Whole-exome sequencing of pediatric acute lymphoblastic leukemia.

Lilljebjörn, Henrik; Rissler, Marianne; Lassen, Carin; Heldrup, Jesper; Behrendtz, M; Mitelman, Felix; Johansson, Bertil; Fioretos, Thoas

Published in:
Leukemia

DOI:
10.1038/leu.2011.333

2012

Whole exome sequencing of pediatric acute lymphoblastic leukemia

Henrik Lilljebjörn,[1] Marianne Rissler,[1] Carin Lassen,[1] Jesper Heldrup,[2] Mikael Behrendtz,[3] Felix Mitelman,[1] Bertil Johansson,[1] and Thoas Fioretos.[1]

[1]Department of Clinical Genetics, University and Regional Laboratories, Skåne University Hospital, Lund University, Lund, Sweden. [2]Department of Pediatrics, Skåne University Hospital, Lund University, Lund, Sweden. [3]Department of Pediatrics, Linköping University Hospital, Linköping, Sweden.

Running title: Whole exome sequencing of pediatric ALL

Correspondence: Henrik Lilljebjörn, Department of Clinical Genetics, University Hospital, SE-221 85 Lund, Sweden, Tel.: +46 46-173398, Fax.: +46 46-131061, Email: henrik.lilljebjorn@med.lu.se; Thoas Fioretos, Department of Clinical Genetics, University Hospital, SE-221 85 Lund, Sweden, Tel.: +46 46-173367, Fax.: +46 46-131061, Email: thoas.fioretos@med.lu.se

Abstract

Acute lymphoblastic leukemia (ALL), the most common malignant disorder in childhood, is typically associated with numerical chromosome aberrations, fusion genes or small focal deletions, thought to represent important pathogenetic events in the development of the leukemia. Mutations, such as single nucleotide changes, have also been reported in childhood ALL, but these have only been studied by sequencing a small number of candidate genes. Herein, we report the first unbiased sequencing of the whole exome of two cases of pediatric ALL carrying the *ETV6/RUNX1* (*TEL/AML1*) fusion gene (the most common genetic subtype) and corresponding normal samples. A total of 14 somatic mutations were identified, including four and seven protein altering nucleotide substitutions in each ALL. Twelve mutations (86%) occurred in genes previously described to be mutated in other types of cancer, but none was found to be recurrent in an extended series of 29 *ETV6/RUNX1*-positive ALLs. The number of single nucleotide mutations was similar to the number of copy number alterations as detected by single nucleotide polymorphism arrays. Although the true pathogenetic significance of the mutations must await future functional evaluations, this study provides a first estimate of the mutational burden at the genetic level of t(12;21)-positive childhood ALL.

Keywords: *ETV6*/*RUNX1*, childhood acute lymphoblastic leukemia, exome sequencing, next generation sequencing

Introduction

ALL is, like all cancers, a clonal disease that arises from a cell that has acquired a set of features that allows it to escape the rules governing proliferation and differentiation in normal cells.[1] With the increased resolution in techniques for studying genetic changes, a large number of mutations that contribute to these acquired capabilities has been discovered in neoplastic cells.[2,3] The increased knowledge of the underlying genetic changes has had profound implications for improved diagnostic accuracy, prognostication, and the development of targeted treatment. Hence, detailed knowledge of the mutations contributing to ALL is clinically important. Recently, technological advancements have enabled whole genome sequencing of individual tumors and tumor cell lines;[4-8] however, this technology is still relatively expensive. An alternative that is both cheaper and requires less handling of data is to sequence only the known protein coding regions, the exome, which constitutes ~1% of the total human genome. This subset can be highly enriched using hybridization-based techniques; a strategy that has successfully been used in combination with highly parallel sequencing to identify the disease causing mutations in several Mendelian disorders,[9-11] and recently also to identify somatic mutations in three cases of acute promyelocytic leukemia.[12] We here report, for the first time, how this strategy can be used to explore fully the exome mutation profiles associated with childhood ALL.

Material and methods

Patient material

In total, four DNA samples were subjected to exon enrichment followed by highly parallel sequencing. The samples were from two *ETV6/RUNX1*-positive ALL cases, each with a matched *ETV6/RUNX1*-negative follow-up sample. Bone marrow blast counts were 85% for case 1 and 96% for case 2. DNA was extracted using standard methods from bone marrow (BM) at ALL diagnosis and from blood or BM for the normal samples. The normal samples were taken at two and 30

months after diagnosis, respectively, and in both cases the patient's BM was *ETV6/RUNX1*-negative by reverse transcriptase PCR at that time. This study was reviewed and approved by the Research Ethics Committees of Lund and Linköping Universities.

Exon enrichment and sequencing

Prior to sequencing, the DNA was enriched for exonic sequences using 2.1M sequence capture human exome arrays (Roche, Madison, USA). Highly parallel sequencing was performed on the exon-enriched DNA using the Genome Analyzer II (Illumina, San Diego, USA). The samples were prepared according to the standard Illumina protocol for paired end sequencing, modified to include a sequence capture step. In brief, five microgram of genomic DNA was fragmented to an average size of 300 bp using sonication. The fragment ends were repaired and phosphorylated using Klenow, T4 DNA polymerase, and T4 polynucleotide kinase. Next, Illumina paired end adapters were ligated to the fragments by T-A mediated ligation. The fragments were then hybridized to a capture array for 64 hours at 42 °C. The arrays were washed and the captured DNA was eluted using Nimblegen elution system (Roche). The eluted DNA was amplified by PCR using Illumina paired end primers and size selected for 300 bp fragments using gel electrophoresis. Each eluted DNA library was seeded onto between 4 and 6 lanes of a genome analyzer flowcell at 16-18 pM. The libraries were subjected to 42 (read 1 for one flow cell) or 55 (remaining reads) sequencing cycles. Exon capture and highly parallel sequencing were performed in collaboration with Ambry Genetics, Aliso Viejo, USA.

Data analysis

Cluster intensities were extracted from the raw image data and base calling was performed using RTA (Illumina). The base called reads were quality filtered using the Illumina pipeline software (Illumina). The filtered reads were aligned to the hg18 build of the human genome using bwa.[13]

Read pairs with identical start and end positions were assumed to be PCR duplicates and were removed using picard (http://picard.sourceforge.net/); the software was altered to keep the most common sequence variant (instead of the variant with the highest total quality value). Genotyping was performed using the maq consensus model implemented in samtools.[14,15]

PCR, capillary sequencing and restriction enzyme assay

Primers for PCR amplifying all regions with candidate mutations were designed using Primer3 (http://frodo.wi.mit.edu/primer3/). The amplified DNA was sequenced using the Bigdye terminator cycle sequencing kit (Life Technologies, Carlsbad, USA) and analyzed using an ABI3130 genetic analyzer (Life Technologies). Mutations were identified using a version of the SeqDoc[16] software modified to run standalone. The *FLT3* D835 mutation was studied with PCR using primers 5'-ATC ATC ATG GCC GCT CAC-3' and 5'-GCA CTC AAA GGC CCC TAA CT-3' followed by restriction cleavage using *Eco*RV; the uncleaved fragment was then sequenced. A similar approach was used to determine the proportion of mutated alleles. A forward primer modified to include a 5-carboxyfluorescein fluorophore was used for the PCR and the fragments were analyzed using an ABI3130 genetic analyzer after cleavage.

SNP arrays

DNA from both leukemia and remission samples were analyzed using HumanCNV370-quad arrays (Illumina) and cytogenetics whole-genome 2.7M arrays (Affymetrix, Santa Clara, USA) according to the manufacturer's instructions. Genotypes were determined from the HumanCNV370-quad arrays using Genome studio (Illumina) and the copy number states across the genome were determined from the cytogenetics whole-genome 2.7M arrays using Chromosome analysis suite (Affymetrix). For regions determined to be mosaic by the software, the mosaic copy number state

was used. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE25117.

Targeted gene enrichment and sequencing

Coding regions from the selected genes were enriched using a custom Selector technology target enrichment kit (Halo genomics, Uppsala, Sweden). The kit was used according to the manufacturer's specifications. In brief, DNA from *ETV6/RUNX1*-positive ALLs and *ETV6/RUNX1*-negative follow-up samples were digested in eight different restriction enzyme mixes per sample. The digested DNA samples were pooled and hybridized to custom biotin labeled probes. The hybrid DNA containing biotin labeled probes and digested fragments were enriched using streptavidin coated magnetic beads and then ligated into circular fragments using DNA ligase. The circular fragments were amplified by rolling circle amplification using the illustra Templiphi DNA amplification kit (GE Healthcare, Little Chalfont, UK). The target enriched DNA samples were prepared and sequenced on an Illumina HiscanSQ by using the Truseq DNA sample prep kit (Illumina), the Truseq PE cluster kit (Illumina), and the Truseq SBS kit (Illumina) according to the manufacturer's instructions.

Results

Exome sequencing performance

The exome sequencing was performed on leukemic and constitutional DNA from two children with ALL, by combining sequence capture using Nimblegen 2.1M human exome arrays (Roche) and second generation sequencing using the Genome Analyzer II (Illumina). The sequence capture array targets 551 microRNAs and 180,641 exons, corresponding to 25.4 Mb of genomic sequence and representing more than 18,000 consensus coding sequence (CCDS) transcripts.[17] The sequencing was performed using 54 bp paired end reads and generated between 6.5 and 9.4 gigabases (Gb) of

raw data for each case (Table 1). Of the raw reads, 92% could be aligned to the human reference genome build hg18. PCR duplicates were removed from the raw data, leaving 17.4–27.7 million unique reads mapping to the genome. Of the unique reads, 41% were mapped to the exome with a mean coverage of 10.4–19.4 reads and a median coverage of 9–18 reads. All cases were also genotyped using HumanCNV370-quad arrays (Illumina) and the sequence- and array-based genotypes were compared for all coding single nucleotide polymorphisms (SNPs) present on the array (in total 4,497 SNPs). We found that for SNPs meeting our minimum quality values (at least 5 reads and a maq[14] genotype quality score or SNP quality score over 35) there was a 98.9% concordance between genotyping based on sequence and array data. The proportions of the exome covered at this level were 80 and 85% for the two ALLs. Our minimum quality values have a lower amount of minimum reads but a more stringent genotype quality compared with other groups that have performed exome sequencing.[9] The limit of five reads was chosen based on the assumption that the reads follow a binomial distribution; in that case, 97% of positions carrying a non-reference nucleotide on one allele should have at least one read from the alternate allele when covered by five reads. At least 81% of the positions should have two or more reads from the alternate allele, which is sufficient for genotyping using the maq consensus model.[14]

Somatic exome mutations and copy number changes

Putative somatic mutations were separated from constitutional variants using a filtering strategy (Figure 1). We identified a total of 96 such mutations. Of these, 14 variants were confirmed to be genuine somatic mutations; seven in each ALL. Of the remaining 82 candidates, ten were constitutional variants that had been missed by exome sequencing of the normal sample; the remaining 72 were false positive findings. Notably, while 72 false positives in 96 candidates might seem high, this number should be compared with the ~23,000 variants identified in the first step of the filtering strategy. Most of the errors will not be removed by filtering against dbSNP, and only

positions with systematic read errors will be removed by filtering against the reads in the matched normal sample. Hence, sequence errors will be overrepresented in the list of candidate somatic mutations, which necessitates validation using an alternative sequencing approach as the last step in this powerful strategy.

The 14 confirmed somatic mutations (Table 2, Figure 2) included eleven (79%) mutations that would alter the amino acid composition of the resulting protein, comprising nine missense mutations, one nonsense mutation and one mutation of a splice donor site predicted to alter splicing. Along with the single nucleotide mutations, the two *ETV6/RUNX1*-positive cases also carried copy number changes likely to alter the expression of affected genes. The ALLs and the matched normal material were analyzed using cytogenetics whole-genome 2.7M arrays (Affymetrix) to identify somatic copy number aberrations. Apart from germline variants and somatic rearrangements of immunoglobulin genes and T-cell receptors, the ALLs contained 22 regions, in total, with acquired copy number aberrations (Figure 2, Table 3). Only one of these – gain of chromosome 16 in case 1 – overlapped with a somatic mutation identified by exome sequencing. Ten of the regions with acquired copy number aberrations have previously been described by us to be recurrent in *ETV6/RUNX1*-positive ALLs,[20] and are likely to constitute driver mutations, *i.e.* mutations that contribute functionally to leukemia development.[21]

Features of the exome mutations

All 14 mutations identified by exome sequencing were studied in a panel of 27 primary *ETV6/RUNX1*-positive ALLs and two cell lines using capillary sequencing of a 200–500 bp region surrounding the mutation. In addition to this, a target enrichment strategy was used in combination with highly parallel sequencing in an effort to determine the presence of somatic mutations in the entire coding regions of the 14 genes in a subpanel of seven primary *ETV6/RUNX1*-positive ALLs.

This strategy produced high quality genotype information (*i.e.* more than 10 reads and a genotype quality above 50) for almost the entire coding regions of *FLT3*, *GKN1*, *SERPINB1*, and *TP53INP1* (range 68-99% covered, median 87% covered). Approximately half of the coding regions for the genes *CNTN2*, *CSMD2*, *KRT79*, *RUNX1T1*, *RYR1*, and *ZNF546* were successfully genotyped (range 11-91% covered, median coverage 48%). No reliable genotype data were produced for the remaining genes (*MCAM*, *P2RY6*, *PPL*, and *SARDH*). None of the mutations was found to be recurrent and no new mutations were identified using these two strategies. The presence of additional somatic events in the identified genes was also studied using the Catalogue of Somatic Mutations in Cancer (COSMIC).[2] Notably, this revealed that 12 (86%) of the 14 affected genes had previously been described as mutated in cancer. The two tools PolyPhen-2[22] and SIFT[23] were used to predict the effect of amino acid changes on the resulting proteins. Five of the nine missense mutations were predicted by at least one of the tools to disturb the normal function of the protein (Table 2).

Discussion

In the present study we have sequenced the whole coding exome of two *ETV6/RUNX1*-positive ALLs and corresponding normal samples. This allowed us, for the first time, to obtain an overview of the total mutational burden in the coding regions of childhood ALLs. We identified 14 somatic exome mutations (Table 2, Figure 2) and 22 copy number aberrations (Table 3) in the two ALLs. This rather low number of mutations suggests that ALLs are more genetically stable than previously anticipated and that the critical genetic alterations, underlying leukemia development, should constitute a high proportion of the identified changes. It should be noted, however, that although the current approach will identify a large proportion of the somatic mutations in coding regions, a few somatic mutations are likely to remain undetected. The 15-20% of the exome that did not receive a

high quality genotype could harbor 1-2 additional somatic mutations in each case, assuming these regions are mutated to a similar degree as the successfully genotyped part of the exome.

The contribution to leukemia development is currently unknown for all single nucleotide mutations except the *FLT3* mutation, for which there is ample evidence that it is a driving mutation in both lymphoid and myeloid leukemias.[24] Admittedly, a functional evaluation of the outcome of each mutation would be desirable to determine their contribution to leukemogenesis. However, the current lack of reliable assays for determining an additive effect for mutations in *ETV6/RUNX1*-positive cells makes this a challenging task. In the lack of functional data, the best indicator for mutations contributing to leukemia development would be recurrence of the mutation in ALL or other types of cancer. This concept also forms an important basis for cancer mutation cataloging efforts such as COSMIC and the Mitelman database of chromosome aberrations in cancer.[2,3]

The screen for recurring mutations in 29 *ETV6/RUNX1*-positive ALLs together with the screen for additional somatic mutations elsewhere in the implicated genes in seven *ETV6/RUNX1*-positive ALLs did not reveal any new somatic mutations. Strikingly, however, 12 of the 14 genes were described to be mutated in other types of cancer in COSMIC. While these 12 genes are candidate driver mutations, the high proportion of COSMIC genes does not deviate significantly from the expected proportion if the mutations were completely random (data not shown). Hence, they could theoretically constitute passenger mutations.

The outcome of the mutations was also modeled using the prediction tools PolyPhen-2 and SIFT (Table 2). These analyses, combined with studies of the described function of the affected genes, highlighted seven of the mutations as potential driver mutations. For example, eight of the genes with protein altering mutations (besides *FLT3*) are known from COSMIC to display somatic mutations in other types of cancer. Five of these carried mutations that were considered likely to affect the function of the translated protein, either because the amino acid change was predicted to be “damaging” or “probably damaging” by at least one prediction tool (the case for *MCAM*,

CNTN2, *KRT79*, and *RYR1*) or because the mutation affected a splice donor site (the case for *RUNX1T1*) which most likely results in a truncated protein. Two of these genes, *MCAM* and *RUNX1T1*, are known to be perturbed in cancer by other means than those described in COSMIC. *MCAM* encodes a cell adhesion molecule (also known as CD146) that is used as a marker for candidate mesenchymal stem cells, and whose expression has been found to correlate with melanoma progression.[25] *MCAM* also displays aberrant expression levels in prostate and breast tumors.[25] *RUNX1T1* encodes a transcription factor that is commonly rearranged in acute myeloid leukemia as part of the fusion gene *RUNX1/RUNX1T1* (*AML1/ETO*) resulting from the t(8;21)(q22;q22), indicating that disruption of *RUNX1T1* can be an oncogenic event.

The mutation in *CSMD2*, a gene encoding either a transmembrane receptor or an adhesion protein of unknown function, was predicted to be "benign" and "tolerated" by the two prediction models, which could indicate that this is a passenger mutation. However, several somatic mutations are described in this gene in COSMIC. In addition to this, the highly related genes *CSMD1* and *CSMD3* are known to be rearranged in a variety of tumors and suggested to be tumor suppressor genes.[2,26] For *CSMD2* we could also show that only the mutated version of the transcript was expressed, despite the presence of a normal allele at the genomic level (Figure S1), indeed suggesting that this gene may act as a tumor suppressor in leukemia.

One of the genes that was not described as mutated in COSMIC, *TP53INP1*, encodes a protein that acts downstream of TP53. Interestingly, although this gene had not previously been found to be mutated in cancer, it has been suggested to function as a tumor suppressor.[27] The mutation in this gene is predicted by both SIFT and PolyPhen-2 to affect the function of the protein. It is also interesting to note that all three genes with synonymous mutations in this study (*SERPINB1*, *PPL*, and *ZNF546*) are known from COSMIC to be mutated in cancer. This indicates high mutation rates in somatic cells, or that the silent mutations are important, for example by affecting splicing or translation of the proteins.

The single nucleotide variant identified in *FLT3* (Table 2) is a well known driving mutation in both lymphoid and myeloid leukemias.[24] This mutation was only detected in 3 of 30 reads in the affected case and did not receive a high quality genotype; hence, it was not detected using the primary filtering strategy. It was, however, identified when low quality positions were compared with known mutations in COSMIC (Figure 1b). The *FLT3* mutation could initially not be confirmed using capillary sequencing. However, using a PCR and restriction enzyme-based assay, combined with sequencing (Figure S2), we could show that the mutation was indeed present, but not in more than 15% of the *FLT3* alleles; this would explain the low number of reads with the mutation in the exome sequencing. That a mutation present in such a minor clone (approximately 30% of cells assuming a heterozygote mutation) could be detected at all shows the strength of highly parallel sequencing over capillary sequencing, although a higher read depth would have been required to permit such sensitivity throughout the exome.

In conclusion, we have performed the first exome sequencing of ALL samples paired with matched normal samples using sequence capture and highly parallel sequencing. This analysis revealed that seven somatic single nucleotide changes were present in each of the two analyzed ALLs; a similar number to the total number of copy number changes present in these ALLs (eight and fourteen, respectively). Moreover, the number of mutations identified in the exome of the ALLs is similar to what has previously been described in the coding region of acute myeloid leukemia and acute promyelocytic leukemia genomes.[4,5,12] None of the identified mutations was present in an extended collection of *ETV6/RUNX1*-positive ALLs. Despite this, 13 of the 14 mutations affected genes previously implicated in cancer, or previously speculated to be important in cancer development. While a number of these could be highlighted as potential driver mutations, it seems that the majority of single nucleotide mutations in *ETV6/RUNX1*-positive ALLs are rare within the subgroup. The importance of these rare mutations remains to be determined by functional evaluation and by sequencing larger series of *ETV6/RUNX1*-positive ALLs.

Acknowledgements

This work was supported by grants from the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, the Swedish Research Council (personal project grant to T.F.; Hemato-Linne and BioCARE strategic research program grants), the Inga-Britt and Arne Lundberg Foundation, the Gunnar Nilsson Cancer Foundation, and the Medical Faculty of Lund University.

Conflict of interest

The authors declare no conflict of interest.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000; **100**: 57-70.
2. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, *et al.* The catalogue of somatic mutations in cancer (COSMIC). *Curr Protoc Hum Genet* 2008; **57**: 10.11.1–10.11.26.
3. Mitelman F, Johansson B, Mertens F. Mitelman Database of Chromosome Aberrations in Cancer. 2010. http://cgap.nci.nih.gov/Chromosomes/Mitelman
4. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature* 2008; **456**: 66-72.
5. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* 2009; **361**: 1058-1066.
6. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009; **461**: 809-813.
7. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 2010; **463**: 184-190.
8. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 2010; **463**: 191-196.
9. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010; **42**: 30-35.
10. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 2009; **106**: 19096-19101.
11. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, *et al.* Exome sequencing identifies *WDR35* variants involved in sensenbrenner syndrome. *Am J Hum Genet* 2010; **87**: 418-423.

12. Greif PA, Yaghmaie M, Konstandin NP, Ksienzyk B, Alimoghaddam K, Ghavamzadeh A, *et al.* Somatic mutations in acute promyelocytic leukemia (APL) identified by exome sequencing. *Leukemia* 2011;
13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754-1760.
14. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008; **18**: 1851-1858.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078-2079.
16. Crowe M. SeqDoC: rapid SNP and mutation detection by direct comparison of DNA sequence chromatograms. *BMC Bioinformatics* 2005; **6**: 133.
17. Hedges DJ, Hedges D, Burges D, Powell E, Almonte C, Huang J, *et al.* Exome sequencing of a multigenerational human pedigree. *PLoS ONE* 2009; **4**: e8232.
18. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001; **29**: 308-311.
19. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* 2009; **19**: 1639-1645.
20. Lilljebjörn H, Soneson C, Andersson A, Heldrup J, Behrendtz M, Kawamata N, *et al.* The correlation pattern of acquired copy number changes in 164 *ETV6/RUNX1*-positive childhood acute lymphoblastic leukemias. *Hum Mol Genet* 2010; **19**: 3150-3158.
21. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009; **458**: 719-724.
22. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248-249.
23. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003; **31**: 3812-3814.

24. Stirewalt DL, Radich JP. The role of FLT3 in haematopoietic malignancies. *Nat Rev Cancer* 2003; **3**: 650-665.

25. Ouhtit A, Gaur RL, Abd Elmageed ZY, Fernando A, Thouta R, Trappey AK, *et al.* Towards understanding the mode of action of the multifaceted cell adhesion receptor CD146. *Biochim Biophys Acta* 2009; **1795**: 130-136.

26. Ma C, Quesnelle KM, Sparano A, Rao S, Park MS, Cohen MA, *et al.* Characterization *CSMD1* in a large set of primary lung, head and neck, breast and skin cancer tissues. *Cancer Biol Ther* 2009; **8**: 907-916.

27. Cano CE, Gommeaux J, Pietri S, Culcasi M, Garcia S, Seux M, *et al.* Tumor protein 53–induced nuclear protein 1 is a major mediator of p53 antioxidant function. *Cancer Res* 2009; **69**: 219-226.

Figures

Figure 1. Filtering strategies used to identify somatic mutations in the exomes of two *ETV6/RUNX1*-positive ALLs. (A) Positions with five or more reads and a maq genotype quality or maq SNP quality higher than 35 were considered high quality positions. All single nucleotide variants (SNVs) and insertions/deletions (InDels) of high maq quality were included as potential somatic mutations. Positions present in dbSNP 130[18] were considered constitutional variants and removed from the list of potential somatic mutations. Variants present in the matched normal sample were also removed from the list along with the small number of variants that did not have enough reads in the normal sample. Finally, variants present only in reads with identical start points were removed since such reads are likely to be left-over PCR-duplicates not removed by the initial PCR-duplicate filter. All potential somatic mutations left at this stage were examined using PCR followed by capillary sequencing. (B) To determine if positions with five or more reads and a maq genotype quality below 35 contained mutations in known cancer genes, the list of low quality positions were compared to sites of known mutations in COSMIC.[2] All variants seen in more than one read but not present in the matched normal sample were examined using PCR followed by capillary sequencing. The mutation in *FLT3* was also studied using PCR followed by *Eco*RV digestion.

Figure 2. Somatic exome mutations and copy number changes in two *ETV6/RUNX1*-positive ALLs illustrated using circos.[19] Individual chromosomes are indicated in different colors in the outer circle, chromosome numbers are indicated inside. Chromosome bands are indicated in the next circle together with chromosome positions according to genome build hg18. The inferred copy number state, detected by cytogenetics whole-genome 2.7M arrays (Affymetrix), is indicated by the blue line in the innermost circle. All germline copy number variants were excluded from the data. The position of each somatic single nucleotide mutation detected by exome sequencing is indicated

by a red dot together with the name of the affected gene. The *ETV6/RUNX1* fusion, generated by t(12;21)(p13;q22), is indicated by a red line.

Tables

Table 1. Summary of sequencing statistics

Sample	Sex	Flowcell lanes used	Raw output (Gbp)	Reads mapping to genome	Unique reads	Unique reads mapping to exome	High quality exome positions[a]	Mean/median coverage
Case 1 leukemia	**M**	**4**	**7.4**	**129,785,195 (91%)**	**20,076,398 (15%)**	**8,603,304 (43%)**	**21,586,544 (85%)**	**14.3 / 12**
Case 1 normal	M	6	9.4	170,904,800 (94%)	27,715,078 (16%)	11,928,488 (43%)	23,930,258 (94%)	19.4 / 18
Case 2 leukemia	**F**	**4**	**6.5**	**116,114,067 (93%)**	**17,421,572 (15%)**	**6,307,372 (36%)**	**20,285,243 (80%)**	**10.4 / 9**
Case 2 normal	F	4	6.9	117,163,703 (88%)	18,281,641 (16%)	7,409,842 (41%)	21,111,593 (83%)	12.3 / 11

[a]High quality exome positions are defined as exon-, splice site-, or miRNA positions targeted by the Nimblegen 2.1M human exome array that are covered by at least 5 reads and genotyped with a maq[14] genotype quality score or a maq SNP quality score over 35.

Table 2. Confirmed somatic mutations identified by exome sequencing

Case	Chr	Position	Gene symbol (name)	Base change	Predicted protein change	PolyPhen-2 prediction	SIFT prediction	Mutations in COSMIC (cases studied)	Primary tumor tissue, COSMIC	Gene ontology, molecular function
1	1	34,270,835	*CSMD2* (CUB and Sushi mutiple domains 2)	C>T	p.R115H	Benign	Tolerated	6 (75)	Ovary, pancreas	Protein binding
1	2	69,061,401	*GKN1* (gastrokine 1)	C>T	p.T181M	Benign	Tolerated	1 (131)	Breast	Growth factor activity
1	6	2,785,787	*SERPINB1* (serpin peptidase inhibitor, clade B, member 1)	G>A	Synonymous p.F11	ND	ND	5 (260)	Breast, ovary	Peptidase inhibitor activity
1	11	72,686,034	*P2RY6* (pyrimidinergic receptor P2Y, G-protein coupled, 6)	G>A	p.V275I	Benign	Tolerated	1 (69)	Ovary	UDP-activated nucleotide receptor activity
1	11	118,688,836	*MCAM* (melanoma cell adhesion molecule)	G>A	p.S198L	Probably damaging	Tolerated	3 (115)	Breast, ovary	-
1	16	4,873,644	*PPL* (periplakin)	C>T	Synonymous p.P1671	ND	ND	4 (71)	CNS, ovary	Structural constituent of cytoskeleton
1	19	45,212,610	*ZNF546* (zinc finger protein 546)	A>T	Synonymous p.G531	ND	ND	3 (133)	Breast, ovary, skin	DNA binding
2	1	203,298,249	*CNTN2* (contactin 2)	C>T	p.S390L	Probably damaging	Damaging	1 (69)	Ovary	Identical protein binding
2	8	93,068,290	*RUNX1T1* (runt-related transcription factor 1; translocated to, 1)	C>T	Splice site	ND	ND	5 (683)	CNS, LI, pancreas	Sequence-specific DNA binding transcription factor activity
2	8	96,021,348	*TP53INP1* (tumor protein p53 inducible nuclear protein 1)	G>C	p.S130C	Probably damaging	Damaging	0 (204)	-	-
2	9	135,559,916	*SARDH* (sarcosine dehydrogenase)	C>T	p.W510X	ND	ND	0 (68)	-	Sarcosine dehydrogenase activity
2	12	51,504,323	*KRT79* (keratin 79)	C>T	p.A316T	Benign	Damaging	2 (48)	Ovary	Structural molecule activity
2	13	27,490,642	*FLT3* (fms-related tyrosine kinase 3)	C>G	p.D835H	Probably damaging	Damaging	7930 (38617)	Hematopoietic and lymphoid	Receptor activity
2	19	43,701,915	*RYR1* (ryanodine receptor 1)	C>T	p.R3414C	Probably damaging	ND	10 (77)	Breast, pancreas	Ryanodine-sensitive calcium-release channel activity

Abbreviations: Chr, chromosome; ND, not determined; CNS, central nervous system; LI, large intestine; and -, not present in database.

Table 3. Somatic copy number changes in two *ETV6/RUNX1*-positive ALLs

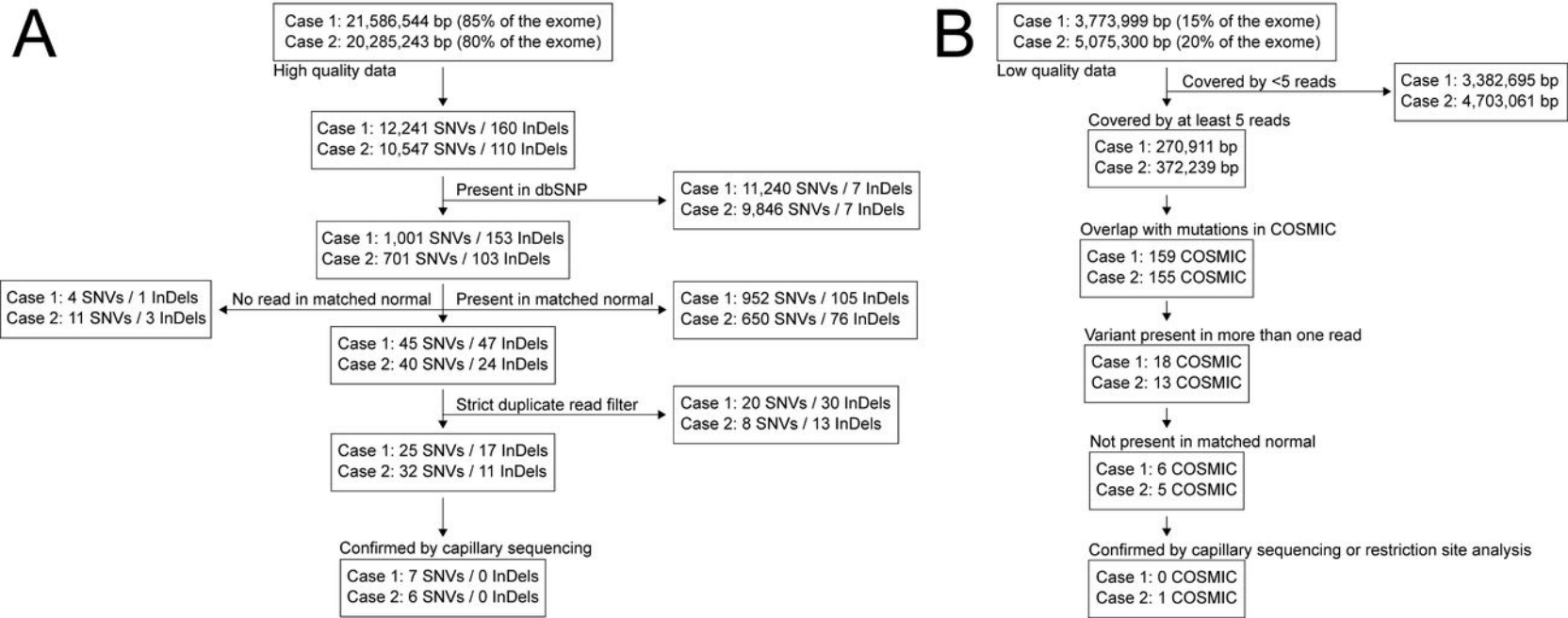
Case	Chr	Start[a]	End	Gain/ Loss	Affected genes or region	Recurrent in *ETV6/RUNX1*-positive ALL[b]
1	3	180,857,114	181,146,460	Loss	*USP13*, *PEX5L*	No
1	6	26,270,052	26,430,479	Loss	17 *HIST1* genes	Yes
1	6	82,291,115	170,756,892	Loss	403 genes	Yes
1	9	9,557,230	27,071,399	Loss	54 genes, including *CDKN2A*	Yes
1	10	111,760,107	111,858,206	Loss	*ADD3*	Yes
1	16	33,002	88,677,424	Gain	Entire chromosome 16	Yes
1	21	23,252,877	23,308,444	Loss	No gene	No
1	X	89,163,432	154,582,607	Gain	401 genes	Yes
2	1	189,084,802	189,288,937	Loss	No gene	Yes
2	2	30,622,624	30,671,918	Loss	*LCLAT1*	No
2	2	30,677,994	30,807,795	Gain	*LCLAT1*, *CAPN13*	No
2	6	32,518,630	32,691,111	Loss	*HLA-DRA*, *HLA-DRB5*	No
2	12	10,712,601	17,006,113	Loss	57 genes, including *ETV6*	Yes
2	17	25,108,115	25,127,275	Gain	*SSH2*	No
2	19	52,060,761	54,543,298	Loss	80 genes	Yes
2	21	13,527,259	46,922,328	Gain	Entire chromosome 21[c]	Yes
2	X	33,520,232	33,740,348	Loss	No gene	No
2	X	36,679,367	36,714,844	Loss	No gene	No
2	X	49,439,236	49,448,229	Loss	No gene	No
2	X	58,108,341	61,963,074	Loss	No gene	No
2	X	105,484,236	105,524,221	Loss	No gene	No
2	X	149,943,095	149,945,912	Loss	No gene	No

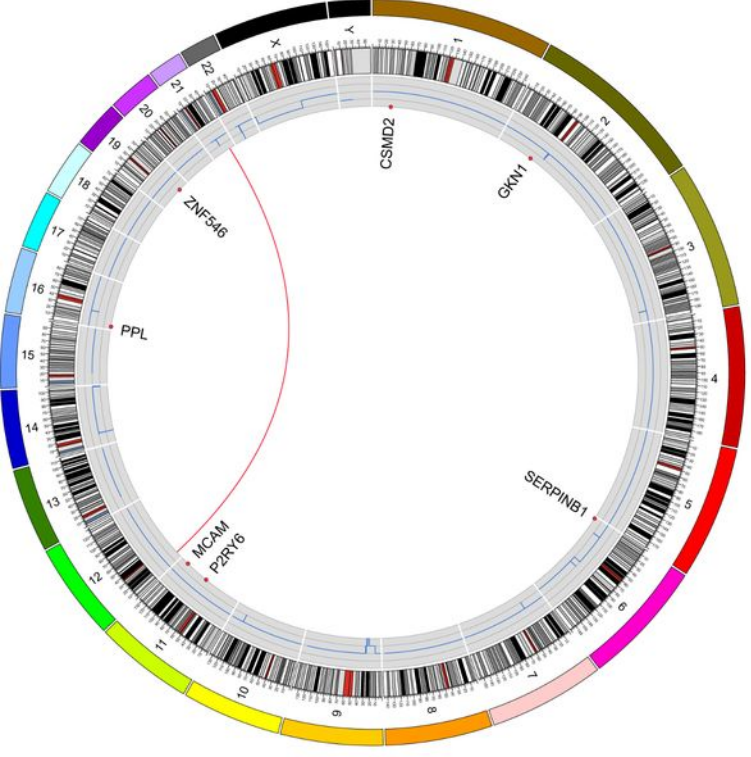
Germline changes and somatic recombinations of immunoglobulin and T-cell receptor loci are not included in the list. Abbreviations: Chr, chromosome.

[a]Genomic positions according to the hg18 genome assembly (NCBI Build 36.3).

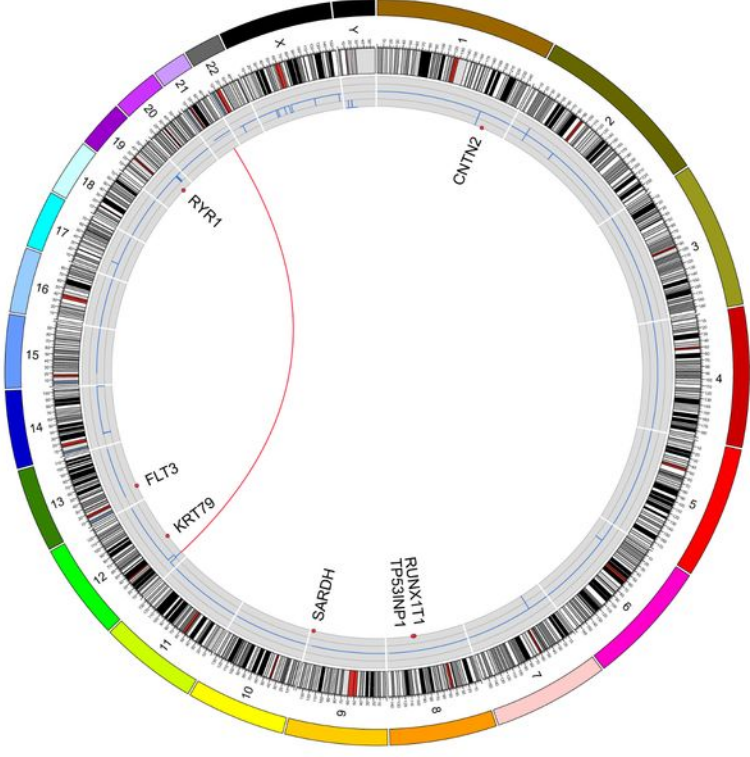
[b]Previously described by us to be recurrent in *ETV6/RUNX1*-positive ALL.[20]

[c]A mosaic copy number of 2.3 was detected, compatible with gain of one copy of chromosome 21 in 30% of the cells.


A
Case 1: 21,586,544 bp (85% of the exome)
Case 2: 20,285,243 bp (80% of the exome)
High quality data
Case 1: 12,241 SNVs / 160 InDels
Case 2: 10,547 SNVs / 110 InDels
Present in dbSNP
Case 1: 11,240 SNVs / 7 InDels
Case 2: 9,846 SNVs / 7 InDels
Case 1: 1,001 SNVs / 153 InDels
Case 2: 701 SNVs / 103 InDels
No read in matched normal
Case 1: 4 SNVs / 1 InDels
Case 2: 11 SNVs / 3 InDels
Present in matched normal
Case 1: 952 SNVs / 105 InDels
Case 2: 650 SNVs / 76 InDels
Case 1: 45 SNVs / 47 InDels
Case 2: 40 SNVs / 24 InDels
Strict duplicate read filter
Case 1: 20 SNVs / 30 InDels
Case 2: 8 SNVs / 13 InDels
Case 1: 25 SNVs / 17 InDels
Case 2: 32 SNVs / 11 InDels
Confirmed by capillary sequencing
Case 1: 7 SNVs / 0 InDels
Case 2: 6 SNVs / 0 InDels
B
Case 1: 3,773,999 bp (15% of the exome)
Case 2: 5,075,300 bp (20% of the exome)
Low quality data
Covered by <5 reads
Case 1: 3,382,695 bp
Case 2: 4,703,061 bp
Covered by at least 5 reads
Case 1: 270,911 bp
Case 2: 372,239 bp
Overlap with mutations in COSMIC
Case 1: 159 COSMIC
Case 2: 155 COSMIC
Variant present in more than one read
Case 1: 18 COSMIC
Case 2: 13 COSMIC
Not present in matched normal
Case 1: 6 COSMIC
Case 2: 5 COSMIC
Confirmed by capillary sequencing or restriction site analysis
Case 1: 0 COSMIC
Case 2: 1 COSMIC


CSMD2
GKN1
ZNF546
PPL
SERPINB1
MCAM
P2RY6

Case 1


CNTN2
RYR1
FLT3
KRT79
SARDH
RUNX1T1
TP53INP1

Case 2