in tumorigenesis and already defines the main identity of the malignancy. Therefore,

cancer types with localized phenomenon on single chromosomes or aneuploid cancers both fit into the punctuated evolution model in cases where there is clonal stability and lack of intermediate states showing gradual evolution **(156)**. Further exploring the idea of a single tripolar division from a diploid cell as an origin for high hyperdiploidy, it is noteworthy that tripolar mitoses are recurrent in tumors and can in fact continue to proliferate. A study on primary tumor samples showed that only a minority of tripolar mitoses result in three daughter cells, whereas a single multinucleated daughter cell occurred more often, and the majority of tripolar divisions resulted in two daughter cells – one binucleated and one mononucleated. Additionally, they observed that such binucleated daughter cells could undergo mitosis again with the formation of a single metaphase plate **(157)**. Once more, it is necessary to make a distinction between a stable aneuploid cell and an aneuploid cell featuring CIN that shows signs of progressive clonal heterogeneity. In ***Article II***, we performed immunofluorescence in RAD21-KD cells, and showed that 2.9% of analyzed mitoses were tripolar (3.89% incidence of multipolarity in total). A similar percentage of multipolar mitoses was observed in another study that analyzed *in vivo* expanded primary hyperdiploid samples, where it was noted that there was no significant difference in frequency of such mitotic aberrations between hyperdiploid and non-hyperdiploid cells **(147)**. This statement, however, considers observation of already established hyperdiploid leukemic cells, which have been shown by our group in ***Article IV*** and by previous cytogenetic studies **(62, 105, 155)** to be stable, with low proliferation rates and low heterogeneity. Rather than that, the diploid/tripolar model for the origin of HeH ALL implies that after the initial formation, the chromosomal gains in successive events will be minimal, and limited to chromosomes that are advantageous for the leukemic cell, hence agreeing with low frequency of tripolar divisions in diagnostic samples.

Ultimately, we have brought substantial evidence to support a punctuated evolution model for the origin of HeH ALL, where a tripolar division from a diploid cell generates the bulk pattern of chromosomal gains, followed by low-grade clonal evolution. Furthermore, we have extensively explored clonal heterogeneity in primary HeH ALL samples, combining cytogenetic and genomic techniques. As we tried to explain why we observe different levels of heterogeneity from case to case – although always towards lower complexity – it seems that the underlying reason has yet to be elucidated. It is likely that future studies involving single-cell genomics and transcriptomics could bring new insights regarding this, and perhaps prior screening for cohesion defects could help select a cohort with a higher chance of displaying copy number variation. Understanding what causes clonal heterogeneity in HeH ALL and the role of sister chromatid cohesion defects could pave the way for developing new targeted treatments, and this would help us understand whether this subtype requires further treatment stratification.

Concluding remarks

The four articles included in this thesis have extensively explored the effects of chromosomal gains in HeH ALL, their chromosome organization features and their origins and clonal development. Our studies revealed that HeH samples display massive transcriptional dysregulation associated with dosage effects and *trans* effects from the chromosomal gains, as well as wide-spread abnormal chromatin architecture likely caused by low levels of CTCF and the cohesin complex. We further observed interphase and mitotic chromosome disorganization in this subtype, where HeH ALL displays fewer and weakened TAD boundaries, impacting on transcriptional regulation, and both poor mitotic chromosome morphology and sister chromatid cohesion defects associated with low expression of cohesin- and condensin-related genes. Additionally, we described the 3D chromatin landscape of the main genetic subtypes of pediatric BCP ALL, showing three different clusters with distinct 3D genomic signatures, revealing that cases with aneuploidy have similar topological characteristics, and shedding light to the involvement of chromatin architecture in dysregulation of leukemia-related genes, such as *IKZF1*, *FLT3* and *PAX5*. We showed that high hyperdiploid cells are overall stable, harboring low-level clonal heterogeneity, but with relatively increased chromosome copy number variation in some cases, partially coinciding with the incidence of cohesion defects. Finally, our results point to a punctuated event as the origin of HeH ALL, with a diploid cell undergoing a tripolar division causing the bulk chromosomal gains, followed by low-grade clonal evolution.

As to future perspectives, we believe that the complex web connecting aneuploidy, aberrant chromatin organization at interphase/metaphase and levels of CTCF, cohesin and condensin require further investigation. More specifically, we would benefit from additional chromatin studies on a diverse cohort of aneuploid cancer-types to help us understand whether specific chromosome gains are responsible for the abnormal chromosome organization we have seen. It is also important to continue investigating CTCF, cohesin, condensin and other factors involved in chromatin folding in HeH ALL to have a clearer picture of their part in chromatin disorganization of this subtype, especially since different members of cohesin and condensin are correlating with our data in each study. Knockdown experiments of major players of chromatin organization in various aneuploid cell lines, followed by Micro-C, RNA-seq and cytogenetic studies, could help us to understand the role of these factors and whether we can reproduce the scenery seen in primary HeH samples. Prior screening of cohesion defects in primary patient samples followed by Micro-C and scWGS would also help clarify how strong is the association between sister chromatid cohesion defects and increased clonal heterogeneity in HeH ALL. To the best of our knowledge, chromosome territories have not been investigated in aneuploid BCP ALL samples yet, and exploring this level of genome

folding with matched 3D-FISH, Hi-C or Micro-C and transcriptomic data could bring more insight into the 3D genomic landscape of these leukemias. Likewise, applying methylome analysis to BCP ALL primary samples, such as assay for transposase accessible chromatin sequencing (ATAC-seq) for detecting open chromatin and whole genome bisulfite sequencing (WGBS) for detecting DNA methylation could bring additional in-depth information of epigenetic factors in this type of leukemia. Ultimately, this thesis took us one step further into understanding the molecular pathogenesis of the largest subtype of pediatric BCP ALL, paving the way for future investigations of the multiple layers of high hyperdiploidy in leukemia.

Popular scientific summary

Leukemia is a blood cancer marked by proliferation of white blood cells, being classified into acute or chronic depending on how fast it develops and further divided based on the type of blood cells that are spreading. Thus, B-cell precursor (BCP) acute lymphoblastic leukemia (ALL) is a cancer of rapid progression and involves lymphocytes known as B-cells. Normally, B-cells undergo many changes to reach maturation and be ready to produce antibodies for our immune system, but in BCP ALL the B-cells are stuck at the beginning of their development, and they spread and renew themselves without ever acquiring their normal functions. BCP ALL is the most common cancer type in children and has an overall survival rate of more than 90%. Nevertheless, it is very important to increase our knowledge of this type of leukemia to improve the outcome of those who relapse, and to reduce the overtreatment, when possible, since the intensity and toxicity of leukemia treatments can bring long-term collateral effects, both physically and mentally.

In our studies, we focus on the largest group of BCP ALL that accounts for 30% of all pediatric cases, the high hyperdiploid (HeH) subtype. While human cells normally have 46 chromosomes that contain all of our genetic material, or DNA, the HeH cells have between 51 and 67 chromosomes. It is still not completely understood how the leukemic cells end up gaining all this extra genetic material, or what the implications of having much more DNA than they should are. Therefore, the four studies included in this thesis explore the origins of the extra chromosomes, their consequences for regulation of genes in our DNA and how the chromosomes are organizing their genetic material.

In ***Articles I-III*** we found that the genetic material with extra copies in HeH ALL have higher expression than normal, and the extra chromosomes also affect the regulation of other parts of the DNA. We also showed that certain proteins that are responsible for how our DNA is compacted and organized have lower levels than usual in this leukemic subtype, namely CTCF and members of the cohesin protein complex. As a result, chromosomes from HeH ALL are very disorganized and fail to keep parts of the DNA separated from each other. This lack of proper separation also affects gene regulation, including genes that are important for leukemia

development, such as *FLT3* and *IKZF1*. We also discovered that the chromosomes in HeH cells are defective during cell division, or mitosis, being shorter than normal and having fuzzy-looking shape even compared to other types of leukemia, and often separating the two halves of the chromosomes before the right time (referred to as cohesion defects). The chromosome defects during mitosis also coincide with low levels of the cohesin complex, as well as the condensin complex – another group of proteins that organize mitotic chromosomes.

In ***Articles II*** and ***IV***, we investigate how the extra chromosomes are gained and if the HeH cells usually keep the exact same chromosomes from cell to cell, or if there is a big variation in chromosome copy number. Many cancer types have unstable genetic material coinciding with gains of extra chromosomes, a phenomenon known as chromosomal instability, and this high rate of variation from cell to cell boosts cancer evolution and hinders treatment success. We observed that HeH cells with more severe cohesion defects in mitosis also had a higher variation in copy number of certain chromosomes from cell to cell. However, most HeH ALL cases do not have very severe cohesion defects, and by analyzing the DNA of each individual leukemic cell in nine patient samples we found that the chromosome number and content had little variation per cell. We showed that most HeH ALL cases have few subclones – leukemic cell populations with different DNA aberrations or chromosome number, and that certain chromosomes are gained more often in newer subclones than others, probably because they bring advantages to the cancer cells. Moreover, we used computer simulations to try and find what type of errors during mitosis could cause normal cells to acquire the same extra chromosomes we see in HeH ALL. By comparing the simulation results with patient data, we concluded that the great majority of HeH cells originate from a normal cell that tries to divide into three daughter cells instead of two – a tripolar mitosis.

In sum, this thesis contributes to a better understanding of the biology and genetics of one of the most common cancer types in children. We answered many questions about how HeH ALL arises and develops, and how their DNA organization and chromosome structures affect gene regulation. In the future, we would like to continue exploring how chromosomes are organized in this subtype, why so many chromosome organizing factors are at low levels and if there are other main players dysregulating the DNA in HeH ALL.

Populärvetenskaplig sammanfattning

Leukemi är en blodcancer som kännetecknas av ökad tillväxt av vita blodkroppar. Den klassificeras som antingen akut eller kronisk beroende på hur snabbt den utvecklas och kan vidare delas upp baserat på vilken typ av blodceller som sprider sig. Således är B-cellprekursor (BCP) akut lymfatisk leukemi (ALL) en cancer med snabb progression och innefattar lymfocyter som kallas B-celler. Normalt genomgår B-celler många förändringar för att mogna och vara redo att producera antikroppar för vårt immunsystem, men i BCP ALL är B-cellerna fast i början av sin utveckling och de sprider sig och förnyar sig själva utan att någonsin få sina normala funktioner. BCP ALL är den vanligaste cancerformen hos barn och har en överlevnad på över 90%. Trots detta är det mycket viktigt att öka vår kunskap om denna typ av leukemi för att förbättra utfallet för dem som får återfall, och för att minska överbehandlingen när det är möjligt, eftersom intensiteten och toxiciteten hos leukemibehandlingar kan ge långsiktiga biverkningar, både fysiskt och psykiskt.

I våra studier fokuserar vi på den största gruppen av BCP ALL som står för 30% av alla pediatriska fall, den hyperdiploida (HeH) subtypen. Medan mänskliga celler normalt har 46 kromosomer som innehåller allt vårt genetiska material, eller DNA, har HeH-cellerna mellan 51 och 67 kromosomer. Det är fortfarande inte helt klart hur leukemi cellerna fått allt detta extra genetiska material, eller vilka konsekvenserna är av att de har mycket mer DNA än de borde. Därför utforskar vi i de fyra studier som ingår i denna avhandling ursprunget till de extra kromosomerna, deras konsekvenser för regleringen av gener i vårt DNA och hur kromosomerna organiserar sitt genetiska material.

I ***Artikel I-III*** fann vi att det genetiska materialet med extra kopior i HeH ALL uttrycks högre än normalt, och de extra kromosomerna påverkar också regleringen av andra delar av DNA:t. Vi visade också att vissa proteiner som är ansvariga för hur vårt DNA är organiserat, nämligen CTCF och medlemmar av cohesin-proteincomplexet, har lägre nivåer än vanligt i denna leukemiska subtyp. Som ett resultat är kromosomerna från HeH ALL mycket oorganiserade och misslyckas med att hålla delar av DNA:t separerade från varandra. Denna brist på korrekt separation påverkar också genregleringen, inklusive gener som är viktiga för

leukemiutveckling, såsom *FLT3* och *IKZF1*. Vi upptäckte också att kromosomerna i HeH-celler är defekta under celldelning, eller mitos, eftersom de är kortare än normalt och ett mer diffust utseende jämfört med andra typer av leukemi, och ofta separerar de två halvorna av kromosomerna förtid (kallat sammanhållningsdefekter). Kromosomdefekter under mitosen är också associerade med låga nivåer av cohesin-komplexet, liksom kondensin-komplexet - en annan grupp proteiner som organiserar mitotiska kromosomer.

I ***Artikel II*** och ***IV*** undersöker vi hur de extra kromosomerna tillkommer och om HeH-celler vanligtvis behåller exakt samma kromosomer från cell till cell, eller om det finns variation i kopietalet av kromosomerna. Många cancerformer har instabilt genetiskt material tillsammans tillskott av extra kromosomer, en fenomen som kallas kromosominstabilitet, och denna stora variation från cell till cell ökar cancerevolutionen och hindrar behandlingsframgång. Vi observerade att HeH-celler med svårare sammanhållningsdefekter under mitosen också hade en större variation i kopietalet av vissa kromosomer mellan celler. De flesta HeH ALL-fall har emellertid inte mycket allvarliga sammanhållningsdefekter, och genom att analysera DNA från varje enskild leukemi cell i nio patientprover fann vi att kromosomantalet och innehållet uppvisade liten variation mellan celler. Vi visade att de flesta HeH ALL-fall har få subkloner - leukemiska cellpopulationer med olika DNA avvikelser eller kromosomantal, och att vissa kromosomer oftare tillkommer i nya subkloner än andra, förmodligen eftersom de ger fördelar till cancercellerna. Dessutom använde vi datorsimuleringar för att försöka hitta vilken typ av fel under mitos som kunde få normala celler att förvärva samma extra kromosomer som vi ser i HeH ALL. Genom att jämföra simuleringsresultaten med patientdata drog vi slutsatsen att den stora majoriteten av HeH-celler härstammar från en normal cell som försöker dela sig i tre dotterceller istället för två - en tripolär mitos.

Sammanfattningsvis bidrar denna avhandling till en bättre förståelse för biologin och genetiken hos en av de vanligaste cancerformerna hos barn. Vi besvarade många frågor om hur HeH ALL uppstår och utvecklas, och hur dess DNA-organisation och kromosomstruktur påverkar genregleringen. I framtiden skulle vi vilja fortsätta att utforska hur kromosomer organiseras i denna subtyp, varför så många kromosomorganiseringsfaktorer har låga nivåer och om det finns andra nyckelaktörer som dysreglerar DNA:t i HeH ALL.

Resumo de divulgação científica

A leucemia é um câncer caracterizado pela proliferação de células brancas do sangue, sendo classificada como aguda ou crônica dependendo da rapidez com que se desenvolve e ainda dividida com base no tipo de células sanguíneas que estão se espalhando. Assim, a leucemia linfoblástica aguda (LLA) tipo B é um câncer de rápida progressão e envolve linfócitos conhecidos como células B. Normalmente, as células B passam por muitas mudanças para atingir a maturação e estar prontas para produzir anticorpos para o nosso sistema imunológico, mas na LLA tipo B, as células B ficam presas no início de seu desenvolvimento, e se espalham e se renovam sem adquirir nunca suas funções normais. A LLA tipo B é o tipo mais comum de câncer em crianças e tem uma taxa de sobrevida global de mais de 90%. No entanto, é muito importante aumentar nosso conhecimento sobre esse tipo de leucemia para melhorar o resultado daqueles onde o câncer volta e para reduzir o supertratamento quando possível, já que a intensidade e toxicidade dos tratamentos de leucemia podem trazer efeitos colaterais de longo prazo, tanto físicos quanto mentais.

Em nossos estudos, focamos no maior grupo de LLA tipo B que representa 30% de todos os casos pediátricos, o subtipo hiperdiploide (HeH). Enquanto as células humanas normalmente têm 46 cromossomos que contêm todo o nosso material genético, ou DNA, as células HeH têm entre 51 e 67 cromossomos. Ainda não se compreende completamente como as células leucêmicas acabam adquirindo todo esse material genético extra, ou quais são as implicações de ter muito mais DNA do que deveriam. Portanto, os quatro estudos incluídos nesta tese exploram as origens dos cromossomos extras, suas consequências para a regulação dos genes em nosso DNA e como os cromossomos estão organizando seu material genético. Nos ***Artigos I-III***, descobrimos que o material genético com cópias extras na LLA HeH tem uma expressão mais alta do que o normal, e os cromossomos extras também afetam a regulação de outras partes do DNA. Também mostramos que certas proteínas que são responsáveis por como o nosso DNA é organizado, CTCF e membros do complexo de proteínas coesina, têm níveis mais baixos do que o normal nesse subtipo de leucemia. Como resultado, os cromossomos da LLA HeH são muito

desorganizados e não conseguem manter partes do DNA separadas umas das outras. Essa falta de separação adequada também afeta a regulação de genes, incluindo genes que são importantes para o desenvolvimento da leucemia, como *FLT3* e *IKZF1*. Descobrimos também que os cromossomos nas células HeH são defeituosos durante a divisão celular, ou mitose, sendo mais curtos do que o normal e tendo uma aparência borrada mesmo em comparação com outros tipos de leucemia, e frequentemente separando as duas metades dos cromossomos antes do momento certo (referido como defeitos de coesão). Os defeitos cromossômicos durante a mitose também coincidem com níveis baixos do complexo de coesina, bem como do complexo condensador - outro grupo de proteínas que organizam os cromossomos mitóticos. Nos ***Artigos II e IV***, investigamos como os cromossomos extras são adquiridos e se as células HeH geralmente mantêm exatamente os mesmos cromossomos de célula para célula, ou se há uma grande variação no número de cópias cromossômicas. Muitos tipos de câncer têm material genético instável coincidindo com ganhos de cromossomos extras, um fenômeno conhecido como instabilidade cromossômica, e essa alta taxa de variação de célula para célula impulsiona a evolução do câncer e dificulta o sucesso do tratamento. Observamos que as células HeH com defeitos de coesão mais graves também tinham uma maior variação no número de cópias de certos cromossomos de célula para célula. No entanto, a maioria dos casos de LLA HeH não tem defeitos de coesão muito graves, e ao analisar o DNA de cada célula leucêmica individualmente em nove amostras de pacientes, descobrimos que o número e o conteúdo dos cromossomos tinham pouca variação por célula. Mostramos que a maioria dos casos de LLA HeH tem poucos subclones - populações de células leucêmicas com aberrações de DNA ou número de cromossomos diferentes, e que certos cromossomos são adquiridos com mais frequência em subclones mais recentes do que outros, provavelmente porque trazem vantagens para as células cancerosas. Além disso, usamos simulações computacionais para tentar encontrar que tipo de erros durante a mitose poderiam fazer com que células normais adquirissem os mesmos cromossomos extras que vemos na LLA HeH. Ao comparar os resultados da simulação com os dados dos pacientes, concluímos que a grande maioria das células HeH se originam de uma célula normal que tenta se dividir em três células filhas em vez de duas - uma mitose tripolar.

Concluindo, esta tese contribui para uma melhor compreensão da biologia e genética de um dos tipos mais comuns de câncer em crianças. Respondemos a muitas perguntas sobre como a LLA HeH surge e se desenvolve, e como a organização de seu DNA e estruturas cromossômicas afetam a regulação de genes. No futuro,

gostaríamos de continuar explorando como os cromossomos estão sendo organizados nesse subtipo, por que tantos fatores de organização cromossômica estão em baixos níveis e se existem outros principais agentes que desregulam o DNA na LLA HeH.

Acknowledgements

When I moved to Sweden in 2016, I had no idea of how difficult this journey would be, but I am grateful for every single moment of it. I've met many incredible people during my master's program and PhD studies, and I am so thankful for everyone who's been in my life ever since and for those who've been involved in the coming of this thesis.

First of all, I would like to thank **Kajsa** for being the best supervisor I could have wished for. I still remember how excited I was when you agreed to supervise my master thesis. Working with leukemia and chromosomes was all I wanted, and I found a research group where I could do it. Thank you for teaching me so much, for being honest and kind, for believing in me and for letting me be creative. Thank you for always having the time to listen to me and for telling me not to worry whenever I was stressed. I truly admire your way of thinking and how you conduct research. I also like that armchair in the corner of your office very much, it is very comfortable. Thanks for everything!

I would also like to thank my co-supervisor **Anders** for his expertise in the clinical aspects of my thesis. A special thank you to all my **co-authors and collaborators** as well, I am grateful for the hard work and expertise you shared with me. This thesis would not have been completed otherwise. On the same note, I'm grateful to **Nils** and **Felix** for sharing their immense knowledge in cytogenetics with me when I needed advice, I feel very privileged.

Ellie, you are so welcoming, skilled, and humble. You taught me so much at the lab, you were always there to help me, and you showed me how it feels to be part of a team. You are also a little odd, like me, which makes me feel pretty comfortable. Thank you for all the long lab days talking about ~~The Walking Dead and American Horror Story~~ science, and for sharing your expertise in ~~motherhood, life in Sweden and sarcasm~~ science again. Also, thanks to you I now understand that Gilson pipettes are the best.

Minjun, one thing is to be patient, another thing is having to deal with me during my final PhD years. Working with you is a great experience, you are kind and fun, and so knowledgeable in ~~witchcraft~~ Bioinformatics. Thank you for all the brainstorming in group meetings, all the troubleshooting, and for being so patient with me when explaining our data. You are an amazing ~~wizard~~ bioinformatician.

Efe, you are so smart, and fun, and chill. Plus, you like heavy metal and video games, so it's impossible not to like you. I'm glad we are part of the same team, and thank you so much for all the science, gaming, and music talks. **Gladys**, thanks for your expertise and critical thinking, and thank you **Charlotte** for being so kind and a great companion in conferences. I feel lucky to have met you all.

Andrea, thank you for always finding the time to help me, for always knowing where things are, and for being so fun to work with. I am very grateful to have someone to go to when I need to discuss cytogenetics, and who would be better than the very own person who taught me how to FISH? And of course, I'm thankful to all of our technical staff, **Linda M**, **Jenny N**, **Marianne**, **Helena S**, **Carro** and **Tina** for answering my silly questions and for being kind when I make mistakes.

A special shout out to my former and current office buddies **Ram**, **Ludvig**, **Valeriia**, **Josephine**, **Bahar**, **Somadri** and **Mattias**. Thanks for the company and for all the ~~random jokes and philosophical discussions during worktime~~ nice conversations.

And of course, I cannot forget the fellow PhD students, **Hanna**, **Louise**, **Natalie**, **Saskia**, **Valeria** and **Vendela**, who are either going through the same things I am facing right now, or soon will be (my prayers and thoughts), and all of the amazing students, post-docs and PIs that make the division of Clinical Genetics a fun environment and a great place to do research. And speaking of fun environment, thanks for the protein people, especially **Hannah** and **Sibel** for the hangouts and good memories.

A very special thanks to my partners in crime **Karim**, **Katrin** and former-C13 **Lexi** for being such great weirdos. You've seen me in my best and my worst, and for some reason you still talk to me, which means you like me, I guess? Thanks for ~~being there during my breakdowns~~ all the support, ~~being part of my weirdest memories~~ hangouts, ~~sharing memes~~ great conversations, and all the play dates with dogs, cats, and now there's a toddler too. You all have a special place in my heart.

Oh yes, the husband! Which also happens to be my favorite bioinformatics and cytogenetics consultant and my number one gaming buddy. **Arthur**, *you are the wind beneath my wings*. You've been by my side since our sophomore year field trip to Cardoso Island in 2010. We became biologists and cytogeneticists together, we were even part of a metal band and a hard rock band together (good times...). Thank you for following me to Sweden, marrying me, and for having **Giovanni** with me (**Gio**, you still don't know how to read, but we love you so much!). You are a great fellow PhD student, a great friend and husband, and a great father. Plus, you always buy me a croissant for fika, which is great too.

Mamma e papi, vocês sempre me ensinaram que família é a coisa mais importante das nossas vidas, e que nós sempre podemos contar uns com os outros. Eu não tenho palavras o suficiente para agradecer tudo o que vocês já fizeram e ainda fazem por

mim. Obrigada por me ensinar a ser eu mesma, a trabalhar duro e a não desistir dos meus objetivos. Vocês sempre acreditaram em mim, e sem o apoio de vocês eu não teria conseguido chegar aqui. **Leo e Lucas**, por culpa de vocês eu aprendi a cuidar de recém-nascidos com 11 anos de idade, muito obrigada. Por outro lado, eu vi vocês crescerem e se tornarem pessoas incríveis, inteligentes, e tão companheiras. Eu tenho muito orgulho de vocês, e mal posso esperar pelo dia que serei eu lendo suas dissertações (sim, eu vou fazer quiz para ver se vocês leram a minha). Amo muito todos vocês!

And now, as we love games at C13, please enjoy this nerd-themed mini quiz.

1. Your fellow co-worker is about to defend their PhD thesis. Which of the following potions would you give them?
   A) Pepperup potion
   B) Edurus potion
   C) Felix felicis
2. A wild Lucario appears right outside Stamstället. Which pokemon would you NOT use to defeat him?
   A) The Bug/Poison type Beedrill
   B) The Fire/Psychic type Armarouge
   C) The Dragon/Ground type Garchomp
3. It is your Biobank week and 3 samples come at the same time. You wish you had the powers of:
   A) Star-Lord
   B) Quicksilver
   C) Enchantress
4. Who could easily convince the BMC reception to give you access to every room in the building?
   A) Obi-Wan Kenobi
   B) Senator Amidala
   C) C-3PO

Answers: 1C; 2A; 3B; 4A

References

1. Wolpert L. Evolution of the cell theory. Philos Trans R Soc Lond B *Biol Sci*. 1995;**349**(1329):227-33.
2. Wang Z. Cell cycle progression and synchronization: An Overview. *Methods Mol Biol*. 2022;**2579**:3-23.
3. Rieder CL. Mitosis in vertebrates: the G2/M and M/A transitions and their associated checkpoints. *Chromosome Res*. 2011;**19**(3):291-306.
4. Moreno-Andres D, Holl K, Antonin W. The second half of mitosis and its implications in cancer biology. *Semin Cancer Biol*. 2023;**88**:1-17.
5. Merkenschlager M, Nora EP. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu Rev Genomics Hum Genet*. 2016;**17**:17-43.
6. Yuen KC, Gerton JL. Taking cohesin and condensin in context. *PLoS Genet*. 2018;**14**(1):e1007118.
7. Losada A. Cohesin in cancer: chromosome segregation and beyond. *Nat Rev Cancer*. 2014;**14**(6):389-93.
8. Nasmyth K. Cohesin: a catenase with separate entry and exit gates? *Nat Cell Biol*. 2011;**13**(10):1170-7.
9. Hauf S, Waizenegger IC, Peters JM. Cohesin cleavage by separase required for anaphase and cytokinesis in human cells. *Science*. 2001;**293**(5533):1320-3.
10. Batty P, Gerlich DW. Mitotic chromosome mechanics: how cells segregate their genome. *Trends Cell Biol*. 2019;**29**(9):717-26.
11. Ghosh RP, Meyer BJ. Spatial organization of chromatin: emergence of chromatin structure during development. *Annu Rev Cell Dev Biol*. 2021;**37**:199-232.
12. Hirota T, Gerlich D, Koch B, Ellenberg J, Peters JM. Distinct functions of condensin I and II in mitotic chromosome assembly. *J Cell Sci*. 2004;**117**(26):6435-45.
13. Zhiteneva A, Bonfiglio JJ, Makarov A, Colby T, Vagnarelli P, Schirmer EC, et al. Mitotic post-translational modifications of histones promote chromatin compaction in vitro. *Open Biol*. 2017;**7**(9):170076.
14. Cuylen S, Blaukopf C, Politi AZ, Muller-Reichert T, Neumann B, Poser I, et al. Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. *Nature*. 2016;**535**(7611):308-12.
15. Samwer M, Schneider MWG, Hoefler R, Schmalhorst PS, Jude JG, Zuber J, et al. DNA Cross-bridging shapes a single nucleus from a set of mitotic chromosomes. *Cell*. 2017;**170**(5):956-72.

16. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. *Science*. 1974;**184**(4139):868-71.

17. Kosak ST, Groudine M. Form follows function: the genomic organization of cellular differentiation. *Genes Dev*. 2004;**18**(12):1371-84.

18. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;**326**(5950):289-93.

19. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*. 2010;**467**(7314):430-5.

20. Eskeland R, Leeb M, Grimes GR, Kress C, Boyle S, Sproul D, et al. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol Cell*. 2010;**38**(3):452-64.

21. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;**485**(7398):376-80.

22. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-Mediated Human 3D Genome architecture reveals chromatin topology for transcription. *Cell*. 2015;**163**(7):1611-27.

23. Furlong EEM, Levine M. Developmental enhancers and chromosome topology. *Science*. 2018;**361**(6409):1341-5.

24. Robson MI, Ringel AR, Mundlos S. Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. *Mol Cell*. 2019;**74**(6):1110-22.

25. Hsieh TS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, et al. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol Cell*. 2020;**78**(3):539-53.

26. Hsieh TS, Cattoglio C, Slobodyanyuk E, Hansen AS, Darzacq X, Tjian R. Enhancer-promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat Genet*. 2022;**54**(12):1919-32.

27. Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*. 2015;**112**(47):6456-65.

28. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop extrusion. *Cell Rep*. 2016;**15**(9):2038-49.

29. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;**100**(1):57-70.

30. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;**144**(5):646-74.

31. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov*. 2022;**12**(1):31-46.

32. Sinha D, Duijf PHG, Khanna KK. Mitotic slippage: an old tale with a new twist. *Cell Cycle*. 2019;**18**(1):7-15.

33. Jaiswal S, Singh P. Centrosome dysfunction in human diseases. *Semin Cell Dev Biol*. 2021;**110**:113-22.

34. Barber TD, McManus K, Yuen KW, Reis M, Parmigiani G, Shen D, et al. Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proc Natl Acad Sci U S A*. 2008;**105**(9):3443-8.

35. Lukow DA, Sausville EL, Suri P, Chunduri NK, Wieland A, Leu J, et al. Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. *Dev Cell*. 2021;**56**(17):2427-39.

36. Lakhani AA, Thompson SL, Sheltzer JM. Aneuploidy in human cancer: new tools and perspectives. *Trends Genet*. 2023;**39**(12):968-80.

37. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. *Nat Rev Genet*. 2020;**21**(1):44-62.

38. Duijf PH, Schultz N, Benezra R. Cancer cells preferentially lose small chromosomes. *Int J Cancer*. 2013;**132**(10):2316-26.

39. Boveri T. Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J Cell Sci*. 2008;**121**(1):1-84.

40. Paulsson K, Lilljebjorn H, Biloglav A, Olsson L, Rissler M, Castor A, et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Genet*. 2015;**47**(6):672-6.

41. Thompson SL, Compton DA. Chromosomes and cancer cells. *Chromosome Res*. 2011;**19**(3):433-44.

42. Dahiya R, Hu Q, Ly P. Mechanistic origins of diverse genome rearrangements in cancer. *Semin Cell Dev Biol*. 2022;**123**:100-9.

43. Abramowicz A, Gos M. Correction to: Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet*. 2019;**60**(2):231.

44. Bailey SF, Alonso Morales LA, Kassen R. Effects of synonymous mutations beyond codon bias: the evidence for adaptive synonymous substitutions from microbial evolution experiments. *Genome Biol Evol*. 2021;13(9):141.

45. Gerbes AL, Caselmann WH. Point mutations of the P53 gene, human hepatocellular carcinoma and aflatoxins. *J Hepatol*. 1993;**19**(2):312-5.

46. Kontomanolis EN, Koutras A, Syllaios A, Schizas D, Mastoraki A, Garmpis N, et al. Role of oncogenes and tumor-suppressor genes in carcinogenesis: a review. *Anticancer Res*. 2020;**40**(11):6009-15.

47. Jacobson LO, Marks EK, et al. The role of the spleen in radiation injury. *Proc Soc Exp Biol Med*. 1949;**70**(4):740-2.

48. Morrison SJ, Uchida N, Weissman IL. The biology of hematopoietic stem cells. *Annu Rev Cell Dev Biol*. 1995;**11**:35-71.

49. Till JE, Mc CE. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res*. 1961;**14**:213-22.

50. Wu AM, Till JE, Siminovitch L, McCulloch EA. A cytological study of the capacity for differentiation of normal hemopoietic colony-forming cells. *J Cell Physiol*. 1967;**69**(2):177-84.

51. Siminovitch L, McCulloch EA, Till JE. The distribution of colony-forming cells among spleen colonies. *J Cell Comp Physiol*. 1963;**62**:327-36.

52. Seita J, Weissman IL. Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip Rev Syst Biol Med*. 2010;**2**(6):640-53.

53. Orkin SH, Zon LI. Snapshot: hematopoiesis. *Cell*. 2008;**132**(4):712.

54. Bao EL, Cheng AN, Sankaran VG. The genetics of human hematopoiesis and its disruption in disease. *EMBO Mol Med*. 2019;**11**(8):10316.

55. Luc S, Buza-Vidas N, Jacobsen SE. Biological and molecular evidence for existence of lymphoid-primed multipotent progenitors. *Ann N Y Acad Sci*. 2007;**1106**:89-94.

56. Akashi K, Reya T, Dalma-Weiszhausz D, Weissman IL. Lymphoid precursors. *Curr Opin Immunol*. 2000;**12**(2):144-50.

57. Zhang Q, Iida R, Yokota T, Kincade PW. Early events in lymphopoiesis: an update. *Curr Opin Hematol*. 2013;**20**(4):265-72.

58. Ramirez J, Lukin K, Hagman J. From hematopoietic progenitors to B cells: mechanisms of lineage restriction and commitment. *Curr Opin Immunol*. 2010;**22**(2):177-84.

59. Rothenberg EV. Transcriptional drivers of the T-cell lineage program. *Curr Opin Immunol*. 2012;**24**(2):132-8.

60. Corces-Zimmerman MR, Majeti R. Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia*. 2014;**28**(12):2276-82.

61. Irons RD, Stillman WS. The process of leukemogenesis. *Environ Health Perspect*. 1996;**104** Suppl 6(Suppl 6):1239-46.

62. Paulsson K, Johansson B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2009;**48**(8):637-60.

63. Rowley JD, Le Beau MM. Cytogenetic and molecular analysis of therapy-related leukemia. *Ann N Y Acad Sci*. 1989;**567**:130-40.

64. Bedi A, Zehnbauer BA, Collector MI, Barber JP, Zicha MS, Sharkis SJ, et al. BCR-ABL gene rearrangement and expression of primitive hematopoietic progenitors in chronic myeloid leukemia. *Blood*. 1993;**81**(11):2898-902.

65. Jonas D, Lubbert M, Kawasaki ES, Henke M, Bross KJ, Mertelsmann R, et al. Clonal analysis of bcr-abl rearrangement in T lymphocytes from patients with chronic myelogenous leukemia. *Blood*. 1992;**79**(4):1017-23.

66. Hong D, Gupta R, Ancliff P, Atzberger A, Brown J, Soneji S, et al. Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science*. 2008;**319**(5861):336-9.

67. Knuutila S, Teerenhovi L, Larramendy ML, Elonen E, Franssila KO, Nylund SJ, et al. Cell lineage involvement of recurrent chromosomal abnormalities in hematologic neoplasms. *Genes Chromosomes Cancer*. 1994;**10**(2):95-102.

68. Nitta M, Kato Y, Strife A, Wachter M, Fried J, Perez A, et al. Incidence of involvement of the B and T lymphocyte lineages in chronic myelogenous leukemia. *Blood*. 1985;**66**(5):1053-61.

69. Haferlach T, Winkemann M, Nickenig C, Meeder M, Ramm-Petersen L, Schoch R, et al. Which compartments are involved in Philadelphia-chromosome positive chronic myeloid leukaemia? An answer at the single cell level by combining May-Grunwald-Giemsa staining and fluorescence in situ hybridization techniques. *Br J Haematol*. 1997;**97**(1):99-106.

70. Fialkow PJ, Jacobson RJ, Papayannopoulou T. Chronic myelocytic leukemia: clonal origin in a stem cell common to the granulocyte, erythrocyte, platelet and monocyte/macrophage. *Am J Med*. 1977;**63**(1):125-30.

71. Juneja HS, Weiner R. Presence of the Philadelphia chromosome (Ph1) in pokeweed mitogen stimulated lymphocytes during chronic phase of chronic myelocytic leukemia (CML). *Cancer Genet Cytogenet*. 1981;**4**(1):39-44.

72. Kikushige Y, Ishikawa F, Miyamoto T, Shima T, Urata S, Yoshimoto G, et al. Self-renewing hematopoietic stem cell is the primary target in pathogenesis of human chronic lymphocytic leukemia. *Cancer Cell*. 2011;**20**(2):246-59.

73. Brown G, Ceredig R, Tsapogas P. The making of hematopoiesis: developmental ancestry and environmental nurture. *Int J Mol Sci*. 2018;**19**(7):2122.

74. Greaves M. In utero origins of childhood leukaemia. *Early Hum Dev*. 2005;**81**(1):123-9.

75. Hutter JJ. Childhood leukemia. *Pediatr Rev*. 2010;**31**(6):234-41.

76. Yagi T, Hibi S, Tabata Y, Kuriyama K, Teramura T, Hashida T, et al. Detection of clonotypic IGH and TCR rearrangements in the neonatal blood spots of infants and children with B-cell precursor acute lymphoblastic leukemia. *Blood*. 2000;**96**(1):264-8.

77. Gale KB, Ford AM, Repp R, Borkhardt A, Keller C, Eden OB, et al. Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. *Proc Natl Acad Sci U S A*. 1997;**94**(25):13950-4.

78. Greaves M. Pre-natal origins of childhood leukemia. *Rev Clin Exp Hematol*. 2003;**7**(3):233-45.

79. Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, et al. Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet*. 1999;**354**(9189):1499-503.

80. Wiemels JL, Xiao Z, Buffler PA, Maia AT, Ma X, Dicks BM, et al. In utero origin of t(8;21) AML1-ETO translocations in childhood acute myeloid leukemia. *Blood*. 2002;**99**(10):3801-5.

81. Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci U S A*. 2002;**99**(12):8242-7.

82. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. *Lancet*. 2013;**381**(9881):1943-55.

83. Duffy C, Graetz DE, Lopez AMZ, Carrillo AK, Job G, Chen Y, et al. Retrospective analysis of outcomes for pediatric acute lymphoblastic leukemia in South American centers. *Front Oncol*. 2023;**13**:1254233.

84. Inaba H, Mullighan CG. Pediatric acute lymphoblastic leukemia. *Haematologica*. 2020;**105**(11):2524-39.

85. A treatment protocol for participants 0-45 years with acute lymphoblastic leukaemia. https://classic.clinicaltrials.gov/show/NCT03911128.

86. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. *Lancet*. 2008;**371**(9617):1030-43.

87. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*. 2008;**322**(5906):1377-80.

88. Li J, Dai Y, Wu L, Zhang M, Ouyang W, Huang J, et al. Emerging molecular subtypes and therapeutic targets in B-cell precursor acute lymphoblastic leukemia. *Front Med*. 2021;**15**(3):347-71.

89. Mullighan CG. The molecular genetic makeup of acute lymphoblastic leukemia. *Hematology Am Soc Hematol Educ Program*. 2012;**2012**:389-96.

90. Harrison CJ, Foroni L. Cytogenetics and molecular genetics of acute lymphoblastic leukemia. *Rev Clin Exp Hematol*. 2002;**6**(2):91-113.

91. Malinowska-Ozdowy K, Frech C, Schonegger A, Eckert C, Cazzaniga G, Stanulla M, et al. KRAS and CREBBP mutations: a relapse-linked malicious liaison in childhood high hyperdiploid acute lymphoblastic leukemia. *Leukemia*. 2015;29(8):1656-67.

92. Ribeiro RC, Abromowitch M, Raimondi SC, Murphy SB, Behm F, Williams DL. Clinical and biologic hallmarks of the Philadelphia chromosome in childhood acute lymphoblastic leukemia. *Blood*. 1987;**70**(4):948-53.

93. Lilljebjorn H, Henningsson R, Hyrenius-Wittsten A, Olsson L, Orsmark-Pietras C, von Palffy S, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat Commun*. 2016;**7**:11790.

94. Sun C, Chang L, Zhu X. Pathogenesis of ETV6/RUNX1-positive childhood acute lymphoblastic leukemia and mechanisms underlying its relapse. *Oncotarget*. 2017;**8**(21):35445-59.

95. Moorman AV, Robinson H, Schwab C, Richards SM, Hancock J, Mitchell CD, et al. Risk-directed treatment intensification significantly reduces the risk of relapse among children and adolescents with acute lymphoblastic leukemia and intrachromosomal amplification of chromosome 21: a comparison of the MRC ALL97/99 and UKALL2003 trials. *J Clin Oncol*. 2013;**31**(27):3389-96.

96. Heerema NA, Carroll AJ, Devidas M, Loh ML, Borowitz MJ, Gastier-Foster JM, et al. Intrachromosomal amplification of chromosome 21 is associated with inferior outcomes in children with acute lymphoblastic leukemia treated in contemporary standard-risk children's oncology group studies: a report from the children's oncology group. *J Clin Oncol*. 2013;**31**(27):3397-402.

97. Wen J, Zhou M, Shen Y, Long Y, Guo Y, Song L, et al. Poor treatment responses were related to poor outcomes in pediatric B cell acute lymphoblastic leukemia with KMT2A rearrangements. *BMC Cancer*. 2022;**22**(1):859.

98. Raimondi SC, Zhou Y, Shurtleff SA, Rubnitz JE, Pui CH, Behm FG. Near-triploidy and near-tetraploidy in childhood acute lymphoblastic leukemia: association with B-lineage blast cells carrying the ETV6-RUNX1 fusion, T-lineage immunophenotype, and favorable outcome. *Cancer Genet Cytogenet*. 2006;**169**(1):50-7.

99. Stark B, Jeison M, Gobuzov R, Krug H, Glaser-Gabay L, Luria D, et al. Near haploid childhood acute lymphoblastic leukemia masked by hyperdiploid line: detection by fluorescence in situ hybridization. *Cancer Genet Cytogenet*. 2001;**128**(2):108-13.

100. Haas OA, Borkhardt A. Hyperdiploidy: the longest known, most prevalent, and most enigmatic form of acute lymphoblastic leukemia in children. *Leukemia*. 2022;**36**(12):2769-83.

101. Heerema NA, Raimondi SC, Anderson JR, Biegel J, Camitta BM, Cooley LD, et al. Specific extra chromosomes occur in a modal number dependent pattern in pediatric acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2007;**46**(7):684-93.

102. Paulsson K, Forestier E, Lilljebjorn H, Heldrup J, Behrendtz M, Young BD, et al. Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A*. 2010;**107**(50):21719-24.

103. Maia AT, van der Velden VH, Harrison CJ, Szczepanski T, Williams MD, Griffiths MJ, et al. Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins. *Leukemia*. 2003;**17**(11):2202-6.

104. Song Y, Bi Z, Liu Y, Qin F, Wei Y, Wei X. Targeting RAS-RAF-MEK-ERK signaling pathway in human cancer: current status in clinical trials. *Genes Dis*. 2023;**10**(1):76-88.

105. Paulsson K, Panagopoulos I, Knuutila S, Jee KJ, Garwicz S, Fioretos T, et al. Formation of trisomies and their parental origin in hyperdiploid childhood acute lymphoblastic leukemia. *Blood*. 2003;**102**(8):3010-5.

106. Paulsson K. Chromosomal gains as a favorable prognostic factor in pediatric ALL. *J Clin Oncol*. 2023;**41**(35):5433-6.

107. Lee SHR, Ashcraft E, Yang W, Roberts KG, Gocho Y, Rowland L, et al. Prognostic and pharmacotypic heterogeneity of hyperdiploidy in childhood ALL. *J Clin Oncol*. 2023;**41**(35):5422-32.

108. Tjio JH. The chromosome number of man. *Am J Obstet Gynecol*. 1978;**130**(6):723-4.

109. Spinner NB. Chromosome banding. Maloy SH, K., editor. *Academic Press*. 2013.

110. Caspersson T, Farber S, Foley GE, Kudynowski J, Modest EJ, Simonsson E, et al. Chemical differentiation along metaphase chromosomes. *Exp Cell Res*. 1968;**49**(1):219-22.

111. McGowan-Jordan J, Hastings R, Moore S. Re: international system for human cytogenetic or cytogenomic nomenclature (ISCN): some thoughts, by T. Liehr. *Cytogenet Genome Res*. 2021;**161**(5):225-6.

112. Shen C. Molecular diagnosis of chromosomal disorders. In: Shen C, editor. Diagnostic molecular biology (second edition). *Academic Press*. 2023. p. 393-423.

113. Drets ME, Shaw MW. Specific banding patterns of human chromosomes. *Proc Natl Acad Sci U S A*. 1971;**68**(9):2073-7.

114. Pardue ML, Gall JG. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc Natl Acad Sci U S A*. 1969;**64**(2):600-4.

115. Sajesh BV, Lichtensztejn Z, McManus KJ. Sister chromatid cohesion defects are associated with chromosome instability in Hodgkin lymphoma cells. *BMC Cancer*. 2013;**13**:391.

116. Im K, Mareninov S, Diaz MFP, Yong WH. An introduction to performing immunofluorescence staining. *Methods Mol Biol*. 2019;**1897**:299-311.

117. Moore CB, Guthrie EH, Huang MT, Taxman DJ. Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown. *Methods Mol Biol*. 2010;**629**:141-58.

118. Ali N, Karlsson C, Aspling M, Hu G, Hacohen N, Scadden DT, et al. Forward RNAi screens in primary human hematopoietic stem/progenitor cells. *Blood*. 2009;**113**(16):3690-5.

119. Garg E, Zubair M. Mass spectrometer. *StatPearls*. 2024.

120. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science*. 2006;**312**(5771):212-7.

121. Sato-Otsubo A, Sanada M, Ogawa S. Single-nucleotide polymorphism array karyotyping in clinical practice: where, when, and how? *Semin Oncol*. 2012;**39**(1):13-25.

122. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;**74**(12):5463-7.

123. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Hum Immunol*. 2021;**82**(11):801-11.

124. Rizzo JM, Buck MJ. Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prev Res*. 2012;**5**(7):887-900.

125. Vergult S, Van Binsbergen E, Sante T, Nowak S, Vanakker O, Claes K, et al. Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations. *Eur J Hum Genet*. 2014;**22**(5):652-9.

126. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*. 2013;**8**(10):77910.

127. Zhao EY, Jones M, Jones SJM. Whole-genome sequencing in cancer. *Cold Spring Harb Perspect Med*. 2019;9(3):034579.

128. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet*. 2017;**18**(1):14.

129. Yasen A, Aini A, Wang H, Li W, Zhang C, Ran B, et al. Progress and applications of single-cell sequencing techniques. *Infect Genet Evol*. 2020;**80**:104198.

130. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet*. 2019;**20**(11):631-56.

131. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh TS, et al. Ultrastructural details of mammalian chromosome architecture. *Mol Cell*. 2020;**78**(3):554-65 e7.

132. Andersson A, Olofsson T, Lindgren D, Nilsson B, Ritz C, Eden P, et al. Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations. *Proc Natl Acad Sci U S A*. 2005;**102**(52):19069-74.

133. Zaliova M, Hovorkova L, Vaskova M, Hrusak O, Stary J, Zuna J. Slower early response to treatment and distinct expression profile of childhood high hyperdiploid acute lymphoblastic leukaemia with DNA index < 1.16. *Genes Chromosomes Cancer*. 2016;**55**(9):727-37.

134. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Corrigendum: Global quantification of mammalian gene expression control. *Nature*. 2013;**495**(7439):126-7.

135. Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. 2013;**153**(3):654-65.

136. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell*. 2011;**44**(2):325-40.

137. Valton AL, Dekker J. TAD disruption as oncogenic driver. *Curr Opin Genet Dev*. 2016;**36**:34-40.

138. Yang M, Safavi S, Woodward EL, Duployez N, Olsson-Arvidsson L, Ungerback J, et al. 13q12.2 deletions in acute lymphoblastic leukemia lead to upregulation of FLT3 through enhancer hijacking. *Blood*. 2020;**136**(8):946-56.

139. Kloetgen A, Thandapani P, Ntziachristos P, Ghebrechristos Y, Nomikou S, Lazaris C, et al. Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nat Genet*. 2020;**52**(4):388-400.

140. Yang L, Chen F, Zhu H, Chen Y, Dong B, Shi M, et al. 3D genome alterations associated with dysregulated HOXA13 expression in high-risk T-lineage acute lymphoblastic leukemia. *Nat Commun*. 2021;**12**(1):3708.

141. Solomon DA, Kim JS, Bondaruk J, Shariat SF, Wang ZF, Elkahloun AG, et al. Frequent truncating mutations of STAG2 in bladder cancer. *Nat Genet*. 2013;**45**(12):1428-30.

142. Galeev R, Baudet A, Kumar P, Rundberg Nilsson A, Nilsson B, Soneji S, et al. Genome-wide RNAi screen identifies cohesin genes as modifiers of renewal and differentiation in human HSCs. *Cell Rep*. 2016;**14**(12):2988-3000.

143. Fisher JB, McNulty M, Burke MJ, Crispino JD, Rao S. Cohesin mutations in myeloid malignancies. *Trends Cancer*. 2017;**3**(4):282-93.

144. Braun R, Ronquist S, Wangsa D, Chen H, Anthuber L, Gemoll T, et al. Single chromosome aneuploidy induces genome-wide perturbation of nuclear organization and gene expression. *Neoplasia*. 2019;**21**(4):401-12.

145. Kemeny S, Tatout C, Salaun G, Pebrel-Richard C, Goumy C, Ollier N, et al. Spatial organization of chromosome territories in the interphase nucleus of trisomy 21 cells. *Chromosoma*. 2018;**127**(2):247-59.

146. Xu J, Song F, Lyu H, Kobayashi M, Zhang B, Zhao Z, et al. Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature*. 2022;**611**(7935):387-98.

147. Molina O, Vinyoles M, Granada I, Roca-Ho H, Gutierrez-Aguera F, Valledor L, et al. Impaired condensin complex and aurora B kinase underlie mitotic and chromosomal defects in hyperdiploid B-cell ALL. *Blood*. 2020;**136**(3):313-27.

148. Betts DR, Riesch M, Grotzer MA, Niggli FK. The investigation of karyotypic instability in the high-hyperdiploidy subgroup of acute lymphoblastic leukemia. *Leuk Lymphoma*. 2001;**42**(1-2):187-93.

149. Blandin AT, Muhlematter D, Bougeon S, Gogniat C, Porter S, Beyer V, et al. Automated four-color interphase fluorescence in situ hybridization approach for the simultaneous detection of specific aneuploidies of diagnostic and prognostic significance in high hyperdiploid acute lymphoblastic leukemia. *Cancer Genet Cytogenet*. 2008;**186**(2):69-77.

150. Alpar D, Pajor G, Varga P, Kajtar B, Poto L, Matics R, et al. Sequential and hierarchical chromosomal changes and chromosome instability are distinct features of high hyperdiploid pediatric acute lymphoblastic leukemia. *Pediatr Blood Cancer*. 2014;**61**(12):2208-14.

151. Holland AJ, Cleveland DW. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat Rev Mol Cell Biol*. 2009;**10**(7):478-87.

152. Gisselsson D. Aneuploidy in cancer: sudden or sequential? *Cell Cycle*. 2011;**10**(3):359-61.

153. Haas OA. Somatic sex: on the origin of neoplasms with chromosome counts in uneven ploidy ranges. *Front Cell Dev Biol*. 2021;**9**:631946.

154. Onodera N, McCabe NR, Rubin CM. Formation of a hyperdiploid karyotype in childhood acute lymphoblastic leukemia. *Blood*. 1992;**80**(1):203-8.

155. Paulsson K, Morse H, Fioretos T, Behrendtz M, Strombeck B, Johansson B. Evidence for a single-step mechanism in the origin of hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2005;**44**(2):113-22.

156. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim Biophys Acta Rev Cancer*. 2017;**1867**(2):151-61.

157. Gisselsson D, Jin Y, Lindgren D, Persson J, Gisselsson L, Hanks S, et al. Generation of trisomies in cancer cells by multipolar mitosis and incomplete cytokinesis. *Proc Natl Acad Sci U S A*. 2010;**107**(47):20489-93.

Article I

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09469-3 OPEN

Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia

Minjun Yang[1], Mattias Vesterlund[2], Ioannis Siavelis[2], Larissa H. Moura-Castro[1], Anders Castor[3], Thoas Fioretos[1], Rozbeh Jafari[2], Henrik Lilljebjörn[1], Duncan T. Odom[4,5], Linda Olsson[1,6], Naveen Ravi[1], Eleanor L. Woodward[1], Louise Harewood[4,7], Janne Lehtiö[2] & Kajsa Paulsson[1]

Hyperdiploidy, i.e. gain of whole chromosomes, is one of the most common genetic features of childhood acute lymphoblastic leukemia (ALL), but its pathogenetic impact is poorly understood. Here, we report a proteogenomic analysis on matched datasets from genomic profiling, RNA-sequencing, and mass spectrometry-based analysis of >8,000 genes and proteins as well as Hi-C of primary patient samples from hyperdiploid and *ETV6/RUNX1*-positive pediatric ALL. We show that CTCF and cohesin, which are master regulators of chromatin architecture, display low expression in hyperdiploid ALL. In line with this, a general genome-wide dysregulation of gene expression in relation to topologically associating domain (TAD) borders were seen in the hyperdiploid group. Furthermore, Hi-C of a limited number of hyperdiploid childhood ALL cases revealed that 2/4 cases displayed a clear loss of TAD boundary strength and 3/4 showed reduced insulation at TAD borders, with putative leukemogenic effects.

[1] Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, SE-221 84 Lund, Sweden. [2] Department of Oncology-Pathology, Science for Life Laboratory and Karolinska Institute, Clinical Proteomics Mass Spectrometry, SE-171 21 Stockholm, Sweden. [3] Department of Pediatrics, Skåne University Hospital, Lund University, SE-221 85 Lund, Sweden. [4] Cancer Research UK Cambridge Institute (CRUK-CI), University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, UK. [5] German Cancer Research Center (DKFZ), Division of Signaling and Functional Genomics, 69120 Heidelberg, Germany. [6] Department of Clinical Genetics and Pathology, Office for Medical Services, Division of Laboratory Medicine, SE-221 85 Lund, Sweden. [7] Precision Medicine Centre of Excellence, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7AE, UK. These authors contributed equally: Minjun Yang, Mattias Vesterlund. Correspondence and requests for materials should be addressed to J.L. (email: janne.lehtio@ki.se) or to K.P. (email: kajsa.paulsson@med.lu.se)

Aneuploidy, i.e., changes in chromosome numbers, is one of the most common phenomena in cancer cells. In spite of the huge efforts that have gone into understanding the impact of somatic genetic events in cancer, the effects of aneuploidy in tumorigenesis remain poorly understood. In fact, it is even debated whether aneuploidy in itself may be a driver event or if it is a passenger event without consequences in tumor development[1].

High hyperdiploid (51–67 chromosomes) pediatric B-cell precursor acute lymphoblastic leukemia (BCP ALL) is one of the most common malignancies in early childhood and is associated with a median age at diagnosis of 3–5 years, a low white blood cell count, and a favorable prognosis on contemporary treatment protocols[2]. Genetically, its defining feature is a non-random aneuploidy consisting of extra chromosomes, most commonly X, 4, 6, 10, 14, 17, 18, and 21. Approximately half of cases also harbor mutations in the RTK–RAS pathway, primarily *KRAS*, and 20% have mutations in histone modifiers such as *CREBBP*, in addition to microdeletions of various genes involved in B-cell differentiation/cell cycle control[3,4]. However, these additional aberrations are seen only in a subset of the cases, are sometimes gained or lost at relapse, and, when occurring, are frequently subclonal, whereas the aneuploidy is uniformly present[3,4]. Furthermore, we and others have shown that the chromosomal gains in these cases are early and likely leukemia-initiating aberrations, often arising several years before overt disease[3,5]. Taken together, available data strongly indicate that the aneuploidy is the main driver event in this type of leukemia, but the underlying leukemogenic mechanism remains unclear.

Previous studies of the RNA expression pattern in high hyperdiploid ALL have revealed a general upregulation of genes on the gained chromosomes, hinting that dosage effects may occur[3,6,7]. However, no detailed analysis of how this may affect leukemogenesis has yet been published and it remains unknown how the chromosomal gains may cause the development of leukemia. Additionally, since genes are subject to post-transcriptional control, the RNA expression level of a gene may not be directly transferable to the protein level. To address this, we performed a proteogenomic analysis of a series of pediatric BCP-ALL, including high hyperdiploid and diploid/near-diploid *ETV6/RUNX1*-positive cases, aiming to determine the effects of aneuploidy. Besides demonstrating that the characteristic extra chromosomes have an impact on the transcriptome and proteome, we also present data suggesting that hyperdiploid leukemia cases harbor aberrant chromatin organization that causes genome-wide transcriptional dysregulation. Taken together, our data give insight into the leukemogenesis of this common and clinically important pediatric leukemia.

Results

Proteogenomic analysis of childhood BCP ALL. This study comprised mass spectrometry (MS)-based analysis of the proteome, whole genome and/or whole exome sequencing (WGS/WES) for somatic mutations and structural events, SNP array analysis for copy number assessment, and RNA-sequencing (RNA-seq) for RNA expression in childhood BCP ALL. In total, 48 high hyperdiploid and 41 *ETV6/RUNX1*-positive cases were investigated; the cohort analyzed with MS comprised eighteen high hyperdiploid and nine *ETV6/RUNX1*-positive pediatric ALL cases, as ascertained by chromosome banding, fluorescence in situ hybridization (FISH), SNP array analysis and reverse transcriptase-PCR for the fusion transcript (Fig. 1a and Supplementary Data 1 and 2).

MS data were generated via high-resolution isoelectric focusing liquid chromatography mass spectrometry (HiRIEF LC-MS/MS) workflow together with isobaric labeling (TMT10) for relative quantification between tumors[8]. In total, 10,981 proteins originating from 10,138 genes were identified at 1% protein false discovery rate (FDR) based on 174,966 unique peptides (Fig. 1b, Supplementary Table 1 and Supplementary Data 3). For all quantitative proteome analyses, we used a gene symbol centric subset of 8480 genes that were quantified in each of the 27 tumors (Fig. 1b).

Genomic and transcriptomic variation was observed at the peptide level by searching HiRIEF LC-MS/MS spectra against a customized sequence database, which included both human RefSeq protein as well as somatic mutant and fusion sequences derived from WGS/WES and RNA-seq. Although many SNPs were seen in the protein dataset, none of the somatic mutations could be detected. The *ETV6/RUNX1* fusion could be identified at the protein level in all nine cases from this subgroup.

Proteome analyses give improved biological insight in cancer. For 8222 (97%) of the 8480 proteins detected by HiRIEF LC-MS/MS, the expression of the corresponding mRNA could be ascertained by rRNA-depleted RNA-seq (RiboZero RNA-seq) of the same samples (Supplementary Data 4 and 5). Expression levels were positively correlated for most (75%) mRNA–protein pairs across the 27 samples, with 22% showing significant correlation (multiple-test adjusted $P \leq 0.05$) and a mean Spearman's correlation coefficient of 0.24 (Fig. 1c). This is similar to what has previously been reported in colorectal cancer[9], but lower than in ovarian cancer and breast cancer[10,11]. When the correlation scores were ascertained for different KEGG pathways, scores were highest for specialized pathways, such as hematopoietic cell lineage and amino acid metabolism, and lowest for house-keeping functions, e.g., ribosomal and spliceosomal processes (Fig. 1d). This is in line with the previous studies[9–11] and demonstrates that the level of expression of mRNA is not always directly translatable to the protein level. To further test the post-translational gene regulation effect on leukemia samples, we performed pairwise correlation of gene/protein abundance for all 8222 proteins. Similar to results previously obtained from the TCGA and CPTAC datasets[12], the correlation score for pairs of proteins involved in the same protein complex displayed a degree of co-regulation (mean Spearman's correlation coefficient = 0.19) that was significantly higher than that observed for random pairs (mean Spearman's correlation coefficient = 0) (Supplementary Fig. 1). A similar co-regulation effect could be seen at the transcript level (mean Spearman's correlation coefficient = 0.16), but the correlation was significantly lower than at the protein level (two-tailed Mann–Whitney *U*-test $P = 3.80e{-}20$).

To explore the details of this relatively low correlation between mRNA expression and protein abundance, we investigated several aspects of mRNA and protein regulation. A global comparison of stable and unstable mRNAs and their corresponding proteins[13] revealed significantly higher correlation for genes with similar stability on both the mRNA and protein levels (Supplementary Fig. 2). Next, we investigated the potential impact of miRNA-targeting. Genes regulated by miRNAs[14] displayed significantly lower mRNA and protein correlations, showing a role for post-transcriptional RNA regulation (Supplementary Fig. 2). We also observed that protein level regulation by the ubiquitin–proteasome pathway[15] affected the correlations since genes with low mRNA–protein correlations were significantly more frequently targeted by the proteasome (Supplementary Fig. 2). Consistent with this result, an analysis of the protein degradation rate also showed that genes with low mRNA–protein correlations were enriched among rapidly degrading proteins[16] (Supplementary Fig. 2). Interestingly, we observed that protein

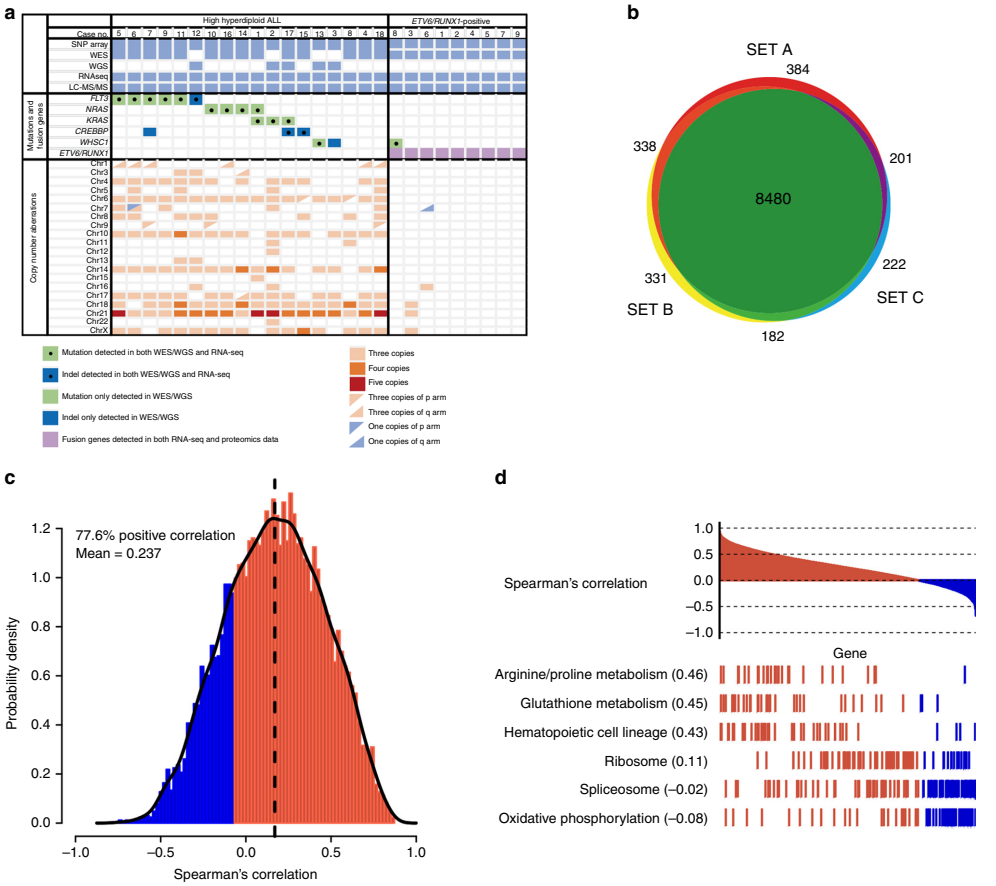


Fig. 1 Proteogenomic study of childhood B-cell precursor acute lymphoblastic leukemia (ALL). **a** Genomic landscape of 27 childhood ALL included in the proteogenomic analysis. All cases were disomic for chromosomes 2, 19, and 20. **b** Numbers of proteins overlapping across the 3 TMT-sets. **c** Spearman's rank order correlation between mRNA and protein abundance. The correlation was positive for 77.6% mRNA-protein pairs in the whole cohort of cases with a mean Spearman's correlation coefficient of 0.24. Approximately 23% mRNA-protein pairs showed significant correlation (multiple-test adjusted $P \leq 0.05$). **d** When investigating different biological processes, mRNA and protein levels displayed the highest correlation for specialized pathways, such as hematopoietic cell lineage and amino acid metabolism, and lowest for house-keeping functions, e.g., ribosomal and spliceosomal processes

subcellular localization had an effect as cytosolic and proteins residing in the plasma membrane, endoplasmic reticulum and the Golgi (i.e., secretory proteins) exhibited increased mRNA-protein correlations whilst nuclear and mitochondrial proteins did not (Supplementary Fig 2). Finally, we also observed that mRNA and proteins that were differentially expressed between hyperdiploid and *ETV6/RUNX1*-positive leukemia had higher mRNA-protein correlations (Supplementary Fig. 2). This fits with the observation in Orre et al.[17] that the secretory protein subset provides a better separation of cell lineages and cell types compared to nuclear proteins. Phenotypic genes thus seem to be more highly correlated on the mRNA–protein levels. Taken together, our analyses show that multiple factors contribute to lowering the correlation between mRNA and protein levels. Thus, proteome analyses are likely to give more biologically relevant data on dysregulated pathways in cancer than RNA expression analyses alone.

Impact of copy number events. To study the impact of the extra chromosomes in high hyperdiploid ALL, we first compared the mean RNA and protein expression according to copy number in high hyperdiploid ALL. This clearly showed that the hyperdiploidy is associated with dosage effects, i.e., a generally increasing expression of genes and proteins with higher copy number (also termed *cis* effects; Fig. 2a). Notably, however, not all genes and proteins were affected in this way; approximately 16% (283/2,080) of genes and 25% (523/2,080) of proteins instead showed negative correlation with copy number. Thus, copy number gain does not always lead to increased expression, in particular at the protein

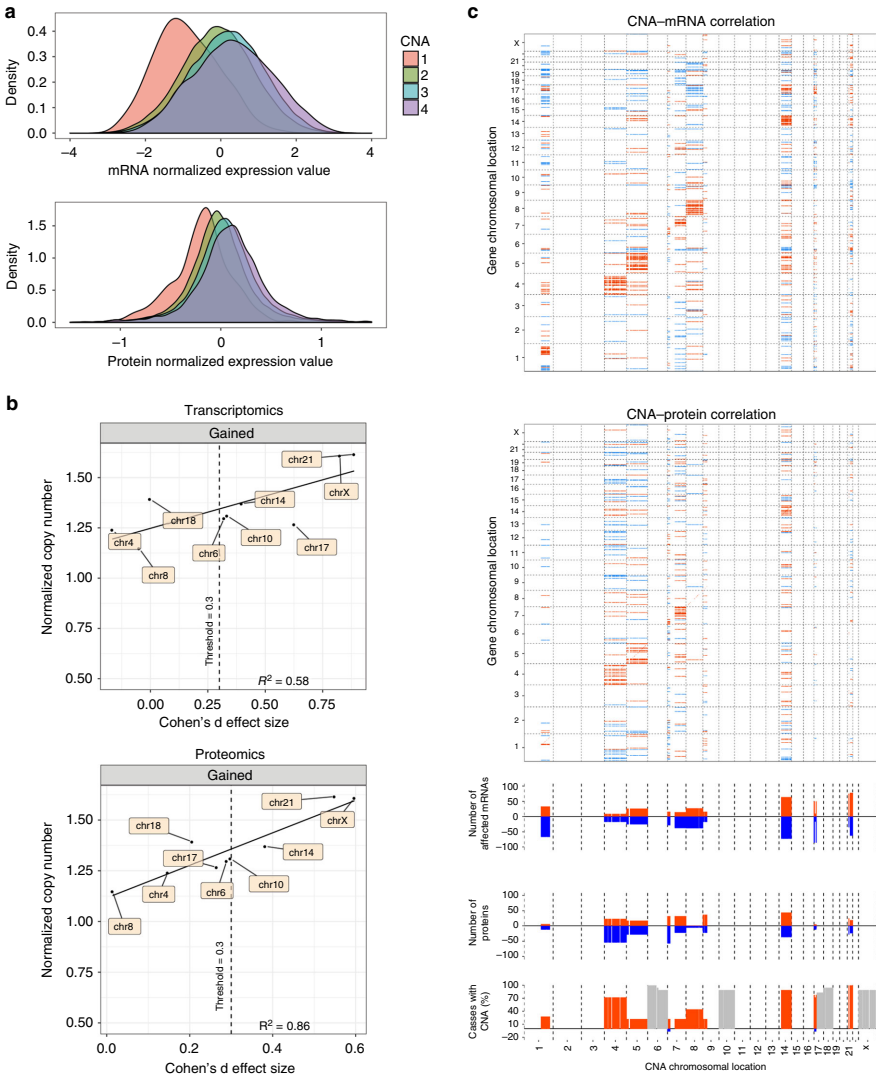


Fig. 2 Effects of copy number alterations on mRNA and protein abundance. **a** Dosage effects in 18 high hyperdiploid ALL at the RNA and protein levels. The effects were lower on the protein level than on RNA level, showing additional layers of control for protein expression. **b** Cohen's d effect size analysis of gained chromosomes in high hyperdiploid vs. *ETV6/RUNX1*-positive leukemia. **c** *cis* and *trans* effects of copy number changes in 18 cases of hyperdiploid childhood B-cell precursor acute lymphoblastic leukemia. Correlations of copy number aberrations (CNA) (*x*-axes) to RNA (top) and protein (bottom) expression levels (*y*-axes) are shown. Note that a large fraction of the genome was not included in the analysis since there was no copy number variance, either because all cases had two copies or because all cases had three copies. Significant (multiple-test adjusted $P < 0.05$) positive (red) and negative (blue) correlations between CNA and mRNAs/proteins are indicated. CNA *cis* effects appear as a red diagonal line, CNA *trans* effects as vertical stripes. The fraction (%) of significant CNA *trans* effects (positive in red and negative in blue) for each CNA gene is shown below. The bottom panel shows the fraction (%) of leukemias harboring CNA (copy number gain in red and copy number loss in blue). Chromosomes that were gained in more than 16 cases were not informative; their copy number is shown in gray

level, presumably because of feedback loops controlling expression and protein turnover. Cohen's d effect size analysis showed that gain of chromosomes X, 14, and 21 was associated with stronger dosage effects compared with the other commonly gained chromosomes in both the RNA-seq and proteomics datasets, with a linear relationship between normalized copy number and effect size (Fig. 2b). Thus, our data clearly support previous studies showing a general—albeit not ubiquitous—upregulation of genes in high hyperdiploid ALL[3,6]. Furthermore, we demonstrate that these dosage effects also are seen at the protein level, with proteins encoded on the gained chromosomes generally being more highly expressed.

To further investigate *cis* as well as *trans* (genes/proteins in other genomic regions) effects of copy number changes, we used a linear regression model to study the correlation between copy number and expression (Fig. 2c and Supplementary Fig. 3)[18]. In addition to the MS and RiboZero RNA-seq datasets, RNA expression from a previously published RNA-seq study was analyzed, comprising 83 cases (oligo(dT) RNA-seq; European Genome-phenome Archive accession number EGAD00001002112; Supplementary Data 1 and 6)[19]. In order to avoid outlier-driven results, only 2080 genes displaying copy number variation involving more than three cases were retained in the *cis*-effect analysis. Again, *cis* dosage effects were seen, involving 25% (524/2080) of genes and 12% (245/2080) of proteins at a significance level of $P < 0.05$ (Fig. 2c, Supplementary Fig. 3). Furthermore, *trans* effects on the whole transcriptome and proteome were seen for all informative regions. These were generally lower in the proteome data, in particular for the long arm of chromosome 1 and chromosomes 8, 17, and 21, which displayed very few *trans* effects in the protein dataset. To further investigate the leukemogenic impact of individual chromosomal gains, we mapped known cancer driver genes to see whether they were associated with the chromosomal pattern of high hyperdiploid ALL. That is, whether oncogenes were more and tumor suppressor genes less commonly located on the frequently gained chromosomes. However, no such association was seen (Supplementary Fig. 3). Taken together, the analysis of copy number and gene/protein expression confirmed that the extra chromosomes in high hyperdiploid ALL have a large impact at RNA and protein levels in both *cis* and *trans*.

Protein expression differences between leukemic subtypes. Next, we focused on expression differences between high hyperdiploid and *ETV6/RUNX1*-positive ALL. To investigate whether proteomics could be used to distinguish between high hyperdiploid and *ETV6/RUNX1*-positive leukemia, hierarchical cluster and principal component analyses were performed. The two subtypes clustered separately in unsupervised analyses, both by RNA and protein expression, in 27 cases (Fig. 3a). In supervised analysis, 2423 genes and 1286 proteins were upregulated and 2222 genes and 1127 proteins were downregulated in high hyperdiploid cases compared with *ETV6/RUNX1*-positive cases (multiple-test adjusted $P \leq 0.05$) (Supplementary Data 4 and 5). Of these, 684 upregulated and 624 downregulated genes and proteins overlapped (Supplementary Data 4 and 5). Overall, there was a linear relationship between the log2 fold changes of RNA-seq and proteomics data (Spearman's correlation coefficient = 0.54, $P < 2.2e{-}16$) (Fig. 3b), with the correlation being stronger for gene/protein pairs with high fold changes.

Several of the top differentially expressed proteins have previously been reported to play a role in leukemogenesis or to be associated with ALL. These include, for example, CD44 and FLT3 (Supplementary Data 4 and Supplementary Fig. 4), which were highly expressed in high hyperdiploid cases. *ETV6/RUNX1*-positive BCP-ALL has previously been reported to display a $CD44^{low\text{-}negative}$ immunophenotype[20], agreeing well with this protein being differentially expressed by proteomics. In regards to FLT3, the *FLT3* gene harbors activating mutations in approximately 10–20% of high hyperdiploid ALL and has previously been reported to be highly expressed in high hyperdiploid ALL regardless of mutational status[3,21]. Here we show that this high expression is maintained at the protein level, suggesting that FLT3 may be involved in the leukemogenesis of high hyperdiploid childhood ALL also in the absence of mutations. A comparison with six sorted pro-B/pre-B samples—the normal cells considered to be closest to ALL blasts—in the oligo(dT) RNA-seq dataset confirmed that *CD44* and *FLT3* were highly expressed in the hyperdiploid leukemias (Supplementary Fig. 4). Among the top-downregulated proteins in high hyperdiploid cases were IGF2BP1, CLIC5, RAG1, and RAG2, which also showed low RNA expression compared with the normal pro-B/pre-B dataset (Supplementary Data 4 and Supplementary Fig. 4). *IGF2BP1* is recurrently involved in fusions with *IGH@* in BCP ALL and has previously been reported to be highly expressed in *ETV6/RUNX1*-positive cases[22,23]. *CLIC5* has been shown to be a target of ETV6 and loss of ETV6 leads to its upregulation, providing the cells with higher resistance to lysosome-mediated apoptosis[24]. As regards RAG1 and RAG2, these are key components of somatic V(D)J recombination and this process has previously been shown to be involved in *ETV6/RUNX1*-mediated leukemogenesis[25], agreeing well with the high expression seen in our cohort. Taken together, the top differentially expressed proteins obtained by MS agree well with previously reported RNA expression results, supporting the validity of our proteomics approach.

Gene set enrichment analysis (GSEA) was performed to identify dysregulated pathways in BCP-ALL. Pathways that were enriched in high hyperdiploid ALL in the protein analysis could be divided into six different categories: (1) translation and ribosomes, (2) innate immunity, (3) cell adhesion, (4) cytokines and activated signaling, (5) protein folding and proteolysis, and (6) the endosome (Fig. 3c and Supplementary Data 7). Pathways that were enriched in *ETV6/RUNX1*-positive cases comprised those related to: (1) chromatin organization, modification, and structure, (2) the G2/M checkpoint, and (3) mitochondria (Fig. 3c and Supplementary Data 8). Support for enrichment at the RNA level was seen for all these processes expect pathways related to mitochondria, both in the RiboZero and the oligo(dT) datasets (Supplementary Data 9–12). Taken together, the GSEA results suggest an upregulation of translation and protein metabolism, including proteolysis and the endosome, in high hyperdiploid ALL. This may be explained by the additional transcription from the extra chromosomes, which would be expected to result in a general increase in translation. Furthermore, aneuploidy, in particular hyperdiploidy, has been reported to be associated with a proteotoxic stress response related to increased strain on the protein folding pathways of the cell[26], which could explain the enrichment for protein folding and proteolysis seen here. The relative downregulation of pathways related to chromatin organization, modifications and structure may be related to a higher proliferative capacity of *ETV6/RUNX1*-positive cases, but could also be associated with epigenetic events in high hyperdiploid ALL, in particular in light of the changes in chromatin organization that we found by high-resolution chromosome conformation capture (Hi-C) in this subtype (see below). That the G2/M checkpoint is enriched in *ETV6/RUNX1*-positive cases, on the other hand, agrees well with the previously reported importance of DNA recombination in such cases[25].

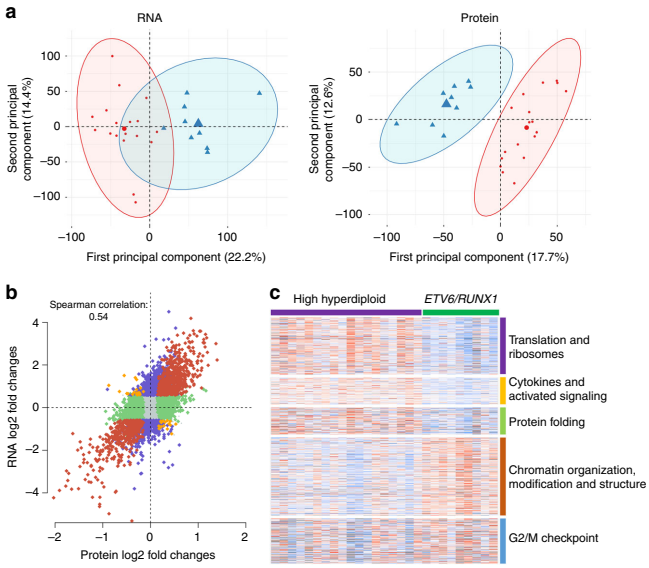


Fig. 3 Clustering and enriched pathways in high hyperdiploid and *ETV6/RUNX1*-positive leukemia. **a** Principal component analyses of 27 B-cell precursor acute lymphoblastic leukemias showed that high hyperdiploid (red) and *ETV6/RUNX1*-positive (blue) cases clustered separately in unsupervised analyses by scaled RNA (left) and protein (right) abundance. **b** There was a linear relationship between the log2 fold changes of RNA-sequencing and proteomics data (Spearman's correlation coefficient = 0.54) between the high hyperdiploid and *ETV6/RUNX1*-positive subtypes, demonstrating a high correlation between changes on the transcript and translation levels of an individual gene product. The correlation was stronger for gene/protein pairs with high fold changes. Significant changes found in both RNA-seq and LC-MS/MS are shown in red, inverse changes found in RNA-seq and LC-MS/MS in yellow, significant changes found only in LC-MS/MS in green, and changes only found in RNA-seq in purple. **c** Gene set enrichment analysis of protein data highlighted sets of pathways that were significantly different between high hyperdiploid and *ETV6/RUNX1*-positive cases

Transcriptional dysregulation in high hyperdiploid ALL. We further found that CTCF, as well as several members of the cohesin complex, were significantly lower expressed at both the RNA and protein levels in our high hyperdiploid cases compared with *ETV6/RUNX1*-positive cases as well as compared with normal pro-B/pre-B cells (Fig. 4a, Supplementary Fig. 5). Analyses of the oligo(dT) RNA-seq dataset and two different publicly available array-based gene expression datasets (GEO accession numbers GSE13351 and GSE13425) confirmed that the expression of both CTCF and cohesin seems to be particularly low in high hyperdiploid ALL compared with other types of childhood ALL (Supplementary Fig. 5). We did not observe any differences in complex formation or correlation between the cohesin complex members between the high hyperdiploid and the *ETV6/RUNX1*-positive cases (Supplementary Fig. 5), indicating that there was a general downregulation of gene/protein expression and not a disturbance of the complex formation. *CTCF* has recently been identified as a putative tumor suppressor gene in ALL and we have previously reported a *CTCF/PARD6A* fusion that presumably results in disruption of the normal function of CTCF in one case of high hyperdiploid ALL[3,27]. Besides being a transcription factor, CTCF binds to chromatin at interphase and, together with the cohesin complex, forms the basis for the formation of topologically associating domains (TADs); chromatin loops <1 Mb in size containing DNA sequences that interact more frequently with each other than with external sequences[28]. TADs are generally conserved between different tissues and their disruption leads to changes in gene expression when the insulating function of the TAD boundaries is lost[28]. Thus, CTCF and cohesin are master regulators of transcription.

We hypothesized that this low expression of CTCF and cohesin could have genome-wide effects on the transcriptional regulation in high hyperdiploid ALL. To address this, we first investigated whether the differences in gene expression between high hyperdiploid and *ETV6/RUNX1*-positive ALL were associated with the number of CTCF binding sites in gene bodies and the flanking 5 kb, i.e., genes regulated by CTCF binding. We found that differentially expressed genes in both the oligo(dT) and RiboZero RNA-seq datasets were strongly enriched for more CTCF binding sites compared with genes that showed similar expression in the two ALL subtypes (chi-square test; $P = 3.41e-05$ and $P = 1.549e-05$, respectively; Supplementary Fig. 6), in line with the previous experimental data from a mouse model with reduced CTCF expression[29]. Furthermore, genes with higher numbers of CTCF binding sites showed larger fold changes in both datasets (Supplementary Fig. 6). We also classified genes as anchor genes or background genes based on the distance of their transcription start sites to the closest CTCF/cohesin anchors, forming the basis for chromatin loops, according to published chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) data[30]. We found that a significantly higher proportion of the genes that were differentially expressed between high hyperdiploid and *ETV6/RUNX1*-positive leukemias were anchor genes in both the oligo(dT) and the RiboZero RNA-seq datasets (hypergeometric test; $P = 0.0139$ and $P = 0.00513$, respectively; Supplementary Fig. 7). Furthermore, differentially

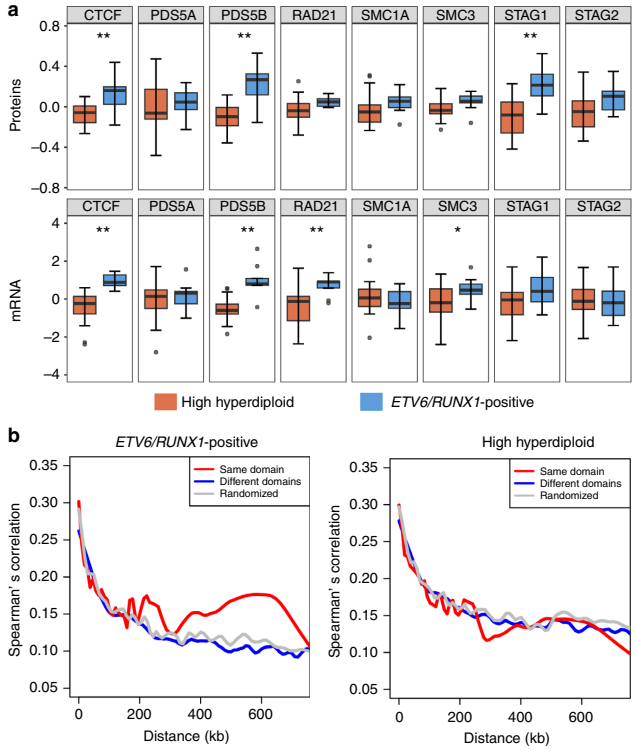


Fig. 4 Low CTCF/cohesin expression and transcriptional dysregulation in high hyperdiploid leukemia. **a** Boxplots of the expression of CTCF and members of the cohesin complex in proteomics (top) and RiboZero RNA-sequencing datasets (bottom). Low expression of CTCF/cohesin complex members was seen in the high hyperdiploid subgroup at both the RNA and protein levels. The center of the boxplot is the median and lower/upper hinges correspond to the first/third quartiles; whiskers are 1.5 times the interquartile range and data beyond this range are plotted as individual points. **b** Spearman's correlation coefficient between gene pairs as a function of distance across the oligo(dT) RNA-sequencing dataset for *ETV6/RUNX1*-positive cases (left; $n = 39$) and high hyperdiploid ALL (right; $n = 44$). The analysis showed that the expression of gene pairs in the same topologically associating domain (TAD; red) displayed higher correlation than those in different domains (blue) or randomly selected regions (gray) in *ETV6/RUNX1*-positive cases, whereas no difference was seen in high hyperdiploid ALL, suggesting that transcriptional dysregulation in hyperdiploid cases is related to TAD borders

expressed anchor genes showed significantly higher fold changes (Mann–Whitney *U*-test; $P = 6.6e{-}6$ and $P = 1.31e{-}4$, respectively; Supplementary Fig. 7), suggesting that a portion of differentially expressed genes between hyperdiploid and *ETV6/RUNX1*-positive cases were the result of changes in CTCF binding.

To further investigate how the low levels of CTCF and cohesin affected genome-wide transcription, we used publicly available data to classify gene pairs according to whether they should be divided by a TAD boundary or not, since the overall TAD structure is generally conserved in human tissues[31]. We then investigated whether their expression was correlated. First, we analyzed RNA-seq data from a large cohort of childhood ALL ($n = 201$) including all genetic subtypes[19], from normal bone marrow ($n = 20$)[19], from acute myeloid leukemia (TCGA-LAML, https://portal.gdc.cancer.gov/projects/TCGA-LAML; $n = 151$)[32] and from papillary renal-cell carcinoma (TCGA-KIRP, https://portal.gdc.cancer.gov/projects/TCGA-KIRP; $n = 270$)[33]. These datasets all clearly displayed higher correlation between the expression of gene pairs within the same TAD compared with gene pairs separated by a TAD boundary (Supplementary Fig. 8), showing that the TAD structure used in the analysis corresponded well with actual transcriptional regulation and was in line with previous studies[34]. We then performed the same analysis in high hyperdiploid ($n = 44$) and *ETV6/RUNX1*-positive ($n = 39$) cases separately. Whereas the *ETV6/RUNX1*-positive leukemias showed a difference between intra- and inter-TAD gene pairs, similar to the other datasets investigated, no correlation with the expected TAD structure was observed in the high hyperdiploid samples (Fig. 4b). Analysis of two publicly available array-based gene expression datasets from childhood ALL (GEO accession numbers GSE13351 and GSE13425) confirmed that high hyperdiploid ALL displayed aberrant expression in relation to the expected TAD structure (Supplementary Fig. 8). Restricting the analysis to only the commonly trisomic or only the commonly disomic chromosomes did not change the result (Supplementary Fig. 8), suggesting that the phenomenon is not directly associated with the copy number of

individual chromosomes. Taken together, the analyses suggest that high hyperdiploid ALL exhibits an aberrant gene expression pattern associated with changes in DNA looping.

TAD boundaries in high hyperdiploid ALL. To investigate further the TAD organization in childhood ALL, we performed in situ Hi-C analysis on four high hyperdiploid and two *ETV6/RUNX1*-positive cases (Supplementary Data 13). Raw sequencing data were processed using the HiCUP pipeline[35], resulting in 230–320 million unique valid sequence tags per sample. An average of 3450 TADs (range 2965–3960) was identified per case by Domaincaller at 25 kb resolution with a mean size of ~740 kb (range 645–845 kb) (Supplementary Data 13). The average number of boundaries was 4230 (range 3853–4659) (Supplementary Data 13). The majority of TAD boundaries were expected to be bound by the insulator protein CTCF (86%) and the cohesin subunit RAD21 (80%), in line with previous reports (Supplementary Data 13)[31]. Comparing the TAD boundaries of our samples with the high resolution Hi-C dataset from the human lymphoblastoid cell line GM12878 [31] showed that approximately 70% of the TAD boundaries we found were also present in GM12878, indicating that the overall TAD structure was intact in the leukemia samples (Fig. 5, Supplementary Data 13).

We then investigated whether the low expression of CTCF/cohesin affects higher-order segregation of active and inactive chromosome domains into A and B compartments. We determined the compartment types of the genome at 500 kb resolution in leukemia samples as well as GM12878 cell line using Juicer eigenvector[36]. Overall, most (>90%) genomic regions were in the same compartment in the leukemia samples as in the GM12878 cell line and more than 98% of genomic regions were in the same compartment in high hyperdiploid ALL and *ETV6/RUNX1*-positive cases, in line with previous studies showing that depletion of CTCF does not lead to compartment switching[37].

Comparing *ETV6/RUNX1*-positive and high hyperdiploid cases, the number of TAD boundaries were reduced and the average TAD structure length was increased by 21–120 kb in three of four high hyperdiploid ALL. The exception (case HeH_42) had an average TAD length of 645 kb (Supplementary Data 13). This increase in TAD lengths in the three hyperdiploid cases resulted from the partial fusion of multiple TADs into one, in line with the decrease in the number of TAD boundaries. In order to analyze changes in chromatin organization, we focused on recurrent changes in boundaries detected in the different subgroups of leukemia. One hundred thirty-one boundaries were weakened or absent in at least two high hyperdiploid samples whereas only 14 boundaries were absent in both *ETV6/RUNX1*-positive cases (Fig. 5 and Supplementary Data 14). Most of the corresponding boundaries overlapped CTCF-cohesin binding sites (127/131, 97%), indicating that the loss of TAD boundaries was associated with loss of a functional CTCF/cohesin complex in high hyperdiploid samples (Supplementary Data 14). We then checked the expression of mRNA and proteins in the RiboZero RNA-seq and MS datasets. Of the 298 expressed mRNAs and 210 expressed proteins encoded within 1 Mb of the lost boundaries, significant (multiple-test adjusted $P \leq 0.05$) expression differences between high hyperdiploid and *ETV6/RUNX1*-positive ALL could be seen for 134 (45%) and 65 (31%), respectively (Supplementary Data 15). Of these differentially expressed genes/proteins, 98/134 (73%) genes and 42/64 (66%) proteins were downregulated in high hyperdiploid samples, which was more often than by chance (chi-square test, $P = 4.6e{-}9$ for RNA-seq, and $P = 0.0032$ for proteomics). This indicates that changes in chromatin organization caused by CTCF/cohesin complex depletion tend to downregulate gene expression in high hyperdiploid cases.

Although the global chromosomal interaction pattern appeared largely unchanged between high hyperdiploid, *ETV6/RUNX1*-positive cases and the cell line GM12878, closer inspection showed differences in the strength of internal interactions within TADs. To test whether the TAD interaction strength was affected, we used directionality index and insulation score analyses to calculate the ratio of interactions found within TADs versus those spanning a boundary[38,39]. We found a genome-wide change in directionality index as well as in insulation score between cases: two high hyperdiploid cases (HeH_9 and HeH_10) showed a clear loss of boundary strength based on directionality index analysis and three (HeH_9, HeH_10, and HeH_48) showed reduced insulation based on insulation score analysis compared with the GM12878 cell line (Fig. 6). Thus, whereas the position of TAD boundaries remained largely unchanged in high hyperdiploid ALL samples, their quality was affected by changes in local and distal interactions, with a fraction of TADs losing insulation strength. This suggests that at least a subset of high hyperdiploid ALL have significant loss of insulation at TAD borders, agreeing well with the observed transcriptional dysregulation in this subgroup.

Poor metaphase chromosome morphology in hyperdiploid ALL. To further explore the possibility of an aberrant chromatin organization in high hyperdiploid ALL, we next focused on the metaphase chromosomes. Metaphase chromosome morphology, corresponding to the number of bands obtained with banding techniques as well as the size and general appearance of the chromosomes, varies between different cells and tissues. CTCF is bound to chromatin throughout the cell cycle and it has been suggested that it also affects the metaphase chromosome architecture[40]. We, therefore, hypothesized that high hyperdiploid ALL may have aberrant chromosome morphology. In fact, a common opinion among hematological cytogeneticists is that this genetic subtype displays particularly poor chromosome morphology, although this has not, to the best of our knowledge, been properly investigated. To address this issue, we developed a scale from 1 to 3 (1 corresponding to poor and 3 to good morphology; Fig. 7) for scoring chromosome morphology (chromosome morphology score; CMS) in a consistent manner and applied it to 37 cases of high hyperdiploid ALL and 33 cases of *ETV6/RUNX1*-positive ALL. Although the CMS varied between cells within the same case, there were clear differences in the mean values between cases (Supplementary Data 16). Furthermore, the investigation revealed a difference in mean CMS between the two genetic subtypes, with high hyperdiploid ALL displaying significantly lower CMS, corresponding to poorer chromosome morphology (Mann–Whitney one-sided test; $P = 0.0075$; mean CMS 1.8 vs. 2.1; Fig. 7; Supplementary Data 16). Thus, metaphase chromosomes of high hyperdiploid ALL show signs of an aberrant chromatin organization, in line with our RNA-seq and Hi-C results.

Discussion

We here report a full-scale proteogenomic analysis of childhood ALL, including the two largest subtypes of this disease that together constitutes more than half of cases. The investigation encompassed more than 8000 proteins and 12,000 RNAs in genetically well-characterized cases. To the best of our knowledge, only primary tumor samples from rhabdomyosarcoma, colon and rectal, prostate, and breast cancer have previously been subjected to proteogenomic analyses to this level[9–11,41]. Although many previous studies of childhood ALL have utilized RNA expression,

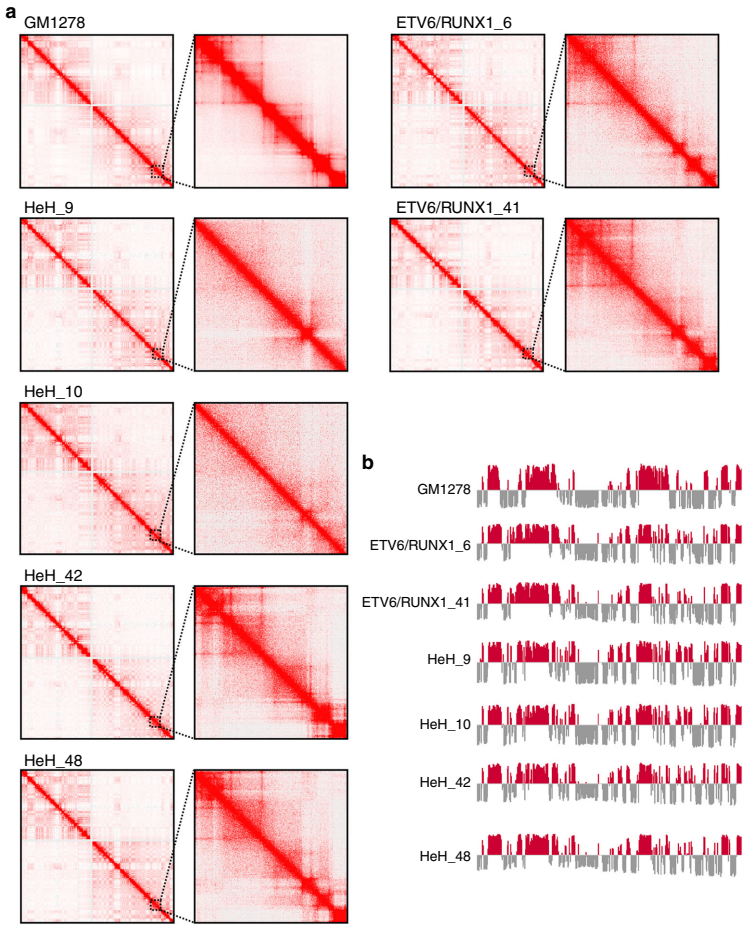


Fig. 5 Hi-C of high hyperdiploid and *ETV6/RUNX1*-positive cases. **a** Contact matrices from chromosome 3, selected because it is disomic in all cases and displays no structural aberrations. The whole chromosome at 250 kb resolution is shown to the left and the 161–172 Mb region at 25 kb resolution to the right. At a resolution of 250 kb, the interaction profile is similar, showing that the general chromatin architecture is intact. However, at a resolution of 25 kb, it can clearly be seen that two of the high hyperdiploid cases (HeH_9 and HeH_10) have lost some topologically associating domains. **b** A/B compartment profile of chromosome 3 in cell line GM12878 and the six leukemia samples at 500 kb resolution. The profiles were similar between the cell line and the leukemias

first with microarrays and more recently with RNA-seq[3,6,7], the protein levels are expected to have a more direct impact on the phenotype and our proteome analysis thus provides an improved insight into leukemogenesis.

Previous studies of RNA expression in high hyperdiploid ALL have shown clear dosage effects[3,6,7], i.e. a general upregulation of genes on the gained chromosomes, corresponding to *cis* effects. However, no data on the effect on protein expression or in-depth analyses of *cis* and *trans* effects in relation to the gained chromosomes have been published to date. Here, we show that the gained chromosomes in high hyperdiploid ALL are also associated with *cis* effects at the protein level, but that those effects are weaker than at the RNA level. In addition, we also identified a general dysregulation of gene expression in relation to TAD boundaries in high hyperdiploid ALL. This corresponds to a likely disturbance in the insulation between regulatory elements and gene promoters that should be separated by a TAD boundary. Depleting CTCF experimentally has been associated with loss of insulation at TAD borders in a dose-dependent manner, whereas cohesin loss has been linked to lower insulation between TADs as well as depletion of TADs[37,42,43]. Thus, it is feasible that the relatively low expression of CTCF and cohesin that is seen in high hyperdiploid ALL could affect TAD border insulation and/or TAD structure. In line with this, our Hi-C analysis revealed that

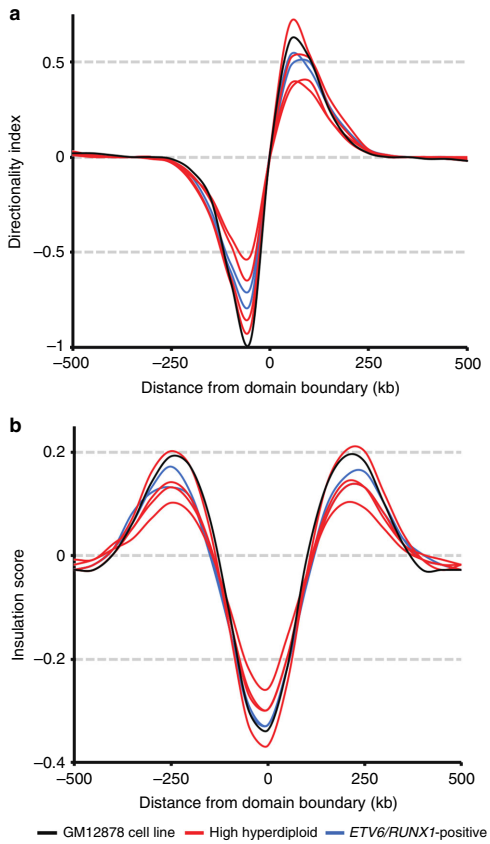


Fig. 6 Weaker topologically associating domain (TAD) boundaries in high hyperdiploid samples. **a** Median standardized directionality index profiles around TAD boundaries identified in high hyperdiploid cases (red), *ETV6/RUNX1*-positive cases (blue) and the GM12878 cell line (black). Two high hyperdiploid cases (HeH_9 and HeH_10) showed markedly decreased boundary strength, indicating permissive TAD boundaries. **b** Median insulation score around the TAD boundaries identified in high hyperdiploid cases (red), *ETV6/RUNX1*-positive cases (blue) and the GM12878 cell line (black). Three high hyperdiploid cases (HeH_9, HeH_10, and HeH_48) showed decreased insulation signal amplitude suggesting weaker insulation between TADs compared to the remaining samples

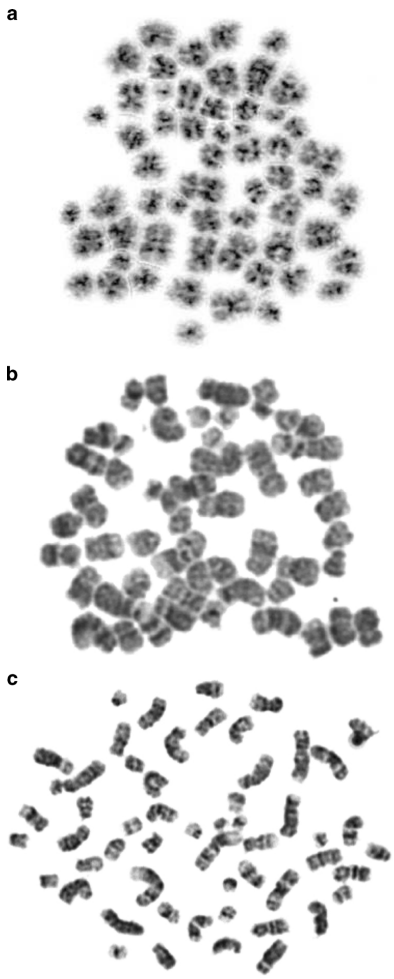


Fig. 7 Metaphase chromosome morphology in hyperdiploid leukemia. High hyperdiploid childhood acute lymphoblastic leukemia samples displayed varying metaphase chromosome morphology, but the majority of cases had poor morphology. **a** Example of metaphase with score 1—poor morphology (case HeH_33). **b** Example of metaphase with score 2—fair morphology (case HeH_48). **c** Example of metaphase with score 3—good morphology (case HeH_48)

three of four investigated high hyperdiploid samples displayed weaker TAD boundaries than the remaining leukemias and control cell lines, although the relatively small number of samples investigated prevents definite conclusions. Additionally, a higher number of TAD boundaries were recurrently lost in high hyperdiploid cases compared with the *ETV6/RUNX1*-positive cases and the cell lines. Furthermore, we found that high hyperdiploid ALL display an aberrant chromosome metaphase morphology, also suggesting an aberrant chromatin architecture. The underlying cause of the low CTCF and cohesin expression in high hyperdiploid ALL is currently unknown but mutations are unlikely to be the general cause; although these do occur, the frequency is relatively low[3]. *CTCF* is encoded on chromosome 16, which is rarely gained[2], whereas the core members of the cohesin complex are all encoded on commonly gained chromosomes (8q24 for *RAD21*, 10q25 for *SMC3*, Xp11 for *SMC1A* and Xq25 for *STAG2*); thus, the specific aneuploidy in high hyperdiploid ALL may cause relatively low expression of CTCF but not of cohesin.

Taken together, we show that the chromosomal gains in high hyperdiploid ALL are associated with genome-wide effects on

transcription. Furthermore, we show genome-wide transcriptional dysregulation with putative leukemogenic effects in relation to TAD borders in the hyperdiploid subtype suggestive of changes in chromatin architecture; such changes could also be seen by Hi-C in a subset of cases. Whether aberrant chromatin architecture is a common phenomenon in aneuploid tumors remains an open question.

Methods

Patients. The study comprised a total of 48 high hyperdiploid and 41 *ETV6/RUNX1*-positive pediatric BCP-ALL cases that had been treated at Skåne University Hospital, Lund, Sweden, selected based on samples being available from diagnosis (Supplementary Data 1). The cohort included 52 boys and 37 girls, with a median age at diagnosis of 4 years (range 0–16) and a median white blood cell count of 7.8×10^9/l (range 0.9–164). All cases had been tested for *BCR/ABL1*, *ETV6/RUNX1*, *PBX1/TCF3* and *KMT2A* (previously *MLL*) rearrangements by reverse-transcriptase PCR, fluorescence in situ hybridization, or Southern blot as part of the clinical analyses and were found to be negative for these fusion genes, with the exception of the *ETV6/RUNX1* fusion in that subgroup. Informed consent was obtained according to the Declaration of Helsinki and the study was approved by the Ethics Committee of Lund University.

DNA and RNA extraction. Details for samples subjected to oligo(dT)-based RNA-seq have been published elsewhere[19]. For samples subjected to mass spectrometry analysis, DNA, total RNA and proteins were extracted from bone marrow or peripheral blood samples obtained at diagnosis and stored in TRIzol (Thermo-Fisher Scientific, Waltham, MA) at −80 °C for 4–17 years. After addition of chloroform, RNA and DNA were precipitated according to the manufacturer's instructions. The remaining fractions containing proteins were stored at −80 °C.

Sample preparation for mass spectrometry. The stored TRIzol fractions were thawed and loaded into Slide-A-lyzer cassettes (ThermoFisher Scientific, 3.5-kDa cut-off, cat no 87722) and dialyzed according to the manufacturer's instructions against an aqueous solution containing 0.25% SDS and 25 mM Hepes pH 7.6 overnight at 4 °C with two changes of the dialysis buffer. The dialyzed samples were digested by a modified FASP-protocol[5]. Protein concentrations were estimated by gel staining and approximately 250 µg of each sample was mixed with 1 mM DTT, 8 M urea, 25 mM HEPES, pH 7.6 and transferred to a 10-kDa cut-off centrifugation filtering unit (Pall, Nanosep®, Merck, Darmstadt, Germany), and centrifuged at 14,000 × *g* for 15 min. Proteins were alkylated by 50 mM iodoacetamide (IAA) in 8 M urea, 25 mM HEPES for 10 min. The proteins were then centrifuged at 14,000 × *g* for 15 min followed by two more additions and centrifugations with 8 M urea, 25 mM HEPES. Proteins were digested at 37 °C with gentle shaking overnight by addition of Lys-C (enzyme:protein = 1:50, Wako Pure Chemical Industries, Ltd.) in 500 mM Urea, 50 mM HEPES pH 7.6 followed by an additional overnight digestion with trypsin (enzyme:protein = 1:50, ThermoFisher Scientific) in 50 mM HEPES, pH 7.6. The filter units were centrifuged at 14,000 × *g* for 15 min followed by another centrifugation with MilliQ water and the flow-through was collected. Peptides were cleaned up by a modified sp3-protocol[44]. Briefly, carboxylate-modified paramagnetic beads (ThermoFisher Scientific; CAT No. 09-981-121 and ThermoFisher Scientific; CAT No. 09-981-123) were mixed 1:1 and washed with MilliQ water. 10 µl of the bead-mixture was added to each peptide sample. Acetonitrile was added so that the final concentration was >95% and beads were incubated at room temperature for 8 minutes. Next, beads were placed on a magnetic rack, the supernatant discarded and the beads washed twice with 180 µl of acetonitrile. Beads were re-suspended in 100 µl of MilliQ water and sonicated to release the peptides, supernatants were collected and stored at −20 °C. Peptide concentration was determined by the Bio-Rad DCC assay and 30 µg of peptides from each digested sample was labeled with TMT 10-plex reagent according to the manufacturer's protocol (ThermoFisher Scientific). A small portion of unlabeled peptides were pooled from all samples to generate an internal standard that was labeled with TMT-channel 131 and included in all sets. Labeled samples were pooled, cleaned by strata-X-C-cartridges (Phenomenex, Torrance, CA) and dried in a Speed-Vac.

Peptide level sample fractionation through HiRIEF. The TMT labeled peptides, 300 µg, were separated by immobilized pH gradient - isoelectric focusing (IPG-IEF) on pH 3–10 strips using the HiRIEF method[45]. Peptides were extracted from the strips by a prototype liquid handling robot, supplied by GE Healthcare Bio-Sciences AB. A plastic device with 72 wells was put onto each strip and 50 µl of MilliQ water was added to each well. After 30 minutes incubation, the liquid was transferred to a 96 well plate and the extraction was repeated two more times with 35% acetonitrile (ACN) and 35% ACN, 0.1% formic acid in MilliQ water, respectively. The extracted peptides were dried in Speed-Vac and dissolved in 3% ACN, 0.1 % formic acid.

Mass spectrometry based quantitative proteomics. Extracted peptide fractions were separated using an Ultimate 3000 RSLCnano system coupled to a Q Exactive (ThermoFisher Scientific). Samples were trapped on an Acclaim PepMap nanotrap column (C18, 3 µm, 100 Å, 75 µm × 20 mm, ThermoFisher Scientific), and separated on an Acclaim PepMap RSLC column (C18, 2 µm, 100 Å, 75 µm x 50 cm, ThermoFisher Scientific). Peptides were separated using a gradient of mobile phase A (5% DMSO, 0.1% FA) and B (90% ACN, 5% DMSO, 0.1% FA), ranging from 6 to 37 % B in 60 min (depending on IPG-IEF fraction complexity) with a flow of 0.25 µl/min. The Q Exactive was operated in a data-dependent manner, selecting top 10 precursors for fragmentation by HCD. The survey scan was performed at 70,000 resolution from 400–1600 m/z, with a max injection time of 100 ms and target of 1×10^6 ions. For generation of HCD fragmentation spectra, a max ion injection time of 140 ms and AGC of 1×10^5 were used before fragmentation at 30% normalized collision energy, 35,000 resolution. Precursors were isolated with a width of 2 m/z and put on the exclusion list for 70 s. Single and unassigned charge states were rejected from precursor selection.

Peptide and protein identification. Orbitrap raw MS/MS files were converted to mzML format using msConvert from the ProteoWizard tool suite[45]. Spectra were then searched using MSGF+ (v10072)[46] and Percolator (v2.08)[47], where search results from eight subsequent fraction were grouped for Percolator target/decoy analysis. All searches were done against the human protein subset of Ensembl 75 in the Galaxy platform. MSGF + settings included precursor mass tolerance of 10 ppm, fully-tryptic peptides, maximum peptide length of 50 amino acids and a maximum charge of 6. Fixed modifications were TMT-10plex on lysines and peptide N-termini, and carbamidomethylation on cysteine residues, a variable modification was used for oxidation on methionine residues. Quantification of TMT-10plex reporter ions was done using OpenMS project's IsobaricAnalyzer (v2.0)[48]. Peptide-spectrum matches (PSMs) found at 1% false discovery rate (FDR) were used to infer gene identities. Protein quantification by TMT 10-plex reporter ions was calculated using TMT PSM ratios to the entire sample set (all 10 TMT-channels) and normalized to the sample median. The median PSM TMT reporter ratio from peptides unique to a gene symbol was used for quantification. Protein false discovery rates were calculated using the picked-FDR method using gene symbols as protein groups and limited to 1% FDR[49].

DNA sequencing analyses. WGS results for cases HeH_2, HeH_3, HeH_12, HeH_13, HeH_17, HeH_22, HeH_24, HeH_25, HeH_27, HeH_32, HeH_35, HeH_38, and HeH_43 and WES results for cases HeH_1, HeH_4, HeH_5, HeH_7, HeH_8, HeH_16, HeH_19, HeH_26, HeH_34, HeH_37, and HeH_41 have been previously published[3]. Briefly, for WGS, matched diagnostic and remission bone marrow or peripheral blood samples were sequenced to ~100x coverage on the Complete Genomics platform. Somatic events were identified using the Complete Genomics Cancer Sequencing v2.0 pipeline with CGA tools. For WES, libraries were constructed using the SureSelectXT2 Human All Exon V4 kit (Agilent Technologies, Santa Clara, CA) from matched diagnostic and remission bone marrow or peripheral blood samples and paired-end sequencing were done to ~120x coverage on an Illumina HiSeq2000. Somatic mutations were detected with MuTect[50]. WES for cases HeH_6, HeH_9-HeH_11, HeH_14, HeH_15, HeH_18, and ETV6/RUNX1_1-ETV6/RUNX1_9 were done by Nextera Rapid Capture Expanded Exome Kit (Illumina, San Diego, CA, USA), with ~110x coverage on an Illumina NextSeq 500. Paired remission samples were available from cases HeH_6, HeH_10, HeH15, ETV6/RUNX1_1-ETV6/RUNX1_5, and ETV6/RUNX7_1-ETV6/RUNX1_9. Pair-end sequence reads were aligned to the human_g1k_v37 by Burrows-Wheeler Aligner (BWA)[51]. Duplicate reads were marked with Picard and Indel realignment was performed with GATK[52]. Somatic mutations were identified using MuTect[50] and MuSE[53] whereas somatic indels were identified by manta[54] and strelka[55] with default settings. Mutations that passed the internal filters of the variation caller were further filtered by a minimum depth of 10 reads. For tumor samples without matched normal (HeH_9, HeH_11, HeH_14, HeH_18 and ETV6/RUNX1_6), variations were identified by GATK UnifiedGenotyper and annotation parameters QD (variant confidence/quality by depth) <2.0, MQ (RMS mapping quality) <40.0, FS (Fisher strand) 60.0, HaploTypeScore >13.0, MQRankSum < −12.5 and ReadPosRankSum <−8.0 were used to filter low quality variations. High-quality variants were further filtered by 1000 Genomes (20110521 release), ESP6500, ExAC, CG46 (popfreq_max_20150413) and 170 million variants (kaviar_20150923) provided by ANNOVAR[56] to remove potential SNP sites. Functional annotation was performed by ANNOVAR.

SNP array analyses. SNP array analysis was done on DNA extracted from diagnostic bone marrow or peripheral blood samples on the Human1M-Duo, Human-Omni1-Quad, Human-Omni5-4v (Illumina, San Diego, CA), or CytoScan HD platforms (Applied Biosystems, Thermo Fisher); data have been published previously[57].

RNA-sequencing. Details on the oligo(dT) RNA-seq dataset have been previously published[19]. Briefly, cDNA sequencing libraries were constructed from poly-A-selected RNA using the Truseq RNA library preparation kit v2 (Illumina, San Diego, CA) and sequenced on an Illumina HiScan SQ or an Illumina NextSeq 500.

For the RiboZero RNA-seq, RNA from cases HeH_1-HeH_18 and ETV6/RUNX1_1-ETV6/RUNX1_9 were constructed using the Human Ribo-Zero rRNA Removal Kit (Illumina, San Diego, CA) and sequenced on an Illumina NextSeq 500.

Expression analyses. RNA sequencing data were processed using the TCGA mRNA-seq pipeline (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#mrna-analysis-pipeline). Briefly, sequencing reads were aligned to the human GRCh38 genome assembly using STAR[58] and read counts for each gene were obtained by HTSeq-count[59].

Genes with count-per-million (CPM) value greater than 1 were defined as expressed genes and only genes expressed in more than 80% of samples in at least one sample group were used for further analyses. For differential expression analysis, batch effects were adjusted by using RUVg function in RUVSeq[60] and differentially expressed genes were identified by using edgeR[61]. Benjamini–Hochberg adjusted (BH-adjusted) $P \leq 0.05$ were used as cutoff. In the oligo(dT) dataset, six samples representing sorted pro-B (CD34+CD38+CD19+CD10+) or pre-B (CD34−CD38+CD19+CD20−CD10+) cells were included and used to compare leukemic samples to the closest normal cell.

For the proteomics dataset, absolute intensity values of the PSMs were converted to ratios based on the pool reference and log2 transformed. Spectra mapping to unique gene symbols were retained and aggregated to proteins using the median value of the PSM ratios. Proteins identified in all sets were used for subsequent analysis. To remove the batch effects of proteomics data, iterative RUV4 algorithm was used[62]. In brief, we initialized the search with the median normalized dataset to detect the proteins with BH-adjusted P greater than 0.9 as the first controls by using limma[63]. In each of the 10 iterations, we applied the RUV4 algorithm on the un-normalized data to obtain the residuals and calculate ab initio the control proteins as those with differential abundance of BH-adjusted P greater than 0.9. These proteins were used as control proteins in the next iteration. Control proteins identified in the last iteration constituted the empirical controls which were uncorrelated to the tumor category. Protein differential expression analysis was performed by limma and BH-adjusted $P \leq 0.05$ was applied as cutoff for identifying differentially expressed proteins.

Two gene expression datasets obtained from NCBI Gene Expression Omnibus (accession numbers GSE13351 and GSE13425) were analyzed. The expression data were analyzed using Transcriptome Analysis Console (Affymetrix, Santa Clara, CA, USA) with default settings.

Correlation between mRNA and protein variation. To compare mRNA and protein variations across samples, we focused on 8222 genes/proteins that were detected in both the Ribozero RNA-seq dataset and the proteomics dataset. We first calculated the Spearman's correlation coefficient between RNA-seq FPKM values and RUV4-normalized values from the proteomics dataset across samples ($n = 27$) and P-values corresponding to the coefficients were computed and adjusted by Benjamini–Hochberg procedure. Significant calls were made based on BH-adjusted $P \leq 0.05$. Functional enrichment analysis was performed by GSEA-Pre-ranked algorithm[64] and Spearman's correlation coefficient was used as the ranking variable. Pairwise correlation analysis of protein pairs, which are present within the same complex of known protein complexes acquired from the CORUM database[65], were performed by using Spearman correlation's coefficient and the same analysis was performed on the RiboZero RNA-seq dataset. For the mRNA-protein stability analysis mRNA and protein half-lives from mouse fibroblast cell lines were extracted from Schwanhäusser et al.[13] and analyzed as per the original manuscript and Zhang et al.[9], i.e., stable (unstable) mRNAs and proteins were categorized according to their rank in the top (bottom) one third of half-lives, respectively. For micro-RNA targeting analysis miRNA-mRNA interactome data was downloaded from Helwak et al.[14]. For the analysis of the impact of ubiquitination and proteasomal degradation time-series ubiquitination data from HCT116 and 293T human cell lines upon bortezomib treatment was used[15]. We chose the 8-hour point as a proxy for steady-state and divided significantly/non-significantly ubiquitinated proteins according to an absolute log2 fold change greater/smaller than 1. We also investigated the impact of protein degradation profiles, namely exponential (ED) and non-exponential (NED) decay, using mouse fibroblast cell data from a click-chemistry assisted pulsed SILAC study[16]. Information on protein subcellular localization was downloaded from a MS-based study on global subcellular localization[17] and used to subset our dataset. For the analysis of the impact of differential expression/abundance on mRNA-protein correlations transcripts/proteins were divided to significant/non-significant based on BH adjusted P-value <0.05 and log2 fold changes higher (lower) than the 90% (10%) percentile. To avoid correlations being driven by tumor subtype differences (Simpson's paradox), partial correlations were estimated after regressing the data on tumor subtype and calculating Spearman correlations on the residuals. For the above analyses, we used overlapped gene symbols between datasets. Two-group and multi-group comparisons were assessed with two-sided Wilcoxon rank sum test and Kruskal–Wallis test, respectively.

Analysis of *cis*- and *trans*-effects. Matched copy number aberrations (CNA) based on SNP-array analysis, WES and/or WGS, proteomics and RNA-seq measurements of 18 high hyperdiploid samples were used to study the impact of CNAs on mRNA and protein expression. In order to avoid outlier-driven results, only genes displaying CNAs involving more than 3 cases in each comparison group were retained (CNA genes, $n = 2080$). To analyze genome-wide *cis* effects of high hyperdiploid ALL samples, Spearman's correlation coefficient between genes/proteins abundance and copy number of 2080 informative CNA genes was calculated, respectively. To analyze genome-wide *trans* effects, the correlation between CNA genes and all 8222 mRNA and proteins detected in both RNA-seq and proteomics of 18 high hyperdiploid samples were determined using the MatrixeQTL R package[18]. Subsequently, P-values corresponding to the coefficient were calculated and significant CNA-mRNA and CNA-protein correlations were identified using BH-adjusted P-value 0.05 as cutoff.

To assess the impact of copy number aberrations on the hyperdiploid vs. *ETV6/RUNX1*-positive ALL differential expression, a normalized copy number per chromosome across the samples was calculated according to the formula:

$$\begin{aligned} &\textit{Normalized copy number per chromosome} \\ &= \textit{Average}_{samples}\left(\frac{\sum_{i=1}^{\#\ chromosomal\ segments} copy\ number\ (i) * length\ (i)\ in\ bp}{total\ length\ of\ chromosome} \Big/ Normal\ ploidy\ (for\ somatic\ chromosomes{-}2,\ for\ sex\ chromosomes{-}1,2,NA)\right) \end{aligned} \quad (1)$$

Cohen's d effect size was calculated per chromosome at the mRNA and protein level and linearly regressed on the normalized copy number. We denoted significantly affected chromosomes as those with effect size >0.3.

To estimate the differential expression of known cancer driver genes[66] on the mRNA and protein level based on edgeR and limma, respectively, fold changes were overlaid on the ALL copy number landscape and gene symbols with BH-adjusted P-values ≤0.05 and fold change greater (lower) than the 90th (10th) percentile were displayed.

Gene set enrichment analysis. Gene set enrichment analysis[64] was done using the GSEA-pre-ranked algorithm with lists of all expressed genes ($n = 12{,}313$ for Ribozero and $n = 13{,}951$ for oligo(dT), respectively) and all expressed proteins ($n = 8480$), by using predictive log fold changes between high hyperdiploid ALL and *ETV6/RUNX1*-positive cases generated by edgeR (RiboZero and oligo(dT) RNA-seq) and log-fold changes values generated by limma (LC-MS/MS) as the ranking variable, respectively. We performed the analysis using the GSEA standalone software with default settings. Family-wise error rate (FWER) $P < 0.05$ was considered significant.

CTCF binding site and ChIA-PET data analysis. CTCF binding sites analysis was done according to Aitken et al.[29]. We downloaded the positions of CTCF binding sites from the ENCODE database (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/wgEncodeAwgTfbsBroadGm12878CtcfUniPk.narrowPeak.gz) and the number of CTCF binding sites in each gene (plus 5 kb on either side) were obtained by using BEDTools (https://bedtools.readthedocs.io/en/latest/) command intersect. The proportion difference between the differentially expressed genes group and the remaining genes group was tested with the chi-square test and the difference between fold changes according to number of CTCF binding sites was tested by Mann–Whitney U-test. ChIA-PET data for CTCF and RAD21 were downloaded from the NCBI GEO database under accession numbers GSM1872886 and GSM1436265, respectively. To get high-confidence chromatin interactions, ChIA-PET interactions with low PET-count (less than ten reads coverage) were removed. BEDTools was used to find the overlapping interactions between CTCF and RAD21 ChIA-PET datasets and only interactions detected in both datasets were used. BEDTools command closest was used to determine the distance between the transcription start sites of expressed genes and CTCF/cohesin anchors. Genes located within 5 kb of a CTCF/cohesin anchor were defined as anchor genes. Statistical differences in the proportion of differentially expressed genes between the anchor genes group and the background group were tested by hypergeometric test and the differences between fold change values were tested with the Mann–Whitney U-test.

Gene pair correlation analysis. Gene pair correlation analysis was done according to Flavahan et al.[34]. Briefly, TADs of the IMR90 and GM12878 cell lines were downloaded from published Hi-C data[31] (Gene Expression Omnibus accession number GSE63525) and genes were assigned to the inner-most domain in which the transcription start site of the canonical transcript fell within. Genes were assigned to the same domain if they were assigned to the same domain in both GM12878 and IMR90 datasets. Ten thousand randomly generated domains were obtained by using BEDTools command random with 1 Mb as interval size. Spearman's correlation coefficient for all relevant gene pairs within the same TAD, different TADs and randomly generated domains were calculated and the correlation plot was smoothed by locally weighted scatterplot smoothing with weighted linear least squares (LOESS).

Hi-C library preparation and sequencing. Hi-C was done on cases HeH_9, HeH-10, HeH_42, HeH_48, ETV6/RUNX1_6, and ETV6/RUNX1_41, selected on the

basis of sample availability. Cell pellets approximately 5 mm in size containing mononuclear bone marrow or peripheral blood cells obtained at leukemia diagnosis were resuspended in 10 ml room temperature 1× PBS. The cells were fixed by the addition of 37% formaldehyde to a final concentration of 2% and gentle mixing on a rocker for 10 min at room temperature. The reaction was quenched by the addition of 1.5 ml cold glycine (0.125 M). Following incubation for 5 min at room temperature and 15 min on ice, the cells were pelleted by centrifugation at 400×*g* for 10 min at 4 °C. The pellet was resuspended in 1 ml cold 1× PBS by pipetting and made up to a final volume of 10 ml with cold 1× PBS. Finally, the cells were pelleted by centrifugation at 400×*g* for 10 min at 4 °C, and the pellet snap-frozen and stored at −80 °C until further analysis. In nucleus Hi-C on the crosslinked mononuclear cells was performed as outlined in Nagano et al.[67]. Each sample was split into two and processed separately to provide a technical replicate. One lane of 150 base pair paired-end sequencing was performed on the Illumina HiSeq 4000 instrument per replicate (12 lanes in all).

Hi-C data analysis. The sequences from the Hi-C libraries were mapped to reference human_g1k_v37 using HICUP with default settings[35]. Redundant reads and short-range Hi-C artifacts were removed from all downstream analyses. Filtered read pairs were then aggregated into 25, 50 , and 100 kb genomic bins to generate Hi-C contact matrices. Low-coverage bins were filtered by using maximum allowed median absolute deviation (MAD-max) and low-coverage bins with MAD-max values higher than 2 were removed. Reads mapped to the same bin or adjacent bins were also removed. For the GM12878 cell line datasets, we downloaded.hic data from the NCBI GEO database (accession number GSE63525, GM12878_insitu_primary+replicate_combined_30.hic.gz) and converted.hic format into 25, 50, and 100 kb contact matrices by using the dump option of Juicertools[36]. The same filtering strategy was applied to the GM12878 cell line dataset. The filtered contact matrices were then normalized using the chromosome-adjusted iterative correction procedure (caICB) to eliminate copy number bias[68].

TAD and boundaries calling. The normalized 25 kb contact matrices of six leukemia cases and GM12878 cell line were used to predict TAD structures by DomainCaller[39], as this showed the best agreement with manual annotation in a previous study[69]. To find the optimal threshold for TAD calling, the window size parameter of DomainCaller was varied from 250 kb to 2 Mb and finally 500 kb window size was used, which showed about 80% of detected boundaries co-aligned with the previously detected boundaries of the GM12878 cell line[31] (accession number GSE63525, GM12878_primary+replicate_Arrowhead_domainlist.txt.gz) with ±100 kb precision. Standardized genome-wide directionality index value (z-score value) was used for TAD analysis. InsulationScore package was also used to identify TAD boundaries[38]. For insulation boundaries analysis, insulation square was set to 250 kb and insulation delta span was set to 125 kb. Insulation score was calculated for each chromosome and then normalized by the genome-wide median.

TAD analysis. When comparing TAD boundaries between high hyperdiploid ALL and *ETV6/RUNX1*-positive cases, boundaries detected in different samples within ±100 kb were called overlapped boundaries while the recurrent boundaries found only in one subgroup of leukemia samples were defined as subgroup-specific boundaries. To investigate loss of TAD boundaries in high hyperdiploid ALL, the multiinter command in BEDTools software was used to identify the overlapped boundaries and subgroup-specific boundaries. To further correlate the presence of boundaries in the different subgroups of leukemia, the subgroup-specific boundaries were manually traced. Briefly, balanced-corrected Hi-C matrices were plotted using Juicebox[36] and subgroup-specific boundaries were stratified into three categories, (i) boundaries showing sharp visual contrast between within and across TAD interaction frequencies were classified as strong boundaries, (ii) boundaries showing little visual contrast were classified as weak boundaries and (iii) boundaries that totally disappeared in one of subgroup of leukemia cases were classified as lost boundaries. To detect CTCF and RAD21 binding sites occupancy over the subgroup-specific boundaries, peak files were downloaded from the UCSC database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/) and the intersectBed command in BEDTools software was used to identify all CTCF/RAD21 peaks located within a ±50 kb window around the boundary.

Chromosome morphology analysis. Bone marrow or peripheral blood preparations and G-banding was performed according to standard methods from cells obtained at diagnosis and stored in fixative (methanol:acetic acid; 3:1) at −20 °C. The slides were analyzed using an Eclipse 80i microscope (Nikon, Tokyo, Japan) equipped with a progressive scan camera (JAI, Copenhagen, Denmark) and a ×100 oil immersion Plan Apo VC lens (Nikon). Metaphases were captured, edited and karyotyped with the CytoVision software (Leica Biosystems, Wetzlar, Germany). Each metaphase was scored from 1–3 according to chromosome morphology, as judged by the level of chromosome condensation, the band resolution, overall chromosome shape and clearness, and how easily the chromosome pairs could be identified.

The criteria used for scoring were: 1—poor chromosome morphology, where no substantial banding pattern could be observed, chromosomes were very condensed, chromosomes presented a fuzzy appearance, i.e., chromosome shape was poor, and homolog pairs were difficult to identify; 2—fair morphology, where band level was at 200–300, chromosomes were less constricted and presented a sharp appearance, which made it easier to karyotype; and 3—good morphology, where the band levels was at least 350–400, chromosomes were elongated and presented an ideal appearance for cytogenetic analysis (Fig. 7). The person doing the chromosome morphology investigation was blinded to the results from Hi-C analysis.

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

Data availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange[70] Consortium (http://proteomecentral.proteomexchange.org) with the dataset identifier PXD010175. RNA-seq data have been deposited to the European Genome-phenome Archive (EGA) under the accession code EGAS00001003079. The remaining data will be available for academic research on somatic variants only by contacting the authors. Publicly available data used in this study can be found as deposited in the following datasets: Oligo(dT) RNA-seq data for ALL patients, accession number EGAD00001002112. Expression data from ALL patients, accession numbers GSE13351 and GSE13425. RNA-seq dataset for AML, accession number TCGA-LAML. RNA-seq dataset for papillary renal cell carcinoma, accession number TCGA-KIRP. Hi-C datasets for GM12878 cell line and IMR90 cell line, accession number GSE63525. GM12878 CTCF ChIA-PET dataset, accession number GSM1872886. GM12878 RAD21 ChIA-PET dataset, accession number GSM1436265.

Received: 18 July 2018 Accepted: 11 March 2019
Published online: 03 April 2019

References

1. Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **10**, 478–487 (2009).
2. Paulsson, K. & Johansson, B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **48**, 637–660 (2009).
3. Paulsson, K. et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Genet.* **47**, 672–676 (2015).
4. Malinowska-Ozdowy, K. et al. *KRAS* and *CREBBP* mutations: a relapse-linked malicious liaison in childhood high hyperdiploid acute lymphoblastic leukemia. *Leukemia* **29**, 1656–1667 (2015).
5. Bateman, C. M. et al. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. *Leukemia* **29**, 58–65 (2015).
6. Andersson, A. et al. Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations. *Proc. Natl Acad. Sci. USA* **102**, 19069–19074 (2005).
7. Zalíova, M. et al. Slower early response to treatment and distinct expression profile of childhood high hyperdiploid acute lymphoblastic leukaemia with DNA index <1.16. *Genes Chromosomes Cancer* **55**, 727–737 (2016).
8. Branca, R. M. M. et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11**, 59–62 (2014).
9. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014).
10. Zhang, H. et al. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166**, 755–765 (2016).
11. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).
12. Goncalves, E. et al. Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. *Cell Syst.* **5**, 386–398 e384 (2017).
13. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
14. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153**, 654–665 (2013).
15. Kim, W. et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol. Cell* **44**, 325–340 (2011).
16. McShane, E. et al. Kinetic analysis of protein stability reveals age-dependent degradation. *Cell* **167**, 803–815 e821 (2016).
17. Orre, L. M. et al. SubCellBarcode: Proteome-wide mapping of protein localization and relocalization. *Mol. Cell* **73**, 166–182 (2019).
18. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

19. Lilljebjörn, H. et al. Identification of *ETV6-RUNX1*-like and *DUX4*-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 11790 (2016).
20. Zaliova, M. et al. *ETV6/RUNX1*-like acute lymphoblastic leukemia: A novel B-cell precursor leukemia subtype associated with the CD27/CD44 immunophenotype. *Genes Chromosomes Cancer* **56**, 608–616 (2017).
21. Armstrong, S. A. et al. FLT3 mutations in childhood acute lymphoblastic leukemia. *Blood* **103**, 3544–3546 (2004).
22. Stoskus, M. et al. Identification of characteristic *IGF2BP* expression patterns in distinct B-ALL entities. *Blood. Cells Mol. Dis.* **46**, 321–326 (2011).
23. Jeffries, S. J., Jones, L., Harrison, C. J. & Russell, L. J. IGH@ translocations co-exist with other primary rearrangements in B-cell precursor acute lymphoblastic leukemia. *Haematologica* **99**, 1334–1342 (2014).
24. Neveu, B. et al. *CLIC5*: a novel ETV6 target gene in childhood acute lymphoblastic leukemia. *Haematologica* **101**, 1534–1543 (2016).
25. Papaemmanuil, E. et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6-RUNX1* acute lymphoblastic leukemia. *Nat. Genet.* **46**, 116–125 (2014).
26. Santaguida, S. & Amon, A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell Biol.* **16**, 473–485 (2015).
27. Ding, L. W. et al. Mutational landscape of pediatric acute lymphoblastic leukemia. *Cancer Res.* **77**, 390–400 (2017).
28. Merkenschlager, M. & Nora, E. P. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genom. Hum. Genet.* **17**, 17–43 (2016).
29. Aitken, S. J. et al. CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biol.* **19**, 106 (2018).
30. Tang, Z. et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–1627 (2015).
31. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
32. Ley, T. J. et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
33. Linehan, W. M. et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).
34. Flavahan, W. A. et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
35. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data [version 1; referees: 2 approved, 1 approved with reservations]. *F1000Res.* **4**, 1310 (2015).
36. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
37. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 e922 (2017).
38. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
39. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
40. Bernardi, G. Chromosome architecture and genome organization. *PLoS One* **10**, e0143739 (2015).
41. Stewart, E. et al. Identification of therapeutic targets in rhabdomyosarcoma through integrated genomic, epigenomic, and proteomic analyses. *Cancer Cell.* **34**, 411–426 e419 (2018).
42. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320 e324 (2017).
43. Zuin, J. et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl Acad. Sci. USA* **111**, 996–1001 (2014).
44. Hughes, C. S. et al. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **10**, 757 (2014).
45. Holman, J. D., Tabb, D. L. & Mallick, P. Employing ProteoWizard to convert raw mass spectrometry data. *Curr. Protoc. Bioinforma.* **46**, 11–19 (2014).
46. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
47. Granholm, V. et al. Fast and accurate database searches with MS-GF+ Percolator. *J. Proteome Res.* **13**, 890–897 (2014).
48. Sturm, M. et al. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinforma.* **9**, 163 (2008).
49. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteom.* **14**, 2394–2404 (2015).
50. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
53. Fan, Y. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
54. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
55. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
56. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
57. Olsson, L. et al. Improved cytogenetic characterization and risk stratification of pediatric acute lymphoblastic leuekmia using single nucleotide polymorphism array analysis: a single center experience of 296 cases. *Genes Chromosomes Cancer* **57**, 604–607 (2018).
58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Anders, S., Pyl, P. T. & Huber, W. HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
60. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
61. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
62. Gagnon-Bartsch J. A., Jacob L. & Speed T. P. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech. Reports from Dep. Stat. Univ. California*, 1–112 (2013).
63. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
64. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
65. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes - 2009. *Nucleic Acids Res.* **38**, D497–D501 (2010).
66. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
67. Nagano, T. et al. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.* **16**, 175 (2015).
68. Wu, H.-J. & Michor, F. A computational strategy to adjust for copy number in tumor Hi-C data. *Bioinformatics* **32**, 3695–3701 (2016).
69. Dali, R. & Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.* **45**, 2994–3005 (2017).
70. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106 (2017).

Acknowledgements

This study was supported by grants from the Swedish Cancer Society (K.P., grant references CAN 2014/1258 and CAN 2016/497), the Swedish Childhood Cancer Foundation (M.Y., grant reference TJ2016-0063; M.V. grant reference TJ2014-0063; R.J., grant references TJ2016-0035 and PR2016-0019; J.L., grant reference PR2016-0059; K.P., grant reference PR2015-0012), the Swedish Research Council (R.J., grant reference 2017-01653; J.L., grant reference 2015-04622; K.P., grant reference 2016-01459), the Royal Physiographic Society of Lund (M.Y.), Felix Mindus Contribution to Leukemia research (M.V. and R.J.), Dr Åke Olsson Foundation for Hematological Research (R.J., grant reference 2017-00437), Governmental Funding of Clinical Research within the National Health Service (K.P., grant reference ALFSKANE-623431), Cancer Research UK (L.H., D.T.O., grant reference 20412), the Wellcome Trust (L.H., D.T.O., grant reference 202878/A/16/Z) and the European Research Council (D.T.O. grant reference 615584).

Author contributions

M.Y., M.V., J.L. and K.P. conceived the study. M.Y. analyzed RNA-seq, proteome, WES, and Hi-C data. M.V. and R.J. performed the MS experiments. M.V. and I.S analyzed proteome and RNA-seq data. L.H.M.-C. performed metaphase chromosome experiments. A.C. provided clinical data and input. T.F. supervised RNA-seq. H.L. performed RNA-seq. D.T.O. supervised Hi-C experiments. L.O. performed SNP array analyses. N.R. and E.L.W. performed RNA-seq and WES. L.H. performed Hi-C experiments and analyzed data. J.L. and K.P analyzed data and supervised the study. M.Y., M.V., J.L. and K.P. wrote the manuscript with input from all authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09469-3.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

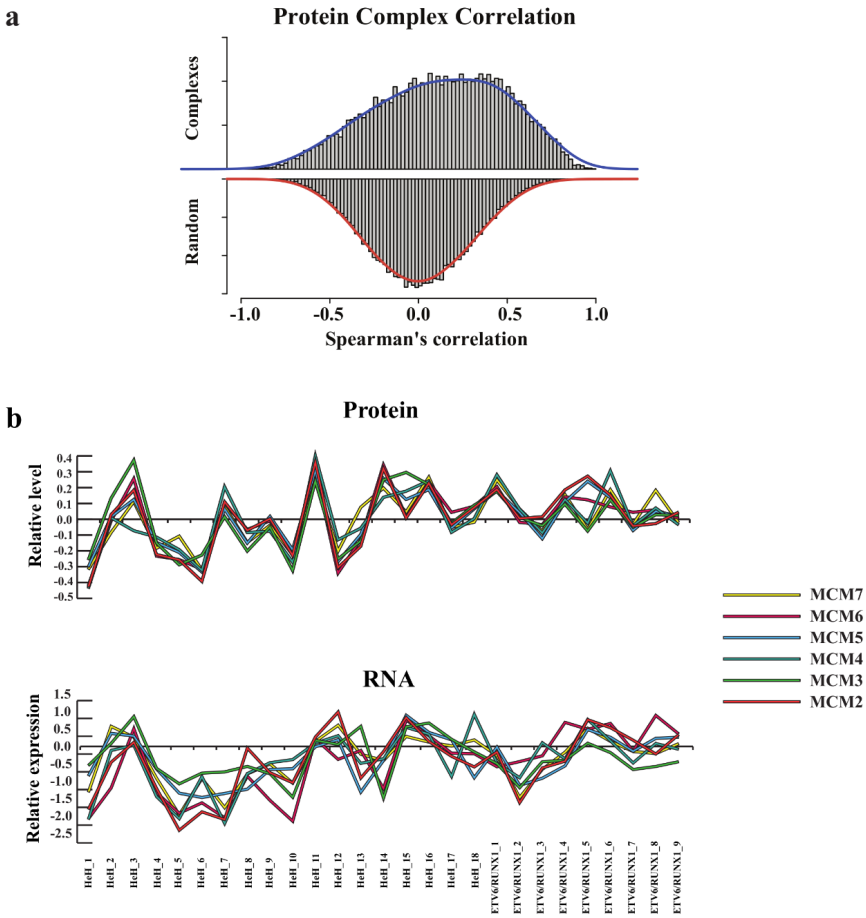
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

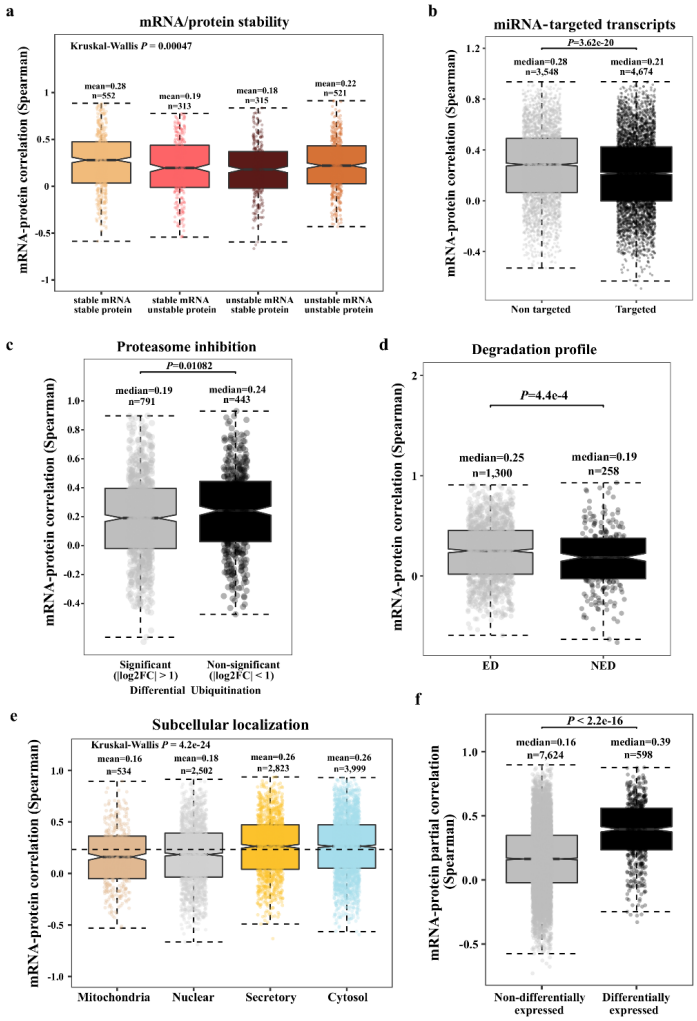
SUPPLEMENTARY INFORMATION FOR

Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia

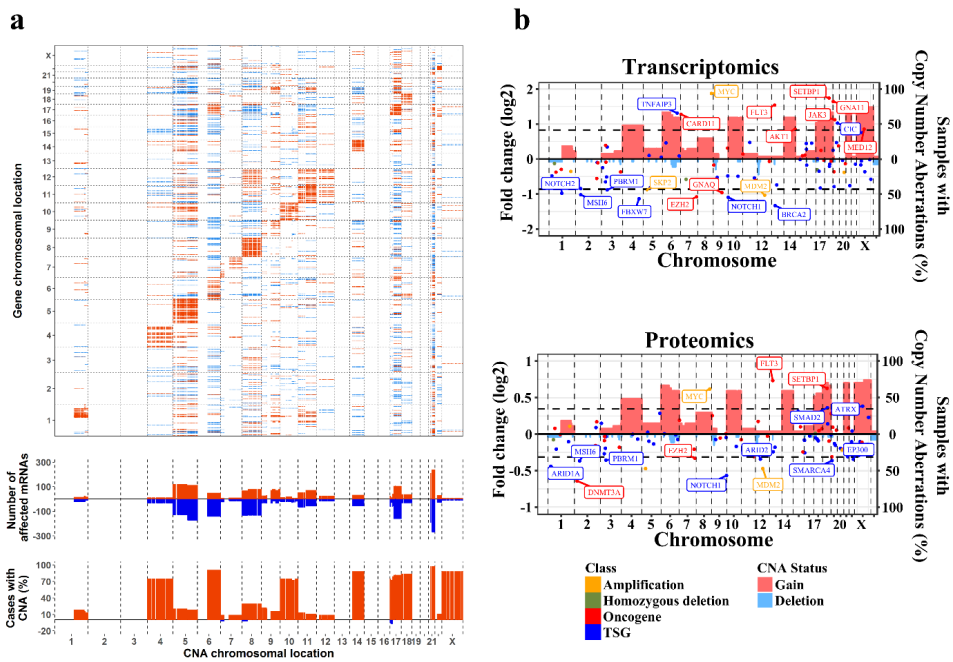
Yang et al.



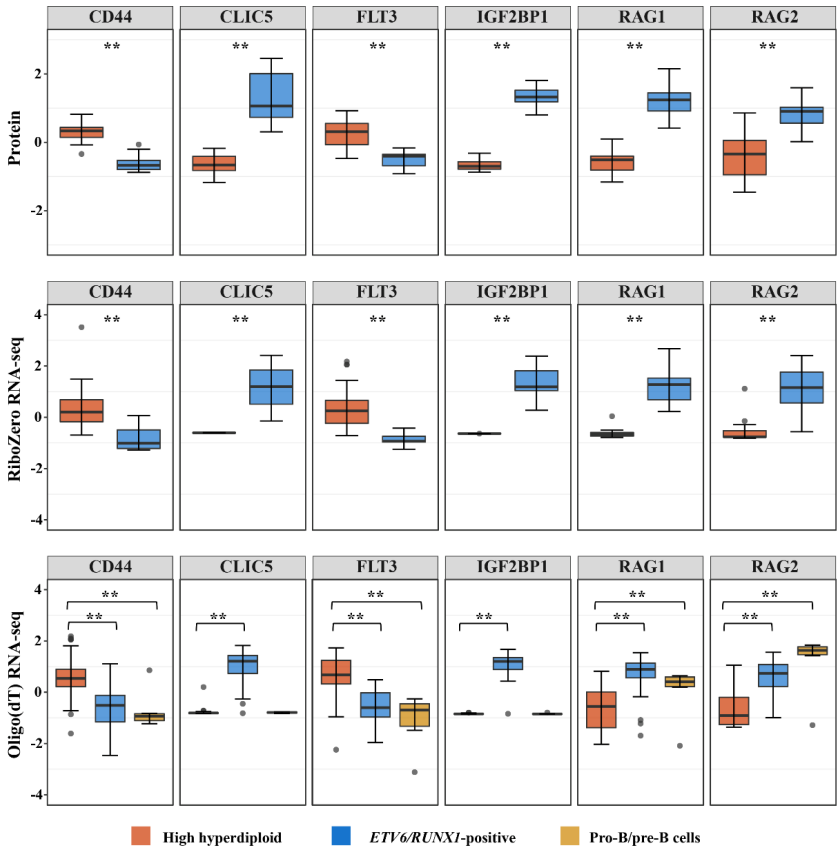
Supplementary Fig. 1. Analysis of protein complex formation a. Distribution of Spearman's correlations between protein-pairs known to form or partake in the same protein complex (CORUM) (top) compared to the distribution of Spearman's correlations of random protein pairs in the proteomics data. **b.** Example of relative levels and relative expression of members of the MCM complex in the proteomics and the RiboZero RNA-sequencing dataset, respectively.



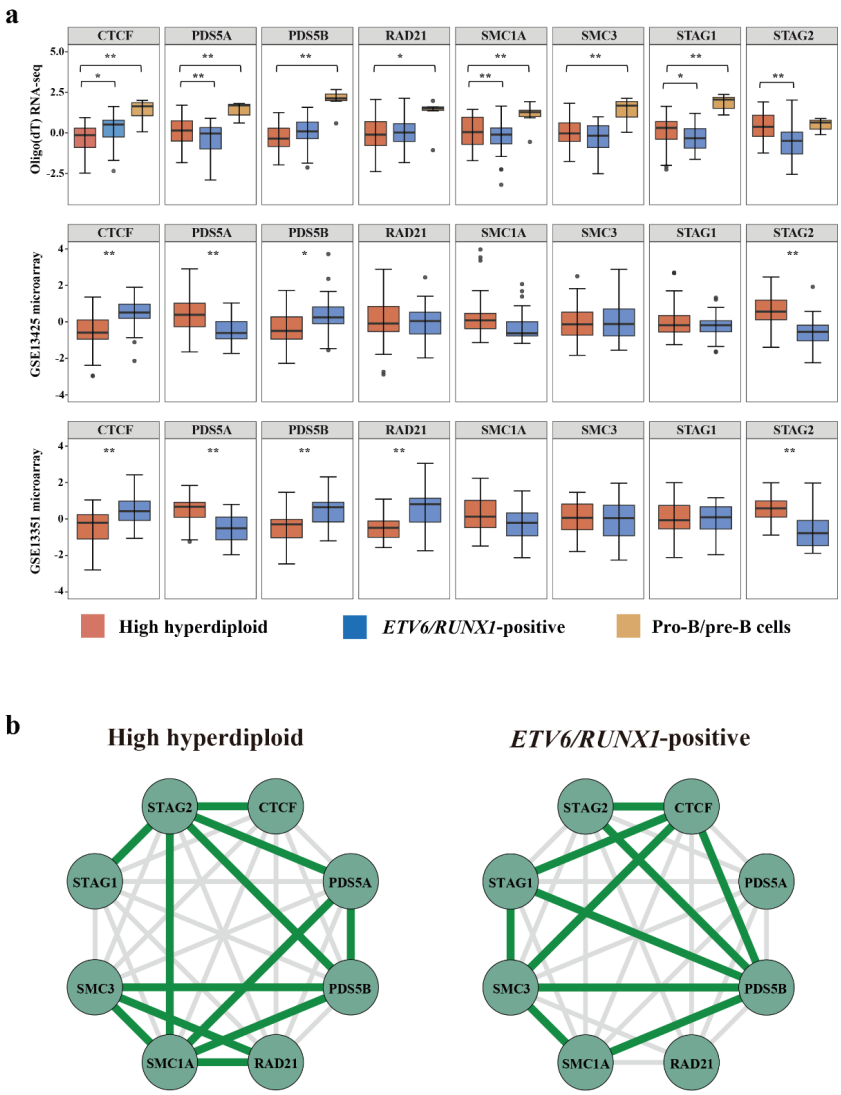
Supplementary Fig. 2. mRNA-protein correlation analysis. a. mRNA-protein correlation in relation to protein and mRNA stability, showing higher correlation for genes with similar stability on both the mRNA and protein levels. **b.** Transcripts reported to be targeted by miRNAs were compared to non-targeted transcripts, showing higher correlations for non-targeted transcripts. **c.** Association of mRNA-protein correlation to ubiquitination. mRNA-protein correlations are categorized into low (black) and high (grey) ubiquitination based on time series data following proteasomal inhibition by bortezomib. Proteins targeted by the proteasome displayed lower correlations. **d.** Comparison of proteins with an exponential (ED) and non-exponential degradation profile (NED), showing enrichment of genes with low mRNA-protein correlations among rapidly degrading proteins. **e.** Impact of protein subcellular localization on mRNA-protein correlations, showing higher correlations for secretory and cytosolic proteins. **f.** Comparison of mRNA-protein correlation distribution for differentially expressed and non-differentially expressed proteins and mRNAs. Genes that were differentially expressed between hyperdiploid and *ETV6/RUNX*1-positive leukemia displayed higher mRNA-protein correlations. For all panels, number of observations, medians, first and third quartiles, and whiskers extending to 1.5 times the interquartile range are displayed. Non-parametric Wilcoxon rank sum test (b-d, f) or Kruskal-Wallis test (a, e) was used to calculate *P*-values.



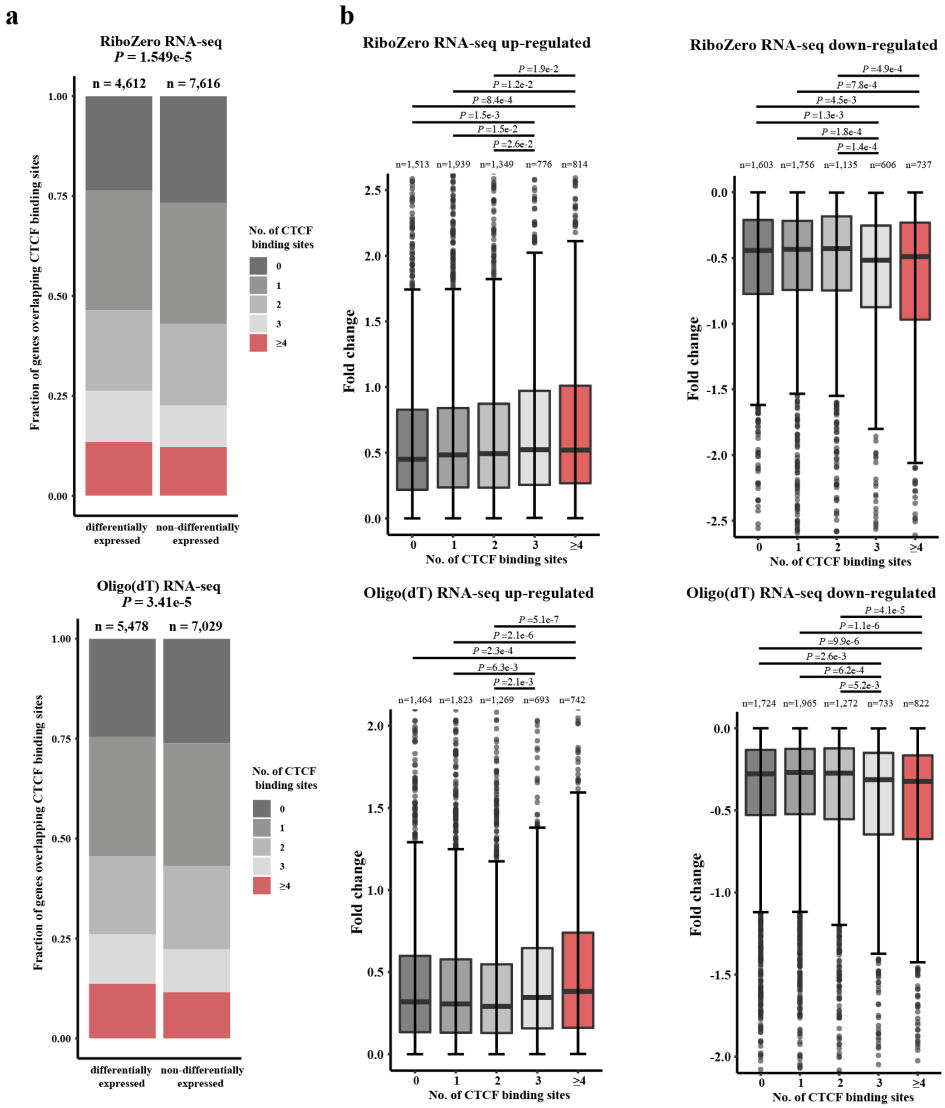
Supplementary Fig. 3. Analysis of the impact of copy number on expression in childhood acute lymphoblastic leukemia (ALL). **a.** Correlations of copy number aberration (CNA) (x-axes) to RNA expression levels (y-axes) based on oligo(dT) RNA-seq from high hyperdiploid cases are shown. Significant (multiple-test adjusted $P < 0.05$) positive (red) and negative (blue) correlations between CNA and mRNAs are indicated. CNA *cis* effects appear as a red diagonal line, CNA *trans* effects as vertical stripes. The fraction (%) of significant CNA *trans* effects (positive in red and negative in blue) for each CNA gene is shown. The bottom panel show the fraction (%) of leukemias harboring CNA (copy number gain in red and copy number loss in blue). **b.** Location and expression of oncogenes and tumor suppressor genes in relation to chromosomal gains in high hyperdiploid vs. *ETV6/RUNX1*-positive ALL. No association was seen between oncogenes and copy number gains or tumor suppressor genes and non-gained chromosomes.



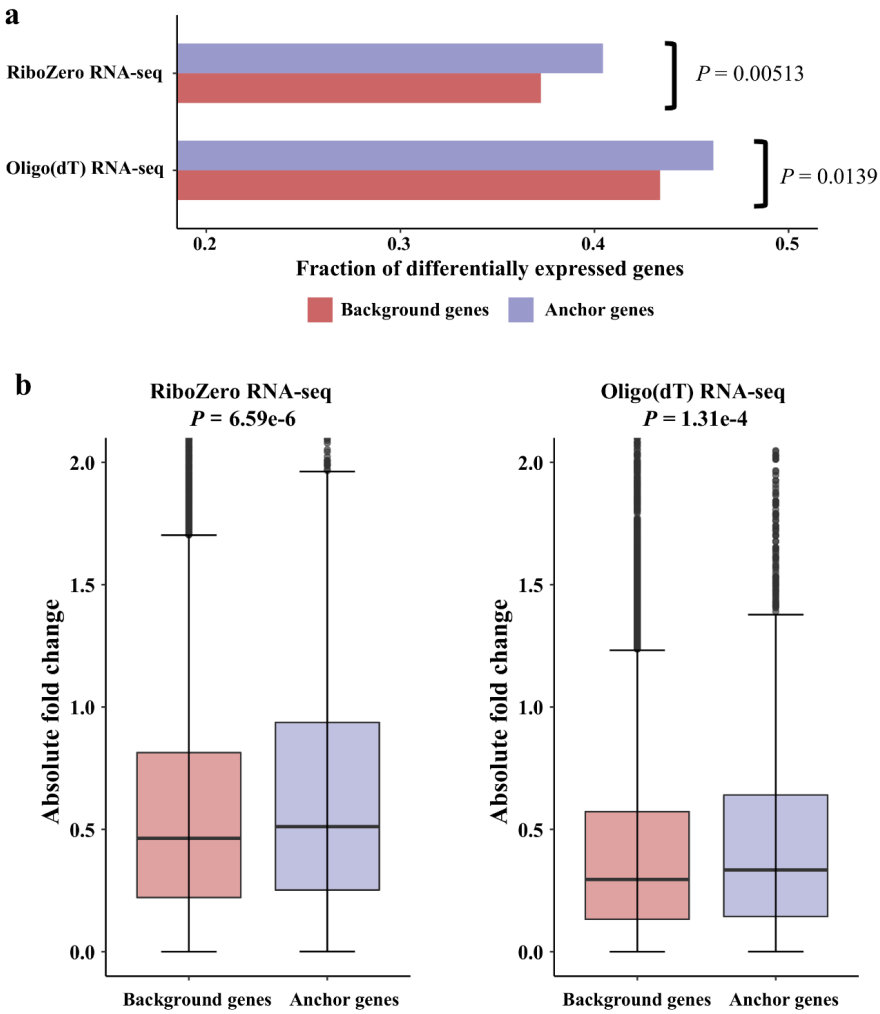
Supplementary Fig. 4. Examples of expression data from differentially expressed genes between high hyperdiploid and *ETV6/RUNX1*-positive acute lymphoblastic leukemia. The center of the boxplot is the median and lower/upper hinges correspond to the first/third quartiles; whiskers are 1.5 times the interquartile range and data beyond this range are plotted as individual points. Non-parametric Wilcoxon rank sum test was used to calculate *P*-values.



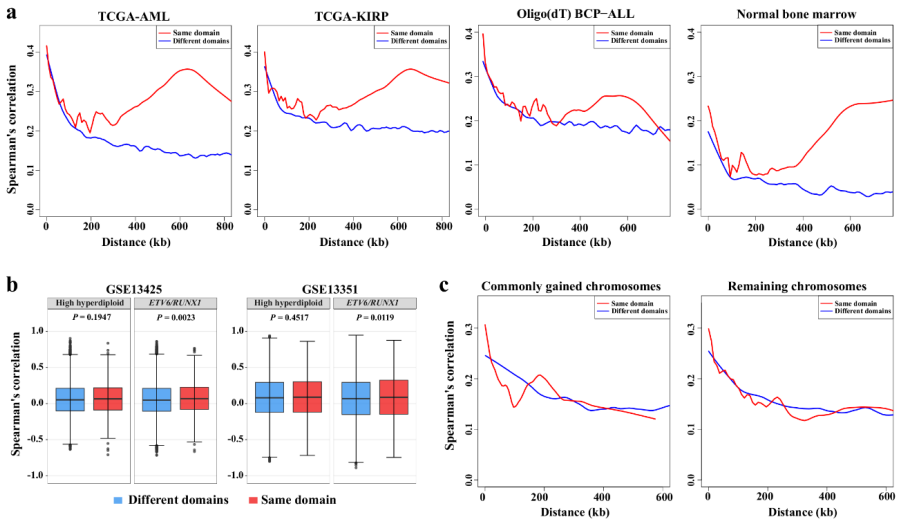
Supplementary Fig. 5. *CTCF* and members of the cohesin complex in high hyperdiploid and *ETV6/RUNX1*-positive leukemia. a. mRNA expression in acute lymphoblastic leukemia datasets. The center of the boxplot is the median and lower/upper hinges correspond to the first/third quartiles; whiskers are 1.5 times the interquartile range and data beyond this range are plotted as individual points. Non-parametric Wilcoxon rank sum test was used to calculate *P*-values. **b.** Cohesin complex members correlation. Green lines represent a Spearman's correlation coefficient of > 0.5 and grey lines are a correlation > 0 but < 0.5.



Supplementary Fig. 6. Gene expression changes and number of CTCF binding sites in high hyperdiploid leukemia. a. Genes that were differentially expressed between high hyperdiploid and *ETV6/RUNX1*-positive leukemias were strongly enriched for more CTCF binding sites in both the RiboZero (chi-squared test, P = 1.55e-5) and oligo(dT) (chi-squared test, P = 3.41e-5) RNA-seq datasets. **b.** Genes with higher numbers of CTCF binding sites in their bodies or flanking 5 kb showed significantly larger fold changes in both datasets. Two-sided Mann-Whitney U test was used to calculate P-values.



Supplementary Fig.7. Gene expression changes and CTCF/cohesin-mediated chromatin structures in high hyperdiploid leukemia. a. Fraction of anchor genes (n=1,825 and n=1,910, respectively) and background genes (n=10,403 and n=10,597, respectively) that were differentially expressed between high hyperdiploid and *ETV6/RUNX1*-positive leukemia. A significantly higher proportion of anchor genes were differentially expressed in both the RiboZero (hypergeometric test, $P = 0.00513$) and oligo(dT) (hypergeometric test, $P = 0.0139$) RNA-seq datasets. **b.** Anchor genes (n=1,825 and n=1,910, respectively) showed significantly higher absolute fold changes than background genes (n= 10,403 and n=10,597, respectively) in both the RiboZero (two-sided Mann-Whitney U test, $P = 6.59e-6$) and oligo(dT) (two-sided Mann-Whitney U test, $P = 1.31e-4$) RNA-seq datasets.



Supplementary Fig. 8. Spearman's correlation score between gene pairs as a function of distance for genes in the same or different topologically associating domains (TADs), showing higher correlation between the expression of gene pairs within the same TAD compared with gene pairs separated by a TAD boundary. a. RNA-sequencing data from acute myeloid leukemia (TCGA-LAML; n=151), papillary renal-cell carcinoma (TCGA-KIRP; n=270), childhood acute lymphoblastic leukemia (OligoT BCP-ALL; n=201), and from normal bone marrow (n=20) **b.** Microarray-based gene expression data from high hyperdiploid and *ETV6/RUNX1*-positive ALL. The center of the boxplot is the median and lower/upper hinges correspond to the first/third quartiles; whiskers are 1.5 times the interquartile range and data beyond this range are plotted as individual points. **c.** RNA-seq data from high hyperdiploid cases in the oligo(dT) dataset, analyzing commonly gained chromosomes separately from the remaining chromosomes.

Supplementary Table 1. Overview of mass-spectrometry data

	Proteins (Gene symbol, 1% FDR)	Peptides (Unique, 1% FDR)	Peptide-spectrum matches (total)
SET A	9,403	136,521	236,759
SET B	9,331	141,510	217,436
SET C	9,085	120,575	190,932
Total	10,138	174,966	645,127

FDR, false discovery rate

Supplementary Data 1. Clinical and cytogenetic data from 89 cases of childhood B-cell precursor acute lymphoblastic leukemia

Case No.	Group	Gender	Age	WBC count ($x10^9$/l)	Bone marrow blast %	Karyotype[a,b]	WGS	WES	SNP array analysis	RiboZero RNA-seq	Oligo(dT) RNA-seq	Proteomics	Hi-C	Chromosome morphology
HeH_1	High hyperdiploid	M	1	36	67	54,XY,+X,+6,+14,+15,+17,+18,+21,+21	n	y	y	y	y	y	n	n
HeH_2	High hyperdiploid	M	1	30	65	62,XY,+X,-Y,t(X;4)(q28;q35.2),t(2;19)(p11.2;q13.32),+4,+5,+6,+7,+10,+11,+12,+14,+14,+16,+der(17)t(4;17)(q26;q22),+18,+21,+21,+21,+22	y	n	y	y	y	y	n	y
HeH_3	High hyperdiploid	M	2	14	70	56,XY,+X,+4,+6,+10,+14,+17,+18,?+der(19)t(1;19)(q11;p13),+21,+21	y	n	y	y	y	y	n	y
HeH_4	High hyperdiploid	M	3	5.4	94	55,XY,+X,dup(1)(q12q25),+4,+6,+10,+14,+17,+18,+21,+21 [based on FISH and SNP array analysis]	n	y	y	y	y	y	n	n
HeH_5	High hyperdiploid	M	3	98	95	56,XY,+X,der(1)?ins(1;?)(q21;?),+4,+6,+8,+10,+14,+17,+18,+21,+21	n	y	y	y	n	y	n	y
HeH_6	High hyperdiploid	M	3	51	NK	55,XY,+X,+4,+5,+6,idic(7)(p11),+8,+10,+14,+17,+21	n	y	y	y	y	y	n	y
HeH_7	High hyperdiploid	M	3	6.6	NK	54,XY,+X,+4,+6,+10,+14,+17,+18,+21	n	y	y	y	y	y	n	y
HeH_8	High hyperdiploid	F	3	10	50	53,XX,+X,+6,+10,+11,+18,+18,+21/54,idem,+8	n	y	y	y	y	y	n	y
HeH_9	High hyperdiploid	F	3	19	NK	55,XX,+X,+4,+6,+7,+10,+14,+17,+18,+21	n	y	y	y	y	y	y	y
HeH_10	High hyperdiploid	M	3	90	90	56,XY,+X,+4,+6,+8,+10,+14,+17,+18,+21,+21	n	y	y	y	n	y	y	y
HeH_11	High hyperdiploid	M	4	4.2	NK	61,XX,-Y,+3,+4,+5,+6,+8,+10,+10,+13,+14,+14,+17,+18,+18,+21,+21	n	y	y	y	y	y	n	y
HeH_12	High hyperdiploid	M	4	17	80	56,XY,+X,+3,+4,+6,+8,+10,+14,+16,+18,+21/56,idem,i(7)(q10)	y	n	y	y	y	y	n	y
HeH_13	High hyperdiploid	F	4	16	84	52,XX,+X,+6,+14,+17,+18,+21/53,idem,+21	y	n	y	y	y	y	n	y
HeH_14	High hyperdiploid	M	7	8.1	64	56,XY,+X,+4,+6,+10,+14,+14,+der(17)t(3;17)(q21;p11.1),+18,+18,+21	n	y	y	y	y	y	n	y
HeH_15	High hyperdiploid	M	8	1.8	91	54,XY,+X,+4,+6,+8,+10,+18,+21,+21	n	y	y	y	n	y	n	y
HeH_16	High hyperdiploid	M	10	4.1	NK	54,XY,+X,+6,+10,+14,+17,+18,+21,+21	n	y	y	y	y	y	n	y
HeH_17	High hyperdiploid	M	13	66	90	55,XY,+X,t(2;8)(p11.2;q21.13),+4,+6,+10,+14,+der(17)t(17;19)(q?;?)del(17)(p11p13),+18,der(19)t(17;19),+21,+21	y	y	y	y	y	y	n	y
HeH_18	High hyperdiploid	F	16	2	NK	60,XX,+X,+4,+5,+6,+7,+8,+10,+14,+14,+17,+18,+21,+21,+21	n	y	y	y	y	y	n	n
HeH_19	High hyperdiploid	M	2	54	NK	56,XY,+X,+Y,+4,t(?5;9)(q?;p13-21),+6,+10,+14,+17,+18,+21,+22	n	y	y	n	y	n	n	y
HeH_20	High hyperdiploid	F	4	6.8	NK	52,XX,+X,+4,+8,+10,+18,+21	n	n	y	n	y	n	n	n
HeH_21	High hyperdiploid	F	6	6.2	NK	53,XX,+X,+4,+6,+14,+17,+21,+21	n	n	y	n	y	n	n	y
HeH_22	High hyperdiploid	F	4	4.8	93	54,XX,+X,+4,+6,+10,+17,+18,+21,+21 [based on FISH and SNP array analysis]	y	n	y	n	y	n	n	y
HeH_23	High hyperdiploid	F	11	4.7	75	57,XX,+X,+6,+10,+10,+add(14)(p1?),+15,+17,+18,+21,+21,+mar,inc	n	n	y	n	y	n	n	y
HeH_24	High hyperdiploid	F	8	5	60	59,XX,+X,+X,+4,+6,+9,+10,+10,+14,+15,idic(17)(p11),+18,+20,+21,+21	y	n	y	n	y	n	n	y
HeH_25	High hyperdiploid	F	4	3	90	54,XX,+X,+4,+6,+14,+17,+18,der(19)t(18;19)(q22.3;p13.3),+21,+21	y	n	y	n	y	n	n	y
HeH_26	High hyperdiploid	F	15	3.5	NK	57,XX,+X,+X,+4,+6,der(8)t(8;14)(p11;q12),+10,+14,+14,+17,+18,+21,+21	n	y	y	n	y	n	n	y
HeH_27	High hyperdiploid	M	5	16	80	57,XY,+X,+4,+5,+6,+9,+10,+14,+der(16)t(11;16)(p11.2;p11.2),+18,+21,+21	y	y	y	n	y	n	n	y
HeH_28	High hyperdiploid	F	2	129	NK	53,XX,+X,+8,+14,+15,+17,+21,+21	n	n	y	n	y	n	n	y
HeH_29	High hyperdiploid	F	4	7.4	70	54-57,XX,+X,+X,+4,+5,+6,*der*(?6)t(1;6)(q21;p25),+10,+14,i(17)(q10),+18,+21,+21	n	n	y	n	y	n	n	y
HeH_30	High hyperdiploid	M	5	2.2	NK	55,XY,+X,+6,+9,+14,+14,+17,+18,+21,+21	n	n	y	n	y	n	n	y
HeH_31	High hyperdiploid	M	3	15	75	57,XY,+X,+Y,+4,+5,+6,+10,+14,+17,+21,+21,+21	n	n	y	n	y	n	n	y
HeH_32	High hyperdiploid	F	2	17	NK	56,XX,+X,+4,+6,+8,+10,+14,+17,+18,+21,+21	y	y	y	n	y	n	n	y
HeH_33	High hyperdiploid	F	3	5	NK	57,XX,+X,+4,+6,+8,+8,+12,+14,+17,+18,+21,+21/58,idem,+10 [based on SNP array analysis]	n	n	y	n	y	n	n	y
HeH_34	High hyperdiploid	F	3	3.8	40	58,XX,+X,+4,+6,+9,+10,+11,+12,+14,+17,+18,+21,+21	n	y	y	n	y	n	n	n
HeH_35	High hyperdiploid	F	5	6.3	NK	53,XX,+4,+6,+10,+14,+18,+21,+21	y	n	y	n	y	n	n	n
HeH_36	High hyperdiploid	F	5	15	85	55,XX,+X,+2,+4,+6,+8,+14,+21,+21,+21 [based on SNP array analysis]	n	n	y	n	y	n	n	n
HeH_37	High hyperdiploid	M	1	85	90	55,XY,+X,+4,+6,+14,+17,+18,+21,+21,+22	n	y	y	n	y	n	n	n
HeH_38	High hyperdiploid	M	3	2.1	NK	52,XY,+X,dup(1)(q21q42),+6,+10,+11,+21,+21	y	n	y	n	y	n	n	y
HeH_39	High hyperdiploid	M	3	30	NK	50-56,XY,+X,+4,+6,+14,+17,+18,+add(21)(q?),+der(?)t(?;21),inc	n	n	y	n	y	n	n	n
HeH_40	High hyperdiploid	F	1	3.4	NK	54,XX,+X,+6,+8,+14,+17,+18,+21,+21	n	n	y	n	y	n	n	y
HeH_41	High hyperdiploid	F	2	9.7	35	54,XX,+X,+6,+10,+14,+17,+18,+21,+21 [based on FISH and SNP array analysis]	n	y	y	n	y	n	n	n
HeH_42	High hyperdiploid	M	7	1.7	NK	52-59,XY,+X,+4,+5,+6,+8,+10,+10,der(13;14)(q10;q10)c,+14,+17,+18,+18,+20,+21,+22	n	n	y	n	y	n	y	y
HeH_43	High hyperdiploid	M	2	6.2	NK	63,XY,+X,+Y,+4,+der(5)t(1;5)(q12;q21),+6,+8,+9,+10,+11,+12,+14,+14,+17,+18,+21,+21,+22	y	n	y	n	y	n	n	y
HeH_44	High hyperdiploid	M	2	8.8	36	54,XY,+X,+4,+6,+10,+14,+17,+18,+21/55,idem,+21	n	n	y	n	y	n	n	y
HeH_45	High hyperdiploid	M	6	2.6	62	57-58,XY,+X,+4,+6,+9,+10,+14,+17,+18,+18,+21,+21	n	y	y	n	y	n	n	y
HeH_46	High hyperdiploid	M	6	3.2	87	55,XY,+X,+4,*i*(7)(q10),+8,+9,+10,+14,+18,+21,+21	n	n	y	n	y	n	n	n
HeH_47	High hyperdiploid	M	2	12	NK	57,XY,+X,+Y,+4,+6,+10,+14,+17,+18,+21,+21,+mar	n	n	y	n	y	n	n	y
HeH_48	High hyperdiploid	F	4	23	85	52-55,XX,+X,+6,+10,+10,+14,+14,+18,+21,+21	n	n	y	n	n	n	y	y
ETV6/RUNX1_1	*ETV6 /RUNX1*	F	3	30	78	46,XX,t(12;21)(p13;q22)	n	y	y	y	y	y	n	y
ETV6/RUNX1_2	*ETV6 /RUNX1*	F	3	16	75	??,X?,del(12)(p13p13),t(12;21)(p13;q22)	n	y	y	y	y	y	n	n
ETV6/RUNX1_3	*ETV6 /RUNX1*	F	3	20	NK	49,XX,+X,t(12;21)(p13;q22),+18,+21	n	y	y	y	y	y	n	y
ETV6/RUNX1_4	*ETV6 /RUNX1*	M	4	164	93	??,X?,t(12;21)(p13;q22)	n	y	y	y	y	y	n	y
ETV6/RUNX1_5	*ETV6 /RUNX1*	M	4	7.8	86	46,XY,del(12)(p13p13),t(12;21)(p13;q22)	n	y	y	y	n	y	n	y
ETV6/RUNX1_6	*ETV6 /RUNX1*	M	5	13	48	46,XY,t(12;21)(p13;q22)/47,idem,+16/47,idem,der(6)t(X;6)(?;q1?),+16	n	y	y	y	y	y	y	y
ETV6/RUNX1_7	*ETV6 /RUNX1*	M	6	31	95	45-46,XY,del(6)(q?),add(12)(p11),t(12;21)(p13;q22),inc	n	y	y	y	y	y	n	y
ETV6/RUNX1_8	*ETV6 /RUNX1*	M	6	5	86	46,XY,del(12)(p13p13)t(12;21)(p13;q22)	n	y	y	y	y	y	n	y
ETV6/RUNX1_9	*ETV6 /RUNX1*	F	7	2.8	NK	46,XX,t(12;21)(p13;q22)	n	n	y	y	y	y	n	y
ETV6/RUNX1_10	*ETV6 /RUNX1*	M	4	55	NK	46,XY,del(2)(p11),add(3)(p11),add(6)(q21),add(12)(p13),der(12)t(12;21)(p13;q22),ider(21)(q10)t(12;21),+2-3mar,inc	n	n	y	n	y	n	n	y
ETV6/RUNX1_11	*ETV6 /RUNX1*	F	7	2.1	53	46,X?X,-5,add(7)(q31),add(12)(p11),t(12:21)(p13;q22),-13,-13,?der(21;21)(q10;q10),+21,+2-3mar	n	n	y	n	y	n	n	y
ETV6/RUNX1_12	*ETV6 /RUNX1*	M	3	4.1	NK	??,X?,t(12;21)(p13;q22),+der(?)t(?;21)(?;q?)/??,X?,t(12;21),+21,+der(?)t(?;12)t(12;21)	n	n	y	n	y	n	n	y
ETV6/RUNX1_13	*ETV6 /RUNX1*	F	6	8.1	NK	46,XX,del(6)(q21),t(12;21)(p13;q22)/47,XX,t(12;21),+21/47,XX,t(12;21),+der(21)t(12;21)	n	n	y	n	y	n	n	y
ETV6/RUNX1_14	*ETV6 /RUNX1*	F	10	2.6	30	??,X,-X,+4,+4,+6,+6,+8,+8,+10,+10,t(12;21)(p13;q22)x1-2,+14,+14,+17,+17,+18,+18,+21,+21,+21	n	y	y	n	y	n	n	y
ETV6/RUNX1_15	*ETV6 /RUNX1*	M	4	25	NK	??,X?,der(11)t(X;11),t(12;21)(p13;q22),+der(21)t(12;21)	n	n	y	n	y	n	n	n
ETV6/RUNX1_16	*ETV6 /RUNX1*	F	1	30.7	68	??,X?,del(12)(p13p13),t(12;21)(p13;q22),+21,+21	n	n	y	n	y	n	n	y
ETV6/RUNX1_17	*ETV6 /RUNX1*	M	6	2.8	NK	??,X?,+X,t(12;21)(p13;q22)	n	n	y	n	y	n	n	n
ETV6/RUNX1_18	*ETV6 /RUNX1*	M	10	5.5	NK	46,XX,t(12;21)(p13;q22)	n	n	y	n	y	n	n	y
ETV6/RUNX1_19	*ETV6 /RUNX1*	M	6	14	81	46,XY,t(12;21)(p13;q22)/46,idem,,add(12)(p13)	n	n	y	n	y	n	n	y
ETV6/RUNX1_20	*ETV6 /RUNX1*	M	8	77	98	48,XY,dup(10)(p11p15),t(12;21)(p13;q22),+16,+21	n	n	y	n	y	n	n	y
ETV6/RUNX1_21	*ETV6 /RUNX1*	M	6	4.5	72	??,X?,t(12;21)(p13;q22),+der(21)t(12;21)	n	n	y	n	y	n	n	y
ETV6/RUNX1_22	*ETV6 /RUNX1*	M	3		83	46,XY,?add(7)(p21),t(12:21)(p13;q22),del(12)(p13p13),add(15)(q21),add(22)(q13)	n	n	y	n	y	n	n	y

ETV6/RUNX1_23	*ETV6/RUNX1*	M	2	5.4	76	47-48,XY,add(12)(p11),t(12;21)(p13;q22),+21	n	n	y	n	y	n	n	y
ETV6/RUNX1_24	*ETV6/RUNX1*	F	3	8.9	75	48,XX,t(12;21)(p13;q22),+21,+22/48,idem,del(6)(q21q27?5)	n	n	y	n	y	n	n	y
ETV6/RUNX1_25	*ETV6/RUNX1*	F	5	5.9	41	45,XX,add(6)(q15),del(12)(p11),t(12;21)(p13;q22),-13,add(15)(q22)	n	n	y	n	y	n	n	y
ETV6/RUNX1_26	*ETV6/RUNX1*	F	3	1.8	NK	46,XX,t(12;21)(p13;q22)	n	n	y	n	y	n	n	n
ETV6/RUNX1_27	*ETV6/RUNX1*	M	3		43	??,?X,t(12:21)(p13;q22),+der(21)t(12:21)p13;q22)	n	n	y	n	y	n	n	n
ETV6/RUNX1_28	*ETV6/RUNX1*	M	5	71	88	Failure	n	n	y	n	y	n	n	n
ETV6/RUNX1_29	*ETV6/RUNX1*	M	5	8	NK	46,XY,dup(5)(q34q35),der(6)t(X;6)(q21;q14),del(11)(q14q25),t(12;21)(p13;q22)	n	n	y	n	y	n	n	n
ETV6/RUNX1_30	*ETV6/RUNX1*	M	6	7.5	18	46,XY,t(12;21)p13;q22)	n	n	y	n	y	n	n	y
ETV6/RUNX1_31	*ETV6/RUNX1*	F	6	14	71	??,XX,t(12;21)(p13;q22),inc	n	n	y	n	y	n	n	y
ETV6/RUNX1_32	*ETV6/RUNX1*	M	9	0.9	NK	47,XY,t(12;21)p13;q22),+der(21)t(12;21)	n	n	y	n	y	n	n	n
ETV6/RUNX1_33	*ETV6/RUNX1*	M	5	39	86	46,XY,dup(X)(q25q28),t(12;21)p13;q22),add(16)(q21)	n	n	y	n	y	n	n	y
ETV6/RUNX1_34	*ETV6/RUNX1*	M	3	3.1	52	46,XY,del(12)(p13p13),t(12;21)(p13;q22)/47,idem,+der(21)t(12;21)	n	n	y	n	y	n	n	y
ETV6/RUNX1_35	*ETV6/RUNX1*	M	1	3.4	74	46,XY,t(12;21)p13;q22)	n	n	y	n	y	n	n	y
ETV6/RUNX1_36	*ETV6/RUNX1*	F	6	4.7	NK	46,XX,der(12)t(12;21)(p17;q?)t(12;16)(q1?;p11),del(16)(p11),del(21)(q21q22),der(21)t(12;21)(p13;q22)t(12;12)(p13;q13)	n	n	y	n	y	n	n	y
ETV6/RUNX1_37	*ETV6/RUNX1*	M	7	2.5	78	46,XY,del(6)(q14q22),del(6)(q24q27),dupB)(q13q24),der(12)t(12;21)(p13;q22),del(13)(q13q31),ider(21)(q10)t(12;21)	n	n	y	n	y	n	n	n
ETV6/RUNX1_38	*ETV6/RUNX1*	F	4	3.3	NK	46,XX,del(12)(p12p13),t(12;21)(p13;q22)	n	n	y	n	y	n	n	y
ETV6/RUNX1_39	*ETV6/RUNX1*	F	3	4	NK	46-47,XX,der(2)t(2;5)(p13;q13),del(4)(q11),del(5)(q13),der(6)t(2;6)(p13;p22),-9,del(12)(p11),der(12)t(4;12)(q11;p12),?add(13)	n	n	y	n	y	n	n	y
ETV6/RUNX1_40	*ETV6/RUNX1*	M	0	15	65	52,XY,+X,+4,der(6)t(6;12)(p17?;q15),add(8)(p?21),+9,+10,del(12)(q15),der(12)t(6;12)(p17?;p13),ins(12;21)(p13;q22),+der(21)t(12;21)(p13;q22)	n	n	y	n	y	n	n	y
ETV6/RUNX1_41	*ETV6/RUNX1*	F	9	47	95	45-47,X,-X,del(11)(q?14),del(12(p11)(t(12;21)(p13;q22),add(15)p11),inc	n	n	y	n	n	n	y	y

Abbreviations: F, female; FISH, fluorescence in situ hybridization; Hi-C, M, male; n, no; NK, not known; SNP, single nucleotide polymorphism; WBC, white blood cell; WES, whole exome sequencing; WGS, whole genome sequencing; y, yes.

[a]Karyotypes based G-banding, SNP array analysis and/or whole genome sequencing.

[b]The karyotypes of all cases except HeH_10 and ETV6/RUNX1_5 have been previously published in Paulsson et al (2010), Paulsson et al (2015), or Lilljebjörn et al (2016).

Supplementary Data 2. Somatic mutations detected in 27 cases of childhood acute lymphoblastic leukemia

Too large for printing

Supplementary Data 3. Log2-values of relative levels of 10,138 gene-centric proteins detected and fully quantified in any one of the 3 TMT-sets

Too large for printing

Supplementary Data 4. Expression values of 8,480 proteins detected in 27 childhood acute lymphoblastic leukemias

Too large for printing

Supplementary Data 5. Expression values of 12,313 mRNAs detected in 27 childhood acute lymphoblastic leukemias analyzed by RiboZero RNA-seq

Too large for printing

Supplementary Data 6. Expression values of 12,594 mRNAs detected in 83 childhood acute lymphoblastic leukemias analyzed by oligo(dT) RNA-seq

Too large for printing

Supplementary Data 7. Enriched gene sets in high hyperdiploid vs *ETV6/RUNX1* -positive acute lymphoblastic leukemias based on proteomics data

Too large for printing

Supplementary Data 8. Enriched gene sets in *ETV6/RUNX1* -positive vs high hyperdiploid acute lymphoblastic leukemias based on proteomics data

Gene set	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
GO_CHROMATIN_MODIFICATION	409	-0.20822823	-4.711899	0	0	0	4515	tags=73%, list=53%, signal=149%
GO_CHROMATIN_ORGANIZATION	446	-0.19487198	-4.563794	0	0	0	4515	tags=72%, list=53%, signal=146%
GO_COVALENT_CHROMATIN_MODIFICATION	268	-0.20880376	-3.9299788	0	0	0	4515	tags=74%, list=53%, signal=152%
GO_PEPTIDYL_LYSINE_MODIFICATION	249	-0.20713407	-3.7435987	0	0	0	5170	tags=81%, list=61%, signal=202%
GO_NUCLEAR_CHROMOSOME	340	-0.19142391	-4.133247	0	0	0	4150	tags=67%, list=49%, signal=127%
GO_CHROMATIN	284	-0.19142766	-3.742697	0	0	0	4101	tags=67%, list=48%, signal=125%
GO_NUCLEAR_CHROMATIN	180	-0.20982257	-3.2515132	0	0	0	4144	tags=69%, list=49%, signal=133%
GO_RESPIRATORY_CHAIN	63	-0.3199164	-3.0175676	0	0	0	4572	tags=86%, list=54%, signal=185%
GO_NUCLEIC_ACID_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY	484	-0.19176981	-4.6464334	0	0	0	4105	tags=67%, list=48%, signal=122%
GO_SEQUENCE_SPECIFIC_DNA_BINDING	425	-0.18665822	-4.3051186	0	0	0	2924	tags=52%, list=35%, signal=76%
GO_REGULATORY_REGION_NUCLEIC_ACID_BINDING	379	-0.18015118	-3.9834127	0	0	0	4198	tags=67%, list=50%, signal=126%
GO_CHROMATIN_BINDING	277	-0.19498594	-3.7340603	0	0	0	3755	tags=63%, list=44%, signal=110%
GO_DOUBLE_STRANDED_DNA_BINDING	369	-0.16288863	-3.557147	0	0	0	4122	tags=64%, list=49%, signal=120%
GO_TRANSCRIPTION_FACTOR_ACTIVITY_PROTEIN_BINDING	364	-0.14537601	-3.1549299	0	0	0	3919	tags=60%, list=46%, signal=107%
GO_RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_ACTIVITY_SEQUENCE_SPECIFIC_DNA_BINDING	233	-0.17601933	-3.132979	0	0	0	2332	tags=45%, list=28%, signal=60%
REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	141	-0.32496488	-4.4203773	0	0	0	3422	tags=72%, list=40%, signal=119%
REACTOME_G2_M_CHECKPOINTS	37	-0.41370836	-2.9813194	0	0	0	4067	tags=89%, list=48%, signal=171%
REACTOME_CELL_CYCLE	313	-0.14381069	-2.9727776	0	0	0	5269	tags=76%, list=62%, signal=194%
GO_CHROMATIN_REMODELING	103	-0.2643032	-3.1186914	0	1.62E-04	0.001	4122	tags=75%, list=49%, signal=144%
GO_CORE_PROMOTER_PROXIMAL_REGION_DNA_BINDING	156	-0.20468259	-2.9279506	0	1.25E-04	0.001	4055	tags=68%, list=48%, signal=128%
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS	31	-0.43459332	-2.872156	0	2.28E-04	0.001	3711	tags=87%, list=44%, signal=154%
GO_DNA_CONFORMATION_CHANGE	171	-0.19402522	-2.9805303	0	2.67E-04	0.002	4137	tags=68%, list=49%, signal=130%
GO_NADH_DEHYDROGENASE_COMPLEX	40	-0.38722274	-2.8776546	0	3.69E-04	0.002	4572	tags=93%, list=54%, signal=200%
GO_INTEGRATOR_COMPLEX	12	-0.65682507	-2.8708763	0	3.08E-04	0.002	2210	tags=92%, list=26%, signal=124%
KEGG_CELL_CYCLE	96	-0.2298955	-2.6562445	0	0.002027473	0.002	3370	tags=63%, list=40%, signal=103%
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	57	-0.3096948	-2.7782145	0	3.81E-04	0.002	5569	tags=96%, list=66%, signal=280%
GO_PROTEIN_DEALKYLATION	24	-0.48988795	-2.892131	0	4.63E-04	0.004	3628	tags=92%, list=43%, signal=160%
GO_CHROMOSOMAL_REGION	249	-0.15560369	-2.8253012	0	5.14E-04	0.004	4437	tags=67%, list=52%, signal=138%
GO_HISTONE_METHYLTRANSFERASE_COMPLEX	59	-0.31242496	-2.8045702	0	4.50E-04	0.004	5126	tags=92%, list=61%, signal=230%
GO_HISTONE_DEMETHYLASE_ACTIVITY	22	-0.5035776	-2.7920105	0	4.35E-04	0.004	1907	tags=73%, list=23%, signal=94%
GO_HISTONE_BINDING	117	-0.22299363	-2.7615194	0	3.91E-04	0.004	1903	tags=44%, list=22%, signal=57%
GO_LIGASE_ACTIVITY	271	-0.14284177	-2.7571561	0	3.56E-04	0.004	4675	tags=69%, list=55%, signal=149%
GO_OXIDOREDUCTASE_COMPLEX	79	-0.26770523	-2.7602043	0	5.06E-04	0.005	4617	tags=81%, list=55%, signal=176%
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS_	67	-0.2734732	-2.655756	0	8.00E-04	0.005	5668	tags=94%, list=67%, signal=282%
GO_PROTEIN_UBIQUITINATION	388	-0.13083683	-2.870379	0	6.15E-04	0.006	6039	tags=84%, list=71%, signal=278%
GO_N_ACYLTRANSFERASE_ACTIVITY	71	-0.27377766	-2.7244554	0	4.92E-04	0.006	4144	tags=76%, list=49%, signal=148%
GO_CHROMOSOME_TELOMERIC_REGION	112	-0.2189791	-2.688354	0	8.92E-04	0.01	4070	tags=70%, list=48%, signal=132%
GO_DNA_DEPENDENT_ATPASE_ACTIVITY	67	-0.275116	-2.662945	0	7.44E-04	0.01	4137	tags=76%, list=49%, signal=148%
KEGG_PARKINSONS_DISEASE	83	-0.23051004	-2.5006206	0	0.005242439	0.01	5722	tags=90%, list=68%, signal=276%
GO_METHYLATED_HISTONE_BINDING	40	-0.35414836	-2.6468227	0	8.36E-04	0.012	1886	tags=58%, list=22%, signal=74%
REACTOME_MRNA_PROCESSING	142	-0.182872	-2.594382	0	0.001932084	0.014	6233	tags=92%, list=74%, signal=341%
GO_DIOXYGENASE_ACTIVITY	55	-0.29380834	-2.569609	0	0.001558365	0.024	3381	tags=69%, list=40%, signal=114%
GO_ATP_DEPENDENT_DNA_HELICASE_ACTIVITY	31	-0.3841194	-2.5673566	0	0.001460967	0.024	4137	tags=87%, list=49%, signal=170%
GO_PRC1_COMPLEX	10	-0.6620967	-2.5375957	0	0.002016454	0.025	2870	tags=100%, list=34%, signal=151%
GO_MLL1_2_COMPLEX	26	-0.4140183	-2.5317245	0	0.001925577	0.026	4324	tags=92%, list=51%, signal=188%
GO_SWI_SNF_SUPERFAMILY_TYPE_COMPLEX	61	-0.27817976	-2.5243893	0	0.001777455	0.026	3494	tags=69%, list=41%, signal=116%
GO_DEMETHYLASE_ACTIVITY	27	-0.42045632	-2.5569108	0	0.001546482	0.027	2098	tags=67%, list=25%, signal=88%
GO_SNRNA_PROCESSING	21	-0.4999549	-2.7454185	0	0.002823366	0.031	3844	tags=95%, list=45%, signal=174%
GO_PROTEIN_METHYLTRANSFERASE_ACTIVITY	57	-0.29218847	-2.547581	0	0.001679306	0.031	4676	tags=84%, list=55%, signal=187%
GO_TRANSFERASE_ACTIVITY_TRANSFERRING_ACYL_GROUPS_OTHER_THAN_AMINO_ACYL_GROUPS	129	-0.19607608	-2.5466044	0	0.001590922	0.031	4144	tags=68%, list=49%, signal=132%
REACTOME_CELL_CYCLE_MITOTIC	268	-0.13229915	-2.4961467	0	0.004681622	0.037	5269	tags=75%, list=62%, signal=192%
GO_REGULATION_OF_CELL_CYCLE_PROCESS	370	-0.122890405	-2.7303624	0	0.003115489	0.038	6057	tags=83%, list=72%, signal=279%
GO_PROTEIN_ACETYLATION	97	-0.23788531	-2.7273772	0	0.002832263	0.038	5170	tags=85%, list=61%, signal=214%
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	129	-0.18181092	-2.4658682	0	0.004707857	0.042	6233	tags=91%, list=74%, signal=341%
GO_NUCLEAR_CHROMOSOME_TELOMERIC_REGION	88	-0.22882618	-2.4684777	0	0.003026905	0.048	4436	tags=75%, list=52%, signal=156%
KEGG_HUNTINGTONS_DISEASE	124	-0.17821258	-2.3442972	0	0.016149033	0.048	6233	tags=91%, list=74%, signal=340%

Supplementary Data 9. Enriched gene sets in high hyperdiploid vs *ETV6/RUNX1* -positive acute lymphoblastic leukemias based on RiboZero RNA-seq data

Too large for printing

Supplementary Data 10. Enriched gene sets in *ETV6/RUNX1* -positive vs high hyperdiploid acute lymphoblastic leukemias based on RiboZero RNA-seq data

Gene set	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
GO_NUCLEAR_CHROMOSOME_SEGREGATION	189	-0.22923005	-3.7736554	0	0	0	4779	tags=61%, list=39%, signal=99%
GO_HOMOPHILIC_CELL_ADHESION_VIA_PLASMA_MEMBRANE_ADHESION_MOLECULES	68	-0.3713941	-3.6778214	0	0	0	2153	tags=54%, list=17%, signal=66%
GO_SISTER_CHROMATID_SEGREGATION	161	-0.23011418	-3.4397635	0	0	0	4776	tags=61%, list=39%, signal=99%
GO_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_ADHESION_MOLECULES	87	-0.30793336	-3.3347015	0	0	0	2605	tags=52%, list=21%, signal=65%
GO_ORGANELLE_FISSION	404	-0.14292921	-3.320054	0	0	0	4943	tags=54%, list=40%, signal=87%
GO_CELL_DIVISION	393	-0.14451775	-3.3179195	0	0	0	6675	tags=68%, list=54%, signal=144%
GO_DNA_REPLICATION	190	-0.19870014	-3.19962	0	0	0	7961	tags=84%, list=65%, signal=235%
GO_CHROMOSOME_SEGREGATION	231	-0.18247779	-3.1942134	0	0	0	4779	tags=57%, list=39%, signal=91%
GO_CHROMOSOME_CENTROMERIC_REGION	158	-0.24284896	-3.5777931	0	0	0	5076	tags=65%, list=41%, signal=109%
GO_CONDENSED_CHROMOSOME	163	-0.24331744	-3.5497098	0	0	0	5203	tags=66%, list=42%, signal=113%
GO_CENTROSOME	425	-0.1419371	-3.3600364	0	0	0	6165	tags=64%, list=50%, signal=123%
GO_CHROMOSOMAL_REGION	299	-0.16845706	-3.277777	0	0	0	6378	tags=68%, list=52%, signal=138%
GO_KINETOCHORE	107	-0.27244452	-3.2519007	0	0	0	5076	tags=68%, list=41%, signal=115%
GO_UBIQUITIN_LIKE_PROTEIN_TRANSFERASE_ACTIVITY	361	-0.13600431	-2.9283817	0	0	0	6118	tags=63%, list=50%, signal=121%
REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	294	-0.25508022	-5.063831	0	0	0	6149	tags=75%, list=50%, signal=146%
GO_MITOTIC_NUCLEAR_DIVISION	325	-0.14703423	-3.088663	0	8.76E-05	0.001	4679	tags=52%, list=38%, signal=82%
REACTOME_G2_M_CHECKPOINTS	41	-0.39166096	-2.968715	0.002016129	0.001182141	0.003	5105	tags=80%, list=41%, signal=137%
GO_MICROTUBULE_CYTOSKELETON_ORGANIZATION	261	-0.16103497	-3.0029573	0	3.20E-04	0.004	4429	tags=52%, list=36%, signal=79%
GO_CONDENSED_CHROMOSOME_CENTROMERIC_REGION	88	-0.27354988	-2.9262466	0	5.94E-04	0.004	5192	tags=69%, list=42%, signal=119%
GO_SPINDLE	251	-0.15763737	-2.9065053	0	5.09E-04	0.004	5507	tags=60%, list=45%, signal=107%
KEGG_CELL_CYCLE	117	-0.20907022	-2.6706586	0.001941748	0.002068068	0.004	6607	tags=74%, list=54%, signal=159%
REACTOME_CELL_CYCLE_MITOTIC	299	-0.14688495	-2.886092	0	0.001059651	0.004	6143	tags=64%, list=50%, signal=125%
REACTOME_MITOTIC_PROMETAPHASE	82	-0.26510605	-2.8809922	0	7.95E-04	0.004	5768	tags=73%, list=47%, signal=137%
GO_CELL_MORPHOGENESIS_INVOLVED_IN_NEURON_DIFFERENTIATION	208	-0.18215217	-2.9635642	0	4.34E-04	0.006	2887	tags=41%, list=23%, signal=53%
GO_SISTER_CHROMATID_COHESION	103	-0.25218946	-2.950217	0	4.64E-04	0.007	5768	tags=72%, list=47%, signal=134%
GO_REGULATION_OF_CELL_DIVISION	189	-0.18263744	-2.9470596	0	4.28E-04	0.007	4953	tags=58%, list=40%, signal=96%
GO_CELL_PART_MORPHOGENESIS	386	-0.12743823	-2.8958368	0	4.55E-04	0.008	2915	tags=36%, list=24%, signal=46%
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS	35	-0.3865776	-2.7242105	0	0.00125596	0.008	5105	tags=80%, list=41%, signal=136%
REACTOME_CELL_CYCLE	383	-0.12040669	-2.6680768	0	0.00129085	0.01	9463	tags=89%, list=77%, signal=371%
GO_REGULATION_OF_SYNAPSE_ORGANIZATION	52	-0.34102657	-2.8697016	0	5.82E-04	0.011	5528	tags=79%, list=45%, signal=142%
GO_CALCIUM_ION_BINDING	352	-0.1213069	-2.6610134	0	0.004746517	0.011	2153	tags=29%, list=17%, signal=34%
REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	29	-0.41184217	-2.645802	0	0.001333416	0.012	5132	tags=83%, list=42%, signal=142%
GO_CORE_PROMOTER_PROXIMAL_REGION_DNA_BINDING	228	-0.15145174	-2.651243	0	0.003747146	0.013	1789	tags=29%, list=15%, signal=34%
GO_MRNA_PROCESSING	382	-0.12812349	-2.825878	0	7.95E-04	0.016	10044	tags=94%, list=82%, signal=494%
GO_NEURON_PROJECTION_GUIDANCE	108	-0.23592031	-2.8161657	0	8.89E-04	0.019	2138	tags=41%, list=17%, signal=49%
GO_CELL_CYCLE_PHASE_TRANSITION	235	-0.15793233	-2.802743	0	9.72E-04	0.022	6634	tags=69%, list=54%, signal=148%
GO_DNA_DEPENDENT_ATPASE_ACTIVITY	76	-0.25511605	-2.5920396	0	0.005192522	0.024	7410	tags=86%, list=60%, signal=213%
GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY	118	-0.21892205	-2.7796605	0	0.001172515	0.028	5366	tags=65%, list=44%, signal=115%
GO_NEURON_PROJECTION_MORPHOGENESIS	225	-0.15964186	-2.7558901	0	0.001394536	0.035	2887	tags=39%, list=23%, signal=50%
GO_POSITIVE_REGULATION_OF_CELL_CYCLE_PROCESS	193	-0.16862465	-2.7556095	0	0.00132813	0.035	6251	tags=67%, list=51%, signal=135%
GO_CONDENSED_NUCLEAR_CHROMOSOME	64	-0.26370007	-2.4751134	0	0.005260806	0.046	3697	tags=56%, list=30%, signal=80%

Supplementary Data 11. Enriched gene sets in high hyperdiploid vs *ETV6/RUNX1* -positive acute lymphoblastic leukemias based on oligo(dT) RNA-seq data

Gene set	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
GO_RIBOSOME_BIOGENESIS	296	0.25070602	4.99918	0	0	0	7425	tags=83%, list=59%, signal=199%
GO_RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS	418	0.21212487	4.848896	0	0	0	8805	tags=90%, list=70%, signal=291%
GO_NCRNA_PROCESSING	365	0.21557494	4.7545476	0	0	0	7990	tags=84%, list=63%, signal=224%
GO_RRNA_METABOLIC_PROCESS	247	0.25013962	4.5609345	0	0	0	7516	tags=84%, list=60%, signal=205%
GO_NCRNA_METABOLIC_PROCESS	489	0.18218802	4.559885	0	0	0	8353	tags=84%, list=66%, signal=239%
GO_AMIDE_BIOSYNTHETIC_PROCESS	450	0.15983313	3.851609	0	0	0	8525	tags=83%, list=68%, signal=248%
GO_PEPTIDE_METABOLIC_PROCESS	485	0.15072544	3.791181	0	0	0	8301	tags=80%, list=66%, signal=227%
GO_TRANSLATIONAL_INITIATION	141	0.2593308	3.5440028	0	0	0	8649	tags=94%, list=69%, signal=298%
GO_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICULUM	117	0.2657012	3.2867408	0	0	0	8632	tags=95%, list=69%, signal=299%
GO_RIBOSOME	209	0.250487	4.2095027	0	0	0	7924	tags=88%, list=63%, signal=232%
GO_RIBOSOMAL_SUBUNIT	156	0.2744567	4.0097294	0	0	0	8372	tags=94%, list=66%, signal=276%
GO_CYTOSOLIC_RIBOSOME	104	0.3145486	3.7615237	0	0	0	8301	tags=97%, list=66%, signal=283%
GO_CYTOSOLIC_PART	196	0.19580589	3.2069528	0	0	0	8302	tags=85%, list=66%, signal=246%
GO_LARGE_RIBOSOMAL_SUBUNIT	89	0.28260428	3.1365097	0	0	0	8210	tags=93%, list=65%, signal=266%
GO_ORGANELLE_INNER_MEMBRANE	439	0.13027835	3.0900455	0	0	0	9030	tags=84%, list=72%, signal=287%
GO_CYTOSOLIC_LARGE_RIBOSOMAL_SUBUNIT	56	0.34957737	3.0744233	0	0	0	8210	tags=100%, list=65%, signal=286%
GO_STRUCTURAL_CONSTITUENT_OF_RIBOSOME	191	0.1942428	3.054154	0	0	0	8931	tags=90%, list=71%, signal=305%
KEGG_RIBOSOME	85	0.3194237	3.4482667	0	0	0	8301	tags=98%, list=66%, signal=285%
REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	104	0.3074383	3.7112532	0	0	0	8632	tags=99%, list=69%, signal=312%
REACTOME_TRANSLATION	144	0.26761383	3.7057164	0	0	0	8649	tags=95%, list=69%, signal=300%
REACTOME_METABOLISM_OF_RNA	252	0.19945833	3.6244328	0	0	0	8632	tags=88%, list=69%, signal=274%
REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION	100	0.2989483	3.5026033	0	0	0	8732	tags=99%, list=69%, signal=320%
REACTOME_INFLUENZA_LIFE_CYCLE	134	0.25755367	3.4797208	0	0	0	8632	tags=94%, list=69%, signal=296%
REACTOME_PEPTIDE_CHAIN_ELONGATION	84	0.31911618	3.4627454	0	0	0	8301	tags=98%, list=66%, signal=284%
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	107	0.2777363	3.3296747	0	0	0	8301	tags=93%, list=66%, signal=272%
REACTOME_METABOLISM_OF_MRNA	207	0.19429906	3.260686	0	0	0	8361	tags=86%, list=66%, signal=250%
REACTOME_METABOLISM_OF_PROTEINS	375	0.13962121	3.0524178	0	0	0	8301	tags=79%, list=66%, signal=226%
REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX	105	0.25953582	3.0285935	0	0	0	8632	tags=94%, list=69%, signal=297%
GO_MITOCHONDRIAL_MATRIX	363	0.13507476	2.9350283	0	1.14E-04	0.001	8547	tags=81%, list=68%, signal=245%
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	47	0.34737128	2.8270206	0	7.17E-05	0.001	5287	tags=77%, list=42%, signal=132%
GO_MULTI_ORGANISM_METABOLIC_PROCESS	136	0.23618302	3.1626823	0	2.10E-04	0.002	8632	tags=92%, list=69%, signal=289%
GO_MITOCHONDRIAL_MEMBRANE_PART	149	0.19511405	2.8355687	0	2.05E-04	0.002	9066	tags=91%, list=72%, signal=322%
GO_HEMOGLOBIN_COMPLEX	10	0.73129374	2.795606	0	1.84E-04	0.002	2131	tags=90%, list=17%, signal=108%
GO_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICULUM	102	0.2743921	3.1598682	0	2.85E-04	0.003	8301	tags=93%, list=66%, signal=271%
GO_MITOCHONDRIAL_PROTEIN_COMPLEX	124	0.21223837	2.781828	0	3.39E-04	0.004	9337	tags=95%, list=74%, signal=364%
GO_MRNA_PROCESSING	383	0.13717589	3.0695462	0	4.33E-04	0.005	10063	tags=93%, list=80%, signal=450%
GO_SMALL_RIBOSOMAL_SUBUNIT	67	0.2861866	2.7310307	0	3.93E-04	0.005	8632	tags=97%, list=69%, signal=307%
KEGG_OXIDATIVE_PHOSPHORYLATION	102	0.2139487	2.5607004	0	0.003368789	0.006	9303	tags=95%, list=74%, signal=361%
GO_PRERIBOSOME	58	0.30402923	2.689826	0	5.96E-04	0.008	7696	tags=91%, list=61%, signal=234%
GO_INNER_MITOCHONDRIAL_MEMBRANE_PROTEIN_COMPLEX	96	0.22972177	2.6643124	0	5.53E-04	0.008	9066	tags=95%, list=72%, signal=336%
GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY	117	0.2404671	3.02104	0	7.13E-04	0.009	8301	tags=90%, list=66%, signal=261%
GO_RIBONUCLEOPROTEIN_COMPLEX_SUBUNIT_ORGANIZATION	182	0.19229972	3.0143461	0	6.62E-04	0.009	9445	tags=94%, list=75%, signal=370%
GO_OXYGEN_TRANSPORT	11	0.74045074	2.9910965	0	6.18E-04	0.009	2131	tags=91%, list=17%, signal=109%
GO_RNA_CATABOLIC_PROCESS	215	0.17328711	2.8959625	0	7.72E-04	0.012	8632	tags=86%, list=69%, signal=267%
REACTOME_ACTIVATION_OF_THE_MRNA_UPON_BINDING_OF_THE_CAP_BINDING_COMPLEX_AND_EIFS_AND_SUBSEQUENT_BINDING_TO_43S	56	0.29798278	2.624525	0	9.57E-04	0.013	8632	tags=98%, list=69%, signal=311%
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	140	0.17823075	2.427844	0	0.008128124	0.022	2637	tags=39%, list=21%, signal=48%
GO_SPLICEOSOMAL_COMPLEX	163	0.17006661	2.5132518	0	0.001633782	0.026	9629	tags=93%, list=76%, signal=391%
GO_RNA_SPLICING_VIA_TRANSESTERIFICATION_REACTIONS	253	0.1480691	2.785585	0	0.001759659	0.029	10019	tags=94%, list=80%, signal=451%
GO_RNA_SPLICING	333	0.13281989	2.782664	0	0.0016619	0.029	10019	tags=92%, list=80%, signal=440%
KEGG_PARKINSONS_DISEASE	97	0.20950261	2.400622	0	0.008696155	0.031	9066	tags=93%, list=72%, signal=329%
GO_CYTOSOLIC_SMALL_RIBOSOMAL_SUBUNIT	43	0.3155924	2.4801562	0	0.002181961	0.037	8632	tags=100%, list=69%, signal=317%
GO_VIRAL_LIFE_CYCLE	255	0.14622794	2.7357216	0	0.002390676	0.044	8665	tags=83%, list=69%, signal=261%
GO_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ORGANELLE	330	0.1289272	2.7315574	0	0.002271142	0.044	8951	tags=84%, list=71%, signal=282%
GO_PROTEIN_CATABOLIC_PROCESS	489	0.107858844	2.7223642	0	0.002312098	0.047	9227	tags=84%, list=73%, signal=301%
GO_NITRIC_OXIDE_METABOLIC_PROCESS	10	0.70486337	2.7198806	0	0.002207002	0.047	3723	tags=100%, list=30%, signal=142%

Supplementary Data 12. Enriched gene sets in *ETV6/RUNX1* -positive vs high hyperdiploid acute lymphoblastic leukemias based on oligo(dT) RNA-seq data

Gene set	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
GO_G_PROTEIN_COUPLED_RECEPTOR_SIGNALING_PATHWAY	313	-0.17501	-3.5808058	0	0	0	2157	tags=34%, list=17%, signal=40%
GO_DISRUPTION_OF_CELLS_OF_OTHER_ORGANISM	16	-0.738571	-3.5547235	0	0	0	1731	tags=88%, list=14%, signal=101%
GO_INFLAMMATORY_RESPONSE	299	-0.160442	-3.165028	0	0	0	1398	tags=27%, list=11%, signal=29%
GO_EXTRACELLULAR_MATRIX	177	-0.239577	-3.7715929	0	0	0	2078	tags=40%, list=16%, signal=47%
GO_DNA_PACKAGING_COMPLEX	69	-0.339398	-3.2557814	0	0	0	3781	tags=64%, list=30%, signal=91%
GO_ANCHORED_COMPONENT_OF_MEMBRANE	69	-0.31682	-3.038409	0	0	0	1691	tags=45%, list=13%, signal=52%
GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX	134	-0.220318	-2.9606977	0	0	0	2519	tags=42%, list=20%, signal=52%
GO_SIGNALING_RECEPTOR_ACTIVITY	462	-0.167844	-3.9972274	0	0	0	1563	tags=29%, list=12%, signal=31%
GO_CALCIUM_ION_BINDING	357	-0.173327	-3.7630732	0	0	0	2078	tags=33%, list=16%, signal=39%
GO_G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	157	-0.224869	-3.354219	0	0	0	1536	tags=34%, list=12%, signal=39%
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	90	-0.270521	-3.0557005	0	0	0	3755	tags=57%, list=30%, signal=80%
REACTOME_AMYLOIDS	49	-0.417833	-3.4394717	0	0	0	3755	tags=71%, list=30%, signal=101%
GO_REGULATION_OF_SYNAPSE_ASSEMBLY	32	-0.461889	-3.1488125	0	2.41E-04	0.001	2070	tags=63%, list=16%, signal=75%
REACTOME_RNA_POL_I_PROMOTER_OPENING	40	-0.428282	-3.2008202	0	5.97E-04	0.001	3755	tags=73%, list=30%, signal=103%
REACTOME_GPCR_LIGAND_BINDING	127	-0.229186	-2.9897523	0	3.98E-04	0.001	2102	tags=39%, list=17%, signal=47%
GO_TAXIS	266	-0.1637	-3.0084414	0	3.86E-04	0.002	2102	tags=33%, list=17%, signal=38%
GO_REGULATION_OF_SYNAPSE_ORGANIZATION	50	-0.363144	-2.9774964	0	6.47E-04	0.004	3254	tags=62%, list=26%, signal=83%
GO_PROTEIN_DNA_COMPLEX	127	-0.208275	-2.707367	0	0.001696138	0.008	4941	tags=60%, list=39%, signal=97%
REACTOME_MEIOTIC_RECOMBINATION	58	-0.306783	-2.776687	0	0.002315712	0.008	3755	tags=60%, list=30%, signal=86%
REACTOME_PACKAGING_OF_TELOMERE_ENDS	32	-0.394871	-2.7068706	0	0.002538668	0.011	3699	tags=69%, list=29%, signal=97%
GO_CELL_CELL_ADHESION_VIA_PLASMA_MEMBRANE_ADHESION_MOLECULES	79	-0.272136	-2.8705292	0	0.001945603	0.014	1856	tags=42%, list=15%, signal=49%
REACTOME_MEIOSIS	82	-0.252345	-2.682399	0	0.002677392	0.014	3755	tags=55%, list=30%, signal=78%
GO_NUCLEAR_NUCLEOSOME	24	-0.452784	-2.6364243	0	0.002646855	0.015	3755	tags=75%, list=30%, signal=107%
REACTOME_CELL_CYCLE	370	-0.120322	-2.6698403	0	0.00245556	0.015	8265	tags=77%, list=66%, signal=218%
REACTOME_HEMOSTASIS	343	-0.126474	-2.6555524	0	0.002433309	0.017	3849	tags=43%, list=31%, signal=60%
GO_POSITIVE_REGULATION_OF_SYNAPSE_ASSEMBLY	25	-0.476662	-2.8528705	0	0.002312046	0.019	2070	tags=64%, list=16%, signal=76%
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	92	-0.233907	-2.5954976	0	0.00242138	0.019	1731	tags=37%, list=14%, signal=43%
GO_HEPARIN_BINDING	62	-0.274595	-2.6055458	0	0.005637264	0.025	1638	tags=40%, list=13%, signal=46%
REACTOME_SIGNALING_BY_GPCR	271	-0.13639	-2.5551717	0	0.002972317	0.026	2131	tags=30%, list=17%, signal=36%
GO_ANCHORED_COMPONENT_OF_PLASMA_MEMBRANE	22	-0.453179	-2.562505	0	0.004228907	0.027	2318	tags=64%, list=18%, signal=78%
GO_GLYCOSAMINOGLYCAN_BINDING	84	-0.238655	-2.601601	0	0.004840542	0.027	2113	tags=40%, list=17%, signal=48%
REACTOME_G2_M_CHECKPOINTS	41	-0.34341	-2.5478773	0	0.00302334	0.028	7977	tags=98%, list=63%, signal=265%
GO_CELL_CHEMOTAXIS	97	-0.237062	-2.7748	0	0.003550655	0.032	2102	tags=40%, list=17%, signal=48%
GO_RECEPTOR_COMPLEX	188	-0.156088	-2.511116	0	0.004924349	0.035	1682	tags=29%, list=13%, signal=33%
REACTOME_PEPTIDE_LIGAND_BINDING_RECEPTORS	51	-0.295201	-2.4922652	0	0.004039824	0.041	1731	tags=43%, list=14%, signal=50%
GO_MYELOID_LEUKOCYTE_MIGRATION	57	-0.308254	-2.7479885	0	0.004351344	0.044	2102	tags=47%, list=17%, signal=57%
GO_REGULATION_OF_SYNAPSE_STRUCTURE_OR_ACTIVITY	120	-0.212743	-2.747738	0	0.003955768	0.044	2070	tags=38%, list=16%, signal=44%

Supplementary Data 13. Hi-C sequencing data

Case	total reads	unique valid tags	TAD_number	TAD_average_length (bp)	boundaries number	Number of overlapped boundaries between leukemia samples and cell line[a]	Fraction of overlapped boundaries between leukemia samples and cell line[a]	bound by the insulator protein CTCF[b]	bound by the insulator protein RAD21[c]	bound by CTCF and RAD21
ETV6/RUNX1_6	799 073 552	229 217 960	3483	726292	4347	3117	0.717046	3755 / 86.38%	3455 / 79.48%	3438 / 79.09%
ETV6/RUNX1_41	804 732 730	319 634 311	3576	724547	4225	3146	0.744615	3676 / 87.01%	3401 / 80.50%	3386 / 80.14%
HeH_9	812 778 809	302 548 808	3341	752862	4129	3024	0.732381	3534 / 85.59%	3263 / 79.03%	3248 / 78.66%
HeH_10	785 004 048	230 833 554	2965	843753	3853	2688	0.697638	3256 / 84.51%	2961 / 76.85%	2953 / 76.64%
HeH_42	803 222 486	283 504 248	3960	644322	4659	3488	0.748659	4078 / 87.53%	3812 / 81.82%	3796 / 81.48%
HeH_48	794 859 422	264 227 875	3372	746147	4166	3138	0.753241	3637 / 87.30%	3402 / 81.66%	3388 / 81.33%

[a]GEO accession number GSE63525, https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE63525&format=file&file=GSE63525%5FGM12878%5Fprimary%2Breplicate%5FArrowhead%5Fdomainlist%2Etxt%2Egz

[b]TAD boundaries containing CTCF binding sites (within ±50kb around the boundary sites) downloaded from wgEncodeAwgTfbsSydhGm12878Ctcfsc15914c20UniPk.narrowPeak, wgEncodeAwgTfbsBroadGm12878CtcfUniPk.narrowPeak, wgEncodeAwgTfbsUtaGm12878CtcfUniPk.narrowPeak, wgEncodeAwgTfbsUwGm12878CtcfUniPk.narrowPeak

[c]TAD boundaries containing RAD21 binding sites (within ±50kb around the boundary sites) downloaded from wgEncodeAwgTfbsSydhGm12878Rad21IggrabUniPk.narrowPeak, RAD21_wgEncodeAwgTfbsHaibGm12878Rad21V0416101UniPk.narrowPeak

Supplementary Data 14. Topologically associated domain boundaries lost based on Hi-C

Chromosome	Start position	End position	CTCF and RAD21 binding sitea	ETV6/RUNX1_6	ETV6/RUNX1_41	HeH_9	HeH_10	HeH_42	HeH_48	Genes located within 500 kb upstream or downstrem of boundary site	Genes located within 1000 kb upstream or downstrem of boundary site
chr1	61 925 000	61 950 000	yes	Y	Y	Y	Y	N	N	NFIA,TM2D1,INADL,L1TD1,KANK4,USP1	NFIA,TM2D1,INADL,L1TD1,KANK4,USP1,ANGPTL3,DOCK7,ATG4C
chr1	68 925 000	68 975 000	yes	Y	Y	Y	Y	N	N	RPE65,DEPDC1	DIRAS3,WLS,RPE65,DEPDC1,LRRC7
chr1	75 700 000	75 750 000	yes	Y	Y	Y	Y	N	N	SLC44A5,ACADM,RABGGTB,MSH4,ASB17,ST6GALNAC3 LMNA,SEMA4A,SLC25A44,PMF1-	CRYZ,TYW3,LHX8,SLC44A5,ACADM,RABGGTB,MSH4,ASB17,ST6GALNAC3 YY1AP1,DAP3,GON4L,SYT11,RIT1,KIAA0907,RXFP4,ARHGEF2,SSR2,UBQLN4,LAMTOR2,RAB25,MEX3A,LMNA,SEMA4A,SLC25A4
chr1	156 625 000	156 675 000	yes	Y	Y	W	Y	N	N	BGLAP,PMF1,BGLAP,PAQR6,SMG5,TMEM79,GLMP,VHLL,CCT3,TSACC,RHBG,C1orf61,MEF2D,JQGAP3,TTC24,APOA1BP,GPA	4,PMF1,PMF1-
chr1	164 750 000	164 775 000	yes	Y	Y	Y	Y	N	W	PBX1,LMX1A	PBX1,LMX1A,RXRG,LRRC52,MGST3,ALDH9A1,TMCO1
chr1	177 150 000	177 175 000	yes	Y	Y	Y	Y	Y	W	PAPPA2,ASTN1,BRINP2	RFWD2,PAPPA2,ASTN1,BRINP2,SEC16B,RASAL2
chr1	193 150 000	193 175 000	yes	Y	Y	Y	Y	N	N	RGS13,RGS2,UCHL5,TROVE2,GLRX2,CDC73,B3GALT2	RGS18,RGS21,RGS1,RGS13,RGS2,TROVE2,UCHL5,GLRX2,CDC73,B3GALT2
chr1	194 750 000	194 775 000	yes	Y	Y	Y	Y	W	Y		
chr2	66 650 000	66 675 000	yes	Y	Y	Y	Y	Y	N	MEIS1	MEIS1,ETAA1
chr2	79 450 000	79 475 000	yes	Y	Y	Y	Y	W	Y	REG3G,REG1B,REG1A,REG3A,CTNNA2	REG3G,REG1B,REG1A,REG3A,CTNNA2,LRRTM1
chr2	80 100 000	80 125 000	yes	Y	Y	Y	Y	N	Y	CTNNA2,LRRTM1	REG1A,REG3A,CTNNA2,LRRTM1
chr2	222 825 000	222 850 000	yes	Y	Y	Y	Y	N	N	SGPP2,FARSB,MOGAT1,ACSL3,KCNE4	CCDC140,PAX3,SGPP2,FARSB,MOGAT1,ACSL3,KCNE4,SCG2,AP1S3
chr2	229 500 000	229 525 000	yes	Y	Y	Y	Y	Y	Y	PID1,DNER,TRIP12,FBXO36	PID1,DNER,TRIP12,FBXO36,SLC16A14,SP110,SP140,SP140L,SP100
chr2	236 000 000	236 075 000	yes	Y	Y	Y	Y	N	Y	AGAP1,GBX2,ASB18,IQCA1,ACKR3	AGAP1,GBX2,ASB18,IQCA1,ACKR3
chr3	34 725 000	34 750 000	yes	Y	Y	Y	Y	Y	Y	ARPP21	PDCD6IP,ARPP21
chr3	35 825 000	35 850 000	no	Y	Y	Y	Y	Y	Y		ARPP21,STAC,DCLK3
chr3	55 425 000	55 500 000	yes	Y	Y	Y	Y	N	Y	CACNA2D3,LRTM1,WNT5A,ERC2	LRTM1,WNT5A,ERC2
chr3	59 050 000	59 075 000	yes	Y	Y	Y	Y	N	N	FAM107A,FAM3D,C3orf67	DNASE1L3,ABHD6,RPP14,PXK,PDHB,KCTD6,ACOX2,FAM107A,FAM3D,C3orf67
chr3	61 400 000	61 450 000	yes	Y	Y	Y	Y	N	W	FHIT,PTPRG	FHIT,PTPRG,C3orf14,FEZF2
chr3	78 700 000	78 725 000	yes	Y	Y	Y	Y	N	Y	ROBO1	ROBO1
chr3	114 250 000	114 275 000	yes	Y	Y	Y	W	N	N	ATP6V1A,GRAMD1C,ZDHHC23,KIAA1407,QTRTD1,DRD3,ZNF80,TIGIT,ZBTB20	069063.2,ZBTB20
chr3	131 750 000	131 775 000	yes	Y	Y	Y	Y	N	Y	N.EK11,NUDT16,MRPL3,CPNE4	AC055733.1,ATP2C1,ASTE1,NEK11,NUDT16,MRPL3,CPNE4,ACPP,DNAJC13,ACKR4,UBA5,ACAD11,NPHP3
chr3	139 100 000	139 150 000	yes	Y	Y	Y	Y	Y	W	FAIM,PIK3CB,FOXL2,FOXL2NB,PRR23A,MRPS22,PRR23B,PRR23C,COPB2,RBP2,RBP1,NMNAT3	B2,RBP2,RBP1,NMNAT3,CLSTN2
chr3	161 350 000	161 375 000	yes	Y	Y	Y	Y	N	W	PPM1L,B3GALNT1,NMD3,SPTSSB,OTOL1	SMC4,IFT80,TRIM59,KPNA4,ARL14,PPM1L,B3GALNT1,NMD3,SPTSSB,OTOL1
chr3	165 000 000	165 025 000	yes	Y	Y	Y	Y	N	N	SI,SLITRK3	SI,SLITRK3,BCHE
chr3	168 600 000	168 625 000	yes	Y	Y	Y	Y	N	W	MECOM	WDR49,PDCD10,SERPINI1,GOLIM4,MECOM,AC074033.1
chr3	181 425 000	181 450 000	yes	Y	Y	W	Y	N	W	FXR1,DNAJC19,SOX2	TTC14,CCDC39,FXR1,DNAJC19,SOX2
chr4	25 750 000	25 775 000	yes	Y	Y	Y	W	N	N	PI4K2B,ZCCHC4,ANAPC4,SLC34A2,SEL1L3,SMIM20,RBPJ	SOD3,CCDC149,LGI2,SEPSECS,PI4K2B,ZCCHC4,ANAPC4,SLC34A2,SEL1L3,SMIM20,RBPJ,CCKAR,TBC1D19
chr4	42 275 000	42 325 000	yes	Y	Y	W	Y	N	W	TMEM33,DCAF4L1,SLC30A9,BEND4,SHISA3,ATP8A1	LIMCH1,PHOX2B,TMEM33,DCAF4L1,SLC30A9,BEND4,SHISA3,ATP8A1,GRXCR1
chr4	91 250 000	91 275 000	yes	Y	Y	Y	Y	N	N	CCSER1	
chr4	95 125 000	95 200 000	yes	Y	Y	Y	Y	N	N	PDLIM5,BMPR1B,UNC5C	SMARCAD1,HPGDS,PDLIM5,BMPR1B,UNC5C,PDHA2
chr4	96 050 000	96 100 000	yes	Y	Y	Y	Y	N	N	PDHA2	UNC5C,PDHA2
chr4	107 225 000	107 250 000	yes	Y	Y	W	Y	Y	W	DKK2,PAPSS1	AIMP1,TBCK,DKK2,PAPSS1,SGMS2,CYP2U1,HADH,LEF1
chr4	131 950 000	132 025 000	yes	Y	Y	W	W	W	W		
chr4	138 750 000	138 775 000	yes	Y	Y	Y	Y	N	Y	CCRN4L,ELF2,MGARP,NDUFC1	SLC7A11,CCRN4L,ELF2,MGARP,NAA15,NDUFC1,RAB33B,SETD7,MGST2
chr4	150 750 000	150 775 000	yes	Y	Y	Y	Y	N	N	DCLK2,LRBA,MAB21L2,RPS3A,SH3D19	DCLK2,MAB21L2,LRBA,RPS3A,SH3D19,PRSS48,FAM160A1,GATB
chr4	158 100 000	158 125 000	yes	Y	Y	Y	Y	Y	Y	FAM198B,TMEM144,RXFP1	GRIA2,FAM198B,TMEM144,RXFP1,C4orf46,ETFDH,PPID,FNIP2,C4orf45
chr4	164 500 000	164 525 000	yes	N	N	Y	Y	N	Y	MARCH1,APELA,TRIM61,FAM218A	MARCH1,APELA,FAM218A,TRIM61,TRIM60,TMEM192,KLHL2,MSMO1,CPE
chr4	190 150 000	190 175 000	yes	Y	Y	Y	Y	N	Y	FRG1,FRG2,DUX4	FRG1,FRG2,DUX4
chr5	5 000 000	5 025 000	yes	Y	Y	Y	Y	N	N	ADAMTS16,ICE1	ADAMTS16,ICE1
chr5	28 500 000	28 525 000	yes	Y	Y	Y	Y	N	Y		
chr5	66 125 000	66 175 000	yes	Y	Y	W	Y	N	W	TRIM23,TRAPPC13,CTC-534A2.2,SGTB,NLN,ERBB2IP,SREK1,MAST4	ADAMTS6,CENPK,PPWD1,TRIM23,CTC-534A2.2,TRAPPC13,SGTB,NLN,ERBB2IP,SREK1,MAST4
chr5	82 000 000	82 025 000	no	Y	Y	Y	Y	W	N	SSBP2,ATG10,RPS23,ATP6AP1L	CKMT2,ZCCHC9,ACOT12,SSBP2,ATG10,RPS23,ATP6AP1L
chr5	88 750 000	88 775 000	yes	N	N	Y	W	W	Y	TMEM161B,MEF2C	TMEM161B,MEF2C
chr5	110 500 000	110 525 000	yes	Y	Y	Y	W	Y	Y	TMEM232,SLC25A46	MAN2A1,TMEM232,SLC25A46,TSLP,WDR36,CAMK4,STARD4
chr5	123 675 000	123 700 000	yes	Y	Y	W	Y	N	Y	PRDM6,CEP120,CSNK1G3	SNX2,SNX24,PPIC,PRDM6,CEP120,CSNK1G3
chr5	123 975 000	124 025 000	yes	Y	Y	Y	Y	Y	W	CSNK1G3	PPIC,PRDM6,CEP120,CSNK1G3,ZNF608
chr5	167 000 000	167 025 000	yes	Y	Y	W	Y	Y	Y	TENM2	TENM2
chr5	168 750 000	168 775 000	yes	Y	Y	Y	Y	N	N	TENM2,WWC1,RARS,FBLL1,PANK3,SLIT3	WWC1,RARS,FBLL1,PANK3,SLIT3,SPDL1,DOCK2
chr5	173 500 000	173 525 000	yes	Y	Y	Y	Y	Y	N	ATP6V0E1,CREBRF,BNIP1,NKX2-5,STC2,BOD1,CPEB4,C5orf47	NEURL1B,DUSP1,ERGIC1,RPL26L1,ATP6V0E1,CREBRF,BNIP1,NKX2-5,STC2,BOD1,CPEB4,C5orf47,HMP19
chr6	8 925 000	8 950 000	yes	Y	Y	W	Y	Y	W	SLC35B3	BLOC1S5,EEF1E1,SLC35B3,OFCC1
chr6	40 500 000	40 525 000	yes	Y	Y	Y	W	Y	Y	LRFN2	KIF6,DAAM2,MOCS1,LRFN2,UNC5CL,TSPO2,APOBEC2,OARD1,NFYA,TREML1,TREM2,TREML2,TREML4,TREM1,NCR2
chr6	93 500 000	93 525 000	yes	N	W	Y	Y	Y	Y	EPHA7	EPHA7
chr6	125 125 000	125 150 000	yes	Y	Y	Y	Y	N	N	NKAIN2,RNF217,TPD52L1,HDDC2	RNF217,TPD52L1,HDDC2,HEY2,NCOA7,HINT3,TRMT11
chr6	129 050 000	129 075 000	yes	Y	Y	Y	Y	Y	Y	LAMA2	PTPRK,LAMA2,ARHGAP18,TMEM244,L3MBTL3
chr6	133 525 000	133 550 000	yes	Y	Y	Y	Y	N	N	EYA4,TCF21,TBPL1,SLC2A12	TAAR9,TAAR8,TAAR6,TAAR5,TAAR2,TAAR1,VNN1,VN N3,VNN2,SLC18B1,RPS12,EYA4,TCF21,TBPL1,SLC2A12,SGK1
chr6	133 925 000	133 950 000	yes	Y	Y	Y	Y	N	N	EYA4,TCF21,TBPL1,SLC2A12,SGK1	EYA4,TCF21,TBPL1,SLC2A12,SGK1
chr6	160 500 000	160 525 000	yes	Y	Y	Y	W	N	N	IGF2R,SLC22A1,SLC22A2,SLC22A3,LPA,PLG,MAP3K4	SOD2,WTAP,ACAT2,TCP1,MRPL18,PNLDC1,MAS1,IGF2R,SLC22A1,SLC22A2,SLC22A3,LPA,PLG,MAP3K4,AGPAT4
chr7	21 000 000	21 025 000	yes	Y	Y	Y	W	N	N	ABCB5,SP8,SP4	MACC1,ITGB8,ABCB5,SP8,SP4,DNAH11,CDCA7L
chr7	78 925 000	78 950 000	yes	Y	Y	Y	Y	N	N	MAGI2	MAGI2
chr7	85 100 000	85 200 000	yes	Y	Y	Y	Y	N	N	SEMA3D	SEMA3A,SEMA3D
chr7	115 750 000	115 775 000	yes	Y	Y	Y	W	W	W	TFEC,TES,AC073130.2	MDFIC,TFEC,TES,AC073130.2,CAV2,CAV1,MET
chr7	118 750 000	118 775 000	yes	Y	Y	Y	Y	W	W		CTTNBP2,LSM8,ANKRD7
chr7	119 900 000	119 950 000	yes	Y	Y	Y	Y	Y	Y	KCND2	KCND2,TSPAN12
chr7	156 275 000	156 300 000	yes	Y	Y	W	Y	Y	Y	RBM33,SHH,RNF32,LMBR1	INSIG1,BLACE,EN2,CNPY1,RBM33,SHH,RNF32,LMBR1,NOM1,MNX1,UBE3C
chr8	4 900 000	4 925 000	yes	Y	Y	W	Y	N	W	CSMD1	CSMD1
chr8	12 750 000	12 775 000	yes	Y	Y	Y	Y	N	Y	DEFB130,FAM86B2,LONRF1,KIAA1456,DLC1	F1,KIAA1456,DLC1,C8orf48
chr8	13 500 000	13 525 000	yes	Y	Y	Y	Y	N	W	KIAA1456,DLC1,C8orf48	LONRF1,KIAA1456,DLC1,C8orf48
chr8	56 250 000	56 275 000	yes	Y	Y	Y	W	N	N	TMEM68,TGS1,LYN,RPS20,MOS,PLAG1,CHCHD7,SDR16C5,PENK	TMEM68,TGS1,LYN,RPS20,MOS,PLAG1,CHCHD7,SDR16C5,PENK,IMPAD1
chr8	84 725 000	84 750 000	yes	W	N	Y	Y	W	Y	RALYL,LRRCC1,E2F5,C8orf59,CA13	RALYL,LRRCC1,E2F5,C8orf59,CA13,CA1,CA3,CA2
chr8	105 500 000	105 525 000	yes	N	N	Y	Y	Y	W	ZFPM2	LRP12,ZFPM2
chr8	113 750 000	113 775 000	yes	W	N	Y	Y	Y	Y	CSMD3	CSMD3
chr8	114 500 000	114 525 000	yes	W	N	Y	W	Y	Y		

chr8	115 250 000	115 275 000	yes	W	N	Y	Y	Y	Y	TRPS1	TRPS1
chr8	139 150 000	139 175 000	yes	W	N	W	Y	Y	Y	COL22A1,KCNK9	FAM135B,COL22A1,KCNK9
chr9	4 300 000	4 350 000	yes	Y	Y	W	Y	N	W	GLIS3,SLC1A1,SPATA6L,PPAPDC2,CDC37L1,AK3,RCL1	AL365202.1,RFX3,GLIS3,SLC1A1,PPAPDC2,SPATA6L,CDC37L1,AK3,RCL1,JAK2,INSL6,INSL4,RLN2,RLN1
chr9	8 875 000	8 900 000	yes	Y	Y	Y	Y	N	W	PTPRD	
chr9	13 125 000	13 150 000	yes	Y	Y	Y	Y	W	Y	TYRP1,LURAP1L,MPDZ	TYRP1,LURAP1L,MPDZ
chr9	20 350 000	20 400 000	yes	Y	Y	W	Y	N	W	MLLT3,FOCAD	9,IFNA6,IFNA13,IFNA2
chr9	78 525 000	78 575 000	yes	Y	Y	Y	Y	N	W	GNAQ,CEP78,PSAT1	GNA14,GNAQ,CEP78,PSAT1,TLE4
chr9	108 500 000	108 525 000	yes	Y	Y	Y	Y	N	N	ACTL7B,ACTL7A,IKBKAP,FAM206A,CTNNAL1,TMEM245	ACTL7B,ACTL7A,IKBKAP,FAM206A,CTNNAL1,TMEM245,FRRS1L,EPB41L4B,PTPN3
chr9	118 500 000	118 525 000	yes	Y	Y	Y	Y	N	N	BRINP1	TLR4,BRINP1
chr9	119 475 000	119 500 000	yes	Y	Y	Y	Y	N	N	CDK5RAP2,MEGF9,FBXW2,PSMD5,PHF19	BRINP1
chr9	120 350 000	120 375 000	yes	Y	Y	W	Y	N	N		BRINP1,CDK5RAP2,MEGF9,FBXW2,PSMD5,PHF19,TRAF1,C5,CNTRL,RAB14,GSN,STOM
chr9	122 750 000	122 775 000	yes	Y	Y	Y	Y	N	Y	RBM18,MRRF,PTGS1,OR1J1,OR1J2,OR1J4,OR1N1,OR1N2,OR1L8,OR1Q1,OR1B1,OR1L1,OR1L3,OR1L4,OR1L6,OR5C1,PDCL,	R1L4,OR1L6,OR5C1,OR1K1,PDCL,RC3H2,ZBTB6,ZBTB26,RABGAP1,GPR21,STRBP,CRB2
chr10	20 150 000	20 175 000	yes	Y	Y	Y	Y	Y	Y	MALRD1,PLXDC2	PLXDC2,C10orf113,NEBL
chr10	21 000 000	21 025 000	yes	Y	Y	Y	Y	Y	Y	NEBL,C10orf113,CASC10,SKIDA1	C10orf113,NEBL,CASC10,SKIDA1,MLLT10,DNAJC1
chr10	107 500 000	107 525 000	yes	Y	Y	Y	Y	Y	Y	SORCS1	SORCS1
chr10	124 000 000	124 025 000	yes	Y	Y	Y	W	N	Y	GPR26,CPXM2,CHST15,OAT,NKX1-2,LHPP	IKZF5,ACADSB,HMX3,HMX2,BUB3,GPR26,CPXM2,CHST15,OAT,NKX1-2,LHPP,FAM53B,METTL10,FAM175B,ZRANB1
chr11	29 425 000	29 450 000	yes	Y	Y	Y	Y	N	N		KCNA4,FSHB,ARL14EP
chr11	81 250 000	81 275 000	yes	Y	Y	Y	Y	N	N		
chr11	129 250 000	129 300 000	yes	Y	Y	Y	Y	N	Y	FLI1,KCNJ1,KCNJ5,C11orf45,TP53AIP1,ARHGAP32,BARX2	ETS1,FLI1,KCNJ1,KCNJ5,C11orf45,TP53AIP1,ARHGAP32,BARX2,TMEM45B,NFRKB,PRDM10,APLP2,ST14
chr11	131 500 000	131 525 000	yes	Y	Y	Y	Y	N	N	NTM	SNX19,NTM
chr11	133 750 000	133 825 000	yes	Y	Y	W	Y	N	N	OPCML,SPATA19,IGSF9B,JAM3,NCAPD3,VPS26B,THYN1,ACAD8,GLB1L3	OPCML,SPATA19,IGSF9B,JAM3,NCAPD3,VPS26B,THYN1,ACAD8,GLB1L3,GLB1L2,B3GAT1
chr12	79 800 000	79 850 000	yes	Y	Y	W	Y	N	N	SYT1,PAWR,PPP1R12A,OTOGL	SYT1,PAWR,PPP1R12A,OTOGL,PTPRQ,MYF6,MYF5
chr12	98 400 000	98 425 000	yes	Y	Y	Y	Y	Y	Y	TMPO,SLC25A3,IKBIP,APAF1,ANKS1B	TMPO,SLC25A3,IKBIP,APAF1
chr13	26 150 000	26 175 000	yes	Y	Y	Y	Y	Y	N	ATP8A2,SHISA2,RNF6,CDK8,WASF3	AMER2,MTMR6,NUP1,ATP8A2,SHISA2,RNF6,CDK8,WASF3,GPR12,USP12
chr13	33 275 000	33 300 000	yes	Y	Y	W	Y	N	Y	PDS5B,KL,STARD13	ZAR1L,BRCA2,N4BP2L1,N4BP2L2,PDS5B,KL,STARD13,RFC3
chr13	67 250 000	67 275 000	yes	Y	Y	Y	Y	Y	Y	PCDH9	PCDH9
chr13	71 175 000	71 225 000	yes	Y	Y	Y	W	N	W	DACH1	DACH1
chr13	74 250 000	74 275 000	yes	Y	Y	Y	W	N	N	KLF12	KLF12
chr13	82 450 000	82 475 000	yes	Y	Y	Y	Y	N	W		
chr13	92 475 000	92 500 000	yes	Y	Y	Y	Y	N	Y	GPC5	GPC6
chr13	102 175 000	102 200 000	yes	Y	Y	Y	Y	N	N	ITGBL1,FGF14,TPP2,METTL21C	NALCN,ITGBL1,FGF14,TPP2,METTL21C,CCDC168,TEX30,KDELC1,BIVM,BIVM-ERCC5,ERCC5,SLC10A2
chr13	102 400 000	102 425 000	yes	Y	Y	Y	Y	N	N	FGF14,TPP2,METTL21C,CCDC168,TEX30,KDELC1,BIVM,BIVM-ERCC5,ERCC5	NALCN,ITGBL1,FGF14,TPP2,METTL21C,CCDC168,TEX30,KDELC1,BIVM,BIVM-ERCC5,ERCC5,SLC10A2
chr13	107 275 000	107 325 000	yes	Y	Y	W	Y	N	W	FAM155A	EFNB2,ARGLU1,FAM155A,LIG4,ABHD13,TNFSF13B
chr14	26 350 000	26 375 000	no	Y	Y	W	Y	W	N	NOVA1	NOVA1
chr14	30 025 000	30 050 000	yes	Y	Y	Y	Y	N	W	PRKD1	PRKD1,G2E3,SCFD1,COCH,AP4S1,STRN3
chr14	34 250 000	34 275 000	yes	Y	Y	Y	Y	N	N	NPAS3,EGLN3,SPTSSA,EAPP,SNX6,CFL2,BAZ1A	EGLN3,SPTSSA,EAPP,SNX6,CFL2,BAZ1A,SRP54,FAM177A1,PPP2R3C,KIAA0391
chr14	37 050 000	37 075 000	yes	W	N	W	W	Y	W	NKX2-8,PAX9,SLC25A21,MIPOL1	MBIP,SFTA3,NKX2-1,NKX2-8,PAX9,SLC25A21,MIPOL1,FOXA1,TTC6
chr14	86 500 000	86 525 000	yes	Y	Y	Y	N	N	Y		FLRT2
chr14	86 525 000	86 550 000	yes	Y	Y	Y	Y	N	Y		FLRT2
chr14	98 450 000	98 475 000	yes	Y	Y	Y	Y	N	W	C14orf177	C14orf177,BCL11B
chr15	27 100 000	27 150 000	yes	Y	Y	Y	Y	W	W	GABRB3,GABRA5,GABRG3	GABRB3,GABRA5,GABRG3,OCA2
chr15	54 075 000	54 100 000	yes	Y	Y	Y	W	N	W	WDR72,UNC13C	WDR72,UNC13C
chr15	61 250 000	61 275 000	yes	Y	Y	W	Y	N	W	RORA	ANXA2,ICE2,RORA,VPS13C,C2CD4A,C2CD4B
chr15	62 625 000	62 700 000	yes	Y	Y	Y	Y	N	Y	C2CD4B,TLN2,TPM1,LACTB,RPS27L,RAB8B	VPS13C,C2CD4A,C2CD4B,TLN2,TPM1,LACTB,RPS27L,RAB8B,APH1B,CA12,LISP3,FBXL22
chr15	71 350 000	71 375 000	yes	W	N	Y	N	Y	Y	LARP6,LRRC49,THAP10,THSD4,CT62,NR2E3,MYO9A	UACA,LARP6,LRRC49,THAP10,THSD4,CT62,NR2E3,SENP8,MYO9A,GRAMD2,PKM,PARP6,CELF6
chr15	100 900 000	100 925 000	yes	Y	Y	Y	Y	N	Y	CERS3,LINS,ASB7,AC087762.1,ALDH1A3,LRRK1,CHSY1,VIMP,SNRPA1,PCSK6	4
chr16	8 250 000	8 275 000	yes	Y	Y	Y	Y	N	Y	TMEM114,METTL22,ABAT	RBFOX1,TMEM114,METTL22,ABAT,TMEM186,PMM2,CARHSP1,USP7,C16orf72
chr16	12 250 000	12 275 000	yes	Y	Y	Y	W	N	W	ZC3H7A,RSL1D1,GSPT1,RP11-166B2.1,TNFRSF17,SNX29,CPPED1	166B2.1,TNFRSF17,SNX29,CPPED1,SHISA9
chr16	14 000 000	14 025 000	yes	Y	Y	Y	Y	N	N	ERCC4,MKL2,PARN	ERCC4,MKL2,PARN,BFAR,PLA2G10,NPIPA3,NPIPA2,NOMO1,NPIPA1,PDXDC1
chr16	55 250 000	55 275 000	yes	N	Y	Y	Y	Y	W	IRX5,IRX6,MMP2,LPCAT2,CAPNS2,SLC6A2	IRX3,IRX5,IRX6,MMP2,LPCAT2,CAPNS2,SLC6A2,CES1,CES5A,GNAO1
chr16	72 875 000	72 925 000	yes	Y	Y	Y	Y	N	N	AC004158.1,ZFHX3	ZNF821,IST1,PKD1L3,DHODH,HP,HPR,DHX38,TXNL4B,PMFBP1,AC004158.1,ZFHX3
chr17	12 150 000	12 175 000	yes	Y	Y	Y	Y	Y	W	DNAH9,ZNF18,MAP2K4,MYOCD	SHISA6,DNAH9,ZNF18,MAP2K4,MYOCD,ARHGAP44,ELAC2
chr18	5 875 000	5 900 000	yes	Y	Y	Y	Y	Y	Y	EPB41L3,TMEM200C,L3MBTL4	C18orf42,ZBTB14,EPB41L3,TMEM200C,L3MBTL4,ARHGAP28
chr18	7 250 000	7 275 000	yes	Y	Y	Y	Y	W	W	ARHGAP28,LAMA1,LRRC30,PTPRM	L3MBTL4,ARHGAP28,LAMA1,LRRC30,PTPRM
chr18	24 250 000	24 275 000	yes	W	N	Y	N	Y	Y	LAMA3,TTC39C,CABYR,OSBPL1A,IMPACT,HRH4	TMEM241,RIOK3,C18orf8,NPC1,ANKRD29,LAMA3,TTC39C,CABYR,OSBPL1A,IMPACT,HRH4
chr18	25 750 000	25 775 000	yes	Y	Y	Y	Y	Y	Y	ZNF521,SS18,PSMA8,TAF4B	ZNF521,SS18,PSMA8,TAF4B,KCTD1
chr18	28 750 000	28 775 000	yes	Y	Y	Y	Y	N	W		CDH2
chr18	30 725 000	30 800 000	yes	Y	Y	Y	Y	N	W	DSC3,DSC2,DSC1	DSC3,DSC2,DSC1,DSG1,DSG4,DSG3,DSG2,TTR,B4GALT6,SLC25A52
chr18	35 175 000	35 225 000	yes	Y	Y	Y	Y	N	N	DTNA,MAPRE2,ZNF397,ZSCAN30,ZNF24,ZNF396,RP11-322E11.6,JN080C,GALNT1	322E11.6,JN080C,GALNT1,C18orf21,RPRD1A,SLC39A6,ELP2,MOCOS
chr18	50 700 000	50 725 000	yes	N	N	Y	N	Y	Y	CFAP53,MBD1,CXXC1,SKA1,MAPK4,MRO,ME2,ELAC1,SMAD4,MEX3C	ACAA2,MYO5B,CFAP53,MBD1,CXXC1,SKA1,MAPK4,MRO,ME2,ELAC1,SMAD4,MEX3C
chr18	55 750 000	55 775 000	yes	W	N	Y	N	Y	Y	TCF4	RAB27B,CCDC68,TCF4,TXNL1,WDR7
chr18	75 425 000	75 475 000	yes	Y	Y	Y	Y	Y	Y	ZNF407,ZADH2,TSHZ1,SMIM21	FAM69C,CNDP2,CNDP1,ZNF407,ZADH2,TSHZ1,SMIM21
chr19	31 900 000	31 925 000	no	Y	Y	Y	Y	N	N	ZNF507,DPY19L3	TSHZ3,ZNF507,DPY19L3,PDCD5,ANKRD27,RGS9BP,NUDT19,TDRD12,SLC7A9
chr20	9 375 000	9 425 000	yes	Y	Y	Y	Y	N	Y	PLCB1,PLCB4,LAMP5,PAK7	PLCB4,LAMP5,PAK7,ANKEF1,SNAP25
chr20	10 650 000	10 700 000	yes	Y	Y	W	Y	N	W	SNAP25,MKKS,SLX4IP,JAG1	PAK7,ANKEF1,SNAP25,MKKS,SLX4IP,JAG1
chr20	12 850 000	12 875 000	yes	Y	Y	Y	Y	N	Y	SPTLC3,ISM1	BTBD3,SPTLC3,ISM1,TASP1,ESF1,NDUFAF5
chr20	14 325 000	14 350 000	yes	Y	Y	W	Y	N	Y	SEL1L2,MACROD2,FLRT3	TASP1,ESF1,NDUFAF5,SEL1L2,MACROD2,FLRT3
chr20	51 875 000	51 925 000	yes	Y	Y	Y	Y	Y	Y	NFATC2,ATP9A,SALL4,ZFP64	ADNP,DPM1,MOCS3,KCNG1,NFATC2,ATP9A,SALL4,ZFP64
chr20	59 825 000	59 850 000	yes	Y	Y	Y	Y	W	Y	EDN3,PHACTR3,SYCP2,FAM217B,PPP1R3D,CDH26	GNAS,NELFCD,CTSZ,TUBB1,ATP5E,SLMO2,ZNF831,EDN3,PHACTR3,SYCP2,FAM217B,PPP1R3D,CDH26
chr21	41 500 000	41 525 000	yes	Y	Y	Y	Y	N	W	BACE2,FAM3B,MX2,MX1,TMPRSS2,RIPK4,PRDM15,C2CD2,ZBTB21	A,RSPH1,SLC37A1
chr21	42 500 000	42 525 000	yes	Y	Y	Y	Y	W	N	ZBTB21,UMODL1,ABCG1,TFF3,TFF2,TFF1,TMPRSS3,UBASH3A,RSPH1,SLC37A1,PDE9A,WDR4,NDUFV3,PKNOX1	NDUFV3,PKNOX1,CBS,U2AF1,CRYAA,SIK1
chrX	3 550 000	3 575 000	yes	Y	Y	W	W	N	N	ARSF,MXRA5,PRKX	CD99,XG,GYG2,ARSD,ARSE,ARSH,ARSF,MXRA5,PRKX

Abbreviations: N, TAD bounday lost; W, weak boundary; Y, strong boundary

[a]CTCF/RAD21 binding sites within ±50kb around the boundary

Supplementary Data 15. Genes/proteins close to lost TAD boundaries that displayed significant differences in expression between high hyperdiploid and *ETV6/RUNX1* -positive acute lymphoblastic leukemias

Too large for printing

Supplementary Data 16. Chromosome morphology scores in childhood acute lymphoblastic leukemia

Case	Total number of cells	Average score	Number of cells with score 1 (%)	Number of cells with score 2 (%)	Number of cells with score 3 (%)
HeH_2	23	1.6	10 (43)	13 (57)	0
HeH_3	16	1.9	1 (6)	15 (94)	0
HeH_5	17	1.2	14 (82)	3 (18)	0
HeH_6	11	1.9	1 (9)	10 (91)	0
HeH_7	10	2	0	10 (100)	0
HeH_8	24	2.2	0	19 (79)	5 (21)
HeH_9	16	1.3	12 (75)	4 (25)	0
HeH_10	13	2.2	0	10 (77)	3 (23)
HeH_11	6	2	0	6 (100)	0
HeH_12	25	2.2	0	19 (76)	6 (24)
HeH_13	24	2.1	1 (4)	19 (79)	4 (17)
HeH_14	7	1.7	2 (29)	5 (71)	0
HeH_15	8	1	8 (100)	0	0
HeH_16	7	2	0	7 (100)	0
HeH_17	18	2.3	0	13 (72)	5 (28)
HeH_19	11	2	3 (27)	5 (46)	3 (27)
HeH_21	4	1.5	2 (50)	2 (50)	0
HeH_22	20	1.3	14 (70)	6 (30)	0
HeH_23	11	1.3	8 (73)	3 (27)	0
HeH_24	15	1.4	9 (60)	6 (40)	0
HeH_25	5	2	0	5 (100)	0
HeH_26	11	1.7	3 (27)	8 (73)	0
HeH_27	10	1.4	6 (60)	4 (40)	0
HeH_28	5	1.8	1 (20)	4 (80)	0
HeH_29	13	2.2	0	11 (85)	2 (15)
HeH_30	5	1.6	2 (40)	3 (60)	0
HeH_31	4	1.5	2 (50)	2 (50)	0
HeH_32	4	1.3	3 (75)	1 (25)	0
HeH_33	6	1.5	3 (50)	3 (50)	0
HeH_38	12	1.4	7 (59)	5 (41)	0
HeH_40	9	2.3	0	6 (66)	3 (34)
HeH_42	7	2.7	0	2 (29)	5 (71)
HeH_43	7	2.1	1 (14)	4 (58)	2 (28)
HeH_44	9	2.2	0	7 (78)	2 (22)
HeH_45	23	2.6	0	10 (48)	13 (52)
HeH_47	18	1.4	11 (61)	7 (39)	0
HeH_48	29	2.8	0	7 (24)	22 (76)
ETV6/RUNX1_1	6	2.2	1 (17)	3 (50)	2 (33)
ETV6/RUNX1_3	5	2.8	0	1 (20)	4 (80)
ETV6/RUNX1_4	17	1.4	10 (59)	7 (41)	0
ETV6/RUNX1_5	31	2.1	1 (3)	27 (87)	3 (10)
ETV6/RUNX1_6	4	2.3	0	3 (75)	1 (25)
ETV6/RUNX1_7	17	1.9	2 (12)	15 (88)	0
ETV6/RUNX1_8	27	1.7	9 (33)	18 (67)	0
ETV6/RUNX1_9	24	2.1	0	21 (88)	3 (12)
ETV6/RUNX1_10	6	1.7	2 (33)	4 (67)	0
ETV6/RUNX1_11	25	2.8	0	4 (16)	21 (84)
ETV6/RUNX1_12	17	1.5	8 (47)	9 (53)	0
ETV6/RUNX1_13	7	2	0	7 (100)	0
ETV6/RUNX1_14	8	2.3	1 (12)	4 (50)	3 (38)
ETV6/RUNX1_16	8	2	0	8 (100)	0
ETV6/RUNX1_18	20	2.1	1 (5)	16 (80)	3 (15)
ETV6/RUNX1_19	6	2.8	0	1 (17)	5 (83)
ETV6/RUNX1_20	7	1.6	3 (43)	4 (58)	0
ETV6/RUNX1_21	11	2.4	0	7 (64)	4 (36)
ETV6/RUNX1_22	6	2.8	0	1 (17)	5 (83)
ETV6/RUNX1_23	9	2.1	0	8 (88)	1 (12)
ETV6/RUNX1_24	16	1.8	4 (25)	11 (69)	1 (6)
ETV6/RUNX1_25	7	2.7	0	2 (29)	5 (71)
ETV6/RUNX1_27	8	2.5	0	4 (50)	4 (50)
ETV6/RUNX1_30	25	1.6	10 (40)	14 (56)	1 (4)
ETV6/RUNX1_31	20	1.9	2 (10)	18 (90)	0
ETV6/RUNX1_33	12	2	1 (8)	10 (83)	1 (9)
ETV6/RUNX1_34	8	2.5	0	4 (50)	4 (50)
ETV6/RUNX1_35	21	2.1	0	19 (91)	2 (9)
ETV6/RUNX1_36	4	2.3	0	3 (75)	1 (25)
ETV6/RUNX1_38	16	1.9	4 (25)	10 (63)	2 (12)
ETV6/RUNX1_39	5	1.6	2 (40)	3 (60)	0
ETV6/RUNX1_40	11	1.7	3 (28)	8 (72)	0
ETV6/RUNX1_41	26	1.9	3 (11)	23 (89)	0

Article II

Received: 19 October 2020 | Revised: 18 December 2020 | Accepted: 20 December 2020
DOI: 10.1002/gcc.22933

RESEARCH ARTICLE

WILEY

Sister chromatid cohesion defects are associated with chromosomal copy number heterogeneity in high hyperdiploid childhood acute lymphoblastic leukemia

Larissa H. Moura-Castro[1] | Pablo Peña-Martínez[1] | Anders Castor[2] | Roman Galeev[3] | Jonas Larsson[3] | Marcus Järås[1] | Minjun Yang[1] | Kajsa Paulsson[1]

[1]Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, Lund, Sweden

[2]Department of Pediatrics, Skåne University Hospital, Lund University, Lund, Sweden

[3]Division of Molecular Medicine and Gene Therapy, Lund Stem Cell Center, Lund University, Lund, Sweden

Correspondence
Dr. Kajsa Paulsson, Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, BMC C13, SE-221 84 Lund, Sweden.
Email: kajsa.paulsson@med.lu.se

Funding information
Barncancerfonden, Grant/Award Numbers: PR2015-0012, TJ2016-0063; Cancerfonden, Grant/Award Number: CAN 2016/497; Crafoordska Stiftelsen, Grant/Award Number: 20180529; Governmental funding of clinical research within the National Health Service, Grant/Award Number: ALFSKANE-623431; Vetenskapsrådet, Grant/Award Number: 2016-01459

Abstract

High hyperdiploid acute lymphoblastic leukemia (ALL) is one of the most common malignancies in children. The main driver event of this disease is a nonrandom aneuploidy consisting of gains of whole chromosomes but without overt evidence of chromosomal instability (CIN). Here, we investigated the frequency and severity of defective sister chromatid cohesion—a phenomenon related to CIN—in primary pediatric ALL. We found that a large proportion (86%) of hyperdiploid cases displayed aberrant cohesion, frequently severe, to compare with 49% of *ETV6/RUNX1*-positive ALL, which mostly displayed mild defects. In hyperdiploid ALL, cohesion defects were associated with increased chromosomal copy number heterogeneity, which could indicate increased CIN. Furthermore, cohesion defects correlated with *RAD21* and *NCAPG* mRNA expression, suggesting a link to reduced cohesin and condensin levels in hyperdiploid ALL. Knockdown of *RAD21* in an ALL cell line led to sister chromatid cohesion defects, aberrant mitoses, and increased heterogeneity in chromosomal copy numbers, similar to what was seen in primary hyperdiploid ALL. In summary, our study shows that aberrant sister chromatid cohesion is frequent but heterogeneous in pediatric high hyperdiploid ALL, ranging from mild to very severe defects, and possibly due to low cohesin or condensin levels. Cases with high levels of aberrant chromosome cohesion displayed increased chromosomal copy number heterogeneity, possibly indicative of increased CIN. These abnormalities may play a role in the clonal evolution of hyperdiploid pediatric ALL.

KEYWORDS
acute lymphoblastic leukemia, aneuploidy, chromosomal instability, hyperdiploidy, sister chromatid cohesion

1 | INTRODUCTION

Changes in the number of chromosomes, termed aneuploidy, is one of the most common genetic aberrations in cancer cells. In spite of this, many questions remain regarding its impact on the cell and its role in tumorigenesis. One such controversial issue is whether aneuploidy is always associated with chromosomal instability (CIN), that is an increased rate of missegregation of chromosomes at mitosis.[1]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2020 The Authors. *Genes, Chromosomes & Cancer* published by Wiley Periodicals LLC.

The high hyperdiploid (HeH; 51-67 chromosomes) subgroup of B-cell precursor acute lymphoblastic leukemia (BCP ALL) comprises 25% to 30% of all pediatric cases and is associated with young age (3-5 years) at diagnosis and a superior prognosis.[2] Genetically, HeH leukemia is characterized by a specific aneuploidy, comprising gains of chromosomes X, 4, 6, 10, 14, 17, 18 and 21, which is believed to be the main driver event because it occurs very early in leukemogenesis and is always present in all leukemic cells.[3-7] In contrast to many other aneuploid malignancies, there is little evidence for CIN in HeH ALL; extra chromosomes are rarely subclonal and the chromosomal pattern generally does not change over the course of the disease.[3,6,8,9] Thus, HeH ALL appears to be a chromosomally stable aneuploid disease, suggesting that aneuploidy is not necessarily associated with CIN per se. However, underlying CIN that is masked by stable dividing major clones cannot be excluded.

We recently reported that HeH childhood BCP ALL displays relatively low expression of cohesin.[10] Cohesin is a ring-like multi-protein complex that includes SMC1/SMC3 heterodimers, RAD21, and a STAG1 or STAG2 subunit.[11,12] One of its main functions is to mediate sister chromatid cohesion at metaphase; a process that, if disrupted, may cause CIN.[1] Individuals with constitutional mutations in cohesin components, which cause Cornelia de Lange syndrome, have increased levels of cohesion defects[13] and knockdown experiments in human cell lines have shown that loss of expression of *SMC1A*, *SMC3*, *STAG2*, and *RAD21* are all associated with cohesion defects to different degrees.[14-22] In cancer, mutations in components of the cohesin complex are recurrent in a wide variety of malignancies, including acute myeloid leukemia, but conflicting results have been reported regarding their link to cohesion defects and aneuploidy.[23-27] In line with this, a recent study has shown that primary HeH ALL cases present chromatid cohesion defects in addition to delays in early mitosis and chromosome-alignment defects.[28] The authors found an association between such aberrations and defective condensin complexes, aurora B kinase and the spindle assembly checkpoint[28] but did not address the possible link with cohesin levels.

In this study, we have investigated the severity and frequency of aberrant sister chromatid cohesion in primary HeH ALL, and how it may affect chromosomal heterogeneity within the HeH subgroup. We found that primary HeH frequently have severe cohesion defects in metaphase chromosomes that are associated with increased chromosomal copy number heterogeneity, indicating that a subset of HeH ALL may possibly harbor CIN. Our data point to a novel opportunity for targeted therapy in HeH ALL, in line with other cancers with cohesion defects.

2 | MATERIALS AND METHODS

2.1 | Patient samples

Eighty-two childhood ALL cases were included in the study (Supporting Information Table S1), selected on the basis of material being available. Forty-five cases displayed HeH as ascertained by G-banding, fluorescence in situ hybridization (FISH) and/or single nucleotide polymorphism (SNP) array analysis, whereas 37 were *ETV6/RUNX1*-positive by reverse transcriptase quantitative PCR (RT-PCR) and/or FISH. We have previously shown that the latter generally display normal levels of cohesin.[10] Diagnostic samples obtained at ALL diagnosis and stored in fixative (methanol:acetic acid, 3:1) for 1 to 31 years were utilized. Informed consent was obtained according to the Declaration of Helsinki and the study was approved by the Ethics Committee of Lund University. SNP array data from samples obtained at diagnosis were available for all cases except HeH_29 and have been previously published.[3,4] These data were reanalyzed to identify subclonal chromosome gains using TAPS.[29] The lower limit of detecting subclonality using this technique was estimated to be approximately 10% to 20% of the cells.

2.2 | *RAD21* knockdown

To investigate the effect of lower levels of cohesin, we performed shRNA-mediated knockdown of *RAD21* (RAD21-KD) in the *ETV6/RUNX1*-positive ALL REH cell line (ACC-22, DSMZ, Braunschweig, Germany) according to methods previously described.[26] Briefly, lentivirus carrying the shRNA construct and expressing GFP were produced using the human cell line 293 T.[26,30] REH cells were cultured in standard cell medium (RPMI, 20% FBS, 1% P/S) and transduction was done with two different shRNAs targeting *RAD21* (RAD21 shRNA-1 and RAD21 shRNA-2, respectively) as well as a non-targeting shRNA control, with three replicates for each. GFP+ cells were sorted by FACS 48 hours post-transduction. Gene expression levels were determined 1 week after transduction by RT-PCR (7500 Real-Time PCR system; Applied Biosystems, Waltham, MS), using probes from Taqman (Life Technologies, Carlsbad, CA) for RAD21 (Hs00366721_mH) and HPRT1 (Hs02800695_m1). Cytogenetic analysis was performed according to standard methods 2 to 3 and immunofluorescence 4 to 5 weeks post-transduction.

2.3 | Cohesion assay

Sister chromatid cohesion assay was performed according to Sajesh et al[15] in the RAD21-KD REH cells and controls (blinded analysis) and in the primary patient samples. For the latter, the analysis was done without prior knowledge of expression levels of members of the cohesin or condensin complexes. FISH or standard G-banding preparations were analyzed using a Z2 fluorescence microscope (Zeiss, Oberkochen, Germany). Images were captured and enhanced using the CytoVision software (Leica, Wetzlar, Germany). Aberrant cohesion was defined as the presence of primary constriction gaps (PCGs), that is, visible gaps between the sister chromatids at the centromeres.[15] The severity of PCGs was classified based on literature,[15] with the addition of a fourth category: (a) PCG-I (mild), 1-4 chromosomes with PCGs; (b) PCG-II (moderate) 5-19 chromosomes with PCGs; (c) PCG-III (severe) ≥20 chromosomes with PCGs, but not all; and (d) PCG-IV (very severe), complete loss of cohesion. Since *ETV6/RUNX1* is cryptic, leukemic blasts were identified by simultaneous detection of the translocation by FISH or by additional chromosomal aberrations identifiable by chromosome banding. The frequency of cohesion defects in the different subgroups was compared using the Mann-Whitney

U test; *P*-values <.05 were considered significant. Furthermore, the overall frequency of chromosomes with cohesion defects per patient was calculated taking the number of chromosomes with cohesion defects and dividing it by the total number of chromosomes times the number of analyzed cells.[15]

2.4 | Immunoflourescence

The frequency of aberrant mitoses in REH cells under *RAD21* knockdown was investigated in a blinded manner using the RAD21-KD cells and controls. Slide preparation was performed according to the guidelines "Cell Staining for Immunofluorescence Microscopy" provided by BD Biosciences (Franklin Lakes, NJ), with minor modifications. A minimum of 100 mitoses were analyzed for each replicate, and statistical significance was evaluated by the Mann-Whitney *U* test; *P*-values <.05 were considered significant. Primary antibodies anti-tubulin alpha (produced in rabbit, SAB4500087; Sigma–Aldrich, St. Louis, MO) and anti-tubulin gamma (produced in mouse, T6557; Sigma–Aldrich) were used to label microtubuli and centrosomes, respectively. Samples were counterstained with anti-rabbit IgG (FITC green, F1262; Sigma–Aldrich) and anti-mouse IgG (Cy3 orange, C2181; Sigma–Aldrich) and slides were mounted with Vectashield medium with DAPI (H-1200; Vector Laboratories, Burlingame, CA).

2.5 | Interphase FISH

Interphase FISH was done in a blinded manner on RAD21-KD cells and controls according to standard methods. Slides from each replicate were hybridized with FISH probes for chromosomes X, 2, 3 and 21 (Vysis, Abbott Laboratories, Chicago, IL). A total of 300 nuclei were analyzed for each replicate, with each probe counted separately. Interphase FISH on HeH primary samples was done in a blinded manner in >300 nuclei in five cases with high percentage and five cases with low percentage of cohesion defects, all chosen according to availability of material, using probes for chromosomes X, 2, 3, 6, 10, and 21 (Vysis). To minimize technical artefacts, only cases with the same copy number for the analyzed chromosome in the major clone were included in each analysis. To ensure that only leukemic blasts were analyzed, only nuclei where the other probes confirmed hyperdiploidy were included.

2.6 | Gene expression correlation

Data from RNA-sequencing were available for 36 of the HeH and 32 *ETV6/RUNX1*-postive cases (Supporting Information Table S1).[31] For each cohesin- and condensin-related gene, HeH cases were divided into two groups: high expression and low expression (top and bottom 50% of cases for the given gene); one-sided Mann-Whitney *U* test was applied to inquire whether the low expression groups presented higher levels of cohesion defects. Genes included in the analysis were core subunits of the cohesin and condensin complexes, that is, *RAD21*, *SMC1A*, *SMC3*, *STAG1*, *STAG2* (cohesin), *NCAPD2*, *NCAPD3*, *NCAPG*, *NCAPG2*, *NCAPH*, *NCAPH2*, *SMC2* and *SMC4* (condensin I and II); $P < .05$ was considered significant.

3 | RESULTS

3.1 | Primary patient samples of HeH ALL have aberrant sister chromatid cohesion

To investigate whether primary ALL samples display cohesion defects, we first ensured that we could detect cohesion defects by analyzing the *RAD21*-knockdown cells (2 to 2.5-fold decrease in *RAD21* expression; Supporting Information Figure S1). The proportion of mitotic cells displaying cohesion defects was clearly higher in *RAD21*-knockdown cells: 36% and 33%, respectively, of two technical replicates vs 4% of control cells ($P = 0.0119$, Mann-Whitney *U* test; Supporting Information Figure S2A-B and Table S2). RAD21-KD cells also displayed more severe defects, ranging from PCG I-III, while controls presented only PCG I (Supporting Information Table S2). These results are in line with previous reports of cohesin knockdown,[14,16] and shows that our analysis is able to detect sister chromatid cohesion defects.

Cohesion defects were detected in 86% of HeH cases vs 49% of *ETV6/RUNX1*-positive cases ($P = 3.02 \times 10^{-8}$; Mann-Whitney *U* test; Figure 1, Supporting Information Table S3). In HeH cases, 0% to 85% of the cells displayed cohesion defects (Figure 1E), with mild defects (PCG I) seen in 27%, moderate defects (PCG II) in 40%, and severe defects (PCG III) in 11% of cases (Figure 1). Complete loss of cohesion was observed in two cells from case HeH_14 and in one cell each from cases HeH_18 and HeH_45 (7% of cases). In *ETV6/RUNX1*-positive cases, cohesion defects were seen in 0% to 18% of cells, with 30% of cases classified as mild, 14% as moderate and 5% as severe. Complete loss of cohesion was not observed. Taken together, the distribution between the categories (PCG I-PCG IV) clearly differed between the two subtypes (Figure 1F). Furthermore, the frequency of chromosomes displaying cohesion defects per case (based on each case's modal chromosome number) was significantly higher in HeH ALL than in *ETV6/RUNX1*-positive ALL ($P = 4.74 \times 10^{-8}$, Mann-Whitney *U* test; Supporting Information Table S3). Remission samples from twelve HeH and eight *ETV6/RUNX1*-positive cases were also analyzed to investigate the frequency of cohesion defects in normal cells. Of 210 analyzed normal cells, only one cell in one remission sample had mild cohesion defects (0.48%). Thus, sister chromatid cohesion defects were both significantly more common and more severe in primary HeH as compared with primary *ETV6/RUNX1*-positive ALL and normal bone marrow cells.

3.2 | Cohesion defects are associated with increased chromosomal heterogeneity in primary HeH ALL

Next, we investigated whether cohesion defects lead to increased chromosomal heterogeneity, which is indicative of CIN, in primary

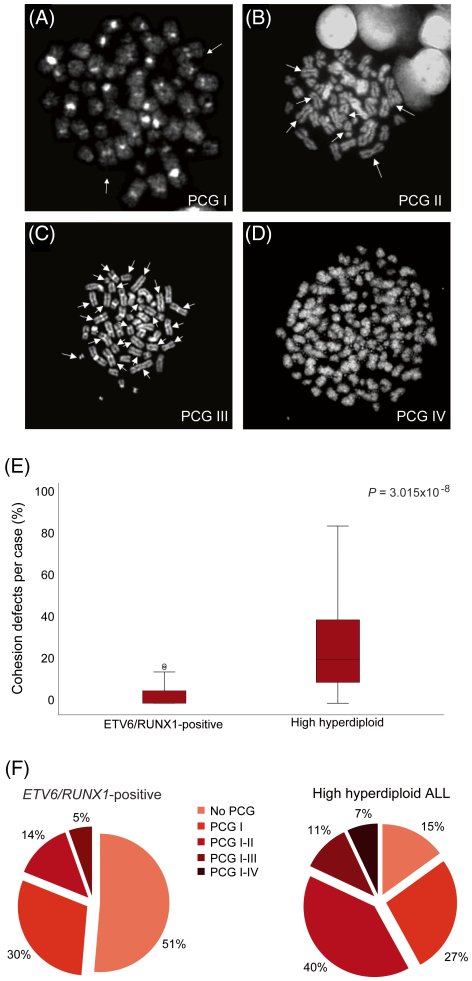


FIGURE 1 Analysis of sister chromatid cohesion, detected as primary constriction gaps (PCGs), in primary patient samples from high hyperdiploid and *ETV6/RUNX1*-positive ALL cases. A, Metaphase from case HeH_3, classified as PCG I, where two chromosomes (arrows) displayed cohesion defects; B, metaphase from case HeH_40, classified as PCG II, where seven chromosomes (arrows) displayed cohesion defects; C, metaphase from case HeH_26, classified as PCG III, where 22 chromosomes (arrows) displayed cohesion defects; D, metaphase from case HeH_14, classified as PCG IV, that is, complete loss of cohesion. E, Incidence of cohesion defects, comparing high hyperdiploid and *ETV6/RUNX1*-positive ALL cases: cohesion defects, shown as percentage of total cells presenting PCGs per case, where the boxes show the interquartile range and median (line) values, whiskers show minimum and maximum values in the cohort and outliers are shown as circles; F, classification of *ETV6/RUNX1*-positive and high hyperdiploid cases according to the cohesion assay criteria, from "no PCG" to "PCG I-IV"

ALL by interphase-FISH. There was an increase in copy number variation for all the six assessed chromosomes in cases with high levels of cohesion defects, being statistically significant for chromosomes 3, 6, 10 and 21 (P = .0357, P = .00794, P = .0286, and P = .0179, respectively; Figure 2, Supporting Information Table S4). Thus, sister chromatid cohesion defects were associated with increased chromosomal copy number heterogeneity in HeH ALL, indicating that cohesion defects may lead to increased CIN.

To investigate whether sister chromatid cohesion defects were also associated with the number of subclones involving whole chromosomes in primary HeH ALL, SNP array data was analyzed in 44 of the HeH cases. Subclonality was detected for 1 to 4 chromosomes in 18 cases (40%). No difference in the frequency of cohesion defects was seen between cases with and without subclonal chromosome changes (P = .747; Mann-Whitney U test).

3.3 | Cohesion defects are associated with decreased *RAD21* and *NCAPG* expression in primary HeH ALL

To investigate whether sister chromatid cohesion defects could be linked to cohesin or condensin levels, we analyzed whether mRNA expression of core subunits from both protein complexes correlated with the percentage of cells displaying cohesion defects. Consistent with our hypothesis, the expression of *RAD21* was negatively correlated with the number of cells displaying cohesion defects in HeH ALL (P = .00111; one-sided Mann-Whitney U test; Supporting Information Figure S3C). Moreover, low *NCAPG* levels also correlated with aberrant cohesion (P = .00424; one-sided Mann-Whitney U test; Supporting Information Figure S4C). No other correlations were seen in the HeH cases only (Supporting Information Figure S3, Supporting Information Figure S4), the *ETV6/RUNX1*-positive cases only, or in both groups combined.

3.4 | Knockdown of *RAD21* leads to increased chromosomal heterogeneity and aberrant mitoses in leukemic cells

To assess further whether the low levels of cohesin affect chromosomal stability in hematopoietic cells, we performed interphase FISH in the RAD21-KD cells. For all four investigated chromosomes, RAD21-KD cells displayed increased variation in copy number; however, it was only statistically significant for chromosome 21 (Figure 3, Supporting Information Table S5). Taken together, the interphase FISH showed increased chromosomal heterogeneity in RAD21-KD cells, indicating that decreased levels of cohesin lead to increased CIN.

Next, we investigated whether low expression of *RAD21* affects cell division in ALL cells by analyzing for spindle defects—monopolar, tripolar, and tetrapolar mitoses—and chromatin bridges/lagging

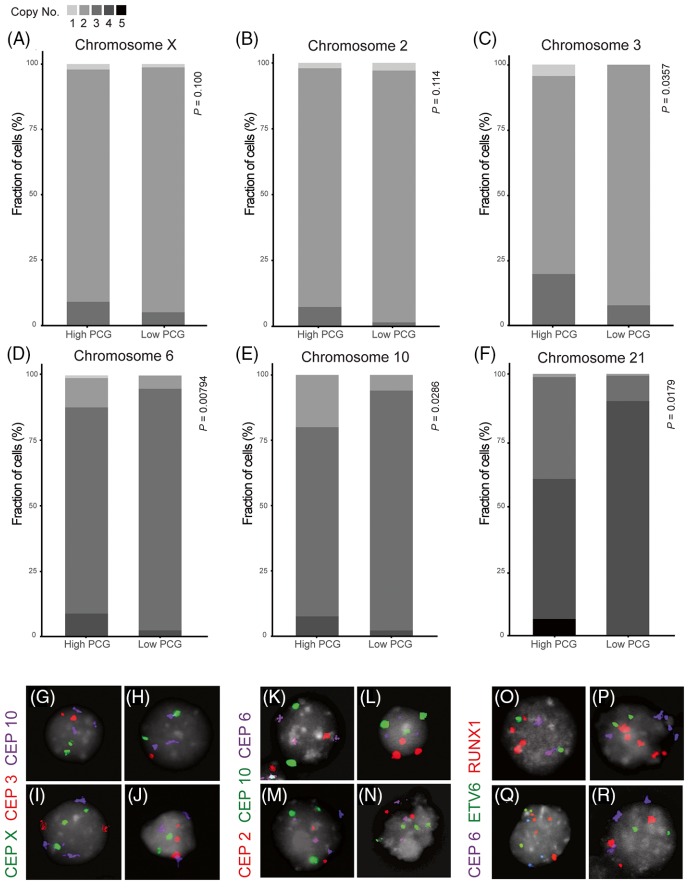


FIGURE 2 Copy number in HeH ALL primary patient cells with high PCG (primary constriction gaps) (HeH_4, HeH_14, HeH_18, HeH_40 and HeH_41) and low PCG (HeH_21, HeH_23, HeH_30, HeH_39 and HeH_42), analyzed by interphase fluorescence in situ hybridization for chromosomes X, 2, 3, 6, 10 and 21. A, Overall copy number of chromosome X, disomy expected; B, overall copy number of chromosome 2, disomy expected; C, overall copy number of chromosome 3, disomy expected; D, overall copy number of chromosome 6, trisomy expected; E, overall copy number of chromosome 10, trisomy expected; F, overall copy number of chromosome 21, tetrasomy expected; G, nucleus from HeH_42, showing disomy of chromosome X, disomy 3 and trisomy 10; H, nucleus from HeH_14, showing disomy X, monosomy 3 and trisomy 10; I, nucleus from HeH_40, showing trisomy X, disomy 3 and trisomy 10; J, nucleus from HeH_14, showing disomy X, trisomy 3 and trisomy 10; K, nucleus from HeH_14, showing monosomy 2,trisomy 6 and disomy 10; L, nucleus from HeH_41, showing trisomy 2, trisomy 6 and trisomy 10; M, nucleus from HeH_30, showing disomy 2, tetrasomy 6 and trisomy 10; N, nucleus from HeH_30, showing disomy 2, trisomy 6 and tetrasomy 10; O, nucleus from HeH_18, showing disomy 6, two copies of *ETV6* (chromosome 12) and four copies of *RUNX1* (chromosome 21); P, nucleus from HeH_40, showing tetrasomy 6, two copies of *ETV6* and five copies of *RUNX1*; Q, nucleus from HeH_18, showing trisomy 6, three copies of chromosome *ETV6* and five copies of *RUNX1*; R, nucleus from HeH_40, showing trisomy 6, two copies of *ETV6* and three copies of *RUNX1*

chromosomes (Supporting Information Figure S2C-F, Table S6). The overall frequency of mitotic aberrations in RAD21-KD cells was 6.5%, while no control cells had such aberrations (P = .0119, Mann-Whitney U test; Supporting Information Figure S2G). Spindle defects were detected in 4.8% of RAD21-KD cells; in particular, tripolar mitoses were more frequent (P = .0119; Mann-Whitney U test). Chromatin bridges/lagging chromosomes were only seen in RAD21-KD cells (1.6% of the cells). Taken together, lower expression of RAD21 increased the frequency of spindle defects, in particular tripolar mitoses.

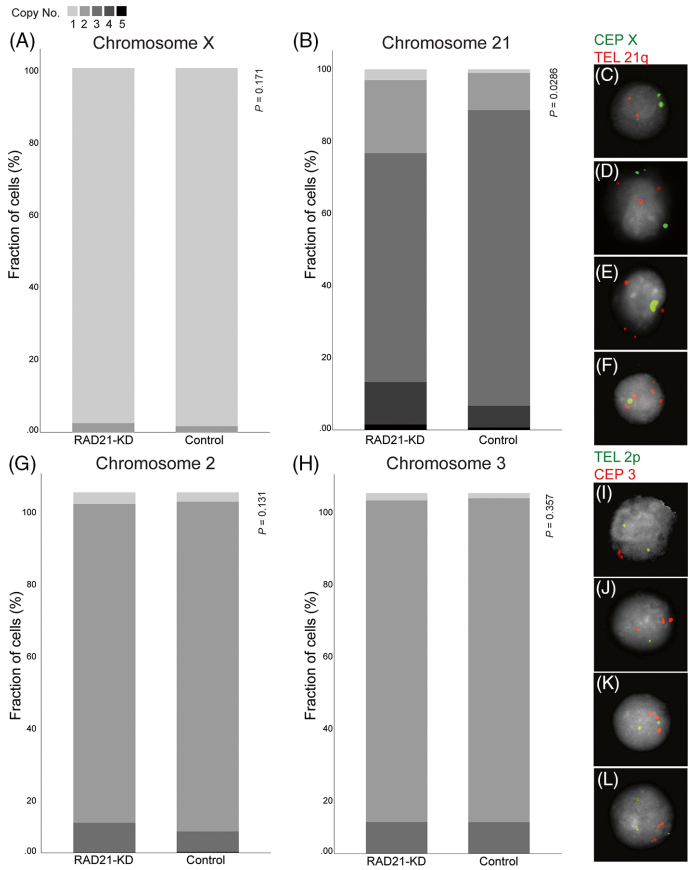


FIGURE 3 Copy number in REH cells with low expression of *RAD21* (RAD21-KD cells) and controls, analyzed by interphase fluorescence in situ hybridization for chromosomes X, 2, 3 and 21. A, Overall copy number of chromosome X, monosomy expected; B, overall copy number of chromosome 21, trisomy expected; C, nucleus from replicate RAD21.1-3, showing one copy of chromosome X and two copies of chromosome 21; D, nucleus from replicate RAD21.2-3, showing two copies of chromosome X and three copies of chromosome 21; E, nucleus from replicate RAD21.2-2, showing one copy of chromosome X and four copies of chromosome 21; F, nucleus from replicate RAD21.1-3, showing one copy of chromosome X and five copies of chromosome 21; G, overall copy number of chromosome 2, disomy expected; H, overall copy number of chromosome 3, disomy expected; I, nucleus from replicate RAD21.2-4, showing two copies of chromosome 2 and one copy of chromosome 3; J, nucleus from replicate RAD21.2-5, showing one copy of chromosome 2 and three copies of chromosome 3; K, nucleus from replicate control 5, showing two copies of chromosome 2 and three copies of chromosome 3; L, nucleus from replicate control 5, showing two copies of chromosome 2 and two copies of chromosome 3

4 | DISCUSSION

In this study, we found that HeH ALL cells frequently harbor aberrant cohesion, in line with the recent investigation by Molina et al.[28] In a further expansion of their findings, we report that the incidence and severity of defects vary within the subgroup. HeH ALL displayed both a higher frequency and more severe cohesion defects compared to *ETV6/RUNX1*-positive cases. Notably, the percentage of aberrant cohesion ranged from 0% to 85% in primary HeH ALL, where 40% of the cases presented moderate and 18% presented severe or very severe cohesion defects, showing that aberrant sister chromatid cohesion is a widespread, yet heterogeneous, phenomenon in HeH childhood ALL.

Sister chromatid cohesion defects may be associated with CIN, which in turn may give rise to increased heterogeneity in chromosomal copy numbers. Whether HeH ALL display CIN or not is a controversial issue. Some investigators have reported widespread chromosomal heterogeneity when using interphase FISH to analyze commonly gained chromosomes.[28,32-34] However, a major problem with these studies is that they have compared HeH samples containing trisomies with normal or other BCP ALL samples containing disomies. Since the baseline number of the chromosomes differ between these two groups, appropriate cut-off levels for the probes cannot be determined and solid data are thus lacking. Here, we circumvented this problem by comparing two groups of HeH cases, only including cases with the same copy number for the analyzed chromosomes. We detected a clear difference between cases with high levels of sister chromatid cohesion defects and those with low levels/no cohesion defects. Thus, we can conclude that aberrant sister chromatid cohesion result in increased levels of chromosomal copy number heterogeneity. Although the link between chromosomal copy number heterogeneity and CIN is not absolute, these findings suggest that cohesion defects are associated with increased CIN. However, this did not translate into an increased number of subclones detectable by SNP array analysis in cases with high levels of cohesion defects, likely due to the lower limit of detection of subclonal trisomies with this

method being approximately 10-20%, preventing detection of smaller subclones. Molina et al[28] reported that inhibition of AURKB, suggested to be the functional outcome of defective condensin in their study, in CD34-positive hematopoietic stem/progenitor cells led to an increase in the number of hyperdiploid cells, although no further characterization of the exact chromosomal content of these cells was done. However, considering that we show that not all HeH cases display cohesion defects, regardless of the underlying cause, it does seem unlikely that these are directly causative of the aneuploidy, in particular as the allelic patterns in HeH ALL suggest that the majority of extra chromosomes are gained in one abnormal cell division.[35-37] Rather, sister chromatid cohesion defects may promote clonal evolution in HeH ALL through increased chromosomal heterogeneity. Taken together, since the incidence of cohesion defects as well as the level of chromosomal heterogeneity varies among HeH cases, these phenomena are likely not early events in leukemogenesis, but could rather have a role in later optimization of the chromosomal gains once the initial hyperdiploidy has been established.

We recently reported that HeH ALL display low levels of cohesin compared to *ETV6/RUNX1*-positive cases and normal BCP cells.[10] This was shown on the protein level as well in multiple mRNA datasets and impacted the overall gene expression. Since low cohesin levels would be expected to also result in aberrant cohesion, we investigated whether the cohesion defects described here in primary HeH ALL correlated with the expression of cohesin subunits in HeH ALL. We found a negative correlation between *RAD21* mRNA expression and the incidence of cohesion defects, in agreement with prior in vitro studies,[14,15] whereas no statistically significant correlation was seen for the remaining genes. Another recent study investigated the incidence of mitotic and chromosomal defects in HeH ALL compared to other B-ALL subgroups, suggesting that impairment of aurora B kinase and the condensin complex were underlying such abnormalities.[28] Although the authors observed no correlation at the mRNA expression level of condensin, suggesting that posttranslational modifications were likely to be the cause of condensin impairment, we here found that *NCAPG* mRNA expression correlates with cohesion defects within the HeH ALL subgroup. Taken together, we cannot definitely state whether dysregulation of cohesin, condensin, or a combination of both causes cohesion defects in primary ALL.

There are conflicting data in the literature on the link between cohesin dysregulation and aneuploidy. Solomon et al[18,25] reported increased variability in chromosome numbers in cell lines with knockdown of *STAG2*, whereas Balbás-Martínez et al[38] did not observe such effects. Using interphase FISH, we detected increased chromosomal heterogeneity for chromosome 21, the only investigated trisomy, in REH leukemic cells with knockdown of *RAD21*, but not for chromosomes X, 2 and 3. Whether this discrepancy is due to an underlying chromosome-specific effect—as recently shown to exist for certain CIN-associated phenomena[39]—or to the fact that more copy number variation can be expected for trisomic chromosomes, simply because there are more copies that can be affected, remain to be investigated. Taken together, our data support that low expression of *RAD21* compromises the integrity of chromosome segregation in BCP ALL cells, at least for some chromosomes.

Recent studies have suggested possible agents for targeted therapy in cohesion-defective cancers, based on synthetic lethality experiments in cells with aberrant cohesion. In particular, inhibition of the anaphase promoting complex in the presence of aberrant cohesion has been shown to have synthetic lethality, leading to mitotic death.[40] Furthermore, synthetic lethality has been described for cohesin defects and poly-ADP ribose polymerases (PARP)—a protein involved in double-stranded DNA repair—where cell lines under siRNA-mediated depletion of *SMC1*, *SMC3* or *RAD21* showed increased sensitivity to the PARP-inhibitor olaparib.[41] Thus, considering our data showing that primary samples have cohesion defects, such treatments could be a possible future option in at least a subset of HeH ALL.

CONFLICT OF INTEREST

The authors declare no potential conflicts of interest.

DATA AVAILABILITY STATEMENT

The RNA dataset used in this study is available at the European Genome-phenome archive under accession number EGAD00001002112. SNP array data are not publicly available due to privacy concerns but are available from the corresponding author on reasonable request.

ORCID

Larissa H. Moura-Castro https://orcid.org/0000-0001-9063-5592
Pablo Peña-Martínez https://orcid.org/0000-0002-0789-6431
Kajsa Paulsson https://orcid.org/0000-0001-7950-222X

REFERENCES

1. Soto M, Raaijmakers JA, Medema RH. Consequences of genomic diversification induced by segregation errors. *Trends Genet*. 2019;35(4):279-291.
2. Paulsson K, Johansson B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2009;48:637-660.
3. Davidsson J, Paulsson K, Lindgren D, et al. Relapsed childhood high hyperdiploid acute lymphoblastic leukemia: presence of preleukemic ancestral clones and the secondary nature of microdeletions and RTK-RAS mutations. *Leukemia*. 2010;24(5):924-931.
4. Olsson L, Lundin-Strom KB, Castor A, et al. Improved cytogenetic characterization and risk stratification of pediatric acute lymphoblastic leukemia using single nucleotide polymorphism array analysis: a single center experience of 296 cases. *Genes Chromosomes Cancer*. 2018;57(11):604-607.
5. Szczepański T, Willemse MJ, Van Dongen JJM, et al. Precursor-B-ALL with DH-JH gene rearrangements have an immature immunogenotype with a high frequency of oligoclonality and hyperdiploidy of chromosome 14. *Leukemia*. 2001;15(9):1415-1423.
6. Raimondi SC, Pui CH, Hancock ML, Behm FG, Filatov L, Rivera GK. Heterogeneity of hyperdiploid (51-67) childhood acute lymphoblastic leukemia. *Leukemia*. 1996;10(2):213-224.
7. Bateman CM, Alpar D, Ford AM, et al. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. *Leukemia*. 2015;29(1):58-65.
8. Yang JJ, Bhojwani D, Yang W, et al. Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia. *Blood*. 2008;112(10):4178-4183.

9. Mullighan CG, Phillips LA, Su X, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*. 2008;322 (5906):1377-1380.
10. Yang M, Vesterlund M, Siavelis I, et al. Proteogenomics and hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Commun*. 2019;10(1):1519.
11. Nasmyth K. Cohesin: a catenase with separate entry and exit gates? *Nat Cell Biol*. 2011;13(10):1170-1177.
12. Morales C, Losada A. Establishing and dissolving cohesion during the vertebrate cell cycle. *Curr Opin Cell Biol*. 2018;52(9):51-57.
13. Kaur M, DeScipio C, McCallum J, et al. Precocious sister chromatid separation (PSCS) in Cornelia de Lange syndrome. *Am J Med Genet A*. 2005;138A(1):27-31.
14. Losada A, Yokochi T, Hirano T. Functional contribution of Pds5 to cohesin-mediated cohesion in human cells and Xenopus egg extracts. *J Cell Sci*. 2005;118(10):2133-2141.
15. Sajesh BV, Lichtensztejn Z, McManus KJ. Sister chromatid cohesion defects are associated with chromosome instability in Hodgkin lymphoma cells. *BMC Cancer*. 2013;13(1):391.
16. Stoepker C, Ameziane N, van der Lelij P, et al. Defects in the Fanconi anemia pathway and chromatid cohesion in head and neck cancer. *Cancer Res*. 2015;75(17):3543-3553.
17. Barber TD, McManus K, Yuen KWY, et al. Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proc Natl Acad Sci U S A*. 2008;105(9):3443-3448.
18. Solomon DA, Kim T, Diaz-Martinez LA, et al. Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science*. 2011;333 (6045):1039-1043.
19. Kleyman M, Kabeche L, Compton DA. STAG2 promotes error correction in mitosis by regulating kinetochore-microtubule attachments. *J Cell Sci*. 2014;127(19):4225-4233.
20. Hoque MT, Ishikawa F. Cohesin defects lead to premature sister chromatid separation, kinetochore dysfunction, and spindle-assembly checkpoint activation. *J Biol Chem*. 2002;277(44):42306-42314.
21. Toyoda Y, Yanagida M. Coordinated requirements of human topo II and cohesin for metaphase centromere alignment under Mad2-dependent spindle checkpoint surveillance. *Mol Biol Cell*. 2006; 17(5):2287-2302.
22. Watrin E, Schleiffer A, Tanaka K, Eisenhaber F, Nasmyth K, Peters JM. Human Scc4 is required for cohesin binding to chromatin, sister-chromatid cohesion, and mitotic progression. *Curr Biol*. 2006;16 (9):863-874.
23. Yan J, Enge M, Whitington T, et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*. 2013;154(4):801-813.
24. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505 (7484):495-501.
25. Solomon DA, Kim J-S, Bondaruk J, et al. Frequent truncating mutations of STAG2 in bladder cancer. *Nat Genet*. 2013;45(12):1428-1430.
26. Galeev R, Baudet A, Kumar P, et al. Genome-wide RNAi screen identifies cohesin genes as modifiers of renewal and differentiation in human HSCs. *Cell Rep*. 2016;14(12):2988-3000.
27. Fisher JB, McNulty M, Burke MJ, Crispino JD, Rao S. Cohesin mutations in myeloid malignancies. *Trends Cancer*. 2017;3(4):282-293.
28. Molina O, Vinyoles M, Granada I, et al. Impaired condensin complex and Aurora B kinase underlie mitotic and chromosomal defects in hyperdiploid B-cell ALL. *Blood*. 2020;136(3):313-327.
29. Rasmussen M, Sundstrom M, Goransson Kultima H, et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol*. 2011;12(10):R108.
30. Ali N, Karlsson C, Aspling M, et al. Forward RNAi screens in primary human hematopoietic stem/progenitor cells. *Blood*. 2009;113(16): 3690-3695.
31. Lilljebjorn H, Henningsson R, Hyrenius-Wittsten A, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat Commun*. 2016;7(1):11790.
32. Betts DR, Riesch M, Grotzer MA, Niggli FK. The investigation of karyotypic instability in the high-hyperdiploidy subgroup of acute lymphoblastic leukemia. *Leuk Lymphoma*. 2001;42(1–2):187-193.
33. Blandin AT, Muhlematter D, Bougeon S, et al. Automated four-color interphase fluorescence in situ hybridization approach for the simultaneous detection of specific aneuploidies of diagnostic and prognostic significance in high hyperdiploid acute lymphoblastic leukemia. *Cancer Genet Cytogenet*. 2008;186(2):69-77.
34. Alpar D, Pajor G, Varga P, et al. Sequential and hierarchical chromosomal changes and chromosome instability are distinct features of high hyperdiploid pediatric acute lymphoblastic leukemia. *Pediatr Blood Cancer*. 2014;61(12):2208-2214.
35. Onodera N, McCabe NR, Rubin CM. Formation of a hyperdiploid karyotype in childhood acute lymphoblastic leukemia. *Blood*. 1992;80(1): 203-208.
36. Paulsson K, Panagopoulos I, Knuutila S, et al. Formation of trisomies and their parental origin in hyperdiploid childhood acute lymphoblastic leukemia. *Blood*. 2003;102(8):3010-3015.
37. Paulsson K, Morse H, Fioretos T, et al. Evidence for a single-step mechanism in the origin of hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2005;44(2):113-122.
38. Balbas-Martinez C, Sagrera A, Carrillo-de-Santa-Pau E, et al. Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. *Nat Genet*. 2013;45(12):1464-1469.
39. Worrall JT, Tamura N, Mazzagatti A, et al. Non-random mis-segregation of human chromosomes. *Cell Rep*. 2018;23(11):3366-3380.
40. de Lange J, Faramarz A, Oostra AB, et al. Defective sister chromatid cohesion is synthetically lethal with impaired APC/C function. *Nat Commun*. 2015;6(1):8399.
41. McLellan JL, O'Neil NJ, Barrett I, et al. Synthetic lethality of cohesins with PARPs and replication fork mediators. *PLoS Genet*. 2012;8(3): e1002574.

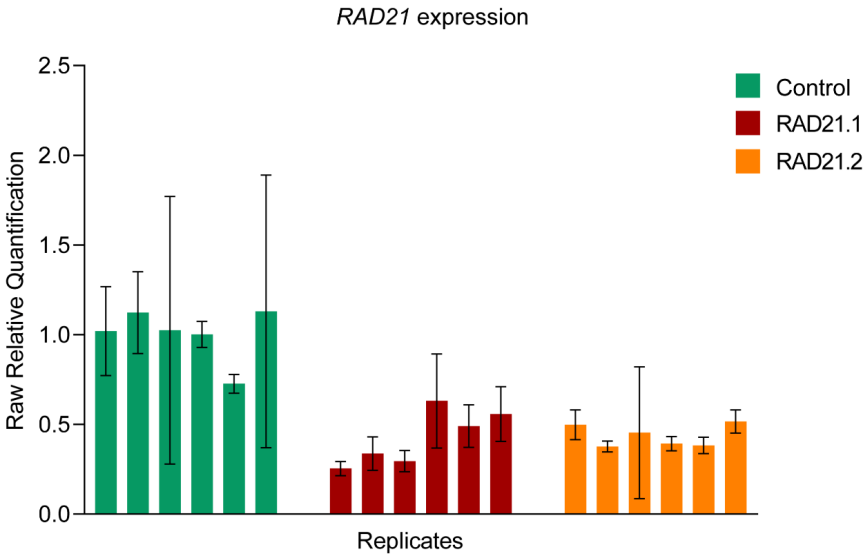
SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

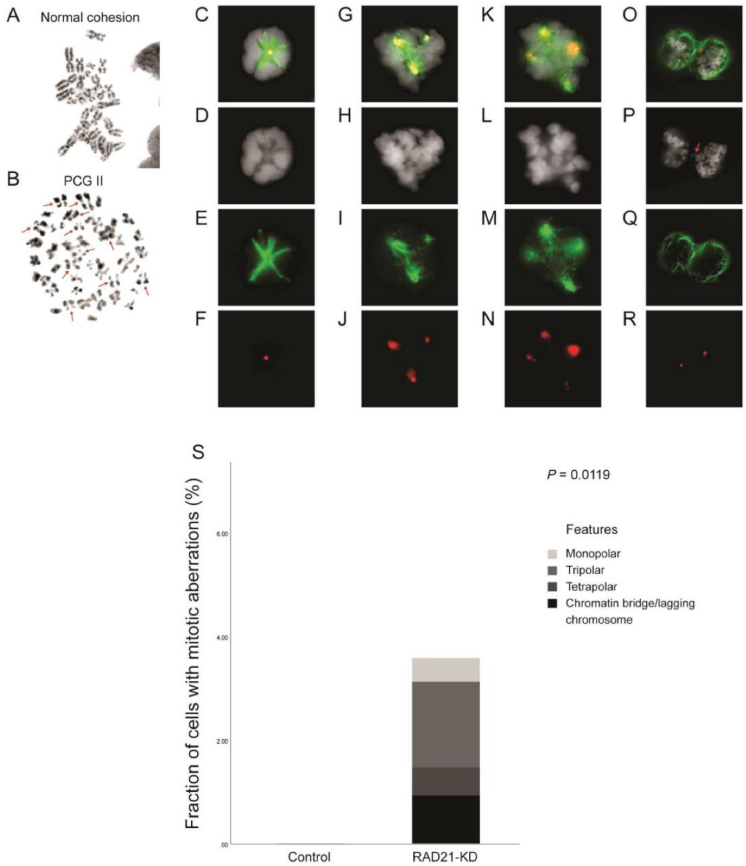
How to cite this article: Moura-Castro LH, Peña-Martínez P, Castor A, et al. Sister chromatid cohesion defects are associated with chromosomal copy number heterogeneity in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2021;60:410–417. https://doi.org/10.1002/gcc.22933

Supporting Information

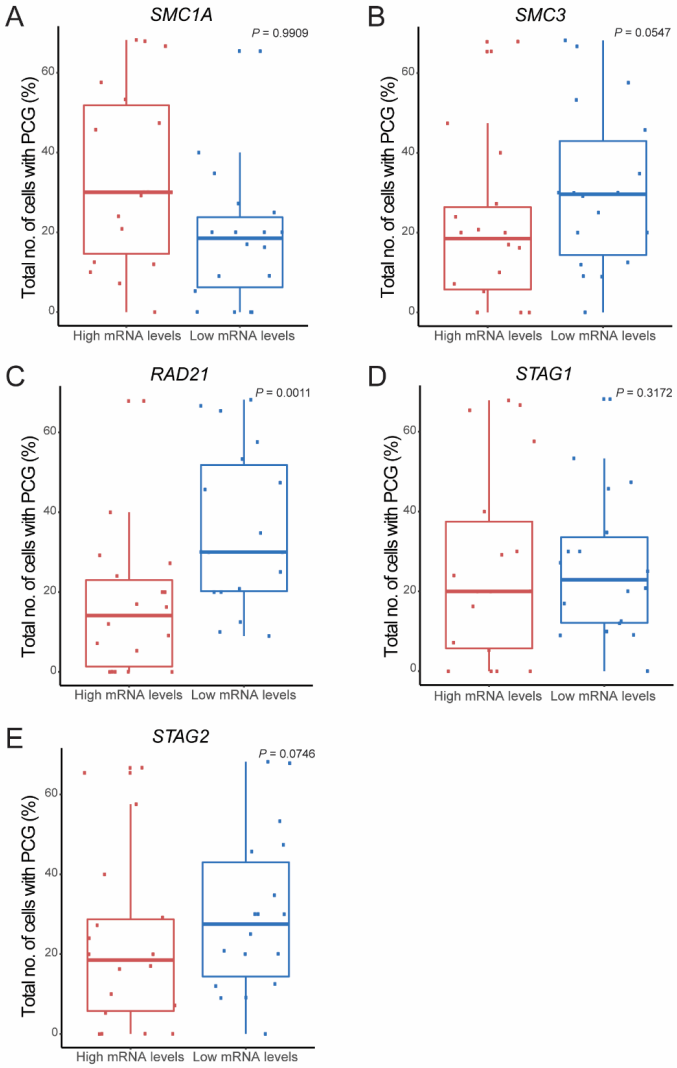
Table of contents



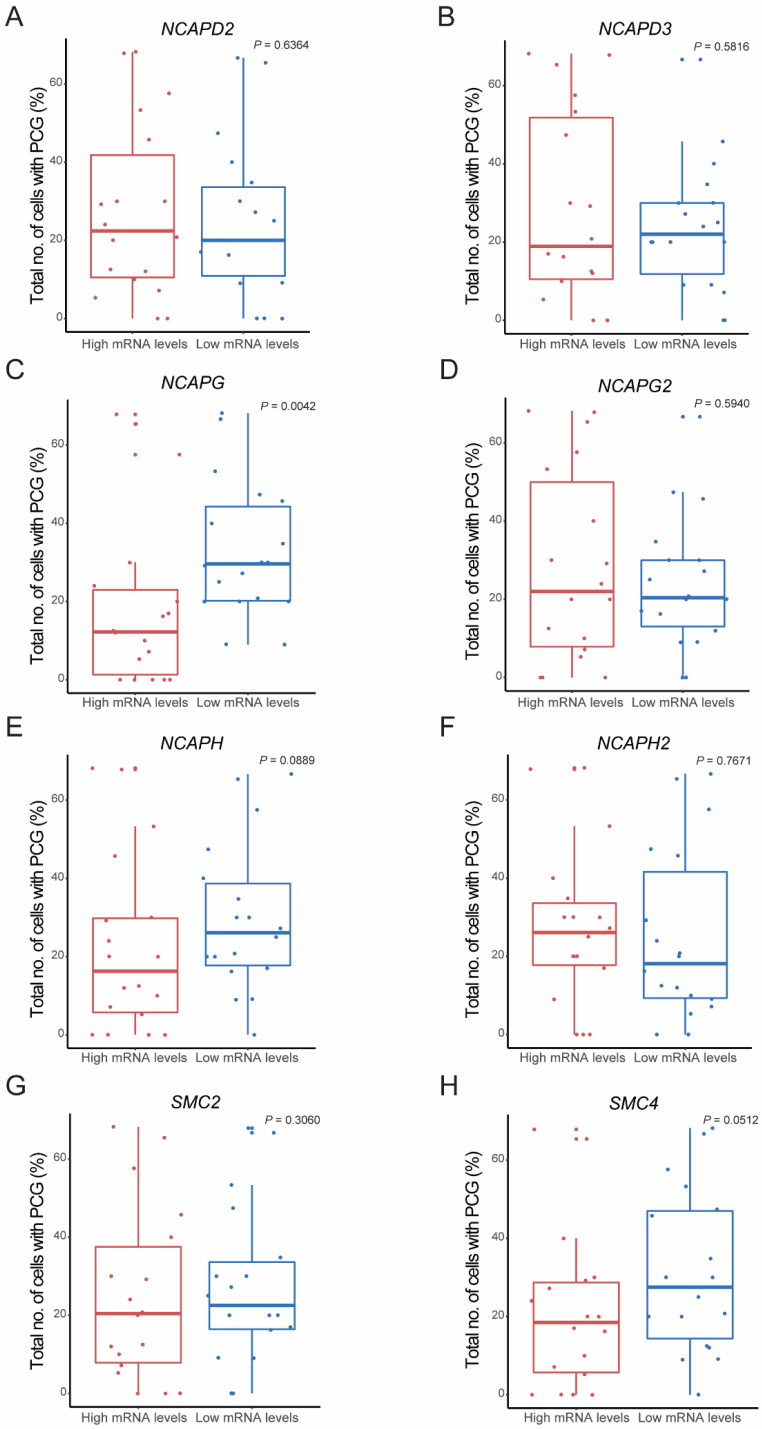
Supporting Information Fig. 1. *RAD21* expression in REH cells submitted to knockdown (RAD21.1 replicates 1-6, RAD21.2 replicates 1-6) and REH control cells (Control 1-6), measured as relative quantification by qPCR, where the expression of the gene of interest is compared to expression of a reference gene (*HPRT1* in the present study). Calculations were performed using the comparative *C*t method (i.e., $\Delta\Delta C$t) and standard deviation was calculated from the three replicates of each sample (whiskers). *RAD21* against *HPRT1* expression in each replicate.



Supporting Information Fig. 2. Cohesion defects and aberrant mitoses in cells with knockdown of RAD21. (**A**-**B**) Examples of sister chromatid cohesion detected as number of primary constriction gaps (PCGs) (primary in REH cells with low expression of *RAD21* (RAD21-KD cells) and controls. **(A)** Metaphase with normal chromosome cohesion from replicate Control 2; **(B)** metaphase classified as PCG II from replicate RAD21.2-2. (**C**-**F**) Examples of aberrant mitoses detected by immunofluorescence microscopy on RAD21-KD cells and Control cells; microtubuli structures were stained for α-Tubulin (green – FITC), centrosomes were stained for γ-Tubulin (orange – Cy3) and DNA was stained with DAPI. **(C)** Monopolar mitosis in replicate RAD21.1-6; **(D)** tripolar mitosis in replicate RAD21.1-6; **(E)** tetrapolar mitosis in replicate RAD21.1-4; **(F)** chromatin bridge/lagging chromosomes (arrow) during mitosis in replicate RAD21.2-4. **(G)** Fraction of cells with aberrant mitoses in RAD21-KD and Control cells.



Supporting Information Fig. 3. mRNA expression of cohesin-related genes in relation to number of cells with cohesion defects, measured as percentage of cells with primary constriction gaps (PCGs), in high hyperdiploid ALL cases. **(A)** *SMC1A*, **(B)** *SMC3*, **(C)** *RAD21*, **(D)** *STAG1* and **(E)** *STAG2*. Boxes show the interquartile range and median (line) values, whiskers show minimum and maximum values in the cohort and dots show cases individually. Whether low mRNA expression was associated with more PCGs were investigated with the one-sided Mann-Whitney U test. *RAD21* showed statistically significant association ($P<0.05$) with the number of cells with PCGs.


A
NCAPD2
P = 0.6364
B
NCAPD3
P = 0.5816
C
NCAPG
P = 0.0042
D
NCAPG2
P = 0.5940
E
NCAPH
P = 0.0889
F
NCAPH2
P = 0.7671
G
SMC2
P = 0.3060
H
SMC4
P = 0.0512
Total no. of cells with PCG (%)
0
20
40
60
High mRNA levels
Low mRNA levels

Supporting Information Fig. 4. mRNA expression of condensin-related genes in relation to number of cells with cohesion defects, measured as percentage of cells with primary constriction gaps (PCGs), in high hyperdiploid ALL cases. **(A)** *NCAPD2*, **(B)** *NCAPD3*, **(C)** *NCAPG*, **(D)** *NCAPG2*, **(E)** *NCAPH*, **(F)** *NCAPH2*, **(G)** *SMC2* and **(H)** *SMC4*. Boxes show the interquartile range and median (line) values, whiskers show minimum and maximum values in the cohort and dots show cases individually. Whether low mRNA expression was associated with more PCGs were investigated with the one-sided Mann-Whitney U test. *NCAPG* showed statistically significant association ($P<0.05$) with the number of cells with PCGs.

Supporting Information Table 1. Gender, age at diagnosis, karyotype information and RNA-seq for 82 primary acute lymphoblastic leukemia.

Case	Gender	Age	Karyotype	RNA-seq
HeH_1	M	2	56,XY,+X,+Y,+4,t(?5;9)(q?;p13-21),+6,+10,+14,+17,+18,+21,+22	yes
HeH_2	F	16	60,XX,+X,+4,+5,+6,+7,+8,+10,+14,+14,+17,+18,+21,+21,+21	yes
HeH_3	F	6	53,XX,+X,+4,+6,+14,+17,+21,+21	yes
HeH_4	F	11	57,XX,+X,+6,+10,+10,+add(14)(p1?),+15,+17,+18,+21,+21,+mar,inc	yes
HeH_5	F	8	59,XX,+X,+X,+4,+6,+9,+10,+10,+14,+15,idic(17)(p11),+18,+20,+21,+21	yes
HeH_6	F	4	54,XX,+X,+4,+6,+14,+17,+18,der(19)t(18;19)(q22.3;p13.3),+21,+21	yes
HeH_7	F	15	57,XX,+X,+X,+4,+6,der(8)t(8;14)(p11;q12),+10,+14,+14,+17,+18,+21,+21	yes
HeH_8	F	2	53,XX,+X,+8,+14,+15,+17,+21,+21	yes
HeH_9	M	2	56,XY,+X,+4,+6,+10,+14,+17,+18,?+der(19)t(1;19)(q11;p13),+21,+21	yes
HeH_10	M	5	55,XY,+X,+6,+9,+14,+14,+17,+18,+21,+21	yes
HeH_11	M	3	60-63,XY,inc/46,XY	no
HeH_12	F	2	56,XX,+X,+4,+6,+8,+10,+14,+17,+18,+21,+21	yes
HeH_13	F	3	57,XX,+X,+4,+6,+8,+8,+12,+14,+17,+18,+21,+21/58,idem,+10	yes
HeH_14	M	3	56,XY,+X,der(1)?ins(1;?)(q21;?),+4,+6,+8,+10,+14,+17,+18,+21,+21	no
HeH_15	M	3	55,XY,+X,+4,+5,+6,idic(7)(p11),+8,+10,+14,+17,+21	yes
HeH_16	M	7	56,XY,+X,+4,+6,+10,+14,+14,+der(17)t(3;17)(q21;p11.1),+18,+18,+21	yes
HeH_17	M	3	54,XY,+X,+4,+6,+10,+14,+17,+18,+21	yes
HeH_18	M	3	52,XY,+X,dup(1)(q21q42),+6,+10,+11,+21,+21	yes
HeH_19	M	12	52-53,XY,+6,+?14,+18,+21,+22,inc/46,XY	no
HeH_20	F	1	54,XX,+X,+6,+8,+14,+17,+18,+21,+21	yes
HeH_21	M	7	52-59,XY,+X,+4,+5,+6,+8,+10,+10,der(13;14)(q10;q10)c,+14,+17,+18,+18,+20,+21,+22	yes
HeH_22	F	3	52-56,XX,add(1)(q3?),+6,+21,+21,inc	no
HeH_23	M	2	63,XY,+X,+Y,+4,+der(5)t(1;5)(q12;q21),+6,+8,+9,+10,+11,+12,+14,+14,+17,+18,+21,+21,+22	yes
HeH_24	M	4	61,XX,-Y,+3,+4,+5,+6,+8,+10,+10,+13,+14,+14,+17,+18,+18,+21,+21	yes
HeH_25	M	2	54,XY,+X,+4,+6,+10,+14,+17,+18,+21/55,idem,+21	yes
HeH_26	M	13	55,XY,+X,t(2;8)(p11.2;q21.13),+4,+6,+10,+14,+der(17)t(17;19)(q?;?)del(17)(p11p13),+18, der(19)t(17;19),+21,+21	yes
HeH_27	M	2	57,XY,+X,+Y,+4,+6,+10,+14,+17,+18,+21,+21,+mar	yes
HeH_28	M	6	57-58,XY,+X,+4,+6,+9,+10,+14,+17,+18,+18,+21,+21	yes
HeH_29	F	5	54,XX,+X,+4,+6,del(6)(q16),+14,+17,+18,+21,+21/56,idem,+8,+10	no
HeH_30	M	4	56,XY,+X,+3,+4,+6,+8,+10,+14,+16,+18,+21/56,idem,i(7)(q10)	yes
HeH_31	M	3	57,XY,+X,+Y,+4,+5,+6,+10,+14,+17,+21,+21,+21	yes
HeH_32	F	3	53,XX,+X,+6,+10,+11,+18,+18,+21/54,idem,+8	yes
HeH_33	F	4	54-57,XX,+X,+X,+4,+5,+6,der(?6)t(1;6)(q21;p25),+10,+14,i(17)(q10),+18,+21,+21	yes

HeH_34	M	9	??,?X,+8[9.2%],+8[10.2%],+21[13.7%],+21[11.7%] (FISH-based)	no
HeH_35	F	3	55,XX,+X,+4,+6,+7,+10,+14,+17,+18,+21	yes
HeH_36	M	5	57,XY,+X,+4,+5,+6,+9,+10,+14,+der(16)t(11;16)(p11.2;p11.2),+18,+21,+21	yes
HeH_37	F	4	52,XX,+X,+6,+14,+17,+18,+21/53,idem,+21	yes
HeH_38	M	1	62,XY,+X,-Y,t(X;4)(q28;q35.2),t(2;19)(p11.2;q13.32),+4,+5,+6,+7,+10,+11,+12,+14,+14,+16,+der(17)t(4;17)(q26;q22),+18,+21,+21,+21,+22	yes
HeH_39	M	3	56,XY,+X,+4,+6,+8,+10,+14,+17,+18,+21,+21	no
HeH_40	F	4	54,XX,+X,+4,+6,+10,+17,+18,+21,+21	yes
HeH_41	M	8	54,XY,+X,+4,+6,+8,+10,+18,+21,+21	no
HeH_42	M	10	54,XY,+X,+6,+10,+14,+17,+18,+21,+21	yes
HeH_43	F	4	52-55,XX,+6,+10,+10,+14,+14,+18,+21,+21[9]/46,XX[1]	no
HeH_44	M	3	55,XY,+X,dup(1)(q12q25),+4,+6,+10,+14,+17,+18,+21,+21	yes
HeH_45	M	3	50-56,XY,+X,+4,+6,+14,+17,+18,+add(21)(q?),+der(?)t(?;21),inc	yes
ETV6/RUNX1_1	M	4	46,XY,del(2)(p11),add(3)(p11),add(6)(q21),add(12)(p13),der(12)t(12;21)(p13;q22),ider(21)(q10)t(12;21),+2-3mar,inc	yes
ETV6/RUNX1_2	F	7	46,X?X,-5,add(7)(q31),add(12)(p11),t(12:21)(p13;q22),-13,-13,?der(21;21)(q10;q10),+21,+2-3mar	yes
ETV6/RUNX1_3	M	5	46,XY,t(12;21)(p13;q22)/47,idem,+16/47,idem,der(6)t(X;6)(?;q1?),+16	yes
ETV6/RUNX1_4	F	3	46,XX,t(12;21)(p13;q22)	yes
ETV6/RUNX1_5	M	3	??,X?,t(12;21)(p13;q22),+der(?)t(?;21)(?;q?)/??,X?,t(12;21),+21,+der(?)t(?;12)t(12;21)	yes
ETV6/RUNX1_6	F	10	??,X,-X,+4,+4,+6,+6,+8,+8,+10,+10,t(12;21)(p13;q22)x12,+14,+14,+17,+17,+18,+18,+21,+21,+21	yes
ETV6/RUNX1_7	F	7	46,XX,t(12;21)(p13;q22)	yes
ETV6/RUNX1_8	M	10	47,XY,+X,t(12;21)(p13;q22)	yes
ETV6/RUNX1_9	M	6	46,XY,t(12;21)(p13;q22)	no
ETV6/RUNX1_10	F	10	??,?X,del(12)(p13p13),t(12;21)(p13q22)[23.5%]/??,?X[68.0%]	no
ETV6/RUNX1_11	F	3	??,X?,del(12)(p13p13),t(12;21)(p13;q22)	yes
ETV6/EUNX1_12	M	6	??,X?,t(12;21)(p13;q22),+der(21)t(12;21)	yes
ETV6/RUNX1_13	F	3	48,XX,t(12;21)(p13;q22),+21,+22/48,idem,del(6)(q21q2?5)	yes
ETV6/RUNX1_14	F	5	45,XX,add(6)(q15),del(12)(p11),t(12;21)(p13;q22),-13,add(15)(q22)	yes
ETV6/RUNX1_15	M	3	??,?X,t(12:21)(p13;q22),+der(21)t(12:21)(p13;q22)	yes
ETV6/RUNX1_16	M	6	46,XY,t(12;21)(p13;q22)	yes
ETV6/RUNX1_17	F	6	??,XX,t(12;21)(p13;q22),inc	yes
ETV6/RUNX1_18	M	5	46,XY,dup(X)(q25q28),t(12;21)(p13;q22),add(16)(q21)	yes
ETV6/RUNX1_19	M	3	46,XY,del(12)(p13p13),t(12;21)(p13;q22)/47,idem,+der(21)t(12;21)	yes
ETV6/RUNX1_20	F	4	46,XX,del(12)(p12p13),t(12;21)(p13;q22)	yes
ETV6/RUNX1_21	F	3	46-47,XX,der(2)t(2;5)(p13;q13),del(4)(q11),del(5)(q13),der(6)t(2;6)(p13;p22),-9,del(12)(p11),der(12)t(4;12)(q11;p12),?add(13)	yes

			52,XY,+X,+4,der(6)t(6;12)(p1?2;q15),add(8)(p?21),+9,+10,del(12)(q15),der(12)t(6;12)	yes
ETV6/RUNX1_22	M	0	(p1?2;p13), ins(12;21)(p13;q22),+der(21)t(12;21)(p13;q22)	
ETV6/RUNX1_23	M	6	46,XY,del(12)(p13p13)t(12:21)(p13;q22)	yes
ETV6/RUNX1_24	M	4	??,X?,t(12;21)(p13;q22)	yes
ETV6/RUNX1_25	M	1	46,XY,t(12;21)(p13;q22)	yes
ETV6/RUNX1_26	M	5	46,XY,del(12)(p11),t(12;21)(p13;q22)/46,XY	no
ETV6/RUNX1_27	M	2	47-48,XY,add(12)(p11),t(12;21)(p13;q22),+21	yes
ETV6/RUNX1_28	M	4	46,XY,del(12)(p13p13),t(12;21)(p13;q22)	no
ETV6/RUNX1_29	M	3	46,XY,?add(7)(p21),t(12:21)(p13;q22),del(12)(p13p13),add(15)(q21),add(22)(q13)	yes
ETV6/RUNX1_30	M	6	45-46,XY,del(6)(q?),add(12)(p11),t(12;21)(p13;q22),inc	yes
			47,XY,dup(5)(p12p15),del(6)(q14q27),del(8)(p11),dup(11)(q24q25),del(12)(p12p13),t(12;21)	no
ETV6/RUNX1_31	M	7	(p13;q22),+der(21)t(12;21)	
ETV6/RUNX1_32	M	8	48,XY,dup(10)(p11p15),t(12;21)(p13;q22),+16,+21	yes
ETV6/RUNX1_33	M	6	46,XY,t(12;21)(p13;q22)/46,idem,add(12)(p13)	yes
ETV6/RUNX1_34	F	1	??,X?,del(12)(p13p13),t(12;21)(p13;q22),+21,+21	yes
ETV6/RUNX1_35	F	6	46,XX,del(6)(q21),t(12;21)(p13;q22)/47,XX,t(12;21),+21/47,XX,t(12;21),+der(21)t(12;21)	yes
			46,XX,der(12)t(12;21)(p1?;q?)t(12;16)(q1?;p11),del(16)(p11),del(21)(q21q22),der(21)t(12;21)	yes
ETV6/RUNX1_36	F	6	(p13;q22)t(12;12)(p13;q13)	
ETV6/RUNX1_37	F	3	49,XX,+X,t(12;21)(p13;q22),+18,+21	yes

Abbreviations: F, female; HeH, high hyperdiploid; M, male; RNA-seq, RNA-sequencing.

Supporting Information Table 2. Sister chromatid cohesion defects in REH cells with knockdown of *RAD21* (RAD21.1 and RAD21.2) and controls.

Replicate	No. of cells	Total no. of cells with PCG (%)	No. of cells with PCG I (%)	No. of cells with PCG II (%)	No. of cells with PCG III (%)	No. of cells with PCG IV (%)	Overall frequency of chromosomes with PCGs
Control 1	44	1 (2)	1 (100)	0	0	0	0.00193
Control 2	52	4 (8)	4 (100)	0	0	0	0.00286
Control 3	38	1 (3)	1 (100)	0	0	0	0.00112
RAD21.1-1	23	8 (35)	5 (62)	3 (38)	0	0	0.0250
RAD21.1-2	34	13 (38)	3 (23)	8 (62)	2 (15)	0	0.0645
RAD21.1-3	37	13 (35)	9 (69)	4 (31)	0	0	0.0322
RAD21.2-1	48	17 (35)	12 (71)	5 (29)	0	0	0.0328
RAD21.2-2	36	8 (22)	4 (5)	3 (38)	1 (12)	0	0.0721
RAD21.2-3	44	18 (41)	6 (33)	12 (67)	0	0	0.0653

Abbreviations: PCG, primary constriction gap.

Supporting Information Table 3. Sister chromatid cohesion defects in high hyperdiploid (HeH) and *ETV6/RUNX1*-positive childhood acute lymphoblastic leukemia.

Case	No of cells	Total No of cells with PCG (%)	No of cells with PCG I (%)	No of cells with PCG II (%)	No of cells with PCG III (%)	No of cells with PCG IV (%)	Overall frequency of chromosomes with PCGs
HeH_1	11	1 (9.1)	1 (100)	0	0	0	0.00649
HeH_2	25	5 (20)	2 (40)	3 (60)	0	0	0.0353
HeH_3	23	8 (35)	5 (62)	3 (38)	0	0	0.0623
HeH_4	28	19 (68)	6 (32)	11 (58)	2 (10)	0	0.0886
HeH_5	33	19 (58)	9 (47)	10 (53)	0	0	0.0515
HeH_6	14	1 (7.1)	1 (100)	0	0	0	0.0225
HeH_7	25	6 (24)	5 (83)	1 (17)	0	0	0.0182
HeH_8	5	2 (40)	1 (50)	1 (50)	0	0	0.0453
HeH_9	27	8 (30)	4 (50)	4 (50)	0	0	0.0298
HeH_10	5	0	0	0	0	0	0
HeH_11	21	4 (19)	3 (75)	1 (25)	0	0	0.0261
HeH_12	4	1 (25)	1 (100)	0	0	0	0.0178
HeH_13	6	1 (17)	1 (100)	0	0	0	0.0113
HeH_14	39	33 (85)	4 (12)	23 (70)	4 (12)	2 (6.1)	0.174
HeH_15	11	3 (27)	3 (100)	0	0	0	0.0165
HeH_16	19	1 (5.3)	1 (100)	0	0	0	0.00369
HeH_17	10	3 (30)	2 (67)	1 (33)	0	0	0.0277
HeH_18	26	17 (65)	2 (11)	5 (30)	9 (53)	1 (5.8)	0.237
HeH_19	26	3 (11)	3 (100)	0	0	0	0.0377
HeH_20	9	0	0	0	0	0	0
HeH_21	7	0	0	0	0	0	0
HeH_22	33	7 (21)	4 (57)	3 (43)	0	0	0.0222
HeH_23	22	2 (9.1)	2 (100)	0	0	0	0.00577
HeH_24	6	0	0	0	0	0	0
HeH_25	9	0	0	0	0	0	0
HeH_26	35	16 (46)	9 (56)	6 (38)	1 (6.3)	0	0.0535
HeH_27	30	6 (20)	5 (83)	1 (17)	0	0	0.0131
HeH_28	23	3 (12)	2 (67)	1 (33)	0	0	0.0107
HeH_29	29	7 (24)	3 (43)	4 (57)	0	0	0.0240
HeH_30	25	3 (12)	3 (100)	0	0	0	0.00772
HeH_31	4	1 (20)	1 (100)	0	0	0	0.0132

HeH_32	24	5 (21)	2 (40)	3 (60)	0	0	0.0285
HeH_33	13	4 (30)	4 (100)	0	0	0	0.0159
HeH_34	35	14 (40)	11 (79)	3 (21)	0	0	0.0270
HeH_35	37	6 (16)	5 (83)	1 (17)	0	0	0.0128
HeH_36	10	2 (20)	1 (50)	1 (50)	0	0	0.0263
HeH_37	24	7 (29)	4 (57)	3 (43)	0	0	0.0393
HeH_38	45	24 (53)	12 (50)	10 (42)	2 (8.3)	0	0.0550
HeH_39	13	0	0	0	0	0	0
HeH_40	44	30 (68)	9 (30)	17 (57)	4 (13)	0	0.102
HeH_41	16	13 (81)	2 (16)	8 (61)	3 (23)	0	0.166
HeH_42	7	1 (10)	1 (100)	0	0	0	0.0104
HeH_43	29	0	0	0	0	0	0
HeH_44	19	9 (47)	6 (67)	3 (33)	0	0	0.0373
HeH_45	15	10 (67)	4 (40)	5 (50)	0	1(10)	0.126
ETV6/RUNX1_1	6	1 (14)	1 (100)	0	0	0	0.00362
ETV6/RUNX1_2	25	1 (4.0)	1 (100)	0	0	0	0.000870
ETV6/RUNX1_3	4	0	0	0	0	0	0
ETV6/RUNX1_4	6	0	0	0	0	0	0
ETV6/RUNX1_5	27	1 (3.7)	1 (100)	0	0	0	0.00161
ETV6/RUNX1_6	8	0	0	0	0	0	0
ETV6/RUNX1_7	24	4 (17)	3 (75)	1 (25)	0	0	0.0118
ETV6/RUNX1_8	20	0	0	0	0	0	0
ETV6/RUNX1_9	27	2 (7.4)	2 (100)	0	0	0	0.00644
ETV6/RUNX1_10	26	0	0	0	0	0	0
ETV6/RUNX1_11	7	1 (14)	0	1 (100)	0	0	0.0466
ETV6/RUNX1_12	11	0	0	0	0	0	0
ETV6/RUNX1_13	16	1 (6.3)	0	0	1 (100)	0	0.0286
ETV6/RUNX1_14	16	1 (6.3)	0	1 (100)	0	0	0.00833
ETV6/RUNX1_15	8	0	0	0	0	0	0
ETV6/RUNX1_16	25	2 (8.0)	1 (50)	1 (50)	0	0	0.00522
ETV6/RUNX1_17	20	1 (5.0)	1 (100)	0	0	0	0.00326
ETV6/RUNX1_18	12	0	0	0	0	0	0
ETV6/RUNX1_19	8	0	0	0	0	0	0
ETV6/RUNX1_20	16	0	0	0	0	0	0
ETV6/RUNX1_21	5	0	0	0	0	0	0
ETV6/RUNX1_22	11	2 (18)	1 (50)	0	1 (50)	0	0.00524
ETV6/RUNX1_23	46	2 (4.3)	2 (100)	0	0	0	0.00189

ETV6/RUNX1_24	32	2 (18)	5 (83)	1 (17)	0	0	0.0122
ETV6/RUNX1_25	21	0	0	0	0	0	0
ETV6/RUNX1_26	13	2 (15)	2 (100)	0	0	0	0.00669
ETV6/RUNX1_27	17	0	0	0	0	0	0
ETV6/RUNX1_28	51	2 (3.9)	2 (100)	0	0	0	0.000853
ETV6/RUNX1_29	6	0	0	0	0	0	0
ETV6/RUNX1_30	17	1 (5.9)	1 (100)	0	0	0	0.00256
ETV6/RUNX1_31	23	1 (4.3)	1 (100)	0	0	0	0.000925
ETV6/RUNX1_32	7	0	0	0	0	0	0
ETV6/RUNX1_33	6	0	0	0	0	0	0
ETV6/RUNX1_34	8	0	0	0	0	0	0
ETV6/RUNX1_35	25	3 (12)	3 (100)	0	0	0	0.00522
ETV6/RUNX1_36	4	0	0	0	0	0	0
ETV6/RUNX1_37	5	0	0	0	0	0	0

Abbreviations: PCG, primary constriction gap.

Supporting Information Table 4. Copy number analysis by interphase-FISH of chromosomes X,2, 3, 6, 10 and 21 in HeH ALL primary cases with high and low levels of cohesion defects, measured as PCG percentage per case.

		No. of cells (%)					
	Cohesin defects (%)	**Monosomy**	**Disomy**	**Trisomy**	**Tetrasomy**	**Pentasomy**	**P-value[b]**
Chromosome X			**a**				
HeH_14	85%	10 (3.08)	287 (88.6)	27 (8.33)	0	0	
HeH_18	65%	3 (0.99)	270 (89.4)	29 (9.60)	0	0	
							0.100
HeH_30	12%	7 (2.02)	335 (96.8)	4 (1.16)	0	0	
HeH_42	10%	2 (0.67)	283 (94.3)	15 (5.00)	0	0	
HeH_39	0%	3 (0.98)	274 (89.84)	26 (8.52)	2 (0.66)	0	
Chromosome 2			**a**				
HeH_14	85%	7 (2.31)	282 (93.1)	14 (4.62)	0	0	
HeH_41	81%	10 (3.00)	309 (92.8)	14 (4.20)	0	0	
HeH_18	65%	2 (0.66)	262 (85.9)	41 (13.4)	0	0	
							0.114
HeH_30	12%	14 (4.39)	296 (92.8)	9 (2.82)	0	0	
HeH_42	10%	5 (1.61)	301 (96.8)	1 (0.32)	0	0	
HeH_23	9%	10 (3.16)	304 (96.2)	2 (0.63)	0	0	
HeH_39	0%	8 (2.35)	329 (96.8)	3 (0.88)	0	0	
Chromosome 3			**a**				
HeH_14	85%	9 (2.83)	297 (92.2)	16 (4.97)	0	0	
HeH_41	81%	9 (2.84)	223 (70.3)	85 (26.8)	0	0	
HeH_4	68%	16 (5.26)	210 (69.5)	76 (25.2)	0	0	
HeH_40	68%	31 (9.71)	224 (70.2)	64 (20.1)	0	0	
HeH_18	65%	4 (1.30)	234 (76.2)	69 (22.5)	0	0	
							0.0357
HeH_42	10%	1 (0.33)	279 (93.0)	20 (6.67)	0	0	
HeH_23	9%	0	279 (93.0)	21 (7.00)	0	0	
HeH_21	0%	0	284 (90.4)	30 (9.60)	0	0	
Chromosome 6				**a**			
HeH_14	85%	1 (0.29)	61 (18.1)	257 (76.3)	18 (5.34)	0	
HeH_4	68%	0	34 (10.3)	254 (76.7)	43 (12.9)	0	
HeH_40	68%	0	7 (2.16)	279 (86.1)	38 (11.7)	0	
HeH_18	65%	12 (4.00)	43 (14.3)	230 (76.7)	15 (5.00)	0	
							0.00794
HeH_30	12%	0	31 (8.93)	305 (87.9)	11 (3.17)	0	
HeH_42	10%	0	7 (2.21)	305 (96.2)	5 (1.58)	0	
HeH_23	9%	0	12 (3.81)	294 (93.3)	9 (2.86)	0	
HeH_21	0%	1 (0.32)	4 (1.28)	295 (94.5)	12 (3.85)	0	
HeH_39	0%	0	28 (8.21)	311 (91.2)	2 (0.59)	0	
Chromosome 10				**a**			

HeH_14	85%	0	25 (7.81)	279 (87.2)	16 (5.00)	0	
HeH_41	81%	0	96 (29.3)	196 (59.7)	36 (10.9)	0	
HeH_18	65%	0	72 (22.6)	225 (70.7)	21 (6.60)	0	
							0.0286
HeH_30	12%	0	35 (8.06)	358 (82.5)	14 (3.23)	0	
HeH_42	10%	0	13 (4.11)	302 (95.6)	1 (0.32)	0	
HeH_23	9%	0	6 (1.97)	295 (97.0)	3 (0.99)	0	
HeH_39	0%	0	28 (8.78)	279 (87.5)	12 (3.76)	0	
Chromosome 21					**a**		
HeH_14	85%	0	2 (0.62)	115 (35.6)	183 (56.7)	26 (8.05)	
HeH_41	81%	0	6 (1.79)	94 (28.1)	201 (60.2)	33 (9.88)	
HeH_4	68%	0	2 (0.58)	147 (42.9)	189 (55.3)	4 (1.17)	
HeH_40	68%	0	4 (1.23)	127 (39.1)	185 (56.9)	9 (2.77)	
HeH_18	65%	0	6 (1.85)	160 (49.4)	128 (39.5)	30 (9.26)	
							0.0179
HeH_42	10%	0	2 (0.66)	29 (9.60)	271 (89.7)	0	
HeH_23	9%	0	3 (0.97)	27 (8.77)	278 (90.2)	0	
HeH_39	0%	0	1 (0.33)	33 (10.9)	272 (89.8)	0	

Abbreviations: FISH, fluorescence in situ hybridization; HeH ALL, high hyperdiploid childhood acute lymphoblastic leukemia; PCG, primary constriction gaps.

a Expected copy number for the given chromosome based on karyotype information and SNP array data.

b Mann-Whitney U test.

Supporting Information Table 5. Copy number analysis by interphase-FISH of chromosomes X, 2, 3 and 21 in REH cells with knock-down of RAD21 (RAD21.1 and RAD21.2) and controls.

	Monosomy	Disomy	Trisomy	Tetrasomy	Pentasomy	P-value[b]
Chromosome X	**a**					
Controls - No. of cells (%)	921 (98.1)	18 (1.92)	0	0	0	0.171
RAD21-KD - No. of cells (%)	1224 (97.2)	35 (2.78)	0	0	0	
Chromosome 2		**a**				
Controls - No. of cells (%)	24 (2.59)	849 (91.5)	52 (5.60)	3 (0.0323)	0	0.131
RAD21-KD - No. of cells (%)	63 (3.25)	1714 (88.4)	161 (8.29)	2 (0.0103)	0	
Chromosome 3		**a**				
Controls - No. of cells (%)	14 (1.41)	888 (89.7)	88 (8.89)	0	0	0.357
RAD21-KD - No. of cells (%)	40 (2.03)	1753 (89.1)	174 (8.85)	0	0	
Chromosome 21			**a**			
Controls - No. of cells (%)	7 (0.0757)	99 (10.7)	758 (81.9)	58 (6.27)	3 (0.0324)	0.0286
RAD21-KD - No. of cells (%)	41 (2.92)	288 (20.5)	887 (63.1)	169 (12.0)	21 (1.49)	

Abbreviations: FISH, fluorescence in situ hybridization; KD, knock-down

a Expected copy number for the given chromosome based on karyotype information and SNP array data.

b Mann-Whitney U test.

Supporting Information Table 6. Analysis of mitosic cells by immunofluorescence microscopy in REH cells with knock-down of *RAD21* (RAD21.1 and RAD21.2) and controls.

Categories	Control No of cells (%)	RAD21-KD No of cells (%)	*P*-value[a]
Bipolar (normal)	300 (100)	565 (93.5)	
All aberrations	0	39 (6.5)	0.0119
Spindle defects	0	29 (4.8)	
Monopolar	0	5 (0.83)	0.2381
Tripolar	0	18 (2.9)	0.0119
Tetrapolar	0	6 (0.99)	0.2381
Chromatin bridges/ Lagging chromosomes	0	10 (1.6)	0.119

[a]Mann-Whitney U Test.

Article IV

nature communications

Article

https://doi.org/10.1038/s41467-023-37356-5

Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukaemia

Received: 19 April 2022

Accepted: 14 March 2023

Published online: 25 March 2023

Check for updates

Eleanor L. Woodward[1,9], Minjun Yang[1,9], Larissa H. Moura-Castro[1], Hilda van den Bos[2], Rebeqa Gunnarsson[1], Linda Olsson-Arvidsson[1,3], Diana C. J. Spierings[2], Anders Castor[4], Nicolas Duployez[5,6], Marketa Zaliova[7,8], Jan Zuna[7,8], Bertil Johansson[1,3], Floris Foijer[2] & Kajsa Paulsson[1]

High hyperdiploid acute lymphoblastic leukemia (HeH ALL), one of the most common childhood malignancies, is driven by nonrandom aneuploidy (abnormal chromosome numbers) mainly comprising chromosomal gains. In this study, we investigate how aneuploidy in HeH ALL arises. Single cell whole genome sequencing of 2847 cells from nine primary cases and one normal bone marrow reveals that HeH ALL generally display low chromosomal heterogeneity, indicating that they are not characterized by chromosomal instability and showing that aneuploidy-driven malignancies are not necessarily chromosomally heterogeneous. Furthermore, most chromosomal gains are present in all leukemic cells, suggesting that they arose early during leukemogenesis. Copy number data from 577 primary cases reveals selective pressures that were used for in silico modeling of aneuploidy development. This shows that the aneuploidy in HeH ALL likely arises by an initial tripolar mitosis in a diploid cell followed by clonal evolution, in line with a punctuated evolution model.

The genetic origin of tumours remains obscure as the earliest stages of tumorigenesis cannot be observed. In the classic view of tumour development, cells acquire mutations in a stepwise manner, with clonal selection shaping the tumour genome over time and genomic heterogeneity arising by branching of different subclones[1,2]. However, in recent years this view has been challenged by data showing that some tumours arise by punctuated evolution, where the bulk of genetic aberrations occur within a short time frame at tumour initiation, followed by proliferation during which only little additional genomic heterogeneity is added[1,3,4]. The punctuated evolution model appears to fit particularly well with copy number aberrations, both intrachromosomal and those involving whole chromosomes[1].

The high hyperdiploid (HeH; 51-67 chromosomes) subtype comprises 25–30% of all paediatric B-cell precursor acute lymphoblastic leukaemia (ALL). HeH ALL is characterized by nonrandom chromosomal gains predominately involving 1–2 extra copies of chromosomes X, 4, 6, 10, 14, 17, 18, and 21, whereas chromosomal losses are very rare[5]. Several lines of evidence suggest that the aneuploidy arises early in

[1]Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, Lund, Sweden. [2]European Research Institute for the Biology of Ageing (ERIBA), University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [3]Department of Clinical Genetics, Pathology, and Molecular Diagnostics, Office for Medical Services, Region Skåne, Lund, Sweden. [4]Department of Pediatrics, Skåne University Hospital, Lund University, Lund, Sweden. [5]Laboratory of Hematology, Centre Hospitalier Universitaire (CHU) Lille, Lille, France. [6]Unité Mixte de Recherche en Santé (UMR-S) 1172, INSERM/University of Lille, Lille, France. [7]Department of Pediatric Hematology and Oncology, Second Faculty of Medicine, Charles University/University Hospital Motol, Prague, Czech Republic. [8]Childhood Leukaemia Investigation Prague (CLIP), Prague, Czech Republic. [9]These authors contributed equally: E. L. Woodward, M. Yang. e-mail: kajsa.paulsson@med.lu.se

HeH ALL, possibly already before birth[6–10], although overt leukaemia does not occur until several years later. Furthermore, analyses of allelic ratios in tetrasomic chromosomes have suggested that the extra chromosomes are gained at the same time in one abnormal cell division[11–13]. However, the details on how HeH ALL develops genetically remain unknown.

We have addressed the origin of HeH ALL using single cell whole genome sequencing (scWGS), analyses of selection pressures in a large patient cohort, and through in silico modelling. We find that stable aneuploid karyotypes that we observe in HeH ALL likely arise during a single tripolar mitosis followed by low-level clonal evolution. Our findings shed light into the earliest stages of tumorigenesis of the most common malignancy in childhood.

Results

HeH ALL displays little genomic heterogeneity

To understand how the aneuploidy arises in HeH ALL, we first set out to determine the degree of genomic heterogeneity, in particular chromosomal heterogeneity as a readout of chromosomal instability (CIN). We performed low-pass scWGS of 257–348 individual bone marrow cells/case, in total 2847 cells, from nine primary hyperdiploid ALL cases (2–13 years old at diagnosis; median 5 years) and one normal bone marrow sample (Supplementary Table 1). Copy number analysis for each individual cell was carried out with a resolution of approximately 5 Mb. For some chromosomes, we investigated which chromosomal homologue that was gained, lost, or displayed uniparental isodisomy (UPID; disomies involving two copies of the same chromosomal homologue) taking advantage of heterozygous variants identified through bulk WGS of matched samples. Phylogenetic trees were then constructed based on the combined data from scWGS, bulk WGS, and fluorescence in situ hybridization (FISH).

The normal bone marrow displayed diploidy in 269/270 cells (99.6%), with only one cell deviating by loss of chromosome 21, showing the high quality of the scWGS (Fig. 1). Of the 2577 cells in the leukaemic samples, five were normal diploid cells and the rest showed copy number changes agreeing with leukaemic cells. Overall, highly homogeneous genomes were seen for most of the leukaemias (Fig. 1), with predominantly whole chromosome gains being present in all cells. When assessing whole chromosome changes, 5/9 cases had the same chromosomal content in >99% of the cells, with only 1–2 cells displaying gains or losses of single chromosomes that were not seen in the other cells, suggesting a chromosome missegregation rate highly similar as observed for the normal bone marrow. The remaining four cases had 3–5 numerical subclones each (a clone being defined as at least two cells with the same genetic aberrations), with the major clone making up 55–88% of the cells (Table 1). For 3/4 cases, at least one of these subclones was also detectable in copy number analysis of bulk DNA; i.e. they would appear to harbour subclones also by this method. Case 2, however, displayed three minor subclones, each corresponding to 2.7–3.9% of the cells, which analysis of bulk DNA failed to detect so that it appeared to have only one clone. Analysis of chromosomal homologues revealed hidden heterogeneity in #9, where trisomy 17 involved different homologues in two distinct cell populations (Supplementary Fig. 1); this was, however, the only case of haplotype heterogeneity found among the 62 chromosomal gains/UPIDs that could be investigated.

Next, we calculated heterogeneity scores for each case (Table 1). There was no correlation between the heterogeneity scores and the number of cells sequenced, showing that the results were not skewed based on the number of cells included ($r_s = -0.38$, $P = 0.32$; two-sided Spearman's correlation test; Supplementary Fig. 2). Cases with relatively few subclones (#1, #5, #6, #7, and #8) had lower scores than cases with more subclones (#2, #3, #4, and #9). To investigate whether the observed differences in heterogeneity were due to mutations in genes affecting genomic stability, we screened bulk WGS data, but no such correlation was seen (Supplementary Table 1). We further investigated whether increased heterogeneity correlated with the presence of sister chromatid cohesion defects in metaphase chromosomes, which we have recently reported to be associated with increased chromosomal heterogeneity in HeH ALL[14]. Indeed, #2, which had the second highest heterogeneity score, had a very high frequency of cohesion defects in metaphase cells (85%; Table 1). However, #3, #4 and #9, which also had high heterogeneity scores, had relatively few cells with cohesion defects. Overall, however, although the heterogeneity scores varied between cases, all had relatively low levels of heterogeneity, with non-clonal changes only seen in 0–2.6% of the cells.

In conclusion, the scWGS analysis revealed very low to low chromosomal heterogeneity in HeH childhood ALL. Thus, these leukaemias appear to have relatively stable genomes, despite being aneuploid.

The chromosomal gains are early and ubiquitously present in HeH ALL

To understand how hyperdiploidy develops in the absence of CIN, we studied the phylogenetic trees of the chromosomal changes (Fig. 2). In all cases, the majority of chromosomal gains were seen at the roots of the trees, with most remaining stable and unchanging. The pattern of chromosomal gains in the inferred initial leukaemic cells resembled the one usually seen in HeH ALL: chromosomes X (100%), 21 (100%), 4 (89%), 14 (89%), 18 (89%), 6 (67%), 10 (67%), 17 (67%), 8 (44%), 9 (33%), 5 (22%), 16 (22%), 3 (11%), 11 (11%), and 12 (11%). Looking at chromosomal gains only, those in the earliest clone and in the major clone were identical in 5/9 (56%) of the cases, with the remaining four cases differing by gain or loss of 1–2 chromosomes. Thus, most extra chromosomes found at diagnosis were acquired early in leukemogenesis, in line with previous studies of HeH ALL[6–10]. Calculation of phylogenetic distances showed long truncal and short branching distances, suggesting punctuated evolution (Supplementary Fig. 3)[1]. Chromosomes that changed in copy number during clonal evolution comprised X, 8, 9, 14, 16, 17, and 21 (Fig. 2). Several cases displayed more than one instance of a particular chromosomal copy number change during their clonal evolution, comprising losses of 9 (two events in #2), gains of 17 (two events in #9), and gains of 21 (two events in #4), indicating strong clonal selection for these changes.

Only few clonal structural changes leading to copy number changes were detected by scWGS, in line with such events being relatively rare in hyperdiploid ALL[10]. In 7/9 cases, no structural changes were present in the inferred earliest cell, indicating that such abnormalities typically arose after the bulk of the chromosomal gains (Fig. 2). Duplication of 1q [dup(1q)] was seen in subclones in three different cases; one of which (#2) had dup(1q) with different breakpoints between two subclones. scWGS also revealed more complex patterns of copy number changes associated with structural events in #3 and #4. Further analysis with FISH and bulk WGS confirmed that these structural abnormalities involved complex rearrangements of chromosomes 16 and 14, respectively (Fig. 2, Supplementary Fig. 4). Thus, scWGS can also be used to delineate complex structural events.

Taken together, phylogenetic analysis of the scWGS data showed that most of the chromosomal gains were present at the root of the phylogenetic trees, with clonal evolution involving gains or losses of 1–2 chromosomes in approximately half of the cases. Structural changes, on the other hand, generally occurred later during leukemogenesis.

Aneuploid pattern based on copy number changes in 577 cases reveals selective pressures

To elucidate further the aneuploid pattern in HeH ALL, we next studied copy number data derived from single nucleotide polymorphism (SNP) arrays, whole exome sequencing (WES), or WGS for 577 primary

https://doi.org/10.1038/s41467-023-37356-5

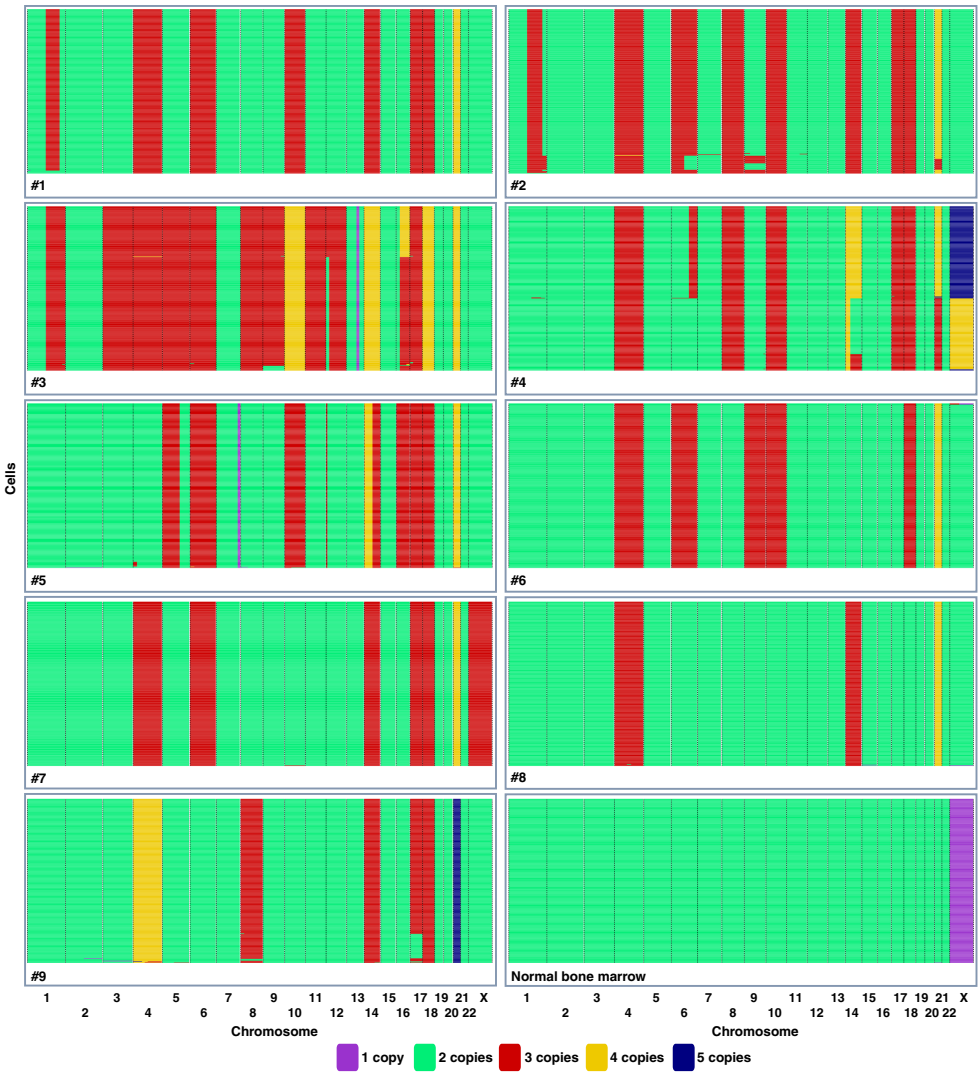


Fig. 1 | Single cell whole genome sequencing results from nine primary high hyperdiploid childhood acute lymphoblastic leukaemia cases and one normal bone marrow. The heatmaps show the genome-wide copy number of each individual cell with a resolution of 5 Mb (the Y chromosome is not included). Overall, only low to very low levels of copy number heterogeneity was seen. Created with BioRender.com. Source data are provided as a Source Data file.

cases (Supplementary Data 1, Supplementary Fig. 5), in total encompassing 13271 chromosomal pairs. Of these, 6 (0.045%) were monosomic, 8410 (63%) disomic, 3997 (30%) trisomic, 829 (6.2%) tetrasomic, and 29 (0.22%) pentasomic. Together, these data corroborate the view that HeH ALL is primarily characterized by trisomies and tetrasomies[5], with monosomies being exceedingly rare.

To be able to model HeH development, we utilized this copy number data to better understand selective pressures, reasoning that chromosomal gains providing a selective advantage are more common. Eight chromosomes were gained in more than 70% of cases: chromosomes 21 (100%), X (97%), 14 (95%), 6 (89%), 18 (83%), 4 (82%), 17 (78%), and 10 (74%), indicating a strong selection for extra copies of these chromosomes and suggesting that these gains are highly likely driver events. Six additional gains were relatively common: chromosomes 8 (38%), 5 (23%), 9 (19%), 11 (14%), 12 (14%), and 22 (11%). These copy number alterations might also be (co-)driving events, at least occasionally. The remaining autosomal chromosomes were gained in <10% of the cases and hence unlikely to provide a selective advantage; some, such as chromosomes 13 and 20, which were recurrently monosomic, may even be selected against. Chromosome Y displayed

Table 1 | Genetic heterogeneity in nine high hyperdiploid childhood acute lymphoblastic leukaemia cases based on single cell whole genome sequencing

Case	Number of cells sequenced	Number of clones (% of cells)[a]	Number of clones—numerical changes only (% of cells)[a]	Number of cells with a unique genome	Genome-wide heterogeneity score	% of cells with PCG
1	269	2 (A, 98%; C, 1.9%)	1 (A, C, 100%)	1	0.07	22
2	257	6 (A, 88%; E, 3.9%; D, 2.7%; I, 1.9%; F, 1.2%; H, 0.8%)	4 (A, 88%; E, 3.9%; F, H, I, 3.9%; D, 2.7%)	5	1.20	85
3	272	4 (E, 64%; A, 30%; H, 2.6%; G, 0.7%)	3 (E, G, 65%; A, 30%; H, 2.6%)	7	1.16	14
4	348	5 (H, 55%; C, 34%; A, 8.9%; E, 0.9%; I, 0.6%)	5 (H, 55%; C, 34%; A, 8.9%; E, 0.9%; I, 0.6%)	5	5.11	19
5	347	2 (A, 96%; D, 2.9%)	1 (A, D, 99%)	3	0.18	20
6	271	2 (A, 99%; B, 0.7%)	1 (A, B, 100%)	0	0.04	5
7	273	1 (A, 100%)	1 (A, 100%)	1	0.12	10
8	266	1 (A, 99%)	1 (A, 99%)	3	0.16	0
9	269	4 (B, 75%; F, 15%; A, 7.1%; I, 0.7%)	4 (B, 75%; F, 15%; A, 7.1%; I, 0.7%)	7	0.69	10

PCG primary constriction gap.
[a]Letters correspond to different clones as denoted in Fig. 2.

both gains (21% of male cases)—always as XXYY or XXXYY—and nullisomy (4% of male cases), indicating that it is neutral to selection.

Recurrent tetrasomies were seen for chromosomes 21 (81%), X/Y (20%; including XXXX in females and XXXY/XXYY in males), 14 (17%), 18 (12%), 10 (8.3%), 8 (2.6%), and 4 (1.7%). The majority (787/829; 95%) of tetrasomies were of the 2:2 type, i.e. showed duplication of both chromosomal homologues. Of the 42 3:1 tetrasomies (triplication of one homologue and retention of the other), 31 (74%) were for chromosome 21 and five (12%) were XXXY. Apart from one case with XXXYY, pentasomy was only seen for chromosome 21 (4.5% of cases), indicating that the selection for extra copies of this chromosome is particularly strong.

UPIDs were seen in 208/577 (36%) of the cases (median 1/case, range 1–6). The UPIDs/all disomies ratio was 0–5% for all chromosomes except for chromosome 9, where it was 17%. This rather constant frequency (except for chromosome 9) suggests that, in general, UPIDs are passenger events.

Subclonality indicates selective pressures

Copy number analysis based on bulk samples has a limited resolution in detecting subclones, with an approximate detection limit of subclones corresponding to 20–30% of the cells (Supplementary Fig. 6). Nevertheless, subclonality involving relatively large clones that are detectable with this method can indicate ongoing clonal evolution and may reveal selective pressures in the leukaemic population. The majority (72%) of the 577 HeH ALLs did not have detectable subclonality involving whole chromosomes, agreeing well with the scWGS data. Most chromosomes displayed subclonality in <3% of cases, but higher levels were seen for chromosomes 8 (4.5%), 9 (8.7%), 21 (3.6%), and X in females (5.1%) (Supplementary Table 2). For chromosomes 8, 9, and X, subclonality was mainly seen between two and three copies, either in the form of (hetero)disomy/trisomy or in the form of UPID/trisomy; two forms of subclonality that have approximately the same detection limits in the HeH scenario. Whereas the former of these could arise either by an initial disomy becoming a trisomy or vice versa, the latter can only arise from initial trisomy by loss of one chromosomal homologue (Supplementary Fig. 7). Then, the likelihood is 2/3 that it becomes a heterodisomy (normal disomy with retained heterozygosity) and 1/3 that it becomes a UPID. Most chromosomes conformed to the expected ratio of subclonal disomy/trisomy to UPID/trisomy (Supplementary Table 2), suggesting loss from trisomy. For chromosome X in females, however, trisomy/UPID subclonality was significantly more common than expected ($P = 2.60 \times 10^{-4}$; two-sided exact binomial test), which is likely explained by preferential loss of the inactive X, as it is usually the active X that is duplicated in HeH ALL with trisomy X[12]. Chromosome 8 displayed borderline significance ($P = 0.0529$; two-sided exact binomial test) for fewer cases with subclonal UPID/trisomy than expected (Supplementary Table 2), possibly indicating that some cases were gaining an extra chromosome from a disomy, in line with positive selection. Chromosome 9 displayed frequencies of subclonal disomy/trisomy and UPID/trisomy agreeing with loss from a trisomic state, indicating selection against trisomy. Finally, subclonality for chromosome 21 was mainly seen for trisomy/tetrasomy and tetrasomy/pentasomy, indicating selection for extra chromosomal copies. Altogether, selection against extra copies of chromosome 9 and for extra copies of chromosome 21 and possibly chromosome 8 was apparent, with the reservation that subclones corresponding to less than 20-30% of the cells could not be analyzed.

Comparison of diagnostic and relapse samples shows positive selection for trisomy 8 and negative for trisomy 9

Selective pressures can also be inferred from comparing paired samples obtained at different time points. We studied chromosomal copy number and ascertained whether trisomies and UPIDs involved the same chromosomal homologue in paired diagnostic/relapse samples from 23 cases. Such samples have previously been shown to be clonally related and display overall very similar karyotypes[15,16]. In total, 4.4% of 529 chromosomal pairs differed in copy number between the diagnostic and relapse samples (Supplementary Data 2). Of the 171 investigated trisomies and UPIDs, only one trisomy 8 involved different chromosomal homologues in the diagnostic and relapse sample, indicating that heterogeneity of this type is rare in HeH ALL.

https://doi.org/10.1038/s41467-023-37356-5

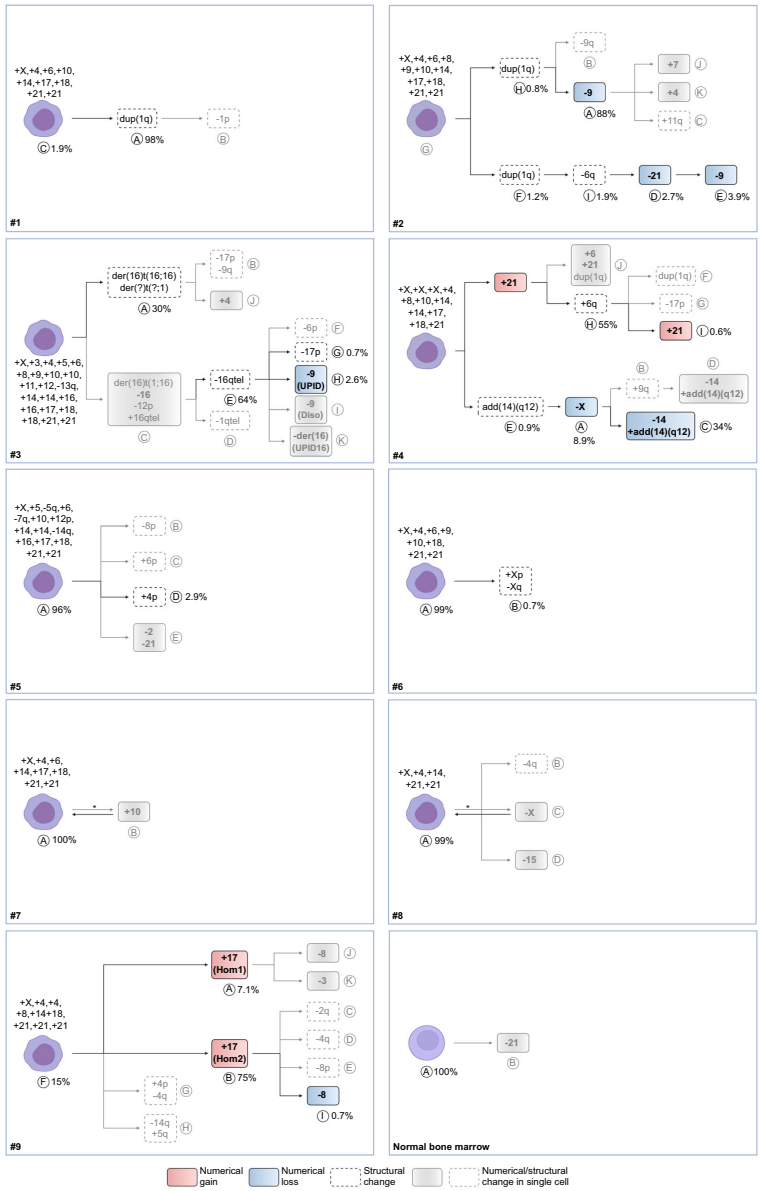


Fig. 2 | Phylogenetic trees showing the most probable course of genetic evolution, based on single cell whole genome sequencing (scWGS), bulk WGS, and fluorescence in situ hybridization in nine primary childhood acute lymphoblastic leukaemia cases and one normal bone marrow. The bulk of the chromosomal gains was present already in the inferred earliest cells, with 1–2 chromosomes being gained or lost during clonal evolution in some of the cases. *indicates that the direction of the clonal evolution cannot be determined. Diso heterodisomy, Hom1 homologue 1, Hom2 homologue 2, UPID uniparental isodisomy. Created with BioRender.com.

Chromosomes that recurrently differed between diagnostic and relapse samples were chromosomes 8 (17%), 4, 9, 21 (13%), and X, 7, 10, and 15 (8.7%). Trisomy 8 displayed signs of positive selection, as it never went from trisomy to UPID and as the trisomy involved different homologues in one case. Chromosome 9, on the other hand, displayed UPID in one sample and heterodisomy in the other in three cases, indicating an original clone with trisomy 9 that was selected against.

Altogether, analyses of the frequencies of chromosomal gains, subclonality patterns, and paired diagnostic/relapse samples suggested that the chromosomal gains in HeH ALL can be divided into three groups based on the selective pressures: chromosomes X, 4, 6, 10, 14, 17, 18, and 21, which are associated with strong positive selection (group strong-pos), chromosomes 5, 8, 11, 12, and 22, which are associated with weaker positive selection (group weak-pos), and chromosomes Y, 1–3, 7, 9, 13, 15, 16, 19, and 20, which are neutral or associated with negative selection (group neg).

Simulation of HeH development suggests formation by a tripolar mitosis

To understand further how the aneuploidy in HeH ALL arises, we next simulated hyperdiploidy development in silico under different scenarios. We included five possible routes to aneuploidy that have been reported to occur in cancer[17]: (1) sequential gains in a diploid cell (diploid/sequential), (2) initial tetraploidy followed by chromosomal losses (tetraploid/sequential), (3) tripolar division in a diploid cell (diploid/tripolar), (4) tripolar division in a tetraploid cell (tetraploid/tripolar), and (5) mitotic catastrophe resulting from complete loss of sister chromatid cohesion (mitotic catastrophe) (Supplementary Fig. 8). For models 3, 4, and 5, the simulation started with an abnormal mitosis directly resulting in aneuploid daughter cells according to the respective mechanism, followed by a low likelihood of nondisjunction of individual chromosomes, whereas mechanisms 1 and 2 started with a diploid or tetraploid cell, respectively, followed by individual nondisjunction events. First, we only included positive selection for the strong-pos group of chromosomes, i.e. X, 4, 6, 10, 14, 17, 18, and 21, with gains of chromosome 21 given the highest selective advantage based on its ubiquitous presence in these leukaemias. Briefly, 50,000 virtual cells were followed over multiple generations, with gain of strong-pos chromosomes increasing survival probability in the daughter cells and other nondisjunction events lowering it. Since the UPID frequency in the patient cohort was constant at 2.5% for non-strong-pos chromosomes (except chromosome 9), simulations were stopped when this level was reached. The resulting virtual cell populations were then compared with the chromosomal patterns in the 577 primary HeH ALLs.

All models resulted in a continuous increase in the UPID frequency over generations (Supplementary Fig. 9A). For the diploid/tripolar and diploid/sequential models, UPID frequencies of 2.5% were reached after 50–800 generations (median 72.5 and 485, respectively) and for the tetraploid/sequential model within 10 generations. For the tetraploid/tripolar and mitotic catastrophe models, the initial UPID frequency was >2.5% (18.2% and 9.9%, respectively) and plateaued at >30% after 1000 generations. Since this was inconsistent with the patient data, they were removed from further testing.

Next, we investigated the average number of trisomies/tetrasomies at different modal chromosome numbers (MCN). Interestingly, a marked elevation change was observed at MCN 62 for trisomies in the patient cohort (Fig. 3a), indicating that there may be two subgroups with different trisomy:tetrasomy ratios: MCN 51–61 ($n = 545$) and MCN 62–67 ($n = 32$), respectively. This suggests that HeH ALL with lower and higher MCN could arise through different mechanisms. Therefore, we investigated these groups separately in the following analyses.

Starting with MCN 51–61, we observed that the tetraploid/sequential model resulted in very few such cells (Supplementary Fig. 9B). We therefore concluded that this model could not give rise to HeH with MCN 51–61 and excluded it from further testing. We then compared the pattern of trisomies and tetrasomies at different MCN (Fig. 3a) and the pattern of trisomies and tetrasomies for each chromosome (Fig. 3b) between the HeH ALL patient data and the simulations results by the root mean squared error (RMSE) method (Supplementary Table 3). The diploid/tripolar and diploid/sequential models both showed low RMSE values, indicating that they fit relatively well with the patient data. We next looked at the frequency of tetrasomy 21 of the 2:2 type (duplication of both homologues) and 3:1 type (triplication of one homologue). In the diploid/sequential model, the 3:1 type was enriched during the simulation process, resulting in 64% tetrasomy 21 of this type. However, the patient data and the diploid/tripolar model both showed lower proportions of tetrasomy 3:1 (6.6% and 21%, respectively), supporting a diploid/tripolar origin. Notably, 3:1 tetrasomies were not an indication of one homologue being selected for, but rather resulted from the strong overall selection for extra copies of chromosome 21 in both models. To see if we could fine-tune the diploid/tripolar further, we included positive selection also for the weak-pos chromosomes. The modified version yielded even lower RMSE values than the original one (Supplementary Table 3). Hence, our simulations showed that the diploid/tripolar model consistently resulted in virtual cells with karyotypes similar to those seen in HeH ALL with MCN 51–61. Furthermore, sampling of the simulation results over consecutive generations showed that the diploid/tripolar model displayed whole chromosome copy number evolution consistent with a punctuated evolution model, with an initial sharp rise in chromosome numbers (Supplementary Fig. 10).

Next, we turned to the MCN 62–67 group, again studying the pattern of trisomies and tetrasomies at different MCN and for different chromosomes. Here, both the diploid/tripolar and the tetraploid/sequential models agreed well with the patient data (Supplementary Table 3), with the tetraploid/sequential model more closely following the distribution of average number of trisomies and tetrasomies across MCNs (Fig. 3a). We included selection for the weak-pos chromosomes also here and, since the UPID frequency is higher at higher MCNs, let the simulations run to a UPID frequency of 5%. Both the diploid/tripolar and tetraploid/sequential models resulted in virtual cells that fit well with the patient data (Supplementary Table 3). Interestingly, looking at the patient copy number data, cases with MCN 62–67 had more subclonality, with half of these cases (16/32) harbouring ≥1 subclonal chromosome; significantly higher than observed in the other HeH ALLs ($P = 0.0006$; two-sided Fisher's exact test). Furthermore, the only case in the scWGS analysis with MCN in this range (#3) also had a relatively high heterogeneity score.

Taken together, our modelling in conjunction with the patient data suggested that, in most instances, high hyperdiploidy in paediatric ALL arises by a tripolar division in a diploid cell. However, HeH ALL with MCN 62–67 (comprising around 5% of cases) may possibly arise by initial tetraploidy followed by chromosomal losses.

Chromosomal age pattern validates a tripolar division origin

In the diploid/tripolar model, most chromosomes are gained in the initial mitosis but some are gained and fixed during clonal selection. Seeking to validate our results from the in silico modelling, we reasoned that chromosomes that are gained later during clonal evolution should primarily be those that give a selective advantage, i.e. the strong-pos and weak-pos groups. In contrast, neg chromosomes would all have been gained at the initial division since they would not be selected for, although some may arise later due to drift. Therefore, we hypothesized that strong-pos and weak-pos trisomies should, on average, be newer than neg trisomies. If the hyperdiploidy instead arose by sequential gains, there would be no difference in the ages of the trisomies between these groups, as they could arise in any order (Fig. 4a).

https://doi.org/10.1038/s41467-023-37356-5

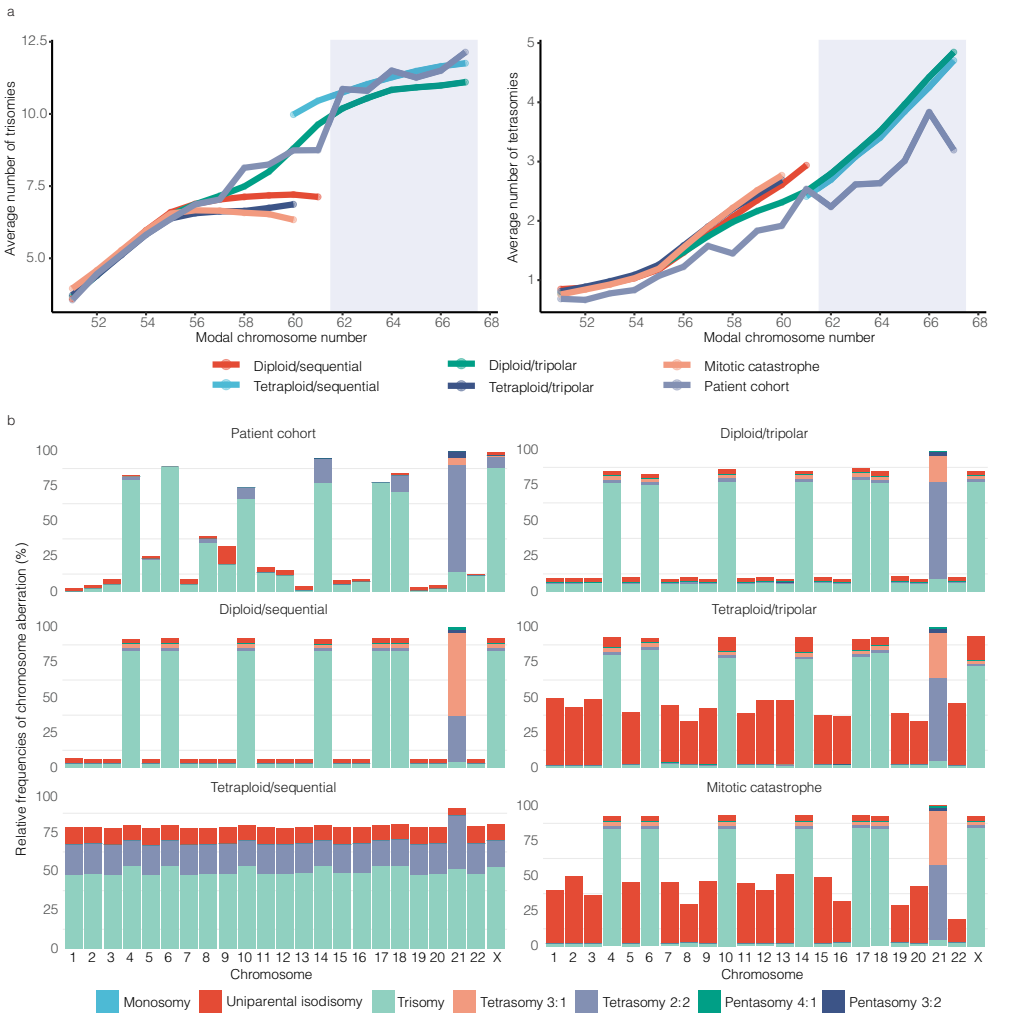


Fig. 3 | Simulation of high hyperdiploidy development in childhood acute lymphoblastic leukaemia according to five different models: (1) sequential gains in diploid cell (diploid/sequential), (2) initial tetraploidy followed by chromosomal losses (tetraploid/sequential), (3) tripolar division in a diploid cell (diploid/tripolar), (4) tripolar division in a tetraploid cell (tetraploid/tripolar), and (5) mitotic catastrophe resulting from complete loss of sister chromatid cohesion (mitotic catastrophe). Data shown are from the end point in the simulations. **a** Correlation between the average number of trisomies/tetrasomies and the modal chromosome number (MCN) in the simulation results and the patient cohort of 577 cases of high hyperdiploid ALL. The average number of trisomies at each modal number in the diploid/tripolar model closely follows what is seen in the patient cohort at MCN 51–61, indicating a very good fit of the model to patient data. At MCN 62–67, there is a sharp increase in the average number of trisomies per modal number in the patient cohort (indicated by a grey square), and it follows the tetraploid/sequential model more closely, possibly indicating a different mechanism. The average number of tetrasomies at each modal number in the patient cohort is based on fewer chromosomes (since tetrasomies are less common than trisomies) and follows most closely the diploid/tripolar and the tetraploid/sequential models for MCN 51–61 and MCN 62–27, respectively. **b** Pattern of chromosomal copy number changes and uniparental isodisomies resulting from the simulations according to each model and in the patient cohort. Frequency of each type of aberration (as given in the legend) is seen on the Y axis and each chromosome (except Y) on the X axis. Whereas the tetraploid/sequential, tetraploid/tripolar, and mitotic catastrophe model all result in chromosomal patterns very different from the one seen in the patient cohort, the diploid/sequential model, diploid/tripolar model, and patient cohort display relatively similar patterns. However, based on the high frequency of 3:1 tetrasomies in the diploid/sequential model it could be excluded, leaving the chromosomal pattern resulting from the diploid/tripolar model most similar to the one seen in the primary cases. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-023-37356-5

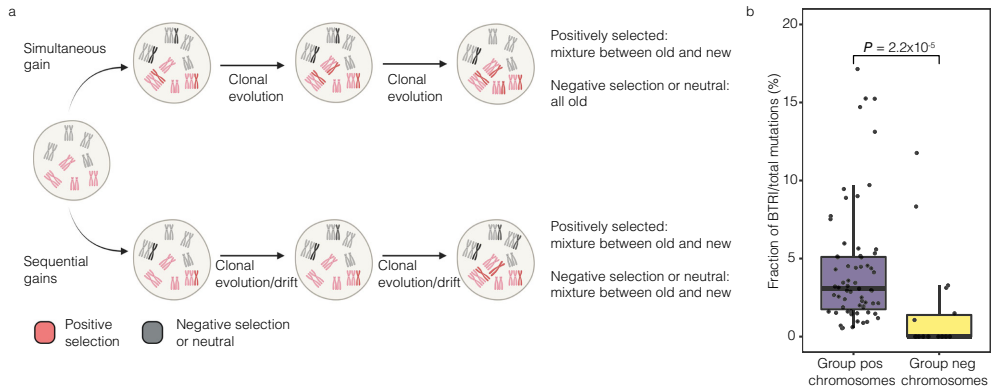


Fig. 4 | Analysis of somatic single nucleotide variants (SNVs) in trisomic chromosomes supports a simultaneous gain of most chromosomes followed by clonal evolution in high hyperdiploid (HeH) childhood acute lymphoblastic leukaemia (ALL). a Schematics of chromosomal gains in a scenario where chromosomes are either gained predominately by a first hit or via sequential non-disjunction. Trisomies subjected to positive selection (strong-pos and weak-pos trisomies) are shown in red and trisomies subjected to negative selection or neutral (neg trisomies) in grey. These two scenarios are expected to result in different mixtures of older and newer chromosomes. **b** Boxplots of the fraction of BTRI mutations in groups strong-pos/weak-pos versus group neg trisomies in 67 cases of HeH ALL. Groups strong pos/weak trisomies have a higher fraction of BTRI mutations ($P = 2.2 \times 10^{-5}$ Mann–Whitney two-sided test), indicating that they are on average newer, consistent with an initial tripolar cell division followed by clonal evolution. The centre of the boxplot is the median and lower/upper hinges correspond to the first/third quartiles; whiskers are 1.5 times the interquartile range and data beyond this range are plotted as individual points. Figure 4a was created with BioRender.com. Source data are provided as a Source Data file.

To investigate the age of different trisomies, we studied somatic single nucleotide variants (SNVs) in trisomies based on WGS in 67 HeH ALL. We utilized that SNVs that are present already before the trisomy forms (BTRI mutations) will be duplicated if they are in the gained homologue and display variant allele frequencies (VAFs) of ~0.67. Conversely, SNVs that arise after the trisomy or in the homologue that is not duplicated will only be present in one of the homologues and display VAFs of ~0.33 (B/ATRI mutations) (Fig. 5a, b). Hence, the proportion of SNVs that are of the BTRI type will be higher the newer the trisomy is, since the chromosome will have spent a longer time as not duplicated, allowing time for more mutations to arise. Among the 67 investigated cases, 536 of 15,828 SNVs (3.39%) in groups strong-pos and weak-pos chromosomes were of the BTRI type and 9 of 819 (1.09%) in group neg chromosomes ($P = 2.2 \times 10^{-5}$; Mann–Whitney two-sided test) (Fig. 4b). Thus, the former chromosomal gains were on average newer than the remaining trisomies, in line with what would be expected from a diploid/tripolar origin.

Mutational signatures show different etiological factors during leukemogenesis

To gain further insight into the leukemogenesis of HeH ALL, we studied mutational signatures in 67 cases with bulk WGS data. We investigated BTRI and B/ATRI mutations in trisomic chromosomes and relapse-specific mutations, since these groups can be put into a distinct timeline (Fig. 5a, b). BTRI mutations were predominantly associated with mutational signatures SBS1 and SBS5 (Fig. 5c); known clock-like signatures likely caused by intrinsic mutational processes[18,19]. Their high frequency at the earliest time point, before the hyperdiploidy arises, agrees well with an early origin devoid of environmental exposure. B/ATRI mutations displayed a wider range of mutational signatures, with SBS1, SBS5, SBS7a, SBS8, SBS18, SBS19, and SBS39 all contributing (Fig. 5c). Of these, SBS7a has been associated with ultraviolet light exposure[20]; this signature has previously been reported to dominate in some cases of aneuploid childhood ALLs[10,21] and it was present in six (9.0%) cases. SBS8 has been suggested to be associated with late replication errors[22], whereas SBS18 has been linked to mutagenesis by reactive oxygen species[23]. SBS19 and SBS39 have unknown etiologies[20]. Mutations specific for the relapse samples, which represent the latest mutations, were similar to the B/ATRI mutations, but with addition of signatures SBS15, SBS26, and SBS87 (Fig. 5c), as has previously been reported for the TARGET cohort[24]. SBS15 and SBS26 are associated with defective DNA mismatch repair[24], whereas SBS87 is associated with thiopurine treatment and hence likely induced by chemotherapy[24].

Temporal analysis of additional somatic events shows that the chromosomal gains are early

To determine when other somatic genetic events occur in relation to the chromosomal gains, we analyzed structural rearrangements, deletions, and mutations, focusing on (1) subclonality and (2) events occurring in gained chromosomes or UPIDs, where the temporal order could be investigated by looking at the allelic patterns.

For structural rearrangements, the analysis comprised known drivers that can be identified from copy number data: dup(1q), deletions of 6q [del(6q)], isochromosomes 7q [i(7q)], and partial gains of 17q (gain_17q)[25,26]. Of these, dup(1q), del(6q), and gain_17q were frequently subclonal (30–40% of cases), whereas i(7q) was generally present in the main clone (Supplementary Table 4). One case had two different subclonal dup(1q), similar to #2 in the scWGS analysis (Fig. 2). Furthermore, analysis of BTRI and B/ATRI mutations showed a significantly higher proportion of BTRI mutations in dup(1q) than in trisomies (median 29.4% vs. 4.3%; $P = 4.9 \times 10^{-4}$; Mann–Whitney two-sided test), indicating a later origin (Supplementary Fig. 11). Temporal order could be determined for dup(1q) and del(6q), showing that 8/8 and 22/22 informative cases, respectively, arose after the UPID or chromosomal gain (Supplementary Table 4).

Deletions of *IKZF1*, *CDKN2A*, *PAX5*, *ETV6*, *CREBBP*, and *TCF3*[26,27] were subclonal in 10–40% of the cases (Supplementary Table 4). Temporal analysis showed that 15/16 *CDKN2A* deletions, 1/1 *PAX5* deletion, 8/9 *ETV6* deletions, and 1/1 *CREBBP* deletion occurred after the respective UPID or trisomy. Thus, most informative deletions happened after the respective chromosome became

https://doi.org/10.1038/s41467-023-37356-5

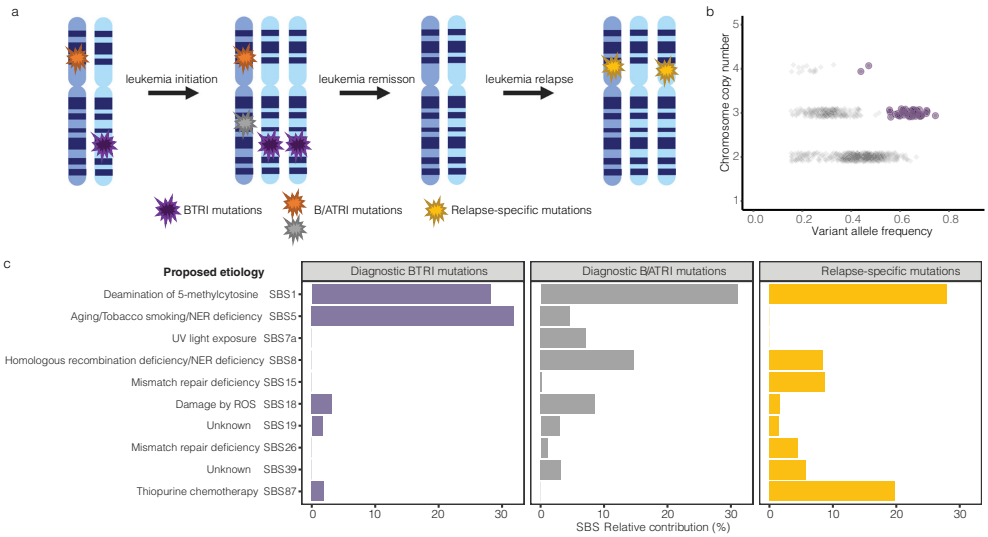


Fig. 5 | Patterns of somatic mutations in high hyperdiploid acute lymphoblastic leukaemia. a Timeline of mutations arising either before the trisomies (BTRI mutations), before/after the trisomies (B/ATRI mutations), and at relapse. **b** Variant allele frequencies (VAFs) in bulk sequencing data of scWGS case 1 for all mutations in relation to copy number based on the number of reads for that particular chromosome segment. BTRI mutations, which are present in 2/3 chromosomal homologues and have VAFs of ~0.67, are shown in purple and B/ATRI mutations, which are present in 1/3 chromosomal homologues and have VAFs of ~0.33, are shown in grey. **c** Relative contribution of the ten most common SBS signatures in BTRI mutations in trisomic chromosomes, B/ATRI mutations in trisomic chromosomes, and mutations specific for relapse samples. NER nucleotide excision repair, ROS reactive oxygen species, SBS single base substitutions, UV ultraviolet. Figure 5a was created with BioRender.com. Source data are provided as a Source Data file.

trisomic, but one somatic *CDKN2A* deletion and one *ETV6* deletion (which we have previously shown to be constitutional[26]) occurred at a disomic state.

Finally, we looked at 338 driver mutations in 218 cases where WES or WGS data were available. Of these, 150 (44%) were subclonal (Supplementary Tables 4 and Supplementary Data 3). For clonal mutations in trisomies, tetrasomies, or UPIDs, mutations were also classified as B/ATRI or BTRI. Sixty (92%) B/ATRI mutations and five (8.7%) BTRI mutations were found, including one *IKZF1* mutation in a case with UPID7 (Supplementary Data 3).

Taken together, the analysis showed that structural rearrangements, deletions, and mutations were frequently subclonal and generally occurred after the chromosomal event, supporting an overall scenario where the hyperdiploidy arises first and other somatic aberrations occur at later stages.

Discussion

We have performed a detailed analysis of the genetic mechanisms and the temporal order of different genetic events in HeH ALL. Using a combination of scWGS and in-depth analysis of SNP array and sequencing data from a large cohort of cases, we show that the chromosomal gains are early events and relatively stable throughout leukaemia development, whereas structural rearrangements and mutations generally occur later. In silico simulations of high hyperdiploidy development suggested that an initial tripolar division in a diploid cell, followed by clonal selection, best recapitulated the chromosomal patterns seen in patient samples.

Whether HeH ALL exhibits CIN has been debated. Cytogenetic data as well as bulk copy number analysis with SNP arrays have suggested that these leukaemias generally are chromosomally stable, with most cells displaying the same chromosomal gains[5,26,28]. However, cytogenetic analyses only comprise the dividing cells and misclassification of chromosomes can lead to underestimation of heterogeneity, whereas SNP arrays cannot detect all minor clones. Several previous studies have also used interphase FISH to investigate chromosomal heterogeneity[14,29–31], but with variable results and conclusions, likely due to a high degree of technical artifacts[5]. Here, we used scWGS to circumvent the above problems. This method is superior for characterizing copy number heterogeneity by including all cells—also non-dividing—and due to unequivocal identification of chromosomes[32]. We found relatively little chromosomal heterogeneity, with non-clonal numerical changes seen in only 12 (0.47%) of all 2572 leukaemic cells sequenced and 5/9 cases having identical chromosomal content in >99% of the cells (Fig. 1). Of the remaining four cases, three displayed subclones that were also detectable by SNP array analysis, and only one appeared to have a single clone by SNP array analysis when in fact it had several minor clones. Thus, scWGS strongly supports that HeH ALL is chromosomally stable, in line with cytogenetic and SNP array data. Notably, this also shows that aneuploidy in cancer does not lead to CIN per se; something that has also been debated[33,34].

Although HeH ALL overall appeared stable, there was nevertheless some variation in heterogeneity between cases. We recently reported a high but varying frequency of sister chromatid cohesion defects in HeH ALL, possibly associated with low levels of cohesin and/or condensin[14]. Case 2 had cohesion defects in 85% of the metaphase cells, possibly explaining the high heterogeneity in this case. However, the remaining eight cases all had percentages of cells displaying cohesion defects that were at or below the median value (21%) in our previous study[14] (Table 1). Thus, it is possible that we would have found more chromosomal heterogeneity by scWGS if more cases with severe cohesion defects had been included in this study.

Several mechanisms have been suggested for how the extra chromosomes in HeH ALL are gained, including one abnormal mitosis, loss of chromosomes from a tetraploid cell, sequential gains due to CIN, and, most recently, fusion of a mitotic cell and a G0/G1 cell[11–13,29,35]. Any such mechanism should conform to/explain a number of features of HeH ALL genomes: (1) the specific pattern of trisomies, tetrasomies, and low-level UPIDs, including why 2:2 tetrasomies are much more common than 3:1 tetrasomies, (2) the presence not only of the common trisomies but also of gains (at low frequency) of all chromosomes, (3) the relative chromosomal stability shown by our scWGS analysis, and (4) that strong-pos and weak-pos chromosomes are on average newer (occur later during leukemogenesis) than neg chromosomes, as evidenced by our analysis of B/ATRI and ATRI mutations. To test which of the proposed mechanism(s) that conformed to the first of these features, we performed in silico modelling (excluding the fusion model since its outcome could not be statistically predicted) and compared the outcome with the chromosomal patterns seen in a large cohort of HeH ALL. We found that an initial tripolar mitosis that leads to gain of the bulk of the extra chromosomes, followed by clonal evolution over multiple generations of cells, recapitulated the chromosomal and allelic patterns seen in the patient samples. Furthermore, this mechanism can also explain why the low frequency trisomies occur, as they are passenger events that are gained in the initial tripolar division, as well as why they are on average older than the high frequency trisomies, which may also arise and be fixated later due to positive selection pressure. Finally, the diploid/tripolar model does not require chromosomal instability for aneuploidy to occur within a reasonable (considering the young age of the patients) time frame, as the bulk of the chromosomal gains occur very early (also in line with previous data showing hyperdiploidy years before overt diagnosis of HeH ALL[6–10]). Notably, tripolar cell divisions have been reported to occur in cancer and lead to viable daughter cells[4,36] that potentially could regain mitotic stability by clustering or loss of supernumerary centrosomes. Thus, no evidence of this initial mitotic error apart from the allelic patterns would still be visible at the time of diagnosis.

The punctuated evolution model in cancer states that somatic aberrations arise in short bursts of time very early in tumour evolution[1]. By scWGS, all cases showed phylogeny in line with this, with few intermediate cells indicating gradual evolution, long truncal distances, and short branching distances (Supplementary Fig. 3). The diploid/tripolar model that we suggest underlies the extra chromosomes in HeH ALL is a clear example of a way that such punctuated evolution for whole chromosome copy number changes could occur. Our results thus support previous studies showing frequent punctuated evolution for copy number changes in malignancies[3,4].

We found a possible difference in the chromosome distribution in HeH cases with MCN 62–67. Heerema et al.[37] reported that cases with MCN 63-67 have different chromosomal gains than HeH ALL with lower MCN, in line with them being a separate entity genetically. Furthermore, we and others have previously shown that cases with higher MCN have a significantly better prognosis[38,39], indicating that they also differ clinically. However, it should be noted that due to the rarity of cases with MCN in this span, we cannot exclude that the observed differences in chromosome distribution were due to chance only. Both a diploid/tripolar and a tetraploid/sequential mechanism agreed relatively well with the chromosomal patterns in the patient cohort, and further studies are needed to ascertain how HeH ALL with MCN 62–67 arises.

In conclusion, we present a model for the leukemogenesis of HeH paediatric ALL where most cases are initiated by an erroneous tripolar mitosis, after which they undergo low-level clonal evolution to optimize their chromosomal pattern and gain additional driver events that eventually leads to overt leukaemia several years later. This model agrees well with a wealth of previous observations, including the early occurrence of the chromosomal gains[6–10], chromosomal and allelic patterns[11–13], and general genomic stability[5,26,28] in this disease. Furthermore, it strengthens the evidence that copy number changes and aneuploidy frequently arise by punctuated evolution at the early stages of tumorigenesis and that aneuploidy-driven malignancies do not necessarily have high levels of chromosomal copy number heterogeneity and CIN.

Methods

Single cell WGS

All investigations complied with relevant ethical regulations. Written informed consent was obtained from the patients and/or their guardians according to the Declaration of Helsinki and the study was approved by the Ethics Committee of Lund University, Sweden. No monetary compensation was offered for patient participation. Viable bone marrow cells obtained at diagnosis from nine patients with high hyperdiploid ALL and one healthy individual, selected on the basis of sample availability, were subjected to low-pass scWGS. Single nuclei in G_0/G_1 phase were isolated using a fluorescence-activated cell sorting (FACS) cytometer and DNA libraries were constructed for multiplexed whole genome sequencing with average sequencing depth between 0.006x to 0.089x per cell (median 0.02x)[40]. Sequencing reads were aligned to the UCSC human reference genome (hg19, [http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/]) using the Burrows-Wheeler Aligner (BWA, v0.7.17)[41]. The aligned reads were sorted and merged with SAMtools (v1.9)[42]. The copy number state of each chromosome was determined using AneuFinder (v1.14)[32]. Briefly, duplicate reads, low-quality alignments (MAPQ < 20), and reads falling into the regions specified by the blacklists provided by AneuFinder were discarded. Read counts in 2.5 Mb, 5 Mb, and 10 Mb variable-width bins were GC-corrected and copy number states were determined using the edivisive algorithm with copy-number states nulli-, mono-, di-, tri-, tetra-, penta-, and hexasomy. The copy number state was also determined by Ginkgo[43] using default settings with a bin size of 1 Mb. All data were manually curated and in the final heatmaps, breakpoints were aggregated depending on supportive data from WGS, SNP array, and FISH. scWGS phylogenetic trees were constructed using MEDICC2[44] and subsequently manually curated to accommodate structural rearrangements by combining scWGS, WGS, SNP array, and FISH results for some cases. Pairwise distances of single cells and simulated normal diploid cells were calculated using Manhattan distance by R (version 4.1.2) to obtain a distance matrix for each tumour. Phylogenetic inference for single cell trees and consensus trees were performed with the balanced minimum evolution algorithm from R package ape (v5.6)[45]. Normal diploid nodes for phylogenetic tress were constructed from simulated variable binning profiles in which bins presented an integer copy number equal to 2 for autosomes and 1/2 for chromosome X depending on patient sex. Clones were defined as ≥2 cells presenting with the same numerical and/or structural aberrations. Genome-wide heterogeneity scores were obtained from AneuFinder. Homologue inheritance of chromosomes gained or lost was determined by screening for heterozygous variants identified from bulk WGS data. Briefly, heterozygous variants were called by GATK (v4.0.11.0) haplotypecaller[46] and the variants from trisomies and tetrasomies of 3:1 type and UPIDs were extracted. For trisomies/tetrasomies 3:1, heterozygous variants were assigned to different homologues based on the alternative allele frequency obtained from bulk WGS data. Variants with alternative allele frequency higher than 0.6 were assigned to one chromosomal homologue and variants with alternative allele frequency less than 0.4 were assigned to the other. For UPIDs that were found in diagnostic samples, remission-specific variants from the same chromosome were assigned to one chromosomal homologue and variants that showed heterozygosity in the remission sample but homozygosity in the matched diagnostic sample were assigned to the other homologue. Then variants informative for

chromosomal homologue were screened in scWGS data and the homologue inheritance of chromosomes gained or lost was determined by the ratio between the number of each type of variant in the single cells.

Copy number analysis of bulk data

Log R ratio (LRR) and B allele frequency (BAF) of SNP array data from Illumina (.idat files) and Affymetrix (.CEL files) intensity files were analyzed by Illumina GenomeStudio (v2.0, Illumina, San Diego, CA) and Affymetrix Analysis Power Tools (v2.10.0, Thermo Fisher Scientific Inc., Waltham, MA), respectively. Copy number alterations were called using TAPS[47] and manually reviewed in GenomeStudio or Chromosome Analysis Suite (v3.3, Thermo Fisher Scientific Inc., Waltham, MA). Subclonality of whole chromosomes was assessed using the TAPS software from SNP array, WES, or WGS data, considering LRR, BAF, and tumour purity. Depending on the type of subclonality (disomy/trisomy, UPID/trisomy, etc.), the lower limit of detection of subclones was estimated to 20-30% of the cells. The dataset included four different cohorts: from our Department[26], Zaliova et al.[48], Duployez et al.[49], and The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program (dbGAP accession number phs00464) (Supplementary Data 1). Of those 577 cases, 253 (44%) were females and 324 (56%) were males, based on the absence or presence of a Y chromosome.

For WES data from TARGET, paired-end reads were aligned to the human reference genome hg19 by the bwa[41]. Duplicate reads marking and local realignment were performed by GATK[46]. Constitutional variants of matched tumour/normal pairs were called by GATK HaplotypeCaller and the bedtools (v2.27.1) intersect was used to extract the variants in the regions targeted by the exome sequencing kit. After normalizing read counts of constitutional mutation sites to the sequencing depth, the LRR of the constitutional variants was then calculated by the log-odds ratio of the variant allele count in the tumour versus in the normal. Reference allele frequency of constitutional variant sites was defined by the reference allele count versus total sequencing depth of the constitutional variant site in the tumour sample.

Paired diagnostic and relapse samples have been previously published[16] or were from TARGET. To investigate the chromosomal homologue involved in paired diagnostic and relapse samples, heterozygous variants from trisomies and tetrasomies 3:1 were extracted and assigned to different homologues based on the BAF of the diagnostic sample. Variants with BAF higher than 0.6 were assigned to one chromosomal homologue and variants with BAF less than 0.4 were assigned to the other. Then variants informative for chromosomal homologues were screened in the relapse sample to determine the involved chromosomal homologue. For UPIDs that were found in diagnostic samples, variants with BAF higher than 0.8 were screened in the paired relapse sample and homologue inheritance of chromosomes was determined by the BAF of corresponding variants in the relapse sample.

WGS data analysis and identification of BTRI and B/ATRI mutations

WGS data from 14 BCP ALL cases have been previously published[10]. The initial putative somatic mutations were identified by the Complete Genomics Cancer Sequencing pipeline and the data were further filtered for Somatic Score ≥0 and number of unique reads for the mutated allele >10. For Complete Genomics data generated by the TARGET program ($n = 34$), somatic variants were identified by the TARGET WGS analysis pipeline ([https://ocg.cancer.gov/programs/target/target-methods#3233]). Illumina WGS sequencing libraries of nineteen matched diagnostic and remission bone marrow or peripheral blood samples diagnosed at Skåne University Hospital, Sweden, were constructed by the TruSeq Nano DNA sample preparation kit (Illumina, San Diego, CA, USA). Paired-end sequencing (2x150bp) was done to ~60x coverage for diagnostic samples and ~30x coverage for remission. Somatic variants were identified by the GDC DNA-Seq analysis pipeline ([https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/]). Whether mutations occurred before (BTRI) or before/after (B/ATRI) trisomy formation were determined by mutant allele fractions according to Paulsson et al.[10]. Driver genes/mutations were identified by MutsigCV[50] and DriverPower[51]. A literature review focusing on genes identified by bulk WGS sequencing as targeted by non-silent somatic mutations associated with the search terms “aneuploidy”, “instability” and “cohesin” was performed in order to investigate whether mutations in genes affecting genomic stability were responsible for the heterogeneity observed within the nine scWGS cases.

Mutational signatures analysis

The R package MutationalPatterns[52] (v3.4.1) was used to decompose mutational profiles into pre-defined single base substitution (SBS) mutational signatures based on the Sanger mutational signatures (v3.2 - March 2021) and to ascertain the relative contributions of the SBS mutational signatures for BTRI and B/ATRI mutations in trisomic chromosomes at diagnosis, and all informative relapse-specific mutations.

Cohesion assay and FISH

Sister chromatid cohesion was analyzed in metaphase spreads in all nine HeH ALL patient samples subjected to scWGS. The percentage of cells with cohesion defects, measured as visible primary constriction gaps (gaps between the sister chromatids at the centromeres)[14], was counted. FISH metaphase spreads mounted with DAPI were used for the assay, where 20–39 cells were analyzed per case. Images were captured using a Z2 fluorescence microscope (Zeiss, Germany) and the CytoVision software (v7.4, Leica, Germany).

Metaphase FISH was carried out on cases 2, 3, and 4 according to standard methods, with a total of 17–35 cells captured for each analysis. All whole chromosome paint FISH probes were acquired from Applied Spectral Imaging (Carlsbad, CA), and locus-specific probes from Vysis (Abbot Laboratories, Chicago, IL). FISH analysis was performed as follows: slides from case 2 were hybridized with whole chromosome paint probes for chromosomes 1 (Aqua – blue), 6 (Cy3 – red), and 21 (FITC – green); for case 3, whole chromosome paint probes for chromosomes 1 (FITC) and 16 (Aqua) were used together with a telomeric probe for 16q (Cy3); and for case 4, one analysis was performed with whole chromosome paint probes for chromosomes 3 (FITC) and 6 (Cy3), and another analysis for chromosome 14 (Aqua) together with a LSI TRA/D (14q11.2) break-apart dual colour probe (Cy3/FITC).

Simulation of high hyperdiploidy development

To investigate the development of aneuploidy observed in HeH ALL, we constructed an algorithm to simulate the clonal expansion and to trace single-cell karyotypes over two thousand generations using the Python programming language (v2.7.15). For each model, 50,000 virtual cells were created and the copy number of individual chromosomes was defined according to the initial hit based on the simulation model: sequential gains in a diploid cell (diploid/sequential), initial tetraploidy followed by chromosomal losses (tetraploid/sequential), tripolar division in a tetraploid cell (tetraploid/tripolar), tripolar division in a diploid cell (diploid/tripolar), and mitotic catastrophe (mitotic catastrophe). For simplicity, all scenarios started with 46,XX cells (the Y chromosome was not included in the analysis). All virtual cells were represented by a $23 \times 50{,}000$ matrix. For the virtual cells (C_g) at generation g, $C_g(i)$ was the copy number of a virtual cell for each of the 23 chromosomes indexed by i. During cell division, two daughter cells would be formed from the mother cell. The missegregation rate (M_{misseg}) was set to (15×10^{-4}/chromosome/

mitosis)[53] and the probability of missegregation (P_{misseg}) of each chromosome was weighted by the copy number of the given chromosome (N) and $P_{misseg} = M_{misseg}N$. Only one missegregation event of any given chromosome in a single cell division was allowed and the missegregated chromosome was randomly assigned to one of the two daughter cells. For tetraploid/sequential, the probability of chromosome loss was set to 35% according to previously published data[54]. Virtual cells with nullisomy were excluded from subsequent generations. Clonal expansion of virtual cells was altered by positive and negative selection of gain/loss of certain chromosomes. In the algorithm, we employed a survival/proliferation score (S_{score}) to determine the survival probability of virtual cells. Normal diploid cells were given a probability of 50% for proliferative survival. The S_{score} of the virtual cell was determined according to its karyotype. Virtual cells with trisomies X, 4, 6, 10, 14, 17, and 18 (group 1) and gain of chromosome 21 (group 2) were subjected to positive selection, whereas virtual cells with gain/loss of the remaining chromosomes (group 3) were subjected to negative selection. In addition, virtual cells were also subjected to negative selection pressure (aneuploidy penalty score, $S_{aneuploidy}$), which increased with the modal number of chromosomes ($MCN > 46$) of that cell. The S_{score} of the given virtual cell was computed according to:

$$S_{score} = 0.5 + \frac{NT_{g1} + 2NT_{g2} - NT_{g3}}{23} - S_{aneuploidy} \quad (1)$$

where NT_{g1} is the number of trisomic chromosomes in group 1, NT_{g2} is the number of trisomic chromosomes in group 2 and NT_{g3} is the number of trisomic chromosomes in group 3. The $S_{aneuploidy}$ was calculated by using the probability density function of beta distribution from python scipy package (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.beta.html) with the empirically determined location parameter loc 0, scale parameter scale 1.8, shape parameters a and b 0.18, 0.65, respectively. The x parameter was defined as:

$$x = \frac{MCN - 46}{46} \quad (2)$$

For the tetraploid/sequential model, no $S_{aneuploidy}$ was used since no initial tetraploid cell would survive under that condition.

In addition, an extended version (four groups version) of S_{score} was also used by dividing group 3 into two groups: one with negative selection for gain of chromosomes 1–3, 7, 9, 13, 15, 16, 19, and 20 (group 3b) and the other one with weak positive selection for gain of chromosomes 5, 8, 11, 12 and 22 (group 4). Then the S_{score} of the given virtual cell was computed according to:

$$S_{score} = 0.5 + \frac{NT_{g1} + 2NT_{g2} + 0.02N_{g4} - NT_{g3b}}{23} - S_{aneuploidy} \quad (3)$$

where NT_{g3b} is the number of trisomic chromosomes in group 3b and NT_{g4} is the number of trisomic chromosomes in group 4. To save the computational memory requirements for exponential cell growth, virtual cells that died were removed from subsequent generations and 50,000 cells were randomly sampled into subsequent generations. If the number of virtual cells was less than 50,000, cells with aneuploid karyotypes were drawn from the pre-defined model and added to the current generation. Simulations were stopped when the UPID frequency of chromosomes 1–3, 5, 7–8, 11–13, 15, 16, 19, 20, and 22 became 2.5% or terminated after 2000 generations. Fifty parallel runs were performed for each model. After the end of the simulation, one million virtual cells were randomly sampled from each model and the karyotype similarity between the patient cohort and sampled cells was measured using the RMSE method.

To investigate whether the aneuploidy developed by punctuated or gradual evolution in the diploid/tripolar and diploid/sequential models, ten thousand virtual cells were randomly sampled from each simulated generation and the corresponding median modal chromosome number was calculated. One hundred parallel runs were performed and smoothing regression analysis (LOESS) was used to model the relationship between the modal chromosome number and the number of generations.

Statistics and reproducibility

For assessing technical reproducibility, bulk WGS data from technical replicates represented by independent next generation sequencing libraries from the same DNA of 2 HeH samples (case L31 and case L74) were generated. A high correlation between the results from the two replicates was observed and over 97% of mutation sites were identified in the replication datasets. Since the reproducibility was very high, no additional replicates were generated. No statistical method was used to predetermine sample size. All cases with HeH where SNP array/WGS/WES data were available were included in the bulk copy number analysis, except for samples where the technical quality was too poor. The sister chromatid cohesion assay and the copy number variation calling were performed independently in a blinded fashion. All statistical tests were performed in R (version 4.1.2). The detailed statistical tests are indicated in figures or associated legends where applicable. No data were excluded from the analyses. None of the statistical tests used in this study required the assumption of normality or the assumption of equal variance. P values were calculated based on non-parametric tests that do not have degrees of freedom associated with a sampling distribution. A significance threshold of <0.05 was used for all statistical tests.

Reporting summary

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

Data availability

The scWGS data generated in this study have been deposited in the European Genome Archive (EGA) under accession number EGAS00001006347. The scWGS dataset is available under restricted access due to privacy concerns; access can be obtained for academic research by contacting the Data Access Committee via EGA. The processed somatic SNP array data and bulk WGS data are freely available through the following DOIs: https://doi.org/10.17044/scilifelab.21953114 (SNP array dataset) and https://doi.org/10.17044/scilifelab.21953117 (bulk WGS dataset). The raw SNP array data and bulk WGS data generated during the current study have been deposited to EGA under accession numbers EGAS00001007049 and EGAS00001007052, respectively. These datasets are available under restricted access due to privacy concerns; access can be obtained for academic research by contacting the Data Access Committee via EGA. The WGS data generated by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) are available under accession code phs000464. The human reference GRCh37 (hg19) used in this study is available in the UCSC Genome Browser [http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/]. Source data are provided with this paper.

Code availability

The code used to perform the analysis is available as supplementary code, also available on Zenodo[55].

References

1. Davis, A., Gao, R. & Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim. Biophys. Acta Rev. Cancer* **1867**, 151–161 (2017).

2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–8 (1976).
3. Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet* **48**, 1119–30 (2016).
4. Bollen, Y. et al. Reconstructing single-cell karyotype alterations in colorectal cancer identifies punctuated and gradual diversification patterns. *Nat. Genet* **53**, 1187–1195 (2021).
5. Paulsson, K. & Johansson, B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **48**, 637–60 (2009).
6. Szczepanski, T. et al. Precursor-B-ALL with D_H-J_H gene rearrangements have an immature immunogenotype with a high frequency of oligoclonality and hyperdiploidy of chromosome 14. *Leukemia* **15**, 1415–23 (2001).
7. Bateman, C. M. et al. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. *Leukemia* **29**, 58–65 (2015).
8. Maia, A. T. et al. Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins. *Leukemia* **17**, 2202–6 (2003).
9. Maia, A. T. et al. Identification of preleukemic precursors of hyperdiploid acute lymphoblastic leukemia in cord blood. *Genes Chromosomes Cancer* **40**, 38–43 (2004).
10. Paulsson, K. et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat. Genet* **47**, 672–676 (2015).
11. Paulsson, K. et al. Formation of trisomies and their parental origin in hyperdiploid childhood acute lymphoblastic leukemia. *Blood* **102**, 3010–3015 (2003).
12. Paulsson, K. et al. Evidence for a single-step mechanism in the origin of hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **44**, 113–22 (2005).
13. Onodera, N., McCabe, N. R. & Rubin, C. M. Formation of a hyperdiploid karyotype in childhood acute lymphoblastic leukemia. *Blood* **80**, 203–8 (1992).
14. Moura-Castro, L. H. et al. Sister chromatid cohesion defects are associated with chromosomal copy number heterogeneity in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **60**, 410–417 (2021).
15. Mullighan, C. G. et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science* **322**, 1377–80 (2008).
16. Davidsson, J. et al. Relapsed childhood high hyperdiploid acute lymphoblastic leukemia: presence of preleukemic ancestral clones and the secondary nature of microdeletions and RTK-RAS mutations. *Leukemia* **24**, 924–931 (2010).
17. Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **10**, 478–87 (2009).
18. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet* **48**, 126–133 (2016).
19. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).
20. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
21. Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
22. Singh, V. K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun. Biol.* **3**, 421 (2020).
23. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
24. Li, B. et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood* **135**, 41–55 (2020).
25. Herou, E., Biloglav, A., Johansson, B. & Paulsson, K. Partial 17q gain resulting from isochromosomes, unbalanced translocations and complex rearrangements is associated with gene overexpression, older age and shorter overall survival in high hyperdiploid childhood acute lymphoblastic leukemia. *Leukemia* **27**, 493–496 (2013).
26. Paulsson, K. et al. Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. USA* **107**, 21719–21724 (2010).
27. Mullighan, C. G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–64 (2007).
28. Raimondi, S. C. et al. Heterogeneity of hyperdiploid (51-67) childhood acute lymphoblastic leukemia. *Leukemia* **10**, 213–24 (1996).
29. Molina, O. et al. Impaired condensin complex and Aurora B kinase underlie mitotic and chromosomal defects in hyperdiploid B-cell ALL. *Blood* **136**, 313–327 (2020).
30. Betts, D. R., Riesch, M., Grotzer, M. A. & Niggli, F. K. The investigation of karyotypic instability in the high-hyperdiploidy subgroup of acute lymphoblastic leukemia. *Leuk. Lymphoma* **42**, 187–93 (2001).
31. Alpar, D. et al. Sequential and hierarchical chromosomal changes and chromosome instability are distinct features of high hyperdiploid pediatric acute lymphoblastic leukemia. *Pediatr. Blood Cancer* **61**, 2208–14 (2014).
32. Bakker, B. et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* **17**, 115 (2016).
33. Nicholson, J. M. et al. Chromosome mis-segregation and cytokinesis failure in trisomic human cells. *Elife* **4**, e05068 (2015).
34. Valind, A., Jin, Y., Baldetorp, B. & Gisselsson, D. Whole chromosome gain does not in itself confer cancer-like chromosomal instability. *Proc. Natl. Acad. Sci. USA* **110**, 21119–23 (2013).
35. Haas, O. A. Somatic sex: on the origin of neoplasms with chromosome counts in uneven ploidy ranges. *Front Cell Dev. Biol.* **9**, 631946 (2021).
36. Gisselsson, D. et al. Generation of trisomies in cancer cells by multipolar mitosis and incomplete cytokinesis. *Proc. Natl. Acad. Sci. USA* **107**, 20489–93 (2010).
37. Heerema, N. A. et al. Specific extra chromosomes occur in a modal number dependent pattern in pediatric acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **46**, 684–93 (2007).
38. Paulsson, K. et al. High modal number and triple trisomies are highly correlated favorable factors in childhood B-cell precursor high hyperdiploid acute lymphoblastic leukemia treated according to the NOPHO ALL 1992/2000 protocols. *Haematologica* **98**, 1424–1432 (2013).
39. Dastugue, N. et al. Hyperdiploidy with 58-66 chromosomes in childhood B-acute lymphoblastic leukemia is highly curable: 58951 CLG-EORTC results. *Blood* **121**, 2415–23 (2013).
40. van den Bos, H., et al. Quantification of aneuploidy in Mammalian Systems. in *Cellular Senescence: Methods and Protocols* (ed. Demaria, M.) 159-190 (Springer New York, New York, NY, 2019).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
42. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
43. Garvin, T. et al. Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–60 (2015).
44. Kaufmann, T. L., et al. MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. 2021. bioRxiv https://doi.org/10.1101/2021.02.28.433227
45. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
46. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11 10 1–11 10 33 (2013).

47. Rasmussen, M. et al. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol.* **12**, R108 (2011).
48. Zaliova, M. et al. Slower early response to treatment and distinct expression profile of childhood high hyperdiploid acute lymphoblastic leukaemia with DNA index <1.16. *Genes Chromosomes Cancer* **55**, 727–37 (2016).
49. Duployez, N. et al. Detection of a new heterozygous germline *ETV6* mutation in a case with hyperdiploid acute lymphoblastic leukemia. *Eur. J. Haematol.* **100**, 104–107 (2018).
50. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–8 (2013).
51. Shuai, S. et al. Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nat. Commun.* **11**, 734 (2020).
52. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. Mutational patterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
53. Valind, A., Jin, Y. & Gisselsson, D. Elevated tolerance to aneuploidy in cancer cells: estimating the fitness effects of chromosome number alterations by in silico modelling of somatic genome evolution. *PLoS One* **8**, e70445 (2013).
54. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278–82 (2009).
55. Woodward, E. L., et al. Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukemia. HeH_simulation: Simulation of high hyperdiploidy development. Zenodo. https://doi.org/10.5281/zenodo.6340286. 2022.

Acknowledgements

The results published here are in part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (https://ocg.cancer.gov/programs/target) initiative, phs000218. Figures 1, 2, 4, and 5 and Supplementary Figs. 4, 7, and 8 were created with BioRender.com. This study was supported by grants from the Swedish Childhood Cancer Foundation, grant numbers PR2020-0033 (MY), TJ2020-0024 (MY), PR2018-0004 (BJ), and PR2018-0023 (KP); the Swedish Cancer Fund, grant numbers 20 0792 PjF (BJ) and 19-0252-Pj (KP); Governmental funding of clinical research within the National Health Service, grant number ALFSKANE-623431 (KP); the Swedish Research Council, grant numbers 2020-01164 (BJ) and 2020-00997 (KP); IngaBritt och Arne Lundbergs Forskningsstiftelse, grant number LU2019-0100 (KP), the Gunnar Nilsson Cancer Foundation (MY), the Royal Physiographic Society of Lund (EW), the Czech Health Research Council, grant number (NU20-07-00322)) (MZ)), the University Hospital Motol, grant number #00064203 (MZ, JZ), and Program EXCELES, grant number LX22NPO5102 (MZ, JZ).

Author contributions

E.L.W., M.Y., and K.P. conceived the study; E.L.W., M.Y., L.H.M.-C., H.v.d.B., R.G., L.O.-A., and D.C.J.S. performed experiments and analyzed data; A.C., N.D., M.Z., J.Z., and B.J. provided clinical data and samples and analyzed data; F.F. supervised experiments and analyzed data; K.P. supervised the study; E.L.W., M.Y., and K.P. wrote the article with input from all authors.

Funding

Open access funding provided by Lund University.

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/s41467-023-37356-5.

Correspondence and requests for materials should be addressed to Kajsa Paulsson.

Peer review information *Nature Communications* thanks Hamim Zafar, William Carroll, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023

SUPPLEMENTARY INFORMATION FOR

Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukemia

Eleanor L Woodward[1], Minjun Yang[1], Larissa H Moura-Castro[1], Hilda van den Bos[2], Rebeqa Gunnarsson[1], Linda Olsson-Arvidsson[1,3], Diana CJ Spierings[2], Anders Castor[4], Nicolas Duployez[5,6], Marketa Zaliova[7,8], Jan Zuna[7,8], Bertil Johansson[1,3], Floris Foijer[2], Kajsa Paulsson[1]

[1]Department of Laboratory Medicine, Division of Clinical Genetics, Lund University, Lund, Sweden, [2]European Research Institute for the Biology of Ageing (ERIBA), University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, [3]Department of Clinical Genetics, Pathology, and Molecular Diagnostics, Office for Medical Services, Region Skåne, Lund, Sweden, [4]Department of Pediatrics, Skåne University Hospital, Lund University, Lund, Sweden, [5]Laboratory of Hematology, Centre Hospitalier Universitaire (CHU) Lille, Lille, France, [6]Unité Mixte de Recherche en Santé (UMR-S) 1172, INSERM/University of Lille, Lille, France, [7]Department of Pediatric Hematology and Oncology, Second Faculty of Medicine, Charles University/University Hospital Motol, Prague, Czech Republic, [8]Childhood Leukaemia Investigation Prague (CLIP), Prague, Czech Republic.

Supplementary Table 1. Patient and genetic data for nine cases of high hyperdiploid childhood acute lymphoblastic leukemias subjected to single cell whole genome sequencing

Case	Gender	Karyotype[a]	No. of somatic mutations	Genes targeted by non-silent somatic mutations	Other somatic changes
1	M	56,XY,+X,+Y,dup(1)(q21q32),+4,+6,+10,+14,+17,+18,+21,+21	1,045	*CCDC63, MARVELD2, MRAP2, NPAS3, NRAS, OR1D5, SCN4A, TLN2*	
2	M	56,XY,+X,dup(1)(q21q41),+4,+6,+8,+10,+14,+17,+18,+21,+21	1,005	*AC012593.1, BEND5, FLT3, GIF, PCDHGA5, RBFOX1, RP11-347L18.1, SLC7A4, SULF1*	Focal *ETV6* deletion
3	M	66,XY,+X,dup(1)(q21q44)x2,+3,+4,+5,+6,+8,+9,+10,+10,+11,+12, 13,del(13)(q21q31),+14,+14,+16,+16,del(16)(p12)x2,+17,+18,+18, +21,+21	1,043	*ATP8A2, CNTN1, HYDIN, NAV3, NOC2L, POLR3E, ROBO2, SPATA31D1, TCP10, TCTN3, TEKT4, TSEN34, UGT8*	
4	F	57,XX,+X,+X,der(3)t(3;6)(q29;q22),+4,+8,+10,+14,+del(14)(q12),+17, +18,+21,+21	1,108	*CDKN2A, DPP10, GPR4, HSPG2, HUWE1, IKZF1, LL22NC01-81G9.3, PCDH15, TBR1, TEX28, ZSCAN16*	
5	M	57,XY,+X,+del(5)(q23),+6,del(7)(q34),+10,+dup(14)(q11q22), +del(16)(p13),+17,+18,+21,+21,+mar	1,857	*ADAM12, AKAP6, COL4A6, CREBBP, DSCAML1, FLT3, HSD3B2, KLHL36, LAMA1, MAOA, MAPK4, MUC16, MUC17, MYPN, OTOA, PCSK6, RNF130, STON1-GTF2A1L, SULT1B1, TEP1, TGM2, TRRAP, TUBA4A, UBA6, ZNF804A*	UPID11, UPID13, subclonal focal *ETV6* deletion
6	M	54,XY,+X,+4,+6,+9,+10,+18,+21,+21	837	*ACOT1, DOT1L, FLT3, KRAS, SIGLEC1, STOX2, TENM3*	
7	F	54,XX,+X,+4,+6,+14,+17,+18,+21,+21	1,632	*ADH7, AR, ATF7IP, DHCR7, FAM27E1, FAM27E3, GATA5, GRIK2, HYDIN, KRAS, NRAS, OR56A1, PADI1, PTPN11, TBPL2, WDHD1*	Focal *PAX5* deletion
8	M	51,XY,+X,+4,+14,+21,+21	822	*CILP, F11, FAM81A, FLT3, GATA1, HSF5, KCNH7, MATN3, PTK2*	
9	M	56,XY,+X,+4,+4,+8,+14,+17,+18,+21c,+21,+21	1,734	*APOL3, CAMLG, CREBBP, CSMD3, CTD-2128A3.2, DNTT, FN1, IFT122, KRAS, LPHN3, LRP3, LUC7L, M1AP, MYB, PRICKLE2, RP11-763F8.1, SGOL1, SMARCD1, THSD4, TNPO2, ZAR1L*	

[a]Based on G-banding and SNP array analysis on bulk DNA

Abbreviations: F, female; M, male; UPID, uniparental isodisomy

Supplementary Table 2. Subclonality involving whole chromosomes in 577 cases of high hyperdiploid pediatric acute lymphoblastic leukemia

Chromosome	No of cases with subclonality (%)	Type of subclonality (No of cases)										*P* disomy vs UPID[a]
		Trisomy/ disomy	Trisomy/ UPID	Trisomy/ tetrasomy 2:2	UPID/ disomy	Disomy/ tetrasomy 2:2	Disomy/ monosomy	Tetrasomy 2:2/ pentasomy	Trisomy/ tetrasomy 3:1	XXY/XX0	XY/X0	
1	2 (0.35)	1	0	0	1	0	0	0	0	0	0	
2	3 (0.52)	3	0	0	0	0	0	0	0	0	0	0.593
3	5 (0.87)	5	0	0	0	0	0	0	0	0	0	0.263
4	9 (1.6)	7	2	0	0	0	0	0	0	0	0	0.754
5	4 (0.69)	3	1	0	0	0	0	0	0	0	0	1
6	10 (1.7)	9	1	0	0	0	0	0	0	0	0	0.208
7	6 (1.0)	3	3	0	0	0	0	0	0	0	0	0.639
8	26 (4.5)	20	3	2	0	1	0	0	0	0	0	0.0529
9	50 (8.7)	26	22	0	1	0	1	0	0	0	0	0.0969
10	14 (2.4)	9	2	3	0	0	0	0	0	0	0	0.468
11	10 (1.7)	6	2	0	0	1	1	0	0	0	0	0.936
12	5 (0.87)	4	1	0	0	0	0	0	0	0	0	0.922
13	5 (0.87)	2	0	0	0	0	3	0	0	0	0	-
14	8 (1.4)	2	1	5	0	0	0	0	0	0	0	1
15	5 (0.87)	2	3	0	0	0	0	0	0	0	0	0.420
16	7 (1.2)	5	2	0	0	0	0	0	0	0	0	1
17	10 (1.7)	6	3	1	0	0	0	0	0	0	0	1
18	7 (1.2)	0	0	7	0	0	0	0	0	0	0	-
19	0 (0)	0	0	0	0	0	0	0	0	0	0	-
20	1 (0.17)	0	0	0	0	0	1	0	0	0	0	-
21	21 (3.6)	1	0	9	0	0	0	5	6	0	0	-
22	3 (0.52)	2	0	0	0	0	0	0	1	0	0	-
X females	13 (5.1)	1	10	1	0	0	0	1	1	0	0	**0.000260**
X males	0 (0)	0	0	0	0	0	0	0	0	0	0	-
Y	4 (1.2)	0	0	0	0	0	0	0	0	3	1	-
Sum	-	117	56	28	1	2	6	6	8	3	1	0.872

[a]Two-sided exact binomial test for subclonality trisomy/disomy (expected 2/3) vs trisomy/UPID (expected 1/3)

Abbreviations: UPID, uniparental isodisomy

Supplementary Table 3. RMSE values between simulation result and 577 cases of high hyperdiploid pediatric acute lymphoblastic leukemia

UPID frequency at the end of simulation	Modal chromosome number	Comparison group	Trisomy/tetrasomy at different chromosome modal numbers[a]	Trisomy/tetrasomy distribution similarity
2.5%	51-61	Diploid/sequential vs cases	0.81 / 0.37	0.10 / 0.11
		Tetraploid/sequential vs cases	NA / NA[b]	NA / NA[b]
		Diploid/tripolar 3 groups vs cases	0.35 / 0.27	0.087 / 0.035
		Diploid/tripolar 4 groups vs cases	**0.27 / 0.19**	**0.075 / 0.035**
	62-67	Tetraploid/sequential 3 groups vs cases	0.25 / 0.83	0.341 / 0.163
		Tetraploid/sequential 4 groups vs cases	**0.23 / 0.776**	0.339 / 0.166
		Diploid/tripolar 3 groups vs cases	0.61 / 0.94	0.325 / 0.138
		Diploid/tripolar 4 groups vs cases	1.10 / 0.779	**0.271 / 0.128**
5.0%	62-67	Tetraploid/sequential 3 groups vs cases	0.22 / 0.81	0.34 / 0.162
		Tetraploid/sequential 4 groups vs cases	**0.21 / 0.79**	0.339 / 0.166
		Diploid/tripolar 3 groups vs cases	0.47 / 0.81	0.324 / 0.136
		Diploid/tripolar 4 groups vs cases	1.7 / 0.71	**0.275 / 0.13**

[a]The smallest RMSE values are shown in bold

[b]Lack of required virtual cell numbers for RMSE value calculation

Abbreviations: NA, not applicable; RMSE, root mean squared error; UPID, uniparental isodisomy

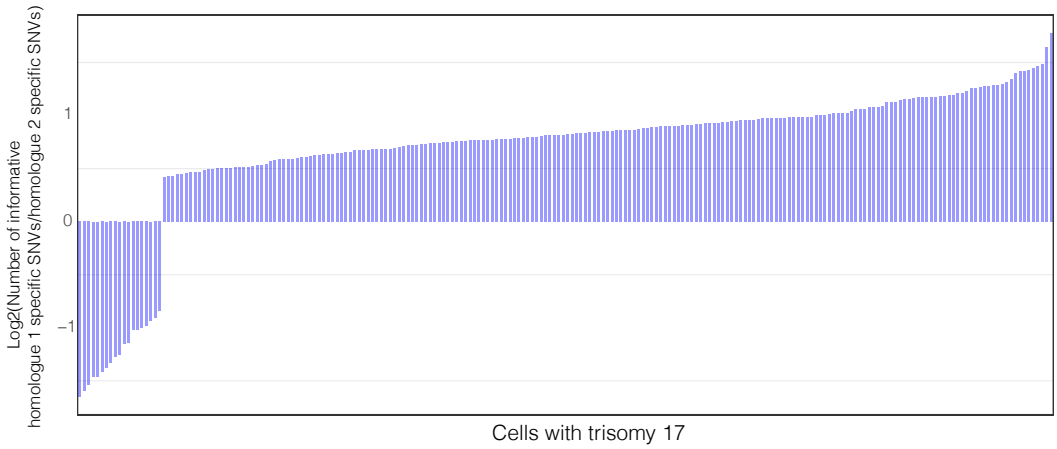
Supplementary Table 4. Frequency and pattern of structural rearrangements, targeted deletions, and mutations in high hyperdiploid childhood acute lymphoblastic leukemia

Somatic aberration	Total no. (%)	No. of clonal (%)	No of subclonal (%)	Occurred before chromosomal event, no (%)	Occurred after chromosomal event, no (%)
dup(1q)	140/577 (24)	82 (59)	58 (41)	0/8 (0)	8/8 (100)
del(6q)	25/577 (4.3)	15 (60)	10 (40)	0/22 (0)	22/22 (100)
i(7q)	15/577 (2.6)	14 (93)	1 (6.7)	N.I.	N.I.
Partial 17q gain	54/577 (9.4)	38 (70)	16 (30)	N.I.	N.I.
IKZF1 del[a]	23/427 (5.4)	20 (87)	3 (13)	N.I.	N.I.
CDKN2A del	53/427 (12)	42 (79)	11 (21)	1/16 (6.3)	15/16 (94)
PAX5 del	19/427 (4.4)	14 (74)	5 (26)	0/1 (0)	1/1 (100)
ETV6 del	45/427 (11)	27 (60)	18 (40)	1/9 (11)[b]	8/9 (89)
CREBBP del	10/427 (2.3)	8 (80)	2 (20)	0/1 (0)	1/1 (100)
TCF3 del	10/427 (2.3)	9 (90)	1 (10)	N.I.	N.I.
CREBBP mut	21/218 (9.6)	12 (57)	9 (43)	N.I.	N.I.
FLT3 mut	32/218 (15)	15 (47)	17 (53)	0/2 (0)	2/2 (100)
IKZF1 mut	8/218 (3.7)	3 (38)	5 (62)	1/1 (100)	0/1 (0)
KRAS mut	61/218 (28)	31 (51)	30 (49)	0/4 (0)	4/4 (100)
NRAS mut	50/218 (23)	21 (42)	29 (58)	0/1 (0)	1/1 (100)
PTPN11 mut	14/218 (6.4)	7 (50)	7 (50)	0/1 (0)	1/1 (100)
Other mut	155	102 (66)	53 (34)	4/56 (7.1)	52/56 (93)

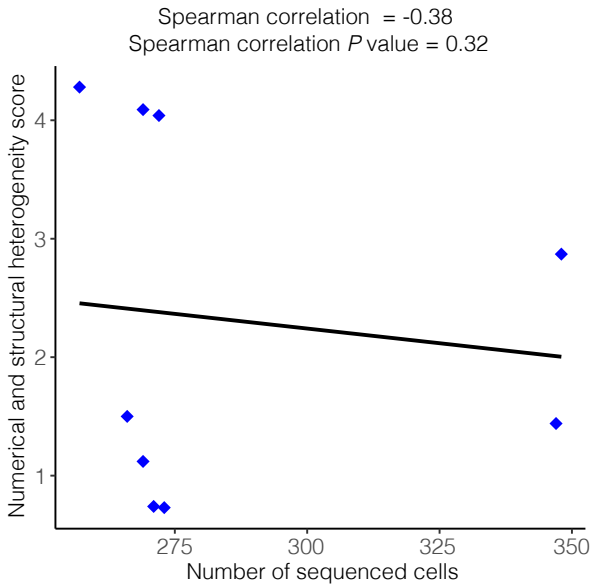
Abbreviations: N.I., no informative cases; del, deletion; mut, mutation

[a]Includes cases with loss of *IKZF1* through isochromosome 7q and monosomy 7

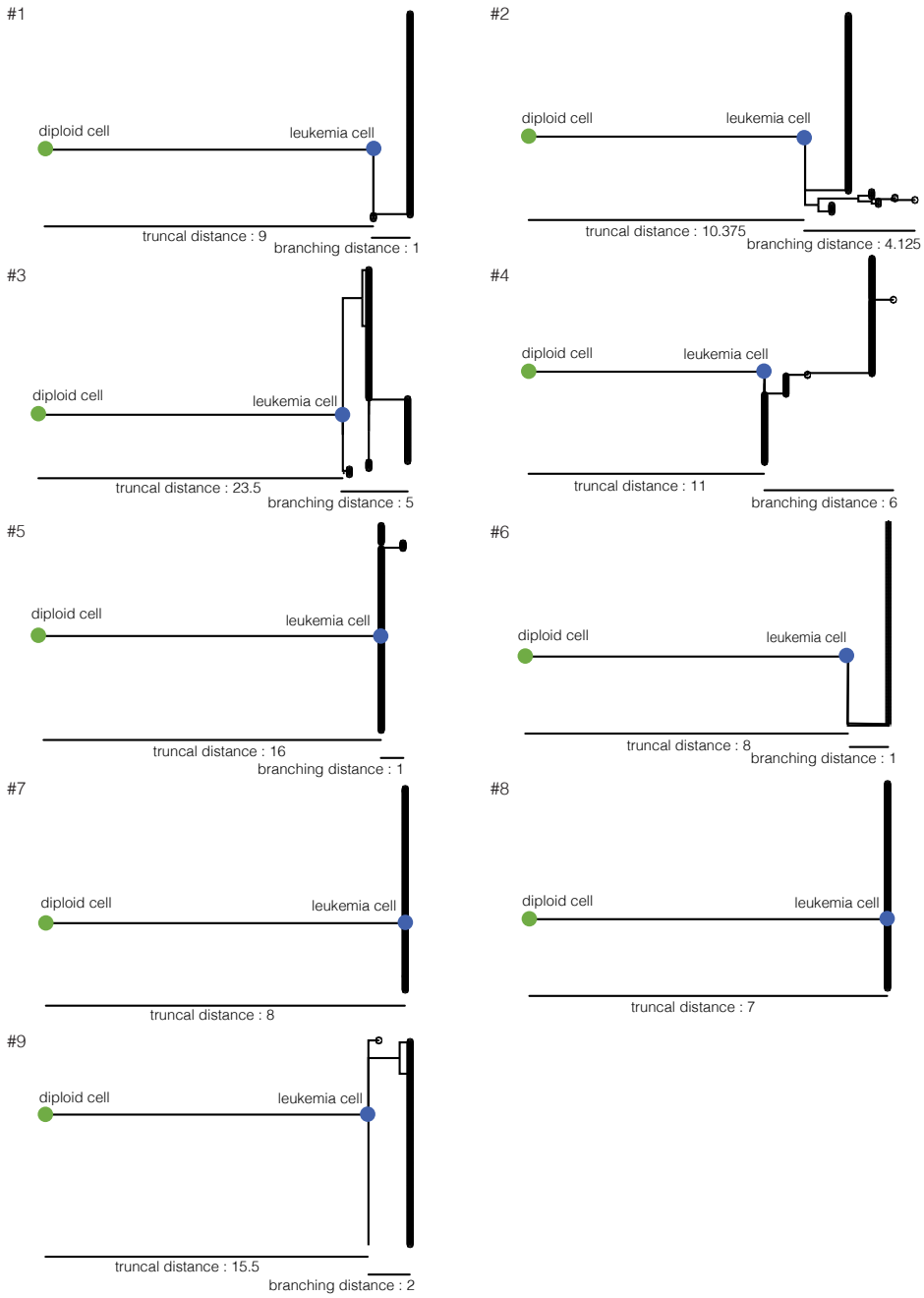
[b]Constitutional *ETV6* deletion



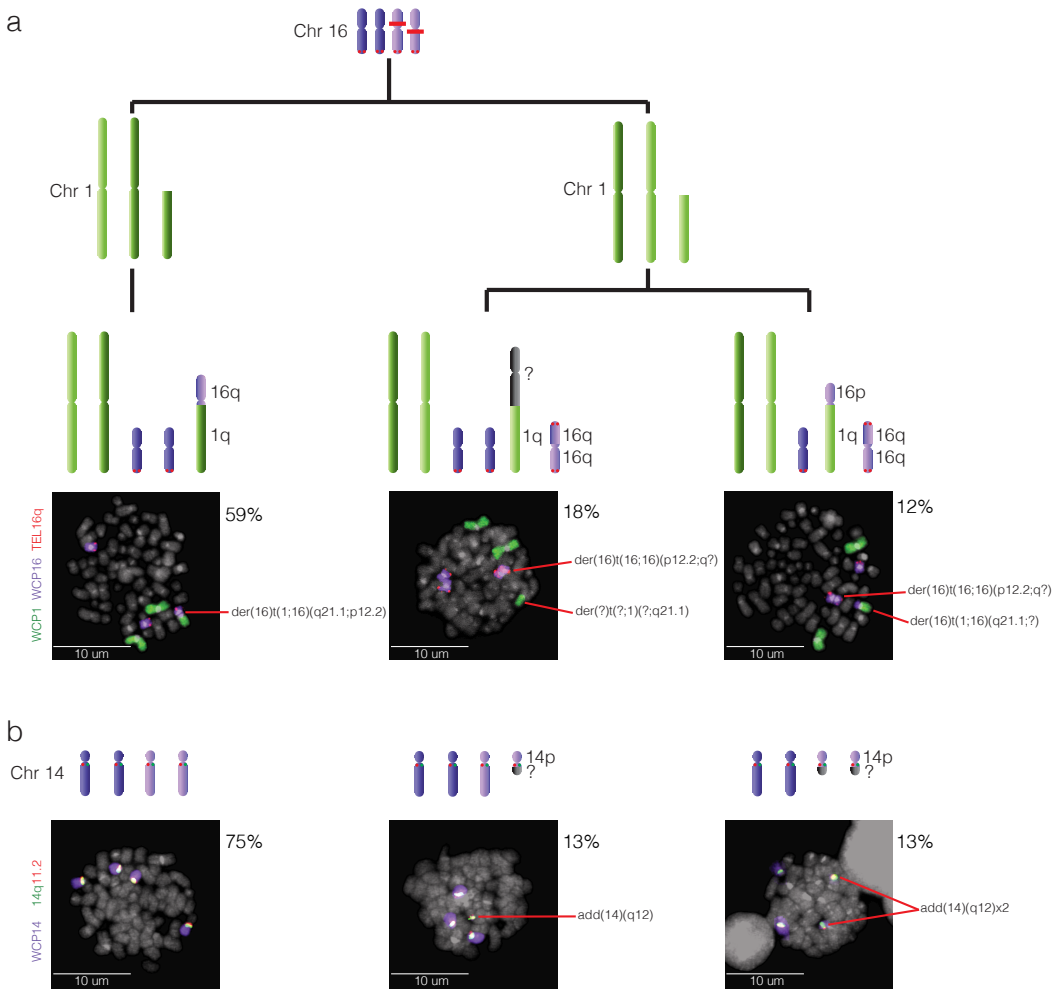
Supplementary Figure 1. Trisomy 17 homologue-specific analysis of scWGS data in case 9, showing gain of homologue 1 in 19 cells and of homologue 2 in 201 cells. Abbreviations: SNV, single nucleotide variant. Source data are provided as a Source Data file.



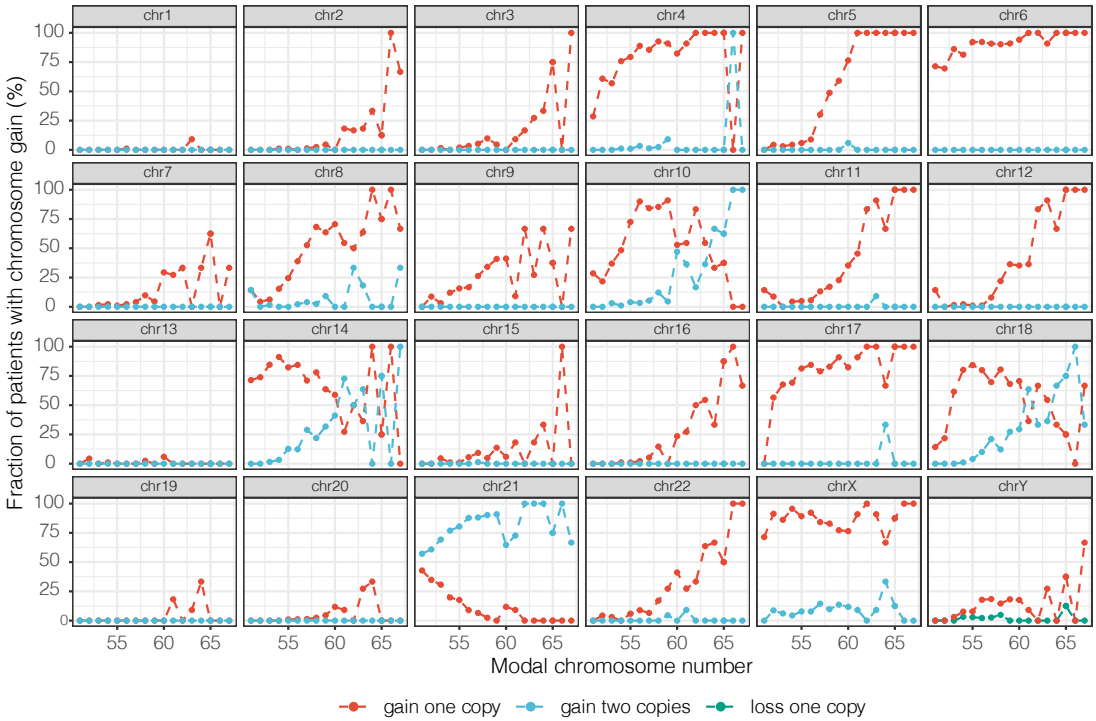
Supplementary Figure 2. Spearman's rank correlation between the number of sequenced cells and the numerical and structural heterogeneity scores. No correlation was seen (P = 0.32; Spearman's correlation two-sided test). Source data are provided as a Source Data file.



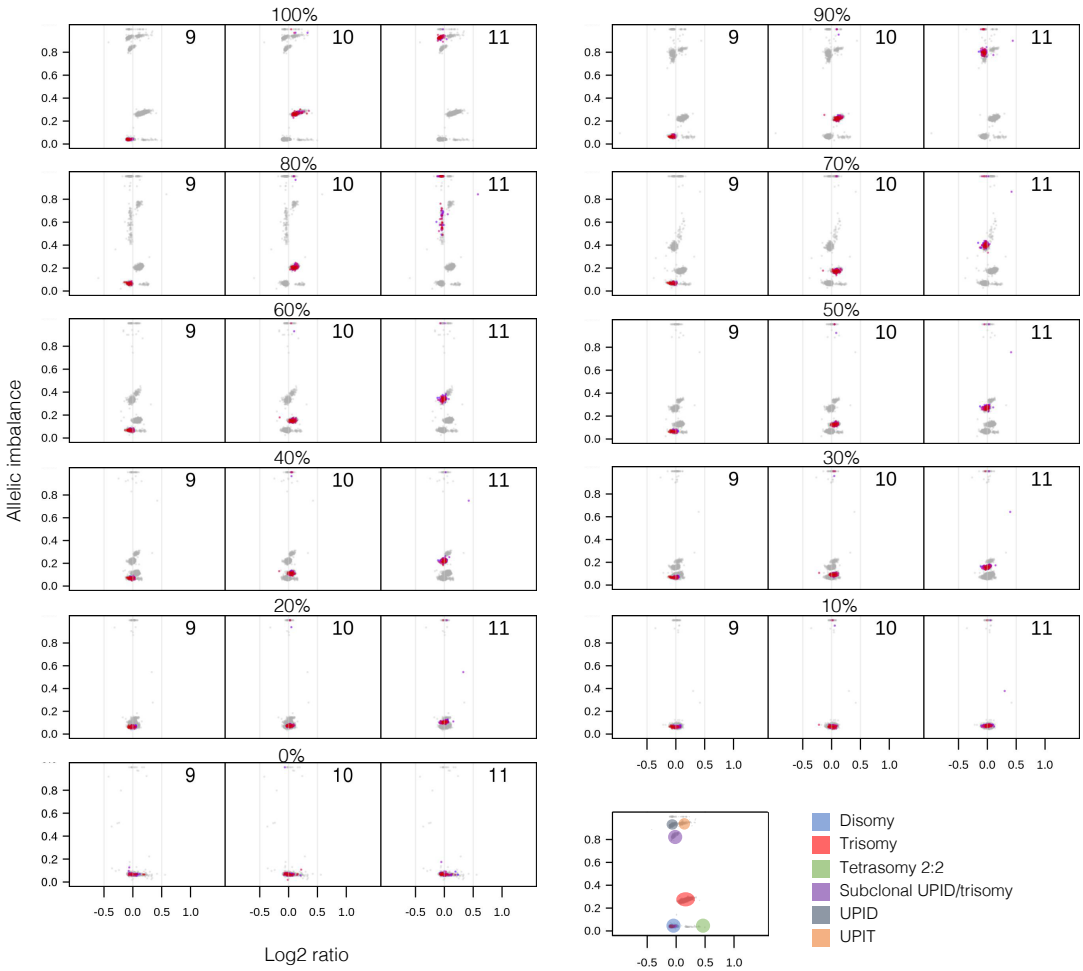
Supplementary Figure 3. Minimum evolution trees of single cell copy number data for nine primary high hyperdiploid childhood acute lymphoblastic leukemia cases. Trees are rooted by simulated normal diploid cells and only the copy number events that were observed in at least two single cells were used. All cases showed relatively long truncal and relatively short branching distances, agreeing with a punctuated evolution model for copy number changes in these malignancies. Source data are provided as a Source Data file.



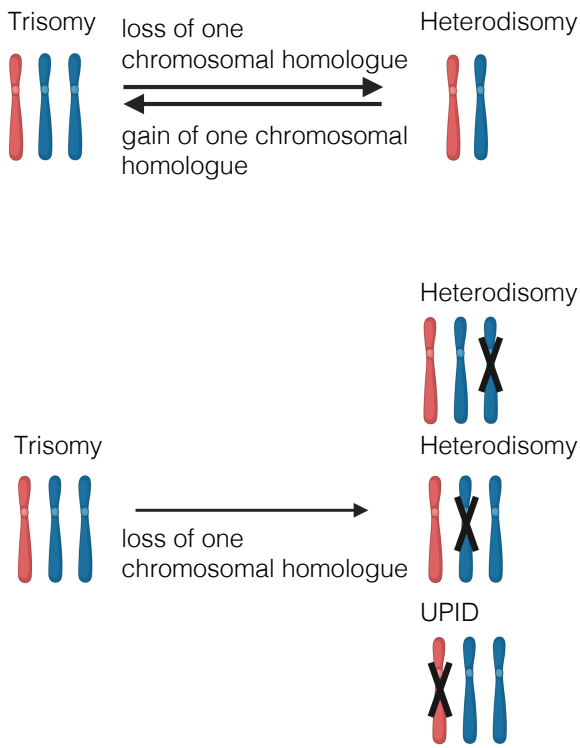
Supplementary Figure 4. Complex patterns of copy number changes involving structural rearrangements in high hyperdiploid ALL cases as elucidated by fluorescence in situ hybridization; in the schematic drawings, different shades of each color represent different homologues. Twenty to thirty-nine metaphase cells were independently analyzed for each case, with percentages of cells corresponding to each subclone shown in the figure. A) Multiple rearrangements involving chromosome 16 in case 3. A traumatic event likely caused breakage in two different chromosomes 16 (at 16p12.2 and 16q) in a cell with tetrasomy 16 as well as two separate gains of 1q, resulting in three subclones with different chromosome 1 and 16 rearrangements. B) Chromosome 14 rearrangements in case 4. Two events of loss of 14q resulted in three subclones with a tetrasomy 14, three normal copies of chromosome 14, and an add(14)(q12), and two normal copies of chromosome 14 and two identical add(14)(q12), respectively. Abbreviations: TEL16q, telomere 16q; WCP, whole chromosome paint. Schematics created with BioRender.com.



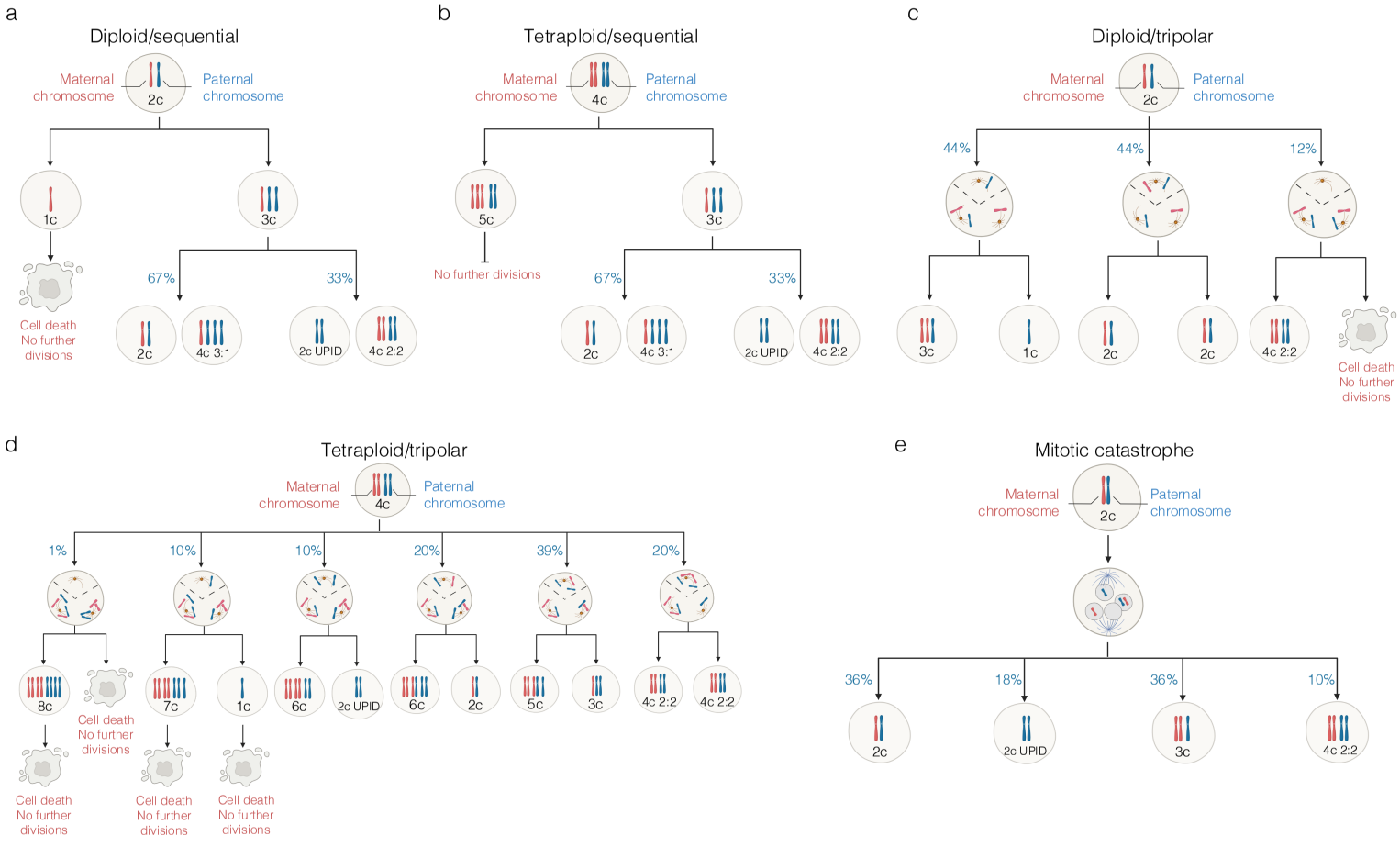
Supplementary Figure 5. Gain of each chromosome per modal number. Source data are provided as a Source Data file.



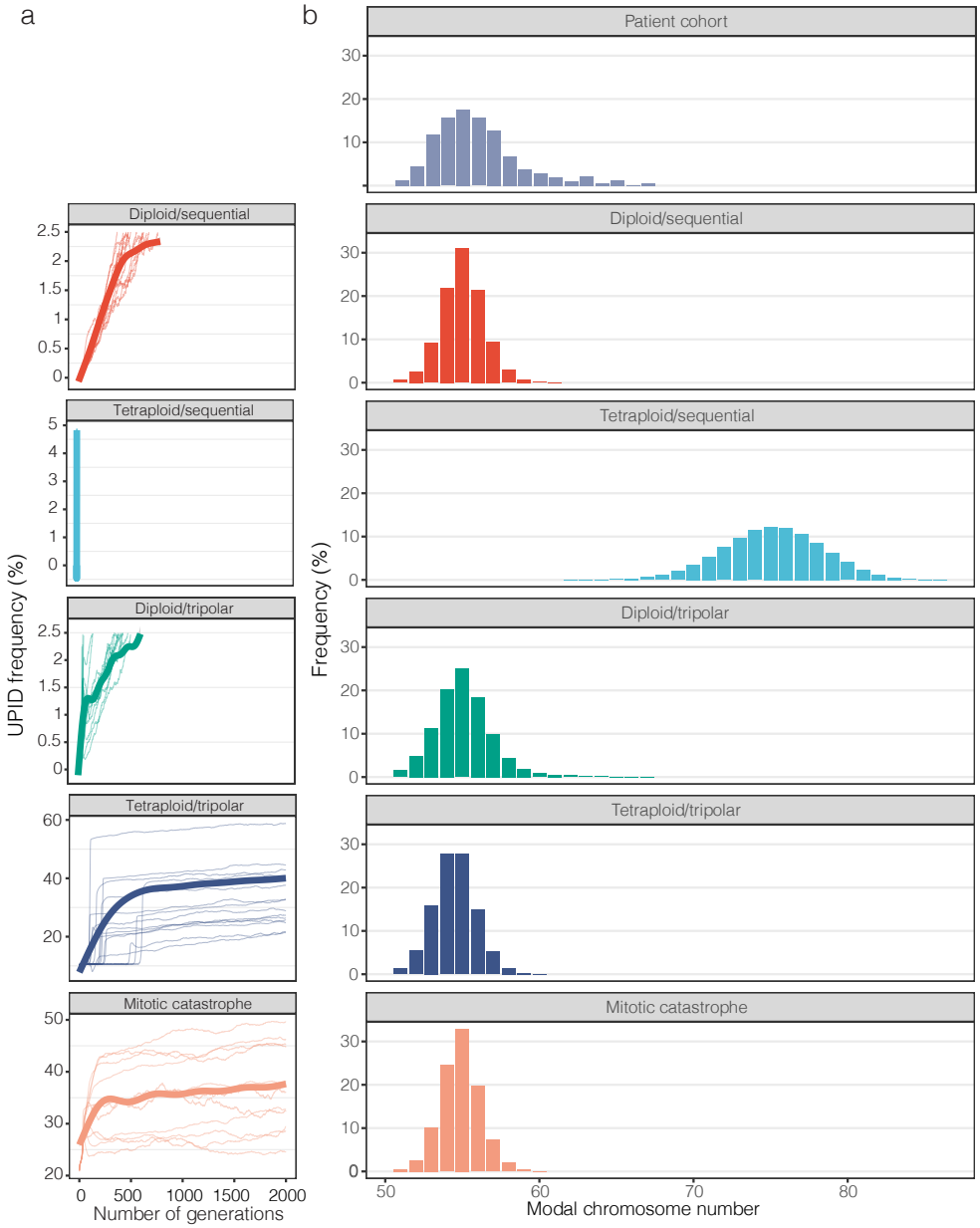
Supplementary Figure 6. SNP array analysis of a dilution series with 0-100% leukemic cells. The series was made using a leukemic sample with close to 100% leukemic blast cells and its corresponding remission sample, with 0% blast cells. Results are from TAPS[1] and graphs show the allelic imbalance versus the log2 ratio. Signals from chromosomes 9 (disomic), 10 (trisomic) and 11 (uniparental isodisomy; UPID) are shown in red. The legend shows where the signal from chromosomes with a specific copy number clusters in the pure leukemic sample. UPID11 can be detected at 20% and trisomy 10 at 30% leukemic cells, corresponding to these clone sizes. The dilution series SNP array data have previously been published in Paulsson et al.[2] Abbreviations: UPID, uniparental isodisomy; UPIT, uniparental isotrisomy.



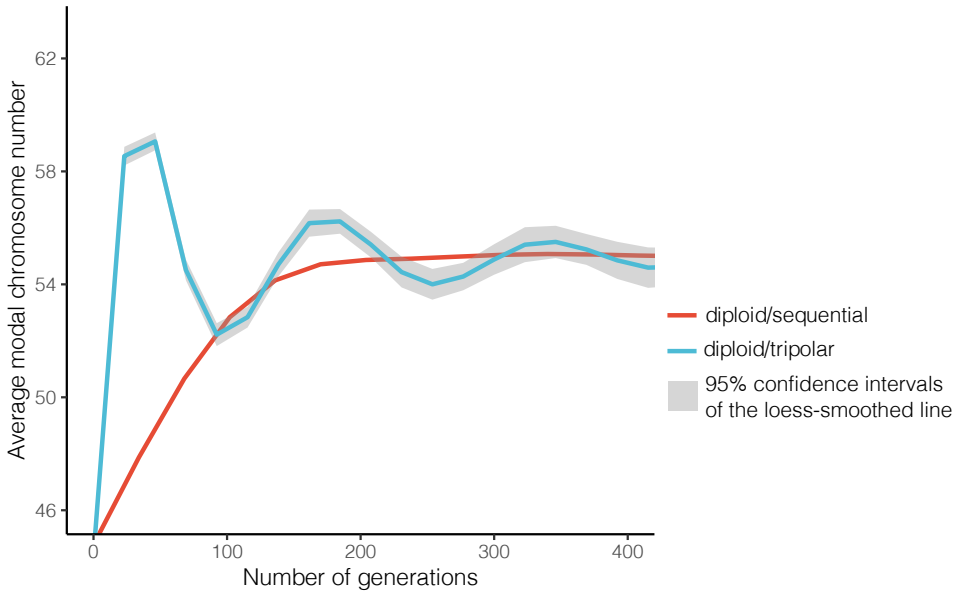
Supplementary Figure 7. Schematics of how subclonal populations may arise. The top panel shows how heterodisomy/trisomy could arise either by an initial disomy becoming a trisomy or vice versa, i.e. the direction of the change cannot be inferred. The bottom panel shows that uniparental isodisomy(UPID)/trisomy can only arise from initial trisomy by loss of one chromosomal homologue. Here, 2/3 cells become heterodisomies and 1/3 cells becomes a UPID. Abbreviations: UPID, uniparental isodisomy. Created with BioRender.com.



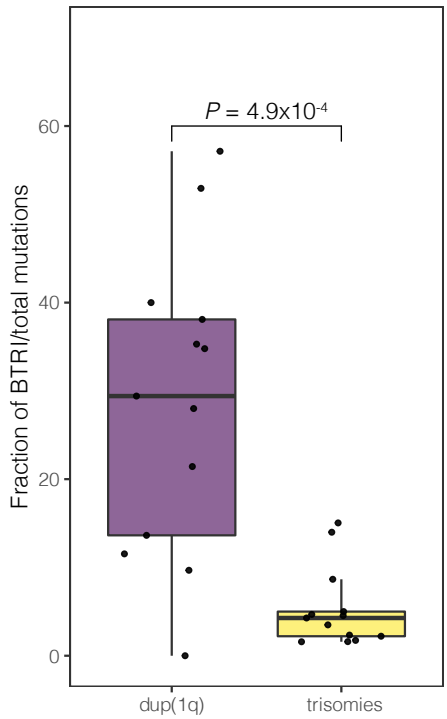
Supplementary Figure 8. Schematics of how five different models for hyperdiploidy development will lead to different patterns of tetrasomies 2:2 and 3:1 and uniparental isodisomies. Abbreviations: c, copy/copies; UPID, uniparental isodisomy. Created with BioRender.com.



Supplementary Figure 9. Uniparental isodisomy (UPID) frequencies and distribution of modal chromosome numbers (MCN). A. UPID frequencies for groups weak pos and neg chromosomes over 2,000 generations in each simulation model. Thick lines show the generalized additive model regression of UPID frequencies of all simulations and thin lines show the UPID frequencies of 15 randomly selected simulations. The diploid/tripolar and diploid/sequential models reached 2.5% UPIDs after 50-800 generations, consistent with the patient data. B. Distribution of MCN in the patient cohort and in the five simulation models. All simulation models but the tetraploid/sequential model resulted in a similar MCN distribution to the patient cohort. The tetraploid/sequential model resulted in very few cells that had MCN 51-67; most of the virtual cells showed MCN around 75. Abbreviations: UPID, uniparental isodisomy. Source data are provided as a Source Data file.



Supplementary Figure 10. Models of chromosome copy number evolution during high hyperdiploidy development. The LOESS-smoothed lines show the correlation between the average modal chromosome number (Y axis) and the number of simulated generations (X-axis) of 100 randomly sampled simulation results from the diploid/sequential model and diploid/tripolar model. The gray ribbon shows the 95% confidence intervals of the loess-smoothed line. An initial burst of whole chromosome gain events was observed in the diploid/tripolar model. These events were followed by a period of transient instability and stable expansions during the high hyperdiploidy development, in line with the punctuated copy number evolution model. In the diploid/sequential model, chromosomes were acquired sequentially throughout high hyperdiploidy development, indicating gradual copy number evolution. Abbreviations: LOESS, locally weighted scatterplot smoothing. Source data are provided as a Source Data file.



Supplementary Figure 11. Fraction of BTRI mutations (occurring before the shift to three copies) in dup(1q) compared with trisomies in the same cases. The fraction of BTRI mutations is significantly higher in dup(1q) (P = 4.9x10^{-4}; Mann-Whitney two-sided test), indicating that this rearrangement is formed subsequently to the trisomies. The centre of the boxplot is the median and lower/upper hinges correspond to the first/third quartiles; whiskers are 1.5 times the interquartile range and data beyond this range are plotted as individual points. Source data are provided as a Source Data file.

References

1. Rasmussen, M., Sundström, M., Göransson Kultima, H., Botling, J., Micke, P., Birgisson, H., Glimelius, B. & Isaksson, A. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol* **12**, R108 (2011).
2. Paulsson, K., Lilljebjörn, H., Biloglav, A., Olsson, L., Rissler, M., Castor, A., Barbany, G., Fogelstrand, L., Nordgren, A., Sjögren, H., Fioretos, T. & Johansson, B. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nat Genet* **47**, 672-676 (2015).

" Still round the corner there may wait
A new road or a secret gate,
And though we pass them by today,
Tomorrow we may come this way
And take the hidden paths that run
Towards the Moon or to the Sun."

...

"Home behind, the world ahead,
And there are many paths to tread
Through shadows to the edge of night,
Until the stars are all alight
Then world behind and home ahead,
We'll wander back to home and bed."

The Fellowship of the Ring, Book 1, Chapter 3,
J. R. R. Tolkien

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Laboratory Medicine

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2024:43
ISBN 978-91-8021-536-7
ISSN 1652-8220