RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL
HGVS HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

The Mutational Landscape in Pediatric Acute Lymphoblastic Leukemia Deciphered by Whole Genome Sequencing

Carl Mårten Lindqvist,[1] Jessica Nordlund,[1] Diana Ekman,[2] Anna Johansson,[3] Behrooz Torabi Moghadam,[1] Amanda Raine,[1] Elin Övernäs,[1] Johan Dahlberg,[1] Per Wahlberg,[1] Niklas Henriksson,[1] Jonas Abrahamsson,[4†] Britt-Marie Frost,[5†] Dan Grandér,[6] Mats Heyman,[7†] Rolf Larsson,[8] Josefine Palle,[1,5†] Stefan Söderhäll,[7†] Erik Forestier,[9†] Gudmar Lönnerholm,[5†] Ann-Christine Syvänen,[1] and Eva C. Berglund[1]*

[1]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden; [2]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden; [3]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden; [4]Department of Pediatrics, Queen Silvia Children's Hospital, Gothenburg, Sweden; [5]Department of Women's and Children's Health, University Children's Hospital, Uppsala, Sweden; [6]Department of Oncology and Pathology, Karolinska Institutet, Stockholm, Sweden; [7]Childhood Cancer Research Unit, Department of Women and Child Health, Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm, Sweden; [8]Department of Medical Sciences, Cancer Pharmacology and Computational Medicine, Uppsala University, Uppsala, Sweden; [9]Department of Medical Biosciences, University of Umeå, Umeå, Sweden

Communicated by Georgia Chenevix-Trench

Received 18 July 2014; accepted revised manuscript 22 October 2014.
Published online 30 October 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22719

ABSTRACT: Genomic characterization of pediatric acute lymphoblastic leukemia (ALL) has identified distinct patterns of genes and pathways altered in patients with well-defined genetic aberrations. To extend the spectrum of known somatic variants in ALL, we performed whole genome and transcriptome sequencing of three B-cell precursor patients, of which one carried the t(12;21)*ETV6-RUNX1* translocation and two lacked a known primary genetic aberration, and one T-ALL patient. We found that each patient had a unique genome, with a combination of well-known and previously undetected genomic aberrations. By targeted sequencing in 168 patients, we identified *KMT2D* and *KIF1B* as novel putative driver genes. We also identified a putative regulatory non-coding variant that coincided with overexpression of the growth factor *MDK*. Our results contribute to an increased understanding of the biological mechanisms that lead to ALL and suggest that regulatory variants may be more important for cancer development than recognized to date. The heterogeneity of the genetic aberrations in ALL renders whole genome sequencing particularly well suited for analysis of somatic variants in both research and diagnostic applications.

Hum Mutat 36:118–128, 2015. Published 2014 Wiley Periodicals, Inc.*

KEY WORDS: clonal heterogeneity; acute lymphoblastic leukemia; whole genome sequencing; RNA sequencing

Introduction

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer, which arises from the malignant transformation of lymphocyte progenitor cells to leukemic cells in the B- and T-cell lineages. B-cell precursor ALL (BCP-ALL) is the most common immunophenotype, which is divided into genetic subtypes with therapeutic and prognostic importance based on recurrent large-scale chromosomal aberrations that are detected in about 75% of the patients. Hyperdiploidy and t(12;21)*ETV6-RUNX1* (MIM#s 600618, 151385) rearrangement characterize the most common subtypes, which are associated with a favorable outcome (Pui et al., 2011). The high-risk T-cell immunophenotype represents about 12% of the patients.

Genomic characterization of T-ALL, BCP-ALL samples carrying the t(12;21) rearrangement, and several of the high-risk BCP-ALL subtypes using microarrays and next generation sequencing has revealed distinct patterns of genetic lesions in these subtypes [Mullighan, 2013; Papaemmanuil et al., 2014], whereas few studies have addressed the genetics of BCP-ALL patients without a known primary genetic aberration. The focus of most studies has been on protein-coding regions, where the lesions have been found to affect hematopoietic development, cell cycle regulation, Ras and tyrosine signaling, cytokine receptors, tumor suppression, and epigenetic regulation [Mullighan, 2013]. The findings that most driver genes in acute myeloid leukemia are involved in gene regulation [The Cancer Genome Atlas Research Network, 2013],

Additional Supporting Information may be found in the online version of this article.

†For the Nordic Society of Pediatric Hematology and Oncology.

*Correspondence to: Eva C. Berglund, Department of Medical Sciences, Molecular Medicine, Uppsala University, BMC, Box 1432, 751 44 Uppsala, Sweden. E-mail: eva.berglund@medsci.uu.se

Contract grant sponsors: The Swedish Cancer Society (CAN2010/592); The Swedish Childhood Cancer Foundation (11098); The Swedish Research Council for Science and Technology (90559401); The Swedish Foundation for Strategic Research (RBc08-008); The EU FP7-Funded European Sequencing and Genotyping Infrastructure (FP7 262055); The Erik, Karin and Gösta Selanders Stiftelse; The Knut and Alice Wallenberg Foundation (to the Wallenberg Advanced Bioinformatics Infrastructure).

Table 1. Clinical Characteristics of Whole Genome Sequenced ALL Patients

Patient	Sex	Age[a]	WBC[b]	% Blasts	Immunophenotype	Genetic subtype[c]	Treatment group[d]	Clinical follow-up[e]	Remission tissue
ALL_458	Male	3.6	12.3	90	BCP-ALL	t(12;21)	IR	CCR1 (8.5)	PB
ALL_559	Male	5.9	128.0	95	T-ALL	T-ALL	HR	CCR1 (7)	PB
ALL_707	Male	1.6	9.6	80–90	BCP-ALL	Other	SR	CCR1 (5)	PB
ALL_501	Female	6.7	1.4	80	BCP-ALL	Normal	SR	CCR1 (8)	BM

[a] Age at diagnosis in years.
[b] White blood cell count at diagnosis (10^9 cells/l).
[c] The full karyotypes are:
ALL_458: 46,XY.ish.t(12;21)(p13;q22),del(12)(p13p13),del(21)(q22q22)
ALL_559: 46,XY,t(7;9)(q3?4;q3?2)[10].ish.del(9)(p21p21)x2,der(11)t(7;11)(q3?4;p1?3)/46,XY[15]. The t(7;9) was detected by G-banding in 10 of 25 metaphases. Remaining aberrations, and t(7;9), were detected by FISH. The question marks indicate that the breakpoint on sub-band level is uncertain.
ALL_707: 46,XY,der(7)t(7;9)(q11;p13)del(9)(p21p24),der(9)t(7;9)(q11;p13),del(19)(q13)[24]/46,XY[1]
ALL_501: 46,XX[20]. Hyperdiploidy and the most common rearrangements (*BCR-ABL1*, *PBX1-TCF3*, *ETV6-RUNX1*, and *MLL*) were excluded by FISH and DNA index analysis.
[d] The patients were treated according to the Nordic Society for Pediatric Haematology and Oncology (NOPHO) protocols [Schmiegelow et al., 2010].
[e] Within parenthesis is the follow-up time in years.
BCP-ALL, B-cell precursor ALL; SR, standard risk; IR, intermediate risk; HR, high risk; CCR1, first continuous complete remission; PB, peripheral blood; BM, bone marrow.

and that the *TERT* promoter is recurrently mutated across several types of human cancer [Vinagre et al., 2013] suggest that variants in non-coding regulatory regions might be more important for cancer development than recognized to date.

To determine the full range of genetic lesions in pediatric ALL, we sequenced the whole genomes and transcriptomes of two patients belonging to well-characterized groups (t(12;21) and T-ALL) and two patients without a known primary genetic aberration. We analyzed coding and non-coding regions of the genome and identified putative driver genes by targeted sequencing in 168 additional patients.

Table 2. Cytogenetic Representation of ALL Patients Included in the Validation Cohort

Immunophenotype	Genetic subtype[a]	Number of patients[b]	Population (%)[c]
BCP-ALL	HeH	47 (28.0)	26.3
	t(12;21)	35 (20.8)	16.7
	Other	21 (12.5)	13.4
	Normal / no result	18 (10.7)	21.7
	t(9;22)	8 (4.8)	2.2
	11q23/MLL	4 (2.4)	3.6
	iAMP21	4 (2.4)	0.5
	t(1;19)	4 (2.4)	1.9
	dic(9;20)	3 (1.8)	1.8
	> 67 chr	1 (0.6)	0.4
T-ALL	T-ALL	23 (13.7)	10.5
Total		168 (100)	99.0

[a] HeH, high hyperdiploidy (51–67 chromosomes); t(12;21), translocation between the chromosomes (12;21)(p13;q22)*ETV6-RUNX1*; t(9;22), translocation between the chromosomes (9;22)(q11;q34)*BCR-ABL1*; 11q23/MLL, translocation between *MLL* and various other genes; iAMP21, intrachromosomal amplification of chromosome 21; dic(9;20), dicentric chromosome (9;20)(p13;q11); > 67 chr, > 67 chromosomes; Other, other clonal aberrations; Normal, no genetic aberrations detected and a normal karyotype observed in at least 5 of 25 metaphases; No result, no karyotype reported or the cytogenetic analysis failed.
[b] Within parenthesis is the percentage of samples of each subtype in the validation cohort.
[c] Percentage of each subtype in 2367 patients diagnosed with ALL in the Nordic countries during 1996–2008. The frequency of subtypes changes with time, since new subtypes are discovered and new analysis methods are added. The reason why the population does not sum to 100% is that the rare subtype hypodiploidy is not represented in the validation cohort.

Materials and Methods

Patient Samples

The pediatric ALL patients analyzed in this study were diagnosed and treated at Swedish centers (Uppsala, Umeå, Stockholm and Gothenburg) according to the Nordic Society for Pediatric Haematology and Oncology (NOPHO) protocols [Schmiegelow et al., 2010]. ALL diagnosis was established by analysis of leukemic cells with respect to morphology, immunophenotype, and cytogenetics. ALL lineage (BCP-ALL or T-ALL) was defined according to the European Group for the Immunological Characterization of Leukemias. Fluorescence *in situ* hybridization (FISH) or reverse transcriptase PCR (RT-PCR) analyses were used to screen for gene fusions. Karyotypes were based on the International System for Human Cytogenetic Nomenclature [Shaffer et al., 2013]. Bone marrow aspirates collected at diagnosis of ALL and matched germline peripheral blood or bone marrow samples collected in first continuous complete remission (CCR1) from four patients were subjected to whole genome sequencing (WGS) (Table 1). Targeted sequencing of 168 additional samples collected at ALL diagnosis and 159 matched CCR1 samples from the same patients was performed (Table 2 and Supp. Table S1). For nine patients, no CCR1 sample was available. An in-house RNA-seq dataset containing 27 BCP-ALL samples and 18 T-ALL samples (Nordlund, Dahlberg et al., unpublished data, Supp. Table S2) was used as control to assess potential effects of somatic variants on gene expression. This dataset provides a better control than matched remission samples, which are comprised of a mixture of mononuclear cells (T-cells, B-cells, monocytes, etc.) and therefore are expected to display substantial expression differences compared with ALL cells that are unrelated to somatic variants. The study was approved by the Regional Ethical Review Board in Uppsala, Sweden. The study was conducted according to the guidelines of the Declaration of Helsinki, and all patients and/or guardians provided written informed consent.

Mononuclear cells were isolated with 1.077 g/ml Ficoll-Isopage (Pharmacia, Uppsala, Sweden) density-gradient centrifugation. The proportion of leukemic cells was estimated to be ≥80% in all samples by light microscopy in May-Grünwald-Giemsa-stained cytocentrifugate preparations. DNA and RNA were extracted from vital frozen cells or cell pellets containing 1–15 million cells using the QIAamp DNA Blood or AllPrep DNA/RNA Mini Kit (Qiagen, GmbH, Hilden, Germany). RNA samples were treated with DNase using the RNase-Free DNase Set (Qiagen). DNA and RNA were quantified using the Qubit dsDNA Broad-Range assay and Qubit RNA Broad-Range assay, respectively, on a Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA, USA). The integrity of the RNA was examined by capillary electrophoresis with a Bioanalyzer using RNA 6000 Nano Labchips (Agilent, Technologies, Santa Clara, CA, USA).

WGS and Analysis of WGS Data

Four to eight WGS libraries were prepared from diagnostic and remission DNA from each of the four selected ALL patients. The libraries were sequenced paired-end with 100 or 150 bp reads using a HiSeq2000 or GAIIx instrument (Illumina Inc, San Diego, CA, USA). Sequence reads were aligned to the human reference genome hg19 using BWA [Li and Durbin, 2009]. Single nucleotide variants (SNVs) were called with SomaticSniper [Larson et al., 2012] and insertion-deletions (indels) were called with GATK [DePristo et al., 2011]. A variant was considered conserved if it overlapped a conserved non-exonic element (CNEE) [Lowe et al., 2011] or a region present in the phastCons track from the UCSC table browser. DNase hypersensitive (DHS) regions and active histone marks (H3K4me1, H3K27ac, H3K4me3, H3K36me3) in lymphoid cells were determined using data from the NIH Epigenomics project [Bernstein et al., 2010] as previously described [Nordlund et al., 2013]. The experiment IDs are listed in Supp. Table S3. Somatic copy number alterations (CNAs) were called using BIC-Seq [Xi et al., 2011], ControlFREEC [Boeva et al., 2011], dwac-seq (https://github.com/Vityay/DWAC-Seq), and Patchwork [Mayrhofer et al., 2013]. CNAs detected by at least two programs were retained. All variants were annotated against the Ensembl database.

RNA Sequencing and Analysis of RNA Sequence Data

Strand-specific libraries for RNA-sequencing (RNA-seq) were prepared from 1 μg RNA using the ScriptSeq V2 RNA-seq Library Preparation Kit (Epicentre, Madison, WI, USA) after ribosomal RNA depletion with the Ribo-Zero method (Epicentre). Paired-end sequencing was performed on a HiSeq2000/2500 (Illumina). The read length was 100 bp for the four whole genome sequenced samples and 50 bp for the samples in the control dataset. Alignment of sequence reads and quantification of expression levels was performed with TopHat and Cufflinks [Trapnell et al., 2012]. The expression levels were normalized to fragments per kilobase per million mapped reads (FPKM) with Cufflinks. Fusion genes were identified using FusionCatcher (Nicorici et al., submitted manuscript).

HumanOmni2.5 BeadChip Genotyping

Two hundred and fifty nanograms of DNA from diagnostic and remission samples from the four whole genome sequenced patients was genotyped using the HumanOmni2.5 BeadChip (Illumina). The diploid coverage of the WGS data was estimated as the percentage of heterozygous sites detected using genotyping that were also heterozygous in the sequence data. CNAs were predicted using ASCAT [Van Loo et al., 2010].

Target Capture Experiment

The following categories of genomic regions were selected for target capture and resequencing (Supp. Fig. S1): (1) 51 bp regions flanking all candidate SNVs identified by WGS. (2) Exons of genes with an SNV or indel in an exon or an untranslated region (UTR) (Supp. Table S4). (3) Non-coding conserved or DHS regions containing a candidate SNV and regions flanking candidate SNVs with score ≤2 in RegulomeDB [Boyle et al., 2012]. Target capture was performed using 200 ng DNA and reagents from a HaloPlex Target Enrichment kit (Agilent), according to the HaloPlex Target Enrichment System Automation Protocol Version D.3. All leukemic samples ($n = 172$) and the four whole genome sequenced remission samples were enriched individually. Remaining remission samples were enriched in pools of 10 samples. In addition, 84 samples from healthy Swedish blood donors, enriched in pools of 21 samples, were included as population controls. Paired-end sequencing with read length of 100 bp was performed on a HiSeq2000/2500 (Illumina). The average sequence depth in the target region was 638× for ALL samples in the validation cohort, 162× for remission samples, and 133× for Swedish blood donors (Supp. Fig. S2).

Validation of Somatic Variants Identified by WGS

Exonic indels (n = 2 of 6 tested) and a subset of candidate SNVs (n = 50 of 60 tested) identified by WGS were validated by PCR and Sanger sequencing. Fifty nanograms of genomic DNA was whole genome amplified using the REPLI-g Midi Kit (Qiagen). PCR primers were designed using Primer3Plus [Untergasser et al., 2007]. ALL and remission samples were amplified by PCR and the products were sequenced with BigDye Terminator 3.1 chemistry using an Applied Biosystems 3730XL DNA sequencer. The sequence traces were analyzed with the Sequencher software (Applied Biosystems, Foster City, CA, USA). For validation of remaining SNVs, the allele fraction (AF) from the target capture experiment was calculated with a custom Python script (available at https://github.com/Molmed/Berglund-Lindqvist-2013). A candidate SNV was considered validated if the AF was ≥ 0.1 in the ALL sample, the AF was < 0.01 in the matched remission sample, and the sequence depth was ≥10 in both samples. Bases with a Phred quality score < 20 and bases in a read with mapping quality = 0 were disregarded. Positions where an SNV was called in more than one patient were manually inspected in the Integrative Genomics Viewer [Thorvaldsdottir et al., 2013].

Variant Calling in the Validation Cohort

SNVs and indels in the validation cohort were called with FreeBayes (http://arxiv.org/abs/1207.3907) and the GATK HaplotypeCaller [DePristo et al., 2011], respectively. Variants were filtered based on sequence coverage and quality scores. Germline variants were excluded using remission samples and population variation.

Accessibility of Reported Variants

All variants identified in the whole genome sequenced patients and the validation cohort are listed in the Supporting Information. Putative driver variants have been submitted to COSMIC with COSP ID 37259.

Further details on the materials and methods are available in the Supporting Information.

Results

Somatic Variants in Whole Genome Sequenced ALL Patients

In this study, four patients selected to be representative of pediatric ALL were subjected to WGS. ALL_458 (BCP-ALL) carried the recurrent t(12;21)(p13;q22)*ETV6-RUNX1* translocation, ALL_559 (T-ALL) had two translocations involving chromosome 7, ALL_707 (BCP-ALL) had several cytogenetic aberrations, including a t(7;9)(q11;p13) translocation, and ALL_501 (BCP-ALL) showed a normal karyotype (Table 1). Each of the patients responded to treatment and had remained in first continuous complete remission (CCR1) for at least 5 years. We sequenced the genomes of diagnostic and remission samples from these patients to an average depth of 31× high quality aligned data and the transcriptomes from the diagnostic samples (Supp. Table S5).

In the genomes of the four patients, we identified between 713 and 851 candidate somatic SNVs and exonic insertion–deletions (indels) in non-repeated regions (Supp. Table S6). We validated close to 200 somatic SNVs and indels per BCP-ALL patient, and 305 in the T-ALL patient (Supp. Tables S7–S8) using PCR and Sanger sequencing and/or target capture and deep sequencing. A relatively low validation rate of 29% was obtained for SNVs (Supp. Table S7), which can be attributed to the non-stringent criteria for inclusion of candidate SNVs in the HaloPlex experiment. In addition, most comparable studies have focused on validation in exons, which are more conserved than the remaining part of the genome and less prone to alignment artifacts. Indeed, exonic SNVs were associated with both higher somatic scores, as defined by SomaticSniper, (averages of 87 and 72 for exonic and non-exonic candidates, respectively) and a substantially higher validation rate of 66% (Supp. Table S7). The candidate SNVs that failed to validate were either false positives, or had similar AFs in the leukemic and remission samples, suggesting that they were germline variants or alignment artifacts (Supp. Fig. S3). All further analysis includes only the validated variants. The validated SNVs were evenly distributed over the genome with no evidence of hypermutated regions (Fig. 1A). The most common somatic mutation in all patients except ALL_707 (subtype “other”) was C>T, whereas C>A was most frequent in ALL_707 (Fig. 1B).

We validated 23 exonic variants in the four patients, including three loss of function mutations, 18 nonsynonymous SNVs (nsSNVs), and two synonymous SNVs (Table 3). Prediction of functional effects suggested that 11 out of 18 nsSNVs were damaging, including nsSNVs in *KRAS* (MIM# 190070) and *NOTCH1* (MIM# 190198), which are known drivers of BCP-ALL [Liang et al., 2006] and T-ALL [Weng et al., 2004], respectively. We validated 74 SNVs in non-coding putative regulatory regions, of which 50 were located in conserved regions, 16 in DNase hypersensitive (DHS) regions and eight had a high regulatory potential according to RegulomeDB (Supp. Table S7 and Supp. Fig. S4). In comparison with the remaining non-coding SNVs, these putatively functional SNVs were more often located in introns and within 1 kb from a gene and less often in intergenic regions (Supp. Fig. S4).

Using a combination of the WGS data and high-density genotype data, we identified between two and five somatic CNAs and copy-neutral loss of heterozygosity (LOH) events in the four patients (Fig. 1A and Table 4). The majority of the events (10/15) were not detected by cytogenetic analysis at diagnosis. The portion of the genome affected by these events varied between 428 kb in ALL_501 (normal karyotype) and 176 Mb in ALL_458 (t(12;21)). Several of the aberrations, including deletions of the wild type *ETV6* [12p13] and *VPREB1* (MIM# 605141) [22q11] in ALL_458, *CDKN2A* (MIM# 600160) [9p21] in ALL_559 (T-ALL) and ALL_707 (subtype “other”), and *IKZF1* (MIM# 603023) [7p12] in ALL_501, and duplication of chromosome 10 in ALL_458, are known to be recurrent in pediatric ALL [Mullighan et al., 2007; Lilljebjorn et al., 2010; Mangum et al., 2014]. Both of the *CDKN2A* deletions were homozygous and they were flanked by two LOH events in ALL_559 and by two hemizygous deletions in ALL_707. Other deletions include, for example, the oncogenes *MDM2* (MIM# 164785) and *RAP1B* (MIM# 179530) in ALL_458 and the putative tumor suppressors *FOXP1* (MIM# 605515), *RYBP* (MIM# 607535), and *SHQ1* (MIM# 613663) in ALL_559 (Table 4).

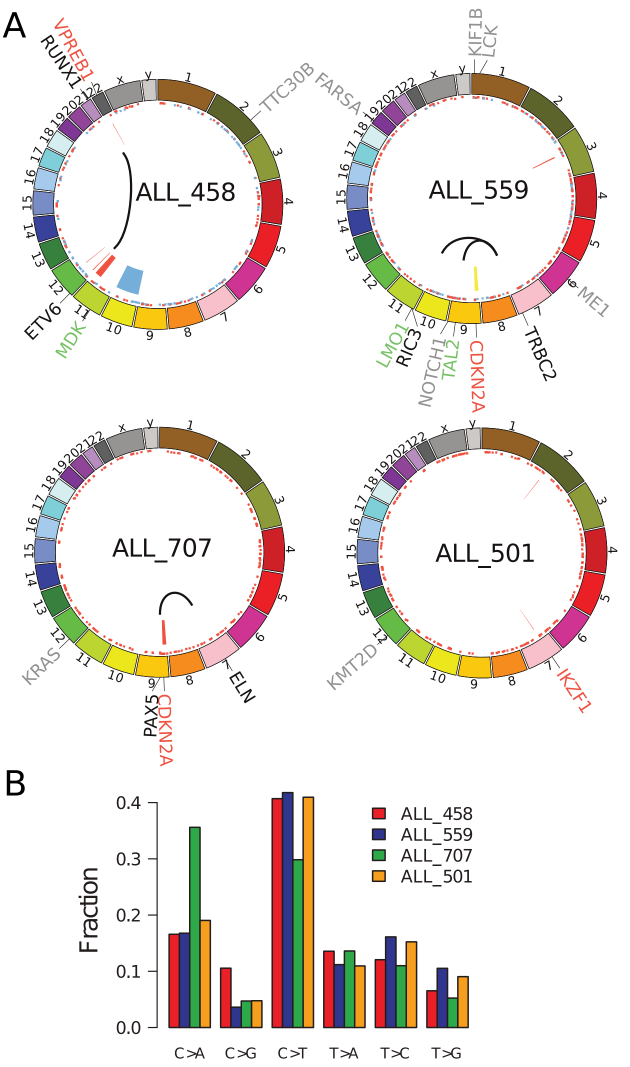

Figure 1. **A**: Circos [Krzywinski et al., 2009] plots showing the genomic location of validated somatic single nucleotide variants (SNVs), insertion-deletions (indels), copy number alterations (CNAs), copy neutral loss of heterozygosity (LOH) events, and translocations in the whole genome sequenced ALL patients. SNVs and indels are shown as red (original clone) or blue (subclone) dots in the circle closest to the chromosomes. Inside the SNVs and indels, deletions are shown with red, duplications with blue and LOH with yellow circle segments. Black arcs indicate translocations. Gene names are color-coded as follows: gray, expressed genes with exonic indels or nsSNVs that were predicted to be damaging and genes that were highlighted as putative drivers in the validation cohort; black, genes involved in translocations; red, selected genes in CNA or LOH regions; green, selected differentially expressed genes that are located near breakpoints for translocations or putatively regulatory SNVs. **B**: Mutational patterns in the whole genome sequenced patients. The higher frequency of C>A mutations compared to C>T mutations in ALL_707 is significantly different from the other patients (chi-square test, $P < 0.001$).

Allele Fractions of Somatic Variants Reveal Clonal Heterogeneity

As we have shown before, target capture followed by deep sequencing allows accurate estimation of the AF of somatic SNVs and enables detection of clonal heterogeneity [Berglund et al., 2013]. Analysis of the validated somatic SNVs revealed a large density peak with AF 0.41–0.46 in each of the four whole genome sequenced

Table 3. Validated Exonic Single Nucleotide Variants and Insertion–Deletions in Whole Genome Sequenced ALL Patients

Sample	Chr	Position	cDNA change[a]	Protein change	Gene	Gene description	Effect	SIFT[b]	PP2[c]	AF DNA[d]	AF RNA	FPKM sample	FPKM control[e]
ALL_458	2	178416648	c.844G>A	p.E282K	TTC30B	Tetratricopeptide repeat domain 30B	nsSNV	D	D	0.12	0.14	2.2	0.8
ALL_458	3	51749797	c.2008C>T	p.R670W	GRM2	Glutamate receptor, metabotropic 2 isoform a	nsSNV	D	D	0.44	NA	0.1	0.1
ALL_458	12	11285961	c.883C>T	p.R295W	TAS2R30	Type 2 taste receptor member 30	nsSNV	T	B	0.33	0.00	0.2	1.6
ALL_559	1	10363472	c.2229T>G	p.I743M	KIF1B	Kinesin family member 1B isoform b	nsSNV	D	D	0.34	0.20	2.5	3.3
ALL_559	1	11186751	c.1069C>T	p.R357C	MTOR	FK506 binding protein 12-rapamycin associated	nsSNV	T	B	0.56	0.60	12.3	10.8
ALL_559	1	32745467	c.1067A>C	p.E356A	LCK	Lymphocyte-specific protein tyrosine kinase	nsSNV	D	D	0.61	0.64	306.8	67.2
ALL_559	3	48617469	c.5119C>T	p.R1707W	COL7A1	Alpha 1 type VII collagen precursor	nsSNV	D	P	0.47	1.00	0.0	0.3
ALL_559	5	180053029	c.1261C>T	p.P421S	FLT4	fms-related tyrosine kinase 4 isoform 2	nsSNV	D	P	0.35	NA	0.0	1.0
ALL_559	6	39864708	c.2462G>A	p.R821Q	DAAM2	Dishevelled associated activator of	nsSNV	T	P	0.40	0.00	0.1	1.6
ALL_559	6	84055977	c.515G>T	p.G172V	ME1	Cytosolic malic enzyme 1	nsSNV	D	D	0.39	0.00	0.1	0.0
ALL_559	9	139397768	c.5033T>C	p.L1678P	NOTCH1	Notch1 preproprotein	nsSNV	D	D	0.31	0.29	9.5	7.8
ALL_559	15	41165511	c.456C>T	p.N152N	RHOV	ras homolog gene family, member V	sSNV	NA	NA	0.22	NA	0.1	0.1
ALL_559	17	4536237	c.1459G>A	p.V487M	ALOX15	Arachidonate 15-lipoxygenase	nsSNV	D	P	0.45	NA	0.0	0.0
ALL_559	19	13041102	c.438_439insG	p.G146fs	FARSA	Phenylalanyl-tRNA synthetase, alpha subunit	frameshift ins	NA	NA	0.62	0.36	8.8	9.5
ALL_707	12	25398281	c.38G>A	p.G13D	KRAS	c-K-ras2 protein isoform a precursor	nsSNV	D	NA	0.38	0.25	14.5	10.9
ALL_707	20	10030811	c.1594G>T	p.A532S	ANKEF1	Ankyrin repeat and EF-hand domain containing 1	nsSNV	D	D	0.39	NA	0.1	0.2
ALL_707	21	42609582	c.544G>T	p.E182X	BACE2	Beta-site APP-cleaving enzyme 2 isoform A	nonsense SNV	NA	NA	0.41	0.00	NA	NA
ALL_501	X	151092993	c.857A>G	p.K286R	MAGEA4	Melanoma antigen family A, 4	nsSNV	T	P	0.32	NA	0.0	0.0
ALL_501	X	48382171	c.12C>T	p.N4N	EBP	Emopamil binding protein (sterol isomerase)	sSNV	NA	NA	0.46	0.00	30.4	16.0
ALL_501	1	157659674	c.1724C>T	p.A575V	FCRL3	Fc receptor-like 3 precursor	nsSNV	T	B	0.31	0.00	0.6	0.7
ALL_501	2	61711139	c.2610G>C	p.Q870H	XPO1	exportin 1	nsSNV	T	B	0.46	0.48	115.1	107.1
ALL_501	11	112123109	c.410C>T	p.A137V	PLET1	Hypothetical protein LOC349633 precursor	nsSNV	T	B	0.38	0.00	0.2	0.2
ALL_501	12	49426841	c.11647_11648 insGCTC	p.H3883fs	KMT2D	Lysine (K)-specific methyltransferase 2D	frameshift ins	NA	NA	0.60	0.32	5.2	15.4

[a] Nucleotide numbering uses +1 as the A of the ATG translation initiation codon in the reference sequence, with the initiation codon as codon 1.
[b] SIFT predictions: D, damaging; T, tolerated.
[c] PolyPhen2 (PP2) predictions: D, probably damaging; P, possibly damaging; B, benign.
[d] AF from deep-sequencing data. The AF for the SNV in *TAS2R30* is from WGS data, as this SNV was not covered in the deep-sequencing data.
[e] The shown expression value represents the mean of 27 BCP-ALL (for ALL_458, ALL_707 and ALL_501) or 18 T-ALL (for ALL_559) samples.
Chr, chromosome; nsSNV, nonsynonymous SNV; sSNV, synonymous SNV; ins, insertion; AF, allele fraction; FPKM: fragments per kilobase of transcript per million mapped reads.

Table 4. Somatic Copy Number Alterations in Whole Genome Sequenced ALL Patients

Sample	Subtype	Chr	Start	End	Size (kb)[a]	Type[b]	Cytoband	Cytogenetic prediction[c]	Affected genes	FPKM sample[d]	FPKM control[e]
ALL_458	t(12;21)	10	1	135534747	135,535	Dup*	all chr 10	NA	691 genes	21.9	13.3
ALL_458	t(12;21)	11	96600814	134944379	38,344	Del*	q21-q25	NA	261 genes	11	10.6
ALL_458	t(12;21)	12	11253986	13357791	2,104	Del	p13.2-p13.1	p13	25 genes including ETV6	3.1	6.5
ALL_458	t(12;21)	12	68835364	69203100	368	Del	q15	NA	MDM2, NUP107, RAP1B, SLC35E3	85.4	87.6
ALL_458	t(12;21)	22	22569501	22600200	31	Del	q11.22	NA	VPREB1	39.3	72.3
ALL_559	T-ALL	3	70233601	74494601	4,261	Del	p13-p12.3	NA	11 genes including FOXP1, RYBP, SHQ1	6	9.8
ALL_559	T-ALL	9	46587	21295942	21,249	LOH	p24.3-p21.3	NA	74 genes	46.5	28.5
ALL_559	T-ALL	9	21300000	22100401	800	Del	p21.3	p21×2	11 genes including CDKN2A, CDKN2B	0.1	0.9
ALL_559	T-ALL	9	22110997	32243981	10,133	LOH	p21.3-p21.1	NA	13 genes	1.5	1.5
ALL_707	Other	9	197421	21980801	21,783	Del	p24.3-p21.3	p21-p24	82 genes	25.7	19.9
ALL_707	Other	9	21980802	22021001	40	Del	p21.3	p21-p24	CDKN2A, CDKN2B	0	1.2
ALL_707	Other	9	22021002	36908001	14,887	Del	p21.3-p13.2	p21-p24	92 genes	28.3	34.9
ALL_707	Other	19	58519186	59089786	571	Del	q13.43	q13	23 genes	18.9	16.8
ALL_501	Normal	2	89165816	89554016	388	Del	p11.2	NA	NA	NA	NA
ALL_501	Normal	7	50412894	50463634	51	Del	p12.2	NA	IKZF1	84.6	71.1

[a] The size represents the minimal overlap between different predictions. The breakpoints of the *IKZF1* deletion were determined by analysis of softclipped reads in the WGS data.
[b] A * indicates that the CNA is subclonal.
[c] Results of cytogenetic analysis at diagnosis. NA indicates that no aberration was observed in the region by cytogenetic analysis.
[d] Mean expression of genes located within the CNA in the sample harboring the CNA.
[e] Mean expression of genes located within the CNA in the control data set, where each gene is represented by the mean of 27 BCP-ALL (for ALL_458, ALL_707 and ALL_501) or 18 T-ALL (for ALL_559) samples.
Chr, chromosome; Dup, duplication; Del, deletion; LOH, loss of heterozygosity; FPKM, fragments per kilobase of transcript per million mapped reads; CNA, copy number alteration.

patients (Fig. 2A), in agreement with the estimation that the samples contained 80%–95% leukemic blasts. In addition, density peaks indicative of subclones with AF of 0.19 and 0.27, were observed in ALL_458 (t(12;21)) and ALL_559 (T-ALL), respectively. Based on the AFs, we estimate that approximately half of the SNVs in ALL_458 belong to the subclone, and that 40% of the leukemic cells carry these SNVs. For comparison, an SNV present in the original clone is expected to be present in all leukemic cells, including those that belong to the subclone. The subclone in ALL_559 contains fewer (30%) of the SNVs, but they are present in a larger proportion (60%) of the leukemic cells. Comparison of the mutational patterns of the SNVs that were present in the original clones and those arising in

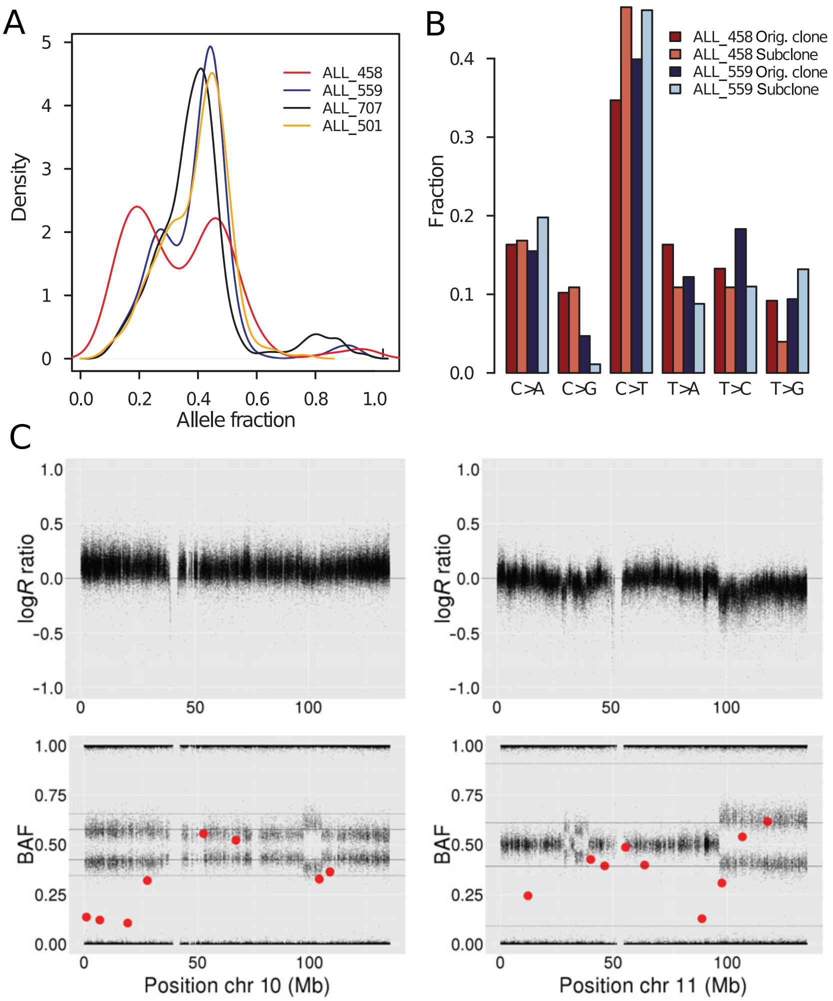

Figure 2. **A**: Density plot showing the allele fraction (AF) distribution of validated somatic single nucleotide variants (SNVs) in the four whole genome sequenced ALL patients. Each sample displays a density peak with AF between 0.41 and 0.46. In addition, ALL_458 and ALL_559 display density peaks with AF of 0.19 and 0.27, respectively, indicative of subclones. **B**: Mutational patterns of SNVs belonging to the original clone and the subclone for ALL_458 and ALL_559. **C**: Subclonal copy number alterations (CNAs) in ALL_458 visualized using Omni2.5 BeadChip data. The top panel shows the log *R* ratio (LRR) and the bottom panel shows the B-allele frequency (BAF) for the duplication of chromosome 10 (left) and the deletion of chromosome 11q21–25 (right). R corresponds to the total intensity of each probe, and LRR is the $\log_2$ of the ratio of the measured normalized *R*-value in the ALL sample and the normalized *R*-value of the reference. LRR = 0 indicates no change in copy number. BAF represents the AF, with values of 0 and 1 indicating homozygosity and 0.5 indicating heterozygosity in a diploid genomic region. The solid and dashed lines correspond to the estimated BAF if the CNA is present in 40% and 100% of the leukemic cells, respectively. The red dots show the genomic location of somatic SNVs, with the position on the *y*-axis corresponding to the AF.

the subclones showed no major differences, although the subclones contained a larger proportion of C>T substitutions than the original clone in both patients (Fig. 2B). ALL_707 and ALL_501 did not display additional density peaks, however, both patients harbored a substantial number of SNVs with relatively low AF, which could indicate the presence of minor clones with few SNVs. It should also be noted that because of the limited sequence depth in the WGS data, only subclones present in a relatively large proportion of the cells would have been detected.

Analysis of the AF of SNPs from genotyping data suggested that the duplication of chromosome 10 and the deletion of chromosome 11q21–25 in ALL_458 were subclonal (Fig. 2C). Interestingly, the somatic SNVs on chromosome 10 fell into three AF clusters, which provide clues to when and where these mutations occurred. The first cluster (n = 3 SNVs) showed low AF (0.10–0.13), suggesting that these SNVs occurred in the subclone, either before duplication on the non-duplicated allele or after duplication on any allele. The second cluster of SNVs (n = 3, AF 0.32–0.36) likely occurred on the non-duplicated allele in the original clone. The third set of SNVs (n = 2) had high AF (0.52–0.55), and probably occurred in the original clone, on the allele that was subsequently duplicated in the subclone. The SNVs in the deleted region on chromosome 11 also have varying AFs, however, the low number of SNVs (n = 3) hinders inference of their origin.

Expression of Genes Affected by Somatic Variants

Using RNA-seq data, we found that 10 of the 23 genes with validated somatic exonic variants were expressed at ≥ 1 fragment per kilobase of transcript per million mapped reads (FPKM) (Table 3). The variant allele was expressed in 9/10 genes, but never overexpressed in comparison to the AF in DNA (Table 3). To assess the putative effect of the somatic variants on gene expression, we compared to immunophenotype-matched control RNA-seq datasets, consisting of 27 BCP-ALL samples or 18 T-ALL samples (Supp. Table S2). In comparison with the control dataset, *LCK* (MIM# 153390) was 4.6-fold overexpressed and *KMT2D* (MIM# 602113, previously named *MLL2*) was threefold underexpressed in the samples harboring the variants (Table 3). Of the 14 genes with damaging or loss of function mutations, 50% ($n = 7$) were expressed, in comparison with 33% (3/9) of the genes with predicted benign variants. We also observed at least 1.6-fold decreased expression of the genes in three of the nine hemizygous deletions that affected at least one gene, almost complete loss of expression in the two homozygous deletions, and 1.6-fold increased expression of the genes on the duplicated chromosome 10 (Table 4). The expression of the genes on chromosome 9p24–21 with copy-neutral LOH was 1.6-fold increased in ALL_559 compared with the control dataset, whereas there was no difference in expression of the genes in the other LOH region on the same chromosome (Table 4).

Next, we analyzed putative effects of non-coding SNVs on the expression of nearby genes. A total of 125 SNVs that were located in a conserved or DHS region, had a RegulomeDB score ≤ 4, or a score from FunSeq [Khurana et al., 2013] ≥ 3, were included in the analysis. In order to call a gene differentially expressed, we required a fourfold relative difference and an absolute difference of at least 3 standard deviations between the sample and the control. Using these criteria, we identified 31 genes that were differentially expressed in comparison to the control dataset, all of which were overexpressed, corresponding to 19 putatively regulatory SNVs (Supp. Table S9). These 19 SNVs were located in introns (n = 8), intergenic regions (n = 8) or the 5′ flank of a gene (n = 3). Three of the intronic SNVs were covered by at least two reads in the RNA-seq data, putatively representing unspliced mRNAs (Supp. Table S9). To further investigate the putative function of these variants, we determined the overlap between SNVs and known histone marks in lymphoid cells (Supp. Table S9). Eight of the 19 putative regulatory SNVs, associated with differential expression of 12 genes, overlapped with histone marks. Notably, an SNV located 53 kb upstream of the growth factor *MDK* (MIM# 162096), which was overexpressed in ALL_458 (t(12;21)) overlapped with markers for enhancer elements (H3K27ac and H3K4me1 modifications) and active transcription (H3K4me3 modification).

We also used the RNA-seq data to identify expressed fusion genes (Supp. Table S10). In addition to the canonical *ETV6-RUNX1* fusion in ALL_458, which was detected by RT-PCR at diagnosis, the RNA-seq data revealed expression of the reciprocal fusion gene *RUNX1-ETV6* that contains the first exon of *RUNX1* and the last three exons of *ETV6*. Cytogenetic analysis of ALL_559 (T-ALL) detected the two translocations t(7;9) and t(7;11). RNA-seq demonstrated that both translocations result in expressed fusion genes with a common 3′ partner *TRBC2* (MIM# 615445), which is part of the T-cell receptor beta locus [7q34]. The 5′ partners were *RIC3* (MIM# 610509) [11p15] and a non-annotated gene located 500 bp upstream of *TMEM38B* (MIM# 611236) [9q31]. We observed overexpression of *LMO1* (MIM# 186921) and *TUB* (MIM# 601197), which flank *RIC3,* and *TAL2* (MIM# 186855), which is located downstream of *TMEM38B,* in ALL_559 compared with other T-ALL samples (Fig. 3). ALL_707 also had a t(7;9) based on karyotype data. RNA-seq demonstrated that this fusion resulted in a highly expressed *PAX5-ELN* (MIM#s 167414, 130161) fusion gene (Fig. 3). We did not detect any expressed fusion genes apart from those resulting from the translocations detected at diagnosis in the three patients mentioned above and no fusions were detected in ALL_501 with normal karyotype.

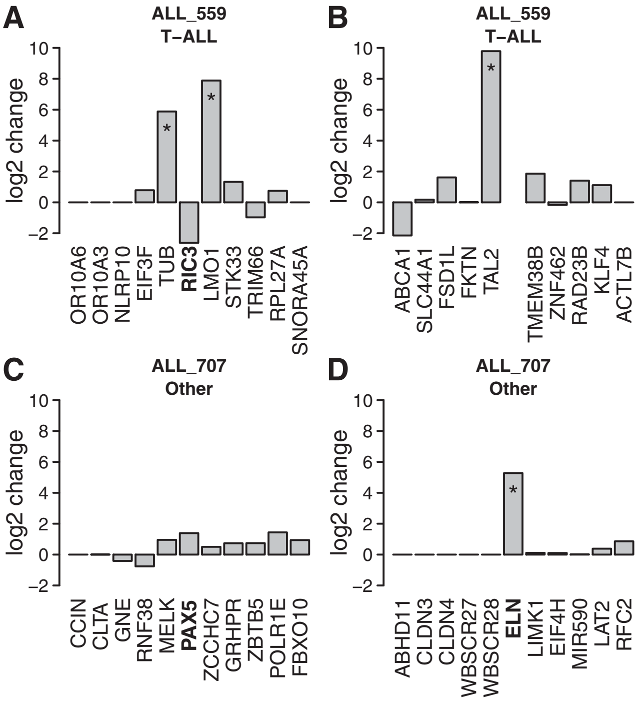

Figure 3. Bar plots showing the expression of five genes up- and downstream of genes involved in chromosomal rearrangements. The gene involved in the fusion, if annotated, is highlighted in bold. Genes exhibiting a fourfold relative difference and an absolute difference of at least 3 standard deviations between the sample and the control are marked with a *. **A and B:** Expression changes in ALL_559 associated with the fusions of *TRBC2* with *RIC3* (**A**) and a non-annotated gene close to *TMEM38B* (**B**). *LMO1*, *TUB*, and *TAL2* are overexpressed in ALL_559 compared with 18 T-ALL samples. **C and D:** Expression changes in ALL_707 associated with the *PAX5-ELN* fusion. *ELN*, which constitutes the major part of the fusion gene, is overexpressed in ALL_707 compared with 27 BCP-ALL samples suggesting that the *PAX5-ELN* fusion gene is highly expressed.

Recurrently Mutated Genes and Regions Identified by Targeted Sequencing

To ascertain if any of the genes or putative regulatory regions identified by WGS of four patients were recurrently mutated in ALL, we performed target capture and deep sequencing of these regions in a cohort of 145 BCP-ALL and 23 T-ALL samples (Table 2 and Supp. Tables S1 and S4). Matched DNA from remission was sequenced in pools for 159 of the patients, and analysis of the sequence data demonstrated that 139 of the remission samples were well represented in the pools and could be used to filter out germline variants (Supp. Methods). In these 139 patients, we detected on average 0.7 SNVs, whereas in the remaining 29 patients, we detected on average 3.0 SNVs. This result suggests that the majority of the SNVs called in the samples without matched remission sample are germline variants. To avoid false positives we only report variants included in the COSMIC database for these samples.

We detected 107 SNVs (Supp. Table S11) and 15 indels (Supp. Table S12) in the validation cohort, including 43 SNVs and 10 indels in exons or UTRs of the 30 resequenced genes (Fig. 4). The most frequently mutated genes were *KRAS* (18 SNVs affecting 16 patients) and *NOTCH1* (9 SNVs and 7 indels, affecting 12 patients). 17/18 *KRAS* mutations were found in BCP-ALL patients, and all *NOTCH1* mutations were found in T-ALL patients. In addition to

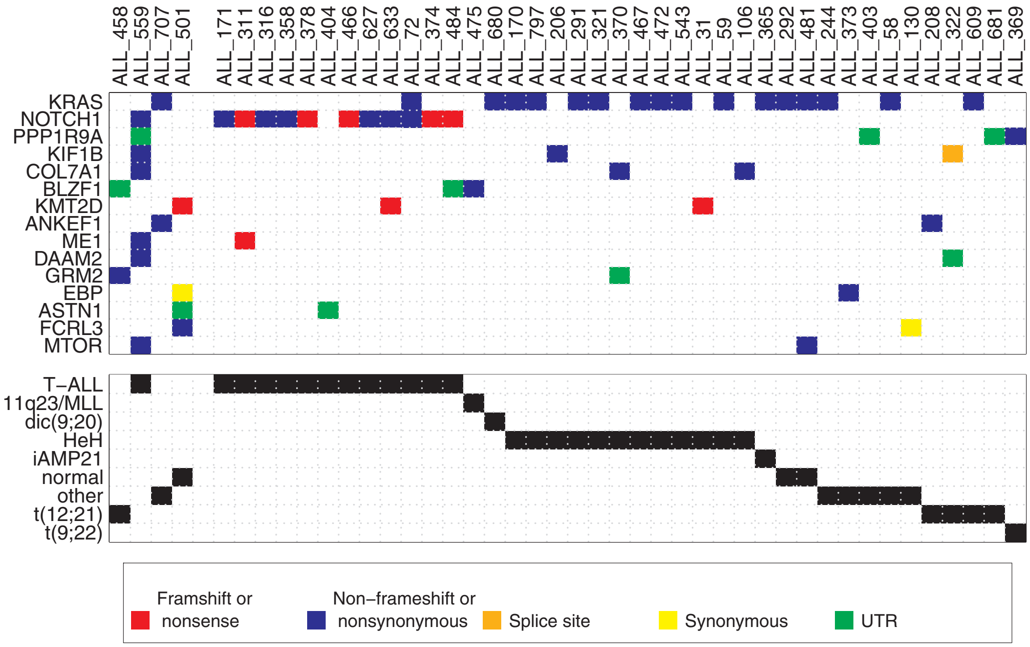

Figure 4. Recurrent somatic mutations detected in the validation cohort in the genes that were identified by WGS. Each column represents one patient, with the whole genome sequenced samples in the four leftmost columns. In the upper panel, each row represents one gene. Only samples and genes with at least one mutation in an exon, splice site, or untranslated region (UTR) in the validation cohort are shown. Each colored box indicates a mutation. For patients with more than one variant in the same gene, the color is prioritized according to the order shown in the legend. In the lower panel, the genetic subtype of each sample is shown.

KRAS and *NOTCH1*, we identified *KMT2D*, *KIF1B* (MIM# 605995) and *ME1* (MIM# 154250) as novel putative driver genes using three complementary tools (MutSigCV, Lawrence et al. (2013) Oncodrive-fm, Gonzalez-Perez, Lopez-Bigas (2012) and OncodriveCLUST, Tamborero et al. (2013)). To find subtype-specific patterns, we analyzed the T-ALL samples and the largest BCP-ALL subtypes (HeH, t(12;21), normal and other) individually. *KRAS* was highlighted as driver in all BCP-ALL subtypes except t(12;21) and *NOTCH1* was highlighted in T-ALL. No subtype-specific pattern was observed for the novel genes.

Recurrent mutations in UTRs were identified in several genes (Fig. 4), however, the mutations were co-located in either the 3′ or the 5′ UTR only for *ASTN1* (MIM# 600904), which is not expressed in ALL according to our RNA-seq dataset. For identification of recurrently mutated non-coding regions in the validation cohort, we defined "super-regions" consisting of all non-coding regions that were selected because of the same original SNV (Supp. Methods). In most cases, the super-regions consisted of a contiguous genomic region, however, in cases where the SNV was located in a conserved region, they contained all conserved regions within 2 kb. Twelve non-coding super-regions were found to harbor recurrent mutations in the validation cohort (Supp. Table S13). The original SNVs detected by WGS were located in a conserved (n = 4) or DHS region (n = 1), had a RegulomeDB hit (n = 1), were located in a DHS region and had a RegulomeDB hit (n = 1), or lacked functional annotation (n = 5). Two SNVs overlapped histone marks for enhancer elements (H3K27ac and H3K4me1) and active transcription (H3K4me3). One specific SNV (chr4:157006979C>T) was identified in two t(12;21) patients, including the whole genome sequenced ALL_458. This SNV did not overlap any of the annotated regions or histone marks, and the possible functional implications are unclear.

Discussion

In this study, we performed a thorough characterization of the genomes of four representative pediatric ALL patients using WGS and RNA-seq. We validated and determined the AFs of somatic variants genome-wide and identified recurrently mutated coding and non-coding regions by targeted sequencing of 168 additional ALL patients.

In ALL_458 (t(12;21)), we found deletions of *ETV6* and *VPREB1*, in line with previous observations that patients with the *ETV6-RUNX1* translocation often harbor deletions that target genes involved in B-cell development [Mullighan et al., 2007; Papaemmanuil et al., 2014]. We also observed expression of the rarely reported reciprocal *RUNX1-ETV6* fusion gene, which has been suggested to be involved in cellular regrowth [Stams et al., 2005; Al-Shehhi et al., 2013], and overexpression of *MDK* coinciding with a putative regulatory SNV located in an enhancer element and DHS region 53 kb upstream of the gene. *MDK* is involved in cancer development and has previously been shown to be upregulated in BCP-ALL compared to normal peripheral blood and bone marrow [Hidaka et al., 2007]. ALL_458 also contained a subclone with a large number of SNVs, suggesting that these cells have an increased mutation rate and/or a growth advantage during leukemic progression compared to the cells in the original clone. Clonal heterogeneity has previously been observed in ALL by copy number profiling [Jan and Majeti, 2013] or analysis of AFs from exome sequencing [Papaemmanuil et al., 2014] and is one of the most important challenges for the successful application of targeted therapies [Landau et al., 2014]. Despite the potentially rapidly growing subclone and the presence of two lesions that have been suggested to have a negative impact on clinical outcome, namely the deletion of *VPREB* [Mangum et al., 2014] and the expression of *RUNX1-ETV6* [Stams et al., 2005], this patient responded well to treatment and has remained in CCR1. The absence of SNVs or indels in putative driver genes in ALL_458, the few *KRAS* mutations in t(12;21) patients, and the previous failure to identify recurrently mutated genes by exome sequencing of t(12;21) patients [Lilljebjorn et al., 2012] suggest that point mutation in exons of protein-coding genes is not a dominant force for leukemic development in this subtype.

We identified several lesions that are characteristic of T-ALL in ALL_559, including a *NOTCH1* mutation, two translocations involving the T-cell receptor beta locus [Le Noir et al., 2012] that resulted in overexpression of *LMO1* [Atak et al., 2013] and *TAL2* [Marculescu et al., 2003], and deletion of the tumor suppressor *CDKN2A* [Mullighan et al., 2007]. ALL_559 also displayed mutation and overexpression of the proto-oncogene *LC*K, which is involved in T-cell development. *LCK* has previously been found to be mutated and overexpressed in fusions with the T-cell receptor region in T-ALL and it was suggested that oncogenic transformation of *LCK* requires two mutations, one that deregulates gene transcription and one that activates protein function [Wright et al., 1994]. Although we did not identify any regulatory SNV that could cause the aberrant expression of *LCK* in ALL_559, it is possible that there is a regulatory SNV that is located at a larger distance than 1 Mb from *LCK* or that another type of genetic lesion is involved. Findings in our study that are novel in T-ALL include deletion of the putative tumor suppressors *FOXP1, RYBP* and *SHQ1*, which has frequently been observed in prostate cancer and has been suggested to exert a tumor-promoting effect [Krohn et al., 2013], the novel fusion partner for *TRBC2* in a non-annotated gene on 9q31, and the identification of *ME1* and *KIF1B* as putative driver genes. Although *ME1* is not expressed in any of the T-ALL samples in our RNA-seq dataset and probably does not play a major role in leukemogenesis, *KIF1B* is expressed and has been suggested to be a tumor suppressor that contributes to cancer development by dosage reduction [Henrich et al., 2012].

Recurrent lesions in ALL_707 (subtype “other”) included the *KRAS* mutation [Liang et al., 2006], the *CDKN2A* deletion [Mullighan et al., 2007], and the *PAX5-ELN* fusion gene [Bousquet et al., 2007]. We did not identify any novel putative driver events in this patient, however, we observed a large number of C>A mutations, in contrast to the other patients where the dominant mutation was C>T. Although C>T mutations are common in most cancer types, and can be caused by UV-light or deamination of 5-methylcytosines [Alexandrov et al., 2013], excessive C>A mutations have mainly been observed in lung cancer, and have been attributed to tobacco exposure [Pleasance et al., 2010]. Future studies will reveal whether this mutational signature, which has not previously been observed in ALL, will be detected in other ALL patients.

ALL_501 was selected for this study because its karyotype was completely normal, and we were especially interested in finding the driver events in this patient. In agreement with the cytogenetic results, we found no fusion gene or large CNA. We detected a focal deletion of exons 3–6 in *IKZF1*, which encodes the transcription factor Ikaros that plays key roles in lymphoid development and tumor suppression. The resulting transcript, which lacks four zinc fingers that are required for DNA binding and therefore is unable to bind transcriptional targets, is known as Ik6 and acts as a dominant negative inhibitor of Ikaros function [Mullighan and Downing, 2008]. The second finding in ALL_501 was a frameshift insertion and reduced expression of *KMT2D* (*MLL2*), which was highlighted as a putative driver gene in the validation cohort. Frequent loss-of-function mutations in *KMT2D*, which encodes a histone methyltransferase involved in regulation of gene transcription, have been found in a range of cancers, and this gene has been proposed to be a tumor suppressor and a putative therapeutic target [Guo et al., 2013]. An intriguing question is whether these two mutations are sufficient to induce leukemia. Mouse models have shown that expression of Ik6 can induce T-cell leukemia [Winandy et al., 1995], however, expression exclusively in B-cells does not result in B-lineage leukemia [Wojcik et al., 2007]. Point mutations in *KMT2D* have, to our knowledge, not been reported as putative drivers before in ALL, and their role in leukemogenesis is yet to be determined.

In summary, we provide a high-resolution map of the genomes of four representative pediatric ALL patients. We found that each patient had a unique genome, with a combination of well known and previously undetected genomic aberrations, including SNVs, CNAs, and chromosomal rearrangements. Despite the limited size of the discovery cohort we identified *KMT2D* and *KIF1B* as novel putative driver genes in ALL, which suggests that analysis of more samples would enable identification of additional genes. The non-annotated fusion partner to *TRBC2* is an example of a novel finding enabled by RNA-seq. Our finding that overexpression of *MDK* coincide with a non-coding putative regulatory SNV suggests that regulatory variants may be more important for the development of ALL and other cancers than recognized to date, and that future WGS and RNA-seq studies in larger cohorts combined with functional experiments will be useful to explore this area. The results from our study as well as earlier sequencing studies [Roberts et al., 2012; Zhang et al., 2012; Holmfeldt et al., 2013; Papaemmanuil et al., 2014] contribute to an increased understanding of the biological mechanisms that lead to ALL. The heterogeneity of the genetic aberrations in ALL, and lack of large numbers of recurrent mutations renders WGS particularly well suited for diagnosis and stratification of ALL patients into subgroups for new treatment protocols. Analysis of serially collected samples from the 20% of patients that relapse has already revealed mechanisms of clonal evolution that provide clues to the cause of treatment failure [Meyer et al., 2013; Tzoneva et al., 2013], and more extensive studies are likely to further increase the understanding of the biology behind relapse in ALL. Next generation sequencing technology has developed fast in recent years, and today it is a reality to apply WGS at costs and speed that are acceptable for routine clinical genetic diagnostics of ALL.

Acknowledgments

Illumina sequencing and SNP genotyping was performed at the SNP&SEQ Technology Platform in Uppsala, which is part of the National Genomics Infrastructure (NGI) funded by the Swedish Council for Research Infrastructures and Science for Life Laboratory. Sanger sequencing was performed at the Uppsala Genome Centre. Computational analysis was performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We thank Anna Kiialainen, Christina Leek, Anna Karin Lannergård, and Katja Pokrovskaja for assistance with sample procurement, our colleagues from NOPHO and the ALL patients who contributed samples to this study.

Disclosure Statement: The authors declare no conflict of interest.

References

Al-Shehhi H, Konn ZJ, Schwab CJ, Erhorn A, Barber KE, Wright SL, Gabriel AS, Harrison CJ, Moorman AV. 2013. Abnormalities of the der(12)t(12;21) in ETV6-RUNX1 acute lymphoblastic leukemia. Genes Chromosomes Cancer 52:202–213.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, et al. 2013. Signatures of mutational processes in human cancer. Nature 500:415–421.

Atak ZK, Gianfelici V, Hulselmans G, De Keersmaecker K, Devasia AG, Geerdens E, Mentens N, Chiaretti S, Durinck K, Uyttebroeck A, Vandenberghe P, Wlodarska I, et al. 2013. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. PLoS Genet 9:e1003997.

Berglund EC, Lindqvist CM, Hayat S, Overnas E, Henriksson N, Nordlund J, Wahlberg P, Forestier E, Lonnerholm G, Syvanen AC. 2013. Accurate detection of subclonal single nucleotide variants in whole genome amplified and pooled cancer samples using HaloPlex target enrichment. BMC Genomics 14:856.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28:1045–1048.

Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics 27:268–269.

Bousquet M, Broccardo C, Quelen C, Meggetto F, Kuhlein E, Delsol G, Dastugue N, Brousset P. 2007. A novel PAX5-ELN fusion protein identified in B-cell acute lymphoblastic leukemia acts as a dominant negative on wild-type PAX5. Blood 109:3417–3423.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498.

Gonzalez-Perez A, Lopez-Bigas N. 2012. Functional impact bias reveals cancer drivers. Nucleic Acids Res 40:e169.

Guo C, Chen LH, Huang Y, Chang CC, Wang P, Pirozzi CJ, Qin X, Bao X, Greer PK, McLendon RE, Yan H, Keir ST, et al. 2013. KMT2D maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation. Oncotarget 4:2144–2153.

Henrich KO, Schwab M, Westermann F. 2012. 1p36 tumor suppression–a matter of dosage? Cancer Res 72:6079–6088.

Hidaka H, Yagasaki H, Takahashi Y, Hama A, Nishio N, Tanaka M, Yoshida N, Villalobos IB, Wang Y, Xu Y, Horibe K, Chen S, et al. 2007. Increased midkine gene expression in childhood B-precursor acute lymphoblastic leukemia. Leuk Res 31:1045–1051.

Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, Payne-Turner D, Churchman M, Andersson A, Chen SC, McCastlain K, Becksfort J, et al. 2013. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet 45:242–252.

Jan M, Majeti R. 2013. Clonal evolution of acute leukemia genomes. Oncogene 32:135–140.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. Genome Res 12:656–664.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, et al. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342:1235587.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315.

Krohn A, Seidel A, Burkhardt L, Bachmann F, Mader M, Grupp K, Eichenauer T, Becker A, Adam M, Graefen M, Huland H, Kurtz S, et al. 2013. Recurrent deletion of 3p13 targets multiple tumour suppressor genes and defines a distinct subgroup of aggressive ERG fusion-positive prostate cancers. J Pathol 231:130–141.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645.

Landau DA, Carter SL, Getz G, Wu CJ. 2014. Clonal evolution in hematological malignancies and therapeutic implications. Leukemia 28:34–43.

Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28:311–317.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218.

Le Noir S, Ben Abdelali R, Lelorch M, Bergeron J, Sungalee S, Payet-Bornet D, Villarese P, Petit A, Callens C, Lhermitte L, Baranger L, Radford-Weiss I, et al. 2012. Extensive molecular mapping of TCRalpha/delta- and TCRbeta-involved chromosomal translocations reveals distinct mechanisms of oncogene activation in T-ALL. Blood 120:3298–3309.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Liang DC, Shih LY, Fu JF, Li HY, Wang HI, Hung IJ, Yang CP, Jaing TH, Chen SH, Liu HC. 2006. K-Ras mutations and N-Ras mutations in childhood acute leukemias with or without mixed-lineage leukemia gene rearrangements. Cancer 106:950–956.

Lilljebjorn H, Rissler M, Lassen C, Heldrup J, Behrendtz M, Mitelman F, Johansson B, Fioretos T. 2012. Whole-exome sequencing of pediatric acute lymphoblastic leukemia. Leukemia 26:1602–1607.

Lilljebjorn H, Soneson C, Andersson A, Heldrup J, Behrendtz M, Kawamata N, Ogawa S, Koeffler HP, Mitelman F, Johansson B, Fontes M, Fioretos T. 2010. The correlation pattern of acquired copy number changes in 164 ETV6/RUNX1-positive childhood acute lymphoblastic leukemias. Hum Mol Genet 19:3150–3158.

Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. Science 333:1019–1024.

Mangum DS, Downie J, Mason CC, Jahromi MS, Joshi D, Rodic V, Muschen M, Meeker N, Trede N, Frazer JK, Zhou Y, Cheng C, et al. 2014. VPREB1 deletions occur independent of lambda light chain rearrangement in childhood acute lymphoblastic leukemia. Leukemia 28:216–220.

Marculescu R, Vanura K, Le T, Simon P, Jager U, Nadel B. 2003. Distinct t(7;9)(q34;q32) breakpoints in healthy individuals and individuals with T-ALL. Nat Genet 33:342–344.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing. EMBnet.journal 17:10–12.

Mayrhofer M, Dilorenzo S, Isaksson A. 2013. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol 14:R24.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069–2070.

Meyer JA, Wang J, Hogan LE, Yang JJ, Dandekar S, Patel JP, Tang Z, Zumbo P, Li S, Zavadil J, Levine RL, Cardozo T, et al. 2013. Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat Genet 45:290–294.

Mullighan C, Downing J. 2008. Ikaros and acute leukemia. Leuk Lymphoma 49:847–849.

Mullighan CG. 2013. Genomic characterization of childhood acute lymphoblastic leukemia. Semin Hematol 50:314–324.

Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB, Su X, Pui CH, et al. 2007. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature 446:758–764.

Nordlund J, Backlin CL, Wahlberg P, Busche S, Berglund EC, Eloranta ML, Flaegstad T, Forestier E, Frost BM, Harila-Saari A, Heyman M, Jonsson OG, et al. 2013. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biol 14:r105.

Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, Alexandrov LB, Van Loo P, Cooke SL, Marshall J, Martincorena I, Hinton J, et al. 2014. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat Genet 46:116–125.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, et al. 2010. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463:184–190.

Pui CH, Carroll WL, Meshinchi S, Arceci RJ. 2011. Biology, risk stratification, and therapy of pediatric acute leukemias: an update. J Clin Oncol 29:551–565.

Roberts KG, Morin RD, Zhang J, Hirst M, Zhao Y, Su X, Chen SC, Payne-Turner D, Churchman ML, Harvey RC, Chen X, Kasap C, et al. 2012. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell 22:153–166.

Schmiegelow K, Forestier E, Hellebostad M, Heyman M, Kristinsson J, Soderhall S, Taskinen M. 2010. Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. Leukemia 24:345–354.

Shaffer LG, McGowan-Jordan J, Schmid M, editors. 2013. *ISCN (2013): An international system for human cytogenetic nomenclature*. Basel: S. Karger.

Stams WA, den Boer ML, Beverloo HB, Meijerink JP, van Wering ER, Janka-Schaub GE, Pieters R. 2005. Expression levels of TEL, AML1, and the fusion products TEL-AML1 and AML1-TEL versus drug sensitivity and clinical outcome in t(12;21)-positive pediatric acute lymphoblastic leukemia. Clin Cancer Res 11:2974–2980.

Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29:2238–2244.

The Cancer Genome Atlas Research Network. 2013. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med 368:2059–2074.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578.

Tzoneva G, Perez-Garcia A, Carpenter Z, Khiabanian H, Tosello V, Allegretta M, Paietta E, Racevskis J, Rowe JM, Tallman MS, Paganin M, Basso G, et al. 2013. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat Med 19:368–371.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35(Web Server issue):W71–W74.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Borresen-Dale AL, et al. 2010. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci USA 107:16910–1695.

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

Weng AP, Ferrando AA, Lee W, Morris JPt, Silverman LB, Sanchez-Irizarry C, Blacklow SC, Look AT, Aster JC. 2004. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. Science 306:269–271.

Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, Melo M, da Rocha AG, et al. 2013. Frequency of TERT promoter mutations in human cancers. Nat Commun 4:2185.

Winandy S, Wu P, Georgopoulos K. 1995. A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. Cell 83:289–299.

Wojcik H, Griffiths E, Staggs S, Hagman J, Winandy S. 2007. Expression of a non-DNA-binding Ikaros isoform exclusively in B cells leads to autoimmunity but not leukemogenesis. Eur J Immunol 37:1022–1032.

Wright DD, Sefton BM, Kamps MP. 1994. Oncogenic activation of the Lck protein accompanies translocation of the LCK gene in the human HSB2 T-cell leukemia. Mol Cell Biol 14:2429–2437.

Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, Kucherlapati R, Park PJ. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci USA 108:E1128–E1136.

Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, Easton J, Chen X, Wang J, Rusch M, Lu C, Chen SC, et al. 2012. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature 481:157–163.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137.