主　論　文
(Accepted Final Manuscript)

Recurrent SPI1 (PU.1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia

(高リスクの小児 T 細胞性急性リンパ芽球性白血病において繰り返し認められる SPI1（PU.1）融合遺伝子)

Nature Genetics, 49(8):1274-1281,2017.
DOI : https://doi.org/10.1038/ng.3900

木村　俊介
Shunsuke Kimura
(医歯薬保健学研究科　医歯薬学専攻　小児科学)

Recurrent SPI1(PU.1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia

Masafumi Seki[1,*], Shunsuke Kimura[1,2,*], Tomoya Isobe[1], Kenichi Yoshida[3], Hiroo Ueno[3], Yaeko Nakajima-Takagi[4], Changshan Wang[4], Lin Lin[5], Ayana Kon[3], Hiromichi Suzuki[3], Yusuke Shiozawa[1], Keisuke Kataoka[3], Yoichi Fujii[3], Yuichi Shiraishi[6], Kenichi Chiba[6], Hiroko Tanaka[6], Teppei Shimamura[7], Kyoko Masuda[8], Hiroshi Kawamoto[8], Kentaro Ohki[9], Motohiro Kato[9], Yuki Arakawa[10], Katsuyoshi Koh[10], Ryoji Hanada[10], Hiroshi Moritake[11], Masaharu Akiyama[12], Ryoji Kobayashi[13], Takao Deguchi[14], Yoshiko Hashii[15], Toshihiko Imamura[16], Atsushi Sato[17], Nobutaka Kiyokawa[9], Akira Oka[1], Yasuhide Hayashi[18], Masatoshi Takagi[5], Atsushi Manabe[19], Akira Ohara[20], Keizo Horibe[21], Masashi Sanada[21], Atsushi Iwama[4], Hiroyuki Mano[22], Satoru Miyano[6], Seishi Ogawa[3,**] & Junko Takita[1,**]

[1]Department of Pediatrics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; [2]Department of Pediatrics, Hiroshima University Graduate School of Biomedical Sciences, Hiroshima, Japan; [3]Department of Pathology and Tumor Biology, Graduate School of Medicine, Kyoto University, Kyoto, Japan; [4]Department of Cellular and Molecular Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan; [5]Department of Pediatrics and Developmental Biology, Tokyo Medical and Dental University, Tokyo, Japan; [6]Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan; [7]Division of Systems Biology, Nagoya University Graduate School of Medicine, Nagoya, Japan; [8]Laboratory of Immunology, Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto, Japan; [9]Department of Pediatric Hematology and Oncology Research, National Research Institute for Child Health and Development, Tokyo, Japan; [10]Department of Hematology/Oncology, Saitama Children's Medical Center, Saitama, Japan; [11]Division of Pediatrics, Faculty of Medicine, University of Miyazaki, Miyazaki, Japan; [12]Department of Pediatrics, The Jikei University School of Medicine, [13]Department of Pediatrics, Sapporo Hokuyu Hospital, Sapporo, Hokkaido, Japan; [14]Department of Pediatrics, Mie University Graduate School of Medicine, Tsu, Japan; [15]Department of Pediatrics, Osaka University Graduate School of Medicine, Suita, Japan; [16]Department of Pediatrics, Kyoto Prefectural University of Medicine, Graduate School of Medical Science, Kyoto, Japan; [17]Department of Hematology and Oncology, Miyagi Children's Hospital; [18]Gunma Children's Medical Center, Shibukawa, Gunma, Japan; [19]Department of Pediatrics, St. Luke's International Hospital, Tokyo, Japan, [20]Department of Pediatrics, Toho University, Tokyo, Japan; [21]Clinical Research Center,

National Hospital Organization Nagoya Medical Center, Nagoya, Japan; [22]Department of Cellular Signaling, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan.

* These authors contributed equally to this work.
** Correspondence should be addressed to J.T (jtakita-tky@umin.ac.jp) and S.O (sogawa-tky@umin.ac.jp)

Summary

The outcome of refractory/relapsed patients with pediatric T-cell acute lymphoblastic leukemia (T-ALL) is extremely poor[1], the genetic basis of which remains poorly understood. Here we report a comprehensive profiling of 121 cases with pediatric T-ALL using transcriptome/targeted-capture sequencing, through which we identified novel recurrent gene fusions involving *SPI1* (*STMN1-SPI1*/*TCF7-SPI1*). *SPI1* fusion-positive cases, accounting for 3.9% (7/181) of pediatric T-ALL examined, showed a double-negative (DN; $CD4^{-}CD8^{-}$) or CD8 single-positive (SP) phenotype and had a uniformly poor overall survival. These cases represent a subset of pediatric T-ALL distinguishable from the known T-ALL subsets[2], in terms of expression of genes involved in T-cell precommitment, establishment of T-cell identity, and post ß-selection and of mutational profile. *SPI1* fusions retained transcriptional activities, and when constitutively expressed in mouse stem/progenitor cells, they induced an enhanced cell proliferation and a maturation block. Our findings highlight a unique role of *SPI1* fusions in high-risk pediatric T-ALL.

We first searched recurrent gene fusions based on RNA sequencing (RNA-seq) data from 121 T-ALL samples, using our in-house pipeline (Genomon2.3.0: https://github.com/Genomon-Project/GenomonPipeline) (**Supplementary Table 1**)[3]. In addition to previously reported fusions[4-8] (**Supplementary Table 2**), we newly identified recurrent fusions involving *SPI1* (also known as PU.1 encoding gene) in 5 samples, in which 3' exons of *SPI1* (exons 3–5 or exons 2–5) were fused in-frame to 5' exons of *STMN1* (exons 1–2) (n=2) or *TCF7* (exons 1–3 or exons 1–2) (n=3), generating *STMN1-SPI1 or TCF7-SPI1* fusion gene, respectively (**Fig. 1a** and **Supplementary Table 3**). An additional 2 cases with *TCF7-SPI1* fusion were also found using RT-PCR in an independent validation cohort of 60 T-ALL cases, which were confirmed by RNA-seq. Overall, *SPI1* gene fusions were detected in 7 (3.9%) of 181 cases with pediatric T-ALL examined. Genomic breakpoints were confirmed by targeted-capture sequencing in all *SPI1* fusion-positive cases **(Fig. 1b)**. The fusion-positive cases were either DN or CD8 SP and uniformly positive for CD2, CD3, CD5, and CD7. Of note, HLA-DR expression, which is mostly negative in other T-ALL[9], was positive in all *SPI1* fusion-positive cases (**Supplementary Fig. 1**). No cytogenetic abnormalities involving the *SPI1* locus (11p11.2) had been reported in the fusion-positive cases, except for t(1;11)(p36;p11) in a patient with an *STMN1-SPI1* fusion.

SPI1 is a member of the ETS family of transcription factors. Originally identified as a transforming gene activated via a proviral integration of Friend erythroleukemia virus

(*Spi1*)[10,11], *SPI1* is expressed in various hematopoietic compartments, in which tightly regulated *SPI1* expression is essential for the normal development of hematopoietic stem cells[12,13-17,18-20]. During T-cell development, like other phase 1 genes involved in the precommitment to T-cells (such as *LYL1, LMO2, GATA2, MEIS1,* and *BCL11A*)[21], *SPI1* is expressed in prethymic progenitors and early T-cell precursor (ETP)[22]. *SPI1* expression then is shut off and replaced by the expression of *TCF7* and other phase 2 genes (such as *NOTCH1, RUNX1, GATA3, E2A,* and *BCL11B*) during the transition from $CD44^{+}CD25^{+}$ (DN2) to $CD44^{+}CD25^{-}$ T-cells (DN3), followed by the expression of phase 3 genes, such as *ETS2, LEF1*, and *ID3*[21,23].

The fusion-positive samples invariably showed markedly elevated *SPI1* expression (**Fig. 1c**), most likely reflecting high-level fusion transcripts from the rearranged allele under the control of a heterologous promoter from *TCF7* or *STMN1*, which was demonstrated to be highly expressed in the fusion-positive samples (**Supplementary Fig. 2**). The latter scenario was suggested by the disproportionally higher numbers of RNA-seq reads from 3' *SPI1* exons compared to those from 5' *SPI1* exons (**Supplementary Fig. 2**) and directly confirmed by the significantly increased number of fusion point-containing reads specifically found in the fusion-positive samples (**Supplementary Fig. 3** and **Supplementary Table 4**).

Irrespective of the fusion partners, all predicted *SPI1* fusion proteins retain the domains for DNA binding (ETS domain) but lack varying proportions of the 3' end of *SPI1* [24]. Thus, to assess the transcriptional potential of *SPI1* fusions, we performed luciferase assay using Hela cells carrying a reporter containing a *SPI1*-responsive *CSF1R* (M-CSF receptor) promoter sequence[25]. Compared to the Mock construct, *STMN1*- and *TCF7-SPI1* fusion constructs showed markedly elevated (8- to 10-fold) luciferase activities, which were almost completely suppressed in mutants lacking the DNA binding ETS domain (**Fig. 2a** and **Supplementary Fig. 4**), suggesting that both *SPI1* fusions retain the transcriptional activity inherent to *SPI1*. Next, we evaluated the effect of *SPI1* fusions on T-cell proliferation; DN T-cells from wild-type mice were transduced with fusion or wild-type *SPI1* in an IRES-GFP construct, and 3.0×10^{3} GFP-positive transduced cells were subjected to *in vitro* cell culture to examine the effect of both *SPI1* fusions on cell proliferation. As shown in **Figure 2b**, both wild-type and fusion *SPI1*-transduced cells showed significantly higher proliferation compared to Mock-transduced cells ($P=2.1\times10^{-2}$, 1.8×10^{-3}, and 1.3×10^{-2}, respectively, the Student's *t*-test). We also investigated the effects of the *SPI1* fusion on T-cell development using the TSt4-DLL1 co-culture system[26,27], in which mouse DN1/2 thymocytes ($Lin^{-}CD44^{+}$) were transduced with either

wild-type or each fusion *SPI1* and cultured on TSt4-DLL1 stromal cells. As shown in **Figure 2c-f**, both *STMN1*- and *TCF7-SPI1*-transduced cells and wild-type *SPI1*-transduced cells showed differentiation block during DN T-cell development: there were significantly reduced numbers of DP, SP (CD4+ and CD8+), and DN4 T-cells, which were accompanied by an increase in DN, especially immature DN T-cells (DN1/2), although the effects of *STMN1-SPI1* fusion were milder compared to wild-type *SPI1* and *TCF7-SPI1* fusion. To see the *in vivo* effects of *SPI1* fusions on T-cell differentiation, mouse stem cells ($CD105^{+}Sca\text{-}1^{+}$; LTR-HSCs) were transduced with either a wild-type *SPI1* or a *TCF7-SPI1* fusion construct, transplanted into lethally irradiated mice, and analyzed for their differentiation potential within the thymus. In accordance with *in vitro* analysis, recipients of wild-type *SPI1*- and *TCF7-SPI1* fusion-transduced cells showed significantly increased DN T-cells and reduced CD4-SP and CD8-SP T-cells, compared to the control mice, although the reduction in CD4-SP cells was not statistically significant in the animals transplanted with wild-type *SPI1*-transduced cells. Differential counts of DN T-cells also showed similar trends found in *in vitro* analysis; an increase in immature DN fractions (DN1/DN2) and a reduction in a more mature fraction (DN4). Moreover, the recipients of fusion-transduced cells exhibited a significantly higher proportion of immature CD8 single-positive T-cells ($CD4^{-}CD8^{+}CD24^{+}TCR\beta^{-}$) (**Supplementary Fig. 5**).

To further characterize *SPI1* fusion-positive T-ALL, we performed gene expression profiling of 123 T-ALL patients, including 2 additional fusion-positive cases. Using consecutive two-step unsupervised consensus clustering, we obtained 5 stable clusters (**Fig. 3** and **Supplementary Figs. 6** and **7**). Among these, 4 clusters largely recapitulated distinct T-ALL subtypes characterized in previous studies by an ETP, high *TAL1* expression (*TAL1*-related), and mutually exclusive expression of *TLX1* and *TLX*3 (*TLX*-related) (**Fig. 3** and **Supplementary Figs. 6** and **7**)[28,29]. The *TAL1*-related cluster was further classified into two stable subclusters, which conform to *TAL1*-RA and *TAL1*-RB according to the previous literature[29], in regard to their immunophenotypic characteristics. However, the remaining cluster was newly identified and exclusively consisted of the 7 *SPI1* fusion-positive cases. In terms of their characteristic gene expression pattern, *SPI1* fusion-positive cases showed a phenotype mimicking the normal DN3-4 T-cells[30], which was distinct from that of other clusters (**Supplementary Fig. 8**). Compared to ETP-ALL cases, these *SPI1* fusion-positive patients typically showed reduced expression of the phase 1-specific transcription factor genes implicated in early T-cell development[31]. An exception for this was the aberrant expression of *SPI1* and *SPI1*-regulated genes (*MEF2C* and *HHEX*)[32-34], which should be ascribed to the relevant gene fusions. Of interest in this regard was aberrantly elevated expression of *FLT3* and

KIT in the *SPI1* fusion cluster, which have also been reported to be regulated by *SPI1* (**Supplementary Figs. 9** and **10**)[35,36]. In contrast, ETP-ALL typically lacks expression of the phase 2/3-related genes, which were largely upregulated in *SPI1* fusion-positive cases. Unlike *TLX*-related and *TAL1*-RA/RB cases, which are largely thought to conform to DP T-cells, none of the *SPI1* fusion-positive cases showed overexpression of *HOXA9, TLX1*/*TLX3,* and *TAL1* (**Supplementary Fig. 8**).

Genetic alterations in 121 T-ALL cases including 7 *SPI1* fusion-positive cases were also investigated, using targeted-capture sequencing of 158 commonly mutated genes in pediatric ALL (see Online Methods), which were determined based on the analysis of the whole exome sequencing data from the Therapeutically Applicable Research to Generate Effective Treatment project (**Supplementary Table 5-7**). In line with previous reports[2,37-39], *NOTCH1* (74%) and *CDKN2A* (73%) were by far the most frequently affected genes in T-ALL, although the frequency of *CDKN2A* lesions was characteristically lower in ETP-ALL than other T-ALL subtypes (**Supplementary Fig. 9**) (P=1.8×10^{-9}, the Fisher's exact test). Other frequently mutated genes (>10%) included *PHF6, FBXW7, PTEN, DNM2, USP7, NRAS, JAK3, WT1,* and *RUNX1* (**Fig. 4a**), which had previously been reported in T-ALL[8,31,40-47] except for *USP7*. The distribution of these genetic alterations substantially differed between T-ALL subtypes, conferring a unique mutational profile to each subtype (**Fig. 4b**). As shown in previous reports[31,44], ETP-ALL was characterized by frequent mutations in *GATA3, RUNX1, NRAS, ETV6,* and the epigenetic regulators, and rare *CDKN2A* deletions. Frequent mutations in these epigenetic regulators were also found in *TLX*-related T-ALL, in which an especially high frequency of *PHF6* and *WT1* mutations were among the unique mutational features. The *TAL1*-RA and -RB[29] were distinguished not only by their unique gene expression, but also by gene mutations; *TAL1*-RA is characterized by the paucity of these mutations commonly found in *TAL1*-RB, except for frequent *PTEN* and *USP7* abnormalities. Frequent mutations in *USP7,* which is involved in deubiquitination of PTEN[48], were newly identified in this particular subtype of T-ALL. In contrast, *TAL1*-RB showed frequent mutations in *NOTCH1*/*FBXW7*, transcription factors, *RAS* pathway genes, and epigenetic regulators. Although their small sample size precluded accurate evaluation, *SPI1* fusion-positive T-ALL shared several genetic abnormalities with other T-ALL subtypes[2,37,49], such as frequent *NOTCH1*/*FBXW7* mutations (6/7) and *CDKN2A* deletions (6/7), while other mutations commonly found in other T-ALL subtypes did not seem to be frequent, except for altered *RAS* pathway genes, which were found in 4 (57%) cases with *SPI1* fusions (**Supplementary Fig. 9**).

Finally, we evaluated the effect of *SPI1* fusions on the patients' clinical outcomes

(**Fig. 5)**. Compared to *SPI1* fusion-negative cases, *SPI1* fusion-positive cases showed a significantly shorter overall survival with a median survival of 2.2 years (hazard ratio [HR]=6.05, 95% confidential interval [95%CI]: 2.4−15.1, log-rank P=1.12×10^{-8}); 6 of 7 cases died within 3 years after diagnosis due to early relapse (**Fig. 5a,d)**. The survival was shorter compared to that of patients with other fusions or other clusters, respectively (log-rank P=5.12×10^{-4}, 3.17×10^{-5}) (**Fig. 5b,c**). In multivariate analysis, *SPI1* fusion status was shown to be an independent predictor of survival in pediatric T-ALL (**Table 1**), indicating a possible link between *SPI1* fusions and an aggressive disease phenotype.

SPI1 needs to be downregulated in the DN3 stage for normal thymocyte differentiation[50]. Constitutive *SPI1* expression in fetal thymocyte organ culture leads to a block in the development at the DN3 stage and favored differentiation into macrophage[51]. In *SPI1* fusion-positive cells, functionally active *SPI1* (fusion) expression is thought to be maintained during the transition from DN2/DN3[21,52] and DN/DP[52], where *TCF7* and *STMN1* are expressed, respectively**.** In fact, RNA-seq data from these fusion-positive samples confirmed a markedly elevated expression of full length *TCF7* and *STMN1* transcripts, together with those fused with *SPI1* sequences, while the transcription from the wild-type *SPI1* allele was silenced (**Supplementary Fig. 2**). Also interesting in this point of view is a recent report describing a T-ALL case, in which *SPI1* was translocated to the *BCL11B* locus and overexpressed from an a promoter of *BCL11B*, a gene expressed at the DN2/DN3 stage (**Supplementary Fig. 11**)[33]. Therefore, it is postulated that ectopic *SPI1* expression during the DN2/DN3 transition should cause a differentiation block at the DN3 stage, contributing to T-cell leukemogenesis. In accordance with this, *SPI1* fusion-positive leukemic cells showed a DN or a CD8-SP phenotype and a gene expression profile that conformed to DN3−DN4 T-cells, and therefore is thought to be more immature than *TAL1*- and *TLX*-related subtypes but more differentiated than ETP (**Supplementary Fig. 9**)**.** To support this, upregulated *SPI1* expression in the mice lacking the *SPI1* upstream regulatory element, which normally suppresses *SPI1* expression in T-cells, also causes an impaired transition to the DN3 stage and even resulted in a high frequency of T-cell lymphoma[50,53].

In conclusion, we have described novel recurrent fusions involving *SPI1* in pediatric T-ALL. Exhibiting unique cytological and gene expression profiles, *SPI1* fusion-positive T-ALL had a uniformly dismal clinical outcome. To the best of our knowledge, these are the first genetic lesions associated with a very poor prognosis in pediatric T-ALL, although their impacts on survival need to be confirmed in additional cases. Patients with *SPI1* fusions seem to be incurable with current standard chemotherapy, which underscores the importance of detecting this subset of patients for more intensive or alternative

therapies. In this regard, *RAS* pathway mutations and aberrant expression of *FLT3* and *KIT* associated with *SPI1* fusion-positive T-ALL are of interest, which could be potentially targeted for therapeutics, although further studies would be warranted.

URLs. Genomon2, http://genomon.readthedocs.io/ja/latest/index.html; Integrative Genomics Viewer (IGV), https:// www.broadinstitute.org/igv/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing data have been deposited in the DNA Data Bank of Japan (DDBJ) under accession JGAS00000000090.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGEMENTS

We gratefully acknowledge the TCGA Consortium and all its members for making their invaluable data publically available. The results published here are in whole or part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatment (TARGET) initiative managed by the NCI. The data used for this analysis are available [dbGaP accession number; phs000464]. Information about TARGET can be found at http://ocg.cancer.gov/programs/target. We are also grateful to Ms. Matsumura, Ms. Hoshino, Ms. Yin, Ms. Saito, Ms. Mori, Ms. Mizota, and Ms. Nakamura for their excellent technical assistance. We also wish to express our appreciation to Dr. M-J. Park, Gunma Children's Medical Hospital, Dr. K. Nomura, Toyama university, Dr. H. Kanegane, Tokyo Medical and Dental University, and Dr. K. Kato, Ibaraki Children's Hospital for collecting samples and Dr. T. Yasuda, Ms. Takeyama, Dr. J. Mitsui, and Dr. S. Tsuji The University of Tokyo, for next generation sequencing. This work was supported by KAKENHI (17H04224 [J.T.], 26713037 [M.K.], and 15H05909 [S.O.]) of Japan Society of Promotion of Science; Research on Measures for Intractable Diseases, Health, and Labor Sciences Research Grants, Ministry of Health, Labor and Welfare; by Research on Health Sciences focusing on Drug Innovation (J.T.); by the Japan Health Sciences Foundation (J.T.); by Core Research for Evolutional Science and Technology, Japan Science and Technology Agency (J.T.); Friends of Leukemia Research Fund (M. Seki); and by P-CREATE (J.T.). This research also used computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research

project (hp140230 [S.M.], hp160219 [S.M.], and hp150232 [S.M.]).

AUTHOR CONTRIBUTIONS

Y. Shiraishi, K.C., H.T., T.S, and S.M. developed bioinformatics pipelines. M. Seki, S.K., T.I, K.Y., H.S., Y.F., H.U., K. Kataoka, and Y. Shiozawa performed sequencing data analyses. M. Seki, S.K., T.I., K.Y., M. Sanada, and H. Mano performed sequencing experiments. M. Seki, S.K., L.L., K.M., M.T., C.W., Y.N., A.K., H.K., and A.I. performed functional assays. K.O., T.D., Y. Hashii, and N.K performed FACS analyses. M. Seki, S.K., T.I., J.T., and S.O. interpreted the results. M.K., Y.A., K. Koh, R.H., M.A., H. Moritake, R.K, T. Imamura, A.S., A.M., K.H., and A. Ohara collected specimens. M. Seki, S.K., J.T. and S.O. generated figures and tables and wrote the manuscript. A. Oka, Y. Hayashi, S.O. and J.T. co-led the entire project. All authors participated in discussions and interpretation of the data and results.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

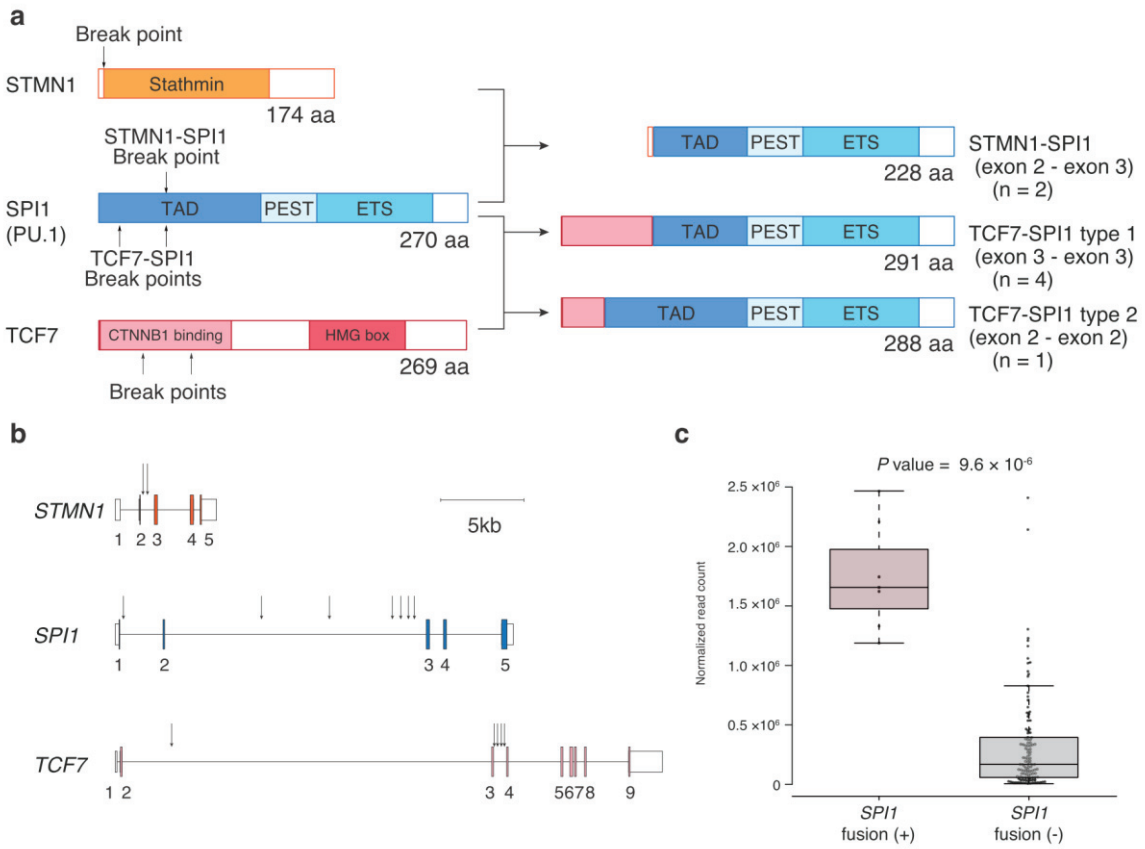

Figure 1. *SPI1* fusions in pediatric T-ALL.

(**a**) Schematic representation of *SPI1* fusions. TAD: transcription activation domain; PEST: proline, glutamic acid, serine, and threonine-rich domain; ETS: E26 transformation-specific domain. (**b**) Breakpoints of *SPI1* rearrangements detected by target capture sequencing are indicated by arrows. (**c**) Comparison of *SPI1* expression levels between *SPI1* fusion-positive (n=7) and negative (n=116) cases, where normalized read counts mapped to *SPI1* exons 1-5 are plotted. The *P*-value was calculated using the Wilcoxon rank-sum test. The mean and the 25th and 75th percentiles are also indicated in box plots and the whiskers extend up to 1.5 times the interquartile range.

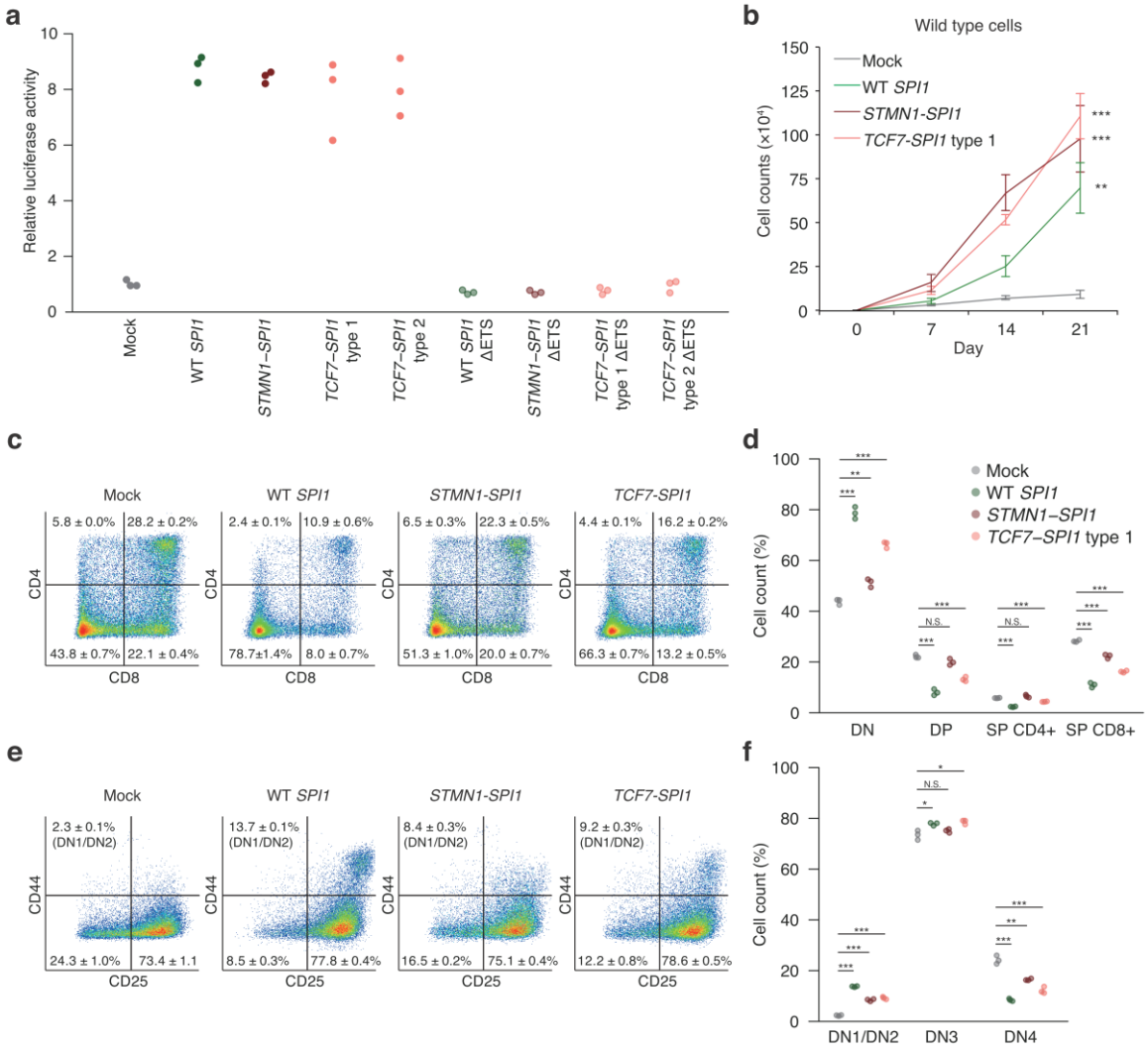

Figure 2. Functional impacts of *SPI1* fusions.

(**a**) *SPI1* transcriptional activities for different *SPI1* constructs (see Supplementary Fig. 4 for each construct, n=3 each) as measured by luciferase reporter assay using the reporter vector contained the *CSF1R* (M-CSF receptor) promoter region with PU.1 binding site. (**b**) *In vitro* growth of DN1 T-cells transduced with *SPI1*, *STMN1-SPI1*, and *TCF7-SPI1* (type 1). Cell counts at indicated time-points are presented (n=3 each) with *P*-values (the Student *t*-test). (**c-f**) The effect of *SPI1* fusions on *in vitro* T-cell differentiation as evaluated by TSt-4/DLL1 co-culture assays. Cell surface expression of CD4/CD8 in total GFP-positive cells (**c**) and that of CD25/CD44 in GFP-positive DN T-cells (**e**) are plotted for each *SPI1* construct as indicated. The fractions (in %) of DN, DP, CD4 and CD8 SP cells among total GFP-positive cells and those of DN1/2, DN3, and DN4 cells (**f**) are compared between different *SPI1* constructs with corresponding *P*-values (the Student's *t* test): * $P<0.05$, ** $P<0.01$, *** $P<0.001$.

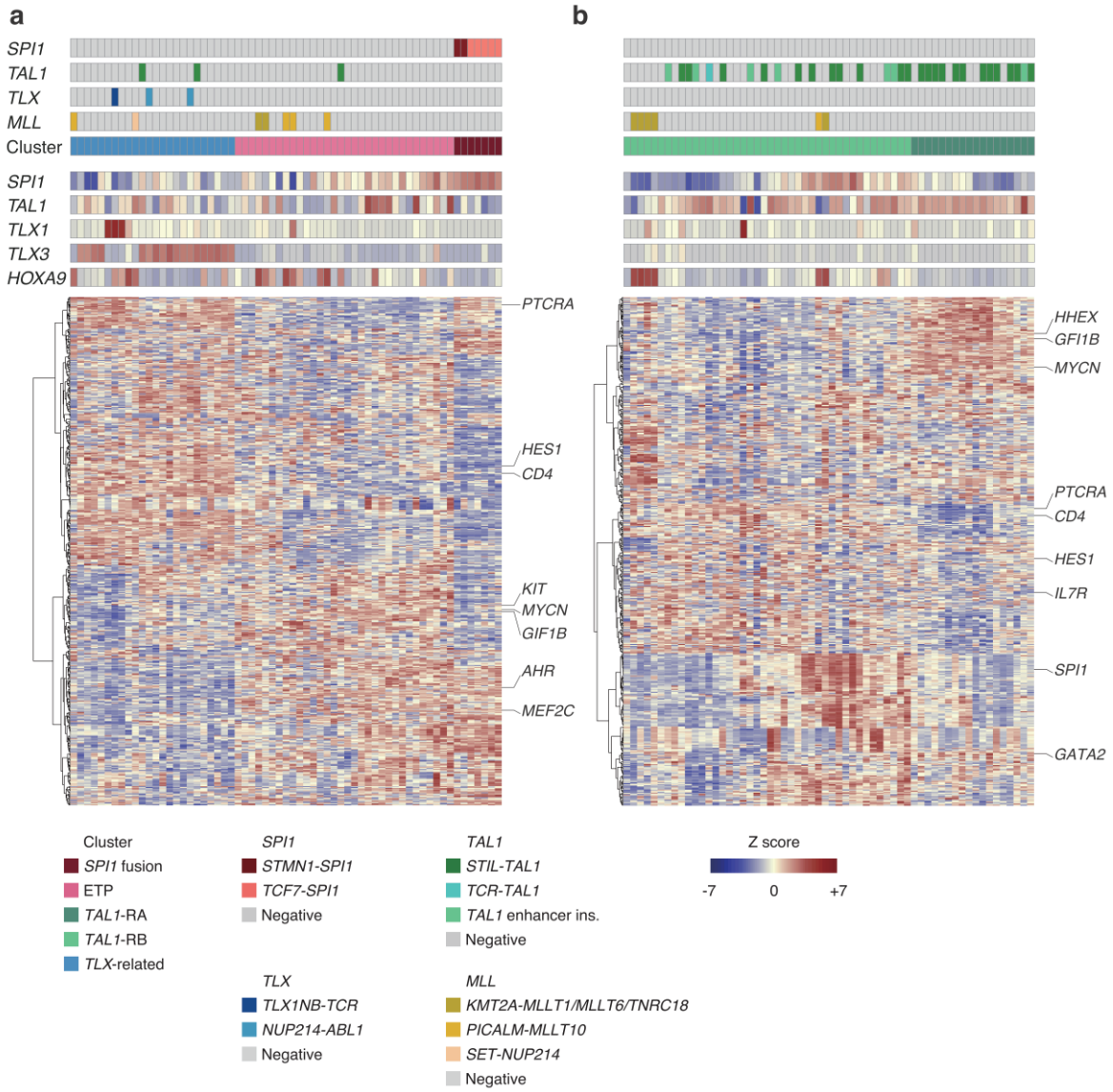

Figure 3. Gene expression clusters in 123 T-ALL based on the consecutive two-step unsupervised consensus clustering.

Two stable clusters (*TAL1* and non-*TAL1*) identified by the initial analysis (in Supplementary Fig. 6) are further divided into 5 subclusters; three (*SPI1* fusion, ETP, and *TLX*-related) from the non-*TAL1* cluster (**a**) and two (*TAL1*-RA and *TAL1*-RB) from the high *TAL1* expression cluster (*TAL1* cluster) (**b**). Gene expression profile of these 5 clusters are shown in the bottom panels. The expression levels of core transcription factors relevant to T-cell development/differentiation are shown in the top panels and also indicated on the right. Genetic alterations involving *SPI1*, *TAL1*, *TLX1* and *TLX3*, as well as *MLL* family genes characteristic of these clusters are also indicated.

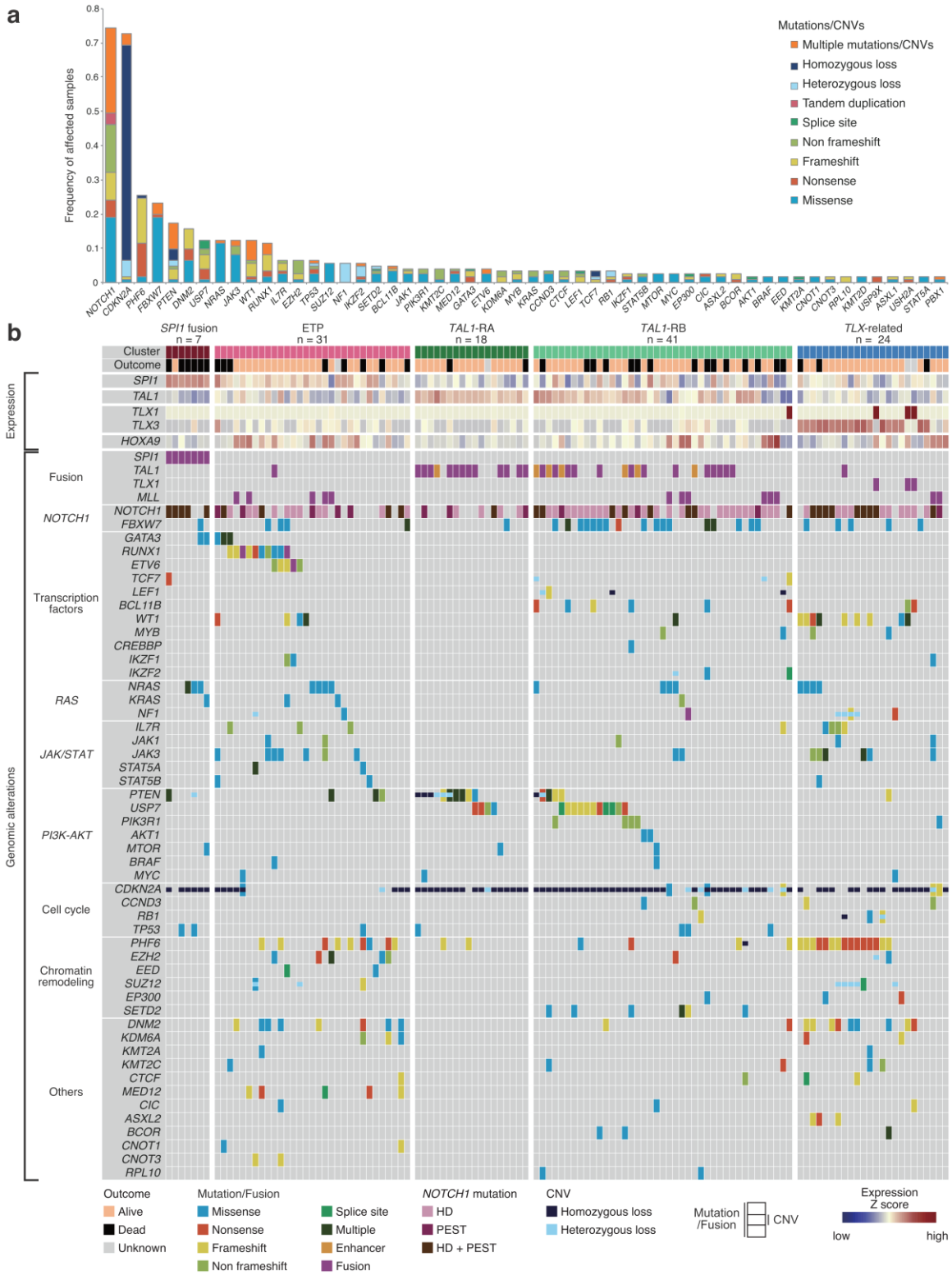

Figure 4. Gene mutations in different T-ALL expression subtypes.

(**a**) Frequencies of recurrent mutations and copy number variations (CNVs) identified by targeted capture sequencing. Mutation are categorized in indicated colors, where "Multiple" means gene abnormalities of two mutations or a mutation with a CNV. **(b)** Gene mutations and CNVs among different T-ALL expression subtypes (n = 121). Cases

are grouped according to the expression clusters as indicated. Expression signatures of core transcription factors (TF), fusion status, significant mutations of known leukemia-related genes, and focal CNVs are also indicated together with clinical outcomes. *MLL* in "Fusion" lanes includes *KMT2A-MLLT1/MLLT6/TNRC18*, *PICALM-MLLT10*, and *SET-NUP214*. The color for each type of genetic alteration is defined below, where "Multiple" means gene abnormalities of two mutations or a mutation with a fusion.

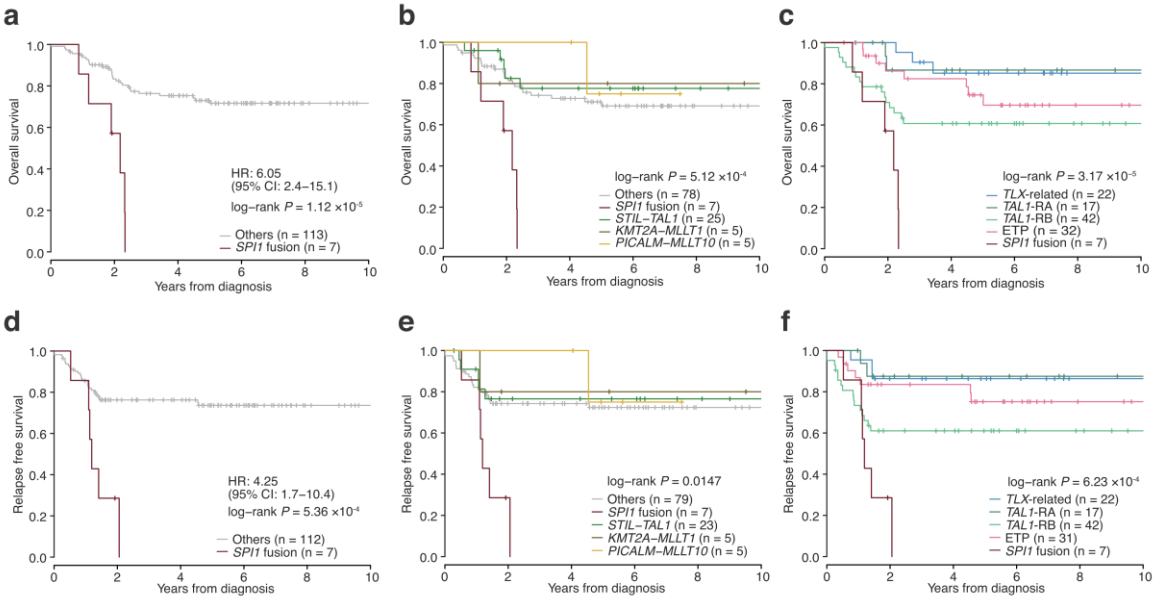

Figure 5. Clinical impact of *SPI1* fusions in T-ALL.

Kaplan-Meier survival curves of overall survival (OS) (**a-c**) and relapse free survival (RFS) (**e-f**) for patients with and without a *SPI1* fusion (**a, d**), for those having indicated gene fusions and those with no known fusions (**b, e**), and for different expression clusters. *P*-values are based on the log-rank test (**c, f**).

Table 1. Univariate and multivariate analysis of overall survival according to the *SPI1* fusion group and selected variables

	Univariate			Multivariate		
	HR	95% CI	*P* value	HR	95% CI	*P* value
***SPI1* fusion**						
SPI1 fusion-positive vs. *SPI1* fusion-negative	6.05	2.42-15.13	1.16×10^{-4}	8.66	2.60-28.78	4.28×10^{-4}
Age (y.o.)						
≥10 vs. <10	0.67	0.32-1.40	0.290	0.97	0.38-2.48	0.95
Sex						
Male vs. female	2.72	1.05-7.01	0.039	3.63	1.07-12.33	0.04
WBC (/μl)						
≥ 10×10^6 vs. < 10×10^6	1.88	0.92-3.86	0.083	1.24	0.50-3.33	0.64
PSL response						
PPR vs. PGR	1.30	0.61-2.76	0.50	1.72	0.74-3.99	0.21
CNS involvement						
CNS 3 vs. CNS 1 or 2	1.64	0.63-4.29	0.31	1.19	0.41-3.46	0.75
Mediastinal mass						
Yes vs. No	1.41	0.67-2.97	0.46	1.30	0.56-2.98	0.54

HR, hazard ratio; WBC, white blood cell count; PSL, prednisolone; PPR, PSL poor responder; PGR, PSL good responder; CNS, central nervous system

References

1. Herold, R., von Stackelberg, A., Hartmann, R., Eisenreich, B. & Henze, G. Acute lymphoblastic leukemia-relapse study of the Berlin-Frankfurt-Munster Group (ALL-REZ BFM) experience: early treatment intensity makes the difference. *J Clin Oncol* **22**, 569-70; author reply 570-1 (2004).
2. Belver, L. & Ferrando, A. The genetics and mechanisms of T cell acute lymphoblastic leukaemia. *Nature Reviews Cancer* **16**, 494-507 (2016).
3. Kataoka, K. *et al*. Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat Genet* **47**, 1304-15 (2015).
4. Aplan, P.D. *et al*. Disruption of the human SCL locus by "illegitimate" V-(D)-J recombinase activity. *Science* **250**, 1426-9 (1990).
5. Rubnitz, J.E. *et al*. Childhood acute lymphoblastic leukemia with the MLL-ENL fusion and t(11;19)(q23;p13.3) translocation. *J Clin Oncol* **17**, 191-6 (1999).
6. Prasad, R. *et al*. Leucine-zipper dimerization motif encoded by the AF17 gene fused to ALL-1 (MLL) in acute leukemia. *Proc Natl Acad Sci U S A* **91**, 8107-11 (1994).
7. (GFCH), G.F.d.C.H. t(10;11)(p13-14;q14-21): a new recurrent translocation in T-cell acute lymphoblastic leukemias. Groupe Francais de Cytogenetique Hematologique (GFCH). *Genes Chromosomes Cancer* **3**, 411-5 (1991).
8. Meyer, C. *et al*. New insights to the MLL recombinome of acute leukemias. *Leukemia* **23**, 1490-9 (2009).
9. Babusikova, O., Stevulova, L. & Fajtova, M. Immunophenotyping parameters as prognostic factors in T-acute leukemia patients. *Neoplasma* **56**, 508-13 (2009).
10. Moreau-Gachelin, F., Tavitian, A. & Tambourin, P. Spi-1 is a putative oncogene in virally induced murine erythroleukaemias. *Nature* **331**, 277-80 (1988).
11. Moreau-Gachelin, F. *et al*. Spi-1 oncogene activation in Rauscher and Friend murine virus-induced acute erythroleukemias. *Leukemia* **4**, 20-3 (1990).
12. Kodandapani, R. *et al*. A new pattern for helix-turn-helix recognition revealed by

the PU.1 ETS-domain-DNA complex. *Nature* **380**, 456-60 (1996).

13. Carotta, S., Wu, L. & Nutt, S.L. Surprising new roles for PU.1 in the adaptive immune response. *Immunol Rev* **238**, 63-75 (2010).
14. Mak, K.S., Funnell, A.P., Pearson, R.C. & Crossley, M. PU.1 and Haematopoietic Cell Fate: Dosage Matters. *Int J Cell Biol* **2011**, 808524 (2011).
15. McKercher, S.R. *et al*. Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *Embo j* **15**, 5647-58 (1996).
16. Scott, E.W., Simon, M.C., Anastasi, J. & Singh, H. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science* **265**, 1573-7 (1994).
17. Back, J., Dierich, A., Bronn, C., Kastner, P. & Chan, S. PU.1 determines the self-renewal capacity of erythroid progenitor cells. *Blood* **103**, 3615-23 (2004).
18. Nutt, S.L. & Kee, B.L. The transcriptional regulation of B cell lineage commitment. *Immunity* **26**, 715-25 (2007).
19. Laslo, P., Pongubala, J.M., Lancki, D.W. & Singh, H. Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Semin Immunol* **20**, 228-35 (2008).
20. Champhekar, A. *et al*. Regulation of early T-lineage gene expression and developmental progression by the progenitor cell transcription factor PU.1. *Genes Dev* **29**, 832-48 (2015).
21. Rothenberg, E.V., Moore, J.E. & Yui, M.A. Launching the T-cell-lineage developmental programme. *Nature Reviews Immunology* **8**, 9-21 (2008).
22. Nutt, S.L., Metcalf, D., D'Amico, A., Polli, M. & Wu, L. Dynamic regulation of PU.1 expression in multipotent hematopoietic progenitors. *J Exp Med* **201**, 221-31 (2005).
23. Del Real, M.M. & Rothenberg, E.V. Architecture of a lymphomyeloid developmental switch controlled by PU.1, Notch and Gata3. *Development* **140**,

1207-19 (2013).

24. Dik, W.A. *et al*. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. *J Exp Med* **201**, 1715-23 (2005).

25. Zhang, D.E., Hetherington, C.J., Chen, H.M. & Tenen, D.G. The macrophage transcription factor PU.1 directs tissue-specific expression of the macrophage colony-stimulating factor receptor. *Mol Cell Biol* **14**, 373-81 (1994).

26. Miyazaki, M. *et al*. Polycomb group gene mel-18 regulates early T progenitor expansion by maintaining the expression of Hes-1, a target of the Notch pathway. *J Immunol* **174**, 2507-16 (2005).

27. Masuda, K. *et al*. Prethymic T-cell development defined by the expression of paired immunoglobulin-like receptors. *EMBO J* **24**, 4052-60 (2005).

28. Ferrando, A.A. *et al*. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**, 75-87 (2002).

29. Soulier, J. *et al*. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood* **106**, 274-86 (2005).

30. Yui, M.A. & Rothenberg, E.V. Developmental gene networks: a triathlon on the course to T cell identity. *Nat Rev Immunol* **14**, 529-45 (2014).

31. Zhang, J. *et al*. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157-63 (2012).

32. Stehling-Sun, S., Dade, J., Nutt, S.L., DeKoter, R.P. & Camargo, F.D. Regulation of lymphoid versus myeloid fate 'choice' by the transcription factor Mef2c. *Nat Immunol* **10**, 289-96 (2009).

33. Homminga, I. *et al*. Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. *Cancer Cell* **19**, 484-97 (2011).

34. Zhang, J.A., Mortazavi, A., Williams, B.A., Wold, B.J. & Rothenberg, E.V. Dynamic

transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* **149**, 467-82 (2012).

35. Neumann, M. *et al*. FLT3 mutations in early T-cell precursor ALL characterize a stem cell like leukemia and imply the clinical use of tyrosine kinase inhibitors. *PLoS One* **8**, e53190 (2013).

36. Zhou, J. *et al*. PU.1 is essential for MLL leukemia partially via crosstalk with the MEIS/HOX pathway. *Leukemia* **28**, 1436-48 (2014).

37. Vlierberghe, P.V. & Ferrando, A. The molecular basis of T cell acute lymphoblastic leukemia. (2012).

38. Weng, A.P. *et al*. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269-71 (2004).

39. Hebert, J., Cayuela, J.M., Berkeley, J. & Sigaux, F. Candidate tumor-suppressor genes MTS1 (p16INK4A) and MTS2 (p15INK4B) display frequent homozygous deletions in primary cells from T- but not from B-cell lineage acute lymphoblastic leukemias. *Blood* **84**, 4038-44 (1994).

40. Van Vlierberghe, P. *et al*. PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat Genet* **42**, 338-42 (2010).

41. O'Neil, J. *et al*. FBW7 mutations in leukemic cells mediate NOTCH pathway activation and resistance to gamma-secretase inhibitors. *J Exp Med* **204**, 1813-24 (2007).

42. Thompson, B.J. *et al*. The SCFFBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. *J Exp Med* **204**, 1825-35 (2007).

43. Palomero, T. *et al*. Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. *Nat Med* **13**, 1203-10 (2007).

44. Van Vlierberghe, P. *et al*. ETV6 mutations in early immature human T cell leukemias. *J Exp Med* **208**, 2571-9 (2011).

45. Bar-Eli, M., Ahuja, H., Foti, A. & Cline, M.J. N-RAS mutations in T-cell acute

lymphocytic leukaemia: analysis by direct sequencing detects a novel mutation. *Br J Haematol* **72**, 36-9 (1989).

46. Tosello, V. *et al*. WT1 mutations in T-ALL. *Blood* **114**, 1038-45 (2009).

47. Grossmann, V. *et al*. Prognostic relevance of RUNX1 mutations in T-cell acute lymphoblastic leukemia. *Haematologica* **96**, 1874-7 (2011).

48. Song, M.S. *et al*. The deubiquitinylation and localization of PTEN are regulated by a HAUSP-PML network. *Nature* **455**, 813-7 (2008).

49. Aifantis, I., Raetz, E. & Buonamici, S. Molecular pathogenesis of T-cell leukaemia and lymphoma. *Nat Rev Immunol* **8**, 380-90 (2008).

50. Anderson, M.K., Weiss, A.H., Hernandez-Hoyos, G., Dionne, C.J. & Rothenberg, E.V. Constitutive expression of PU.1 in fetal hematopoietic progenitors blocks T cell development at the pro-T cell stage. *Immunity* **16**, 285-96 (2002).

51. Laiosa, C.V., Stadtfeld, M., Xie, H., de Andres-Aguayo, L. & Graf, T. Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBP alpha and PU.1 transcription factors. *Immunity* **25**, 731-44 (2006).

52. Tydell, C.C. *et al*. Molecular dissection of prethymic progenitor entry into the T lymphocyte developmental pathway. *J Immunol* **179**, 421-38 (2007).

53. Rosenbauer, F. *et al*. Lymphoid cell growth and transformation are suppressed by a key regulatory element of the gene encoding PU.1. *Nat Genet* **38**, 27-37 (2006).

ONLINE METHODS

Patients and materials. A total of 181 cases with pediatric T-ALL were enrolled in this study (**Supplementary Table 1**). These cases included two large cohorts from Tokyo Children's Cancer Study Group (TCCSG) and Japan Association of Childhood Leukemia Study (JACLS) and cases from various hospitals. Peripheral blood (PB), bone marrow blood (BM), lymph node, or other infiltrated tissue samples were collected from patients after receiving written informed consent according to protocols approved by the Human Genome, Gene Analysis Research Ethics Committee of the University of Tokyo, and other participating institutes. Matched normal samples from BM or PB at complete remission were available in 31 cases.

Next-generation sequencing. Next-generation sequencing was performed using the Illumina HiSeq 2000 or 2500 platform with a standard 100-bp paired-end read protocol, according to the manufacturer's instruction[54].

Targeted-capture sequencing. Among the 123 cases of RNA-seq performed, DNA was available in 121 cases and following targeted capture sequencing was performed. Targeted capture was performed using a SureSelect custom kit (Agilent Technologies), according to the manufacturer's protocol[3,55,56], for which 158 genes and regions were selected for the custom bait library (**Supplementary Table 5**), according to the following criteria on the basis of the whole exome sequencing data obtained from TARGET data of 566 ALL cases (T-ALL 196 cases, B-ALL 367 cases) (dbGaP accession number; phs000464): (i) genes significantly mutated (q value < 0.1) in MutSigCV analysis[57]; (ii) genes or regions reported to be mutated in T-ALL or other T-cell lymphoma; (iii) TCR signaling–associated genes with somatic mutations occurring in one or more patients; (iv) intron or intergenic region of *TCF7*, *STMN1* and *SPI1*; (v) enhancer region of *TAL1*; and (vi) 3′UTR region of *NOTCH1*. Sequence alignment and detection of mutations and structural variations were performed using our in-house pipelines (Genomon v.2.3.0), where reads that had either a Mapping Quality score of <25, a BaseQuality score of <30 or five or more mismatched bases were excluded from the analysis. Relevant somatic mutations were called and filtered by excluding (i) synonymous mutations and variants without complete ORF information; (ii) known variants listed in the 1000 Genomes Project (May 2011 release), NCBI dbSNP build 131, National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) 5400, the Human Genome Variation Database (HGVD; October 2013 release) or our in-house SNP database; (iii) variants present only in unidirectional reads; (iv) variants occurring in repetitive genomic regions; (v) variants with a VAF <0.1;

(vi) variants with less than five supporting reads in tumor samples; (vii) all variants found in non-paired normal samples (n=30) showing allele frequencies >0.0025; and (xiii) all variants in paired samples showing VAF ≥0.1 in the normal sample. We used stringent criteria for mutation calling, requiring a *P* value (by EBCall[58]) of $<10^{-4}$ and the Fisher's *P* value of $<10^{-2}$. Candidate mutations were further filtered by removing all missense SNVs with a VAF of 0.4–0.6 in copy-neutral regions, except for some pathogenic SNVs registered in the Catalogue of Somatic Mutations in Cancer (COSMIC) v70 as hematopoietic and lymphoid tissues or showed "H" in Mutation Assessor. Putative structural variations were manually curated and further filtered by removing those with (i) less than four supporting reads in tumor samples; (ii) a VAF of <0.10 in tumor samples; and (iii) all variants found in normal samples. Finally, mapping errors were removed by visual inspection on the Integrative Genomics Viewer (IGV) browser.

Validation of mutations detected by Targeted-capture sequencing. To confirm the somatic mutations detected in targeted-captured sequencing, we conducted PCR-based deep sequencing (depth≥1,500×) using matched normal samples (n=31). To enable the random reading of targeted sequences, we first amplified the target sequences using primers tagged with NotI cleavage sites, and the products were then ligated and fragmented for deep sequencing as described previously[59,60]. In total, 95 of 104 (91.3%) true-positive somatic variants were detected with a sensitivity of 96.9% (95/98).

Mutation analysis. The significance of the mutations in each gene was evaluated by calculating type I error under the null hypothesis that all the non-silent mutations were passenger changes, assuming a Poisson distribution with a uniform background mutation rate (λ). We employed λ=0.21 mutation/Mb observed in whole exome sequencing data from the TARGET project. After calculating the type I errors for 158 genes, significantly mutated genes (q<0.1) were estimated according to standard methods of Benjamini-Hochberg.

RNA sequencing. In 123 samples for which high-quality RNA (RNA integrity number>5.5) were available, libraries for RNA sequencing were prepared using the NEBNext Ultra RNA Library Prep kit for Illumina (New England BioLabs) or TruSeq RNA Library Preparation Kit v2 (Illumina). Fusion transcripts were detected by Genomon-fusion or Genomon v.2.3.0 and filtered by excluding fusions (i) mapped to repetitive regions; (ii) with less than four spanning reads; (iii) occurring out of frame; or (iv) with junctions not located at known exon–intron boundaries. Normalized count data obtained by the variance-stabilizing

transformation (VST) of the R package DESeq2 were subjected to clustering analysis. Cluster stability was ascertained via consensus clustering with 1,000 iterations using the R package ConsensusClusterPlus. Heatmaps were generated by pheatmap 1.0.7 using VST count data. Differentially expressed genes were extracted using the R package DESeq2.

Fusion transcript validation and screening. All candidate in-frame fusion transcripts were validated by RT-PCR (**Supplementary Table 2**). In an additional 60 cases as discovery cohort, *TCF7-SPI1* and *STMN1-SPI1* fusion transcripts were screened by RT-PCR and Sanger sequencing.

Genomic copy number variations (CNVs). All samples (n=121 samples) examined by targeted-capture sequencing were analyzed for genomic copy numbers using signal data calculated from sequence reads of T-ALL and control samples. These analyses were implemented by in-house pipeline CNACS[61]. The depths were calculated from the weighted sum of the fragments accounting for length and GC-biases during sequencing library amplification. The depths were compared to those of pooled controls. Genomic copy number was estimated from the obtained depth ratios using circular binary segmentation method with the DNA copy package in R. Allelic imbalance was assessed by the allele frequencies of the heterozygous SNPs covered by >50 reads. When these criteria were applied to *CDKN2A*, *CDKN2B*, and *PTEN* for validation in matched 78 samples of available SNP array data (data not shown), 54 of 60 (90.0%) samples with true-positive homo/heterozygous loss were detected with a sensitivity of 87.1% (54/62).

Flow cytometry analysis of leukemic cell samples. FACS data of 91 cases were available in 123 cases with RNA sequencing analysis (**Supplementary Fig. 1**). FACS analyses were performed at central institutes of TCCSG and JACLS or at each institute.

Survival analysis. Survival data were available for 120 patients with pediatric T-ALL having RNA sequencing data. Overall survival (OS) was calculated from the date of diagnosis to death by any cause, and relapse-free survival (RFS) was calculated from diagnosis until first event (failure to achieve remission (calculated as an event at time=0 day), relapse, secondary malignancy, or death by any cause) or last follow-up. The Kaplan–Meier method was used to estimate OS and RFS, and the log-rank test was used to assess the differences in OS between the patient groups (R package survival).

Statistical analysis. Statistical analyses were performed using R 3.2.0 software. For

functional assays, statistical significance was assessed by Student's two-tailed *t*-test, and values were considered statistically significant at $P<0.05$.

Luciferase assays. Hela cells were collected 24 h after transfection and assayed for M-CSF receptor luciferase activity[25] using the Dual-Luciferase Reporter Assay System (Promega). The M-CSF receptor reporter vector was kindly provided by H. Hirai (Department of Transfusion Medicine and Cell Therapy, Graduate School of Medicine, Kyoto University, Kyoto, Japan). Firefly luciferase activity was normalized by *Renilla* luciferase activity (pRL-SV40 vector) in each sample and was presented relative to the activity in Mock-transfected cells. Constructs for luciferase assay are shown in **Supplementary Figure 4**.

DN thymocyte culture. We purified Lineage marker (Gr-1, Mac-1, Ter-119, B220, CD4, CD8α, γδTCR, CD3, CD11c, and NK1.1)$^-$ CD44$^+$ CD25$^-$ DN1 and CD44$^+$ CD25$^+$ DN2 cells from mouse thymus and transduced them with indicated retroviruses on TSt-4/DLL1 stromal cells supplemented with 20 ng/ml of SCF, IL-7, and Flt3L. At day 3 post-transduction, we sorted transduced cells expressing GFP, plated them on TSt-4/DLL1 stromal cells and monitored their growth in the presence of 20 ng/ml of SCF, IL-7, and Flt3L in triplicate and their growth was monitored at day 7, 14 and 21. To evaluate the impact of transduced genes on differentiation of DN1/2 cells, sorted GFP$^+$ cells were cultured on TSt-4/DLL1 stromal cells in the presence of 1 ng/ml of SCF, IL-7, and Flt3L for 7 days, stained with anti-CD45.2, anti-CD4, and anti-CD8, or anti-CD45.2, Lineage marker mixture, anti-CD25, and anti-CD44, and subjected to flow cytometric analyses. Fraction of GFP positive cells is shown in **Supplementary Figure 12**

Isolation of adult long-term repopulating hematopoietic stem cells (LTR-HSCs). According to previously described methods[62,63], hematopoietic progenitors were harvested from 6- to 10-week-old female C57BL/6 mice 4 days after i.p. administration of 150 mg/kg 5-fluorouracil (5-FU) and isolated using lymphoprep (Axis-Shield). The MSCV IRES GFP, pMYs *TCF7-SPI1*, or wild-type *SPI1* vector was transfected using polyethylenimine into Plat-E cells, an ecotropic packaging cell line. Supernatants containing high titers of retrovirus were collected at 48 and 72 h and concentrated using Retro-X Concentrator (TAKARA-Clontech) and used to infect the isolated LTR-HSCs. Infection was performed in Retronectin-coated plate with 4 μg/mL protamine. Hematopoietic progenitors were cultured overnight in αMEM supplemented with 20% FCS and 50 ng/ml of each of mouse SCF (R&D systems #455-MC-010), human IL-6

(PEPROTECH#200-06), human FLT3 ligand (PEPROTECH#250-31L), and human thrombopoietin and infected with retrovirus particles at second day in αMEM supplemented with the same FCS and cytokines, using 24-well dishes coated according to the manufacturer's recommendations with Retronectin. 60 h after infection with retrovirus, LTR-HSCs were transplanted into 6- to 10-week-old C57BL/6 male mice that received myeloablative conditioning with 9.5 Gy of total body irradiation 6 h before BMT. The mice received ad libitum drinking water containing 1 mg/ml neomycin trisulfate salt hydrate (SIGMA) and 100 U/ml polymyxin B sulfate salt (SIGMA). Samples from the thymus were stained using 5–6 color combination of antibodies. Data were acquired on BD LSRFORTESSA (Becton Dickinson) and analyzed using FACS DIVA software. The following antibodies were purchased from Biolegend and eBioscience: PE Anti-Mouse CD19(1D3), Ly-6G/Ly-6C(Gr-1)(RB6-8C5), TER-119/Erythroid Cells (TER-119), NK-1.1(PK136), CD11b(M1/70), and CD45R/B220(RA3-6B2). Fraction of GFP positive cells is shown in **Supplementary Figure 13**. Experiments were performed in a specific pathogen–free unit in the vivarium of Tokyo Medical and Dental University. Experimental and animal care protocols were approved by the Tokyo Medical and Dental University animal care and use committee (protocol numbers 017012A).

Expression array analysis in normal T cell development. For the analysis of selected gene expression profiling in normal T cell development, public expression data submitted under GEO accessions GSE22601 (Affymetrix Human Genome U133A Array) was used. Replicate data in Series Matrix File (RMA normalized processed data) were converted to log 2 mean value.

Data availability. All sequencing data that support the findings of this study have been deposited in the in the DNA Data Bank of Japan (DDBJ) and are accessible through the DDBJ accession number JGAS00000000090. All other relevant data are available from the corresponding authors on request.

References

54. Sato, Y. *et al*. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* **45**, 860-7 (2013).

55. Suzuki, H. *et al*. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat Genet* **47**, 458-68 (2015).

56. Haferlach, T. *et al*. Landscape of genetic lesions in 944 patients with

myelodysplastic syndromes. *Leukemia* **28**, 241-7 (2014).

57. Lawrence, M.S. *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).

58. Shiraishi, Y. *et al*. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic acids research* **41**, e89-e89 (2013).

59. Yoshida, K. *et al*. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-9 (2011).

60. Seki, M. *et al*. Integrated genetic and epigenetic analysis defines novel molecular subgroups in rhabdomyosarcoma. *Nat Commun* **6**, 7557 (2015).

61. Saiki R., *et al*. NGS-based copy number analysis in 1,185 patients with myeloid neoplasms. *Blood* **128**, Abstract 955 (2016).

62. Ono, R. *et al*. Dimerization of MLL fusion proteins and FLT3 activation synergize to induce multiple-lineage leukemogenesis. *J Clin Invest* **115**, 919-29 (2005).

63. Isoda, T. *et al*. Process for immune defect and chromosomal translocation during early thymocyte development lacking ATM. *Blood* **120**, 789-99 (2012).