A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia

Andrea Grioni[1,2], Grazia Fazio[1], Silvia Rigamonti[1], Vojtech Bystry[2], Giulia Daniele[3], Zuzana Dostalova[2], Manuel Quadri[1], Claudia Saitta[1,6], Daniela Silvestri[7,8], Simona Songia[1], Clelia T. Storlazzi[3], Andrea Biondi[1,5], Nikos Darzentas[2,4], Giovanni Cazzaniga[1]

Correspondence: Giovanni Cazzaniga (e-mail: gianni.cazzaniga@hsgerardo.org).

Abstract

Acute lymphoblastic leukemia (ALL) is the most frequent pediatric cancer. Fusion genes are hallmarks of ALL, and they are used as biomarkers for risk stratification as well as targets for precision medicine. Hence, clinical diagnostics pursues broad and comprehensive strategies for accurate discovery of fusion genes. Currently, the gold standard methodologies for fusion gene detection are fluorescence in situ hybridization and polymerase chain reaction; these, however, lack sensitivity for the identification of new fusion genes and breakpoints. In this study, we implemented a simple operating procedure (OP) for detecting fusion genes. The OP employs RNA CaptureSeq, a versatile and effortless next-generation sequencing assay, and an in-house as well as a purpose-built bioinformatics pipeline for the subsequent data analysis. The OP was evaluated on a cohort of 89 B-cell precursor ALL (BCP-ALL) pediatric samples annotated as negative for fusion genes by the standard techniques. The OP confirmed 51 samples as negative for fusion genes, and, more importantly, it identified known (*KMT2A* rearrangements) as well as new fusion events (*JAK2* rearrangements) in the remaining 38 investigated samples, of which 16 fusion genes had prognostic significance. Herein, we describe the OP and its deployment into routine ALL diagnostics, which will allow substantial improvements in both patient risk stratification and precision medicine.

Introduction

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer.[1] The 5-year survival rate exceeds 85% in children, but the survival following relapse is poor.[2] Analysis of paired diagnosis/relapse ALL samples shows clonal diversity that arises from the accumulation of new deletions and mutations over time. Despite that, the founding fusion genes are usually conserved from diagnosis to relapse, indicating that the predominant clones observed at diagnosis and relapse are clones derived from a common 'preleukemic' clone.[3] Fusion genes arise from chromosomal translocations and intrachromosomal

This work was partially supported by the program 'Passaporto genetico' of the Parents Committee 'Comitato Maria Letizia Verga ONLUS', as well as by the IG17593 and IG20564 Grants of the Associazione Italiana per la Ricerca sul Cancro (AIRC) to GC and AB, respectively; the IG2014 grant no. 15413 to CTS; the fellowship program Brno PhD talent 2017 as well as fellowship AIRC 2018 no. 22620 to AG; by the DIMET PhD program - University of Milano-Bicocca to CS. Bioinformatics of CEITEC Masaryk University is gratefully acknowledged for the obtaining of the scientific data presented in this paper

AG, ND and VB contributed new analytic tools, and participated in the data analysis and writing of the paper. AB, AG, GF and GC participated in research design and writing of the paper. GF, SR, GD and CTS participated in the performance of the research. CS and MQ performed PCR validations. DS provided samples information. ZD participated in the writing and correction of the paper.

The authors declare no conflicts of interest.

[1] *Centro Ricerca Tettamanti, Clinica Pediatrica, Università degli Studi di Milano-Bicocca, Fondazione MBBM/Ospedale S. Gerardo, Monza, Italy*
[2] *Central European Institute of Technology, Masaryk University, Brno, Czech Republic*
[3] *Department of Biology, University of Bari "Aldo Moro", Bari, Italy*
[4] *Department of Hematology, University Hospital Schleswig-Holstein, Kiel, Germany*
[5] *Clinica Pediatrica, Università degli Studi di Milano-Bicocca, Fondazione MBBM/Ospedale S. Gerardo, Monza, Italy*
[6] *Cancer Center, Humanitas Research Hospital, Humanitas University, Rozzano, Milan, Italy*
[7] *Center of Biostatistics for Clinical Epidemiology, Department of Health Science, University of Milano-Bicocca, Milan, Italy*
[8] *Pediatric Hematology-Oncology Unit, Department of Pediatrics, University of Milano-Bicocca, MBBM Foundation/ASST Monza, Monza, Italy.*

HemaSphere (2019) 3:3(e250)

Received: 27 December 2018 / Received in final form: 4 April 2019 / Accepted: 4 April 2019

Citation: Grioni A, Fazio G, Rigamonti S, Bystry V, Daniele G, Dostalova Z, Quadri M, Saitta C, Silvestri D, Songia S, Storlazzi CT, Biondi A, Darzentas N, Cazzaniga G. A Simple RNA Target Capture NGS Strategy for Fusion Genes Assessment in the Diagnostics of Pediatric B-cell Acute Lymphoblastic Leukemia. *HemaSphere*, 2019;3:3. http://dx.doi.org/10.1097/HS9.0000000000000250

rearrangements that mainly disrupt genetic regulators of normal hematopoiesis as well as lymphoid development (e.g., those involving *RUNX1* and *ETV6*) and constitutively activate tyrosine kinases[4] (e.g., *ABL1* chimeras). Thus, fusion genes are hallmarks of ALL that play a pivotal role in leukemogenesis, and their identification is crucial for patient risk stratification.[5]

Common fusion genes in B-lineage ALL are: t(12;21)(p13;q22), encoding ETV6-RUNX1 (TEL-AML); t(1;19)(q23;p13), encoding TCF3-PBX1 (E2A-PBX1)[6]; t(9;22)(q34;q11.2), resulting in formation of the "Philadelphia" chromosome, encoding BCR-ABL1; rearrangements of *KMT2A* (*MLL*) at 11q23 to a range of fusion partners[7]; and rearrangements of the cytokine receptor gene *CRLF2* at the pseudo autosomal region 1 (PAR1) at Xp22.3/Yp11.3.[8,9] Fusion genes correlate with the clinical outcome, and they are used as biomarkers for patient risk stratification[10]: for example, patients positive for t(12;21)/ETV6-RUNX1 have the most favorable prognosis, whereas t(9;22)/BCR-ABL1, t(1;19)/TCF3-PBX1, and KMT2A-AFF1 correlate with a brief disease latency and have a poor prognosis.[10,11] Moreover, specific drug inhibitors antagonizing the fusion proteins provide a more efficient and less toxic tool for disease eradication (precision medicine): for example, the imatinib tyrosine kinase inhibitor inhibits the oncogenic deregulation caused by the (9;22)/BCR-ABL1 fusion protein.[12]

Before the next generation sequencing (NGS) era, elaborate and extensive cytogenetic studies lead to the description of few recurrent and highly expressed fusion genes,[13] such as BCR-ABL1 and ETV6-RUNX1. The characterization of their breakpoint coordinates enabled the design of diagnostic screening by both quantitative multiplex polymerase chain reaction (qPCR) and fluorescence in situ hybridization (FISH).[14] The recent introduction of NGS allowed a fast and accurate screening of the patient's genome at the nucleotide level, which lead to the discovery of a broad array of previously unknown fusion genes.[15] This reflects the increased capability of NGS to recognize subtle chromosomal rearrangements. On the contrary, FISH may only detect exchanges of considerably larger chromosome segments, without nucleotide precision, while qPCR screenings can identify already known fusion gene breakpoints only.[16]

Whole transcriptome sequencing (RNAseq), together with open-source bioinformatics tools, has already been applied to identifying fusion genes.[17] Whole RNAseq performs well in the detection and quantification of highly and medium abundant transcripts, but it may fail in cases of low abundance transcripts.[18] The RNA capture sequencing (RNA CaptureSeq) is a probe-based assay for capturing, amplifying, and sequencing genomic regions of interest only (targets). The RNA CaptureSeq generates libraries of small fragments (250–300 bp) in a short time (2.5 days) compared to whole RNAseq, and it is compatible with the well-known MiSeq and NextSeq Illumina NGS platforms. RNA CaptureSeq is sensitive to low abundance transcript variants of targeted genes[19]; however, the detection of fusion transcripts may be compromised when the fusion partner gene is not part of the capture procedure (unknown partner). This scenario reduces discoverability of fusion transcripts to only those fragments that span the target gene breakpoint.

We have developed and herein present a simple, efficient, and ready-to-use operating procedure (OP) for the clinical identification of fusion genes in B-cell ALL. The OP is based on RNA CaptureSeq, and it is supported by an in-house bioinformatics pipeline that is purpose-built to detect and extend fragments spanning the fusion gene breakpoint. We applied the OP to a cohort of 89 B-cell ALL pediatric patients enrolled in the AIEOP-BFM ALL clinical protocol[20] that were annotated as negative to fusion genes by the standard screening methods. This paper summarizes the results of the OP applied to clinical diagnostics and discusses its implications for patient risk stratification.

Results

Comparison of available bioinformatics pipelines

We developed a bioinformatic method for fusion gene assessment from RNA CaptureSeq datasets and evaluated it on a training dataset composed of 23 samples evaluated as positive to 6 different fusion genes, namely t(9;22)/BCR-ABL1, t(12;21)/ETV6-RUNX1, t(4;11)/KMT2A-AFF1, del(X)/P2RY8-CRLF2, t(1;19)/TCF3-PBX, and t(9;11)/KMT2A-MLLT3, by standard methods. Our method distinguished all 6 sample-specific fusion genes within the dataset. In addition, we analyzed the same training dataset through Illumina BaseSpace, STAR-Fusion,[21] and the customized pipeline described by Jennifer L. Winters et al.[22] The STAR-Fusion tool did not detected 1 out of 6 fusion genes (del(X)/P2RY8-CRLF2), while the Illumina BaseSpace did not detect 2 out of 6 fusion genes (t (9;11)/KMT2A-MLLT3 and t(4;11)/ KMT2A- AFF1). The method described by Jennifer L. Winters et al. did not detect 3 out of 6 fusion genes (t(1;19)/TCF3-PBX, t(9;11)/KMT2A-MLLT3, and del(X)/P2RY8-CRLF2) (Table 1).

The ability of our procedure to detect all fusion transcripts derives from the fine-tuning of the bioinformatics pipeline to cover the specific RNA target–capture scenario, where both genes involved in the fusion are not always captured (see Material and Methods and Fig. 1). For these reasons, we applied only our method in the subsequent analyses.

Evaluation of the OP in clinical diagnosis

RNA material obtained from patient bone marrow mononuclear cells at the onset or relapse of the disease was sequenced using the RNA PanCancer (Illumina, San Diego, CA). Raw FASTQ files underwent quality control and were afterwards analyzed through our system. A detailed description of the OP strategy is available in the Materials and Methods section. The time required for the procedure from library preparation to obtaining results was 2.5 days.

We screened a cohort of 89 samples of B-cell ALL leukemia (test set) for positivity to fusion genes. All samples were negative for the fusion genes t(12;21)/ETV6-RUNX1, t(9;22)/BCR-ABL1, t(4;11)/KMT2A-AFF1, and t(1;19)/TCF3-PBX1 by the standard screening methods. The test set was divided into 3 groups: frontline high-risk (HR), relapse (RL), and patients with a high value of minimal residual disease (MRD) at day 33 of chemotherapy induction (TP1 +). Overall, the OP identified 26 different fusion genes in 38 out of the 89 investigated samples, with the transcripts of 16 of them being of prognostic value (Table 2 and Suppl. Table 1, Supplemental Digital Content, http://links.lww.com/HS/A34). New fusion genes in B-cell ALL and not recorded in public databases were validated through reverse transcription PCR (RT-PCR) or FISH to discern between false and true positives (Supplementary Table 2, Supplemental Digital Content, http://links.lww.com/HS/A34).

OP applied to the frontline HR group

Seven out of 16 samples (43%) resulted as positive for fusion genes (Fig. 2a). Four samples carried fusion genes recurrently associated to B-cell ALL: t(5;5)/EBF1-PDGFRB (n = 2), t(9;9)/PAX5-JAK2 (n = 1), and t(12;19)/ZNF384-TCF3 (n = 1) and 3 samples were positive for t(19;19)/TCF3-OAZ1 (n = 1), t(7;7)/IKZF1-DDC (n = 1), t(2;9)/ZEB2-JAK2 (n = 1), and t(9;17)/MPRIP-JAK2 (n = 1)

Table 1

Comparison of available bioinformatics pipelines.

		Metadata					Bioinformatics Pipeline		
Sample	Blast%	Fusion gene	Raw-reads	FASTQC	Probes	Internal	BaseSpace	TopHat	Star-Fusion
KN1	90	t(9;22) BCR-ABL1	3.22E+06	+	t/p	+	+	+	+
KN2	90	t(12;21) ETV6-RUNX1	5.19E+06	+	t/p	+	+	ND	+
KN3	92	t(4;11) KMT2A-AFF1	5.60E+06	+	t/p	+	+	ND	+
KN4	90	t(9;22) BCR-ABL1	4.76E+06	+	t/p	+	+	ND	+
KN5	93	t(9;22) BCR-ABL1	6.06E+06	+	t/p	+	+	ND	+
KN5	93	t(12;21) ETV6-RUNX1	6.06E+06	+	t/p	+	+	ND	+
KN6	98	del(X) P2RY8-CRLF2	3.81E+06	+	t/p	+	+	ND	ND
KN7	NA	t(9;22) BCR-ABL1	2.41E+06	+	t/p	+	+	ND	+
KN8	NA	t(4;11) KMT2A-AFF1	2.57E+06	+	t/p	+	+	ND	+
KN9	91	t(1;19) TCF3-PBX	2.46E+06	+	t/p	+	+	ND	+
KN10	64	t(12;21) ETV6-RUNX1	2.30E+06	+	t/p	+	+	ND	+
KN11	NA	t(9;11) KMT2A-MLLT3	2.50E+06	+	t/p	+	+	ND	+
KN12	NA	t(9;22) BCR-ABL1	1.62E+06	+	t/p	+	+	+	+
KN13	NA	t(4;11) KMT2A-AFF1	2.40E+06	+	t/p	+	+	+	+
KN14	91	t(1;19) TCF3-PBX	6.53E+05	+	t/p	+	+	ND	+
KN15	64	t(12;21) ETV6-RUNX1	2.47E+06	+	t/p	+	+	ND	+
KN16	NA	t(9;11) KMT2A-MLLT3	6.21E+06	+	t/p	+	ND	ND	+
KN17	NA	t(9;22) BCR-ABL1	5.29E+06	+	t/p	+	+	ND	+
KN18	93	t(4;11) KMT2A-AFF1	3.17E+06	+	t/p	+	ND	ND	+
KN19	90	t(4;11) KMT2A-AFF1	6.56E+06	+	t/p	+	+	+	+
KN20	93	t(1;19) TCF3-PBX	6.73E+06	+	t/p	+	+	ND	+
KN21	94	t(4;11) KMT2A-AFF1	4.50E+06	+	t/p	+	+	ND	+
KN22	70	t(12;21) ETV6-RUNX1	4.66E+06	+	t/p	+	+	+	+
KN23	97	t(9;22) BCR-ABL1	5.44E+06	+	t/p	+	+	+	+

fusion genes. All fusion transcripts were confirmed by RT-PCR, while the novel fusion genes t(2;9)/ZEB2-JAK2 (n=1) and t(9;17)/MPRIP-JAK2 were validated through FISH (Suppl. Fig. 1, Supplemental Digital Content, http://links.lww.com/HS/A34).

OP applied to the TP1+ group

The OP identified fusion genes in 19 out of 49 samples (38.8%) (Fig. 2b). Nine samples were evaluated as positive for fusion genes that are frequent in B-cell ALL: t(17;19)/TCF3-HLF (n=2), del(X)/P2RY8-CRLF2 (n=3), t(5;5)/EBF1-PDGFRB (N=2), t(12;19)/ETV6-JAK3 (n=1), t(12;22)/ZNF384-EP300 (n=1). We also identified a novel inter-chromosomal rearrangement, t(9;20)/PAX5-C20orf112 (n=1), and a variety of intra-chromosomal fusion genes (n=9) that were already annotated in public databases, and we validated them by RT-PCR (Suppl. Table 1, Supplemental Digital Content, http://links.lww.com/HS/A34).

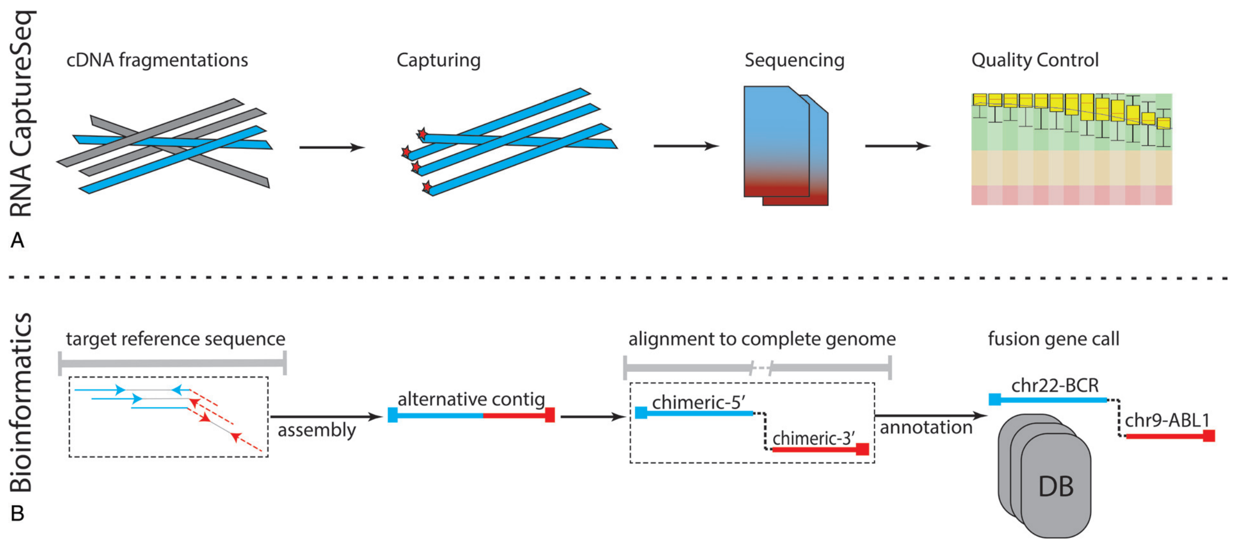

FIGURE 1. The standard operating procedure: (A) RNA CaptureSeq protocol allows the isolation of specific genomic regions (targets) through complementary probes; then, the captured fragments are sequenced, and the FASTQ file quality is evaluated. (B) The bioinformatics pipeline includes four sequential steps, which allows the identification of fusion genes through the identification of putative break-points on the genomic sequences of targeted genes.

Table 2

RNAseq Fusion transcripts identified by our OP.

fz	Fusion gene	Probes	Progn.	PCR	FusionHub
6	t(8;8) NDRG1-ST3GAL1	t	–	+	['CHIMERSEQ', 'Tumor_Fusion_GDP','HPA','Banned_dataset','Known_Fusions']
5	t(5;5) CAMK2A-CD74	t/p	–	+	['Known_Fusions']
5	del(X) P2RY8-CRLF2	t/p	+	+	['CHIMERPUB', 'FARE-CAFE', 'TICDB']
4	t(5;5) PDGFRB-EBF1	t/p	+	+	['CHIMERSEQ', 'CHITARS', 'Known_Fusions']
3	t(13;13) PSPC1-ZMYM2	p	–	+	['Banned_Dataset','GTEx']
3	t(19;19) DOT1L-OAZ1	t	–	+	['HPA', 'Banned_Dataset']
2	t(10;10) PTEN-RNLS	t	–	+	['Tumor_Fusion_GDP']
2	5(13;13) RB1-RCBTB2	t	–	+	['GTEx']
2	t(17;19) TCF3-HLF	t/p	+	+	['CHIMERKB', 'CHIMERPUB', 'FARE-CAFE','TICDB']
2	t(19;19) TCF3-OAZ1	t	–	+	NOVEL
2	t(5;5) ARHGAP26-NR3C1	t/p	–	+	['HPA', 'Banned_Dataset','GTEx']
1	t(10;11) MLLT10-KMT2A	t/p	+	+	['CHIMERKB', 'CHIMERPUB']
1	5(11;11) KMT2A-USP2	t/p	+	+	['Known_Fusions']
1	t(12;12) BCL7A-NCOR2	t/p	–	+	['Known_Fusions']
1	t(12;19) ETV6-JAK3	t/p	+	+	NOVEL
1	t(12;19) ZNF384-TCF3	t/p	+	+	['CHIMERSEQ', 'CHITARS', 'FARE-CAFE', 'TICDB', 'Known_Fusions']
1	t(12;22) ZNF384-EP300	t/p	+	+	['CHIMERPUB']
1	t(17;17) SUZ12P1-CRLF3	t	–	+	['18_Cancers']
1	t(9;17) MPRIP-JAK2	p	+	+	NOVEL
1	t(21;21) RUNX1-DYRK1A	t	+	+	['GTEx']
1	t(2;9) ZEB2-JAK2	p	+	+	NOVEL
1	t(3;9) MBNL1-PAX5	t/p	+	+	['Known_Fusions']
1	t(7;7) IKZF1-DDC	t	–	+	NOVEL
1	t(9;20) PAX5-C20orf112	t	+	+	['CHIMERSEQ', 'CHITARS', 'FARE-CAFE', 'TICDB']
1	t(9;9) NUP214-ABL1	t/p	+	+	['COSMIC','CHIMERAKB','CHIMERPUB','CHIMERSEQ', 'FARE-CAFE', ' TICDB','TUMOR_Fusion_GDP','Oesophagus_Dataset]
1	t(9;9) PAX5-JAK2	t/p	+	+	['COSMIC', 'CHIMERKB', 'FARE-CAFE', 'TICDB']

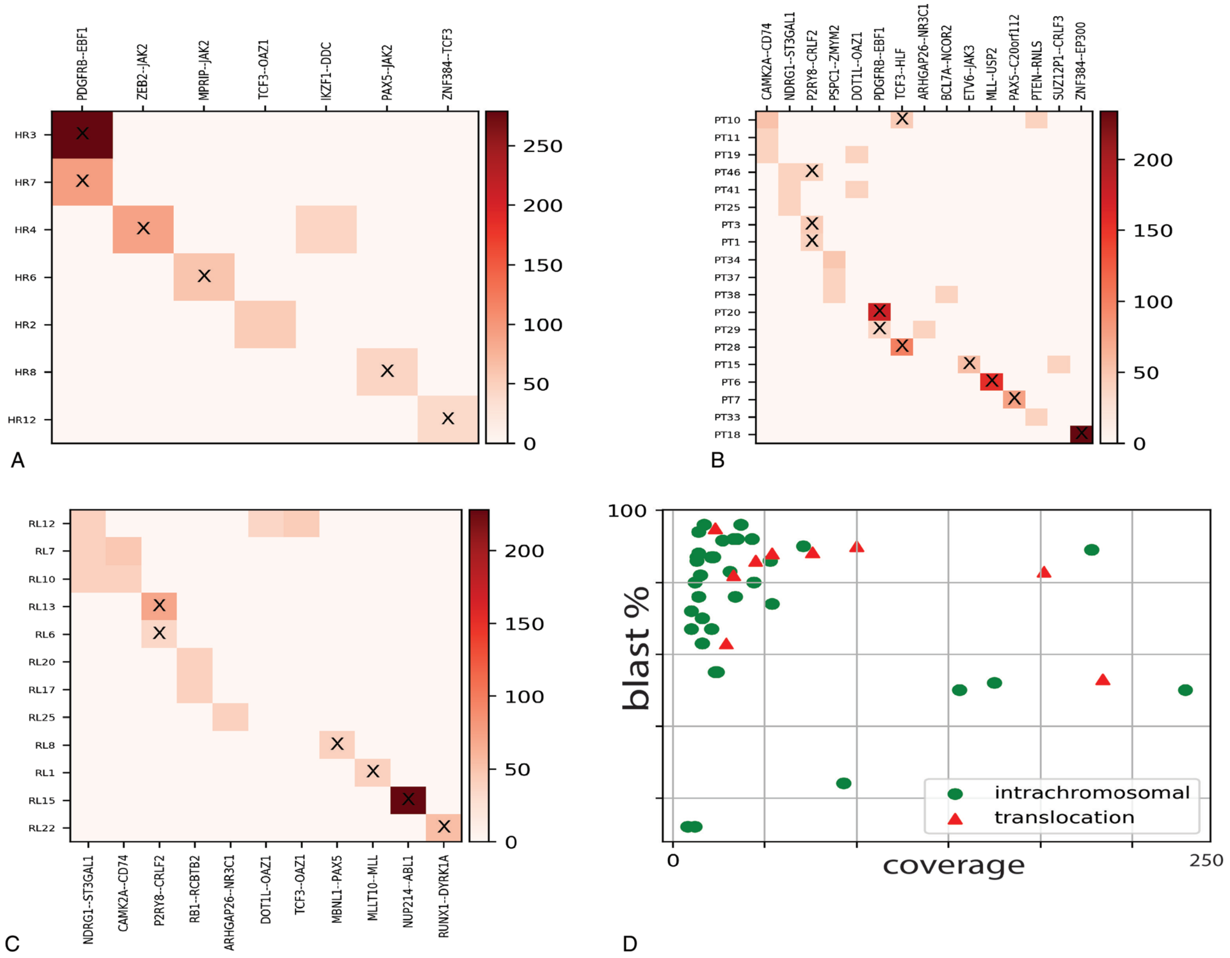

FIGURE 2. (A), (B), and (C) Heatmaps of detected fusion genes among different risk groups. The axes correspond to the detected fusion genes (X) and sample names (Y). The color code represents the coverage on the fusion gene breakpoint as reported by the scale on the right. The 'X' tag highlights fusion genes of prognostics relevance. (D) Fusion genes distribution in terms of intrachromosomal (green dots) or interchromosomal translocations (red triangles) in relations to the breakpoint read coverage and percentage of blast cells.

OP applied to the RL group

The OP identified fusion genes in 12 out of 24 samples of the RL group (~50%) (Fig. 2c): t(9;9)/NUP214-ABL1 (n=1), del(X)/P2RY8-CRLF2 (n=2), t(10;11)/MLLT10-KMT2A (n=1), t(21;21)/RUNX1-DYRK1A (n=1), and t(3;9)/PAX5-MBLN1 (n=1) fusion genes were associated with ALL and of clinical relevance for the patients and were hence immediately validated by RT-PCR. On the other hand, the OP identified additional fusion genes derived from intra-chromosomal rearrangements, such as t(8;8)/NDRG1-ST3GAL1 (n=3), t(13;13)/RB1-RCBTB2 (n=2), t(19;19)/DOT1L-OAZ1 (n=1), t(19;19)/TCF3-OAZ1 (n=1), t(5;5)/ARHGAP26-NR3C1 (n=1), and t(5;5)/CAMK2A-CD74 (n=2), which were already annotated in public databases.

Enrichment of intra-chromosomal fusion genes

The OP identified 26 fusion genes in 38 investigated patients (HR, RL, and TP1+ groups). Among them, 17 (65%) fusion genes derived from intra-chromosomal rearrangements and were supported by a low read coverage (~20× to ~50×) in coexistence with high levels of blast cells in the BM (~70% to ~96%) (Fig. 2d). We did not observe a correlation between intra-chromosomal fusion genes associated with recurrent chromosomal translocations in B-cell ALL (Table 3). RT-PCR confirmed frequent B-cell ALL intra-chromosomal fusion genes, such as PDGFRB-EBF1, NUP214-ABL1, and PAX5-JAK2 (Suppl. Table 2, Supplemental Digital Content, http://links.lww.com/HS/A34). P2RY8-CRLF2 fusions were not confirmed by RT-PCR since those samples correlated with del(X)(p22p22) detected by multiplex ligation-dependent probe amplification and highly expressed CRLF2 detected by gene expression profile (data not presented). We further investigated gene expression levels in healthy whole-blood samples for genes involved in intra-chromosomic fusions as well as those not known in B-cell ALL (n=21, gene set) through the GTEx portal.[23] Sixteen genes had transcript per million (TPM) expression levels from medium to high (TPM greater than 5.4), while 5 of them had low levels (TPM between 1 and 5.4) (Fig. 3). Also, some intra-chromosome fusion transcripts involved genes spatially close, within a range of

Table 3

Sample-specific fusion transcripts.

Sample	Fusion gene	Chromosome	% Leukemic cell in BM	Sex	Karyotype
HR2	TCF3-OAZ1	t(19;19)	98	F	
HR3	PDGFRB-EBF1	t(5;5)	60	M	
HR4	ZEB2-JAK2|IKZF1-DDC	t(2;9)|t(7;7)	NA	M	
HR6	MPRIP-JAK2	t(9;17)	NA	M	
HR7	PDGFRB-EBF1	t(5;5)	53	M	46,XY,der(1)inv(1)(q21q31)dup (1)(q31q32)[8]/46,XY[14]
HR8	PAX5-JAK2	t(9;9)	NA	M	
HR12	ZNF384-TCF3	t(12;19)	90	F	
PT1	P2RY8-CRLF2	del(X)	91	M	46,XY, der(9)T(9;?)(p13;?), -13, add (13)(q34), +21 [10]/47,XY,+21[4]
PT3	P2RY8-CRLF2	del(X)	95	M	
PT6	KMT2A-USP2	t(11;11)	NA	M	
PT7	PAX5-C20orf112	t(9;20)	NA	M	
PT10	TCF3-HLF|CAMK2A-CD74|PTEN-RNLS	t(17;19)|t(5;5)|t(10;10)	NA	F	
PT11	CAMK2A-CD74	t(5;5)	90	M	
PT15	ETV6-JAK3|SUZ12P1-CRLF3	t(12;19)|t(17;17)	90	F	
PT18	ZNF384-EP300	chr12-chr22	85	M	
PT19	CAMK2A-CD74|DOT1L-OAZ1	t(5;5)|t(19;19)	NA	M	
PT20	PDGFRB-EBF1	t(5;5)	80	F	
PT25	NDRG1-ST3GAL1	t(8;8)	NA	F	
PT28	TCF3-HLF	t(17;19)	95	F	
PT29	PDGFRB-EBF1|ARHGAP26-NR3C1	t(5;5)|t(5;5)	98	F	
PT33	PTEN-RNLS	t(10;10)	NA	M	
PT34	PSPC1-ZMYM2	t(13;13)	NA	F	
PT37	PSPC1-ZMYM2	t(13;13)	80	M	
PT38	BCL7A-NCOR2|PSPC1-ZMYM2	t(12;12)|t(13;13)	NA	F	
PT41	DOT1L-OAZ1|NDRG1-ST3GAL1	t(19;19)|t(8;8)	NA	M	
PT46	P2RY8-CRLF2|NDRG1-ST3GAL1	del(X)|t(8;8)	NA	F	
RL1	MLLT10-KMT2A	t(10;11)	90	M	
RL6	P2RY8-CRLF2	del(X)	76	M	
RL7	CAMK2A-CD74|NDRG1-ST3GAL1	t(5;5)|t(8;8)	70	M	
RL8	MBNL1-PAX5	t(3;9)	NA	M	
RL10	CAMK2A-CD74|NDRG1-ST3GAL1	t(5;5)|t(8;8)	NA	M	
RL12	NDRG1-ST3GAL1|TCF3-OAZ1|DOT1L-OAZ1	t(8;8)|t(19;19)|t(19;19)	97	M	
RL13	P2RY8-CRLF2	del(X)	98	F	47,XX,+21c[14]
RL15	NUP214-ABL1	t(9;9)	92	F	
RL17	RB1-RCBTB2	t(13;13)	40	M	
RL20	RB1-RCBTB2	t(13;13)	NA	M	
RL22	RUNX1-DYRK1A	t(21;21)	NA	M	
RL25	ARHGAP26-NR3C1	t(5;5)	99	F	

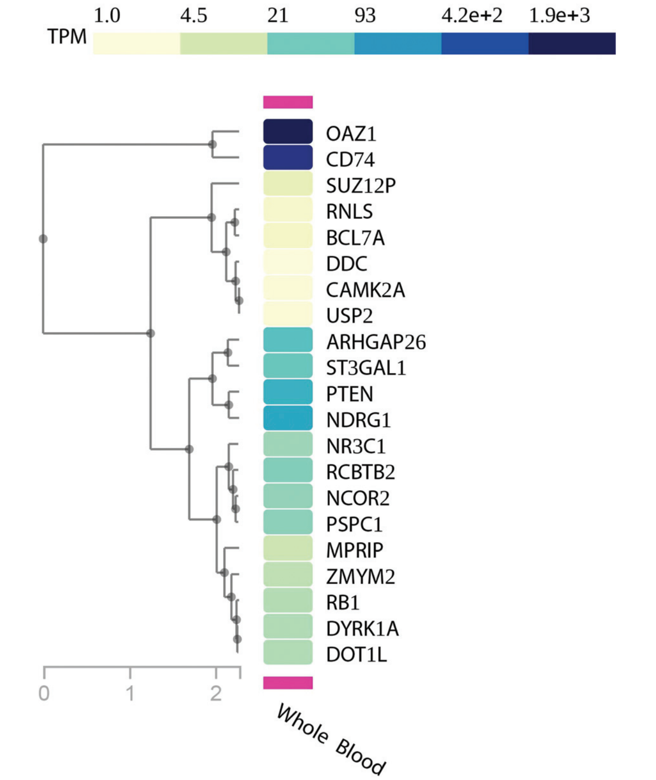

FIGURE 3. Gene expression profile of genes involved in intra-chromosomal fusion genes but not associated to ALL.

150 to 250 kb, and annotated as conjoined genes. Indeed, we validated those fusion gene events by RT-PCR and confirmed their nucleotide sequences by Sanger sequencing (Suppl. Table 2, Supplemental Digital Content, http://links.lww.com/HS/A34).

Discussion

Fusion genes are hallmarks of ALL both in pediatric and adult patients; their identification is crucial to design a risk-reducing-driven chemotherapy treatment (precision medicine). Precision medicine allows either very low-risk patients to proceed with standard therapy or very high-risk patients to be candidates for experimental and/or targeted therapies. For this purpose, sensitive, specific, and comprehensive screening of selected genomic regions prone to chromosomic breaks are needed in routine diagnostics to identify the increasing variety of fusion genes.

We built a versatile and straightforward OP to recognize fusion genes at nucleotide resolution without any a priori knowledge, which overcomes the limitations of qPCR and FISH. The OP employs an RNA CaptureSeq panel that allows targeted transcriptome sequencing through a simple library preparation protocol. For the subsequent data analysis, we fine-tuned a bioinformatics pipeline that deploys robust and stable tools, which can be easily set up on any operative system through the Anaconda Platform. Our bioinformatics pipeline recognized all fusion genes harbored by samples within the training dataset, while the Star-Fusion, Illumina BaseSpace, and the strategy proposed by Winter et al reached 83%, 66%, and 50% success in fusion transcripts identification, respectively. Prognostically significant and frequent B-cell precursor ALL fusion genes such as *KMT2A* rearrangements and P2RY8-CRLF2 were not fully detected by the external tools. Patients harboring *KMT2A* rearrangements have a particularly unfavorable prognosis.[10,24,25] *KMT2A* is prone to breaks in various genomic location with several partners, thus making the detection of its resulting fusion genes challenging. On the other hand, the repetitive nature of the chromosome X may compromise read alignment and the identification of the P2RY8-CRLF2 fusion gene. Our results indicated that our purpose-built, disease- and NGS-strategy specific bioinformatics pipeline is required for covering many possible scenarios causing fusion genes. The evaluation of the OP through the analysis of 89 pediatric B-cell precursor ALL samples identified 26 different fusion genes among 38 samples that were undetectable by the standard routine diagnostics. Sixteen of those fusion transcripts have prognostic value since they involved rearrangements in genes driving leukemogenesis (*KMT2A*, *JAK2*, and *PAX5*). Moreover, the newly identified fusion genes t(2;9)/ZEB2-JAK2 and t(9;17)/MPRIP-JAK2, which are possibly targetable by JAK/STAT inhibitors, highlight the potential of our OP for precision medicine and biomarker discovery. Additionally, we detected a case of NUP214/ABL1 fusion genes in B-cell ALL, which only 2 cases were previously reported.[26] We confirmed the increased capability provided by RNA CaptureSeq to detect small local structural variants through the identification of a variety of intra-chromosomal fusion genes (n = 17). Multiple intra-chromosomal fusion genes were the only detected in the sample within our set of genes (n = 1385); hence, it is not possible to state any functional correlation between those rearrangements and the recurrent fusion genes (such as BCR-ABL1, ETV6-RUNX1, and *KMT2A* rearrangements). Some intra-chromosomal fusion transcripts, namely PSPC1-ZMYM2, DOT1L-OAZ1, RB1-RCBTB2, ARHGAP26-NR3C1, were also observed in NGS studies[27,28,29] of healthy populations (e.g., GTEx, Banned_dataset, and HPA), or annotated as conjoined genes.[30,31] We also detected intra-chromosomal fusion transcripts involving recurrent leukemogenic genes (IKZF1-DDC, P2RY8-CRLF2, KMT2A-UPS2, MLLT10-KMT2A) that are prone to deletions and with a prognostic value (such as IKZF1,[32] and KMT2A[33]). Despite RNA CaptureSeq cannot discerns between inter- and intra- chromosome fusion genes when the same chromosomes are involved, these previous studies suggested an intra-chromosome origin.

In conclusion, herein we have described an NGS-based approach suitable for the detection of fusion genes, regardless of their expression levels, that may be incorporated into routine ALL diagnostics, with the advantage of a substantial improvement of precision medicine. Despite the OP lacks ISO certification, our finding highlights its potential and the need to develop bioinformatics tools addressing fusion genes detections from the RNA CaptureSeq scenario with precision. For this purpose, our OP may offer an idea for their implementation. Nonetheless, further studies are required to understand the biological significance and the potential therapeutic implication of the additional discoveries allowed by this tool.

Materials and methods

Patient cohort

A cohort of 89 B-cell precursor (BCP) ALL patients enrolled in the AIEOP-BFM ALL2009 protocol in Italy was sequenced by Illumina RNA CaptureSeq PanCancer to discern prognostic fusion genes. The cohort was composed of: 16 patients from the frontline HR group, with a level of MRD above 5×10^{-4} at day +78 (TP2), who were shown as fusion gene-negative during the screening; 49 patients TP1+, that is, with a high level of PCR-MRD ($>5 \times 10^{-4}$ compared to diagnostic value) at day +33 from the start of the induction therapy; and 24 patients from the RL (defined as having at least 5×10^{-2} blast cells after complete remission, CR). See Suppl. Table 3 (Supplemental Digital Content, http://links.lww.com/HS/A34).

Training dataset

A subgroup of 23 pediatric ALL patients enrolled in the AIEOP-BFM ALL2009 protocol, who were positive for fusion genes by standard clinical diagnosis, were selected. We used this subgroup as a training dataset for the development and evaluation of our bioinformatics pipeline of analysis for the assessment of fusion genes.

FISH analysis for validating the identified fusion genes

The experiments were performed on BM metaphases from archival methanol:acetic acid-fixed chromosome suspensions, as previously described.[17] Bacterial Artificial Chromosome (BAC) clones were opportunely selected according to the NGS data from the University of California Santa Cruz (UCSC) database (release of December 2013, GRCh38/hg38) and previously tested on normal human metaphases. Briefly, chromosome preparations from BM cells were hybridized in situ with 1 μg of each BAC probe labeled by nick translation. Hybridization was performed at 37°C in 2× saline–sodium citrate (SSC), 50% (vol/vol) formamide, 10% (w/vol) dextran sulfate, 5 μg Cot-1 DNA (Bethesda Research Laboratories, Gaithersburg, MD, USA), and 3 μg sonicated salmon sperm DNA in a volume of 10 μL. Post-hybridization washings were performed at 60°C in 0.1× SSC (3 times). In co-hybridization experiments, the probes were directly labeled with fluorescein, Cy3, and Cy5 or indirectly with biotin–dUTP and subsequently detected by 7-(diethylamino)coumarin-3-carboxylic acid *N*-succinimidyl ester-conjugated streptavidin. Chromosomes were identified by DAPI staining. Digital images were obtained using a Leica DMRXA epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments, Boston, MA). All fluorescence signals that were detected using specific filters were recorded separately as gray-scale images. Pseudo-coloring and merging of images were performed with Adobe Photoshop software.

Enrichment analysis

Ensembl gene IDs were extracted through the BioMart API (https://www.ensembl.org/biomart). Gene expression profile data from non-diseased samples were obtained from the GTEx portal through submission of the corresponding ENSEMBL gene ID (https://gtexportal.org/home/).

External tools for fusion gene assessment

The Illumina BaseSpace pipeline for the identification of fusion genes first aligns filtered FASTQ files to the reference human genome through the TopHat[34] (v. 2.1.0) or STAR[35] aligner (v. 2.5.0a). Then, the STAR aligner supports Manta-fusion and the TopHat aligner supports the TopHat-fusion[36] to identify candidate fusion genes. For the purpose of our analysis, we required the Illumina BaseSpace to recognize the sample-specific fusion gene by at least one application. The STAR-Fusion tool, v. 1.5.0, was utilized with standard parameters on the GRCh38.p12 genome reference and the corresponding Gencode[37] annotation set. We simulated the customized pipeline described by Jennifer L. Winters et al by deploying TopHat v. 2.1.1, which included TopHat-Fusion, and running the TopHat-Fusion pipeline with the Bowtie1[38] flag activated.

Operating procedure

The OP consists of a laboratory and a bioinformatics module that has been built to both maximize the efficiency and minimize the time of ALL clinical diagnostics. Each element of the laboratory module is fully customizable and commercially available, whereas each tool deployed for the bioinformatics module is freely available through the Anaconda Platform (https://www.anaconda.com/).

Laboratory module

RNA extraction protocol. Total RNA was extracted during diagnosis from bone marrow mononuclear cells by the guanidinium thiocyanate–phenol–chloroform method. Guanidine methods were used for total RNA preparation, as described by Sacchi et al.[39]

RNA CaptureSeq and sample sequencing. The RNA CaptureSeq 'TruSight RNA PanCancer' (Illumina), which includes 57,010 probes complementary to 21,043 coding regions for a total of 1385 cancer-related RNA transcripts, was applied (Fig. 1a). The protocol required 2.5 days, from library preparation to NGS sequencing. The sample libraries were prepared per the manufacturer's protocol using 10 ng of total RNA. Batches of 8 samples per run were sequenced through cartridge V3 on the Illumina MiSeq platform in a 75 bp paired-end setting for a total of 25 million paired-end reads (PE reads). The cost per sample was about 250 USD. A detailed list of targeted regions can be obtained from Illumina (https://support.illumina.com/sequencing/sequencing_kits/trusight-rna-pan-cancer-panel/downloads.html).

Bioinformatics module

FASTQ file quality control. The raw FASTQ quality control was performed using the FASTQC tool (https://www.bioinformatics.babraham.ac.uk/), which provided information on reads in terms of sequence duplication levels, per base and per sequence average quality score, sequence length distribution, and adapter content.

Fusion gene assessment. A purpose-built bioinformatics pipeline was developed to detect fusion genes from RNA CaptureSeq datasets. The pipeline deploys stable and open-source bioinformatics tools in a sequential mode (Fig. 1b):

- *Alignment to targets*. BWA-MEM[40] v. 0.7.15-r1140 aligned PE reads to the genomic sequences of the targeted genes. The PE reads that did not map entirely on the reference genome through SAMTOOLS[41] v. 1.8 were isolated; these PE reads (informative) may derive from fragments of the fusion gene breakpoint.
- *Assembly*. The informative reads are assembled into longer sequences (contigs) through the SPAdes[42] v. 3.12.0 tool.

SPAdes was run with 3 different settings of k-mer size (25, 31, and 51) to cover any possible contig scenarios, thus maximizing the sensitivity of our strategy. This step is critical since more extended sequences have a higher chance of correctly aligning on the fusion gene partner at the genomic level.

- *Alignment to the complete genome*. BWA-MEM aligned contig sequences to the complete human genome (GRCh38.p12). SAMTOOLS then retrieved contig sequences that showed chimeric features, thus mapping the 5′- and 3′-sides of different genomic locations.
- *Gene annotation and fusion gene assessment*. The chimeric sequences were annotated with BEDTOOLS[43] v. 2.27.0 and GENCODE[37] release 29 (GRCh38.p12) annotation. Any chimeric sequence with different gene annotation between the 5′- and 3′-side were termed fusion genes. These were queried to the web-application FusionHub[44] to highlight fusion genes already described in other studies.
- Description of public databases is provided by the FusionHub's authors (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5929557/table/pone.0196588.t001/?report=objectonly).

REFERENCES

1. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. *Lancet.* 2013;381:1943–1955. doi:10.1016/S0140-6736(12)62187-4.
2. Nguyen K, Devidas M, Cheng S-C, et al. Factors influencing survival after relapse from acute lymphoblastic leukemia: a Children's Oncology Group study. *Leukemia.* 2008;22:2142–2150. doi:10.1038/leu.2008.251.
3. Hunger SP, Mullighan CG. Redefining ALL classification: toward detecting high-risk ALL and implementing precision medicine. *Blood.* 2015;125:3977–3987. doi:10.1182/blood -2015- 02-580043.
4. Iacobucci I, Mullighan CG. Genetic basis of acute lymphoblastic leukemia. *J Clin Oncol.* 2017;35:975–983. doi:10.1200/JCO.2016.70.7836.
5. Harrison CJ. Cytogenetics of paediatric and adolescent acute lymphoblastic leukaemia. *Br J Haematol.* 2009;144:147–156. doi:10.1111/j.1365-2141.2008.07417.x.
6. Felice MS, Gallego MS, Alonso CN, et al. Prognostic impact of t(1;19)/TCF3-PBX1 in childhood acute lymphoblastic leukemia in the context of Berlin-Frankfurt-Münster-based protocols. *Leuk Lymphoma.* 2011;52:1215–1221. doi:10.3109/10428194.2011.565436.
7. Winters AC, Bernt KM. MLL-rearranged leukemias-an update on science and clinical approaches. *Front Pediatr.* 2017;5:4doi:10.3389/fped.2017.00004.
8. Harvey RC, Mullighan CG, Chen I-M, et al. Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia. *Blood.* 2010;115:5312–5321. doi:10.1182/blood -2009- 09-245944.
9. Russell LJ, Capasso M, Vater I, et al. Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia. *Blood.* 2009;114:2688–2698. doi:10.1182/blood-2009-03-208397.
10. Pui C-H, Robison LL, Look AT. Acute lymphoblastic leukaemia. *Lancet.* 2008;371:1030–1043. doi:10.1016/S0140-6736(08)60457-2.
11. Stam RW. MLL-AF4 driven leukemogenesis: what are we missing? *Cell Res.* 2012;22:948–949. doi:10.1038/cr.2012.16.
12. Iqbal N, Iqbal N. Imatinib: a breakthrough of targeted therapy in cancer. *Chemother Res Pract.* 2014;2014:357027doi:10.1155/2014/357027.
13. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst.* 1960;25:85–109. http://www.ncbi.nlm.nih.gov/pubmed/14427847. Accessed February 19, 2019.
14. Iijima-Yamashita Y, Matsuo H, Yamada M, et al. Multiplex fusion gene testing in pediatric acute myeloid leukemia. *Pediatr Int.* 2018;60:47–51. doi:10.1111/ped.13451.
15. Mertens F, Johansson B, Fioretos T, et al. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer.* 2015;15:371–381. doi:10.1038/nrc3947.
16. Bacher U, Shumilov E, Flach J, et al. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. *Blood Cancer J.* 2018;8:113doi:10.1038/s41408-018-0148-6.
17. Kumar S, Vo AD, Qin F, et al. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep.* 2016;6:21597doi:10.1038/srep21597.
18. Mercer TR, Clark MB, Crawford J, et al. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat Protoc.* 2014;9:989–1009. doi:10.1038/nprot.2014.058.
19. Clark MB, Mercer TR, Bussotti G, et al. Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nat Methods.* 2015;12:339–342. doi:10.1038/nmeth.3321.
20. Conter V, Bartram CR, Valsecchi MG, et al. Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. *Blood.* 2010;115:3206–3214. doi:10.1182/blood -2009- 10-248146.
21. Haas B, Dobin A, Stransky N, et al. STAR-fusion: fast and accurate fusion transcript detection from RNA-seq. *bioRxiv.* 2017;120295. doi: https://doi.org/10.1101/120295.
22. Winters JL, Davila JI, McDonald AM, et al. Development and verification of an RNA sequencing (RNA-Seq) assay for the detection of gene fusions in tumors. *J Mol Diagn.* 2018;20:495–511. doi:10.1016/J.JMOLDX.2018.03.007.
23. GTEx Consortium TGteThe genotype-tissue expression (GTEx) project. *Nat Genet.* 2013;45:580–585. doi:10.1038/ng.2653.
24. van der Linden MH, Valsecchi MG, De Lorenzo P, et al. Outcome of congenital acute lymphoblastic leukemia treated on the Interfant-99 protocol. *Blood.* 2009;114:3764–3768. doi:10.1182/blood -2009- 02-204214.
25. Pieters R, Schrappe M, De Lorenzo P, et al. A treatment protocol for infants younger than 1 year with acute lymphoblastic leukaemia (Interfant-99): an observational study and a multicentre randomised trial. *Lancet.* 2007;370:240–250. doi:10.1016/S0140-6736(07)61126-X.
26. Roberts KG, Morin RD, Zhang J, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell.* 2012;22:153–166. doi:10.1016/j.ccr.2012.06.005.
27. Puig-Oliveras A, Revilla M, Castelló A, et al. Expression-based GWAS identifies variants, gene interactions and key regulators affecting intramuscular fatty acid content and composition in porcine meat. *Sci Rep.* 2016;6:31803doi:10.1038/srep31803.
28. Babiceanu M, Qin F, Xie Z, et al. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res.* 2016;44:2859–2872. doi:10.1093/nar/gkw032.
29. Nicorici D, Şatalan M, Edgren H, et al. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv.* 2014;011650. doi:10.1101/011650.
30. Kim RN, Kim A, Choi S-H, et al. Novel mechanism of conjoined gene formation in the human genome. *Funct Integr Genomics.* 2012;12:45–61. doi:10.1007/s10142-011-0260-1.
31. Prakash T, Sharma VK, Adati N, et al. Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. *PLoS One.* 2010;5:e13284doi:10.1371/journal.pone.0013284.
32. Mullighan CG, Su X, Zhang J, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N Engl J Med.* 2009;360:470–480. doi:10.1056/NEJMoa0808253.
33. Sevov M, Bunikis I, Häggqvist S, et al. Targeted RNA sequencing assay efficiently identifies cryptic KMT2A (MLL)-fusions in acute leukemia patients. *Blood.* 2014;124: http://www.bloodjournal.org/content/124/21/2406?sso-checked=true. Accessed March 2, 2019.
34. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25:1105–1111. doi:10.1093/bioinformatics/btp120.
35. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21. doi:10.1093/bioinformatics/bts635.
36. Kim D, Salzberg SL. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 2011;12:R72doi:10.1186/gb -2011- 12-8-r72.
37. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* 2012;22:1760–1774. doi:10.1101/gr.135350.111.
38. Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10:R25doi:10.1186/gb-2009-10-3-r25.

39. Chomzynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal Biochem.* 1987;162:156–159. doi:10.1006/abio.1987.9999.
40. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2010;26:589–595. doi:10.1093/bioinformatics/btp698.
41. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078–2079. doi:10.1093/bioinformatics/btp352.
42. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455–477. doi:10.1089/cmb.2012.0021.
43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–842. doi:10.1093/bioinformatics/btq033.
44. Panigrahi P, Jere A, Anamika K. FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer. Kumar-Sinha C, ed. *PLoS One.* 2018;13:e0196588doi:10.1371/journal.pone.0196588.