ARTICLES
https://doi.org/10.1038/s41588-018-0315-5
nature genetics

PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia

Zhaohui Gu[1,34], Michelle L. Churchman[1,34], Kathryn G. Roberts[1,34], Ian Moore[1], Xin Zhou[2], Joy Nakitandwe[1], Kohei Hagiwara[2], Stephane Pelletier[3], Sebastien Gingras[4], Hartmut Berns[5], Debbie Payne-Turner[1], Ashley Hill[1], Ilaria Iacobucci[1], Lei Shi[6], Stanley Pounds[6], Cheng Cheng[6], Deqing Pei[6], Chunxu Qu[1], Scott Newman[2], Meenakshi Devidas[7], Yunfeng Dai[7], Shalini C. Reshmi[8], Julie Gastier-Foster[8], Elizabeth A. Raetz[9], Michael J. Borowitz[10], Brent L. Wood[11], William L. Carroll[12], Patrick A. Zweidler-McKay[13], Karen R. Rabin[14], Leonard A. Mattano[15], Kelly W. Maloney[16], Alessandro Rambaldi[17], Orietta Spinelli[17], Jerald P. Radich[18], Mark D. Minden[19], Jacob M. Rowe[20], Selina Luger[21], Mark R. Litzow[22], Martin S. Tallman[23], Janis Racevskis[24], Yanming Zhang[25], Ravi Bhatia[26], Jessica Kohlschmidt[27], Krzysztof Mrózek[27], Clara D. Bloomfield[27], Wendy Stock[28], Steven Kornblau[29], Hagop M. Kantarjian[29], Marina Konopleva[29], Williams E. Evans[30], Sima Jeha[31], Ching-Hon Pui[31], Jun Yang[30], Elisabeth Paietta[24], James R. Downing[1], Mary V. Relling[30], Jinghui Zhang[2], Mignon L. Loh[32], Stephen P. Hunger[33] and Charles G. Mullighan[1*]

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); 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.

B-progenitor acute lymphoblastic leukemia (B-ALL) is the most common pediatric malignancy[1], and it consists of multiple subtypes with distinct constellations of inherited and somatic genetic alterations[2]. Genomic analyses, especially transcriptome sequencing (RNA-seq), have identified recurrent chromosomal rearrangements that lead to 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 resulting in expression of chimeric fusion oncoproteins (for example, *ETV6-RUNX1*, *BCR-ABL1* or *TCF3-PBX1*), rearrangements

[1]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, USA. [2]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA. [3]Department of Immunology, St. Jude Children's Research Hospital, Memphis, TN, USA. [4]Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA. [5]Department of Transgenic/Gene Knockout Shared Resource, St. Jude Children's Research Hospital, Memphis, TN, USA. [6]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA. [7]Department of Biostatistics, University of Florida, Gainesville, FL, USA. [8]Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA. [9]Division of Pediatric Hematology-Oncology, New York University, New York, NY, USA. [10]Division of Hematologic Pathology, Johns Hopkins University, Baltimore, MD, USA. [11]Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA, USA. [12]Perlmutter Cancer Center, NYU-Langone Health, New York, NY, USA. [13]ImmunoGen, Inc, Waltham, MA, USA. [14]Baylor College of Medicine, Houston, TX, USA. [15]HARP Pharma Consulting, Mystic, CT, USA. [16]University of Colorado School of Medicine and Children's Hospital, Aurora,, CO, USA. [17]Hematology and Bone Marrow Transplant Unit, Ospedale Papa Giovanni XXIII, Bergamo, Italy. [18]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [19]Princess Margaret Cancer Centre, University Health Network, Toronto, Canada. [20]Hematology, Shaare Zedek Medical Center, Jerusalem, Israel. [21]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA. [22]Division of Hematology, Department of Medicine, Mayo Clinic, Rochester, MN, USA. [23]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [24]Cancer Center, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, USA. [25]Cytogenetics Laboratory, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [26]Division of Hematology-Oncology, University of Birmingham, Birmingham, AL, USA. [27]Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [28]University of Chicago Medical Center, Chicago, IL, USA. [29]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [30]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN, USA. [31]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, USA. [32]Department of Pediatrics, UCSF Benioff Children's Hospital and the Helen Diller Family, San Francisco, CA, USA. [33]Children's Hospital of Philadelphia and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [34]These authors contributed equally: Z. Gu, M. L. Churchman, K. G. Roberts. *e-mail: charles.mullighan@stjude.org

in these new subtypes are commonly not evident in conventional cytogenetic analysis (for example, *DUX4*-rearranged ALL) or involve diverse chromosomal rearrangements that converge on a single gene (for example, *MEF2D*- and *ZNF384*-rearranged ALL)[6,7,10,11]. Additional cases have common transcriptional profiles but diverse genetic alterations (BCR-ABL1-like, also known as Philadelphia chromosome (Ph)-like[13,14], and *ETV6-RUNX1*-like ALL[11]). This has refined the classification of B-ALL and identified new therapeutic targets, such as kinase inhibition in Ph-like ALL[3–5].

Despite these advances, many B-ALL cases cannot be categorized in any of the currently established subtypes. Such cases commonly relapse and lack targeted therapeutic approaches[2]. 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 subjects with B-ALL to identify chromosomal rearrangements, gene expression profiles and large-scale copy number alterations (CNAs; Supplementary Tables 1 and 2). Whole-genome sequencing (WGS; $n=17$), whole-exome sequencing (WES; $n=73$) and SNP ($n=1{,}141$) array data were available for a subset of cases (Supplementary Tables 3 and 4). Gene expression profiles evaluated from RNA-seq data were analyzed by using hierarchical clustering, *t*-distributed stochastic neighbor embedding (tSNE) analysis, and predictive modeling[15] using cases of known subtypes (Fig. 1a and Supplementary Figs. 1 and 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, 76.6% of children with high risk, 71.6% of adolescents and young adults, and 76.2% of adults; age-group definitions are 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 intrachromosomal amplification of chromosome 21 (iAMP21; 3.2% of 1,141 cases with SNP array data)[16,17]. Notably, the gene expression profile of near haploid B-ALL was similar to that of 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]. 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. Some 18 cases (0.9%) harbored rearrangements of *BCL2*, *MYC* and/or *BCL6*, predominantly in adults. These alterations are identical to those observed in double- or triple-hit lymphoma, but have rarely been observed in ALL, and here were of B-cell-precursor immunophenotype[7,18]. We identified 11 cases (0.6%) with rearrangements of *NUTM1* to six different partner genes (three with *ACIN1*, three with *CUX1*, two with *BRD9* and one for each of *IKZF1*, *SLC12A6* and *ZNF618*), 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—for example, *MED12-HOXA9* and *AFF1-TMEM156* in *KMT2A*-like cases, and *TCF3-FLI1*, *FUS-ERG*, *IKZF1* and alternate *ETV6* fusions (two cases with *IKZF1-ETV6*, two with *ETV6-ELMO1* and ten others with different fusion partners; Supplementary Table 1) in *ETV6-RUNX1*-like cases. In addition, cases with a modal chromosome number 47–50 and similar gene expression profiles to those in 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 ($n=16$), owing to the importance of *CRLF2* rearrangement in guiding treatment[5] (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 (NCI) criteria. In the PAX5alt group, 57 cases (38.5% of this group) harbored *PAX5* rearrangements involving 24 partner genes that led to 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 it encodes an in-frame chimeric fusion protein that leads to constitutive activation of 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* and *PAX5-JAK2*), *ETV6-RUNX1* and *IGH-DUX4*, thus indicating that *PAX5-ZCCHC7* is not a leukemia-initiating or subtype-defining event (Supplementary Table 1 and Supplementary Fig. 3a). Together, the *PAX5* rearrangements except *PAX5-JAK2* and *PAX5-ZCCHC7* were significantly enriched in the PAX5alt group ($n=56$, 37.8%) compared with the remaining B-ALL cases ($n=24$, 1.3%; two-sided Fisher's exact test, $P<0.0001$).

Some 46 (31.1% of this group) PAX5alt cases harbored nonsilent *PAX5* sequence mutations, compared with 4.4% ($n=79$) of other B-ALL cases, excluding cases defined by *PAX5* p.Pro80Arg (described 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 (mutant-allele frequency (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 two ($n=14$) or three ($n=1$) mutations (Fig. 2a and Supplementary Table 7). Two hotspot missense alterations affecting amino acids Arg38 ($n=20$) and Arg140 ($n=11$) were identified and were highly enriched in the PAX5alt subtype ($n=11$ and 9, respectively). Notably, 10 of 11 Arg140 missense alterations were concomitant with Arg38 alterations, and nine of the cases with these two alterations were classified as PAX5alt, which account for more than half of the PAX5alt cases with multiple alterations (Figs. 2 and 3a). Among the 203 nonsilent *PAX5* mutations identified in this study, 73.9% were missense mutations, especially those involving the DNA-binding domain (94.6%), whereas 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). We also identified a cluster of mutations in the PAX5 nuclear-localization sequence across the spectrum of B-ALL

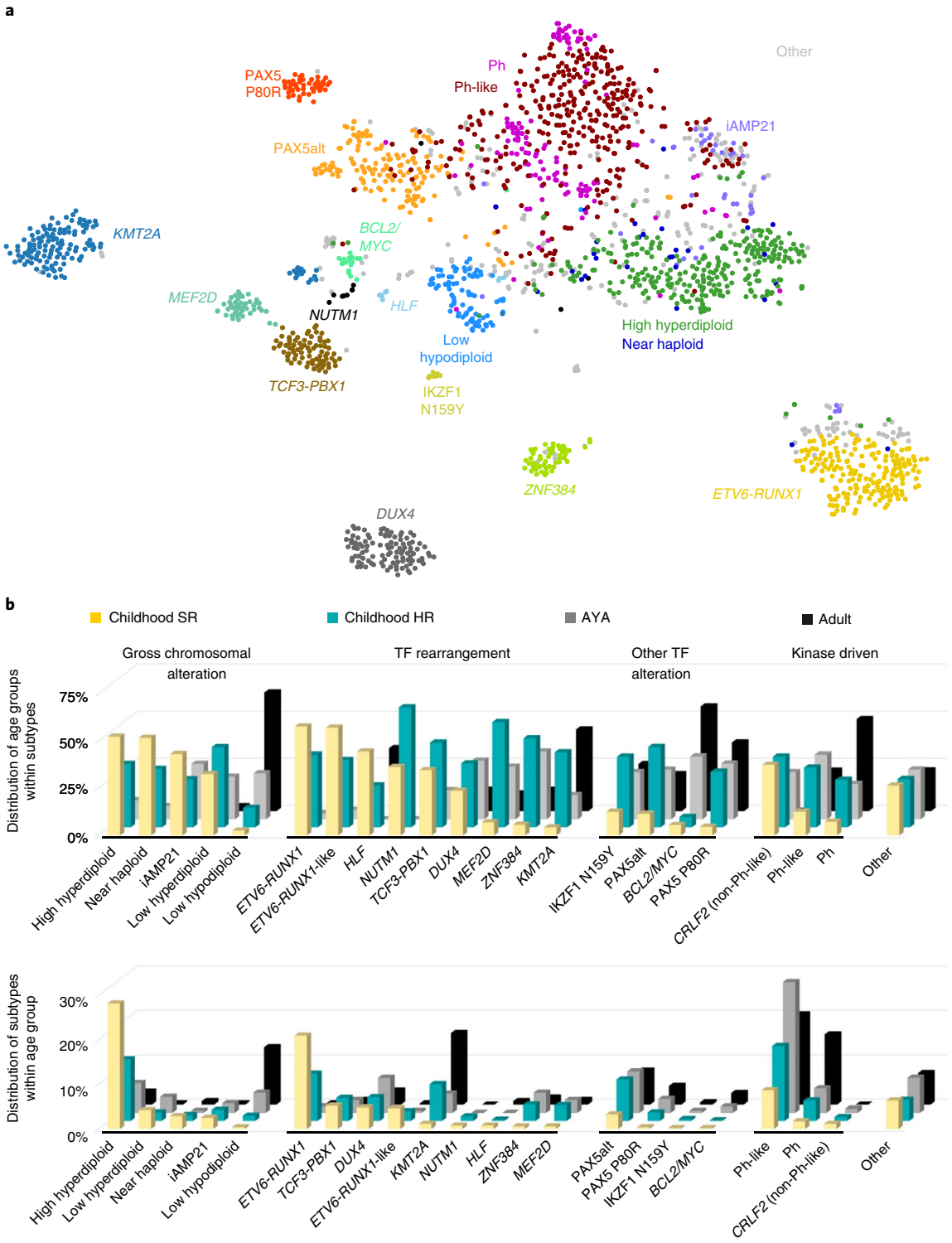

Fig. 1 | Integrative B-ALL subtypes. a, Gene expression profiling of 1,988 cases shown in a two-dimensional tSNE plot. Each dot represents a sample. The top 1,000 most variable genes (on the basis of median absolute deviation) were selected and processed by the tSNE algorithm with a perplexity score 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 gene expression profiling and universal PAX5 p.Pro80Arg (P80R) alteration were observed (PAX5 P80R). A cluster of cases with diverse *PAX5* alterations (PAX5alt) was also observed adjacent to the PAX5 P80R group, with diverse rearrangements, focal intragenic amplifications and non-PAX5 P80R alterations. Eight cases with distinct gene expression profiling were identified with the same *IKZF1* missense alteration encoding 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 are not specifically labeled in the plot. **b**, Distribution of B-ALL subtypes within each subtype (top) or each age group (bottom). Age-group definitions are described in Table 1. Subtypes are grouped as gross chromosomal alteration, transcription factor (TF) rearrangement, other TF alteration, kinase-driven and others. SR, standard risk; HR, high risk; AYA, adolescents and young adults.

Table 1 | Definition of B-ALL subtypes

Subtype	Case no.	Childhood SR[a]	Childhood HR[a]	AYA[a]	Adult[a]	Class	Criteria[b]
ETV6-RUNX1	187	21.0%	10.7%	1.4%	0.3%	A	*ETV6-RUNX1* fusion
KMT2A	136	1.0%	7.8%	4.1%	16.1%	A	*KMT2A* rearrangements, commonly with *AFF1*, *MLLT1*, *MLLT3* and *MLLT10*
Ph	123	1.7%	4.6%	5.5%	15.9%	A	*BCR-ABL1* fusion
DUX4	106	4.9%	5.3%	7.9%	3.2%	A	*DUX4* rearrangements, commonly with IGH region
TCF3-PBX1	78	5.2%	5.2%	2.9%	1.1%	A	*TCF3-PBX1* fusion
ZNF384	49	0.6%	3.7%	3.8%	1.3%	A	*ZNF384* rearrangements, commonly with *EP300*, *TCF3* and *TAF15*
MEF2D	43	0.6%	3.6%	2.9%	1.1%	A	*MEF2D* rearrangements, commonly with *BCL9*, *HNRNPUL1*, *DAZAP1* and *SS18*
BCL2/MYC	18	0.2%	0.1%	1.4%	2.6%	A	*BCL2*, *MYC* or *BCL6* rearrangements, commonly with IGH region
NUTM1	11	0.8%	1.0%	0.0%	0.0%	A	*NUTM1* rearrangements, commonly with *ACIN1*, *CUX1* and *BRD9*
HLF	9	0.8%	0.3%	0.0%	0.8%	A	*HLF* rearrangements, commonly with *TCF3*
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 *BCR-ABL1* fusion
PAX5alt[c]	148	3.3%	9.3%	9.3%	7.7%	C	Hierarchical gene expression profile cluster enriched with *PAX5* alterations
PAX5 P80R	44	0.4%	1.9%	3.1%	4.2%	C	PAX5 p.Pro80Arg (P80R) alteration or clustered with PAX5 P80R subtype
IKZF1 N159Y	8	0.2%	0.4%	0.5%	0.5%	C	IKZF1 p.Asn159Tyr (N159Y) alteration
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	*ETV6-RUNX1* PAM score ≥0.98
KMT2A-like	5	0.2%	0.4%	0.2%	0.0%	C	*KMT2A* PAM score ≥0.95
ZNF384-like	4	0.0%	0.0%	0.7%	0.3%	C	*ZNF384* PAM score ≥0.98
CRLF2 (non-Ph-like)	16	1.2%	0.9%	1.0%	0.0%	C	*CRLF2* fusion; Ph PAM score <0.85
Other	125	6.4%	4.7%	7.9%	7.1%		

[a]Childhood, age range 0.2–15 years, standard risk (SR) and high risk (HR) are defined according to NCI criteria; adolescents and young adults (AYA), age range 16–39 years; adult, age 40–79 years. [b]Prediction analysis of microarrays (PAM)[15] was used to calculate similarity of gene expression profile to specific B-ALL subtypes. [c]PAX5alt group is defined by hierarchical clustering shown in Supplementary Fig. 1. Classes: A, defined by gene rearrangements; B, defined by karyotype; C, defined by integration of expression profile, karyotype and/or genetic mutations.

subtypes; these mutations were predicted to impede translocation of PAX5 to the nucleus (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), which was identified in ten cases, eight of which were in the PAX5alt group (Supplementary Fig. 3b). One of the cases showed that the region of amplification involved exons 2–5 and was predicted to lead to duplication of the PAX5 DNA-binding domain. The structure and consequences of *PAX5* amplification were validated by using WGS, PCR with reverse transcription, Sanger sequencing and fluorescence in situ hybridization, showing in-frame internal tandem duplication from exon 2 to exon 5 of *PAX5* (Supplementary Fig. 4). Together, genetic alterations of *PAX5* were significantly enriched in the PAX5alt group compared with other B-ALL subtypes (73.6% (109 of 148) of PAX5alt versus 5.7% (103 of 1,796) of other samples; two-sided Fisher's exact test, $P<0.0001$), except the PAX5 p.Pro80Arg 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 (for example, *ZFP36L2*, *ETV6* and *LEF1*) and/or epigenetic modification (for example, *KDM6A*, *KMT2A* and *ATRX*; Fig. 2b and Supplementary Tables 8 and 9). Notably, signaling-pathway mutations were observed in 63.1% (41 out of 65 cases with WES or WGS data) cases in this subtype. The gene expression profile of PAX5alt was notable for a preponderance of downregulated genes, as compared with other B-ALL cases (319 upregulated and 2,150 downregulated genes with changes twofold or greater and adjusted $P<0.01$) (Supplementary Tables 10 and 11), thus suggesting that loss of PAX5 transcriptional activation promotes leukemogenesis. Pathway

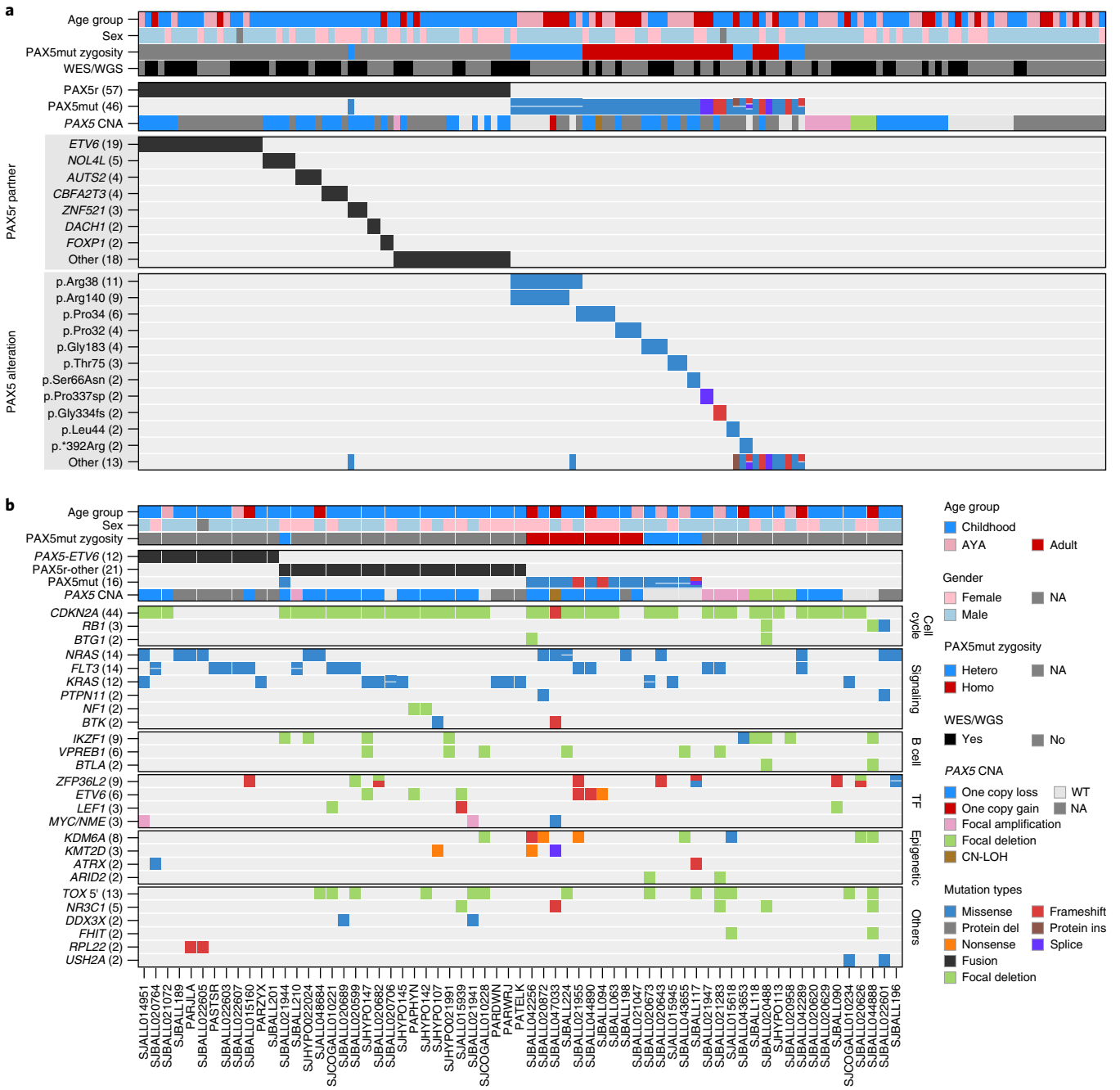

Fig. 2 | Mutational profile of *PAX5*-altered (PAX5alt) B-ALL. a, Genetic alterations including gene rearrangements (PAX5r), sequence mutations (PAX5mut) and focal intragenic amplifications (PAX5amp) observed in the PAX5alt cohort. *PAX5* mutation zygosity is defined as heterozygous (hetero), MAF < 0.8; homozygous (homo), MAF ≥ 0.8. For cases with multiple *PAX5* mutations, the highest MAF was used to define zygosity. *PAX5* CNAs were called from cases with SNP array data. All recurrent *PAX5* fusions and sequence mutations are shown in the heat map and number of cases are indicated in parentheses. Recurrent mutations mean that the same reference amino acids are affected, even with different variant amino acids (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 (for example, 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 WGS or WES data. Samples are ordered primarily on the basis of key *PAX5* alterations (PAX5r, PAX5mut and PAX5amp) and genes are grouped into specific pathways. CN-LOH, copy-neutral loss of heterozygosity; Protein del, in-frame deletion mutation; Protein ins, in-frame insertion mutation; WT, wild type; NA, not available; NME, NOTCH1-driven MYC enhancer.

analysis showed that genes encoding regulators of cytokine receptor signaling were highly enriched, in agreement with the high frequency of mutations in signaling pathways (Supplementary Table 12).

PAX5 p.Pro80Arg defines a distinct subtype of B-ALL. A second group with a distinct gene expression profile was defined by the PAX5 p.Pro80Arg alteration, which was present in all 44 cases, compared with 4 of 1,944 other B-ALL cases (0.2%; Figs. 1a and 3a, Supplementary Table 7 and Supplementary Fig. 1a). In 30 cases, the mutation encoding PAX5 p.Pro80Arg was hemizygous or homozygous, owing to deletion of the wild-type *PAX5* allele or copy-neutral loss of heterozygosity (Fig. 3b,c and Supplementary Tables 13

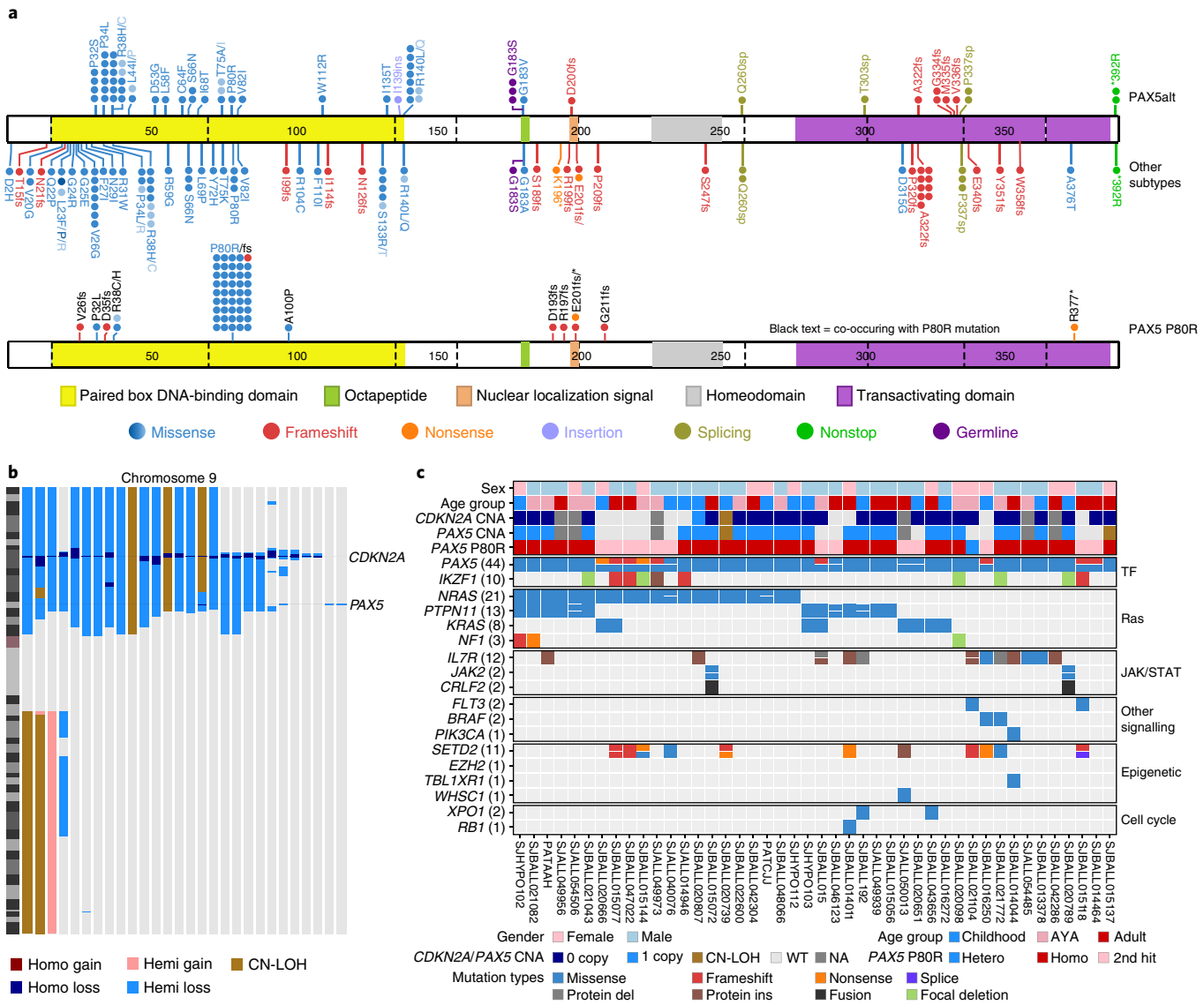

Fig. 3 | Mutational profile of PAX5 P80R B-ALL. a, Protein domain plot of PAX5, showing the 57 alterations detected in 44 subjects in the PAX5 p.Pro80Arg (P80R) subtype (bottom) compared with all the other B-ALL cases (146 mutations in 125 of 1,944 subjects (top), which is further divided for PAX5alt and other B-ALL subtypes). Details of mutations are in Supplementary Table 7. Individual cases are represented by circles; missense mutations affecting the same amino acid residues are shown as graded shades of blue to indicate the number of cases for each substitution. **b**, CNAs identified on chromosome 9 from SNP arrays. Two commonly altered genes (*CDKN2A* and *PAX5*) affected by CNAs are highlighted. Hemi, hemizygous; homo, homozygous. **c**, Genetic mutations including SNVs, indels and CNAs detected from transcriptome sequencing, WGS, WES or SNP array data in the PAX5 P80R group. Genes are ordered according their recurrence and grouped into specific pathways. Zygosity of the PAX5 P80R alteration (PAX5 P80R) is shown between copy number of *PAX5* (*PAX5* CNA) and detailed *PAX5* mutations (*PAX5* (44), indicating 44 cases with *PAX5* mutations) to illustrate that homozygous *PAX5* mutations result from a loss of the wild-type allele of *PAX5*, whereas cases with heterozygous P80R mutations are usually observed with a second hit to disrupt function of the other copy of *PAX5*.

and 14). Of the remaining 14 cases with heterozygous *PAX5* p.Pro80Arg-encoding alterations, 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 p.Pro80Arg alteration, all were heterozygous with preservation of a wild-type *PAX5* allele and had similar gene expression profiles to those of other subtypes (two Ph-like, one BCL2/MYC and one PAX5alt), in agreement with the notion that biallelic *PAX5* alterations, including those encoding p.Pro80Arg, are a hallmark of this subtype.

Collectively, signaling-pathway alterations (in the Ras and JAK/STAT pathways, *FLT3*, *BRAF* and *PIK3CA*) were present in 42 (95.5%) of PAX5 p.Pro80Arg cases, thereby 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 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 nonmutated PAX5 p.Pro80Arg cases harbored alterations in JAK/STAT pathway members, most commonly the interleukin 7 (IL-7) receptor (*IL7R*, $n=7$). We compared the distribution of mutations in each signaling pathway (Ras and JAK-STAT pathway members and *FLT3*) across B-ALL subtype and observed significant enrichment, particularly in Ras and JAK-STAT pathway

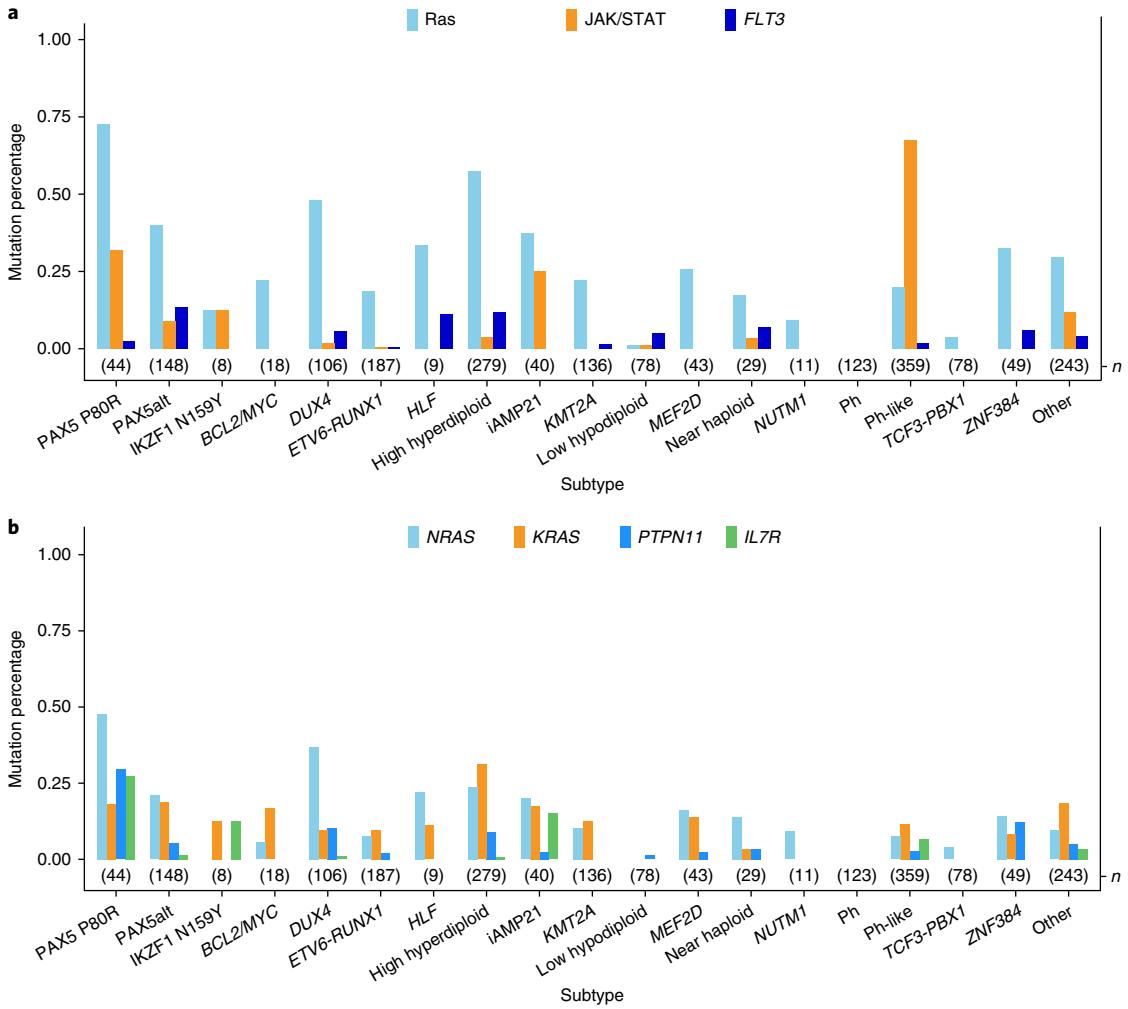

Fig. 4 | Distribution of signaling mutations in B-ALL subtypes. a, Distribution of alterations in members of three key signaling pathways in different B-ALL subtypes. The Ras pathway alterations include sequence mutations from *NRAS*, *KRAS* and *PTPN11*; the JAK/STAT pathway alterations include *JAK1*, *JAK2*, *JAK3* and *IL7R* sequence mutations and *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 the PAX5 P80R group) in different B-ALL subtypes.

members, in the PAX5 p.Pro80Arg group (90.1% of 44 PAX5 p.Pro80Arg versus 35.3% of 1,944 other B-ALL, two-sided Fisher's exact test, $P<0.0001$; Fig. 4a). *NRAS*, *PTPN11* and *IL7R* were most frequently mutated in the PAX5 p.Pro80Arg group compared with other B-ALL subtypes (47.7, 29.5 and 27.3%, respectively; Fig. 4b). We examined coding-sequence mutations genome wide and found an additional target of mutation that was most commonly mutated in PAX5 p.Pro80Arg cases, *SETD2*, which encodes a histone 3 Lys36 trimethylase (25.0% versus 6% of other B-ALL cases)[19,20].

The gene expression profile of PAX5 p.Pro80Arg ALL (334 upregulated and 2,552 downregulated genes with twofold change or greater 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, approximately half ($n=1,224$) of the downregulated genes in these two groups were shared, indicating loss of PAX5 transcriptional activity in both the PAX5 p.Pro80Arg and PAX5alt groups (Fig. 5b). In agreement with the higher frequency of mutations in signaling pathways compared with that in the PAX5alt group, there was greater expression of genes encoding regulators of cytokine-receptor signaling (Supplementary Tables 18–19). Moreover, direct comparison of p.Pro80Arg versus PAX5alt ALL showed negative enrichment of B-lineage genes in p.Pro80Arg ALL, including targets of PAX5 such as BACH2 (ref. [21]), thus indicating that p.Pro80Arg has more profoundly deleterious effects on B cell maturation than the alterations collectively present in PAX5alt ALL (Fig. 5c and Supplementary Table 20). Notably, the *MEGF10* gene (encoding multiple epidermal growth factor–like domains protein 10), which was otherwise silent or poorly expressed in normal B cells and other B-ALL subtypes, was markedly overexpressed in PAX5 p.Pro80Arg cases ($\log_2$ fold change 7.74, adjusted $P=6.13\times10^{-111}$), suggesting that increased expression of this gene may serve as a biomarker and/or driver of this subtype.

IKZF1 p.Asn159Tyr alteration defines a B-ALL subtype. 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. In agreement with this idea, we observed an additional subtype that was defined by a single transcription-factor alteration. Eight cases harbored heterozygous *IKZF1* p.Asn159Tyr-encoding missense alterations, and in contrast to PAX5 p.Pro80Arg ALL, retention of expression of the nonmutated *IKZF1* allele. The gene expression profile was markedly distinct from those of other B-ALL cases, including other *IKZF1*-altered cases (593 upregulated and 1,227 downregulated genes with twofold change or greater and adjusted $P<0.01$; Fig. 1a, Supplementary Fig. 1a and Supplementary Table 21). Asn159 is

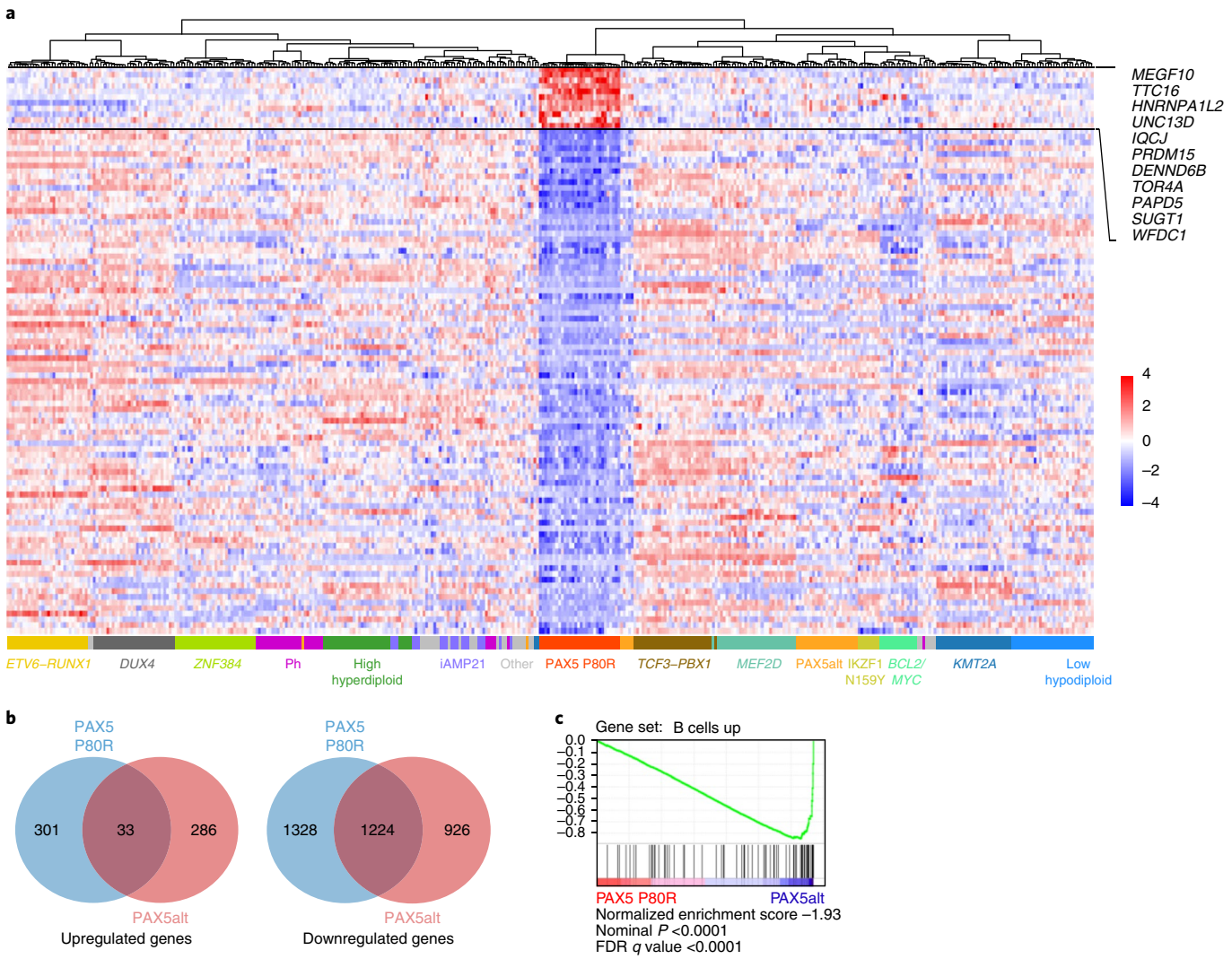

Fig. 5 | Gene expression signature of PAX5 P80R. a, Heat map of the top 100 differentially expressed genes (based on two-sided Wald test and Benjamini–Hochberg adjustment, 11 upregulated and 89 downregulated) in PAX5 P80R ($n = 33$) versus other B-ALL subtypes ($n = 372$). Upregulated genes in the PAX5 P80R subtype are listed in the figure. For subtypes with many available samples, only 30 with top RNA-seq quality (on the basis of 30× coverage) are included. A full list of genes in signature is in Supplementary Table 17. Upregulated and downregulated genes are ordered in the heat map according to the significance of the adjusted *P* value. **b**, Venn diagram of differentially expressed genes (twofold change or greater 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**, GSEA of the PAX5 P80R subtype ($n = 33$) versus PAX5alt group ($n = 85$). The gene set 'B cells up' was derived from gene expression profiling of mouse hematopoietic lineages[34]. The false discovery rate (FDR), nominal *P* value and normalized enrichment score were calculated by GSEA[35].

located in the DNA-binding domain of IKZF1, and we previously showed this mutation to perturb IKZF1 function, with distinctive nuclear mislocalization and induction of aberrant intercellular adhesion that are characteristic of many *IKZF1* alterations[22]. Notably, this subtype exhibited upregulation of genes with roles in oncogenesis (the *IKZF1*-interacting gene *YAP1* (ref. [23])), chromatin remodeling (*SALL1* (ref. [24])), and signaling (*ARHGEF28* (ref. [25])) that were not deregulated in other subgroups of *IKZF1*-altered ALL. Interrogation of exome and DNA copy number data did not identify additional recurrent sequence mutations or focal CNAs, but six of the eight cases had a gain of whole chromosome 21, indicating potential interaction between IKZF1 p.Asn159Tyr and abnormal chromosome 21 in leukemogenesis (Supplementary Tables 1, 4 and 22).

Clinical characteristics and outcomes of novel ALL subtypes. The median ages at diagnosis for subjects in PAX5 p.Pro80Arg and PAX5alt ALL were 22.0 years and 15.4 years, respectively (Supplementary Table 23). Subjects with PAX5 p.Pro80Arg and PAX5alt subtypes had median presenting white-blood-cell counts of $13.0 \times 10^9 l^{-1}$ and $16.9 \times 10^9 l^{-1}$, 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 p.Pro80Arg and PAX5alt cases, respectively (Supplementary Table 24). In children treated in the Children's Oncology Group AALL0232 study of NCI high-risk B-ALL[26], the outcome was intermediate for both PAX5 p.Pro80Arg (5-year event-free survival (EFS) 75.0 ± 14.2%, overall survival (OS) 75.0 ± 14.2%, eight evaluable cases) and PAX5alt (EFS 71.5 ± 7.0%, OS 75.7 ± 6.6%, 46 evaluable cases) compared with DUX4 ALL and other favorable risk subtypes (high hyperdiploid, *ETV6-RUNX1* and *TCF3-PBX1*; Fig. 6a, Supplementary Table 25). In contrast, the outcome for PAX5 p.Pro80Arg in children treated on St. Jude Total Therapy protocols was unfavorable, although few subjects were evaluable and were treated on multiple protocols (Supplementary

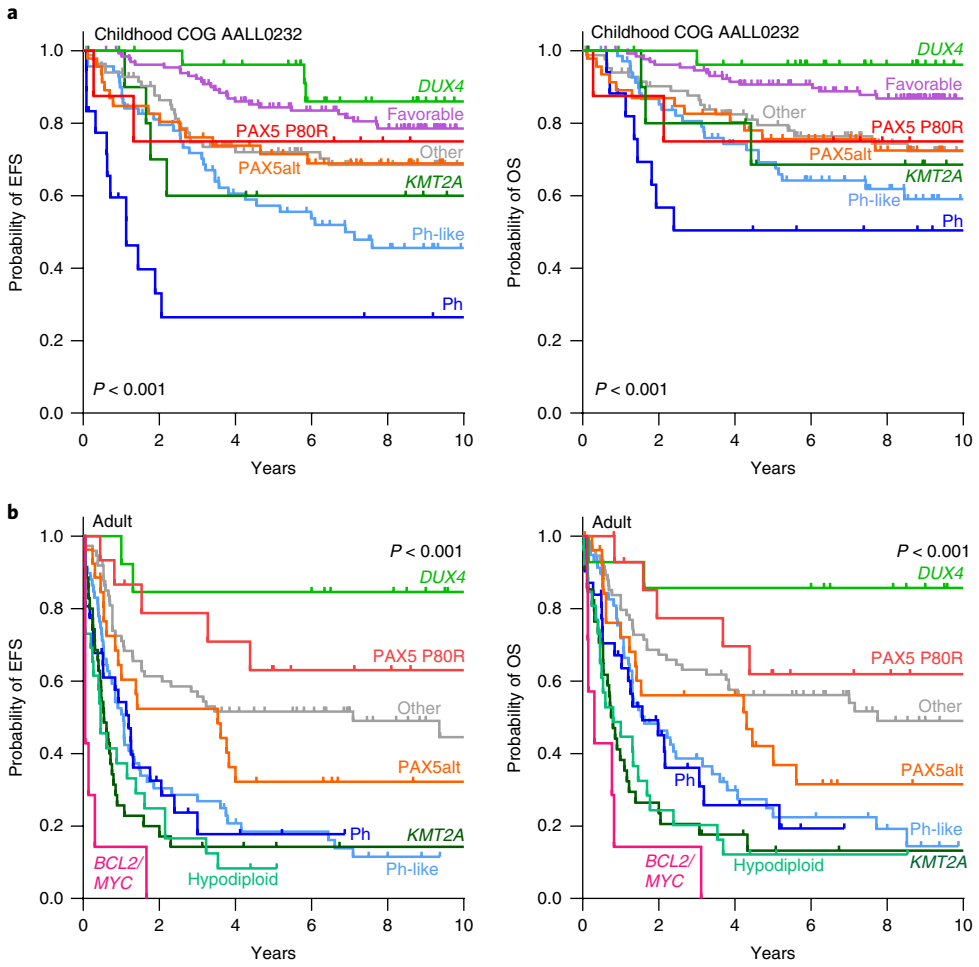

Fig. 6 | Event-free and overall survival of the PAX5 p.Pro80ARg (P80R) subtype. a, Kaplan–Meier estimates of EFS and OS for children with B-ALL treated on the COG NCI HR AALL0232 protocol (favorable subtypes include high hyperdiploid, *ETV6-RUNX1* and *TCF3-PBX1*, 132 subjects; *DUX4*, 28; *KMT2A*, 11; PAX5 p.Pro80Arg (P80R), 8; PAX5alt, 46; Ph, 18; Ph-like, 70; other includes *CRLF2* (non-Ph-like), *ETV6-RUNX1*-like, high hypodiploid, iAMP21, IKZF1 p.Asn159Tyr (N159Y), *MEF2D*, *NUTM1*, *ZNF384* and all others, 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 subjects (>18 years) (*BCL2/MYC*, 7 subjects; *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 others, 75).

Fig. 5 and Supplementary Table 25). Adults with PAX5 p.Pro80Arg ALL (15 of 288 evaluable cases) had intermediate and superior outcomes (EFS 63.0 ± 13.3%, OS 61.9 ± 13.4%) compared with those of subjects with PAX5alt (EFS 32.2 ± 9.4%, OS 42.1 ± 10.2%, 27 evaluable cases). Notably, in adults, the *DUX4* subtype was associated with excellent outcome, and *BCL2*/*MYC* was associated with uniformly early treatment failure (Fig. 6b and Supplementary Table 26).

PAX5 p.Pro80Arg drives B lymphoid leukemogenesis. We previously reported that PAX5 p.Pro80Arg and other point mutations affecting the DNA-binding domain of PAX5 have impaired ability to bind DNA and transcriptionally activate target genes[27]. To examine the effects of PAX5 p.Pro80Arg on B cell maturation, we expressed wild-type PAX5, PAX5 p.Pro80Arg, p.Val26Gly and p.Pro34Gln in *Pax5*$^{-/-}$ lineage-depleted bone marrow cells. Expression of PAX5 p.Val26Gly and p.Pro34Gln led to near complete rescue of B-cell differentiation; however, expression of PAX5 p.Pro80Arg led to 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 p.Pro80Arg or p.Gly183Ser, a germline variant in the octapeptide domain of PAX5 observed in familial B-ALL[28] (Supplementary Fig. 6 and Supplementary Tables 27–28). Heterozygous *Pax5*$^{p.Pro80Arg/+}$ and homozygous *Pax5*$^{p.Pro80Arg/p.Pro80Arg}$ mice developed B220^{+}CD19^{-} B-progenitor leukemia with median latencies of 160 and 83 d, respectively (Fig. 7b–d). Secondary transplantation into sublethally irradiated recipients led to rapid development of leukemia (Fig. 7e). In contrast, the PAX5 p.Gly183Ser octapeptide-domain alteration 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*$^{p.Pro80Arg/+}$ mice (Fig. 7f and Supplementary Table 29). These included a deletion of the entire *Pax5* locus leading to monoallelic expression of *Pax5* p.Pro80Arg, and a truncating frameshift mutation of the wild-type allele (Fig. 7g), recapitulating loss of the wild-type *PAX5* allele observed in human PAX5 p.Pro80Arg ALL. Although no CNA was observed of *Pax5* in one leukemia, the MAF of p.Pro80Arg was 0.98, as determined by RNA-seq analysis, thus suggesting duplication of the mutant allele to homozygosity by copy-neutral loss of heterozygosity. Amplification of the entire *Pax5* locus was observed in a *Pax5*$^{p.Pro80Arg/p.Pro80Arg}$ tumor, accompanied by high expression of

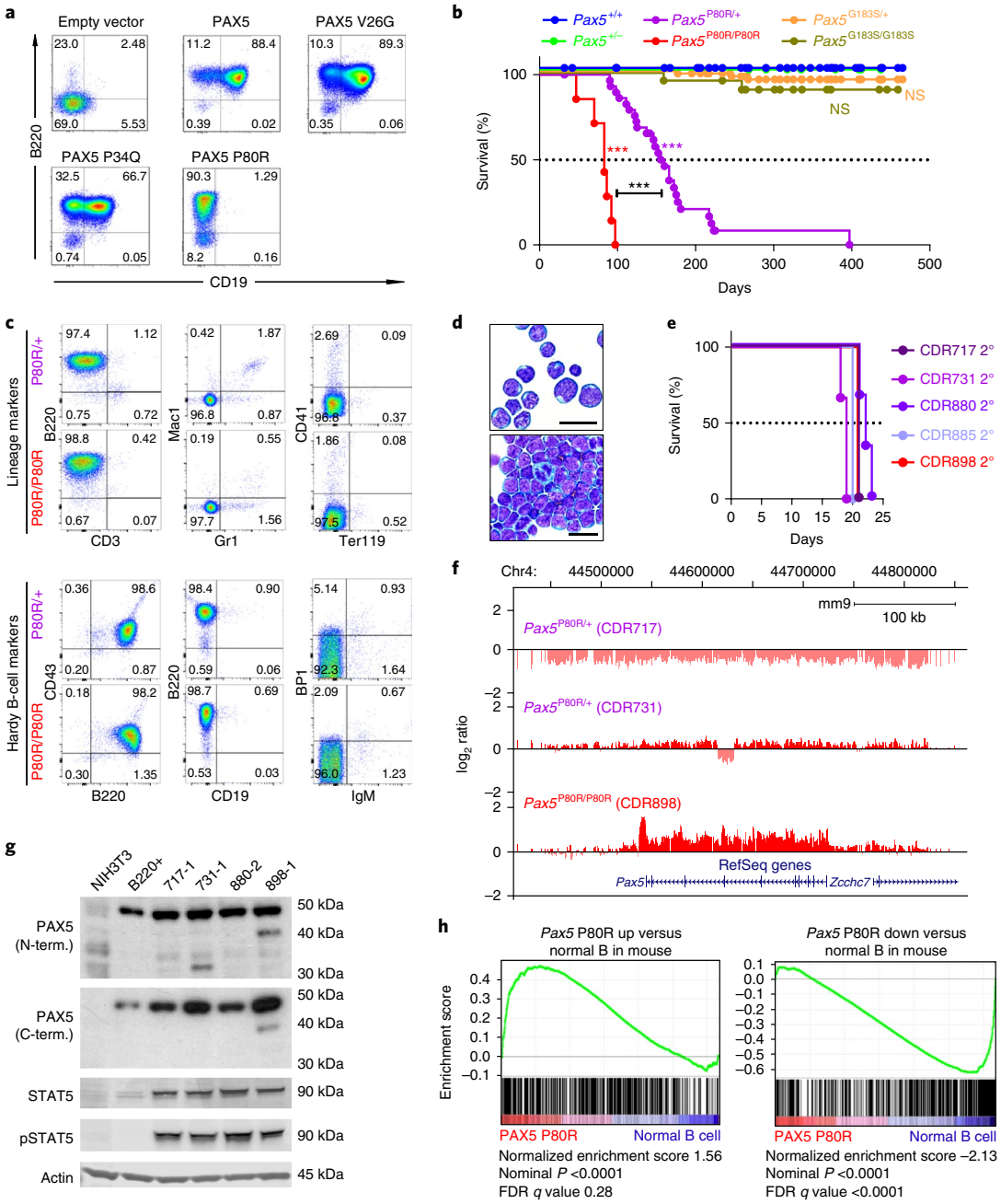

Fig. 7 | PAX5 P80R impairs B cell differentiation and drives development of B-ALL. a, Flow cytometric immunophenotyping of ex vivo cultures derived from *Pax5*$^{-/-}$ lineage-negative bone marrow cells transduced with empty vector, wild-type *PAX5*, or point mutants within the DNA-binding domain of 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 represents at least three identical but independent experiments. **b**, Kaplan–Meier survival curve for mice harboring Pax5 p.Pro80Arg or Pax5 p.Gly183Ser (G183S) point alteration; two-sided log-rank Mantel-Cox test, ***$P < 0.0001$, $n = 212$ total mice; all weaned mice in the colony were included in the study (66 *Pax5*$^{+/+}$, 11 *Pax5*$^{+/-}$, 75 *Pax5*$^{G183S/+}$, 22 *Pax5*$^{G183S/G183S}$, 31 *Pax5*$^{P80R/+}$ and 7 *Pax5*$^{P80R/P80R}$). NS, not significant. **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 Hardy[36] B-cell panel (CD43, B220, CD19, BP1 and IgM) to determine the immunophenotype of leukemic cells. Flow panels represent one mouse of each genotype out of a total of ten *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 µm; 17 independent *Pax5*$^{P80R/+}$ and 5 *Pax5*$^{P80R/P80R}$ samples were analyzed and yielded similar results. **e**, Kaplan–Meier curve of secondary-transplant recipient mice (2°; $n = 3$ mice per group). **f**, Array comparative genomic hybridization data for representative *Pax5*$^{P80R/+}$ and *Pax5*$^{P80R/P80R}$ primary tumors, indicating focal and broad deletions or amplifications affecting the *Pax5* locus. Animal IDs are in parentheses. Copy number alterations were detectable in three of four mice analyzed. Chr, chromosome. **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). Antibodies to PAX5 detecting N or C terminus (N-term. or C-term.) were used to confirm a truncation observed in 731-1. Actin was used as a loading control. Immunoblots were repeated three times. **h**, GSEA for PAX5 P80R human B-ALL subtype versus normal human B cells isolated from the bone marrow. Gene sets were derived from the top 500 upregulated or downregulated genes between Pax5 P80R ($n = 4$) leukemia cells vs normal B samples ($n = 3$) from mouse model. FDR, nominal *P* value and normalized enrichment score were calculated by GSEA[35].

mutant Pax5 p.Pro80Arg (Fig. 7f). Primary tumors harbored multiple *Jak1* and/or *Jak3* mutations (Supplementary Tables 30–32), most of which lead to constitutive activation of JAK/STAT signaling[29]. Cells grown in vitro from secondary transplants showed hyperphosphorylation of STAT5 (Fig. 7g) that was inhibited by ruxolitinib, thus resulting in half-maximal lethal concentration values in cytotoxic assays ranging from 10 to 50 nM (data not shown). Gene set enrichment analysis (GSEA) showed significant similarity of gene expression profiles of mouse and human PAX5 p.Pro80Arg leukemias (Supplementary Table 33 and Fig. 7h). These data support the notion that PAX5 p.Pro80Arg is an initiating lesion and cooperates with activated kinase signaling in leukemogenesis.

Discussion

This study identifies multiple new subtypes of B-ALL that exhibit distinct genomic, clinical and outcome-based features, and variation in prevalence according to age. Although recent studies have identified several new targets of rearrangement in B-ALL (for example, *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 according to gene expression profile clustering, the marked variance in subtype prevalence by age and the observation that transcription-factor missense mutations are initiating, subtype-defining leukemogenic alterations. Moreover, integrative multimodal genomic analysis has distilled the often diverse alterations that define subgroups that had previously defied 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 termed B-other, which lacked a subtype-defining lesion and previously eluded accurate risk stratification. Specifically, the PAX5alt and PAX5 p.Pro80Arg subtypes account for 9.7% of those cases previously termed B-other. Moreover, subtypes associated with unfavorable outcomes, such as *KMT2A*-rearranged, low hypodiploid and kinase-driven ALL account for >65% of adult cases, indicating that 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 have identified frequent *PAX5* alterations in ALL, including rearrangements, focal deletions, sequence mutations and intragenic amplification[27,32]. With the exception of *PAX5* rearrangements and the germline *PAX5* p.Gly183Ser-encoding alteration[28], 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 Pro80 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 activation[27]. In more physiologic assays such as activation of *CD79A* (a PAX5-regulated gene that encodes the Ig-α protein of the B-cell antigen receptor) and expression of surface immunoglobulin, and as shown in this study, B-cell differentiation, PAX5 p.Pro80Arg is profoundly deleterious and leads to 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, by deletion, acquired copy-neutral loss of heterozygosity that duplicates the mutant allele or acquisition of a second *PAX5* sequence mutation that leads to loss of function. The importance of biallelic alterations of *PAX5* in this subgroup is supported by the knock-in *Pax5*$^{p.Pro80Arg}$ mouse model of ALL, in which most tumors acquire second hit alterations of *Pax5*, either by deletion of the wild-type allele or by 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 lead to 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 owing to the genetic heterogeneity of this PAX5alt group. Additional genetic alterations, particularly intragenic *PAX5* amplification and non-p.Pro80Arg sequence alterations, 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 ALL[5]. In this context, the gene expression profile is consistent with that of Ph-like ALL rather than the PAX5alt group. Thus, accurate subgroup assignment requires consideration of transcriptional gene expression profiles 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 p.Pro80Arg and PAX5alt 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$^+$ or 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 p.Pro80Arg ALL. These results are also of immediate diagnostic importance, because they suggest that most 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.

URLs. Interactive portal of this study, https://pecan.stjude.cloud/proteinpaint/study/PanALL; Ensembl, http://www.ensembl.org/; Best practices workflow for calling SNPs and indels from RNA-seq data, http://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq; Picard, http://broadinstitutfor exampleithub.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/; Database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/gap; R version 3.4.3, http://www.r-project.org.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0315-5.

Received: 1 August 2018; Accepted: 13 November 2018;
Published online: 14 January 2019

References

1. Hunger, S. P. & Mullighan, C. G. Acute lymphoblastic leukemia in children. *N. Engl. J. Med.* **373**, 1541–1552 (2015).
2. Iacobucci, I. & Mullighan, C. G. Genetic basis of acute lymphoblastic leukemia. *J. Clin. Oncol.* **35**, 975–983 (2017).
3. Roberts, K. G. et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell* **22**, 153–166 (2012).
4. Iacobucci, I. et al. Truncating erythropoietin receptor rearrangements in acute lymphoblastic leukemia. *Cancer Cell* **29**, 186–200 (2016).
5. Roberts, K. G. et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N. Engl. J. Med.* **371**, 1005–1015 (2014).

6. Zhang, J. et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat. Genet.* **48**, 1481–1489 (2016).
7. Gu, Z. et al. Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 13331 (2016).
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**, 3451–3459 (2016).
9. Gocho, Y. et al. A novel recurrent EP300-ZNF384 gene fusion in B-cell precursor acute lymphoblastic leukemia. *Leukemia* **29**, 2445–2448 (2015).
10. Yasuda, T. et al. Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. *Nat. Genet.* **48**, 569–574 (2016).
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).
12. Lilljebjorn, H. & Fioretos, T. New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood* **130**, 1395–1401 (2017).
13. Den Boer, M. L. et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet. Oncol.* **10**, 125–134 (2009).
14. Mullighan, C. G. et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360**, 470–480 (2009).
15. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002).
16. Harrison, C. J. et al. An international study of intrachromosomal amplification of chromosome 21 (iAMP21): cytogenetic characterization and outcome. *Leukemia* **28**, 1015–1021 (2014).
17. Holmfeldt, L. et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat. Genet.* **45**, 242–252 (2013).
18. Johnson, N. A. et al. Lymphomas with concurrent BCL2 and MYC translocations: the critical factors associated with survival. *Blood* **114**, 2273–2279 (2009).
19. Zhu, X. et al. Identification of functional cooperative mutations of SETD2 in human acute leukemia. *Nat. Genet.* **46**, 287–293 (2014).
20. Mar, B. G. et al. Mutations in epigenetic regulators including SETD2 are gained during relapse in paediatric acute lymphoblastic leukaemia. *Nat. Commun.* **5**, 3469 (2014).
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).
22. Churchman, M. L. et al. Efficacy of retinoids in IKZF1-mutated BCR-ABL1 acute lymphoblastic leukemia. *Cancer Cell* **28**, 343–356 (2015).
23. Hu, Y., Yoshida, T. & Georgopoulos, K. Transcriptional circuits in B cell transformation. *Curr. Opin. Hematol.* **24**, 345–352 (2017).
24. Lauberth, S. M. & Rauchman, M. A conserved 12-amino acid motif in Sall1 recruits the nucleosome remodeling and deacetylase corepressor complex. *J. Biol. Chem.* **281**, 23922–23931 (2006).
25. Miller, N. L. et al. A non-canonical role for Rgnef in promoting integrin-stimulated focal adhesion kinase activation. *J. Cell. Sci.* **126**, 5074–5085 (2013).
26. Larsen, E. C. 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**, 2380–2388 (2016).
27. Mullighan, C. G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
28. Shah, S. et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat. Genet.* **45**, 1226–1231 (2013).
29. Dang, J. et al. Pax5 is a tumor suppressor in mouse mutagenesis models of acute lymphoblastic leukemia. *Blood* **125**, 3609–3617 (2015).
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**, 1589–1607 (1992).
31. Urbanek, P., Wang, Z. Q., Fetka, I., Wagner, E. F. & Busslinger, M. Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking Pax5/BSAP. *Cell* **79**, 901–912 (1994).
32. Kuiper, R. P. 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**, 1258–1266 (2007).
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).
34. Novershtern, N. et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
35. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
36. Hardy, R. R. & Hayakawa, K. B cell development pathways. *Annu. Rev. Immunol.* **19**, 595–621 (2001).

Acknowledgements

We thank the Biorepository, the Genome Sequencing Facility of the Hartwell Center for Bioinformatics and Biotechnology, and the Cytogenetics core facility of 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 Award (to Z.G.), St. Baldrick's Foundation Robert J. Arceci Innovation Award (to C.G.M.), Amgen, Inc. to ECOG-ACRIN, NCI Outstanding Investigator Award R35 CA197695 (to C.G.M.), National Institute of General Medical Sciences grant P50 GM115279 (to C.G.M.), NCI grants P30 CA021765 (St. Jude Cancer Center Support Grant), ECOG-ACRIN Operations Center grants CA180820 (to P. 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).

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.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0315-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.G.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Subjects and clinical treatment protocols. The subjects were enrolled at St. Jude Children's Research Hospital (SJCRH), 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 subjects include the St. Jude Children's Research Hospital Total XV[37]; (ClinicalTrials.gov identifier NCT00137111) and Total XVI protocols (NCT00549848); the COG P9906 high-risk B-ALL study[38]; the COG AALL0232 high-risk ALL study[26] (NCT00075725); the COG AALL0331 standard-risk ALL study (NCT00103285); the ECOG-ACRIN E2993 (ref. [39]); (NCT00002514) and E1910 trials (NCT02003222); the MD Anderson Cancer Center protocols[40–43]; the Alliance–Cancer and Leukemia Group B protocols C19802 (NCT00003700), C10102 (NCT00061945) and C10403 (NCT00558519); and SWOG protocols S0333 (NCT00109837) and S9400 (NCT00002665). Detailed clinical information for each case is provided in Supplementary Table 1. The subjects 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 was performed using TruSeq library preparation and HiSeq 2000 and 2500 sequencers (Illumina). All sequence reads were paired-end, and were performed by using (i) total RNA and stranded RNA-seq (75- or 100-base pair (bp) reads); (ii) poly(A)-selected mRNA (50-, 75- or 100-bp reads). Sequencing reads were mapped to the GRCh37 human genome reference by STAR[44] (version 2.4.2a), through the suggested two-pass mapping pipeline. Gene annotation downloaded from the 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 ± s.d., 37.2 ± 7.5%). CICERO[5] and FusionCatcher[45,46] were used to detect fusions, and all the reported rearrangements were manually reviewed to keep the reliable ones. Owing to the complexity of *DUX4* rearrangements, some of the *DUX4* fusions were manually rescued by checking the aligned reads within IGV browser[47]. RNA extracted from flow-sorted normal B lymphoid cells were used for RNA-seq and details are in our previous study[48].

Gene expression level evaluation from RNA-seq data. To evaluate gene expression level, the read count for each annotated gene was calculated with the HTSeq package[49], and gene expression level normalization and differential expression analysis were carried out with the DESeq2 Bioconductor R package[50]. To evaluate the digital gene expression level, a regularized log-transformed (rlog) value was calculated by DESeq2. The ComBat function in the sva R package[51] was used to correct the batch effect introduced by different library preparation strategies and sequencing lengths. With the rlog gene expression level, the R package Rtsne was used to map the samples to a two-dimensional tSNE plot with the top 1,000 most variable genes (on the basis of median absolute deviation), and the tSNE perplexity parameter was set to 30. Different gene numbers (200, 500, 1,000 and 2,000) and tSNE parameters (perplexity of 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 detection from RNA-seq data. RNA-seq data are not optimal for calling genomic-level CNAs but are still informative for evaluating chromosome or arm-level copy number changes for cases lacking karyotypic or DNA array data. To accomplish the latter, we ordered the genes on the basis of the median absolute deviation of their expression level across the samples, and then picked a subset (one-fourth to one-third) of the genes with least median absolute deviation as stably expressed genes. To assist in the CNA calling, the MAF of single-nucelotide variants (SNVs) detected from RNA-seq data was also plotted against the gene expression levels of the stably expressed genes to double-check whether the CNAs were reliable (Supplementary Fig. 7). This strategy was mainly applied to resolve the B-ALL subtyping issue for potential hyperdiploid 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 and RNA-seq data (Supplementary Table 34). In total, 295 CNAs were called from the SNP array data, and 285 (96.6%) could be faithfully recapitulated by RNA-seq data, indicating that the application of RNA-seq to evaluate chromosomal level CNAs was highly reliable. One false-positive CNA called from RNA-seq was on the X chromosome, and ten false-negative CNAs missed by RNA-seq were mainly on sex chromosome X ($n = 8$) and 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 GATK[52] forum (see URLs). In general, STAR-mapped bam files were processed by Picard (see URLs, version 1.129) to mark duplicate reads, and then 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 the following criteria were applied to retain the high-quality mutations: (i) at least three reads supported the mutation, and the MAF was ≥0.05; (ii) the mutation was not observed in the common SNP database dbSNP 150; (iii) the mutation was not observed in two or more samples from our germline exome sequence cohort (775 samples). After filtering, all the mutations were annotated to RefSeq genes, and nonsilent mutations that have been previously validated according to the COSMIC database V84 (see URLs) and/or PeCan portal (see URLs) were kept for further analysis.

Whole-genome and whole-exome sequencing. WGS of leukemia and paired germline samples was carried out by TARGET (Therapeutically Applicable Research To Generate Effective Treatments program; see URLs) and SJCRH. WES was performed at SJCRH.

For WGS data generated by TARGET, methods for DNA preparation, sequencing and quality control are available at the TARGET Project portal (see URLs). For WGS performed at SJCRH, genomic DNA was quantified by 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 with a 2100 BioAnalyzer High Sensitivity kit (Agilent Technologies) or Caliper LabChip GX DNA High Sensitivity Reagent kit (PerkinElmer). Libraries were quantified by 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 with Nextera rapid exome kit (Illumina), by using the Caliper Biosciences (Perkin Elmer) Sciclone G3. First-round PCR (ten cycles) was performed with Nextera kit reagents (Illumina), and cleanup steps used BC/Agencourt AMPure XP beads. Target capture was done with a Nextera rapid capture exome kit (Illumina) and the supplied hybridization and associated reagents. Library quality control was performed with a Victor fluorescence plate reader with Quant-it dsDNA reagents for prepool quantification, and a Bio-analyzer 2200 (Agilent) was used for final library quantification. Paired-end sequencing was performed with a HiSeq 2000 or 2500 (Illumina) instrument with a read length of 100 bp.

WGS and WES read alignment and quality control. Paired-end WGS and WES reads were mapped to human reference genome GRCh37 by BWA[53] (version 0.7.12). Samtools[54] (version 1.3.1) was used to generate chromosomal coordinate–sorted and indexed bam files, which were then processed by the Picard (see URLs, version 1.129) MarkDuplicates module to mark PCR duplications. Then the reads were realigned around potential indel regions by the GATK[55] (version 3.7) IndelRealigner module. Sequencing depth and coverage was assessed based on coding regions (~34 Mb) defined by RefSeq genes.

WGS and 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 (i) reported common SNPs or indels from dbSNP v150 and (ii) germline mutations detected from matched germline control samples. All the nonsilent SNVs or indels that 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 detection from microarrays. SNP microarray data from two different platforms were used in this study: Illumina Infinium Omni2.5 Exome-8 (2.6 million probes) and 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 the SNP 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 and B allele frequency 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 p.Pro80Arg samples were converted to log R ratio and B allele frequency values by following the pipeline suggested by PennCNV[57] (see URLs). With the input of log R ratios and B allele frequencies, somatic CNA and loss of heterozygosity from paired or unpaired samples were called by OncoSNP[58] (version 2.1) and manually reviewed by ShinyCNV[59]. 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 counts from RNA-seq data were imported to DESeq2 for differential gene expression analysis. To perform GSEA[35], we defined the gene expression profile of PAX5 p.Pro80Arg by comparing gene expression levels between PAX5 p.Pro80Arg ALL and normal B cells purified from human bone marrow, and ranking genes according to fold change and significance. GSEA was performed by using mSigDB C2 genes and curated gene sets from in-house analyses. Significantly regulated genes (twofold change or

greater and adjusted $P < 0.01$) were selected for Kyoto Encyclopedia of Genes and Genomes pathway and gene-ontology enrichment analysis by GO-Elite[60].

Animals. Mice were housed in an American Association of Laboratory Animal Care–accredited facility, and all experiments were approved and were in compliance with the SJCRH Institutional Animal Care and Use Committee–approved protocol in accordance with National Institutes of Health guidelines. *Pax5*$^{-/-}$ mutant mice[31] were provided by M. Busslinger and maintained on the C56BL/6 background. Genotyping was determined by PCR analysis as described[61].

***Pax5*$^{p.Pro80Arg}$ and *Pax5*$^{p.Gly183Ser}$ mouse lines.** *Pax5*$^{p.Pro80Arg}$ mice were generated by using CRISPR–Cas9 technology. Pronuclear-stage C57BL/6NJ zygotes were injected by the SJCRH Transgenic–Gene Knockout Shared Resource with a single-guide RNA (sgRNA) (Pax5_P80R_Guide 01, 50 ng μl^{-1}) 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^{-1}) and a 200-nt-long single-stranded DNA molecule containing the desired mutations (Pax5-P80R-HDR, 2 pmol μl^{-1}, Supplementary Table 27). Approximately 25 injected zygotes were surgically transplanted into an infundibulum of 0.5-d-postcoitus pseudopregnant CD-1 females, and newborn mice carrying the *Pax5*$^{p.Pro80Arg}$ allele were identified by PCR and Sanger sequencing by using primers Pax5-e3-F1 and Pax5-e3-R1. A similar strategy was used to generate *Pax5*$^{p.Gly183Ser}$ mice. *Pax5*$^{p.Gly183Ser}$ sgRNA (Pax5_G183S_Guide 01), repair template (Pax5_G183S_HDR) and PCR primer (Pax5-e5-F1 and Pax5-e5-R1) sequences are shown in Supplementary Table 27. sgRNAs were designed and generated as described[62]. Cas9 mRNA transcripts were also generated as described[62]. The target site for each sgRNA is unique in the mouse genome, and no potential off-target site with fewer than three mismatches was found by using the Cas-OFFinder algorithm[63] (Supplementary Table 28). *Pax5*$^{p.Pro80Arg}$ and *Pax5*$^{p.Gly183Ser}$ 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 killed, 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, B220-eFluor605, CD3-APC, CD41-PE and Ter119-V500) and Hardy panel B-lymphocyte markers (CD43-PerCP-Cy5.5, B220-eFluor605, CD19-APC-Cy7, BP1-APC and IgM-PE-Cy7) to determine the immunophenotype of leukemic cells. For western blotting, cells were lysed in RIPA buffer, subjected to SDS–polyacrylamide gel electrophoresis, and probed with antibody to PAX5 (Millipore; 05–1573, clone 1H9).

Retroviral vectors and retrovirus production. Vector production was performed as described by Mullighan et al[27]. Briefly, the coding regions of the exon 1a isoforms of wild-type and mutant *PAX5* were cloned from B-ALL subjects 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 h after Eco Phoenix cells were plated in complete DMEM medium, the cells were transfected with wild-type *PAX5*, mutant *PAX5* or MIR plasmid DNA by using FuGENE 6 according to the manufacturer's instructions (Roche Diagnostics). 24 h later, the medium was removed and replaced with complete IMDM. Viral supernatant was collected starting at 48 h after transfection, filtered through a 0.45-μm filter (Millipore), divided into aliquots and frozen at −80 °C until use. Virus titration was performed by transduction of NIH3T3 fibroblast cells and quantification of red fluorescent protein (RFP) expression at 48 h. Viral titers ranged from 10^5 to 10^6 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- to 12-d-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) and magnetically separated by using DynaMag-15 (Invitrogen Life Technologies), per the manufacturer's instructions. The lineage-negative (Lin$^-$) cells were then prestimulated for 48 h with 50 ng ml^{-1} stem cell factor, 50 ng ml^{-1} Flt3L, 30 ng ml^{-1} IL-6, 20 ng ml^{-1} IL-3 and 20 ng ml^{-1} IL-7 from PeproTech. Up to 2 million Lin$^-$ cells were transduced in Retronectin (Takara)-coated plates preloaded with viral supernatants and cultured in the presence of cytokines for 2 d. RFP$^+$ cells were isolated by flow-activated sorting and cultured for 13 d 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 monoclonal antibody to B220, CD19, BP-1, IgM, CD5 (Ly-1), Ly-6G (Gr-1), CD45R/B220 or TER119 (BD Biosciences). Staining of cells was performed by using standard protocols, and analysis was done in the presence of 4,6-diamidino-2-phenylindole (DAPI) nucleic acid to exclude dead cells. Cell sorting was done on a FACSVantage or FACSDiva Cell Sorter (BD Biosciences). Data were analyzed by using FlowJo software (Treestar) and are expressed as the percentage of positive cells for the specific B-ell antigens. Each experiment was repeated at least three times.

Statistical analysis. Associations among categorical values were examined using two-sided Fisher's exact test. Associations among B-ALL subtypes and EFS and OS were examined by the Kaplan–Meier estimator, with Peto's estimator of standard deviation and the two-sided time-stratified Cochran–Mantel–Haenszel test[64]. 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 model[65]. Analyses were performed by using Prism version 7.0 (GraphPad Software), R version 3.4.3 (see URLs) and SAS software version 9.4 (SAS Institute).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

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

References

37. Pui, C. H. et al. Treating childhood acute lymphoblastic leukemia without cranial irradiation. *N. Engl. J. Med.* **360**, 2730–2741 (2009).
38. Bowman, W. P. 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**, 569–577 (2011).
39. Goldstone, A. H. 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**, 1827–1833 (2008).
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**, 2788–2801 (2004).
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**, 2070–2077 (2010).
42. Thomas, D. A. et al. Treatment of Philadelphia chromosome-positive acute lymphocytic leukemia with hyper-CVAD and imatinib mesylate. *Blood* **103**, 4396–4407 (2004).
43. Thomas, D. A. 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**, 3880–3889 (2010).
44. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
45. Nicorici, D. et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Preprint at https://www.biorxiv.org/content/early/2014/11/19/011650 (2014).
46. Edgren, H. et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome. Biol.* **12**, R6 (2011).
47. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
48. Alexander, T. B. et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373–379 (2018).
49. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
50. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome. Biol.* **11**, R106 (2010).
51. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
52. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

54. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
55. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
56. Pounds, S. et al. Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* **25**, 315–321 (2009).
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**, 1665–1674 (2007).
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**, 2482–2484 (2013).
59. Gu, Z. & Mullighan, C. G. ShinyCNV: a Shiny/R application to view and annotate DNA copy number variations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/bty546 (2018).
60. Zambon, A. C. et al. Go-elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* **28**, 2209–2210 (2012).
61. Nutt, S. L., 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**, 476–491 (1997).
62. Pelletier, S., Gingras, S. & Green, D. R. Mouse genome engineering via CRISPR-Cas9 for study of immune function. *Immunity* **42**, 18–27 (2015).
63. Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
64. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–170 (1966).
65. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Series B Stat. Methodol.* **34**, 187–220 (1972).

nature research

Corresponding author(s): Charles G. Mullighan

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	See below for software used. The software used for data collection also used for some level of data analysis.
Data analysis	STAR (version 2.4.2a) CICERO (in-house version) FusionCatcher (version 1.0) HTSeq (version 0.6.0) R programming language (version 3.4.3) DESeq2 R package (verison 1.18.1, RNA-seq gene expression analysis) sva R package (version 3.26.0, batch effect correction) Rtsne R package (version 0.13, tSNE analysis) BWA (version 0.7.12) Samtools (version 1.3.1) Picard (version 1.129) GATK (version 3.7) Variant Effect Predictor (VEP https://useast.ensembl.org/Tools/VEP) Pamr R package (versoin 1.55, PAM prediction) Genome Studio(Illumina, version 2.0.3) GSEA2 (version 2.2.4) GOELite (version 1.2.5)

OncoSNP(CNV calling, version 2.1)
PennCNV (http://penncnv.openbioinformatics.org/en/latest/user-guide/affy/, SNP data processing)
ShinyCNV (https://github.com/gzhmat/ShinyCNV, version 1.1)
SAS software (version 9.1.2)
Prism (GraphPad, version 7.0)
BD FACSDiva (version 8.0.1)
FlowJo (version 10)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The raw and analyzed data are provided in a graphical, interactive platform https://pecan.stjude.cloud/proteinpaint/study/PanALL.
Genomic data generated for this study are deposited in the European Genome-phenome Archive (EGA) under accession number EGAS00001003266.
Other legacy data used in this study could be found at EGA under accession number EGAS00001000654, EGAS00001001952, EGAS00001001923, EGAS00001002217 and EGAS00001000447. The TARGET genomic data used in this study could be found through the TARGET website at https://ocg.cancer.gov/programs/target and also available at the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) under accession number phs000218 (TARGET)
The other data supporting this study are available from the corresponding author upon request.
Figures that have associated raw data: Figures 1,2,3,4,5; Supplementary Figures 1,3,7

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was specifically performed. However, to identify novel B-ALL subtypes, even the rare ones that are around 1% of B-ALL, we aim to collect around 2,000 samples, which end up with 1,988 high-quality RNA-seq samples.
Data exclusions	RNA-seq data with low sequencing coverage (30-fold coverage <15%) were excluded since insufficient coverage and depth will lead to unreliable and biased evaluation of gene expression level. We have tested and established this criteria in our previous studies (Gu et al., Nat Commun, 2016; Alexander et al., Nature, 2018).
Replication	No replication of genomic analysis was performed in analysis.
Randomization	Nothing to disclose. This is not an intervention study, so randomization is not applicable.
Blinding	Nothing to disclose. This is not an intervention study, so blinding is not applicable.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

BD Biosciences: Mac1-Alexa700 (cat.# 557960, clone M1/70, 1:50), Gr1-PerCP-Cy5.5 (cat.# 552093, clone RB6-8C5), B220-eFluor605 (cat.# 563708, clone RA3-6B2), CD3-APC (cat.# 553066, clone 145-2C11, 1:50), CD41-PE (cat.# 561850, clone MWReg30, 1:50), Ter119-V500 (cat.# 562120, clone TER-119, 1:50), CD43-PerCP-Cy5.5 (cat.# 562865, clone S7, 1:50), CD19-APC-Cy7 (cat.# 557655, clone 1D3, 1:50), BP1-APC (cat.# 553735, BP-1, 1:50), IgM-PE-Cy7 (cat.# 552867, clone R6-60.2, 1:50), CD5-biotin (cat.# 553019, clone 53-7.3, lot 63282, 1:100) , Ly-6G-biotin (cat.# 553125, clone RB6-8C5, lot 13056, 1:100), CD45R/B220-biotin (cat.# 553086, clone RA3-6B2, lot 08258, 1:100) and TER119-biotin (cat.# 553672, clone TER-199, lot 7001, 1:100); Millipore: Pax5 N-term (cat.# 05-1573, clone 1H9, lot 2930158, 1:500); Santa Cruz: Pax5 C-term (cat.# sc-13146, clone A-11, lot B0818, 1:500)

Validation

All antibodies are validated for detecting mouse proteins by the manufacturer and confirmed for each specific application using cells of known origin and differentiation state and compared to isotype controls and cells that are known to express or lack the antigen. Fluorescence-labeled antibodies used for flow cytometric analysis and validated by the SJCRH Flow core facility. Biotinylated antibodies were used for magnetic separation of labeled lineage-specified bone marrow cells. Pax5 antibody was used for western blotting and was confirmed by lack of staining in NIH3T3 fibroblasts and specific detection of the appropriate size band in hematopoietic cells and cells transduced to overexpress PAX5.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

All weaned mice in the Pax5 P80R and G183S colonies were included on the survival/leukemogenesis study (total of 212 mice; 66 Pax5+/+, 11 Pax5+/-, 75 Pax5G183S/+, 22 Pax5G183S/G183S, 31 Pax5P80R/+, 7 Pax5P80R/P80R). Both male and females were included, ranging in age from 32 to 465 days old at completion of the study (mean = 284.7 days, S.D. 111.7 days; median = 290 days). Donor mice for isolation of lineage-negative bone marrow were 8-10 week old females. Mice for all studies were on a pure C57/Bl6 background.

Wild animals

This study did not involve wild animals.

Field-collected samples

This study did not involve field-collected samples.

Human research participants

Policy information about studies involving human research participants

Population characteristics

This study did not actually involve human subjects. We collected banked, de-identified tumor samples that were previously collected from pediatric and adult patients with confirmed B-ALL.

Recruitment

Not applicable

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Bone marrow was flushed from the hind legs of mice and spleens were physically dissociated into RPMI media. Cells were centrifuged and resuspended in 10%DMSO in fetal calf serum for freezing vials slowly overnight at -80 degrees Celsius in

	cryocontainers filled with isopropanol . For flow analysis, cells were thawed and stained for surface markers. Normal bone marrow was used for isotype and antibody controls for comparison to tumor samples.
Instrument	Flow cytometric analysis and sorting was performed with FACSVantage or FACSDiva Cell Sorter (BD Biosciences)
Software	Data was analyzed using FlowJo10 software (Treestar Inc.)
Cell population abundance	100% of the post-sort fraction were the relevant RFP+ cells.
Gating strategy	Live cells were gated by Dapi vs FSC which were further gated to singlets by FSC vs SSC. Isotype controls and untransduced cells were used for setting gates for negative populations.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

nature research

Corresponding author(s): Charles G. Mullighan

Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data presentation

For all flow cytometry data, confirm that:

☒ 1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ 2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ 3. All plots are contour plots with outliers or pseudocolor plots.

☒ 4. A numerical value for number of cells or percentage (with statistics) is provided.

Methodological details

5. Describe the sample preparation.	Bone marrow was flushed from the hind legs of mice and spleens were physically dissociated into RPMI media. Cells were centrifuged and resuspended in 10%DMSO in fetal calf serum for freezing vials slowly overnight at -80 degrees Celsius in cryocontainers filled with isopropanol . For flow analysis, cells were thawed and stained for surface markers. Normal bone marrow was used for isotype and antibody controls for comparison to tumor samples.
6. Identify the instrument used for data collection.	Flow cytometric analysis and sorting was performed with FACSVantage or FACSDiva Cell Sorter (BD Biosciences)
7. Describe the software used to collect and analyze the flow cytometry data.	Data was analyzed using FlowJo10 software (Treestar Inc.)
8. Describe the abundance of the relevant cell populations within post-sort fractions.	100% of the post-sort fraction were the relevant RFP+ cells.
9. Describe the gating strategy used.	Live cells were gated by Dapi vs FSC which were further gated to singlets by FSC vs SSC. Isotype controls and untransduced cells were used for setting gates for negative populations.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information. ☒