nature genetics

A recurrent germline *PAX5* mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia

Sohela Shah[1,2,34], Kasmintan A Schrader[1,2,34], Esmé Waanders[3,4,34], Andrew E Timms[5,34], Joseph Vijai[1,2,34], Cornelius Miething[1,34], Jeremy Wechsler[5], Jun Yang[6], James Hayes[1], Robert J Klein[1,2], Jinghui Zhang[7], Lei Wei[3,7], Gang Wu[7], Michael Rusch[7], Panduka Nagahawatte[7], Jing Ma[3], Shann-Ching Chen[3], Guangchun Song[3], Jinjun Cheng[3,8], Paul Meyers[9], Deepa Bhojwani[10], Suresh Jhanwar[11], Peter Maslak[12], Martin Fleisher[13], Jason Littman[2], Lily Offit[2], Rohini Rau-Murthy[2], Megan Harlan Fleischut[2], Marina Corines[2], Rajmohan Murali[11], Xiaoni Gao[1], Christopher Manschreck[2], Thomas Kitzing[1], Vundavalli V Murty[14], Susana C Raimondi[3], Roland P Kuiper[4], Annet Simons[4], Joshua D Schiffman[15], Kenan Onel[16], Sharon E Plon[17,18], David A Wheeler[17,18], Deborah Ritter[17,18], David S Ziegler[19,20], Kathy Tucker[21], Rosemary Sutton[20], Georgia Chenevix-Trench[22], Jun Li[22], David G Huntsman[23], Samantha Hansford[23], Janine Senz[23], Tom Walsh[24,25], Ming Lee[24,25], Christopher N Hahn[26], Kathryn G Roberts[3], Mary-Claire King[24,25], Sarah M Lo[27], Ross L Levine[28], Agnes Viale[29], Nicholas D Socci[30], Katherine L Nathanson[31], Hamish S Scott[26], Mark Daly[32], Steven M Lipkin[33], Scott W Lowe[1], James R Downing[3], David Altshuler[32], John T Sandlund[10,35], Marshall S Horwitz[5,35], Charles G Mullighan[3,35] & Kenneth Offit[1,2,33,35]

Somatic alterations of the lymphoid transcription factor gene *PAX5* (also known as *BSAP*) are a hallmark of B cell precursor acute lymphoblastic leukemia (B-ALL)[1–3], but inherited mutations of *PAX5* have not previously been described. Here we report a new heterozygous germline variant, c.547G>A (p.Gly183Ser), affecting the octapeptide domain of PAX5 that was found to segregate with disease in two unrelated kindreds with autosomal dominant B-ALL. Leukemic cells from all affected individuals in both families exhibited 9p deletion, with loss of heterozygosity and retention of the mutant *PAX5* allele at 9p13. Two additional sporadic ALL cases with 9p loss harbored somatic *PAX5* substitutions affecting Gly183. Functional and gene expression analysis of the *PAX5* mutation demonstrated that it had significantly reduced transcriptional activity. These data extend the role of *PAX5* alterations in the pathogenesis of pre-B cell ALL and implicate *PAX5* in a new syndrome of susceptibility to pre-B cell neoplasia.

B cell precursor ALL is the most common pediatric malignancy. Children with affected siblings have 2- to 4-fold greater risk of developing the disease[4], and, in occasional cases, ALL is inherited as a mendelian disorder[5]. *PAX5*, encoding the B cell lineage transcription factor paired box 5, is somatically deleted, rearranged or otherwise mutated in approximately 30% of sporadic B-ALL cases[1–3,6–9]. In *Pax5*-deficient mice, B cell development is arrested at the pro-B cell stage, and these cells can differentiate *in vitro* into other lymphoid and myeloid lineages[10]. PAX5 is also essential for maintaining the identity and function of mature B cells[11], and its deletion in mature B cells results in dedifferentiation to pro-B cells and aggressive lymphomagenesis[12].

We identified a heterozygous germline *PAX5* variant, c.547G>A (NM_016734), encoding p.Gly183Ser (NP_057953), by exome sequencing in two families, one of Puerto Rican ancestry (family 1; **Fig. 1a**) and the other of African-American ancestry (family 2; **Fig. 1b** and **Supplementary Note**). This variant had not previously been described in public databases (Exome Variant Server, 1000 Genomes Project and dbSNP137) or previous sequencing analyses of ALL and cancer genomes[1,2,9]. All affected family members had B-ALL, and all available diagnostic and relapse leukemic samples from both families demonstrated loss of 9p through the formation of an isochromosome of 9q, i(9)(q10), or the presence of dicentric chromosomes involving 9q, both of which resulted in loss of the wild-type *PAX5* allele and retention of the *PAX5* allele encoding p.Gly183Ser (**Fig. 1c**, **Supplementary Fig. 1** and **Supplementary Table 1**).

The germline *PAX5* mutation encoding p.Gly183Ser segregated with leukemia in both kindreds; however, several unaffected obligate carriers (family 1: II3, III2 and III3 and family 2: I1, I2, II2 and II3) were also observed, suggesting incomplete penetrance. Unaffected mutation carriers and affected individuals at the time of diagnosis with ALL had normal immunoglobulin levels and no laboratory or clinical evidence of impaired B cell function. Sanger sequencing of cDNA from the peripheral blood of unaffected carriers indicated biallelic transcription of *PAX5* (data not shown). The only mutated gene common to both families was *PAX5*, and no germline copy

A full list of authors affiliation appears at the end of the paper.

Received 26 March; accepted 9 August; published online 8 September 2013; doi:10.1038/ng.2754

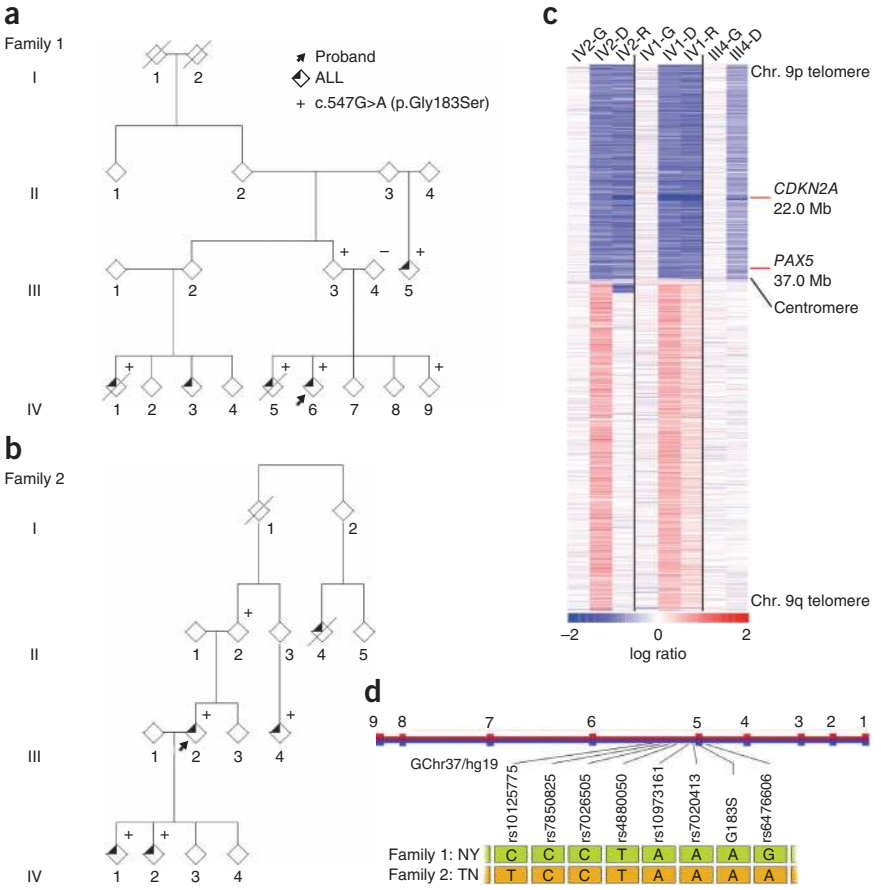

Figure 1 Familial pre-B cell ALL associated with i(9)(q10) and dic(9;v) alterations in two families harboring a new, recurrent germline variant encoding p.Gly183Ser. (**a**) Family 1 of Puerto Rican ancestry. The proband is indicated by an arrow. Exome sequencing was undertaken on germline DNA from all available affected (IV1, IV5, IV6, III5) and unaffected (IV9, III3, III4) individuals as well as on the diagnostic leukemic sample from IV6. (**b**) Family 2 of African-American ancestry. The proband is indicated by an arrow. Exome sequencing was undertaken in diagnostic, remission and relapse leukemic samples from individuals III4, IV1 and IV2. (**c**) Chromosome 9 copy number heatmap for SNP6.0 microarray data of germline and tumor samples from three members of family 2. These data demonstrate the common feature of loss of 9p in the tumor specimens. Note the focal dark-blue band denoting homozygous loss of *CDKN2A* and/or *CDKN2B* in all samples. Blue indicates deletions, and red indicates gains. G, germline; D, diagnostic; R, relapse. (**d**) Haplotype flanking the mutation encoding p.Gly183Ser. A five-SNP haplotype from rs7850825 to rs7020413 (chr. 9: 36.997–37.002 Mb) proximal to the mutation was concordant in both family 1 and family 2. However, the distal end flanking the mutation rs6476606 was discordant.

number aberrations were found to be shared by affected individuals (**Supplementary Tables 2** and **3**).

To determine whether the mutation encoding p.Gly183Ser arose independently in each kindred or instead reflects common ancestry, we compared the risk haplotypes of the families. The families shared a 4.7-kb haplotype spanning five SNPs (**Fig. 1d** and **Supplementary Note**). The relatively small size of this shared haplotype and principal-component analysis of genome-wide SNP genotype data (**Supplementary Fig. 2**) together implied that the two families were not recently related and differed in ancestry. Moreover, given the reduced fitness due to increased susceptibility to childhood ALL, it is unlikely that such a lethal mutation could be propagated over time. Because the identified haplotype is relatively frequent worldwide (**Supplementary Table 4**), it is likely that each family's mutation arose independently.

Genomic profiling of tumor samples demonstrated expression of the mutant *PAX5* allele encoding p.Gly183Ser in diagnostic and relapse tumor specimens from affected members of family 2, with an average of 1 chimeric fusion and 9 non-silent sequence variants per case and homozygous deletion of *CDKN2A* with or without *CDKN2B* in all cases due to loss of 9p and focal deletion of the second allele. Apart from loss of 9p, no other somatic sequence mutations or structural rearrangements were shared by the affected families (**Supplementary Tables 1** and **5–12**).

As somatic i(9)(q10) or dic(9;v) abnormalities were seen in all of the familial leukemias, we sequenced *PAX5* in 44 additional sporadic pre-B-ALL cases with i(9)(q10) or dic(9;v) aberrations to assess whether *PAX5* mutations frequently co-occur with loss of 9p. Two leukemic samples had mutations encoding p.Gly183Ser and p.Gly183Val substitutions in the octapeptide domain, and, in others, previously reported variants including p.Pro80Arg and p.Val26Gly[1] were observed (**Fig. 2** and **Table 1**). We examined the frequency of non-silent *PAX5* somatic sequence mutations in a cohort of B-ALL cases with 9p loss through i(9)(q10) or dic(9;v) aberrations (n = 28) and in 2 cohorts of B-ALL without i(9)(q10) or dic(9;v) aberrations (n = 183 and 221; refs. 1,2). We observed a significantly higher frequency of *PAX5* mutations in the cohort with isochromosomal or dicentric aberrations of chromosome 9 (P = 0.0001). No germline *PAX5* mutations were detected in 39 families with a history of 2 or more cases of cancer, including at least 1 childhood hematological cancer, although 1 familial case of ALL harbored a dic(9;20)(p11;q11.1) alteration and a somatic variant encoding p.Pro80Arg (**Table 1** and **Supplementary Note**).

Previously identified *PAX5* somatic mutations commonly result in marked reduction in the transcriptional activation mediated by PAX5. Downstream targets of PAX5 include *CD19* and *CD79A* (also known as *IGA* and *MB-1*)[13]. We examined the transactivating activity of the proteins encoded by the wild-type and mutant *PAX5* alleles using a PAX5-dependent reporter gene assay containing copies of a high-affinity PAX5-binding site derived from the *CD19* promoter[14]. Both the p.Gly183Ser and p.Gly183Val alterations resulted in partial but significant reduction in transcriptional activation compared to wild-type PAX5 (P < 0.0001 for both alterations; **Fig. 3a**). Additionally, there was no detectable difference in the subcellular localization of wild-type and p.Gly183Ser PAX5 (**Supplementary Fig. 3**). To study the effect of the p.Gly183Ser alteration on CD79A expression, we expressed mutant and wild-type PAX5 in J558 and J558LμM, mouse plasmacytoma cell lines that do not express PAX5 or CD79A. Enforced expression of PAX5 results in expression of CD79A and assembly of the surface immunoglobulin M (sIgM) complex. The amount of sIgM expression may be used to assess the transcriptional activity of *PAX5* alleles on the *CD79A* promoter[13]. Both alleles encoding alterations to Gly183

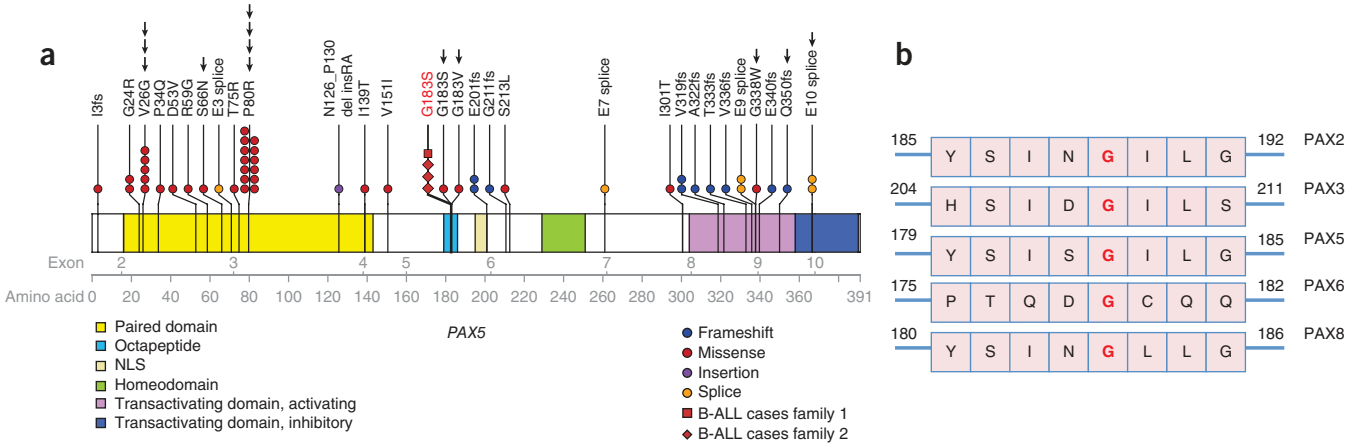

Figure 2 Recurrent *PAX5* mutations in ALL. (**a**) Gene schematic of *PAX5* (NM_016734) showing the exons, amino acid residues, protein domains and position of the germline variant encoding p.Gly183Ser (red) in relation to the somatic *PAX5* mutations described in this study ($n = 13$, arrows) and somatic mutations described previously in B-ALL[1,2,20]. Primary leukemic samples with confirmed retention of the germline variant encoding p.Gly183Ser are denoted by squares (family 1) and diamonds (family 2). In one case of ALL with a dicentric aberration of chromosome 9, we found both a heterozygous mutation encoding p.Val26Gly and a heterozygous mutation encoding p.Gln350fs, indicating polyclonality of the tumor. (**b**) Conservation of the octapeptide domain in selected PAX family members.

resulted in a significant reduction in sIgM expression compared to the wild-type *PAX5* allele ($P < 0.0001$; **Fig. 3b**). These results suggest that *PAX5* mutations affecting Gly183 result in partial loss of PAX5 activity.

The identified missense variant p.Gly183Ser is located at a conserved residue in the octapeptide domain of PAX5 that mediates interaction with Groucho transcriptional corepressors[15] (**Fig. 2b**). Previous studies have shown that GRG4 (also known as TLE4) represses PAX5-dependent luciferase activity in cells expressing wild-type PAX5 but not in cells expressing PAX5 octapeptide-domain mutants[15]. We observed GRG4-mediated repression of the transcriptional activity of wild-type and p.Gly183Ser PAX5 (**Fig. 3c**), suggesting that the effect of the alteration is not mediated by altered interaction with GRG4.

To further explore the effect of the p.Gly183Ser variant on downstream targets, we performed genome-wide transcriptional profiling of J558LμM cells transduced with empty vector or with vector expressing wild-type or mutant *PAX5* alleles (examining either all transduced cells marked by red fluorescent protein (RFP) expression or the subset of cells expressing sIgM) and analyzed the expression of genes activated and repressed by PAX5 as previously defined in *Pax5*$^{-/-}$ mouse pro-B cells and mature B cells[16–19] and in human *ETV6-RUNX1*–positive B-ALL[1]. Examining all PAX5-expressing cells, we observed profound deregulation of genes activated and repressed by PAX5 in J558LμM cells expressing known loss-of-function alleles (for example, the common exon 2–6 deletion that results in a truncating frameshift *PAX5* allele, Δ2–6) or strongly hypomorphic alleles (for example, the *PAX5* allele encoding p.Pro80Arg) and less marked deregulation in cells expressing p.Gly183Ser or p.Gly183Val mutant PAX5 (*P* values for each mutant protein versus wild type were all $P < 0.001$; **Supplementary Figs. 4** and **5**). Comparing sorted sIgM-positive cells expressing p.Gly183Ser PAX5 to those expressing wild-type protein, we observed reduced expression of genes activated by PAX5 in pro-B cells and mature B cells ($P = 1.4 \times 10^{-4}$ and 3.8×10^{-4}, respectively; **Supplementary Tables 13–15**).

We next examined the transcriptional consequences of the *PAX5* mutation encoding p.Gly183Ser by performing transcriptome sequencing (mRNA-seq) of diagnostic and relapse samples obtained from 2 affected individuals in kindred 2 and from 139 sporadic childhood B-ALL samples. We performed gene-set enrichment analysis incorporating gene sets of *PAX5*-mutated, *ETV6-RUNX1*–positive

Table 1 ***PAX5*** **mutations found in familial and sporadic B-ALL samples with i(9)(q10) or dic(9;v) aberrations**

Inheritance	Subject	Mutation	Protein alteration	Tumor status	Germline status
Family 1	IV6	c.547G>A	p.Gly183Ser	Homozygous	Heterozygous
Family 2	III4	c.547G>A	p.Gly183Ser	Homozygous	Heterozygous
Family 2	IV1 D	c.547G>A	p.Gly183Ser	Homozygous	Heterozygous
Family 2	IV1 R	c.547G>A	p.Gly183Ser	Homozygous	Heterozygous
Family 2	IV2 D	c.547G>A	p.Gly183Ser	Homozygous	Heterozygous
Family 2	IV2 R	c.547G>A	p.Gly183Ser	Homozygous	Heterozygous
Familial[a]		c.239C>G (tumor shows dic(9;20)(p11;q11.1))	p.Pro80Arg	Homozygous	Wild type
Sporadic		c.77T>G	p.Val26Gly	Heterozygous	Wild type
Sporadic		c.77T>G	p.Val26Gly	Heterozygous	Wild type
Sporadic		c.77T>G	p.Val26Gly	Heterozygous	Wild type
Sporadic		c.197G>A	p.Ser66Asn	Homozygous	ND
Sporadic		c.239C>G	p.Pro80Arg	Homozygous	Wild type
Sporadic		c.239C>G	p.Pro80Arg	Homozygous	Wild type
Sporadic		c.239C>G	p.Pro80Arg	Homozygous	Wild type
Sporadic		c.547G>A	p.Gly183Ser	Homozygous	ND
Sporadic		c.548G>T	p.Gly183Val	Heterozygous	Wild type
Sporadic		c.1012G>T	p.Gly338Trp	Heterozygous	Wild type
Sporadic		c.1049_1051delAGTinsGTCCG =	p.Gln350fs	Heterozygous	Wild type
Sporadic		c.1100_1100+15 del16 (IVS9 splice)		Heterozygous	ND

ND, not determinable; germline DNA either not tested or not available. *PAX5* mRNA sequence is available under accession NM_016734.
[a]This family included a case of pediatric ALL (analyzed here) and a case of breast cancer.

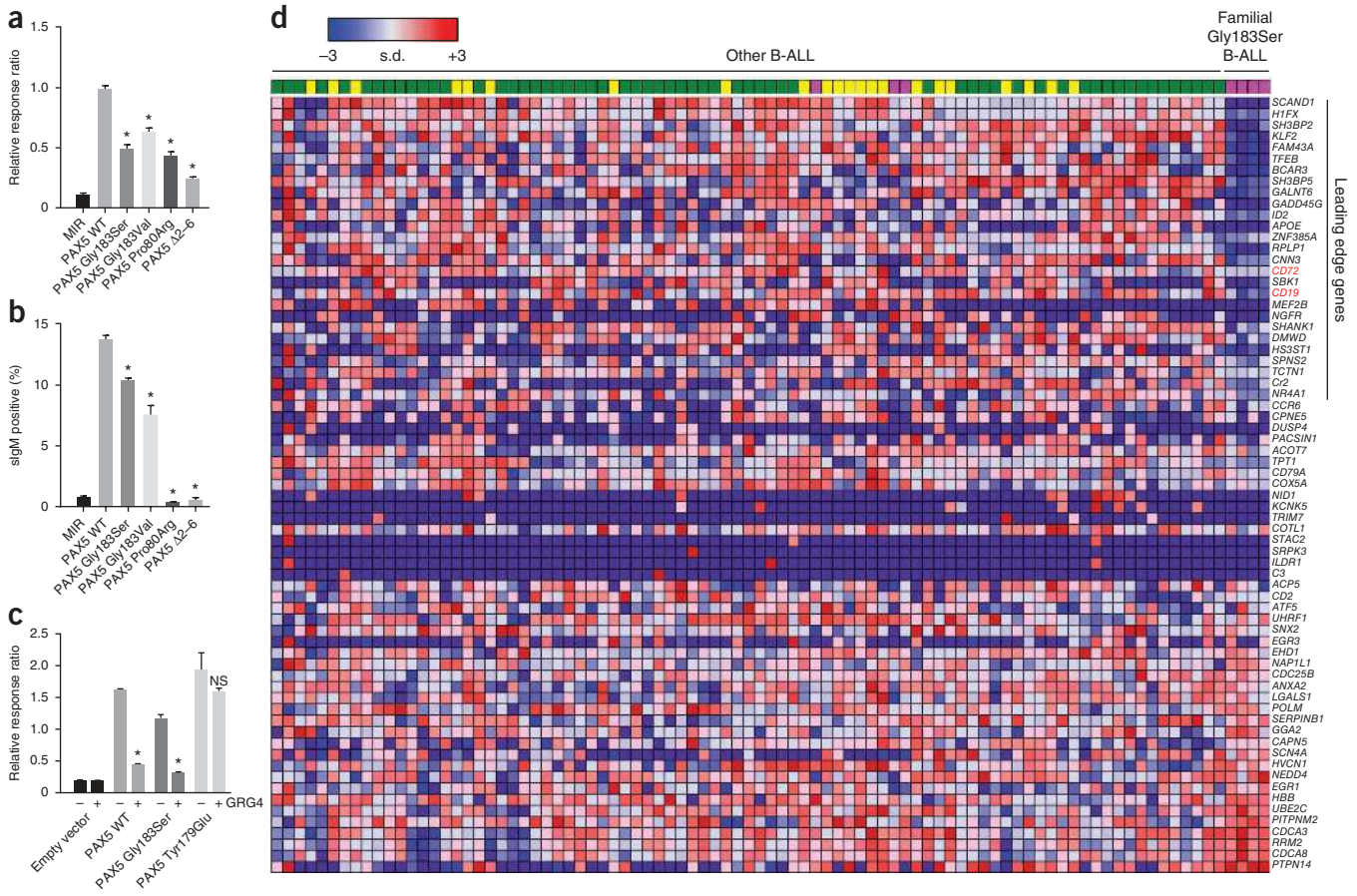

Figure 3 Attenuated transcriptional activity of p.Gly183Ser PAX5. (**a**) Transcriptional activity of PAX5 variants compared to wild-type protein determined using a PAX5-dependent reporter gene assay in 293T cells. Bars show mean (± s.e.m.) luciferase activity from six individual experiments with triplicate measurements (for PAX5 p.Gly183Val and PAX5 Δ2–6, four experiments with triplicate measurements). Asterisks indicate significant differences calculated by Dunnett's test ($P < 0.0001$). MIR, MSCV-IRES-mRFP empty vector; WT, wild type. (**b**) Transcriptional activity of PAX5 variants determined using CD79A-dependent sIgM expression in the mouse J558LμM plasmacytoma cell line. Percentages indicate the proportion of mRFP-positive cells that show sIgM expression. Bars show mean (± s.e.m.) sIgM expression in two individual experiments with three replicates each. Asterisks indicate significant differences calculated by Dunnett's test ($P < 0.0001$). (**c**) PAX5-dependent reporter gene assay of wild-type and p.Gly183Ser PAX5 run in triplicate with or without cotransfection with 0.05 μg of vector encoding GRG4 as indicated. A p.Tyr179Glu PAX5 mutant that is deficient in binding to GRG4 and empty vector were used as controls. Asterisks indicate significant differences as determined by two-tailed *t* test ($P < 0.0001$). NS, not significant. (**d**) GSEA examining enrichment of genes known to be activated or repressed by PAX5 in experimental systems in the transcriptional profile of familial ALL. A representative heatmap is presented of genes shown to be activated by PAX5 in mouse B cells[17], which were negatively enriched in the transcriptional signature of familial ALL compared to B-ALL cases (excluding *ETV6-RUNX1* ALL; $P < 0.01$, FDR = 0.09; see also **Supplementary Tables 19–21**). Leading-edge genes in this gene set responsible for enrichment are *SCAND1* to *NR4A1*. Four samples from family 2 (diagnostic and relapse samples from individuals IV1 and IV2) show differential expression of PAX5-activated genes compared to a group of 139 sporadic B-ALL cases. This indicates an effect of the mutation encoding p.Gly183Ser on PAX5 function. Red indicates high expression, blue represents low expression. *PAX5* mutational status is indicated by the top row of colored boxes: green, wild-type *PAX5*; yellow, heterozygosity for a *PAX5* mutation; magenta, biallelic *PAX5* mutation.

ALL cases (one-third of which harbor focal *PAX5* deletions)[1], PAX5-regulated genes in *Pax5*$^{-/-}$ mice[16–19] and genes regulated during mouse B-lymphoid development[20]. As a limited set of genes is known to be regulated in both mouse pro-B cells and mature B cells and as the overlap between mouse and human PAX5-regulated genes is unknown, we used all previously published PAX5-regulated genes and genes regulated during mouse B cell development[16–20] in an unbiased approach to explore the effects of the *PAX5* mutations affecting Gly183 on direct and indirect transcriptional targets of PAX5. This analysis showed striking enrichment of genes deregulated in *PAX5*-mutated, *ETV6-RUNX1*–positive ALL, genes activated and repressed by PAX5 (including *CD19*, *CD72* and *CD79A*), and genes regulated during mouse B-lymphoid development in the signature of familial B-ALL with the *PAX5* mutation encoding p.Gly183Ser versus sporadic B-ALL (**Fig. 3d** and **Supplementary Figs. 6** and **7**). We also analyzed the overlap of previously published data and the expression differences between the familial ALL tumor samples and other B-ALL cases stratified by *PAX5* mutation status (**Supplementary Fig. 8** and **Supplementary Table 16**). Together, our results suggest that the *PAX5* mutation encoding p.Gly183Ser results in attenuation of PAX5 function and deregulation of PAX5 target genes that is less severe than for the previously reported p.Pro80Arg and Δ2–6 alterations that result in marked or complete loss of PAX5 activity.

The *PAX5* deletions, translocations and sequence mutations identified as somatic events in B-ALL commonly affect the DNA-binding and transactivation domains and result in complete loss or marked attenuation of PAX5 transcriptional activity but are rarely homozygous and are not observed as inherited variants. Moreover, PAX5 loss

promotes the development of B-ALL in experimental models that are commonly affected by the acquisition of accompanying second hits in *PAX5* (ref. 21), indicating that profound loss of PAX5 activity is commonly a central event in leukemogenesis. In contrast, the inherited *PAX5* mutation encoding p.Gly183Ser results in modest attenuation of PAX5 activity in transcriptional reporter assays and is accompanied by somatic loss of the wild-type *PAX5* allele due to 9p alterations during leukemogenesis. This model is also consistent with the finding of a significant association of somatic *PAX5* hypomorphic mutations coincident with complete loss of the normal *PAX5* allele in leukemic cells with absent 9p. These observations suggest that a severe reduction in PAX5 activity is incompatible with normal B-lymphoid development and is deleterious in carriers; by contrast, the partial hypomorphic allele encoding p.Gly183Ser is tolerated as a germline allele, but additional genetic events further reducing PAX5 activity are required to establish the leukemic clone. The universal finding of deletion of wild-type *PAX5* in all familial ALL cases, rather than the acquisition of additional hypomorphic *PAX5* mutations, suggests that a complete loss of wild-type PAX5 activity is required for developmental arrest and loss of maturation. This notion is supported by our transcriptional profiling of J558LμM cells expressing p.Gly183Ser PAX5 and by familial leukemias showing deregulation of PAX5 target gene expression that is significant but less marked than that observed with known loss-of-function mutations. The differences in the transcriptional profiles of some target gene panels were not as robust as in mouse model systems, presumably owing to inherent germline and somatic genetic and epigenetic variability in human leukemias. In addition, ongoing studies will be of interest to fully characterize the functional consequences of PAX5 octapeptide-domain mutations.

Our findings have clinical implications with regard to options for pre-implantation genetic diagnosis and the possible relevance of somatic 9p alterations as a harbinger of a germline *PAX5* mutation. The recent identification of germline *TP53* mutations in familial ALL[20,22] and the data presented here strongly implicating *PAX5* mutations in a new syndrome of inherited susceptibility to pre-B cell ALL indicate that further sequencing of affected kindreds is required to define the full spectrum of germline variations contributing to ALL pathogenesis.

URLs. dbSNP137, http://www.ncbi.nlm.nih.gov/projects/SNP/; National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project, http://evs.gs.washington.edu/EVS/; European Genome-phenome Archive (EGA), http://www.ebi.ac.uk/ega/; public data portal for results from the St. Jude–Washington University Pediatric Cancer Genome Project, http://explore.pediatriccancergenomeproject.org/.

METHODS

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

Accession codes. Transcriptome and whole-exome sequencing data and SNP microarray data have been deposited in the European Genome-phenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI), under accession EGAS00001000447. Mouse Affymetrix gene expression data are deposited in the Gene Expression Omnibus (GEO) under accession GSE45260.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M.A.S. Moore and S. Jae-Hung for their contributions to ongoing tumor studies (Sloan-Kettering Institute); G. Dressler (University of Michigan) for the mouse GRG4 construct; M. Busslinger (The Research Institute for Molecular Pathology) for the *luc-CD19* reporter construct; J. Hagman (National Jewish) for providing a *PAX5* vector and the J558LμM cell line; D. Payne-Turner (St. Jude Children's Research Hospital) for technical assistance; W. Yang and C. Smith (Pharmaceutical Sciences, St. Jude Children's Research Hospital) for their assistance in the haplotype analyses; and the Tissue Resources Core Facility, Pediatric Cancer Genome Project Core Facility and Flow Cytometry and Cell Sorting Core Facility of St. Jude Children's Research Hospital. We thank the families for their generous participation in these studies. This project was supported by grant I5-A523 from the Starr Cancer Consortium, the Robert and Kate Niehaus Clinical Cancer Genetics Initiative, the Sabin Family Research Fund, the Lymphoma Foundation, Geoffrey Beene Cancer Research Center grant 78730, the Sharon Levine Corzine Foundation, the Barbara L. Goldsmith Genetics Research Fund, Cancer Prevention and Research Institute of Texas grant RP101089, the New South Wales Priory of the Knights of the Order of Saint John, the Matthew Bell Foundation, National Cancer Institute of the US National Institutes of Health (NIH) Comprehensive Cancer Center Core grant CA21765, the American Lebanese Syrian Associated Charities of St. Jude Children's Research Hospital and grant R01DK58161 from the US NIH. R.P.K. is funded by a grant from the Dutch Cancer Society (KUN2009-4298). T.K. is supported by a German Research Foundation Postdoctoral Fellowship (KI1605/1-1). C.G.M. is a Pew Scholar in the Biomedical Sciences and a St. Baldrick's Scholar. K.G.R. is supported by a National Health and Medical Research Council (NHMRC, Australia) CJ Martin Postdoctoral Fellowship. K.A.S. is funded by the Canadian Institutes of Health Research. A.E.T. is supported by T32GM007454 from the National Institute of General Medical Sciences (NIGMS). G.C.-T. is a Senior Principal Research Fellow of the NHMRC. E.W. is funded by the Dutch Cancer Society, project number KUN2012-5366. H.S.S. is a Principal Research Fellow of the NHMRC (APP1023059), and the work was supported by grant APP1024215.

AUTHOR CONTRIBUTIONS

K. Offit, C.G.M., M.S.H., J.T.S., S.S., K.A.S., E.W., A.E.T., J.V., C. Miething, S.M. Lipkin, R.J.K., M.D., D.A. and S.W.L. conceived and designed the experiment. S.S., K.A.S., E.W., A.E.T., J.V., C. Miething, J.W., J.Y., X.G., C. Manschreck, R.J.K., A.V., N.D.S., D.A., M.S.H., C.G.M., K. Offit, M.-C.K., T.W., M.L., T.K., D.B., J. Littman, L.O., S.C.R., P. Maslak, M.F., K.G.R. and J.C. performed the experiments. S.S., K.A.S., E.W., A.E.T., J.V., C. Miething, J.W., J.Y., R.J.K., N.D.S., M.S.H., C.G.M., K. Offit, L.W., J.Z., G.W., M.R., P.N., J.M., S.-C.C., G.S. and J.C. performed statistical analysis. S.S., K.A.S., E.W., A.E.T., J.V., C. Miething, J.W., J.Y., C. Manschreck, R.R.-M., M.C., R.M., M.H.F., S.M. Lipkin, R.J.K., A.V., N.D.S., D.A., C.N.H., H.S.S., S.W.L., M.S.H., C.G.M., K. Offit, L.W., J.H., J.Z., G.W., M.R., P.N., J.M., S.-C.C., G.S. and J.C. analyzed the data. E.W., C. Miething, J.T.S., S.J., M.H.F., J.S., V.V.M., S.E.P., D.G.H., D.S.Z., G.C.-T., S.M. Lipkin, S.M. Lo, R.L.L., A.V., K.L.N., M.D., D.A., C.N.H., H.S.S., S.W.L., M.S.H., C.G.M., K. Onel, R.P.K., A.S., J. Li, K.T., R.S., S.H., J.D.S., D.A.W., D.R., P. Meyers, J.Z., G.W., J.M., S.-C.C., J.R.D. and K. Offit contributed reagents, materials and analysis tools. K. Offit, C.G.M., M.S.H., S.S., K.A.S., E.W., A.E.T., J.V., C. Miething, J.Y., R.J.K. and S.W.L. wrote the manuscript. K. Offit, C.G.M., M.S.H., J.T.S., R.J.K., M.D., D.A. and S.W.L. jointly supervised the research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mullighan, C.G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
2. Mullighan, C.G. *et al.* Deletion of *IKZF1* and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360**, 470–480 (2009).
3. Kuiper, R.P. *et al.* High-resolution genomic profiling of childhood ALL reveals novel recurrent genetic lesions affecting pathways involved in lymphocyte differentiation and cell cycle progression. *Leukemia* **21**, 1258–1266 (2007).
4. Hemminki, K. & Jiang, Y. Risks among siblings and twins for childhood acute lymphoid leukaemia: results from the Swedish Family-Cancer Database. *Leukemia* **16**, 297–298 (2002).
5. Pui, C.H., Robison, L.L. & Look, A.T. Acute lymphoblastic leukaemia. *Lancet* **371**, 1030–1043 (2008).
6. Mullighan, C.G. & Downing, J.R. Global genomic characterization of acute lymphoblastic leukemia. *Semin. Hematol.* **46**, 3–15 (2009).
7. Mullighan, C.G. *et al. BCR-ABL1* lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453**, 110–114 (2008).
8. Nebral, K. *et al.* Incidence and diversity of *PAX5* fusion genes in childhood acute lymphoblastic leukemia. *Leukemia* **23**, 134–143 (2009).

9. Zhang, J. *et al.* Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* **118**, 3080–3087 (2011).
10. Nutt, S.L., Heavey, B., Rolink, A.G. & Busslinger, M. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature* **401**, 556–562 (1999).
11. Horcher, M., Souabni, A. & Busslinger, M. Pax5/BSAP maintains the identity of B cells in late B lymphopoiesis. *Immunity* **14**, 779–790 (2001).
12. Cobaleda, C., Jochum, W. & Busslinger, M. Conversion of mature B cells into T cells by dedifferentiation to uncommitted progenitors. *Nature* **449**, 473–477 (2007).
13. Maier, H., Colbert, J., Fitzsimmons, D., Clark, D.R. & Hagman, J. Activation of the early B-cell-specific *mb-1* (Ig-α) gene by Pax-5 is dependent on an unmethylated Ets binding site. *Mol. Cell. Biol.* **23**, 1946–1960 (2003).
14. Czerny, T. & Busslinger, M. DNA-binding and transactivation properties of Pax-6: three amino acids in the paired domain are responsible for the different sequence recognition of Pax-6 and BSAP (Pax-5). *Mol. Cell. Biol.* **15**, 2858–2871 (1995).
15. Eberhard, D., Jimenez, G., Heavey, B. & Busslinger, M. Transcriptional repression by Pax5 (BSAP) through interaction with corepressors of the Groucho family. *EMBO J.* **19**, 2292–2303 (2000).
16. Pridans, C. *et al.* Identification of Pax5 target genes in early B cell differentiation. *J. Immunol.* **180**, 1719–1728 (2008).
17. Revilla-I-Domingo, R. *et al.* The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. *EMBO J.* **31**, 3130–3146 (2012).
18. Delogu, A. *et al.* Gene repression by Pax5 in B cells is essential for blood cell homeostasis and is reversed in plasma cells. *Immunity* **24**, 269–281 (2006).
19. Schebesta, A. *et al.* Transcription factor Pax5 activates the chromatin of key genes involved in B cell signaling, adhesion, migration, and immune function. *Immunity* **27**, 49–63 (2007).
20. Holmfeldt, L. *et al.* The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat. Genet.* **45**, 242–252 (2013).
21. Dang, J., Mullighan, C.G., Phillips, L.A., Mehta, P. & Downing, J.R. Retroviral and chemical mutagenesis identifies Pax5 as a tumor suppressor in B-progenitor acute lymphoblastic leukemia. *Blood (ASH Annual Meeting Abstracts)* **112**, 1789 (2008).
22. Powell, B.C. *et al.* Identification of *TP53* as an acute lymphocytic leukemia susceptibility gene through exome sequencing. *Pediatr. Blood Cancer* **60**, E1–E3 (2013).

[1]Cancer Biology and Genetics Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [2]Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [3]Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [4]Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences and Radboud Institute for Oncology, Radboud University Medical Centre, Nijmegen, The Netherlands. [5]Department of Pathology, University of Washington, Seattle, Washington, USA. [6]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [7]Department of Computational Biology and Bioinformatics, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [8]Pediatric Cancer Genome Project Laboratory, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [9]Department of Pediatrics, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [10]Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA. [11]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [12]Hematology Laboratory Service, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [13]Clinical Chemistry Service, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [14]Department of Pathology and Cell Biology, Columbia University, New York, New York, USA. [15]High-Risk Pediatric Cancer Clinic, Huntsman Cancer Institute/Primary Children's Medical Center, University of Utah, Salt Lake City, Utah, USA. [16]Department of Pediatrics, University of Chicago, Chicago, Illinois, USA. [17]Texas Children's Cancer Center, Baylor College of Medicine, Houston, Texas, USA. [18]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [19]Kids Cancer Centre, Sydney Children's Hospital, Sydney, New South Wales, Australia. [20]Children's Cancer Institute Australia for Medical Research, University of New South Wales, Randwick, New South Wales, Australia. [21]Hereditary Cancer Clinic, Prince of Wales Hospital, Randwick, New South Wales, Australia. [22]Cancer Genetics Laboratory, The Queensland Institute of Medical Research, Herston, Queensland, Australia. [23]Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [24]Department of Medicine, University of Washington, Seattle, Washington, USA. [25]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [26]Department of Molecular Pathology, SA Pathology and Centre for Cancer Biology, Adelaide, South Australia, Australia. [27]Department of Pediatrics, Weill Cornell College of Medicine, New York, New York, USA. [28]Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [29]Genomics Core Laboratory, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [30]Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [31]Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [32]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [33]Department of Medicine, Weill Cornell College of Medicine, New York, New York, USA. [34]These authors contributed equally to this work. [35]These authors jointly directed this work. Correspondence should be addressed to K. Offit (offitk@mskcc.org), C.G.M. (charles.mullighan@stjude.org) or M.S.H. (horwitz@uw.edu).

ONLINE METHODS

Subjects and samples. Family 1 was ascertained from the Memorial Sloan-Kettering Cancer Center Clinical Genetics Service. Study subjects provided written informed consent as part of a study to define genomic causes of lymphoid malignancies, and the study was approved by the local research ethics board. Family 2 from St. Jude Children's Research Hospital was ascertained in accord with local institutional review board approval. To protect subject identity, pedigrees were anonymized by alterations that do not affect genetic analysis.

Exome sequencing. Germline DNA (1 µg) from the peripheral leukocytes of affected individuals in remission and unaffected family members was used for whole-exome capture using an Agilent SureSelect 45Mb or 50Mb kit and paired-end sequencing with the Illumina HiSeq 2000 (ref. 23). Family 1 exome data were analyzed using Burrows-Wheeler Aligner (BWA)[24] to align fastq files and generate BAM files, and the Genome Analysis Tool Kit (GATK)[23,25] was used for variant calling. SNP clustering and proximity to indels and the proportion of aligned reads at a site with mapping quality of zero were used for filtering variants. Variant quality score–recalibrated (VQSR) data were then processed using the SNPEff program for functional annotation. Samples from family 2 underwent variant analysis as previously described[20]. Downstream analysis consisted of filtering out low-quality variant calls and those already reported in public databases. The downstream processing of sequence data, variant annotation and the filtering strategy based on a presumed autosomal dominant mode of inheritance with incomplete penetrance are detailed in the **Supplementary Note**.

Principal-component analysis. From the exome-sequenced samples, single-nucleotide variants seen at a frequency above 5% in the dbSNP database were selected for principal-component analysis. These data were then combined together with 1000 Genomes Project SNP data. SNPs were pruned on the basis of pairwise linkage disequilibrium within a 50-kb window. Data were transformed to calculate eigenvectors and eigenvalues for each sample, and the first two principal components were plotted.

SNP array genotyping. SNP array genotyping was performed using Affymetrix SNP 6.0 microarrays on the diagnostic leukemic sample from individual IV6 from family 1 and on germline DNA from unaffected individuals III3, III4 and IV9 and analyzed using the Genotyping Console (Affymetrix). SNP 6.0 arrays were also performed for diagnostic leukemic and remission samples from individuals IV1, IV2 and III4 from family 2, as well as on relapse samples from IV1 and IV2, and data were analyzed by optimal reference normalization[26] and circular binary segmentation[27,28] as previously described[29] using R and dChip[30]. Haplotype analysis was conducted using germline samples from III3, III4 and IV9 and the diagnostic leukemic sample from IV6 from family 1 and the diagnostic and remission samples from IV1, IV2 and III4 from family 2.

In view of the cytogenetic abnormalities in each of the leukemic samples resulting in monosomy 9p, for which Sanger sequencing of the variant encoding p.Gly183Ser demonstrated loss of heterozygosity with retention of the mutant allele, we were able to biologically phase the SNP risk haplotype containing the mutant allele. Beagle phased haplotypes from the 1000 Genomes Project were analyzed for the five-SNP shared haplotype, and frequencies were estimated among the populations in HapMap.

***PAX5* sequencing.** Sanger sequencing (primer sequences available upon request) of the entire ORF of *PAX5* was performed in 44 cases of sporadic ALL characterized by i(9) or dic(9;v) and 31 cases of familial cancer. We also reviewed the coding regions of *PAX5* in an additional 8 families that had been exome sequenced or B-ALL cases that had been Sanger sequenced ($n = 87$ treatment-resistant adult-onset ALLs) as part of other studies. Cases were acquired from St. Jude Children's Research Hospital (Memphis, Tennessee; $n = 34$ i(9) or dic(9;v) and 28 familial cases), Memorial Sloan-Kettering Cancer Center/Columbia University (New York, New York; $n = 2$ i(9) or dic(9;v) and 87 treatment-resistant adult-onset ALLs), Radboud University Nijmegen Medical Centre (Nijmegen, The Netherlands; $n = 6$ i(9) or dic(9;v)), Texas Children's Cancer Center and Human Genome Sequencing Center (Houston, Texas; 7 familial cases), Children's Cancer Institute Australia for Medical Research (Sydney, Australia; $n = 2$ i(9) or dic(9;v) and 3 familial cases) and the Huntsman Cancer Institute/Primary Children's Medical Center (Salt Lake City, Utah; 1 familial case).

DNA constructs. The *CD19* luciferase construct used for PAX5-dependent reporter gene assays contains copies of a high-affinity PAX5-binding site (derived from the *CD19* promoter)[14] and was a kind gift from M. Busslinger. The pFLAG-CMV2-*Grg4* construct was a kind gift from G. Dressler[31]. The mutations encoding p.Gly183Ser and p.Gly183Val were introduced into the pSG5_*PAX5*-WT, MSCV-IRES-mRFP–*PAX5*-WT and pMSCV-Puro-IRES-GFP–*PAX5*-WT vectors by site-directed mutagenesis (QuikChange, Agilent Technologies). For retroviral expression, wild-type *PAX5* and other mutant cDNAs were subcloned as an XhoI-EcoRI fragment into MSCV-Puro-IRES-GFP (MSCV-PIG) or MSCV-IRES-RFP vector.

Cells and antibodies. HEK293 (ATCC CRL-1573) and HEK293T (ATCC CRL-11268) cells were maintained in Iscoves Modified Dulbecco's medium supplemented with 10% FCS and streptomycin. Parental J558 cells (ATCC TIB-6) were grown in DMEM with 10% horse serum[32]. J558LµM cells have been generated from a subline (J558L) that had lost immunoglobulin heavy chain expression by infection with virus encoding a cDNA of the membrane-bound heavy-chain isoform[33] and were grown in RPMI 1640 medium (Invitrogen) supplemented with 10% FBS (Hyclone), 2 mM L-glutamine (Invitrogen), 50 mg/ml gentamicin (Invitrogen), 0.3 µg/ml xanthine (Sigma) and 1 µg/ml mycophenolic acid (Sigma) as previously described[1,34]. Both lines (parental J558 and J558LµM) do not normally express sIgM because they lack expression of CD79A[35], but partial expression of CD79A can be induced by exogenous expression of *PAX5*, leading to the upregulation of sIgM[13]. Retroviral supernatants were produced by transient transfection of Phoenix Eco cells with MSCV-PIG-*PAX5* constructs and were used to infect J558 cells by spinoculation in the presence of 4 µg/ml polybrene. Rabbit monoclonal antibody to PAX5 (ab109443) and mouse monoclonal antibody to Flag (ab18230) were purchased from Abcam and were used at a 1:250 and 1:500 dilution, respectively. Mouse monoclonal antibodies to β-actin (sc-1615) were purchased from Santa Cruz Biotechnology and to SF2 were purchased from Zymed (32-4500) and were used at a 1:1,000 dilution. Antibodies to IgM conjugated to R-phycoerythrin (PE) (553409) or allophycocyanin (APC) (550676) were obtained from BD Pharmingen (BD Biosciences).

Subcellular fractionation. Protein expression and subcellular localization of the wild-type and p.Gly183Ser PAX5 proteins were examined using lysates from transiently transfected HEK293 cells separated by sucrose density gradient. The protocol for the separation of nuclei by sucrose gradient was adapted from the one for the Nuclei PurePrep Isolation kit (Sigma). CF buffer (10 mM Tris-HCl, 1 mM $MgCl_2$, 1 mM DTT, 10 µM PMSF) and 1.8 M Sucrose Solution (Sigma) were used to create density layers for resolved separation by ultracentrifugation. Fractions were then subjected to SDS-PAGE and immunoblotting with various antibodies to confirm adequate separation of nuclear and cytosolic fractions and to determine localization of recombinant PAX5.

Luciferase assays. We transfected 293T cells with MIR/MSCV-PIG^{WT} or MIR/MSCV-PIG^{mutant} along with *luc-CD19* and pRL-TK *Renilla* luciferase plasmid DNA (Promega) using FuGene 6 (Roche Diagnostics). For GRG4 repression assays, 500 ng of either the MSCV-PIG empty vector or of MSCV-PIG–*PAX5*-WT, MSCV-PIG–*PAX5*-Gly183Ser or MSCV-PIG–*PAX5*-Tyr179Glu, 2 µg of *luc-CD19* construct and 0.1 µg of pRL-TK *Renilla* luciferase plasmid were cotransfected with or without 50 ng of cDNA for GRG4 in pFLAG-CMV2 into HEK293T cells using X-tremeGENE HP DNA Transfection Reagent (Roche Diagnostics). Forty-eight hours after transfection, cell lysis and measurement of firefly and *Renilla* luciferase activity was performed using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions. All transfections were performed in triplicate in at least two independent experiments. Firefly luciferase activity was normalized according to corresponding *Renilla* luciferase activity and reported as mean relative luciferase units (RLU) ± s.e.m.

Flow cytometry analysis. J558LµM cells transduced with MIR-*PAX5* vectors or cells selected with puromycin after transduction with pMSCV-PIG vectors

doi:10.1038/ng.2754

were analyzed for RFP (MIR) or GFP (pMSCV-PIG) and sIgM expression after staining with PE- or APC-conjugated antibodies to IgM (1:20, BD Pharmingen) using LSRII or Fortessa flow cytometers (Becton Dickinson).

Gene expression profiling. RFP-positive fractions of J558LμM cells transduced with empty MIR vector or vector expressing wild-type, Δ2–6, p.Pro80Arg, p.Gly183Ser or p.Gly183Val PAX5 ($n \geq 6$ replicate transductions) were flow sorted, expanded and purity checked, and mRNA was extracted from 5–10 × 10^6 cells using TRIzol (Invitrogen). mRNA was quantified by spectrophotometry, and integrity was assessed using a 2200 Tapestation instrument (Agilent Technologies). Expression of wild-type and mutant *PAX5* alleles was verified by RT-PCR and sequencing, and by immunoblotting. Gene expression profiling was performed using Mouse 430v2 PM arrays (Affymetrix) as previously described[20]. Statistical analyses, principal-component analysis and unsupervised hierarchical clustering were performed using R 2.15.2 (ref. 36), Bioconductor 2.6 (ref. 37) and Spotfire Decision Site 9.1.1 (Tibco), and Partek Genomics Suite version 6.5 (6.11.0207). Data were normalized upon import using the Robust Multi-array Average algorithm[38]. To adjust for batch effects introduced by isolation and plate batches, we further corrected probe set signals with ComBat[39], which applies an empirical Bayes framework for adjusting data for batch effects. Probe sets with signal not above the background level (twice the average signal for the control probes with different GC content) across all samples were excluded from differential expression analysis using limma[40] with estimation of false discovery rate (FDR)[41]. For Gene Set Enrichment Analysis (GSEA)[42], we used gene sets obtained from the Molecular Signatures Database v3.0, Hardy Fraction (GSE38463)[20] and previous *PAX5* studies[16–19]. Gene sets with less than 10 or more than 500 genes were excluded, and significantly enriched gene sets after 1,000 permutations at an FDR of <0.25 are reported. *P* values were calculated using ANOVA with Dunnett's *post hoc* comparing each mutant allele to the wild-type allele. We also analyzed the overlap with the data from Revilla-i-Domingo *et al.*[17] and the expression differences between wild-type PAX5, empty vector (MIR) and the PAX5 mutants p.Gly183Ser, p.Gly183Val, p.Pro80Arg and Δ2–6 (**Supplementary Tables 17** and **18**). For the analysis of the sIgM-expressing subset, J558 cells infected with retroviruses expressing either wild-type PAX5 ($n = 3$), the p.Gly183Ser mutant ($n = 4$) or control MSCV-PIG empty vector ($n = 3$) were stained with PE-conjugated antibody to IgM and sorted for GFP and IgM (PE) double-positive cells (only single-positive GFP^+ cells in the case of cells infected with the empty vector) on a FACSAria II cell sorter (BD Biosciences). Total mRNA was extracted using TRIzol according to the manufacturer's instruction, and gene expression profiling was performed using Affymetrix GeneCHIP Mouse Gene 1.0 ST arrays. All data analysis was performed using Partek Genomics Suite 6.5 (6.11.0207). Data were normalized upon import using RMA, and an ANOVA analysis was then performed to determine any differences between the conditions. One mutant sample was removed from the analysis because of its appearance as an outlier during principal-component analysis. The genes used for analysis were constrained to those genes previously identified as being PAX5 targets in mouse pro-B cells[17]. Genes were classified as being either activated or repressed by PAX5 ($n = 122$ and 237, respectively). Revilla-i-Domingo *et al.* also identified the most significant subsets of activated or repressed PAX5 targets ($n = 20$ and 21, respectively)[17]. *P* values were calculated using the pbinom function in R 2.12.1.

Transcriptome sequencing. Transcriptome sequencing was performed on diagnostic and relapse samples obtained from 2 affected individuals in kindred 2 and on 139 sporadic childhood B-ALL samples as described previously[1,2,20]. The 139 sporadic childhood B-ALL samples included ALLs with *ETV6-RUNX1* fusion ($n = 54$), alteration of *ERG* ($n = 22$), hyperdiploidy with greater than 50 chromosomes ($n = 1$), hypodiploidy ($n = 8$), *BCR-ABL1* fusion ($n = 27$) and *BCR-ABL1*–like characteristics ($n = 27$). *PAX5* deletion and mutation status was derived from whole-exome and/or whole-genome sequencing of all cases.

mRNA-seq library construction. Total RNA was extracted using TRIzol. Total RNA quality and quantity were assessed on Agilent RNA6000 Chips (Agilent Technologies) and Qubit (Invitrogen). A standard mRNA-seq library was prepared from 1 μg of total RNA following Illumina's RNA-seq protocols, including DNase treatment and phenol purification, PolyA+ RNA selection using oligo(dT) beads, cDNA conversion, fragmentation by Covaris Ultrasonicator, end repair, deoxyadenosine tailing, adaptor ligation and PCR amplification (ten cycles). 10 pM of the library was clustered on an Illumina cBot and loaded into a flow cell on a HiSeq instrument for sequencing using the Illumina 2 × 100-bp sequencing kit.

Bioinformatic analysis. mRNA-seq paired-end reads were mapped to the human hg19 genome, RefSeq transcripts and known splice junctions using an in-house modified BWA mapping pipeline[43]. Transcript expression levels were estimated as fragments per kilobase of transcript per million mapped reads (FPKM), and gene FPKM values were computed by summing the transcript FPKM values for each gene using the Cuffdiff program available from the Cufflinks package[44]. We called a gene 'expressed' in a given sample if it had an FPKM value of ≥0.35 based on the distribution of FPKM gene expression levels and removed genes that were not expressed in any sample from the final gene expression data matrix for downstream analysis. We used FPKM[44] and limma[40] to derive a gene expression profile for ALL with p.Gly183Ser-mutated PAX5 compared to *ETV6-RUNX1*–negative B-ALL (**Supplementary Table 19**) and gene-set enrichment analysis to interrogate these expression profiles, including gene sets representing previously described genes activated and repressed by PAX5 (refs. 16–19) and genes regulated during mouse B-lymphoid development. As one-third of *ETV6-RUNX1*–positive ALL cases harbor focal *PAX5* deletions (but not sequence mutations) that influence the expression of PAX5 target genes, we also incorporated gene sets of up- and downregulated genes in *PAX5*-mutated *ETV6-RUNX1*–positive ALL derived from the analysis of RNA-seq and whole-genome sequencing data for *ETV6-RUNX1*–positive cases (**Supplementary Tables 20** and **21**). We also analyzed the overlap with the data from Revilla-i-Domingo *et al.*[17] and the expression differences between the familial ALL tumor samples (FAMALL) and non-*ETV* B ALL wild type for *PAX5* (nonETVBALL.*PAX5*WT), other B-ALLs wild type for *PAX5* (OtherBALL.*PAX5*WT) and all B-ALL cases, including those with *PAX5* mutations (OtherBALL) (**Supplementary Table 16**).

23. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
25. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
26. Pounds, S. *et al.* Reference alignment of SNP microarray signals for copy number analysis of tumors. *Bioinformatics* **25**, 315–321 (2009).
27. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
28. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
29. Mullighan, C.G. Single nucleotide polymorphism microarray analysis of genetic alterations in cancer. *Methods Mol. Biol.* **730**, 235–258 (2011).
30. Lin, M. *et al.* dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics* **20**, 1233–1240 (2004).
31. Cai, Y., Brophy, P.D., Levitan, I., Stifani, S. & Dressler, G.R. Groucho suppresses Pax2 transactivation by inhibition of JNK-mediated phosphorylation. *EMBO J.* **22**, 5522–5529 (2003).
32. Lundblad, A. *et al.* Immunochemical studies on mouse myeloma proteins with specificity for dextran or for levan. *Immunochemistry* **9**, 535–544 (1972).
33. Sitia, R., Neuberger, M.S. & Milstein, C. Regulation of membrane IgM expression in secretory B cells: translational and post-translational events. *EMBO J.* **6**, 3969–3977 (1987).
34. Maier, H. *et al.* Requirements for selective recruitment of Ets proteins and activation of *mb-1*/Ig-α gene transcription by Pax-5 (BSAP). *Nucleic Acids Res.* **31**, 5483–5489 (2003).
35. Hombach, J., Tsubata, T., Leclercq, L., Stappert, H. & Reth, M. Molecular components of the B-cell antigen receptor complex of the IgM class. *Nature* **343**, 760–762 (1990).
36. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2009).
37. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
38. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).

39. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
40. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
41. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).
42. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
43. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157–163 (2012).
44. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

doi:10.1038/ng.2754