

1 **Activation of the *LMO2* oncogene through a somatically acquired**
2 **neomorphic promoter in T-Cell Acute Lymphoblastic Leukemia**

3
4 Sunniyat Rahman¹, Michael Magnussen¹, Theresa E. León¹, Nadine Farah¹, Zhaodong Li²,
5 Brian J Abraham³, Krisztina Z. Alapi¹, Rachel J. Mitchell¹, Tom Naughton¹, Adele K.
6 Fielding¹, Arnold Pizzey¹, Sophia Bustraan¹, Christopher Allen¹, Teodora Popa , Karin Pike-
7 Overzet⁵, Laura Garcia-Perez⁵, Rosemary E. Gale¹, David C. Linch¹, Frank J.T. Staal⁵,
8 Richard A. Young^{3,4}, A. Thomas Look^{2,6}, Marc R. Mansour¹

9
10 ^{1.} University College London Cancer Institute, Department of Haematology, 72 Huntley
11 Street, London. WC1E 6DD. United Kingdom.

12 ^{2.} Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical
13 School, Boston, MA 02215

14 ^{3.} Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA
15 02142

16 ^{4.} Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

17 ^{5.} Department of Immunohematology, Leiden University Medical Center, Albinusdreef 2,
18 Building 1, L3-35. 2300 Leiden.

19 ^{6.} Division of Hematology/Oncology, Children's Hospital, Boston, MA 02115

20
21 **Running title:** Somaticly acquired activation of *LMO2* in T-ALL.

22

23

24 **Key points**

- 25 1. Recurrent intronic mutations that create probable MYB, ETS1, and RUNX1 binding
26 sites occur at the *LMO2* promoter in some T-ALL patients
- 27 2. CRISPR/Cas9-mediated disruption of the mutant MYB site in PF-382 cells markedly
28 downregulates *LMO2* expression.

29

30 **Abstract**

31 Somatic mutations within non-coding genomic regions that aberrantly activate oncogenes
32 have remained poorly characterized. Here we describe recurrent activating intronic mutations
33 of *LMO2*, a prominent oncogene in T-cell acute lymphoblastic leukemia (T-ALL).
34 Heterozygous mutations were identified in PF-382 and DU.528 T-ALL cell lines, in addition
35 to 3.7% (6/160) of pediatric and 5.5% (9/163) of adult T-ALL patient samples. The majority
36 of indels harbour putative *de novo* MYB, ETS1 or RUNX1 consensus binding sites. Analysis
37 of 5'-capped RNA transcripts in mutant cell lines identified the usage of an intermediate
38 promoter site, with consequential monoallelic *LMO2* overexpression. CRISPR/Cas9-
39 mediated disruption of the mutant allele in PF-382 cells markedly downregulated *LMO2*
40 expression, establishing clear causality between the mutation and oncogene dysregulation.
41 Furthermore, the spectrum of CRISPR/Cas9-derived mutations provide important insights
42 into the interconnected contributions of functional transcription factor binding. Finally, these
43 mutations occur in the same intron as retroviral integration sites in gene therapy induced T-
44 ALL, suggesting that such events occur at preferential sites in the non-coding genome.

45

46 **Introduction**

47

48 LIM-domain-only protein 2 (*LMO2*) plays a crucial bridging role in the formation of a large
49 multimeric transcriptional complex, that includes TAL1, LDB1, GATA, RUNX1, ETS1 and
50 MYB¹. In mice, *Lmo2* is progressively silenced after the early T-cell progenitor (ETP) stage
51 of thymic development, and leads to T-cell acute lymphoblastic leukemia (T-ALL) when
52 overexpressed in transgenic models²⁻⁴. In human thymi, *LMO2* is similarly downregulated
53 after commitment to the T cell lineage as indicated by DNA microarray analyses⁵.
54 Overexpression of *LMO2* in human hematopoietic stem cells also leads exclusively to pre-
55 leukemic alterations in thymocytes and T cells, but not in other lineages⁶. Reported
56 mechanisms of aberrant *LMO2* expression in human T-ALL include i) recurrent
57 chromosomal translocations, such as t(11;14)(p13;q11) and t(7;11)(q35;p13); ii) cryptic
58 deletions of an upstream negative regulatory region, as in del(11)(p12p13); and iii) retroviral
59 insertional mutagenesis at the *LMO2* locus during gene therapy⁷⁻¹¹. While approximately
60 50% of T-ALL patients overexpress *LMO2*, only about 10% of patients have a detectable
61 cytogenetic lesion¹². Notably, many of these patients will overexpress *LMO2* from a single
62 allele, a feature reminiscent of *TAL1* overexpressing T-ALL cases driven by small somatic
63 indel mutations that create binding sites for MYB, generating a neomorphic enhancer^{13,14}. We
64 thus hypothesized that *cis*-acting mechanisms may account for T-ALL cases with monoallelic
65 *LMO2* expression that lack abnormalities of the *LMO2* locus^{15,16}.

66

67 **Methods**

68

69 Detailed methods are described in the supplementary section. Chromatin
70 immunoprecipitation (ChIP)-sequencing was performed on T-ALL cell lines following
71 immunoprecipitation with antibodies against MYB and acetylated H3K27 (H3K27ac).
72 Analysis of Motif Enrichment (AME) was used to confirm enrichment of MYB motifs in the
73 MYB ChIP-seq data (Table S1 and S2). *LMO2* mRNA levels were quantified by qRT-PCR.
74 Mutation screening of primary T-ALL samples was achieved by denaturing HPLC of *LMO2*
75 intron 1 PCR products. Luciferase reporter constructs, consisting of 469 bp PCR products
76 inserted upstream of a SV40 promoter and firefly luciferase gene, were electroporated into
77 Jurkat cells. CRISPR/Cas9 genome editing was used to target the *LMO2* intron 1 mutations in
78 the PF-382 T-ALL cell line.

79

80 **Results and Discussion**

81

82 To test this hypothesis, we first assessed *LMO2* expression by quantitative RT-PCR (qRT-
83 PCR) in several T-ALL cell lines arrested at different stages of thymic differentiation. The
84 ETP-like T-ALL cell line Loucy expressed *LMO2* at levels significantly higher than the more
85 mature T-ALL cell lines (DND-41, ALL-SIL, Jurkat), reflecting physiological expression of
86 *LMO2* at the ETP stage of thymic development (Figure 1A). The *TALI*-positive cell lines
87 DU.528 and PF-382 both exhibited upregulated *LMO2* expression, yet crucially have no
88 reported chromosomal lesions affecting this locus (Figure 1A)^{17,18}. In contrast to Loucy cells,
89 aberrant H3K27ac marks, indicative of active chromatin, were identified prior to and
90 encompassing the non-coding exon 2 of the *LMO2* gene by ChIP-seq in PF-382 and DU.528
91 T-ALL cell lines (Figure 1B and S1). Sequencing across these peaks revealed a heterozygous
92 20bp duplication in PF-382 cells and a heterozygous 1bp deletion in DU.528 cells, located
93 close to a region recently described as an intermediate promoter for reasons that were not
94 then apparent (Figure 1B)¹⁹. Notably, the mutations were not described as normal germline
95 variants in dbSNP. *In silico* analysis of the reference sequence identified a high confidence
96 primary MYB binding motif (AACCGTT) that was duplicated in the PF-382 cell line, while
97 the single bp deletion in DU.528 cells creates a CAACCGC sequence that closely resembles
98 a secondary MYB binding motif (Figure 1B; Table S3 and S4).

99

100 To assess whether the mutations form aberrant sites of MYB binding, we performed ChIP-
101 seq for MYB and analyzed peaks of MYB enrichment at the *LMO2* locus. There was a
102 complete absence of MYB binding at the intermediate promoter in cells that were wild-type
103 at this locus, suggesting that the presence of the single native MYB motif in itself is
104 insufficient to recruit MYB. In contrast, both PF-382 and DU.528 cells that harbor dual MYB

105 motifs displayed precisely aligned MYB binding at the mutation site (Figure 1B). To
106 determine whether the mutations affected promoter usage, we performed 5'RACE in *LMO2*
107 mutant and wild-type cell lines using a common primer in exon 6 capable of capturing the
108 transcription start site (TSS) of all *LMO2* isoforms. While the majority (73%) of 5' capped
109 transcripts in Loucy cells originated from the proximal promoter, both PF-382 and DU.528
110 cells demonstrated preferential usage of the recently-described intermediate promoter (75%
111 and 67% of transcripts respectively; Figure 1C).

112

113 Our observations were not limited to T-ALL cell lines as heterozygous mutations at *LMO2*
114 intron 1 were detected in diagnostic samples from 3.7% (6/160) of pediatric and 5.5% (9/163)
115 of adult T-ALL patients (Figure 1D). Absence of the mutations in 7 available patient-
116 matched remission samples confirmed that they were somatic (Figure S2). Notably, the
117 mutations were densely distributed around highly conserved native ETS1, MYB and GATA
118 motifs (Figure S3). Including the cell lines, seven mutations introduced an additional MYB
119 site, resulting in two MYB motifs spaced 10 or 20 bp apart, equivalent to one or two helical
120 coils of DNA respectively (Figure 1E). Three mutations created potential binding sites for
121 both MYB and ETS1, three formed potential ETS1 sites, and three produced potential new
122 RUNX1 binding sites (Figure 1E; Table S3 and S4). Given *NOTCH* and *TALI* have been
123 shown to collaborate with *LMO2* to promote leukemogenesis in murine models of T-ALL, it
124 is noteworthy that of the 15 patients with *LMO2* promoter mutations, 7 had *NOTCH-1*
125 mutations and 8 had *TALI* activating lesions, including two with *TALI*-enhancer mutations
126 (both creating new MYB motifs; Table S5)^{21,22}. Such collaboration between *TALI*, *LMO2*
127 and *NOTCH-1* has also been described in gene therapy-induced T-ALL, including one patient
128 that harbored both a retroviral integration upstream of *LMO2* and an episomal reintegration at
129 the *TALI* locus^{9,13,23}.

130

131 To ascertain whether *LMO2* promoter mutations in T-ALL led to aberrant expression
132 compared to its matched thymic counterpart, we assessed *LMO2* expression by qRT-PCR in
133 thymic subsets sorted for different levels of thymic differentiation⁵. Validating earlier reports
134 using microarrays, *LMO2* expression was highest in the most immature, pre-commitment
135 stages of T cell development, and expressed at low levels from the double-negative (DN)
136 stage onwards, when thymocytes have undergone biallelic TCR- γ rearrangement (Figure
137 2A)⁵. To determine the level of differentiation arrest of the 15 mutant patient samples, we
138 analyzed the TCR- γ locus by q-PCR (Figure S4); twelve of the 15 samples (including 5 of the
139 6 patients with available RNA) had biallelic TCR- γ deletion (Figure S4; Table S5), indicating
140 maturation arrest occurred after the pro-T-cell stage of differentiation, and that the majority
141 of patients were not of the ETP-ALL phenotype. Thus, compared to their physiological
142 counterpart, those patients with RNA available for *LMO2* qRT-PCR, exhibited aberrant
143 *LMO2* overexpression (Figure 2A; $P < 0.002$ vs DN and DP subsets). Although we were
144 unable to confirm *LMO2* overexpression in all mutant samples due to availability of RNA, all
145 classes of mutation (additional MYB, ETS1, RUNX1 or MYB+ETS1 sites) were represented
146 in the 6 patients with *LMO2* overexpression. Exploiting a heterozygous germline SNP
147 (rs3740617), DU.528 cells and 3 of 4 informative patient samples displayed skewed allelic
148 expression of *LMO2* (Figure 2B). The observation of biallelic expression in sample A1
149 suggests a potential lesion on the second allele that remains undefined. Consistent with their
150 *cis*-activating potential, $\geq 96\%$ of reads from MYB ChIP-seq performed in DU.528 and PF-
151 382 cells aligned to the mutant rather than wild-type allele (Figure 2C and S5). Furthermore,
152 the gain-of-function nature of the mutations was confirmed by luciferase reporter assays
153 conducted in Jurkat cells where all mutations markedly activated luciferase activity compared
154 to the wild-type sequence (Figure 2D and S6A).

155

156 To assess causality between the mutations and *LMO2* dysregulation, we used CRISPR/Cas9
157 genome-editing with a guide RNA designed to target the duplicated MYB site in PF-382 cells
158 (Figure S6B). Crucially, clone 4F11 that had a single T>C substitution disrupting the MYB
159 binding site, and clone 1A8 where the mutant allele had been reverted to wild-type, resulted
160 in the most dramatic downregulation of *LMO2* (Figure 2E, 2F and S7). Interestingly, two
161 clones (4H12 and 6D4) that increased the distance between the native and the mutant MYB
162 sites resulted in a marked reduction in *LMO2* expression, supporting the hypothesis that
163 MYB binding is augmented when additional motifs are orientated on the same side of the
164 DNA helix²⁴. This was further validated by the lack of reduction in *LMO2* expression in a
165 clone (5F10) where the sequence between the two MYB sites was altered but the spacing
166 distance was unchanged.

167

168 In conclusion, we identified and functionally validated a novel recurrent mutation hotspot
169 occurring in a non-coding site that drives *LMO2* overexpression from a neomorphic promoter
170 in a substantial proportion of both adult and pediatric T-ALL patients. Remarkably, the
171 mutations create potential binding sites for MYB, ETS1 or RUNX1, all of which are
172 members of a highly oncogenic TAL1-LMO2 complex in T-ALL, indicating that LMO2 is a
173 component of an autoregulatory self-sustaining positive feedback loop in these cells,
174 analogous to autoregulation of *TAL1* we recently described in Jurkat cells^{14,25}. To prove the
175 newly formed ETS1 and RUNX1 sites are sufficient to drive *LMO2* expression, we attempted
176 but ultimately were unable to knockin these mutations *in vitro*. Thus, the oncogenic potential
177 of these particular mutations are an area of ongoing study. It has remained obscure as to
178 exactly how various members of the TAL1 complex orient themselves on DNA with regards
179 spacing, orientation and order of motifs, so called syntax²⁶. Thus, identification of gain-of-
180 function non-coding mutations that have been selected for during tumorigenesis *in vivo*,

181 offers important insights into the optimal DNA syntax required for nucleation of such multi-
182 protein transcription factor complexes. For instance, it may become apparent why a single
183 MYB binding site is sufficient to drive expression from certain loci, such as at the *TALI*
184 enhancer, while others require dual MYB motifs. Lastly, we note that these mutations occur
185 within the same intron as retroviral integration sites described in two cases of gene therapy-
186 induced T-ALL (Figure S8)^{23,27}. This raises the possibility that formation of aberrant
187 promoters and enhancers, either by mutation or retroviral insertion, occur at preferred, rather
188 than random sites in the non-coding genome.
189

190 **Acknowledgements**

191

192 M.R.M and S.R. are funded by Bloodwise, and receive support from the Gabrielle's Angels
193 Foundation. M.M. is funded by the Freemason's Grand Charity. Z.L. was supported by
194 Alex's Lemonade Stand Foundation for Childhood Cancer. B.J.A. is the Hope Funds for
195 Cancer Research Grillo-Marxuach Family Fellow. The UKALL2003 trial was supported by
196 grants from Bloodwise (formerly known as Leukaemia and Lymphoma Research, UK) and
197 the Medical Research Council (UK), with trial number ISRCTN07355119. Primary
198 childhood leukemia samples used in this study were provided by the Bloodwise Childhood
199 Leukemia Cell Bank, working with the laboratory teams in the Bristol Genetics Laboratory,
200 Southmead Hospital, Bristol: Molecular Biology Laboratory, Royal Hospital for Sick
201 Children, Glasgow: Molecular Haematology Laboratory, Royal London Hospital, London:
202 Molecular Genetics Service and Sheffield Children's Hospital, Sheffield. The UKALL14 trial
203 is supported by Cancer Research UK (UK) with the trial number ISRCTN66541317. This
204 work was funded by NIH grants 1R01CA176746-01, 5P01CA109901-08, and 5P01CA68484
205 (A.T.L.). R.A.Y. is a founder and member of the Board of Directors of Syros
206 Pharmaceuticals, a company developing therapies that target gene regulatory elements. We
207 thank the patients, families and clinical teams who have been involved in both trials. This
208 work was undertaken at UCL, which receives a proportion of funding from the Department of
209 Health's NIHR Biomedical Research Centre's funding scheme.

210

211 **Authorship Contributions**

212

213 S.R, M.M, T.E.L, N.F, A.P, Z.L, S.B, C.A, T.P, K.P.O, L.G.P and B.J.A performed

214 experimental work.

215 K.Z.A, R.J.M, T.N, A.K.F, R.E.G, K.P.O, L.G.P, F.J.T.S provided primary samples.

216 S.R, M.M, T.E.L, N.F, D.C.L, R.A.Y, F.J.T.S and A.T.L analyzed data.

217 S.R, R.E.G, F.J.T.S, D.C.L, M.R.M wrote the manuscript.

218 M.R.M designed the study.

219 All authors approved the final manuscript.

220

221 **Disclosure of Conflicts of Interest**

222 The authors declare no competing financial interests.

223

224 REFERENCES

- 225 1. Chambers J, Rabbitts TH. LMO2 at 25 years: a paradigm of chromosomal
 226 translocation proteins. *Open Biol.* 2015;5(6):150062.
- 227 2. Boehm T, Feroni L, Kaneko Y, Perutz MF, Rabbitts TH. The rhombotin family of
 228 cysteine-rich LIM-domain oncogenes: distinct members are involved in T-cell translocations
 229 to human chromosomes 11p15 and 11p13. *Proceedings of the National Academy of Sciences*
 230 *of the United States of America.* 1991;88(10):4367-4371.
- 231 3. Fisch P, Boehm T, Lavenir I, et al. T-cell acute lymphoblastic lymphoma induced in
 232 transgenic mice by the RBTN1 and RBTN2 LIM-domain genes. *Oncogene.* 1992;7(12):2389-
 233 2397.
- 234 4. Herblot S, Steff AM, Hugo P, Aplan PD, Hoang T. SCL and LMO1 alter thymocyte
 235 differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression. *Nat*
 236 *Immunol.* 2000;1(2):138-144.
- 237 5. Dik WA, Pike-Overzet K, Weerkamp F, et al. New insights on human T cell
 238 development by quantitative T cell receptor gene rearrangement studies and gene
 239 expression profiling. *J Exp Med.* 2005;201(11):1715-1723.
- 240 6. Wiekmeijer AS, Pike-Overzet K, Brugman MH, et al. Overexpression of LMO2 causes
 241 aberrant human T-Cell development in vivo by three potentially distinct cellular
 242 mechanisms. *Exp Hematol.* 2016;44(9):838-849 e839.
- 243 7. Hacein-Bey-Abina S, von Kalle C, Schmidt M, et al. A serious adverse event after
 244 successful gene therapy for X-linked severe combined immunodeficiency. *The New England*
 245 *journal of medicine.* 2003;348(3):255-256.
- 246 8. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, et al. LMO2-associated clonal T cell
 247 proliferation in two patients after gene therapy for SCID-X1. *Science (New York, NY).*
 248 2003;302(5644):415-419.
- 249 9. Howe SJ, Mansour MR, Schwarzwaelder K, et al. Insertional mutagenesis combined
 250 with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1
 251 patients. *J Clin Invest.* 2008;118(9):3143-3150.
- 252 10. Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic
 253 leukemia. *The Journal of clinical investigation.* 2012;122(10):3398-3406.
- 254 11. Van Vlierberghe P, van Grotel M, Beverloo HB, et al. The cryptic chromosomal
 255 deletion del(11)(p12p13) as a new activation mechanism of LMO2 in pediatric T-cell acute
 256 lymphoblastic leukemia. *Blood.* 2006;108(10):3520-3529.
- 257 12. Ferrando AA, Look AT. Gene expression profiling in T-cell acute lymphoblastic
 258 leukemia. *Semin Hematol.* 2003;40(4):274-280.
- 259 13. Navarro JM, Touzart A, Pradel LC, et al. Site- and allele-specific polycomb
 260 dysregulation in T-cell leukaemia. *Nat Commun.* 2015;6:6094.
- 261 14. Mansour MR, Abraham BJ, Anders L, et al. Oncogene regulation. An oncogenic
 262 super-enhancer formed through somatic mutation of a noncoding intergenic element.
 263 *Science.* 2014;346(6215):1373-1377.
- 264 15. Ferrando AA, Herblot S, Palomero T, et al. Biallelic transcriptional activation of
 265 oncogenic transcription factors in T-cell acute lymphoblastic leukemia. *Blood.*
 266 2004;103(5):1909-1911.
- 267 16. Van Vlierberghe P, Beverloo HB, Buijs-Gladdines J, et al. Monoallelic or biallelic
 268 LMO2 expression in relation to the LMO2 rearrangement status in pediatric T-cell acute
 269 lymphoblastic leukemia. *Leukemia.* 2008;22(7):1434-1437.

- 270 17. Chen S, Nagel S, Schneider B, et al. Novel non-TCR chromosome translocations
271 t(3;11)(q25;p13) and t(X;11)(q25;p13) activating LMO2 by juxtaposition with MBNL1 and
272 STAG2. *Leukemia*. 2011;25(10):1632-1635.
- 273 18. Dong WF, Xu Y, Hu QL, et al. Molecular characterization of a chromosome
274 translocation breakpoint t(11;14)(p13;q11) from the cell line KOPT-K1. *Leukemia*.
275 1995;9(11):1812-1817.
- 276 19. Oram SH, Thoms JAI, Pridans C, et al. A previously unrecognized promoter of LMO2
277 forms part of a transcriptional regulatory circuit mediating LMO2 expression in a subset of
278 T-acute lymphoblastic leukaemia patients. *Oncogene*. 2010;29(43):5796-5808.
- 279 20. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity
280 between motifs. *Genome Biol*. 2007;8(2):R24.
- 281 21. Larson RC, Lavenir I, Larson TA, et al. Protein dimerization between Lmo2 (Rbtn2)
282 and Tal1 alters thymocyte development and potentiates T cell tumorigenesis in transgenic
283 mice. *EMBO J*. 1996;15(5):1021-1027.
- 284 22. O'Neil J, Calvo J, McKenna K, et al. Activating Notch1 mutations in mouse models of
285 T-ALL. *Blood*. 2006;107(2):781-785.
- 286 23. Hacein-Bey-Abina S, Garrigue A, Wang GP, et al. Insertional oncogenesis in 4 patients
287 after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest*. 2008;118(9):3132-3142.
- 288 24. Molvaersmyr AK, Saether T, Gilfillan S, et al. A SUMO-regulated activation function
289 controls synergy of c-Myb through a repressor-activator switch leading to differential p300
290 recruitment. *Nucleic Acids Res*. 2010;38(15):4970-4984.
- 291 25. Sanda T, Lawton LN, Barrasa MI, et al. Core transcriptional regulatory circuit
292 controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell*.
293 2012;22(2):209-221.
- 294 26. Farley EK, Olson KM, Zhang W, Rokhsar DS, Levine MS. Syntax compensates for poor
295 binding sites to encode tissue specificity of developmental enhancers. *Proceedings of the*
296 *National Academy of Sciences of the United States of America*. 2016;113(23):6508-6513.
- 297 27. Braun CJ, Boztug K, Paruzynski A, et al. Gene therapy for Wiskott-Aldrich syndrome--
298 long-term efficacy and genotoxicity. *Sci Transl Med*. 2014;6(227):227ra233.
- 299 28. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools
300 and content for the online database of protein-binding microarray data on protein-DNA
301 interactions. *Nucleic Acids Res*. 2015;43(Database issue):D117-122.

302
303

304

305 **Figure Legends**

306

307 **Figure 1: *LMO2* intron 1 mutations in pediatric and adult human T-cell acute**
308 **lymphoblastic leukemia (T-ALL). (a) *LMO2* expression as determined by qRT-PCR in**
309 ***LMO2* translocated T-ALL cell lines – KOPT-K1 and P12-Ichikawa, and non-translocated T-**
310 **ALL cell lines, DU.528, PF-382, Loucy, DND41, Jurkat and ALL-SIL. (b) ChIP-Seq tracks**
311 **at the *LMO2* locus for MYB and H3K27ac in PF-382, DU.528, Loucy and Jurkat T-ALL cell**
312 **lines. Y-axis values are reads per bin per million mapped reads (RPM). Below, mutations are**
313 **shown as identified by Sanger sequencing of PF-382 and DU.528 DNA, with inserted**
314 **sequences shown in red, and MYB motifs underlined. The position weight matrices (PWM)**
315 **for the primary and secondary MYB binding sites are from UniPROBE²⁸. (c) Pie chart**
316 **summarising the percentage of *LMO2* transcripts identified by 5'RACE that start from the**
317 **distal, intermediate and proximal promoters, for the PF-382, DU.528 and Loucy T-ALL cell**
318 **lines. A total of 20, 21 and 22 *LMO2* transcripts was examined respectively for PF-382,**
319 **DU.528 and Loucy T-ALL cell lines. (d) Pie chart summarising mutation recurrence within**
320 **pediatric and adult human T-ALL cohorts. (e) Indels mapped to the *LMO2* intron 1 mutation**
321 **hotspot, labelled with the associated *de novo* consensus site as aligned to the UniPROBE or**
322 **HOCOMOCO PWMs, where MYB, ETS1 and RUNX1 sites are marked as a triangle, square**
323 **and diamond respectively. Below, motif analysis of the region shows the native binding sites**
324 **for members of the TAL1 complex including, RUNX1, E-box (for TAL1 binding), ETS1,**
325 **MYB and GATA.**

326

327 **Figure 2: *LMO2* intron 1 indels are predominantly monoallelically activating and**
328 **CRISPR/Cas9 mediated knockout of the PF-382 mutant allele downregulates *LMO2***
329 **expression (a) *LMO2* expression as determined by qRT-PCR in human sorted thymic**

330 subsets, primary patient samples with *LMO2* intron 1 indels, and the wild-type Jurkat cell
331 line. $P < 0.002$ for samples A1, A2, A3, A9, and P6, vs DN and DP by two-tailed t test.
332 Primary patient samples were assessed for the absence of bi-allelic TCR- γ deletion (ABD), of
333 which patient sample A4 (orange bar) exhibited ABD, whilst all other patients were non-
334 ABD. **(b)** The informative SNP, rs3740617 was amplified in 4 patient samples and the
335 DU.528 cell line from both gDNA and cDNA templates to infer monoallelic expression. To
336 do this, if one chromatogram peak is detected at a heterozygous SNP within the cDNA, the
337 expression can be interpreted as coming from one allele **(c)** Quantification of the number of
338 reads mapped to the wild type (WT) or mutant (MUT) allele where 54 of 56 reads, and 85 of
339 85 reads mapped to the mutant alleles for DU.528 and PF-382 respectively **(d)** Firefly
340 luciferase activity following renilla and no-insert vector normalisation for patient-derived
341 indels. Data shown is from ≥ 3 independent experiments performed in triplicate. Values
342 shown are mean \pm SD and p-values (where $p \leq 0.05$ is denoted by *) were calculated by a
343 two-tailed Student's t-test. **(e)** The yellow highlighted sequence is the target region for the
344 CRISPR/Cas9 guide RNA. Aligned sequences are from CRISPR/Cas9-edited PF-382 single
345 cell clones showing the associated genomic edits generated. Red sequences are inserted
346 sequences, blue are altered, and dashes represent deleted bases. Underlined region shows the
347 presence of the native and mutant MYB binding sites. **(f)** Gene expression of *LMO2* for each
348 PF-382 clone, as determined by qRT-PCR. Data is expressed as fold change relative to the
349 mean expression of the unedited clones in arbitrary units (AU). Clones are labelled as
350 "unedited", where CRISPR/Cas9 did not edit region targeted by the guide RNA, and "edited"
351 where successfully targeting led to the formation of an indel.
352

Figure 1. Rahman et al.

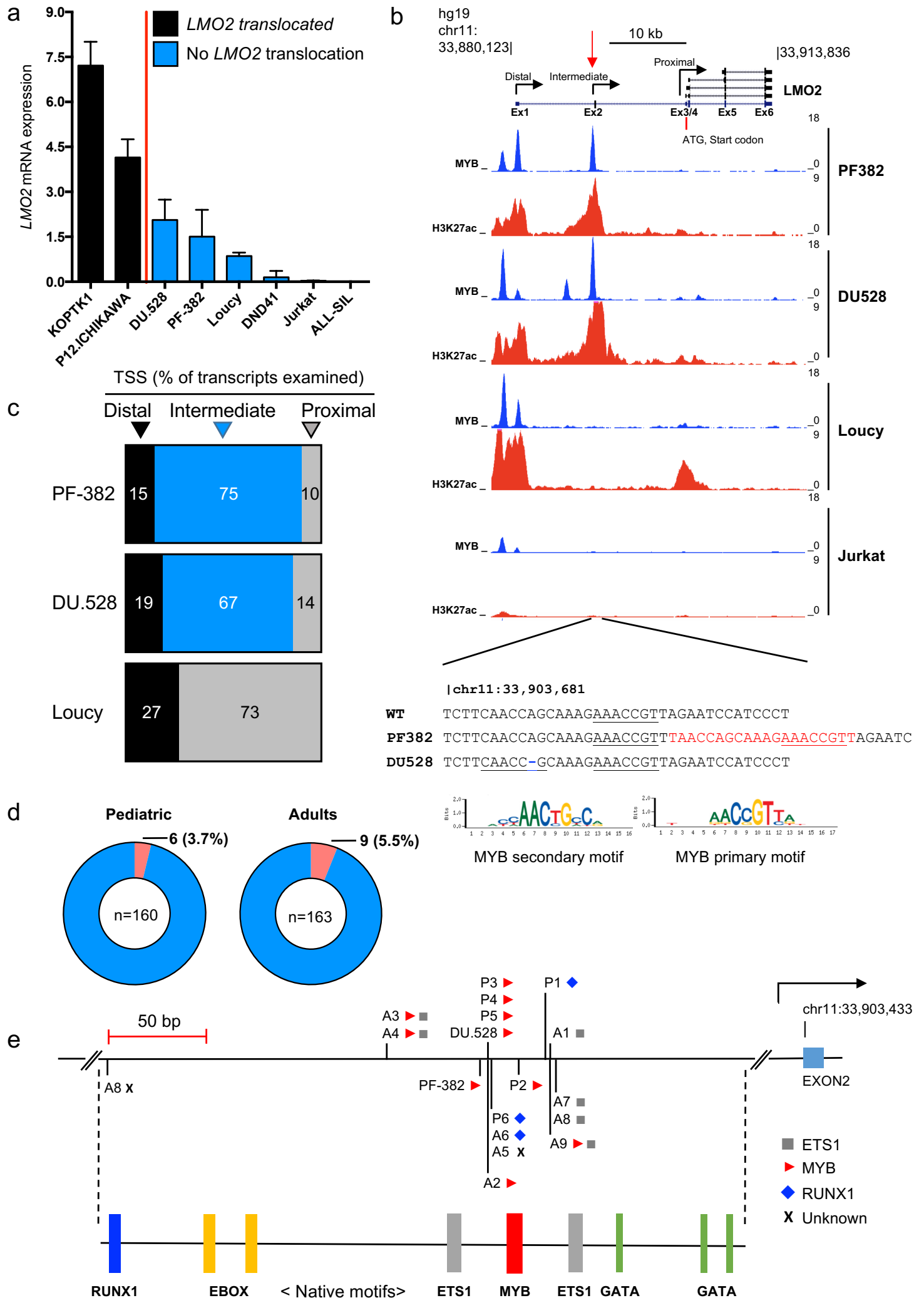
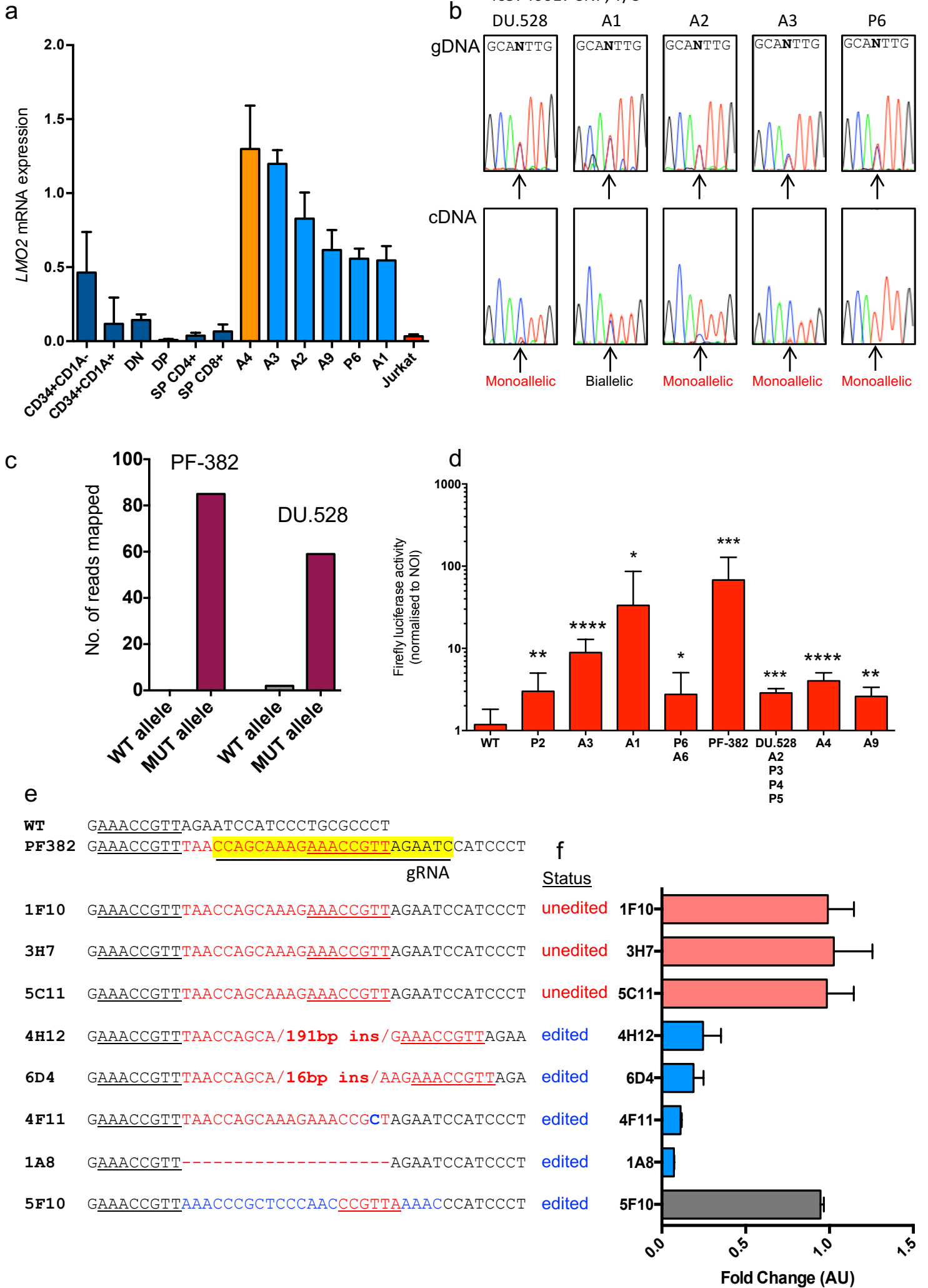


Figure 2. Rahman et al.



Activation of the LMO2 oncogene through a somatically acquired neomorphic promoter in T-Cell Acute Lymphoblastic Leukemia.

Supplemental Material and Methods, Figures and Tables for Rahman et al.

Supplemental Material and Methods

ChIP-Seq of T-ALL cell lines.

ChIP was performed as described by Lee et al. previously with a few adjustments¹. Suspension cultures were grown to a density of ~1-10 million cells/ml prior to crosslinking, and adherent cell lines were crosslinked directly on the culture vessel. Crosslinking was performed for 10-15 min at room temperature by the addition of one-tenth of the volume of 11% formaldehyde solution (11% formaldehyde, 50 mM HEPES pH 7.3, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0) to the growth media followed by 5 min quenching with 125 mM glycine or 1M Tris pH7.5. Cells were washed twice with PBS, then the supernatant was aspirated and the cell pellet was flash frozen in liquid nitrogen. Frozen crosslinked cells were stored at -80°C. 100µl of Protein G Dynabeads (Life Technologies) were blocked with 0.5% BSA (w/v) in PBS. Magnetic beads were bound with 10 µg of anti-H3K27Ac antibody (Abcam ab4729). Additional antibodies used included anti-MYB (Abcam ab45150). Nuclei were isolated as previously described (Lee et al., 2006), and sonicated in lysis buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.1% SDS, and 1% Triton X-100) on a Misonix 3000 sonicator for 10 cycles at 30s each on ice (18-21 W) with 60 s on ice between cycles. Sonicated lysates were

cleared once by centrifugation and incubated overnight at 4°C with magnetic beads bound with antibody to enrich for DNA fragments bound by the indicated factor. Beads were washed with wash buffer A (50 mM HEPES-KOH pH7.9, 140 mM NaCl, 1 mM EDTA pH 8.0, 0.1% Na-Deoxycholate, 1% Triton X-100, 0.1% SDS), B (50 mM HEPES-KOH pH7.9, 500 mM NaCl, 1 mM EDTA pH 8.0, 0.1% Na-Deoxycholate, 1% Triton X-100, 0.1% SDS), C (20 mM Tris-HCl pH8.0, 250 mM LiCl, 1 mM EDTA pH 8.0, 0.5% Na-Deoxycholate, 0.5% IGEPAL C-630 0.1% SDS) and D (TE with 50 mM NaCl) sequentially. DNA was eluted in elution buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS). Cross-links were reversed overnight. RNA and protein were digested using RNase A and Proteinase K, respectively and DNA was purified with phenol chloroform extraction and ethanol precipitation. Additional cell line-specific details in the ChIP protocol are available upon request. Purified ChIP DNA was used to prepare Illumina multiplexed sequencing libraries. Libraries for Illumina sequencing were prepared following the Illumina TruSeq DNA Sample Preparation v2 kit. Amplified libraries were size-selected using a 2% gel cassette in the Pippin Prep system from Sage Science set to capture fragments between 200 and 400 bp. Libraries were quantified by qPCR using the KAPA Biosystems Illumina Library Quantification kit according to kit protocols. Libraries were sequenced on the Illumina HiSeq 2500 for 40 bases in single read mode. Reads were aligned to the hg19 revision of the human reference genome using bowtie with parameters `-best -k 2 -m 2 -sam` and `-l` set to read length 37^2 . Read pileup in 50bp bins was determined using MACS with parameters `-w -S -space=50 -shiftsize=200 -nomodel` ⁴⁹ ³. WIG file output from MACS was visualized in the

UCSC genome browser 50⁴. ChIP-Seq data has been submitted to GEO, accession number pending.

Allelic ChIP quantification

To quantify binding of proteins to different alleles, we aligned ChIP-Seq reads for MYB to custom small reference genomes for the reference sequence and mutant sequence at the known genomic loci. Bowtie was used to align reads with parameters `-best -chunkmbs 256 -l 40 -strata -m 1 -n 0 -S` to minimize mismatches with the small custom reference genomes. Reads that mapped with these parameters to these references were counted and plotted. Small custom genomes are listed below.

DU528:

AAAAAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACCGCAAAGAAACCGT
TAGAATCCATCCCTGCGCCCTGA

DU528 REF:

AAAAAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACCGCAAAGAAACCG
TTAGAATCCATCCCTGCGCCCTGA

PF382:

CAGGAAGCAGCCTCTTCAACCGCAAAGAAACCGTTTAACCGCAAAGA
AACCGTTAGAATCCATCCCTGCGCCCT

PF382 REF:

CAGGAAGCAGCCTCTTCAACCGCAAAGAAACCGTTAGAATCCATCCCTG
CGCCCT

Quantitative Real-Time Polymerase Chain Reaction (qRT-PCR).

For primary samples and cell lines, total RNA was extracted with a RNeasy Mini Kit (Qiagen) as per manufacturer's protocol and concentrations were measured on a Nanodrop 1000 spectrophotometer (Thermo Scientific). For two-step qRT-PCR, cDNA was synthesised initially with the Omniscript RT Kit (Qiagen) and 200 ng input RNA was used for each reaction. For the sorted thymic subsets, cDNA was provided by our collaborators where the methods for thymocyte isolation, RNA extraction, and cDNA synthesis have been described previously⁵. All qPCR reactions used FastStart Universal SYBR Green Master (ROX) mix as per manufacturer's protocol and samples were run on a Mastercycler epgradient S thermocycler (Eppendorf). Primer pairs for *LMO2* were 5'- ATTGGGGACCGCTACTTCCT -3' (forward) and 5'- TCTTGCCCAAAAAGCCTGAGAT-3' (reverse). Primer pairs for the housekeeping gene *GAPDH* were 5' - TGCACCACCAACTGCTTAGC -3' (forward) and 5' - GGCATGGACTGTGGTCATGAG - 3' (reverse). *LMO2* expression was considered as absent if no signal was detected after 40 cycles of PCR amplification. Normalised expression ratios were calculated by the efficiency-corrected ΔC_t method whilst using *GAPDH* as the endogenous reference mRNA as described at length by Bookout et al⁶.

Characterization of transcript start position by rapid amplification of cDNA to the 5' end (5'RACE)

Amplification of mature *LMO2* transcripts to the 5' end in PF-382, DU.528 and Loucy cell lines was achieved by using the SMARTer RACE 5'3' Kit (Clontech) as

per manufacturer's guidelines. Briefly, a gene-specific primer (GSP) was designed against the final exon of *LMO2* to capture all isoforms, appended with a 15 bp overlap sequence to the 5' end to allow for cloning. The following GSP was used for the reaction: 5'-GATTACGCCAAGCTTCCCTTACCCACCCCTCAAACCCCA-3'. First, RACE-ready cDNA was synthesised with SMARTScribe Reverse Transcriptase coupled with a proprietary 5' specific SMARTer II A oligonucleotide. Then, RACE-ready cDNA was used as the template for RACE PCR reactions run with 10X Universal Primer Short, and the aforementioned 5' GSP. RACE products were cloned into the pRACE vector and used to transform Stellar Competant Cells. Colonies picked and plasmid DNA was isolated by QIAprep Spin Miniprep Kit (Qiagen). Isolated DNA was analysed by Sanger sequencing off an M13 primer and mapped to the *LMO2* locus by using the UCSC blat tool to determine the transcript start positions.

Mutation screening at LMO2 intron 1 by denaturing high-performance liquid chromatography (dHPLC).

Genomic DNA extracts were amplified by PCR with Phusion High-Fidelity PCR Master Mix and HF Buffer (New England Biolabs, UK) as per manufacturer's instructions. Primers were designed against *LMO2* intron 1 giving a total amplicon size of 204 base pairs. The primer pairs used were 5'-CAGGCGGGTGTCCCTTGATA-3' (forward) and 5'-ACACCAGTCCTGTTCATTTGG-3' (reverse). Final PCR products were denatured and allowed to re-anneal through a step-wise cooling program to allow for the formation of a heteroduplex for those samples with mutations. All products were then

analyzed on the WAVE dHPLC equipment (Transgenomic, UK) and samples with positive chromatograms were subject to Sanger sequencing. Large and complex indels were confirmed by TOPO cloning and sequencing.

Allelic discrimination via SNP analysis.

RNA samples were subjected to on-column DNase treatment (Qiagen) prior to cDNA synthesis with the Omniscript RT Kit (Qiagen). First, genomic DNA was amplified by PCR to ensure amplification of the rs3740617 SNP (T/C) within LMO2 with the following primers 5'- GTCCTTCTGTCACCTTGAAGTG -3' (forward) and 5' – TATGCCAGATCCAAATGCCAG- 3' (reverse). Samples that were informative i.e. heterozygous for the SNP, were then analyzed at the sample position by PCR with a paired cDNA template and were called monoallelic if only one of the two possible bases were observed at the SNP position.

Motif analysis

Patient and cell line-derived mutant sequences were analyzed using UniPROBE, a database generated through universal protein binding microarray (PBM) technology⁷. For patients P1, A1 and A6, where no motif was identified in UniPROBE, sequences were analyzed in Tfbind⁸. Note binding data for RUNX1 is not included in the UniPROBE database. To test whether potential motifs would reach significance when tested against multiple databases, *P* and *E* values were generated using Tomtom; *E* values <10 are considered to meet the match threshold when accounting for multiple testing⁹.

Luciferase reporter constructs and assays.

Genomic DNA extracts were amplified by PCR with Phusion High-Fidelity PCR Master Mix and HF Buffer (New England Biolabs, UK) as per manufacturer's instructions, using primers flanking the mutation hotspot, giving an approximate 469 base pairs product, depending on the size of the indel. Primers used were as follows 5'- TATATAGGTACCCACTTGCTTTCTCAGACCGG-3' (forward) and 5'-TATATACTCGAGCCTGCCTCTCCACTAGCTAC-3' (reverse) both of which included the restriction enzymes sites for KpnI and XhoI respectively. PCR products were cloned into the pGL3-promoter vector (Promega – E1761) into a multi clonal site upstream of a SV40 promoter and the firefly luciferase gene. For the luciferase assay, a total of 1×10^6 Jurkat cells were resuspended in 100 μ L of Ingenio Electroporation Solution (Mirus) along with 1.5 μ g of pGL3-promoter vector containing each respective cloned insert and 250 ng of renilla control plasmid (pTK). Cells were electroporated on the D-23 program (Amaxa) and allowed to recover for 48 hours in 1000 μ L RPMI supplemented with 10% FCS and incubated at under standard tissue culture conditions (37°C and 5% CO₂). Cells were harvested and luciferase activity was assessed using the Dual-Glo Luciferase Assay System (Promega – E2920) in triplicate. Firefly luciferase activity was normalised to renilla luciferase and data shown was the ratio relative to the no-insert (empty) vector.

Retroviral transduction of PF-382 with LMO1

We anticipated that loss of *LMO2* expression through successful genome editing of the aberrant promoter would result in loss of cell viability and inability to expand

single cell clones. We thus expressed LMO1 in PF-382 cells through retroviral infection, given it can replace LMO2 in the LMO-TAL1 complex. *LMO1* was amplified from PCS2-LMO1 (a gift from Takaomi Sanda) by PCR with Phusion High-Fidelity PCR Master Mix and HF Buffer (NEB) as per manufacturer's instructions. Primers were designed to include digest sites for restriction enzymes BglIII on the forward sequence (BglIII-LMO1-F), and EcoRI-HF on the reverse (EcoRI-LMO1-R). The primer pairs used were BglIII-LMO1-F 5'-TATATAGATCTGCCACCATGATGGTGCTGGACAAGGAGGACGGCGTG - 3' and EcoRI-LMO1-R 5'-ATATAGAATTCTTACTGAACTTGGGATTCAAAGGTGCCATTGAGC. - 3' The PCR product was digested with BglIII and EcoRI-HF and cloned into the corresponding digest sites of MSCV-puro plasmid. The retrovirus was generated in human embryonic kidney 293T (HEK293T) cells, which were chemically transfected with 18µl of FUGENE and 222µl of OPTIMEM supplemented with 4 µg of MSCV-LMO1-puromycin, 2µg of VSVG (pMD2.G) and 4µg of pMD.MLV. The mixture was added dropwise to the HEK293T cells. After 48 hours, the retrovirus was collected by harvesting the culture medium and concentrated by using an Amicon filter (Milipore) as per manufacturer's instructions. PF-382 cells were infected with the MSCV-LMO1-puromycin retrovirus, by resuspending 1x10⁶ cells in 3 ml of the aforementioned viral media along with polybrene at 8 µg/ml and transferred to a 24-well culture plate. The plate was centrifuged at 2,500g for 1.5 hours at 37°C and incubated overnight to assist in the infection process. The next day, cells were centrifuged, the viral media aspirated off and resuspended in fresh RPMI.

PF-382 cells constitutively expressing *LMO1* were then selected by puromycin after 48 hours at a concentration of 2 µg/ml.

CRISPR/Cas9 genome editing of PF-382

Knock out of the *LMO2* intron 1 mutation in the PF-382 *LMO1* positive cell line was achieved by using CRISPR/Cas9 genome editing technology. Guide RNAs were designed against the PF-382 mutation by using the CRISPR design tool (<http://crispr.mit.edu>)¹⁰. Two guides were annealed and cloned into the BbsI sites found within the pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid (Addgene plasmid # 42230)¹¹. The guides used are as follows: guide#1-up 5'-CACCGATTCTAACGGTTTCTTTGC-3' and guide#1-down 5'-AAACGCAAAGAAACCGTTAGAATC-3'. Single cells were sorted by exploiting a BFP selectivity marker within the pX330 plasmid by fluorescent activated cell sorting into 96 well plates, and incubated under standard tissue culture conditions (37°C and 5% CO₂) in RPMI supplemented with 10% FCS. Once single cells had grown into colonies, gDNA was extracted by using the QuickExtract DNA Extraction solution (Epicentre) as per manufacturer's instructions and clones were screened for mutations by Sanger sequencing.

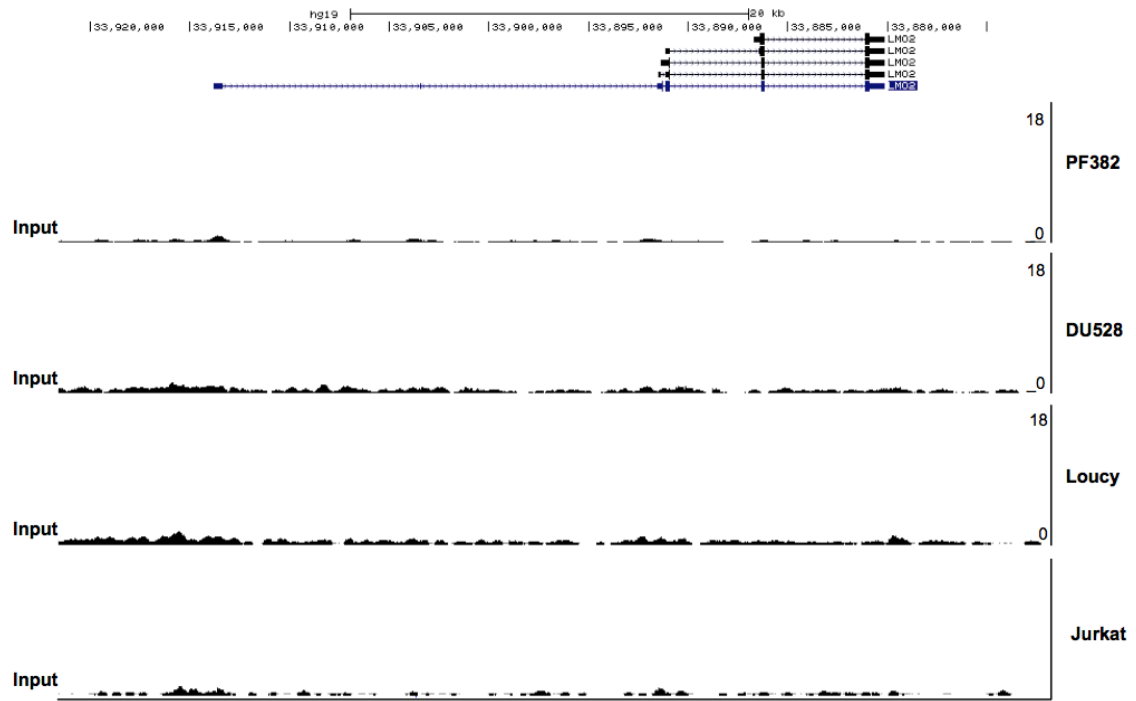
Identification of TCR-γ biallelic deletion and characterisation of genetic mutations in primary T-ALL samples.

Absence of Biallelic Deletion (ABD) at the T cell receptor gamma (TCR-γ) gene locus was determined for all the patients using genomic DNA from diagnostic

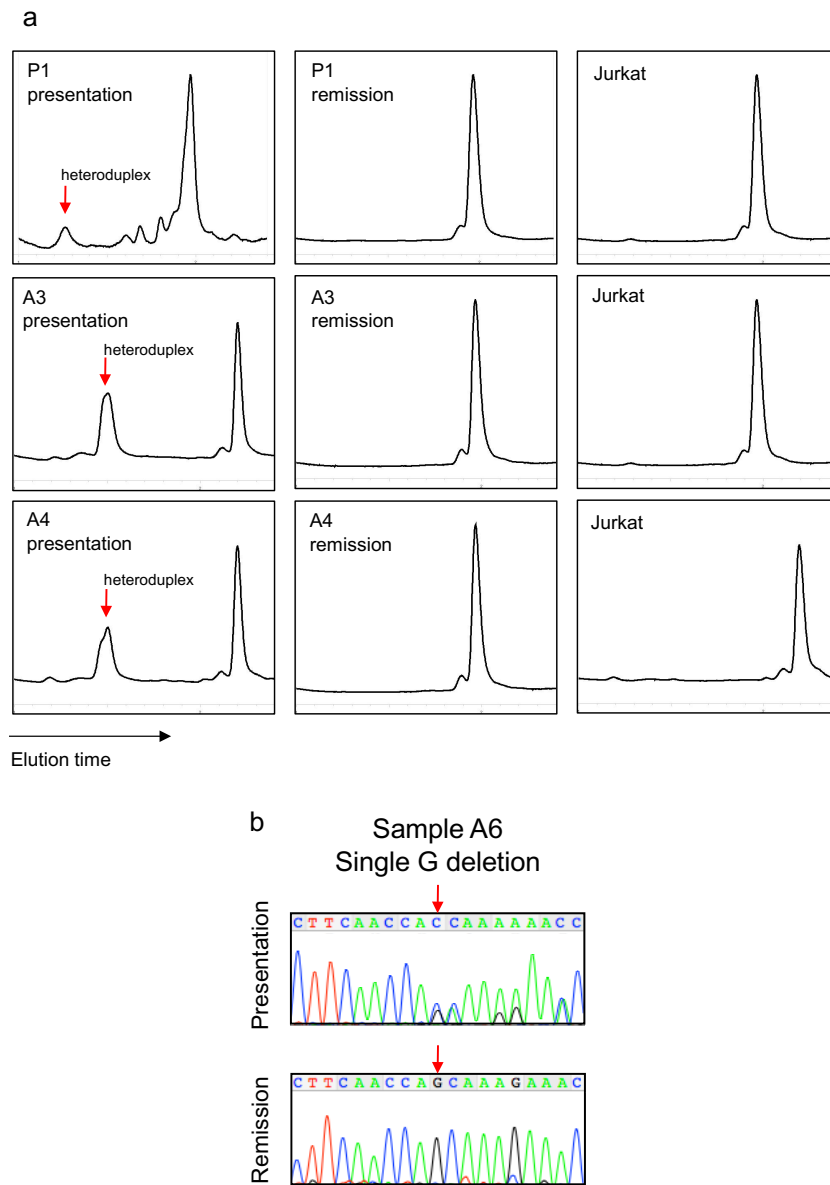
samples followed by qPCR. Notably, ABD is concomitant with early thymic progenitors that would not have rearranged the TCR- γ locus. Determination of ABD by this method has been previously outlined, and the same primers were used in the present study¹². All qPCR reactions were set up in triplicate with FastStart Universal SYBR Green Master (ROX) mix as per manufacturer's protocol and samples were run on a Mastercycler epgradient S thermocycler (Eppendorf). Mean Ct values were calculated and reactions were repeated if the standard deviation of the reference gene *ANLN* Ct values was greater than 0.5.

FBXW7 and *NOTCH1* mutations were identified by PCR followed by denaturing high-performance liquid chromatography or Sanger sequencing. The following genomic regions of *NOTCH1* were amplified for mutation analysis: HD-N (exon 26), HD-C (exon 27), and PEST domains (exon 34). For *FBXW7* the WD40 domain (exons 9, 10 and 12) were amplified for mutation screening. These methods including the primers used have been described previously¹³.

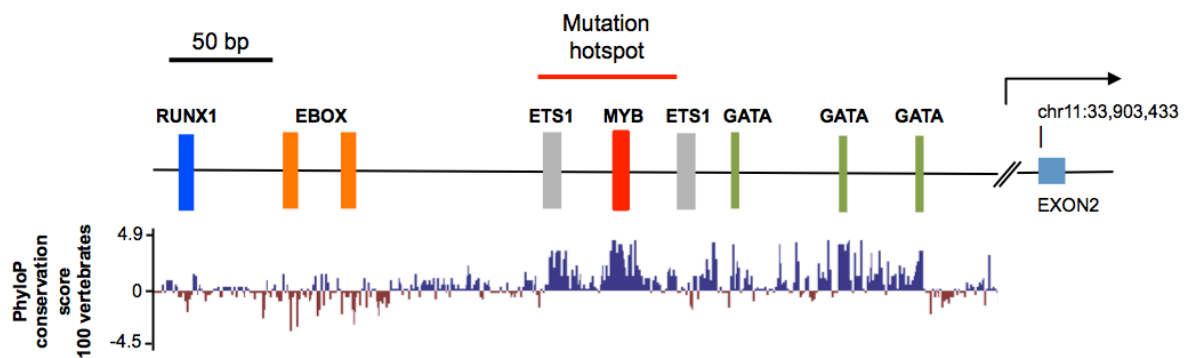
SIL-TAL1 deletions were detected primarily by PCR of genomic DNA with the forward primer Sildb.F 5'-AAGGGGAGCTAGTGGGAGAAA-3' coupled with reverse primer Tal1db1-R 5'-AGAGCCTGTCGCCAAGAA-3' yielding a 300 bp product when the deletion is present. A secondary form was detected by using the aforementioned Sildb.F primer with the reverse primer Tal1db2-R 5'-TTGTAAAATGGGGAGATAATGTCGAC-3' giving a 359 bp product when the deletion is present. Both PCRs have been described previously¹⁴.



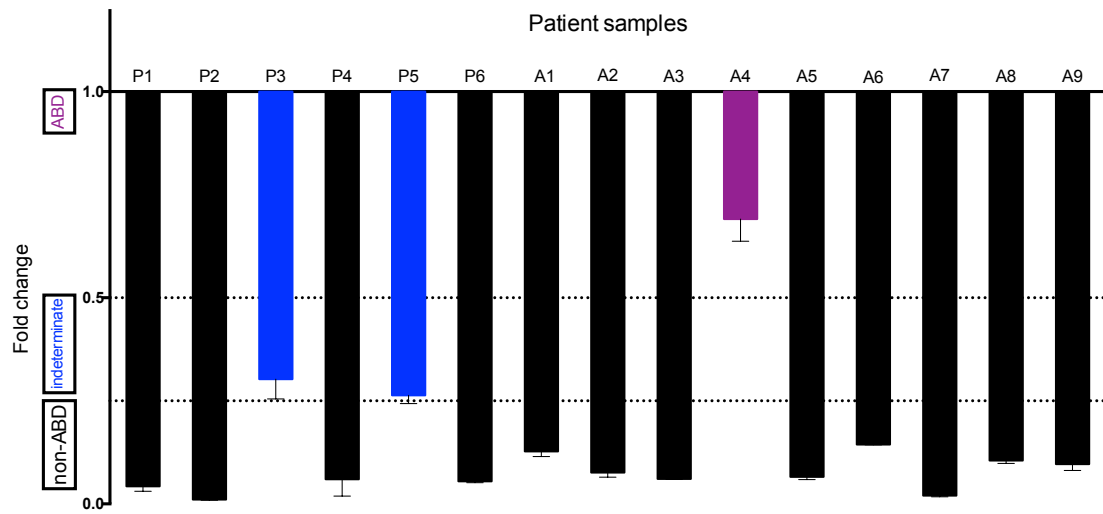
Supplementary Figure 1. Input ChIP-Seq controls for PF-382, DU.528, Loucy and Jurkat T-ALL cell lines. Control tracks for data presented in Fig. 1, b. Y-axis values are reads per bin per million mapped reads (RPM).



Supplementary Figure 2. Representative examples of presentation and remission gDNA at *LMO2* intron 1 mutational hotspot as analyzed by dHPLC and Sanger sequencing (a) Comparison of dHPLC traces following PCR of presentation gDNA and patient-matched remission gDNA at the *LMO2* intron 1 locus, hg19, chr11: 33,903,787 – 33,903,584. Jurkat is shown as the negative control with elution time along the x-axis. Mutant heteroduplexes are labelled with a red arrow (b) Sequence trace comparison of the *LMO2* intron 1 mutation observed in patient A6 at presentation and remission.



Supplementary Figure 3. The ETS1, GATA and MYB binding sequences at the LMO2 intron 1 mutation hotspot are highly conserved in vertebrates. To scale schematic of the LMO2 intron 1 locus showing binding sites for the TAL1 complex aligned to the conservation score from 100 vertebrates as determined by PhyloP using the UCSC genome browser.

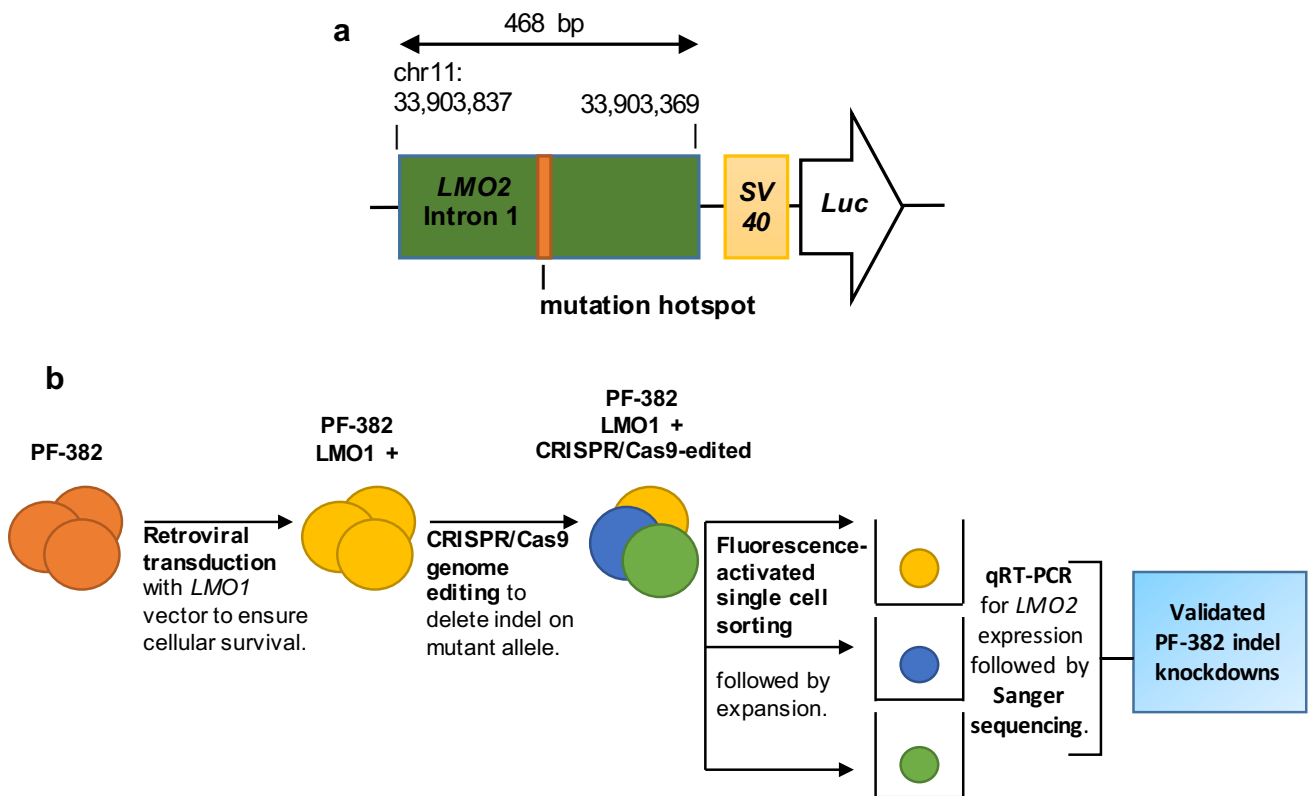


Supplementary Figure 4. Absence of biallelic deletion at TCR- γ (ABD) by qPCR for primary T-ALL samples. Fold change was calculated using the comparative delta Ct method using gDNA from HEK293T cells (that do not have rearrangement at the TCR γ locus) as a calibrator. ABD and non-ABD status was assigned if fold change was above 0.5 and less than 0.25 respectively. Samples with a fold change between 0.25 and 0.5 were assigned an indeterminate ABD status.

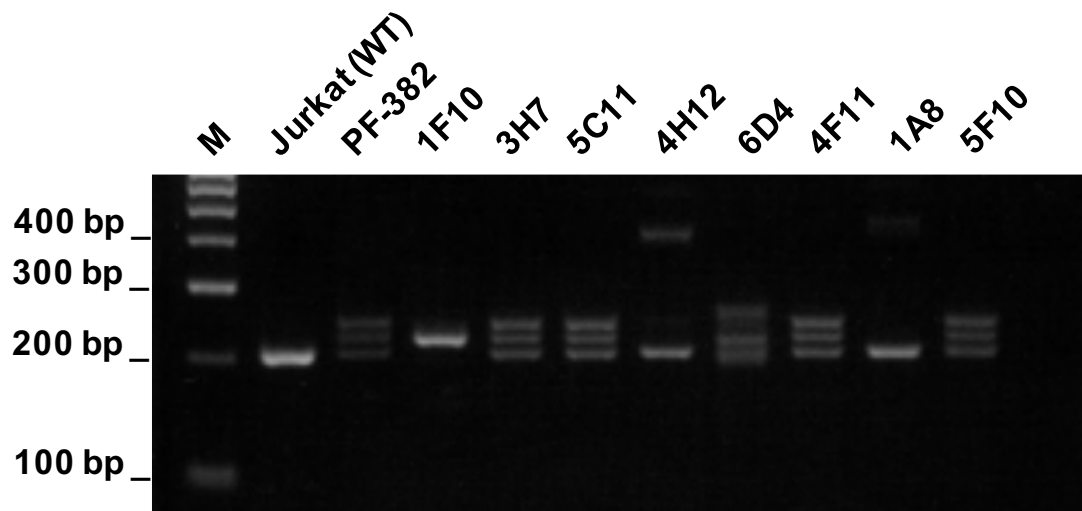
MYB ChIP-seq reads – DU.528

AAAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACCAGCAAAGAAACCGTTAGAAATCC	REF
-----AAAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAG-----	MUT
-----AAAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAG-----	MUT
-----AAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGA-----	MUT
-----AAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGA-----	MUT
-----AAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGA-----	MUT
-----AAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGA-----	MUT
-----AAAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGA-----	MUT
-----AAAAGAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGA-----	MUT
-----AAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAA-----	MUT
-----AAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAA-----	MUT
-----AAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAA-----	MUT
-----AAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAA-----	MUT
-----AAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAA-----	MUT
-----AAAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAA-----	MUT
-----AAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAA-----	MUT
-----AAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAA-----	MUT
-----AAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAA-----	MUT
-----AAGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAA-----	MUT
----- AAGAAGTCGGCAGGAAGCAGCCTCTTCAACCAGCAAAGAA -----	WT
-----AGAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAAC-----	MUT
-----GAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACC-----	MUT
-----GAAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACC-----	MUT
-----AAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCG-----	MUT
-----AAGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCG-----	MUT
-----AGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGT-----	MUT
-----AGTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGT-----	MUT
-----GTCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGTT-----	MUT
-----TCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGTTA-----	MUT
-----TCGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGTTA-----	MUT
-----CGGCAGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGTTAG-----	MUT
-----AGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGTTAGAAATC-----	MUT
-----AGGAAGCAGCCTCTTCAACC-GCAAAGAAACCGTTAGAAATC-----	MUT

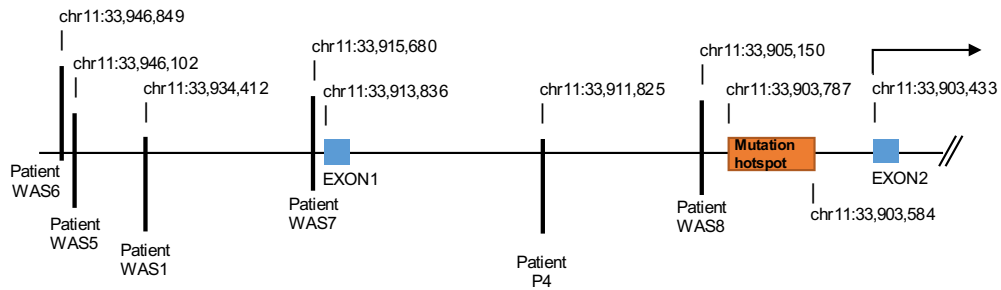
Supplementary Figure 5. Selected allele-specific ChIP-Seq mapped reads ChIP-seq reads mapped to mutant allele in the DU.528 cell line to the wild type (WT) or mutant (MUT) allele.



Supplementary Figure 6. Schematic showing the design of the luciferase reporter and the workflow of CRISPR/Cas9 experiments. (a) Schematic of the luciferase reporter construct, which includes a 469 bp stretch across the mutational hotspot of *LMO2* intron 1 inserted upstream of a minimal SV40 promoter and the luciferase gene. **(b)** PF-382 cells were first retrovirally infected to stably express the closely related *LMO1* gene, to counteract the possibility of cell death following knockout of the *LMO2* intron 1 mutation. Following single cell sorting and expansion, clones were screened for mutations by Sanger sequencing.



Supplementary Figure 7. Gel electrophoresis of the *LMO2* Intron 1 hotspot in PF-382 CRISPR/Cas9-edited clones. Gel electrophoresis following PCR of the *LMO2* intron 1 region using gDNA isolated from Jurkat, PF-382 and PF-382 CRISPR/Cas9 edited clones, 1F10, 3H7, 5C11, 4H12, 6D4, 4F11, 1A8 and 5F10.



Supplementary Figure 8. Retroviral integration sites in gene therapy-induced T-ALL about *LMO2* as reported by Braun et al. and Haccin-Bey-Abina et al^{15,16}. Schematic demonstrating the integration sites following patient treatment with WASP-expressing retroviral vectors for the treatment of Wiskott-Aldrich Syndrome (Patient WAS#) and MFG- γ c (encoding the *IL2R* common gamma chain) for severe combined immunodeficiency (Patient P4). All are plotted in relation to the hotspot of somatic mutation within *LMO2* intron 1.

RANK	JURKAT	DU.528	PF-382	Loucy
1.	UP00092_2 Myb_secondary 4.35e-61	UP00080_1 Gata5_primary 1.80e-12	UP00081_2 Mybl1_secondary 4.90e-37	UP00081_2 Mybl1_secondary 3.12e-09
2.	UP00081_2 Mybl1_secondary 3.72e-53	UP00100_1 Gata6_primary 9.05e-12	UP00092_2 Myb_secondary 9.42e-32	UP00002_1 Sp4_primary 1.27e-07
3.	UP00081_1 Mybl1_primary 5.56e-20	UP00092_1 Myb_primary 7.18e-08	UP00279_1 Rsc30 1.25e-12	UP00013_1 Gabpa_primary 4.83e-06
4.	UP00092_1 Myb_primary 8.70e-17	UP00287_1 Gat1 1.22e-07	UP00000_2 Smad3_secondary 2.25e-11	UP00092_2 Myb_secondary 7.96e-06
5.	UP00002_1 Sp4_primary 1.09e-15	UP00092_2 Myb_secondary 2.15e-07	UP00081_1 Mybl1_primary 8.15e-11	UP00080_1 Gata5_primary 0.00044
6.	UP00000_2 Smad3_secondary 2.91e-12	UP00347_1 Gzf3 2.28e-07	UP00099_1 Ascl2_primary 3.07e-09	UP00021_1 Zfp281_primary 0.00061
7.	UP00093_1 Klf7_primary 3.66e-12	UP00081_2 Mybl1_secondary 3.81e-07	UP00065_2 Zfp161_secondary 2.53e-08	UP00100_1 Gata6_primary 0.00074
8.	UP00065_2 Zfp161_secondary 1.17e-08	UP00081_1 Mybl1_primary 2.62e-06	UP00043_2 Bcl6b_secondary 3.19e-08	UP00081_1 Mybl1_primary 0.0034
9.	UP00279_1 Rsc30 1.82e-08	UP00318_1 Gln3 4.75e-06	UP00092_1 Myb_primary 6.08e-08	UP00093_1 Klf7_primary 0.0061
10.	UP00043_2 Bcl6b_secondary 2.81e-08	UP00032_1 Gata3_primary 0.00025	UP00002_1 Sp4_primary 1.64e-07	UP00033_2 Zfp410_secondary 0.013

Supplementary Table S1. AME (Analysis of Motif Enrichment) performed for MYB ChIP-seq data from T-ALL cell lines¹⁷. Most enriched motif IDs are shown together with a P value with Bonferroni correction for multiple testing (number of motifs x number of thresholds tested).

CHROM	START	END	NAME	TRANSFORMED_P	RANK	ACTUAL_P
PF382_MYB						
chr11	33903044	33904242	MACS_peak_3088	2204.93	1044	3.21E-221
chr11	33912852	33914188	MACS_peak_3089	1560.04	1784	9.91E-157
chr11	33914910	33916135	MACS_peak_3090	758.37	4238	1.46E-76
DU528_MYB						
chr11	33902372	33904415	MACS_peak_3512	3213.47	16	4.496e-322
chr11	33914670	33916188	MACS_peak_3515	2025.57	2021	2.77E-203
chr11	33906426	33907585	MACS_peak_3513	574.45	7133	3.59E-58
chr11	33913001	33914127	MACS_peak_3514	118.1	19885	1.55E-12
Loucy_MYB						
chr11	33914613	33917117	MACS_peak_3673	2435.96	2519	2.54E-244
chr11	33912252	33914417	MACS_peak_3672	726.99	7903	2.00E-73
Jurkat_MYB						
chr11	33914880	33916121	MACS_peak_5137	857.99	7978	1.59E-86
chr11	33913049	33913949	MACS_peak_5136	194.25	21862	3.76E-20
PF382_H3K27ac						
chr11	33902008	33907187	MACS_peak_2953	3100	95	0
chr11	33912047	33916915	MACS_peak_2954	991.12	3533	7.73E-100
DU528_H3K27ac						
chr11	33899100	33909513	MACS_peak_3076	3100	191	0
chr11	33910575	33918245	MACS_peak_3077	2484.48	1946	3.56E-249
Loucy_H3K27ac						
chr11	33912047	33918539	MACS_peak_2487	3100	80	0
chr11	33918606	33919909	MACS_peak_2488	135.59	11908	2.76E-14
Jurkat_H3K27ac						

Supplementary Table S2. ChIP-Seq peak calling. Peaks were defined using MACS³ with parameters --keep-dup=auto -p 1e-9 and input control, and peaks between chr11 33870000 and 33920000 are reported.

Mutation Start Coordinate (hg19, chr11)	Sample	Type	Mutation	Mutant sequence	WT sequence	TF binding site
33,903,641	P1	Pediatric	GTGGGGCTC 9 bp ins CCCTGATGCCAA 12 bp del	GTTAGAATCCATCCCTGCGGT GGGGCTCAGTTCCGCCT	GTTAGAATCCATCCCTGCGCC CTGATGCCAAAGTTCCGCCT	RUNX1
33,903,656	P2	Pediatric	AC 2 bp ins GAATCCATCCCTG 13 bp del	GAAACCGTTAACCGCCCTGAT GCCAAAG	CAGCCTCTTCAACCAGCAAAG AAACCGTTAGAATCCATCCCTG CGCCCTGATGCCAAAG	MYB (secondary motif)
33,903,672	P3	Pediatric	A 1 bp del	CTTCAACCGCAAAGAAACCGT TAGAATCCATCCCTGCGCCCT GATG	CTCTTCAACCAAGCAAAGAAAC CGTTAGAATCCATCCCTGCGC CCTGATG	MYB (secondary motif)
33,903,672	P4	Pediatric	A 1 bp del	CTTCAACCGCAAAGAAACCGT TAGAATCCATCCCTGCGCCCT GATG	CTCTTCAACCAAGCAAAGAAAC CGTTAGAATCCATCCCTGCGC CCTGATG	MYB (secondary motif)
33,903,672	P5	Pediatric	A 1 bp del	CTTCAACCGCAAAGAAACCGT TAGAATCCATCCCTGCGCCCT GATG	CTCTTCAACCAAGCAAAGAAAC CGTTAGAATCCATCCCTGCGC CCTGATG	MYB (secondary motif)
33,903,670	P6	Adult	G 1 bp del	CCTCTTCAACCACAAAGAAAC CGTTAG	CCTCTTCAACCAAGCAAAGAAA CCGTTAG	RUNX1
33,903,672	C1	Cell Line DU.528	A 1 bp del	GAAGCAGCCTCTTCAACCGCA AAGAAACCGTTAGAATCCATC CCT	GGCAGGAAGCAGCCTCTTCAA CCAGCAAAGAAACCGTTAGAA TCCATCCCT	MYB (secondary motif)

Table cont...

Mutation Start Co-ordinate (hg19, chr11)	Sample	Type	Mutation	Mutant sequence	WT sequence	TF binding site
33,903,676	C2	Cell Line PF-382	AACCAGCAAAGAA ACCGTTT 20 bp ins	AGCAGCCTCTTCAACCAGCAA AGAAACCGTTTAAACCAGCAA GAAACCGTTAGAATCCATCCC T	AGCAGCCTCTTCAACCAGCAA AGAAACCGTTAGAATCCATCC CT	MYB (primary motif)
33,903,639	A1	Adult	C>G 1 bp substitution T 1 bp del	CCATCCCTGCGCCGATGCC AAAGTTCCGCCTGCC	CCATCCCTGCGCCCTGATGCC AAAGTTCCGCCTGCC	ETS1
33,903,672	A2	Adult	A 1 bp del	CTTCAACCGCAAAGAAACCGT TAGAATCCATCCCTGCGCCCT GATG	CTCTTCAACCAAGCAAAGAAAC CGTTAGAATCCATCCCTGCGC CCTGATG	MYB (secondary motif)
33,903,724	A3	Adult	GAAGAATAAGAAG AAAAAAAAAAGAA GTCGGCAGGAAG CAGCCTCTTCAAC CAGCAAAGAAACC GTTA 68 bp ins	TTCACATTACAAGCTGGGCT GGTAAGTGAAGAATAAGAAGA AAAAAAAAAGAAGTCGGCAGG AAGCAGCCTCTTCAACCAGCA AAGAAACCGTTAGAAGAA	TTCACATTACAAGCTGGGCT GGTAAGTGAAGAA	ETS1 MYB (primary motif)

Table cont...

Mutation Start Co-ordinate (hg19, chr11)	Sample	Type	Mutation	Mutant sequence	WT sequence	TF binding site
33,903,724	A4	Adult	GAAGAATAAGAAG AAAAAAAAAAGAA GTCGGCAGGAAG CAGCCTCTTCAAC CAGCAAAGAAACC GTTAGAATC 73 bp ins	AAGCTGGGCTGGTAAGTGAAG AATAAGAAGAAAAAAAAAGA AGTCGGCAGGAAGCAGCCTC TTCAACCAGCAAAGAAACCGT TAGAATCGAAGAATAAGAAGA AAAAAAAAAAG	AAGCTGGGCTGGTAAGTGAAG AATAAGAAGAAAAAAAAAAG	ETS1 MYB (primary motif)
33,903,671	A5	Adult	A>C 1 bp substitution	GCAGCCTCTTCAACCCGCAA GAAA	GCAGCCTCTTCAACCAGCAA GAAA	UNKNOWN
33,903,670	A6	Adult	G 1 bp del	CCTCTTCAACCACAAAGAAAC CGTTAG	CCTCTTCAACCA G CAAAGAAA CCGTTAGAATCCATCCCTG	RUNX1
33,903,637	A7	Adult	T>G 1 bp substitution	CATCCCTGCGCCC G GATGCC AAAGTTC	CATCCCTGCGCCCTGATGCCA AAGTTCCG	ETS1
33,903,885	A8 1 st mutation	Adult	A>G 1 bp substitution	ACTCAGAGGGGATAGGAGATTT GCAA	ACTCAGAGGGGATAAGAGATTT GCAAAGCGTGAGACA	Unknown
33,903,637	A8 2 nd mutation	Adult	T>G 1 bp substitution	CATCCCTGCGCCC G GATGCC AAAGTTC	CATCCCTGCGCCCTGATGCCA AAGTTCCG	ETS1
33,903,640	A9	Adult	TAAGAAGAAAAA AAAAGAAGTCGGC AGGAAGCAGCCTC TTCAACCAGCAA GAAACCGTTAGAA TCCATCCCTGCG 77 bp ins	CCCTGCGTAAGAAGAAAAA AAAGAAGTCGGCAGGAAGCA GCCTCTTCAACCAGCAAAGAA ACCGTTAGAATCCATCCCTGC GCCTGATT	CCCTGCGCCCTGATGCCAAAG TTCCGCCTGCCCCACCCGTCA CGCTATCAAGGACACCC	ETS1 MYB (primary motif)

Table S3. Mutations identified in primary T-ALL samples and cell lines, PF-382, and DU.528. Mutation start points are given as hg19 co-ordinates, and the nature of the indels are described. Underlined sequences show the consensus sites for the transcription factors (TF) that were identified by *in silico* analysis.



Sample ID	Primary MYB motif	Uniprobe E.S. (Top Kmer 0.49)	Tomtom Motif ID	Tomtom P value	Tomtom E value (<10)	Secondary MYB motif	Uniprobe E.S. (Top Kmer 0.49)	Tomtom Motif ID	Tomtom P value	Tomtom E value (<10)
	(Uniprobe)					(Uniprobe)				
										
P1										
P2						TAACCGCC	0.47	MA0100.2 (Myb)	4.20E-03	6.00E+00
P3						CAACCGCAA	0.37	MA0100.2 (Myb)	5.70E-03	8.10E+00
P4						CAACCGCAA	0.37	MA0100.2 (Myb)	5.70E-03	8.10E+00
P5						CAACCGCAA	0.37	MA0100.2 (Myb)	5.70E-03	8.10E+00
P6										
A1										
A2						CAACCGCAA	0.37	MA0100.2 (Myb)	5.70E-03	8.10E+00
A3	AACCGTTA	0.49	UP00092_1	2.29E-05	3.28E-02					
A4	AACCGTTA	0.49	UP00092_1	2.29E-05	3.28E-02					
A6										
A7										
A8										
A9	AACCGTTA	0.49	UP00092_1	2.29E-05	3.28E-02					
DU.528						CAACCGCAA	0.37	MA0100.2 (Myb)	5.70E-03	8.10E+00
PF-382	AACCGTTT	0.48	UP00092_1	1.43E-03	2.06E+00					
REFERENCE	AACCGTTA	0.49	UP00092_1	2.29E-05	3.28E-02					

Table Cont....

Table cont....



Sample ID	ETS1 motif (Uniprobe) 	Uniprobe E.S. (Top Kmer 0.50)	Tomtom Motif ID	Tomtom P value	Tomtom E value (<10)	RUNX1 motif (HOCOMOCO) 	Tfbind score (>0.83)	Tomtom Motif ID	Tomtom P value	Tomtom E value (<10)
P1						CTGCGGT	0.959	MA0002.2	2.36E-03	3.38E+00
P2										
P3										
P4										
P5										
P6						TTGTGGTT	1	MA0002.2	1.70E-04	2.50E+00
A1	GCCGGATGC	0.49	ETS1_full_2 (HumanTF 1.0)	8.10E-03	7.80E+00					
A2										
A3	GCAGGAAGC	0.47	MA0098.2	5.50E-04	7.90E-01					
A4	GCAGGAAGC	0.47	MA0098.2	5.50E-04	7.90E-01					
A6						TTGTGGTT	1	MA0002.2	1.70E-04	2.50E+00
A7	CCCGGATG	0.48	ETS1_full_2 (HumanTF 1.0)	1.19E-03	1.70E+00					
A8	CCCGGATG	0.48	ETS1_full_2 (HumanTF 1.0)	1.19E-03	1.70E+00					
A9	GCAGGAAGC	0.47	MA0098.2	5.50E-04	7.90E-01					
DU.528										
PF-382										
REFERENCE										

Table S4. Motif analysis of patient and cell line-derived mutations. Sequences containing mutations from Table S2 were interrogated using UniPROBE. Note RUNX1 is not included in the UniPROBE database, thus sequences were also analyzed using Tfbind for samples P1, P6 and A6 where no match was identified with UniPROBE. The closer the UniPROBE enrichment scores (E.S.) to the top scoring Kmer, the more significant the alignment. The Tfbind significance threshold score for RUNX1 is >0.83. Statistics for motif alignment using Tomtom are also shown, where *E* values <10 are considered to meet the match threshold accounting for multiple testing.

Sample	Age	Sex	Presenting WCC (x10 ⁹ / L)	Extramedullary disease	ABD of TCR- γ by qPCR	NOTCH	FBXW7	MuTE	SIL-TAL deletion
P1	7	M	104	-	Non-ABD	HD-N	WT	WT*	Positive
P2	16	M	33.2	-	Non-ABD	HD-N	WD40	WT*	Positive
P3	9	M	248	-	Indeterminate	WT	WT	Mutant	Negative
P4	10	M	157	-	Non-ABD	HD-C	WT	WT	Negative
P5	16	M	313	-	Indeterminate	WT	WT	WT*	Negative
P6	15	M	320	-	Non-ABD	WT	WT	WT	Positive
A1	53	F	47	Mediastinal mass	Non-ABD	WT	WT	WT	Positive
A2	31	M	56	No	Non-ABD	HD-N	WT	Mutant	Negative
A3	27	M	53	Spleen/nodes	Non-ABD	PEST	WT	WT	Negative
A4	25	M	147	No	ABD	WT	WT	WT	Negative
A5	24	M	264	No	Non-ABD	WT	WT	WT	Positive
A6	21	M	400	Mediastinal mass	Non-ABD	HD-N	WT	WT	Negative
A7	34	M		UNK	Non-ABD	WT	WT	WT	Positive
A8	22	M	140	Mediastinal mass	Non-ABD	WT	WT	WT	Negative
A9	17	M	354	NO	Non-ABD	HD-N;PEST	ND	WT	Negative

Table S5. Clinical and genetic features of primary T-ALL samples. Mutation screening for samples marked with an * was achieved by dHPLC analysis. For all other samples mutation screening was achieved by Sanger sequencing. MuTE: mutation of the *TAL1* enhancer (Mansour et al., 2014). ABD: Absence of biallelic TCR gamma deletion.

Supplementary References

1. Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols*. 2006;1(2):729-748.
2. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009;10(3):R25.
3. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. 2008;9(9):R137.
4. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Research*. 2002;12(6):996-1006.
5. Dik WA, Pike-Overzet K, Weerkamp F, et al. New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. *J Exp Med*. 2005;201(11):1715-1723.
6. Bookout AL, Cummins CL, Mangelsdorf DJ, Pesola JM, Kramer MF. High-throughput real-time quantitative reverse transcription PCR. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]*. 2006;Chapter 15:Unit- Uni8.
7. Badis G, Berger MF, Philippakis AA, et al. Diversity and complexity in DNA recognition by transcription factors. *Science*. 2009;324(5935):1720-1723.
8. Tsunoda T, Takagi T. Estimating transcription factor bindability on DNA. *Bioinformatics*. 1999;15(7-8):622-630.
9. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24.
10. Hsu PD, Scott DA, Weinstein JA, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*. 2013;31(9):827-832.
11. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, NY)*. 2013;339(6121):819-823.
12. Gutierrez A, Dahlberg SE, Neuberg DS, et al. Absence of biallelic TCRgamma deletion predicts early treatment failure in pediatric T-cell acute lymphoblastic leukemia. *Journal of Clinical Oncology*. 2010;28(24):3816-3823.
13. Mansour MR, Sulis ML, Duke V, et al. Prognostic implications of NOTCH1 and FBXW7 mutations in adults with T-cell acute lymphoblastic leukemia treated on the MRC UKALLXII/ECOG E2993 protocol. *J Clin Oncol*. 2009;27(26):4352-4356.
14. Pongers-Willemse MJ, Seriu T, Stolz F, et al. Primers and protocols for standardized detection of minimal residual disease in acute lymphoblastic leukemia using immunoglobulin and T cell receptor gene rearrangements and TAL1 deletions as PCR targets: report of the BIOMED-1 CONCERTED ACTION: investigation of minimal residual disease in acute leukemia. *Leukemia*. 1999;13(1):110-118.
15. Braun CJ, Boztug K, Paruzynski A, et al. Gene therapy for Wiskott-Aldrich syndrome--long-term efficacy and genotoxicity. *Sci Transl Med*. 2014;6(227):227ra233.

16. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science (New York, NY)*. 2003;302(5644):415-419.
17. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*. 2010;11:165.