Published in final edited form as:

Nat Genet. 2019 February ; 51(2): 296–307. doi:10.1038/s41588-018-0315-5.

PAX5-driven Subtypes of B-progenitor Acute Lymphoblastic Leukemia

A full list of authors and affiliations appears at the end of the article.

Abstract

Recent genomic studies have identified chromosomal rearrangements defining new subtypes of B-progenitor acute lymphoblastic leukemia (B-ALL), however many cases lack a known initiating genetic alteration. Using integrated genomic analysis of 1,988 childhood and adult cases, we describe a revised taxonomy of B-ALL, incorporating 23 subtypes defined by chromosomal rearrangements, sequence mutations, or heterogeneous genomic alterations, many of which show marked variation in prevalence according to age. Two subtypes have frequent alterations of the B lymphoid transcription factor gene *PAX5*. One, *PAX5alt* (7.4%), has diverse *PAX5* alterations (rearrangements, intragenic amplifications or mutations), and a second subtype is defined by *PAX5* p.Pro80Arg and biallelic *PAX5* alterations. We show that p.Pro80Arg impairs B lymphoid development and promotes the development of B-ALL with biallelic *Pax5* alteration in vivo. These results demonstrate the utility of transcriptome sequencing to classify B-ALL and reinforce the central role of *PAX5* as a checkpoint in B lymphoid maturation and leukemogenesis.

INTRODUCTION

B-cell acute lymphoblastic leukemia (B-ALL) is the most common pediatric malignancy¹, and consists of multiple subtypes with distinct constellations of inherited and somatic genetic alterations². Genomic analyses, especially transcriptome sequencing (RNA-seq), have identified recurrent chromosomal rearrangements that result in expression of chimeric fusion transcripts that define new subtypes of ALL^{3–12}. In contrast to subtypes characterized by aneuploidy or a single chromosomal rearrangement (e.g., *ETV6-RUNX1*, *BCR-ABL1* or *TCF3-PBX1*), rearrangements in these new subtypes are commonly not evident by conventional cytogenetic analysis (e.g., *DUX4*-rearranged ALL), or involve a diverse range of chromosomal rearrangements that converge on a single gene (e.g., *MEF2D*, *ZNF384*-rearranged ALL)^{6,7,10,11}. Additional cases have common transcriptional profiles but diverse

*Correspondence should be addressed to C.G.M. (charles.mullighan@stjude.org).

*These authors contributed equally.

AUTHOR CONTRIBUTIONS

Z.G. and C.G.M. designed the study, analyzed the data, and wrote the manuscript. M.L.C. performed experiments, analyzed the data and wrote the manuscript. K.G.R. performed sample preparation and data analysis. K.G.R., D.P. and C.-H.P. analyzed survival data. I.M., S.Pelletier, S.G., H.B., D.P.-T., A.H. and I.I. performed experiments. J.N. and J.D. provided in vitro modeling data. X.Z. developed the data portal webpage. K.H., L.S., S.Pounds, C.Q., S.N. and J.Z. analyzed genomic data. C.C., M.D. and Y.D. performed biostatistical analysis. S.R., J.G.-F., E.A.R., M.J.B., B.L.W., W.L.C., P.A.Z.-M., K.R.R., L.A.M., K.W.M., A.R., O.S., J.P.R., M.D.M., J.M.R., S.L., M.R.L., M.S.T., J.R., Y.Z., R.B., J.K., K.M., C.D.B., W.S., S.K., H.M.K., M.K., W.E., S.J., J.Y., E.P., J.D., M.V.R., M.L.L. and S.P.H. provided clinical samples and data.

COMPETING INTERESTS

The authors declare no competing financial interests.

genetic alterations (Phlike^{13,14} and *ETV6-RUNX1*-like ALL¹¹). This has refined the classification of B-ALL and identified new therapeutic targets, such as kinase inhibition in Ph-like ALL^{3–6}.

Despite these advances, many B-ALL cases cannot be categorized into any of the currently established subtypes. Such cases commonly relapse and lack targeted therapeutic approaches². Here we used integrated genomic analysis of a large B-ALL cohort to systematically define the nature, prevalence and prognostic significance of subtypes across the age spectrum.

RESULTS

Integrative genetic and genomic classification of B-ALL

We analyzed RNA-seq data from leukemic cells of 1,988 patients with B-ALL to identify chromosomal rearrangements, gene expression profiles and large-scale copy number alterations (CNA; Supplementary Tables 1–2). Whole genome (WGS; N=17) and exome (WES; N=73) sequencing, and single nucleotide polymorphism (SNP; N=1,141) array data were available for a subset of cases (Supplementary Tables 3–4). Gene expression profiles evaluated from RNA-seq data was analyzed using hierarchical clustering, t-distributed stochastic neighbor embedding (tSNE) analysis, and predictive modeling¹⁵ using cases of known subtypes (Fig. 1a and Supplementary Fig. 1–2). In conjunction with cytogenetic data, we classified the cohort into 23 subtypes (Table 1 and Supplementary Table 1). Twelve previously recognized subtypes accounted for 75.8% of the cohort (77.9% of children with standard risk (SR), 76.6% of children with high risk (HR), 71.6% of adolescent and young adult (AYA) and 76.2% of adults; age group definition is shown in Table 1). These include subtypes characterized by gross chromosomal alterations, including high hyperdiploidy (14.0%), low hypodiploidy (3.9%), near haploidy (1.5%), and intra-chromosomal amplification of chromosome 21 (iAMP21; 3.2% of 1,141 cases with SNP array data)¹⁶. Notably, the gene expression profile of near haploid B-ALL was similar to high hyperdiploid ALL, suggesting a common pathogenesis (Fig. 1a). Subtypes defined by rearrangements or gene expression profile include *BCR-ABL1* (Ph; 6.2%), Ph-like (18.1%), *ETV6-RUNX1* (9.4%), *KMT2A (MLL)* -rearranged (KMT2A; 6.8%), *DUX4*-rearranged (DUX4; 5.3%), *TCF3-PBX1* (3.9%), *ZNF384*-rearranged (ZNF384; 2.5%) and *MEF2D*-rearranged (MEF2D; 2.2%)^{3,17}. There was marked variation in the prevalence of subtypes according to age, with *ETV6-RUNX1* and high hyperdiploid ALL being most common in children, and low hypodiploid, KMT2A and kinase-activated (Ph and Ph-like) ALL being more common in adults (Fig. 1b and Table 1).

An additional eight subtypes with distinct gene expression profiles and/or common genetic lesions were identified. Eighteen cases (0.9%) harbored rearrangements of *BCL2*, *MYC* and/or *BCL6*, predominantly in adults. These alterations are identical to those observed in “double/triple hit” lymphoma, but have rarely been observed in ALL, and here were of B cell precursor immunophenotype^{7,18}. Eleven cases (0.6%) with rearrangements of *NUTM1* to six different partner genes (3 with *ACIN1*, 3 *CUX1*, 2 *BRD9*, and one for each of *IKZF1*, *SLC12A6* and *ZNF618*) were identified, and nine cases (0.5%) with rearrangement of *HLF* to *TCF3* (N=8) or *TCF4* (N=1; Fig. 1a, Supplementary Fig. 1 and Supplementary Table 1).

Three groups included cases with similar gene expression profiles to established subtypes but lacking the founding rearrangements: ETV6-RUNX1-like (N=42), KMT2A-like (N=5) and ZNF384-like (N=4). Such cases commonly harbored alternate chromosomal rearrangements – e.g. *MED12-HOXA9* and *AFF1-TMEM156* in KMT2A-like cases, and *TCF3-FLI1*, *FUS-ERG*, *IKZF1* and alternate *ETV6* fusions (2 cases with *IKZF1-ETV6*, 2 with *ETV6-ELMO1* and 10 others with different fusion partners; Supplementary Table 1) in ETV6-RUNX1-like. In addition, cases with a modal chromosome number 47–50 and similar gene expression profile as high hyperdiploid ALL were defined as a low hyperdiploid subtype (N=51), and cases with *CRLF2* rearrangement lacking the gene expression profile of Ph-like ALL were assigned as an individual subtype (N=16) due to the importance of *CRLF2* rearrangement in guiding treatment⁵ (Supplementary Table 1 and Supplementary Fig. 1).

PAX5 alterations define two subtypes of B-ALL

Two subtypes were characterized by distinct gene expression profiles (defined by hierarchical clustering shown in Supplementary Fig. 1a) and different types of *PAX5* alterations. One, herein termed *PAX5*-altered (*PAX5alt*), comprised 148 (7.4%) cases, 109 (73.6%) of which were found with diverse *PAX5* alterations, including rearrangements, sequence mutations and focal/intragenic amplifications (Fig. 1, Supplementary Fig. 1 and Supplementary Table 5). Children in this subtype were more commonly classified as high risk (N=63) rather than standard risk (N=17) according to National Cancer Institute criteria. In the *PAX5alt* group, 57 cases (38.5% of this group) harbored *PAX5* rearrangements involving 24 partner genes that result in the expression of chimeric in-frame fusion proteins, the most frequent of which were *PAX5-ETV6* (N=19), *PAX5-NOL4L* (N=5), *PAX5-AUTS2* (N=4) and *PAX5-CBFA2T3* (N=4) (Fig. 2a and Supplementary Table 6). Two recurrent *PAX5* rearrangements were observed in non-*PAX5alt* cases. *PAX5-JAK2* (N=17) was exclusively observed in Ph-like ALL, and encodes an in frame chimeric fusion protein that results in constitutive JAK-STAT signaling. Rearrangement of *PAX5* with *ZCCHC7* immediately 5' of *PAX5* (N=18, 14 of which were in-frame) was observed in cases with other subtype-defining alterations, including kinase-activating rearrangements (*CRLF2*, *EBF1-PDGFRB*, *PAX5-JAK2*), *ETV6-RUNX1* and *IGH-DUX4*, indicating that *PAX5-ZCCHC7* is not a leukemia-initiating or subtype-defining event (Supplementary Table 1 and Supplementary Fig. 3a). Taken together, the *PAX5*-rearrangements except *PAX5-JAK2* and *PAX5-ZCCHC7* were significantly enriched in *PAX5alt* group (N=56, 37.8%) compared to other B-ALL (N=24, 1.3%; two-sided Fisher's exact test, $P<0.0001$).

Forty-six (31.1% of this group) *PAX5alt* cases harbored non-silent *PAX5* sequence mutations, compared to 4.4% (N=79) of other B-ALL excluding cases defined by *PAX5* p.Pro80Arg (P80R, see below; two-sided Fisher's exact test, $P<0.0001$; Supplementary Fig. 3a). Among the 62 sequence mutations identified within the *PAX5alt* group, 27 were homozygous (MAF range 0.87–1.00, median 0.96), which was commonly due to the loss of the wild type allele. The remaining 35 heterozygous mutations (MAF range 0.11–0.78, median 0.46) were observed in 19 cases, of which 15 were with 2 (N=14) or 3 (N=1) mutations. (Fig. 2a and Supplementary Table 7). Two hotspot missense mutations affecting amino acid p.Arg38 (N=20) and p.Arg140 (N=11) were identified, and were highly enriched

in the PAX5alt subtype (N=11 and 9, respectively). Notably, 10 of 11 p.Arg140 missense mutations were concomitant with p.Arg38 mutations, and 9 of the cases with these two mutations were classified as PAX5alt, which account for over half of the PAX5alt cases with multiple mutations (Fig. 2a and Fig. 3a). Among the 203 non-silent *PAX5* mutations identified in this study, 73.9% were missense mutations, especially those involving the DNA-binding domain (94.6%), while the more disruptive mutations including frameshift (N=36), nonsense (N=4) and splice-site mutations (N=9) were more commonly observed on the distal region of the PAX5 protein (Fig. 3a). Notably, a cluster of mutations in the PAX5 nuclear localization sequence, predicted to impede translocation of PAX5 to the nucleus, is observed across the spectrum of B-ALL subtypes (Fig. 3a).

Of the 1,141 cases with SNP array data, 368 (32.2%) had *PAX5* CNAs. However, these were not more frequent in the PAX5alt group, except for focal intragenic amplification of *PAX5* (PAX5amp) that was identified in 10 cases, eight of which were in the PAX5alt group (Supplementary Fig. 3b). One of the cases showed the region of amplification involved exons 2–5, and was predicted to result in duplication of the PAX5 DNA-binding domain. The structure and consequences of *PAX5* amplification was validated using WGS, RT-PCR, Sanger sequencing and fluorescent in-situ hybridization, showing in-frame internal tandem duplication from exon 2 to 5 on *PAX5* (Supplementary Fig. 4). Taken together, genetic alterations on *PAX5* were significantly enriched in PAX5alt group compared with other B-ALL subtypes (73.6% (109 out of 148) of PAX5alt vs. 5.7% (103/1796) of other samples; two-sided Fisher's exact test, $P<0.0001$), except the PAX5 P80R group, as described below (Supplementary Fig. 3c). Among the 96 PAX5alt cases with WGS and/or SNP array data, 11.5% (N=11) lacked an identifiable *PAX5* alteration, highlighting the need for complementary data to fully identify the genetic drivers of the PAX5alt gene expression profile in these cases.

In addition to *PAX5* alterations, recurrent genetic alterations observed in PAX5alt cases included those affecting cell cycle regulation (*CDKN2A*, *RB1* and *BTG1* deletions), B cell development (*IKZF1*, *VPREB1* and *BTLA* deletions), transcriptional regulation (e.g. *ZFP36L2*, *ETV6*, and *LEF1*) and/or epigenetic modification (e.g. *KDM6A*, *KMT2A*, *ATRX*; Fig. 2b and Supplementary Tables 8–9). Of note, signaling pathway mutations were observed in 63.1% (41 out of 65 cases with WES/WGS data) cases in this subtype. The distinct gene expression profile of PAX5alt was notable for a preponderance of down-regulated genes versus other B-ALL (319 up and 2,150 downregulated genes with 2 fold-change and adjusted $P<0.01$) (Supplementary Table 10), suggesting that loss of PAX5 transcriptional activation promotes leukemogenesis. Pathway analysis showed that genes encoding regulators of cytokine receptor signaling were highly enriched, consistent with the high frequency of mutations in signaling pathways (Supplementary Tables 11–12).

PAX5 P80R defines a distinct subtype of B-ALL

A second group with distinct gene expression profile was defined by the *PAX5* P80R mutation, which was present in all 44 cases, compared with four out of 1,944 other B-ALL cases (0.2%; Fig. 1a, Fig. 3a and Supplementary Table 7). In 30 cases, *PAX5* P80R was homozygous due to deletion of the wild-type *PAX5* allele or copy-neutral loss of

heterozygosity (Fig. 3b,c and Supplementary Tables 13–14). Of the remaining 14 cases with heterozygous *PAX5*P80R mutations, 13 harbored a second frameshift (N=7), nonsense (N=2) or deleterious missense (N=4) *PAX5* mutation. Although four of the remaining 1,944 cases also harbored the *PAX5*P80R mutation, all were heterozygous with preservation of a wild type *PAX5* allele and observed with similar gene expression profile with other subtypes (2 Ph-like, 1 BCL2/MYC and 1 PAX5alt), consistent with the notion that biallelic *PAX5* mutations, including P80R, are a hallmark of this subtype.

Collectively, signaling pathway mutations (Ras, JAK/STAT, *FLT3*, *BRAF* and *PIK3CA*) were present in 42 (95.5%) of PAX5 P80R cases, suggesting cooperativity between deregulated PAX5 activity and kinase signaling in leukemogenesis. The Ras pathway was particularly frequently mutated, most commonly with *NRAS*, *KRAS*, *PTPN11*, and *NF1* alterations (N=33, 75.0% vs 27.7% (538 out of 1,944) in other B-ALL, two-sided Fisher's exact test, $P<0.0001$; Fig. 3c and Supplementary Tables 15–16). Most Ras non-mutated PAX5 P80R cases harbored JAK/STAT mutations, most commonly interleukin 7 receptor (*IL7R*, N=7). We compared the distribution of mutations in each signaling pathway (Ras, JAK-STAT and *FLT3*) across B-ALL subtype and significant enrichment, particularly of the Ras and JAK-STAT pathways, was observed in PAX5 P80R group (90.1% of 44 PAX5 P80R versus 35.3% of 1944 other B-ALL, two-sided Fisher's exact test, $P<0.0001$; Fig. 4a). *NRAS*, *PTPN11* and *IL7R* were most frequently mutated in PAX5 P80R group compared with other B-ALL subtypes (47.7%, 29.5%, and 27.3%, respectively; Fig. 4b). Examining coding sequence mutations genome-wide, an additional target of mutation that was most commonly mutated in PAX5 P80R cases was *SETD2*, which encodes a histone 3 lysine 36 trimethylase (25.0% versus 6% of other B-ALL cases)^{19,20}.

The gene expression profile of PAX5 P80R ALL (334 up- and 2,552 downregulated genes with 2 fold-change and adjusted $P<0.01$; Fig. 5a and Supplementary Table 17) showed limited overlap of upregulated genes with the signature of PAX5alt ALL (9.9%, N=33 genes). However, around half (N=1,224) of the down regulated genes in these two groups were shared, indicating loss of PAX5 transcriptional activity in both PAX5 P80R and PAX5alt group (Fig. 5b). Consistent with the higher frequency of mutations in signaling pathways compared to PAX5alt group, expression of genes encoding regulators of cytokine receptor signaling were also enriched (Supplementary Tables 18–19). Moreover, direct comparison of P80R versus PAX5alt ALL revealed negative enrichment of B-lineage genes in P80R ALL, including targets of PAX5 such as BACH2²¹, indicating that P80R has more profoundly deleterious effects on B cell maturation than the alterations collectively present in PAX5alt ALL (Fig. 5c and Supplementary Table 20). Of note, the *MEGF10* gene (Multiple Epidermal Growth Factor-Like Domains Protein 10) otherwise silent or with low expression in normal B cells and other B-ALL subtypes was markedly overexpressed in PAX5 P80R cases (log2 fold-change 7.74, adjusted $P=6.13\times10^{-111}$), suggesting increased expression of this gene may serve as a biomarker and/or driver of this subtype.

***IKZF1* p.Asn159Tyr (N159Y) mutation defines a subtype of B-ALL**

The data described above suggest that sequence mutations may serve as initiating, subtype-defining events in B-ALL, rather than being secondary, cooperating events in

leukemogenesis. Consistent with this, we observed an additional subtype defined by a single transcription factor mutation. Eight cases harbored heterozygous *IKZF1* N159Y missense mutation, and in contrast to PAX5 P80R ALL, retention of expression of the non-mutated *IKZF1* allele. The gene expression profile was strikingly distinct compared to other B-ALL cases, including other *IKZF1*-altered cases (593 up and 1227 downregulated genes with 2 fold-change and adjusted $P < 0.01$; Fig. 1a, Supplementary Fig. 1a and Supplementary Table 21). N159 is located in the DNA-binding domain of IKZF1, and we have previously shown this mutation to perturb IKZF1 function, with distinctive nuclear mislocalization and induction of aberrant intercellular adhesion that is characteristic of many *IKZF1* alterations²². Notably, this subtype exhibited increased expression genes with roles in oncogenesis (the *IKZF1* interacting gene *YAP1*²³), chromatin remodeling (*SALL1*²⁴), and signaling (*ARHGEF28*²⁵) that were not deregulated in other subgroups of *IKZF1*-altered ALL. Interrogation of exome and DNA copy number data failed to identify recurrent sequence mutations or focal CNAs, but observed 6 out of the 8 cases have gain of whole chromosome 21, indicating potential interaction between IKZF1 N159Y and abnormal chromosome 21 in leukemogenesis (Supplementary Tables 1, 4 and 22).

Clinical characteristics and outcome of novel ALL subtypes

The median ages at diagnosis for patients in PAX5 P80R and PAX5alt ALL were 22.0 years and 15.4 years, respectively (Supplementary Table 23). Patients in PAX5 P80R and PAX5alt subtypes had median presenting white blood cell counts of $13.0 \times 10^9/L$ and $16.9 \times 10^9/L$, respectively, and were more likely to be male (65.9% and 68.9%) (Supplementary Table 23). Positive minimal residual disease ($< 0.01\%$) at the end of induction was detected in 7.2% and 29.4% of PAX5 P80R and PAX5alt cases, respectively (Supplementary Table 24). In children treated on the Children's Oncology Group AALL0232 study of NCI high risk B-ALL²⁶, the outcome was intermediate for both PAX5 P80R (5-year event-free survival (EFS) $75.0 \pm 14.2\%$, overall survival (OS) $75.0 \pm 14.2\%$, 8 evaluable cases) and PAX5alt (EFS $71.5 \pm 7.0\%$, OS $75.7 \pm 6.6\%$, 46 evaluable cases) compared to DUX4 ALL and other favorable risk subtypes (high hyperdiploid, ETV6-RUNX1 and TCF3-PBX1; Fig. 6a and Supplementary Table 25). In contrast, the outcome for PAX5 P80R in children treated on St. Jude Total Therapy protocols was unfavorable, although few patients were evaluable and were treated over multiple protocols (Supplementary Fig. 5 and Supplementary Table 25). Adults with PAX5 P80R ALL (15 of 288 evaluable cases) had intermediate and superior outcomes (EFS $63.0 \pm 13.3\%$, OS $61.9 \pm 13.4\%$) compared to patients with PAX5alt (EFS $32.2 \pm 9.4\%$, OS $42.1 \pm 10.2\%$, 27 evaluable cases). Of note in adults, DUX4 subtype was associated with excellent outcome, and BCL2/MYC with uniformly early treatment failure (Fig. 6b and Supplementary Table 26).

PAX5 P80R drives B-lymphoid leukemogenesis

We previously reported that PAX5 P80R and other point mutations affecting the DNA-binding domain of PAX5 have impaired ability to bind DNA and transcriptionally activate target genes²⁷. To examine the effects of PAX5 P80R on B cell maturation, we expressed wild-type PAX5, PAX5 P80R, p.Val26Gly (V26G) and p.Pro34Gln (P34Q) in *Pax5*^{-/-} lineage-depleted bone marrow cells. Expression of PAX5 V26G and P34Q resulted in near

complete rescue of B cell differentiation; however, expression of PAX5 P80R resulted in a block in differentiation at the pre-pro-B stage of B cell maturation (B220⁺ CD19⁻; Fig. 7a).

To investigate the oncogenic potential of *PAX5* mutations in B-ALL, we generated knock-in mouse strains harboring P80R or p.Gly183Ser (G183S), a germline variant in the octapeptide domain of PAX5 observed in familial B-ALL²⁸ (Supplementary Tables 27–28). Heterozygous *Pax5*^{P80R/+} and homozygous *Pax5*^{P80R/P80R} mice developed B220⁺CD19⁻ B-progenitor leukemia with median latencies of 160 and 83 days, respectively (Fig. 7b–d). Secondary transplantation into sub-lethally irradiated recipients resulted in rapid development of leukemia (Fig. 7e). In contrast, the PAX5 G183S octapeptide domain mutation did not induce leukemia. Array comparative genomic hybridization analysis of the mouse tumors identified CNAs of *Pax5* in two of three leukemias that arose in *Pax5*^{P80R/+} mice (Fig. 7f and Supplementary Table 29). These included a deletion of the entire *Pax5* locus resulting in monoallelic expression of *Pax5* P80R, and a truncating frameshift mutation of the wild type allele (Fig. 7g), recapitulating loss of the wild-type *PAX5* allele observed in human PAX5 P80R ALL. Although no CNA was observed on *Pax5* in one leukemia, the mutant allele frequency of P80R was 0.98 by RNA-seq analysis. Amplification of the entire *Pax5* locus was observed in a *Pax5*^{P80R/P80R} tumor, accompanied by high level expression of mutant *Pax5* P80R (Fig. 7f). Primary tumors harbored multiple *Jak1* and/or *Jak3* mutations (Supplementary Tables 30–32), the majority of which are known to result in constitutive activation of JAK/STAT signaling²⁹. Cells grown in vitro from secondary transplants displayed hyper-phosphorylation of STAT5 (Fig. 7g) which was inhibited by ruxolitinib, resulting in LC50 values in cytotoxic assays ranging from 10–50nM (data not shown). Gene set enrichment analysis demonstrated significant similarity of gene expression profiles of mouse and human PAX5 P80R leukemias (Supplementary Table 33 and Fig. 7h). These data support the notion that PAX5 P80R is an initiating lesion and cooperates with activated kinase signaling in leukemogenesis.

DISCUSSION

This study identified multiple new subtypes of B-ALL that exhibit distinct genomic, clinical and outcome features, and variation in prevalence according to age. While recent studies have identified several new targets of rearrangement in B-ALL (e.g. *DUX4*, *ZNF384* and *MEF2D*), here we show the power of transcriptome sequencing of large cohorts of ALL to identify new subtypes of heterogeneous genetic basis by clustering of gene expression profiles, the marked variance in subtype prevalence by age, and the striking observation that transcription factor missense mutations are initiating, subtype defining leukemogenic alterations. Moreover, integrative multimodal genomic analysis has enabled distillation of the often diverse alterations that define subgroups that had previously defined classification, particularly PAX5alt ALL, with its characteristic *PAX5* rearrangements, sequence mutations and focal amplifications. Identification and description of these subtypes account for many of the cases that lacked a subtype defining lesion and were termed “B-other” that previously eluded accurate risk stratification. Specifically, the PAX5alt and PAX5 P80R subtypes account for 9.7% of those cases previously termed B-other. Moreover, subtypes associated with unfavorable outcomes – *KMT2A*-rearranged, low hypodiploid, and kinase driven ALL

– account for over 65% of adult cases, indicating the genomic subtype is a central determinant of the poor outcome characteristic of ALL in older individuals.

Our results also highlight the importance of *PAX5* in regulating B cell lineage differentiation, and of *PAX5* alterations as central events in B lymphoid leukemogenesis. *PAX5* encodes a paired box DNA-binding transcription factor that regulates B lymphoid lineage commitment and maturation^{30,31}. Prior studies had identified frequent *PAX5* alterations in ALL, including rearrangements, focal deletions, sequence mutations and intragenic amplification in ALL^{27,32}. With the exception of *PAX5* rearrangements and the germline *PAX5* G183S mutation²⁸, these alterations have been considered secondary events that contribute to the arrest in lymphoid maturation characteristic of the disease. The current study indicates that *PAX5* alterations may be initiating, subtype-defining events in B-ALL. The *PAX5* P80 residue is located at a region of the paired domain that directly contacts the minor groove of DNA, impairs binding of *PAX5* to targets, and partly attenuates transcriptional activation²⁷. In more physiologic assays such as activation of *CD79A* (a *PAX5*-regulated gene that encodes the Ig-alpha protein of the B-cell antigen receptor) and expression of surface immunoglobulin, and as shown in this study, B-cell differentiation, *PAX5* P80R is profoundly deleterious and results in arrest in maturation at the pro- to pre-B cell stage. This subtype of B-ALL is also notable for the near universal inactivation of the wild-type *PAX5* allele, either by deletion, acquired copy-neutral loss of heterozygosity that duplicates the mutant allele, or acquisition of a second *PAX5* sequence mutation that results in loss-of-function. The importance of biallelic alterations of *PAX5* in this subgroup is supported by the knock-in *Pax5*^{P80R} mouse model of ALL, in which the majority of tumors acquire second hit alterations of *Pax5*, either by deletion of the wild type allele or amplification of mutant *Pax5*.

The utility of transcriptomic gene expression profiling to classify B-ALL and identify new subtypes is further supported by the identification of the more genetically diverse *PAX5*-altered group. Although multiple prior studies have identified *PAX5* rearrangements that result in a chimeric fusion that retains the paired domain of *PAX5* at the N-terminus with variable loss of the distal, transcriptional regulatory domains^{27,33}, a B-ALL subtype enriched for *PAX5* alterations had not previously been recognized, in part due to the genetic heterogeneity of this *PAX5*alt group. Additional genetic alterations, particularly intragenic *PAX5* amplification and non-P80R sequence mutations are also significantly enriched in this group. Such alterations are identified at much lower frequency in other subtypes, except rearrangements of *PAX5* to *JAK2* and *ZCCHC7* in Ph-like ALL⁵. In this context, the gene expression profile is consistent with Ph-like ALL rather than the *PAX5*alt group. Thus, accurate subgroup assignment requires consideration of transcriptional gene expression profile in addition to identification of *PAX5* alterations.

These results have important clinical implications for diagnosis and risk stratification. We show diversity in treatment outcomes according to subtype, with *PAX5* P80R and *PAX5*alt cases having intermediate to poor outcome in children and adults with B-ALL. This is consistent with prior observations that forms of leukemia with more stem-cell like features, such as *KMT2A*-rearranged and *IKZF1*-mutated Ph⁺/Ph-like ALL have inferior outcomes. Our findings suggest that new therapeutic approaches should be explored, including

targeting of the deregulated signaling pathways in PAX5 P80R ALL. These results are also of immediate diagnostic significance, as they suggest that the majority of B-ALL cases, and their underlying driver alterations, may be rapidly detected by analysis of transcriptome sequencing to guide classification, risk stratification, and tailored therapy.

ONLINE METHODS

Patients and clinical treatment protocols

The patients were enrolled on St. Jude Children's Research Hospital (St. Jude), Children's Oncology Group (COG), ECOG-ACRIN Cancer Research Group (ECOG-ACRIN), the Alliance for Clinical Trials in Oncology (Cancer and Leukemia Group B), M.D. Anderson Cancer Center (MDACC), University of Toronto, Northern Italian Leukemia Group (NILG), Southwestern Oncology Group (SWOG), Medical Research Council UK, and City of Hope treatment protocols. The treatment protocols for the patients include the St. Jude Children's Research Hospital Total XV³⁷; ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00137111) Identifier [NCT00137111](https://clinicaltrials.gov/ct2/show/study/NCT00137111)) and Total XVI protocols ([NCT00549848](https://clinicaltrials.gov/ct2/show/study/NCT00549848)); the COG P9906 high-risk B-ALL study³⁸; the COG AALL0232 high-risk ALL study²⁶ ([NCT00075725](https://clinicaltrials.gov/ct2/show/study/NCT00075725)); the COG AALL0331 standard-risk ALL study ([NCT00103285](https://clinicaltrials.gov/ct2/show/study/NCT00103285)); the ECOG-ACRIN E2993³⁹ ([NCT00002514](https://clinicaltrials.gov/ct2/show/study/NCT00002514)) and E1910 trials ([NCT02003222](https://clinicaltrials.gov/ct2/show/study/NCT02003222)); the MD Anderson Cancer Centre protocols^{40–43}, the Alliance – Cancer and Leukemia Group B protocols C19802 ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00003700) Identifier [NCT00003700](https://clinicaltrials.gov/ct2/show/study/NCT00003700)), C10102 ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00061945) Identifier [NCT00061945](https://clinicaltrials.gov/ct2/show/study/NCT00061945)) and C10403 ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00558519) Identifier [NCT00558519](https://clinicaltrials.gov/ct2/show/study/NCT00558519)), and SWOG protocols S0333 ([NCT00109837](https://clinicaltrials.gov/ct2/show/study/NCT00109837)) and S9400 ([NCT00002665](https://clinicaltrials.gov/ct2/show/study/NCT00002665)). Detailed clinical information for each case is provided in Supplementary Table 1. The patients enrolled in this study have provided written informed consent, assent (as appropriate), or parental consent (as appropriate) as part of their protocols, for research, including genetic research. All relevant ethical regulations were strictly followed during this study.

Transcriptome sequencing (RNA-seq)

RNA-seq was performed using TruSeq library preparation and HiSeq 2000 and 2500 sequencers (Illumina). All sequence reads were paired-end, and was performed using (1) total RNA and stranded RNA-seq [75 or 100 base-pair (bp) reads]; (2) polyA-selected mRNA (50, 75 or 100bp reads). Sequencing reads were mapped to the GRCh37 human genome reference by STAR⁴⁴ (version 2.4.2a) through the suggested two-pass mapping pipeline. Gene annotation downloaded from Ensembl website (see URLs) was used for STAR mapping and the following read-count evaluation. All the samples were sequenced with RefSeq coding region covered with 30-fold coverage 15% (median $\pm$ standard deviation, $37.2 \pm 7.5\%$). CICERO⁵ and FusionCatcher^{45,46} were used to detect fusions, and all the reported rearrangements were manually reviewed to keep the reliable ones. Due to the

URLs

Interactive portal of this study, <https://pecan.stjude.cloud/proteinpaint/study/PanALL>; Ensembl, <http://www.ensembl.org/>; The best practice workflow for calling SNVs and Indels from RNAseq data, <http://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq>; Picard, <http://broadinstitute.github.io/picard/>; COSMIC database, <https://cancer.sanger.ac.uk/cosmic/>; PeCan portal, <https://pecan.stjude.cloud/home>; TARGET Project, <https://ocg.cancer.gov/programs/target>; TARGET Project methods, <https://ocg.cancer.gov/programs/target/target-methods#3233>; PennCNV for Affymetrix data, <http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/>; dbGaP, <http://www.ncbi.nlm.nih.gov/gap>.

complexity of *DUX4* rearrangements, some of the *DUX4* fusions were manually rescued by checking the aligned reads within IGV browser⁴⁷. RNA extracted from flow sorted normal B lymphoid cells were used for RNA-seq and details were provided in our previous study⁴⁸.

Gene expression level evaluation from RNA-seq data

To evaluate gene expression level, read-count for each annotated gene was calculated by HTSeq package⁴⁹, and gene expression level normalization and differential expression analysis was carried out by DESeq2 Bioconductor R package⁵⁰. To evaluate digital gene expression level, regularized log transformed (rlog) value was calculated by DESeq2. ComBat function in sva R package⁵¹ was used to correct the batch effect introduced by different library preparation strategies and sequencing lengths. With the rlog gene expression level, R package Rtsne was used to map the samples to 2-dimensional t-Distributed Stochastic Neighbor Embedding (tSNE) plot with the top 1,000 most variable genes (base on median absolute deviation) and the tSNE perplexity parameter was set 30. Different gene numbers (200, 500, 1,000, and 2,000) and tSNE parameters (perplexity 20, 30, 40 and 50) were explored and stable clusters were observed. Gene signature analysis was also carried out by DESeq2 with default parameters to evaluate the differentially expressed genes.

Copy number alteration (CNA) detection from RNA-seq data

RNA-seq data is not optimal for calling genomic level CNAs, but still informative to evaluate chromosome or arm level copy number changes for cases lacking karyotypic or DNA array data. To accomplish this, we ordered the genes based on the median absolute deviation (MAD) of their expression level across the samples, and then picked a subset ($\frac{1}{4}$ to $\frac{1}{3}$) of the genes with least MAD as stably expressed genes. To assist the CNA calling, mutant allele frequency of SNVs detected from RNA-seq data were also plotted against the gene expression level of the stably expressed genes to double check if the CNAs were reliable (Supplementary Fig. 7). This strategy is mainly applied to resolve the B-ALL subtyping issue for potential hyper- and hypodiploid cases without karyotypic information. An evaluation was performed in 30 aneuploid cases (7 high hyperdiploid, 21 low hypodiploid and 2 near haploid) with chromosomal level CNAs called from both SNP array RNAseq data (Supplementary Table 34). In total, 295 CNAs were called from SNP array data, and 285 (96.6%) could be faithfully recapitulated by RNA-seq data, indicating the application of RNA-seq to evaluate chromosomal level CNAs was highly reliable. One false positive CNA called from RNA-seq was on X chromosome, and 10 false negative CNAs missed by RNA-seq were mainly on sex chromosome X (N=8), which may be explained by X chromosome inactivation.

Mutation detection from RNA-seq data

The SNVs and Indels were called from RNA-seq data by following the best practice workflow from the GATK⁵² forum (see URLs). In general, STAR mapped bam files were processed by Picard (see URLs, version 1.129) to mark duplicate reads, then the GATK module SplitNCigarReads was used to split reads into exon segments and hard-clip any sequences overhanging into the intronic regions. Mutations were called by the HaplotypeCaller module in GATK and then filtered by the following steps: 1. At least 3 reads support the mutation and the mutant allele frequency is ≥ 0.05 ; 2. Not observed in

common SNP database dbSNP 150; 3. Not observed in 2 samples from our germline Exome sequence cohort (775 samples). After filtering, all the mutations were annotated to RefSeq genes and the non-silent mutations have been previously validated according to COSMIC database V84 (see URLs) and/or PeCan portal (see URLs) were kept for further analysis.

Whole genome/exome sequencing

Whole genome sequencing (WGS) of leukemia and paired germline samples were carried out by TARGET (Therapeutically Applicable Research To Generate Effective Treatments program, see URLs) and St Jude Children's Research Hospital (SJCRH). Whole exome sequencing (WES) were performed at SJCRH.

For WGS data generated by TARGET, methods for DNA preparation, sequencing, and quality control are available at TARGET Project portal (see URLs). For WGS performed at SJCRH, Genomic DNA was quantified using the Quant-iT PicoGreen assay (Life Technologies). Genomic DNA was sheared on an LE220 ultrasonicator (Covaris). Libraries were prepared from sheared DNA with HyperPrep Library Preparation Kits (Kapa Biosystems). Libraries were analyzed for insert size distribution on a 2100 BioAnalyzer High Sensitivity kit (Agilent Technologies) or Caliper LabChip GX DNA High Sensitivity Reagent Kit (PerkinElmer). Libraries were quantified using the Quant-iT PicoGreen dsDNA assay (Life Technologies). Paired end 150 cycle sequencing was performed on a NovaSeq 6000 (Illumina). For WES at SJCRH, library preparation was performed using Nextera rapid exome kit (Illumina), and was performed using the Caliper Biosciences (Perkin Elmer) Sciclone G3. First-round PCR (10 cycles) was performed using Nextera kit reagents (Illumina), and clean-up steps employed BC/Agencourt AMPure XP beads. Target capture utilized Nextera rapid capture exome kit (Illumina) and supplied hybridization and associated reagents. Library quality control was performed using a Victor fluorescence plate reader with Quant-it dsDNA reagents for pre-pool quantitation, and Bio-analyzer 2200 (Agilent) for final library quantitation. Paired-end sequencing was performed using HiSeq 2000 or 2500 (Illumina) with read length 100bp.

WGS/WES reads alignment and quality control

Paired-end WGS and WES reads were mapped to human reference genome GRCh37 by BWA⁵³ (version 0.7.12). Samtools⁵⁴ (version 1.3.1) were used to generate chromosomal coordinate-sorted and indexed bam files, and then processed by Picard (see URLs, version 1.129) MarkDuplicates module to mark PCR duplications. Then the reads were realigned around potential indel regions by GATK⁵⁵ (version 3.7) IndelRealigner module. Sequencing depth and coverage was assessed based on coding regions (~34Mb) defined by RefSeq genes.

WGS/WES mutation calling and filtering workflow

UnifiedGenotyper (within GATK v3.7) and muTect2 (Beta version within GATK v4) modules were applied to call SNVs and Indels from leukemia and paired germline samples. The raw mutations were filtered by a homemade pipeline to exclude: 1) reported common SNPs/Indels from dbSNP v150; 2) germline mutations detected from matched germline

control samples. All the non-silent SNVs/Indels passed the filtering pipeline were manually reviewed and only the highly reliable somatic ones were reported. Meanwhile, adjacent nucleotide changes observed on the same allele were merged into one mutation.

CNA and loss of heterozygosity (LOH) detection from microarray

Single nucleotide polymorphism (SNP) microarray data from two different platforms were used in this study: one is Illumina Infinium Omni2.5 Exome-8 (2.6 million probes), and the other one is from Affymetrix Genome-Wide Human SNP Array 6.0 (1.8 million probes). For the Illumina platform, DNA extracted from leukemia and matched germline samples was hybridized to SNPs array according to the manufacturer's protocol. The raw intensity data (*.idat files) were analyzed by the Genotyping Module from Illumina GenomeStudio software (version 2.0.3). Normalized Log R Ratio (LRR) and B Allele Frequency (BAF) for all the available probes were evaluated. All the Affymetrix SNP data are from our previous study and have been thoroughly analyzed^{5,56}. Affymetrix SNP data for the PAX5 P80R samples are converted to LRR and BAF value following the pipeline suggested by PennCNV⁵⁷ (see URLs). With the input of LRR and BAF, somatic CNA and LOH from paired or unpaired samples were called by OncoSNP⁵⁸ (version 2.1) and manually reviewed by ShinyCNV⁵⁹. Only the somatic alterations meeting the criteria proposed by OncoSNP and PennCNV were kept for further analysis.

Gene set enrichment and pathway analysis

Raw read-count from RNA-seq data was imported to DESeq2 for differential gene expression analysis. To perform Gene Set Enrichment Analysis³⁵ (GSEA), the gene expression profile of PAX5 P80R was defined by comparing gene expression levels between PAX5 P80R ALL and normal B cells purified from human bone marrow, and ranking genes according to fold-change and significance. GSEA was performed using mSigDB C2 genes and curated gene sets from in house analyses. Significantly regulated genes (≥ 2 fold-change and adjusted $P < 0.01$) were selected to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and gene ontology (GO) enrichment analysis by GO-Elite⁶⁰.

Animals

Mice were housed in an American Association of Laboratory Animal Care (AALAC)-accredited facility and all experiments were approved and in compliance with the SJCRH Institutional Animal Care and Use Committee (IACUC)-approved protocol in accordance with NIH guidelines. *Pax5*^{-/-} mutant mice³¹ were provided by Meinrad Busslinger and maintained on the C56BL/6 background. Genotyping was determined by PCR analysis as previously described⁶¹.

Pax5^{P80R} and *Pax5*^{G183S} mouse lines

Pax5^{P80R} mice were generated using CRISPR-Cas9 technology. Pronuclear-stage C57BL/6N/J zygotes were injected by the SJCRH Transgenic/Gene Knockout Shared Resource with a sgRNA [*Pax5*_P80R_Guide 01 (50 ng/μL)] designed to introduce a DNA double strand break into exon 3 of the *Pax5* gene (gene ID 18507), a human codon-optimized Cas9 mRNA transcript (100 ng/μL) and a 200-nucleotide-long single stranded

DNA molecule containing the desired mutations (Pax5-P80R-HDR, 2 pmol/μL, Supplementary Table 27). Approximately 25 injected zygotes were surgically transplanted into an infundibulum of 0.5 days post coitus (dpc) pseudo pregnant CD-1 females and newborn mice carrying the *Pax5*^{P80R} allele were identified by PCR and Sanger sequencing using primers Pax5-e3-F1 and Pax5-e3-R1. A similar strategy was used to generate *Pax5*^{G183S} mice. *Pax5*^{G183S} sgRNA (Pax5_G183S_Guide 01), repair template (Pax5_G183S_HDR) and PCR primers (Pax5-e5-F1 and Pax5-e5-R1) sequences are shown in Supplementary Table 27. sgRNAs were designed and generated as described previously⁶². Cas9 mRNA transcripts were also generated as described previously⁶². The target site for each sgRNA is unique in the mouse genome and no potential off-target site with fewer than three mismatches were found using the Cas-OFFinder algorithm⁶³ (Supplementary Table 28). *Pax5*^{P80R} and *Pax5*^{G183S} loci were genotyped by PCR using primers Pax5-e3-F1 and Pax5-e3-R1 or Pax5e5-F1 and Pax5e5-R1 and subsequent Sanger sequencing of the PCR amplicon (Supplementary Table 27).

For leukemia studies, heterozygous mice for each allele were interbred and the offspring were monitored daily for signs of illness. Moribund mice were sacrificed and complete blood counts were taken from peripheral blood. Bone marrow and spleen samples were collected and analyzed by flow cytometry for lineage markers (Mac1-Alexa700, Gr1-PerCP-Cy5.5, B220eFluor605, CD3-APC, CD41-PE, Ter119-V500) and Hardy panel B-lymphocyte markers (CD43PerCP-Cy5.5, B220-eFluor605, CD19-APC-Cy7, BP1-APC, IgM-PE-Cy7) to determine the immunophenotype of leukemic cells. For western blotting, cells were lysed in RIPA buffer, subjected to SDS-PAGE, and probed with anti-PAX5 antibody (Millipore; 05–1573, clone 1H9).

Retroviral vectors and retrovirus production

Vector production was performed as previously described by Mullighan et al²⁷. Briefly, the coding regions of the exon 1a isoforms of wild type and mutant *PAX5* were cloned from B-ALL patients into the *XhoI* site of the retroviral MSCV-IRES-mRFP (MIR) vector. The Eco Phoenix packaging system was used to produce ecotropic retrovirus. Briefly, 24 hours after plating Eco Phoenix cells in complete DMEM medium, the cells were transfected with WT *PAX5*, mutant *PAX5* or MIR plasmid DNA using FuGENE 6 according to the manufacturer's instructions (Roche Diagnostics, Alameda, CA). Twenty-four hours later, media was removed and replaced with complete IMDM. Viral supernatant was collected starting 48 hours post-transfection, filtered through a 0.45-μm filter (Millipore, Billerica, MA, USA), aliquoted and frozen at -80°C until use. Virus titration was performed by transduction of NIH3T3 fibroblast cells and quantitation of RFP expression at 48 hours. Viral titers ranged from 10⁵ to 10⁶ virus particles per milliliter, depending on the construct and these titers were highly reproducible.

Retroviral transduction and ex-vivo culture of *Pax5*^{-/-} progenitors

Bone marrow cells were obtained from the long bones of 9–12 day old mice. Mononuclear cells were stained with biotinylated anti-mouse CD5, Ly-6G, CD45R/B220 and TER119 antibodies (BD Biosciences) and lineage-positive (Lin⁺) cells were labeled with Streptavidin dynabeads M-280 (Invitrogen Life Technologies, Grand Island, NY, USA) and magnetically

separated using DynaMag-15 (Invitrogen Life Technologies, Grand Island, NY, USA) per manufacturer's instructions. The Lin- cells were then pre-stimulated for 48 hours with 50ng/ml SCF, 50ng/ml Flt3L, 30ng/ml IL6, 20ng/ml IL-3 and 20ng/ml IL-7 from PeproTech Inc (Rocky Hill, NJ, USA). Up to 2 million Lin- cells were transduced in Retronectin (Takara, Shuzo, Otsu, Japan) coated plates preloaded with viral supernatants and cultured in the presence of cytokines for two days. RFP positive cells were isolated by flow activated sorting and cultured for 13 days on an IL-7 producing irradiated T220 stromal cell line. Cells were then harvested for immunophenotyping using Allophycocyanin (APC), APC-Cy7, Fluorescein isothiocyanate (FITC)-, Phycoerythrin (PE)-, Peridinin-chlorophyll-protein (PerCP) cy5.5, PE-cy7, or Biotin-conjugated mAb specific for B220, CD19, BP-1, IgM, CD5 (Ly-1), Ly-6G (Gr-1), CD45R/B220, TER119 (BD Biosciences). Staining of cells was performed using standard protocols and analysis was done in the presence of DAPI nucleic acid to exclude dead cells. Cell sorting was done on a FACS Vantage or FACSDiva Cell Sorter (BD Biosciences). Data was analyzed using then FlowJo software (Treestar Inc.) and expressed as the percentage of positive cells for the specific B cell antigens. Each experiment was repeated at least three times.

Statistical analysis

Associations between categorical values were examined using two-sided Fisher's exact test. Associations between B-ALL subtypes, and event-free survival (EFS) and overall survival (OS) were examined by the Kaplan-Meier estimator, with Peto's estimator of standard deviation and the two-sided time-stratified Cochran-Mantel-Haenszel test⁶⁴. An event was defined as a failure to achieve remission, a relapse after remission, or the development of a second malignant neoplasm. A multivariate analysis of event-free and overall survival was performed with the Cox proportional hazards regression model⁶⁵. Analyses were performed using Prism version 7.0 (GraphPad Software), R version 3.4.3 (www.r-project.org), and SAS software, version 9.4 (SAS Institute).

Reporting summary

Further information on experimental design is available in the Life Sciences Reporting Summary linked to this paper.

Data availability

The raw and analyzed data are provided in a graphical, interactive platform (see URLs). Genomic data generated for this study are deposited to the European Genome-phenome Archive (EGA) under accession number EGAS00001003266. Other legacy data used in this study have been deposited to EGA in previous projects under accession number EGAS00001000654, EGAS00001001952, EGAS00001001923, EGAS00001002217 and EGAS00001000447. The TARGET genomic data used in this study are available through the TARGET website (see URLs) and also at the database of Genotypes and Phenotypes (see URLs) under accession number phs000218 (TARGET). The other data supporting this study are available from the corresponding author upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Zhaohui Gu^{1,*}, Michelle L. Churchman^{1,*}, Kathryn G. Roberts^{1,*}, Ian Moore¹, Xin Zhou², Joy Nakitandwe¹, Kohei Hagiwara², Stephane Pelletier³, Sebastien Gingras⁴, Hartmut Berns⁵, Debbie Payne-Turner¹, Ashley Hill¹, Ilaria Iacobucci¹, Lei Shi⁶, Stanley Pounds⁶, Cheng Cheng⁶, Deqing Pei⁶, Chunxu Qu¹, Scott Newman², Meenakshi Devidas⁷, Yunfeng Dai⁷, Shalini C. Reshmi⁸, Julie Gastier-Foster⁸, Elizabeth A. Raetz⁹, Michael J. Borowitz¹⁰, Brent L. Wood¹¹, William L. Carroll¹², Patrick A. Zweidler-McKay¹³, Karen R. Rabin¹⁴, Leonard A. Mattano¹⁵, Kelly W. Maloney¹⁶, Alessandro Rambaldi¹⁷, Orietta Spinelli¹⁷, Jerald P. Radich¹⁸, Mark D. Minden¹⁹, Jacob M. Rowe²⁰, Selina Luger²¹, Mark R. Litzow²², Martin S. Tallman²³, Janis Racevskis²⁴, Yanming Zhang²⁵, Ravi Bhatia²⁶, Jessica Kohlschmidt²⁷, Krzysztof Mrózek²⁷, Clara D. Bloomfield²⁷, Wendy Stock²⁸, Steven Kornblau²⁹, Hagop M. Kantarjian²⁹, Marina Konopleva²⁹, Williams Evans³⁰, Sima Jeha³¹, Ching-Hon Pui³¹, Jun Yang³⁰, Elisabeth Paietta²⁴, James Downing¹, Mary V. Relling³⁰, Jinghui Zhang², Mignon L. Loh³², Stephen P. Hunger³³, and Charles G. Mullighan^{1,+}

Affiliations

¹Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN

²Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN

³Department of Immunology, St. Jude Children's Research Hospital, Memphis, TN

⁴Department of Immunology, University of Pittsburgh, Pittsburgh, PA

⁵Department of Transgenic/Gene Knockout Shared Resource, St. Jude Children's Research Hospital, Memphis, TN

⁶Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN

⁷Department of Biostatistics, University of Florida, Gainesville, FL

⁸Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH

⁹Division of Pediatric Hematology-Oncology, New York University, New York, NY

¹⁰Division of Hematologic Pathology, Johns Hopkins University, Baltimore, MD

¹¹Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA

¹²Perlmutter Cancer Center, NYU-Langone Health, New York, NY

¹³ImmunoGen, Inc, Waltham, MA

¹⁴Baylor College of Medicine, Houston, TX

- ¹⁵HARP Pharma Consulting, Mystic, CT
- ¹⁶University of Colorado School of Medicine and Children's Hospital, Aurora CO
- ¹⁷Hematology and Bone Marrow Transplant Unit, Ospedale Papa Giovanni XXIII, Bergamo, Italy
- ¹⁸Fred Hutchinson Cancer Research Center, Seattle, WA
- ¹⁹Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada
- ²⁰Hematology, Shaare Zedek Medical Center, Jerlem, Israel
- ²¹Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA
- ²²Division of Hematology, Department of Medicine, Mayo Clinic, Rochester, MN
- ²³Memorial Sloan Kettering Cancer Center, New York, NY
- ²⁴Cancer Center, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY
- ²⁵Cytogenetics Laboratory, Memorial Sloan Kettering Cancer Center
- ²⁶Division of Hematology-Oncology, University of Birmingham, Alabama
- ²⁷Comprehensive Cancer Center, The Ohio State University, Columbus, OH
- ²⁸University of Chicago Medical Center, Chicago, IL
- ²⁹Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX
- ³⁰Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN
- ³¹Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN
- ³²Department of Pediatrics, UCSF Benioff Children's Hospital and the Helen Diller Family, San Francisco, CA
- ³³Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA

ACKNOWLEDGEMENTS

We thank the Biorepository, the Genome Sequencing Facility of the Hartwell Center for Bioinformatics and Biotechnology, and the Cytogenetics core facility of St. Jude Children's Research Hospital (SJCRH). This work was supported by the American Lebanese Syrian Associated Charities of SJCRH, American Society of Hematology Scholar Award (to Z.G. and K.G.R.), The Leukemia & Lymphoma Society's Career Development Program Special Fellow (to Z.G.), St. Baldrick's Foundation Robert J. Arceci Innovation Award (to C.G.M.), National Cancer Institute (NCI) Outstanding Investigator Award R35 CA197695 (to C.G.M.), National Institute of General Medical Sciences (NIGMS) P50 GM115279 (to C.G.M.), the NCI grants P30 CA021765 (St. Jude Cancer Center Support Grant), the ECOG-ACRIN Operations Center grant CA180820 (to Dr. Peter O'Dwyer from University of Pennsylvania and the Abramson Cancer Center), CA189859 (to E.P.), CA180790 (to M.R.L.) and CA180791 (to M.S.T. and Y.Z.).

REFERENCES

1. Hunger SP & Mullighan CG Acute Lymphoblastic Leukemia in Children. *N Engl J Med* 373, 1541–52 (2015). [PubMed: 26465987]
2. Iacobucci I & Mullighan CG Genetic Basis of Acute Lymphoblastic Leukemia. *J Clin Oncol* 35, 975–983 (2017). [PubMed: 28297628]
3. Roberts KG et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell* 22, 153–66 (2012). [PubMed: 22897847]
4. Iacobucci I et al. Truncating Erythropoietin Receptor Rearrangements in Acute Lymphoblastic Leukemia. *Cancer Cell* 29, 186–200 (2016). [PubMed: 26859458]
5. Roberts KG et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N Engl J Med* 371, 1005–15 (2014). [PubMed: 25207766]
6. Zhang J et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat Genet* 48, 1481–1489 (2016). [PubMed: 27776115]
7. Gu Z et al. Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat Commun* 7, 13331 (2016). [PubMed: 27824051]
8. Suzuki K et al. MEF2D-BCL9 Fusion Gene Is Associated With High-Risk Acute B-Cell Precursor Lymphoblastic Leukemia in Adolescents. *J Clin Oncol* 34, 3451–9 (2016). [PubMed: 27507882]
9. Gocho Y et al. A novel recurrent EP300-ZNF384 gene fusion in B-cell precursor acute lymphoblastic leukemia. *Leukemia* 29, 2445–8 (2015). [PubMed: 25943178]
10. Yasuda T et al. Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. *Nat Genet* 48, 569–74 (2016). [PubMed: 27019113]
11. Lilljebjorn H et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat Commun* 7, 11790 (2016). [PubMed: 27265895]
12. Lilljebjorn H & Fioretos T New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood* 130, 1395–1401 (2017). [PubMed: 28778863]
13. Den Boer ML et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol* 10, 125–34 (2009). [PubMed: 19138562]
14. Mullighan CG et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N Engl J Med* 360, 470–80 (2009). [PubMed: 19129520]
15. Tibshirani R, Hastie T, Narasimhan B & Chu G Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 99, 6567–72 (2002). [PubMed: 12011421]
16. Harrison CJ et al. An international study of intrachromosomal amplification of chromosome 21 (iAMP21): cytogenetic characterization and outcome. *Leukemia* 28, 1015–21 (2014). [PubMed: 24166298]
17. Holmfeldt L et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet* 45, 242–52 (2013). [PubMed: 23334668]
18. Johnson NA et al. Lymphomas with concurrent BCL2 and MYC translocations: the critical factors associated with survival. *Blood* 114, 2273–9 (2009). [PubMed: 19597184]
19. Zhu X et al. Identification of functional cooperative mutations of SETD2 in human acute leukemia. *Nat Genet* 46, 287–93 (2014). [PubMed: 24509477]
20. Mar BG et al. Mutations in epigenetic regulators including SETD2 are gained during relapse in paediatric acute lymphoblastic leukaemia. *Nat Commun* 5, 3469 (2014). [PubMed: 24662245]
21. Schebesta A et al. Transcription factor Pax5 activates the chromatin of key genes involved in B cell signaling, adhesion, migration, and immune function. *Immunity* 27, 49–63 (2007). [PubMed: 17658281]
22. Churchman ML et al. Efficacy of Retinoids in IKZF1-Mutated BCR-ABL1 Acute Lymphoblastic Leukemia. *Cancer Cell* 28, 343–56 (2015). [PubMed: 26321221]
23. Hu Y, Yoshida T & Georgopoulos K Transcriptional circuits in B cell transformation. *Curr Opin Hematol* 24, 345–352 (2017). [PubMed: 28463873]

24. Lauberth SM & Rauchman M A conserved 12-amino acid motif in Sall1 recruits the nucleosome remodeling and deacetylase corepressor complex. *J Biol Chem* 281, 23922–31 (2006). [PubMed: 16707490]
25. Miller NL et al. A non-canonical role for Rgnef in promoting integrin-stimulated focal adhesion kinase activation. *J Cell Sci* 126, 5074–85 (2013). [PubMed: 24006257]
26. Larsen EC et al. Dexamethasone and High-Dose Methotrexate Improve Outcome for Children and Young Adults With High-Risk B-Acute Lymphoblastic Leukemia: A Report From Children's Oncology Group Study AALL0232. *J Clin Oncol* 34, 2380–8 (2016). [PubMed: 27114587]
27. Mullighan CG et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* 446, 758–64 (2007). [PubMed: 17344859]
28. Shah S et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat Genet* 45, 1226–31 (2013). [PubMed: 24013638]
29. Dang J et al. Pax5 is a tumor suppressor in mouse mutagenesis models of acute lymphoblastic leukemia. *Blood* 125, 3609–3617 (2015). [PubMed: 25855603]
30. Adams B et al. Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes Dev* 6, 1589–607 (1992). [PubMed: 1516825]
31. Urbanek P, Wang ZQ, Fetka I, Wagner EF & Busslinger M Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking Pax5/BSAP. *Cell* 79, 901–12 (1994). [PubMed: 8001127]
32. Kuiper RP et al. High-resolution genomic profiling of childhood ALL reveals novel recurrent genetic lesions affecting pathways involved in lymphocyte differentiation and cell cycle progression. *Leukemia* 21, 1258–66 (2007). [PubMed: 17443227]
33. Fortschegger K, Anderl S, Denk D & Strehl S Functional heterogeneity of PAX5 chimeras reveals insight for leukemia development. *Mol Cancer Res* 12, 595–606 (2014). [PubMed: 24435167]
34. Novershtern N et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* 144, 296–309 (2011). [PubMed: 21241896]
35. Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545–50 (2005). [PubMed: 16199517]
36. Hardy RR & Hayakawa K B cell development pathways. *Annu Rev Immunol* 19, 595–621 (2001). [PubMed: 11244048]
37. Pui CH et al. Treating childhood acute lymphoblastic leukemia without cranial irradiation. *N Engl J Med* 360, 2730–41 (2009). [PubMed: 19553647]
38. Bowman WP et al. Augmented therapy improves outcome for pediatric high risk acute lymphocytic leukemia: results of Children's Oncology Group trial P9906. *Pediatr Blood Cancer* 57, 569–77 (2011). [PubMed: 21360654]
39. Goldstone AH et al. In adults with standard-risk acute lymphoblastic leukemia, the greatest benefit is achieved from a matched sibling allogeneic transplantation in first complete remission, and an autologous transplantation is less effective than conventional consolidation/maintenance chemotherapy in all patients: final results of the International ALL Trial (MRC UKALL XII/ECOG E2993). *Blood* 111, 1827–33 (2008). [PubMed: 18048644]
40. Kantarjian H et al. Long-term follow-up results of hyperfractionated cyclophosphamide, vincristine, doxorubicin, and dexamethasone (Hyper-CVAD), a dose-intensive regimen, in adult acute lymphocytic leukemia. *Cancer* 101, 2788–801 (2004). [PubMed: 15481055]
41. Ravandi F et al. First report of phase 2 study of dasatinib with hyper-CVAD for the frontline treatment of patients with Philadelphia chromosome-positive (Ph+) acute lymphoblastic leukemia. *Blood* 116, 2070–7 (2010). [PubMed: 20466853]
42. Thomas DA et al. Treatment of Philadelphia chromosome-positive acute lymphocytic leukemia with hyper-CVAD and imatinib mesylate. *Blood* 103, 4396–407 (2004). [PubMed: 14551133]
43. Thomas DA et al. Chemoimmunotherapy with a modified hyper-CVAD and rituximab regimen improves outcome in de novo Philadelphia chromosome-negative precursor B-lineage acute lymphoblastic leukemia. *J Clin Oncol* 28, 3880–9 (2010). [PubMed: 20660823]
44. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013). [PubMed: 23104886]

45. Nicorici D et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNAsequencing data. *bioRxiv* (2014).
46. Edgren H et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* 12, R6 (2011). [PubMed: 21247443]
47. Robinson JT et al. Integrative genomics viewer. *Nat Biotechnol* 29, 24–6 (2011). [PubMed: 21221095]
48. Alexander TB et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* (2018).
49. Anders S, Pyl PT & Huber W HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–9 (2015). [PubMed: 25260700]
50. Anders S & Huber W Differential expression analysis for sequence count data. *Genome Biol* 11, R106 (2010). [PubMed: 20979621]
51. Leek JT, Johnson WE, Parker HS, Jaffe AE & Storey JD The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–3 (2012). [PubMed: 22257669]
52. McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing nextgeneration DNA sequencing data. *Genome Res* 20, 1297–303 (2010). [PubMed: 20644199]
53. Li H & Durbin R Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–60 (2009). [PubMed: 19451168]
54. Li H et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–9 (2009). [PubMed: 19505943]
55. DePristo MA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491–8 (2011). [PubMed: 21478889]
56. Pounds S et al. Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* 25, 315–21 (2009). [PubMed: 19052058]
57. Wang K et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17, 1665–74 (2007). [PubMed: 17921354]
58. Yau C OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* 29, 2482–4 (2013). [PubMed: 23926227]
59. Gu Z & Mullighan CG ShinyCNV: a Shiny/R application to view and annotate DNA copy number variations. *Bioinformatics*, bty546-bty546 (2018).
60. Zambon AC et al. GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* 28, 2209–10 (2012). [PubMed: 22743224]
61. Nutt SL, Urbanek P, Rolink A & Busslinger M Essential functions of Pax5 (BSAP) in pro-B cell development: difference between fetal and adult B lymphopoiesis and reduced V-to-DJ recombination at the IgH locus. *Genes Dev* 11, 476–91 (1997). [PubMed: 9042861]
62. Pelletier S, Gingras S & Green DR Mouse genome engineering via CRISPR-Cas9 for study of immune function. *Immunity* 42, 18–27 (2015). [PubMed: 25607456]
63. Bae S, Park J & Kim JS Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* 30, 1473–5 (2014). [PubMed: 24463181]
64. Mantel N Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50, 163–70 (1966). [PubMed: 5910392]
65. COX D Regression Models and Life-Tables. *Journal of* (1972).

Figure 1. Integrative B-ALL subtypes.

a. Gene expression profiling (GEP) of 1,988 cases shown in a two-dimensional t-distributed stochastic neighbor embedding (tSNE) plot. Each dot represents a sample. The top 1,000 most variable genes (based on median absolute deviation) were selected and processed by the tSNE algorithm with perplexity score of 30. Major B-ALL subtypes are highlighted in different colors, which include *ETV6-RUNX1*, *KMT2A*- (*MLL*-) rearranged (*KMT2A*), *TCF3-PBX1*, *DUX4*-rearranged (*DUX4*), *ZNF384*-rearranged (*ZNF384*), *MEF2D*-rearranged (*MEF2D*), *BCR-ABL1* (Ph), Ph-like, High hyperdiploid, Low hypodiploid, Near

haploid and cases with intrachromosomal amplification of chromosome 21 (iAMP21). Three uncommon subtypes are also shown: *BCL2/MYC*-rearranged (BCL2/MYC), *TCF3/TCF4-HLF* (HLF) and *NUTM1*-rearranged (NUTM1). A group of samples with distinct GEP and universal *PAX5* p.Pro80Arg (P80R) mutation were observed (PAX5 P80R). A cluster of cases with diverse *PAX5* alterations (PAX5alt) is also observed adjacent to the PAX5 P80R group, with diverse rearrangements, focal/intragenic amplifications and non-*PAX5* P80R mutations. Eight cases with distinct GEP were identified with the same *IKZF1* missense mutation p.Asn159Tyr (N159Y). Cases in five subtypes including low hyperdiploid, ETV6-RUNX1-like, KMT2A-like, ZNF384-like and CRLF2(non-Ph-like) are shown as gray dots, but not specifically labeled in the plot. **b**, The distribution of B-ALL subtypes within each subtype (upper) or each age group (lower). The definition for age groups is described in Table 1. The subtypes are grouped as gross chromosomal alteration, transcription factor (TF) rearrangement, other TF alteration, kinase driven and others.

Figure 2. Mutational profile of PAX5-altered (PAX5alt) B-ALL

a. Genetic alterations including gene rearrangements (PAX5r), sequence mutations (PAX5mut) and focal intragenic amplifications (PAX5amp) observed in PAX5alt cohort. PAX5 mutation zygosity is defined as: heterozygous (Hetero), MAF <0.8; homozygous (Homo), MAF ≥ 0.8. For cases with multiple PAX5 mutations, the highest MAF was used to define zygosity. PAX5 copy number alterations (CNA) were called from cases with SNP array data. All the recurrent PAX5 fusions and sequence mutations are shown in the heatmap and the number of cases are indicated in parentheses. The recurrent mutations mean the same reference amino acids are affected, even with different variant amino acids, like

p.Arg38Cys and p.Arg38His, are shown as p.Arg38; if the variant amino acids are the same, then the full amino acid changes are shown, e.g. p.Ser66Asn. p.*392Arg is a stop loss mutation. fs, frameshift; sp, canonical splice site mutation. **b**, Genetic mutation spectrum of 65 PAX5alt cases with whole genome/exome sequencing data. Samples are ordered primarily based on the key *PAX5* alterations (PAX5r, PAX5mut and PAX5amp) and genes are grouped into specific pathways. AYA, adolescent and young adult; TF, transcription factor; CN-LOH, copy-neutral loss of heterozygosity; WT, wild type; NA, not available; NME, NOTCH1-driven MYC enhancer.

P80R”) is shown between copy number of *PAX5* (marked as “*PAX5* CNA”) and detailed *PAX5* mutations (marked as “*PAX5* (44)”, indicating 44 cases with *PAX5* mutations) to illustrate the fact that homozygous (Homo.) *PAX5* mutations result from loss of the wild-type allele of *PAX5*, while cases with heterozygous (Hetero.) P80R mutations are usually observed with a second hit to disrupt function of the other copy of *PAX5*. WT, wild type; NA, not available.

Figure 4. Distribution of signaling mutations in B-ALL subtypes.

a, Distribution of mutations in 3 key signaling pathways in different B-ALL subtypes. Ras pathway includes sequence mutations from *NRAS*, *KRAS* and *PTPN11*; JAK/STAT pathway includes *JAK1/2/3* and *IL7R* sequence mutations, *JAK2/TYK2*, *EPOR* and *CRLF2* rearrangements. The mutations were called from RNA-seq data. The total sample number (N) for each subtype is indicated in parentheses. **b**, Distribution of frequently mutated signaling genes (according to PAX5 P80R group) in different B-ALL subtypes.

Figure 5. Gene expression signature of PAX5 P80R.

a, Heatmap of top 100 differentially expressed genes (based on two-sided Wald test and Benjamini-Hochberg adjustment, 11 up-regulated and 89 down-regulated) in PAX5 P80R (N=33) vs. other B-ALL subtypes (N=372). Up-regulated genes in PAX5 P80R subtype are listed in the figure. For subtypes with many available samples, only 30 with top RNAseq quality (based on 30X coverage) are included. Full list of the genes in the signature is provided in Supplementary Table 17. Up- and down-regulated genes are ordered in the heatmap according to the significance of the adjusted *P* value. **b**, Venn diagram of differentially expressed genes (2 fold-change and two-sided Wald test and Benjamini-Hochberg adjusted *P*<0.01) in PAX5 P80R (N=33) and PAX5alt groups (N=85) versus other B-ALL (N=372). **c**, Gene set enrichment analysis (GSEA) of PAX5 P80R subtype (N=33) versus PAX5alt group (N=85). Gene set “B cells up” was derived from gene expression profiling of mouse hematopoietic lineages³⁴. False discovery rate (FDR), nominal p-value and normalized enrichment score are calculated by GSEA³⁵.

Figure 6. Event-free (EFS) and overall survival (OS) of PAX5 P80R subtype.

a, Kaplan-Meier estimates of EFS and OS for children with B-ALL treated on COG NCI HR AALL0232 protocol (favorable subtypes includes High hyperdiploid, ETV6-RUNX1 and TCF3PBX1, 132 patients; DUX4, 28; KMT2A, 11; PAX5 P80R, 8; PAX5alt, 46; Ph, 18; Ph-like, 70; Other includes CRLF2 (non-Ph-like), ETV6-RUNX1-like, High hypodiploid, iAMP21, IKZF1 N159Y, MEF2D, NUTM1, ZNF384 and all other, 85). P values were calculated by the two-sided time-stratified Cochran–Mantel–Haenszel test across all the subtypes in each panel. **b**, Kaplan-Meier estimates of EFS and OS for adult B-ALL patients (>18 years) (BCL2/MYC, 7 patients; DUX4, 13; Hypodiploid, 26; KMT2A,

35; PAX5 P80R, 15; PAX5alt, 27; Ph, 31; Ph-like, 59; Other includes High hyperdiploid, CRLF2 (non-Ph-like), ETV6-RUNX1, iAMP21, IKZF1 N159Y, MEF2D, ZNF384 and all other, 75).

Figure 7. PAX5 P80R impairs B cell differentiation and drives development of B-ALL.

a, Flow cytometric immunophenotyping of ex vivo cultures derived from *Pax5*^{-/-} lineage-negative bone marrow cells transduced with empty vector, wild-type *PAX5*, or point mutants within the DNA-binding domain of *PAX5* (p.Pro80Arg (P80R), p.Val26Gly (V26G) and p.Pro34Gln (P34Q)). Cultures were grown on IL7-secreting supportive T220 stromal cells to promote differentiation to B220⁺ CD19⁺ pre-B cells. Each flow panel is a representation of at least three identical but independent experiments. **b**, Kaplan-Meier survival curve for mice harboring *Pax5*^{P80R} or *Pax5*^{G183S} point mutations; ***denotes two-sided log-rank Mantel-

Cox test $P < 0.0001$, $N = 212$ total mice, all weaned mice in the colony were included on study (66 *Pax5*^{+/+}, 11 *Pax5*^{+/-}, 75 *Pax5G183S*^{+/+}, 22 *Pax5G183S/G183S*, 31 *Pax5P80R*^{+/+}, 7 *Pax5P80R/P80R*). **c**, Flow cytometric analysis of bone marrow samples from moribund *Pax5*^{P80R/+} and *Pax5*^{P80R/P80R} mice for lineage markers B220 (B lymphocyte), CD3 (T lymphocyte), Mac 1 (monocyte) and Gr1 (granulocyte), CD41 (megakaryocyte) and Ter119 (erythrocyte) and a Hardy³⁶ B cell panel (CD43, B220, CD19, BP1, IgM) to determine the immunophenotype of leukemic cells. Flow panels are representative of one mouse of each genotype out of a total of 10 *Pax5*^{P80R/+} and three *Pax5*^{P80R/P80R} mice analyzed. **d**, Representative Giemsa-Wright stained bone marrow samples from moribund *Pax5*^{P80R/+} mice; scale bars = 20 μ m; 17 independent *Pax5*^{P80R/+} and 5 *Pax5*^{P80R/P80R} samples were analyzed with similar results **e**, Kaplan-Meier curve of secondary transplant recipient mice ($N = 3$ mice per group). **f**, Array comparative genomic hybridization data for representative *Pax5*^{P80R/+} and *Pax5*^{P80R/P80R} primary tumors, indicating focal and broad deletions/amplifications affecting the *Pax5* locus. Animal IDs are in parentheses. Copy number alterations were detectable in three out of four mice analyzed. **g**, Immunoblot for PAX5, STAT5, and pSTAT5 in mouse fibroblasts (NIH3T3, negative control), B220+ splenocytes, and in vitro cultures of bone marrow cells collected from secondary transplant recipients (717-1, 731-1, 880-2, and 898-1). PAX5 antibodies detecting the N- or C-terminus were used to confirm a truncation observed in 731-1. ACTIN was used as a loading control. Immunoblots were repeated three times. **h**, Gene set enrichment analysis for the PAX5 P80R human B-ALL subtype versus normal human B cells isolated from bone marrow. Gene sets were derived from top 500 up or down regulated genes between *Pax5* P80R ($N = 4$) leukemia cells vs normal B samples ($N = 3$) from mouse model. False discovery rate (FDR), nominal p-value and normalized enrichment score are calculated by GSEA³⁵.

Table 1.

Definition of B-ALL subtypes

Subtype	Case no.	Childhood SR [†]	Childhood HR [†]	AYA [†]	Adult [†]	Class	Criteria [‡]
ETV6-RUNX1	187	21.0%	10.7%	1.4%	0.3%	A	<i>ETV6-RUNX1</i> fusion
KMT2A	136	1.0%	7.8%	4.1%	16.1%	A	<i>KMT2A</i> rearrangements, commonly with <i>AFF1</i> , <i>MLLT1</i> , <i>MLLT3</i> , <i>MLLT10</i> etc.
Ph	123	1.7%	4.6%	5.5%	15.9%	A	<i>BCR-ABL1</i> fusion
DUX4	106	4.9%	5.3%	7.9%	3.2%	A	<i>DUX4</i> rearrangements, commonly with IGH region
TCF3-PBX1	78	5.2%	5.2%	2.9%	1.1%	A	<i>TCF3-PBX1</i> fusion
ZNF384	49	0.6%	3.7%	3.8%	1.3%	A	<i>ZNF384</i> rearrangements, commonly with <i>EP300</i> , <i>TCF3</i> , <i>TAF15</i> , etc.
MEF2D	43	0.6%	3.6%	2.9%	1.1%	A	<i>MEF2D</i> rearrangements, commonly with <i>BCL9</i> , <i>HNRNPUL1</i> , <i>DAZAPI1</i> , <i>SS18</i> , etc.
BCL2/MYC	18	0.2%	0.1%	1.4%	2.6%	A	<i>BCL2</i> , <i>MYC</i> or <i>BCL6</i> rearrangements, commonly with IGH region
NUTM1	11	0.8%	1.0%	0.0%	0.0%	A	<i>NUTM1</i> rearrangements, commonly with <i>ACIN1</i> , <i>CUX1</i> , <i>BRD9</i> , etc.
HLF	9	0.8%	0.3%	0.0%	0.8%	A	<i>HLF</i> rearrangements, commonly with <i>TCF3</i>
High hyperdiploid	279	28.3%	13.9%	6.7%	2.9%	B	Chromosome number 51
Low hypodiploid	78	0.4%	1.2%	4.5%	13.0%	B	Chromosome number 31–39
Near haploid	29	2.9%	1.3%	0.5%	0.8%	B	Chromosome number 24–30
iAMP21	40	2.5%	2.5%	2.1%	0.3%	B	Intrachromosomal amplification of chromosome 21, based on SNP array
Ph-like	359	8.7%	16.9%	29.4%	20.4%	C	Ph PAM score 0.85; no <i>BCR-ABL1</i> fusion
PAX5alt [¶]	148	3.3%	9.3%	9.3%	7.7%	C	Hierarchical gene expression profile cluster enriched with <i>PAX5</i> alterations
PAX5 P80R	44	0.4%	1.9%	3.1%	4.2%	C	<i>PAX5</i> p.Pro80Arg (P80R) mutation or clustered with <i>PAX5</i> P80R subtype
IKZF1 N159Y	8	0.2%	0.4%	0.5%	0.5%	C	<i>IKZF1</i> p.Asn159Tyr (N159Y) mutation
Low hyperdiploid	51	4.3%	1.9%	3.6%	0.3%	C	Hyperdiploid PAM score 0.9; Chromosome number 47–50
ETV6-RUNX1-like	42	4.5%	2.2%	0.7%	0.3%	C	<i>ETV6-RUNX1</i> PAM score 0.98
KMT2A-like	5	0.2%	0.4%	0.2%	0.0%	C	<i>KMT2A</i> PAM score 0.95
ZNF384-like	4	0.0%	0.0%	0.7%	0.3%	C	<i>ZNF384</i> PAM score 0.98
CRLF2(non-Ph-like)	16	1.2%	0.9%	1.0%	0.0%	C	<i>CRLF2</i> fusion; Ph PAM score <0.85
Other	125	6.4%	4.7%	7.9%	7.1%		

[†] Childhood, age range 0.2 to 15 years, standard risk (SR) and high risk (HR) are defined according to National Cancer Institute criteria; adolescent and young adult (AYA), age range 16 to 39 years; adult, age 40 to 79 years.

PAM¹⁵ (Prediction Analysis of Microarrays) is used to calculate the similarity of gene expression profile to specific B-ALL subtypes.

PAX5alt group is defined by hierarchical clustering shown in Supplementary Fig. 1.

Class:

A, defined by gene rearrangements

B, defined by karyotype

C, defined by integration of expression profile, karyotype and/or genetic mutations