LISBOA

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA LISBOA

TRABALHO FINAL
MESTRADO INTEGRADO EM MEDICINA

Instituto de Biologia Molecular

Identification of novel potential therapeutic targets for T-cell acute lymphoblastic leukaemia

Catarina Sofia Micaelo Fernandes

Julho'2019

LISBOA

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA LISBOA

TRABALHO FINAL
MESTRADO INTEGRADO EM MEDICINA

Instituto de Biologia Molecular

Identification of novel potential therapeutic targets for T-cell acute lymphoblastic leukaemia

Catarina Sofia Micaelo Fernandes

Orientado por:

Doutora Maria Joana Pinto Desterro

Julho'2019

ABSTRACT

T-cell acute lymphoblastic leukaemia (T-ALL) is a rare haematological neoplasm. Although several genetic lesions and cytogenetic abnormalities have been identified, uncertainty around their predictive value has limited therapeutic advances and aggressive high-dose combination chemotherapy regimens are still employed.

In this project, we performed a comprehensive cross-species oncogenomic comparison, by integrating transcriptome and exome data from a mouse model of T-ALL driven by the *TLX3* transcription factor with data on human *TLX3*+ T-ALL patients, to identify potential therapeutic target genes and surrogate targetable signalling pathways.

Analysis of RNA-sequencing and whole-exome sequencing data revealed that both the differential gene expression profiles and the mutational landscape of the murine *TLX3*+ T-ALL samples show little resemblance with the human counterpart, with a reduced number of the subtype-enriched alterations selected *a priori* upon bibliographic search being observed. At the transcriptomic level, only 12 of 59 (20%) candidate genes were consistently deregulated in all samples (S1-S4) and at the genomic level, of the 654 genes harbouring 933 somatic non-synonymous single-nucleotide variants in the leukaemic samples (S2-S4), only 4 were among the 37 (11%) selected recurrently mutated genes in *TLX3*+ T-ALL.

Interestingly, while *TLX3* overexpression was confirmed in all samples, experimental data does not corroborate *TCRA* downregulation, the key mechanism governing differentiation blockage in *TLX3*+ T-ALL, with *TCRA* being overexpressed in all samples.

As a result, the *TLX3*+T-ALL mouse model could not be validated against clinical data. Different molecular and cellular contexts and discrepancies in tumour-microenvironment interactions may have conditioned the appearance of non-representative clones, resulting in a model that does not reliably reproduce the human tumour biology. Overall, this work highlights the need for careful design and characterization of murine disease models to assess the translational relevance of the results obtained in this setting.

Keywords: T-cell acute lymphoblastic leukaemia; *TLX3*; transcriptome analysis; whole-exome sequencing analysis; molecular targeted therapy

This work reflects the author's opinion and not the opinion of the Faculty of Medicine of the University of Lisbon.

RESUMO

A leucemia linfoblástica aguda de células T (LLA-T) é uma neoplasia hematológica rara. Apesar da identificação de várias lesões genéticas e anomalias citogenéticas, incertezas acerca do seu valor preditivo têm impedido avanços terapêuticos, continuando a utilizar-se regimes agressivos de quimioterapia combinada em alta dose.

Neste projecto realizou-se uma comparação oncogenómica interespécies abrangente, integrando dados de transcriptoma e exoma dum modelo de ratinho de LLA-T promovida pelo factor de transcrição *TLX3* com dados de doentes com LLA-T *TLX3*+, para identificar potenciais alvos terapêuticos e vias de sinalização celular análogas.

A análise dos dados de sequenciação de RNA e do exoma completo revelou que os perfis de expressão génica diferencial e o espectro mutacional das amostras murinas de LLA-T *TLX3*+ apresentam escassas semelhanças com os equivalentes humanos, observando-se poucas das alterações recorrentes neste subtipo selecionadas *a priori* mediante pesquisa bibliográfica. A nível transcriptómico, apenas 12 de 59 (20%) genes candidatos estão consistentemente desregulados em todas as amostras (S1-S4) e a nível genómico, dos 654 genes que contêm as 933 variantes de nucleótido único somáticas e não-sinónimas encontradas nas amostras leucémicas (S2-S4), apenas 4 se incluem nos 37 (11%) genes frequentemente mutados em LLA-T *TLX3*+ selecionados.

Conquanto se tenha confirmado sobrexpressão de *TLX3* em todas as amostras, os resultados experimentais não corroboram a infra-regulação do *TCRA*, mecanismo chave que rege o bloqueio de diferenciação na LLA-T *TLX3*+, verificando-se sobrexpressão do *TCRA* em todas as amostras.

Com efeito, o modelo de ratinho de LLA-T *TLX3*+ não foi validado em função de dados clínicos. Diferenças nos contextos moleculares e celulares e nas interacções tumor-microambiente poderão ter condicionado o aparecimento de clones não representativos, culminando num modelo que não reproduz fiavelmente a biologia dos tumores humanos. Assim, este trabalho salienta a necessidade de conceber e caracterizar cuidadosamente os modelos de ratinho de doença a fim de avaliar a relevância clínica dos resultados obtidos neste contexto.

Palavras-chave: leucemia linfoblástica aguda de células T; *TLX3*; análise de transcriptoma; análise de sequenciação total do exoma; terapia molecular dirigida

O Trabalho Final exprime a opinião do autor e não da Faculdade de Medicina da Universidade de Lisboa.

RESUMO EXPANDIDO

Introdução:

A leucemia linfoblástica aguda de células T (LLA-T) é uma neoplasia de linfoblastos comprometidos com a linhagem T e representa 10-15% e 20-25% dos casos de LLA pediátrica e do adulto, respectivamente.

A activação de fatores de transcrição específicos dos linfócitos T constitui o cerne do processo de transformação maligna na LLA-T. Geralmente esta activação resulta de translocações que colocam estes oncogenes sob a influência de promotores de genes expressos durante o normal desenvolvimento das células T no timo e tem como consequência um bloqueio de diferenciação em diferentes estádios, definindo subgrupos moleculares da doença associados a padrões transcripcionais e imunofenótipos distintos.

Apesar dos avanços na compreensão da biologia da doença, os achados genéticos e citogenéticos carecem de valor prognóstico e preditivo e as abordagens terapêuticas são ainda baseadas em protocolos de associação de vários agentes de quimioterapia em alta dose, com efeitos adversos consideráveis. Além disso, embora estes esquemas terapêuticos tenham permitido atingir taxas de cura de aproximadamente 80% na população pediátrica e 60% na população adulta, o prognóstico de doentes de alto risco com doença primariamente resistente ou recidivante permanece muito reservado. Deste modo, o desenvolvimento de terapias-alvo baseadas em lesões genéticas específicas certamente beneficiará os doentes.

Este projecto foca-se num subgrupo molecular em particular: a LLA-T promovida pela sobrexpressão factor de transcrição da família *homeobox TLX3*, que perfaz cerca de 20-25% dos casos de LLA-T pediátrica e 5% dos casos no adulto. Neste trabalho utilizaram-se dados transcriptómicos e de sequenciação total do exoma dum modelo de ratinho de LLA-T com o objectivo de identificar eventos recorrentes na LLA-T *TLX3*+, que, por provavelmente representarem alterações em genes e/ou vias de sinalização celular críticos para o estádio de maturação das células transformadas, se revelam potenciais alvos terapêuticos.

Métodos:

Modelo experimental de LLA-T *TLX3*+: Produziram-se células em cultura que sobrexpressam *TLX3* (S1, estado pré-maligno) através da transdução retroviral do oncogene *TLX3* em timócitos *wild-type* de ratinho no estádio duplamente negativo (CD4-/CD8-) (S0, controlo). Estas células foram injectadas por via endovenosa em ratinhos imunodeprimidos *Rag2-/- γc-/-* que desenvolveram tumores em órgãos linfoides. Coletaram-se duas amostras de tecido esplénico, os tumores 7 (S2) e 8 (S3), e estabeleceu-se uma linha celular a partir de outro enxerto tumoral (S4, linha celular TAP). Este modelo foi estabelecido no Laboratório Pierre Ferrier no Centre d'Immunologie de Marseille-Luminy (Ciml).

Pesquisa bibliográfica e selecção de genes candidatos: Os motores de busca/bases de dados National Center for Biotechnology Information (NCBI) PubMed e Gene Expression Omnibus, European Bioinformatics Institute (EMBL-EBI) Ensembl e Google Scholar foram utilizados para conduzir uma pesquisa alargada de genes diferencialmente expressos ou mutados na LLA-T *TLX3*$^+$ em doentes humanos, utilizando os termos MeSH *T-cell acute lymphoblastic leukaemia, gene expression, transcriptome, exome* e *mutation.* Subsequentemente, caracterizaram-se as funções biológicas destes genes com recurso às bases de dados Catalogue Of Somatic Mutations In Cancer (COSMIC) Cancer Gene Census, Candidate Cancer Gene Database (CCGD), Atlas of Genetics and Cytogenetics in Oncology and Haematology, NCBI Gene, EMBL-EBI UniProt e Gene Ontology (GO) para uma selecção mais refinada de candidatos presumivelmente envolvidos na iniciação e progressão tumoral.

Análise bioinformática de dados: Os dados de transcriptoma foram obtidos mediante sequenciação de todo o RNA das amostras S0-S4 e cálculo da expressão diferencial entre os timócitos *wild-type* (S0, controlo) e as quatro amostras restantes (S1-S4), utilizando o *software* DESeq2. A ferramenta Strelka2 foi utilizada para realizar uma detectar variantes somáticas (*i.e., variant calling*) nas amostras S2-S4, tomando o exoma de S1 como sequência de referência.

Resultados:

Análise de dados de transcriptoma: A pesquisa bibliográfica culminou numa selecção de 59 genes com funções biológicas associadas a cancro e expressão anormal reportada em doentes com LLA-T *TLX3*$^+$. A análise comparativa de expressão diferencial entre o modelo de ratinho de LLA-T e a LLA-T em humanos revelou pouca consistência, com apenas 12 genes identicamente desregulados em todas as amostras (S1-S4). A comparação entre amostras permitiu aferir agrupamento (*clustering*) entre os perfis de expressão génica dos pares S1/S4 e S2/S3. Os dados transcriptómicos corroboram a sobrexpressão de *TLX3* transversal a todas as amostras, mas, surpreendentemente, o nível de expressão em S4 é 100 vezes menor do que em S1.

A sobrexpressão de *TLX3* induz um bloqueio de diferenciação num estádio cortical precoce, devido à supressão da recombinação e expressão do gene que codifica subunidade alfa do receptor de células T (*TCRA*). Porém, a infra-regulação do *TCRA* não se verificou no modelo experimental, constatando-se, pelo contrário, sobrexpressão do gene em todas as amostras.

A delecção do *locus CDKN2A/B*, com consequente inactivação dos genes supressores de tumor p16$^{\text{INK4A}}$, p14$^{\text{ARF}}$ e p15$^{\text{INK4B}}$, está presente na maioria dos casos de LLA-T, independentemente do subgrupo molecular. Contudo, não há evidência que indicie a reprodução deste evento pelo modelo experimental, dada a sobrexpressão deste *locus* em todas as amostras.

A activação da via IL7R/JAK1/3/STAT5 é predominantemente observada entre os doentes com LLA-T *TLX3+*, pelo que se avaliou sumariamente a actividade de sinalização desta via com base nas alterações a nível transcripcional expectáveis perante a hipótese de um aumento de actividade.

Globalmente, os resultados foram muito heterogéneos e incongruentes, não fundamentando a hipótese colocada.

Análise de dados de sequenciação total do exoma: Um total de 43056 variantes de nucleótido único somáticas foram obtidas entre as amostras leucémicas S2-S4. Após exclusão de todas as variantes previamente descritas/polimorfismos e variantes sinónimas, 933 variantes exónicas não-sinónimas em 654 genes diferentes foram analisadas.

Numa primeira abordagem, procuraram-se variantes num conjunto de 37 genes frequentemente mutados em doentes com LLA-T *TLX3*+, selecionados a partir da Literatura. Somente quatro destes genes - *NOTCH1*, *KMT2C*, *NF1* e *EP300* - se encontram mutados nas amostras de origem murina.

Numa segunda fase, uma análise não enviesada revelou que 210 dos 654 genes supramencionados estão potencialmente envolvidos em neoplasias hematológicas. O estudo detalhado das 38 variantes com impacto funcional moderado ou alto que afetam estes genes revelou a existência de alinhamento de sequências entre o genoma do ratinho e o genoma humano a nível nucleotídico em 21 variantes. Quatro destas variantes são mutações reportadas em doentes oncológicos, estando descritas na base de dados COSMIC. Ocorrem nos genes *NOTCH1, PTEN, KLHL42* e *BRD4.* À data, a mutação que ocorre no gene *NOTCH1*, c.5033T>C / p.L1678P, foi detetada em neoplasias do foro hematológico, sobretudo LLA-T. As restantes mutações foram relatadas apenas em doentes com neoplasias não-hematológicas.

Discussão:

O modelo experimental de LLA-T *TLX3*+ não aparenta recapitular a doença em humanos, dada a ausência das principais características moleculares que definem este subtipo de LLA-T, nomeadamente, a supressão da expressão do *TCRA*, o mecanismo chave conducente ao bloqueio de diferenciação durante a transformação maligna. Numa perspectiva mais abrangente, a integração e análise comparativa dos dados de transcriptoma e de sequenciação de exoma também não conferem robustez a este modelo.

Vários aspectos importantes devem ser considerados na interpretação destes resultados. Primeiramente, este é um modelo que contém informação genética apenas de origem murina. Em segundo lugar, o microambiente tumoral poderá ser bastante distinto da doença humana, não só devido à premissa anteriormente enunciada, mas também ao facto de se utilizarem ratinhos imunodeprimidos, sendo, portanto, um microambiente privado da normal interacção entre as células neoplásicas e o sistema imunitário. Outra observação interessante prende-se nas semelhanças encontradas entre as duas linhas celulares por oposição às duas amostras de tumor sólido. Tendo este modelo origem numa linha celular, a questão da adaptação às pressões selectivas *in vitro* é incontornável, na medida em que estas, certamente, diferem das encontradas *in vivo*, e algumas das alterações adquiridas pelas células em cultura possivelmente não reverterão mesmo após a transferência para o organismo do ratinho. Note-se também que o modo pelo qual se atingiu a sobrexpressão de *TLX3* na linha celular – expressão ectópica mediada por transdução com retrovírus – diverge do observado na LLA-T em humanos, sendo difícil estimar as consequências que tal poderá arrecadar, desde ao nível do papel do *TLX3* na iniciação tumoral, destacando-se a influência no bloqueio maturativo, ao nível do papel na progressão tumoral.

Em conjunto, estas diferenças intrínsecas e extrínsecas no contexto celular terão provavelmente contribuído para uma evolução clonal marcadamente heterogénea e que não mimetiza a doença humana, a tal ponto que, à vista da grande diminuição na expressão diferencial de *TLX3* entre o estado pré-leucémico S1 e a linha celular leucémica S4, alguns clones poderão mesmo ter-se tornado independentes da alteração fundadora (*driver*) para a manutenção do fenótipo maligno. Acresce que a LLA-T é uma doença biologicamente complexa e que, apesar do contributo inegável dos estudos genómicos e transcriptómicos, é cada vez mais reconhecido o papel de outros intervenientes, como os factores epigenéticos e o genoma não codificante, que não foram objecto deste estudo.

Em suma, este trabalho reflecte a necessidade de proceder a uma conceptualização e caracterização/validação cuidadosas dos modelos de ratinho e de se interpretar e extrapolar com prudência os resultados obtidos na investigação em modelos animais de doença para o contexto clínico.

TABLE OF CONTENTS

INTRODUCTION

T-cell acute lymphoblastic leukaemia (T-ALL) is a rare haematological neoplasm of lymphoid progenitors committed to the T-cell lineage, accounting for 10-15% of paediatric and 20-25% of adult cases of ALL (1,2,).

Evidence on T-ALL's biology has grown over the last decade. Impaired differentiation occurring at distinct stages of thymocyte maturation is considered the core of T-ALL's pathogenesis and is usually driven by chromosomal translocations that lead to the overexpression of several oncogenic T-cell-specific transcription factors, defining disease molecular subgroups with unique gene expression signatures and immunophenotypes (3). T-ALL then arises from a multistep oncogenic transformation process, where cumulative genetic lesions disrupt key molecular pathways regulating T-cell development, proliferation and survival (1,2). Recent genomic studies have identified an average of 10 to 20 biologically relevant genomic lesions present in each T-ALL case that can be grouped into targetable pathways, including NOTCH, JAK/STAT, PI3K/Akt/mTOR and RAS-MAPK (4,5,6). Amongst these are *NOTCH1* activating mutations and *CDKN2A/B* inactivating deletions, two hallmark features of T-ALL, with a noteworthy prevalence of 60% and 70%, respectively, and often co-occurring (7,8). *NOTCH1* encodes a transmembrane receptor that acts as a ligand-activated transcription factor responsible for development and fate specification of thymocytes and the *CDKN2A/B* locus at chromosome 9p21 encodes for the tumour suppressors p16$^{\mathrm{INK4A}}$ and p14$^{\mathrm{ARF}}$ and p15$^{\mathrm{INK4B}}$.

However, the prognostic and predictive value of these abnormal genetic and cytogenetic findings is unclear and treatment still relies in aggressive high-dose combination chemotherapy regimens, with considerable short and long-term side effects (4). Moreover, despite overall survival improvement, with cure rates reaching around 80% in children and 60% in adults, the prognosis of patients with high-risk primary-resistant or relapsed disease remains very poor, with scarce therapeutic options available (9). Thereby, and given the great biological heterogeneity of T-ALL, the development of targeted therapy driven approaches based on defined molecular features would likely benefit patients, who strive for more effective and less toxic anti-leukaemic drugs. Likewise, these features could ultimately be used for patient risk stratification and treatment adjustment, avoiding overtreatment in low-risk patients and sparing them from treatment-associated toxicities.

In this project we focus on a specific subgroup: T-ALL driven by *TLX3*, a transcription factor of the homeobox family, which accounts for 20-25% of paediatric and 5% of adult T-ALL cases (10). In the majority of cases, *TLX3* overexpression is the result of a translocation juxtaposing the *TLX3* oncogene to the *BCL11B* locus, under the control of strong regulatory sequences, as *BCL11B* is a gene normally expressed during T-cell development (10). Less often, the *TLX3* locus is aberrantly activated by other chromosomal rearrangements or by translocation-independent mechanisms.

This T-ALL subgroup is characterised by early cortical thymic maturation arrest as a consequence of *TLX3*-mediated suppression of T-cell receptor alpha *(TCRA*) gene rearrangement and expression (11). Importantly, tumours that overexpress another homeobox factor, *TLX1*, share this differentiation blockage mechanism and, thereby, exhibit a great transcriptional overlap with *TLX3*-expressing tumours (12).

Controversy exists on whether *TLX3* overexpression can independently predict the outcome in T-ALL, with some studies claiming association with poorer prognosis and higher incidence of relapse and others concluding for no clear impact on prognosis or even the opposite (13).

Subtype-enriched genomic alterations can be particularly interesting therapeutic targets as they probably represent genes/pathways critical to specific stages of T-cell development, since it has been observed that different signalling pathways are preferentially activated depending on the maturational stage of the leukaemia cells and, similarly, mutated genes also vary across T-ALL subtypes (5,6,14,15). For example, activation of the IL7R/JAK1/3/STAT5 and, to a lesser extent, of the RAS-MAPK oncogenic signalling pathways is predominantly found amid *TLX3*+ T-ALL patients (5,14,15).

In this project, we compared transcriptomic and whole-exome sequencing (WES) data from a *TLX3*+ T-ALL mouse model with human clinical data to identify genes involved in the malignant transformation of *TLX3*+ T-ALL with the main goal of disclosing novel potential therapeutic targets and surrogate targetable pathways.

METHODS

Experimental mouse model of *TLX3*+ T-ALL: *TLX3*-overexpressing cultured cells (S1, pre-malignant state) were generated by retroviral transduction of the *TLX3* oncogene into primary wild-type double-negative (CD4-/CD8-) murine thymocytes (S0, control). These cells were intravenously injected into Rag2-/- γc-/- immunodeficient mice, that developed tumours in lymphoid organs. Two splenic tissue samples, tumours 7 (S2) and 8 (S3), were collected and a cell line was established from another tumour graft (S4, TAP cell line) (Figure 1). These experiments were performed in Pierre Ferrier's Lab at the Centre d'Immunologie de Marseille-Luminy (Ciml).

Bibliographic Search: The National Center for Biotechnology Information (NCBI) PubMed and Gene Expression Omnibus, the European Bioinformatics Institute (EMBL-EBI) Ensembl and the Google Scholar databases were used to search for genes differentially expressed and/or mutated in human *TLX3*+ T-ALL, using the MeSH terms *T-cell acute lymphoblastic leukaemia, gene expression, transcriptome, exome* and *mutation.* The studies included used biological samples from T-ALL patients and allowed for the identification of *TLX3*+ T-ALL patients/subgroup among the studied cohort. Studies also provided a clear description of the methods applied. Studies that did not meet these criteria were excluded.

Candidate gene selection: The Catalogue Of Somatic Mutations In Cancer (COSMIC) Cancer Gene Census, the Candidate Cancer Gene Database (CCGD), the Atlas of Genetics and Cytogenetics in Oncology and Haematology, Gene Ontology (GO), the NCBI Gene and the EMBL-EBI UniProt databases were then used to characterise the genes biological functions and to assess their involvement in cancer for a more refined selection of candidates owning a putative or potential role in cancer initiation and progression. Depending on the specificities of the selected studies, additional criteria were applied (Supplementary Methods 1 and 2).

Transcriptomic data analysis / Bioinformatics tools: Total cell mRNA was extracted and sequenced by paired-end RNA-Seq and whole-transcriptome data from samples S0-S4 was obtained (Pierre Ferrier's Lab). DESeq2 [(16)] was used to calculate differential expression between wild-type thymocytes (control, S0) and the four remaining samples (S1-S4). The gene expression profiles of the different experimental model samples were then compared to the current literature on human *TLX3*+ T-ALL. Additionally, the expression levels of the main components and transcriptional targets of the IL7R/JAK/STAT pathway were separately analysed.

RT-PCR and gel electrophoresis: Transcriptional levels of *TLX3* in TAP cell line (S4) were assessed by RT-PCR using two specific pairs of primers and *GADPH* as a control endogenous gene (Supplementary Figure 1). RT-PCR products were analysed by agarose gel electrophoresis.

Whole-exome sequencing data analysis / Bioinformatics tools: Exome sequencing reads were mapped to the reference mouse genome assembly MGSCv37 (mm9) using Bowtie2 aligner (17). Processing of alignments was performed as outlined in the Genome Analysis toolkit (GATK) best practices (18,19), including duplicate reads removal, realignment around indels and base quality score recalibration (BQSR). Strelka2 (20) in was used for somatic variant calling from samples S2-S4, using S1 as the reference exome (control). Only variants with 'PASS' in the filter field of the VCF were selected. Picard tool LiftoverVcf (http://broadinstitute.github.io/picard/) was used to convert variants to the latest genome version, GRCm38 (mm10). Finally, Ensembl Variant Effect Predictor (VEP) (21) was used to annotate all the obtained VCFs. The single-nucleotide variants (SNVs) obtained were submitted to a filtering process by excluding all previously reported variations/polymorphisms and synonymous variants. The remaining SNVs were considered for mutational analysis that combined a search for mutations in known recurrently mutated genes in human *TLX3*+ T-ALL and an unbiased analysis of all the mutated genes that were potentially involved in haematological malignancies, as reported in the Candidate Cancer Gene Database. For the variants with an estimated moderate to high functional impact, according to the Ensembl VEP tool, the analysis additionally comprised the verification of pairwise genome alignment between mouse and human. In cases where the sequence of interest was conserved interspecies, with the exact same nucleotide at the position where the SNV occurred, the COSMIC Genome Browser and the NCBI ClinVar databases were used to search for disease-specific human SNVs.

RESULTS

Transcriptomic data analysis

Three studies provided data on gene expression in human *TLX3*+ T-ALL and another study focused specifically on chromosomal imbalances and uncovered 5 recurrent genomic deletions in *TLX3*+ T-ALL patients (Table 1). A total of 78 genes with reported abnormal expression in *TLX3*$^+$ T-ALL patients were obtained (Supplementary Table 1). After an extensive filtering process and combining evidence on CNVs coming from several studies (Supplementary Tables 1 and 2, Supplementary Methods 1 and 2), 59 cancer-associated genes (36 downregulated and 23 upregulated) were selected (Table 2). The selected genes are listed in the first row of Figure 2 and gene expression profiling results are shown in Figure 2-A.

Transcriptomic data of the mouse orthologs for the 59 selected human genes shows very little consistency with the gene expression profile observed in *TLX3*$^+$ T-ALL patients, with only 12 genes (20%, blue box in Figure 2-B) genes in fully agreement in all samples (S1-S4). Even when each sample is analysed pairwise with the human *TLX3*+ T-ALL data, the consistency is below 50% - 44% (26/59) in S1, 42% (25/59) in S2, 37% (22/59) in S3 and 46% (27/59) in S4 (Figure 2-B). Pairwise comparison was done between all mouse samples (Supplementary Figure 2) and, as it can be seen in Figure 2-C, the greatest number of identically deregulated genes occurs between mouse samples S1/S4 (47/59 genes, 80%) and S2/S3 (48/59 genes, 81%). Comparative expression analysis between the pre-leukaemic *TLX3*-overexpressing state S1 and samples S2-S4 shows acquisition of opposite deregulation for 30 genes during leukaemic transformation in the mouse recipient in at least one of the leukaemic samples S2-S4, but the expression shifts are reproducing the expression deregulation described in the literature for human *TLX3*+ T-ALL in only 16 out of these 30 genes (53%) (Figure 2-D) – 10 in a single sample, 5 in two samples and 1 in all leukaemic samples.

TLX3 overexpression and *TCRA* downregulation, core features of human *TLX3*+ T-ALL, were analysed in detail. *TLX3* overexpression was validated, with transcriptional upregulation confirmed in all samples, although to different extents, the highest fold-change (FC), of 20577, being found in S1 and the lowest, of 225, in S4 (Figure 2-E). Further evaluation of *TLX3* expression levels in S4 by RT-PCR revealed that none DNA product was amplified from the cDNA extracted from S4 cells (Supplementary Figure 1). In contrast, data does not corroborate *TCRA* downregulation in the experimental mouse model, with *TCRA* being overexpressed in all samples (FC ranging from 83 in S2 to 659 in S1) (Figure 2-E). As for the prevalence of *CDKN2A/B* deletions in the majority of T-ALL cases, irrespective of the molecular subgroup, the *CDKN2A/B* locus expression levels were also evaluated, revealing overexpression in all samples (Figure 2-E).

In the *TLX3*+ T-ALL mouse model it is difficult to determine the activation status of the IL7R/JAK1/3/STAT5 pathway based on the expected IL-7 receptor activation induced transcriptional changes, either *STAT5B*-mediated, as for *SOCS2, PIM1* and *CISH* upregulation, or via crosstalk with the PI3K/Akt/mTOR pathway, in the case of *BCL-2* upregulation and *CDKN1B* (p27^{kip}) downregulation. For sample S1 there is some evidence of signalling activation, given that all *STAT5B* transcriptional targets are upregulated, with a FC greater than 2, and that both the *IL7R* and the main signalling effectors are also being overexpressed. For the other samples, S2, S3 and S4, it is hard to evaluate and to conclude for pathway activation due to very incongruent inter and intra-samples results. For example, all three samples overexpress one or more components of the signalling pathway, but only one of the target genes is overexpressed in samples S2 and S3. Oddly, S4 overexpresses all the pathway components except for *JAK3*, shows transcriptional downregulation of *DNM2*, which encodes for a protein that intervenes in clathrin-dependent endocytosis of the IL-7 receptor - and thereby its loss usually typically leads to increased IL-7 receptor surface expression and signalling –, and yet none of the target genes is upregulated (Figure 2-F).

Whole-exome sequencing data analysis

Altogether, 43056 somatic SNVs were identified by somatic variant calling – 6375 in S2, 36488 in S3 and 193 in S4 - of which 42013 (97,6%) were reported SNVs, with associated Reference SNP (rs) ID accession numbers, and were thereby excluded from further analysis. An additional set of 110 synonymous SNVs were also rejected, leaving a total of 933 exonic non-synonymous SNVs – 176 in S2, 602 in S3 and 155 in S4 – affecting 654 different genes (data not shown).

Ninety-six genes were found to be mutated and/or deleted/amplified in *TLX3*+ T-ALL patients over seven different cohorts (Table 3, Supplementary Table 2).

The first phase of the WES data analysis was focused in a subset of 37 genes (Table 4), in which mutational events were enriched in the *TLX3*+ T-ALL subgroup or occurred in at least 3 patients, assigned to two or more different studies. Only four of the target genes were found to be mutated amongst the three leukaemic samples: *NOTCH1* in S2, *KMT2C* in S2 and S4, *NF1* in S3 and *EP300* in S4 (Figure 3, in yellow). Subsequently, a more comprehensive unbiased analysis was done including all of the 654 mutated genes. Remarkably, we found that about one third (n=210) of these genes (listed in the rows of Figure 3) are recorded as potentially involved in the biology of haematological malignancies in the CCGD. From the 269 SNVs affecting these 210 genes, the 38 SNVs with moderate and high functional impact, *i.e.*, protein-altering SNVs, were examined in depth, in order to unveil any possible similarity to the human disease (Table 5). There were 25 SNVs occurring at conserved genomic regions and 21 SNVs displayed the exact same nucleotide as of the human sequence. Nevertheless, only 4 SNVs were variants already reported for patients suffering from different types of malignancies (first four genes listed in Table 5). Strikingly, one of the four genes harbouring these mutations is *NOTCH1*. This variant, c.5033T>C / p.L1678P, is likely an oncogenic activating mutation and is reported in 80 T-ALL patients. So far, the remaining genes, *BRD4, KLHL42* and *PTEN*, harbour variants reported only for non-haematological malignancies.

DISCUSSION

The aim of this work was to identify secondary cooperative leukaemogenic events subsequent to a founding driver alteration, namely, the overexpression of the *TLX3* transcription factor, through an integrative multi-modal approach combining genomics and transcriptomics.

Overall, our model does not appear to be recapitulating the human disease, and this is greatly illustrated by the overexpression of *TCRA* in all experimental samples, a finding that undermines the reliability of the model from its conception, as the key molecular mechanism of differentiation blockage leading to malignant transformation in *TLX3*+ T-ALL is absent. Regarding this particular finding, we can wonder if the means by which *TLX3* overexpression is achieved can influence its function. While in in our model, murine cells *TLX3* overexpression was ectopically-induced by retroviral transduction, in human leukaemia cells *TLX3* upregulation typically results from translocations that put this oncogene under the influence of strongly active promoters. Therefore, it would be interesting to assess the maturation status of the T-cells/leukaemia cells.

In our work, we identified gene expression clustering between cell lines (S1/S4) versus tumour grafts (S3/S2). The use of cell-line derived tumour grafts brings up all the concerns about adaptation to the selective pressures of *in vitro* culture and the uncertainties on whether some changes will not revert upon engraftment in the mouse organism. Indeed, studies that included both patient-derived cell-lines and primary tumour samples have reported substantial differences in terms of the amount of mutations, with cultured cells harbouring significantly more mutations [(4,26,27)]. Evidence also points to the acquisition of mutations in certain genes during the course of in vitro culture, which are underrepresented or even lacking *in vivo* [(26)].

Mutations in genes encoding for proteins that, directly or indirectly, cause IL7R/JAK1/3/STAT5 pathway activation have been extensively documented and are found in as much as one third of paediatric T-ALL cases [(14,15)]. *TLX3*-driven T-ALL happens to be particularly prone to activate this signalling pathway, which is meaningful from a clinical perspective, since IL7R/JAK1/3/STAT5 pathway activation has been linked to steroid treatment resistance and poor outcome and is thought to be overrepresented in relapsed T-ALL cases [(15)]. Regarding the role IL7R/JAK1/3/STAT5 signalling in our model, as aforementioned, results are very erratic and inconsistent, and a judicious interpretation would be to say that this pathway is not having an essential role in driving tumour progression. Although S1 shows an expression profile more coherent with IL7R/JAK1/3/STAT5 pathway activation it is noteworthy that this pattern seems more compatible to what has been described for the activation in normal T-cells. In fact, IL-7 is an essential cytokine for normal thymocyte development and homeostasis, being a positive regulator of cellular proliferation. However, in T-ALL cells, besides the proliferative effects, the activation of this signalling network promotes survival due to the downstream activation of the PI3K/Akt/mTOR pathway and subsequent upregulation of anti-apoptotic *BCL-2* and downregulation of *CDKN1B* (p27$^{\mathrm{kip}}$), a cyclin-dependent kinase inhibitor [(28,29)]. In S1, *CDKN1B* expression is not affected and *BCL-2* overexpression could be *STAT5B*-mediated, just as it is has been postulated for normal T-cells [30].

WES data analysis had also limitations. The first big drawback is the absence of control germline matched wild-type DNA. The use of S1 as the source of the genetic material to serve as reference is an implicit flaw present from the beginning of the analysis that affected all subsequent results. The pre-leukaemic sample S1 already owns the initial leukaemia driver event and, furthermore, these are immortalized cultured cells, and so the odds of this sample having accumulated genetic lesions throughout the time that preceded engraftment in the murine model are virtually 100%. If some of these alterations confer advantages to the proliferating cells and represent novel drivers in malignant transformation most likely they would be passed to the derived leukaemic samples and would be missed in our WES data analysis. Thereby, some of the most important leukaemogenic events could have passed unnoticed due to the study design.

Even though we did not find many *TLX3*+ T-ALL recurrently mutated genes to be mutated in our leukaemic samples (only 4/37 amongst S2-S4), we found several mutated genes that could be contributing to tumour progression and maintenance in the experimental model. It is important not only to recognize the co-occurrence of certain genetic lesions, but also the order by which those defects are acquired, since, as stated before, the context, with emphasis on the hematopoietic developmental stage, where the defect arises is likely very relevant for its leukaemogenic action. In this view, data on allelic frequency would have been critical to gain insights on the clonality of the variants found in our samples, since mutations shared by the majority of the clones tend to be acquired earlier in tumour progression, advocating for a role in malignant transformation and cancer initiation.

Moreover, data on patients suggests that subclonal evolution by acquisition of subclonal driver mutations is a trait frequently present in T-ALL and is probably related to treatment resistance, influencing therapeutic outcomes [5,31]. In effect, the different cell-intrinsic and extrinsic contexts might have conditioned clonal evolution in our model in a direction that does not mimic the human disease, in such an extent that some samples may have become *TLX3*-independent. The decrease in *TLX3* expression of approximately 100 times between S1 and S4 and the fact that *TLX3* expression was not detected by RT-PCR of cDNA extracted from S4 cells, strongly suggests that in the S4 cell line other events than *TLX3* overexpression are keeping the malignant phenotype.

This model is not reproducing the main molecular features currently described for human *TLX3*+ T-ALL and a series of factors can be influencing these results. Adding to the previous considerations, is the use of a model that only carries murine genetical information and with a defective immune system, interdicting the interplay between the immune system and neoplastic cells that naturally occurs in the human leukaemic niche.

In fact, the molecular understanding of T-ALL poses important challenges. Despite the undeniable contribution of genomic and transcriptomic studies to the current understanding of the biology of

T-ALL, more and more attention is being paid to other players in this disease. For instance, T-ALL is a neoplasm where epigenetic modulation has a substantial role (32), moreover in *TLX3*+ T-ALL, a subgroup where alterations in epigenetic regulators are particularly common, in some studies exceeding 90% (5). This is in line with a study that found association between mutations in genes that encode for components of the IL7R/JAK1/3/STAT5 pathway, a signalling pathway preferentially activated in *TLX3*+ T-ALL, and epigenetic regulators (14). An extra layer of complexity is added when we consider the non-protein coding genome, since miRNAs and lncRNAs have also shown to be active players in the pathogenesis of T-ALL.

Another challenge ahead in the development of targeted therapies is the apparent difference in the genetic background of adult and paediatric T-ALL, and its possible repercussions on the tumour's clinico-biological behaviour. Disparities are noticed not only regarding the average number of genetic defects, being higher in the adult population, but also in the spectrum of affected genes (4,6,14). However, it is not well understood if this truly implies the existence of more driver events in adult T-ALL or whether this discrepancy in the number of mutations is achieved at the expense of non-driver events, since the distinction between driver and passenger mutations cannot rely solely on sequencing studies and functional assays are warranted.

After all, this work highlights the need for careful design and characterisation of murine disease models and cautious interpretation and extrapolation of results to the clinics. Different types of models are suitable for different purposes, but even the most successfully established disease models have dismal outcomes, as success in preclinical drug trials poorly translates to the clinical set, with less than 8% of the tested drugs advancing into clinical trials (33).

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my supervisor, Professor Joana Desterro for her time, support and dedication to this work. Besides the constructive comments and advice that greatly contributed to improve this project throughout time, Professor Joana also gave me warm encouragement to take the best profit out of my work and I am exceedingly thankful for having had the opportunity to join her and present my work at the 1st International Conference on Acute Lymphoblastic Leukaemia organized by the European School of Haematology.

I would also like to express my greatest appreciation to Professor Carmo-Fonseca who motivated me and followed my work closely and who was always available to help me and to share her expertise. Her guidance and personal example were crucial for defining my journey at the Faculty of Medicine of the University of Lisbon and her constant enthusiasm reinforced my will to pursue a career in Medical Research.

I feel very privileged for having had the chance to develop this project at Professor's Carmo Lab and I have to extend my sincere gratitude to all the people who contributed to this work: to Pedro Barbosa and Marta Ferreira, whose assistance with the Bioinformatics analysis was of the utmost importance, and to Professor Noélia Custódio, who besides helping me with the wet-lab experiments, provided me valuable insights on the design of the experimental model.

Special thanks also to Professor João Barata for his feedback as an expert on the activation of the IL-7R signalling pathway in T-cell acute lymphoblastic leukaemia.

I gratefully acknowledge the funding received from the Gabinete de Apoio à Investigação Científica, Tecnológica e Inovação (Faculty of Medicine of the University of Lisbon).

Finally, I dedicate this work to my family - my parents Rosa and Hilário, my brother Pedro and my grandparents Lucília and Manuel -, who nurtured me with love and care, to Marcos and to my dearest friends, who cheered me up in times of self-doubt, always believing I would be able to overcome all the challenges I was faced with. I owe all that I am and all I have achieved to them.

REFERENCES

1. Belver, L. and Ferrando, A. (2016). The genetics and mechanisms of T cell acute lymphoblastic leukaemia. *Nature Reviews Cancer*. 16: 494-507;

2. Van Vlierberghe, P. and Ferrando, A. (2012). The molecular basis of T cell acute lymphoblastic leukemia. *The Journal of Clinical Investigation*. 122(10): 3398–3406

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

4. De Keersmaecker, K. *et al*. (2013). Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet*. 45(2):186-190

5. Liu, Y. *et al.* (2017). The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nature Genetics*. 49 (8):1211-1218

6. Chen, B. *et al*. (2018). Identification of fusion genes and characterization of transcriptome features in T-cell acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. USA*. 115:373–378

7. Hebert, J. *et al.* (1994). Candidate tumor-suppressor genes MTS1 (p16INK4A) and MTS2 (p15INK4B) display frequent homozygous deletions in primary cells from T- but not from B-cell lineage acute lymphoblastic leukemias. *Blood*. 84:4038–44.

8. Weng, A.P. *et al*. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science*. 306(5694):269–271

9. Pui, C. H., Robison, L. L. & Look A. T. (2008). Acute lymphoblastic leukaemia. *Lancet.* 371(9617):1030-1043.

10**.** Bernard, O. A. *et al.* (2001). A new recurrent and specific cryptic translocation, t(5;14)(q35;q32), is associated with expression of the *Hox11L2* gene in T acute lymphoblastic leukemia. *Leukemia.* 15, 1495–1504

11. Dadi, S. *et al.* (2012). TLX Homeodomain Oncogenes Mediate T Cell Maturation Arrest in T-ALL via Interaction with ETS1 and Suppression of TCRα Gene Expression. *Cancer Cell*. 21(4):563–576.

12. Della Gatta, G. *et al.* (2012). Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nature Medicine*. 18, 436–440

13. Ma, J. *et al.* (2014). The effect of TLX3 expression on the prognosis of pediatric T cell acute lymphocytic leukemia—a systematic review. *Tumor Biology*. 35: 8439–8443

14. Vicente C, *et al.* (2015). Targeted sequencing identifies associations between IL7R-JAK mutations and epigenetic modulators in T-cell acute lymphoblastic leukemia. *Haematologica.* 100:1301–1310.

15. Li, Y. *et al*. (2016). IL-7 Receptor Mutations and Steroid Resistance in Pediatric T cell Acute Lymphoblastic Leukemia: A Genome Sequencing Study. *PLoS Med*. 13(12): e1002200

16. Love, M.I. *et al.* (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*. 15(12): 550

17. Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9(4): 357–359

18. DePristo, M. A. *et al*. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet*. 43, 491–498

19. Van der Auwera, G.A. *et al*. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*. 43: 10.11.1–33

20. Kim, S. *et al.* (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*. 15, 591–594

21. McLaren, W. *et al*. (2016). The Ensembl Variant Effect Predictor. *Genome Biol*. 17:122

22. Van Vlierberghe, P. *et al*. (2008). The recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia. *Blood.* 111(9):4668–4680

23. Homminga, I. *et al.* (2011). Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. *Cancer Cell.* 19:484–497

24. Van Vlierberghe, P. *et al*. (2008). Cooperative genetic defects in TLX3 rearranged pediatric T-ALL. *Leukemia*. 22:762–770

25. Seki, M. *et al.* (2017). Recurrent SPI1 (PU.1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia. *Nat Genet*. 49:1274–1281

26. Kalender Atak, Z. *et al.* (2012). High Accuracy Mutation Detection in Leukemia on a Selected Panel of Cancer Genes. *PLoS ONE.* 7(6): e38463

27. Kalender Atak, Z. *et al.* (2013). Comprehensive Analysis of Transcriptome Variation Uncovers Known and Novel Driver Events in T-Cell Acute Lymphoblastic Leukemia. *PLoS Genet.* 9(12): e1003997

28. Barata, J.T. *et al.* (2001). Interleukin-7 promotes survival and cell cycle progression of T-cell acute lymphoblastic leukemia cells by down-regulating the cyclin-dependent kinase inhibitor p27(kip1). *Blood*. 98, 1524–1531

29. Barata, J. T, *et al.* (2004). Activation of PI3K is indispensable for interleukin 7-mediated viability, proliferation, glucose use, and growth of T cell acute lymphoblastic leukemia cells. *J. Exp. Med*. 200(5): 659–669

30. Rathmell, J. C. *et al.* (2001). IL-7 enhances the survival and maintains the size of naive T cells. *J Immunol*. 167:6869–76

31. Neumann, M.et al. (2015). Mutational spectrum of adult T-ALL. Oncotarget. 6(5):2754-2766

32. Peirs, S. *et al.* (2015). Epigenetics in T-cell acute lymphoblastic leukemia. *Immunol Rev*. 263:50–67

33. Mak, I. W., Evaniew, N. & Ghert, M. (2014). Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res*. 6(2):11

FIGURES and TABLES

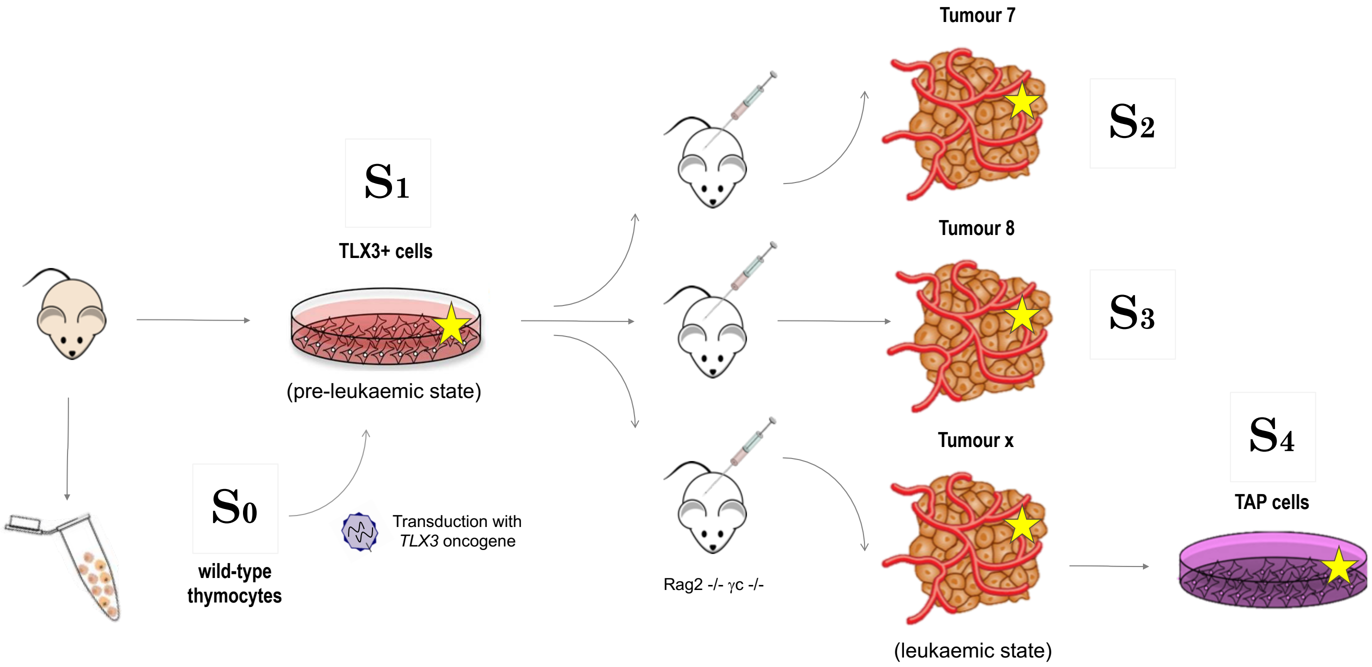


Figure 1. Schematic representation of the experimental mouse model of *TLX3*-overexpressing T-ALL. (see *Methods)*

S0 – primary wild-type thymocytes; S1 - *TLX3* oncogene overexpressing cell line; S2 – tumour 7; S3 – tumour 8; S4 – TAP tumour-derived cell line.

Study / Author	Year	Subjects	Samples	Methods	REF
Van Vlierberghe, P. *et al*.	2008	94 pediatric patients (22 *TLX3*+)	Tumour (diagnostic only, BM / PB)	Gene expression array	22
Della Gatta, G. *et al*.	2012	82 pediatric patients (22 *TLX3*+)	Tumour (diagnostic only, BM / PB)	Gene expression array	12
Homminga, I. *et al*.	2011	117 pediatric patients (22 *TLX3*+)	Tumour (diagnostic only, BM / PB) + 7 normal BM	Gene expression array	23
Van Vlierberghe, P. *et al*.	2008	146 pediatric patients (21 *TLX3*+)	Tumour (diagnostic only, BM / PB)	array-CGH	24

Table 1. Studies included for candidate gene selection for transcriptomic data analysis.

BM, bone marrow; PB, peripheral blood.

Gene	Evidence in TLX3+ T-ALL	Candidate Cancer Gene Database[1]		Cancer Gene Census[2]	Hallmark gene[a]	Curated gene[b]
		Haematological malignancies	**Non-haematological malignancies**			
TCRA	Downregulation			yes (fusion)		
CDKN2A	Deletion, majority of T-ALL cases	yes	yes	yes (TSG)	yes	yes
CDKN2B	Deletion, majority of T-ALL cases		yes			
MYB	Copy number gain / overexpression	yes	yes	yes (oncogene, fusion)		
ABL1	Episomal amplification	yes	yes	yes (oncogene, fusion)	yes	yes
JAK3	Copy number gain			yes (oncogene)	yes	yes
WT1	Deletion			yes (TSG, oncogene, fusion)		yes
BCL11B	Deletion		yes	yes (TSG, oncogene, fusion)	yes	
RB1	Deletion	yes	yes	yes (TSG)	yes	yes
NF1	Deletion	yes	yes	yes (TSG, fusion)	yes	yes
PTPN2	Deletion	yes	yes			
PHF6	Deletion		yes	yes (TSG)		yes
SMARCA4	Deletion	yes	yes	yes (TSG)		yes
SUZ12	Deletion	yes	yes	yes (TSG, oncogene, fusion)	yes	
EED	Deletion	yes	yes	yes (TSG)		
IKZF1	Deletion	yes		yes (TSG fusion)		yes
BCOR	Deletion	yes	yes	yes (TSG, fusion)	yes	yes
ABHD12B	Upregulation	yes				
ARMH1	Upregulation	yes				
C1ORF21	Upregulation	yes	yes			
CCND2	Upregulation	yes		yes (oncogene, fusion)	yes	
DNM1	Upregulation	yes	yes			
ECPAS	Upregulation	yes	yes			
FLT3	Upregulation	yes	yes	yes (oncogene)	yes	yes
GAS2	Upregulation	yes				
JUP	Upregulation	yes	yes			
LAT2	Upregulation	yes	yes			

LST1	Upregulation	yes	yes			
MSRB3	Upregulation		yes			
NDFIP2	Upregulation		yes			
PLXND1	Upregulation		yes			
PTPN14	Upregulation		yes			
SCGB3A1	Upregulation	yes				
SOCS2	Upregulation		yes			
SPON1	Upregulation	yes				
TASP1	Upregulation	yes				
TRIO	Upregulation	yes	yes			
ALDH1A2	Downregulation	yes				
BACH2	Downregulation	yes	yes			
CAMTA1	Downregulation	yes	yes	yes (TSG, fusion)	yes	
CTCF	Downregulation	yes	yes	yes (TSG)	yes	yes
DLG3	Downregulation		yes			
FHL2	Downregulation		yes			
IL17RA	Downregulation	yes				
ITPKB	Downregulation	yes				
MBNL2	Downregulation	yes	yes			
MFHAS1	Downregulation		yes			
NEGR1	Downregulation		yes			
NSD1	Downregulation	yes	yes	yes		
PCDH10	Downregulation		yes			
PGM2L1	Downregulation	yes	yes			
PLEKHB1	Downregulation	yes	yes			
PLXDC1	Downregulation	yes				
PROM1	Downregulation		yes			
SEMA4A	Downregulation	yes				
SLC38A1	Downregulation	yes	yes			
STXBP2	Downregulation	yes				
TMEM106B	Downregulation		yes			
ZNF423	Downregulation	yes				

Table 2. Selected genes for comparative gene expression profiling analysis.

[1] The Candidate Cancer Gene Database indicates if transposon mutagenesis-based forward genetic screens in mice support a potential role in cancer for a gene/set of genes, detailing information on the specific types of cancer where the experiments were performed.

[2] The COSMIC Cancer Gene Census is a high-confidence evidence-based list of genes causally involved in cancer. [a] Some of the genes included in this list are classified for their role in promoting and/or suppressing the Hallmarks of Cancer, *i.e.*, the fundamental biological proprieties shared by cancer cells proposed by D. Hanahan and R.A. Weinberg (2000, 2011) (adapted). [b] In addition, for some genes, more exhaustive expert curated data is available.

TSG, tumour suppressor gene.

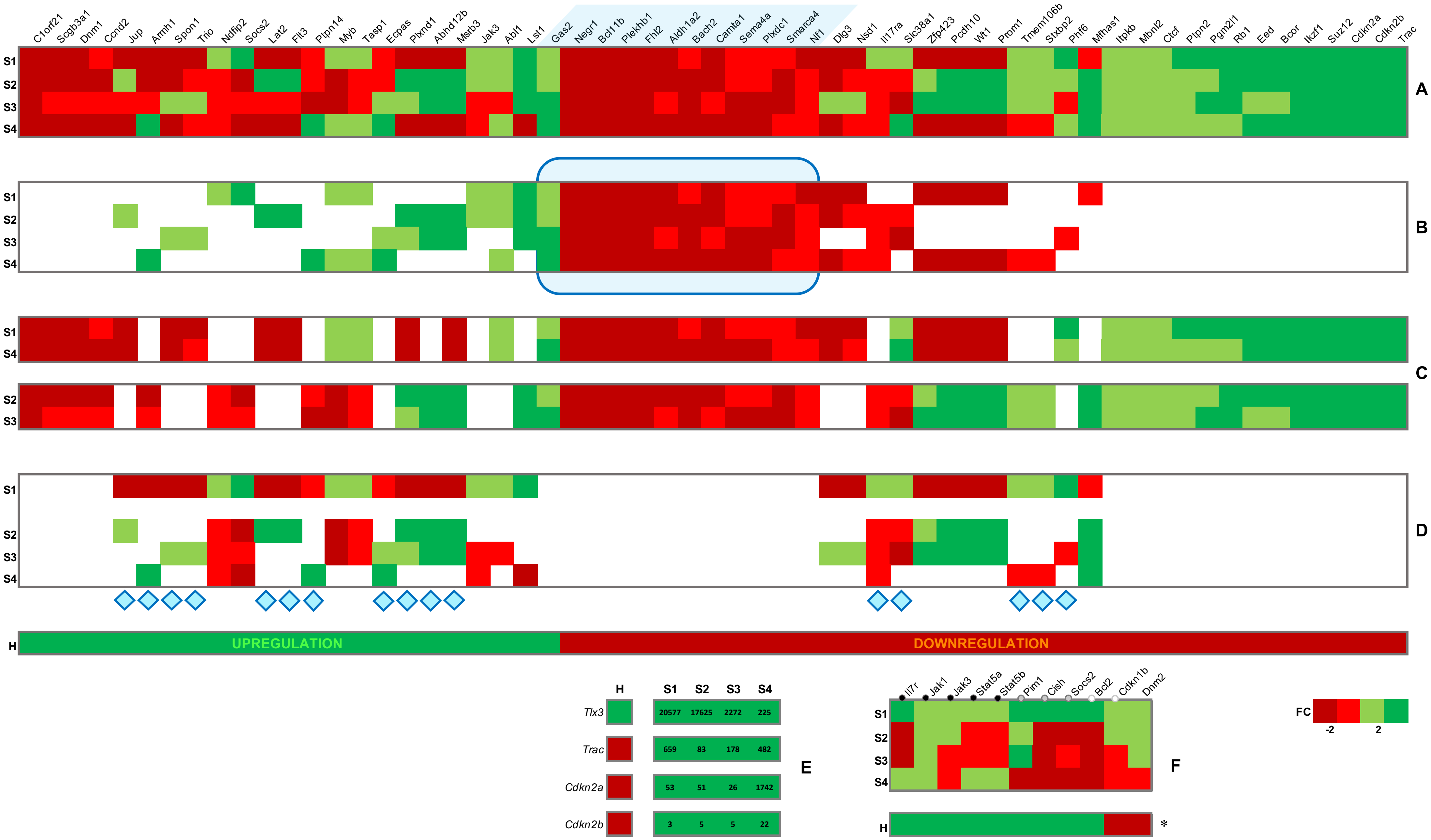

C1orf21
Scgb3a1
Dnm1
Ccnd2
Jup
Armh1
Spon1
Trio
Ndfip2
Socs2
Lat2
Flt3
Ptpn14
Myb
Tasp1
Ecpas
Plxnd1
Abhd12b
Msrb3
Jak3
Abl1
Lst1
Gas2
Negr1
Bcl11b
Plekhb1
Fhl2
Aldh1a2
Bach2
Camta1
Sema4a
Plxdc1
Smarca4
Nf1
Dlg3
Nsd1
Il17ra
Slc38a1
Zfp423
Pcdh10
Wt1
Prom1
Tmem106b
Stxbp2
Phf6
Mfhas1
Itpkb
Mbnl2
Ctcf
Ptpn2
Pgm2l1
Rb1
Eed
Bcor
Ikzf1
Suz12
Cdkn2a
Cdkn2b
Trac
S1
S2
S3
S4
A
S1
S2
S3
S4
B
S1
S4
C
S2
S3
S1
S2
S3
S4
D
H
UPREGULATION
DOWNREGULATION
H
S1
S2
S3
S4
Tlx3
20577
17625
2272
225
Trac
659
83
178
482
Cdkn2a
53
51
26
1742
Cdkn2b
3
5
5
22
E
Il7r
Jak1
Jak3
Stat5a
Stat5b
Pim1
Cish
Socs2
Bcl2
Cdkn1b
Dnm2
S1
S2
S3
S4
F
H
*
FC
-2
2

Figure 2. Comparative transcriptomic analysis between the *TLX3*$^{+}$ T-ALL mouse model (S1-S4) and *TLX3*$^{+}$ T-A LL patients (H).

Transcriptomic data was analysed using DESeq2. Fold-changes (FC) were calculated between each sample S1-S4 and the control S0. The genes selected for the comparison (n=59) are listed in the first row.

Panel A. Differential gene expression profiles of samples S1-S4.

Panel B. Transcriptional overlap between the *TLX3+* T-ALL mouse model and human *TLX3+* T-ALL. The blue box highlights the genes consistently deregulated amongst mouse samples and human data.

Panel C. Transcriptional overlap between unsupervised clustered samples. Pairwise analysis was done between all mouse samples (see Supplementary Figure 2), and these pairs show the greatest number of genes clustered.

Panel D. Differences in gene expression profiles between the pre-leukaemic sample S1 and leukaemic samples S2, S3 and S4 (see Figure 1).

◇ Genes acquired the deregulation phenotype described for human T-ALL.

Panel E. Differential expression levels of *Tlx3*, *Tcra, Cdkn2a* and *Cdkn2b* genes (FC values in bold).

Panel F. Differential expression levels of the IL7R/JAK1/3/STAT5 signalling pathway components (●) and of its main direct (via *Stat5b*) (◉) and indirect (via PI3K/Akt/mTOR crosstalk) (○) transcriptional targets.

The expression levels of the *Dnm2* gene, which encodes for a protein involved in the internalization of the IL-7 receptor, are also shown.

* Transcriptional pattern that could reflect IL7R/JAK1/3/STAT5 pathway activation in T-ALL.

S0 – primary wild-type thymocytes; S1 - *TLX3* oncogene overexpressing cell line; S2 – tumour 7; S3 – tumour 8; S4 – TAP tumour-derived cell line; H - human *TLX3+* T-ALL transcriptional signature.

Study / Author	Year	Subjects	Samples	Methods	REF
Liu, Y. et *al.*	2017	264 pediatric and young adult patients (ages 1-29)	Tumour (diagnostic and matched-remission, BM / PB)	WES and RNA-Seq	5
Vicente, C. *et al.*	2015	155 pediatric (n=111) and adult (n=44) patients (ages 1-66)	Tumour (diagnostic only, BM)	Targeted sequencing of 115 genes recurrently mutated in T-ALL or other hematological malignancies and candidate driver genes	14
Chen, B. *et al.*	2018	130 pediatric (n=69) and adult (n=61) patients (ages 1-62)	Tumour (diagnostic only, BM / PB)	RNA-Seq (validated by WES on 36 patients)	6
Seki, M. *et al.*	2017	121 pediatric patients (ages 1-15)	Tumour (diagnostic only, BM / PB / LN / others)	Targeted sequencing of 158 genes recurrently mutated in ALL	25
Li, Y. *et al.*	2016	69 pediatric patients (ages 1-18)	Tumour (diagnostic and matched-remission, BM / PB)	Targeted sequencing of 254 genes (144 identified by WGS in a cohort of 13 patients + 110 recurrently mutated in leukemia)	15
De Keersmaecker, K. *et al.*	2013	67 pediatric and adult patients (ages 2-72)	Tumour (diagnostic and matched-remission, non-specified)	WES	4
Kalender Atak, Z. *et al.*	2013	31 pediatric (n=11) and adult (n=20) patients (ages 1-72)	Tumour (diagnostic only, non-specified)	RNA-Seq	26

Table 3. Studies included for candidate gene selection for whole-exome sequencing data analysis.

BM, bone marrow; PB, peripheral blood; LN, lymph nodes.

WES, whole-exome sequencing; RNA-seq, RNA sequencing; WGS, whole-genome sequencing.

Gene	**Evidence in TLX3+ T-ALL**	**Candidate Cancer Gene Database[1]**		**Cancer Gene Census[2]**	**Hallmark gene[a]**	**Curated gene[b]**
		Haematological malignancies	**Non-haematological malignancies**			
NOTCH1	Mutated	yes	yes	yes (oncogene, TSG, fusion)	yes	yes
FBXW7	Mutated	yes	yes	yes (TSG)	yes	yes
RB1	Mutated	yes	yes	Yes (TSG)	yes	yes
CCND3	Mutated	yes		yes (oncogene, fusion)		
BCL11B	Mutated		yes	yes (TSG, oncogene, fusion)	yes	
WT1	Mutated			yes (TSG, oncogene, fusion)		yes
RUNX1	Mutated	yes	yes	yes (TSG, oncogene, fusion)		yes
MYB	Mutated	yes	yes	yes (oncogene, fusion)		
IKZF1	Mutated	yes	yes	yes (TSG fusion)		yes
NF1	Mutated	yes	yes	yes (TSG, fusion)	yes	yes
NRAS	Mutated	yes	yes	yes (oncogene)	yes	yes
KRAS	Mutated	yes	yes	yes (oncogene)	yes	yes
IL7R	Mutated			yes (oncogene)		yes
JAK1	Mutated	yes	yes	yes (oncogene, TSG)	yes	yes

JAK3	Mutated			yes (oncogene)	yes	yes
STAT5B	Mutated	yes	yes	yes (oncogene, TSG, fusion)		yes
DNM2	Mutated		yes	yes (TSG)	yes	yes
FLT3	Mutated	yes	yes	yes (oncogene)	yes	yes
ADGRL2	Mutated		yes			
GNB1	Mutated	yes	yes			
PHF6	Mutated		yes	yes (TSG)		yes
SMARCA4	Mutated	yes	yes	yes (TSG)		yes
CTCF	Mutated	yes	yes	yes (TSG)	yes	yes
KDM6A	Mutated	yes	yes	yes (TSG, oncogene)		yes
KMT2A	Mutated	yes	yes	yes (oncogene, fusion)		
KMT2C	Mutated	yes	yes	yes (TSG)	yes	yes
KMT2E	Mutated	yes	yes			
SUZ12	Mutated	yes	yes	yes (TSG, oncogene, fusion)	yes	
EZH2	Mutated	yes	yes	yes (TSG, oncogene)		yes
EED	Mutated	yes	yes	yes (TSG)		
EP300	Mutated	yes	yes	yes (TSG, fusion)	yes	yes
BCOR	Mutated	yes	yes	yes (TSG, fusion)	yes	yes
DNMT3A	Mutated		yes	yes (TSG)		yes
ASXL2	Mutated	yes	yes	yes (TSG)		
RBBP4	Mutated	yes	yes			
RPL5	Mutated	yes	yes	yes (TSG)		yes
USP9X	Mutated		yes			

Table 4. Selected genes for targeted mutational analysis.

[1] The Candidate Cancer Gene Database indicates if transposon mutagenesis-based forward genetic screens in mice support a potential role in cancer for a gene/set of genes, detailing information on the specific types of cancer where the experiments were performed.

[2] The COSMIC Cancer Gene Census is a high-confidence evidence-based list of genes causally involved in cancer. [a] Some of the genes included in this list are classified for their role in promoting and/or suppressing the Hallmarks of Cancer, i.e., the fundamental biological proprieties shared by cancer cells proposed by D. Hanahan and R.A. Weinberg (2000, 2011) (adapted). [b] In addition, for some genes, more exhaustive expert curated data is available.

TSG, tumour suppressor gene.

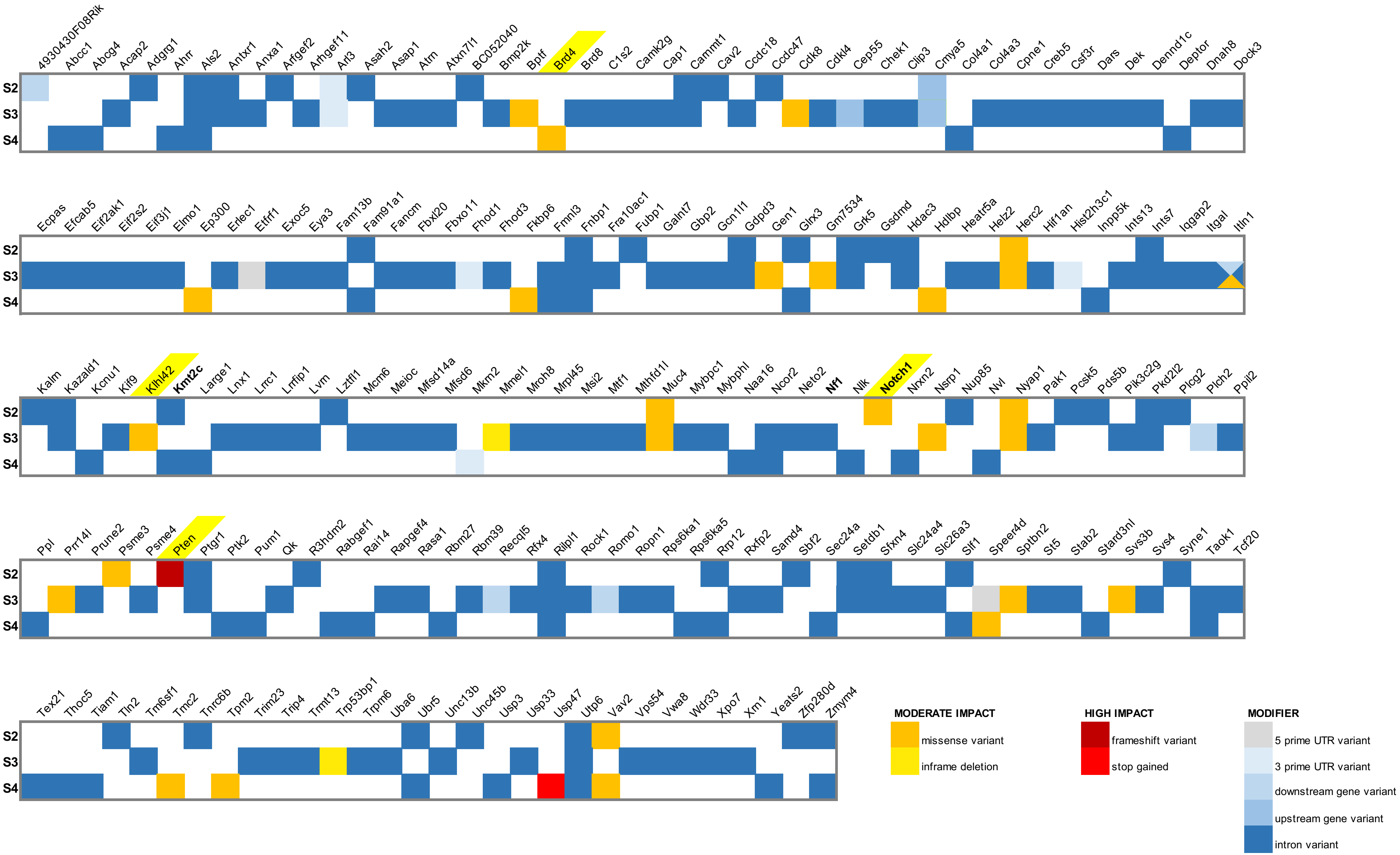

4930430F08Rik
Abcc1
Abcg4
Acap2
Adgrg1
Ahrr
Als2
Antxr1
Anxa1
Arfgef2
Arhgef11
Arl3
Asah2
Asap1
Atm
Atxn7l1
BC052040
Bmp2k
Bptf
Brd4
Brd8
C1s2
Camk2g
Cap1
Cammt1
Cav2
Ccdc18
Ccdc47
Cdk8
Cdkl4
Cep55
Chek1
Clip3
Cmya5
Col4a1
Col4a3
Cpne1
Creb5
Csf3r
Dars
Dek
Dennd1c
Deptor
Dnah8
Dock3
Ecpas
Efcab5
Eif2ak1
Eif2s2
Eif3j1
Elmo1
Ep300
Erlec1
Etfrf1
Exoc5
Eya3
Fam13b
Fam91a1
Fancm
Fbxl20
Fbxo11
Fhod1
Fhod3
Fkbp6
Fmnl3
Fnbp1
Fra10ac1
Fubp1
Galnt7
Gbp2
Gcn1l1
Gdpd3
Gen1
Glrx3
Gm7534
Grk5
Gsdmd
Hdac3
Hdlbp
Heatr5a
Helz2
Herc2
Hif1an
Hist2h3c1
Inpp5k
Ints13
Ints7
Iqgap2
Itgal
Itln1
Kalrn
Kazald1
Kcnu1
Kif9
Klhl42
Kmt2c
Large1
Lnx1
Lrrc1
Lrrfip1
Lvrn
Lztfl1
Mcm6
Meioc
Mfsd14a
Mfsd6
Mkrn2
Mmel1
Mroh8
Mrpl45
Msi2
Mtf1
Mthfd1l
Muc4
Mybpc1
Mybphl
Naa16
Ncor2
Neto2
Nf1
Nlk
Notch1
Nrxn2
Nsrp1
Nup85
Nvl
Nyap1
Pak1
Pcsk5
Pds5b
Pik3c2g
Pkd2l2
Plcg2
Plch2
Ppil2
Ppl
Prr14l
Prune2
Psme3
Psme4
Pten
Ptgr1
Ptk2
Pum1
Qk
R3hdm2
Rabgef1
Rai14
Rapgef4
Rasa1
Rbm27
Rbm39
Recql5
Rfx4
Rilpl1
Rock1
Romo1
Ropn1
Rps6ka1
Rps6ka5
Rrp12
Rxfp2
Samd4
Sbf2
Sec24a
Setdb1
Sfxn4
Slc24a4
Slc26a3
Slf1
Speer4d
Sptbn2
St5
Stab2
Stard3nl
Svs3b
Svs4
Syne1
Taok1
Tcf20
Tex21
Thoc5
Tiam1
Tln2
Tm6sf1
Tmc2
Tnrc6b
Tpm2
Trim23
Trip4
Trmt13
Trp53bp1
Trpm6
Uba6
Ubr5
Unc13b
Unc45b
Usp3
Usp33
Usp47
Utp6
Vav2
Vps54
Vwa8
Wdr33
Xpo7
Xrn1
Yeats2
Zfp280d
Zmym4
S2
S3
S4
MODERATE IMPACT
missense variant
inframe deletion
HIGH IMPACT
frameshift variant
stop gained
MODIFIER
5 prime UTR variant
3 prime UTR variant
downstream gene variant
upstream gene variant
intron variant

Figure 3. Single Nucleotide Variants (SNVs) affecting candidate blood cancer genes in leukaemic samples (S2-S4).

Strelka2 was used for whole-exome variant calling from samples S2-S4, using S1 as reference.

The figure shows the 210 genes (listed in the upper rows) identified in the Candidate Cancer Gene Database (CCGD) as potentially involved in haematological malignancies. The 269 non-synonymous SNVs occurring in these genes are rated for functional impact into three categories (moderate, high, modifier), according to the Ensembl Variant Effect Predictor tool, each comprising different calculated consequences, described in terms defined by the Sequence Ontology. Four genes - *Notch1, Brd4, Klhl42* and *Pten* (in yellow) harbour variants already reported for patients suffering from different types of malignancies (see Table 5 for detailed description).

S1 - *TLX3* oncogene overexpressing cell line; S2 – tumour 7; S3 – tumour 8; S4 – TAP tumour-derived cell line

Gene	Sample	Mouse Genome Coordinate	Human Genome Coordinate	Wild-type allele	Mutant allele	Molecular consequence	Impact	COSMIC / ClinVar reference
Pten	S2	chr19:32811748:1	chr10:87952170:1	T	TA	Frameshift	High	**COSM921111** **c.545_546insA/ p.N184fs*6**
Notch1	S2	chr2:26466601:1	chr9:136503316:1	A	G	Missense	Moderate	**COSM13048** **c.5033T>C / p.L1678P**
Klhl42	S3	chr6:147107848:1	chr12:27797868:1	G	A	Missense	Moderate	**COSM300077** **c.1220G>A / p.R407H**
Brd4	S4	chr17:32212991:1	chr19:15255449:1	C	T	Missense	Moderate	**COSM4666105** **c.1895G>A/ p.R632H**
Usp47	S4	chr7:112074547:1	chr11:11920231:1	A	T	Stop gained	High	n/a
Nyap1	S2, S3	chr5:137734908:1	chr7:100489595:-1	G	T	Missense	Moderate	n/a
Vav2	S2	chr2:27288673:1	chr9:133791850:1	T	C	Missense	Moderate	n/a
Psme3	S2	chr11:101321843:1	chr17:42841549:1	C	T	Missense	Moderate	n/a
Bptf	S3	chr11:107078605:1	chr17:67903833:-1	A	C	Missense	Moderate	n/a
Bptf	S3	chr11:107078606:1	chr17:67903832:-1	C	A	Missense	Moderate	n/a
Bptf	S3	chr11:107078607:1	chr17:67903831:-1	T	G	Missense	Moderate	n/a
Nsrp1	S3	chr11:77049348:1	chr17:30179222:-1	T	C	Missense	Moderate	n/a
Gen1	S3	chr12:11241642:1	chr2:17781573:-1	C	G	Missense	Moderate	n/a
Sptbn2	S3	chr19:4750632:1	chr11:66687016:-1	C	T	Missense	Moderate	n/a
Gm7534	S3	chr4:134202685:1	chr1:26210192:-1	A	T	Missense	Moderate	n/a
Cdk8	S3	chr5:146299800:1	chr13:26401574:1	G	C	Missense	Moderate	n/a
Hdlbp	S4	chr1:93408422:1	chr2:241229886:1	A	G	Missense	Moderate	n/a
Tpm2	S4	chr4:43523335:1	chr9:35689798:1	T	C	Missense	Moderate	n/a
Ep300	S4	chr15:81649350:1	chr22:41177321:1	G	C	Missense	Moderate	n/a
Vav2	S4	chr2:27288673:1	chr9:133791850:1	T	C	Missense	Moderate	n/a

Fkbp6	S4	chr5:135344683:1	chr7:73331742:-1	T	C	Missense	Moderate	n/a
Herc2	S2, S3	chr7:56132560:1	Different nucleotide	G	A	Missense	Moderate	n/a
Mmel1	S3	chr4:154871716:1	Different nucleotide	GTGC	G	Deletion	Moderate	n/a
Prr14l	S3	chr5:32844355:1	Different nucleotide	G	A	Missense	Moderate	n/a
Tmc2	S4	chr2:130202185:1	Different nucleotide	G	T	Missense	Moderate	n/a
Muc4	S2	chr16:32753678:1	No alignment	C	G	Missense	Moderate	n/a
Muc4	S3	chr16:32752305:1	No alignment	C	G	Missense	Moderate	n/a
Muc4	S3	chr16:32752369:1	No alignment	T	G	Missense	Moderate	n/a
Muc4	S3	chr16:32753285:1	No alignment	C	G	Missense	Moderate	n/a
Muc4	S3	chr16:32753841:1	No alignment	G	A	Missense	Moderate	n/a
Muc4	S3	chr16:32753958:1	No alignment	T	A	Missense	Moderate	n/a
Muc4	S3	chr16:32754331:1	No alignment	G	T	Missense	Moderate	n/a
Muc4	S3	chr16:32754437:1	No alignment	C	A	Missense	Moderate	n/a
Muc4	S3	chr16:32755039:1	No alignment	G	T	Missense	Moderate	n/a
Itln1	S3	chr1:171530587:1	No alignment	T	G	Missense	Moderate	n/a
Trp53bp1	S3	chr2:121259195:1	No alignment	TCTC	T	Deletion	Moderate	n/a
Svs3b	S3	chr2:164256155:1	No alignment	T	C	Missense	Moderate	n/a
Speer4d	S4	chr5:15620430:1	No alignment	A	G	Missense	Moderate	n/a

Table 5. Pairwise genomic alignment between mouse and human for the 38 protein-altering SNVs found in genes potentially involved in haematological malignancies.

Comparative genomic alignments were obtained in the Ensembl database, using the mouse and human genome assembly version GRCh38. When sequence alignment was detected at the nucleotide level, the COSMIC Genome Browser and the NCBI ClinVar were used to search for reported mutations in human cancer. Four variants were found to have been previously described, in *Pten, Notch1*, *Klhl42* and *Brd4* (the first four genes listed in the table). For each of these genes, specific information on the corresponding variant (COSMIC ID and annotation at the coding DNA and protein aminoacidic sequence levels) can be found in the last column of the table.

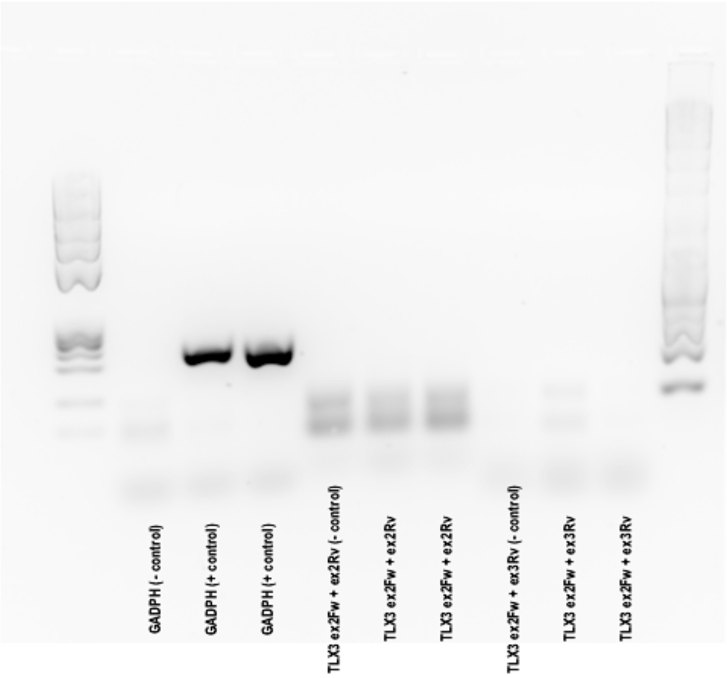
SUPPLEMENTARY MATERIAL

Supplementary Figure 1. RT-PCR *TLX3* primers (A) and *TLX3* RT-PCR results from S4 cells (B).

A

Primer	Sequence (5'-3')
TLX3 ex2Fw	GCCACCCAAGCGTAAGAAG
TLX3 ex2Rv	CACTTGGTCCTCCGATTTTG
TLX3 ex3Rv	GTGGCTCACACCAGAGAGGT

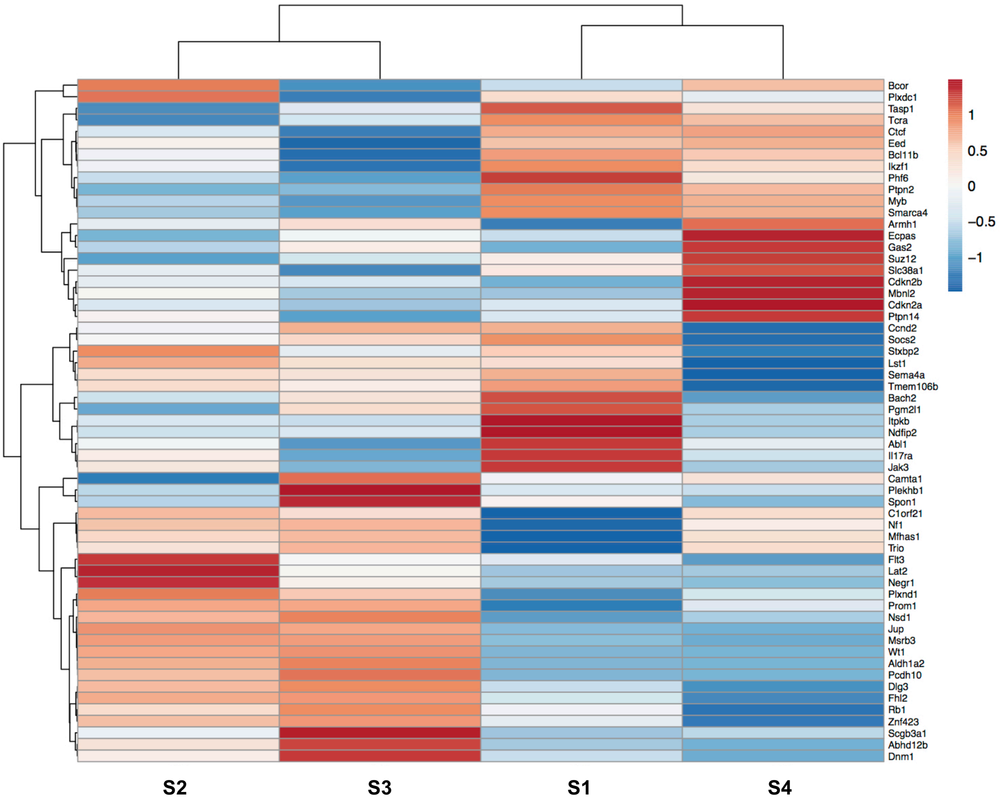
B



Supplementary Figure 2. Unsupervised hierarchical clustering heatmap of samples S1-S4.

Log2-transformed fold-change values were plotted as a heatmap using the ClustVis web tool (https://biit.cs.ut.ee/clustvis/).

S1 - *TLX3* oncogene overexpressing cell line; S2 – tumour 7; S3 – tumour 8; S4 – TAP tumour-derived cell line.



Gene	**1** (Ref. 22) (*TLX3*+ vs non-*TLX3*)	**2** (Ref. 12) (*TLX3*+ vs non-*TLX* and *TLX3*+ vs *TLX1*)	**3** (Ref. 23) (*TLX3*+ vs non-*TLX3*)	**4** (Ref. 24) (29 *TLX3*+)
NDFIP2	Upregulation	Upregulation	Upregulation	
CCND2	Upregulation	Upregulation	Upregulation	
OBSL1	Upregulation	Upregulation	Upregulation	
SIGLEC17P	Upregulation	Upregulation	Upregulation	
SNAI2	Upregulation	Upregulation	Upregulation	
CFH	Upregulation	Upregulation	Upregulation *	
TRIO	Upregulation		Upregulation	
ECPAS	Upregulation		Upregulation	
TASP1	Upregulation		Upregulation	
FLT3	Upregulation		Upregulation	
PAM	Upregulation		Upregulation	
TNFRSF4	Upregulation		Upregulation	
DNM1	Upregulation		Upregulation *	
ABHD12B	Upregulation		Upregulation *	
C1orf21	Upregulation		Upregulation *	
TNFRSF18	Upregulation		Upregulation *	
LAT2		Upregulation	Upregulation	
GAS2		Upregulation	Upregulation	
SCGB3A1		Upregulation	Upregulation	
PLXND1		Upregulation	Upregulation	
PTPN14		Upregulation	Upregulation	
MSRB3		Upregulation	Upregulation	
SPON1		Upregulation	Upregulation	
SOCS2		Upregulation	Upregulation	
GBP5		Upregulation	Upregulation	
IRX3		Upregulation	Upregulation	
NCF4		Upregulation	Upregulation	
FAM26F		Upregulation	Upregulation	
ARMH1		Upregulation	-	
JUP		Upregulation	-	
LST1		Upregulation	-	
TRBC2		Upregulation	-	
HES4		Upregulation	-	
MBNL2	Downregulation		Downregulation	
TMEM106B	Downregulation		Downregulation	
IL17RA	Downregulation		Downregulation	
SLC38A1	Downregulation		Downregulation	
ITPKB	Downregulation		Downregulation	
STXBP2	Downregulation		Downregulation	
PGM2L1	Downregulation		Downregulation	
CTHRC1	Downregulation		Downregulation	
CHST12	Downregulation		Downregulation	
LAMP3	Downregulation		Downregulation	
CLEC2B	Downregulation		Downregulation	

SLFN5	Downregulation		Downregulation	
TSPAN5	Downregulation		Downregulation	
CECR1	Downregulation		Downregulation	
LRP10	Downregulation		Downregulation	
FHL2	Downregulation		Downregulation *	
PLEKHB1	Downregulation		Downregulation *	
DLG3	Downregulation		Downregulation *	
SEMA4A	Downregulation		Downregulation *	
MAGED4	Downregulation		Downregulation *	
PLXNA1	Downregulation		Downregulation *	
BACH2	Downregulation *		Downregulation *	
PLXDC1	Downregulation *		Downregulation *	
PROM1	Downregulation		-	
PCDH10		Downregulation	Downregulation	
ALDH1A2		Downregulation	Downregulation	
MFHAS1		Downregulation	Downregulation	
ZNF423		Downregulation	Downregulation	
TCRA		Downregulation	Downregulation	
BEX2		Downregulation	Downregulation	
HPGD		Downregulation	Downregulation	
TMSB15A		Downregulation	Downregulation	
SIX6		Downregulation	Downregulation	
NEGR1		Downregulation	-	
TRD		Downregulation	-	
SCG2		Downregulation	-	
ECRG4		Downregulation	-	
ABTB2	?			
LOC150166	?			
RNF152	?			
NSD1				5/21 24% del(5)(q35)
CAMTA1				4/21 19% del(1)(p36.31)
HES2				4/21 19% del(1)(p36.31)
HES3				4/21 19% del(1)(p36.31)
CTCF				3/21 14% del(16)(q22.1)

Supplementary Table 1. Genes with altered expression in *TLX3*+ T-ALL.

The table summarizes all the results of the studies cited in Table 1. Genes selected for comparative transcriptome analysis are shown in blue.
All the results shown are statistically significant (p < 0.05). Fold-changes are >2 or <-2, for upregulated and downregulated genes, respectively, except when indicated by an asterisk (*).

Supplementary Methods 1. For study number 1, data was available on the significantly differentially expressed probsets between T-ALL subgroups, without explicit information on either the gene was over or underexpressed. Using the NCBI GEO series accession code number GSE10609 provided by the authors, microarray data was analysed using the NCBI GEO2R online tool (https://www.ncbi.nlm.nih.gov/geo/geo2r) Comparison was made between the 22 *TLX3*+ T-ALL patients and the 45 non-*TLX3*+ patients with a clearly identified molecular subgroup. Unclassified patients were not included. By this means, it was possible to characterise the deregulation in the expression of 40 out of the 43 genes pointed by the authors (16 upregulated, 24 downregulated, 3 unknown). Study 2 also analysed the microarray data gathered by study 1, applying slightly different criteria for the comparisons. Study 3 collected microarray data available under the accession number GSE26713 but no specific comparison that suited the purposes of our study was made. Microarray data was thereby analysed, to confirm the results obtained by the two studies previously mentioned. Comparison was made between the 22 *TLX3*+ T-ALL patients and the 55 non-TLX3+ patients included in this dataset, once more rejecting unclassified patients. The results were validated for 60 out of the 70 deregulated genes, with no contradictory results (for the remaining 10 genes no results of differential expression were found at a statistically significant level, p < 0.05).

		1 (Ref. 5)		2 (Ref. 14)		3 (Ref. 6)		4 (Ref. 25)		5 (Ref. 15)	6 (Ref. 4)	7 (Ref. 26)
		TLX3+ (n=46)	All (n=264)	*TLX3*+ (n=28)	All (n=155)	*TLX3*+ (n=18)	All (n=130)	*TLX3*+ (n=18)	All (n=121)	*TLX3*+ (n=11)	*TLX3*+ (n=5)	*TLX3*+ (n=3)
NOTCH1 pathway	***NOTCH1***	39/46 (85%) **	Mut 75%	21/28 (75%) **	Mut 59%	16/18 (89%)	Mut 75%	15/18 (83%) **	Mut 67%	6/11 (54%)	4/5 (80%)	3/3 (100%)
	FBXW7	17/46 (37%) *	Mut 24%	5/28 (18%)	Mut 17%	8/18 (44%)	Mut 32%	5/18 (28%)	Mut 23%	1/11 (9%)		1/3 (33%)
Cell Cycle and Apoptosis	***CDKN2A***		Mut <1%	1/28 (4%) 26/28 (93%) *	Mut 2% CNV(-) 72%	----- 14/18 (78%) **	Mut 5% CNV(-) 58%	----- 12/18 (67%)	Mut 5% CNV(-) 73%			
	CDKN2B	1/46 (2%)	Mut <1%	----- 24/28 (86%) *	Mut <1% CNV(-) 57%	12/18 (67%)	CNV(-) 49%					
	CDKN2C	1/46 (2%)	Mut <1%									
	CCND3	2/46 (4%)	Mut 6%			1/18 (5%)	Mut 6%	1/18 (5%)	Mut 4%			
	CDKN1B	1/46 (2%)	Mut 2%	-----	Mut 1%							
	RB1		Mut <1%	----- 1/28 3%	Mut 1% CNV(-) 1%			2/18 (11%) ** 2/18 (11%) *	Mut 2% CNV(-) 2%			
	STAG2	1/46 (2%)	Mut <1%									
	PAK4					1/18 (5%)	Mut 3%					
Transcription Factors / Regulators	***BCL11B***	7/46 (15%) 3/46 (6%)	Mut 10% CNV(+) 3%	----- 1/28 (3%)	Mut 10% CNV(+) 2%	-----	Mut 8%	-----	Mut 7%	2/11 (18%)		1/3 (33%)
	WT1	11/46 (24%) * 5/46 (11%) *	Mut 9% CNV(-) 4%	10/28 (36%) * 4/28 (14%) *	Mut 12% CNV(-) 5%	3/18 (17%) **	Mut 6%	8/18 (44%) *	Mut 12%	3/11 (27%)	4/5 (80%)	
	RUNX1	2/46 (4%)	Mut 5%	-----	Mut 5%	1/18 (5%)	Mut 5%	-----	Mut 7%			
	MYB	4/46 (9%) 3/46 (6%)	Mut 5% CNV(+) 13%	2/28 (7%) ** 4/28 (14%)	Mut 2% CNV(+) 9%			2/18 (11%) **	Mut 3%			----- 1/3 (33%)
	MYC	1/46 (2%) 6/46 (13%)	Mut <1% CNV(+) 9%					1/18 (5%)	Mut 2%			
	IKZF1	----- 3/46 (6%) **	Mut 2% CNV(-) 2%	1/28 (3%) -----	Mut 1% CNV(-) 1%	-----	Mut 3%	-----	Mut 2%			
	ZNF217	1/46 (2%)	Mut 1%	1/28 (3%)	Mut 3%							
	GLI3	1/46 (2%)	Mut <1%	1/28 (3%)	Mut 3%							
	MED12	1/46 (2%)	Mut 3%			1/18 (5%)	Mut 3%	-----	Mut 4%			
	ETV6	1/46 (2%)	Mut 3%	-----	Mut 3%	-----	Mut 4%	-----	Mut 2%			
	LEF1	----- 2/46 (4%)	Mut 4% CNV(-) 14%	----- -----	Mut 4% CNV(-) 3%			-----	Mut <1%			
	TSPYL2	2/46 (4%)	Mut 2%									
	EWSR1	1/46 (2%)	Mut <1%									
	MGA	2/46 (4%)	Mut 1%									
	ZBTB7A	----- 1/46 (2%)	Mut <1% CNV(-) 2%									
	CIC							-----	Mut 2%			1/3 (33%)

Epigenetics / chromatin remodeling	***PHF6***	20/46 (43%) * 9/46 (20%) *	Mut 19% CNV(-) 5%	6/28 (21%) 1/28 (3%)	Mut 16% CNV(-) 3%	13/18 (72%) *	Mut 24%	14/18 (77%) *	Mut 25%	2/11 (18%)	1/5 (20%)	
	SMARCA4	3/46 (6%) ** 5/46 (11%) *	Mut 3% CNV(-) 3%									
	CTCF	6/46 (13%) * 8/46 (17%) *	Mut 5% CNV(-) 5%	2/28 (7%) 2/28 (7%)	Mut 4% CNV(-) 3%	4/18 (22%) *	Mut 5%	2/18 (11%) **	Mut 3%		1/5 (20%)	
	KDM6A	7/46 (15%) * 2/46 (4%)	Mut 5% CNV(-) 3%	----- -----	Mut 5% CNV(-) <1%	4/18 (22%) *	Mut 4%	2/18 (11%)	Mut 4%			
	KMT2A	1/46 (2%) 1/46 (2%)	Mut 2% CNV(-) 2%			2/18 (11%)	Mut 4%	1/18 (5%)	Mut 2%			
	KMT2C	1/46 (2%)	Mut 2%			2/18 (11%)	Mut 7%	2/18 (11%)	Mut 5%			
	KMT2D	1/46 (2%)	Mut 2%			-----	Mut 5%	-----	Mut 2%			
	KMT2E	2/46 (4%) **	Mut 1%									
	CREBBP	1/46 (2%)	Mut 2%	1/28 (3%)	Mut 5%	-----	Mut 5%	-----	Mut <1%			
	EZH2	1/46 (2%)	Mut 5%	1/28 (3%) 2/28 (7%)	Mut 5% CNV(-) 5%	2/18 (11%)	Mut 8%	1/18 (5%)	Mut 6%			
	SUZ12	5/46 (11%) *	Mut 3%	1/28 (3%) 1/28 (3%)	Mut 5% CNV(-) 1%	2/18 (11%)	Mut 8%	1/18 (5%) 5/18 (28%)**	Mut 2% CNV(-) 6%			
	EED	----- 3/46 (6%) **	Mut 1% CNV(-) 3%	1/28 (3%) 1/28 (3%)	Mut 3% CNV(-) 3%			-----	Mut 2%			1/3 (33%)
	EP300	1/46 (2%)	Mut <1%	2/28 (7%)	Mut 4%	1/18 (5%)	Mut 3%	2/18 (11%) **	Mut 2%			
	DNMT3A	1/46 (2%) 1/46 (2%)	Mut <1% CNV(-) <1%	1/28 (3%)	Mut 3%	1/18 (5%)	Mut 7%					
	BCOR	2/46 (4%) ** 3/46 (6%) *	Mut 1% CNV(-) 2%					1/18 (5%)	Mut 2%			
	BAZ1A	1/46 (2%)	Mut 2%									
	ARID1A	1/46 (2%) 1/46 (2%)	Mut <1% CNV(-) 2%									
	ASXL1	-----	Mut <1%			1/18 (5%)	Mut 6%					
	ASXL2	1/46 (2%) 1/46 (2%)	Mut 2% CNV(-) <1%					3/18 (17%) *	Mut 2%			
	SATB1	2/46 (4%)	Mut 2%									
	RBBP4	2/46 (4%) **	Mut 1%									
	SP140L	1/46 (2%)	Mut 1%									
	UTY	1/46 (2%)	Mut <1%									
	ZMYM3	1/46 (2%)	Mut <1%									
	TET3			1/28 (3%)	Mut 5%							
	CHD4					1/18 (5%)	Mut 6%					
	H3F3A											1/3 (33%)

Jak-STAT pathway	***IL7R***	5/46 (11%)	Mut 7%	2/28 (7%)	Mut 8%	1/18 (5%)	Mut 5%	4/18 (22%) *	Mut 5%			
	JAK1	-----	Mut 3%	5/28 (18%) *	Mut 6%	2/18 (11%)	Mut 10%	1/18 (5%)	Mut 4%			
	JAK3	3/46 (6%)	Mut 8%	6/28 (21%) 2/28 (7%) **	Mut 15% CNV(+) 2%	2/18 (11%)	Mut 14%	5/18 (28%) *	Mut 12%		2/5 (40%)	
	STAT5B	3/46 (6%)	Mut 4%	1/28 (3%)	Mut 2%	3/18 (17%) **	Mut 5%	1/18 (5%)	Mut 2%			1/3 (33%)
	DNM2	15/46 (33%) *	Mut 11%	8/28 (29%) *	Mut 12%	6/18 (33%) *	Mut 8%	7/18 (39%) *	Mut 16%		1/5 (20%)	1/3 (33%)
	PTPN2	5/46 (11%) **	CNV(-) 5%	4/28 (14%) **	CNV(-) 6%							
Ras-MAPK pathway	***NF1***	----- 1/46 (2%)	Mut 1% CNV(-) 2%	1/28 (3%) 1/28 (3%)	Mut 2% CNV(-) 1%			2/18 (11%) ** 4/18 (22%) *	Mut 2% CNV(-) 4%			1/3 (33%)
	NRAS	6/46 (13%)	Mut 8%			2/18 (11%)	Mut 9%	4/18 (22%)	Mut 12%	2/11 (18%)		
	KRAS	3/46 (6%)	Mut 3%	1/28 (3%) 1/28 (3%)	Mut 3% CNV(+) <1%	1/18 (5%)	Mut 8%	-----	Mut 3%	2/11 (18%)		
	BRAF	1/46 (2%)	Mut 1%	-----	Mut 3%			-----	Mut 2%			
	PTPN11	1/46 (2%)	Mut <1%	-----	Mut <1%							
PI3K-Akt	*PTEN*	1/46 (2%) 1/46 (2%)	Mut 14% CNV(-) 9%	----- 1/28 (3%)	Mut 10% CNV(-) 3%	-----	Mut 10%	1/18 (5%)	Mut 12%			
	PIK3R1	-----	Mut 6%			1/18 (5%)	Mut 8%	-----	Mut 4%			
Signaling - Others	***GNB1***	2/46 (4%) **	Mut 1%									
	FLT3	7/46 (15%) *	Mut 5%	1/28 (3%)	Mut 3%					1/11 (9%)		
	ADGRL2	2/46 (4%) **	Mut 1%	-----	Mut 1%						1/5 (20%)	
	KIT	1/46 (2%) 1/46 (2%)	Mut <1% CNV(+) <1%									
	ABL1	3/46 (6%)	CNV(+) 6%			3/18 (17%) *	CNV(+) 4%	2/18 (11%) *	CNV(+) 2%		1/5 (20%)	
	PTCH1	-----	Mut <1%	2/28 (7%) 1/28 (3%)	Mut 3% CNV(-) <1%							
	PTCH2	-----	Mut <1%	1/28 (3%)	Mut 4%							
	SH2B3	-----	Mut 1%	1/28 (3%) 1/28 (3%)	Mut 3% CNV(-) 1%							
	PHIP	1/46 (2%)	Mut 1%									
	PTPRC	1/46 (2%)	Mut 1%	-----	Mut 2%	-----	Mut 4%					
	PTPRD			1/28 (3%)	Mut 3%							
	PTK2B											1/3 (33%)
RNA and protein synthesis	***RPL5***	2/46 (4%)	Mut 2%	1/28 (3%)	Mut 3%							
	RPL10	1/46 (2%)	Mut 6%	----- 1/28 (3%)	Mut 5% CNV(+) 1%			-----	Mut <1%			

	RPL22	1/46 (2%)	Mut <1%	----- 1/28 (3%)	Mut 1% CNV(-) <1%							
	CNOT3	-----	Mut 3%	1/28 (3%) 1/28 (3%)	Mut 4% CNV(+) 1%	-----	Mut 5%	-----	Mut 2%			
	TSR1			2/28 (7%)	Mut 3%							
Protein ubiquitination	***USP9X***	6/46 (13%) *	Mut 3%	3/28 (11%) **	Mut 4%							
	FBXO28	1/46 (2%)	Mut 1%									
	HUWE1	2/46 (4%)	Mut 2%									
Cellular adhesion, motility, cytoskeleton, organelles, transporters, metabolic processes, DNA repair, splicing, immune response, others.	*BRCA2*	1/46 (2%)	Mut <1%									
	LETM1	1/46 (2%)	Mut 1%									
	B2M	1/46 (2%)	Mut <1%									
	U2AF1	1/46 (2%)	Mut 2%									
	ODZ2	1/46 (2%)	Mut 1%	1/28 (3%)	Mut 6%							
	RELN	-----	Mut <1%	1/28 (3%)	Mut 9%							
	NPM1			2/28 (7%)	CNV(-) 4%							
	JAKMIP2			1/28 (3%)	Mut 3%							
	HADHA											1/3 (33%)
	CELSR3					1/18 (5%)	Mut 3%					
	VCP	-----	Mut <1%			1/18 (5%)	Mut 3%					

Supplementary Table 2. Gene mutations and copy-number variations (CNVs) in *TLX3*+ T-ALL.

The table summarizes all the results of the studies cited in Table 3. Genes selected for comparative transcriptome analysis are shown in blue and genes selected for comparative mutational analysis are shown in red. Genes included in both analysis are underlined and depicted in bold font. For studies numbers 1 to 4, given the reasonable cohort size (264, 155, 130 and 121 patients, respectively), relative frequencies of the reported alterations are presented for the whole study population, besides the *TLX3*+ subgroup. For the remaining studies (5 to 7), only the frequencies of events occurring in *TLX3*+ T-ALL patients are shown. Statistically significant associations ($p < 0.05$) between mutations/CNVs and the *TLX3*+ subgroup are indicated by one asterisk (*). Weaker associations ($0.05 > p < 0.1$) are labelled with two asterisks (**).

Mut, mutation (includes missense, nonsense, frameshift and indels); CNV(+), copy-number variation gain; CNV(-), copy-number variation loss

Supplementary Methods 2. Studies 1, 2, 3 and 4: For the selection of genetic alterations enriched in *TLX3*+ T-ALL, the absolute frequencies (not shown) of mutations and copy number variations in each gene in patients from the *TLX3*+ subgroup were compared to the frequencies in non-*TLX3*+ patients, within the same cohort, by means of a two-tailed Fisher's exact test, given the small number of individuals enrolled in each study. Statistically significant association was considered when the p-value was <0.05 and weaker association was set at a p-value <0.1. Genes with mutations or copy number variations that showed likely or possible predominance in the *TLX3*+ subgroup were chosen for the exome and transcriptome analysis, respectively. All studies: Genes that were altered in at least 3 patients across a minimum of two different studies were also selected.