The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia.

Paulsson, Kajsa; Lilljebjörn, Henrik; Biloglav, Andrea; Olsson, Linda; Rissler, Marianne; Castor, Anders; Barbany, Gisela; Fogelstrand, Linda; Nordgren, Ann; Sjögren, Helene; Fioretos, Thoas; Johansson, Bertil

Published in:
Nature Genetics

DOI:
10.1038/ng.3301

2015

The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia

Kajsa Paulsson[1], Henrik Lilljebjörn[1], Andrea Biloglav[1], Linda Olsson[1], Marianne Rissler[1], Anders Castor[2], Gisela Barbany[3], Linda Fogelstrand[4,5], Ann Nordgren[3], Helene Sjögren[4], Thoas Fioretos[1,6] & Bertil Johansson[1,6]

[1]Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, Lund, Sweden. [2]Department of Pediatrics, Skåne University Hospital, Lund University, Lund, Sweden. [3]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden. [4]Department of Clinical Chemistry and Transfusion Medicine, Institute of Biomedicine, University of Gothenburg, Göteborg, Sweden. [5]Laboratory of Clinical Chemistry, Sahlgrenska University Hospital, Göteborg, Sweden. [6]Department of Clinical Genetics, University and Regional Laboratories Region Skåne, Lund, Sweden.

Correspondence should be addressed to K.P. (kajsa.paulsson@med.lu.se).

High hyperdiploid (51-67 chromosomes) acute lymphoblastic leukemia (ALL) is one of the most common childhood malignancies, comprising 30% of all pediatric B-cell precursor ALL. Its characteristic genetic feature is the nonrandom gain of chromosomes X, 4, 6, 10, 14, 17, 18, and 21, with individual trisomies/tetrasomies being seen in over 75% of cases, but the pathogenesis remains poorly understood[1]. We performed whole genome sequencing (WGS) (n=16) and/or whole exome sequencing (WES) (n=39) of diagnostic and remissions samples from 51 cases of high hyperdiploid ALL to define further the genomic landscape of this malignancy. The majority of cases showed involvement of the RTK-RAS pathway and of histone modifiers. No recurrent fusion gene-forming rearrangement was found and an analysis of mutations on trisomic chromosomes indicated that the chromosomal gains were early events, strengthening the notion that the high hyperdiploid pattern as such is the main driver event in this common pediatric malignancy.

The mean coverage of the WGS and WES analyses of the 51 cases was 103x and 123x, respectively (**Supplementary Tables 1-3**). The median chromosome number, based on copy number analysis of data obtained from sequence coverage, was 55-56 and 8/16 cases (50%) investigated by WGS harbored 1-4 whole chromosome uniparental disomies (UPDs) (**Fig.1** and **Supplementary Fig.1**); this agrees well with previous data on the chromosomal content of high hyperdiploid childhood ALL[1]. Subclonal gains of one or two whole chromosomes were seen in four cases, including trisomy/tetrasomy 18 (2:2) in case 10, trisomy/tetrasomy 21 (3:1) in case 11, disomy/trisomy 4 in case 13, and UPD/trisomy 16 and disomy/monosomy 13 in case 14. Furthermore, subclonality was detected for the majority of structural aberrations leading to copy number changes, including 4/4 cases with dup(1q), 1/1 idic(7)(p11), and 1/1 idic(17)(p11.2).

The mean and median number of somatic mutations, including single nucleotide variants (SNVs), insertions and deletions (indels), and substitutions involving more than one nucleotide in cases analyzed with WGS were 1,292 and 801 (range 123-7,862), respectively. The corresponding numbers for mutations in coding regions, excluding synonymous variants, were 7.5 and 6 (range 0-55), respectively, for all 51 cases (**Supplementary Tables 4** and **5**). There was no correlation between the total number of mutations or of coding mutations and modal chromosome number, gender, or white blood cell count. However, the total number of mutations was significantly higher in older patients (mean 425 [range 123-703] in those aged 1-3 years vs. mean 1,812 [range 678-7,862] in those aged >3 years; P=0.029), as was the number of coding mutations (mean 3.6 [range 0-9] vs. mean 11 [range 1-55]; P<0.0001]. This most likely reflects the higher number of cell divisions that the leukemia-initiating cell has gone through in older patients; similar increases in mutation frequency with age have previously been observed in pediatric neuroblastoma and T-ALL[2,3]. The four cases that relapsed had on average 5.5 mutations in coding regions, to compare with 7.7 in those that did not relapse (**Supplementary Table 1**).

Analysis of the genomic context of SNVs in coding regions for all cases except no. 12 (discussed below) revealed that the majority (274/415; 66%) were C>T transitions, most of which involved CpG dinucleotides (188/274; 69%; **Fig. 2**). This is the most common mutational signature in the human genome and is a sign of endogenous mutagenic mechanisms, i.e., deamination of 5-methylcytosine to thymidine[4,5]. In contrast to what was recently reported for *ETV6*/*RUNX1*-positive childhood ALL[6], we did not detect a signature corresponding to transitions and transversions at cytosines in TpC dinucleotides, which would have indicated an involvement of the APOBEC family of enzymes. Thus, in spite of the clinical similarities between high hyperdiploid and *ETV6*/*RUNX1*-positive childhood ALL, such as favorable outcome, prenatal origin, and age peak, their mutational signatures differ,

indicating different etiologic and pathogenetic mechanisms. Case 12 harbored the highest number of both total (n=7,862) and non-synonymous coding (n=55) mutations in the investigated cohort. Notably, it also displayed a different mutational pattern, with 77/89 (87%) of mutations in coding regions (silent and non-silent) being C>T transitions; of those, only 36% were at CpGs (**Supplementary Fig. 2**). An investigation of the previously identified regions in the *IKZF1*, *ARID5B*, and *CEBPE* genes that have been linked to increased risk of high hyperdiploid childhood ALL did not reveal any novel constitutional variants predisposing to leukemia development in any of the 51 cases (**Supplementary Note** and **Supplementary Tables 6-10**).

A total of 75 somatic structural joints resulting from translocations, deletions, duplications, and complex rearrangements (**Supplementary Table 11**) were detected by WGS. Multiple deletions and duplications involving single genes were seen, including known targets such as *ADD3*, *ETV6*, *IKZF1*, and *PAX5*, as well as not previously implicated genes (**Supplementary Table 11**). There was no evidence of chromothripsis in any of the cases[7]. Two cases had reciprocal translocations involving the *IGK* locus at 2p11.2 – t(2;8)(p11.2;q21.13) in case 15 and t(2;19)(p11.2;q13.32) in case 2. RNA-sequencing (RNA-seq) of the breakpoint region on chromosome 8 in case 15 revealed a high expression of *TPD52*, residing ~400 kb from the breakpoint, compared with 35 other high hyperdiploid childhood ALL (**Supplementary Fig. 3**). *TPD52*, which is frequently upregulated in neoplasia, encodes a protein that may play a role in cytokinesis by supporting membrane trafficking events[8], but is rarely expressed in high hyperdiploid childhood ALL[9]. No RNA was available from case 2, precluding identification of an upregulated gene at 19q13. Case 16 had a 40 kb deletion in 16q22.1 that resulted in an in-frame fusion between exon 5 of *CTCF* and exon 2 of *PARD6A*; this was validated by RNA-seq and RT-PCR (**Supplementary Table 11** and **Supplementary Fig. 4**). PARD6A regulates centrosomal protein recruitment and

cytokinesis[10], whereas CTCF is a regulator of higher-order chromatin structure[11]. The 16q22 deletion in case 16 leads to complete loss of zinc fingers 5-11 in *CTCF*. Zinc fingers 4-7 bind to the core motif of CTCF genomic target sites[12], and it may thus be surmised that the *CTCF-PARD6A* fusion severely impacts the normal function of CTCF. Case 16 also harbored an interstitial deletion on Xp leading to a *P2RY8*/*CRLF2* fusion, which has been previously reported in childhood ALL[13]. None of the other structural rearrangements identified by WGS (**Supplementary Table 11**) could be shown to result in expressed fusion genes.

The number of mutations in coding regions correlated with the total number of mutations (Pearson correlation coefficient=0.975; $P<0.0001$), suggesting a random distribution of mutations throughout the genomes and indicating a large number of passenger mutations. A total of 399 SNVs, indels, and substitutions in coding regions were detected in the 51 cases; for a subset of these, Sanger sequencing was performed for validation (see Online Methods). Nine genes were recurrently mutated and also either mutated more frequently than expected by chance according to MutSigCV[14] or targeted by structural events. These included six well-known leukemia-associated genes: *KRAS* (13/51 cases; 25%), *FLT3* (6/51 cases; 12%), *CREBBP* (5/51 cases; 9.8%), *NRAS* (5/51 cases; 9.8%), *WHSC1* (3/51 cases; 5.9%), and *PTPN11* (3/51 cases; 5.9%), of which *CREBBP* and *WHSC1* were also targeted by small deletions in one case each (**Supplementary Tables 4, 5,** and **11**). The mutated alleles of the above-mentioned genes were all expressed, except one with a frameshift insertion in *CREBBP*, in cases with available RNA-seq data (**Supplementary Tables 4** and **5**). In total, mutations in the RTK-RAS signaling pathway, including the *FLT3*, *NRAS*, *KRAS*, and *PTPN11* genes, were seen in 26/51 cases (51%) (**Fig. 3** and **Supplementary Tables 4** and **5**); a frequency notably higher than the ~30% previously suggested[15-18]. In addition to *KRAS* mutations in the known hotspot regions, there were three p.Lys117Asn and two p.Ala146Thr mutations in *KRAS* (**Supplementary Tables 4** and **5**); both of these have been reported in

colorectal cancer and shown to lead to downstream phosphorylation of ERK[19]. The p.K117N has also been identified in a single case of high hyperdiploid childhood ALL at relapse[20] and the p.A146T has been shown to occur in other types of hematologic malignancies[21,22], including two cases of non-hyperdiploid childhood ALL[23]. Thus, *KRAS* codons 117 and 146 may be novel recurrent mutational hotspots in high hyperdiploid ALL. Other genes frequently targeted by mutations were those encoding histone modifiers, including *CREBBP* in six cases (12%), *WHSC1* in four (7.8%), and *SUV420H1*, *SETD2*, and *EZH2* in one case each (2.0%; **Fig. 3**); in total, 12/51 (24%) cases harbored mutations or deletions in one of these genes. *CREBBP* has been reported to be mutated in a high proportion of relapsing high hyperdiploid childhood ALL[24]; however, none of our six cases with a *CREBBP* mutation/deletion has relapsed (**Supplementary Table 1**). The mutations in the remaining three genes that were identified as possible drivers – *DPP6* (4/51 cases; 7.8%), *MLLT3* (2/51 cases; 3.9%), and *PRB2* (2/51 cases; 3.9%) – were not damaging according to Provean and SIFT, have not been previously reported to be mutated in ALL, and were not expressed according to RNA-seq (**Supplementary Tables 4** and **5**); their significance is hence unclear.

To investigate how early in the leukemogenic process the chromosomal gains occurred, we compared the number of somatic mutations present in 1/3 copies of trisomic chromosomes with the number of mutations present in 2/3 copies in each case subjected to WGS (**Fig. 4** and **Supplementary Figs. 5** and **6**). Whereas mutations present in 1/3 homologues could have occurred either before or after the chromosomal gain, mutations present in 2/3 homologues must have preceded the formation of the trisomy (**Supplementary Fig. 5**). Based on these data, we calculated the number of mutations occurring before the trisomies (BTRI mutations) and after the trisomies (ATRI mutations). For 14/16 (88%) cases, ATRI was much (3-33x) higher than BTRI (**Supplementary Table 12**), indicating either a longer time period between trisomy occurrence and diagnosis than between the initial hematopoietic stem cell and the

trisomy occurrence, or an increased mutation rate after trisomy occurrence. Similar results were obtained when we looked at UPDs in four of these 14 cases, identifying a high number of heterozygous mutations; these must have occurred after the UPD. It has been shown that the chromosomal gains in high hyperdiploid ALL, at least sometimes, arise prenatally[25-28]. Hence, we deem it likely that the high frequency of mutations occurring after compared with before the trisomy indicates a long latency period after the high hyperdiploid pattern was established in most patients. Two cases displayed an equal or lower number of ATRI mutations compared with BTRI mutations, indicating a later occurrence of the trisomies (**Supplementary Table 12**). These were the two oldest patients in the cohort – both were thirteen years old – suggesting that the etiology of high hyperdiploid childhood ALL may differ in older patients.

To ascertain whether high hyperdiploidy is associated with gene dosage effects, we investigated the association between copy number changes based on SNP array analysis and gene expression in a cohort of 29 high hyperdiploid cases with available RNA-seq data. There was a clear correlation between gene copy number and expression levels (**Supplementary Fig. 7**), agreeing well with previous studies reporting an increased expression of genes on the gained chromosomes in high hyperdiploid childhood ALL[29-31] and indicating that the chromosomal gains result in gene dosage effects.

Considering the absence of a recurrent fusion gene or common mutation and the present evidence that the chromosomal gains are early events, we conclude that the initiating pathogenetic step in the leukemogenesis of high hyperdiploid childhood ALL is the high hyperdiploid pattern in itself. It is noteworthy that several cases in our study harbored changes involving centrosome-related proteins, such as mutations of *CEP290*, *NRK*, and *AKAP9* and deletion of *CEP76* and *CEP192* (**Supplementary Tables 4**, **5**, and **11**), which could have promoted abnormal cell division. However, since none of these genes was recurrently

targeted, their impact remains unclear. The consequence of the chromosomal gains is probably dosage effects (**Supplementary Fig. 7**), but additional genetic aberrations are most likely needed for overt leukemia. Our results suggest that although these secondary hits are heterogeneous in high hyperdiploid childhood ALL, mutations targeting the RTK-RAS pathway and histone modifiers are particularly common. Therefore, these could be attractive targets for novel therapies in high hyperdiploid pediatric ALL.

ACKNOWLEDGMENTS

This study was supported by grants from the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, and the Swedish Research Council.

AUTHOR CONTRIBUTIONS

K.P. performed the whole genome and exome sequencing analyses. H.L., M.R. and T.F. performed the RNA-seq. A.B. and L.O performed validation experiments. A.C. provided clinical data. G.B., L.F., A.N. and H.S. provided samples and clinical data. K.P and B.J. conceived the study and wrote the manuscript, which was reviewed and edited by the other co-authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

DATA ACCESS

Sequence data may be obtained for academic purposes by contacting the corresponding author.

REFERENCES

1. Paulsson, K. & Johansson, B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **48**, 637-660 (2009).
2. Molenaar, J.J. *et al.* Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**, 589-593 (2012).
3. De Keersmaecker, K. *et al.* Exome sequencing identifies mutation in *CNOT3* and ribosomal genes *RPL5* and *RPL10* in T-cell acute lymphoblastic leukemia. *Nat. Genet.* **45**, 186-190 (2013).
4. Biggs, P.J., Warren, W., Venitt, S. & Stratton, M.R. Does a genotoxic carcinogen contribute to human breast cancer? The value of mutational spectra in unravelling the aetiology of cancer. *Mutagenesis* **8**, 275-283 (1993).
5. Collins, A.R. Molecular epidemiology in cancer research. *Mol. Aspects Med.* **19**, 359-432 (1998).
6. 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).
7. Stephens, P.J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
8. Thomas, D.D.H., Frey, C.L., Messenger, S.W., August, B.K. & Groblewski, G.E. A role for tumor protein TPD52 phosphorylation in endo-membrane trafficking during cytokinesis. *Biochem. Biophys. Res. Commun.* **402**, 583-587 (2010).
9. Barbaric, D., Byth, K., Dalla-Pozza, L. & Byrne, J.A. Expression of tumor protein D52-like genes in childhood leukemia at diagnosis: clinical and sample considerations. *Leuk. Res.* **30**, 1355-1363 (2006).

10. Kodani, A., Tonthat, V., Wu, B. & Sütterlin, C. Par6α interacts with the dynactin subunit p150^{Glued} and is a critical regulator of centrosomal protein recruitment. *Mol. Biol. Cell* **21**, 3376-3385 (2010).

11. Merkenschlager, M. & Odom, D.T. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285-1297 (2013).

12. Nakahashi, H. *et al.* A Genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3**, 1678-1689 (2013).

13. Mullighan, C.G. *et al.* Rearrangement of *CRLF2* in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. *Nat. Genet.* **41**, 1243-1246 (2009).

14. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).

15. Paulsson, K. *et al.* Mutations of *FLT3, NRAS, KRAS,* and *PTPN11* are frequent and possibly mutually exclusive in high hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **47**, 26-33 (2008).

16. Armstrong, S.A. *et al.* FLT3 mutations in childhood acute lymphoblastic leukemia. *Blood* **103**, 3544-3546 (2004).

17. Tartaglia, M. *et al.* Genetic evidence for lineage-related and differentiation stage-related contribution of somatic *PTPN11* mutations to leukemogenesis in childhood acute leukemia. *Blood* **104**, 307-313 (2004).

18. Perentesis, J.P. *et al.* RAS oncogene mutations and outcome of therapy for childhood acute lymphoblastic leukemia. *Leukemia* **18**, 685-692 (2004).

19. Janakiraman, M. *et al.* Genomic and biological characterization of exon 4 KRAS mutations in human cancer. *Cancer Res.* **70**, 5901-5911 (2010).

20. Mullighan, C.G. *et al.* *CREBBP* mutations in relapsed acute lymphoblastic leukaemia. *Nature* **471**, 235-239 (2011).

21. Gelsi-Boyer, V. *et al.* Genome profiling of chronic myelomonocytic leukemia: frequent alterations of *RAS* and *RUNX1* genes. *BMC Cancer* **8**, 299 (2008).

22. Tyner, J.W. *et al.* High-throughput sequencing screen reveals novel, transforming *RAS* mutations in myeloid leukemia patients. *Blood* **113**, 1749-1755 (2009).

23. Zhang, J. *et al.* Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* **118**, 3080-3087 (2011).

24. Inthal, A. *et al.* *CREBBP* HAT domain mutations prevail in relapse cases of high hyperdiploid childhood acute lymphoblastic leukemia. *Leukemia* **26**, 1797-1803 (2012).

25. 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**, 1415-1423 (2001).

26. Bateman, C.M. *et al.* Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. *Leukemia* **29**, 58-65 (2015).

27. Maia, A.T. *et al.* Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins. *Leukemia* **17**, 2202-2206 (2003).

28. Maia, A.T. *et al.* Identification of preleukemic precursors of hyperdiploid acute lymphoblastic leukemia in cord blood. *Genes Chromosomes Cancer* **40**, 38-43 (2004).

29. Andersson, A. *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.* **102**, 19069-19074 (2005).

30. Gruszka-Westwood, A.M. *et al.* Comparative expressed sequence hybridization studies of high-hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **41**, 191-202 (2004).

31. Ross, M.E. *et al.* Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **102**, 2951-2959 (2003).

FIGURE LEGENDS

Figure 1 Circos plots showing all somatic genetic events in three representative cases. The outer ring shows all single nucleotide variants (SNV), indels, and substitutions detected in the sample, with mutations affecting coding regions (non-silent only) labeled by gene names. The second ring shows the chromosomal positions. The third ring shows inferred copy number with one copy depicted in purple, two copies in blue, three copies in light red, and four or more copies in dark red. The fourth ring shows the lesser allele frequency in a 100 kb window; loss of heterozygosity is labeled green. The fifth (inner) ring shows small deletions and duplications, with affected genes labeled by gene names. Blue traversing lines correspond to structural rearrangements. (**a**) Case 9 had a total of 123 somatic mutations, with one SNV affecting a coding region, and a structural rearrangement leading to gain of 1q. (**b**) Case 4 had a total of 419 somatic mutations including SNVs targeting the RTK-RAS pathway (*NRAS*) and a histone modifier (*CREBBP*). (**c**) Case 12 had a total of 7,862 somatic mutations, including 55 in coding regions, in addition to microdeletions affecting single genes.

Figure 2 Mutational signature of high hyperdiploid childhood ALL. The plot is done according to Lawrence et al.[14] and shows the type of mutation and the neighboring nucleotides. All mutations in coding regions of 50 high hyperdiploid pediatric ALL cases (all except #12) are included, showing a predominance of C>T transitions in a CpG context.

Figure 3 Overview of mutations and structural events affecting genes involved in the RTK-RAS pathway and histone modifications. Red boxes denote single nucleotide variants; blue boxes denote structural changes. Cases that did not have any mutations in such genes were excluded from the figure.

Figure 4 Mutant allele fraction (MAF) in relation to relative copy number in trisomic chromosomes in case 11. (**a**) MAF for all mutations in relation to relative copy number based on the number of reads for that particular chromosome segment, with trisomic chromosomes having relative copy numbers of ~1.4. Mutations in coding regions, including both non-silent and silent mutations, are shown in red. (**b**) Number of mutations in trisomic chromosomes with a certain MAF. There are two peaks corresponding to mutations present in 1/3 homologues (MAFs of ~0.33) and mutations present in 2/3 copies (MAFs of ~0.67), respectively. Based on these data, the number of mutations occurring before and after the trisomy formation may be inferred. The number of mutations that occurred after the trisomies is much higher than the number of mutations that occurred before the trisomies, indicating that the trisomies arose early in leukemogenesis.

ONLINE METHODS

Patients

The study comprised a total of 51 cases of high hyperdiploid childhood acute lymphoblastic leukemia (ALL) diagnosed and treated at Skåne University Hospital, Lund, Sweden, Sahlgrenska University Hospital, Göteborg, Sweden, or Karolinska Institutet, Stockholm, Sweden (**Supplementary Table 1**). There were 27 boys and 24 girls, and the median age at diagnosis was 4 years (range 1-17, **Supplementary Table 1**). The median bone marrow blast percentage was 88.5% (range 35-98%; **Supplementary Table 1**). All cases were high hyperdiploid as ascertained by standard G-banding, interphase fluorescence in situ hybridization (FISH), or single nucleotide polymorphism (SNP) array analysis, and all were negative for *TCF3*/*PBX1*, *ETV6*/*RUNX1*, and *BCR*/*ABL1* fusions by RT-PCR analysis and *MLL* rearrangements by Southern blot or FISH analyses. Informed consent was obtained according to the Declaration of Helsinki and the study was approved by the Ethics Committee of Lund University.

Whole genome sequencing

DNA from cases 1-16 was extracted from bone marrow or peripheral blood samples obtained at diagnosis and remission using the Gentra Puregene Blood Kit (Qiagen). Cases were subjected to paired-end next-generation sequencing (NGS) of the whole genome using the Complete Genomics technology[32]. Initial detection of somatic mutations, copy number changes, and structural variants was done with the Complete Genomics Cancer Sequencing v2.0 pipeline using CGA tools (Complete Genomics). Subsequent analyses and figure preparations were performed using R v2.15.2 and Circos v0.64[33,34].

Whole exome sequencing

DNA from cases 17-51 was extracted from bone marrow or peripheral blood samples obtained at diagnosis and remission using the Gentra Puregene Blood Kit (Qiagen). Libraries were constructed using the SureSelectXT2 Human All Exon V4 kit (Agilent Technologies) and cases were subjected to paired-end NGS with an Illumina Hiseq2000 (Illumina) by BGI Tech Solutions (Hong Kong). WES was also performed on DNA from cases 3, 4, 8, and 15.

Somatic mutations

For WGS, the initial putative somatic mutations identified by the Complete Genomics Cancer Sequencing pipeline, including SNVs, indels, and substitutions, were validated using standard Sanger sequencing. A total of 109 mutations from 5 cases were investigated. Based on these experiments, data were filtered on Somatic Score ≥0 and number of unique reads for the mutated allele >10, resulting in a validation rate of 89%. For WES, SNVs were identified with MuTect and indels with Indelocator[35]. Only mutations with a mutation allele frequency (MAF) >0.22 were kept, in order to limit the analysis to mutations present in the major clone. The identified mutations in the four cases that were analyzed by both WGS and WES were compared with those detected by WGS. Mutations that had not been seen by WGS were further investigated with Sanger sequencing. The final validation rate for the WES data was 98%. The data were then filtered using Annovar v2013Feb21[36], excluding variants present in the 1000 genomes project (1000g2012apr), dbSNP129, 6,500 exomes from the NHLBI-ESP project, and for cases subjected to WGS, in 69 normal genomes run on the Complete Genomics sequencing platform. The remaining mutations were subjected to gene-based annotation using Annovar. Driver mutations were identified using MutSigCV[14]. The mutational signature was investigated with MutSigCV using all coding mutations, including silent and non-silent, according to Lawrence et al[14].

Large-scale copy number and allelic fraction analyses

Assessment of large-scale copy number changes was performed using the relative coverage, based on the average number of reads, over a 100 kb window with a non-diploid model in the Complete Genomics Cancer Sequencing pipeline in the 16 cases subjected to WGS. Average lesser allele fractions (LAF) over a 100 kb window were used to identify partial and whole chromosome uniparental disomies as well as to ascertain which homologues that had been duplicated in the tetrasomies. To identify subclonality of whole chromosome copy numbers, boxplots were constructed in R using the relative coverage and the LAF per chromosome.

Structural rearrangements

"High confidence" somatic structural rearrangements identified by the Complete Genomics Cancer Sequencing pipeline in cases 1-16 were filtered for no presence in a baseline set of 69 normal genomes sequenced by the Complete Genomics technology and further analyzed using BLAT (http://genome-euro.ucsc.edu/index.html). Variants that showed high homology to multiple genomic regions were considered to be artifacts and excluded. Rearrangements of the immunoglobulin and T-cell receptor loci were assumed to be the result of somatic recombinations and were also excluded.

SNP array analysis

SNP array analysis of cases 1-16 was performed using the Illumina 1M-duo bead or 1M-quad Infinium BD Beadchip platforms according to the manufacturer's instructions; the results for cases 1-12 and 15 have been previously published[37,38]. When comparing copy number changes based on SNP array analysis and WGS, 162/163 (99%) whole chromosome changes seen by SNP array analysis were also detected by WGS. All 26 deletions and duplications found by SNP array analysis were identified also by WGS. In addition, WGS detected 26

deletions or duplications that had been missed by SNP array analysis; all but two of these were below the resolution limit of the array platform (**Supplementary Table 6**).

RNA sequencing

RNA libraries from 35 high hyperdiploid ALLs were produced using the Truseq RNA sample prep kit v2 (Illumina) according to the manufacturer's instructions and sequenced on a HiScanSQ (Illumina); this cohort partly overlapped with cases that were subjected to WGS or WES. Raw reads were aligned to human genome GRCh37 using Tophat 2.0.7 (http://ccb.jhu.edu/software/tophat/) and STAR 2.4.0j (https://github.com/alexdobin/STAR). Gene expression values were estimated from the aligned reads using Cufflinks 2.1.0 (http://cole-trapnell-lab.github.io/cufflinks/) and visualized using Qlucore Omics Explorer 2.3 (Qlucore). Fusion genes were detected using ChimeraScan 0.4.5 (http://code.google.com/p/chimerascan/). Normalized gene expression values per copy number state (within single cases) were determined for 20,855 genes, excluding genes on the X and Y chromosomes, in 29 cases with available SNP array and RNA-seq data. The copy numbers were ascertained from Illumina 1M-duo Beadchips (Illumina) using the cnvPartition analysis plugin v3.2.0 for Genome Studio 2011.1 (Illumina). Gene expression values were determined from RNA-seq data as fragments per kilobase of transcript per million reads and normalized (per gene) to mean 0 and variance 1. The copy number state of a gene within a single case was set to the lowest detected copy number of a SNP within that gene. All genes with intragenic SNPs present on the array were included in the analysis. The cumulative distribution of gene expression values for each copy number state was visualized in R using ggplot2.

Mutations and gene copies

To calculate the number of mutations occurring before (BTRI) or after (ATRI) trisomy formation, we used the WGS data from cases 1-16 to investigate the number of mutations present in one (y) or two (x) of the homologues in trisomic chromosomes, as determined by the MAF (~0.33 or 0.67, respectively). If BTRI mutations are assumed to be distributed equally on both homologues, mutations present on two homologues at the time of sequencing (x) will represent half of the BTRI mutations, i.e., BTRI = 2x. The number of mutations detected in one homologue at the time of sequencing (y) is the sum of the nonduplicated BTRI mutations and ATRI mutations, i.e., ATRI = y-x. The ratio between BTRI and ATRI was calculated as ATRI/(1.5BTRI) to adjust for the extra chromosome.

Statistical analysis

Differences between groups were calculated with two-sided Mann-Whitney or Fisher's exact tests (http://www.vassarstats.net/). *P*-values <0.05 were considered significant.

METHODS-ONLY REFERENCES

32. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).

33. R core team. *A language and environment for statistical computing.* R foundation for statistical computing, Vienna, Austria (2012).

34. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).

35. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213-219 (2013).

36. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

37. 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).

38. Paulsson, K. *et al.* Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21719-21724 (2010).

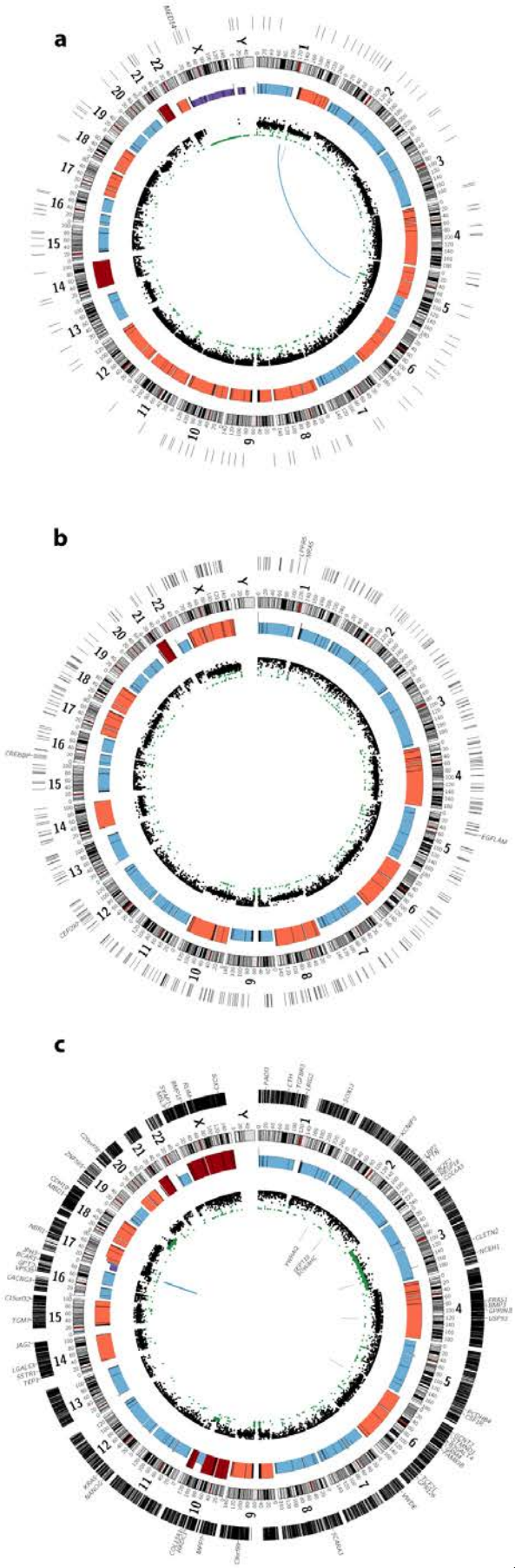

Fig 1.

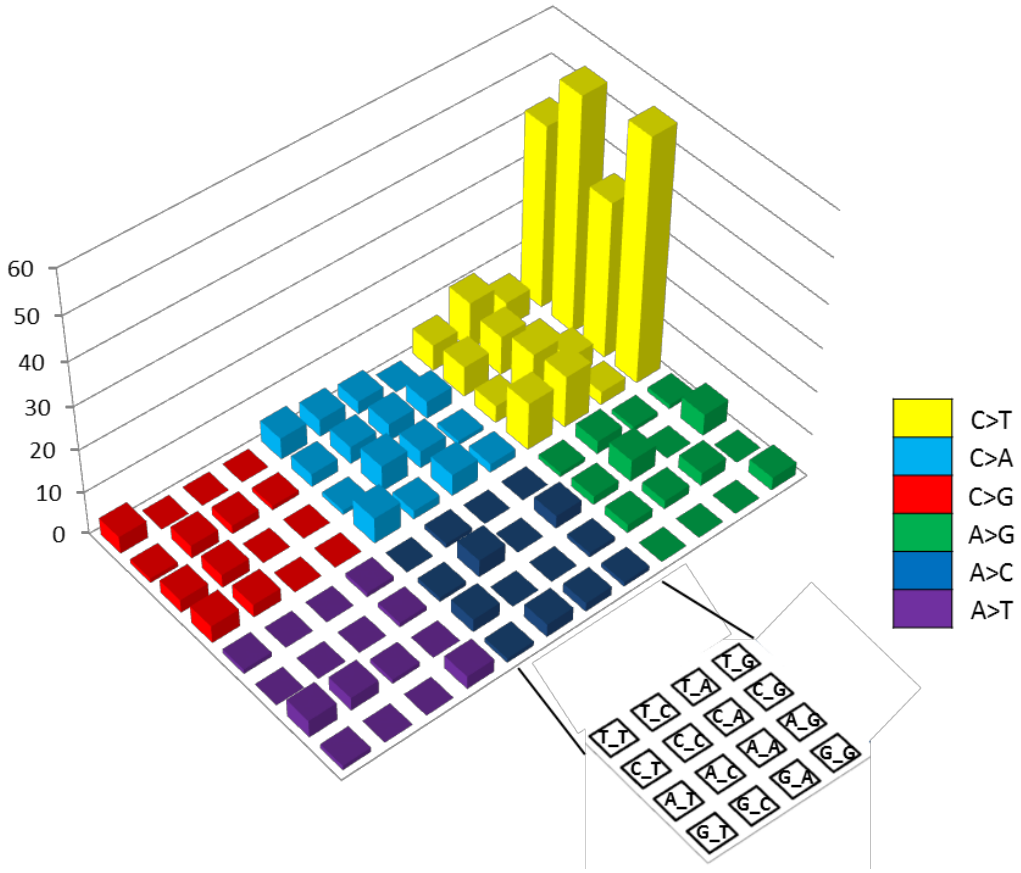
60
50
40
30
20
10
0
C>T
C>A
C>G
A>G
A>C
A>T
T_G
T_A
C_G
T_C
C_A
A_G
T_T
C_C
A_A
C_T
A_C
G_G
G_A
A_T
G_C
G_T

Fig 2

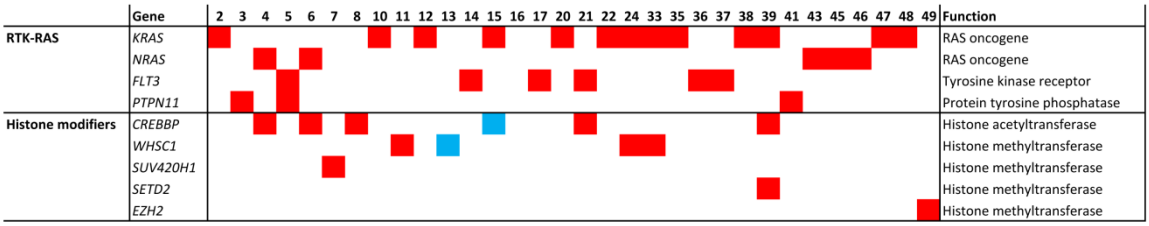
Gene
2 3 4 5 6 7 8 10 11 12 13 14 15 16 17 20 21 22 24 33 35 36 37 38 39 41 43 45 46 47 48 49
Function
RTK-RAS
KRAS
RAS oncogene
NRAS
RAS oncogene
FLT3
Tyrosine kinase receptor
PTPN11
Protein tyrosine phosphatase
Histone modifiers
CREBBP
Histone acetyltransferase
WHSC1
Histone methyltransferase
SUV420H1
Histone methyltransferase
SETD2
Histone methyltransferase
EZH2
Histone methyltransferase

Fig 3

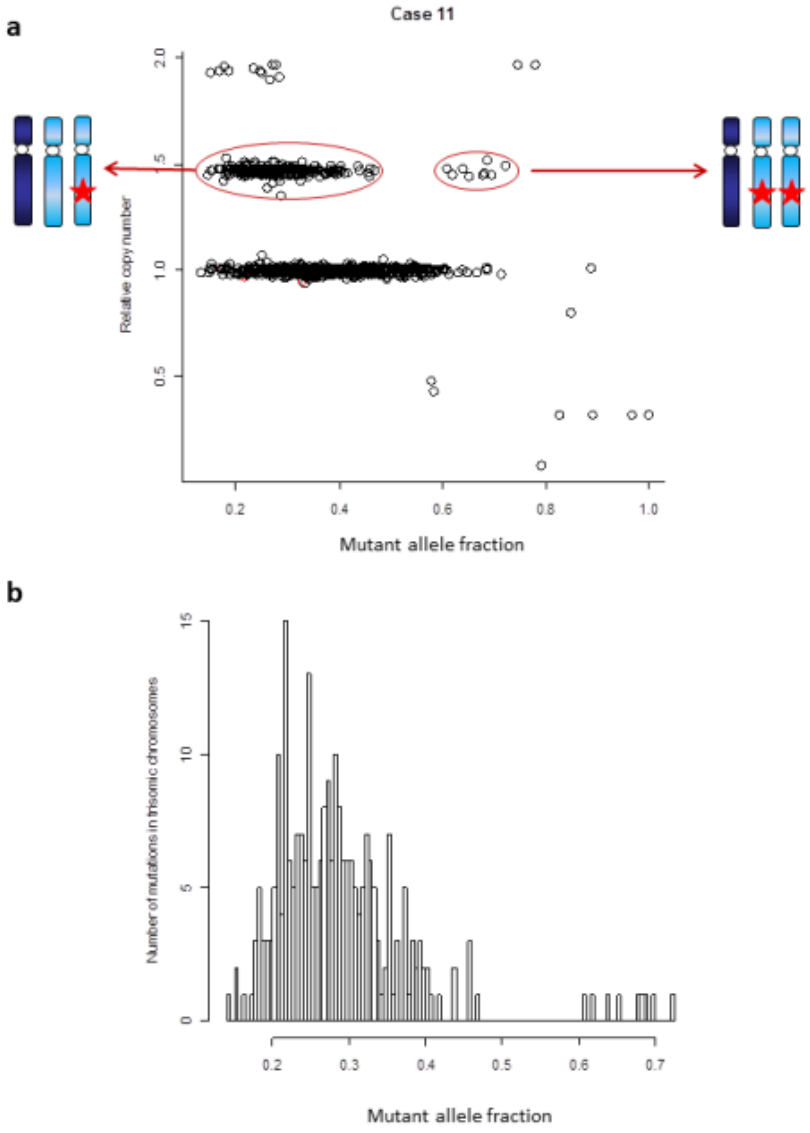
a
Case 11
Relative copy number
2.0
1.5
1.0
0.5
0.2
0.4
0.6
0.8
1.0
Mutant allele fraction
b
Number of mutations in trisomic chromosomes
15
10
5
0
0.2
0.3
0.4
0.5
0.6
0.7
Mutant allele fraction

Fig 4