Université de Montréal

Paysage génomique de la
leucémie aiguë lymphoblastique de l'enfant

par

Jean-François Spinella

Programmes de biologie moléculaire
Faculté de médecine

Thèse présentée à la Faculté de médecine
en vue de l'obtention du grade de Doctorat
en biologie moléculaire

Novembre, 2016

RÉSUMÉ

La leucémie aiguë lymphoblastique (LAL) est une maladie complexe à l'étiologie multifactorielle. Elle représente la forme la plus commune de cancer pédiatrique et malgré une augmentation significative du taux de survie des patients, près de 15% d'entre eux ne répondent pas aux traitements classiques et plus de 2/3 subissent les effets du traitement à long terme. Réduire ces chiffres passe par une meilleure compréhension des causes sous-jacentes de la LAL.

À travers l'analyse des données de séquençage de nouvelle génération (SNG) de la cohorte QcALL du CHU Sainte-Justine, je me suis intéressé aux déterminants génomiques contribuant aux différents aspects de la LAL (prédispositions, développement/progression et rechutes). Dans un premier temps, j'ai développé un outil d'analyse (SNooPer) basé sur un algorithme d'apprentissage intégrant les données SNG normales et tumorales des patients, permettant d'identifier les mutations somatiques au sein de données à faible couverture (*low-pass*). Cet outil, couplé aux analyses prédictives *in silico* et aux validations fonctionnelles adéquates, nous a permis de caractériser les événements rares ou récurrents impliqués dans le processus leucémogène.

En analysant les données de LALs pré-B, j'ai pu mettre en évidence une série de mutations *drivers* rares au niveau de gènes (*ACD, DOT1L, HCFC1*) qui n'avaient jamais été associés à la LAL. L'étude fonctionnelle de la mutation identifiée au niveau d'*ACD*, membre du complexe shelterin, a démontré qu'elle conduit à une réduction de l'apoptose et une augmentation de la taille des télomères. Outre l'intérêt de la découverte de ces nouveaux *drivers*, je souhaitais démontrer l'importance des mutations somatiques rares afin d'établir la spécificité interindividuelle, généralement sous-estimée, et d'identifier l'ensemble des fonctions cellulaires impliquées.

Au cours de ces travaux, j'ai également mis en évidence de nouveaux évènements récurrents de la LAL à cellules T (LAL-T), en particulier au niveau de patients présentant un phénotype immature encore mal caractérisé. J'ai démontré l'influence d'une mutation dans le gène codant pour U2AF1, membre de la machinerie d'épissage (*spliceosom*e), sur l'épissage de gènes d'intérêt et ainsi confirmer l'importance du dysfonctionnement de l'épissage dans le développement de la leucémie. J'ai également identifié deux suppresseurs de tumeurs portés par le chromosome X, *MED12* et *USP9X*, qui n'avaient jamais été associés à la LAL-T auparavant et qui représentent un intérêt particulier étant donné le débalancement de l'incidence en fonction du sexe (ratio garçon:fille =1.22).

Enfin, grâce à l'étude longitudinale de patients LAL-B ayant subi une ou plusieurs rechutes, j'ai analysé l'architecture et l'évolution clonales des tumeurs. J'ai ainsi identifié 2 profils évolutifs distincts gouvernant les rechutes précoces et tardives: d'un côté, une dynamique élevée alimentée par un dysfonctionnement des mécanismes de réparation de l'ADN et conduisant à l'émergence rapide de clones mieux adaptés – de l'autre, une dynamique réduite, quasi-inerte, suggérant l'échappement de cellules en dormance épargnée par la chimiothérapie.

De manière générale, cette thèse a permis de contribuer à la caractérisation des déterminants génomiques qui constituent la variabilité inter- et intra-tumorale, participent au processus leucémogène et/ou aux mécanismes de résistance au traitement. Ces nouvelles connaissances contribueront à un raffinement de la stratification des patients et leur prise en charge personnalisée.

Mots-clés: leucémie aiguë lymphoblastique de l'enfant, génomique, *drivers,* mutations rares et récurrentes, mutations somatiques, séquençage de nouvelle génération, algorithme d'apprentissage, rechute, architecture et dynamique clonales, susceptibilité génétique.

ABSTRACT

Acute lymphoblastic leukemia (ALL) is a complex disease with a multi-factorial etiology. It represents the most frequent pediatric cancer and despite a significant increase of survival rate, about 15% of the patients still do not respond to current treatment protocols and over 2/3 of survivors experience long-term treatment related side effects. To reduce these numbers, a better understanding of the underlying causes of ALL is needed.

Through the analysis of next-generation sequencing (NGS) data obtained from the established Quebec cALL (QcALL) cohort of the Sainte-Justine hospital, I have been particularly concerned about the genomic determinants that contribute to different phases of ALL (predispositions, onset/progress and relapses). First, I developed an analysis tool (SNooPer) based on a machine learning algorithm integrating both normal and tumor NGS data of the patient to identify somatic mutations from low-pass sequencing. This tool, combined to *in silico* predictive analysis and to adequate functional validations, allowed us to characterize rare or recurrent events involved in the leukemogenesis process.

Through the analysis of pre-B ALLs, I have been able to identify several rare driver genes which had never been associated to ALL before (*ACD, DOT1L, HCFC1*). The functional study of the identified mutation in ACD, a member of the shelterin complex, showed a concomitant lengthening of the telomeres and decreased apoptosis levels in leukemia cells. Besides the interest aroused by the discovery of these new drivers, I wanted to demonstrate the importance of low-frequency somatic events to establish the generally underestimated interindividual specificity and identify all cellular functions involved.

During this work, I also identified new recurrent driver events in T-cell ALL (T-ALL), particularly among poorly characterized immature T-ALL patients. For example, I demonstrated the impact of a recurrent mutation in U2AF1, member of the spliceosome, on alternative splicing of cancer-relevant genes, further suggesting the importance of aberrant splicing in leukemogenesis. I also identified two new X-linked tumor suppressors, *MED12* and *USP9X,* never associated to T-ALL before and obtained results supporting a potential role for these genes in the male-biased sex ratio observed in T-ALL (ratio male:female =1.22).

Finally, through the longitudinal study of pre-B cALLs who suffered one or multiple relapses, I analyzed the clonal architecture and evolution of the tumors. I identified two distinct evolution patterns governing either early or late relapses: on one hand a highly dynamic pattern, sustained by a defect of DNA repair processes, illustrating the quick emergence of fitter clones - and on the other hand, a quasi-inert evolution pattern suggesting the escape from dormancy of neoplastic stem cells likely spared from initial cytoreductive therapy.

Overall, this thesis contributed to the characterization of genomic determinants that constitute the inter- and intra-tumor variability, participate in leukemogenesis and/or in resistance mechanisms. This new knowledge will contribute to refine patient stratification and treatment.

Keywords: Childhood acute lymphoblastic leukemia, genomic, rare or recurrent driver mutations, somatic mutations, next-generation sequencing, machine learning, relapse, clonal architecture and dynamics, genetic susceptibility.

TABLE DES MATIÈRES

Liste des tableaux

ARTICLE I - SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing

ARTICLE II - Whole-exome sequencing of a rare case of familial childhood acute lymphoblastic leukemia reveals putative predisposing mutations in Fanconi anemia genes

ARTICLE III - A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells

ARTICLE IV - Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations

ARTICLE V - Mutational dynamics of early and late relapsed childhood B-cell acute lymphoblastic leukemia: rapid clonal expansion and long-term dormancy

Liste des figures

ARTICLE I - SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing

ARTICLE II - Whole-exome sequencing of a rare case of familial childhood acute lymphoblastic leukemia reveals putative predisposing mutations in Fanconi anemia genes

ARTICLE III - A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells

ARTICLE IV - Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations

ARTICLE V - Mutational dynamics of early and late relapsed childhood B-cell acute lymphoblastic leukemia: rapid clonal expansion and long-term dormancy

Liste des abréviations

ALL/LAL: acute lymphoblastic leukemia/leucémie aiguë lymphoblastique

AF: allele frequency

AA: amino acid

AnV: annexin V

AGM : aorte-gonade-mésonéphros

AUC: area under the curve

RS domain: arginine-serine domain

ARNi: ARN interférence

L-ASP: L-asparaginase

BMR: background mutation rate

BQV: base quality value

BM: bone marrow

bp/pb: base pair/paire de bases

CPT: camptothecin

CAF: cancer-associated fibroblast

CIS: carcinoma in situ

COSMIC: Catalogue Of Somatic Mutations In Cancer

CSCs : cellules souches cancéreuses

CSH: cellules souches hématopoïétiques

CNS: central nervous system

CSF: cerebrospinal fluid

CLP: common lymphoid progenitor

CMP: common myeloid progenitor

cDNA: complementary DNA

CMH: complexe majeur d'histocompatibilité

CTX: cyclophosphamide

CN: cytogenetically normal

DFCI: Dana Farber Cancer Institute

DC: dendritic cell

DS: Down syndrome

DOX: doxorubicin

ETP-ALL: early T-cell precursor ALL

ECs: endothelial cells

ECs: epithelial cells

EMT: epithelial-mesenchymal transition

EA: European American

ExAC: Exome Aggregation Consortium

ESP: Exome Sequencing Project

ECM: extracellular matrix

FDR: false discovery rate

FP: false positive

FA: Fanconi anemia

FISH: fluorescent in situ hybridization

FI: frameshift indel

FAB classification: French-American-British classification

GWAS: genome-wide association study

GM-CSF: granulocyte macrophage colony-stimulating factor

G-CSF: granulocyte-colony stimulating factor

GMP: granulocyte/macrophage progenitor

HSC: hematopoietic stem cell

HD domain: heterodimerization domain

HD: hyperdiploid

IG: information gain

TKI: inhibiteur de tyrosine kinase

Indel: insertion and deletion

IGV: Integrative Genomics Viewer

JmjC domain: jumanji C domain

PQL domain: leucine-rich domain

LFS: leukemia free-survival

LT-HSC: long-term hematopoietic stem cell

LOH: loss of heterozygosity

M-CSF: macrophage colony-stimulating factor

MQV: mapping quality value

Mb: megabase

MEP: megakaryocyte/erythroid progenitor

MSC: mesenchymal stem cell

Me: methylation

MAF: minor allele frequency

MS: missense

MAPK: mitogen-activated protein kinase

MDS: myelodysplastic syndrome

MDSC: myeloid-derived suppressor cell

NGS/SNG: next-generation sequencing/séquençage de nouvelle génération

NO: nitric oxide

OB domain: oligonucleotide/oligosaccharide-binding domain

ORF: open reading frame

OPA domain: opposite paired domain

Oob error: out-of-bag error

PB: peripheral blood

PCR: polymerase chain reaction

PR: precision-recall

Pred: prednisone

PT: primary tumor

PI: propidium iodide

qPCR: quantitative PCR

QcALL: Quebec cALL

RF: random Forest

RPKM: reads per kilobase per million mapped reads

ROC curve: receiver operating characteristic curve

RBC: red blood cell

Treg: regulatory T cell

rRNA: ribosomal ribonucleic acid

RNA-seq: RNA-sequencing

ST-HSC: short-term hematopoietic stem cell

SNV: single nucleotide variant

shRNA: small hairpin RNA

SS: splice site

SD: standard deviation

SG: stop gain

TRF assay: telomere restriction fragment assay

TRPs: tetratricopeptide repeat domains

TCGA: The Cancer Genome Atlas

PEST domain: threonine-rich domain

TFBS: transcription factor-bind

TAD domain: transcriptional activation domain

TP: true positive

TME: tumor microenvironment

TSG: tumor suppressor gene

TAM: tumor-associated macrophage

TAN: tumour-associated neutrophil

Ubl: ubiquitin-like module

VAF: variant allele frequency

CNV: copy number variant

VCR: vincristine

WBC: white blood cell count

WES: whole-exome sequencing

WGG: whole-genome genotyping

WGS: whole-genome sequencing

WT: wild-type

À mes parents, Chantal et François

Remerciements

Je souhaite tout d'abord remercier Daniel pour m'avoir fait confiance, même lorsque j'ai voulu me spécialiser dans un domaine qui n'était pas le mien, celui de la bioinformatique. Je le remercie également de m'avoir toujours laissé mener mes recherches avec une grande liberté.

Merci à mes collègues du laboratoire, avec une pensée particulière pour Virginie sans qui l'entrée dans le monde de l'informatique aurait été bien plus complexe. Je remercie Sonia Cellot avec qui ce fut un véritable plaisir de collaborer et de discuter. Un merci particulier à Jasmine pour m'avoir accompagné tout au long de cette thèse, pour ses conseils et son amitié.

Merci à toutes les personnes que j'ai croisées au cours de mes études au Québec et qui sont devenues de véritables amis: Joseph mon premier ami québécois, Ramon qui m'a tant fait rire au cours de nos voyages, Dorival et ses aventures racontées autour d'une bière, Alan et Kate avec qui c'est toujours un plaisir de raviver la rivalité franco-britannique, Diane et nos débats sans fin, Benjamin pour les soirées à discuter et à refaire le monde, Karine ma petite soeur de coeur.

Merci à Pamela (Er Kapish), le rayon de soleil à côté de qui il est impossible d'être de mauvaise humeur.

Je souhaite également remercier Marie-Line Daveran-Mingot, Pascal Le Bourgeois, Michèle Coddeville et Paul Ritzenthaler pour m'avoir fait découvir et apprécier le monde de la recherche.

Merci à mon plus vieil ami Thomas, à Élodie, Ulysse et Heidi que je retrouve dès que je rentre en France et qui agissent comme si je n'étais jamais parti. Merci à Anne pour nos discussions. Merci à Pauline pour son soutien.

Merci aux deux Darwin, l'un pour sa théorie, l'autre pour ses ronronnements.

Merci à Mark Knopfler pour sa musique qui m'a accompagné dans l'écriture de la plupart de ces lignes.

Je remercie les membres du jury pour leur lecture et leurs conseils.

Enfin, un immense merci à mes parents, Chantal et François, qui souffrent de l'éloignement et pour qui le monde de la recherche reste un mystère, mais qui pourtant n'ont jamais cessé de me soutenir.

“*Diplôme: signe de science. Ne prouve rien.*”

- Gustave Flaubert

“*La science est un jeu dont la règle du jeu consiste à trouver quelle est la règle du jeu.*”

- François Cavanna

CHAPITRE 1
Introduction

1.1. Les cancers pédiatriques

D'après l'agence nationale de santé publique (Centre for Chronic Disease Prevention and Control, Public Health Agency of Canada, Ottawa, Ontario, Canada), en moyenne 850 enfants (<15 ans) sont diagnostiqués avec un cancer chaque année au Canada. Les patients âgés de 1 à 14 ans représentent un peu moins d'1% du total des cas identifiés dans le pays mais avec 135 morts par an, le cancer est la cause principale de décès liés à une maladie chez l'enfant (Canadian Cancer Statistics 2008. Toronto (ON): Canadian Cancer Society/National Cancer Institute of Canada; 2010).

Avec 32.5% du total de cancers diagnostiqués, la leucémie est le type de cancer pédiatrique qui présente le plus grand nombre de cas au cours de la période de l'étude (15 ans de 1992 à 2006) [1], suivi du cancer du système nerveux central (19.9%) et des lymphomes (11.2%, Tableau I).

Tableau I. Nombre de cas de cancer pediatrique (0-14 years) et taux d'incidence moyen standardisés sur l'âge par million, par sexe et par groupe d'âge, Canada, 1992-2006. Tableau reproduit de Mitra *et al*., 2012 [1].

	Groupe	Sexe		Groupes d'âge				
		Garçons	Filles	< 1	1–4	5–9	10–14	Total
	Ensemble des cancers combinés							
	Nombre total de cas	7131	6070	1308	4768	3441	3694	13211
	ASIR, par million	160.7	143.4	245.1	213.1	116.2	120.3	152.3
I	**Leucémies**							
	Nombre total de cas	2345	1940	264	2060	1175	788	4287
	ASIR, par million	53.1	46.1	49.5	92.1	39.7	25.7	49.7
II	**Lymphomes**							
	Nombre total de cas	981	499	46	222	426	786	1480
	ASIR, par million	21.4	11.4	8.6	9.9	14.4	25.6	16.5
III	**Sytème nerveux central**							
	Nombre total de cas	1408	1217	142	820	915	754	2631
	ASIR, par million	31.5	28.5	26.6	36.6	30.9	24.6	30
IV	**Neuroblastomes**							
	Nombre total de cas	498	463	334	475	119	32	960
	ASIR, par million	11.8	11.5	62.6	21.2	4	1	11.7
V	**Rétinoblastomes**							
	Nombre total de cas	167	156	108	200	11	3	322
	ASIR, par million	4.0	3.9	20.2	8.9	0.3	<0.1	3.9
VI	**Tumeurs rénales**							
	Nombre total de cas	343	414	99	436	178	44	757
	ASIR, par million	7.9	10	18.6	19.5	6	1.4	9
VII	**Tumeurs hépatiques**							
	Nombre total de cas	120	79	49	109	20	22	200
	ASIR, par million	2.8	2	9.2	4.9	0.7	0.7	2.4
VIII	**Tumeurs osseuses**							
	Nombre total de cas	304	294	5	50	175	368	598
	ASIR, par million	6.5	6.6	0.9	2.2	5.9	12	6.6
IX	**Sarcomes des tissus mous**							
	Nombre total de cas	451	366	84	210	230	292	816
	ASIR, par million	10.0	8.6	15.7	9.4	7.8	9.5	9.3
X	**Tumeurs des cellules germinales**							
	Nombre total de cas	201	237	75	91	57	216	439
	ASIR, par million	4.6	5.5	14.1	4.1	1.9	7	5
XI	**Carcinomes et tumeurs épithéliales**							
	Nombre total de cas	213	295	40	34	96	338	508
	ASIR, par million	4.6	6.7	7.5	1.5	3.2	11	5.6
XII	**Autre**							
	Nombre total de cas	100	110	62	61	39	50	212
	ASIR, par million	2.4	2.7	11.6	2.7	1.3	1.6	2.5

ASIR, *age-standardized incidence rate*.

Le cancer du cerveau a récemment surpassé la leucémie comme première cause de décès liée au cancer [2]. Une amélioration significative des connaissances sur la leucémie au cours des 60 dernières années a permis de mettre en place des traitements conduisant à une diminution du taux de mortalité de 76%, contre 31% seulement en ce qui concerne le cancer du cerveau et autres systèmes nerveux [2] (Figure I).

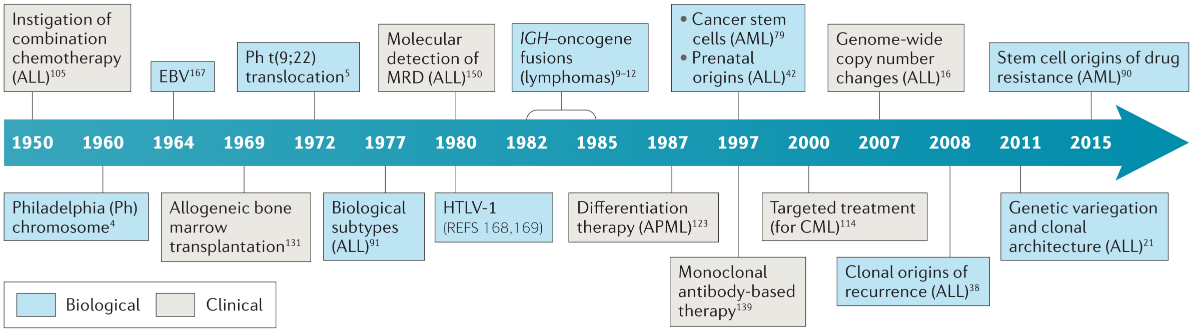

Instigation of combination chemotherapy (ALL)[105]
EBV[167]
Ph t(9;22) translocation[5]
Molecular detection of MRD (ALL)[150]
IGH–oncogene fusions (lymphomas)[9–12]
• Cancer stem cells (AML)[79]
• Prenatal origins (ALL)[42]
Genome-wide copy number changes (ALL)[16]
Stem cell origins of drug resistance (AML)[90]
1950
1960
1964
1969
1972
1977
1980
1982
1985
1987
1997
2000
2007
2008
2011
2015
Philadelphia (Ph) chromosome[4]
Allogeneic bone marrow transplantation[131]
Biological subtypes (ALL)[91]
HTLV-1 (REFS 168,169)
Differentiation therapy (APML)[123]
Targeted treatment (for CML)[114]
Genetic variegation and clonal architecture (ALL)[21]
Monoclonal antibody-based therapy[139]
Clonal origins of recurrence (ALL)[38]
Biological
Clinical

Figure I. Chronologie de la recherche sur les cancers hématologiques. Principales découvertes biologiques (bleu) et cliniques (gris) de 1950 à nos jours conduisant à l'amélioration des connaissances sur la leucémie et les lymphomes et une meilleure prise en charge des patients. ALL, *acute lymphoblastic leukemia*; AML, *acute myeloid leukemia*; APML, *acute promyelocytic leukemia*; CML, *chronic myeloid leukemia*; EBV, *Epstein-Barr virus*; HTLV-1, *human T-lymphotropic virus 1*; IGH, *immunoglobulin heavy chain locus*; MRD, *minimal residual disease*; Ph, *Philadelphia chromosome*. Figure adaptée de Greaves, 2016 [3].

Ces recherches ont également permis de mieux caractériser les cancers pédiatriques par rapport à ceux retrouvés chez l'adulte. Outre une agressivité et une capacité d'invasion accrues, plusieurs points les distinguent. Si certains cancers relativement fréquents chez l'enfant comme les médulloblastomes ou les neuroblastomes sont extrêmement rares chez l'adulte [4], au sein même d'un type histologique particulier, adultes et enfants présentent également des sous-types moléculaires et un spectre mutationnel divergents. Par exemple, la translocation t(12;21) (*ETV6-RUNX1/TEL-AML-1*) est retrouvée chez 21% des LAL de l’enfant et y représente l'altération la plus fréquente alors que chez l'adulte, c'est la translocation t(9;22) (*BCR-ABL1*) qui est la plus souvent identifiée (27% des cas) [4]. Ces spécificités s'expliquent en partie par l'origine *in utero* de nombreux cancers pédiatriques qui se développent au sein de tissus en développement présentant une activité cellulaire importante impliquant division rapide et maturation.

1.2. Hématopoïèse et leucémogenèse

L'hématopoïèse représente l'ensemble des processus permettant la production et la maturation des cellules sanguines à partir de cellules souches hématopoïétiques (CSH). Au cours du développement, l'hématopoïèse prend place au sein d'une succession de tissus incluant le sac vitellin, la région AGM (aorte-gonade-mésonéphros), le placenta et le foie fœtal (Figure II) [5]. Après la naissance, l'hématopoïèse prend place au niveau de la moelle osseuse où l'exposition des cellules totipotentes à divers facteurs de croissance ou de transcription permet le maintien de leur capacité d'auto-renouvèlement (ex: RUNX1, LMO2, MLL et GATA2 [5,6]), la génération de progéniteurs multipotents à partir des CSH (ex: GATA1, E2A et EBF [5,6]) et leur différenciation en précurseurs lignée-spécifiques (ex: EKLF, FOG1 et GFI1 [5,6]) qui donneront les cellules sanguines matures de la lignée lymphoïde (lymphocytes B, T et *natural killer*) et myéloïde (granulocytes, monocytes, plaquettes, globules rouges et cellules dendritiques) (Figure II) [5-7].

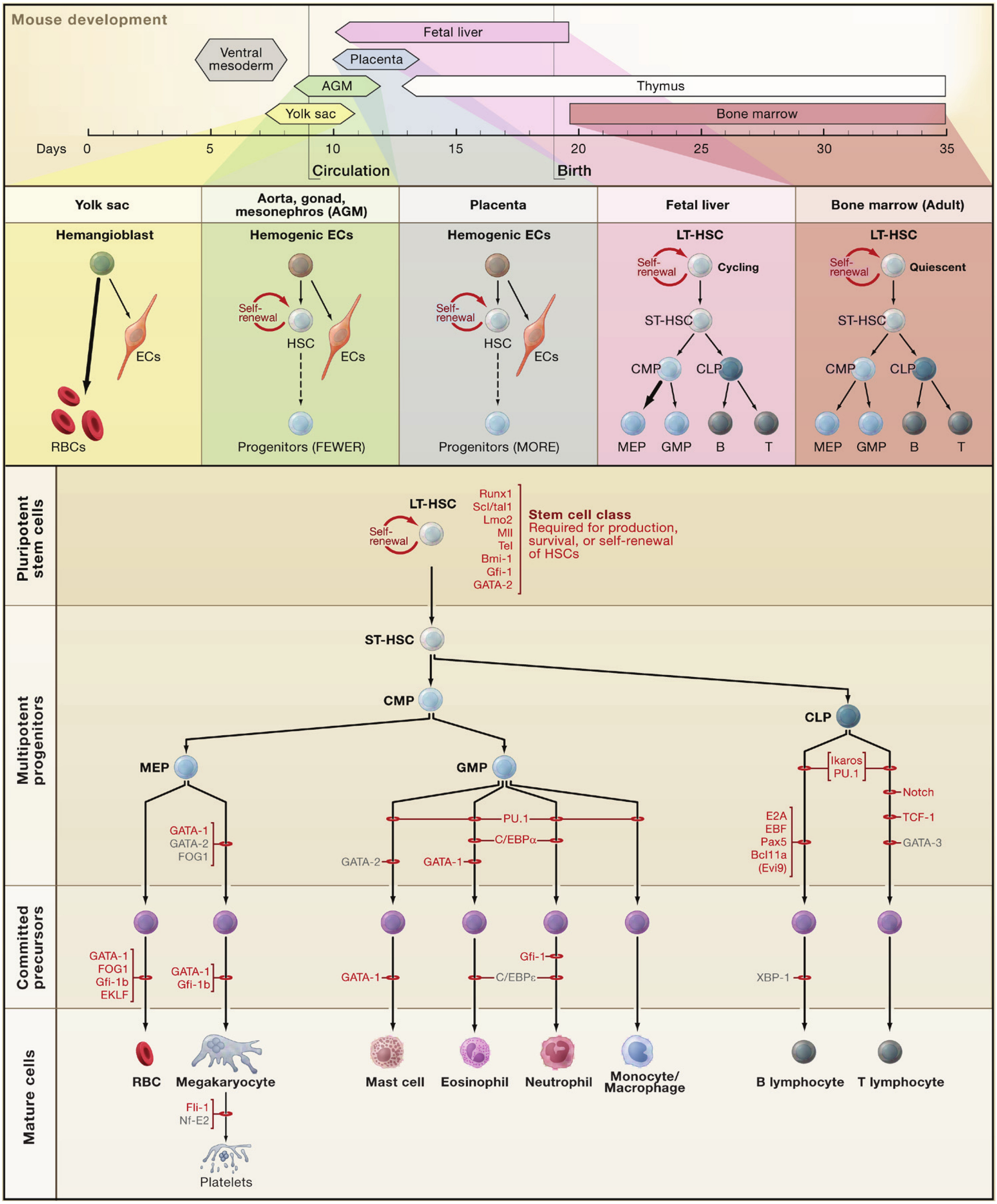


Figure II. Illustration de l'hématopoïèse. Le panneau supérieur illustre les étapes et les tissus de l'hématopoïèse au cours du développement de la souris. Le panneau inférieur décrit les étapes de différenciation conduisant à la génération des cellules sanguines matures chez les mammifères ainsi que les facteurs de transcription gouvernant ces étapes. ECs, *endothelial cells*; RBCs, *red blood cells*; LT-HSC, *long-term hematopoietic stem cell*; ST-

HSC, *short-term hematopoietic stem cell*; CMP, *common myeloid progenitor*; CLP, *common lymphoid progenitor*; MEP, *megakaryocyte/erythroid progenitor*; GMP, *granulocyte/macrophage progenitor*. Figure reproduite d'Orkin and Zon, 2008 [5].

La leucémogenèse est le résultat, entre autre, d'une acquisition d'altérations génétiques (translocations, variations du nombre de copie d'un gène (CNV), petites insertions ou délétions (indels) ou encore mutations ponctuelles (SNV)) et épigénétiques au niveau d'une cellule hématopoïétique normale non mature. Ces modifications transforment une CSH ou cellule progénitrice multipotente en cellule néoplasique capable d'auto-renouvèlement à l'origine de la prolifération clonale maligne [8]. Les évènements génétiques ou épigénétiques contribuant à la transformation cellulaire, à la progression de la maladie, ou de manière générale conférant un avantage sélectif aux cellules mutées par rapport aux cellules normales ou aux autres cellules tumorales, seront appelés évènements "*drivers*". Ces mutations *drivers* sont à distinguer de celles appelées "*passengers*" qui n'ont pas participé à l'émergence de la tumeur et qui ne confèrent (ou n'ont conféré) aucun avantage à la cellule mutée au cours du développement tumoral [9]. Notez que le fait d'écarter une variation à l'issue d'analyses *in vitro* ou *in silico*, faute d'effet (ou d'effet mesurable) dans un contexte expérimental qui ne reproduit à l'identique ni le contexte ni l'historique tumoral, n'exclut en aucun cas que cette variation puisse avoir eu une implication dans l'apparition ou le développement de la tumeur. Faute d'informations suffisantes, il est donc abusif de qualifier ces mutations de *passengers*.

Il existe différents types et sous-types de leucémies de l'enfant selon le lignage cellulaire affecté: la LAL, la leucémie myéloïde aiguë et la leucémie myéloïde chronique représentant respectivement 75-80%, 20-25% et <5% des cas [10]. Ces différentes leucémies partagent les mêmes caractéristiques néoplasiques, tel qu'une résistance accrue à l'apoptose, une augmentation des capacités de prolifération et d'auto-renouvèlement, un blocage de la différenciation, ainsi qu'une instabilité génomique [11] (Figure III).

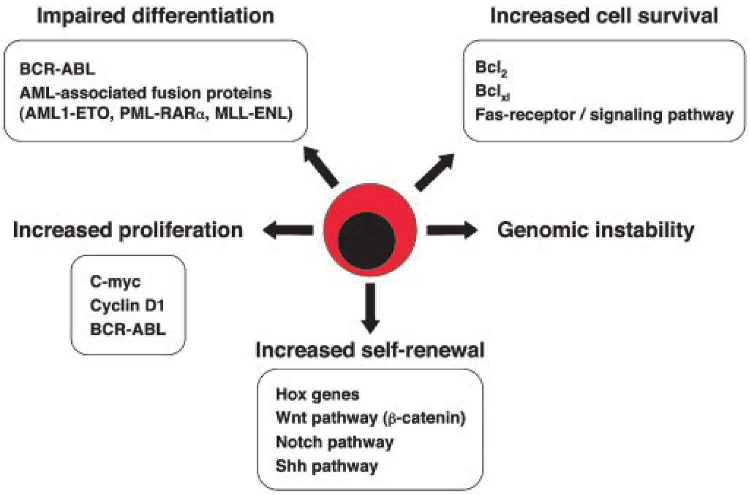


Figure III. Voies de signalisation dérégulées conduisant à la leucémie. Illustration des mécanismes ou voies de signalisations identifiés comme altérés au niveau des différents types de leucémies et permettant l'acquisition des 5 caractéristiques néoplasiques représentées. Figure reproduite de Passegué *et al*., 2003 [11].

1.3. Cellules souches cancéreuses

Les cellules souches cancéreuses (CSCs) sont des cellules aux capacités d'auto-renouvèlement accrues qui possèdent la faculté de propagation clonale à long terme [12]. Si aucun consensus n'est établi en ce qui concerne leurs fréquences et la variabilité de leurs propriétés phénotypiques, la dynamique clonale associée à toute progression tumorale suggère la diversité plutôt qu'une entité fixe [13]. Longtemps considérée comme une population cellulaire distincte à l'origine du maintien, de la croissance et de l'hétérogénéité tumorale (modèle hiérarchique), il semble de plus en plus clair que la plasticité caractérisant toute population de cellules cancéreuses permettrait aussi l'interconversion entre cellules souches et différenciées [14]. Ainsi, en fonction des mutations acquises, des signaux microenvironnementaux et de la pression de sélection, certaines cellules tumorales pourraient se dédifférencier pour venir enrichir le *pool* de CSCs et accroître la variabilité phénotypique [15,16].

L'interaction réciproque qui existe entre la CSC et son microenvironnement, la “niche”, est complexe. Cette dernière stimule le caractère “souche” des CSCs comme l'auto-renouvèlement, la dormance et la plasticité. Elle recrute également des cellules immunitaires qui promeuvent l'invasion des cellules cancéreuses (Figure IV)[16]. La plasticité et la quiescence des CSCs au sein de leurs niches entraînent une résistance à la chimiothérapie. D'ailleurs, la mesure de l'activité des CSCs d'un cancer, par xénotransplantation ou par l'analyse de la signature transcriptomique, corrèle avec l'issue clinique [17]. De plus, la réversibilité de l'état de dormance des CSCs entraîne la possibilité de rechutes jusqu'à plusieurs dizaines d'années après la rémission du patient [18]. Ce sujet est abordé de nouveau dans le chapitre 5 de cette thèse.

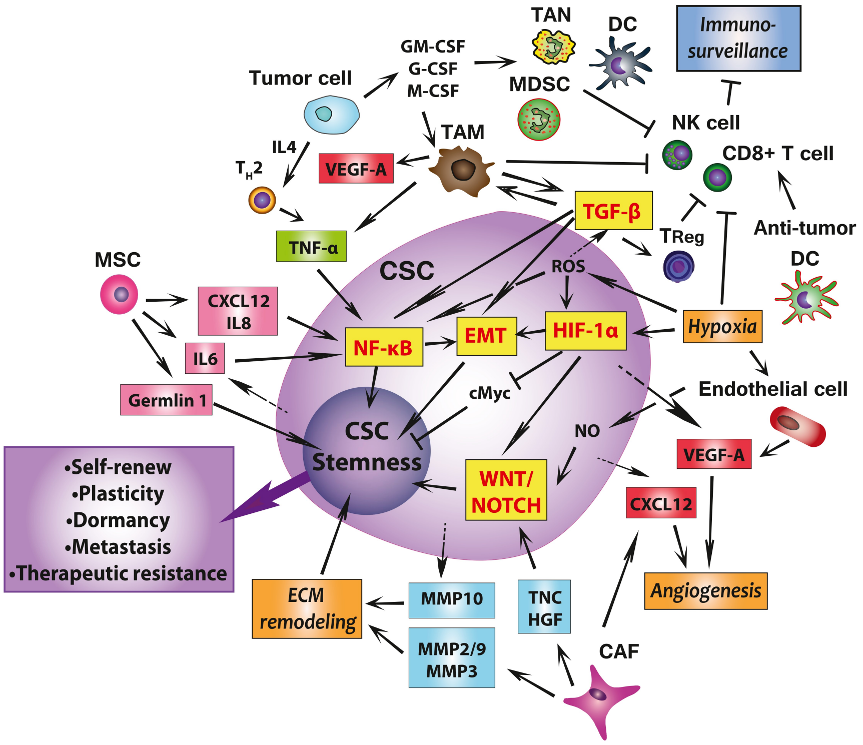


Figure IV. Les bases cellulaires et moléculaires de l'interaction réciproque entre les CSCs et leurs niches. En sécrétant des facteurs comme CXCL12, IL6 et IL8, les MSCs promeuvent le caractère souche des CSCs par stimulation de la voie NF-kB tandis que les CSCs sécrètent de l'IL6 pour attirer plus de MSCs. Les MSCs produisent également l'antagoniste Gremlin 1 qui stimule l'état non différencié. Les cellules tumorales avoisinantes produisent des IL4 pour induire la différenciation des lymphocytes TH2 qui produisent à leur tour du TNFa qui active la voie NF-kB et facilite la mise en place d'un pro-TME. Dans ce contexte, les cellules tumorales produisent les M-CSF, GM-CSF et G-CSF pour induire l'expansion des TAMs, MDSCs, TANs et Dcs. Les TAMs produisent du TNFa et TGF-b pour promouvoir l'EMT et induire la plasticité des CSCs. Le TGFb peut aussi interagir directement avec la voie NF-kB pour promouvoir le caractère souche des cellules cancéreuses. De plus, le TGF-b produit par les TAMs entraîne l'accumulation des cellules Treg. Les TAMs, Treg et l'environnement hypoxique inhibe l'immunosurveillance par l'inhibition des cellules T CD8+, des cellules NK et de la phagocytose des macrophages. Les DCs anti-tumeur nécessaires à la réjection de la tumeur médiée par les cellules T sont maintenues à l'écart de la niche. L'hypoxie augmente les ROS, ce qui promeut la survie cellulaire et induit l'EMT par le biais de la voie TGF-b. Les ROS et l'hypoxie induisent l'expression d'HIF-1a par les CSCs, ce qui promeut directement l'EMT. De plus, l'hypoxie inhibe la prolifération cellulaire en réprimant l'expression de cMyc et renforce le caractère souche des cellules en induisant l'activation de la voie WNT. Les CSCs et les CAFs produisent le facteur CXCL12 qui induit l'angiogenèse. L'hypoxie provoque également la production de VEGF par les CSCs et les ECs, ce qui induit également l'angiogenèse. Les ECs induisent l'auto-renouvellement des CSCs par contact direct et par la production de NO via la voie NOTCH. Les CAFs activent les voies WNT et NOTCH par la production de TNC et HGF ce qui entraîne la maintenance des CSCs. Les CAFs produisent également les MMP2, 3 et 9 qui, avec le MMP19 produit par les CSCs, entraînent la dégradation et le remodelage de l'ECM et favorise l'EMT. CAFs, *cancer-*

*associated fibroblast*s; DCs, *dendritic cell*s; ECs, *epithelial cell*s; ECM, *extracellular matri*x; EMT, *epithelial–mesenchymal transitio*n; G-CSF, *granulocyte-colony stimulating facto*r; GM-CSF, *granulocyte macrophage colony-stimulating facto*r; M-CSF, *macrophage colony-stimulating facto*r; MDSCs, *myeloid-derived suppressor cell*s; MSCs, *mesenchymal stem cell*s; NO, *nitric oxid*e; TAMs, *tumor-associated macrophage*s; TANs, *tumour-associated neutrophil*s; TME, *tumor microenvironmen*t; Treg, *regulatory T cell*s. Figure reproduite de Plaks *et al*., 2015 [16].

1.4. La leucémie aiguë lymphoblastique (LAL) de l'enfant

Mes travaux portent sur la LAL qui est la leucémie la plus fréquemment retrouvée chez les enfants. La LAL peut toucher les lignées lymphocytaires B (LAL-B) et T (LAL-T). La classification de la maladie en différents sous-types de ces deux lignées se fait au moment du diagnostic en se basant sur la morphologie cellulaire (taille de la cellule, chromatine nucléaire, forme du noyau, importance du cytoplasme) et différents marqueurs immunophénotypiques référencés par le système de classification French-American-British (FAB) [19]. La subdivision des deux formes de LAL, B et T, est représentative du niveau de différenciation des lymphoblastes (pro-B/T, pré-B/T)(Figure V).

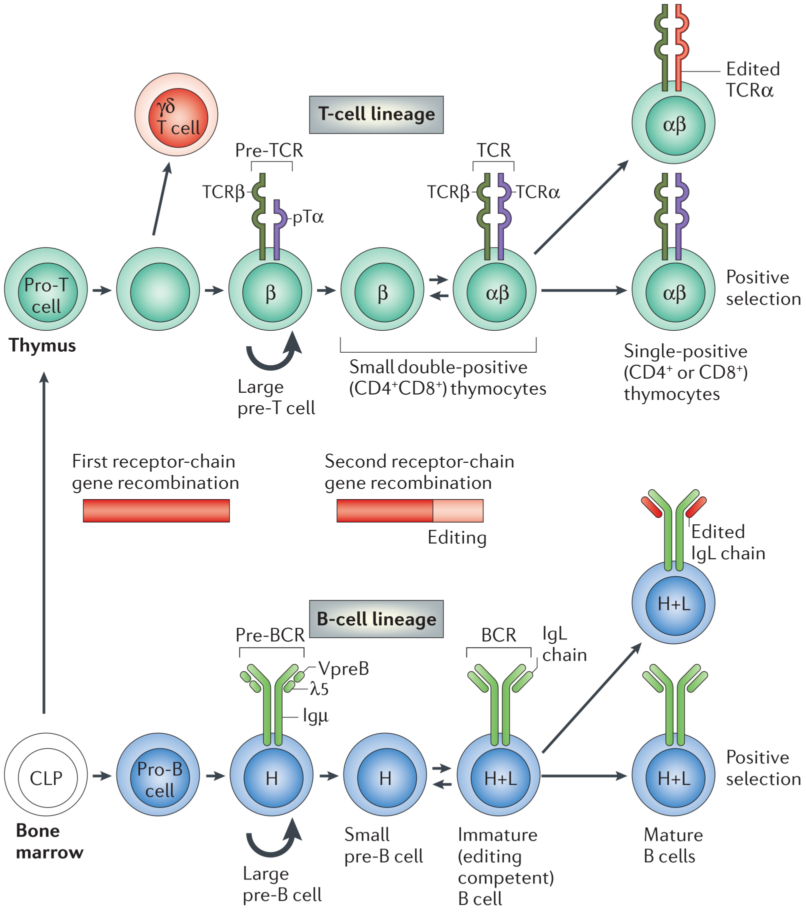


Figure V. Voies de développement des lymphocytes T et B et recombinaisons des gènes codant les récepteurs antigéniques. Le développement des cellules B et T se produit séparément dans la moelle osseuse et le thymus, respectivement. CLP, progéniteurs

lymphoïdes communs. Figure reproduite de Nemazee, 2006 [20].

1.4.1. Épidémiologie

Une large majorité de ces LALs concerne la lignée lymphocytaire B (85%) contre 15% seulement pour la lignée T [21]. Un débalancement de l'incidence est observé en fonction du sexe (ratio garçon:fille =1.22) [22] et du groupe ethnique (aux États-Unis 14.8, 35.6 et 40.9 de cas de LAL chez les enfants et jeunes adultes par million pour la population afro-américaines, d'origine européenne et hispanique, respectivement)[23,24].

1.4.2. Étiologie et facteurs de risque

Bien que plusieurs hypothèses concernant le facteur déclencheur du développement de la LAL ont été émises, aucune n'a été définitivement retenue à ce jour (voir section 1.4.2.4: Environnement et hypothèse virale). Par contre, nous savons que la LAL est la résultante de multiples altérations somatiques acquises successivement par un progéniteur anormal en suivant une chronologie particulière. Ces évènements *drivers* sont divers (réarrangements chromosomiques ou aneuploïdie, SNVs, indels et CNVs), présentent des fréquences variables et perturbent des voies de signalisation multiples. La complexité et la diversité de ces évènements font de la LAL (B ou T) une maladie hétérogène [25].

1.4.2.1. Une origine prénatale

À la fin des années 90, l'analyse rétrospective par RT-PCR de taches de sang (*blood spots)* néonatales (test de Guthrie) a permis de mettre en évidence l'existence de fusion t(12;21) (*ETV6-RUNX1*/*TEL-AML-1*) et t(4;11) (*MLL-AF4*) chez des enfants ayant respectivement développé une LAL-B et une leucémie infantile (diagnostic avant la fin de la 1ère année), confirmant l'apparition précoce des évènements *drivers* mais surtout suggérant leur origine prénatale [26,27].

Par la suite, des travaux portant sur l'étude de jumeaux monozygotes leucémiques ont identifié une concordance (réarrangement génomique clonotypique) concernant 10% des cas âgés de 1 à 15 ans et pouvant atteindre 50% en ce qui concerne les leucémies infantiles [28]. L'analyse de LALs concordantes t(12;21) (*ETV6-RUNX1*/*TEL-AML-1*) a démontré que les lymphoblastes issus des deux patients présentaient à chaque fois une fusion aux points de cassures génomiques identiques, ce qui suggère une migration métastasique intraplacentaire entraînant le partage des clones pré-leucémiques originaires d'un des deux enfants et confirme l'origine *in utero* des premiers événements somatiques contribuant au développement de la maladie [28,29].

1.4.2.2. Un processus multi-étapes

Tel que discuté ci-dessus, des réarrangements chromosomiques montrent une origine *in utero* au cours de l'hématopoïèse fœtale. Toutefois, des expériences menées chez des souris [30-32] ont démontré que ces réarrangements ne sont pas suffisants à la transformation cellulaire conduisant à l'apparition d'une leucémie [29]. Le fait que la fréquence d'identification de la fusion t(12;21) dans le sang de cordon des nouveau-nés excède d'environ 100 fois la prévalence de la LAL-B pédiatrique, vient confirmer ces observations [33]. Bien que la capacité néoplasique soit spécifique à chaque type de translocation [29], l'accumulation d'évènements coopératifs secondaires semble indispensable, d'où la proposition d'un modèle minimal en deux étapes [34] (Figure VI):

- la première étape, *in utero* et commune, impliquerait un événement mutationnel perturbant la maturation cellulaire et entraînant la séquestration du clone pré-leucémique dans les sites principaux de l'hématopoïèse fœtale (ex: placenta, foie). Cette étape, jusqu'à la transformation effective de la cellule, correspond à la phase de latence et peut durer de quelques semaines dans le cas de leucémies infantiles à plusieurs années pour la LAL commune [27,29].

- la deuxième étape, postnatale et rare (donc limitante), met fin à la phase de latence. Elle conduit à la transformation complète et l'émergence de la leucémie par le biais de l'acquisition d'évènements génétiques secondaires (SNVs, indels et CNVs). C'est l'initiation de la leucémogenèse. Les lymphoblastes, bloqués à ce stade précoce de différenciation, envahissent alors la moelle osseuse au dépend de la production du tissu hématopoïétique normal, joignent alors la circulation périphérique et peuvent s'étendre aux ganglions lymphatiques, à la rate, au foie, ainsi qu'au système nerveux central. L'envahissement de ces organes et l'altération de leurs fonctions physiologiques qui en découle correspond à la phase de progression de la LAL.

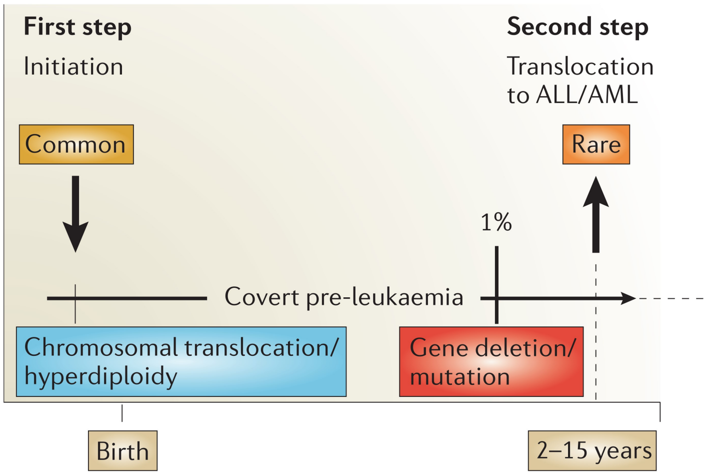


Figure VI. Modèle illustrant les événements séquentiels minimaux nécessaires à l'émergence de la leucémie de l'enfant. Le 1% indiqué représente la fréquence des transformations de “pré-leucémies” (porteuses de réarrangements chromosomiques) en leucémies cliniquement diagnostiquées. Figure reproduite de Greaves *et al.*, 2006 [34].

Si ce modèle est maintenant communément accepté, le nombre d'événement entraînant l'initiation est discuté et probablement variable d'un patient à l'autre. De plus, il est important ici de distinguer la LAL de l'enfant de la leucémie infantile, biologiquement (phénotype pro-B ou monocytaire) et cliniquement différentes. Si l'origine prénatale des deux maladies a été démontrée, le temps de latence très différent implique aussi un mécanisme sous-jacent de transformation différent. Le taux élevé de concordance des leucémies infantiles entre jumeaux (voir section 1.4.2.1: Une origine prénatale), couplé à un temps de latence extrêmement court, suggère que le processus leucémique au moment de la naissance est quasiment complet et que les mutations alors acquises (principalement diverses fusions impliquant le gène *MLL*) sont suffisantes à l'initiation de la maladie [28,35].

1.4.2.3. Facteurs de susceptibilité génétique

L'importance des variations génétiques héritées sur le développement de la LAL reste aujourd'hui encore incertaine. Divers syndromes de prédisposition comme les syndromes de Bloom ou de Down, l'anémie de Fanconi ou la neurofibromatose sont des preuves directes démontrant que la LAL pédiatrique possède une composante génétique, mais ces cas ne représentent collectivement que moins de 5% des diagnostics [36].

Depuis quelques années, plusieurs études ont démontré le rôle possible de polymorphismes rares ou communs sur une augmentation du risque de développement de la maladie [37]. Une analyse de gènes candidats réalisée au sein de notre laboratoire a par exemple montré que des polymorphismes régulant les niveaux d'expression des gènes codant pour les inhibiteurs du cycle cellulaire CDKN2A, CDKN2B et CDKN1B pourraient influencer le risque de développer une LAL-B chez l'enfant [38]. Plus récemment, plusieurs études d'association génétique pangénomiques (*genome-wide association study*, GWAS) utilisant des données issues de génotypage par biopuces ont permis de mettre en évidence la contribution de plusieurs allèles communs de faible pénétrance à une prédisposition génétique de la LAL: 10q21.2 (*ARID5B*), 7p12.2 (*IKZF1*), 14q11.2 (*CEBPE*), 10p12.31-12.2 (*BMI1-PIP4K2A*), 7p12.2 (*DDC*), 9p21.3 (*CDKN2A*), 1q44 (*OR2C3*), 10p12.2 (*PIP4K2A*) et 10p14 (*GATA3*) [39-43]. Plusieurs de ces gènes codent pour des facteurs de transcription impliqués dans la différenciation des progéniteurs de lymphocytes B et/ou sont également les cibles de mutations somatiques identifiées au niveau de lymphoblastes. Certains de ces allèles sont associés avec certains sous-types de LAL, comme par exemple le SNP rs3824662 au niveau de *GATA3* associé aux LALs *Ph-like* [44]. Quelques-unes de ces associations ont pu être confirmées par notre groupe à partir des données de 284 patients issus de la cohorte *Quebec Childhood ALL* (QcALL) de l'hôpital Sainte-Justine [45].

Les rares cas de leucémies familiales ont permis de mettre en évidence des allèles rares de pénétrance plus élevée au niveau de plusieurs gènes comme *TP53*, *RAS* [46], *PAX5* [47] ou plus récemment *ETV6* [48]. Un des projets reportés au niveau de la section 3.1 de cette thèse, également basé sur l'étude d'un cas familial, nous a permis la mise en évidence de variants rares transmis au niveau de gènes de la voie de l'anémie de Fanconi.

1.4.2.4. Environnement et hypothèse virale

Depuis plusieurs décennies, pléthore de causes environnementales ont été étudiées comme facteur de risque, voire élément déclencheur initial expliquant la latence réduite des leucémies pédiatriques. Si plus d'une vingtaine de facteurs ont été identifiés par différentes études épidémiologiques comme contributeurs potentiels au développement de la LAL [49], la plupart des associations ne sont pas reproductibles [21,34]. Seule l'exposition à des doses anormales de radiations ionisantes (10 à ~200 mSv) a démontré un lien causal sans ambiguïté [50,51].

Par contre, bien que datant de 1917 [52], l’hypothèse d’une réponse anormale à une infection virale a fait l'objet de plusieurs études [53,54] et reste aujourd'hui encore l'une des pistes les plus plausibles [21,34]. Le modèle proposé par Greaves [34] suggère qu'une dérégulation de la réponse des cellules T à l'infection pourrait conduire à une puissante réponse inflammatoire et au relargage de chimiokines et cytokines induisant une réduction transitoire de l'hématopoïèse, voire provoquer une apoptose (Figure VII). Cet état particulier constituerait un microenvironnement propice à la survie et à la prolifération d'un clone pré-leucémique déjà porteur d'une mutation lui conférant un avantage sélectif. Une expansion clonale de ces cellules exposées au stress oxydatif associé à l'inflammation contribuerait à l'acquisition d'une mutation secondaire et ainsi conduirait à l'émergence des cellules leucémiques [34].

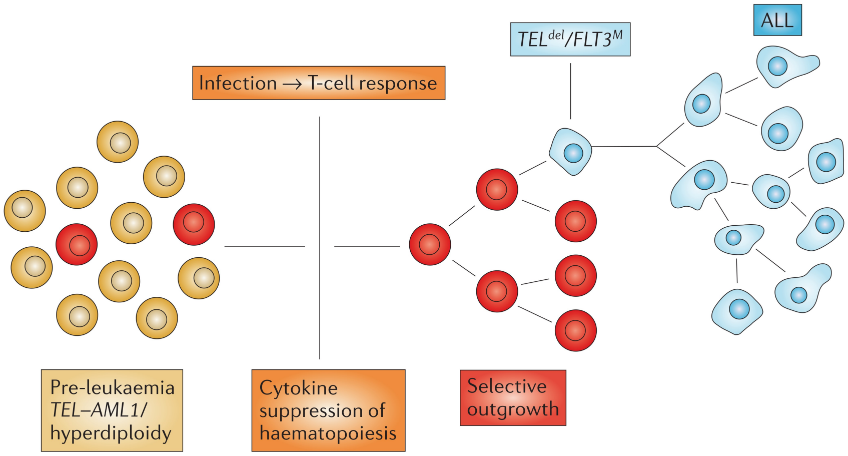

ALL
TEL^del/FLT3^M
Infection → T-cell response
Pre-leukaemia TEL–AML1/ hyperdiploidy
Cytokine suppression of haematopoiesis
Selective outgrowth

Figure VII. Modèle de stress prolifératif médié par l'exposition à un agent infectieux et entraînant la sélection des cellules pré-leucémiques. Figure reproduite de Greaves *et al.*, 2006 [34].

À cet égard, nous avons récemment démontré que la surexpression de *CLIC5* codant pour un canal chlorure, résultant de la perte de fonction de son répresseur transcriptionnel ETV6, entraîne une résistance accrue à l'apoptose médiée par le peroxyde d'hydrogène [55]

(Annexe I). La translocation t(12;21) (*ETV6-RUNX1*/*TEL-AML-1*), qui présente une origine *in utero* [56], représente un candidat intéressant en tant que première mutation somatique conférant un avantage sélectif pouvant correspondre au modèle proposé par Greaves.

L'équipe de Greaves a récemment identifié une expression simultanée des enzymes AID, RAG1 et RAG2 au niveau des cellules pré-BII en présence de forts stimuli inflammatoires [57]. Ces enzymes sont normalement restreintes aux processus de recombinaison VDJ et d'hypermutation somatique des gènes d'immunoglobulines. Ils ont démontré que ces enzymes ne permettent pas seulement la diversification du répertoire des lymphocytes B mais contribuent également à l'acquisition de lésions génétiques pouvant entraîner l'évolution clonale de cellules pré-leucémiques (porteuses par exemple de la translocation t(12;21)) en lymphoblastes.

1.4.3. Les bases génétiques de la LAL de l'enfant

Environ 75% des LALs de l'enfant présentent une anomalie génomique macroscopique à partir de laquelle un classement phénotypique est établi (Figure VIII, Tableau II) [58]. Ces altérations, plus ou moins fréquentes et souvent considérées comme événement primaire de la maladie, sont associées à divers prognostic [37].

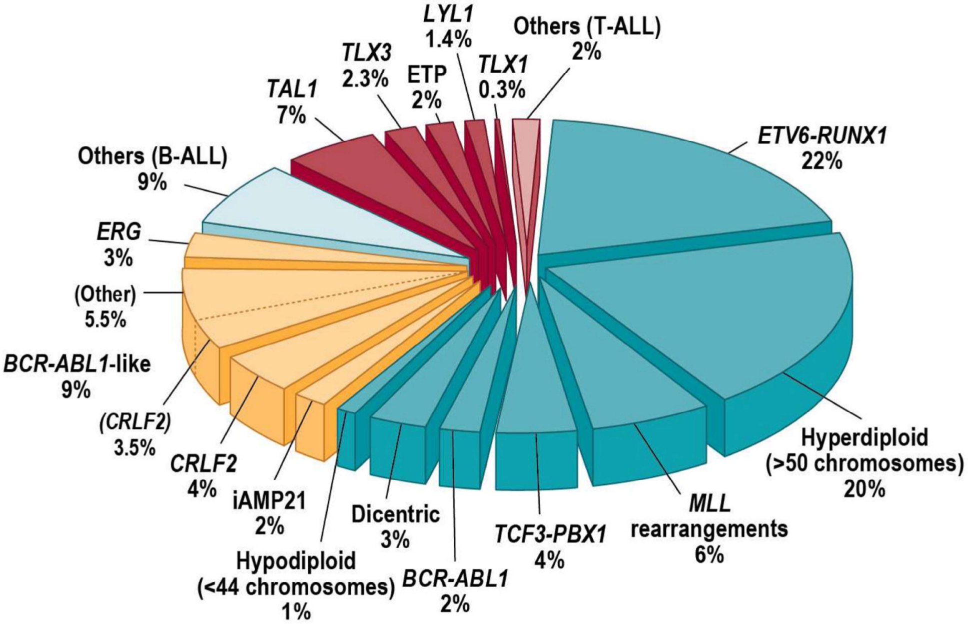


Figure VIII. Anomalies cytogénétiques et moléculaires de la LAL de l'enfant. Les segments en bleu ou jaune concernent la leucémie de type B tandis que les segments en rouge concernent la leucémie de type T. Figure reproduite de Inaba *et al.*, 2013 [21].

Elles peuvent consister en des réarrangements chromosomiques entraînant la création d'un produit de fusion chimérique oncogénique, comme c'est le cas de la translocation t(12;21)(p13;q22) (*ETV6-RUNX1*/*TEL-AML-1*) dont il a été question plus tôt dans cette introduction. Avec 22% des patients LAL-B porteurs, cette translocation représente une des altérations les plus fréquemment retrouvées au niveau de la leucémie de l'enfant [21]. L'expression du produit de fusion, formé à partir de deux facteurs impliqués dans la régulation normale de l'hématopoïèse et du développement des cellules lymphoïdes, perturbe la différenciation des lymphocytes B [59] et favorise l'auto-renouvellement des cellules pré-leucémiques [60]. Un autre exemple important de remaniement chromosomique conduisant à la formation d'une chimère oncogénique est la translocation t(9;22)(q34;q11.2) (*BCR-ABL1*). Aussi connue sous le nom de chromosome de Philadelphie, elle est présente chez environ 2% des patients LAL-B [21]. L'activité tyrosine kinase exacerbée de la chimère

BCR-ABL1 active de manière aberrante les voies de signalisation comme RAS/MAPK, PI-3 kinase, c-CBL et CRKL, JAK-STAT et SRC. Ceci entraîne la transformation néoplasique en modifiant des processus cellulaires de base comme le contrôle de la prolifération et de la différenciation, de l'adhésion et de la survie [61].

Tableau II. **Principaux sous-types de la LAL de l'enfant.** Tableau reproduit de Hunger and Mullighan, 2015 [24].

Sous-types	Prévalence (%)	Commentaires
LAL-B		
Hyperdiploïdie avec> 50 chromosomes	20-30	Excellent pronostic
Hypodiploïdie avec <44 chromosomes	2-3	Mauvais pronostic; Mutations fréquentes dans la voie Ras et les gènes de la famille Ikaros
Translocation t(12;21)(p13;q22) encodant la fusion ETV6-RUNX1	15-25	Excellent pronostic
Translocation t(1;19)(q23;p13) encodant la fusion TCF3-PBX1	2-6	Augmentation de l'incidence chez les Afro-Américains; Généralement excellent pronostic; Association avec une rechute du SNC
Translocation t(9;22)(q34;q11.2) encodant la fusion BCR-ABL1	2-4	Résultats historiquement médiocres; Amélioré avec l'ajout d'imatinib et/ou de dasatinib à une chimiothérapie intensive
LAL *Ph-like*	10-15	Lésions multiples activant les kinases; Associée à un âge plus avancé, un taux élevé de globules blancs et une altération d'*IKZF1* ; Susceptibles d'être traitées par TKI
Translocation t(4;11)(q21;q23) encodant la fusion MLL-AF4	1-2	Fréquent chez le nourrisson ALL (<6 mois); mauvais pronostic
t(8;14)(q24;q32), t(2;8)(q12;q24), t(2;8)(q12;q24); Réarrangement *MYC*	2	Pronostic favorable avec une chimiothérapie à court terme à fortes doses
Réarrangement *CRLF2* (IGH-CRLF2; P2RY8-CRLF2)	5-7	Fréquent chez les patients atteints du syndrome de Down et les LAL *Ph-like* (environ 50%); Associée à une deletion et/ou des mutations d'IKZF1 et des mutations de JAK1/2 et à un pronostic défavorable dans le cas d'une LAL associée au syndrome de Down
LAL avec dérégulation d'ERG	~7	Profil d'expression distinct; La majorité présente des délétions focales d'*ERG* et un résultat favorable en dépit des altérations *IKZF1*
Réarrangement *PAX5*	~2	Partenaires multiples, habituellement au niveau de dic(7;9), dic(9;12) et dic(9;20)
iAMP21	~2	Altérations structurales complexes du chromosome 21; Rarement associé à une translocation Robertsonienne constitutionnelle (15;21)(q10;q10)c; Mauvais pronostic
LAL-T		
Translocations t(1;7)(p32;q35) et t(1;14)(p32;q11) et délétion 1p32; Dérégulation de *TAL1*	15-18	Résultats généralement favorables
Translocation t(11;14)(p15;q11) et délétion *5'LMO2* ; Dérégulation de *LMO2*	10	Résultats généralement favorables
Translocations t(10;14)(q24;q11) and t(7;10)(q35;q24); Dérégulation de *TLX1/HOX11*	7	Bon pronostic
Translocation t(5;14)(q35;q32); Dérégulation de TLX3	20	Couramment fusionné à BCL11B, également une cible de délétion et/ou de mutations; Mauvais pronostic
Translocation t(10;11)(p13;q14); PICALM-MLLT10/CALM-AF10	10	Peut être associé à un mauvais pronostic
MLL-MLLT1/MLL-ENL	2-3	Pronostic supérieur à d'autres types de leucémie avec réarrangement de *MLL*
Amplification de 9q34 encodant NUP214-ABL1	6	Potentiellement tolérable aux TKIs, également identifiés dans le B-ALL à haut risque; Les autres fusions de kinase identifiées dans T-ALL comprennent EML1-ABL1, ETV6-JAK2 et ETV6-ABL1
Translocation t(7;9)(q34;q34)	<1	Réarrangement de *NOTCH1* ; Également des mutations dans> 50% de T-ALL
Early T-cell precursor ALL	10-15	Immunophénotype immature; Expression de marqueurs myéloïdes et/ou de cellules souches; Résultats historiquement médiocres, quoique améliorés dans des études récentes; Génétiquement hétérogènes avec des mutations dans les régulateurs hématopoïétiques, les cytokines, la voie Ras, et les modificateurs épigénétiques

Il existe également des réarrangements chromosomiques entraînant l'expression ectopique de proto-oncogènes à proximité du point de cassure. Par exemple, la surexpression des proto-oncogènes *TAL1*, *TLX1*, *TLX3*, et *LYL1* par leur juxtaposition aux loci des récepteurs antigéniques des lymphocytes T est commune au niveau de la LAL-T [21]. On retrouve aussi des anomalies du nombre de chromosomes (aneuploidie), soit avec une hypodiploïdie (<44 chromosomes) chez environ 2 à 3% des patients [24], ou avec une haute hyperdiploïdie (>50 chromosomes) avec 25 à 30% des patients présentant un gain non aléatoire de chromosomes [24] incluant généralement les chromosomes 4, 6, 10, 14, 17, 18, 21 et X [21].

Toutefois, la plupart de ces altérations précoces ne sont pas suffisantes pour le développement de leucémies lorsqu'elles sont testées au sein de modèles expérimentaux [58]. De plus, environ 25% des LALs de l'enfant ne présentent aucune anomalie génomique de ce type, ce qui implique la présence d'altérations génétiques submicroscopiques contribuant à la leucémogenèse. Moins d'une vingtaine de ces mutations sont généralement présentes chez un patient [62-64,37,46]. Certains sous-types, comme le LAL infantile avec réarrangement *MLL*, présentent un taux mutationnel parmi les plus faibles des cancers humains [37,65]. De manière générale, les mutations identifiées altèrent des processus cellulaires clés, présentent une fréquence dépendante du sous-type de LAL considéré et influencent le pronostic du patient (Tableau III, Figure IX) [37]. Les régulateurs transcriptionnels modulant le développement des cellules lymphocytaires (*PAX5*, *IKZF1* et *EBF1*) sont fréquemment ciblés par des mutations entraînant une perte de fonction ou l'expression d'un allèle dominant négatif au niveau de la LAL-B [66,67]. De nombreuses mutations sont également identifiées au niveau des voies régulatrices du cycle cellulaire (*TP53*, *RB1* et *CDKN2A*), de récepteurs de cytokines, de tyrosines kinases, de la voie de signalisation RAS/MAPK (*ABL1*, *ABL2*, *CRLF2*, *CSF1R*, *EPOR*, *FLT3*, *IL2RB*, *IL7R*, *JAK1,* *JAK*2, *JAK3*, *NTRK3*, et *PDGFRB*) et de régulateurs épigénétiques (*EZH2*, *CREBBP*, *SETD2*, *MLL2*/*KMT2D* et *WHSC1*/*NSD2*) [68-70,37].

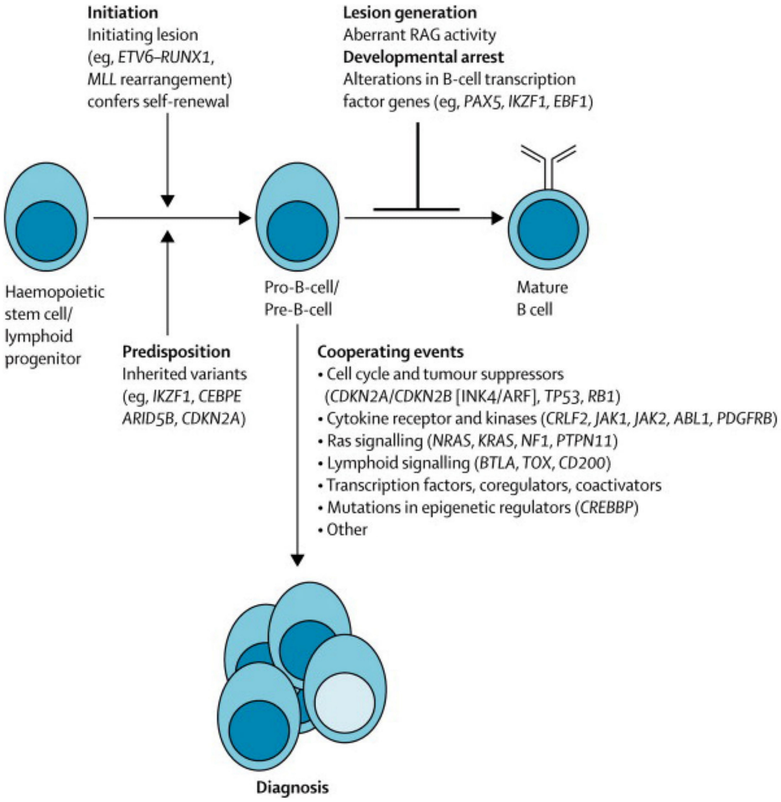

Initiation
Initiating lesion
(eg, ETV6–RUNX1,
MLL rearrangement)
confers self-renewal
Lesion generation
Aberrant RAG activity
Developmental arrest
Alterations in B-cell transcription
factor genes (eg, PAX5, IKZF1, EBF1)
Haemopoietic
stem cell/
lymphoid
progenitor
Pro-B-cell/
Pre-B-cell
Mature
B cell
Predisposition
Inherited variants
(eg, IKZF1, CEBPE
ARID5B, CDKN2A)
Cooperating events
• Cell cycle and tumour suppressors
(CDKN2A/CDKN2B [INK4/ARF], TP53, RB1)
• Cytokine receptor and kinases (CRLF2, JAK1, JAK2, ABL1, PDGFRB)
• Ras signalling (NRAS, KRAS, NF1, PTPN11)
• Lymphoid signalling (BTLA, TOX, CD200)
• Transcription factors, coregulators, coactivators
• Mutations in epigenetic regulators (CREBBP)
• Other
Diagnosis

Figure IX. Représentation schématique des mutations génétiques récurrentes de la LAL-B et des processus cellulaires altérés. Figure reproduite de Hunger and Mullighan, 2016 [37].

Au niveau de la LAL-T, l'activation aberrante de la voie NOTCH1 (*NOTCH1-FBXW7*) ainsi que la perte de fonction du tumeur suppresseur *CDKN2A* représentent les deux altérations les plus fréquentes (> 60% des patients) [71-73]. Les gènes codants pour des transducteurs de signaux (*PTEN*, *JAK1*, *JAK3*, *NF1*, *NRAS*, *IL7R* et *FLT3*), des facteurs de transcription (*WT1*, *LEF1*, *ETV6*, *GATA3* et *BCL11B*) et des régulateurs épigénétiques (*EZH2*, *SUZ12*, *EED* et *KDM6A/UTX*) sont également des cibles récurrentes [73]. La caractérisation du paysage génomique de la LAL-T a fait l'objet d'un projet mené au cours de mon doctorat (chapitre 4).

Tableau III. Altérations génétiques principalement retrouvées dans la LAL de l'enfant. Tableau reproduit de Hunger and Mullighan, 2016 [37].

Gène	Altération	Fréquence	Voie de signalisation et conséquences de l'altération	Pertinence clinique
PAX5	Délétions, translocations et mutations	30% des LAL-B	Facteur de transcription requis pour le développement lymphoïde B; Mutations compromettent la liaison de l'ADN et l'activation transcriptionnelle	Important dans la leucémogenèse, mais pas associé à un résultat défavorable
IKZF1	Délétions, translocations et mutations	15% de tous les cas de LAL-B pédiatrique, incluant 70% -80% de LAL positive pour BCR-ABL1 et un tiers de LAL-B BCR-ABL1 à haut risque	Facteur de transcription requis pour le développement de HSC en précurseur lymphoïde; Des deletions et des mutations entraînent une perte de fonction ou des isoformes à l'effet dominant négatif	Associé à un mauvais pronostic
JAK1/2	Mutations des domaines pseudokinase et kinase	18%-35% des LAL avec syndrome de Down et 10.7% des LAL négative pour BCR-ABL1	Activation constitutive de JAK-STAT	Potentiellement ciblable par les TKIs qui inhibent JAK1/2
CRLF2	Réarrangement IGH-CRLF2 ou P2RY8-CRLF2 conduisant à une surexpression	5%-16% des LAL-B pédiatriques et adultes et >50% des LAL avec syndrome de Down; 50% des LAL *Ph-like*	Associé à des mutation de *JAK* dans jusqu'à 50% des cas; Les mutations de *CRLF2* et *JAK* résultent en une activation constitutive de *STAT* ; Au niveau des LAL-B à haut risque, associé aux altérations d'*IKZF1* , aux mutations *JAK* et à un mauvais pronostic	Détection par cytométrie en flux ou analyse moléculaire; Potentiellement ciblable par des inhibiteurs de JAK
Alterations activatrices des kinases dans LAL *Ph-like*	Réarrangements de 13 récepteurs de cytokine et des tyrosine kinases; Mutations d'*IL7R* et *FLT* ; Délétions de *SH2B3*	10% des LAL-B pédiatriques, jusqu'à 30% des LAL adultes; Associé avec un profile d'expression *Ph-like*	Activation des voies de signalisation des kinases; Peut être soumis à une thérapie TKI	Haut risque et risque accru de rechute; Rapports anecdotiques de réponse à la thérapie TKI
NOTCH1	Mutations, parfois délétions	>60% LAL-T	Activation de la voie *NOTCH1* ; Les mutations dans *FBXW7* et *PTEN* influencent également la signalisation *NOTCH1* et le pronostic	Lésion pathogène clé dans T-ALL; Ciblage direct limité par la toxicité; Associations variables avec le résultat
CREBBP	Délétions et mutations	19% des rechutes; Également muté dans les lymphomes non hodgkinien	Entraînent une dérégulation de l'acétylation des histones et une dérégulation transcriptionnelle	Mutations sélectionnées à la rechute et associées à la résistance aux glucocorticoïdes
NT5C2	Mutations	Jusqu'à 20% des rechutes	Gain de fonction	Mutations sélectionnées à la rechute et associées à la résistance aux thiopurines
TP53	Délétions et mutations	~90% des LAL à faible hypodiploïdie; Rare dans les autres cas	Fréquemment germinale; Confèrent une résistance au traitement	Mutations associées aux rechutes

Outre les mutations communes constituant le paysage génomique classique de chaque type de leucémie, nous croyons que des mutations plus rares, spécifiques au contexte génétique de la tumeur de chaque patient, participent à la déstabilisation de certains processus cellulaires et contribuent à l'initiation ou à la progression de la maladie. Ces altérations, plus difficiles à caractériser, font l'objet d'une partie du chapitre 3 de cette thèse.

1.4.4. Traitement et pronostic

Le protocole de traitement du Dana-Farber Cancer Institute (DFCI) [74], utilisé au CHU Sainte-Justine, dure 24 mois et se divise en 3 phases (le groupe de risque du patient influence principalement le dosage de la chimiothérapie):

- L'induction, qui dure de 4 semaines. L'enfant reçoit une chimiothérapie composée de glucocorticoïdes (prednisone ou dexaméthasone), d'antinéoplasiques (vincristine), et d'asparaginase. L'induction est éventuellement complémentée d'anthracycline comme la doxorubicine. À la fin de cette phase d'induction, le patient est généralement considéré en rémission.
- L'intensification, qui consiste en 6 à 8 mois de chimiothérapie intensive pour maintenir la rémission et empêcher l'invasion du système nerveux central. Le patient reçoit notamment de fortes doses de méthotrexate par intraveineuse et des doses intramusculaires d'asparaginase.
- La maintenance, jusqu'à 24 mois durant lesquels le patient reçoit des doses journalières de méthotrexate. Le protocole est équivalent à celui de la période d'intensification (excepté pour l'asparaginase en ce qui concerne les patients classés en risque standard).

Ce traitement, ainsi que les thérapies équivalentes (ALL au Royaume Uni, DCOG aux Pays-Bas et FRALLE en France), ont permis une amélioration significative des résultats avec une survie sans événement à 5 ans d'environ 85% et un taux de survie globale pouvant atteindre 90% [75].

La caractérisation génétique des patients a mis en évidence l'existence de lésions influençant le pronostic (voir 1.4.3: Les bases génétiques de la LAL de l’enfant). Par contre, seulement une fraction de cette information génétique est présentement considérée dans le calcul de stratification du risque. Les méthodes actuelles tiennent compte de l'âge, du compte de cellules blanches au diagnostic, de l'invasion éventuelle du système nerveux central, de la mesure de maladie résiduelle minimale et se limitent à quelques informations génétiques macroscopiques seulement comme l'aneuploïdie et certains réarrangements chromosomiques [58] (Tableau IV). Cette stratification de patients a permis d'améliorer le taux de survie et de réduire les conséquences du traitement à long terme en optimisant l'utilisation de la chimiothérapie et en limitant l'utilisation de l'irradiation crânienne prophylactique. Par contre, la prise en compte des données génomiques permettra un raffinement de la stratification, l'administration d'une thérapie optimale [37] et une diminution des chances de rechutes qui concernent environ 15% des patients [76] et dont l'issue est généralement dramatique.

Tableau IV. Exemple de stratification du risque d'après le consortium DFCI (1985-2000). Tableau reproduit de Silverman *et al.*, 2010 [74].

	Risque standard	**Risque élevé**	**Risque très élevé**
Âge (années)	1985-95: 2 to <9 1995-2000: 1 to <10	1985-95: ≥9 1995-2000: ≥10	1985-2000: <1
Compte de cellules blanches (x10^9/L)	1985-95: <20 1995-2000: <50	1985-91: 20 to <100 1991-5: ≥20 1995-2000: ≥50	1985-91: ≥100
Phénotype	LAL-B	LAL-T	
Atteinte du système nerveux central	Non	Oui	
Masse médiastinale	Non	Oui	
t(9;22)	Non	Oui	

1.4.5. Théorie de l'évolution et leucémie

L'ensemble des cancers, incluant la LAL, évoluent selon le processus de diversification génétique couplé à une sélection naturelle, c'est-à-dire selon un principe d'évolution tel qu'énoncé par Charles Darwin il y a maintenant plus de 150 ans. Grâce à des études longitudinales [76-79], les équipes de recherches ont contribué à déchiffrer les mécanismes régissant l'évolution intra-tumorale et inférer les relations ancestrales inter-clonales.

L'ensemble de ces données s'accordent aujourd'hui sur le fait que la trajectoire évolutive des tumeurs est complexe et ramifiée, comme l'avait déjà prédit Nowell en 1976 [80], plutôt que linéaire comme certains modèles le proposaient initialement (Figure X) [13]. Comprendre ces processus est essentiel étant donné qu'ils régissent également la sélection des clones leucémiques à l'origine des rechutes [81].

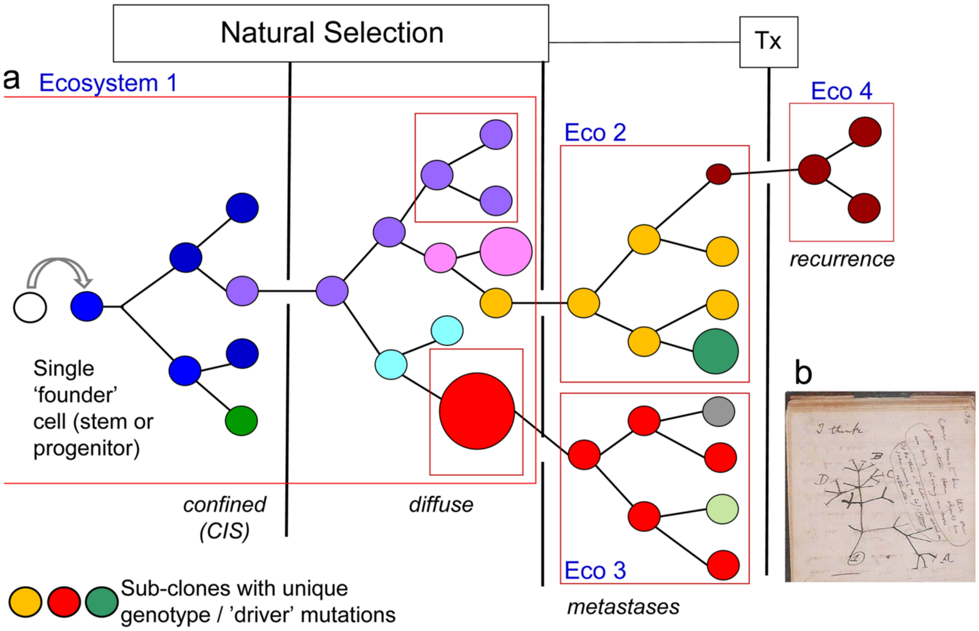

Natural Selection
Tx
a
Ecosystem 1
Eco 2
Eco 4
recurrence
Single 'founder' cell (stem or progenitor)
confined (CIS)
diffuse
b
Eco 3
metastases
Sub-clones with unique genotype / 'driver' mutations

Figure X. Évolution clonale d'une tumeur. (**a**) Architecture clonale en “arbre”. La pression de sélection entraîne l'expansion ou l'extinction de certains sous-clones. (**b**) Arbre phylogénétique de Darwin (tiré de son cahier de notes en 1837). Eco 1 à 4 (boîtes rouges) correspondent à différents écosystèmes ou habitats. Tx, thérapie; CIS, *carcinoma in situ*. Figure reproduite de Greaves and Maley, 2012 [13].

1.4.5.1. Écosystème, compétition et architecture clonale

De la même façon que la capacité d'interaction et d'adaptation avec l'environnement permet l'émergence d'une espèce, les cellules leucémiques ne sont pas autonomes et dépendent de leur faculté à exploiter de façon dynamique leur microenvironnement tissulaire et plus spécifiquement certaines niches [81-84]. Outre la compétition pour les ressources et l'espace du microenvironnement entre cellules normales et tumorales, il existe une compétition inter-clonale qui conduit à la sélection des clones les mieux adaptés [85]. La diversification génétique des cellules tumorales, qui permet à certains sous-clones l'acquisition de mutations *drivers* conférant un avantage sélectif au regard des ressources disponibles dans le milieu, entraîne l'existence de valeurs adaptatives (ou *fitness*) spécifique à chaque sous-clone par rapport à chaque microenvironnement. Cette relation avec le milieu n'est pas unidirectionnelle et souvent l'interaction réciproque cancer/microenvironnement entraîne un remodelage du tissu (ex: angiogenèse des tumeurs solides) et une spécialisation des niches procurant un avantage sélectif à certains clones [86]. L'avantage procuré par une mutation étant contexte-dépendant, une mutation préalablement *passenger*, c'est-à-dire n'offrant aucun avantage sélectif, peut devenir *driver* dans un milieu différent ou modifié. La dynamique du milieu joue d'ailleurs un rôle essentiel dans le processus d'évolution tumorale. Des modèles mathématiques ont démontré que la sélection de phénotypes plus agressifs ou plus robustes serait moins probable au sein de microenvironnements homogènes et stables [13,87]. Outre la modification du milieu, l'adaptation passe également par la migration des cellules tumorales et l'occupation de nouvelles niches (comme le système nerveux central). D'ailleurs, si les différents sous-clones se partagent initialement l'espace du microenvironnement primaire, différents clones peuvent par la suite occuper différents territoires [13]. Ce phénomène d'évolution clonale en parallèle au niveau de plusieurs sites est illustré par les leucémies concordantes chez les jumeaux (comme discuté avant), par l'analyse des lésions métastasiques [88,89] ou encore l'apparition de cancers testiculaires bilatéraux [13,90] (Figure XI). Outre le défi clinique que représente la séparation territoriale de clones génétiquement différents au moment du diagnostic, certains sites ectopiques peuvent également servir de sanctuaires (niches) pour certaines cellules tumorales [91].

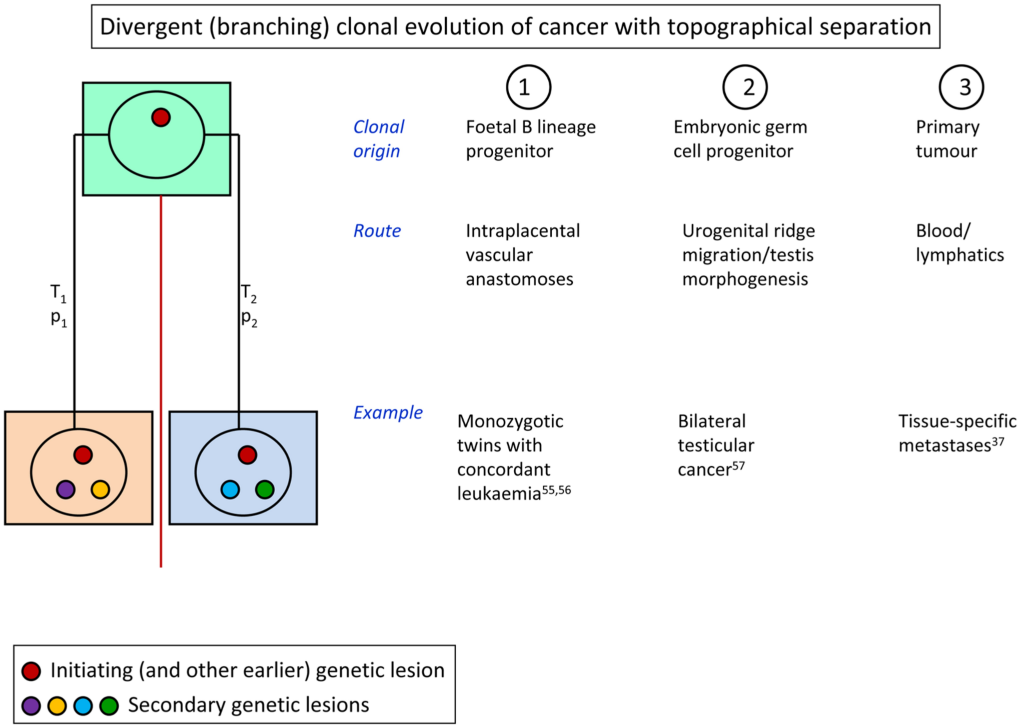

Divergent (branching) clonal evolution of cancer with topographical separation
T1
p1
T2
p2
1
2
3
Clonal origin
Foetal B lineage progenitor
Embryonic germ cell progenitor
Primary tumour
Route
Intraplacental vascular anastomoses
Urogenital ridge migration/testis morphogenesis
Blood/ lymphatics
Example
Monozygotic twins with concordant leukaemia55,56
Bilateral testicular cancer57
Tissue-specific metastases37
Initiating (and other earlier) genetic lesion
Secondary genetic lesions

Figure XI. Évolution clonale divergente d'un cancer avec séparation topographique. Dans chaque exemple, un ancêtre clonal (une seule cellule) est à l'origine du partage de mutations (ex: *ETV6-RUNX1* au niveau des leucémies ou c-kit au niveau des cancers testiculaires). L'évolution des 2 sous-clones (T1 et T2) peut être simultanée ou séparée de plusieurs années [28,90,92,93]. Les probabilités d'émergence des 2 sous-clones sont indépendantes et différentes. Dans la plupart des cas, un seul des jumeaux développe une leucémie (90%). La pénétrance des cancers testiculaires bilatéraux ayant la même origine est inconnue [90]. Figure reproduite de Greaves and Maley, 2012 [13].

1.4.5.2. Évolution clonale et rechute

Malgré le succès thérapeutique, près de 15% des patients LAL présentent des rechutes dont l'issue est généralement fatale [76]. Ainsi la LAL de l'enfant demeure le deuxième cancer pédiatrique en terme de taux de mortalité derrière les cancers du cerveau [2]. Il est donc essentiel de comprendre les mécanismes régissant ces rechutes. Depuis quelques années, l'avènement du séquençage de nouvelle génération a rendu possible l'analyse des paysages génomiques des rechutes. Outre l'identification des mutations qui y sont spécifiquement associées, le séquençage du matériel normal, de la tumeur primaire et de la rechute d'un patient, a permis d'étudier la contribution de l'évolution clonale à ces rechutes [76-79,94-97]. Dans ces études, l'identité clonale est inférée de la fréquence allélique des mutations (nombre de *reads* de séquençage supportant les mutations sur la couverture totale des positions considérées) en extrapolant le fait que la fréquence identifiée est représentative de la proportion de clones mutés dans la population totale de cellules. Cette méthode permet de décortiquer l'architecture clonale pré- et post-traitement, d'identifier les lésions génétiques présentes (de façon sous-clonale) ou non au diagnostic et qui émergent à la rechute et de déterminer la chronologie d'accumulation de ces mutations. L'étude du *turnover* mutationnel permet de caractériser la nature des clones résistants qui survivent à la thérapie. Cette méthode a été appliquée afin d'étudier les rechutes de 19 patients LAL-B issus de la cohorte QcALL du CHU Sainte-Justine (chapitre 5). De manière générale, ces études ont permis de mettre en évidence de nouveaux gènes associés à la résistance au traitement comme *CREBBP*, codant pour une histone acetyltransferase et induisant une résistance aux glucocorticoïdes une fois muté [94,95], ou encore *NT5C2* codant pour une hydrolase et conférant une résistance au 6-mercaptopurine et 6-thioguanine [96,97].

Deux profils majeurs d'évolution clonale ont pu être identifiés (Figure XII)[76,77]:

- un clone dominant au diagnostic reste dominant à la rechute en accumulant des mutations supplémentaires (modèle 1, Figure XII). Les patients présentant ce modèle évolutif ont soit été traités de manière inadéquate, soit présentaient déjà au diagnostic des mutations de résistance au traitement, somatiques ou germinales, au niveau du clone fondateur.
- un clone mineur au diagnostic émerge à la rechute après acquisition de mutations de résistance ou après l'augmentation de sa valeur adaptative due à une mutation préexistante conférant un avantage face à la pression de sélection (chimiothérapie) venant remodeler le microenvironnement tumoral (modèle 2, Figure XII). Comme le soulignait Ding en 2012 [77],

le génome des rechutes des patients leucémiques sont des cibles “en mouvement” qu'il est indispensable de caractériser en profondeur pour adapter les traitements afin de diminuer les chances de résistance en éradiquant totalement clones fondateurs et sous-clones.

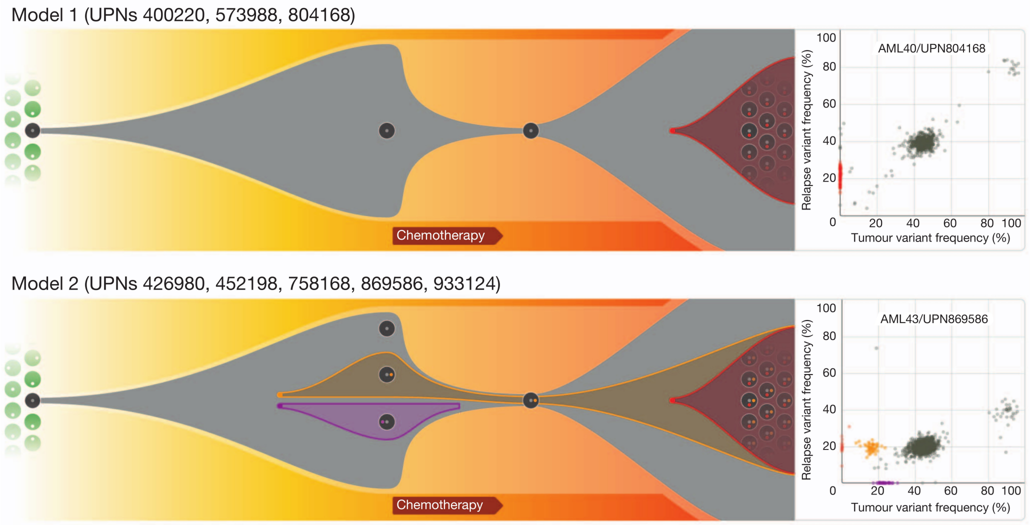

Model 1 (UPNs 400220, 573988, 804168)
Chemotherapy
AML40/UPN804168
Relapse variant frequency (%)
100
80
60
40
20
0
20
40
60
80
100
Tumour variant frequency (%)
Model 2 (UPNs 426980, 452198, 758168, 869586, 933124)
Chemotherapy
AML43/UPN869586
Relapse variant frequency (%)
100
80
60
40
20
0
20
40
60
80
100
Tumour variant frequency (%)

Figure XII. Représentation graphique de l'évolution clonale de la tumeur primaire à la rechute. Les exemples présentés ici sont issus de l'analyse de rechutes de patients atteints de leucémie myéloïde aiguë. Les mêmes *patterns* ont été identifiés au niveau de rechutes LAL. Le modèle 1 montre le clone fondateur au diagnostic resté dominant à la rechute après l'acquisition de mutations rechute-spécifiques. Le modèle 2 montre un clone mineur au diagnostic émerger lors de la rechute. Figure reproduite de Ding *et al.*, 2012 [77].

1.5. L'étude des cancers à l'heure de la génomique

Depuis plusieurs années maintenant, le séquençage de nouvelle génération (SNG, aussi appelé séquençage parallèle de masse) a révolutionné l'étude de la génomique des cancers. Cette technologie permet d'accéder à l'information génétique du patient grâce au séquençage de génome (le plus souvent par le biais d'enrichissement des régions), du transcriptome grâce au séquençage de l'ARN, ainsi qu'aux informations épigénétiques par séquençage bisulfite et *ChIP-sequencing* informant respectivement sur la méthylation de l'ADN et les modifications subies par les histones.

De nombreuses études ont ainsi contribué à la construction d'un répertoire d'altérations concernant des centaines de gènes impliqués dans la susceptibilité aux cancers, l'initiation et la progression de la maladie [98]. Notamment, de nouveaux *drivers* ont été identifiés au niveau de carcinomes de la peau [99,100], cancers de la vessie [101], cancers de la prostate [102,103], cancers colorectaux [104], cancers du sein [23,105-108], medulloblastomes [109] et leucémies/lymphomes [46,63,110-112]. Nous avons également contribué à cet essor en identifiant de nouveaux drivers des LAL-B (chapitre 3: A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells) et LAL-T pédiatriques (chapitre 4: Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations). De plus, par l'intégration de données d'expression et de méthylation, nous avons pu identifier les voies de signalisation dérégulées au niveau de la LAL-B [113,114].

Aujourd'hui, le SNG offre une perspective clinique, non seulement en permettant une meilleure caractérisation moléculaire et stratification du risque des patients, mais aussi offrant la possibilité d’identifier un traitement personnalisé basé sur le paysage génomique spécifique aux patients. Si le SNG est à l'origine d'avancées extraordinaires dans la compréhension de la biologie des cancers, il représente également un défi d'analyse majeur. Outre la gestion d'une quantité de données sans précédent [115], les chercheurs ont été rapidement confrontés à deux difficultés lors de l'interrogation des données SNG afin d'identifier de nouvelles mutations *drivers*:

- Comment distinguer des mutations (SNVs et indels) somatiques de polymorphismes germinaux ou de faux positifs générés par des erreurs de séquençage ou d'alignement sur le génome de référence?

- Comment dissocier les quelques mutations *drivers* des nombreuses mutations *passengers*?

De nombreux algorithmes ont été développés pour répondre à ces problèmes. La détection de SNVs et indels somatiques se fait généralement par l'intégration des données de séquençage des matériels normaux (souvent obtenus après rémission) et tumoraux des patients selon plusieurs approches. Par exemple, Strelka [116], un outil développé par une équipe d'Illumina, utilise une approche Bayésienne qui considère les fréquences alléliques “normales” et tumorales comme des variables continues, avec l'échantillon normal considéré comme un mélange de variations diploïdes germinales et de “bruit de fond” et le tumoral comme un mélange d'échantillon normal et de variations somatiques. Un réalignement autour des indels putatifs est effectué systématiquement par Strelka. Varscan2 [117], qui est couramment utilisé, lit simultanément les échantillons normaux et tumoraux et effectue une comparaison pairée des variations et de la couverture normalisée pour chaque position. Un algorithme heuristique détermine les génotypes normaux et tumoraux indépendamment en se basant sur des filtres de qualité prenant en compte la profondeur du séquençage, la qualité de la base considérée, la fréquence allélique et la significativité statistique des tests effectués. On peut également citer MuTect [118], développé par le Broad Institute, qui utilise un premier classifieur Bayésien pour identifier les variants au sein de l'échantillon tumoral et un deuxième classifieur pour déterminer la valeur somatique des mutations identifiées. Une étape additionnelle permet de filtrer les faux positifs restants en utilisant un panel d'échantillons normaux. Chacun de ces outils fournit des résultats satisfaisants dans des conditions standards avec une profondeur de séquençage élevée (>100X) et une faible contamination de l'échantillon tumoral par des cellules normales. Toutefois, comme le souligne la faible concordance des résultats obtenus sur un même jeu de données par l'utilisation de plusieurs algorithmes incluant les 3 cités [119,120] (Figure XIII), l'hétérogénéité des échantillons tumoraux couplée aux erreurs aléatoires ou systématiques du séquençage entraînent des difficultés d'analyse lorsque la qualité ou le *design* des données ne sont pas optimaux. Notamment, l'identification des mutations présentant de faibles fréquences alléliques est particulièrement complexe lorsque la profondeur de séquençage est réduite [119]. Cette identification est pourtant essentielle dans le cadre de diagnostics précoces, de détection de la maladie résiduelle et pour la caractérisation du processus de rechute (voir section 1.4.5.2: Évolution clonale et rechute). Par exemple, pour un échantillon avec une profondeur de séquençage de ~60X, Varscan2 présente une sensibilité inférieure à 0.5, 0.08 et 0.02 pour des fréquences alléliques de 18, 8 et 4%, respectivement [119]. Bien que le coût du séquençage soit en constante réduction, l'effort de caractérisation de cohortes de plus en plus grandes impose une limitation globale de la profondeur de séquençage dépassant rarement les 50X. Pour tenter de résoudre ce problème, nous avons développé un outil, SNooPer, s'adaptant aux conditions de séquençage sous-optimales et qui est présenté au niveau du chapitre 2 de cette thèse. Nous avons également développé un modèle probabiliste (QUADGT) permettant d'inférer les génotypes au sein du matériel tumoral en intégrant les génotypes du matériel normal du patient mais aussi celui des deux parents lorsqu'ils sont disponibles [121].

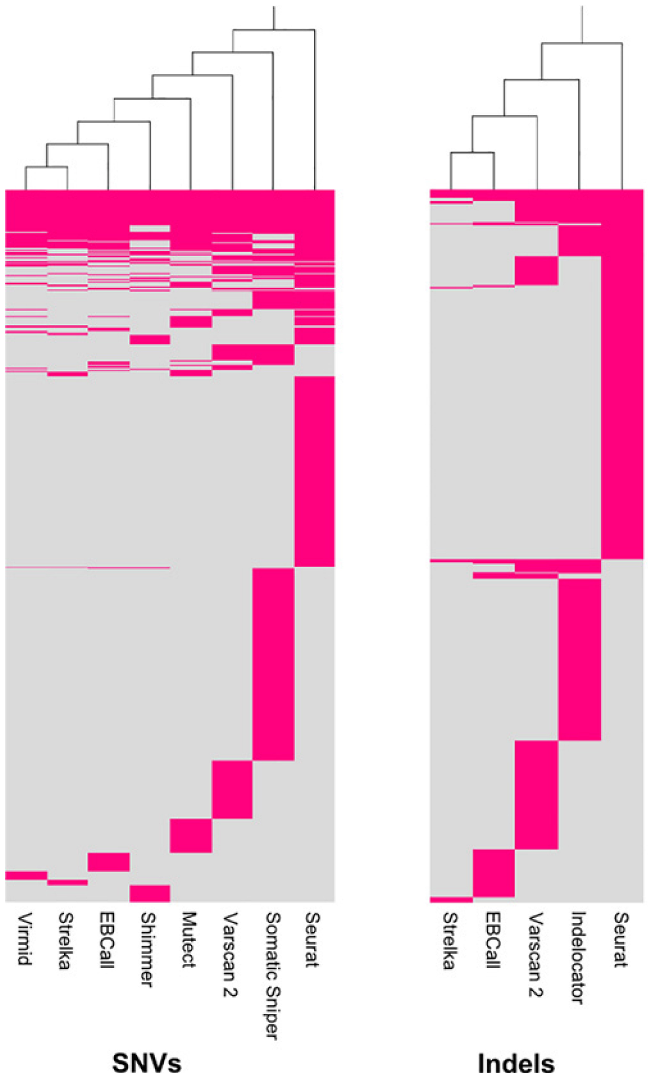


Figure XIII. Profils d'identification des SNVs et indels par différents outils. Regroupements hiérarchiques des mutations somatiques identifiées au niveau d'un séquençage d'exome présentant une couverture de 80X. Chaque ligne représente une mutation. Figure reproduite de Krøigård *et al.*, 2016 [120].

Une fois identifiées, les centaines (voire milliers) de mutations somatiques doivent être filtrées pour ne conserver que celles pouvant correspondre à des évènements *drivers*, c'est-à-dire celles présentant un impact fonctionnel putatif conférant un avantage sélectif aux cellules mutées. Pour cette étape également, une multitude d'algorithmes ont été développés (Figure XIV). Ils sont basés sur 2 principes majeurs:

- l'analyse de récurrences
- la prédiction de l'impact fonctionnel

Ces méthodes et leurs variantes sont autant d'alternatives disponibles mais qui ne présenteront pas la même efficacité en fonction du type de mutation recherché (fréquente ou rare) et du type tumoral (taux mutationnel élevé ou faible).

L'analyse de récurrences se base sur le principe même de la théorie de l'évolution. Étant donné que le processus mutationnel converge vers un phénotype oncogénique commun, les mutations *drivers* contribuant à la transformation et/ou la progression tumorale devraient apparaître plus souvent dans une cohorte de patients qu'attendu par chance étant donné le contexte mutationnel [122]. La méthode se base donc uniquement sur l'identification de régions nommées “*hills*” (pour collines) présentant un taux de mutation au sein de la cohorte étudiée significativement différent des “plaines” présentant un taux mutationnel de base (ou BMR pour *background mutation rate*). Plus spécifiquement, en se basant sur le BMR et le nombre *n* de nucléotides séquencés dans un gène *g*, la probabilité *Pg* qu'une mutation passenger soit observée dans *g* est Pg =1-(1-BMR). Étant donné que chaque mutation apparaît de manière indépendante dans chaque patient, le nombre d'occurrences dans *g* suit la loi de Bernoulli où *pg* est la probabilité de mutation. En mesurant *s* échantillons, le nombre de patients avec *g* muté est décrit par la variable aléatoire binomiale B(s,Pg). Ainsi, il est possible de calculer la probabilité que le nombre de patients observés (ou plus) présentent une mutation *passenger* [122]. Plusieurs outils populaires comme MutSigCV [123] et MuSiC [124] exploitent ce principe. La différence majeure entre les divers algorithmes se situe principalement au niveau de l’estimation du BMR, étape cruciale et particulièrement complexe étant donné l’hétérogénéité des taux mutationnels intra- et inter-tumoraux [123,125]. Par exemple, les gènes présentant l’expression la plus faible ou étant répliqués tardivement au cours du cycle cellulaire montrent un taux mutationnel plus élevé [123]. De manière générale, l'estimation complexe du BMR entraine la limitation majeure de ces méthodes: leur sensibilité. Si certains gènes *drivers* sont effectivement mutés fréquemment, la plupart des *drivers* nouvellement découverts sont mutés dans une petite fraction des tumeurs seulement [126]. Bien qu'il ait été démontré que des mutations somatiques rares (<1% des tumeurs) peuvent également agir en tant que *drivers* [127](voir aussi chapitre 3), ces outils sont pour la plupart incapables de les détecter.

Des méthodes indépendantes de la fréquence mutationnelle ont donc été développées. L'avantage principal de ces approches réside dans le fait qu'elles sont applicables à un seul individu et permettent donc l'identification de mutations plus rares. Beaucoup de ces outils plus ou moins complexes tentent de distinguer les mutations *passengers* des *drivers* en prédisant l'impact fonctionnel des mutations non silencieuses en utilisant diverses informations biologiques. Les plus rudimentaires tentent de prédire un impact éventuel en prenant en compte principalement le degré de conservation de l'acide aminé modifié, comme SIFT [128] et PROVEAN (Protein Variation Effect Analyzer) [129], ou/et l'impact de la modification sur la structure protéique, comme PolyPhen-2 [130] et Mutation Assessor [131]. Certains de ces outils sont en fait des classificateurs basés sur des algorithmes d'apprentissage entrainés à distinguer les mutations causant un impact fonctionnel des autres mutations. Par exemple, PolyPhen-2 utilise une classification naïve bayésienne entrainée en utilisant 2 jeux de données différents: les données HumDiv, contenant tous les allèles avec un effet connu sur une fonction moléculaire à l'origine de maladies mendéliennes, présents dans la base de données UniProtKB, et incluent les différences entre les protéines humaines et des mammifères proches considérées comme non délétères; les données HumVar, composées de l'ensemble des mutations référencées dans UniProtKB comme étant à l'origine de maladies humaines, avec les polymorphismes communs (fréquence allélique >1%) non-associés à une maladie et donc considérés comme non délétères. Pour chaque mutation ensuite testée, PolyPhen-2 calcule la probabilité postérieure bayésienne que cette mutation ait un impact fonctionnel et reporte une estimation de la sensibilité et de la spécificité de la prédiction [130]. Bien qu'ayant été utilisés par la communauté pour l'étude de données tumorales, les outils comme PolyPhen-2 exploitent des données d'entrainement non adaptées et génèrent donc de nombreux faux-positifs dans le cadre de l'analyse de mutations somatiques. Plus récemment, des méthodes spécialisées dans l'étude de ces mutations ont vu le jour. CHASM [126] par exemple utilise un classificateur Random Forest qui apprend au préalable à distinguer les mutations faux-sens *drivers* des *passengers* en se basant sur un ensemble d'entraînement constitué de vrais *drivers* issus de la base de données COSMIC [132] après curation et des *passengers* générés *in silico* en se basant sur l'estimation du BMR spécifique au type tumoral considéré. À chaque mutation est associée une série de caractéristiques incluant le degré de conservation, le changement de structure, l'annotation UniProtKB, etc. Lors de l'analyse, un score est attribué à chaque mutation testée en considérant la fraction d'arbre du Random Forest qui a voté pour une classification en tant que *passenger*.

Enfin, un nouveau principe est de plus en plus exploité pour l'identification de *drivers*, notamment au niveau de gènes rarement mutés. Étant donné que les protéines agissent rarement de manière isolée mais plutôt à l'intérieur de voies de signalisation ou métabolique, en identifiant les voies et les réseaux d'interactions susceptibles d'être altérés, il est possible de mettre en évidence des cibles plus rares leur appartenant [122]. L'algorithme HotNet [133] par exemple intègre les données de fréquence de mutation avec la topologie locale du réseau d'interaction. Il semble capable d'identifier des sous-réseaux contenant des gènes mutés chez un nombre limité de patients et non identifiables par les méthodes se basant sur la fréquence uniquement [133]. L'algorithme MEMo [134], quant à lui, utilise une approche différente où des sous-réseaux sont définis en suivant 3 critères: i) une récurrence de mutations dans les gènes du groupe, ii) des gènes codant pour des protéines participant à la même fonction biologique et iii) des mutations mutuellement exclusives à l'intérieur du même sous-réseau. De manière générale, cette nouvelle approche est prometteuse et a déjà permis de mettre en évidence de nouveaux réseaux impliqués dans le développement de cancers [133], toutefois elles sont encore limitées par la qualité de définition des réseaux d'interaction. Ces réseaux sont d'ailleurs souvent biaisés autour des oncogènes et des suppresseurs de tumeurs connus puisque leurs interactions ont fait l'objet d'études détaillées. En comparaison, les réseaux d'interaction des nouveaux gènes sont flous ou partiels [122].

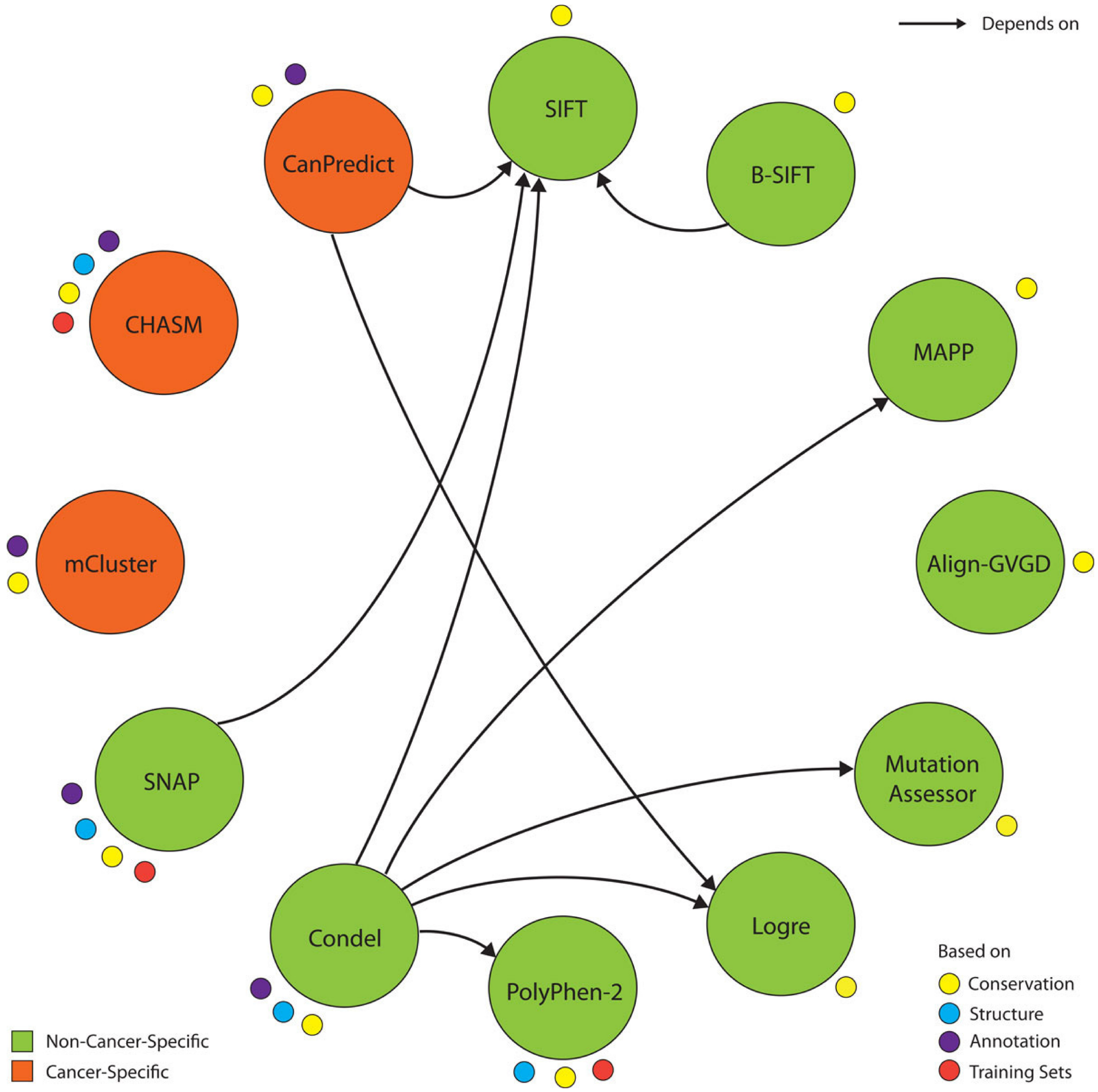


Figure XIV. Vue d'ensemble d'une série d'outils d'identification de mutations faux-sens *drivers*. Les types de données sur lesquels se basent les outils pour leurs prédictions sont indiqués par des petits cercles de couleurs. Les flèches indiquent l'interdépendance des outils. Figure reproduite de Gnad *et al*., 2013 [135].

Bien que ces méthodes se distinguent par leur principe, la plupart se limitent à l'analyse des variations codantes et délaissent l'analyse de nombreuses mutations localisées dans les milliards de nucléotides restants générés par un séquençage de génome complet. Depuis 2014, seulement quelques algorithmes permettant de prioriser les mutations non-codantes ont été développés [136-141].

1.6. Objectifs

La LAL est le cancer pédiatrique le plus courant et une des causes principales de mortalité par maladie chez les enfants de moins de 14 ans. Bien que les avancées thérapeutiques adaptées au pronostic établi par des facteurs cliniques et cytogénétiques ont permis une augmentation du taux de survie, 15 (LAL-B) à 25% (LAL-T) des patients ne répondent pas aux traitements standards et présentent ensuite un taux de mortalité supérieur à 50%. De plus, environ deux tiers de l'ensemble des survivants subissent à long terme les effets secondaires du traitement. Afin d'optimiser le traitement et l'issue de celui-ci, une meilleure classification de chaque tumeur ainsi qu'une thérapie plus personnalisée sont requises. Ceci nécessite une caractérisation en profondeur des bases moléculaires, héritées ou acquises, et des mécanismes liés gouvernant l'apparition et la progression de la maladie ainsi que la résistance au traitement. Il faut non seulement considérer les mutations récurrentes mais aussi être capable de mettre en évidence les mutations somatiques rares, voire tumeur-spécifiques, participant à la spécificité interindividuelle pourtant sous-estimée par la plupart des études actuelles. Ainsi, bien que nombreux, les efforts de caractérisation de la LAL se sont souvent limités aux seules mutations somatiques récurrentes, oubliant les évènements rares mais potentiellement fonctionnels, plus difficiles à mettre en évidence.

Au cours de mon doctorat, j'ai voulu exploiter le potentiel du séquençage de nouvelle génération afin de caractériser le paysage génomique d’une cohorte de LAL de l’enfant diagnostiqués au CHU Sainte-Justine (la cohorte QcALL), intégrant les mutations récurrentes mais aussi ces évènements *drivers* rares. Le SNG offre l’opportunité sans précédent de déchiffrer l'hétérogénéité génétique intratumorale qui contribue à la résistance aux traitements dans le cadre des rechutes.

Les objectifs spécifiques de cette étude étaient:

1. Développer un outil capable d'identifier efficacement les mutations somatiques au sein de données SNG “*low pass*” en intégrant les matériels normal et tumoral du patient;

2. Investiguer l'implication des mutations rares, héritées ou somatiques, dans la prédisposition, l'initiation ou la progression de la LAL-B;

3. Établir un paysage génomique exhaustif de la LAL-T;

4. Définir les schémas d'évolution clonale et identifier les mutations *drivers* de rechute à travers une étude longitudinale de patients LAL-B ayant subi au moins une rechute.

1.7. Impact attendu

Ce projet permettra: i) de fournir à la communauté de recherche en oncogénomique un outil capable d'exploiter les données de séquençage sous-optimales; ii) d'accroître notre compréhension des bases moléculaires des LAL-B et -T; iii) de capturer la dynamique clonale et les évènements associés à la résistance aux traitements.

En déchiffrant les déterminants génomiques qui constituent la variabilité inter- et intra-tumorale et en identifiant les mutations *drivers* de rechutes, nous espérons compléter les connaissances sur le processus leucémogène. À l'heure de la médecine personnalisée, ces données devraient permettre d'améliorer la classification des patients dans les groupes de risques et promouvoir la recherche translationnelle vers un développement de soins individualisés.

CHAPITRE 2

ARTICLE I - SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing

2.1. Avant-propos

Depuis l'avènement du séquençage de nouvelle génération, le nombre d'études oncogénétiques exploitant cette technologie n'a eu de cesse de croître. Il est aujourd'hui clair que les événements les plus fréquents, identifiables au sein de cohortes réduites, sont pour la plupart caractérisés. Afin d'identifier des évènements plus rares ou des effets/associations nécessitant d'étudier les données provenant d'un grand nombre d'individus, la communauté développe de plus en plus de projets exploitant de très larges cohortes. Étant donnée le coût du séquençage, l'augmentation du *n* de ces études se fait souvent au détriment de la qualité du séquençage avec notamment une profondeur de couverture réduite. Ces données de qualité sous-optimale posent de nombreuses difficultées. En particulier, l'identification de mutations somatiques par les outils disponibles conduit à la génération de nombreux faux positifs imposant des coûts supplémentaires de validations infructueuses. C'est dans l'optique de proposer une solution à ce problème que nous avons développé SNooPer, un outil permettant d'identifier SNVs et indels somatiques en s'adaptant à la qualité des données disponibles. Afin de permettre son utilisation par la communauté, l'outil est téléchargeable en ligne sur la plateforme d'"*open source software*“ SourceForge (https://sourceforge.net/projects/snooper/files/) et un site internet dédié a été créé (http://www.somaticsnooper.com/).

SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing.

Jean-François Spinella[1], Pamela Mehanna[1], Ramon Vidal[1], Virginie Saillour[1], Pauline Cassart[1], Chantal Richer[1], Manon Ouimet[1], Jasmine Healy[1], Daniel Sinnett[1,2]

1) CHU Sainte-Justine Research Center, Université de Montréal, Montreal, Qc, Canada; 2) Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, Qc, Canada.

This research was originally published in BMC Genomics. 2016;17(1):912. doi: 10.1186/s12864-016-3281-2.

The article is included with the permission of the authors. As open access journal, BMC Genomics requires no permission from the authors or the publishers to use the published article.

2.2. Authors' contributions

DS is the principal investigator and takes primary responsibility for the paper. JFS designed the somatic variant caller and the study. PC, CR and MO were involved in sample and library preparation for sequencing. JFS wrote SNooPer’s code and performed the analysis. PM, RV and VS were involved in revision of the code. JFS drafted the paper and interpreted the data. JH and DS were involved in critical revision of the manuscript. All authors approved the final version.

2.3. Abstract

Background. Next-generation sequencing (NGS) allows unbiased, in-depth interrogation of cancer genomes. Many somatic variant callers have been developed yet accurate ascertainment of somatic variants remains a considerable challenge as evidenced by the varying mutation call rates and low concordance among callers. Statistical model-based algorithms that are currently available perform well under ideal scenarios, such as high sequencing depth, homogenous tumor samples, high somatic variant allele frequency (VAF), but show limited performance with sub-optimal data such as low-pass whole-exome/genome sequencing data. While the goal of any cancer sequencing project is to identify a relevant, and limited, set of somatic variants for further sequence/functional validation, the inherently complex nature of cancer genomes combined with technical issues directly related to sequencing and alignment can affect either the specificity and/or sensitivity of most callers.

Results. For these reasons, we developed SNooPer, a versatile machine learning approach that uses Random Forest classification models to accurately call somatic variants in low-depth sequencing data. SNooPer uses a subset of variant positions from the sequencing output for which the class, true variation or sequencing error, is known to train the data-specific model. Here, using a real dataset of 40 childhood acute lymphoblastic leukemia patients, we show how the SNooPer algorithm is not affected by low coverage or low VAFs, and can be used to reduce overall sequencing costs while maintaining high specificity and sensitivity to somatic variant calling. When compared to three benchmarked somatic callers, SNooPer demonstrated the best overall performance.

Conclusions. While the goal of any cancer sequencing project is to identify a relevant, and limited, set of somatic variants for further sequence/functional validation, the inherently complex nature of cancer genomes combined with technical issues directly related to sequencing and alignment can affect either the specificity and/or sensitivity of most callers.

The flexibility of SNooPer's random forest protects against technical bias and systematic errors, and is appealing in that it does not rely on user-defined parameters. The code and user guide can be downloaded at https://sourceforge.net/projects/snooper/.

Keywords

Somatic variant; Low-pass next-generation sequencing; Machine learning; Random Forest.

2.4. Background

The advent of next-generation sequencing (NGS) has allowed unbiased in-depth interrogation of cancer genomes and has led to the identification of a number of tumor-specific mutations responsible for driving oncogenesis in multiple cancer types including skin carcinoma [1,2], bladder cancer [3], prostate cancer [4,5], colorectal cancer [6], breast cancer [7-12], medulloblastoma [13] and leukemias/lymphomas [14-18]. Sequencing of matched normal-tumor pairs is routine in cancer research in order to identify a relevant, and limited, set of somatic variants for further functional validation. However, the inherently complex nature of cancer genomes [19], the heterogeneity of tumor samples, as well as random (or systematic) sequencing and alignment errors can affect the specificity and/or sensitivity of most variant callers [20]. Of particular interest is the identification of low-frequency tumor alleles that arise in subclonal tumor cell populations, often contributing to treatment failure and relapse [21-25]. While NGS provides the opportunity to track specific mutations in tumor subclones and potentially uncover mutations with relapse driving potential [26], the identification of such mutations within the primary tumor cell population is often confounded and difficult to distinguish from background noise, as evidenced by the consistently low concordance rates between algorithms [20].

A number of methods have been developed to overcome these challenges in somatic mutation calling in matched normal-tumor samples. These methods are either heuristic, such as VarScan2 [27] that relies on independent analysis of tumor and normal genomes followed by a statistical Fisher's Exact Test of read counts for variant detection, or probabilistic, such as SomaticSniper [28], JointSNVMix [29], Strelka [30] and MuTect [31] that use Bayesian modeling to estimate likely joint normal-tumor genotype probabilities. Yet most somatic variant callers still perform poorly at low sequencing depths [32]. Indeed, large investments in validation efforts are needed to compensate the high false positive rates of most exploratory projects that are aimed at investigating more than a small set of top ranked high confidence somatic variants. And though progressively larger cohorts of individuals are being sequenced, the tendency towards shallow or low-coverage data is still de rigueur, particularly for whole genome sequencing initiatives, due to high sequencing costs.

To address these issues, we developed SNooPer, a versatile data mining approach that uses Random Forest (RF) classification [33] to accurately identify somatic variants in complex, low-depth sequencing data. Unlike available somatic variant callers, SNooPer does not rely on user-defined parameters but builds upon the data itself to construct powerful prediction models and increase calling performances. Using both simulated and real datasets, we evaluated SNooPer's ability to detect true somatic mutations in unbalanced, low-depth datasets while limiting false positive calls, and compared its performance to three benchmarked algorithms - Varscan2 [27], JointSNVmix [29], and MuTect [31].

2.5. Design and implementation

2.5.1. Design

The purpose of SNooPer is to distinguish sequencing errors (false positives - FPs) from actual somatic variants (true positives - TPs) in matched normal-tumor sequencing data. SNooPer uses a Leo Breiman RF classifier [33] which was chosen because of its limited tendency to overfit training data [33], its efficient management of very large datasets and its capability to cope with unbalanced datasets, in which one class (in this case sequencing error) is overrepresented in comparison to the other (somatic variation). RF applies bootstrap aggregation or “bagging” (subsets of the training data are selected with replacement) on multiple decision trees grown without pruning in which each node is split based on the information provided by a subset of randomly selected features. For each variant position, 15 features expected to be informative for the identification of true somatic mutations are extracted and/or calculated from the mpileup files. The complete list of features and their descriptions are presented in Supplementary Table 1 (S1 Table). These features are divided into five main groups: i) quality bias of alternative bases (related to base and mapping phred quality values), ii) coverage and VAF, iii) location along the read, iv) strand bias, and v) others. When appropriate, features are evaluated with respect to reference bases at the same position (vs_ref). To reduce over-fitting on training data and when possible, instead of absolute values, features are normalized using the corresponding median value calculated from randomly extracted subsets of variants from corresponding mpileup files (vs_med). For each model, features are ranked and selected by measuring information gain (IG) or Kullback–Leibler divergence [34] with respect to the class (InfoGainAttributeEval method, Weka suite [35]). Given that the relative importance of features for prediction may vary depending on the dataset or the genomic region of interest, the flexibility of SNooPer allows a new set of features to be selected in the training of each new model. By default, during each training phase and using the remaining bootstrap datasets (unused portion of the bootstrap as a test set), RF estimates the generalization error using the out-of-bag (oob) error as an internal control. Once trained, the model is saved and applicable to any new dataset presented to SNooPer. In the event that validation subsets are not readily available, we have also developed a series of pre-trained classification models that can be used to call variants from most datasets, including those obtained from other cancer types.

2.5.2. Code implementation

SNooPer is written in the Perl programming language and has a few dependencies: Math::CDF, Text::NSP::Measures::2D::Fisher, Statistics::Test::WilcoxonRankSum and Statistics::R. Furthermore, SNooPer uses a RF classifier implemented in Weka suite (3.6.10 or greater) [35] and requires the Java Runtime Environment (1.5 or greater). Additional and optional filters (germline dataset and blacklisted genomic regions) require a Bedtools intersect function [36]. Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves are drawn using the R package 'pracma' (Practical Numerical Math Functions). Detailed information about how to install and run SNooPer, including all available options, are described in the supplementary information.

2.5.3. Workflow

The complete workflow of SNooPer's algorithm is shown in Figure 1 (common keywords between the figure and the following description are indicated in italics).

2.5.3.1. Somatic testing and feature extraction

SNooPer expects both normal and tumor files in SAMtools mpileup format (*Pileup T* vs. *Pileup N* in Figure 1). To call variants as somatic, a *Fisher's exact test* is applied to compare the distribution of reads supporting the reference and the alternative allele between normal and tumor samples. Optionally, SNooPer can integrate two additional filters input as BED format files (*Bedtools* step in Figure 1) to exclude overlaps with any provided germline dataset (e.g. common polymorphisms from 1000 Genomes dataset [37]) or blacklisted genomic regions (e.g. poorly mappable regions from the RepeatMasker sequence [38]). Using the default parameters of *quality filters*, the algorithm only considers positions presenting at least one read (mapping quality value - MQV ≥10) supporting the alternative allele (base quality value - BQV ≥20), and requires a minimum coverage of 8X in both the tumor sample and its normal counterpart. *Features extraction* (S1 Table) is then make for each putative somatic variants that passed these filters.

2.5.3.2. Training phase

During this phase, identified variants are divided into two classes according to *orthogonal validations*: a false positive class (errors) and a true positive class (validated variations). This dataset is then used to train the *RF classifier*. To improve time-effectiveness, the default number of trees used to build the model is limited to 300 (see Results). At each node, Log2(total number of attributes)+1 features are randomly selected. The oob error rate is used as an unbiased estimate of the classification error as trees are added to the forest during training. The classification error rate is also controlled by default using a *10-fold cross-validation*estimator. Informative features for the classification are selected by measuring *information gain* or Kullback–Leibler divergence. ROC and PR curves (*Training curves*) and the related Areas Under the Curves (AUCs) are calculated for each training run (S1 Fig). Furthermore, SNooPer was designed to allow variable VAF intervals for targeted training as well as *cost-sensitive learning* to compensate unbalanced data and allow for high sequencing error rates. For discovery, users can also vary the cost of false negatives and false positives to reflect more liberal or conservative modeling. The trained *model* can be saved and applied to any subsequent dataset.

2.5.3.3. Calling phase

During the calling phase, the trained *model* as well as new tumor and matched normal mpileup files are used as input. A *Fisher's Exact Test* is performed (*Pileup T* vs. *Pileup N*) to identify putative somatic variants. Features that have been used to train the model of interest are calculated from the mpileup files for each of the putative variants and the *model* is applied for classification. The calling phase outputs a VCF file, which includes the somatic p-value from the Fisher's exact test, a categorical annotation of prediction ("PASS" or "REJ") and associated class probability (from 0.5 to 1) for each somatic variant identified, allowing the user to adjust numerical filters with more flexibility than that allowed by categorical predictions.

SNooPer's run-time efficiency is acceptable. For example, to run an entire training phase using 250 TPs and 30,000 FPs from 4 sets of whole-exome sequencing (WES) data as input (12 matched normal-tumor pileup pairs) and 300 trees and a 10-folds cross validation as training parameters, the algorithm runs for about 8 hours on a standard 12-core computer workstation with 24 Gbytes of memory, each core running at 2.667 Ghz. The time taken by the Random Forest increased linearly with the number of trees built: 0.58, 8.43, 24.45, 50.45 and 83.22 minutes were needed to build 10, 100, 300, 600 and 1,000 trees, respectively (these periods of time excluded the time taken for the calculation of features which relies on the size of the training dataset, not on the the number of trees used). Finally, during a standard calling phase, using a single-core (2.667 Ghz), SNooPer analyses approximately 5,000 pileup lines per minute.

2.6. Results and Discussion

2.6.1. Classifier performance assessment

For the development and assessment of SNooPer, we used a series of real NGS datasets from 40 unrelated childhood acute lymphoblastic leukemia (cALL) patients (Fig 2 and S2 Table). All study subjects were French-Canadians of European descent from the established Quebec cALL (QcALL) cohort [39]. For each patient, bone marrow and blood samples were collected at diagnosis prior to treatment (patient tumor) and at remission (matched patient normal). DNA was extracted using standard protocols [40] and sequenced on the Life Technologies SOLiD 4 System to constitute Dataset 1 (mean coverage on targeted region =30X). 12 cALL patient genomes (6 tumor-normal), overlapping Dataset 1, were also sequenced by Illumina, Inc. on the HiSeq 2000 (mean coverage =90X) and considered as orthogonal validation for Dataset 1. Finally, 2 samples sequenced at higher depth on the Illumina system (HiSeq 2500, mean coverage of 200X), overlapping Datasets 1 and 2, were used as validation for Dataset 2 (Fig 2 and Supplementary information for details). To test our somatic caller, we generated 4 model scenarios constructed using these 3 datasets (Fig 2 and Supplementary information for details). These scenarios were constructed to test the effect on training of the RF classifier of either a variation of the number of trees (models 1A vs. 1B), of the skewness of the (unbalanced) datasets (models 1B vs. 1C), or of the sequencing depth and technology (model 2 vs. model 1A).

2.6.1.1. Classifier training

10-fold cross-validation was used to compare the performance of SNooPer’s classification based on the training for each Model (Fig 2, Supplementary information). ROC and PR curves were generated and the related AUC was measured on each training dataset (Fig 3A, 3B, 4A and 4B). Cohen's kappa coefficient [41] was also used to assess the performances of

SNooPer’s RFs under each modeled condition (Fig 3C and 4C). To assess SNooPer’s ability to classify an unbalanced test set while being trained with a reduced and balanced training dataset, we constructed Model 1C using 250 true and false positive calls from the training set. To cope with the unbalanced test set on which the model was applied, we weighted training instances (stronger cost on false positives) using SNooPer's cost sensitive training option.

Evaluation of the oob error rates for Models 1A, 1B, and 1C (0.003, 0.003 and 0.022, respectively), suggested powerful classification performances for SNooPer’s RF. ROC AUCs (0.9724, 0.9783 and 0.9815), PR AUCs (0.7933, 0.8059 and 0.9817) and Kappa coefficients (0.7824, 0.7882 and 0.8600) also showed good agreement for SNooPer’s RF under Models 1A, 1B, and 1C, respectively. Improved training statistics for Model 1C were due to a strong reduction of the number of false positives in the training dataset from 30,000 to 250.

We compared classification performances of SNooPer’s RF to two other decision tree generators: we trained Dataset 1 using the C4.5 algorithm [42] (J48 in Weka suite) and SimpleCart [35]. For C4.5 classification, a confidence factor of 0.25 was used for pruning and we set a minimum of two instances per leaf. For SimpleCart, a minimum number of two observations at the terminal nodes was used with 5-fold internal cross-validation. C4.5 and SimpleCart trainings clearly underperformed RF with ROC AUCs of 0.8834 and 0.8343 respectively (Fig 3A).

To investigate how coverage, sequencing technologies and post-sequencing data processing (e.g. mapping method) may influence SNooPer’s performance, we constructed Model 2 using Dataset 2. The training phase for this Illumina whole-genome sequencing (WGS) dataset (mean coverage of 90X) returned a oob of 0.001, a Kappa coefficient of 0.9344 and ROC and PR AUCs of 0.9643 and 0.8662.

Firstly, to evaluate the influence of coverage on SNooPer's classification performances, we constructed artificial test datasets. The coverage of Dataset 2 was reduced by 10% (~81X) to 80% (~18X), through subsampling using SAMtools [43]. We found no clear decrease in performance except at sequencing depths below 18X (80% reduction) as illustrated by a PR AUC of 0.7982 and a Kappa coefficient of 0.8237 (Fig 4). Interestingly, the best overall performance was observed at 45X (50% reduction) with ROC and PR AUCs of 0.9918 (2nd best) and 0.9297 (best) and a Kappa coefficient of 0.9423 (2nd best). At ~36X (40% reduction in coverage), SNooPer performances were better than those obtained for Dataset 1 (mean 30X depth coverage) with ROC, PR AUCs and a Kappa coefficient of 0.9834, 0.8818 and 0.9130 compared to 0.9724, 0.7933 and 0.7824 obtained using Model 1A on Dataset 1. The improved performance is likely due to differences in sequencing and post-sequencing data processing methods, suggesting that inherent sequencing platform and/or mapping biases can influence SNooPer’s classification. Overall, and despite slight variations between Datasets 1 and 2, evaluation of the performance of the classification model yielded satisfying results across distinct datasets and sequencing technologies, further highlighting the flexibility of SNooPer’s classification model.

2.6.1.2. Comparison with other methods

To achieve an accurate and unbiased estimate of the performance of SNooPer in predicting somatic variants, and to compare SNooPer to other routinely used somatic single nucleotide variant (SNV) callers including Varscan2 255 [27], JointSNVMix [29] and MuTect [31] (Supplementary information), we randomly excluded whole exome sequencing data from

Dataset 1 before training and used it as test set (Fig 2, Supplementary information). This test set is a particularly demanding dataset given its severely unbalanced class distribution, with approximately 1 true somatic variation per million false positives presenting at least one supporting read (TP/FP =9.3E-07).

To accurately compare the performances of different algorithms, recall values were fixed for all callers and we estimated the precision (fraction of real calls) of each algorithm on the test dataset. Data were filtered on numerical values for all callers instead of on categorical variables only (Supplementary information). To evaluate the predictive performance of each somatic SNV calling algorithm, we generated PR curves and assessed the related AUCs (Fig 5). Regardless of the trained model used, SNooPer outperformed all other callers on this test dataset. The lowest AUC obtained for SNooPer (0.5732) was obtained using Model 1C while JointSNVMix, Varscan2 and MuTect reached AUCs of 0.3930, 0.1768, and 0.0491 respectively. SNooPer Models 1A and 1B, trained using 300 and 1,000 trees respectively, showed very similar performances with AUCs of 0.6310 and 0.6517. For Model 1C, reweighting of false positives correctly compensated the bias that was generated from the use of a balanced training set that was not representative of the test set. Overall, the use of SNooPer's RF classification algorithm lead to efficient identification of clonal and subclonal somatic variations with VAFs ranging from 0.16 to 0.58, with low false discovery rates of 0.363, 0.342 and 0.367 for Models 1A, 1B and 1C, respectively (mean false discovery rate - FDR over tested points). Among the 3 models, only Model 1C missed a mutation with low VAF (0.16). Limited performances and high mean FDRs observed for other methods ($FDR_{Varscan2}$ =0.527, $FDR_{JointSNVMix}$ =0.822, FDR_{MuTect} =0.945) were probably due to the suboptimal quality of SOLiD sequencing data with high sequencing/mapping error rates and low coverage (mean coverage of ~30X) for the callers' standards. More specifically, given the limited power of the strand bias feature to discriminate true positive calls from errors in Dataset 1 (see Feature Selection section in Supplementary information and S1 Table), methods such as Varscan2 and MuTect that rely significantly on this feature to call variations were expected to underperform on these data. Varscan2 filters out variants with >90% of the supporting reads originating from the same strand, and MuTect applies a restrictive strand bias filter based on a separate calling step on each strand implemented to avoid variants supported by a biased alignment. As expected, SNVs that were missed by these two algorithms were all positions that were affected by strand bias. Still, despite its strong strand bias filter, Varscan2 showed the best overall performance of the three benchmarked algorithms that were tested here. On the other hand, MuTect is known to be a very sensitive SNV caller that is powered to detect low VAF mutations. However, as illustrated in Supplementary Figure 2 (S2 Fig), the VAF distribution of somatic MuTect variations was clearly skewed toward very low VAFs compared to the distribution of true somatic SNVs present in the test set, leading to a large number of false positives in the MuTect output. A similar pattern with an increase in low VAF calls (<0.2) was observed for JointSNVmix, also resulting in increased false positive calls. Unlike other callers, the SNooPer algorithm involves a training phase where class assignment is directly learned from the dataset at hand, and this translated into a VAF distribution that matched the true positives distribution. Moreover, under Models 1A and 1B, SNooPer identified less than 90 somatic SNVs that included all true somatic SNVs present in the test dataset, while MuTect (power ≥0.16) identified 274 somatic variants, Varscan2 (somatic p-value ≤0.17) 397, and JointSNVMix (P(somatic) ≥0.29) identified 705 somatic SNVs, which included 92%, 83% and 100% of the true somatic variants, respectively. SNooPer's somatic SNV calls under Model 1A and 1B were thus more precise and no true somatic variants were missed, further highlighting its superior performance. With a higher sensitivity and specificity for somatic SNV detection in our low quality test set (mean coverage <30X), SNooPer outperformed commonly used somatic variant callers such as Varscan2, JointSNVMix and MuTect. Importantly, this report was not meant to question the performance of benchmarked callers that have proven to be efficient and that classically show satisfactory results with high coverage datasets.

2.6.1.3. Real data analysis

We then evaluated our trained Model 1A on the remaining data from Dataset 1 that consisted of 34 B- and T-cell cALL patients (S2 Table). To identify somatic variations with high driver potential, only predicted deleterious SNVs with Sift [44] p-values ≤ 0.05 were considered. 50 heterozygous candidate SNVs (VAF <0.6) presenting a class probability over 0.9 and a coverage of at least 15X in the normal sample were randomly selected for validation. These variations showed coverage values ranging from 23 to 115X (mean coverage 51X) and VAFs ranging from 0.10 to 0.57 (median 0.38). For orthogonal validation of this dataset, we used targeted ultra-deep sequencing (Illumina) of the patient's tumor material (>1000X) and of the normal counterpart in order to confirm the somatic nature of each of the identified variants (Fig 6, Supplementary information).

A total of 90% (45/50) of the tested SNVs were confirmed real variants, that is found in the tumor material of the patient following our filtering criteria (see Methods). Among these 45 variations, 80% (36/45) were validated somatic mutations (present in tumor only) and 20% (9/45) were identified as germline. Overall, if the confirmed somatic variations are considered true positives and the errors (no calls in re-sequencing) combined with germline variations are considered as false positives, SNooPer's somatic SNV identification reached a precision of 0.71 (see Methods). As expected, the identified germline mutations had VAFs around the expected clonal heterozygous allele frequency of 0.50 with a mean VAF of 0.47 and a variance (σ^2) of 0.004, while the confirmed somatic mutations had a lower mean VAF of 0.36 associated with a wider distribution (σ^2 =0.009) averaged from the different subclones present in the sample. Importantly, SNooPer showed no bias of performance in calling subclonal SNVs with low VAF with 2 FPs under and 3 FPs over the median VAF of 0.36, and reached a precision of 0.90 for mutations located within the lower 50th percentile.

2.7. Conclusion

Most available somatic SNV calling methods offer user-defined categorical filters or at best, numerical filters to fine-tune or customize SNV calling, however these can have a strong influence on the output. SNooPer does not rely on user-defined parameters and in doing so, allows versatility and flexibility to cope with complex datasets. Here, the model is directly built around the data itself therefore limiting any bias or subjectivity in somatic mutation calling. Firstly, although systematic errors in the training dataset are likely to exist, the use of an independently sequenced (different technology, mapping) validation dataset will teach SNooPer to recognize systematic errors from the original dataset and to classify them as false positive. Therefore, this method leads to a by-default elimination of systematic errors associated to each sequencing platform. Furthermore, rather than using standardized filters, the importance of each feature for variant classification is directly measured from the data. While any RF algorithm includes by default attribute selection, we also provided the possibility for users to perform a dimensionality reduction of features based on information gain. In doing so we reduce the chance of false positive occurrence due to a strong yet biased feature, which may, in part, explain SNooPer's superior performance compared to other tested callers. Moreover, SNooPer can accommodate reduced training datasets, such as the one constituted of 250 false and true positives used here, compensate the balance bias using cost sensitive training, and still outperform other commonly used somatic variant callers. Although not reported here, SNooPer also offers an Indel training algorithm and the corresponding calling option that is available in the latest released version. Finally, given that sequencing errors have been linked to homopolymers or G-rich sequence motifs, an updated version of the software that considers the context of genomic coordinates is under development.

As NGS moves toward the clinic and proves its usefulness as a powerful diagnostic tool, whole-genome approaches remain limited to rapid low-pass whole-genome sequencing as a cost-compatible compromise. Sensitive calling algorithms such as SNooPer that is tailored around the data, will thus be indispensable to weed out true somatic variants and identify potential driver mutations or actionable targets. SNooPer was developed in response to this need and has already proven its utility in identifying novel mutations in childhood leukemia [45-47].

2.8. Supplementary information

2.8.1. Datasets

Dataset 1 consisted of 80 cALL patient exomes (40 tumor and 40 normal) (Fig 2 and S2 Table). Whole exomes were captured in solution with Agilent's SureSelect Human All Exon (50Mb) kit according to the manufacturer's protocol and sequenced on the Life Technologies SOLiD 4 System (mean coverage on targeted region =30X) at the Integrated Clinical Genomic Centre in Pediatrics, CHU Sainte-Justine. Reads were aligned to the Hg19 reference genome using LifeScope Genomic Analysis Software. PCR duplicates were removed using Picard [1]. Genotype quality score recalibration was performed using the Genome Analysis ToolKit (GATK) [2]. After filtering out low quality reads, mpileup files were created from BAM files using SAMtools [3] (Fig 1).

Dataset 2 consisted of a subset of 12 cALL patient genomes (6 tumor and 6 normal) overlapping Dataset 1 (Fig 2 and S2 Table). Whole genomes were sequenced by Illumina, Inc. on the HiSeq 2000 (mean coverage =90X); resulting reads were aligned to the Hg19 reference genome using the Illumina Casava software. Bam files were cleaned and mpileup files created as above. SNVs were called from cleaned BAM files using GATK HaplotypeCaller [2] and filtered according to the Broad Institute recommendations (QD <2.0, MQ <40.0, FS >60.0, MQRankSum <-12.5 and ReadPosRankSum <-8.0).

Dataset 3 was composed of 2 samples sequenced at higher depth on the Illumina system and overlapping Datasets 1 and 2 (Fig 2 and S2 Table). Here, exomes were captured using Illumina's Nextera Exome Enrichment kit following the manufacturer's' protocol and sequenced on the HiSeq 2500 at the Integrated Clinical Genomic Centre in Pediatrics, CHU Sainte-Justine with a mean coverage of 200X. Mapping to the Hg19 reference genome was performed using Bowtie2 [4] and BAM files were cleaned and mpileup files created as above. SNVs were called using GATK HaplotypeCaller as described for Dataset 2.

The sequencing quality reports are available upon request. For all 3 Datasets, the genomic regions considered for further analysis and comparison were limited to the NCBI's Reference Sequence (RefSeq) [5].

2.8.2. Models

To train the 4 distinct models used here (model 1A, 1B, 1C and 2), we considered the GATK HaplotypeCaller [2] output of Dataset 2 as the orthogonal validation of Dataset 1 (Fig 2). Although datasets 2 and 3 shared similar sequencing technologies (Illumina) and were therefore not orthogonal, given the differences of chemistry, platforms, coverage and mapping processes, we considered Dataset 3 to be a reliable validation set for Dataset 2. To construct models 1A, 1B and 2, 30,000 positions presenting alternative bases in the mpileup files of the test Datasets (1 or 2) and not identified by HaplotypeCaller [2] in the validation Datasets (2 or 3) were randomly selected and considered as false positives. Conversely, 250 overlapping mutations between validation and test sets were considered as true positives. 300 trees were used to construct models 1A and 2 and 1,000 trees for model 1B. To construct model 1C, a balanced training set consisting of 250 false and 250 true positives was used. For all our analyses, we performed paired normal/tumor somatic analysis and further filtered out variants that overlapped with 1000 Genomes (2012) [6]. Furthermore, overlaps with the RepeatMasker sequence obtained from UCSC genome browser [7] were excluded to avoid putative miscalled variants located in repetitive elements, including low-complexity sequences and interspersed repeats. To determine the optimal number of trees to be generated, we gradually increased this value (from 100 to 1,000) and determined that, as the number of trees grew beyond 300, classifier performance was only slightly increased at the expense of processing time. Therefore, the default number of trees for SNooPer's RF was set to 300.

2.8.3. Comparison with other methods

Given that classes (TP or FP somatic SNV calls) used to train SNooPer's RF model were based on the GATK HaplotypeCaller analysis of Dataset 2, and in order not to favor SNooPer over the other algorithms tested, we used an independent somatic mutation caller (Strelka [8]) for somatic SNV analysis in the test dataset. Overlapping mutations between Datasets 1 and 2 with a VAF >0.10 in Dataset 1 and confirmed as somatic by Strelka in Dataset 2 were considered as true positives; overlapping non-somatic mutations were omitted. Conversely, variant positions in Dataset 1 that were not identified in Dataset 2 were considered as false positives. Positions matching the described criteria were retained in the original Bam files for further assessment by SNooPer and comparison with other methods.

Using this dataset, we compared SNooPer to 3 benchmarked somatic SNV callers: i) Varscan2 (version 2.3.6) [9] was run in somatic mode using a pipe from SAMtools mpileup with a minimum mapping quality value (minBaseQ) of 10, the strand bias filter turned on (strand-filter = 1), a minimum of tumor (min-coverage-tumor) and normal coverage (min-coverage-normal) of 10; ii) JointSNVMix and JointSNVMix2 (version 0.7.5) [10] were first trained ('train' mode) with default parameters (including 'min_normal_depth' and a 'min_tumour_depth' of 10) to tune the parameters and were then run in 'classify' mode. Optional parameters 'minimum base quality' (min_base_qual) and 'minimum mapping quality' (min_map_qual) were set to 20 and 10, respectively. JointSNVMix yielded more accurate results in terms of sensitivity/specificity and was therefore used in our comparative analysis; we did not consider JointSNVMix2 any further; iii) MuTect (version 1.1.4) was run in high confidence (HC) mode with COSMIC version 54 [11] and dbSNP132 [12] as input. We also ran MuTect in "artifact-detection-mode", which increased sensitivity at the expense of specificity, therefore this option was not considered for the comparative analysis.

SNooPer performance was assessed and compared to the performance of each benchmarked method by measuring the false discovery rate (FDR), as well as the precision, defined as the ratio of the number of real variants retrieved (TP) to the total number of real variants and errors retrieved (TP+FP), and the recall, defined as the ratio of the number of real variants retrieved (TP) to the total number of real variants in the dataset (TP+FN):

$$FDR=\frac{FP}{FP+TP} \quad Precision=\frac{TP}{TP+FP} \quad Recall=\frac{TP}{TP+FN}$$

For SNooPer, we kept somatic SNVs (p-value =0) flagged as ”PASS” and varied the class probabilities (from 0.537 to 0.987, 0.519 to 0.949 and 0.756 to 0.930 for models 1A, 1B and 1C, respectively). For Varscan2, only variants identified as “somatic” were considered and we evaluated the output based on somatic p-value that varied from 0.1669 to 3.2577E-8. JointSNVMix analysis was evaluated based on different somatic P-values that varied from 0.2901 to 1. Finally, we considered mutations annotated as “KEEP” and “COVERED” by MuTect and varied the “power” (from 0.1614 to 0.9994). Mutect is very restrictive on its somatic SNV calls and selection of variants annotated as “NOVEL” only resulted in undercalling of true somatic variants.

2.8.4. Real data analysis

To assess SNooPer’s performance on a real dataset, we called somatic SNVs on a dataset consisting of 34 cALL patient exomes (68 matched normal-tumor samples) that were sequenced as in Dataset 1 (above). For orthogonal validation, somatic SNVs identified using SNooPer were then subjected to ultra-deep targeted resequencing using Illumina TruSeq Custom Amplicon assay as per the manufacturer’s instructions. Illumina DesignStudio was used to design custom oligos targeting 50 randomly selected high confidence SNVs called by

SNooPer. PCR purification was performed with Ampure Beads for 150bp amplicon size selection. Double stranded amplicons were pooled, quantified by qPCR and sequenced on the Illumina HiSeq 2500 system (paired-end: 2x100bp) to reach a minimum of 1000X coverage. Sequenced reads were aligned to the Hg19 reference genome using Bowtie2 [4] and variants were called using Varscan2 "mpileuptosnp" analysis [9]. Somatic SNVs identified in tumor only were considered as validated, while variants present in both tumor and normal were considered germline.

2.8.5. Feature selection

Features calculated from Dataset 1 were ranked according to their IG (S1 Table). Different threshold values of IG were measured to filter out less informative features and best results were obtained by eliminating features with less than 0.001 bits of IG. For this dataset, 3 features presented more than 0.01 bits of IG: high quality VAF (allelic_freq_highqual), IG =0.0419; p-value obtained from a Wilcoxon rank-sum test comparing the base quality value (BQV) of reference bases versus alternative bases (GQBprob), IG =0.0142; mean mapping quality value (MQV) of alternative bases (var_mean_mqv_quality), IG =0.0121. Here, VAF was of particular importance for SNooPer’s classification performance: using a wide range of VAFs rather than a binary cutoff led to more efficient calling of subclonal variations (see below). Interestingly, the first feature belonging to the strand bias subgroup (Sbpbinom) only ranked 9th with an IG of 0.0022. This highlighted the limited importance of the strand bias for classification of these data. This could be explained by limited sampling (low coverage) on captured sequences known for their tendency towards artificial strand biases. Given the importance of strand bias in other algorithms tested here, this could lead to misinterpretation of certain performance comparisons.

2.8.6. Installation and usage

INSTALLATION

To install this module, run the following commands:

```
 perl Makefile.pl
 make
 make test
 make install
```

SYNOPSIS

>SNooPer.pl -help [brief help message] -man [full documentation]

>Training:

>SNooPer.pl -i [input_directory] -o [output_directory] -a1 [type_of_analysis1] -a2 [train] -w [path_to_weka] [options]

>Classify/Evaluate:

>SNooPer.pl -i [input_directory] -o [output_directory] -a1 [type_of_analysis1] -a2 [classify/evaluate] -m [model] -w [path_to_weka] [options]

DESCRIPTION

>SNooPer requires a training phase during which a training dataset (a subset of validated positions) is used to construct a model that can be then applied to call variants on an extended test dataset.

>For the training phase ("train"), the user must provide 2 types of files:

>1.pileup files (.pu) with similar characteristics as the test dataset on which the trained model will be applied.

>Somatic analysis format: tset_T_sample_id.pu and tset_N_sample_id.pu

>Germline analysis format: tset_sample_id.pu

>2.vcf files (.vcf) validation files that are ideally orthogonal validations of the positions contained in the pileup files.

>Somatic analysis format: vset_T_sample_id.vcf

>Germline analysis format: vset_sample_id.vcf

>Each position in the pileup files must be tested a priori so that the class (true variant or sequencing error) is known by comparison with the vcf files.

If a variant is present in the corresponding validation file, it will be considered as an actual variant. If the variant is absent from the validation file, the variant will be considered as an error.

>To be considered as the corresponding validation file of a .pu file, the .vcf file has to present the same sample_id.

>For the classification phase ("classify") or to evaluate a model ("evaluate"), the user simply provides the paths to the model that is to be applied and to the pileup files from the test dataset:

>Somatic analysis format: tset_T_sample_id.pu and tset_N_sample_id.pu

>Germline analysis format: tset_sample_id.pu

>Note that input files must contain the prefix tset_ (for training or test dataset, depending on the context) and the .pu extension or vset_ (for validation dataset) and the .vcf extension.

REQUIREMENTS

>The following programs must be installed:

-Weka; the current version of SNooPer was tested with version weka-3-6-10.

-R; the current version of SNooPer was tested with version R/3.2.1

-Bedtools if BlackList (-r) or germDB_track (-g) options are applied. The current version of SNooPer was tested with version bedtools-2.17.0.

>For the development and testing of SNooPer:

The BlackList track corresponded to the RepeatMasker track downloaded from UCSC. "Assembly" has to be set according to the reference used to map your sequences, "Group" was set to Variation and Repeats, and "Track" was set to RepeatMasker. The track was downloaded in a .bed format.

>The germline database used as germDB_track corresponded to the 1000 Genomes database downloaded from http://www.1000genomes.org/. The track was formated in a .bed format.

OPTIONS

-help <brief help message>

-man <full documentation>

-a1 <type_of_analysis1> Can take the following values: "somatic" or "germline". "somatic" means that the somatic evaluation will be done based on N samples provided (and additional germline data if provided, see germDB_track -g option).

-a2 <type_of_analysis2> Can take the following values: "train", "classify" or "evaluate".

->if "train" is selected, a model will be trained based on the comparison of the training dataset (tset) and the validation dataset (vset). A subset of the data provided (subset chosen with the -v and -nv options or automatically selected) for which the class is known (0/1 = non-validated/validated = not shared by tset and vset / shared by tset and vset) will be used for training. Therefore, a partially overlapping dataset between tset and vset must be provided. Final classification of the complete data will be done base on the trained model. Furthermore, evaluation of the model will be performed using a subset excluded beforehand.

->if "classify" is selected, the provided test dataset (tset) is classified using a model created previously. This model has to be in an .arff format (see Weka documentation for more info).

->if "evaluate" is selected, the provided dataset (tset) is classified using a model created previously. The purpose of this option is to evaluate a previously created model based on the classification of an independent dataset (never used to train the model). To evaluate the model, the class of each variant in the dataset must be known. Therefore, the data from both tset and vset must be provided. These data should be located in a new directory containing these files only.

-i <input_directory> Complete path to your input directory.

-o <output_directory> Complete path to your output directory (input and output can be located in the same directory).

-m <path_to_model> Complete path to the directory of a previously trained model. This option should be set only if the type of analysis 2 is "classify" or "evaluate".

-w <path_to_weka> Complete path to the weka.jar executable.

-a3 <type_of_analysis3> [optional] Can take the following values: "SNP" or "Indel". The default value is "SNP".

-a4 <attributes_selection> [optional] Can take the following value: "off", "MI" or "BestFirst". The default value is "off". If "MI" is selected (Weka InfoGainAttributeEval + Ranker): evaluation the worth of an attribute by measuring the information gain with respect to the class + ranking of attributes by their individual evaluations. Attributes will be discarded if presenting less than 0.001 bits of mutual information. If "BestFirst" is selected (Weka CfsSubsetEval + BestFirst): evaluate the value of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them + evaluate the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility.

-b <path_to_bedtool> [optional] Complete path to bedtools binary file.

-bqv <bqv> Base quality value (phred) of a variation to be considered as "High Quality". Default value is 20.

-c <contamination> [optional] Fraction of normal cells in the tumor sample. Can take a value between 0 and 1. Default value is 0.

-cf <covered_filter_N> [optional] Can take the following values: "on" or "off". If the filter is "on", only positions with a minimum coverage of "coveragefilter_N" in the N will be considered in the T for somatic analysis. Default value is on.

-cm <cost_matrix> [optional] used to adjust the weight of mistakes on a class (see http://weka.wikispaces.com/CostMatrix). The cost matrix has to be define in a single line format using comma to separate values ex: 0.0,5.0,1.0,0.0 here the weight on false positive is 5 and on false negatives is 1.

-cn <coveragefilter_N> [optional] Defines the minimum of coverage for a position to be considered in the N files during a Somatic analysis or the Germline analysis. If a position in the T file doesn't reach the coverage limit in the N file, the position can't be call Somatic and won't be considered. Default value is 8.

-ct <coveragefilter_T> [optional] Defines the minimum coverage required for a position to be considered in the T file during a Somatic analysis. Default value is 8.

-fi <freqinf> [optional] Defines the inferior limit of allele frequency for a variant position to be considered in the T file during a Somatic analysis. Default value is 0.

-fs <freqsup> [optional] Defines the superior limit of allele frequency for a variant position to be considered in the T file during a Somatic analysis. Default value is 1.

-g <path_to_germDB_track> [optional] Complete path to any germline variant database track. If such a file is provided and if the type_of_analysis1 is "somatic", the variations located at these positions will be considered as germline during the somatic variant calling process.

-id <job_id> [optional] The output file name will be: SNooPer_output_job_id_date.

-ind <indel_filter> [optional] Can take the following values: "on" or "off" when type_of_analysis3 is "SNP". If the filter is "on", pileup lines containing indels won't be considered during the SNP calling process. Default value is on.

-k <cross_validation> [optional] Integer to define the k-fold cross-validation used to train the model. This option must be set only if the type of analysis 2 is "train" or "classify". Default value is 10.

-mem <memory> [optional] The user can extend the memory available for the virtual machine by setting appropriate options. Ex: -Xmx2g to set it to 2GB. The user can also redirect temporary JVM files using the format: -Djava.io.tmpdir=/path/to/tmpdir

-mqv <mqv> [optional] Minimum mapping quality value (phred) of a read in order for it to be retained as "High Quality" in the variant calling process. Default value is 20.

-nN <nbvar_N> [optional] Defines the number of supporting variant reads required for a position to be considered in the N files during a Germline or Somatic analysis.

-nT <nbvar_T> [optional] Defines the number of supporting variant reads required for a position to be considered in the T files during a Somatic analysis.

-nv <nb_of_non_validated_var_to_train> [optional] Number of non-validated variants (discordant between tset and vset) used to train your model. If no value is provided, a default value will be calculated from the input file. It prevails over validated_variant_fraction and validated_nonvalidated_ratio.

-p1 <tech> [optional] Technology/chemistry used to produce the data to be classified. Can take the following values: "Solid", "Solexa", "Illumina-1.3", "Illumina-1.5" or ">Illumina-1.8". Default value is the Illumina-1.8 or higher ">Illumina-1.8".

-q <qual_filter> [optional] Can take the following values: "on", "on+", "off" or "off". If the filter is "on" or "on+", only variants matching the selected bqv and mqv values will be considered. If "on+" or "off+" are selected, all attributes will be considered including those that depend on quality. Default value is on.

-r <path_to_blacklist> [optional] Complete path to the BlackList track. This black list usually corresponds to problematic regions in the genome. If such a file is provided, the variations located in these regions won't be considered during the variant calling process.

-s <somatic_pvalue> [optional] Somatic P-value filter based on a one-tailed Fisher's exact test comparing the somatic and germline allele count. Only variants presenting a P-value < to this value will be conserved. The default value is 0.1. The value must be set between 0 and 1.

-t <tree> [optional] Number of trees to build the model. Default value is 300.

-v <nb_of_validated_var_to_train> [optional] Number of validated variants (concordant between tset and vset) used to train your model. If no value is provided, a default value will be calculated from the input file. It prevails over validated_variant_fraction and validated_nonvalidated_ratio.

-vf <validated_variant_fraction> [optional] Fraction of the validated variants to be used for training. The default value is 1. Note that if the number of validated positions is large, the analysis can be time-consuming.

-vr <validated_nonvalidated_ratio> [optional] Ratio (nb of non-validated variants / nb of validated variants) in the training dataset. The default value is 0.1. Note that, if the training dataset is extremely imbalanced, cost sensitive learning can be used to improve the algorithm’s performance.

2.9. Figures

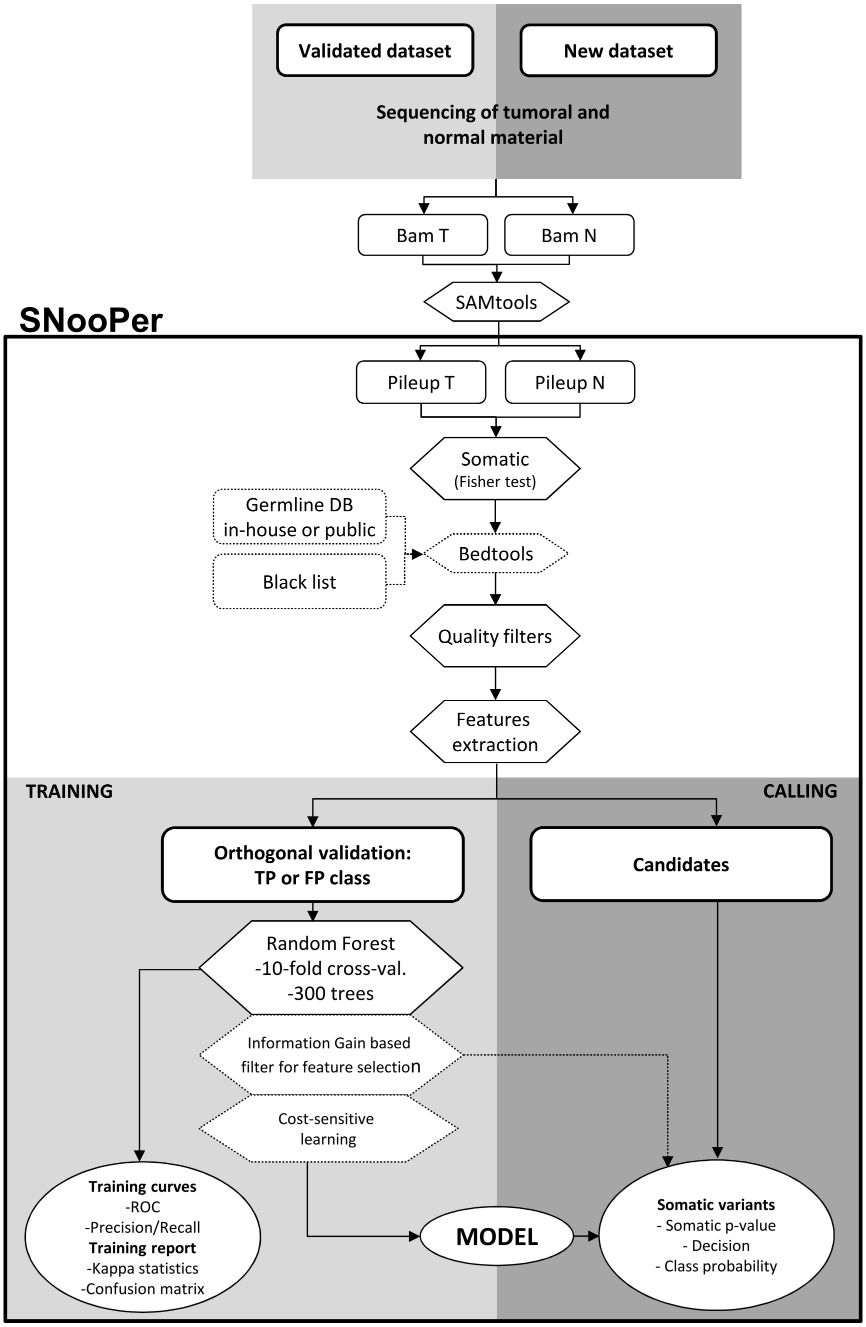
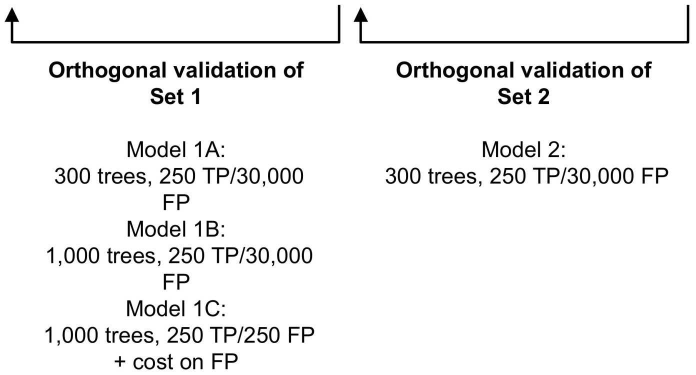


Figure 1. Workflow of SNooPer's algorithm. SNooPer uses both normal and tumor files in a SAMtools mpileup format as input. It requires a training phase in which an orthogonal validation (re-sequencing) dataset is used to train the RF classification model that is subsequently used to call somatic variations in the test dataset. Light grey boxes represent the training steps while dark grey boxes represent calling steps. Dotted boxes represent optional steps in the workflow. Circles represent the output following either the training or calling phases.

	Set 1	Set 2	Set 3
Sequenced region	WES	WGS	WES
Capture	SureSelect 50Mb (Agilent)	NA	Nextera (Illumina)
Mapper	Lifescope (Life Technologies)	CASAVA (Illumina)	Bowtie 2
Run type	Single-end	Paired-end	Paired-end
Read length (bp)	75	100	100
Sample	80	12	2
Patient	40	6	1
Mean coverage	~30X	~90X	~200X
Sequencing technology	SOLiD 4 (Life Technologies)	HiSeq 2000 (Illumina)	HiSeq 2500 (Illumina)
Sequencing platform	CHU Sainte-Justine Research Center	Illumina, Inc.	CHU Sainte-Justine Research Center



Orthogonal validation of Set 1

Model 1A:
300 trees, 250 TP/30,000 FP
Model 1B:
1,000 trees, 250 TP/30,000 FP
Model 1C:
1,000 trees, 250 TP/250 FP + cost on FP

Orthogonal validation of Set 2

Model 2:
300 trees, 250 TP/30,000 FP

Figure 2. Datasets used to develop and assess SNooPer. All 3 datasets were generated from real childhood acute lymphoblastic leukemia samples. Arrows indicate sequencing overlaps between datasets. Re-sequencing was used as orthogonal validation for the training phases of the algorithm. RF Models (1A, 1B, 1C and 2) resulting from these training phases are shown below the corresponding arrows.

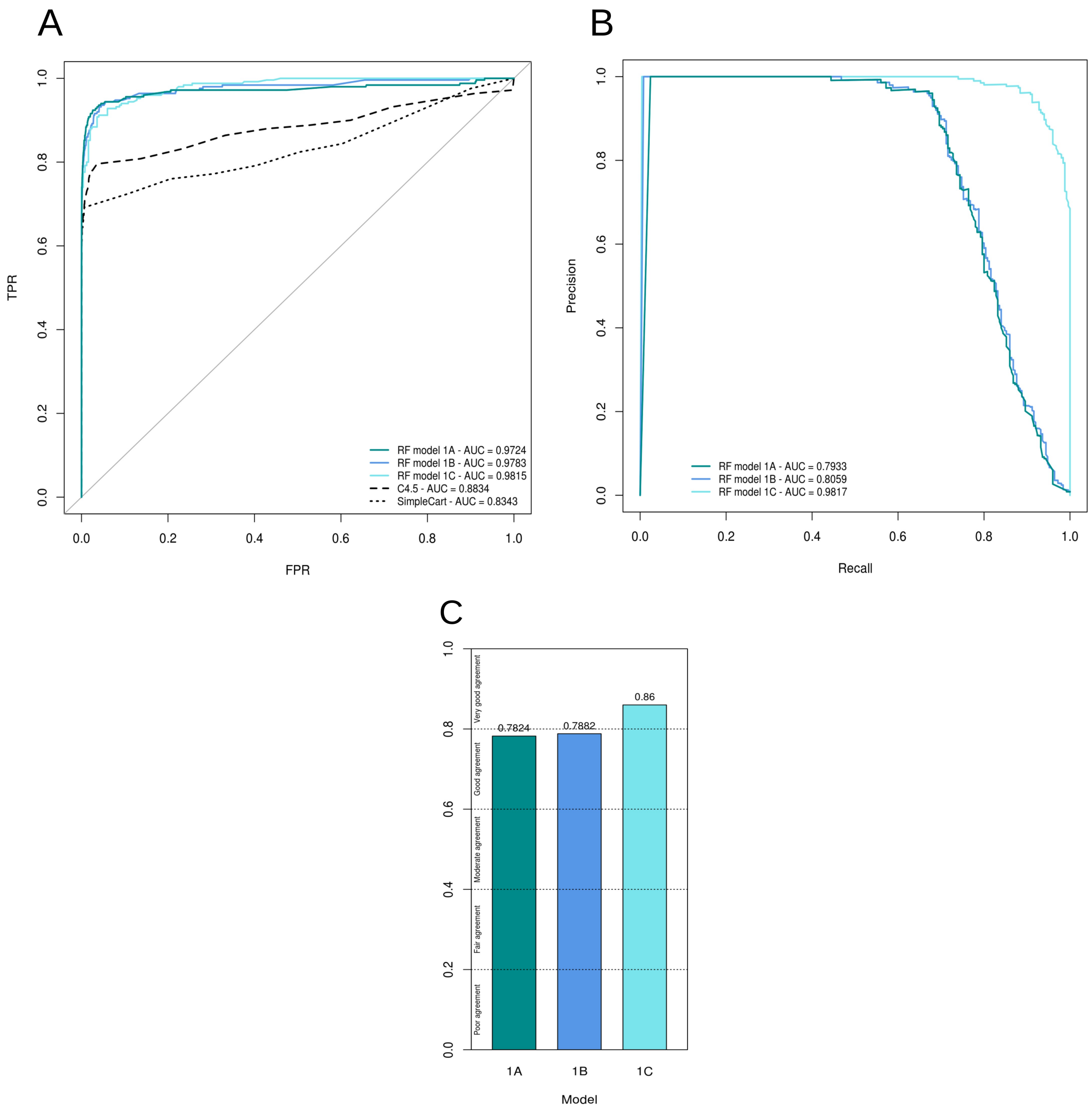


Figure 3. Training assessment of Model 1A, 1B and 1C. Data used to construct these curves were obtained from SNooPer's RF training phase using Dataset 2 as a validation set and a subset of Dataset 1 as training set. Dark cyan, blue and light blue represent SNooPer's Model 1A, 1B and 1C, respectively and AUCs are shown for each model. (**A**) ROC curves. Solid, dashed and dotted lines represent RF, C4.5 (J48) and SimpleCart algorithms respectively. TPR stands for True Positive Rate and FPR for False Positive Rate. (**B**) PR curves. (**C**) Cohen's Kappa coefficient.

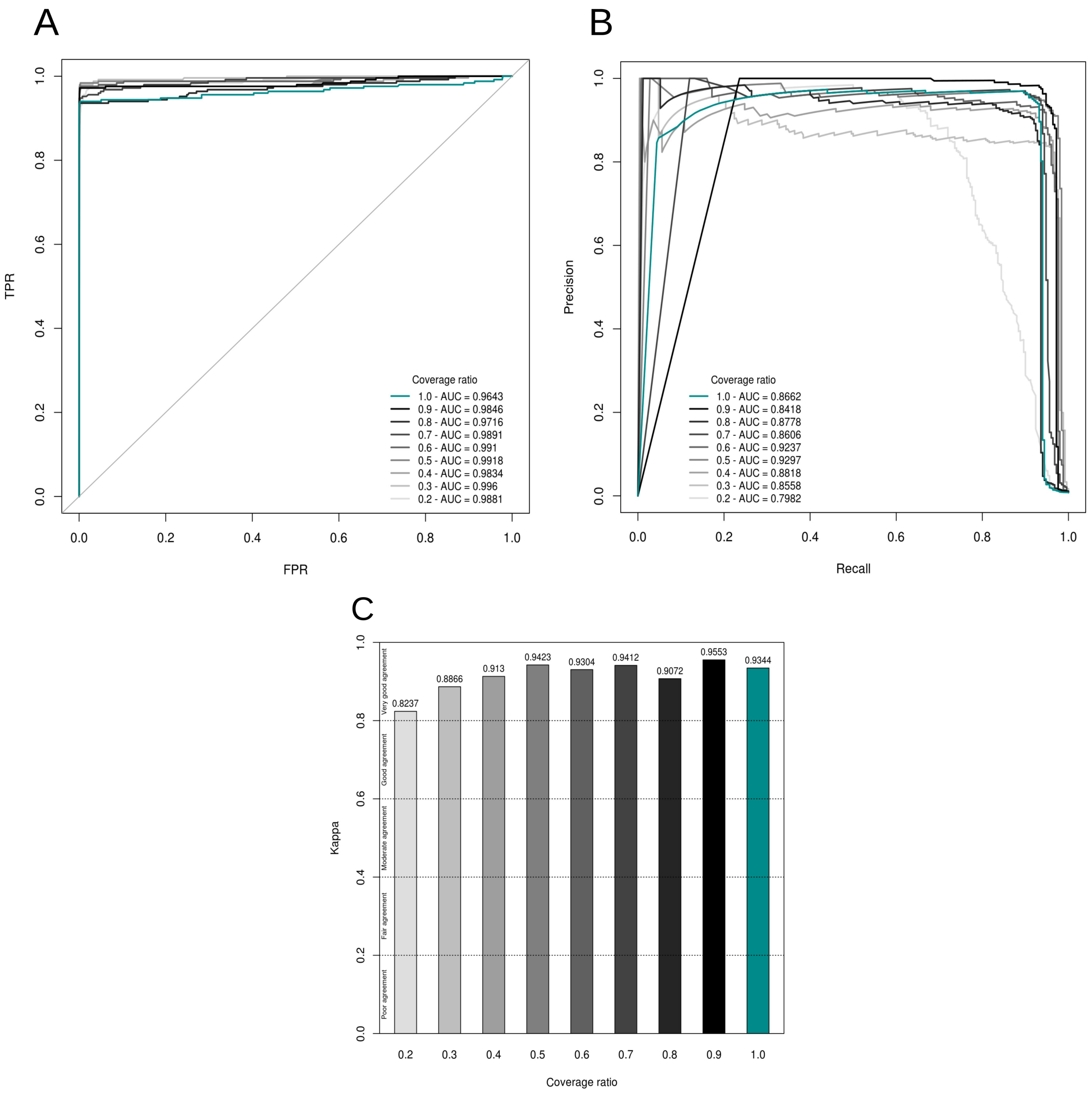


Figure 4. Training assessment of Model 2. The data used to construct these curves were obtained from training phases using Dataset 3 as validation set and either the original Dataset 2 (dark cyan) or an artificial version of Dataset 2 (shades of grey) in which the coverage was gradually subsampled from 10% (ratio of 0.9 ~81X; darkest grey) to 80% (ratio of 0.2 ~18X; lightest grey) as training set. AUCs are shown for each model. (**A**) ROC curves. Solid, dashed and dotted lines represent RF, C4.5 (J48) and SimpleCart algorithms respectively. TPR stands for True Positive Rate and FPR for False Positive Rate. (**B**) PR curves. (**C**) Cohen's Kappa coefficient.

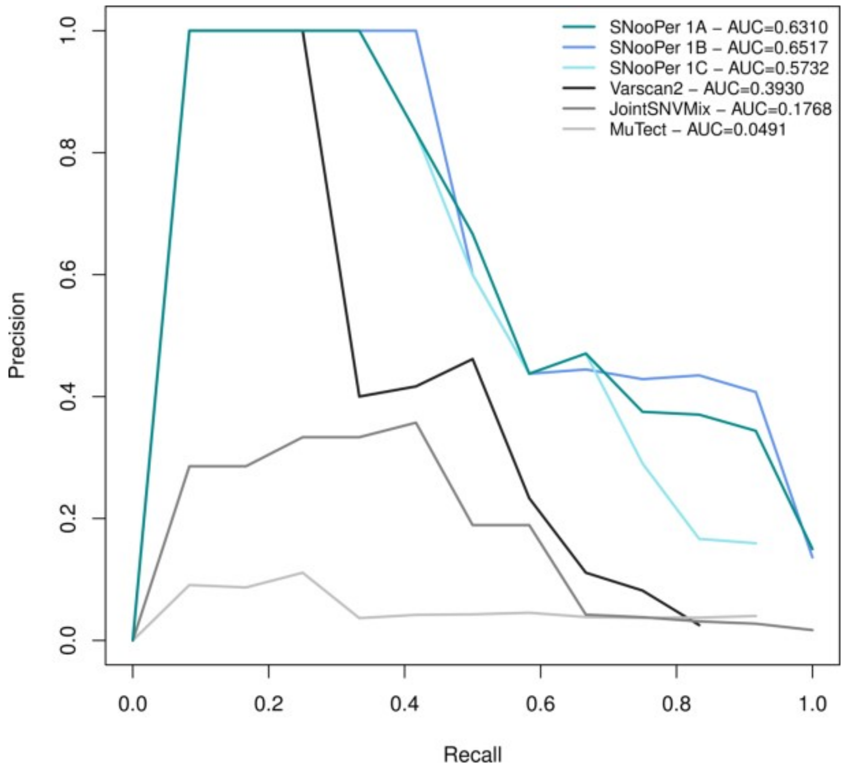


Figure 5. Precision – Recall curves for method comparison. Data used to construct these curves were obtained from SNooPer's calling phase using Model 1A (dark cyan), Model 1B (blue), Model 1C (light blue), Varscan2 (black), JointSNVMix (dark grey) and MuTect (light grey) on the test set. The test set was built using a subset of Dataset 1 that was kept completely separate during the training phase. AUCs are shown for each model.

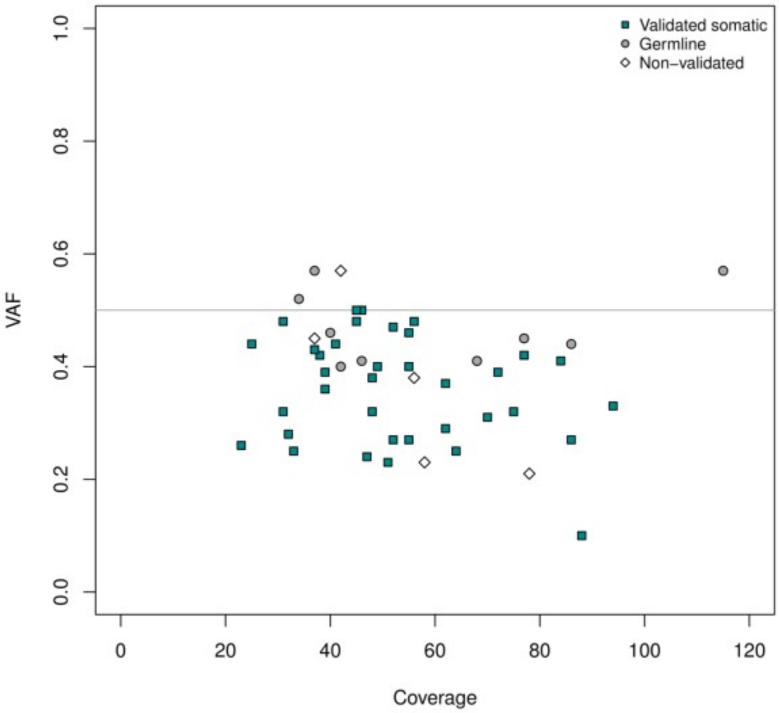
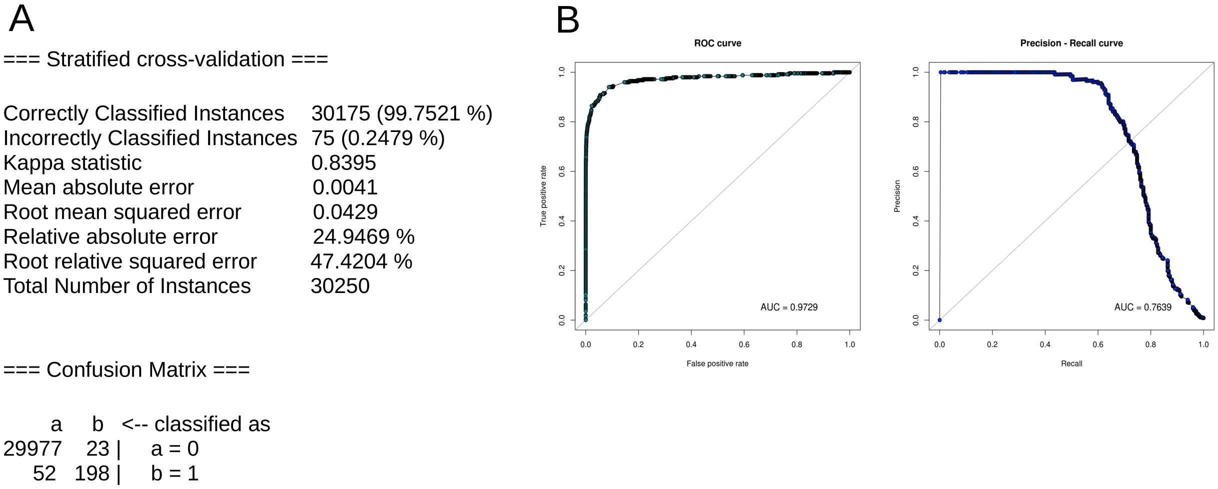


Figure 6. Validation plot. Distribution of 50 randomly selected SNVs called using SNooPer's Model 1A on the independent validation set constituted of samples obtained from 34 childhood acute lymphoblastic leukemia patients (matched normal and tumor). All selected SNVs were heterozygous with a VAF<0.6, predicted as damaging (Sift [42 44] p-values ≤0.05) and presented a class probability >0.9. Each identified SNV was validated by targeted ultra-deep re-sequencing (>1000X). The grey line indicates the expected VAF (50%) for germline or clonal somatic heterozygous variants. Dark cyan squares, grey dots and white diamonds represent validated somatic, germline and non-validated variations respectively.



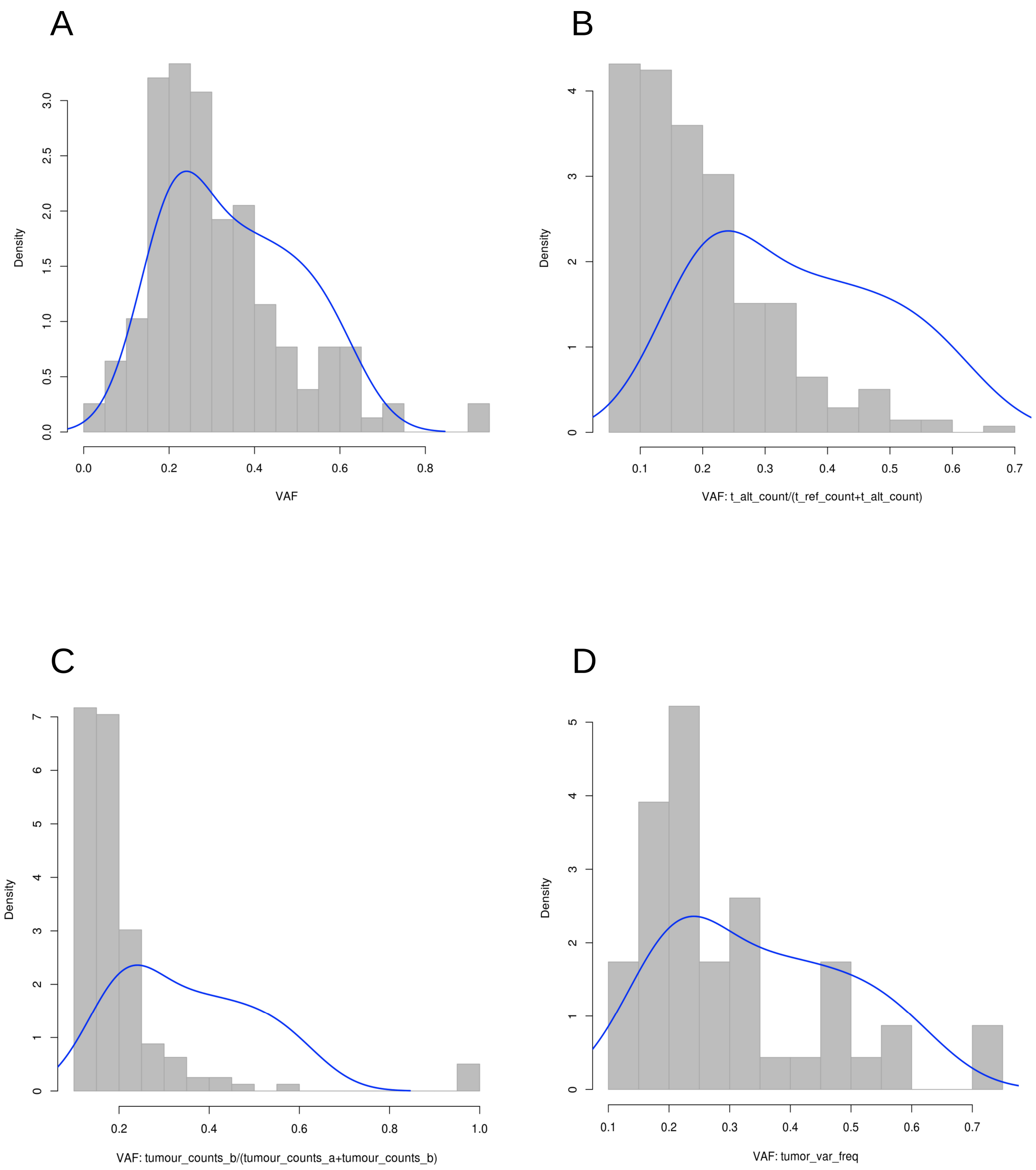
```
=== Stratified cross-validation ===

Correctly Classified Instances    30175 (99.7521 %)
Incorrectly Classified Instances  75 (0.2479 %)
Kappa statistic                   0.8395
Mean absolute error               0.0041
Root mean squared error           0.0429
Relative absolute error           24.9469 %
Root relative squared error       47.4204 %
Total Number of Instances         30250


=== Confusion Matrix ===

     a     b   <-- classified as
 29977    23 |     a = 0
    52   198 |     b = 1
```

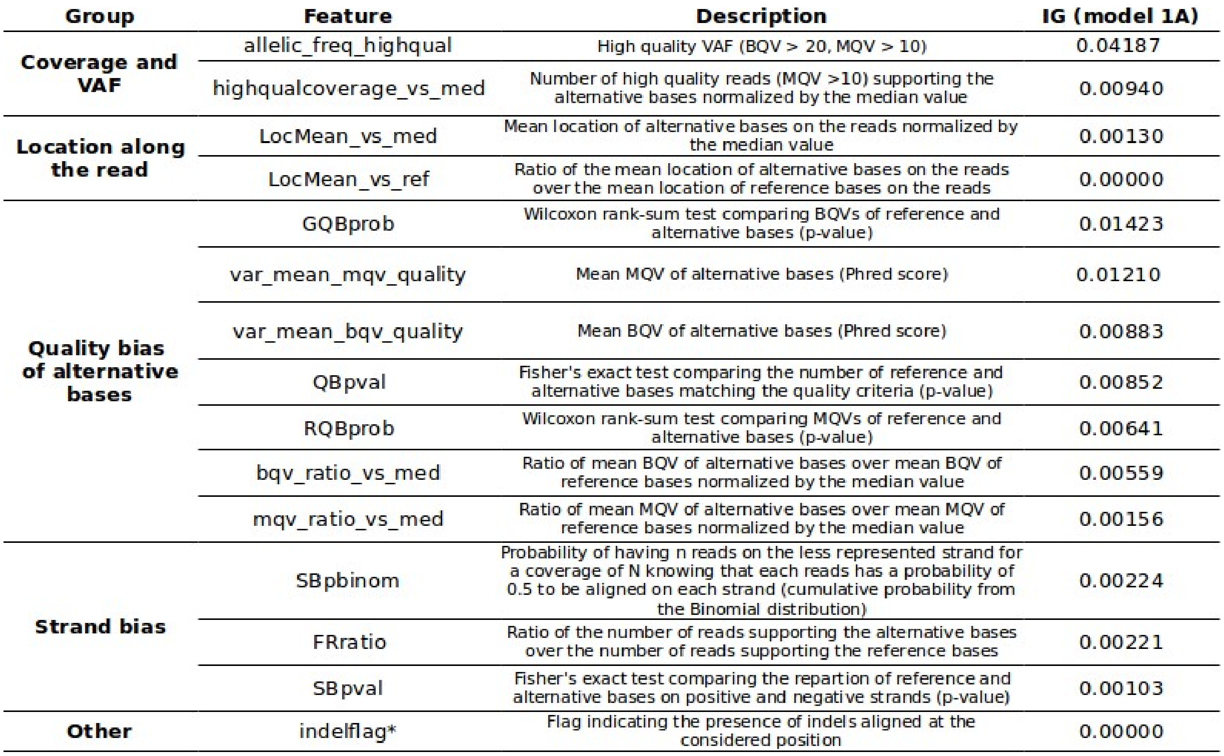
S1 Fig. Snapshot of a SNooPer output from the training phase. (**A**) General statistics including Kappa statistics (top) and the confusion matrix (bottom) obtained from a 10-fold cross validation training phase. (**B**) Receiver operating characteristics (left) and precision-recall (right) curves and the related AUCs calculated during the training phase.



S2 Fig. Distribution of VAFs called by SNooPer (Model 1A), MuTect, JointSNVMix and Varscan2 on the test set. Selected somatic SNVs called by (**A**) SNooPer had to be flagged as "PASS" and to present a class probability >0.5. SNVs called by (**B**) MuTect had to be flagged as "KEEP" or present a somatic p-value <0.05 for (**C**) JointSNVMix and (**D**) Varscan2. In the histogram, bars represent probability densities (relative frequencies) of variants according to VAFs. Blue lines represent the kernel density estimates of true positive variants from the test set.

2.10. Tables

S1 Table. List of SNooPer's features and descriptions.



Group	Feature	Description	IG (model 1A)
Coverage and VAF	allelic_freq_highqual	High quality VAF (BQV > 20, MQV > 10)	0.04187
	highqualcoverage_vs_med	Number of high quality reads (MQV >10) supporting the alternative bases normalized by the median value	0.00940
Location along the read	LocMean_vs_med	Mean location of alternative bases on the reads normalized by the median value	0.00130
	LocMean_vs_ref	Ratio of the mean location of alternative bases on the reads over the mean location of reference bases on the reads	0.00000
Quality bias of alternative bases	GQBprob	Wilcoxon rank-sum test comparing BQVs of reference and alternative bases (p-value)	0.01423
	var_mean_mqv_quality	Mean MQV of alternative bases (Phred score)	0.01210
	var_mean_bqv_quality	Mean BQV of alternative bases (Phred score)	0.00883
	QBpval	Fisher's exact test comparing the number of reference and alternative bases matching the quality criteria (p-value)	0.00852
	RQBprob	Wilcoxon rank-sum test comparing MQVs of reference and alternative bases (p-value)	0.00641
	bqv_ratio_vs_med	Ratio of mean BQV of alternative bases over mean BQV of reference bases normalized by the median value	0.00559
	mqv_ratio_vs_med	Ratio of mean MQV of alternative bases over mean MQV of reference bases normalized by the median value	0.00156
Strand bias	SBpbinom	Probability of having n reads on the less represented strand for a coverage of N knowing that each reads has a probability of 0.5 to be aligned on each strand (cumulative probability from the Binomial distribution)	0.00224
	FRratio	Ratio of the number of reads supporting the alternative bases over the number of reads supporting the reference bases	0.00221
	SBpval	Fisher's exact test comparing the repartion of reference and alternative bases on positive and negative strands (p-value)	0.00103
Other	indelflag*	Flag indicating the presence of indels aligned at the considered position	0.00000

As an example, the information gain (IG) of each feature was measured with respect to the class (InfoGainAttributeEval method, Weka suite [35]) for Model 1A and is indicated in the last column of the table. Features presenting the suffix "vs_med" were normalized using a median value calculated from variants randomly extracted from mpileups files. Features presenting the suffix "vs_ref" were evaluated with respect to reference bases at the same position. (*) Available feature but not used to construct SNooPer's Model 1A.

S2 Table. Childhood ALL patient clinical information.

Patient	Gender	Age at Diagnosis (Months)	Immunophenotype	Karyotype	Interchromosomal recombination	Blast – BM (%)	WBC (.10^9/l)	Platelet (.10^9/l)	Clinical Risk	Dataset	Mean coverage on CDS (Dataset 1)
1	M	35	B	54,XY,+X,+4,+6,+10,+15,+17,+18,+21	-	99.5	29.91	24	Standard	1,2,3	33
2	M	29	B	54,XY,+4,+6,+8,+10,+14,+17,+18,+21	-	94.5	22.3	85	Standard	1,2	28
3	F	23	B	46,XX,inv(4)(p14q27),del(9p)	-	94	99.2	16	High	1,2	33
4	F	129	B	48,XX,+1,+6,-11,+21	-	97.5	61.1	94	High	1,2	27
5	M	75	B	46,XY	-	95.5	183.1	29	High	1,2	13
6	M	55	B	54,XY,+X,+4,+5,+6,+10,+14,+17,+21,del(9p)	-	96	59.2	21	High	1,2	19
7	M	58	B	47,XY,+21	-	NA	56.5	21	High	1	31
8	F	114	B	47,XX,+21	-	NA	32	65	High	1	37
9	M	112	B	NA	NA	NA	19.0	33.0	High	1	31
10	F	78	B	NA	NA	NA	14.7	12.0	Standard	1	28
11	F	117	B	47,XX,+21	-	95.0	72.5	21.0	High	1	32
12	M	133	B	46,XY,del(9p)	-	97.0	94.5	21.0	High	1	30
13	F	26	B	56,XX,+X,+3,+5,+6,+8,+10,+14,+15,+21,+22	-	100.0	47.9	82.0	Standard	1	29
14	F	164	B	46,XX	-	93.5	1.4	28.0	High	1	33
15	F	66	B	56,XX,+X,+4,+5,+6,+10,+15,+17,+18,+21,+22	-	97.0	5.4	38.0	Standard	1	30
16	M	28	B	NA	NA	95.0	6.3	103.0	Standard	1	31
17	F	33	B	NA	NA	97.0	132.0	16.0	High	1	32
18	F	29	B	46,XX	-	99.0	51.8	27.0	High	1	27
19	F	32	B	46,XX	t(12;21)	98.0	141.4	87.0	High	1	13
20	M	33	B	52,XY,+4,+6,+15,+17,+18,+21	-	75.5	4.0	315.0	Standard	1	15
21	M	33	B	47,XY,+21	-	99.0	32.5	16.0	Standard	1	28
22	M	102	B	46,XY	t(2;12)	99.0	14.0	8.0	Standard	1	33
23	M	182	B	57,XY,+Y,+3,+4,+8,+11,+14,+16,+19,+20,+21,+22	-	97.0	3.2	211.0	High	1	24
24	M	47	B	46,XY	t(9;20)	83.0	81.5	26.0	High	1	21
25	M	26	B	49,XY,+10,+18,+21	-	85.5	2.9	139.0	Standard	1	28
26	M	97	B	46,XY	t(12;21)	95.5	8.4	61.0	Standard	1	26
27	F	28	B	54,XX,+X,+6,+9,+10,+14,+17,+18,+21	-	86.0	25.1	85.0	Standard	1	24
28	F	88	B	48,XX,+4,+10	-	99.0	16.0	42.0	Standard	1	35
29	F	63	B	46,XX	t(12;21)	99.0	83.5	16.0	High	1	37
30	M	70	B	46,XY	t(12;21)	100.0	64.7	15.0	High	1	30
31	F	33	B	46,XX	t(12;21)	96.0	38.1	32.0	Standard	1	33
32	M	30	B	56,XY,+4,+5,+6,+8,+10,+14,+16,+17,+18,+21	-	87.7	3.2	17.0	Standard	1	17
33	M	168	T	NA	NA	NA	340.3	55.0	High	1	41
34	M	112	T	50,XY,+Y,+8,+10,+19	-	71.0	52.1	85.0	High	1	28
35	M	191	T	46,XY	t(11;17)	96.5	42.8	62.0	High	1	36
36	M	137	T	NA	NA	99.5	465.6	31.0	High	1	32
37	F	179	T	46,XX	t(10;14)	90.5	151.3	70.0	High	1	37
38	F	162	T	46,XX	-	90.5	253.1	24.0	High	1	40
39	M	152	T	NA	NA	91.5	793.5	18.0	High	1	48
40	M	178	T	46,XY,del(9p)	-	88.0	95.5	27.0	High	1	40

M: male; F: female; B: pre-B cell ALL; T: T-cell ALL; BM: bone marrow; WBC: white blood cell; NA: not applicable (missing data); (-): none.

2.11. Availability of data and materials

The source code is free and available at https://sourceforge.net/projects/snooper/. The program package is released under the GNU General Public License version 3.0 (GPLv3). Support is also provided at http://www.somaticsnooper.com/ (Project home page). The datasets supporting the conclusions of this article are available in the GEO public functional genomics data repository (SuperSeries GSE78786). SNooPer is written in the Perl and uses a RF classifier implemented in Weka suite (3.6.10 or greater) which requires the Java Runtime Environment (1.5 or greater). Additional and optional filters require a Bedtools intersect function. ROC and PR curves are drawn using the R package 'pracma' (Practical Numerical Math Functions).

Furthermore, trained models have been released to fulfill user needs. The community is also invited to develop and release their own models, which will allow opportunities to expand and update the tool and improve analysis performances.

2.12. Competing interests

The authors (JFS, PM, RV, VS, PC, CR, MO, JH and DS) declare no conflict of interest.

2.13. Acknowledgments

The authors are indebted to the patients and their parents for participating in this study. Patient tissue samples were provided by the CHU Sainte-Justine Pediatric Cancer Biobank. Sequencing was performed at the Integrated Clinical Genomics Centre In Pediatrics, CHU Sainte-Justine; Computations were made on the supercomputer Briarée from Université de Montréal, managed by Calcul Québec and Compute Canada.

2.14. References

1. Bonilla X, Parmentier L, King B, Bezrukov F, Kaya G, Zoete V, et al. Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. Nat Genet. 2016;48(4): 398-406.

2. Krauthammer M, Kong Y, Bacchiocchi A, Evans P, Pornputtapong N, Wu C, et al. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. Nat Genet. 2015;47(9): 996-1002.

3. Al-Ahmadie HA, Iyer G, Lee BH, Scott SN, Mehra R, Bagrodia A, et al. Frequent somatic CDH1 loss-of-function mutations in plasmacytoid variant bladder cancer. Nat Genet. 2016;48(4): 356-358.

4. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44(6): 685-689.

5. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012;487(7406): 239-243.

6. Giannakis M, Hodis E, Jasmine Mu X, Yamauchi M, Rosenbluh J, Cibulskis K, et al. RNF43 is frequently mutated in colorectal and endometrial cancers. Nat Genet. 2014;46(12): 1264-1266.

7. Tan J, Ong CK, Lim WK, Ng CC, Thike AA, Ng LM, et al. Genomic landscapes of breast fibroepithelial tumors. Nat Genet. 2015 Nov;47(11): 1341-1345.

8. Lim WK, Ong CK, Tan J, Thike AA, Ng CC, Rajasegaran V, et al. Exome sequencing identifies highly recurrent MED12 somatic mutations in breast fibroadenoma. Nat Genet. 2014;46(8): 877-880.

9. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403): 395-399.

10. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012;486(7403): 353-360. doi: 10.1038/nature11143.

11. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403): 400-404.

12. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486(7403): 405-409.

13. Rausch T, Jones DT, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012;148(1-2): 59-71.

14. Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga J, et al. Integrated molecular analysis of adult T cell leukemia/lymphoma. Nat Genet. 2015;47(11): 1304-1315.

15. Choi J, Goh G, Walradt T, Hong BS, Bunick CG, Chen K, et al. Genomic landscape of cutaneous T cell lymphoma. Nat Genet. 2015;47(9): 1011-1019.

16. De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet. 2013;45(2): 186-190.

17. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013 Mar;45(3): 242-252.
18. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2011;44(1): 47-52.
19. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013;501(7467): 338-345.
20. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244.
21. Ma X, Edmonson M, Yergeau D, Muzny DM, Hampton OA, Rusch M, et al. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat Commun. 2015;6:6604.
22. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013, 152: 714-726.
23. Green MR, Gentles AJ, Nair RV, Irish JM, Kihira S, Liu CL, et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. Blood. 2013;121: 1604-1611.
24. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012, 150: 264-278.
25. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. Science. 2008;322(5906): 1377-1380.
26. Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. Leukemia. 2014;28(1): 34-43.
27. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-576.
28. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2011;28: 311-317.
29. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012;28: 907-913.
30. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14): 1811-1817.
31. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31: 213-219.
32. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 2013;5(10):91.
33. Breiman L. Random Forests. Machine Learning. 2001;45: 5-32.
34. Kullback S. Information theory and statistics. John Wiley and Sons. New York; 1959.
35. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 2009;11(1).
36. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47: 11.12.1-11.12.34.
37. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA,

Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422): 56-65.
38. UCSC. UCSC Genome Informatics Group. 2016. [cited 17 July 2016]. Available: https://genome.ucsc.edu/
39. Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood. 2007;109(2): 683-692.
40. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. Leuk Res. 1997;21(9): 817-823.
41. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20: 37-46.
42. Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. San Francisco; 1993.
43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25: 2078-2079.
44. Ng PC, Henikoffa S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13): 3812-3814.
45. Spinella JF, Healy J, Saillour V, Richer C, Cassart P, Ouimet M, et al. Whole-exome sequencing of a rare case of familial childhood acute lymphoblastic leukemia reveals putative predisposing mutations in Fanconi anemia genes. BMC Cancer. 2015;15: 539.
46. Spinella JF, Cassart P, Garnier N, Rousseau P, Drullion C, Richer C, et al. A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells. BMC Cancer. 2015;15: 621.
47. Spinella JF, Cassart P, Richer C, Saillour V, Ouimet M, Langlois S, et al. Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations. Oncotarget. 2016. doi: 10.18632/oncotarget.11796.

2.15. Supplemental references

1. Picard. Broadinstitute. 2016. [cited 5 Avril 2016]. Available: http://broadinstitute.github.io/picard/

2. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9): 1297-1303.

3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25: 2078-2079.

4. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9: 357-359.

5. RefSeq: Reference Sequence Database. NCBI. 2016 [cited 5 Avril 2016]. Available: http://www.ncbi.nlm.nih.gov/refseq/

6. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422): 56-65.

7. UCSC. UCSC Genome Informatics Group. 2016. [cited 5 Avril 2016]. Available: https://genome.ucsc.edu/

8. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14): 1811-1817.

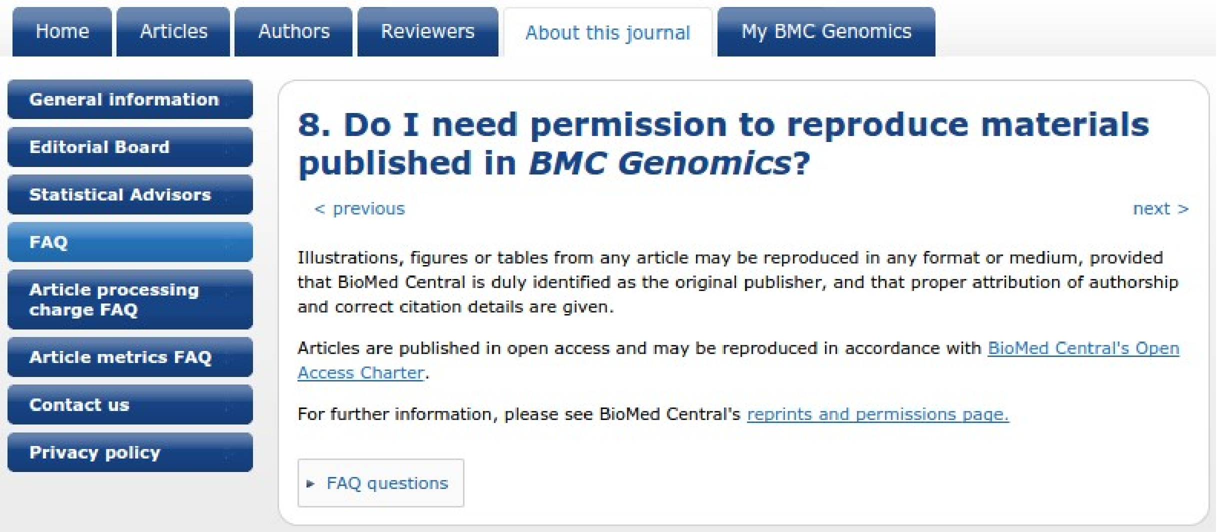
9. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-576.

10. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012;28: 907-913.

11. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10: Unit 10.11.

12. dbSNP Short Genetic Variations. NCBI. 2016 [cited 5 Avril 2016]. Available: http://www.ncbi.nlm.nih.gov/SNP/

2.16. Reprint permission



8. Do I need permission to reproduce materials published in *BMC Genomics*?

< previous

next >

Illustrations, figures or tables from any article may be reproduced in any format or medium, provided that BioMed Central is duly identified as the original publisher, and that proper attribution of authorship and correct citation details are given.

Articles are published in open access and may be reproduced in accordance with BioMed Central's Open Access Charter.

For further information, please see BioMed Central's reprints and permissions page.

▸ FAQ questions

Obtenu de http://old.biomedcentral.com/bmcgenomics/about/faq/reproduce.

CHAPITRE 3

Genomic characterization of pediatric B-cell ALL: importance of rare germline and somatic events

3.1. ARTICLE II - Whole-exome sequencing of a rare case of familial childhood acute lymphoblastic leukemia reveals putative predisposing mutations in Fanconi anemia genes

3.1.1. Avant-propos

Comme expliqué dans l'introduction de cette thèse (voir section 1.4.2.3: Facteurs de susceptibilité génétique), au cours des dernières années plusieurs projets de recherche ont été consacrés à l'étude de l'impact de l'héritage de variants rares ou communs sur la prédisposition à la LAL pédiatrique. Les associations identifiées suivaient généralement le modèle à 2 dimensions “fréquence allélique” vs. “importance de l'effet” (Figure XV): Les allèles à pénétrance fortes sont extrêmement rares mais entraînent un effet important tandis que les variants communs dans la population ne contribuent que par le biais d'effets réduits [142]. Si l'étude de larges populations par GWAS a permis de mettre en évidence plusieurs allèles communs influençant le risque de développement de la maladie, seuls les rares cas de leucémies familiales permettent un *design* expérimental adapté à la découverte de variants rares de pénétrance plus élevée. Nous présentons ici l'étude d'une de ces familles présentant deux frères non jumeaux ayant développé une LAL-B avec concordance du phénotype à quelques années d'intervalle.

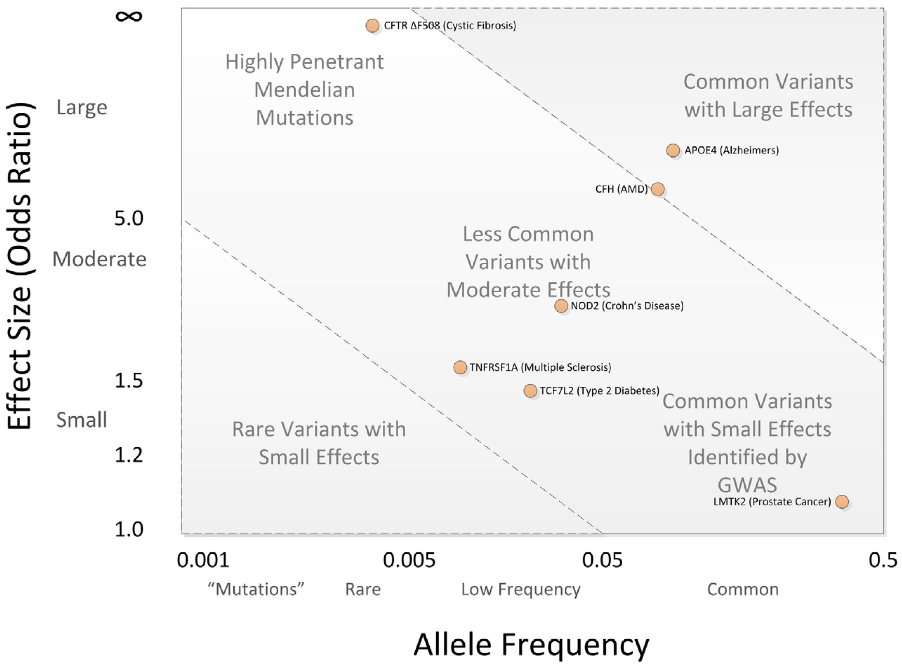


Figure XV. Spectre allélique: fréquence vs. effet. La majeure partie des associations génétiques découvertes se trouvent entre les diagonales pointillées. Figure reproduite de Bush and Moore, 2012 [142].

Whole-exome sequencing of a rare case of familial childhood acute lymphoblastic leukemia reveals putative predisposing mutations in Fanconi anemia genes.

Jean-François Spinella[1], Jasmine Healy[1], Virginie Saillour[1], Chantal Richer[1], Pauline Cassart[1], Manon Ouimet[1], Daniel Sinnett[1,2]

1) Sainte-Justine UHC Research Center, University of Montreal, Montreal, Qc, Canada; 2) Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, Qc, Canada.

This research was originally published in BMC Cancer. 2015;15:539. doi: 10.1186/s12885-015-1549-6.

3.1.2. Authors' contributions

DS is the principle investigator and takes primary responsibility for the paper. JFS, JH, and DS contributed to the conception and design of the study. JFS, PC, MO and CR were involved in sample and library preparation. JFS performed whole-exome and statistical analyses. VS provided bioinformatics support. JFS and JH wrote the paper and DS contributed to the interpretation of the data and was involved in critical manuscript revision. All authors approved the final version.

3.1.3. Abstract

Background

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer. While the multi-step model of pediatric leukemogenesis suggests interplay between constitutional and somatic genomes, the role of inherited genetic variability remains largely undescribed. Nonsyndromic familial ALL, although extremely rare, provides the ideal setting to study inherited contributions to ALL. Toward this goal, we sequenced the exomes of a childhood ALL family consisting of mother, father and two non-twinned siblings diagnosed with concordant pre-B hyperdiploid ALL and previously shown to have inherited a rare form of *PRDM9*, a histone H3 methyltransferase involved in crossing-over at recombination hotspotsand Holliday junctions. We postulated that inheritance of additional rare disadvantaging variants in predisposing cancer genes could affect genomic stability and lead to increased risk of hyperdiploid ALL within this family.

Methods

Whole exomes were captured using Agilent's SureSelect kit and sequenced on the Life Technologies SOLiD System. We applied a data reduction strategy to identify candidate variants shared by both affected siblings. Under a recessive disease model, we focused on rare non-synonymous or frame-shift variants in leukemia predisposing pathways.

Results

Though the family was nonsyndromic, we identified a combination of rare variants in Fanconi anemia (FA) genes *FANCP/SLX4* (compound heterozygote - rs137976282/rs79842542) and *FANCA* (rs61753269) and a rare homozygous variant in the Holliday junction resolvase *GEN1* (rs16981869). These variants, predicted to affect protein function, were previously identified in familial breast cancer cases. Based on our in-house database of 369 childhood ALL exomes, the sibs were the only patients to carry this particularly rare combination and only a single hyperdiploid patient was heterozygote at both *FANCP/SLX4* positions, while no *FANCA* variant allele carriers were identified. *FANCA* is the most commonly mutated gene in FA and is essential for resolving DNA interstrand cross-links during replication. *FANCP/SLX4* and *GEN1* are involved in the cleavage of Holliday junctions and their mutated forms, in combination with the rare allele of *PRDM9*, could alter Holliday junction resolution leading to nondisjunction of chromosomes and segregation defects.

Conclusion

Taken together, these results suggest that concomitant inheritance of rare variants in *FANCA, FANCP/SLX4* and *GEN1* on the specific genetic background of this familial case, could lead to increased genomic instability, hematopoietic dysfunction, and higher risk of childhood leukemia.

Keywords

Familial acute lymphoblastic leukemia, childhood leukemia predisposition, Fanconi anemia genes.

3.1.4. Background

ALL accounts for approximately 25% of all pediatric cancer cases, however its etiology remains elusive [1]. Direct evidence that childhood ALL has a genetic component is provided by the high risk of developing the disease associated with certain inherited cancer-predisposing syndromes such as Bloom's syndrome, Down syndrome, Fanconi anemia, neurofibromatosis and ataxia telangiectasia, however they account for a trivial proportion of cases (collectively <5%) [2]. A heritable basis for ALL outside these syndromes is largely undefined.Genome-wide association studies provided the first unambiguous evidence that common inherited genetic variation increases the risk of developing childhood ALL[3-6]. The identification of low-penetrance susceptibility alleles at 7p12.2 (*IKZF1*), 9p12 (*CDKN2A*/*CDKN2B*), 10q21.2 (*ARID5B*) and 14q11.2 (*CEBPE*) in genes involved in transcriptional regulation and differentiation of B-lymphocyte progenitors, highlights the role of constitutional genetic predisposition in childhood ALL onset. Yet these loci only explain a small proportion of the familial risk associated with childhood ALL[7] suggesting that the underlying genetic architecture likely involves co-inheritance of multiple variants on a wide allelic spectrum with varying penetrance. While large population-based cohorts will be required to identify additional common ALL-predisposing variants, families with multiple non-twinned ALL sibships, though extremely rare[8,9], represent ideal models to investigate the role of rare/private inherited genetic variation in disease etiology.

Through a recent international collaborative effort to identify childhood ALL families, it was reported that ALL sibs exhibit high subtype concordance, likely explained by shared underlying genetic risk[8]. Here we report the case of a nonsyndromic pre-B childhood ALL family with two male non-twinned siblings diagnosed with hyperdiploid pre-B ALL. The prenatal origins of hyperdiploid childhood ALL and the need for additional postnatal mutations to drive overt leukemogenesisare well established [10]. The extent to which inherited genetic variation contributes to the onset of hyperdiploid childhood ALL however is less clear. The sibs were previously shown to have maternally inherited a rare allelic form of *PRDM9*, a meiosis-specific histone H3 methyltransferase that was suggested to influence genomic instability in ALL by potentially controlling the location of genetic crossing-over at recombination hotspots[11] and at Holliday junctions[12]. Based on these data, we postulated that co-inheritance of additional rare disadvantaging DNA variants is likely required to explain this familial case of ALL, the identification of which could allow for better understanding of leukemogenesis and benefit a much broader childhood ALL population. Even though the family was otherwise asymptomatic, because the Fanconi anemia (FA) pathway is a well-known leukemia predisposing disorder and FA-associated gene dysfunction has been linked to genomic instabilities, defects in Holliday junction resolution[13] and aneuploidy[14], we postulated that inherited rare disadvantaging DNA variants in FA cancer predisposing genes/pathway, in combination with *PRDM9*, could contribute to the chromosome instabilities underlying this case of familial hyperdiploid childhood ALL.

3.1.5. Methods

3.1.5.1. Patients

This nonsyndromic pre-B childhood ALL family is of self-reported Moroccan origin (Figure 1); three unaffected sibs (2 females and 1 male) could not be ascertained. Family history includes death due to cancer of both maternal and paternal grandfathers, colon cancer at age 69 and prostate cancer at age 65, respectively. A consanguineous marriage (first cousins) on the paternal side lead to multiple miscarriages and children with polymalformation syndrome, one of which died at 1 week. The probands were diagnosed with childhood ALL and were treated at the Sainte-Justine UHC (SJUHC) in Montreal, Quebec, but were otherwise healthy.

Sibling A, a 2 year old male, had a white blood cell count (WBC) of 4.4 x 10^9/L, 14% and 75.5% lymphoblast cells in the blood and bone marrow respectively, and a platelet count of 315.0 x 10^9/L. Cytogenetic analysis revealed hyperdiploidy with the following karyotype: 53,XY,+4,+6,+12,+15,+17,+18,+21, and fluorescent in situ hybridization (FISH) identified a germline inversion inv(2)(p11.2q13) that was also carried by the mother. This recurrent pericentric inversion is stably inherited without phenotypic or developmental consequences and likely has no clinical relevance[15]. Sib A was classified as standard risk and was enrolled on Dana Farber Cancer Institute (DFCI) ALL Consortium Protocol 95-01. He has been out of treatment for over 60 months with leukemia free-survival (LFS).

Sibling B, a 14 year old male, was diagnosed three years later and was classified as high risk based on his age. He had a WBC of 6.2 x 10^9/L, 18% and 93% lymphoblast cells in the blood and bone marrow respectively, and a platelet count of 57.0 x 10^9/L. Cytogenetic analysis also revealed hyperdiploidy: 54,XY,+X,+5,+8,+10,+14,+17,+18,+21, yet Sib B did not carry his mother's inv(2)(p11.2q13) inversion. Sib B was enrolled on DFCI-ALL protocol 2000-01 for high-risk patients; he has responded well to treatment and is also over 60 months with LFS.

3.1.5.2. Whole exome sequence capture and sequencing

DNA was extracted from peripheral blood samples (obtained after remission) from the sibship, and from both parents using standard protocols as described previously[16]. Whole exomes were captured in solution with Agilent's SureSelect Human All Exon 50Mb kits,and sequenced on the Life Technologies SOLiD System (sibship mean coverage =28.1X, parents mean coverage =19.4X). Reads were aligned to the hg19 reference genome using SOLiD LifeScope software (see Figure 2 for complete sequencing analysis workflow). PCR duplicates were removed using Picard[17]. Base quality score recalibration was performed using the Genome Analysis ToolKit (GATK)[18] and QC Failure reads were removed. Cleaned BAM files were used to create pileup files using SAMtools [19].

3.1.5.3. Variant calling and annotation

Single nucleotide variations (SNVs) and insertion and deletion (indels) were called from pileup files using SNooPer, an in-house variant caller that is based on a machine learning approach and developed to minimize false positive variant calling in low-depth sequencing data (manuscript submitted and software available upon request). Using this familial design, we were able to effectively incorporate parental sequence information to remove Mendelian inconsistencies, reduce false-positive sequencing and alignment errors, and facilitate the identification of candidate disease-predisposing variants shared by both affected siblings. Variant frequencies were assessed using 1000 Genomes [20] and NHLBI GO Exome Sequencing Project (ESP) [21] databases. ANNOVAR [22] was used for non-synonymous SNV annotation. The effect of non-synonymous variants on protein conformation and function was assessed using Sift [23], Polyphen2 [24] and functional analysis through hidden markov models (Fathmm, version 2.3) [25]. Sift, Polyphen2 and Fathmm consider a variant as putatively damaging when it presents a score ≤0.05, ≥0.957 and <-1.5, respectively. SiPhy [26] was used to detect bases under selection using multiple alignment data from 29 mammal genomes; larger is the score, more conserved is the site.

3.1.6. Results and Discussion

The sibs were diagnosed with nonsyndromic childhood ALL three years apart. We previously identified a rare *PRDM9* allele segregating within the family [11]. PRDM9 is a histone H3 methyltransferase involved in crossing-over at recombination hotspots and Holliday junctions. To further characterize the underlying inherited genetic contribution to this childhood ALL family in an unbiased manner, we performed whole exome sequencing of the siblings and both parents. Though the family was nonsyndromic and asymptomatic for FA, this recessive disorder is linked to hematopoietic dysfunction, chromosomal instability and increased susceptibility to childhood ALL. Based on the observed concordant hyperdiploid phenotype of both siblings, we postulated that inherited rare disadvantaging DNA variants in leukemia predisposing pathways like the FA pathway could affect overall genomic instability and, in combination with the rare allelic form of *PRDM9*, favour nondisjunction of chromosomes leading to increased risk of hyperdiploid pre-B ALL within this family. Under a recessive disease model, we interrogated our exome data and identified shared non-synonymous mutations that were either compound heterozygous or homozygous variant (Table 1) and specifically screened genes associated with the leukemia predisposing syndrome FA (F*ANCA, FANCB, FANCC, FANCD1/BRCA2, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCJ, FANCL, FANCM, FANCN/PALB2, FANCO/RAD51C, FANCP/SLX4, FANCQ/XPF* and *FANCS/BRCA1*). Among the identified variants, we identified a combination of missense variants in the FA gene *FANCP/SLX4* (compound heterozygous at rs137976282 and rs79842542), corroborating the assumption of FA pathway destabilization (Figure 2). A more thorough investigation of the other FA pathway genes led then to the identification of a rare heterozygous variant in *FANCA* (rs61753269) that was also shared by the sibs. Although this variant was heterozygous, restricting the analysis to extremely rare variants allowed us to identify potentially deleterious non-synonymous variations in FA genes that could be contributing to inherited susceptibility to ALL in the sibs. For *FANCP/SLX4*, both parents transmitted a putatively damaging allele to their affected offspring who were therefore compound heterozygous at rs137976282 (ESP and 1000 Genomes general population MAF <0.001) and rs79842542 (MAF =0.059 and 0.071 in 1000 Genomes and ESP general populations respectively). While two of the three in silico algorithms predicted that the compound heterozygous variants in *FANCP/SLX4* were likely deleterious (Sift score =0 for both alleles and Polyphen2 score =1 and 0.964 for rs79842542 and rs137976282 respectively), only Fathmm predicted rs61753269 in *FANCA* to be damaging (Fathmm score =-1.78) (Table 1). Nevertheless, the high conservation score at *FANCA* rs61753269 (SiPhy =12.742), combined with its extreme rarity in the population (MAF <0.001 in 1000 Genomes and ESP), suggest that this variant is under strong functional constraint and therefore could have a specific role on protein conformation. Although not a Fanconi anemia gene per se, our exome data also revealed a rare non-synonymous homozygous variant in *GEN1* (rs16981869, MAF =0.145394, ESP general population homozygous frequency q^2 =0.025), that was predicted to be deleterious by all three algorithms. *GEN1* is a member of the *FANCP/SLX4* complex involved in Holliday junction resolution [27], and in conjunction with *PRDM9* and the FA genes identified here, could be contributing to genomic instability in the sibs. Our in-house exome database of 369 individuals from our childhood ALL cohort (103 patient-mother-father trios and 60 patients) from the QcALL cohort [28] (whole exome sequencing performed on Life Technologies SOLiD System or Illumina HiSeq 2500; data available upon request), revealed a single heterozygote patient at both *FANCP/SLX4* positions, 0/369 variant allele carriers at *FANCA* rs61753269 and 3/369 carriers of the homozygous allele at *GEN1* rs16981869 (1 patient and 2 parents). Interestingly, the only 2 other cases harbouring either both variants in *FANCP/SLX4* or the homozygous variant in *GEN1* were also diagnosed with hyperdiploid pre-B ALL, concordant with the sibship. Overall, the sibs were the only two individuals who carried this particularly rare combination of damaging alleles at *FANCA* rs61753269, *FANCP/SLX4* rs137976282, rs79842542 and *GEN1* rs16981869.

Fanconi anemia is a recessive genetic disorder and most frequent cause of inherited bone marrow failure. To date, 17 FA genes have been identified and mutations within these genes have been shown to cause DNA repair defects leading to genomic instability and aneuploidy, characteristic of FA [29]. Given cumulative hematopoietic dysfunction and excess chromosomal instability, FA patients are at higher risk of developing hematopoietic malignancies including leukemia[30]. Interestingly, the rare variants *FANCP/SLX4* rs137976282 and *FANCA* rs61753269 have previously been identified in familial breast cancer cases [31-34], however their pathological effects in cancer predisposition remain unknown. *FANCA*, mutated in over 60% of FA cases, is an essential member of the FA core complex involved in monoubiquitination of the FANCI/D2 complex which in turn guides downstream activation of the DNA repair processes for resolving DNA interstrand cross-links during replication[35]. Mono-allelic deletion of *FANCA* has been suggested to promote genetic instabilities associated with acute myeloid leukemia[36]. *FANCP/SLX4* on the other hand, is a downstream component of the FA pathway that codes for a Holliday junction resolvase. It acts as a docking platform for three structure-specific endonucleases XPF–ERCC1, MUS81–EME1 and SLX1[37]. Recently identified as a FA gene, *FANCP/SLX4* modulates DNA repair and cellular responses to replication fork failure[38]. *GEN*1 codes for an endonuclease, and is a member of the *FANCP/SLX4* complex [27] shown to play a role in the maintenance of centrosome integrity [39]. Along with *PRDM9*, *GEN1* and the *FANCP/SLX4* complex are involved in the definition of Holliday junction branch migration boundaries and the cleavage of static and migrating Holliday junctions [12,27,37]. Efficient

DNA damage repair and simultaneous regulation of cell cycle progression is critical for genomic stability. Interestingly, a rare recessive homozygous variant in *GEN1* has been associated with bilateral breast cancer [40] and the depletion of *GEN1* or *FANCP/SLX4* in Bloom's syndrome cells results in defects in chromosome condensation and severe chromosome abnormalities, such as nondisjunction of sister chromatids and abnormal mitosis leading to aneuploidy [41,42], highlighting their important role in maintaining genome stability. Thus, mutated *FANCP/SLX4* and *GEN1*, in combination with the rare allele of *PRDM9* also segregating within this family, could alter Holliday junction resolution leading to nondisjunction of chromosomes and segregation defects.

While autosomal recessive FA patients are known to present with malformations [43], it has been reported that heterozygous carriers of a FA gene may be predisposed to some of the same congenital malformations or developmental abnormalities that are common among homozygotes [44]. Although the sibs had no apparent physical abnormalities, family history revealed a consanguineous marriage on the paternal side (Figure 1) resulting in multiple miscarriages and polymalformation syndrome in surviving offspring. Given that both rare *FANCP/SLX4* rs137976282 and *FANCA* rs61753269 variants were paternally inherited we could hypothesize an underlying recessive disorder affecting the FA pathway; however this remains highly speculative without further genotype information on the extended family. Overall, these data support a functional role for the rare variants identified in *FANCA, FANCP/SLX4* and *GEN1* in disrupting the FA pathway and Holliday junction resolution, and as a result, they could lead to genomic instability and hematopoietic dysfunction, and increased risk of ALL within this family. However functional assays are required to confirm these observations.

Despite the fact that both siblings were asymptomatic and were not diagnosed with an ALL-linked genetic disorder, the possibility of an underlying FA condition exists and an undiagnosed disorder, although rare, cannot be excluded. One may argue that pure, nonsyndromic ALL families are unlikely and that genetic interrogation of such families will ultimately reveal underlying inherited disorders associated with increased risk of ALL. Indeed, our results show that the study of familial or inherited forms of ALL can further our understanding of the genetic causes underlying more common, sporadic forms and shed light on otherwise asymptomatic genetic syndromes.

Finally, though our rare variant analysis strongly suggests *FANCP/SLX4* and *FANCA* as the most likely candidates, we cannot exclude the possibility that additional inherited genetic variants, rare or common, outside of the FA pathway could contribute to ALL onset within the family. For example, we identified common non-synonymous variants in *PDE4DIP* and *CEP55* (Table 1). Though these centrosomal proteins have been involved in myeloproliferative disorder [45] and carcinogenesis [46] and could promote abnormal cell division and hyperdiploidy, as evidenced recently by Paulsson *et al.* [47], the identified variants had high MAFs and were predicted to have benign effects on protein function, making them unlikely candidates here. Furthermore, the sibs carry common ALL susceptibility alleles at known GWAS loci [3-6,28] (Table 2), that under an additive effects model could lead up to a 2- to 10-fold increase in risk [9]. Given the male-specific inheritance, we also looked for shared deleterious variants on the X chromosome but found no evidence of X-linked genes contributing to ALL in this family. The exomes of the siblings were also screened for shared de novo mutations that could result from gonadal mosaicism. Putative de novo events were defined as private mutations shared by both siblings, and therefore unknown in public databases, and showing no evidence of heritability from either parent, i.e. no reads supporting the variation in the parental exomes considering a minimum coverage of 8X at the given position in the exome sequencing data. Although no candidate de novo mutation fitting our criteria was identified, the limited coverage of parental exomes may have hindered this analysis. The investigation of more complex genetic models including gene-gene and eventually gene-environment interactions could also reveal additional ALL risk factors.

3.1.7. Conclusions

Nonsyndromic families with multiple non-twinned siblings diagnosed with childhood ALL are extremely rare but represent an interesting model to characterize the influence of inherited genetic burden on disease onset. This unique setting can also facilitate the identification of novel genes/pathways involved in driving the leukemic process and further our understanding of the mechanisms involved in childhood pre-B ALL and its subtypes. Here, we used next-generation sequencing technologies to sequence the whole-exomes of a childhood ALL family consisting of mother, father and two male affected sibs. Both brothers were diagnosed with pre-B hyperdiploid childhood ALL and their similar clinical and molecular characteristics suggested shared etiologic factors. Though functional validation studies are required to substantiate the role of these variants in hyperdiploid pre-B childhood ALL, our data suggest that concomitant inheritance of rare variants in FA genes *FANCA*, *FANCP/SLX*4, in combination with rare mutations in the endonuclease *GEN1* and the meiotic recombination gene *PRDM9*, could lead to increased DNA damage and genomic instability, and thus contribute to hyperdiploid leukemia predisposition.

3.1.8. Competing interests

The authors report no competing financial interests.

3.1.9. Acknowledgements

The authors are indebted to the patients and their parents for participating in this study. This study was supported by research funds provided by the Terry Fox Research Institute and the Canadian Institutes for Health Research. JFS is the recipient of a Réseau de médecine génétique appliquée (RMGA) Fellowship. DS holds the François-Karl-Viau Research Chair in Pediatric Oncogenomics. Whole exome sequencing was performed at the Child Health Genomics Platform of the Sainte-Justine UHC Research Center; Computations were made on the supercomputer Briarée from Université de Montréal, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

3.1.10. Figures

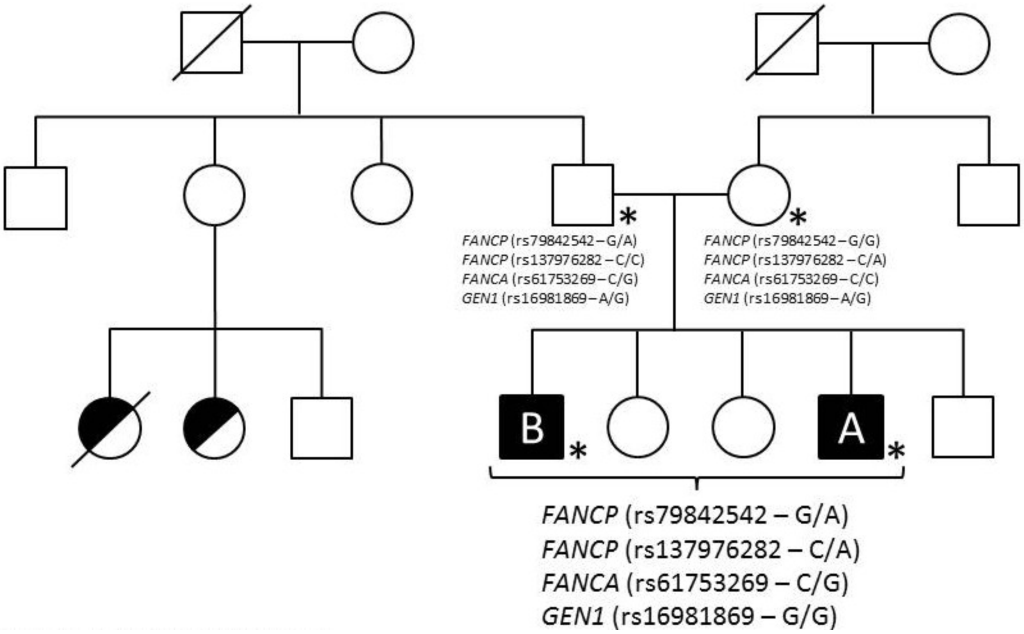


Figure 1. Family pedigree. The family is of self-reported Moroccan origin and consists of five siblings, including two non-twinned brothers diagnosed with pre-B acute lymphoblastic leukemia (A and B) as well as two healthy females and one healthy male. Affected probands are represented by the shaded squares; cousins with poly-malformation syndrome are represented by half-shaded circles. Sequenced individuals are identified by an asterisk.

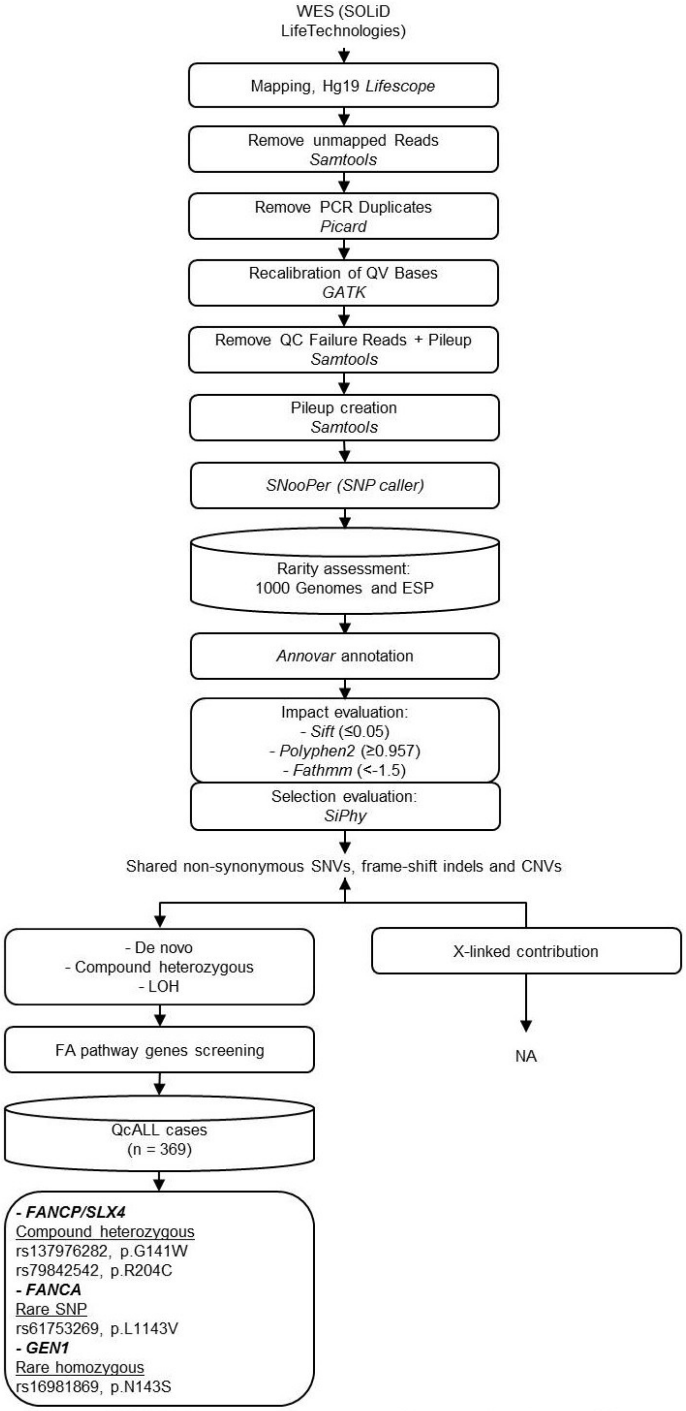


Figure 2. Whole-exome sequencing analysis workflow. Boxes represent the analysis/cleaning steps. Cylinders represent the variant filtering steps used in the data reduction strategy to identify inherited rare mutations shared by both sibs.

3.1.11. Tables

Table 1. Non-synonymous homozygous variants and compound heterozygous shared by both childhood pre-B ALL siblings.

	Gene	SNP ID	Chr	Position	Ref	Sibs	Father	Mother
Compound heterozygous	FANCP/SLX4	rs79842542	16	3656625	GG	AG	AG	GG
		rs137976282	16	3658545	CC	AC	CC	AC
	CEP55	rs75139274	10	95278683	GG	AG	AG	GG
		rs2293277	10	95279506	AA	TA	AA	TA
	DNAH2	rs140035206	17	7673930	AA	GA	GA	AA
		rs79350244	17	7734114	AA	CA	AA	CA
		rs117465420	17	7734476	AA	TA	AA	TA
		rs78354379	17	7736480	TT	AT	AT	TT
	PDE4DIP	rs1778120	1	144879090	CC	CT	CT	TT
		rs1698683	1	144916676	CC	TC	CC	TC
Homozygous	GEN1	rs16981869	2	17946243	AA	GG	GA	GA
	B3GALTL	rs1041073	13	31891746	GG	AA	AG	AG
	CA9	rs2071676	9	35674053	AA	AA	AG	AG
	CHIT1	rs2297950	1	203194186	CC	TT	TC	TC
	CHRNB1	rs17856697	17	7348625	AA	GG	GA	GA
	ERBB2	rs1058808	17	37884037	CC	GG	GC	GC
	ZNF207	rs3795244	17	30692396	GG	TT	TG	TG

	Gene	AA change	1000g MAF	ESP MAF / q^2	Sift	Polyphen2	Fathmm	SiPhy
Compound heterozygous	FANCP/SLX4	R204C	0.0597045	0.071264 / -	0	1	3.49	12.895
		G141W	0.000399361	0.00077 / -	0	0.964	5.2	7.273
	CEP55	R348K	0.0339457	0.074581 / -	0.19	0.214	2.05	11.439
		H378L	0.559704	0.610257 / -	0.13	0.483	2.21	14.69
	DNAH2	Y1385C	0.00219649	0.004075 / -	0	0.999	-0.15	15.101
		I4023L	0.0127796	0.021913 / -	1	0.516	3.81	15.198
		L4062F	0.0127796	0.021759 / -	0.02	0.411	3.06	8.222
		V4357D	0.0467252	0.008073 / -	0.03	0.986	2.95	12.116
	PDE4DIP	K1410E	-	0.124712 / -	0.11	0.996	4.64	11.54
		W626*	-	0.321203 / -	0.16	NA	3.81	18.033
Homozygous	GEN1	N143S	0.127995	0.145394 / 0.025	0.03	0.812	-0.45	8.027
	B3GALTL	E370K	0.666733	0.65539 / 0.442	0.28	0.964	-1.92	7.087
	CA9	V33L	0.323283	0.269107 / 0.560	0	0.815	-0.66	8.009
	CHIT1	G102S	0.290935	0.285253 / 0.065	0	1	3.81	7.755
	CHRNB1	E32G	0.120607	0.25585 / 0.052	0.08	0.772	-1.16	8.739
	ERBB2	P1170A	0.452077	0.513532 / 0.278	0.03	0.953	-0.81	18.007
	ZNF207	A240S	0.0467252	0.045748 / 0.001	0.41	0.748	0.85	20.212

(-) represents missing or not relevant information. For these genes, either or both parents transmitted a putatively damaging allele to their affected offspring, who were therefore compound heterozygous or homozygous, respectively. Genotype calls are provided for each sample (Sibs, Father and Mother) along with corresponding amino acid (AA) changes. Minor allele frequencies (MAF) were derived from the 1000 Genomes (general population, updated in October 2014) and the NHLBI GO Exome Sequencing Project (general population, ESP6500). The frequencies of homozygous variants (q^2) were obtained from ESP6500 and were presented when relevant. The putative effect of these substitutions on the protein function was assessed in silico using Sift (≤0.05) [23], Polyphen2 (≥0.957) [24] and Fathmm (<-1.5) [25]. SiPhy was used to identify bases under selection (larger is the score, more conserved is the site) [26].

Table 2. Childhood ALL susceptibility loci genotyped in siblings A and B.

Gene	SNP ID	Ref	A	B
ARID5B	rs7073837	CC	-	AA
	rs10994982	GG	GA	AA
	rs10740055	AA	-	CC
	rs10821936	TT	-	CC
	rs7089424	TT	-	GG
CEBPE	rs2239633	CC	CT	TT
DDC	rs7809758	AA	AG	AG
	rs880028	TT	TC	TC
	rs3779084	TT	TC	TC
	rs2242041	CC	GG	CG
IKZF1	rs6964823	GG	GA	GA
	rs11978267	AA	-	AG
	rs4132601	TT	-	TG
	rs6944602	GG	GG	GG
OR2C3	rs1881797	TT	TT	-
CDKN2A	rs36228834	TT	TT	TT

(-) represents missing information

3.1.12. References

1. Pui CH, Mullighan CG, Evans WE, Relling MV. Pediatric acute lymphoblastic leukemia: where are we going and how do we get there? Blood. 2012;120(6):1165-1174. doi: 10.1182/blood-2012-05-378943.

2. Horwitz M. The genetics of familial leukemia. Leukemia. 1997;11(8):1347-1359.

3. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JA, Allan JM, Tomlinson IP, Taylor M, Greaves M, Houlston RS. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet. 2009;41(9):1006-1010. doi: 10.1038/ng.430.

4. Treviño LR, Shimasaki N, Yang W, Panetta JC, Cheng C, Pei D, Chan D, Sparreboom A, Giacomini KM, Pui CH, Evans WE, Relling MV. Germline genetic variation in an organic anion transporter polypeptide associated with methotrexate pharmacokinetics and clinical effects. J Clin Oncol. 2009;27(35):5972-5978. doi: 10.1200/JCO.2008.20.4156.

5. Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Koehler R, Greaves M, Sheridan E, Gast A, Kinsey SE, Lightfoot T, Roman E, Taylor M, Pritchard-Jones K, Stanulla M, Schrappe M, Bartram CR, Houlston RS, Kumar R, Hemminki K. Verification of the susceptibility loci on 7p12.2, 10q21.2, and 14q11.2 in precursor B-cell acute lymphoblastic leukemia of childhood. Blood. 2010;115(9):1765-1767. doi: 10.1182/blood-2009-09-241513.

6. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, Papaemmanuil E, Bartram CR, Stanulla M, Schrappe M, Gast A, Dobbins SE, Ma Y, Sheridan E, Taylor M, Kinsey SE, Lightfoot T, Roman E, Irving JA, Allan JM, Moorman AV, Harrison CJ, Tomlinson IP, Richards S, Zimmermann M, Szalai C, Semsei AF, Erdelyi DJ, Krajinovic M, Sinnett D, Healy J, Gonzalez Neira A, Kawamata N, Ogawa S, Koeffler HP, Hemminki K, Greaves M, Houlston RS. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. Nat Genet. 2010;42(6):492-494. doi: 10.1038/ng.585.

7. Kharazmi E, da Silva Filho MI, Pukkala E, Sundquist K, Thomsen H, Hemminki K. Familial risks for childhood acute lymphocytic leukaemia in Sweden and Finland: Far exceeding the effects of known germline variants. Br J Haematol. 2012;159(5):585-588. doi: 10.1111/bjh.12069.

8. Schmiegelow K, Lausten Thomsen U, Baruchel A, Pacheco CE, Pieters R, Pombo-de-Oliveira MS, Andersen EW, Rostgaard K, Hjalgrim H, Pui CH. High concordance of subtypes of childhood acute lymphoblastic leukemia within families: lessons from sibships with multiple cases of leukemia. Leukemia. 2012;26(4):675-681. doi: 10.1038/leu.2011.274.

9. Pombo-de-Oliveira MS, Emerenciano M, Winn AP, Costa I, Mansur MB, Ford AM. Concordant B-cell precursor acute lymphoblastic leukemia in non-twinned siblings. Blood Cells Mol Dis. 2014. doi: 10.1016/j.bcmd.2014.07.011.

10. Bateman CM, Alpar D, Ford AM, Colman SM, Wren D, Morgan M, Kearney L, Greaves M. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. Leukemia. 2014. doi: 10.1038/leu.2014.177.

11. Hussin J, Sinnett D, Casals F, Idaghdour Y, Bruat V, Saillour V, Healy J, Grenier JC, de Malliard T, Busche S, Spinella JF, Larivière M, Gibson G, Andersson A, Holmfeldt L, Ma J,

Wei L, Zhang J, Andelfinger G, Downing JR, Mullighan CG, Awadalla P. Rare allelic forms of PRDM9 associated with childhood leukemogenesis. Genome Res. 2013;23(3):419-430. doi: 10.1101/gr.144188.112.

12. Baker CL, Walker M, Kajita S, Petkov PM, Paigen K. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Res. 2014;24(5):724-732. doi: 10.1101/gr.170167.113.

13. Fekairi S, Scaglione S, Chahwan C, Taylor ER, Tissier A, Coulon S, Dong MQ, Ruse C, Yates JR 3rd, Russell P, Fuchs RP, McGowan CH, Gaillard PH. Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases. Cell. 2009;138(1):78-89. doi: 10.1016/j.cell.2009.06.029.

14. Kim H, Andrea ADD. Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway. Genes Dev. 2012;26(13):1393-1408. doi: 10.1101/gad.195248.112.

15. Hysert M, Bruyère H, Côté GB, Dawson AJ, Dolling JA, Fetni R, Hrynchak M, Lavoie J, McGowan-Jordan J, Tihy F, Duncan AM. Prenatal cytogenetic assessment and inv(2)(p11.2q13). Prenat Diagn. 2006;26(9):810-813.

16. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. Leuk Res. 1997;21(9):817-823.

17. http://picard.sourceforge.net

18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303. doi: 10.1101/gr.107524.110.

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25:2078-2079. doi: 10.1093/bioinformatics/btp352.

20. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-1073. doi: 10.1038/nature09534.

21. http://evs.gs.washington.edu/EVS/

22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. doi: 10.1093/nar/gkq603.

23. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-1081. doi: 10.1038/nprot.2009.86.

24. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20. doi: 10.1002/0471142905.hg0720s76.

25. Shihab HA, Gough J, Cooper DN, Day IN, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29(12):1504-1510. doi: 10.1093/bioinformatics/btt182.

26. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54-62. doi: 10.1093/bioinformatics/btp190.

27. Ip SC, Rass U, Blanco MG, Flynn HR, Skehel JM, West SC. Identification of Holliday junction resolvases from humans and yeast. Nature. 2008;456(7220):357-361. doi: 10.1038/nature07470.

28. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. Haematologica. 2010;95(9):1608-1611. doi: 10.3324/haematol.2010.022459.

29. Wang AT, Smogorzewska A. SnapShot: Fanconi anemia and associated proteins. Cell. 2015;160(1-2):354-354.e1. doi: 10.1016/j.cell.2014.12.031.

30. Wang W. Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. Nat Rev Genet. 2007;8(10):735-748.

31. Fernández-Rodríguez J, Quiles F, Blanco I, Teulé A, Feliubadaló L, Valle JD, Salinas M, Izquierdo A, Darder E, Schindler D, Capellá G, Brunet J, Lázaro C, Pujana MA. Analysis of SLX4/FANCP in non-BRCA1/2-mutated breast cancer families. BMC Cancer. 2012;12:84. doi: 10.1186/1471-2407-12-84.

32. Catucci I1, Colombo M, Verderio P, Bernard L, Ficarazzi F, Mariette F, Barile M, Peissel B, Cattaneo E, Manoukian S, Radice P, Peterlongo P. Sequencing analysis of SLX4/FANCP gene in Italian familial breast cancer cases. PLoS One. 2012;7(2):e31038. doi: 10.1371/journal.pone.0031038.

33. Litim N, Labrie Y, Desjardins S, Ouellette G, Plourde K, Belleau P; INHERIT BRCAs, Durocher F. Polymorphic variations in the FANCA gene in high-risk non-BRCA1/2 breast cancer individuals from the French Canadian population. Mol Oncol. 2013;7(1):85-100. doi: 10.1016/j.molonc.2012.08.002.

34. Seal S, Barfoot R, Jayatilake H, Smith P, Renwick A, Bascombe L, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR, Rahman N; Breast Cancer Susceptibility Collaboration. Evaluation of Fanconi Anemia genes in familial breast cancer predisposition. Cancer Res. 2003;63(24):8596-8599.

35. Kottemann MC, Smogorzewska A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature. 2013;493(7432):356-363. doi: 10.1038/nature11863.

36. Tischkowitz MD, Morgan NV, Grimwade D, Eddy C, Ball S, Vorechovsky I, Langabeer S, Stöger R, Hodgson SV, Mathew CG. Deletion and reduced expression of the Fanconi anemia FANCA gene in sporadic acute myeloid leukemia. Leukemia. 2004;18(3):420-425.

37. Svendsen JM, Smogorzewska A, Sowa ME, O'Connell BC, Gygi SP, Elledge SJ, Harper JW. Mammalian BTBD12/SLX4 assembles a Holliday junction resolvase and is required for DNA repair. Cell. 2009;138(1):63-77. doi: 10.1016/j.cell.2009.06.030.

38. Kim Y, Spitz GS, Veturi U, Lach FP, Auerbach AD, Smogorzewska A. Regulation of multiple DNA repair pathways by the Fanconi anemia protein SLX4. Blood. 2013;121(1):54-63. doi: 10.1182/blood-2012-07-441212.

39. Gao M, Rendtlew Danielsen J, Wei LZ, Zhou DP, Xu Q, Li MM, Wang ZQ, Tong WM, Yang YG. A novel role of human holliday junction resolvase GEN1 in the maintenance of

centrosome integrity. PLoS One. 2012;7(11):e49687. doi: 10.1371/journal.pone.0049687.

40. Kuligina ESh, Sokolenko AP, Mitiushkina NV, Abysheva SN, Preobrazhenskaya EV, Gorodnova TV, Yanus GA, Togo AV, Cherdyntseva NV, Bekhtereva SA, Dixon JM, Larionov AA, Kuznetsov SG, Imyanitov EN. Value of bilateral breast cancer for identification of rare recessive at-risk alleles: evidence for the role of homozygous GEN1 c.2515_2519delAAGTT mutation. Fam Cancer. 2013 Mar;12(1):129-132. doi: 10.1007/s10689-012-9575-x.

41. Wechsler et al., 2011. Wechsler T, Newman S, West SC. Aberrant chromosome morphology in human cells defective for Holliday junction resolution. Nature. 2011;471(7340):642-646. doi: 10.1038/nature09790.

42. Rodrigue A, Coulombe Y, Jacquet K, Gagné JP, Roques C, Gobeil S, Poirier G, Masson JY. The RAD51 paralogs ensure cellular protection against mitotic defects and aneuploidy. J Cell Sci. 2013;126(Pt 1):348-359. doi: 10.1242/jcs.114595.

43. Giampietro PF, Adler-Brecher B, Verlander PC, Pavlakis SG, Davis JG, Auerbach AD. The need for more accurate and timely diagnosis in Fanconi anemia: a report from the International Fanconi Anemia Registry. Pediatrics. 1993;91(6):1116-1120.

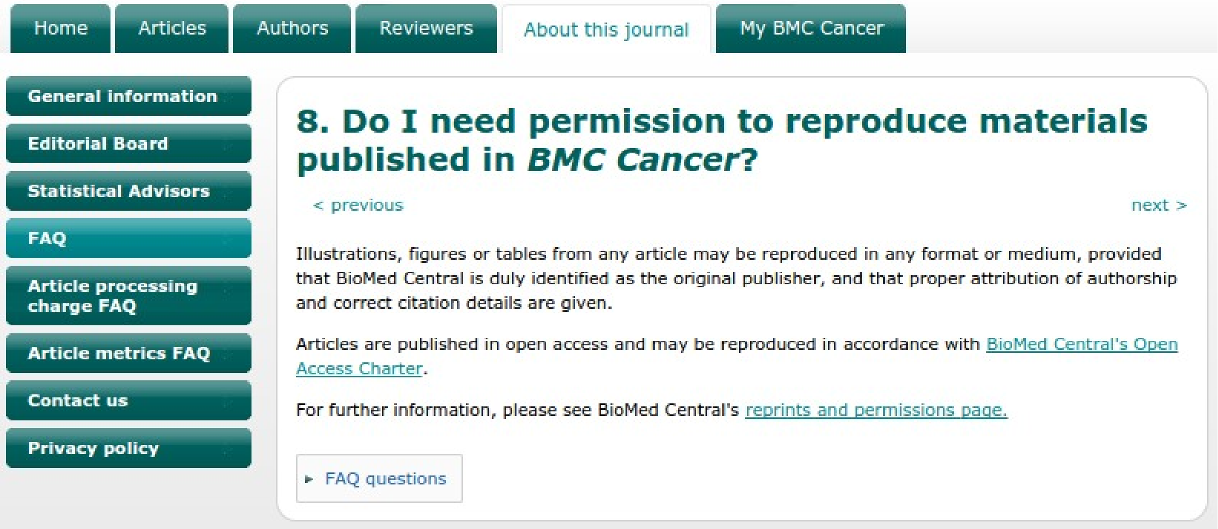
44. Welshimer K, Swift M. Congenital malformations and developmental disabilities in ataxia-telangiectasia, Fanconi anemia, and xeroderma pigmentosum families. Am J Hum Genet. 1982;34(5):781-793.

45. Wilkinson K, Velloso ER, Lopes LF, Lee C, Aster JC, Shipp MA, Aguiar RC. Cloning of the t(1;5)(q23;q33) in a myeloproliferative disorder associated with eosinophilia: involvement of PDGFRB and response to imatinib. Blood. 2003;102(12):4187-4190.

46. Jeffery J, Sinha D, Srihari S, Kalimutho M, Khanna KK. Beyond cytokinesis: the emerging roles of CEP55 in tumorigenesis. Oncogene. 2015. doi: 10.1038/onc.2015.128.

47. Paulsson K, Lilljebjörn H, Biloglav A, Olsson L, Rissler M, Castor A, Barbany G, Fogelstrand L, Nordgren A, Sjögren H, Fioretos T, Johansson B. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. Nat Genet. 2015. doi: 10.1038/ng.3301.

3.1.13. Reprint permission



8. Do I need permission to reproduce materials published in *BMC Cancer*?

Illustrations, figures or tables from any article may be reproduced in any format or medium, provided that BioMed Central is duly identified as the original publisher, and that proper attribution of authorship and correct citation details are given.

Articles are published in open access and may be reproduced in accordance with BioMed Central's Open Access Charter.

For further information, please see BioMed Central's reprints and permissions page.

Obtenu de http://old.biomedcentral.com/bmccancer/about/faq/reproduce.

3.2. ARTICLE III - A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells

3.2.1. Avant-propos

"*Clinical applications of next-generation sequencing and associated methods are emerging from ongoing large-scale discovery projects that have catalogued hundreds of genes as having a role in cancer susceptibility, onset and progression. For example, discovery cancer genomics has confirmed that many of the same genes are altered by mutation, copy number gain or loss, or structural variation across multiple tumor types, resulting in a gain or loss of function that likely contributes to cancer development in these tissues. Beyond these frequently mutated genes, we now know there is a 'long tail' of less frequently mutated, but probably important, genes that play roles in cancer onset or progression.*"

- Elaine Mardis, Genome Medicine 2014 [98]

Nous sommes également convaincus que limiter l'analyse des génomes tumoraux en retenant uniquement les gènes présentant une fréquence mutationnelle se détachant du *background* général nous conduirait à sous-estimer la complexité inhérente à ces génomes et à omettre l'existence de gènes ou de voies de signalisation dérégulés pourtant importants à l'échelle de la tumeur considérée. Toutefois, étant donné la complexité de caractérisation de ces évènements, nécessitant à la fois une analyse *in silico* orientée et des validations fonctionnelles confirmant leur potentiel *driver*, très peu d'études s'y sont intéressées. En conséquence, le nombre de *drivers* rares mis en évidence à ce jour reste très limité. Nous présentons ici une étude sous forme de preuve de concept ayant permis l'identification d'une de ces mutations.

A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells.

Jean-François Spinella[1], Pauline Cassart[1], Nicolas Garnier[1], Philippe Rousseau[2], Claire Drullion[1], Chantal Richer[1], Manon Ouimet[1], Virginie Saillour[1],Jasmine Healy[1], Chantal Autexier[2,3], Daniel Sinnett[1,4]

1) Sainte-Justine UHC Research Center, University of Montreal, Montreal, Qc, Canada; 2) Lady Davis Institute Jewish General Hospital, Montreal, Qc, Canada; 3) Departments of Anatomy, Cell Biology and Medicine, McGill University, Montreal, Qc, Canada; 4) Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, Qc, Canada.

This research was originally published in BMC Cancer. 2015;15:621. doi: 10.1186/s12885-015-1639-5.

3.2.2. Authors' contributions

DS is the principle investigator and takes primary responsibility for the paper. JFS, JH, and DS contributed to the conception and design of the study. JFS, PC, MO and CR were involved in sample and library preparation. VS analyzed raw sequence data and provided bioinformatics support. JFS performed whole-exome and targeted sequencing analysis. PC, PR, NG, and CD performed the functional validations: PC, NG and CD performed the apoptosis resistance assays, PR performed the telomere restriction fragment assays. JFS and JH wrote the paper and PC, NG, CA and DS contributed to the interpretation of the data and were involved in critical manuscript revision. All authors approved the final version.

3.2.3. Abstract

Background

The identification of oncogenic driver mutations has largely relied on the assumption that genes that exhibit more mutations than expected by chance are more likely to play an active role in tumorigenesis. Major cancer sequencing initiatives have therefore focused on recurrent mutations that are more likely to be drivers. However, in specific genetic contexts, low frequency mutations may also be capable of participating in oncogenic processes.Reliable strategies for identifying these rare or even patient-specific (private) mutations are needed in order to elucidate more personalized approaches to cancer diagnosis and treatment.

Methods

Here we performed whole-exome sequencing on three cases of childhood pre-B acute lymphoblastic leukemia (cALL), representing three cytogenetically-defined subgroups (high hyperdiploid, t(12;21) translocation, and cytogenetically normal). We applied a data reduction strategy to identify both common and rare/private somatic events with high functional potential. Top-ranked candidate mutations were subsequently validated at high sequencing depth on an independent platform and in vitro expression assays were performed to evaluate the impact of identified mutations on cell growth and survival.

Results

We identified 6 putatively damaging non-synonymous somatic mutations among the three cALL patients. Three of these mutations were well-characterized common cALL mutations involved in constitutive activation of the mitogen-activated protein kinase pathway (FLT3 p.D835Y, NRAS p.G13D, BRAF p.G466A). The remaining three patient-specific mutations (ACD p.G223V, DOT1L p.V114F, HCFC1 p.Y103H) were novel mutations previously undescribed in public cancer databases. Cytotoxicity assays demonstrated a protective effect of the ACD p.G223V mutation against apoptosis in leukemia cells. ACD plays a key role in protecting telomeres and recruiting telomerase. Using a telomere restriction fragment assay, we also showed that this novel mutation in ACD leads to increased telomere length in leukemia cells.

Conclusion

This study identified ACD as a novel gene involved in cALL and points to a functional role for ACD in enhancing leukemia cell survival. These results highlight the importance of rare/private somatic mutations in understanding cALL etiology, even within well-characterized molecular subgroups.

Keywords

Childhood acute lymphoblastic leukemia, rare/private somatic mutations, ACD/TPP1, telomere.

3.2.4. Background

Childhood acute lymphoblastic leukemia is a heterogeneous disease both biologically and clinically, and is the leading cause of cancer-related deaths among children. Despite significant advances in our understanding of the pathobiology of cALL leading to risk-based treatment regimens and increased survival rates, the etiological causes of this disease remain elusive. Approximately 75% of pre-B cALL cases exhibit hyperdiploidy or a recurring gross chromosomal rearrangement, detection of which is central to disease diagnosis, risk stratification and management [1]. While these chromosomal alterations play an important role in driving the leukemic process by affecting molecular pathways that halt lymphoid progenitor cell differentiation and promote cell proliferation and survival, they are not sufficient for leukemic transformation and are often detected years before leukemia onset. This suggests a need for additional cooperating events in order to achieve overt leukemogenesis [2]. Thorough investigation of cALL genomes is crucial to better understand the underlying genomic complexity of this disease and thus better diagnose and treat it. Toward these goals, recent large-scale sequencing efforts revealedmany somatic driver mutations/genes recurrently mutated in cALL [1]. The identification of these high frequency driver mutations is essential to better understand disease etiology and characterize prognostic subgroups. Yet accumulating evidence has also shown that low-frequency mutations within a cancer type can contribute to onset and progression of the disease [3], and play a role in intra-tumor and inter-patient heterogeneity. The identification of patient-specific mutations could provide crucial information regarding molecular pathways underlying cALL tumorigenesis and thus point to new therapeutic avenues. Here, we performed whole-exome sequencing of 3 pre-B cALL cases. Case 1 was diagnosed with high hyperdiploiy (>50 chromosome) cALL and Case 2 harboured the t(12;21) (*ETV6-RUNX1*) translocation. Together these two molecularly defined subgroups represent over 40% of cALL cases. Case 3 was cytogenetically normal at diagnosis. Using a unique quartet design, we sequenced matched normal (blood following remission) and tumor (bone marrow at diagnosis) patient samples, and the parents of each case. We successfully identified known recurrent drivers, as well as novel patient-specific somatic mutations with high functional potential. Using in vitro assays, we showed that the private p.G223V mutation, adjacent to the TEL patch of the telomere protein ACD (also known as TPP1), leads to apoptosis resistance and may contribute to leukemia cell survival by promoting telomere maintenance and protection.

3.2.5. Methods

3.2.5.1. Patients

All study subjects were self-declared French-Canadians of European descent from the established Quebec cALL (QcALL) cohort [4].The Sainte-Justine UHC Research Ethics Board approved the protocol. Written informed consent was obtained from the participants for publication of this report and any accompanying images. A copy of the written consent is available for review by the Editor of this journal.

Case 1, a 10 year old male, was classified as high risk based on his age. He presented with a platelet count of 15.0 x 10^9/L, a white blood cell count (WBC) of 9.0 x 10^9/L, and 63% and 97% lymphoblast cells in the blood and bone marrow samples respectively. The cytogenetic analysis revealed high hyperdiploidy (karyotype: 53,XY,+3,+4,+6,+10,+14,+17,+18). He was enrolled on the Dana Farber Cancer Institute (DFCI) ALL Consortium protocol 95-01. He achieved complete remission and has been out of treatment for over 60 months with leukemia free-survival (LFS).

Case 2 was a 4 year old female with a platelet count of 35.0 x 10^9/L, a WBC of 33.1 x 10^9/L, and 75% and 96% lymphoblast cells in the blood and bone marrow samples respectively. Cytogenetic analysis revealed a t(12;21) translocation. Despite her karyotype, usually associated with a good prognostic, she was classified as high risk based on the presence of leukemiacells in the cerebrospinal fluid. She was enrolled on DFCI/ALL Consortium protocol 2000-01. She also achieved complete remission and has been out of treatment for over 60 months with LFS.

Case 3 was a 6 year-old male with a WBC of 183.1 x 10^9/L, a platelet count of 29.0 x 10^9/L, 87% and 95% lymphoblast cells in the blood and bone marrow samples respectively, and presented with a normal karyotype according to the cytogenetic analysis. The patient was classified as high risk based on hyperleukocytosis and was treated on DFCI/ALL Consortium

Protocol 2000-01. Case 3 experienced a first relapse 41 months post diagnosis, and following 2 subsequent relapses he died 73 months post diagnosis.

3.2.5.2. Whole exome sequence capture and sequencing

DNA was extracted from bone marrow samples (at diagnosis) and peripheral blood samples (obtained after remission) from the cases and their parents using standard protocols as described previously [5].Whole exomes were captured in solution with Agilent's SureSelect Human All Exon 50Mb kits for SOLiD sequencing (Life Technologies) according to the manufacturer's protocol, were sequenced on the Life Technologies SOLiD System (mean coverage =32X) and aligned to the Hg19 reference genome using LifeScope Genomic Analysis Software (see Figure 1 for complete sequencing analysis workflow). Polymerase chain reaction (PCR) duplicates were removed using Picard [6]. Base quality score recalibration was performed using the Genome Analysis ToolKit (GATK) [7] and reads that failed the quality control were removed. Cleaned BAM files were used to create pileup files using SAMtools [8]. Somatic single nucleotide variants (SNVs) were called from pileup files using SNooPer, an in-house variant caller that is based on a machine learning approach that integrates tumoral and normal data and that was specifically trained for optimal identification of somatic mutations in our low-depth SOLiD sequencing data (manuscript submitted and software available upon request). Furthermore, using the familial design, we were able to use parental sequence information to remove Mendelian inconsistencies, reduce false-positive sequencing and alignment errors and facilitate somatic variant identification. The robustness of this approach was further demonstrated using high-depth sequencing on an independent platform to confirm top-ranked somatic mutations (see below). Resulting somatic SNVs were queried against publically available datasets such as 1000 Genomes [9] and NHLBI GO Exome Sequencing Project (ESP) [10] to filter out common polymorphism eventually remaining (minor allele frequency >0.01).

3.2.5.3. Expression data filter

Publically available microarray expression data [11] were used to filter variants based on the assumption that expressed genes are more likely to carry functionally relevant mutations [12]. Mutated genes identified from each case were compared to the expression profiles of the corresponding cALL subgroup (hyperdiploidy, t(12;21), or “others”) and only expressed genes were subsequently retained in the somatic variant analysis.

3.2.5.4. Ultra-deep targeted re-sequencing

Top-ranked rare/private candidate mutations, as well as common somatic mutations identified in the pre-B cALL cases were validated on the Ion Torrent system. Selected variant regions (~125pb flanking the identified somatic mutation) were amplified and PCR products were sequenced on the Ion PGM Sequencer (Life Technologies) according to the manufacturer’s protocol with a mean coverage >1,800X. Primer sequences and PCR protocols are available upon request.

3.2.5.5. Site-directed mutagenesis and apoptosis assay

Identified mutations were introduced by site-directed mutagenesis into the complementary DNA (cDNA) sequence of each gene cloned into a pDONR221 vector. Using the pLenti-CMV-DEST Gateway Vector (w118-1), we subcloned the wild-type (WT) or mutant coding sequences and generated lentiviruses using a Third Generation Packaging System in 293T cells. Lentiviral particles were then harvested and used to infect Nalm-6 (human leukemia pre-B cells) with 8ug/mL polybrene. Infected cells were selected with puromycin 1 μg/μL, seeded at $5x10^5$ cells/mL in their culture medium and treated with 5 μM of camptothecin for

3h. Apoptosis was measured using conventional Annexin V/Propidium iodide (PI) staining and quantified by flow cytometry (BD Biosciences FACS Fortessa). All experiments were done in triplicate.

3.2.5.6. Telomere restriction fragment (TRF) assay

Parental Nalm-6 cells and Nalm-6 cells overexpressing mutant ACD p.G223V (Nalm-6 ACD G223V), the wild-type ACD (Nalm-6 ACD WT) or Nalm-6 cells infected with the empty vector alone (Nalm-6 pLENTI) were used for TRF assays. For each cell line, extracted DNA at passage 16 (p16) (population doublings ≈50) and p27 (population doublings ≈80) after selection was digested with HinfI and RsaI restriction enzymes and fractionated on an electrophoresis gel apparatus. After drying, the gel was hybridized with a [γ -32P] adenosine triphosphate (ATP)end-labelled (T2AG3)3 probe and exposed on X-ray film. Mean TRF length was calculated as previously described [13]. The TRF assay was performed in duplicate.

3.2.6. Results and Discussion

Assuming that functionally important genes can also be mutated more rarely and in specific tumor contexts, we performed whole-exome sequencing of three pre-B cALL patients and their parents and applied a data reduction strategy to identify both common and novel rare/private events with high functional potential (Figure 1). SNVs were queried against public databases, annotated using ANNOVAR [14], and were binned according to frequency and function (Methods). Non-synonymous mutations that were expressed in cALL subgroups (based on microarray cALL expression data [11]) were then filtered based on a CONsensus DELeteriousness score >0.6 (measure of the degree of coherence of individual methods about the likelihood that a SNV is deleterious) [15]. This led to the identification of 6 expressed damaging non-synonymous somatic mutations among the 3 patients (FLT3 p.D835Y, NRAS p.G13D, BRAF p.G466A, ACD p.G223V, DOT1L p.V114F, HCFC1 p.Y103H) (Table1) that were subsequently confirmed using targeted ultra-deep re-sequencing (mean coverage >1800X). Among these 6 somatic SNVs, FLT3 p.D835Y, NRAS p.G13D and BRAF p.G466A were referenced in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (v71) and previously shown to constitutively activate the mitogen-activated protein kinase (MAPK) pathway and to increase tumor proliferation [16-19]. The identification of previously reported driver mutations validates the robustness of our approach. The NRAS p.G13D gain of function mutation was identified in Case 1, which corroborates previous studies that report postnatal RAS activating mutations in ~30% of pre-B hyperdiploid cALL patients [20]. The BRAF p.G466A mutation identified in Case 2 is associated with a mild increase of ERK activation [21]. While mutations in BRAF are identified in almost 70% of patients with multiple myeloma [17], screening of childhood pre-B cohorts only identified a few cases harbouring these events [22]. Although alterations of the receptor tyrosine kinase FLT3 are frequent in hyperdiploid cALL and rare in other subtypes [23], FLT3 p.D835Y was identified in the cytogenetically normal Case 3.While Case 3 suffered a relapse and activated forms of FLT3 are usually associated with a poor outcome in acute myeloid leukemia patients [24], this association is not confirmed in cALL [25].

Our sequencing analysis pipeline also led to the identification of three candidate novel pathogenic mutations in Cases 1 and 2 (DOT1L p.V114F, HCFC1 p.Y103H, ACD p.G223V), that were neither previously reported nor referenced in public databases. Based on our strict filtering criteria, no private mutations were identified in Case 3. DOT1L is a histone writer and HCFC1 is a broad transcription regulator that plays a critical role in cell proliferation via its involvement in chromatin-modifying activities [26]. These results are consistent with recent cancer sequencing initiatives that highlighted the important role of chromatin remodeling genes in leukemogenesis [27]. Furthermore, DOT1L has recently been implicated in the development of MLL-rearranged leukemia and shown to be essential for leukemic transformation [28]. On the other hand, ACD is a core protein in the shelterin complex and mediates the access of telomerase to the telomere. It is essential for telomere homeostasis in hematopoietic stem and progenitor cell particularly [29,30].

The high variant allele frequency (VAF) of these rare/private mutations (DOT1L, VAF =0.46; HCFC1, VAF =0.51; ACD, VAF =0.51), calculated based on our ultra-deep targeted re-sequencing data (Table 1), was indicative of early clonal selection supporting a possible functional role in leukemia development.Functional validation of the ACD p.G223V mutation (Figure 2A) was further pursued using in vitro expression assays, however our in vitro lentiviral expression system did not permit functional characterization of neither DOT1L nor HCFC1, due to the size of the open reading frames (ORFs) (4.6kb and 6.1kb, respectively) that were beyond the viral packaging capacity of the capside [31]. Further investigation of these two novel mutations using alternative functional screening methods is ongoing. In vitro expression assays of ACD using the topoisomerase I inhibitor camptothecin, showed that leukemic cells overexpressing mutant p.G223V ACD (Nalm-6 ACD G223V) exhibit lower levels of apoptosis compared with cells overexpressing wild-type ACD (Nalm-6 ACD WT) (P =0.014, Mann-Whitney U test) and the empty vector (Nalm-6 pLenti) (P =0.014, Mann-Whitney U test) (Figure 2B). The p.G223V mutation is adjacent to the TEL patch in the oligonucleotide/oligosaccharide-binding (OB) fold domain of ACD that interacts directly with the catalytic subunit of telomerase. Within the shelterin complex, ACD was specifically shown to interact with POT1 to protect telomeres and recruit telomerase at chromosome ends [32-34]. Interestingly, recurrent somatic mutations in the OB domains of POT1 have been shown to cause telomere dysfunction in chronic lymphocytic leukemia suggesting that alteration of shelterin-mediated protein-telomere binding could lead to genomic instability and cancer [35]. Furthermore, very recently, germline mutations in the POT1 binding domain and the TEL patch of ACD were respectively associated with familial melanomas and bone marrow failure disorders [36-38] (Figure 2A). Although further studies are needed to decipher the underlying molecular mechanisms implicating ACD in apoptosis inhibition, taken together, these data strongly support a role for ACD p.G223V in promoting leukemic cell maintenance.

To further investigate the effect of p.G223V mutant ACD on telomere structure, we performed a TRF assay (Figure 2C) and showed that decreased apoptosis correlates with increased telomere length in Nalm-6 ACD G223V at both tested passages compared to Nalm-6 ACD WT and Nalm-6 pLenti (P =0.029, Mann-Whitney U test), confirming stable alteration of telomere homeostasis. ACD mutant induced telomere elongation in leukemia cells is consistent with reports demonstrating disrupted shelterin complex function and telomere lengthening due to mutations in POT1 in chronic lymphocytic leukemia [35,39]. Furthermore, as observed for POT1 at the same population doubling point [35], overexpression of WT ACD did not increase telomere length (Figure 2C), confirming that the observed telomere lengthening is due to the p.G223V mutation and not caused by global overexpression of ACD. Concomitant lengthening of the telomeres and decreased apoptosis levels in NALM-6 ACD G223V following camptothecin treatment corroborates previous findings that altered telomerase activity can lead to hypersensitivity of tumor cells to topoisomerase inhibitors [40,41]. Further investigation is required to characterize the influence of the p.G223V mutation on the recruitment and processivity of telomerase and on telomere-length regulation. Missense or frameshift mutations in ACD, many of which are located in the OB fold, adjacent to the TEL patch, and in the POT1 interaction domain, have been described in multiple cancer types (Supplementary Table ST1) with the highest prevalence found in melanoma (2.7%), suggesting that ACD mutations may participate in a common underlying cancer-promoting pathway that involves telomere dysfunction. Case 1 who carries the ACD p.G223V mutation had high hyperdiploid cALL and although telomere dysfunction can cause chromosome destabilization and aneuploidy [42,43], it is unlikely that this mutation precluded the mitotic event producing the hyperdiploid phenotype. However, our results do support a role for ACD p.G223V in the earlier stages of clonal expansion contributing to telomere maintenance and apoptosis resistance, at least in vitro. Further investigation is required in order to characterize the mechanistic role of ACD p.G223V within this patient’s specific tumor context.

3.2.7. Conclusions

The prenatal origins of cALLs are well established, along with the need for additional postnatal mutations in order to drive overt leukemogenesis [44,45]. The extent to which rare/private genetic events contribute to the onset and progression of cALL however is largely unknown. Through whole-exome sequencing of 3 cALL cases, we successfully identified not only common drivers (NRAS p.G13D, FLT3 p.D835Y, and BRAF p.G466A), but also rare/private somatic mutations (DOT1L p.V114F, HCFC1 p.Y103H, ACD p.G223V), in well-characterized cALL molecular subgroups. The identification of patient-specific events with a functional potential is not surprising and further confirms the underlying genetic complexity of this disease. We went on to demonstrate the functional impact of ACD p.G223V on apoptosis resistance and telomere-length regulation in pre-B ALL cells. The high VAF of this somatic mutation suggests that it is likely present in the major subclone; while this does not necessarily imply functionality, it does support an early role for ACD p.G223V in driving the leukemic process. Though further investigation is needed to fully characterize the influence of the identified mutation on telomere homeostasis, this study is the first to describe the functional implications of a somatic mutation in ACD on leukemic cell behaviour, supporting a role for ACD and telomere regulation in leukemia cell resistance to apoptosis. In conclusion, these results support the need of thorough investigation of rare/private mutations to reveal the underlying complexity of cALL landscapes, including within well-characterized subgroups, and the inter-patient variability that may influence diagnosis and prognosis.

3.2.8. Figures

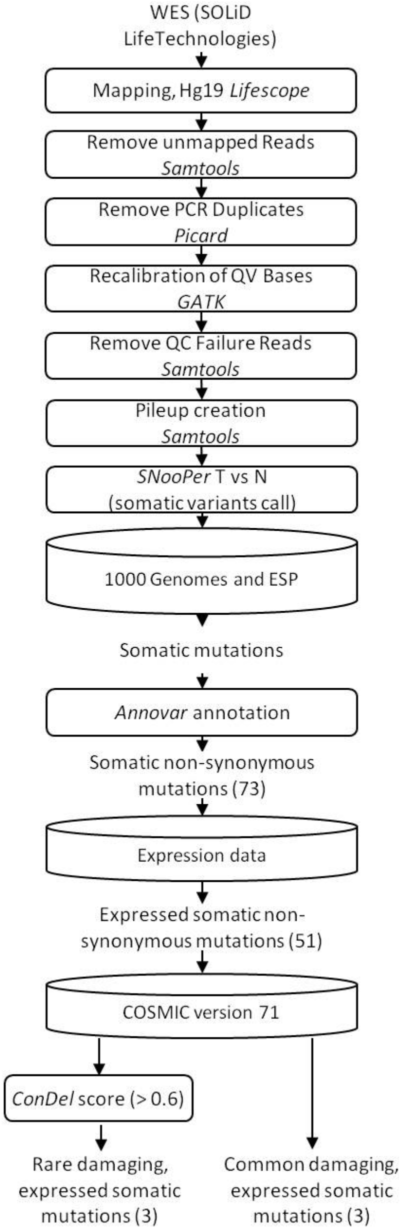

WES (SOLiD LifeTechnologies)
Mapping, Hg19 *Lifescope*
Remove unmapped Reads *Samtools*
Remove PCR Duplicates *Picard*
Recalibration of QV Bases *GATK*
Remove QC Failure Reads *Samtools*
Pileup creation *Samtools*
SNooPer T vs N (somatic variants call)
1000 Genomes and ESP
Somatic mutations
Annovar annotation
Somatic non-synonymous mutations (73)
Expression data
Expressed somatic non-synonymous mutations (51)
COSMIC version 71
ConDel score (> 0.6)
Rare damaging, expressed somatic mutations (3)
Common damaging, expressed somatic mutations (3)

Figure 1. Whole-exome sequencing analysis workflow. Boxes represent the analysis/cleaning steps. Cylinders represent the SNV filtering steps used in the data reduction strategy to identify functional somatic mutations. The number of variations remaining after each step is shown in brackets. Note that only SNVs that passed a given filter were tested for in the subsequent step. Using public databases and variant annotation tools, we identified 6 top-ranked mutations among the pre-B cALL patients, including 3 SNVs referenced in COSMIC v71 and 3 candidate rare/private SNVs in ACD (p.G223V), DOT1L (p.V114F) and HCFC1 (p.Y103H) with a potential functional impact.

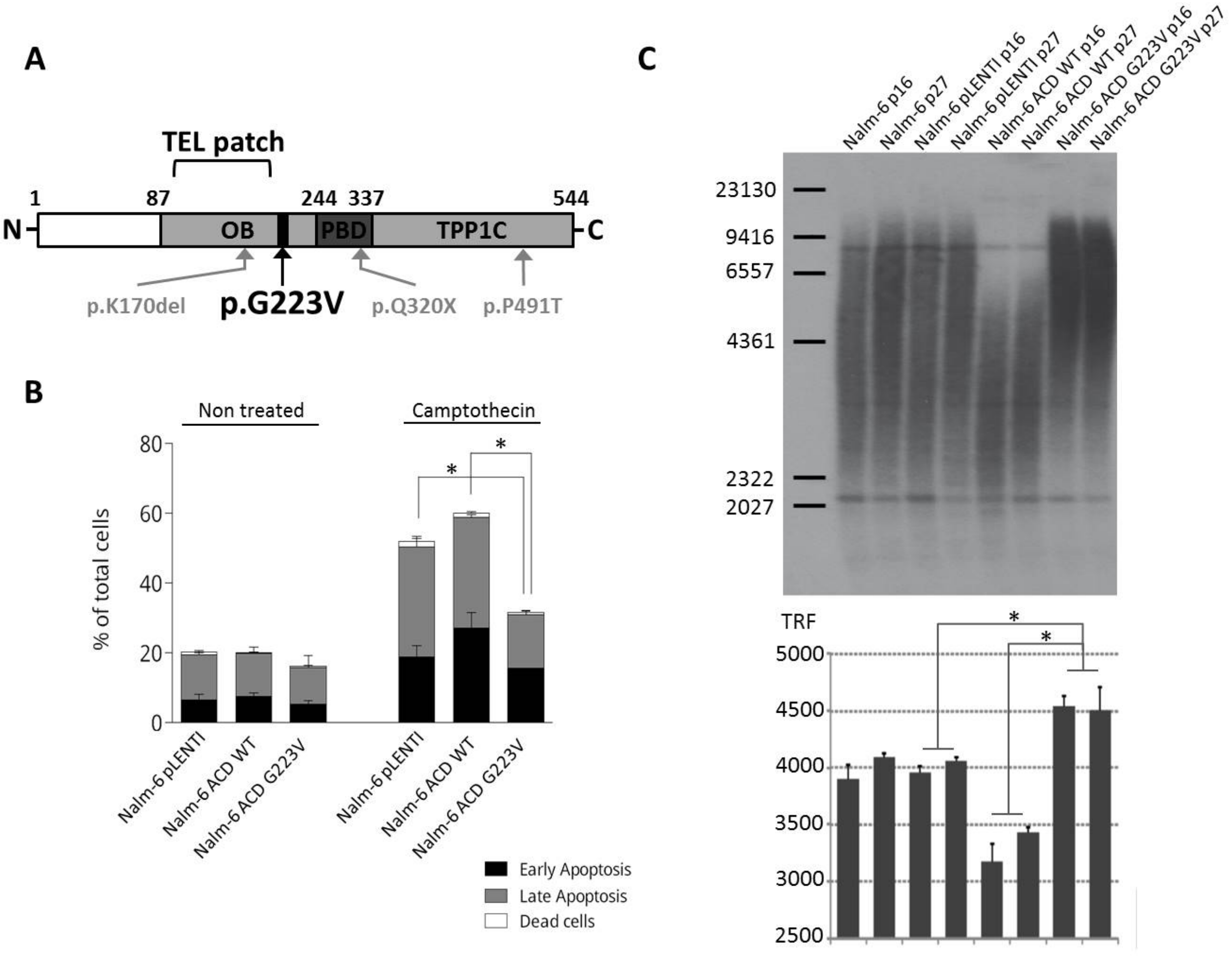


Figure 2. **ACD p.G223V protects from camptothecin-induced apoptosis and increases telomere length.** (**A**) Schematic representation of the ACD protein. The p.G223V mutation, depicted in black, is adjacent to the TEL patch of the OB-fold domain involved in telomerase recruitment and composed of seven critical amino acids located in a region defined by the curly bracket (E168, E169, E171, R180, L183, L212 and E215) [32]. p.Q320X, p.P491T and p.K170del, depicted in grey, are three germline mutations recently identified and associated with familial melanomas and bone marrow failure disorders [36-38]. TPP1C corresponds to the TIN2-binding domain. Together, the OB (oligonucleotide/oligosaccharide-binding) and PBD (POT1 binding domain) domains form the ACDN domain necessary for POT1 binding to telomeric DNA and the stimulation of telomerase processivity. (**B**) In vitro apoptosis assays show overall reduced levels of apoptosis associated with ACD p.G223V. The c.659g>t mutation was introduced into the ACD transgene by site-directed mutagenesis and expressed in Nalm-6 cells. The graph shows annexin V/PI staining for 3h on Nalm-6 pLenti (empty vector), Nalm-6 ACD WT and Nalm-6 ACD G223V cells. (**C**)The telomere restriction fragment assay (TRF) showed a quantitative increase in telomere size for Nalm-6 ACD G223V cells at passage 16 (p16) and p27 after selection. Mean TRF length = ∑ (ODi)/∑ (ODi/Li) where ODi is the radioactive signal, Li is the TRF fragment length at position i. The bar chart of Figure 2C (bottom) represents the mean TRF length for each condition directly quantified from each corresponding lane of the TRF gel presented at the top of Figure 2C. Significance (in B and C) was determined by a Mann-Whitney U test; p-values <0.05 are represented by an asterisk.

3.2.9. Tables

Table1. Candidate somatic mutations identified in each patient.

Patient	Subgroup	Gene	Genomic change	Protein change	Class	VAF	COSMIC v71
Case 1	HD	NRAS	g.chr1:115258744C>T	p.G13D	Missense	0.48	Haematopoietic and Lymphoid tissue
		TPP1/ACD	g.chr16:67693443C>A	p.G223V	Missense	0.51	-
		DOT1L	g.chr19:2191086G>T	p.V114F	Missense	0.46	-
Case 2	t(12;21)	BRAF	g.chr7:140481411C>G	p.G466A	Missense	0.15	Thyroid
		HCFC1	g.chrX:153230064A>G	p.Y103H	Missense	0.51	-
Case 3	CN	FLT3	g.chr13:28592642G>T	p.D835Y	Missense	0.44	Haematopoietic and Lymphoid tissue

Variant allele frequencies (VAF) (number of supporting reads/coverage) were calculated based on ultra-deep targeted re-sequencing data (mean coverage >1800X). Only tissue types harbouring the highest occurrence of the mutation in the COSMIC database (v71) are presented. HD: hyperdiploid; CN: Cytogenetically normal.

S1 Table. Non-synonymous and frameshift mutations in TPP1 referenced in the public cancer database COSMIC version 71.

CDS change	Protein change	COSMIC v71 ID	Count	Class	Domain
c.79G>A	p.G27R	COSM3670197	1	Missense	-
c.100C>T	p.R34*	COSM1519517	1	Nonsense	-
c.125C>T	p.A42V	COSM4141846	1	Missense	-
c.151C>T	p.L51F	COSM3511076	3	Missense	-
c.154C>A	p.L52I	COSM972546	2	Missense	-
c.167C>T	p.P56L	COSM3888782	1	Missense	-
c.181C>T	p.P61S	COSM141471	2	Missense	-
c.184C>T	p.L62F	COSM135719	2	Missense	-
c.187C>T	p.P63S	COSM3888781	1	Missense	-
c.188C>T	p.P63L	COSM3511075	1	Missense	-
c.221A>G	p.N74S	COSM4061942	1	Missense	-
c.229C>G	p.P77A	COSM1479001	1	Missense	-
c.257G>T	p.G86V	COSM4061941	1	Missense	-
c.283C>G	p.L95V	COSM119718	1	Missense	OB
c.294G>A	p.W98*	COSM703839	1	Nonsense	OB
c.335C>T	p.P112L	COSM3511074	1	Missense	OB
c.505T>G	p.F169V	COSM3932389	1	Missense	OB
c.559C>T	p.H187Y	COSM140768	1	Missense	OB
c.562G>A	p.V188I	COSM972544	1	Missense	OB
c.739G>A	p.A247T	COSM1302137	1	Missense	PBD
c.794A>G	p.Q265R	COSM4061939	1	Missense	PBD
c.865delC	p.H289fs*30	COSM972542	1	Frameshift	PBD
c.866A>G	p.H289R	COSM3370509	1	Missense	PBD
c.1054C>A	p.P352T	COSM3511073	1	Missense	TPP1C
c.1066C>A	p.P356T	COSM417319	1	Missense	TPP1C
c.1096G>A	p.G366S	COSM4061938	1	Missense	TPP1C
c.1109_1110CC>TT	p.S370F	COSM143470	1	Missense	TPP1C
c.1141C>T	p.P381S	COSM972540	1	Missense	TPP1C
c.1208G>A	p.C403Y	COSM1378975	1	Missense	TPP1C
c.1214C>T	p.A405V	COSM471951	1	Missense	TPP1C
c.1231C>A	p.P411T	COSM349226	1	Missense	TPP1C
c.1244A>C	p.H415P	COSM3932388	1	Missense	TPP1C
c.1253G>A	p.R418H	COSM1378974	1	Missense	TPP1C
c.1286C>T	p.P429L	COSM3511072	1	Missense	TPP1C
c.1301G>A	p.R434H	COSM1740245	1	Missense	TPP1C
c.1337C>T	p.T446I	COSM1749706	2	Missense	TPP1C
c.1396C>T	p.R466W	COSM1378973	1	Missense	TPP1C
c.1399C>T	p.P467S	COSM4061937	1	Missense	TPP1C
c.1400C>T	p.P467L	COSM1378972	1	Missense	TPP1C
c.1430G>A	p.G477E	COSM417320	1	Missense	TPP1C
c.1455G>T	p.W485C	COSM4061936	1	Missense	TPP1C
c.1474C>T	p.R492C	COSM1644358	1	Missense	TPP1C
c.1603G>T	p.G535W	COSM972538	1	Missense	TPP1C

TPP1C: TIN2-binding domain; OB: oligonucleotide/oligosaccharide-binding; PBD: POT1 binding domain.

3.2.10. Competing interests

The authors report no competing financial interests.

3.2.11. Acknowledgments

The authors are indebted to the patients and their parents for participating in this study. This study was supported by research funds provided by the Terry Fox Research Institute and the Canadian Institutes for Health Research. JFS is the recipient of a Réseau de médecine génétique appliquée (RMGA) Fellowship. NG is the recipient of postdoctoral Cole Foundation Fellowship. DS holds the François-Karl-Viau Research Chair in Pediatric Oncogenomics. Whole exome sequencing was performed at the Child Health Genomics Platform of the Sainte-Justine UHC Research Center; Targeted re-sequencing was performed at the McGill University and Genome Quebec Innovation Centre; Computations were made on the supercomputer Briarée from Université de Montréal, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

3.2.12. References

1. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet. 2013;381(9881):1943-1955. doi: 10.1016/S0140-6736(12)62187-4.
2. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. Nature reviews Cancer. 2003;3(9):639-649.
3. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609-615. doi: 10.1038/nature10166.
4. Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood. 2007;109(2):683-692.
5. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. Leuk Res. 1997;21(9):817-823.
6. http://picard.sourceforge.net
7. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-1303. doi: 10.1101/gr.107524.110.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25:2078-2079. doi: 10.1093/bioinformatics/btp352.
9. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-1073. doi: 10.1038/nature09534.
10. http://evs.gs.washington.edu/EVS/
11. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, Liu HC, Mahfouz R, Raimondi SC, Lenny N, Patel A, Downing JR. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood. 2003;102(8):2951-2959.
12. D'Antonio M, Ciccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 2013;14(5):R52. doi: 10.1186/gb-2013-14-5-r52.
13. D'Souza Y, Lauzon C, Chu TW, Autexier C. Regulation of telomere length and homeostasis by telomerase enzyme processivity. J Cell Sci. 2013;126:676-687. doi: 10.1242/jcs.119297.
14. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. doi: 10.1093/nar/gkq603.
15. Gonzalez-Perez A, Lopez-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-449. doi: 10.1016/j.ajhg.2011.03.004.
16. Yamamoto Y, Kiyoi H, Nakano Y, Suzuki R, Kodera Y, Miyawaki S, Asou N, Kuriyama K, Yagasaki F, Shimazaki C, Akiyama H, Saito K, Nishimura M, Motoji T, Shinagawa K, Takeshita A, Saito H, Ueda R, Ohno R, Naoe T. Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. Blood. 2001;97(8):2434-2439.
17. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson

A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA. Mutations of the BRAF gene in human cancer. Nature. 2002;417(6892):949-954.

18. Boguski MS, McCormick F. Proteins regulating Ras and its relatives. Nature. 1993;366(6456):643-654.

19. Bos JL, Toksoz D, Marshall CJ, Verlaan-de Vries M, Veeneman GH, van der Eb AJ, van Boom JH, Janssen JW, Steenvoorden AC. Amino-acid substitutions at codon 13 of the N-ras oncogene in human acute myeloid leukaemia. Nature. 1985;315(6022):726-730.

20. Wiemels JL, Kang M, Chang JS, Zheng L, Kouyoumji C, Zhang L, Smith MT, Scelo G, Metayer C, Buffler P, Wiencke JK. Backtracking RAS mutations in high hyperdiploid childhood acute lymphoblastic leukemia. Blood Cells Mol Dis. 2010;45(3):186-191. doi: 10.1016/j.bcmd.2010.07.007.

21. Wan PT, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, Good VM, Jones CM, Marshall CJ, Springer CJ, Barford D, Marais R, Cancer Genome Project. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell. 2004;116(6):855-867.

22. Knight T, Irving JA. Ras/Raf/MEK/ERK pathway activation in childhood acute lymphoblastic leukemia and its therapeutic targeting. Front Oncol. 2014;4:160. doi: 10.3389/fonc.2014.00160.

23. Armstrong SA, Mabon ME, Silverman LB, Li A, Gribben JG, Fox EA, Sallan SE, Korsmeyer SJ. FLT3 mutations in childhood acute lymphoblastic leukemia. Blood. 2004;103(9):3544-3546.

24. Ravandi F, Kantarjian H, Faderl S, Garcia-Manero G, O'Brien S, Koller C, Pierce S, Brandt M, Kennedy D, Cortes J, Beran M. Outcome of patients with FLT3-mutated acute myeloid leukemia in first relapse. Leuk Res. 2010;34(6):752-756. doi: 10.1016/j.leukres.2009.10.001.

25. Taketani T, Taki T, Sugita K, Furuichi Y, Ishii E, Hanada R, Tsuchida M, Sugita K, Ida K, Hayashi Y. FLT3 mutations in the activation loop of tyrosine kinase domain are frequently found in infant ALL with MLL rearrangements and pediatric ALL with hyperdiploidy. Blood. 2004;103(3):1085-1088.

26. Wysocka J, Myers MP, Laherty CD, Eisenman RN, Herr W. Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. Genes Dev. 2003;17(7):896-911.

27. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, Payne-Turner D, Churchman M, Andersson A, Chen SC, McCastlain K, Becksfort J, Ma J, Wu G, Patel SN, Heatley SL, Phillips LA, Song G, Easton J, Parker M, Chen X, Rusch M, Boggs K, Vadodaria B, Hedlund E, Drenberg C, Baker S, Pei D, Cheng C, Huether R, Lu C, Fulton RS, Fulton LL, Tabib Y, Dooling DJ, Ochoa K, Minden M, Lewis ID, To LB, Marlton P, Roberts AW, Raca G, Stock W, Neale G, Drexler HG, Dickins RA, Ellison DW, Shurtleff SA, Pui CH, Ribeiro RC, Devidas M, Carroll AJ, Heerema NA, Wood B, Borowitz MJ, Gastier-Foster JM, Raimondi SC, Mardis ER, Wilson RK, Downing JR, Hunger SP, Loh ML, Mullighan CG. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45(3):242-252. doi: 10.1038/ng.2532.

28. McLean CM, Karemaker ID, van Leeuwen F. The emerging roles of DOT1L in leukemia and normal development. Leukemia. 2014;28(11):2131-2138. doi: 10.1038/leu.2014.169.

29. Jones M, Osawa G, Regal JA, Weinberg DN, Taggart J, Kocak H, Friedman A, Ferguson DO, Keegan CE, Maillard I. Hematopoietic stem cells are acutely sensitive to Acd shelterin gene inactivation. J Clin Invest. 2014;124(1):353-366. doi: 10.1172/JCI67871.

30. Nakashima M, Nandakumar J, Sullivan KD, Espinosa JM, Cech TR. Inhibition of telomerase recruitment and cancer cell death. J Biol Chem. 2013;288(46):33171-33180. doi:

10.1074/jbc.M113.518175.

31. Campeau E, Ruhl VE, Rodier F, Smith CL, Rahmberg BL, Fuss JO, Campisi J, Yaswen P, Cooper PK, Kaufman PD. A versatile viral system for expression and depletion of proteins in mammalian cells. PLoS One. 2009;4(8):e6529. doi: 10.1371/journal.pone.0006529.

32. Nandakumar J, Bell CF, Weidenfeld I, Zaug AJ, Leinwand LA, Cech TR. The TEL patch of telomere protein TPP1 mediates telomerase recruitment and processivity. Nature. 2012;492(7428):285-289. doi: 10.1038/nature11648.

33. Zhong FL, Batista LF, Freund A, Pech MF, Venteicher AS, Artandi SE. TPP1 OB-Fold domain controls telomere maintenance by recruiting telomerase to chromosome ends. Cell. 2012;150(3):481-494. doi: 10.1016/j.cell.2012.07.012.

34. Sexton AN, Youmans DT, Collins K. Specificity requirements for human telomere protein interaction with telomerase holoenzyme. J Biol Chem. 2012;287(41):34455-34464. doi: 10.1074/jbc.M112.394767.

35. Ramsay AJ, Quesada V, Foronda M, Conde L, Martínez-Trillos A, Villamor N, Rodríguez D, Kwarciak A, Garabaya C, Gallardo M, López-Guerra M, López-Guillermo A, Puente XS, Blasco MA, Campo E, López-Otín C. POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. Nat Genet. 2013;45(5):526-530. doi: 10.1038/ng.2584.

36. Aoude LG, Pritchard AL, Robles-Espinoza CD, Wadt K, Harland M, Choi J, Gartside M, Quesada V, Johansson P, Palmer JM, Ramsay AJ, Zhang X, Jones K, Symmons J, Holland EA, Schmid H, Bonazzi V, Woods S, Dutton-Regester K, Stark MS, Snowden H, van Doorn R, Montgomery GW, Martin NG, Keane TM, López-Otín C, Gerdes AM, Olsson H, Ingvar C, Borg A, Gruis NA, Trent JM, Jönsson G, Bishop DT, Mann GJ, Newton-Bishop JA, Brown KM, Adams DJ, Hayward NK. Nonsense Mutations in the Shelterin Complex Genes ACD and TERF2IP in Familial Melanoma. J Natl Cancer Inst. 2014;107(2). pii: dju408.

37. Guo Y, Kartawinata M, Li J, Pickett HA, Teo J, Kilo T, Barbaro PM, Keating B, Chen Y, Tian L, Al-Odaib A, Reddel RR, Christodoulou J, Xu X, Hakonarson H, Bryan TM. Inherited bone marrow failure associated with germline mutation of ACD, the gene encoding telomere protein TPP1. Blood. 2014;124(18):2767-2774.

38. Kocak H, Ballew BJ, Bisht K, Eggebeen R, Hicks BD, Suman S, O'Neil A, Giri N; NCI DCEG Cancer Genomics Research Laboratory; NCI DCEG Cancer Sequencing Working Group, Maillard I, Alter BP, Keegan CE, Nandakumar J, Savage SA. Hoyeraal-Hreidarsson syndrome caused by a germline mutation in the TEL patch of the telomere protein TPP1. Genes Dev. 2014;28(19):2090-2102. doi: 10.1101/gad.248567.114.

39. Hockemeyer D, Sfeir AJ, Shay JW, Wright WE, de Lange T. POT1 protects telomeres from a transient DNA damage response and determines how human chromosomes end. EMBO J. 2005;24(14):2667-2678.

40. Cerone MA, Londoño-Vallejo JA, Autexier C. Mutated telomeres sensitize tumor cells to anticancer drugs independently of telomere shortening and mechanisms of telomere maintenance. Oncogene. 2006;25(56):7411-7420.

41. Ludwig A, Saretzki G, Holm PS, Tiemann F, Lorenz M, Emrich T, Harley CB, von Zglinicki T. Ribozyme cleavage of telomerase mRNA sensitizes breast epithelial cells to inhibitors of topoisomerase. Cancer Res. 2001;61(7):3053-3061.

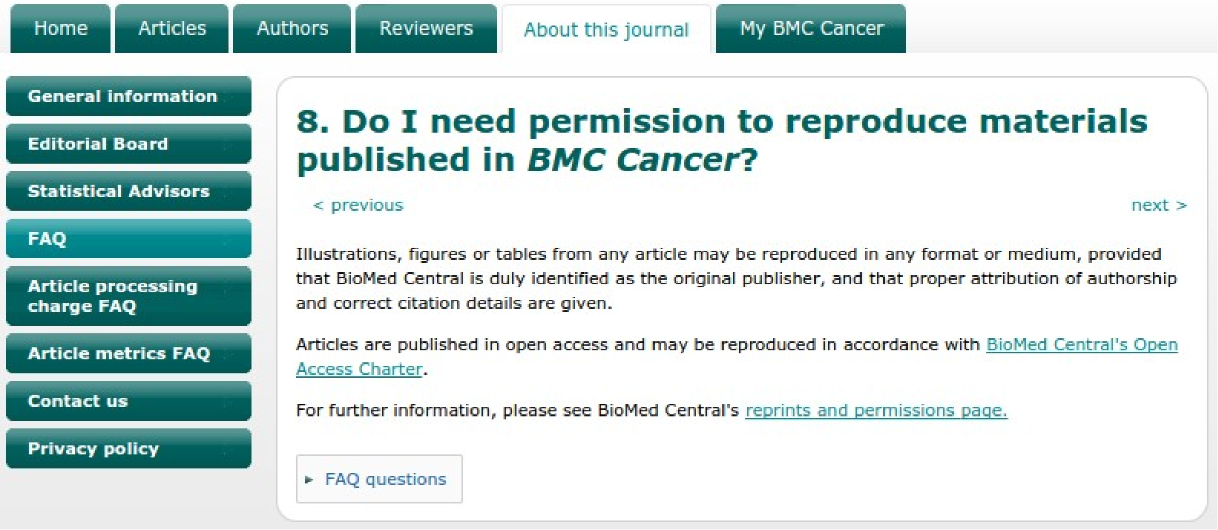
42. Pathak S, Multani AS, Furlong CL, Sohn SH. Telomere dynamics, aneuploidy, stem cells, and cancer (review). Int J Oncol. 2002;20(3):637-641.

43. Martinez P, Blasco MA. Telomeric and extra-telomeric
roles for telomerase and the telomere-binding proteins. Nature Review. 2011;11(3):161-176. doi: 10.1038/nrc3025.

44. Maia AT, van der Velden VH, Harrison CJ, Szczepanski T, Williams MD, Griffiths MJ, van Dongen JJ, Greaves MF. Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins. Leukemia 2003;17:2202-2206.

45. Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, Hows JM, Navarrete C, Greaves M. Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proc Natl Acad Sci U S A. 2002;99(12):8242-8247.

3.2.13. Reprint permission



8. Do I need permission to reproduce materials published in *BMC Cancer*?

< previous next >

Illustrations, figures or tables from any article may be reproduced in any format or medium, provided that BioMed Central is duly identified as the original publisher, and that proper attribution of authorship and correct citation details are given.

Articles are published in open access and may be reproduced in accordance with BioMed Central's Open Access Charter.

For further information, please see BioMed Central's reprints and permissions page.

▸ FAQ questions

Obtenu de http://old.biomedcentral.com/bmccancer/about/faq/reproduce.

3.3. Le paysage génomique de la LAL-B pédiatrique

Au cours de cette thèse, j'ai également eu l'occasion d'analyser les données de séquençage d'exome de plus de 150 patients LAL-B provenant de la cohorte QcALL du CHU Sainte-Justine (Life Technologies SOLiD 4/5500 System, paired-end: 50x35pb, moyenne de profondeur de couverture dans la région ciblée: 35X). Les données brutes ont été nettoyées et préparées comme indiqué dans l'article précédent. Une série de mutations somatiques (SNVs et indels) ont été identifiées par SNooPer et validées par re-séquençage ciblé (Illumina TruSeq Custom Amplicon assay séquencé sur HiSeq2500 device, paired-end: 2x100pb, moyenne de profondeur de couverture: 2,500X). Des gènes *drivers* communs comme *KRAS*, *NRAS*, *FLT3*, *CREBBP* ou *WHSC1* ont été identifiés comme fréquemment mutés au sein de notre cohorte. Nous avons également identifié une série de gènes mutés chez un nombre limité de patients, voire chez un seul d'entre eux, mais présentant comme *ACD* un potentiel fonctionnel.

Afin de prioriser sans à priori les gènes identifiés (par exemple, sans considérer les fréquences mutationnelles au sein de la cohorte), nous avons réalisé un criblage haut-débit “perte de fonction” par l'utilisation d'ARNi (ARN interférence) ciblant ces gènes. Cette approche a déjà été utilisée avec succés afin de caractériser fonctionnellement et à large échelle une série de lignées cellulaires cancéreuses (http://www.broadinstitute.org/achilles) [143,144]. Afin de construire notre *pool* de shARNs lentiviraux (ARNs en épingle à cheveux), nous avons sélectionné 5 shARNs par gène au sein de la librairie MISSION (The RNAi Consortium) pour un total d'environ 1,000 clones shARN ciblant un peu moins de 200 gènes candidats et contrôles. Les criblages ont ensuite été réalisés dans 2 lignées cellulaires LAL pré-B bien caractérisées (697 et Nalm-6) ainsi que dans la lignée cellulaire lymphoblastoïde humaine (GM12878). Les cellules en cultures ont été récoltées à intervalles réguliers: 24h post-transfection et 24h après la sélection à la puromycine (T0), ainsi que tous les 6 jours ensuite (T6, T12). À chaque temps, l'ADN génomique des cellules récoltées était extrait. Nous avons mesuré l'enrichissement ou la déplétion des clones shARNs afin d'identifier les gènes suppresseurs de tumeur (ceux dont la réduction d'expression apporte un avantage sélectif) ou les oncogènes (ceux dont la réduction d'expression entraîne un désavantage). Pour ce faire, les inserts shARNs ont été amplifiés par PCR depuis l'ADN génomique en utilisant des *primers* universels entourant leurs séquences. Les produits de PCR ont ensuite été séquencés (>1000X) sur Illumina HiSeq2500 et les *reads* générées ont été alignées sur les séquences de shARNs. Le ratio Tf:T0 du compte de *reads* normalisé a été mesuré pour chaque shARN (Figure XVI), Tf (T final) étant T12 ici. Les shARNs présentant des ratios s'éloignant de plus de 2 déviations standards au-dessus et en-dessous du ratio moyen ont été considérés comme ciblant des suppresseurs de tumeur et des oncogènes, respectivement (Figure XVI, Tableau V). Les gènes ont été priorisés en fonction du nombre de shARNs significatifs provoquant le même effet. Ces résultats préliminaires montrent que le *knockdown* d'*ACD* entraîne un avantage sélectif suffisamment important pour que celui-ci soit considéré comme suppresseur de tumeur potentiel au sein des lignées cellulaires 697 et GM12878. Bien que non significatif à Tf dans les cellules Nalm-6, le *knockdown* d'*ACD* entraîne également un avantage sélectif qui pourrait devenir significatif dans le cadre d'une prolongation de l'expérience. Ces données confirment les résultats présentés dans le manuscrit précédent (A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells) et supportent l'importance des *drivers* rares dans le processus leucémogène. Les cribles présentés, ainsi que certains des candidats identifiés, sont en phase de validation au moment de la rédaction de cette thèse.

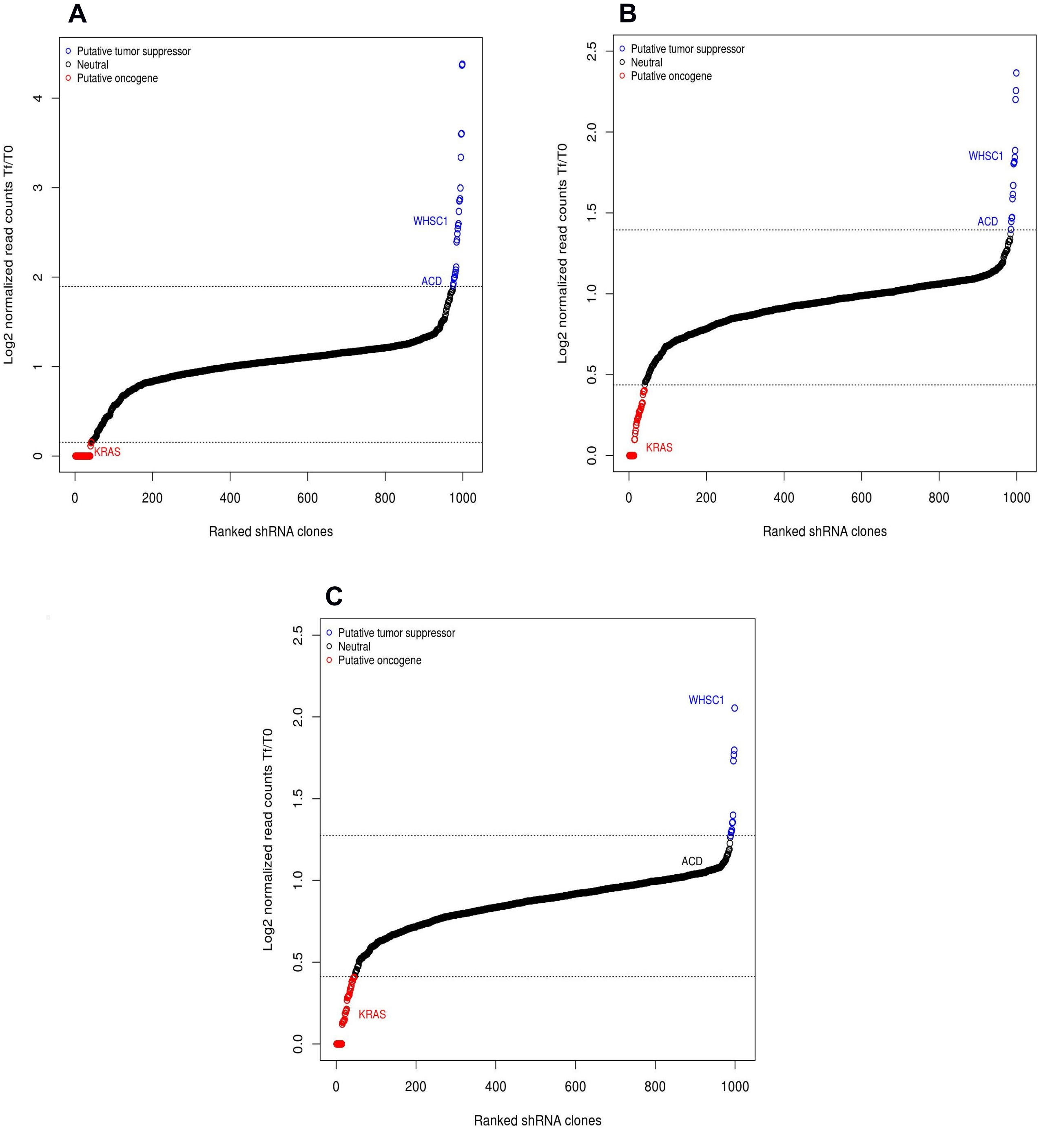


Figure XVI. Criblage fonctionnel shARNs. Représentations ordonnées du ratio Tf:To obtenu pour chaque shARN dans les lignées cellulaires (**A**) GM12878, (**B**) 697 et (**C**) Nalm-6. Les lignes pointillées sont tracées au niveau des valeurs correspondant à la moyenne des ratios Tf:To +/- 2 déviations standards. Les shARNs au-delà de ces valeurs sont considérés comme des oncogènes (rouge) ou suppresseurs de tumeur (bleu) putatifs. *KRAS* et *WHSC1* sont indiqués à titre de contrôles oncogène et tumeur suppresseur, respectivement.

Tableau V. Oncogènes et suppresseurs de tumeur putatifs.

Oncogenes			Tumor suppressors		
GM12878	**697**	**Nalm6**	**GM12878**	**697**	**Nalm6**
ACACA	BCAP31	DDX52	ACD	ACD	ACACA
ASB13	C2orf16	DIAPH3	C2orf16	ALDH1B1	ALDH1B1
CALM1	CALM1	FBXW7	KDM2A	C2orf16	PKP2
CTCF	CTCF	HEY2	MIA3	PDF	WHSC1
KRAS	DDX52	KDM2A	PDF	PTGDR	
NPNT	FLT3	KRAS	STAT3	WHSC1	
PKP2	KDM2A	MED15	TFAP2D		
SFXN3	KRAS	MIA3	UMODL1		
TAF1A	MIA3	NPNT	WHSC1		
TIAM1	RPLP0	PKP2			
ZKSCAN1	SFXN3	SFXN3			
ZNF460	SNCAIP	SNCAIP			
ZNF827	STAT3	TFAP2D			
	STX12	TIAM1			
	TFAP2D	TRIP12			
	TRRAP	TUBGCP2			
	TUBGCP2	U2AF1			
	U2AF1	UMODL1			
	UMODL1	USP9X			
	ZKSCAN1	ZKSCAN1			
	ZNF460	ZNF502			
	ZNF502	ZNF827			

CHAPITRE 4

ARTICLE IV - Genomic characterization of pediatric T-Cell acute lymphoblastic leukemia reveals novel recurrent driver mutations

4.1. Avant-propos

Jusque récemment, la LAL-T était associée à un taux de survie significativement plus faible que son pendant touchant la lignée lymphocytaire B. L'utilisation d'une chimiothérapie intensive a permis d'améliorer drastiquement le nombre de rémission. Toutefois, la LAL-T souffre encore d'une caractérisation insuffisante et les patients atteints présentent toujours un risque important de se montrer réfractaires aux traitements classiques, de subir des rechutes précoces ou une atteinte du système nerveux central [145]. Ceci est particulièrement vrai pour les formes présentant des lymphoblastes bloqués aux stades les plus précoces de différenciation. D'ailleurs, très récemment, un nouveau groupe appelé ETP-ALL (*Early T-cell Precursor ALL)* et associé à un risque particulièrement élevé, a été identifié. Les rares patients appartenant à ce groupe présentent des cellules bloquées à un stade si précoce qu'elles s'apparentent à des cellules souches hématopoïétiques, voire à des progéniteurs myéloïdes [146].

Pour améliorer à terme la prise en charge des patients, un effort de caractérisation génomique en profondeur des différents groupes de LAL-T, incluant les ETP-ALLs, est nécessaire afin de recenser l'ensemble des événements somatiques, rares ou récurrents, participant au développement ou à la progression de la maladie. Nous avons tenté d'y contribuer en menant les travaux qui composent l'article IV de cette thèse.

Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations.

Jean-François Spinella[1], Pauline Cassart[1], Chantal Richer[1], Virginie Saillour[1], Manon Ouimet[1], Sylvie Langlois[1], Pascal St-Onge[1], Thomas Sontag[1], Jasmine Healy[1], Mark D. Minden[2], Daniel Sinnett[1,3]

1) CHU Sainte-Justine Research Center, Université de Montréal, Montreal, Qc, Canada; 2) Princess Margaret Cancer Centre, University Health Network, Toronto, On, Canada; 3) Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, Qc, Canada.

This research was originally published in Oncotarget. 2016. doi: 10.18632/oncotarget.11796.

4.2. Authors' contributions

DS is the principal investigator and takes primary responsibility for the paper. JFS, JH, and DS contributed to the conception and design of the study. TS performed DNA and RNA extractions for patient samples. PC, CR, SL and MO were involved in sample and library preparation for whole-exome, RNA-seq and targeted sequencing. VS and PSO contributed to data processing. JFS performed bioinformatics analysis, data integration and analysis. PC carried out functional assays. MM provided the samples of the additional cohort of adult T-ALL patients. JFS drafted the paper and JFS, JH, PC and DS contributed to interpretation of the data and were involved in critical revision of the manuscript. All authors approved the final version.

4.3. Abstract

T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive hematologic malignancy with variable prognosis. It represents 15% of diagnosed pediatric ALL cases and has a threefold higher incidence among males. Many recurrent alterations have been identified and help define molecular subgroups of T-ALL, however the full range of events involved in driving transformation remain to be defined. Using an integrative approach combining genomic and transcriptomic data, we molecularly characterized 30 pediatric T-ALLs and identified common recurrent T-ALL targets such as *FBXW7,JAK1*, *JAK3*, *PHF6, KDM6A* and *NOTCH1* as well as novel candidate T-ALL driver mutations including the p.R35L missense mutation in splicesome factor *U2AF1* found in 3 patients and loss of function mutations in the X-linked tumor suppressor genes *MED12* (frameshit mutation p.V167fs, splice site mutation g.chrX:70339329T>C, missense mutation p.R1989H) and *USP9X* (nonsense mutation p.Q117*). In vitro functional studies further supported the putative role of these novel T-ALL genes in driving transformation. *U2AF1* p.R35L was shown to induce aberrant splicing of downstream target genes, and shRNA knockdown of *MED12* and *USP9X* was shown to confer resistance to apoptosis following T-ALL relevant chemotherapy drug treatment in Jurkat leukemia cells. Interestingly, nearly 60% of novel candidate driver events were identified among immature T-ALL cases, highlighting the underlying genomic complexity of pediatric T-ALL, and the need for larger integrative studies to decipher the mechanisms that contribute to its various subtypes and provide opportunities to refine patient stratification and treatment.

Keywords

T-cell acute lymphoblastic leukemia, X-linked tumor suppressor, *MED12*, *USP9X*, *U2AF1*.

4.4. Introduction

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer, accounting for 25% of all pediatric tumors [1]. Despite continued refinement of childhood ALL subtype classification and improved risk-based treatment strategies, survival rates remain significantly lower among high-risk patients [2]. Pediatric T-cell ALL (T-ALL) represents 10-15% of ALL cases [3] with a quarter of the patients experiencing relapse, and lower post-relapse survival compared to the more common B-lineage ALL [1]. Interestingly, a threefold higher incidence is observed among males [4], however the biological implications underlying this gender bias remain poorly understood. Despite the introduction of intensified chemotherapy protocols, very few inroads into new therapeutic approaches for these high-risk patients have been made. Recent studies [5-8] have shown that further classification of T-ALL could reveal new diagnostic markers and provide alternative targeted treatment options.

Immunophenotypic and gene expression signature analyses revealed a limited number of T-ALL subtypes based largely on differential expression of surface antigen markers and oncogene expression signatures related to stage-specific T-cell developmental arrest [3, 9, 10]. Gene fusions involving the juxtaposition of transcription factor proto-oncogenes under the control of T-cell specific enhancers located in the *TCRB* (7q34) or *TCRA-TCRD* (14q11) have been shown to be essential driver events in T-ALL and further define molecular subtypes [9]. Additional recurrent, as well as cryptic chromosomal rearrangement events that lead to T-cell specific proto-oncogene activation have also been described and some have shown prognostic significance. For instance, CALM-AF10 resulting from the t(10;11)(p13;q14-21) translocation is one of the most frequent fusion events found in 10% of childhood T-ALL cases and has been associated with poor prognosis, particularly among immature T-ALL patients [11, 12].

Recent studies have used comprehensive genomic approaches to gain further insight into the mutational landscape of T-ALL and have led to the identification of novel disease mechanisms [6, 8] and recurrent somatic alterations with pathogenic relevance. The most prevalent are constitutive activation of NOTCH1 signaling, observed in up to 60% of T-ALL patients [13], and loss of the *CDKN2A/p16INK4a* (chromosome 9p21) locus [14], occurring in up to 70% of cases. Loss of function mutations in *FBXW7* are also frequent in T-ALL (about 15% of cases) and contribute to sustained NOTCH1 activation by preventing its proteasomal degradation in the nucleus [15]. Other frequently altered gene/pathway categories in T-ALL include signal transduction (*PTEN, JAK1, JAK3, NF1, NRAS, IL7R* and *FLT3*), transcription factors (*WT1, LEF1, ETV6, GATA3* and *BCL11B*)as well as chromatin remodeling (*EZH2, SUZ12, EED* and *KDM6A/UTX*) [15]. Such studies have also identified a distinct, very aggressive T-ALL subtype defined by very early arrest in T-cell development [16, 17, 5]. These early T-cell precursor ALLs (ETP-ALLs) were genetically characterized by activating mutations in genes regulating cytokine and RAS signaling, inactivating mutations in hematopoietic development genes and histone-modifying genes [5].

These observations suggest that large, integrative efforts will continue to yield further insight into childhood T-ALL, particularly given that the pathogenesis of this disease and of its various subtypes cannot be entirely explained by the data currently available. In this study, we used a combination of exome and transcriptome sequencing, as well as high-density genotyping to characterize 30 childhood T-ALLs. We identified common recurrent mutations in known T-ALL genes (e.g. *NOTCH1, PHF6, FBXW7* and *JAK3*) as well as novel somatic mutations in genes involved in RNA splicing (*U2AF1*), chromatin remodeling (*KMT2C/MLL3*) and of particular interest given the observed male gender bias, in X-linked genes *MED12* and

USP9X. Additional functional studies provide evidence of the potential implication of these newly identified mutations in childhood T-ALL pathogenesis. Overall, our findings indicate a need for further large-scale genomic investigations to refine patient stratification and optimize treatment strategies in childhood T-ALL.

4.5. Results

4.5.1. The genomic landscape of childhood T-ALL

Partially overlapping data from cytogenetic analysis, whole exome and genome sequencing, ultra-deep targeted re-sequencing, and RNA sequencing were available at diagnosis for 30 childhood T-ALL patients (matched normal-tumor) and at relapse for two of these patients (Figure 1 and Table 1).Immunophenotyping and gene expression data, when available, were used to classify patients according to T-cell maturation status (Table 1, S1 Fig, S1 Table, S2 Table and Supplementary information). Seven patients (324, 432, 706, 716, 748, 791 and 879) clustered as early immature T-ALLs, among whom 2 patients (791 and 879) showed immunophenotype and expression markers indicative of an early T-cell precursor ALL (ETP-ALL) phenotype [16]. Thirteen patients (340, 341, 437, 544, 547, 636, 647, 693, 727, 743, 744, 759 and 849) were classified as mature T-ALLs and 10 patients could not be classified due to insufficient data (Table 1).

We identified several structural chromosomal abnormalities within our cohort (Figure 1 and Supplementary information). Mature T-ALL patients 547, 759 and 636 were shown to carry translocations t(1;14)(p34;q11), t(11;14)(p13;q11) and t(10;14)(q24;q11) respectively, leading to the juxtaposition of the *TAL1*, *LMO2* and *TLX1/HOX11* oncogenes to the T-cell receptor alpha/delta (*TCRA/D*) at locus 14q11 [18-20]. We also identified a rarer translocation t(1;7)(p32;q34)/TRB-TAL1 in the mature T-ALL patient 849 [9] as well as the well-known t(10;11)(p12;q14) CALM-AF10 translocation in both ETP-ALL cases 791 and 879 and a t(9;22)(q34;q11.2)/BCR-ABL in the immature T-ALL patient 748 which is very rare in T-ALL (~1%) [21, 3]. Above >80% of the patients with available expression data showed activation of at least one (proto-)oncogene such as *TAL1, TLX3, FLT3, LMO2, LYL1* and *PIM1* with no evidence of a related fusion event (S1A Fig and S2 Table). For example, *LMO2* was shown to be upregulated in the mature T-ALL patient 547 and four early immature T-ALL patients including both ETP-ALLs (432, 748, 791 and 879), while none of these patients were identified as carriers of a *LMO2* activating translocation. *LYL1* was upregulated in all but one early immature T-ALL patient (716) without an associated translocation.

On average, we identified 29 somatic SNVs/indels and 37 somatic CNVs per tumor (Figure 1 and S3 Table). Based on strict filtering criteria (Methods), we identified a total of 68 candidate driver SNVs/indels (55 distinct mutations) across 28 genes among the 30 pediatric T-ALL patients and all patients harboured at least one candidate driver mutation (Figure 1 and S3 Table). RNA-seq data, when available, confirmed expression of 84% of the mutated alleles (21/25) (Figure 1, S2 Fig). "Hemopoiesis/T-cell differentiation" was the most frequently altered pathway among the cohort with 80% of patients carrying mutations in 9 genes affecting this pathway. "Post/Transcriptional regulation" (14 genes), "Chromatin modification/assembly" (6 genes), "Notch signaling” (6 genes) and "Regulation of cell cycle" (6 genes) were also found to be frequently altered. 34 of the reported candidate driver mutations were previously reported (COSMIC 72) among which 29 in hematopoietic malignancies (COSMIC v72) including 26 in known T-ALL driver genes such as *FBXW7,JAK1*, *JAK3*, *PHF6, KDM6A* and *NOTCH1.* These variations were mostly clonal (mean variant allele frequency - VAF =0.48, standard deviation - SD =0.10), confirming their presence in the majority of tumor cells at diagnosis and their initiating role in T-ALL (S3 Fig and Supplementary information). RAS pathway mutations had significantly lower frequencies compared to these common drivers with a mean VAF =0.33 (SD =0.11) (p =0.006, Mann-Whitney-U test) (S3 Fig and Supplementary information). Subclonality of these mutations corroborates previous reports [22-24] that describe a secondary role for Ras mutations in T-ALL occurring later in tumor progression. Fifteen patients (50%) harbored *NOTCH1* mutations and two thirds (12/18) of identified *NOTCH1* mutations were located in exons 26 and 27, coding for the extracellular heterodimerization domain and were mainly missense mutations (11/12). The remaining 6 *NOTCH1* mutations were located in exon 34 coding for the C-terminal PEST (4/6) and transactivation (2/6) domains and consisted mainly of truncating mutations (5/6), as previously reported [25]. Recurrent events also involved the tumor suppressor locus 9p21 with *CDKN2A/2B* (*p16INK4a* and *p15INK4b*) deletions occurring in 17 (57%) of our childhood T-ALL cases (Figure 1 and S3 Table), of which 14 were biallelic. For 5/17 patients, the 9p21 deletion event included the hematopoiesis regulator MLLT3 [26], 6/17 included the long non-coding RNA MIR31HG recently shown to regulate *CDKN2A* expression [27], and 9/17 included the methylthioadenosine phosphorylase MTAP. Although the associations require validations in larger cohorts, patients mutated for *CDKN2A* had significantly less chance of relapse (p =0.0121, Fisher's Exact test). Interestingly, immature T-ALL patients (early immature T-ALL and ETP-ALL together) had significantly less *CDKN2A* alterations (p =0.0044, Fisher's Exact test) and a higher risk of relapse (p =0.0072, Fisher's Exact test).

Twenty-one variations had not previously been reported and thus were considered as novel T-ALL driver mutations, including 3 in *NOTCH1* (p.N1603K, p.P2475fs and p.L2326fs) as well as novel predicted deleterious mutations in the well-known X-linked *PHF6* (p.E221* and p.G226fs) and *KDM6A/UTX* (p.Q692*) [8, 28] (Figure 1 and S3 Table). Among these novel putative childhood T-ALL drivers, we identified a number of mutations in chromatin remodeling genes *EHMT1* (n =4), *WHSC1* (n =3) and *KMT2C/MLL3* (n =1), a recurrent missense mutation (p.R35L) in the first zinc finger of splicing factor *U2AF1* (n= 3), as well as loss of *ABL1* (n=2) and in *MLH1* (n=2), both involved in DNA repair. Two patients had gain of copy of *JAG2* which functions in the Notch signaling pathway, and two novel missense mutations (p.Q79K and p.R208K) were identified in the oncogenic serine/threonine-protein kinase AKT2 involved in cell cycle regulation. We also identified two novel candidate driver genes on the X chromosome: a novel nonsense mutation in the deubiquitinating protease USP9X (p.Q117*) in male patient 194; and 3 novel mutations in MED12, a member of the Mediator complex involved in regulating RNA polymerase II-dependent transcription, were found in male patient 744 (splice site mutation g.chrX:70339329T>C) and female ETP-ALL patients 791 (missense mutation p.R1989H) and 879 (frameshift insertion p.V167fs). Mutations in *USP9X* and *MED12* presented VAFs (mean =0.97, SD =0.04) that were similar to the known driver mutations in *PHF6* and *KDM6A/UTX* (mean =0.97, SD=0.05) (p =1.0000) (S3 Fig and Supplementary information).

4.5.2. The novel U2AF1p.R35L mutation alters pre-mRNA splicing in human T-cells

The novel recurrent p.R35L missense mutation was located in a zinc finger (ZnF) domain of the *U2AF1* (U2 small nuclear RNA auxiliary factor 1) gene, coding for a member of the spliceosome machinery involved in pre-mRNA processing (Figure 2A and S3 Table). Three patients (200, 324 and 791) carried the predicted damaging mutation; patient 200 was unclassified and the other two were classified as early immature T-ALLs (324 and the ETP-ALL 791) and experienced relapse. An additional cohort consisting of 8 adult relapsed T-ALL patients was used for further screening of newly identified somatic driver candidates and revealed a clonal p.R35L mutation (VAF =0.53) in the relapse genome of a 32 years old female case (patient P6, Methods and S4 Table).

The previously identified *U2AF1* p.S34F mutation (Figure 2A), located in the same ZnF in myelodysplastic syndrome (MDS) patients, was shown to disrupt splicing of a number of cancer-relevant genes leading to overall dysregulation of several downstream pathways including epigenetic regulation and DNA damage response [29, 30-32]. To assess the putative impact of the novel p.R35L mutant on alternative splice site utilization, we tested for the presence of quantifiable isoforms among known U2AF1 targets (*BCOR*, *KMT2D/MLL2*, *KDM6A/UTX* and *PICALM*) [32] in human T lymphocyte (Jurkat) cells. *BCOR* and *KMT2D/MLL2* were the only target genes with two identifiable isoforms (data not shown). Using site-directed mutagenesis, we created mutant T-cell lines overexpressing either WT or p.R35L mutated *U2AF1* (Figure 2B) and showed alternative splice site usage at both *BCOR* and *KMT2D/MLL2* loci that were specific to p.R35L (P =0.0286 for both genes, Mann-Whitney U test) (Figure 2C). These data suggest a comparable effect for the p.R35L mutation leading to disrupted splicing and further support a role for functional *U2AF1* mutations in abnormal hematopoiesis, including childhood T-ALL.

4.5.3. Novel X-linked drivers of T-ALL

In 8 T-ALL patients (74, 194, 544, 636, 727, 744, 791 and 879), we identified 4 known somatic mutations (p.I314T, p.R116*, p.R225*, p.Y303*) and 2 novel mutations (p.E221*, p.G226fs) in the well-known X-linked driver gene *PHF6* (Figure 3A). One mature T-ALL (744) also carried a novel nonsense mutation in *KDM6A/UTX* (p.Q692*) which was recently characterized as a T-ALL X-linked driver gene [8, 28] (Figure 3B). We observed no gender bias in the distribution of *PHF6* mutations in our cohort (p =1.0000) with 25% (5/20) of male patients harboring mutations compared to 30% (3/10) of females. This is in line with recent observations [25, 33] that *PHF6* alone is unlikely to account for the higher incidence of T-ALL among males.

Our genomic investigation revealed 2 new candidate X-linked drivers of T-ALL: *USP9X* and *MED12*. For *USP9X*, we identified a novel truncating mutation (p.Q117*) in exon 5 carried by the unclassified male patient 194 suggesting a tumor suppressor role (Figure 3C). As for *MED12*, we identified 3 novel somatic mutations (Figure 3D). One was located at the donor splice site of exon 2 (g.chrX:70339329T>C) in mature male patient 744 (Figure 4A). RNA-seq performed at diagnosis, as well as RT-PCR assays, confirmed that the mutated donor site was skipped leading to an aberrant form of the mature MED12 mRNA where exon 2 was lost (Figure 4B and 4C). This is the first *MED12* mutation shown to lead to exon skipping in cancer. Interestingly, we also identified a missense mutation (p.R1989H) in exon 41 and a frameshift mutation (p.V167fs) in exon 14 of *MED12* in the two female ETP-ALL cases 791 and 879, respectively. While, *MED12* was previously shown to be subject to X-inactivation, it has also been shown that inactivation can be tissue-specific [34]. Through allelic expression analysis, we showed biallelic expression of a germline synonymous SNP (rs5030619) in the female case 879 (S4 Fig), suggesting that *MED12* could indeed escape X-inactivation in T-ALL.

To further test the impact of the loss of function of *MED12* and *USP9X* in driving T-ALL development or maintenance, we performed *in vitro* small hairpin RNA (shRNA) assays in human T lymphocyte (Jurkat) cells (Figure 3C and 3D) and assessed aberrant proliferation and apoptosis resistance. While no significant changes in proliferation were observed (S5 Fig), knockdown of both genes led to a reproducible and significant reduction of apoptosis following Camptothecin (CPT) and at least one relevant chemotherapy drug treatment, Doxorubicine (DOX) or Vincristine (VCR), compared to control (Figure 3C and 3D and S6 Fig) (p =3.375E-06 and 4.114E-05 after CPT, p =4.095E-04 and 4.114e-05 after DOX, p =3.767E-01 and 5.636E-03 after VCR for shUSP9X and shMED12 cells respectively, two-tailed Mann-Whitney U test). Although the observed effect might be context-dependent or cell-type specific, these results provide evidence that reduced activity of these new candidate

X-linked driver genes could perturb normal T-cell development, confer a treatment resistance and point to a potential contribution to the observed gender bias in T-ALL among males [4].

4.6. Discussion

In this study we performed comprehensive genomic characterization of 30 pediatric T-cell ALL patients, including 7 early immature T-ALLs (including 2 ETP-ALLs), 13 mature T-ALLs, and 10 patients for whom the maturation status of T-ALL cells could not be determined. We identified mutations (novel and known) and analyzed expression profiles of common driver genes, further highlighting the heterogeneity and distinct characteristics of T-ALL. We identified recurrent mutations in novel childhood T-ALL genes and functionally characterized 3 new candidate driver genes (*U2AF1*, *MED12* and *USP9X*). Overall, mutation rates in this childhood T-ALL cohort were similar to those previously reported [14, 33, 35-39]. For example, 15 (50%) and 16 (53%) patients harboured *NOTCH1* mutations and *CDKN2A/2B* deletions, respectively. Immature T-ALL patients harboured significantly less *CDKN2A* deletions and experienced more relapse events compared to mature T-ALL cases, as previously published [16]. Our data also suggested a correlation between *CDKN2A* deletions and positive outcome with only one CDKN2A-deleted patient suffering a relapse event against 15 who did not, suggesting potential clinical utility of *CDKN2A* for risk-based stratification of T-ALL patients.

Another recurrently mutated T-ALL gene in our cohort was *PHF6*. Seven patients harbored *PHF6* mutations with a higher mutational rate compared to those previously reported for pediatric T-ALL (26.7% vs. 16-17.1%) [40, 33]. Interestingly, while previous analysis of the gender distribution of X-linked *PHF6* mutations showed higher prevalence among males (32.0% vs. 2.5% of females) [40], we found here a slightly higher proportion of *PHF6* mutated females (30.0% vs. 25% of males). The E3-ubiquitin ligase FBXW7 had a mutation frequency of 13.3% (n =4) in our childhood T-ALL cohort, similar to previous reports (8.6-19.1%) [33, 41, 42, 35-39]. We also observed recurrent mutations in the *JAK3* oncogene (13.3% vs. 5-

15.3%), which is the most frequently targeted gene of the JAK-STAT pathway in T-ALL [5, 33]. Additional known T-ALL driver genes also found to be mutated in our cohort include *KDM6A/UTX* (3.4% vs. 4.5-14%) [33, 28], *JAK1* (3.4% vs. 4.5%) [33], *PTPN2* (3.4% vs. 3.6%) [33] and *LEF1* (6.9% vs. 7.2-17.0%) [33, 43]. T-ALL genes that were mutated less frequently in our cohort, compared to previous studies, include *WT1* (6.7% vs. 13.2-16.2%) [44, 33], *PTEN* (3.4% vs. 10-11.7%) [45, 46, 5, 33] and *DNM2* (6.9% vs. 10.8-18.0%) [5, 33]. And despite previous reports of childhood T-ALL [5, 8, 47] we found no mutation in the polycomb repressive complex 2 (PRC2), *CNOT3*, *RPL10* and *IL7R* in our cohort.

T-ALL is characteristically more common among males. Recent studies have characterized T-ALL drivers in non-autosomal regions of chromosome X such as *PHF6*, *RPL10* and *KDM6A/UTX* [40, 8, 28]. However, *PHF6* and *RPL10* are subject to X inactivation in females [40, 8], and *KDM6A/UTX* has a relatively low frequency (4.5 to 14%) in T-ALL [33, 28], therefore these genes alone cannot explain the observed skewed male:female ratio [48]. We identified additional novel X-linked candidate driver mutations in *USP9X* and *MED12* that were shown to co-occur with *PHF6* mutations in our childhood T-ALL cohort. Both genes have previously been implicated in diverse cancer types but have never been associated with T-ALL. USP9X belongs to the Ub-specific protease family and targets multiple proteins including SMURF1 [49], MCL1 [50] and Smad4 [51] and has been shown to escape X-inactivation [52]. USP9X has been demonstrated to positively regulate T-cell receptor signaling and to be required for T-cell function [53]. The oncogenic driving potential of *USP9X* was confirmed in several tumor types: loss of function of the gene in chronic myelogenous leukemia [54], hepatocellular [55] or colorectal carcinoma [56], bladder cancer [57] and B-cell ALL [58] leads to increased sensitivity to chemotherapy and to apoptosis. On the other hand, tumor suppressor functions have also been described in pancreatic adenocarcinoma [59, 60] and gingivo-buccal oral squamous cell carcinoma [61], highlighting the context-dependent role of USP9X in oncogenesis. The p.Q117* truncating mutation identified here in a single male patient (194) suggests a tumor suppressor role for *USP9X* in T-ALL, and the strong decrease in apoptosis observed following knockdown of the gene and chemotherapy drug treatment further corroborates the anti-oncogenic role of *USP9X* in childhood T-ALL and suggests a possible involvement on treatment resistance. As for *MED12*, it was altered in 10% of our T-ALL patients and was as frequently mutated as the known T-ALL gene *NRAS*. The loss-of-function mutations in *MED12* (p.V167fs and splice site g.chrX:70339329T>C) were both shown to be expressed, supporting a functional role of these mutations in T-ALL. Somatic mutations in *MED12* exon 2 splice sites have previously been identified in breast cancer and were shown to cause intron retention [62]. MED12, along with MED13, Cyclin C, and CDK8 or CDK19, is a member of the kinase module of Mediator, a multisubunit complex required for regulation of RNA polymerase II-dependent transcription [63]. In line with our observations, *MED12* was recently characterized as a tumor suppressor in uterine leiomyomas, prostate cancer, chronic lymphocytic leukemia and breast fibroadenoma [64-68]. Mutations in exon 2 showed a particularly high rate of somatic alterations (70% of uterine leiomyomas) [64] and have been shown to disrupt the direct interaction of MED12 with the cyclin C-CDK8 leading to reduced Mediator activity [66]. Here, *MED12* appears to escape X-inactivation. In addition we showed that loss of *MED12* in human T-cells leads to decreased apoptosis levels provoked by chemotherapy drugs, further corroborating its role as tumor suppressor in childhood T-ALL and highlighting its putative relapse driving potential in leukemogenesis. This is in line with recent results showing that *MED12* repression induces resistance to multiple cancer drugs through TGF-βR signaling regulation [69].

USP9X and *MED12* mutations were detected in the predominant clone at diagnosis and the two patients for whom we had relapse sequencing data available showed the presence of mutations in the major clone at relapse. These results indicate an early clonal selection supporting a possible functional role in early phases of T-ALL development and a possible implication in relapse. While further investigations are required to fully decipher the mechanisms underlying the observed *USP9X* and *MED12* induced anti-apoptotic effects, the identification of two novel X-linked tumor suppressor genes in pediatric T-ALL and their co-occurrence with known X-linked driver gene mutations (*PHF6*) suggests cooperating effects of X-linked mutations in T-ALL onset. The presence of several T-ALL tumor suppressor genes on the X chromosome, most of which escape X-inactivation, substantially increases the odds for males to develop the disease and could therefore explain the higher incidence of T-ALL among male children, provided that mutated forms of these proteins do not create dominant effects. However the underlying disease mechanisms associated with these alleles remains to be determined.

We also identified a novel recurrent T-ALL mutation (p.R35L) in *U2AF1* in 3 patients, including 2 immature cases. And through screening of additional adult relapsed T-ALL patients we confirmed recurrence of this novel mutation even in adults. *U2AF1* is a common mutational target in several cancer types, such as MDS where 11% of cases are mutated [70, 71]. The importance of the spliceosome machinery in leukemogenesis has been demonstrated [70, 72] and mutations in *U2AF1* were recently shown to predict poor prognosis in patients with *de novo*acute myeloid leukemia [73]. The identified p.R35L mutation was previously reported in 2 cases of myeloid neoplasms [74, 29], but to the best of our knowledge, *U2AF1* has never been associated with T-ALL prior to this study. U2AF1 has a U2AF homology motif allowing the heterodimerization with U2AF2 [75], an arginine-serine (RS) domain required for RNA high-affinity binding [76], and two ZnF domains. The p.R35L mutation identified here, as well as recurrent MDS mutations at residues S34 and Q157, fall within the ZnF domains. Although the role of these domains remains elusive, our functional studies support a function in the splicing mechanism with the p.R35L mutation leading to alternative splice site usage in known target genes such as the transcriptional corepressor BCOR, associated with poor prognosis in MDS when altered [77], or the H3K4 methyltransferase KMT2D/MLL2 that have been show to play a role in hematopoiesis [78]. KMT2C/MLL3 and KMT2D/MLL2 methyltransferases along with the histone demethylase KDM6A/UTX are members of the activating MLL2-KDM6A/UTX complex [79], their deregulation due to cooperating mutations or aberrant splicing events, could provide an alternative mechanism to recurrent alterations of PRC2 members EZH2, SUZ12 or EED in T-ALL. Mutations in *U2AF1* typically occur early in the founding clone and are present in the major clone at diagnosis [80-82], however p.R35L was subclonal in our 3 patients. This *U2AF1* mutation would appear therefore to be a late, secondary event in the developmental history of these T-ALL cases. The fact that this mutation is subclonal could explain its absence in other T-ALL studies, particularly in the event of low tumor purity or extensive intra-tumor heterogeneity. It should be noted that in ETP-ALL patient 791, p.R35L was present in a very minor subclone (VAF =0.09) and was lost at relapse, which in this case, could render its functional role questionable. However, in addition to U2AF1 p.R35L, patient 791 also harbored a frameshift insertion p.Y816fs in the methyltransferase KMT2C/MLL3, that was carried by another subclone at diagnosis (VAF =0.28) and presented a positive shift of frequency at relapse (VAF=0.39). KMT2C/MLL3 and KMT2D/MLL2 methyltransferases, along with the histone demethylase KDM6A/UTX, are members of the activating MLL2-KDM6A/UTX complex involved in promoting chromatin remodeling [79]; cooperating mutations and aberrant splicing events leading to altered histone modification could contribute to disease pathogenesis.

Additional epigenetic regulators that were mutated in this childhood T-ALL cohort include: EHMT1, WHSC1, KTM2C/MLL3, CTCF, CREBBP and KDM6A/UTX. *EHMT1* codes for a H3K9 methyltransferase and is a member of the E2F6 repressor complex. To the best of our knowledge this gene has never been associated with T-ALL, although previous reports demonstrated overexpression of *EHMT1* associated with poor prognosis in esophageal cancer and treatment resistance in chronic myeloid leukemia [83, 84]. We did not identify loss-of-function mutations in known T-ALL genes *EZH2*, *SUZ12* or *EED*, but we identified activating somatic events in the H3K27 methyltransferase WHSC1 (MMSET, NSD2). *WHSC1* is a well-known and recurrent target for mutation in pediatric B-ALL as well as adult T-ALL [85, 86], however somatic disruption of *WHSC1* in pediatric T-ALL is less frequent [86]. *WHSC1* was mutated at higher frequency in our cohort, compared to previous reports in adult T-ALL (10.3% vs. 4.9%) [86]. The gain of copy identified in the ETP-ALL case 791 as well as the recurrent p.E1099K mutation found in the mature case 340 lead to enhanced activation of WHSC1 which correlates with increased H3K36 and decreased H3K27 methylation and an open chromatin state across the genome. Activating *WHSC1* mutations mimic the described PCR2 loss-of-function mutations in ALL and could alter normal lymphoid differentiation and cell survival and support an oncogenic role for *WHSC1* in childhood T-ALL [87]. However the p.S231* loss of function mutation identified in the early immature patient 432 is difficult to interpret. Haploinsufficiency of *WHSC1* accounts for the core phenotypes of Wolf-Hirschhorn syndrome including facial appearance, mental retardation, growth delay and seizures [88]. This stop codon mutation could simply be a passenger event in 432 or provide a specific advantage given this patient's genetic background. The KMT2C/MLL3 frameshift mutation (p.Y816fs) in patient 791 had never been identified in hematological malignancies before. *KMT2C/MLL3* haploinsufficiency was shown to impair differentiation of hematopoietic stem and progenitor cells and to provoke resistance to conventional chemotherapy [89]. This resistance might explain the emergence of subclonal *KMT2C/MLL3* p.Y816fs positive cells in patient 791 at relapse. Finally, as recently reported in T-ALL [33], we identified the loss of a complete copy of the transcriptional repressor CTCF in one ETP-ALL patient, as well as the histone and non-histone acetyltransferase CREBBP in one mature case (636). *CTCF* was recently demonstrated to be a tumor suppressor [90] and its haploinsufficiency to lead to an increased variability in CpG methylation genome-wide. Tumors with hemizygous loss of *CTCF* showed increased aggressiveness, as observed here for ETP-ALL patient 791. The tumor suppressor gene *CREBBP* is a frequent mutational target in hematological malignancies and mutations in this gene are associated with increased risk of relapse in ALL [91], though patient 636 did not suffer relapse. Overall, these results support an important role for chromatin modification in T-ALL [92, 93, 5].

In conclusion, through integrated whole-exome, transcriptome, as well as targeted re-sequencing and genotyping investigation of 30 childhood T-ALL patients, we showed that each patient carried a unique combination of known and novel somatic alterations, including SNVs, indels, CNVs, and chromosomal rearrangements. We observed a number of uncommon and novel mutations in the early immature cases of our cohort, particularly in the two ETP-ALL patients who harbored 80% of newly-identified candidate driver mutations. We characterized a recurrent mutation in the spliceosome member U2AF1 and demonstrated its impact on alternative splicing of cancer-relevant genes, further suggesting the importance of aberrant splicing in leukemogenesis. We also identified *MED12* and *USP9X* as putative new X-linked drivers and provided evidence of the functional impact of their loss in T-cells supporting a potential role for these genes in the male-biased sex ratio observed in T-ALL. These results further highlight the underlying complexity of the genomic landscape of T-ALL, and the pressing need of larger integrative studies in well-defined cohorts to increase understanding of the biological mechanisms that contribute to T-ALL and its various subtypes.

4.7. Materials and Methods

4.7.1. Study subjects

All study subjects were French-Canadians of European descent. Incident cases were diagnosed in the Division of Hematology-Oncology at the Sainte-Justine Hospital (Montreal, Canada) as part of the Quebec childhood ALL cohort (QcALL) [94]. This childhood T-ALL cohort (n=30) consisted of 20 males and 10 females, with a mean age at diagnosis of 11.8 years. All were classified as high-risk patients and treated accordingly under FRALLE and DFCI protocols depending on year of diagnosis (Table 1). Eight patients experienced relapse after a median time of 21 months post-induction, of which five patients did not survive post-relapse, and one case (759) was refractory to induction chemotherapy and died of cerebral hemorrhage at 6 months after diagnosis.

4.7.2. Whole-exome sequencingand variant identification

Whole exome sequencing (WES) was performed on 24 matched normal-tumor T-ALL patients (Figure 1). DNA was extracted from bone marrow samples (at diagnosis) and peripheral blood samples (after remission) (Table 1) using standard protocols [95]. Whole exomes were captured in solution with Agilent’s SureSelect Human All Exon (38Mb or 50Mb), Nextera Rapid Capture Exome Enrichment kit (case 791) or SureSelectXT Clinical Research Exome (case 879) kits according to the manufacturer’s protocol and sequenced on the Life Technologies SOLiD 4/5500 System (paired-end: 50x35 bp, mean coverage on targeted region =35X) and for cases 791 and 879 (paired-end: 100x100 bp, mean coverage on targeted region =120X). Reads obtained from SOLiD 4/5500 and HiSeq 2500 systems were aligned to the Hg19 reference genome using LifeScope Genomic Analysis Software and Bowtie2 (version 2.2.3) [96] respectively. PCR duplicates were removed using Picard [97]. Genotype quality score recalibration was performed using the Genome Analysis ToolKit

(GATK) [98]. Sequencing metrics were obtained using the DepthOfCoverage option in GATK. After filtering out low quality reads, pileup files were created using SAMtools [99]. Somatic single nucleotide variants (SNVs) and small indels were called from pileup files using SNooPer, a highly versatile machine learning approach that uses Random Forest classification models and integrates matched normal-tumor data to accurately call somatic variants in low-depth sequencing data (Spinella *et al*. in revision, software available upon request).

4.7.3. Targeted sequencing

Ultra-deep targeted re-sequencing was performed on all candidate somatic driver mutations using the Illumina TruSeq Custom Amplicon assay as per the manufacturer's instructions (Figure 1). Illumina DesignStudio was used to design custom oligos targeting the select mutated regions (available upon request). Genomic DNA from bone marrow at diagnosis and from blood at remission was used to validate somatic hits identified from WES (24 cases) or to screen identified variant positions in 3 additional cases with insufficient DNA for WES. PCR purification was performed with Ampure Beads for 150bp amplicon size selection. Double stranded amplicons were pooled, quantified by qPCR and sequenced on the Illumina HiSeq2500 device (paired-end: 2x100bp) to reach a mean coverage of 2,500X. Reads were aligned to the Hg19 reference genome using Bowtie2 (version 2.2.3) [96]. Cleaned BAM files were used to create pileup files using SAMtools [99]. A modified version of SNooPer was used to screen mutations at the targeted positions directly in the pileup file, and to compare normal and tumoral information to confirm the somatic origin of the validated mutations (details available upon request).

Because *NOTCH1* mutations were difficult to identify in the exome sequencing data due to insufficient local coverage, Sanger sequencing was performed targeting recurrently mutated regions of *NOTCH1* in T-ALL [13]. These included the N terminal region of the heterodimerization (HD) domain on exon 26; the C terminal region of the HD domain on exon 27; the proline, glutamic acid, serine, threonine-rich (PEST) domain; and the C terminal region of the transcriptional activation domain (TAD) on exon 34. Primers used are listed in S5 Table. Chromatograms were analysed using the Sequencher software (Gene Codes) (S7 Fig).

An additional cohort consisting of 8 adult relapsed T-ALL patients from the Princess Margaret Cancer Centre, University Health Network (Toronto, Canada) (S4 Table) was used for further screening of newly identified somatic driver candidates in *USP9X*, *MED12* and *U2AF1*. Ultra-deep targeted sequencing was performed on the 5 somatic mutations identified in these 3 genes on Illumina MiSeq system (McGill University and Génome Québec Innovation Centre); primers used are available upon request. The analysis of sequencing data was performed as described above.

4.7.4. RNA-sequencing and variant identification

RNA-sequencing (RNA-seq) was performed on 11 T-ALL patients (Figure 1) with suitable RNA quantity and quality. Total RNA was extracted from bone marrow samples at diagnosis for patients 432, 437, 547, 693, 716, 743, 744, 748, 791 and 849 using the mirVana Isolation kit (Ambion) according to the manufacturer's protocol. The Allprep DNA/RNA Mini kit (Qiagen) was used for relapse samples in patients 791 and 879. For patient 744, mature RNA was also purified using the Ambion's MicroPoly(A)Purist kit (Small Scale mRNA Purification Kit P/N AM1919). Following a DNAse I treatment, total or mature RNA samples were quantified by NanoDrop ND1000 (Thermo-Fisher Scientific) and RNA quality was assessed using the

Agilent 2100 Bioanalyzer (Agilent). Ribosomal ribonucleic acid (rRNA) were depleted using the Invitrogen RiboMinus Eukaryote kit (Life Technologies). cDNA libraries were prepared using the SOLiD Total RNA-seq kit (diagnosis samples) and the Illumina TruSeq Stranded Total RNA kit (relapse samples) based on manufacturer's protocol and sequenced on the Life Technologies SOLiD 4/5500 System (paired-end: 50x35 bp) or the Illumina HiSeq 2500 System (paired-end: 100x100 bp). Reads obtained from SOLiD 4/5500 and HiSeq 2500 systems were aligned to the Hg19 reference genome using LifeScope Genomic Analysis Software (Whole Transcriptome Analysis pipeline, default parameters) and STAR aligner (version 2.5) [100] respectively. Remaining ribosomal sequences were filtered out. Recalibration of the genotype quality scores was performed using the Genome Analysis ToolKit (GATK) [98]. Cleaned BAM files were used to create pileup files using SAMtools [99]. SNVs and small indels identified from WES were screened in RNA-seq pileup files using a modified version of SNooPer. Reads Per Kilobase per Million mapped reads (RPKM) were calculated for 22,292 genes using the R bioconductor package edgeR [101]. Heatmaps were constructed using the heatmap.2 library of the gplots R package using RPKM values of genes of interest. Breakdancer (version 1.1.2) [102] with a minimum mapping quality (q) set to 30 was used to confirm rearrangements identified by cytogenetics or to call new events from RNA-seq data.

4.7.5. Whole genome high-density SNV genotyping and copy number variant identification

Whole genome genotyping (WGG) data were available for 23 T-ALL patients (Figure 1). Normal and tumor samples from these patients were genotyped using Illumina's HumanOmni 2.5-Quad or HumanOmni2.5-Octo SNP bead arrays (McGill University and Genome Quebec Innovation Centre, Montreal, Quebec). Extracted genomic DNA was processed according to the Illumina Infinium HD Assay Ultra protocol. BeadChips were imaged on Illumina's iScan

System with iScan Control Software (v3.2.45). The Genotyping Module (Version 1.9.4) of the Illumina GenomeStudio software (V2011.1) was used for raw data normalization, genotype clustering and calling, with default. ASCAT version 2.2 [103] was used to evaluate the sample purity, to evaluate tumor ploidy and to identify tumor-specific copy number variants (CNVs) or copy-neutral loss of heterozygosity (LOH).

CDKN2A gene allelic status was further evaluated in all T-ALL patients at diagnosis using PCR. Each 25 µl reaction contained 50 ng of template DNA, 1X KOD Buffer, 1.5 mM MgSO4, 200 µM dNTPs, 0.3 µM of each primer (listed in S5 Table) and 0.5U of KOD Hot Start DNA Polymerase (Millipore). Cycling parameters used were: 95°C 2 min; 40 cycles (95°C 20 sec, 58°C 10 sec, 70°C 10 sec). Amplified fragments of 368 bp were visualized using standard gel electrophoresis. Electrophoresis gels are available upon request.

4.7.6. Variant annotation and prioritization of cancer driver gene mutations

ANNOVAR (version 2015Jun17) [104] and Oncotator (version 1.8) [105] were used to annotate somatic splice site variants, non-synonymous SNVs and frameshift small indels. Variants were queried against publically available datasets such as 1000 Genomes [106], NHLBI GO Exome Sequencing Project (ESP) [107] and Exome Aggregation Consortium (ExAC) [108] to filter out common polymorphisms (minor allele frequency >0.01). The Catalogue of Somatic Mutations in Cancer (COSMIC, version 72) [109] was used to evaluate prior implications in cancer. We classified each mutated gene as either tumor suppressor genes (TSGs) or oncogenes based on Vogelstein's 20/20 rule [110] on COSMIC v72 data (S3 Table). The predicted functional impact of non-synonymous variants and small indels was assessed using Sift (version 1.03) [111], Polyphen2 (version 2.2.2) [112], MutationTaster 2 [113] and the Cancer-specific High-throughput Annotation of Somatic Mutations tool (CHASM version 3.0) [114] (S3 Table). Variants presenting a score ≤0.05 were considered as damaging by Sift (D). Mutations with a Polyphen2 score between 0.447 and 0.909 were predicted as possibly damaging (P) while a score >0.909 was considered as damaging. Mutations with a MutationTaster score >0.9 were also considered as damaging. Finally, for CHASM classification we used the Blood-Lymphocyte training set and a Benjamini and Hochberg's adjusted false discovery rate (FDR) ≤0.20, to prioritize mutations based on their predicted driver potential. To identify candidate driver mutations we filtered events and kept only somatic alterations that were: i) missense mutations predicted to be driver by CHASM [114]; or to be damaging by at least two of the other three prediction algorithms; or identified as recurrent in COSMIC v72 [109]; ii) splice site and nonsense mutations that were predicted to be damaging by MutationTaster and Sift, respectively; or located in a tumor suppressor gene; or identified as recurrent in COSMIC; iii) frameshift indels and deletions located in a TSG and gain of copies located in an oncogene; iv) genes with more than one mutation (SNV, indel or CNV) in our cohort but previously never associated with pediatric T-ALL were considered as new candidate driver genes. Based on these filtering criteria, we identified somatic gene mutations with putative functional effects in driving T-ALL development, some of which we further investigated (*USP9X, MED12* and *U2AF1*).

4.7.7. Cell lines

Human acute T-cell leukemia-derived Jurkat Tet-On cells (630915, Clontech) and REH human leukemia cells (#CRL-8286, A.T.C.C.) were grown in RPMI-1640 medium (Wisent) supplemented with 10% fetal bovine serum, 100 IU/ml penicillin and 100 µg/ml streptomycin (Wisent). The highly transfectable HEK293T cells were grown in Dulbecco's Modified Eagle's Medium supplemented with 10% fetal bovine serum (Wisent). Cells were routinely maintained at 37°C in a humidified atmosphere composed of 95% air and 5% CO2 and provided with fresh medium every 2 to 3 days.

4.7.8. RT-PCR validation of alternative splicing in MED12

Total RNA was extracted (as above) from the patient's bone marrow at diagnosis, as well as from REH (human B leukemia) cells, mature T cells (CD3+/CD19-) isolated from cord blood samples, and from two g.chrX:70339329TT wild type (WT) patients (727 and 791), used as controls. RNA was reverse transcribed into cDNA using the Ovation® qPCR System (NuGEN Technologies). PCR were performed using KOD Polymerase as described above. Amplified fragments were analyzed on the Agilent 2100 Bioanalyzer Instrument and by Sanger sequencing (McGill University and Genome Quebec Innovation Centre).

4.7.9. ShRNA-mediated gene knockdown

Lentivirus-mediated gene-specific small hairpin RNAs (shRNAs) were used to knockdown expression of 2 candidate driver genes: *USP9X*, and *MED12*, in Jurkat (human T leukemia) cells. The complete list of MISSION® shRNAs (Sigma-Aldrich) used to silence target gene expression are listed in S6 Table (at least 3 for each gene). shRNA target sequences were subcloned into either the lentiviral vector pLKO.1-puro (TRC1 version) or pLKO.5-puro (TRC2 version). Briefly, using Polyethylenimine (PEI; Polysciences), plasmids and packaging vectors (6 µg of pRSV-Rev, 7.8 µg of plasmid pMD2.VSVG, 15 µg of pMDL, and 9 µg of shRNA in pLKO-puro plasmid) were co-transfected into HEK293T cells to generate respective lentivirus. Supernatants containing lentiviruses were harvested 48h post-transfection. 1x10E6 Jurkat cells were infected with 1 ml of supernatant in the presence of 5 µg/mL of polybrene (Sigma-Aldrich). 72 hours post-infection, cells were screened with 2 µg/µl puromycin (Sigma-Aldrich) for two weeks to select for shRNA-knockdown cells. Three biologically independent replicates were carried out for each target gene. The expression of target genes was measured by quantitative PCR (qPCR).

Total RNA was extracted from infected cells using RNeasy mini kit (Qiagen). 500 ng of total RNA were reverse transcribed using the M-MLV reverse transcriptase (Thermo Fisher Scientific) and qPCR amplifications (triplicates) were performed on the ABI PRISM 7000 Sequence Detection System (Thermo Fisher Scientific) in a total volume of 25 µl as follows: 5 µl of cDNA (diluted 1:5), 0.2 µM of each primers (listed in S5 Table), and 1X SYBR Green PCR Master Mix (Thermo Fisher Scientific). The cycling parameters were: 95°C 10 min; 40 cycles [95°C 15 sec, 62°C 1 min], followed by a denaturation curve at 60°C. GAPDH was used as reference gene. Expression values were calculated by the $2^{-(\Delta\Delta Ct)}$ formula previously described [115].

4.7.10. Apoptosis assay

Apoptosis was measured using the Alexa Fluor® 488 Annexin V/Dead Cell Apoptosis Kit (Thermo Fisher Scientific) according to the manufacturer’s instructions. Briefly, cells were seeded at 5x10E5 cells/mL in their culture medium and treated with 2 µM of Camptothecin for 17 hours or 500 nM of Doxorubicine for 18 hours or 50 nM of Vincristine for 24 hours, to promote DNA damage-induced apoptosis. Following incubation with Annexin V Alexa Fluor® and propidium iodide (PI) at room temperature for 30 min, stained cells were immediately analyzed by flow cytometry. The percentage of apoptotic cells was measured on the FACS Fortessa using the BD FACSDiva software (BD Biosciences) according to manufacturer’s guidelines. At least three independent experiments were carried out for each biological replicate.

4.7.11. U2AF1 p.R35L splicing assay

The pOTB7 plasmid containing *U2AF1* cDNA was purchased from Harvard PlasmID Repository (clone # HsCD00321863) and subcloned into the Gateway compatible vector pDONR221. The R35L mutation was introduced into the cDNA sequence of *U2AF1* using the QuickChange II XL Site Directed Mutagenesis kit (Agilent) with primers listed in S5 Table. Sanger sequencing (McGill University and Genome Quebec Innovation Centre) was performed to confirm the presence of the mutation. WT or R35L *U2AF1* coding sequences were then subcloned into the Gateway lentiviral vector pLenti CMV Puro DEST (w118-1) using LR clonase Enzyme mix (Thermo Fisher Scientific). Lentiviruses were generated and Jurkat cells infected (two biologically independent replicates) as described above. Total RNA was extracted from infected cells using RNeasy Mini Kit and treated with RNase-Free DNase Set (Qiagen), and cDNAs were generated using M-MLV reverse transcriptase (Thermo Fisher Scientific). Overexpression of WT or R35L *U2AF1* in Jurkat cells was measured by qPCR amplification as described above. To validate alternative splice site utilization at the *BCOR* and *KMT2D/MLL2*target genes, as described elsewhere [32], RT-PCR was performed on cDNAs in duplicates using KOD Hot Start Polymerase as described previously with primers listed in S5 Table. PCR products were electrophoresed in agarose gel stained with SYBR Safe (Thermo Fisher Scientific), and quantified by densitometry using Image J software (version 1.49).

4.7.12. Statistical tests

Significance of observations was assessed with R using two-tailed Fisher's exact test or Mann-Whitney-U test when appropriate.

4.8. Supplementary information

4.8.1. Classification of T-ALL patients by maturation stage

We used both immunophenotyping and gene expression data when available to classify patients according to T-cell maturation status (Table 1, S1 Fig, S1 Table and S2 Table). Seven patients (324, 432, 706, 716, 748, 791 and 879) clustered as early immature T-ALLs. Five of these tumors strongly expressed CD34 and most had no or weak expression of mature thymocyte markers CD1a, CD4 and CD8. These CD34+/CD1a-/CD4-/CD8- patients also expressed immature T-cell specific markers such as *LMO2, LYL1, BCL2, FLT3, TGFB1* and *MYB* [1]. Two of these patients (791 and 879), had low expression levels of CD5 (<75%) and were positive for myeloid or stem-cell markers CD34, HLA-DR, CD13, CD33, CD11b and positive for the T-cell marker cytoplasmic CD3 (cCD3). They expressed *LYL1*, indicative of an early T-cell precursor ALL (ETP-ALL) phenotype [2]. Twelve patients (340, 341, 437, 544, 547, 636, 693, 727, 743, 744, 759 and 849) positive for CD1a/CD4/CD8 and expressing mature thymocyte markers such as *LCK, RAG1, PTCRA* or *STAT5A* [1] were classified as mature T-ALLs (S1 Fig and S1 Table). Among these, patients 744 and 849 showed the strongest ectopic expression of the homeobox developmental genes *SIX6* and *NKX3*-1 [1], and patient 849 showed strong activation of the TAL1 oncogene, further supporting a more advanced (late cortical) stage of maturation. Of note, CD1a+/CD4+ patient 693 highly expressed *HOXA* genes indicative of T-lineage blockade prior to beta-selection [1], suggesting an early cortical stage of maturation for this mature T-ALL. Some patients were difficult to classify, such as patient 744 who was classified as mature T-ALL and was the only patient who showed activation of *TLX3*, typically associated with early cortical cases. However patient 744 was CD34+ and also expressed immature markers such as *FLT3* and *BMI1*.

4.8.2. Chromosomal rearrangements and cryptic events in T-ALL

We identified translocations using Fluorescent In Situ Hybridization (FISH), molecular cytogenetic analyses and RNA-seq data. RNA-seq data confirmed aberrant transcriptional activation of *TAL1* for patient 849 but surprisingly showed no up-regulation of this gene in patient 547 (S1A Fig). This is possibly due to intra-tumor heterogeneity of the sample leading to reduced tumor-specific gene expression signals. We could not confirm upregulation of *LMO2* and *TLX1/HOX11* in patients 759 and 636 because of the lack of expression data. Transcriptome analysis of the two ETP-ALL cases with the t(10;11)(p12:q14) CALM-AF10 translocation revealed aberrant activation of the *HOXA* cluster [3], as well as activation of its cofactors BMI1 and MEIS1 (S1A Fig and S2 Table). Of note, *BMI1* activation leads to *CDKN2A* inhibition and thus induces cell proliferation, which corroborates with the absence of *CDKN2A* locus deletions in these immature T-ALL patients [4]. Only one patient (744) showed cryptic activation of *TLX3*, typically associated with early cortical cases. This patient also had upregulation of *BMI1*, but no characteristic co-upregulation of *HOXA* genes was observed, nor did we identify CALM-AF10 or MLL rearrangements typically causing *BMI1* activation [3]. Patient 693 presenting the overall strongest activation of the *HOXA* gene cluster also showed upregulation of oncogenes and T-cell differentiation genes such as *NOTCH2, NOTCH3, PTCRA* and *PIM1* as expected. However no expression of *LMO2, LYL1, BCL2* or *FLT3* was observed in this patient. Of note, *TLX1/HOX11* activation, associated with early cortical T-ALLs and with a favorable prognosis [4], was not observed in any of the patients investigated here.

4.8.3. Clonal architecture of childhood T-ALL

We used genomic sequencing data (read counts) to study clonal architecture of these childhood T-ALL tumors. For both ETP-ALL patients, we also investigated clonal dynamics from diagnosis (pre-treatment) to relapse. Variant allele frequencies (VAFs) of 48 somatic SNVs and small indels from 66 candidate drivers were estimated from Illumina WES (Methods, mean coverage on targeted region =120X) and/or ultra-deep targeted re-sequencing data (Methods, mean coverage =2,500X). Variants identified from RNA-seq data and Sanger sequencing were not considered. Tumor purity was determined either from blast counts at diagnosis or from ASCAT profiles (Table 1), and was used to adjust VAFs in each tumor (S3 Fig and S3 Table). Frequencies were not adjusted for copy number variations, which would explain the high somatic variant frequencies at certain loci that overlapped monoallelic deletion events (e.g. *WT1* p.R370H (VAF =0.96)). Once adjusted, 29.2% of identified somatic SNVs and small indels were considered as subclonal as they presented adjusted VAFs ≤0.4, and of VAFs ≤0.8 for X-linked mutations in males (S3A Fig and S3 Table). Variations in common T-ALL drivers (*PHF6, FBXW7, JAK1* and *JAK3*) were mostly clonal (mean VAF =0.48, standard dev. =0.10). Interestingly, both variations in *JAK1* and *JAK3* identified in patient 744 were clonal (VAF =0.59 and 0.56, respectively), suggesting an early cooperative role as opposed to previous reports of a sequential event with *JAK1* acting as first hit and *JAK3* contributing to selection and expansion of the *JAK1*+ subclone [5]. RAS pathway mutations had significantly lower frequencies compared to these common drivers with a mean VAF =0.33 (standard dev. =0.11) (p =0.006, Mann-Whitney-U test). Subclonality of these mutations corroborates previous reports [6-8] that describe a secondary role for Ras mutations in T-ALL occurring later in tumor progression. SNVs and small indels identified in novel T-ALL genes had different modalities of clonal evolution (S3A Fig and S3 Table). For *U2AF1*, p.R35L was subclonal in all three patients carrying the mutation (mean VAF =0.24, standard dev. =0.13), and was therefore likely recently acquired in tumor evolution. On the other hand, mutations in newly identified X chromosome genes *USP9X* and *MED12* (mean =0.97, SD =0.04) in patients 194, 744, 791 and 879 had VAFs that were similar to the known driver mutations in *PHF6* and *KDM6A/UTX* (mean =0.97, SD=0.05) (p =1.0000). Given that we had quality relapse data for the two ETP-ALL patients (791 and 879), we investigated the mutational trajectories of known and novel candidate driver genes (S3B and S3C Fig and S3 Table). The CALM-AF10 translocation carried by both patients at diagnosis was also found in the dominant clone at relapse. In addition to the dominant *PHF6* (p.G226fs, VAF =0.51) and *MED12* (p.R1989H, VAF =0.49) mutations, case 791 also carried subclonal mutations in *KMT2C/MLL3* (p.Y816fs, VAF =0.28) and *JAK3* (p.M511I, VAF =0.13). While patient 791's most subclonal mutations (*NOTCH1*, VAF =0.10; JAK3, VAF =0.13; *U2AF1*, VAF =0.09) were lost at relapse, the subclonal tumor cell population carrying the *KMT2C/MLL3* mutation at diagnosis, was selected at relapse with a positive VAF shift from 0.28 to 0.39. On the other hand, X-linked mutations remained clonal with relapse VAFs of 0.46 and 0.44 for *PHF6* (p.G226fs) and *MED12* (p.R1989H) respectively. This suggested the presence of multiple subclones at diagnosis and subsequent purification at relapse with the emergence of a single founder clone. As for patient 879, almost all relapse mutations were present in the major clone at diagnosis and remained clonal at relapse, including *PHF6* (p.R225*, VAF =0.55 to 0.49), *MED12* (p.V167fs, VAF =0.55 to 0.49), *JAK3* (p.L857P, VAF =0.54 to 0.53) and *FBXW7* (p.W425C, VAF=0.55 to 0.45). Only *WT1* (p.V167fs) was counter-selected at relapse and presented a negative VAF shift from 0.49 to 0.29. Overall, subclones present in both patients at diagnosis underwent a putative selection illustrated by the loss of subclones in 791 with the emergence of a fitter dominant clone at relapse carrying *KMT2C* p.Y816fs, and loss of a *WT1* p.V167fs subclone in 879. Interestingly, no additional relapse-specific events were identified in these cases. Given that *MED12* positive cells showed clonal equilibrium, it was difficult to evaluate their relapse potential, however their maintenance as dominant population in both cases suggests potential involvement in the process.

4.9. Figures

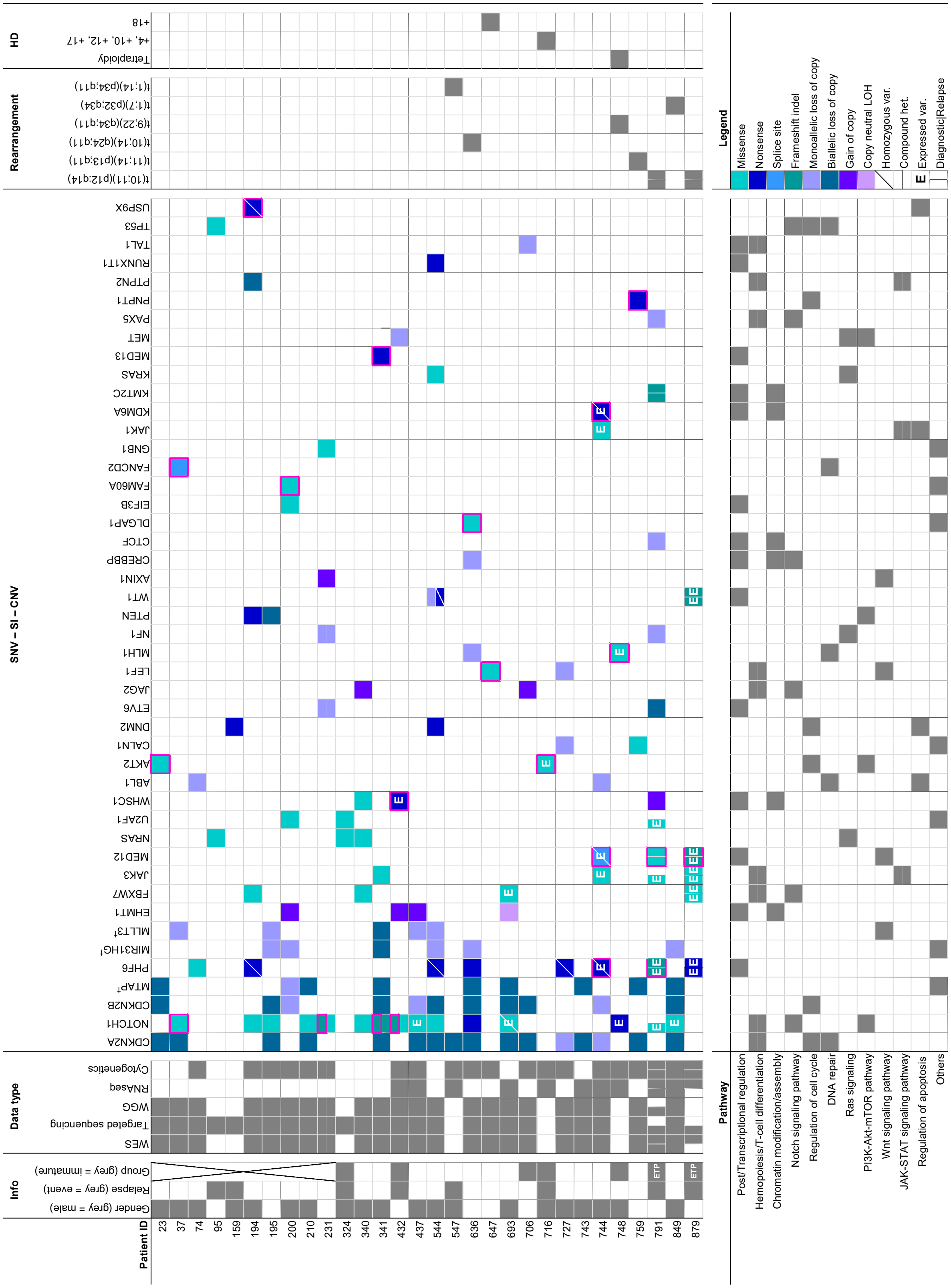

HD
+18
+4, +10, +12, +17
Tetraploidy
Rearrangement
t(1;14)(p34;q11)
t(1;7)(p32;q34)
t(9;22)(q34;q11)
t(10;14)(q24;q11)
t(11;14)(p13;q11)
t(10;11)(p12;q14)
SNV – SI – CNV
USP9X
TP53
TAL1
RUNX1T1
PTPN2
PNPT1
PAX5
MET
MED13
KRAS
KMT2C
KDM6A
JAK1
GNB1
FANCD2
FAM60A
EIF3B
DLGAP1
CTCF
CREBBP
AXIN1
WT1
PTEN
NF1
MLH1
LEF1
JAG2
ETV6
DNM2
CALN1
AKT2
ABL1
WHSC1
U2AF1
NRAS
MED12
JAK3
FBXW7
EHMT1
MLLT3†
MIR31HG†
PHF6
MTAP†
CDKN2B
NOTCH1
CDKN2A
Data type
Cytogenetics
RNAseq
WGS
Targeted sequencing
WES
Info
Group (grey = immature)
Relapse (grey = event)
Gender (grey = male)
Patient ID
23
37
74
95
159
194
195
200
210
231
324
340
341
432
437
544
547
636
647
693
706
716
727
743
744
748
759
791
849
879
E
ETP
Legend
Missense
Nonsense
Splice site
Frameshift indel
Monoallelic loss of copy
Biallelic loss of copy
Gain of copy
Copy neutral LOH
Homozygous var.
Compound het.
E Expressed var.
Diagnostic|Relapse
Pathway
Post/Transcriptional regulation
Hemopoiesis/T-cell differentiation
Chromatin modification/assembly
Notch signaling pathway
Regulation of cell cycle
DNA repair
Ras signaling
PI3K-Akt-mTOR pathway
Wnt signaling pathway
JAK-STAT signaling pathway
Regulation of apoptosis
Others

Figure 1. Overview of the SNVs, small indels and structural variations identified among 30 childhood T-ALL patients. Genes (top) are ordered according to the number of events identified in the cohort (descending order from left to right). Pathways (bottom) are ordered according to the number of genes identified as mutated (descending order from top to bottom). The "Info" and "Data type" sections respectively inform about the clinical data (gender, relapse status and differentiation group classification) and the type of informative data (WES, targeted sequencing, WGG, RNA-seq, cytogenetics) available for each patient. ETP-ALL cases are indicated directly in the corresponding square of the "Group" column in the "Info" section. SNVs and Indels unreferenced in COSMIC v72 are indicated by a pink frame and were considered as novel. The black cross stands for missing data. †: genes included in the deleted 9p region with tumor suppressor genes *CDKN2A/B*. WES: Whole Exome Sequencing; WGG: Whole Genome Genotyping; SNV: Single Nucleotide Variant; SI: Small Indel; CNV: Copy Number Variant; HD: Hyperdiploidy.

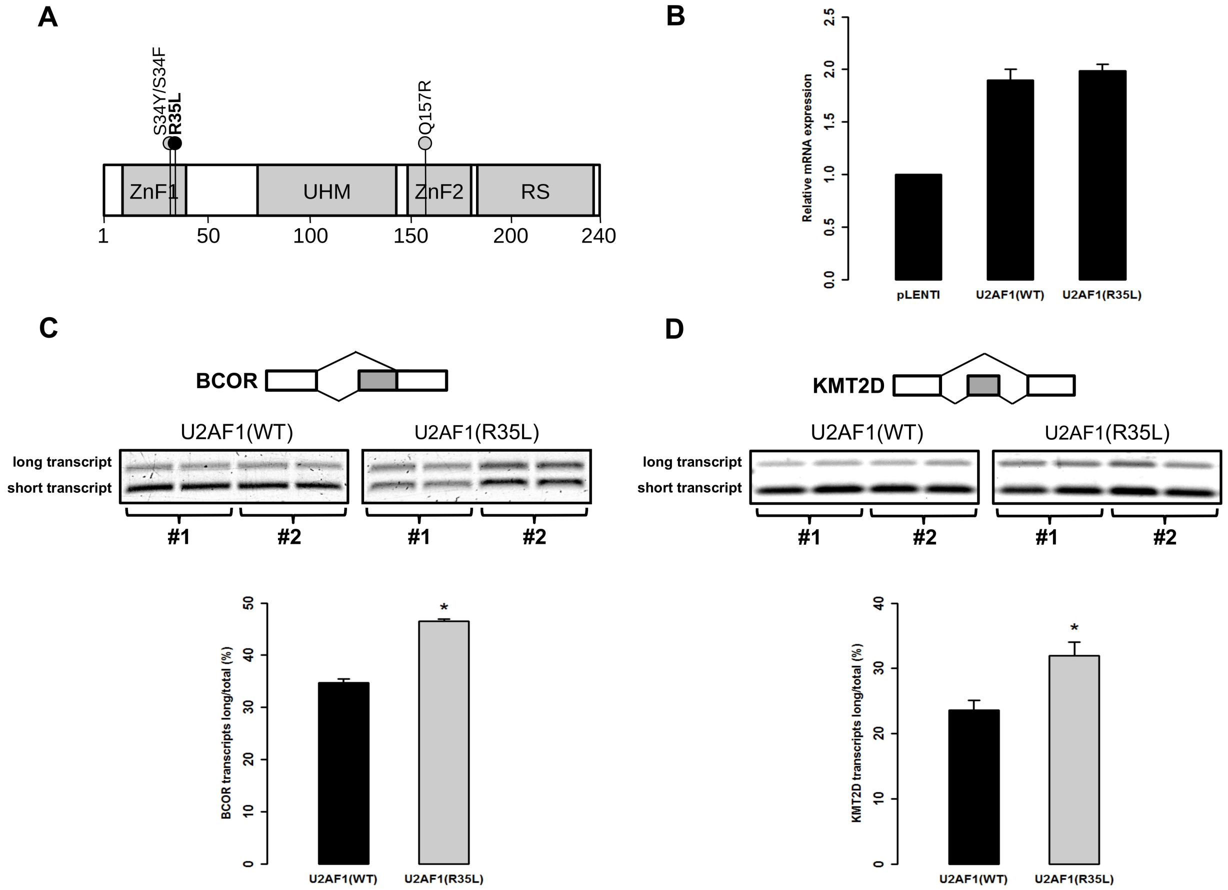


Figure 2. Mutation p.R35L in *U2AF1* alters pre-mRNA splicing in human T-cells. (**A**) Schematic representation of U2AF1 predicted protein including zinc fingers 1 and 2 (ZnF1, ZnF2), the U2AF homology motif (UHM) and the arginine-serine rich domain (RS). The black circle indicates the location of the new mutation p.R35L identified here in 3 T-ALL cases. The grey circles indicate previously identified recurrent mutations in U2AF1. The amino acid scale is indicated at the bottom. (**B**) The expression of WT or R35L *U2AF1* transgenes in Jurkat cells was measured by qPCR and reported to the expression of the endogenous gene measured in cells infected with the empty vector (pLENTI). The alternative splice site utilization of BCOR (**C**) and KMT2D/MLL2 (**D**) mRNAs were measured in infected cells as previously described [32]. RT-PCRs were performed in duplicate on cDNAs obtained from two different infections (#1 and #2) of WT or R35L *U2AF1* transgenes in Jurkat cells and PCR products were electrophoresed on agarose gel stained with SYBR Safe (top left). Quantification of each isoform was done by densitometry using Image J software (version 1.49) and the mean ratio of the long isoforms of BCOR and KMT2D/MLL2 reported to the total of both long and short isoforms (right). Statistical significance was determined by a two-tailed Mann-Whitney U test; P-values <0.05 are represented by one asterisk.

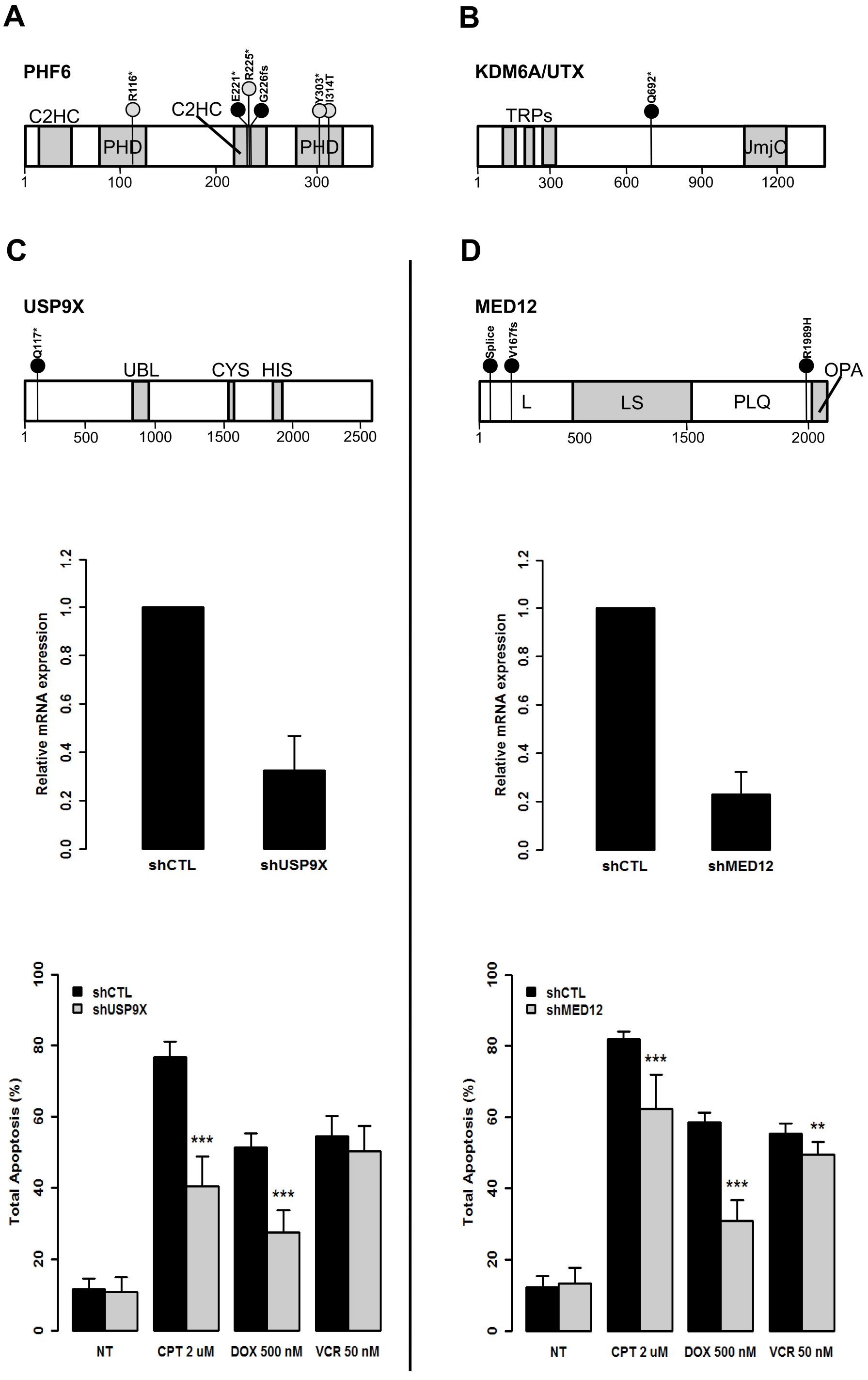


Figure 3. The loss of function of USP9X and MED12 protects from induced apoptosis in leukemic T-cells. (**A**) Schematic representation of PHF6 predicted protein including the two PHD type zinc finger domains (PHD) and the two CysCysHisCys type zinc finger domains (C2HC). Black circles indicate the location of new mutations. Grey circles indicate previously referenced mutations identified in this childhood T-ALL cohort. The size of the bars are representative of mutation frequencies in the cohort (*PHF6* p.R225* was identified in three cases while all others were identified in one case). The amino acid scale is indicated at the

bottom. (**B**) Schematic representation of KDM6A/UTX predicted protein including three tetratricopeptide repeat domains (TRPs) and the Jumanji C domain (JmjC). (**C, top**) Schematic representation of USP9X protein including the ubiquitin-like module (Ubl) and the USP-definitive cysteine and histidine box catalytic motifs (CYS and HIS). (**D, top**) Schematic representation of MED12 including the leucine-rich domain (L), the leucine- and serine-rich domain (LS), the proline-, glutamine-, leucine-rich domain (PQL) and the opposite paired domain (OPA). ShRNAs were used to knockdown expression of: *USP9X* (**C, middle**) and *MED12* (**D, middle**) in Jurkat cells. Residual mRNA levels after shRNA transduction were measured by RT-qPCR. GAPDH was used as calibrator and non-mammalian shRNAs (shCTL) were used as control for normalization. In vitro apoptosis assays show overall reduced levels of apoptosis associated with knockdown of *USP9X* (**C, bottom**) and *MED12* (**D, bottom**). DNA damage-induced apoptosis was provoked with 2 μM of Camptothecin (CPT) for 17 hours or 500 Nm of Doxorubicine (DOX) for 18 hours or 50 nM of Vincristine (VCR) for 24 hours. Non-treated (NT) transduced cells were used as controls. Overall apoptosis levels were measured after 30 minutes of annexin V (AnV) propidium iodide (PI) double staining (AnV+/PI-, AnV+/PI+, and AnV-/PI+). Statistical significance was determined by a two-tailed Mann-Whitney U test; **, P-values <0.01; ***, P-values <0.001.

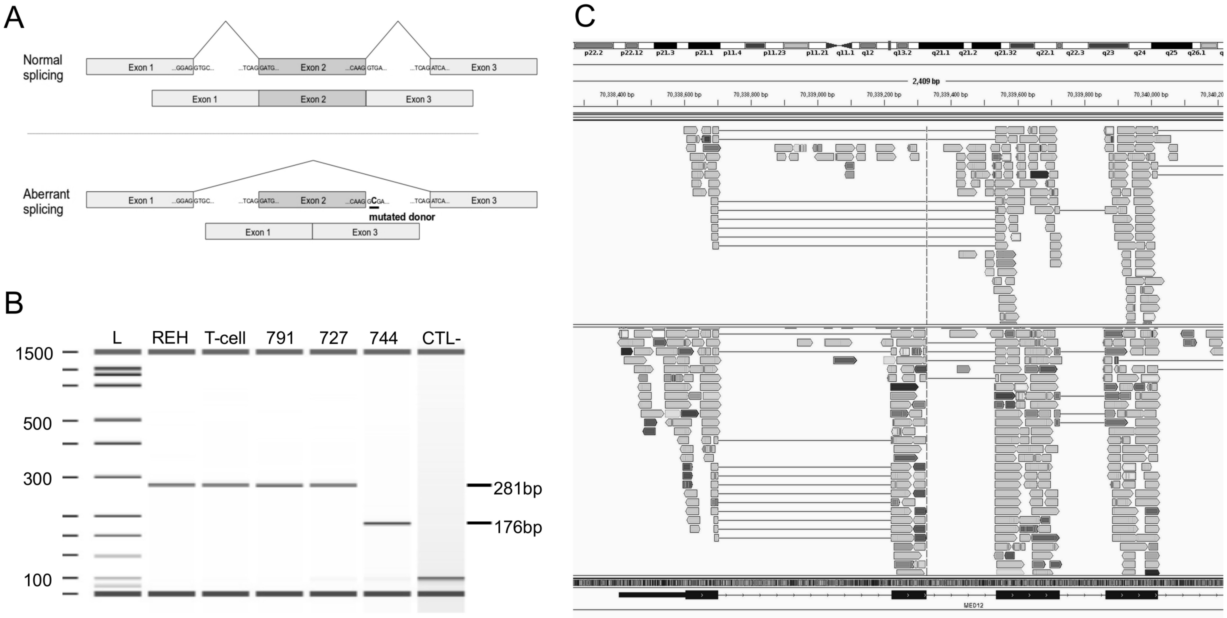
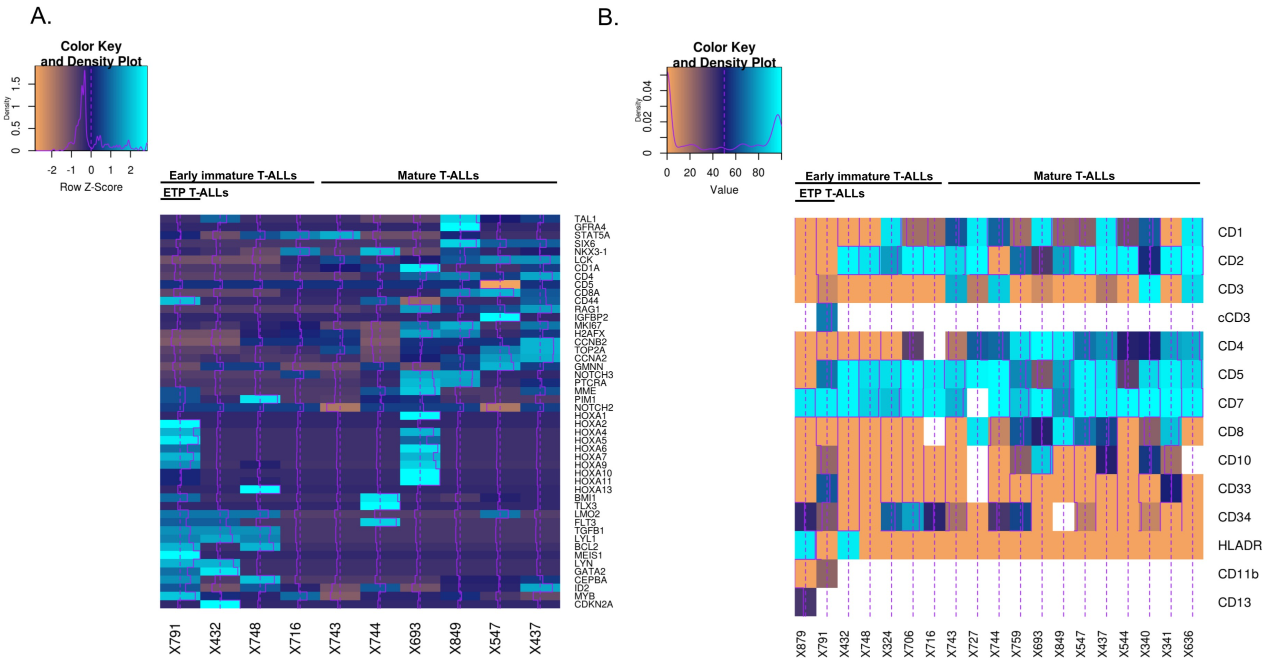
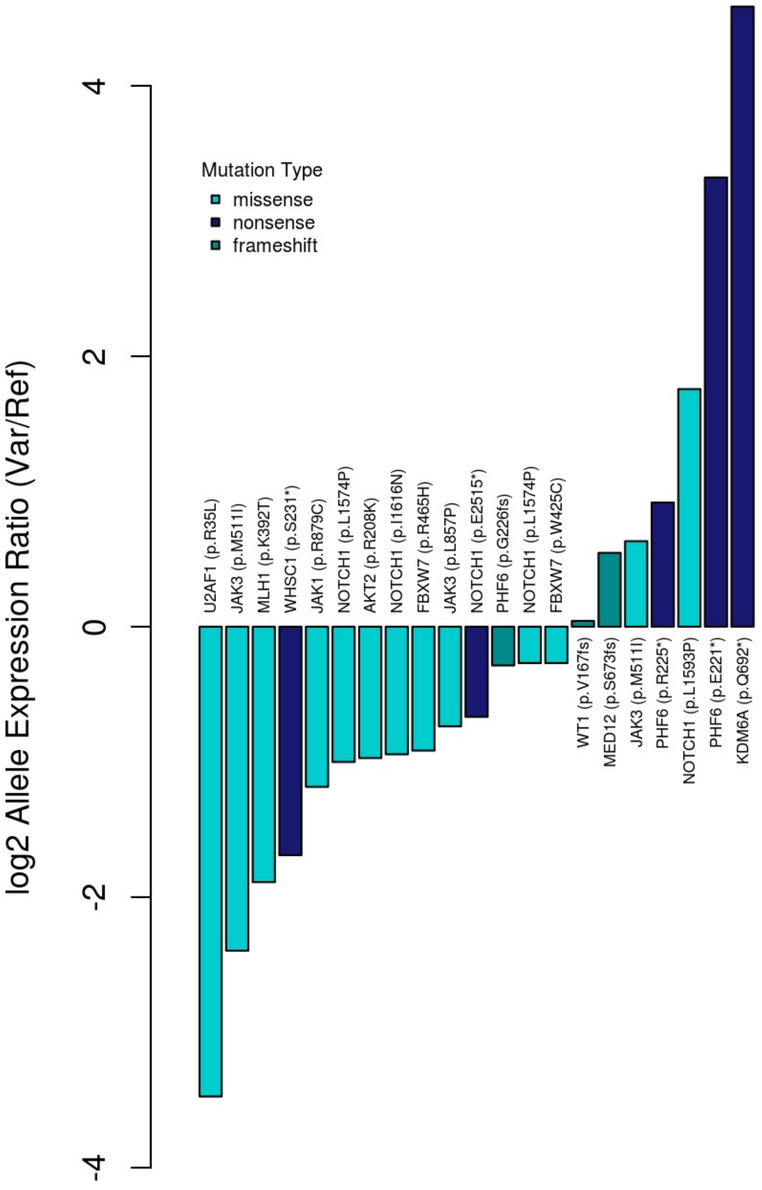


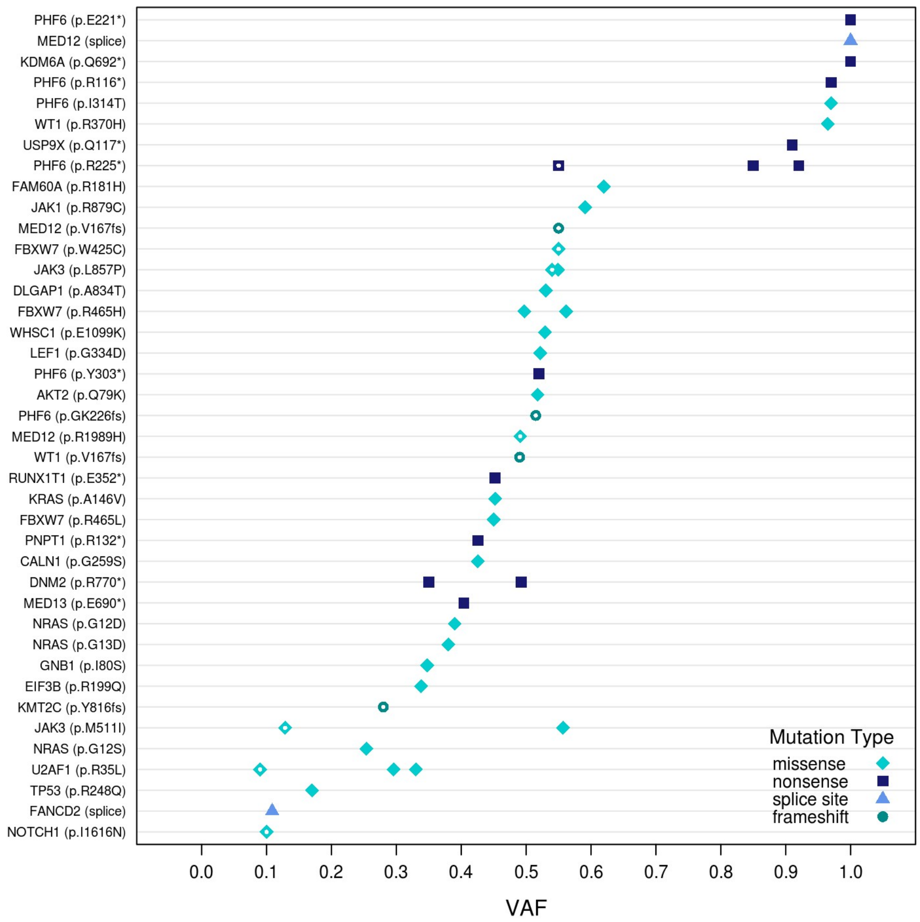
Figure 4. Mutation g.chrX:70339329T>C in *MED12* causes the splicing of exon 2. (**A**) Schematic representation of the normal (top) and aberrant splicing (bottom) of MED12 transcripts from the normal and the identified mutant gene. In the presence of the mutation, the donor splice site of exon 2 (underlined) is skipped and exon 1 is directly spliced to exon 3, leading to an aberrant mature form of MED12 mRNA lacking exon 2. (**B**) Synthetic electrophoresis gel of MED12 PCR products (Methods). cDNAs were synthesized from mature RNA extracted from patient 744 mutated at g.chrX:70339329T>C, REH cells, mature T-cells (CD19-CD3+) isolated from cord blood samples and two T-ALL patients, WT for this position (791 and 727). Amplified fragments of 281 bp (WT samples) and 176 bp (mutated sample) were analyzed using Agilent 2100 Bioanalyzer. The 105 bp difference corresponds to the skipped exon 2. L: ladder; CTL-: negative control. (**C**) Screenshot of the Integrative Genomics Viewer (IGV) window presenting aligned RNA-seq data obtained from cases 744 (top) and a third T-ALL WT patient (693, bottom). The dotted line is centered on the donor splice site of exon 2 of *MED12* (g.chrX:70339329). Case 744 shows aberrant splicing of exon 2, while 693 was WT for this position and shows a mature transcript that includes the exon 2.



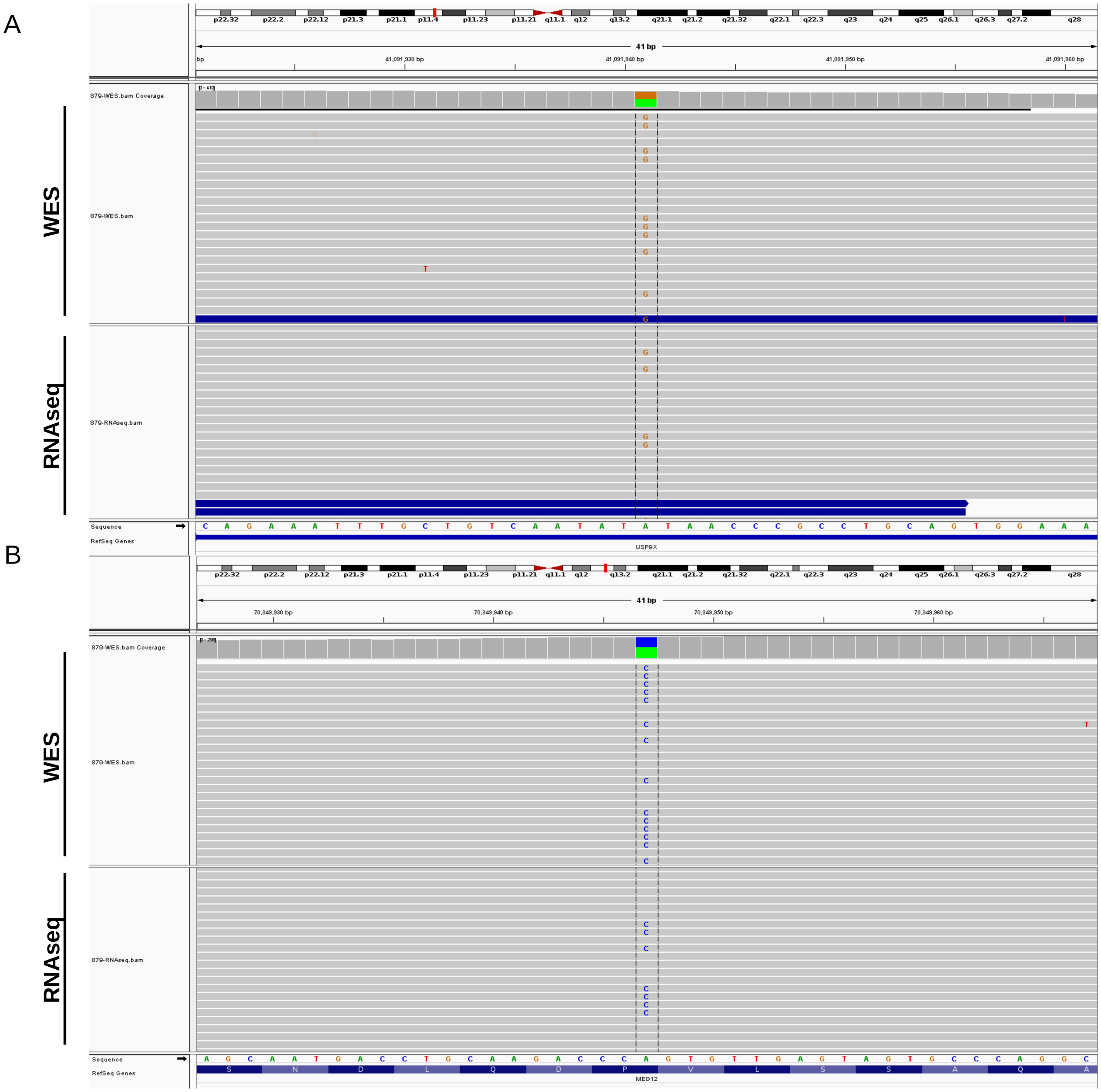
S1 Fig.Genes and antigenic determinant expression for the classification of T-ALL cases. (**A**) Informative markers were selected based on previously published classification criteria [1, 2, 9]. Z-scores were calculated based on Reads Per Kilobase per Million mapped reads (RPKM) obtained from RNA sequencing data (SOLiD 4/5500 System) using the R bioconductor package edgeR (Methods) and scaled per selected marker (red lines). Dotted lines are centered on 0. (**B**) Informative antigenic determinants. Values are percentages of positive leukemic cells for each determinant (black lines). Dotted lines are centered on 50%. White cells represent missing values. Scales for color density are indicated in the top left corner of each heatmap.



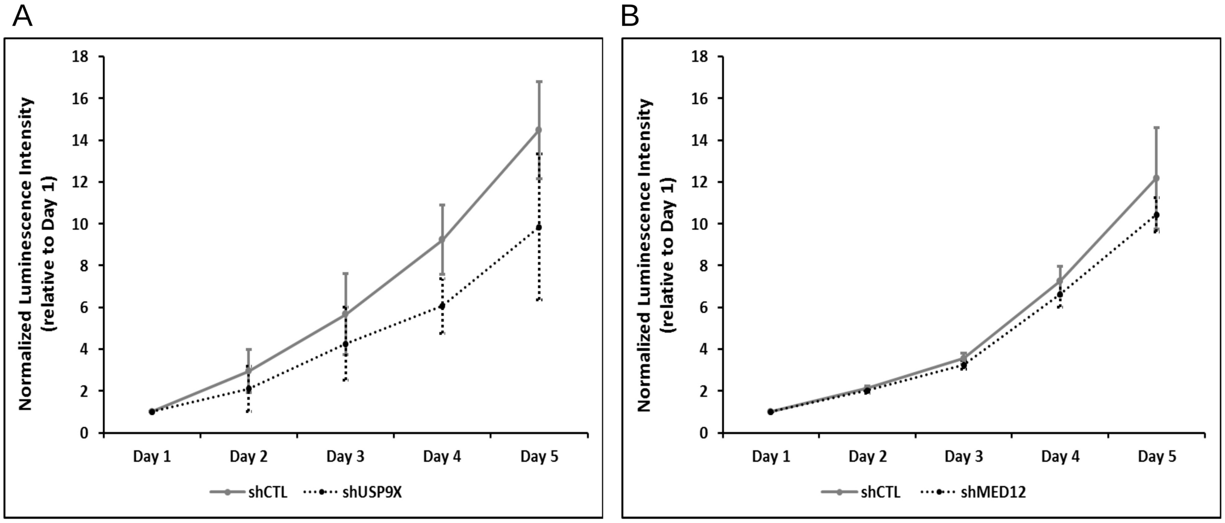
S2 Fig. The analysis of tumor transcriptomes reveal the expression of 84% of identified somatic events (SNVs and small indels). Log2 ratios of the number of RNA-seq reads supporting the variant and the reference allele for each expressed somatic event.



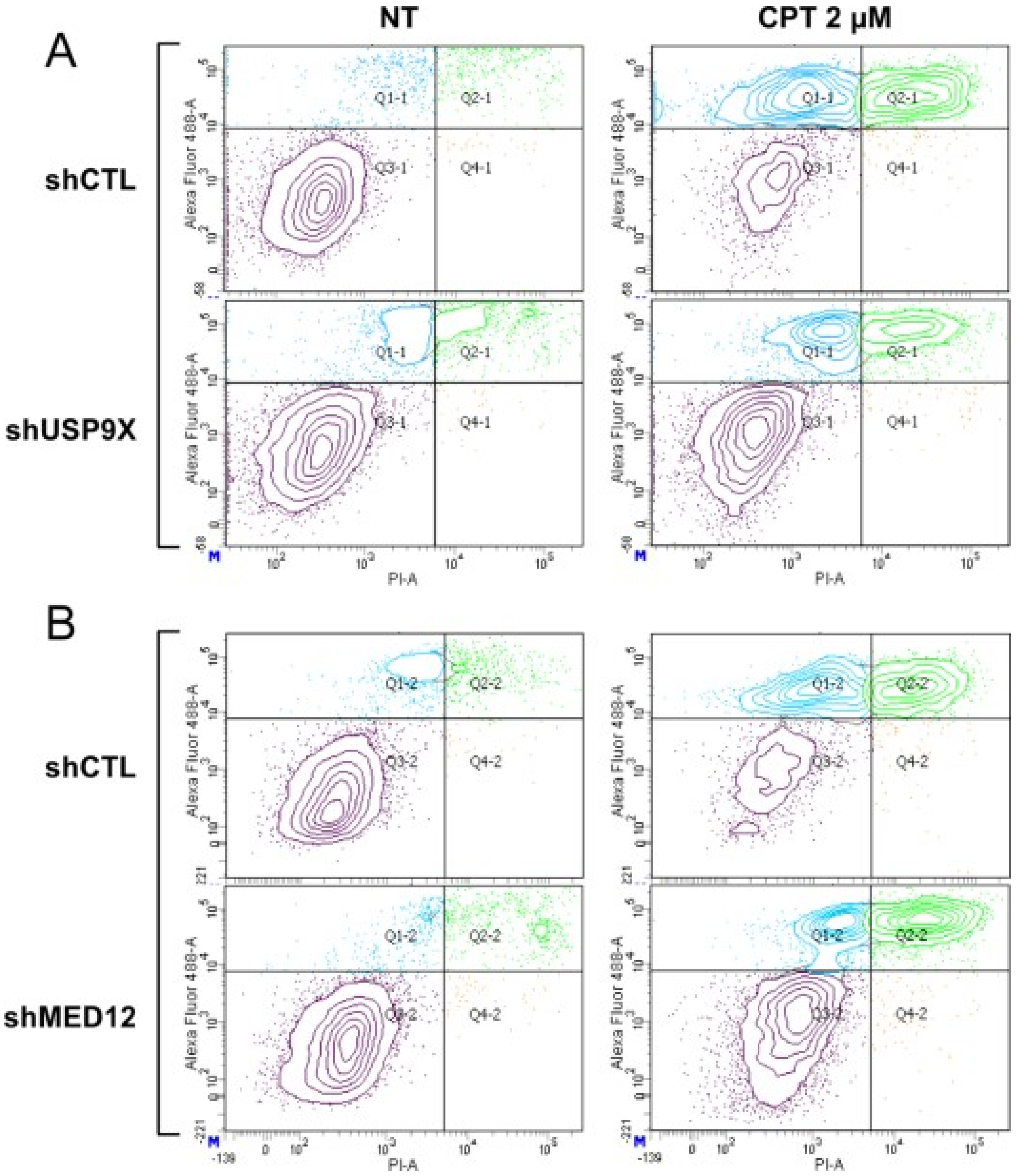
S3 Fig. Variant allele frequency (VAF) analysis at diagnosis and relapse reveals the clonal architecture of somatic events (SNVs and Small Indels) and their evolution under the selection pressure of therapy. (**A**) VAFs of somatic events correspond to the ratio of reads supporting the mutation over the total depth of coverage at the given position. VAFs were calculated from ultra-deep targeted re-sequencing (mean coverage of 2,500X) and adjusted according to either tumor purity (blast count) or from the analysis of genotype profiles using ASCAT. Amino acid changes are indicated in brackets beside gene names. Squares, circles or diamonds with a central white dot indicate events identified at diagnosis from patients with relapse data available (791 and 879). *: stop gain; fs: frameshift. (**B**) Clonal dynamics from diagnosis (pre-treatment) to relapse of the ETP-ALL cases 791 (upper panel) and 879 (lower panel). VAFs at relapse are calculated from WES data and adjusted according to tumor purity. Dark blue, green and black lines stand for clonal equilibrium, positive shift and negative shift respectively.


A
B
WES
RNAseq
p22.32 p22.2 p22.12 p21.3 p21.1 p11.4 p11.23 p11.21 q11.1 q12 q13.2 q21.1 q21.2 q21.32 q22.1 q22.3 q23 q24 q25 q26.1 q26.3 q27.2 q28
41 bp
41,091,930 bp
41,091,940 bp
41,091,950 bp
41,091,960 bp
879-WES.bam Coverage
879-WES.bam
879-RNAseq.bam
Sequence
RefSeq Genes
C A G A A A T T T G C T G T C A A T A T A T A A C C C G C C T G C A G T G G A A A
USP9X
70,349,930 bp
70,349,940 bp
70,349,950 bp
70,349,960 bp
A G C A A T G A C C T G C A A G A C C C A G T G T T G A G T A G T G C C C A G G C
S N D L Q D P V L S S A Q A
MED12

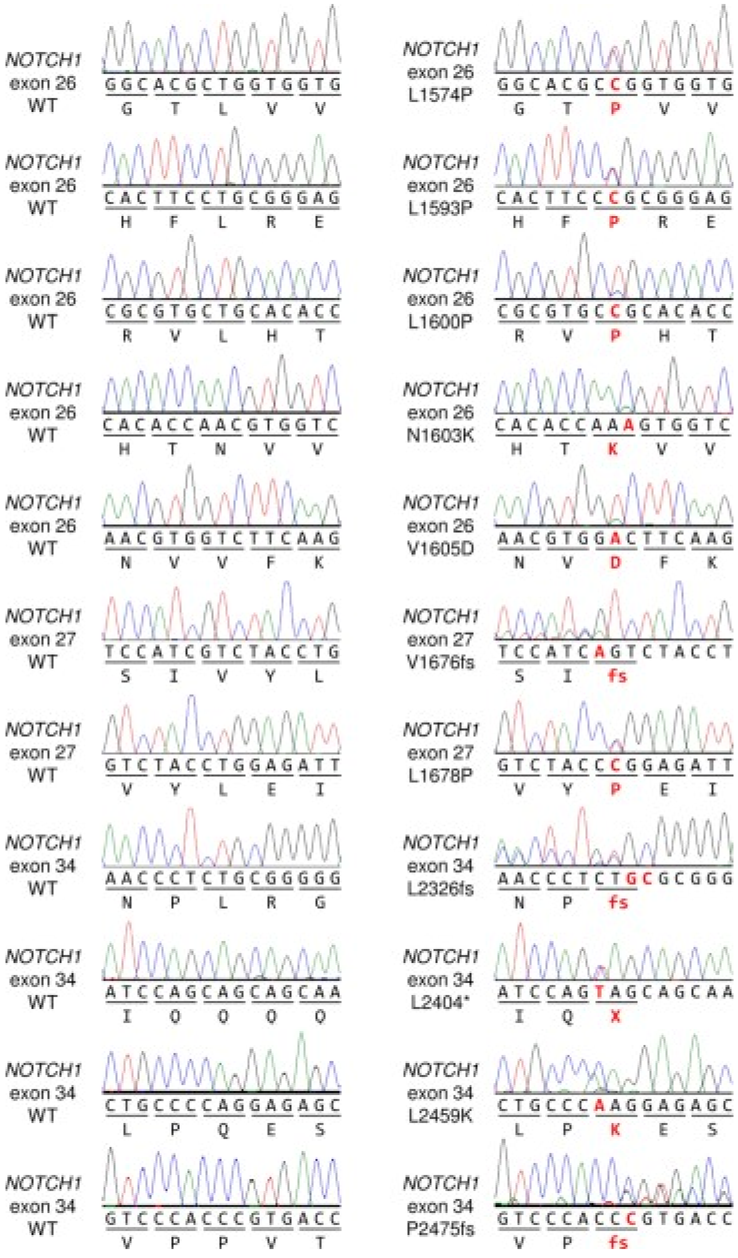
S4 Fig. *USP9X* and *MED12* escape X-inactivation in patient 879. Screenshot of the Integrative Genomics Viewer (IGV) window presenting aligned RNA-seq data obtained from patient 879. The dotted lines are centered on the informative SNPs rs10463 (**A**), located in 3'UTR of *USP9X* as control, and rs5030619 (**B**), located in exon 28 of *MED12* (bottom). Both positions are covered by reference and supporting reads.



S5 Fig. The loss of function USP9X and MED12 has no effect on cell proliferation in Jurkat cells. The experiment was performed using CellTiter-Glo® Luminescent Cell Viability Assay. 1x10E4 cells were seeded in triplicates, harvested daily over 5 days, mixed with CellTiter-Glo solution, and resulting luminescence read on Envision plate reader. Jurkat cells transduced with shUSP9X (**A**) and shMED12 (**B**) were compared to Jurkat cells transduced with non-mammalian shRNAs (shCTL). Normalization was performed daily for each replicate by comparison to data obtained during corresponding day 1.



S6 Fig. The loss of function of USP9X and MED12 protects from camptothecin-induced apoptosis in leukemic T-cells. Representative flow cytometry profiles of apoptosis assays performed on Jurkat cells infected with shRNAs targeting USP9X (**A**) and MED12 (**B**), non-treated (NT) or treated with 2 µM Camptothecin (CPT) for 17h, compared to Jurkat cells transduced with non-mammalian shRNAs (shCTL). Staining was performed using Alexa Fluor® 488-conjugated Annexin V and PI for 30 minutes. Data shown are representative of three independent shRNA infections.



S7 Fig. Sanger sequencing for the identification of mutations in *NOTCH1*. Representative DNA sequencing chromatograms of tumoral genomic DNA samples showing somatic mutations in exons 26, 27 and 34 of *NOTCH1* (Methods).

4.10. Tables

Table 1. Childhood T-ALL patient clinical information.

Patient ID	Gender	Cerebrospinal fluid invasion	Blast cells in the blood sample (%)	Blast cells in the BM sample (%)	WBC (.10^9/l)	Platelet (.10^9/l)
23	M	N	-	85.0[b]	340.3	55.0
37	M	N	-	92.0[b]	361.0	19.0
74	M	N	-	60.0[b]	10.9	119.0
95	F	N	-	-	-	-
159	M	N	-	-	17.5	-
194	M	N	-	100.0[b]	28.8	36.0
195	F	N	-	90.0[b]	163.0	21.0
200	M	N	-	71.0	52.1	85.0
210	M	N	-	93.0[b]	142.0	28.0
231	F	N	-	92.0[b]	29.3	70.0
324	M	N	-	-	22.7	136.0
340	M	N	85.0	94.5	191.5	129.0
341	M	N	33.0	96.5	42.8	62.0
432	F	> 5.10^3 blasts/ml	89.0	83.0	366.7	285.0
437	M	N	32.0	92.0	181.1	52.0
544	M	< 5.10^3 blasts/ml	97.0	99.5	465.6	31.0
547	M	N	-	90.2	195.0	66.0
636	F	N	88.0	90.5	151.3	70.0
647	F	N	68.0	92.0	31.8	15.0
693	M	> 5.10^3 blasts/ml	66.0	85.5	93.6	99.0
706	F	< 5.10^3 blasts/ml	93.0	90.5	253.1	24.0
716	M	< 5.10^3 blasts/ml	89.0	95.0	274.4	34.0
727	M	> 5.10^3 blasts/ml	66.0	90.5	119.4	74.0
743	M	< 5.10^3 blasts/ml	89.0	91.5	793.5	18.0
744	M	< 5.10^3 blasts/ml	97.0	88.0	95.5	27.0
748	M	N	19.0	33.4	7.3	100.0
759	F	N	65.0	70.5	133.0	75.0
791	F	N	0/13.0	77.7/85.0	27.7/4.28	112.0/102.0
849	M	N	61.0	86.0	105.2	31.0
879	F	< 5.10^3 blasts/ml	41.0/18.0	83.0/55.0	3.1/3.9	38.0/49.0

Patient ID	Treatment protocol	Relapse	Death	Diagnosis age (months)	Time before relapse (months)	Relapse free survival[a] (months)	T-ALL group
23	DFCI 91-01	N	N	168	NA	244	-
37	DFCI 91-01	N	N	94	NA	225	-
74	DFCI 95-01	N	N	97	NA	79	-
95	FRALLE	Y	Y	66	28, 34	NA	-
159	FRALLE	Y	Y	81	10, 13	NA	-
194	DFCI 95-01	N	N	152	NA	213	-
195	DFCI 95-01	N	N	105	NA	29	-
200	DFCI 95-01	N	N	112	NA	89	-
210	DFCI 95-01	N	N	205	NA	39	-
231	DFCI 95-01	N	N	216	NA	214	-
324	DFCI 95-01	Y	N	86	0, 135	NA	Immature
340	DFCI 95-01	N	N	117	NA	193	Mature
341	DFCI 95-01	N	N	191	NA	193	Mature
432	DFCI 2000-01	Y	Y	176	14	NA	Immature
437	DFCI 2000-01	N	N	151	NA	125	Mature
544	DFCI 2000-01	N	N	137	NA	113	Mature
547	DFCI 2000-01	Y	Y	180	9	NA	Mature
636	DFCI 2000-01	N	N	179	NA	132	Mature
647	DFCI 2000-01	N	N	68	NA	132	Mature
693	DFCI 2005-01	N	N	162	NA	120	Mature
706	DFCI 2005-01	N	N	162	NA	101	Immature
716	DFCI 2005-01	Y	Y	12	19	NA	Immature
727	DFCI 2005-01	N	N	203	NA	98	Mature
743	DFCI 2005-01	N	N	152	NA	88	Mature
744	DFCI 2005-01	N	N	178	NA	76	Mature
748	DFCI 2005-01	N	N	192	NA	40	Immature
759	DFCI 2005-01	N	Y	192	NA	6	Mature
791	DFCI 2005-01	Y	N	195	57	NA	ETP-ALL
849	DFCI 2005-01	N	N	68	NA	46	Mature
879	DFCI 2011	Y	N	137	32	NA	ETP-ALL

[a]Period during which the patient was followed after diagnosis and presented no relapse event. [b]Blast counts estimated from genotyping data analysis (Methods). BM: bone marrow; WBC: white blood cell; NA: not applicable; N: no; Y: yes; ETP: early T-cell precursor ALL; (-): missing data.

S1 Table. List of Informative antigenic determinants.

	879	791	432	748	324	706	716	743	727	744
CD1	0	0	0	0	94	20	19	63	95	66
CD2	0	0	98	94	78	98	98	92	99	0
CD3	0	10	0	0	0	0	0	82	14	89
cCD3	low	71	NA	NA	NA	NA	NA	NA	NA	NA
CD4	0	2	0	0	0	35	NA	15	65	67
CD5	0	70	98	93	93	98	96	91	98	99
CD7	95	97	98	90	86	97	97	83	NA	99
CD8	0	0	0	0	0	0	NA	0	94	11
CD10	0	22	0	0	0	0	0	0	NA	0
CD33	0	63	0	0	0	0	0	0	NA	0
CD34	45	27	0	0	66	79	46	18	0	40
HLADR	95	0	96	0	0	0	0	0	0	0
CD11b	0	23	NA	NA	NA	NA	NA	NA	NA	NA
CD13	40	0	0	0	0	0	0	0	NA	0

	759	693	849	547	437	544	340	341	636	647
CD1	19	95	23	25	95	32	65	0	95	92
CD2	69	38	80	99	98	98	54	98	99	99
CD3	0	9	0	0	16	0	98	0	92	0
cCD3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD4	93	99	96	70	76	47	49	79	82	0
CD5	74	28	79	98	95	36	91	97	92	98
CD7	88	98	76	98	97	100	99	99	98	99
CD8	64	48	96	70	60	0	25	86	0	94
CD10	25	89	0	0	47	0	57	20	NA	96
CD33	0	0	0	0	0	0	0	52	0	0
CD34	60	0	NA	13	0	0	11	0	0	0
HLADR	0	0	0	0	0	0	0	0	0	0
CD11b	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CD13	0	0	0	0	0	0	0	0	0	0

Values are percentages of positive leukemic cells for each determinant.

S2 Table. Gene expression for the classification of the 2 ETP T-ALL cases.

Gene name	Annotation transcript	RPKM (791)	RPKM (879)
TAL1	ENSG00000162367.6	976.8	2291.2
GFRA4	ENSG00000125861.10	0.0	0.0
STAT5A	ENSG00000126561.11	3364.3	4778.2
SIX6	ENSG00000184302.6	0.0	1.1
NKX3-1	ENSG00000167034.9	840.6	131.6
LCK	ENSG00000182866.12	1239.5	389.5
CD1A	ENSG00000158477.6	202.5	17.3
CD4	ENSG00000010610.5	63.2	308.1
CD5	ENSG00000110448.6	135.2	219.9
CD8A	ENSG00000153563.11	98.0	337.8
CD44	ENSG00000026508.11	882.6	415.8
RAG1	ENSG00000166349.5	32.5	31.4
IGFBP2	ENSG00000115457.5	43.1	21.9
MKI67	ENSG00000148773.8	937.3	862.4
H2AFX	ENSG00000188486.3	4004.2	8216.2
CCNB2	ENSG00000157456.3	468.7	344.0
TOP2A	ENSG00000131747.10	2993.4	663.0
CCNA2	ENSG00000145386.5	1671.5	696.0
GMNN	ENSG00000112312.5	354.3	564.6
NOTCH3	ENSG00000074181.4	17.7	18.5
PTCRA	ENSG00000171611.5	17.2	11.8
MME	ENSG00000196549.6	6.3	23.6
PIM1	ENSG00000137193.8	7148.6	13128.9
NOTCH2	ENSG00000134250.12	294.4	378.6
HOXA1	ENSG00000105991.7	68.7	52.7
HOXA2	ENSG00000105996.5	98.4	134.9
HOXA4	ENSG00000197576.8	5.0	170.1
HOXA5	ENSG00000106004.4	5285.2	2351.6
HOXA6	ENSG00000106006.5	252.6	213.8
HOXA7	ENSG00000122592.6	285.9	130.9
HOXA9	ENSG00000078399.11	1069.4	1270.6
HOXA10	ENSG00000253293.3	1092.9	430.6
HOXA11	ENSG00000005073.5	1221.7	27.0
HOXA13	ENSG00000106031.6	10.8	281.4
BMI1	ENSG00000168283.8	2997.0	978.4
TLX3	ENSG00000164438.5	16.1	0.0
LMO2	ENSG00000135363.7	620.9	2158.2
FLT3	ENSG00000122025.10	152.1	189.1
TGFB1	ENSG00000105329.5	217.5	2276.1
LYL1	ENSG00000104903.4	6740.3	14420.1
BCL2	ENSG00000171791.10	159.3	110.6
MEIS1	ENSG00000143995.13	240.2	210.6
LYN	ENSG00000254087.2	240.4	447.3
GATA2	ENSG00000179348.7	261.6	3431.7
CEPBA	ENSG00000245848.2	22.9	8.5
ID2	ENSG00000115738.5	1510.9	911.0
MYB	ENSG00000118513.14	5741.6	3248.7
CDKN2A	ENSG00000147889.12	40.9	23.5

Informative markers were selected based on previously published classification criteria [1, 2, 9]. Reads Per Kilobase per Million mapped reads (RPKM) obtained from RNA sequencing data (Illumina HiSeq 2500 System) using the R bioconductor package edgeR (Methods).

S3 Table. Identified SNVs, small indels and CNVs among 30 childhood T-ALL patients.

Gene name	Data	Patient	Chr	Start	End	Adj. VAF	Ref	Alt	AA change	Type	Effect	Gene overlap (%)
ABL1	WGG	74	chr9	131476205	134038821	-	-	-	-	CNV	loss	100
	Cytogenetics[b]	744		-	-	-	-	-	-	CNV	loss	100
AKT2	WES - TS - RNAseq	23	chr19	40761117	40761117	0.52	G	T	p.Q79K	SNV	missense	-
AKT2	WES - TS - RNAseq	716	chr19	40745968	40745968	0.32[d]	C	T	p.R208K	SNV	missense	-
AXIN1	WGG	231	chr16	84130	789549	-	-	-	-	CNV	gain	100
CALN1	WES - TS - RNAseq	759	chr7	71252771	71252771	0.43	C	T	p.G259S	SNV	missense	-
CALN1	WGG	727	chr7	71259827	71304752	-	-	-	-	CNV	loss	8
CDKN2A	WGG	743	chr9	21818310	21988896	-	-	-	-	CNV	loss	79
CDKN2A	WGG	210		21826387	21990457	-	-	-	-	CNV	loss	85
CDKN2A	WGG	23		21830479	22005330	-	-	-	-	CNV	loss	100
CDKN2A	WGG	693		21830479	22052068	-	-	-	-	CNV	loss	100
CDKN2A	WGG	759		21859194	21990457	-	-	-	-	CNV	loss	100
CDKN2A	WGG[c]	341, 636, 849		-	-	-	-	-	-	CNV	loss	100
CDKN2A	Cytogenetics[b]	37, 195, 437, 544, 547, 706, 727, 744		-	-	-	-	-	-	CNV	loss	100
CDKN2B	WGG	341	chr9	20175694	22213536	-	-	-	-	CNV	loss	100
CDKN2B	WGG	437	chr9	20368785	22492584	-	-	-	-	CNV	loss	100
CDKN2B	WGG	849	chr9	21393637	22391492	-	-	-	-	CNV	loss	100
CDKN2B	WGG	195	chr9	21576438	23445221	-	-	-	-	CNV	loss	100
CDKN2B	WGG	636	chr9	21686858	22340785	-	-	-	-	CNV	loss	100
CDKN2B	WGG	23	chr9	21830479	22005330	-	-	-	-	CNV	loss	13
CDKN2B	WGG	693	chr9	21830479	22052068	-	-	-	-	CNV	loss	69
CDKN2B	WGG	706	chr9	21853221	23943846	-	-	-	-	CNV	loss	100
CDKN2B	WGG	544	chr9	21862549	22869989	-	-	-	-	CNV	loss	100
CDKN2B	WGG	200	chr9	21965073	22790288	-	-	-	-	CNV	loss	100
CDKN2B	Cytogenetics[b]	744	chr9	-	-	-	-	-	-	CNV	loss	100
CREBBP	WGG	636	chr16	3405875	4235142	-	-	-	-	CNV	loss	100
CTCF	WGG	791	chr16	67357808	68294157	-	-	-	-	CNV	loss	100
DLGAP1	WES - TS - RNAseq	636	chr18	3508639	3508639	0.53	C	T	p.A834T	SNV	missense	-
DNM2	WES - TS - RNAseq	159, 544	chr19	10940819	10940819	0.35,0.49	C	T	p.R770*	SNV	nonsense	-
EHMT1	WGG	437	chr9	138592368	141095050	-	-	-	-	CNV	gain	100
EHMT1	WGG	432	chr9	138875850	141095050	-	-	-	-	CNV	gain	100
EHMT1	WGG	200	chr9	140422452	141095050	-	-	-	-	CNV	gain	100
EHMT1	WGG	693	chr9	140447805	141095050	-	-	-	-	-	copy-neutral-LOH	100
EIF3B	WES - TS - RNAseq	200	chr7	2400441	2400441	0.34	G	A	p.R199Q	SNV	missense	-
ETV6	WGG	231	chr12	11610960	12981519	-	-	-	-	CNV	loss	100
ETV6	WGG	791	chr12	11685519	12526576	-	-	-	-	CNV	loss	100
FAM60A	WES - TS - RNAseq	200	chr12	31435770	31435770	0.62	C	T	p.R181H	SNV	missense	-
FANCD2	WES - TS - RNAseq	37	chr3	10130131	10130131	0.11	A	G	-	SNV	splice site	-
FBXW7	WES - TS - RNAseq	194	chr4	153249384	153249384	0.45	C	A	p.R465L	SNV	missense	-
	WES - TS - RNAseq	340, 693		153249384	153249384	0.50,0.56	C	T	p.R465H	SNV	missense	-
FBXW7	WES - TS - RNAseq	879	chr4	153249503	153249503	0.55/0.45	C	A	p.W425C	SNV	missense	-

Gene name	CHASM p-value	CHASM FDR	Mutation Taster score	Polyphen2 score	SIFT score	Driver genes[a]	Occurrences in COSMIC (v72) primary site	Occurrences in COSMIC (v72) primary site (haematopoietic and lymphoid tissue)	dbSNP_RS	ExAc Aggregated Pop.	ESP Cohort Pop.	1000 Genome
ABL1	-	-	-	-	-	Oncogene	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
AKT2	0.0002	0.05 (D)	1.00 (D)	0.994 (D)	0.00 (D)	-	0	0	-	0	0	0
AKT2			0.999184 (D)	0.04 (T)	0.71 (T)		0	0	rs35817154	0.00028827	0.00065905	0
AXIN1	-	-	-	-	-	Oncogene	-	-	-	0	0	0
CALN1	0.0427	0.10 (D)	0.999998 (D)	0.984 (D)	0.00 (D)	-	2	0	-	0	0	0
CALN1	-	-	-	-	-		-	-	-	0	0	0
CDKN2A	-	-	-	-	-	TSG	-	-	-	0	0	0
CDKN2A	-	-	-	-	-		-	-	-	0	0	0
CDKN2A	-	-	-	-	-		-	-	-	0	0	0
CDKN2A	-	-	-	-	-		-	-	-	0	0	0
CDKN2A	-	-	-	-	-		-	-	-	0	0	0
CDKN2A	-	-	-	-	-		-	-	-	0	0	0
CDKN2A	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-	TSG	-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CDKN2B	-	-	-	-	-		-	-	-	0	0	0
CREBBP	-	-	-	-	-	TSG	-	-	-	0	0	0
CTCF	-	-	-	-	-	TSG	-	-	-	0	0	0
DLGAP1	0.0000	0.05 (D)	1.00 (D)	0.998 (D)	0.00 (D)	-	0	0	-	0	0	0
DNM2	-	-	-	-	1.00 (T)	TSG	6	4	-	0	0	0
EHMT1	-	-	-	-	-	-	-	-	-	0	0	0
EHMT1	-	-	-	-	-		-	-	-	0	0	0
EHMT1	-	-	-	-	-		-	-	-	0	0	0
EHMT1	-	-	-	-	-		-	-	-	0	0	0
EIF3B	0.0396	0.10 (D)	1.00 (D)	0.91 (D)	0.12 (T)	-	1	0	-	0	0	0
ETV6	-	-	-	-	-	TSG	-	-	-	0	0	0
ETV6	-	-	-	-	-		-	-	-	0	0	0
FAM60A	0.0002	0.05 (D)	0.999993 (D)	0.209 (T)	0.27 (T)	-	0	0	-	0	0	0
FANCD2	-	-	1.00 (D)	-	0.00 (D)	-	0	0	-	0	0	0
FBXW7	0.0002	0.05 (D)	1.00 (D)	1.00 (D)	0.14 (T)	TSG	242	65	-	0	0	0
	0.0002	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)				-	0	0	0
FBXW7	0.0074	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)		10	0	-	0	0	0

Gene name	Data	Patient	Chr	Start	End	Adj. VAF	Ref	Alt	AA change	Type	Effect	Gene overlap (%)
GNB1	WES - TS - RNAseq	231	chr1	1737942	1737942	0.35	A	C	p.I80S	SNV	missense	-
JAG2	WGG	340	chr14	105609836	105648277	-	-	-	-	CNV	gain	91
	WGG	706	chr14	105622157	105644421	-	-	-	-	CNV	gain	47
JAK1	WES - TS - RNAseq	744	chr1	65306942	65306942	0.59	G	A	p.R879C	SNV	missense	-
JAK3	WES - TS - RNAseq	341, 879	chr19	17943438	17943438	0.55,0.54/0.53	A	G	p.L857P	SNV	missense	-
	WES - TS - RNAseq	744, 791	chr19	17949108	17949108	0.56,0.13/0	C	T	p.M511I	SNV	missense	-
KDM6A	WES - TS - RNAseq	744	chrX	44928974	44928974	1.00	C	T	p.Q692*	SNV	nonsense	-
KMT2C	WES - TS - RNAseq	791	chr7	151945071	151945072	0.28/0.39	-	T	p.Y816fs	SI	frameshift	-
KRAS	WES - TS - RNAseq	544	chr12	25378561	25378561	0.45	G	A	p.A146V	SNV	missense	-
LEF1	WES - TS - RNAseq	647	chr4	108999383	108999383	0.52	C	T	p.G334D	SNV	missense	-
	WGG	727	chr4	109004975	109085387	-	-	-	-	CNV	loss	67
MED12	WES - TS - RNAseq	744	chrX	70339329	70339329	1.00	T	C	-	SNV	splice site	-
	WES - TS - RNAseq	791		70357715	70357715	0.49/0.44	G	A	p.R1989H	SNV	missense	-
	WES - TS - RNAseq	879		70344655	70344656	0.55/0.49	-	A	p.V167fs	SI	frameshift	-
MED13	WES - TS - RNAseq	341	chr17	60072626	60072626	0.40	C	A	p.E690*	SNV	nonsense	-
MET	WGG	432	chr7	115541339	116815511	-	-	-	-	CNV	loss	100
MIR31HG	WGG	195	chr9	20033087	21574194	-	-	-	-	CNV	loss	100
	WGG	341	chr9	20175694	22213536	-	-	-	-	CNV	loss	100
	WGG	544	chr9	20455616	21862272	-	-	-	-	CNV	loss	100
	WGG	636	chr9	20597576	21686527	-	-	-	-	CNV	loss	100
	WGG	849	chr9	21393637	22391492	-	-	-	-	CNV	loss	100
	WGG	200	chr9	21476348	21962012	-	-	-	-	CNV	loss	79
MLH1	WES - TS - RNAseq	748	chr3	37067264	37067264	0.21[d]	A	C	p.K392T	SNV	missense	-
	WGG	636		37074668	37147069	-	-	-	-	CNV	loss	31
MLLT3	WGG	37	chr9	19688388	21859194	-	-	-	-	CNV	loss	100
	WGG	195	chr9	20033087	21574194	-	-	-	-	CNV	loss	100
	WGG	341	chr9	20175694	22213536	-	-	-	-	CNV	loss	100
	WGG	437	chr9	20368785	22492584	-	-	-	-	CNV	loss	90
	WGG	544	chr9	20455616	21862272	-	-	-	-	CNV	loss	59
MTAP	WGG	849	chr9	21393637	22391492	-	-	-	-	CNV	loss	100
	WGG	200	chr9	21476348	21962012	-	-	-	-	CNV	loss	100
	WGG	636	chr9	21686858	22340785	-	-	-	-	CNV	loss	100
	WGG	743	chr9	21818310	21988896	-	-	-	-	CNV	loss	75
	WGG	210	chr9	21826387	21990457	-	-	-	-	CNV	loss	62
	WGG	23	chr9	21830479	22005330	-	-	-	-	CNV	loss	56
	WGG	693	chr9	21830479	22052068	-	-	-	-	CNV	loss	56
	WGG	759	chr9	21859194	21990457	-	-	-	-	CNV	loss	11
	WGG[c]	341	chr9	-	-	-	-	-	-	CNV	loss	100
NF1	WGG	791	chr17	28853530	30662832	-	-	-	-	CNV	loss	100
	WGG	231		29347759	30306369	-	-	-	-	CNV	loss	100

Gene name	CHASM p-value	CHASM FDR	Mutation Taster score	Polyphen2 score	SIFT score	Driver genes[a]	Occurrences in COSMIC (v72) primary site	Occurrences in COSMIC (v72) primary site (haematopoietic and lymphoid tissue)	dbSNP_RS	ExAc Aggregated Pop.	ESP Cohort Pop.	1000 Genome
GNB1	0.0407	0.10 (D)	1.00 (D)	0.999 (D)	0.03 (D)	-	2	2	-	0	0	0
JAG2	-	-	-	-	-	-	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
JAK1	0.0000	0.05 (D)	1.00 (D)	0.853 (P)	0.00 (D)	Oncogene	2	2	-	0	0	0
JAK3	0.0018	0.05 (D)	1.00 (D)	0.998 (D)	0.00 (D)	Oncogene	6	6	-	0	0	0
	0.0664	0.15 (D)	0.926842 (D)	0.074 (T)	0.12 (T)		32	30	-	0	0	0
KDM6A	-	-	-	-	0.06 (T)	TSG	0	0	-	0	0	0
KMT2C	-	-	-	-	-	TSG	10	0	rs150073007|rs202184064	0	0	0
KRAS	0.0014	0.05 (D)	1.00 (D)	0.909 (P)	0.00 (D)	Oncogene	44	7	-	0	0	0
LEF1	0.0042	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)	-	0	0	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
MED12	-	-	1.00 (D)	-	0.00 (D)	TSG	0	0	-	0	0	0
	0.0235	0.10 (D)	1.00 (D)	0.994 (D)	0.11 (T)		0	0	-	0	0	0
	-	-	-	-	-		0	0	-	0	0	0
MED13	-	-	-	-	0.89 (T)	-	0	0	-	0	0	0
MET	-	-	-	-	-	TSG	-	-	-	0	0	0
MIR31HG	-	-	-	-	-	-	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
MLH1	0.0002	0.05 (D)	1.00 (D)	0.831 (P)	0.13 (T)	TSG	0	0	rs587780678	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
MLLT3	-	-	-	-	-	-	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
MTAP	-	-	-	-	-	-	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
NF1	-	-	-	-	-	TSG	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0

Gene name	Data	Patient	Chr	Start	End	Adj. VAF	Ref	Alt	AA change	Type	Effect	Gene overlap (%)
NOTCH1	Sanger sequencing	195, 340, 544	chr9	139397768	139397768	-	A	G	p.L1678P	SNV	missense	-
	Sanger sequencing	210	chr9	139399344	139399344	-	A	G	p.L1600P	SNV	missense	-
	Sanger sequencing	194, 693	chr9	139399365	139399365	-	A	G	p.L1593P	SNV	missense	-
	Sanger sequencing	437, 849	chr9	139399422	139399422	-	A	G	p.L1574P	SNV	missense	-
	Sanger sequencing	231	chr9	139399329	139399329	-	A	T	p.V1605D	SNV	missense	-
	Sanger sequencing	432	chr9	139390816	139390816	-	G	T	p.Q2459K	SNV	missense	-
	Sanger sequencing	37	chr9	139399334	139399334	-	G	T	p.N1603K	SNV	missense	-
	Sanger sequencing	341	chr9	139390766	139390766	-	G	-	p.P2475fs	SI	frameshift	-
	Sanger sequencing	636	chr9	139390981	139390981	-	G	A	p.Q2404*	SNV	nonsense	-
	Sanger sequencing	231	chr9	139391213	139391213	-	C	-	p.L2326fs	SI	frameshift	-
	Sanger sequencing	432	chr9	139391213	139391214	-	-	G	p.L2326fs	SI	frameshift	-
	Sanger sequencing	341	chr9	139397775	139397775	-	C	-	p.V1676fs	SI	frameshift	-
	WES - TS - RNAseq	748		139390648	139390648	0.38[d]	C	A	p.E2515*	SNV	nonsense	-
	WES - TS - RNAseq	791	chr9	139399296	139399296	0.10/0	A	T	p.I1616N	SNV	missense	-
NRAS	WES - TS - RNAseq	324	chr1	115258744	115258744	0.38	C	T	p.G13D	SNV	missense	-
	WES - TS - RNAseq	95	chr1	115258747	115258747	0.39	C	T	p.G12D	SNV	missense	-
	WES - TS - RNAseq	340	chr1	115258748	115258748	0.25	C	T	p.G12S	SNV	missense	-
PAX5	WGG	791	chr9	36735067	39151736	-	-	-	-	CNV	loss	100
PHF6	WES - TS - RNAseq	74	chrX	133551305	133551305	0.97	T	C	p.I314T	SNV	missense	-
	WES - TS - RNAseq	544		133527636	133527636	0.97	C	T	p.R116*	SNV	nonsense	-
	WES - TS - RNAseq	744		133547928	133547928	1.00	G	T	p.E221*	SNV	nonsense	-
	WES - TS - RNAseq	194, 727, 879		133547940	133547940	0.92,0.85,0.55/0.49	C	T	p.R225*	SNV	nonsense	-
	WES - TS - RNAseq	791		133547944	133547945	0.51/0.46	-	A	p.G226fs	SI	frameshift	-
	WES - TS - RNAseq	636		133551273	133551273	0.52	C	G	p.Y303*	SNV	nonsense	-
PNPT1	WES - TS - RNAseq	759	chr2	55912087	55912087	0.43	G	A	p.R132*	SNV	nonsense	-
PTEN	WGG	195	chr10	89652027	89746568	-	-	-	-	CNV	loss	73
PTPN2	WGG	194	chr18	12782761	12894530	-	-	-	-	CNV	loss	100
PTPRD	WGG	727	chr9	4298955	25659005	-	-	-	-	-	copy-neutral-LOH	100
	WGG	195		46587	18092330	-	-	-	-	CNV	loss	100
	WGG	341		46587	20171664	-	-	-	-	CNV	loss	100
	WGG	849		46587	21392464	-	-	-	-	-	copy-neutral-LOH	100
	WGG	743		46587	21817777	-	-	-	-	-	copy-neutral-LOH	100
	WGG	210		46587	21825996	-	-	-	-	-	copy-neutral-LOH	100
	WGG	23		46587	21827192	-	-	-	-	-	copy-neutral-LOH	100
	WGG	706		46587	21846327	-	-	-	-	-	copy-neutral-LOH	100
	WGG	759		46587	21857566	-	-	-	-	-	copy-neutral-LOH	100
	WGG	432		46587	23367240	-	-	-	-	-	copy-neutral-LOH	100
	WGG	74		46587	25656496	-	-	-	-	-	copy-neutral-LOH	100
	WGG	647		46587	39139095	-	-	-	-	CNV	loss	100
RUNX1T1	WES - TS - RNAseq	544	chr8	92999138	92999138	0.45	C	A	p.E352*	SNV	nonsense	-
TAL1	WGG	706	chr1	47681961	47698007	-	-	-	-	-	loss	49
TP53	WES - TS - RNAseq	95	chr17	7577538	7577538	0.17	C	T	p.R248Q	SNV	missense	-
U2AF1	WES - TS - RNAseq	200, 324, 791	chr21	44524453	44524453	0.30,-,0.09/0	C	A	p.R35L	SNV	missense	-
USP9X	WES - TS - RNAseq	194	chrX	40994004	40994004	0.91	C	T	p.Q117*	SNV	nonsense	-
WHSC1	WES - TS - RNAseq	340	chr4	1962801	1962801	0.53	G	A	p.E1099K	SNV	missense	-
	WES - TS - RNAseq	432		1906037	1906037	0.24[d]	C	A	p.S231*	SNV	nonsense	-
	WGG	759		697371	2071655	-	-	-	-	CNV	gain	100
WT1	WES - TS - RNAseq	544	chr11	32417943	32417943	0.96	C	T	p.R370H	SNV	missense	-
	WES - TS - RNAseq	879		32417916	32417924	0.49/0.29	-	AAGTTCTC	p.V167fs	SI	frameshift	-
	WGG	544		32172037	32556403	-	-	-	-	CNV	loss	100

Gene name	CHASM p-value	CHASM FDR	Mutation Taster score	Polyphen2 score	SIFT score	Driver genes[a]	Occurrences in COSMIC (v72) primary site	Occurrences in COSMIC (v72) primary site (haematopoietic and lymphoid tissue)	dbSNP_RS	ExAc Aggregated Pop.	ESP Cohort Pop.	1000 Genome
NOTCH1	0.0089	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)	TSG	42	42	-	0	0	0
	0.0022	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)		70	70	-	0	0	0
	0.0078	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)		40	40	-	0	0	0
	0.0113	0.05 (D)	1.00 (D)	1.00 (D)	0.00 (D)		39	39	-	0	0	0
	0.0256	0.10 (D)	1.00 (D)	0.959 (D)	0.21 (T)		3	3	-	0	0	0
	0.3137	0.40 (T)	1.00 (D)	0.207 (T)	0.07 (T)		13	13	-	0	0	0
	0.3574	0.40 (T)	0.999989 (D)	1.00 (D)	0.04 (D)		0	0	rs370523171	0.00011689	0	0
	-	-	-	-	-		0	0	-	0	0	0
	-	-	-	-	0.29 (T)		1	1	-	0	0	0
	-	-	-	-	-		0	0	-	0	0	0
	-	-	-	-	-		0	0	-	0	0	0
	-	-	-	-	-		5	3	-	0	0	0
	-	-	-	-	0.02 (D)		1	1	-	0	0	0
	0.0036	0.05	1.00 (D)	0.989 (D)	0.00 (D)		7	7	-	0	0	0
NRAS	0.0012	0.05 (D)	1.00 (D)	0.383 (T)	0.03 (D)	Oncogene	327	263	rs121434596	0.00000824	0	0
	0.0016	0.05 (D)	1.00 (D)	0.399 (T)	0.00 (D)		665	481	rs121913237	0.00001647	0	0
	0.0018	0.05 (D)	1.00 (D)	0.431 (T)	0.04 (D)		309	179	rs121913250	0	0	0
PAX5	-	-	-	-	-	TSG	-	-	-	0	0	0
PHF6	0.1755	0.30 (T)	0.999991 (D)	0.21 (T)	0.39 (T)	TSG	2	2	-	0	0	0
	-	-	-	-	0.44 (T)		20	16	-	0	0	0
	-	-	-	-	0.76 (T)		0	0	-	0	0	0
	-	-	-	-	0.35 (T)		8	7	-	0	0	0
	-	-	-	-	-		0	0	-	0	0	0
	-	-	-	-	0.55 (T)		4	2	-	0	0	0
PNPT1	-	-	-	-	1.00 (T)	-	0	0	-	0	0	0
PTEN	-	-	-	-	-	TSG	-	-	-	0	0	0
PTPN2	-	-	-	-	-	-	-	-	-	0	0	0
PTPRD	-	-	-	-	-	-	-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
RUNX1T1	-	-	-	-	0.00 (D)	-	3	0	-	0	0	0
TAL1	-	-	-	-	-	Oncogene	-	-	-	0	0	0
TP53	0.0020	0.05 (D)	0.999994 (D)	0.999 (D)	0.01 (D)	TSG	1636	105	rs11540652	0.00005765	0	0
U2AF1	0.0014	0.05 (D)	1.00 (D)	0.967 (D)	0.00 (D)	Oncogene	1	1	-	0	0	0
USP9X	-	-	-	-	0.77 (T)	TSG	0	0	-	0	0	0
WHSC1	0.0119	0.05 (D)	1.00 (D)	0.96 (D)	0.01 (D)	Oncogene	33	13	-	0	0	0
	-	-	-	-	1.00 (T)		0	0	-	0	0	0
	-	-	-	-	-		-	-	-	0	0	0
WT1	0.1022	0.20 (D)	1.00 (D)	1.00 (D)	0.00 (D)	TSG/Oncogene	33	33	rs554416372	0.00008237	0	0.0008
	-	-	-	-	-		6	6	rs371168589	0.00002471	0	0
	-	-	-	-	-		-	-	-	0	0	0

To be considered as damaging (D), mutations had to present a CHASM false discovery rate (Benjamini and Hochberg's adjusted FDR) ≤0.2, a MutationTaster score >0.9, a Polyphen2 score >0.909 or a Sift score ≤0.05. Mutations presenting a Polyphen2 score between 0.447 and 0.909 were considered as possibly damaging (P). For each algorithm, mutations that did not meet the respective filtering criteria were annotated as tolerated (T). [a]Genes categorized as tumor suppressor (TSGs) or oncogenes using an adapted version of the 20/20 rule previously published [10];[b]Event identified by cytogenetic assays only; [c]Event imputed from genotyping data. [d]Variant allele frequency (VAF) obtained from RNA-seq data. Adj. VAF: variant allele frequency adjusted according to either tumor purity (blast count) or from genotype profiles using ASCAT (when available). Slash bars separate VAFs at diagnosis and relapse for a variant when appropriate; WGG: whole genome genotyping; WES: whole exome sequencing; TS: targeted sequencing; Chr: chromosome; Ref: reference allele; Alt: Alternative allele; AA: amino acid; SNV: single nucleotide variant; SI: small indel; CNV: copy number variant; LOH: loss of heterozygosity; (-): not applicable.

S4 Table. Additional cohort of 8 adult relapsed T-ALL patients.

ID	Sample	Material	Sex	Age at diagnosis (years)
P1	diagnosis	BM	F	73
P1	relapse	BM	F	73
P2	diagnosis	PB	M	22
P2	relapse	PB	M	22
P3	diagnosis	PB	M	18
P3	relapse/resistant	PB	M	18
P4	diagnosis	PB	F	62
P4	relapse/resistant	BM	F	62
P5	diagnosis	PB	M	22
P5	relapse	PB	M	22
P6	relapse	PB	F	32
P7	relapse	BM	M	37
P8	relapse	PB	M	30

ID	Immunophenotype
P1	cCD3+, CD3-, CD7+, CD11b+,CD13+, CD19-, cCD22-, CD33+ (subset), CD34+, cCD79a+, CD117+
P1	CD1a-, CD2-, cCD3+, CD4-, CD5-, CD7+, CD8-, CD10-, CD33+, CD34+, CD45+
P2	CD1a+, CD2-, cCD3+, CD3+, CD4+, CD5+, CD7+, CD10+
P2	NA
P3	CD1a+, CD2+, cCD3+, CD3+, CD4+, CD5+, CD7+, CD8+, CD34+
P3	CD2+, CD3+, CD5+, CD7+, CD52+, CD34-, CD10-,
P4	CD2+, cCD3+, CD5+, CD7+, CD10+, CD19-, CD34+
P4	CD2+, CD5+, CD7+, CD10+, CD34+
P5	CD1a+, cCD3+, CD4-, CD5+, CD7+, CD8-, CD11b+, CD33+, CD34+, CD79a+, CD117+
P5	CD3-, CD7+, CD33+, CD34+
P6	CD1a-, CD2+ (subset), CD3+, CD4-, CD5-, CD7+, CD8-, CD19+, CD33+ (minor subset), CD34-, CD56-, cCD79a+, CD117-
P7	cCD3+, CD7+, CD13+, CD19+, CD34+
P8	Cd4-, CD5+, CD7+, CD8-, CD10+, CD13+, CD33+, CD34+, CD45+

ID	Karyotype
P1	46,XX
P1	46,XX
P2	NA
P2	NA
P3	46,XY,t(9;9)(p10;p10)[2]/46,XY[10]
P3	NA
P4	NA
P4	NA
P5	46,XY,t(1;2)(p36.1;p21),t(5;18)(q33;p11.2),t(6;11)(q27;q23)[10] t(6;11)(3'MLL+;5'MLL+)[3]
P5	NA
P6	45~46,XX,del(1)(p34p36.1)[2],-8[2],?t(10;11)(p13;q14-22)[2],add(12)(p13)[2][cp2]/46(XX)[18]
P7	74~76,XY,+1,-2,+4,+del(4)(q31.3),del(6)(q21q25)x2,+7,+12,+13,+14,+15,+16,+17,+18,+21,+22[cp7]/46,XY[13]
P8	NA

BM: bone marrow, PB: peripheral blood, NA: not applicable or not assessable.

S5 Table. List of primers.

Name	Forward / Reverse	Sequence 5'-3'	Use
NOTCH1E34aF	Forward	TGTTTTAAAAAGGCTCCTCTGG	Sequencing
NOTCH1E34aR	Reverse	GCAGCATGGCATGGTAGG	Sequencing
NOTCH1E34bF	Forward	GTTTGCTGGCTGCAGGTT	Sequencing
NOTCH1E34bR	Reverse	GAGTCACCCCATGGCTACCT	Sequencing
NOTCH1E26F	Forward	GAGTTGCGGGGATTGACC	Sequencing
NOTCH1E26R	Reverse	GCAGTTCTAAGGCTCTGCTCA	Sequencing
NOTCH1E27F	Forward	GGGTAGCAACTGGCACAAA	Sequencing
NOTCH1E27R	Reverse	GTCCTGACTGTGGCGTCAT	Sequencing
U2AF1_F1	Forward	GCGATGTGGAGATGCAGGAA	Sequencing
U2AF1_R1	Reverse	GCAGCAGGCTTCTCTGAAGT	Sequencing
U2AF1_g104t_sens	Reverse	CGGTTTATTGTGCAACAGAGAGCACCTGTCTCC	Mutagenesis
U2AF1_g104t_as	Forward	GGAGACAGGTGCTCTCTGTTGCACAATAAACCG	Mutagenesis
MED12_EX1_F	Forward	CGGGATCTTGAGCTACGAACA	PCR
MED12_EX3_R	Reverse	GGTTCACTTGGGGCTTCCTG	PCR
CDKN2A_F	Forward	CAGAGACCATGAGCCAAGGA	PCR
CDKN2A_R	Reverse	TCTACTTCCACCATGCAGCA	PCR
BCOR_F[a]	Forward	GACAGCAGCCACACTGAGAC	PCR
BCOR_R[a]	Reverse	TCTTCCGACCAGCTTCTGTT	PCR
KMT2D_F[a]	Forward	GTGCCCGATCAGAGCCTAA	PCR
KMT2D_R[a]	Reverse	GCTGGTGCTGTTCAGGGTAT	PCR
USP9X_RT_F	Forward	TGGGTTATTCCCGCACTGAA	Quantitative PCR
USP9X_RT_R	Reverse	GGGGACTTCGCTGAGTTTGA	Quantitative PCR
PHF6_RT_F	Forward	AGGCACGAAGCTGATGTGTT	Quantitative PCR
PHF6_RT_R	Reverse	TCCTTGTGAAGGTTTCTCTCGT	Quantitative PCR
MED12_RT_F_EX31	Forward	TGCTAAACTGCCCACCTCAG	Quantitative PCR
MED12_RT_R_EX32	Reverse	AAGGGCTGCTGGCTCAATAG	Quantitative PCR
U2AF1_RT_F	Forward	CGGAGTATCTGGCCTCCATC	Quantitative PCR
U2AF1_RT_R	Reverse	TCAAGAGGGCAATGGTCTGG	Quantitative PCR

[a] From Shirai *et al.*, Cancer Cell, 2015 [11].

S6 Table. List of shRNAs.

Gene	MISSION Library Reference	Region	TRC Version	Sequence 5’-3’	Received from
USP9X	TRCN0000007	3UTR	1	CCGGGAGAGTTTATTCACTGTCTTACTCGAGTAAGACAGTGAATAAACTCTCTTTTT	Dr Stéphane Gobeil (Centre Hospitalier de l'Université Laval)
	TRCN0000007	CDS	1	CCGGCGACCCTAAACGTAGACATTACTCGAGTAATGTCTACGTTTAGGGTCGTTTTT	
	TRCN0000007	CDS	1	CCGGCGCCTGATTCTTCCAATGAAACTCGAGTTTCATTGGAAGAATCAGGCGTTTTT	
MED12	TRCN0000235	CDS	2	CCGGTTCGCTGGTCTTTCGATAAATCTCGAGATTTATCGAAAGACCAGCGAATTTTTG	Dr Stéphane Gobeil (Centre Hospitalier de l'Université Laval)
	TRCN0000235	CDS	2	CCGGCACAGCCCAGTACCAACATATCTCGAGATATGTTGGTACTGGGCTGTGTTTTTG	
	TRCN0000235	3UTR	2	CCGGCCCACCCTTTCCTCTTAATTCCTCGAGGAATTAAGAGGAAAGGGTGGGTTTTTG	
	TRCN0000235	CDS	2	CCGGAGATCATCACCAAGTACTTATCTCGAGATAAGTACTTGGTGATGATCTTTTTTG	
	TRCN0000018	CDS	1	CCGGGCAGAGAAATTACGTTGTAATCTCGAGATTACAACGTAATTTCTCTGCTTTTT	
CTL	SHC002	N/A	1	CCGGCAACAAGATGAAGAGCACCAACTCGAGTTGGTGCTCTTCATCTTGTTGTTTTT	Sigma-Aldrich (Saint Louis, Missouri)
	SHC202		2	CCGGCAACAAGATGAAGAGCACCAACTCGAGTTGGTGCTCTTCATCTTGTTGTTTTT	

4.11. Competing interests

The authors (JFS, PC, CR, VS, MO, SL, PSO, TS, JH, MM and DS) declare no conflict of interest.

4.12. Acknowledgments

The authors are indebted to the patients and their parents for participating in this study. We thank Dr. Stéphane Gobeil (Centre Hospitalier de l'Université Laval, Quebec, QC, Canada) and Dr. Christian Beausejour (CHU Sainte-Justine, Montreal, QC, Canada) who respectively provided some of the shRNAs and the HEK293T cells used for this study. Patient tissue samples were provided by the CHU Sainte-Justine Pediatric Cancer biobank. WES and RNA-seq were performed at the Integrated Clinical Genomics Centre In Pediatrics, CHU Sainte-Justine; WGG, Sanger sequencing and part of the ultra-deep targeted sequencing were performed at the McGill University and Genome Quebec Innovation Centre; Computations were made on the supercomputer Briarée from Université de Montréal, managed by Calcul Québec and Compute Canada.

4.13. References

1. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. Lancet. 2008;371(9617):1030-43.

2. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, Downing JR. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. Science. 2008;322(5906):1377-80.

3. Ferrando AA1, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC, Behm FG, Pui CH, Downing JR, Gilliland DG, Lander ES, Golub TR, Look AT. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell. 2002;1(1):75-87.

4. Goldberg JM, Silverman LB, Levy DE, Dalton VK, Gelber RD, Lehmann L, Cohen HJ, Sallan SE, Asselin BL. Childhood T-cell acute lymphoblastic leukemia: the Dana-Farber Cancer Institute acute lymphoblastic leukemia consortium experience. J Clin Oncol. 2003;21(19):3616-22.

5. Zhang J1, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, Easton J, Chen X, Wang J, Rusch M, Lu C, Chen SC, Wei L, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012;481(7380):157-63.

6. Lindqvist CM, Nordlund J, Ekman D, Johansson A, Moghadam BT, Raine A, Övernäs E, Dahlberg J, Wahlberg P, Henriksson N, Abrahamsson J, Frost BM, Grandér D, et al. The mutational landscape in pediatric acute lymphoblastic leukemia deciphered by whole genome sequencing. Hum Mutat. 2015;36(1):118-28.

7. Ntziachristos P, Tsirigos A, Welstead GG, Trimarchi T, Bakogianni S, Xu L, Loizou E, Holmfeldt L, Strikoudis A, King B, Mullenders J, Becksfort J, Nedjic J, et al. Contrasting roles of histone 3 lysine 27 demethylases in acute lymphoblastic leukaemia. Nature. 2014;514(7523):513-7.

8. De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, Gianfelici V, Geerdens E, Clappier E, Porcu M, Lahortiga I, Lucà R, Yan J, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet. 2013;45(2):186-90.

9. Graux C, Cools J, Michaux L, Vandenberghe P, Hagemeijer A. Cytogenetics and molecular genetics of T-cell acute lymphoblastic leukemia: from thymocyte to lymphoblast. Leukemia. 2006;20(9):1496-510.

10. Rothenberg EV, Moore JE, Yui MA. Launching the T-cell-lineage developmental programme. Nat Rev Immunol. 2008;8(1):9-21.

11. Dreyling MH, Schrader K, Fonatsch C, Schlegelberger B, Haase D, Schoch C, Ludwig W, Löffler H, Büchner T, Wörmann B, Hiddemann W, Bohlander SK. MLL and CALM are fused to AF10 in morphologically distinct subsets of acute leukemia with translocation t(10;11): both rearrangements are associated with a poor prognosis. Blood. 1998;91:4662-7.

12. Asnafi V, Radford-Weiss I, Dastugue N, Bayle C, Leboeuf D, Charrin C, Garand R, Lafage-Pochitaloff M, Delabesse E, Buzyn A, Troussard X, Macintyre E. CALM-AF10 is a common fusion transcript in T-ALL and is specific to the TCRgammadelta lineage. Blood. 2003;102:1000-6.

13. Weng AP, Ferrando AA, Lee W, Morris JP 4th, Silverman LB, Sanchez-Irizarry C, Blacklow SC, Look AT, Aster JC. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. Science. 2004;306(5694):269-71.

14. Hebert J, Cayuela JM, Berkeley J, Sigaux F. Candidate tumor-suppressor genes MTS1 (p16INK4A) and MTS2 (p15INK4B) display frequent homozygous deletions in primary cells from T- but not from B-cell lineage acute lymphoblastic leukemias. Blood. 1994;84(12):4038-44.

15. Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic leukemia. J Clin Invest. 2012;122(10):3398-406.
16. Coustan-Smith E, Mullighan CG, Onciu M, Behm FG, Raimondi SC, Pei D, Cheng C, Su X, Rubnitz JE, Basso G, Biondi A, Pui CH, Downing JR, et al. Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. Lancet Oncol. 2009;10(2):147-56.
17. Neumann M, Heesch S, Gökbuget N, Schwartz S, Schlee C, Benlasfer O, Farhadi-Sartangi N, Thibaut J, Burmeister T, Hoelzer D, Hofmann WK, Thiel E, Baldus CD. Clinical and molecular characterization of early T-cell precursor leukemia: a high-risk subgroup in adult T-ALL with a high frequency of FLT3 mutations. Blood Cancer J. 2012;2(1):e55.
18. Xia Y, Brown L, Tsan JT, Yang CY, Siciliano MJ, Crist WM, Carroll AJ, Baer R. The Translocation (I ; I4)(p34;q I I) in Human T-cell Leukemia: Chromosome Breakage 25 Kilobase Pairs Downstream of the TALI Protooncogene. Genes Chromosomes Cancer. 1992;4(3):211-6.
19. Cheng JT1, Yang CY, Hernandez J, Embrey J, Baer R. The chromosome translocation (11;14)(p13;q11) associated with T cell acute leukemia. Asymmetric diversification of the translocational junctions. J Exp Med. 1990;171(2):489-501.
20. Zutter M, Hockett RD, Roberts CW, McGuire EA, Bloomstone J, Morton CC, Deaven LL, Crist WM, Carroll AJ, Korsmeyer SJ. The t(10;14)(q24;q11) of T-cell acute lymphoblastic leukemia juxtaposes the delta T-cell receptor with TCL3, a conserved and activated locus at 10q24. Proc Natl Acad Sci U S A. 1990;87(8):3161-5.
21. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med. 2004;350:1535-48.
22. Tartaglia M, Martinelli S, Cazzaniga G, Cordeddu V, Iavarone I, Spinelli M, Palmi C, Carta C, Pession A, Aricò M, Masera G, Basso G, Sorcini M, et al. Genetic evidence for lineage-related and differentiation stage-related contribution of somatic PTPN11 mutations to leukemogenesis in childhood acute leukemia. Blood. 2004;104(2):307-13.
23. Case M, Matheson E, Minto L, Hassan R, Harrison CJ, Bown N, Bailey S, Vormoor J, Hall AG, Irving JA. Mutation of genes affecting the RAS pathway is common in childhood acute lymphoblastic leukemia. Cancer Res. 2008;68(16):6803-9.
24. Irving J, Matheson E, Minto L, Blair H, Case M, Halsey C, Swidenbank I, Ponthan F, Kirschner-Schwabe R, Groeneveld-Krentz S, Hof J, Allan J, Harrison C, et al. RAS pathway mutations are highly prevalent in relapsed childhood acute lymphoblastic leukaemia, are frequently relapse-drivers and confer sensitivity to MEK Inhibition. Blood. 2013;122(21):823.
25. Wang NJ, Sanborn Z, Arnett KL, Bayston LJ, Liao W, Proby CM, Leigh IM, Collisson EA, Gordon PB, Jakkula L, Pennypacker S, Zou Y, Sharma M, et al. Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. Proc Natl Acad Sci U S A. 2011;108(43):17761-6.
26. Pina C, May G, Soneji S, Hong D, Enver T. MLLT3 regulates early human erythroid and megakaryocytic cell fate. Cell Stem Cell. 2008;2(3):264-73.
27. Montes M, Nielsen MM, Maglieri G, Jacobsen A, Højfeldt J, Agrawal-Singh S, Hansen K, Helin K, van de Werken HJ, Pedersen JS, Lund AH. The lncRNA MIR31HG regulates p16(INK4A) expression to modulate senescence. Nat Commun. 2015;6:6967.
28. Van der Meulen J, Sanghvi V, Mavrakis K, Durinck K, Fang F, Matthijssens F, Rondou P, Rosen M, Pieters T, Vandenberghe P, Delabesse E, Lammens T, De Moerloose B, et al. The H3K27me3 demethylase UTX is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. Blood. 2015;125(1):13-21.
29. Przychodzen B, Jerez A, Guinta K, Sekeres MA, Padgett R, Maciejewski JP, et al. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. Blood. 2013;122(6):999-1006.

30. Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, Bradley RK. U2AF1 mutations alter splice site recognition in hematological malignancies. Genome Res. 2015;25(1):14-26.

31. Okeyo-Owuor T, White BS, Chatrikhi R, Mohan DR, Kim S, Griffith M, Ding L, Ketkar-Kulkarni S, Hundal J, Laird KM, Kielkopf CL, Ley TJ, Walter MJ, et al. U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. Leukemia. 2015;29(4):909-17.

32. Shirai CL, Ley JN, White BS, Kim S, Tibbitts J, Shao J, Ndonwi M, Wadugu B, Duncavage EJ, Okeyo-Owuor T, Liu T, Griffith M, McGrath S, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. Cancer Cell. 2015;27(5):631-43.

33. Vicente C, Schwab C, Broux M, Geerdens E, Degryse S, Demeyer S, Lahortiga I, Elliott A, Chilton L, La Starza R, Mecucci C, Vandenberghe P, Goulden N, et al. Targeted sequencing identifies association between IL7R-JAK mutations and epigenetic modulators in T-cell acute lymphoblastic leukemia. Haematologica. 2015;100(10):1301-10.

34. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005;434:400-4.

35. Clappier E, Collette S, Grardel N, Girard S, Suarez L, Brunie G, Kaltenbach S, Yakouben K, Mazingue F, Robert A, Boutard P, Plantaz D, Rohrlich P, et al. NOTCH1 and FBXW7 mutations have a favorable impact on early response to treatment, but not on outcome, in children with T-cell acute lymphoblastic leukemia (T-ALL) treated on EORTC trials 58881 and 58951. Leukemia. 2010;24(12):2023-31.

36. Kox C, Zimmermann M, Stanulla M, Leible S, Schrappe M, Ludwig WD, Koehler R, Tolle G, Bandapalli OR, Breit S, Muckenthaler MU, Kulozik AE. The favorable effect of activating NOTCH1 receptor mutations on long-term outcome in T-ALL patients treated on the ALL-BFM 2000 protocol can be separated from FBXW7 loss of function. Leukemia. 2010;24(12):2005-13.

37. Zuurbier L, Homminga I, Calvert V, te Winkel ML, Buijs-Gladdines JG, Kooi C, Smits WK, Sonneveld E, Veerman AJ, Kamps WA, Horstmann M, Petricoin EF 3rd, Pieters R, et al. NOTCH1 and/or FBXW7 mutations predict for initial good prednisone response but not for improved outcome in pediatric T-cell acute lymphoblastic leukemia patients treated on DCOG or COALL protocols. Leukemia. 2010;24(12):2014-22.

38. Mansur MB, Hassan R, Barbosa TC, Splendore A, Jotta PY, Yunes JA, Wiemels JL, Pombo-de-Oliveira MS. Impact of complex NOTCH1 mutations on survival in paediatric T-cell leukaemia. BMC Cancer. 2012;12:9.

39. Jenkinson S, Koo K, Mansour MR, Goulden N, Vora A, Mitchell C, Wade R, Richards S, Hancock J, Moorman AV, Linch DC, Gale RE. Impact of NOTCH1/FBXW7 mutations on outcome in pediatric T-cell acute lymphoblastic leukemia patients treated on the MRC UKALL 2003 trial. Leukemia. 2013;27(1):41-7.

40. Van Vlierberghe P, Palomero T, Khiabanian H, Van der Meulen J, Castillo M, Van Roy N, De Moerloose B, Philippé J, González-García S, Toribio ML, Taghon T, Zuurbier L, Cauwelier B, et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. Nat Genet. 2010;42(4):338-42.

41. O'Neil J, Grim J, Strack P, Rao S, Tibbitts D, Winter C, Hardwick J, Welcker M, Meijerink JP, Pieters R, Draetta G, Sears R, Clurman BE, et al. FBW7 mutations in leukemic cells mediate NOTCH pathway activation and resistance to gamma-secretase inhibitors. J Exp Med. 2007;204(8):1813-24.

42. Thompson BJ, Buonamici S, Sulis ML, Palomero T, Vilimas T, Basso G, Ferrando A, Aifantis I. The SCFFBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. J Exp Med. 2007;204(8):1825-35.

43. Gutierrez A, Sanda T, Ma W, Zhang J, Grebliunaite R, Dahlberg S, Neuberg D, Protopopov A, Winter SS, Larson RS, Borowitz MJ, Silverman LB, Chin L, et al. Inactivation

of LEF1 in T-cell acute lymphoblastic leukemia. Blood. 2010;115(14):2845-51.

44. Tosello V, Mansour MR, Barnes K, Paganin M, Sulis ML, Jenkinson S, Allen CG, Gale RE, Linch DC, Palomero T, Real P, Murty V, Yao X, et al. WT1 mutations in T-ALL. Blood. 2009;114(5):1038-45.

45. Palomero T, Sulis ML, Cortina M, Real PJ, Barnes K, Ciofani M, et al. Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia. Nat Med. 2007;13(10):1203-10.

46. Gutierrez A, Sanda T, Grebliunaite R, Carracedo A, Salmena L, Ahn Y, et al. High frequency of PTEN, PI3K, and AKT abnormalities in T-cell acute lymphoblastic leukemia. Blood. 2009;114(3):647-50.

47. Ntziachristos P, Tsirigos A, Van Vlierberghe P, Nedjic J, Trimarchi T, Flaherty MS, Ferres-Marco D, da Ros V, Tang Z, Siegle J, Asp P, Hadler M, Rigo I, et al. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. Nat Med. 2012;18(2):298-301.

48. Aries, Gutierrez. An X-linked tumor suppressor in T-ALL. Blood. 2015;125(1):3-4.

49. Xie Y, Avello M, Schirle M, McWhinnie E, Feng Y, Bric-Furlong E, Wilson C, Nathans R, Zhang J, Kirschner MW, Huang SM, Cong F. Deubiquitinase FAM/USP9X interacts with the E3 ubiquitin ligase SMURF1 protein and protects it from ligase activity-dependent self-degradation. J Biol Chem. 2013;288(5):2976-85.

50. Schwickart M, Huang X, Lill JR, Liu J, Ferrando R, French DM, Maecker H, O'Rourke K, Bazan F, Eastham-Anderson J, Yue P, Dornan D, Huang DC, et al. Deubiquitinase USP9X stabilizes MCL1 and promotes tumour cell survival. Nature. 2010;463(7277):103-7.

51. Dupont S, Mamidi A, Cordenonsi M, Montagner M, Zacchigna L, Adorno M, Martello G, Stinchfield MJ, Soligo S, Morsut L, Inui M, Moro S, Modena N, et al. FAM/USP9x, a deubiquitinating enzyme essential for TGFbeta signaling, controls Smad4 monoubiquitination. Cell. 2009;136(1):123-35.

52. Zhang Y, Castillo-Morales A, Jiang M, Zhu Y, Hu L, Urrutia AO, Kong X, Hurst LD. Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving. Mol Biol Evol. 2013;30(12):2588-601.

53. Park Y, Jin HS, Liu YC. Regulation of T cell function by the ubiquitin-specific protease USP9X via modulating the Carma1-Bcl10-Malt1 complex. Proc Natl Acad Sci U S A. 2013;110(23):9433-8.

54. Sun H, Kapuria V, Peterson LF, Fang D, Bornmann WG, Bartholomeusz G, Talpaz M, Donato NJ. Bcr-Abl ubiquitination and Usp9x inhibition block kinase signaling and promote CML cell apoptosis. Blood. 2011;117(11):3151-62.

55. Hu H, Tang C, Jiang Q, Luo W, Liu J, Wei X, Liu R, Wu Z. Reduced ubiquitin-specific protease 9X expression induced by RNA interference inhibits the bioactivity of hepatocellular carcinoma cells. Oncol Lett. 2015;10(1):268-72.

56. Harris DR, Mims A, Bunz F. Genetic disruption of USP9X sensitizes colorectal cancer cells to 5-fluorouracil. Cancer Biol Ther. 2012;13(13):1319-24.

57. Cui J, Sun W, Hao X, Wei M, Su X, Zhang Y, Su L, Liu X. EHMT2 inhibitor BIX-01294 induces apoptosis through PMAIP1-USP9X-MCL1 axis in human bladder cancer cells. Cancer Cell Int. 2015;15(1):4.

58. Zhou M, Wang T, Lai H, Zhao X, Yu Q, Zhou J, Yang Y. Targeting of the deubiquitinase USP9X attenuates B-cell acute lymphoblastic leukemia cell survival and overcomes glucocorticoid resistance. Biochem Biophys Res Commun. 2015;459(2):333-9.

59. Pérez-Mancera PA, Rust AG, van der Weyden L, Kristiansen G, Li A, Sarver AL, Silverstein KA, Grützmann R, Aust D, Rümmele P, Knösel T, Herd C, Stemple DL, et al. The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. Nature.

2012;486(7402):266-70.
60. Mann KM, Ward JM, Yew CC, Kovochich A, Dawson DW, Black MA, Brett BT, Sheetz TE, Dupuy AJ; Australian Pancreatic Cancer Genome Initiative, Chang DK, Biankin AV, Waddell N, et al. Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. Proc Natl Acad Sci U S A. 2012;109(16):5934-41.
61. India Project Team of the International Cancer Genome Consortium. Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nat Commun. 2013;4:2873.
62. Lim WK, Ong CK, Tan J, Thike AA, Ng CC, Rajasegaran V, Myint SS, Nagarajan S, Nasir ND, McPherson JR, Cutcutache I, Poore G, Tay ST, et al. Exome sequencing identifies highly recurrent MED12 somatic mutations in breast fibroadenoma. Nat Genet. 2014;46(8):877-80.
63. Malik S, Roeder RG. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat Rev Genet. 2010;11(11):761-72.
64. Mäkinen N, Mehine M, Tolvanen J, Kaasinen E, Li Y, Lehtonen HJ, Gentile M, Yan J, Enge M, Taipale M, Aavikko M, Katainen R, Virolainen E, et al. MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. Science. 2011;334(6053):252-5.
65. Je EM, Kim MR, Min KO, Yoo NJ, Lee SH. Mutational analysis of MED12 exon 2 in uterine leiomyoma and other common tumors. Int J Cancer. 2012;131(6):E1044-7.
66. Turunen M, Spaeth JM, Keskitalo S, Park MJ, Kivioja T, Clark AD, Mäkinen N, Gao F, Palin K, Nurkkala H, Vähärautio A, Aavikko M, Kämpjärvi K, et al. Uterine leiomyoma-linked MED12 mutations disrupt mediator-associated CDK activity. Cell Rep. 2014;7(3):654-60.
67. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N, Nickerson E, Chae SS, Boysen G, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44(6):685-9.
68. Kämpjärvi K, Park MJ, Mehine M, Kim NH, Clark AD, Bützow R, Böhling T, Böhm J, Mecklin JP, Järvinen H, Tomlinson IP, van der Spuy ZM, Sjöberg J, et al. Mutations in Exon 1 highlight the role of MED12 in uterine leiomyomas. Hum Mutat. 2014;35(9):1136-41.
69. Huang S, Hölzel M, Knijnenburg T, Schlicker A, Roepman P, McDermott U, Garnett M, Grernrum W, Sun C, Prahallad A, Groenendijk FH, Mittempergher L, Nijkamp W, et al. MED12 controls the response to multiple cancer drugs through regulation of TGF-β receptor signaling. Cell. 2012;151(5):937-50.
70. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, McLellan MD, Dooling DJ, Abbott RM, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nat Genet. 2011;44(1):53-7.
71. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, Chalkidis G, Suzuki Y, Shiosaka M, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011;478(7367):64-9.
72. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, López-Guerra M, Colomer D, Navarro A, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2011;44(1):47-52.
73. Hou HA, Liu CY, Kuo YY, Chou WC, Tsai CH, Lin CC, Lin LI, Tseng MH, Chiang YC, Liu MC, Liu CW, Tang JL, Yao M, Li CC, Huang SY, Ko BS, Hsu SC, Chen CY, Lin CT, Wu SJ, Tsay W, Tien HF. Splicing factor mutations predict poor prognosis in patients with de novo acute myeloid leukemia. Oncotarget. 2016;doi:10.18632/oncotarget.7000.
74. Makishima H, Visconte V, Sakaguchi H, Jankowska AM, Abu Kar S, Jerez A, Przychodzen B, Bupathi M, Guinta K, Afable MG, Sekeres MA, Padgett RA, Tiu RV, et al. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis.

Blood. 2012;119(14):3203-10.
75. Zamore PD, Green MR. Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc Natl Acad Sci. 1989;86: 9243-7.
76. Rudner DZ, Breger KS, Kanaar R, Adams MD, Rio DC. RNA binding activity of heterodimeric splicing factor U2AF: at least one RS domain is required for high-affinity binding. Mol Cell Biol. 1998;18:4004-11.
77. Damm F, Chesnais V, Nagata Y, Yoshida K, Scourzic L, Okuno Y, Itzykson R, Sanada M, Shiraishi Y, Gelsi-Boyer V, Renneville A, Miyano S, Mori H, et al. BCOR and BCORL1 mutations in myelodysplastic syndromes and related disorders. Blood. 2013;122(18): 3169-77.
78. Rao RC, Dou Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. Nat Rev Cancer. 2015;15(6):334-46.
79. Van der Meulen J, Speleman F, Van Vlierberghe P. The H3K27me3 demethylase UTX in normal development and disease. Epigenetics. 2014;9(5): 658-68.
80. Haferlach T, Nagata Y, Grossmann V, Okuno Y, Bacher U, Nagae G, Schnittger S, Sanada M, Kon A, Alpermann T, Yoshida K, Roller A, Nadarajah N, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia. 2014;28(2):241-7.
81. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, Yoon CJ, Ellis P, Wedge DC, Pellagatti A, Shlien A, Groves MJ, Forbes SA, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood. 2013;122(22):3616-27.
82. Walter MJ, Shen D, Shao J, Ding L, White BS, Kandoth C, Miller CA, Niu B, McLellan MD, Dees ND, Fulton R, Elliot K, Heath S, et al. Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. Leukemia. 2013 Jun;27(6):1275-82.
83. Guan X, Zhong X, Men W, Gong S, Zhang L, Han Y. Analysis of EHMT1 expression and its correlations with clinical significance in esophageal squamous cell cancer. Mol Clin Oncol. 2014;2(1):76-80.
84. Loh SW, Ng WL, Yeo KS, Lim YY, Ea CK. Inhibition of euchromatic histone methyltransferase 1 and 2 sensitizes chronic myeloid leukemia cells to interferon treatment. PLoS One. 2014;9(7):e103915.
85. Jaffe JD, Wang Y, Chan HM, Zhang J, Huether R, Kryukov GV, Bhang HE, Taylor JE, Hu M, Englund NP, Yan F, Wang Z, Robert McDonald E 3rd, et al. Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia. Nat Genet. 2013;45(11):1386-91.
86. Neumann M, Vosberg S, Schlee C, Heesch S, Schwartz S, Gökbuget N, Hoelzer D, Graf A, Krebs S, Bartram I, Blum H, Brüggemann M, Hecht J, et al. Mutational spectrum of adult T-ALL. Oncotarget. 2015;6(5):2754-66.
87. Oyer JA, Huang X, Zheng Y, Shim J, Ezponda T, Carpenter Z, Allegretta M, Okot-Kotber CI, Patel JP, Melnick A, Levine RL, Ferrando A, Mackerell AD Jr, et al. Point mutation E1099K in MMSET/NSD2 enhances its methyltranferase activity and leads to altered global chromatin methylation in lymphoid malignancies. Leukemia. 2014;28(1):198-201.
88. Andersen EF, Carey JC, Earl DL, Corzo D, Suttie M, Hammond P, South ST. Deletions involving genes WHSC1 and LETM1 may be necessary, but are not sufficient to cause Wolf-Hirschhorn Syndrome. Eur J Hum Genet. 2014;22(4):464-70.
89. Chen C, Liu Y, Rappaport AR, Kitzing T, Schultz N, Zhao Z, Shroff AS, Dickins RA, Vakoc CR, Bradner JE, Stock W, LeBeau MM, Shannon KM, et al. MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. Cancer Cell. 2014;25(5):652-65.
90. Kemp CJ, Moore JM, Moser R, Bernard B, Teater M, Smith LE, Rabaia NA, Gurley KE, Guinney J, Busch SE, Shaknovich R, Lobanenkov VV, Liggitt D, et al. CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. Cell Rep. 2014;7(4):1020-9.

91. Mullighan CG, Zhang J, Kasper LH, Lerach S, Payne-Turner D, Phillips LA, Heatley SL, Holmfeldt L, Collins-Underwood JR, Ma J, Buetow KH, Pui CH, Baker SD, et al. CREBBP mutations in relapsed acute lymphoblastic leukaemia. Nature. 2011;471(7337):235-9.
92. Andersson AK, Ma J, Wang J, Chen X, Gedman AL, Dang J, Nakitandwe J, Holmfeldt L, Parker M, Easton J, Huether R, Kriwacki R, Rusch M, et al. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. Nat Genet. 2015;47(4):330-7.
93. Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, Kluth S, Bozic I, Lawrence M, Böttcher S, Carter SL, Cibulskis K, Mertens D, et al. Mutations driving CLL and their evolution in progression and relapse. Nature. 2015;526(7574):525-30.
94. Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood. 2007;109(2):683-92.
95. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. Leuk Res. 1997;21(9):817-23.
96. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357-9.
97. Picard. Broadinstitute. 2016. [cited 5 Jan 2016]. Available: http://broadinstitute.github.io/picard/.
98. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
99. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009;25:2078-9.
100. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
101. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
102. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6(9):677-81.
103. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Børresen-Dale AL, Kristensen VN. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107(39):16910-5.
104. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
105. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G. Oncotator: cancer variant annotation tool. Hum Mutat. 2015;36(4):E2423-9.
106. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
107. Exome Variant Server. NHLBI Exome Sequencing Project (ESP). 2016. [cited 5 Jan 2016]. Available: http://evs.gs.washington.edu/EVS/.
108. Exome Aggregation Consortium, Monkol Lek, Konrad Karczewski, Eric Minikel, Kaitlin Samocha, Eric Banks, Timothy Fennell, Anne O'Donnell-Luria, James Ware, Andrew Hill, Beryl Cummings, Taru Tukiainen, Daniel Birnbaum, et al. Analysis of protein-coding genetic variation in 60,706 humans. BioRxiv. 2016. doi:http://dx.doi.org/10.1101/030338.
109. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M,

Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10:Unit 10.11.

110. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339(6127):1546-58.

111. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.

112. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

113. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361-2.

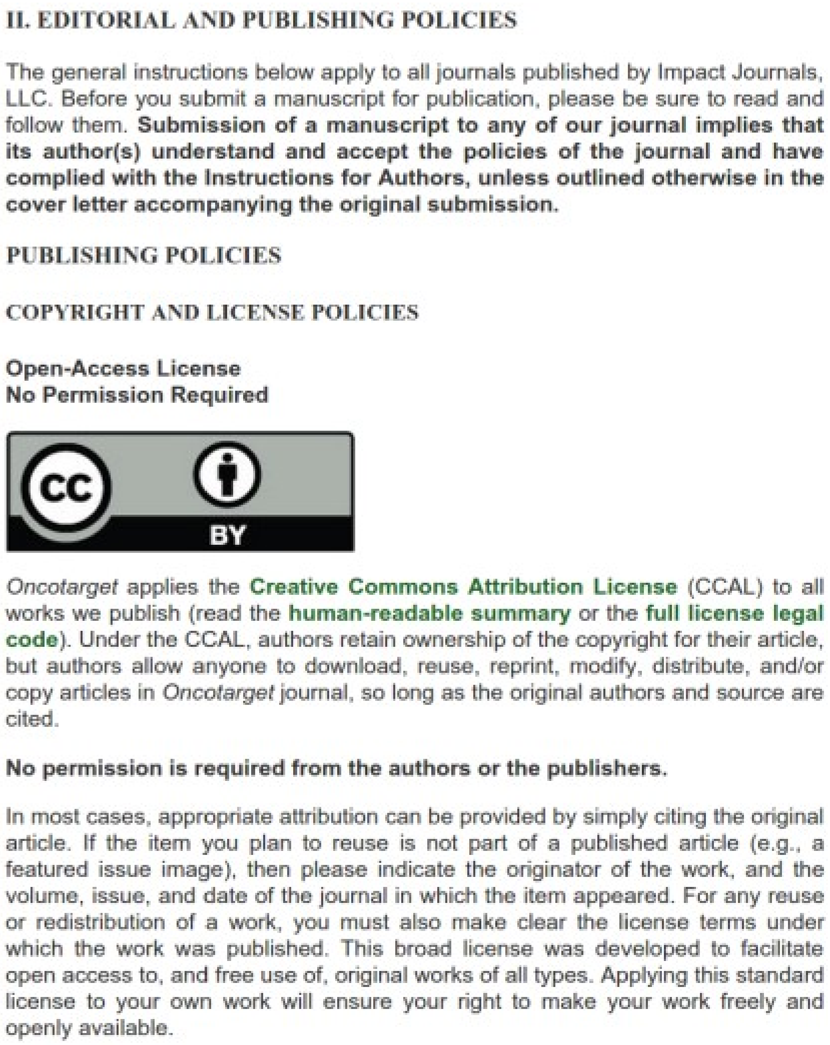
114. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69(16):6660-7.

115. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25(4):402-8.

4.14. Supplemental references

1. Soulier J, Clappier E, Cayuela JM, Regnault A, García-Peydró M, Dombret H, Baruchel A, Toribio ML, Sigaux F. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood. 2005;106(1):274-86.
2. Coustan-Smith E, Mullighan CG, Onciu M, Behm FG, Raimondi SC, Pei D, Cheng C, Su X, Rubnitz JE, Basso G, Biondi A, Pui CH, Downing JR, et al. Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. Lancet Oncol. 2009;10(2):147-56.
3. Caudell D, Aplan PD. The role of CALM-AF10 gene fusion in acute leukemia. Leukemia. 2008;22(4):678-85.
4. Ferrando AA1, Neuberg DS, Staunton J, Loh ML, Huard C, Raimondi SC, Behm FG, Pui CH, Downing JR, Gilliland DG, Lander ES, Golub TR, Look AT. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell. 2002;1(1):75-87.
5. Vicente C, Schwab C, Broux M, Geerdens E, Degryse S, Demeyer S, Lahortiga I, Elliott A, Chilton L, La Starza R, Mecucci C, Vandenberghe P, Goulden N, et al. Targeted sequencing identifies association between IL7R-JAK mutations and epigenetic modulators in T-cell acute lymphoblastic leukemia. Haematologica. 2015;100(10):1301-10.
6. Tartaglia M, Martinelli S, Cazzaniga G, Cordeddu V, Iavarone I, Spinelli M, Palmi C, Carta C, Pession A, Aricò M, Masera G, Basso G, Sorcini M, et al. Genetic evidence for lineage-related and differentiation stage-related contribution of somatic PTPN11 mutations to leukemogenesis in childhood acute leukemia. Blood. 2004;104(2):307-13.
7. Case M, Matheson E, Minto L, Hassan R, Harrison CJ, Bown N, Bailey S, Vormoor J, Hall AG, Irving JA. Mutation of genes affecting the RAS pathway is common in childhood acute lymphoblastic leukemia. Cancer Res. 2008;68(16):6803-9.
8. Irving J, Matheson E, Minto L, Blair H, Case M, Halsey C, Swidenbank I, Ponthan F, Kirschner-Schwabe R, Groeneveld-Krentz S, Hof J, Allan J, Harrison C, et al. RAS pathway mutations are highly prevalent in relapsed childhood acute lymphoblastic leukaemia, are frequently relapse-drivers and confer sensitivity to MEK Inhibition. Blood. 2013;122(21):823.
9. Han X, Bueso-Ramos CE. Precursor T-cell acute lymphoblastic leukemia/lymphoblastic lymphoma and acute biphenotypic leukemias. Am J Clin Pathol. 2007;127(4):528-44.
10. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339(6127):1546-58.
11. Shirai CL, Ley JN, White BS, Kim S, Tibbitts J, Shao J, Ndonwi M, Wadugu B, Duncavage EJ, Okeyo-Owuor T, Liu T, Griffith M, McGrath S, et al. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. Cancer Cell. 2015;27(5):631-43.

4.15. Reprint permission



II. EDITORIAL AND PUBLISHING POLICIES

The general instructions below apply to all journals published by Impact Journals, LLC. Before you submit a manuscript for publication, please be sure to read and follow them. **Submission of a manuscript to any of our journal implies that its author(s) understand and accept the policies of the journal and have complied with the Instructions for Authors, unless outlined otherwise in the cover letter accompanying the original submission.**

PUBLISHING POLICIES

COPYRIGHT AND LICENSE POLICIES

Open-Access License
No Permission Required

Oncotarget applies the **Creative Commons Attribution License** (CCAL) to all works we publish (read the **human-readable summary** or the **full license legal code**). Under the CCAL, authors retain ownership of the copyright for their article, but authors allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in *Oncotarget* journal, so long as the original authors and source are cited.

No permission is required from the authors or the publishers.

In most cases, appropriate attribution can be provided by simply citing the original article. If the item you plan to reuse is not part of a published article (e.g., a featured issue image), then please indicate the originator of the work, and the volume, issue, and date of the journal in which the item appeared. For any reuse or redistribution of a work, you must also make clear the license terms under which the work was published. This broad license was developed to facilitate open access to, and free use of, original works of all types. Applying this standard license to your own work will ensure your right to make your work freely and openly available.

Obtenu de

http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=about.

CHAPITRE 5

ARTICLE V - Mutational dynamics of early and late relapsed childhood B-cell acute lymphoblastic leukemia: rapid clonal expansion and long-term dormancy

5.1. Avant-propos

“*Cancer clone evolution takes place within tissue ecosystem habitats which have themselves evolved over a billion years. Their complex anatomies and networked signals have evolved to optimise and integrate multi-cellular functions whilst restraining renegade clonal expansion. The balance however is delicate as the resilience of multi-cellular and long-lived animals such as ourselves depend upon the very phenotypic properties that, if not tightly regulated, drive or sustain malignancy. […] Tissues provide the context for cancer cell evolution. The usually protracted time required for the clinical emergence of cancer and the resultant mutational complexity often reflect the sequential and random ‘searches’ for phenotypic solutions to micro-environmental constraints. […] Oncologists change cancer clone dynamics dramatically by introducing a new potent source of ‘artificial’ selection – with drugs or radiotherapy. But similar evolutionary principles apply. Massive cell death will usually ensue providing selective pressure for the proliferation of variant cells that can, by one of several mechanisms, resist therapeutic oblivion.*”

- Mel Greaves and Carlo C. Maley, Nature 2012 [13]

Mutational dynamics of early and late relapsed childhood B-cell acute lymphoblastic leukemia: rapid clonal expansion and long-term dormancy

Jean-François Spinella[1], Chantal Richer[1], Pauline Cassart[1], Manon Ouimet[1], Jasmine Healy[1], Daniel Sinnett[1,2]

1) CHU Sainte-Justine Research Center, Université de Montréal, Montreal, Qc, Canada; 2) Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montreal, Qc, Canada.

Cet article sera soumis au journal Leukemia en Avril 2017.

5.2. Authors' contributions

DS is the principal investigator and takes primary responsibility for the paper. JFS and DS contributed to the conception and design of the study. PC, CR, and MO were involved in sample and library preparation for whole-exome. JFS performed bioinformatics analysis, data integration and analysis. JFS drafted the paper and interpreted of the data. JH and DS were involved in critical revision of the manuscript. All authors approved the final version.

5.3. Abstract

Childhood acute lymphoblastic leukemia (cALL) is the most frequent pediatric cancer. Refractory cALL presents a long-term survival rate of about 45% and is one of the leading causes of death by disease among children. Mechanisms such as clonal competition and evolutionary adaptation govern treatment resistance. However, the underlying clonal dynamics leading to multiple relapses and differentiating early (<36 months post-diagnosis) from late relapse events remains elusive. Here, we use an integrative genome-based analysis (whole-genome- and whole-exome-sequencing) combined to serial sampling of relapsed tumors (from primary tumor to up to four relapses events) from 19 cALL patients (8 early and 11 late relapses) to assess the fitness of somatic mutations and infer their ancestral relationships. By quantifying both general clonal dynamics and newly acquired subclonal diversity, we show that two distinct evolutionary patterns govern either early or late relapse: on one hand a highly dynamic pattern, sustained by a putative defect of DNA repair processes, illustrating the quick emergence of fitter clones - and on the other hand, a quasi-inert evolution pattern suggesting the escape from dormancy of neoplastic stem cells likely spared from initial cytoreductive therapy. These results offer new insights into cALL relapse mechanisms and highlight the pressing need for adapted treatment strategies to circumvent resistance mechanisms.

5.4. Introduction

Childhood acute lymphoblastic leukemia (cALL) is the most frequent pediatric cancer, accounting for ~25% of all cases [1]. Despite improved treatment strategies, 10% of patients are either refractory or ultimately relapse, making cALL the second cause of cancer-related mortality amongst children and adolescents [2-5]. This highlights the pressing need for new therapeutic alternatives. These can be achieved through in-depth investigation and thorough understanding of the genetic architecture and mutational dynamics leading to clonal evolution of cALL and its relapse.

Tumor plasticity, based on the random accumulation of somatic mutations combined with external pressures, such as chemotherapy, is a breeding ground for clonal competition and evolutionary adaptation leading to treatment resistance [6,7]. Over the past few years, next-generation sequencing (NGS) data have provided useful information toward the characterization of these evolutionary processes and the identification of somatic events providing selective advantages. In the context of hematological malignancies, mutations in the RAS pathway genes have been associated with early relapse and chemoresistance in cALL [8]. Hyperdiploid leukemia cells mutated for the histone acetyltransferase (HAT) CREB-binding protein (CREBBP) were identified in glucocorticoid resistant patients [9,10] and the mutated cytosolic 5'-nucleotidase II gene (NT5C2) was shown to be involved in resistance to treatment using nucleoside analogue therapies [11,12]. Such studies indicated that treatment could either lead to the eradication of non-resistant clones and the emergence of fitter and more aggressive subclones or, in rarer cases, to the persistence of an already existing resistant and dominant clone [13-17].

Although important, these studies generally offered only snapshots of tumors and therefore failed to capture the complexity of their temporal evolution [6]. The latter is essential to decipher the full spectrum of therapeutic failure determinants. Furthermore, early cALL relapses, occurring during the treatment period, and late relapses, occurring at least 36 months after the initial diagnosis are likely to present different evolutionary mutational processes. However, large scale genomic studies have thus far focused on early events only [13] and did not address the yet elusive origin of long-term cALL relapses.

To decipher the full spectrum of somatic events that lead to treatment failure in cALL and shed light on the spatio-temporal clonal evolution of cALL relapse, we used whole genome (WGS, 80X) and whole-exome sequencing (WES, 200X) on a series of primary tumor (PT) samples and multiple relapses (up to four - R1 to R4) from 19 pre-B cALL patients. Based on VAFs (variant allele frequencies) analysis, we inferred the global clonal population frequencies as well as evolutionary or ancestral relationships. By dividing our cohort in either early (<36 months post-diagnosis, n =8) or late bone marrow relapse event (n =11), we successfully captured significantly different clonal dynamics and showed a variable clonal origin of late relapse events. This study is the first to use serial sampling to show the difference of clonal evolution in late versus early pre-B cALL relapses.

5.5. Results

5.5.1. The genomic landscape of pre- and post-treatment cALL

A collection of 63 samples from 19 pre-B cALL patients who suffered at least one relapse event were collected at diagnosis (primary tumor) and relapse (Table 1). Patients were all treated under DFCI protocols (Table 1). To investigate the temporal evolution of the tumors, we performed serial sampling of 7 of the 19 cases who suffered multiple relapse events (2 to 4 events). Matched normal material, collected at remission, was used as germline reference. Five cases presented a normal karyotype at diagnosis while only 2 remained normal at relapse. Two translocations t(9;14)(p13;q32) and t(11;19)(q23;p13.1), involving *PAX5/IGH* and *KMT2A/ELL* genes respectively, were identified at diagnosis (cases 382 and 808) and persisted at relapse (Table 1). Additional rearrangements (t(7;11)(q32;q14), t(1;7;11)(q21;q32;q14), t(12;14)(q12;q11) and t(16;Y)(q24;q11.2)) were gained at relapse for 3 cases (64, 217 and 382). Finally, 6 patients presented an hyper- or hypodiploidy at both diagnosis and relapse (Table 1).

Using deep WES (200X) and/or WGS (80X) associated with a strict data reduction strategy (S1 Fig, Methods), we generated a comprehensive repertoire of somatic mutations including single nucleotide variants (SNVs), insertions/deletions (indels) (Figure 1 and S1 Table) and copy number variants (CNVs) for each patient (Figure 1 and S2 Table). We completed and/or validated the DNA copy-number alterations data, the tumor purity (percentage of cancer cell in a sample) and the genome ploidy obtained from the clinic using the available whole-genome genotyping (WGG) data (Table 1, S1 Fig, Methods). Notably, recurrent CNVs were identified in the well-known tumor suppressor genes *CDKN2A* (n =4), *WHSC1* (n =4), *PAX5* (n =3), *RUNX1* (n =3) and *IKZF1* (n =2) (Figure 1 and S2 Table).

Variant allele frequencies (VAFs) were calculated from the distribution of reads either supporting the mutated or the reference allele. As suggested by the comparison of VAFs obtained for a series of samples for which both WES and WGS were performed (n =4, 325 PT/R1 and 579 PT/R1), the sequence capture had limited influence on read distributions with a Pearson's correlation coefficient of 0.80, 0.81 and 0.81 for mutations presenting a coverage depth of 20X, 30X and 40X, respectively (S2 Fig). A mean number of 145 (range 5-1,852) and 891 (range 8-9,985) coding somatic mutations (including non-synonymous SNVs, in-frame and frameshift indels as well as mutations located in splicing sites) were identified per primary tumor (PT) and first relapse event, respectively (S3 Table), of which 68% and 52% (mean values) were considered as subclonal (VAFs adjusted for tumor purity <0.30, S3 Table). While a mean of 1,168 mutations (range 7-9,980) were event-specific (S3 Table), on average, 26 and 46% of coding mutations were persistent between PT/R1 and R1/R2, respectively, with ancestral mutations maintained during tumor evolution (S3 Table). All patients presented shared mutations between the PT and subsequent events, suggesting a common ancestral origin of lymphoblasts as previously reported [13,18].

The number of coding mutations, progressively increased at each relapse event. For example, a total number of 36, 48, 60, 911 and 963 mutations were identified in PT, R1, R2, R3 and R4 of the case 325 (S3 Fig and S3 Table). Non-synonymous SNVs, including missense and nonsense variants, were the prevalent class of mutation in all events and for all cases (S3 Fig). No significant differences in the distribution of mutation classes were observed for different subtype of cases (i.e. one vs. multiple events, survival vs. death, early vs. late event and non-refractory vs. refractory (S3 Fig)). Five relapse events (325 R3, 579 R2, 684 R2, 772 R2 and 808 R1) carried more than 10 times the number of coding mutations identified in the precedent event and were considered as hypermutable (S3 Table). All tumors presenting these particular events were either refractory or suffered multiple relapse events (Table 1). The hypermutator phenotype in patient 325 was likely explained by the acquisition at R3 of a homozygous loss of function mutation in the mismatch repair endonuclease PMS2 (p.R134*, VAF =0.77) which presented a modest positive drift at R4 (VAF =0.85, Figure 1 and S1 Table). Of note, hypermutated cases 684, 772 and 808 were the only patients of this cohort to harbor a p.G39E germline polymorphism (rs1042821) in the DNA repair gene *MSH6*, a known biomarker of Lynch syndrome [19-21] that is associated with an increased risk of several cancer types [22]. Case 808 also carried the rs10254120 polymorphism (p.R20Q) in *PMS2* previously associated with reduced apoptotic function [23]. The mutation spectra of the 5 hypermutable cases revealed an important increase of transition mutations (A(T)>G(C) and C(G)>T(A)) from PT (mean percentage of total spectrum =29%) to the hypermutated event (72%) (S4 Fig), further highlighting a defective DNA repair mechanism. Except for these particular cases, no significant shift in mutation number was observed (S3 Table).

Putative driver mutations were grouped into relevant signaling (Methods, Figure 1, S1 Table and S2 Table) and treatment pathways (Methods, S5 Fig, S1 Table and S2 Table). Six signaling pathways were identified as recurrent targets of somatic mutations: Epigenetic regulation (n =18 mutated genes), Hemopoiesis - Immune system development (n =9), MAP kinase - RAS signaling (n =7), DNA repair (n =6), Transcription regulation (n =7) and Cell cycle (n =5) (Figure 1 and S1 Table). Epigenetic regulation and MAP kinase - RAS signaling were the pathways showing the highest mutation rates with, overall, 84.2 and 78.9% of patients presenting at least one altered gene respectively. This included well-known relapse driver genes such as *CREBBP* (n =5 cases mutated), *WHSC1* (n =5), *KRAS* (n =8) and *NRAS* (n =6) [8-10,13]. Interestingly, epigenetic regulation and MAP kinase - RAS signaling showed numerous putative driver events yet counter-selected at relapse such as *KDM6A* p.R519*, *ARID1A* p.L1831V, *KRAS* p.G12D and p.K117N, *NRAS* p.G12S, p.G12A, p.G12D, p.G13D and p.Q61K and *FLT3* p.D835Y (13/47 mutations vs. 3/36 in other pathways, Figure 1 and S1 Table). Other pathways presented a limited number of somatic mutations but mainly selected at relapse (including mutations that positively shifted and relapse-specific mutations, Figure 1 and S1 Table). Of note, altered *NT5C2* was limited to one case only (S5 Fig and S1 Table), which is contradictory with previous reports identifying this gene as frequently mutated in relapsed cALL [11,24]. A mean of 3.0 relapse-specific mutations (range 0-16) was identified per case and while each pathway showed a significant proportion of relapse-specific mutations (range 22.2-75.0%), most mutations identified in MAP kinase - RAS signaling were already present at diagnosis (8.7% - 2/23 relapse-specific mutations, Figure 1 and S1 Table). Multisubclonal hits in 5 genes (*KDM6A*, *SETD2*, *NRAS, NR3C1* and *TP53*) were identified in 4 early relapse cases (325, 382, 447 and 772) and one late relapse patient (659) (Figure 1 and S1 Table). Apart from the mutations in *NR3C1* (p.D724E and p.Y641H) and *KDM6A* (p.R1351*) selected at the expense of co-occurring mutations (p.P626S in *NR3C1* and p.R519* in *KDM6A*), all multisubclonal mutations either persisted (*TP53* p.R273H, p.R273C and p.Y236C in case 325) or underwent counter-selection in following events (*NRAS* p.G12S and p.G13D in case 447, *NRAS* p.G12A and p.G12D in case 659, Figure 1 and S1 Table).

Early and late relapse cases showed comparable distributions of mutations in the pathways of interest. An overall limited excess of early relapse cases were identified as mutated for four of the six pathways: 100 vs. 72.7% of mutated cases for the epigenetic regulation, 75.0 vs. 54.5% for hemopoiesis - immune system development, 87.5 vs. 72.7 for MAP kinase - RAS signaling and 62.5 vs. 36.4% for cell cycle. A strong yet non-significant effect was observed for the DNA repair signaling with 62.5% of early relapse cases mutated compared to only 18.2% of patients with late relapses (odds =6.6, p-Value =0.07, two tailed Fisher's Exact test). Of note, 4 of the 5 early relapse cases that harboured DNA repair pathway mutations carried these mutations within the dominant clonal population. In the late relapse group, one of the 2 mutations identified in this pathway was subclonal.

5.5.2. Early and late cALL relapses show different clonal dynamics

To evaluate the influence of tumor plasticity on outcome, we divided our patients into groups according to their progression-free survival (early vs. late event), their response to treatment (non-refractory vs. refractory) and their survival status. A “dynamic mutation rate” (number of dynamic mutation per Mb per day), considering only mutations showing significant variant allele frequency (VAF) shifts, was calculated for each relapse (Methods). Each value was representative of the clonal dynamics of a tumor between two events (either PT vs. R1 or R(n) vs. R(n+1)) (Figure 2). Based on this method, we identified marked reduction in clonal dynamics in late relapses (>36 months post diagnosis) compared to early relapse events (p =0.0094, WES data, Mann–Whitney U test) and for non-refractory compared to refractory events (p =0.0473) (Figure 2). Regarding early vs. late relapses, the same pattern was observed using WGS data (p =0.0079, S6 Fig). Of note, the absolute number of somatic mutations was not significantly different between early and late relapse groups (p =0.8687 and 1 for primary tumors and relapses respectively, Mann–Whitney U test). These results illustrate the plasticity of the earliest events, with rapid clonal switches over short periods of time (e.g. case 808 with 1.4E-01 mut/Mb/day or case 325 with 4 relapse events presenting a mean dynamic mutation rate of 2.3E-02 mut/Mb/day), while certain late events showed very long-term and quasi-inert relapses (e.g. case 34 with 1.3E-04 dynamic mutations per MB per day). Interestingly, a similar trend of reduced clonal dynamics was observed in WGS data for relapse events of patients who survived compared to those who did not (p =0.0159, S6 Fig). We also observed a progressive increase of dynamics at each new relapse event for the same patient (case 325, Figure 2). To investigate the impact of clonal dynamics on event free survival from diagnosis (PT to R1), we separated our cases into two groups of low and high clonal dynamics based on the median dynamic mutation rate value. Cases with higher clonal dynamics had significantly shorter event free survival (p =0.0067, Chi-square test, Figure 3).

Based on the assumption that subclonal diversity plays a central role in therapeutic failure and relapse [6,15,25], we also quantified the number of newly acquired subclonal mutations at each event. The calculated “subclonal mutation rate”, representative of the general subclonal expansion at each event, revealed the exact same pattern as reported above with significantly reduced subclonal expansion in late events compared to early events (p =0.0053 and p =0.0317 for WES and WGS data, respectively, S7 Fig and S8 Fig) and in non-refractory compared to refractory events (p =0.0022, WES data, S7 Fig). Again, cases who acquired a larger of subclonal mutations were also the ones who did not survive (p =0.0159, WGS data, S8 Fig).

5.5.3. Clonal evolution of early and late cALL relapses

We observed extreme clonal dynamics phenotypes. While certain early relapse cases showed multiple rapid and important switches of predominant clonal populations between events, some late relapse patients retained a clonal architecture that was similar to the structure of their PT. Based on the analysis of somatic VAFs in loss of heterozygosity (LOH)-free regions of the genome [26], we inferred the number of subclones and tracked their dynamic over the time (Figure 4). We also constructed clonal trees to determine the ancestral relationships between somatic mutations (Figure 4) [27]. We illustrated these contrasting clonal evolution phenotypes using three particular cases presenting either a late (34) or an early event (325 and 62, Figure 4).

Both patients 34 and 325 were males diagnosed with a normal 46,XY karyotype pre-B ALL at >10 years of age, and were therefore classified as “high risk” by the clinic. Patient 62 was a male diagnosed with a hyperdiploid karyotype (54,XY,+4,+6,+8,+10,+14,+17,+18,+21) at 29 months old and was considered as presenting a “standard” risk. Although case 34 presented the longest period before relapse (close to 6 years, Table 1), he also showed the slowest evolving tumor of our cohort (Figure 3). He achieved complete remission (bone marrow blasts <5%) after treatment of both PT and relapse. As for case 325, one of the patients presenting a hypermutator phenotype, he presented a first and early relapse event less than a year after PT (day 306) followed by multiple relapse events (R1 to R4) at days 412, 721 and 762 post-diagnosis (Table 1) and did not survive the last relapse event. With a relapse occurring 1,025 days after PT, case 62 was the last patient of the early event group (Table 1). He showed the slowest evolving tumor of this group (S6 Fig) and presented an overall number of mutation comparable to patient 34 (S3 Table). He also achieved complete remission after the second therapy.

Case 34 showed an almost perfect clonal equilibrium with the exact same tumor architecture reoccurring 6 years post-diagnosis (Figure 4A and 4B). Six clusters of clones were detected. An ancestral clone (clone 1), containing the driver mutation *NRAS* p.G12D, was predominant at PT. Clone 2, descendant of clone 1, harbored a nonsense mutation (p.W2006*) in the tumor suppressor gene *MLL2/KMT2D* and became predominant at relapse (VAFPT =0.31, VAFR1 =0.55). This was the only dynamics observed between the two events. A very subclonal mutation in the oncogene *KRAS* was detected at diagnosis and rose again at late relapse with a similar frequency (VAFPT =0.03, VAFR1 =0.05), further highlighting the inertia of this tumor. Clonal evolution in this patient was reminiscent of an escape from long-term dormancy of pre-malignant neoplastic stem cells that also initiated the primary leukemia. While this patient was the most extreme example of long-term clonal equilibrium and the only case that presented the same clonal architecture at diagnosis and relapse, all late relapses presented common mutations with primary tumors (S3 Table).

On the other hand, with more than 25 SNVs/indels identified in tumor suppressors, oncogenes and treatment related genes, case 325 presented a complex clonal architecture composed of 9 subclones spatially evolving over the 5 sequenced time-points (Table 1 and Figure 4C). The PT architecture showed a very close structure to that observed at R1 with a persistent ancestral subclone (clone 4) and a similar distribution of subclones (Figure 4C and 4D). Notably, in clone 4, we identified a MAKP pathway activating mutation p.A72D in the SH2 domain of the oncogene *PTPN11* [28] that was subclonal at diagnosis (VAF =0.23), selected at R1 (VAF =0.49) and remained dominant at the 3 subsequent time-points (VAF =0.52, 0.51 and 0.47 in R2, R3 and R4, respectively). We also observed the loss of several subclonal mutations in driver genes in R1 that were counter-selected after treatment such as *KRAS* p.K117N (VAFPT =0.16 and VAFR1 =0.003) and *SETD2* p.R456* (VAFPT =0.24 and VAFR1 =0). A missense mutation in the glucocorticoid receptor NR3C1 (p.P626S) that emerged at R1 (VAF =0.22) and further expanded at R2 to become predominant (VAF =0.42). A mutation of this proline residue, located in the hydrophobic core after the receptor dimerization, was previously reported to decrease the transactivation by glucocorticoids by 95% [29]. However, *NR3C1* p.P626S was counter-selected at R3 (VAF =0), during the most important changes in this tumor history, and was replaced by two co-occurring mutations in *NR3C1* (p.Y641H, VAF =0.50 and p.D724E, VAF =0.53) that were also located in the steroidbinding domain and limited to the newly dominant clone 5. This clone, descendant of ancestral clone 4, acquired an homozygous loss of function mutation p.R134* in the DNA repair gene *PMS2* likely causing the hypermutable phenotype of R3, which presented over 15 times more somatic mutations than R2 (S3 Table). Clone 5 also carried several mutated driver genes providing a fitness advantage to the cells such as *CREBBP* (p.Y1503H), *IKZF1* (p.F173L), *MLL3/KMT2C* (p.R2609*) and *SRSF2* (p.G12S). This led to the expansion of this clone at R3 which constituted the predominant population at both R3 and R4. Multiple minor subclones, descendant of clone 5 and harboring mutations in *TP53* (p.R273H, p.R273C and p.Y236C), *MSH6* (p.F573L) and *WHSC1* (p.A457T), also emerged at R3 and were maintained as minor clones at R4. While case 325 did not present the highest evolutionary dynamics within this cohort (Figure 2), these multiple relapses in a short period of time illustrated the step-by-step replacement of the dominant clonal population to the benefit of a fitter clonal progeny under particular selective pressures such as chemotherapy.

Finally, case 62 showed an intermediate phenotype with almost 50% of the mutations identified at diagnosis that persisted at relapse (S3 Table). Four clusters were identified (Figure 4E and 4F). The dominant population at relapse, probably pulled up by the driver mutation *KRAS* p.G12D (VAFPT =0.39, VAFR1 =0.55), harbored a series of newly acquired mutations partly belonging to cluster 3. Of note, a subclonal mutation (VAFPT =0.10) predicted to alter splicing in the HAT domain of the histone acetyl transferase CREBBP was identified. While the alteration of this domain is known to cause glucocorticoid resistance in leukemia [9,10], the mutation was lost at relapse (VAFR1 =0). Overall, although modest in comparison with other patients of the early relapse group, the clonal evolution pattern of patient 62 was still clearly distinguishable from the late relapse cases.

5.6. Discussion

We have performed a genomic analysis to build a comprehensive catalogue of somatic variations and to study the relapse evolutionary trajectory of 19 pre-B cALL patients presenting early and late relapses. In addition to the unique serial sampling design (up to five), to our knowledge, this study is the first to capture two clonal dynamics governing late and early relapses through the study of somatic allelic frequencies. The main limitation of this study is the depth of coverage; sequencing whole exomes and genomes as opposed to sequencing a panel of pre-selected genes was useful to avoid bias however it may have precluded full dissection of the subclonal architecture.

Five of the six pathways identified as frequently targeted in our cALL cohort (Epigenetic regulation, Hemopoiesis - Immune system development, MAP kinase - RAS signaling, DNA repair and Cell cycle) confirmed their relapse driving potential [13]. While the JAK-STAT/Hemopoiesis *SH2B3* gene was found mutated in 2 cases (Figure 1 and S1 Table), the recently reported enrichment of mutations in the JAK-STAT pathway in relapsed cALL patients [13] was not observed in our study. With the rise and fall of numerous somatic mutations, epigenetic regulation and MAP kinase - RAS were the two most mutated (84.2 and 78.9% of mutated patients, respectively) and dynamic signaling pathways. They perfectly illustrated the clonal competition for space and resources that occur in each tumor, particularly under external selective pressures such as chemotherapy. For example, in case 325 we identified a p.A72D mutation in *PTPN11* (MAP kinase - RAS pathway) contained in the predominant clone at diagnosis. This clone persisted all along tumor history despite the occurrence of several subclonal driver mutations in *KRAS* (p.K117N) or *HRAS* (p.F156L) that did not succeed to take over the major population. On the other hand, a mutational turnover was observed in case 579 where the dominant cell population harboring a p.D835Y *FLT3* at diagnosis (VAF =0.44) was progressively replaced by a subclone containing a p.G13D mutation in *NRAS* at R1 (VAF =0.16, p.D835Y *FLT3* - VAF =0.30) which became the dominant clone at R2 (p.G13D *NRAS* - VAF =0.44, p.D835Y *FLT3* - VAF =0) with no other mutation in the MAP kinase - RAS pathway. We also identified a mutational turnover of the glucocorticoid receptor *NR3C1* with a succession and a co-occurrence of missense mutations (p.P626S, p.Y641H and p.D724E) in the dominant clone at R2, R3 and R4 of the case 325. Despite its limitation to a single case in this cohort and the need for further mechanistic characterization, the identified reduction of transactivation by glucocorticoids due to a mutation of the residue P626 in CV-1 cells [29] suggests a possible involvement in resistance to therapy and could explain the selection of mutations here.

Interestingly, we found an overrepresentation of cases from the early relapse group with mutation in DNA repair genes (62.5% vs. 18.2% of late relapse group). Furthermore, almost only early relapses presented dominant mutations in this pathway. While this effect requires validations in larger cohorts, a defect in DNA repair process is in line with the fast evolutionary adaptability of most early relapses. We also identified 5 cases presenting a hypermutator phenotype of which, 3 belonged to the early relapse group and 2 to the late relapse group. The limited occurrence of this hypermutator phenotype (26% of the cohort) suggests that a strong accumulation of somatic mutations is not a prerequisite mechanism for the acquisition of resistance, however it obviously increases the odds of acquisition of advantageous mutations and allows a higher tumor plasticity, likely giving rise to highly resistant clones. Two such patients were refractory to therapy and none survived. Of note, all but one hypermutable event were at least second relapses. Only the refractory case 808 presented this phenotype at R1. Loss of function of the DNA repair gene *PMS2*, found mutated in case 325, was likely responsible for this phenotype. This confirmed the recently identified hypermutator potential of the loss of function of the mismatch repair endonuclease PMS2 [13], that until now was mostly known to cause Lynch syndrome via autosomal dominant inheritance of germline mutations [30].

We distinguished two general evolution patterns: a highly dynamic/adaptative clonal pattern specific to early relapses - illustrating the quick emergence of fitter clones and the eradication of other subclones - and a slow to quasi-inert evolution pattern associated to late events. By quantifying both the general clonal dynamics and the newly acquired subclonal diversity, we successfully captured the underlying differences of resistance mechanisms between these two types of pre-B ALL relapse. In the most plastic tumors, the accumulation of somatic mutations allows a rapid subclonal diversification and increases the chance of acquisition of resistant mutations. It leads to a post-chemotherapy turnover of predominant populations and to the selection of the fitter clone. In line with previous findings associating shorter remissions with the early presence of subclonal driver mutations in leukemia samples [15,24] or with the level of minimal residual disease levels on day 19 of remission induction [31], these results demonstrated that rapid subclonal expansion and high tumor plasticity are key determinants of a rapid cALL evolution and predictive factors of early therapeutic escapes.

The significantly reduced dynamics of late events suggests a different mechanism of relapse, less likely relying on a mutation-based chemoresistance. While little is known regarding cell-of-origin and process of long-term cALL relapses, recent studies failed to identify shared mutations (single-nucleotide and copy number variants) between diagnosis and late relapse [32] and have suggested that late relapse cells likely derive from stem cells of a minor clone at diagnosis [32,33]. However, all late events analysed here harbored common somatic alterations (SNVs and indels) with diagnosis. Furthermore, some of these cases presented a majority of mutations with similar allelic frequencies in both primary and relapse leukemias, including clonal mutations. This implies that late relapses can derive from progeny of the predominant cell population at diagnosis. Overall, these results confirm a relapse mechanism based on the re-activation/re-expansion of dormant pre-malignant neoplastic stem cells spared from initial cytoreductive therapy because of their non-proliferative state rather than because of a genetic resistance. Chemosensitive primary tumors that showed no acquisition of resistant mutations at relapse, such as case 34, were also sensitive to second induction, further illustrating this therapeutic escape. In this case, additional investigations are needed to understand how the original clonal architecture was reproduced 6 years following the primary tumor.

Though multiple factors including infection [34] have been proposed to explain the origin and condition of long-term re-emergence of quiescent or dormant cancer cells leading to late relapse in cALL, deciphering the underlying mechanisms remains a challenge of great clinical importance.

5.7. Materials and Methods

5.7.1. Study subjects

All study subjects were French-Canadians of European descent. Incident cases were diagnosed in the Division of Hematology-Oncology at the Sainte-Justine Hospital (Montreal, Canada) as part of the Quebec childhood ALL cohort (QcALL) [35]. This childhood B-ALL cohort (n=19) consisted of twelve males and seven females, with a mean age at diagnosis of 6 years. Eight patients experienced an early relapse (<36 months post-diagnosis) after a median time of 16.7 months post-induction and eleven suffered a late event after a median time of 47.2 months. Overall, seven patients experienced multi-relapse events. Four cases were refractory to the first or the second chemotherapy induction (show no to limited reduction of tumor cells) and were resampled at a median time of 142 days after diagnosis or relapse. Twelve patients did not survive.

5.7.2. Whole-exome/genome sequencingand variant identification

Whole exome or genome sequencing were performed on 26 matched normal-primary tumor-relapse/refractory material. DNA was extracted from bone marrow samples (at diagnosis and relapse) and peripheral blood samples (after remission or samples without blast cells for refractory cases) using standard protocols [36]. Whole exomes of twelve cases (34, 64, 325, 382, 394, 445, 579, 684, 717, 772, 786 and 808) were captured in solution with Nextera Rapid Capture Exome Enrichment kits according to the manufacturer's protocol and sequenced on HiSeq 2000/2500 (paired-end: 100x100 bp, mean coverage on targeted region = 200X). Exome reads obtained from HiSeq 2500 systems were aligned to the Hg19 reference genome using Bowtie2 (version 2.2.3) [37]. Whole genomes of nine cases (62, 217, 325, 391, 447, 579, 659, 690 and 764) were sequenced by Illumina, Inc. on HiSeq 2000 (paired-end: 100x100 bp, mean coverage on targeted region =80X); resulting reads were aligned to the Hg19 reference genome using the Illumina Casava software. PCR duplicates were removed using Picard [38]. Genotype quality score recalibration was performed using the Genome Analysis ToolKit (GATK) [39]. Sequencing metrics were obtained using the *DepthOfCoverage* option in GATK. SNVs and small indels were called from Bam files using Strelka [40] (S1 Fig). CNVs were called using Varscan2 *copyCaller* option [41] (S1 Fig). If necessary, log R ratio distribution of copy neutral regions was recentered on 0 before the calling step. Varscan2 raw output was smoothed and segmented using the DNAcopy library of the BioConductor project. Putative CNVs were manually reviewed and checked against calls made from WGG, when available. Based on this comparison, filtering parameters (minimum read depth, lower and upper bound for log ratio to call amplification and deletion) were adjusted and extended to the whole data.

5.7.3. Whole genome high-density SNV genotyping and copy number variant identification

Normal, primary tumor and relapse of nine cases (62, 217, 325, 391, 447, 579, 659, 690 and 764) were genotyped using Illumina's HumanOmni 2.5-Quad or HumanOmni2.5-Octo SNP bead arrays (McGill University and Genome Quebec Innovation Centre, Montreal, Quebec). Extracted genomic DNA was processed according to the Illumina Infinium HD Assay Ultra protocol. BeadChips were imaged on Illumina's iScan System with iScan Control Software (v3.2.45). The Genotyping Module (Version 1.9.4) of the Illumina GenomeStudio software (V2011.1) was used for raw data normalization, genotype clustering and calling, with default. ASCAT version 2.2 [42] was used to evaluate the sample purity, to evaluate tumor ploidy and to identify tumor-specific copy number variants (CNVs) or copy-neutral loss of heterozygosity (LOH) (S1 Fig).

5.7.4. Variant annotation and prioritization of cancer driver gene mutations

ANNOVAR (version 2015Jun17) [43] and Oncotator (version 1.5.3.0) [44] were used to annotate somatic splice site variants, non-synonymous SNVs and frameshift small indels (S1 Table). Variants were queried against publically available datasets such as 1000 Genomes [45] and NHLBI GO Exome Sequencing Project (ESP) [46] to filter out common polymorphisms (minor allele frequency >0.01). We classified each mutated gene as either tumor suppressor genes (TSGs) or oncogenes based on Vogelstein's 20/20 rule [47] on COSMIC v72 data. The predicted functional impact of non-synonymous variants and small indels was assessed using Sift (version 1.03) [48], Polyphen2 (version 2.2.2) [49], MutationTaster 2 [50] (S1 Table) as previously described [51]. Finally, mutated genes were queried against relevant treatment pathways using the public pharmacogenomics database PharmGKB [52]. To identify candidate driver mutations, we filtered events and kept only somatic alterations that were: i) missense mutations predicted to be damaging by at least one of the prediction algorithms; or identified as recurrent in COSMIC v72; ii) splice site and nonsense mutations that were predicted to be damaging by MutationTaster and Sift, respectively; or located in a tumor suppressor gene; or identified as recurrent in COSMIC; iii) frameshift indels and deletions located in a TSG and gain of copies located in an oncogene; iv) located in recurrent targets (genes and pathways) or in relevant treatment pathways.

5.7.5. Spatial and temporal analysis of clonal populations

Clonal architecture was inferred from VAF measurements using Sciclone [25] with minimumDepth option =40 and 20X for WES and WGS, respectively (S1 Fig). To obtain a highest-confidence quantification of VAF and inference of clonality, CNVs were inputed to Sciclone which focuses on variants located in copy number aberration-free portions of the genome. The evolutionary history and phylogeny of tumors were constructed using AncesTree (“alpha” option ≤0.45) [26] (S1 Fig). Analysis was also conducted on copy-number neutral regions of the genome.

5.7.6. Dynamic and subclonal mutation rates

We measured the “dynamic mutation rate” at each event (n) compared to the previous one (n-1). We calculated the shift of allelic frequency (ΔVAF) between n and n-1 for each mutation (SNVs or indels) located in diploid regions, as determined by CNV analysis: $\Delta VAF = VAF(n)-VAF(n-1)$. To exclude slight clonal drift, we only considered a mutation as “dynamic” if it presented a VAF shift value that was at least 2 standard deviations away from the average shift obtained for the considered event (ΔVAF): $\Delta VAF > \Delta VAF+2\sigma$. A “dynamic mutation rate” per event was then calculated by reporting the number of determined dynamic mutations per megabase (Mb) of sequenced regions (coverage >40X for WES, >20X for WGS) and per day since the last event. The “subclonal mutation rate” was calculated by considering the accumulation of newly acquired subclonal mutations at each event. Mutations with a VAF(n-1) =0 and an adjusted VAF(n) <0.3 (corrected for tumor purity) were considered as new and subclonal. Again these mutations were reported to the number of covered megabases and days since the last event. Kaplan–Meier survival analysis were conducted using the “survival” R package.

5.8. Figures

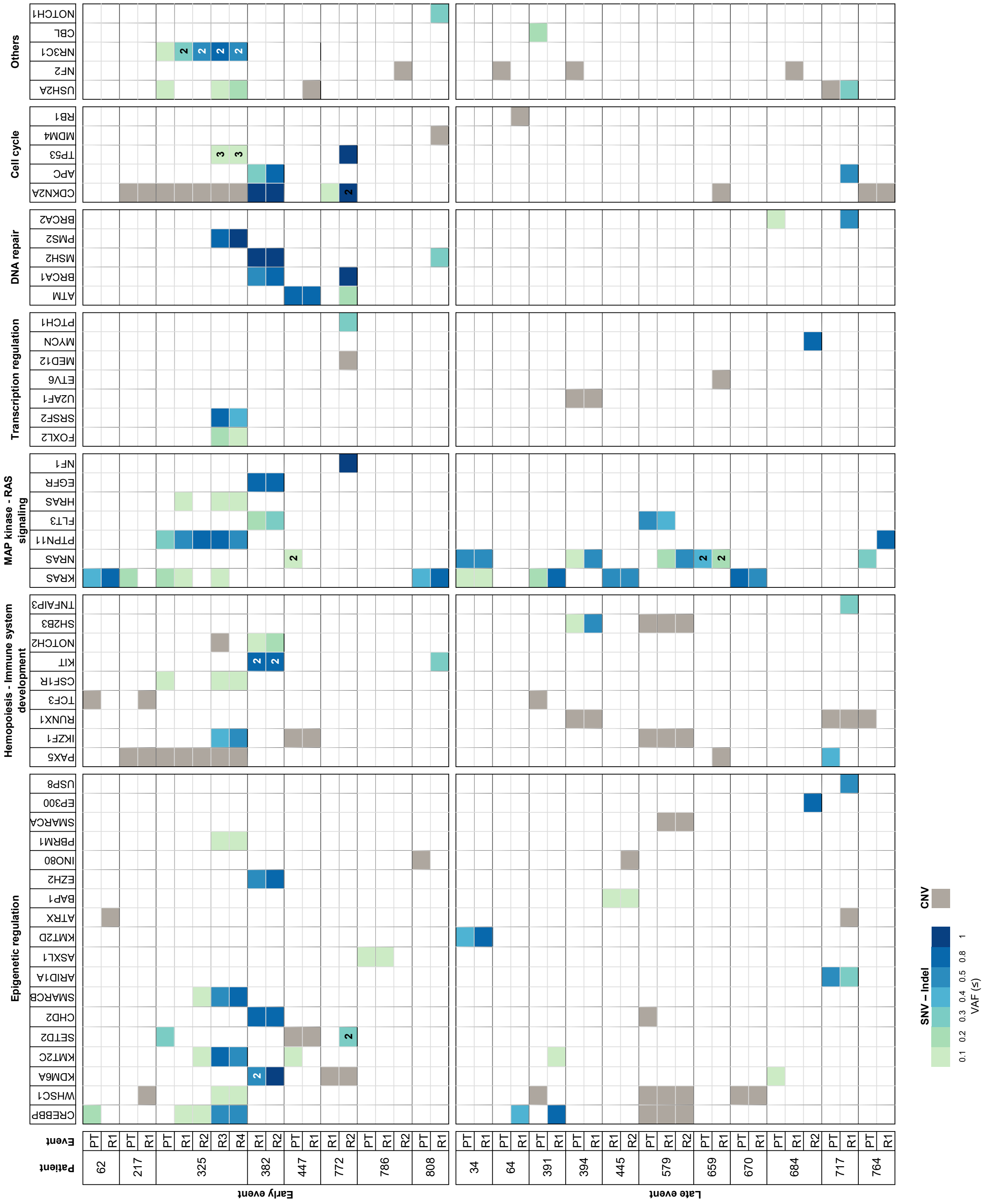

Others
NOTCH1
CBL
NR3C1
NF2
USH2A
Cell cycle
RB1
MDM4
TP53
APC
CDKN2A
DNA repair
BRCA2
PMS2
MSH2
BRCA1
ATM
Transcription regulation
PTCH1
MYCN
MED12
ETV6
U2AF1
SRSF2
FOXL2
MAP kinase - RAS signaling
NF1
EGFR
HRAS
FLT3
PTPN11
NRAS
KRAS
Hemopoiesis - Immune system development
TNFAIP3
SH2B3
NOTCH2
KIT
CSF1R
TCF3
RUNX1
IKZF1
PAX5
Epigenetic regulation
USP8
EP300
SMARCA
PBRM1
INO80
EZH2
BAP1
ATRX
KMT2D
ASXL1
ARID1A
SMARCB
CHD2
SETD2
KMT2C
KDM6A
WHSC1
CREBBP
Event
Patient
PT R1 PT R1 PT R1 R2 R3 R4 R1 R2 PT R1 R1 R2 PT R1 R2 PT R1
62 217 325 382 447 772 786 808
Early event
PT R1 PT R1 PT R1 PT R1 R1 R2 PT R1 R2 PT R1 PT R1 PT R1 R2 PT R1 PT R1
34 64 391 394 445 579 659 670 684 717 764
Late event
2
3
CNV
SNV – Indel
0.1 0.2 0.3 0.4 0.5 0.8 1
VAF (≤)

Figure 1. Overview of putative drivers (somatic SNVs, small indels and CNVs) identified among 19 childhood cALL patients and grouped into relevant signaling pathways. Genes (top) are ordered according to the number of events identified in the cohort (descending order from left to right). Pathways are ordered according to the number of genes identified as mutated (descending order from left to right). The color scale indicates the VAFs of SNVs and indels (from light green - VAF ≤0.1 to dark blue - VAF =1). CNVs are indicated in grey. Numbers in cells indicate the number of multiclonal mutations co-occurring in the same gene. PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse; SNV: single nucleotide variant; CNV: copy number variant; VAF: variant allele frequency.

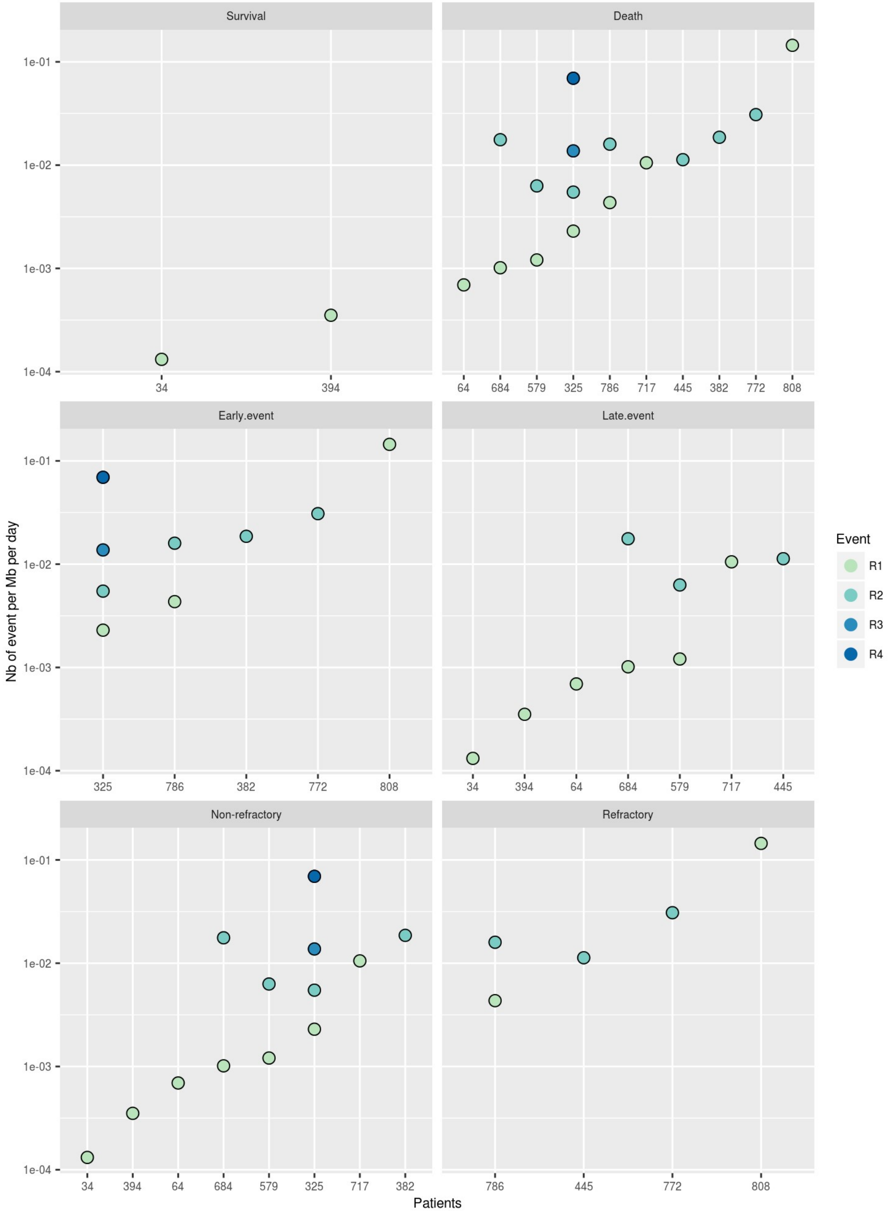


Figure 2. Early and late cALL relapses show different clonal dynamics. The "dynamic mutation rate" (number of dynamic mutation per Mb per day) was calculated from WES data by considering mutations showing a significant VAF shift only (Methods). Each value represents the clonal dynamics of a tumor between two events (either PT vs. R1 or R(n) vs. R(n+1)). Patients are divided into 6 panels depending on their survival or refractory status and on the delay before relapse events. R1 stands for the analysis of PT vs. R1, R2 for R1 vs. R2, R3 for R2 vs. R3 and R4 for R3 vs. R4. The dynamic mutation rate was significantly reduced in late relapses compared to early events (p =0.0094, Mann–Whitney U test) and in non-refractory compared to refractory events (p =0.0473, Mann–Whitney U test). PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse; WES: whole exome sequencing; VAF: variant allele frequency; PT: primary tumor; R: relapse.

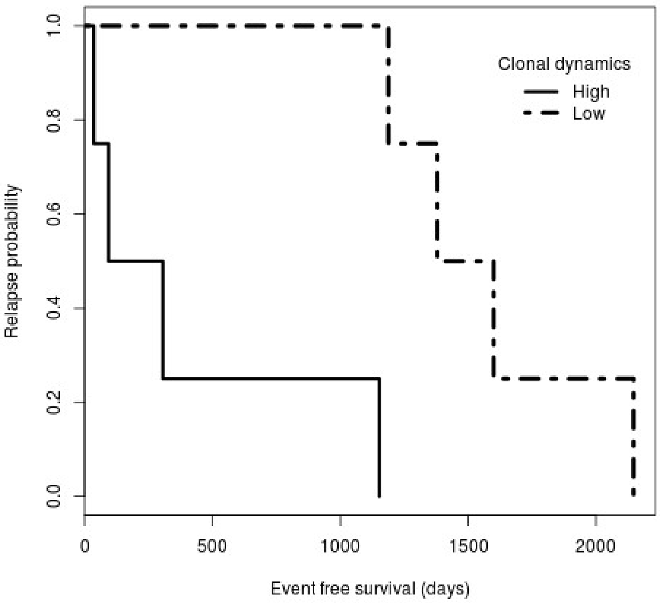
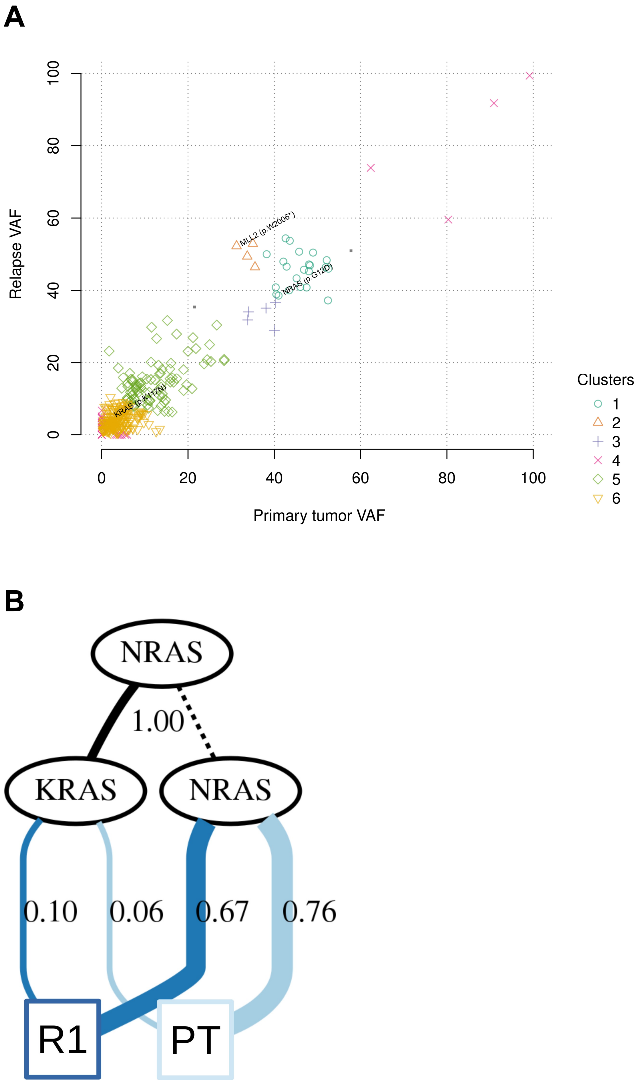
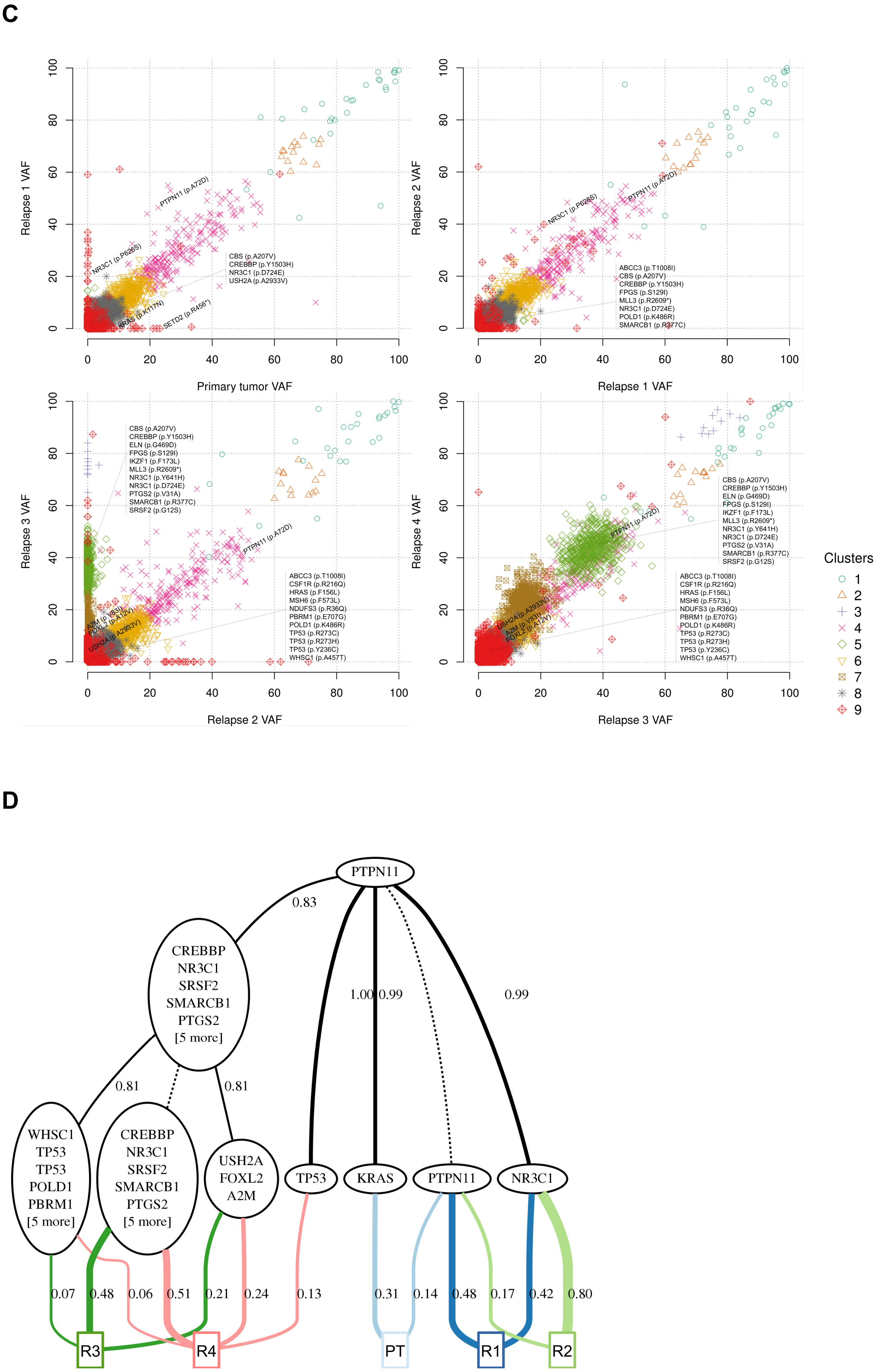


Figure 3. A higher clonal dynamics correlates with a shorter event free survival. Cases were divided into groups of low and high clonal dynamics based on the median dynamic mutation rate value. Kaplan-Meier curves estimating the relapse probability over time after the primary tumor were constructed according to the status of clonal dynamics. Solid and dashed lines stand for clonal dynamics over and under the median value, respectively. Cases with a higher clonal dynamics had a significantly shorter event free survival (p =0.0067, Chi-square test).


A
Relapse VAF
0
20
40
60
80
100
Primary tumor VAF
0
20
40
60
80
100
MLL2 (p.W2006*)
NRAS (p.G12D)
KRAS (p.K117N)
Clusters
○ 1
△ 2
+ 3
× 4
◇ 5
▽ 6
B
NRAS
1.00
KRAS
NRAS
0.10
0.06
0.67
0.76
R1
PT


C
Relapse 1 VAF
Primary tumor VAF
PTPN11 (p.A72D)
NR3C1 (p.P626S)
KRAS (p.K117N)
SETD2 (p.R456*)
CBS (p.A207V)
CREBBP (p.Y1503H)
NR3C1 (p.D724E)
USH2A (p.A2933V)
Relapse 2 VAF
Relapse 1 VAF
NR3C1 (p.P626S)
PTPN11 (p.A72D)
ABCC3 (p.T1008I)
CBS (p.A207V)
CREBBP (p.Y1503H)
FPGS (p.S129I)
MLL3 (p.R2609*)
NR3C1 (p.D724E)
POLD1 (p.K486R)
SMARCB1 (p.R377C)
Relapse 3 VAF
Relapse 2 VAF
CBS (p.A207V)
CREBBP (p.Y1503H)
ELN (p.G469D)
FPGS (p.S129I)
IKZF1 (p.F173L)
MLL3 (p.R2609*)
NR3C1 (p.Y641H)
NR3C1 (p.D724E)
PTGS2 (p.V31A)
SMARCB1 (p.R377C)
SRSF2 (p.G12S)
PTPN11 (p.A72D)
A2M (p.V83I)
FOXL2 (p.A12V)
USH2A (p.A2933V)
ABCC3 (p.T1008I)
CSF1R (p.R216Q)
HRAS (p.F156L)
MSH6 (p.F573L)
NDUFS3 (p.R36Q)
PBRM1 (p.E707G)
POLD1 (p.K486R)
TP53 (p.R273C)
TP53 (p.R273H)
TP53 (p.Y236C)
WHSC1 (p.A457T)
Relapse 4 VAF
Relapse 3 VAF
PTPN11 (p.A72D)
USH2A (p.A2933V)
A2M (p.V83I)
FOXL2 (p.A12V)
0
20
40
60
80
100
Clusters
○ 1
△ 2
+ 3
× 4
◇ 5
▽ 6
⊠ 7
✳ 8
⊕ 9
D
PTPN11
0.83
1.00
0.99
0.99
CREBBP
NR3C1
SRSF2
SMARCB1
PTGS2
[5 more]
0.81
0.81
WHSC1
TP53
TP53
POLD1
PBRM1
[5 more]
CREBBP
NR3C1
SRSF2
SMARCB1
PTGS2
[5 more]
USH2A
FOXL2
A2M
TP53
KRAS
PTPN11
NR3C1
0.07
0.48
0.06
0.51
0.21
0.24
0.13
0.31
0.14
0.48
0.17
0.42
0.80
R3
R4
PT
R1
R2

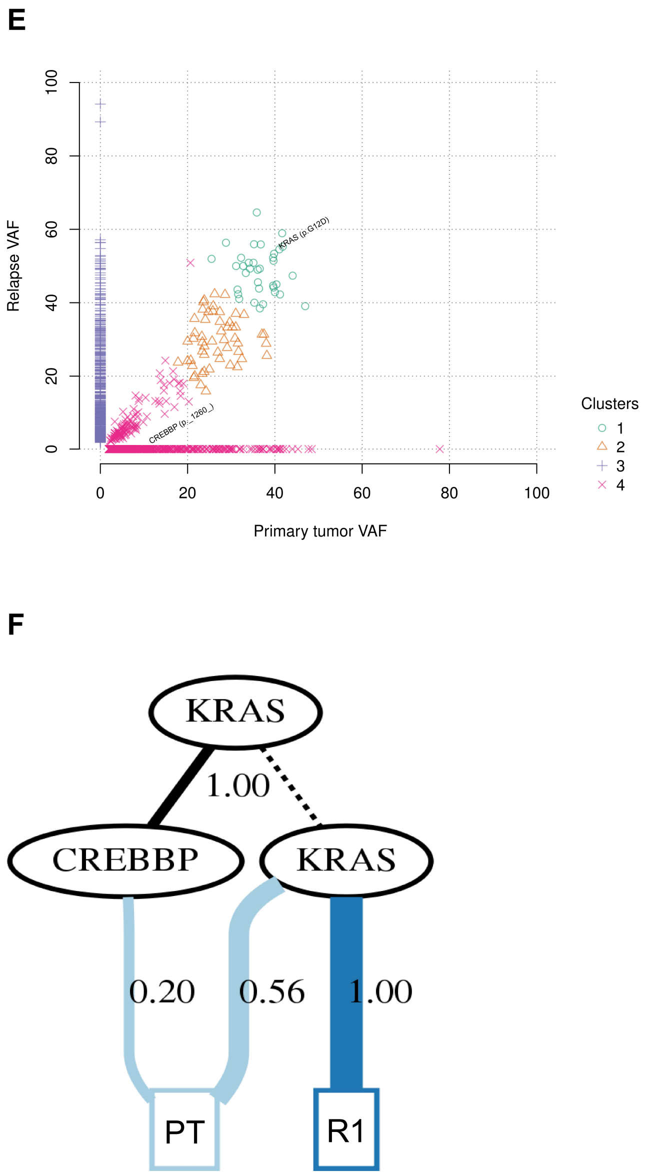
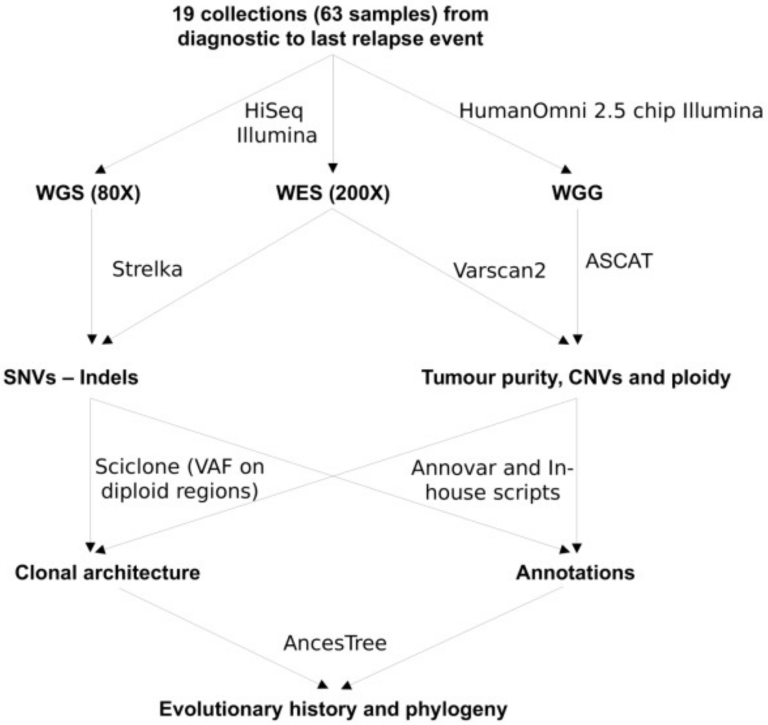
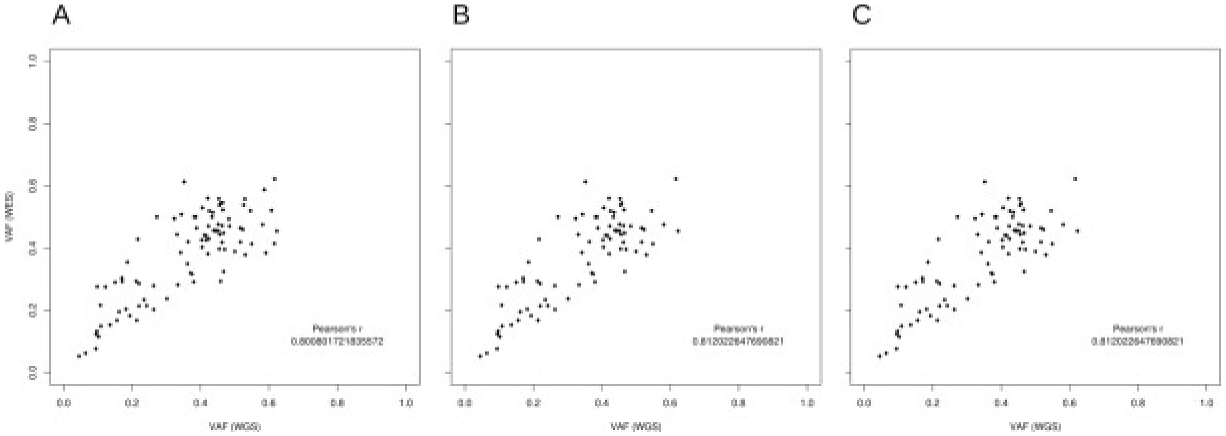


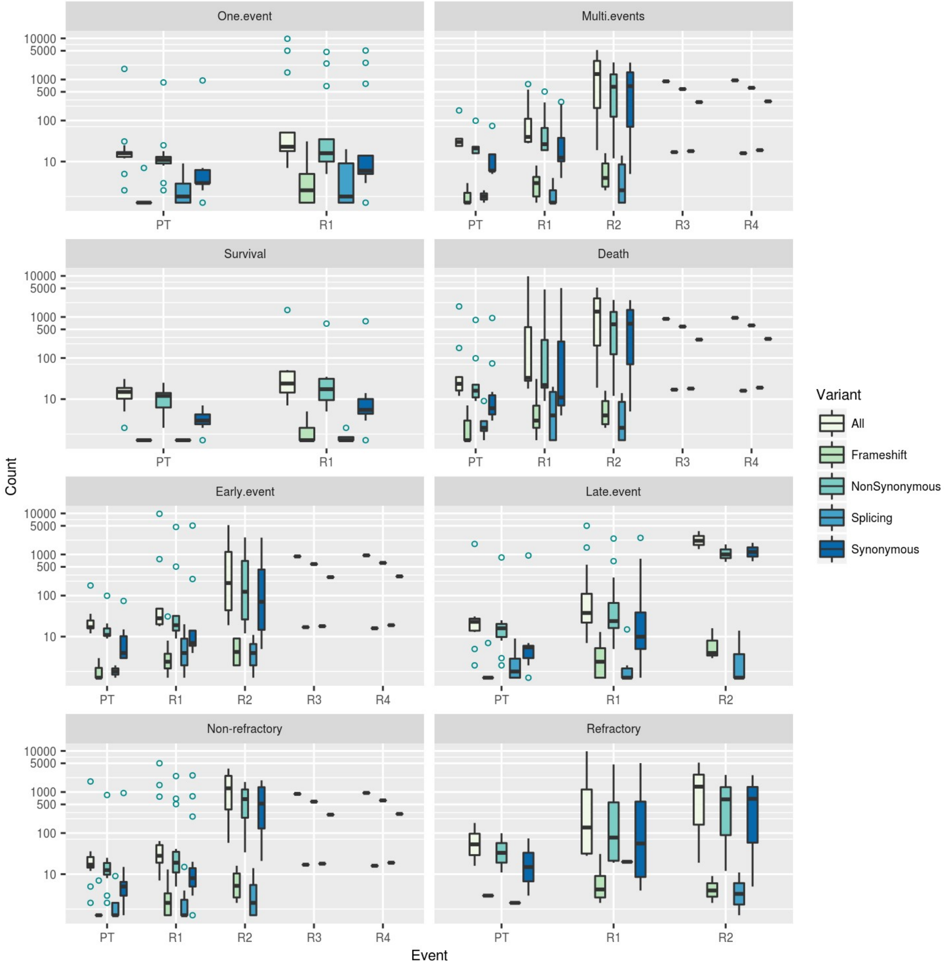
Figure 4. Clonal architectures and ancestral relationships of early and late relapses. Clusters were obtained using Sciclone (variational Bayesian mixture model)[25], for the patients 34 (**A**), 325 (**C**) and 62 (**E**), based on VAFs at PT (x axis) and R1 (y axis) or R(n) and R(n+1). Each dot stands for a mutation (SNVs or indels). Different clusters are depicted by different colors and dot shapes. Mutations of particular interest are annotated. Clonal trees depicting ancestral relationships between somatic mutations in patients 34 (**B**), 325 (**D**) and 62 (**F**) were inferred using AncesTree [26] and are represented by black solid lines. Posterior probabilities of the ancestral relationships are indicated on the side of the black lines. Dashed lines show ancestral clones which existed at the time of sequencing. Each analysed sample is indicated in a colored box at the bottom of the trees. Colored lines indicate the inferred composition of clones and their fraction in each sample. PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse; SNV: single nucleotide variant.


19 collections (63 samples) from diagnostic to last relapse event
HiSeq Illumina
HumanOmni 2.5 chip Illumina
WGS (80X)
WES (200X)
WGG
Strelka
Varscan2
ASCAT
SNVs – Indels
Tumour purity, CNVs and ploidy
Sciclone (VAF on diploid regions)
Annovar and In-house scripts
Clonal architecture
Annotations
AncesTree
Evolutionary history and phylogeny

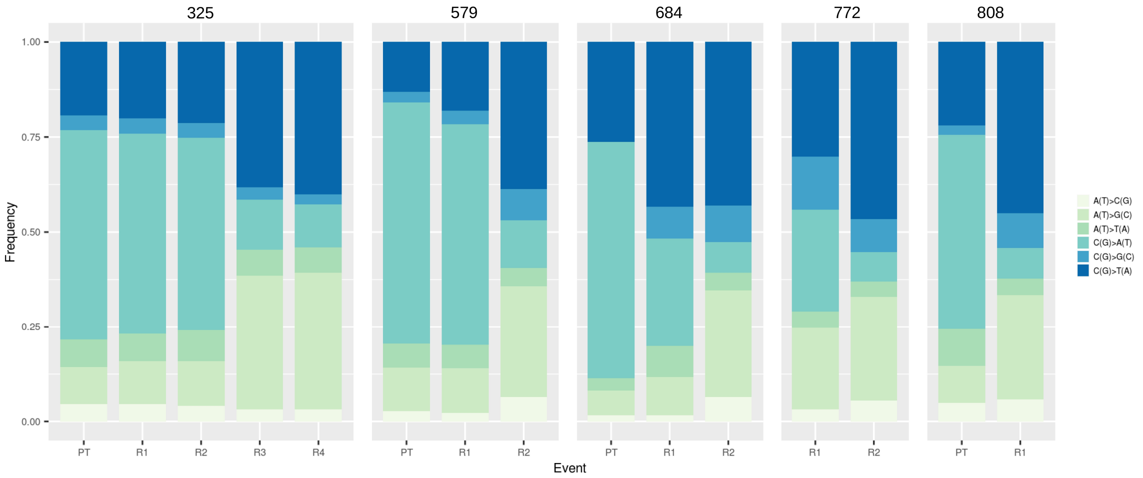
S1 Fig. Analysis workflow. SNV: single nucleotide variant; CNV: copy number variant; VAF: variant allele frequency; WES: whole-exome sequencing; WGS: whole-genome sequencing; WGG: whole-genome genotyping.



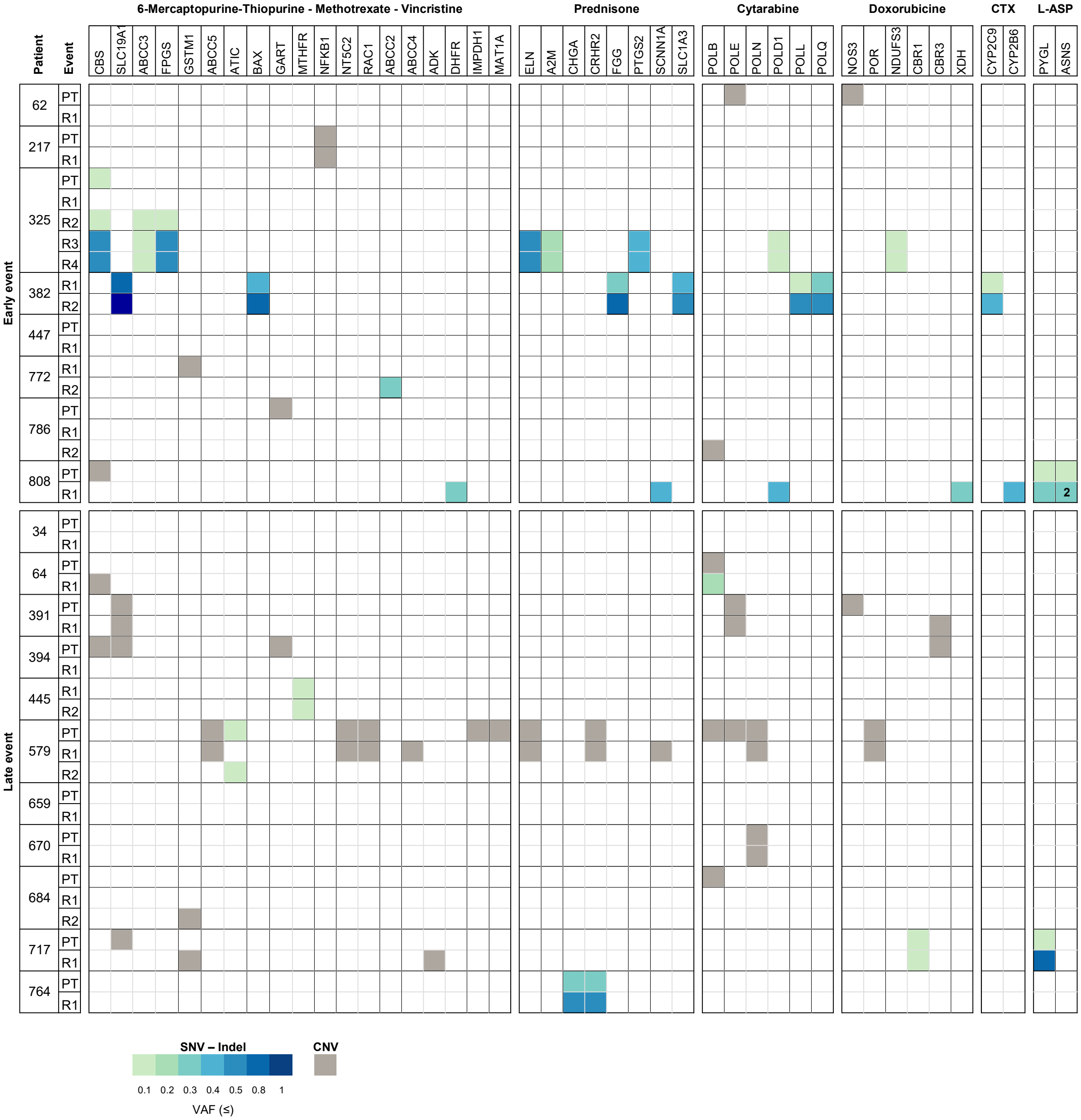
S2 Fig. Correlation of VAFs between WES and WGS. VAFs obtained from a series of samples for which both WES and WGS were performed (n =4, 325 PT/R1 and 579 PT/R1) were compared. We observed a Pearson's correlation coefficient of 0.80, 0.81 and 0.81 for mutations presenting a coverage depth of at least (**A**) 20X, (**B**) 30X and (**C**) 40X, respectively. VAF: variant allele frequency; WES: whole-exome sequencing; WGS: whole-genome sequencing.



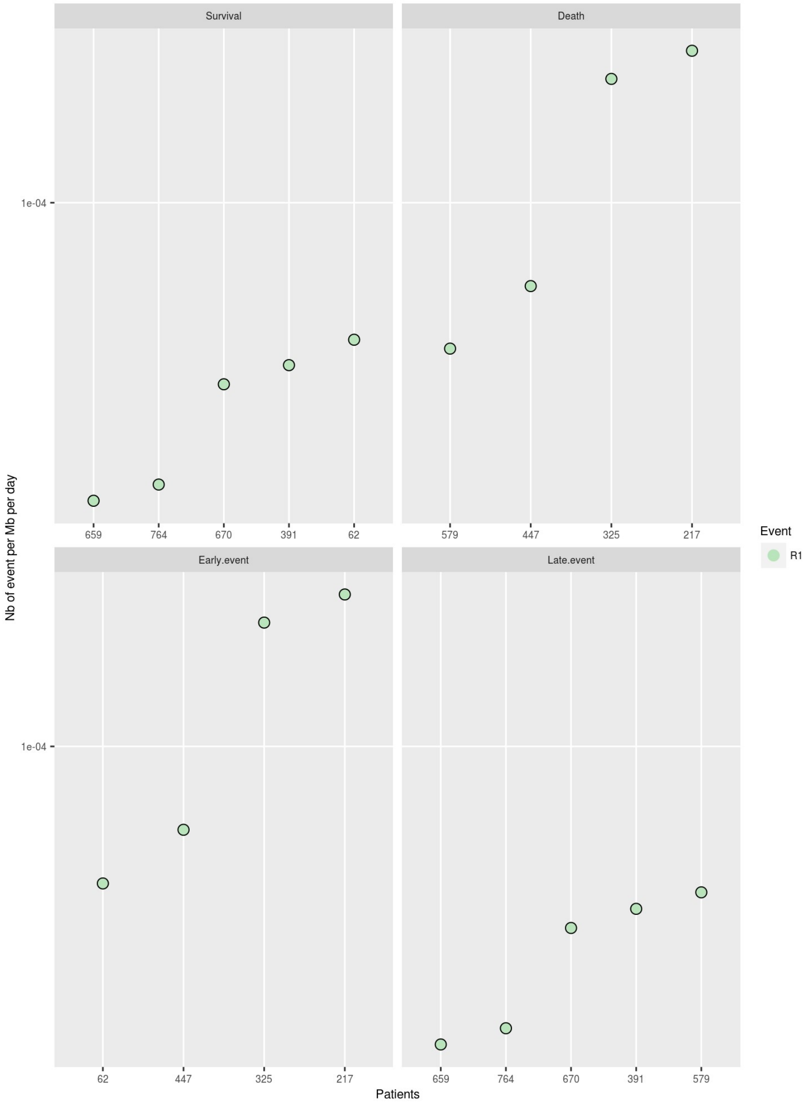
S3 Fig. Distribution of somatic mutations according to their type. Patients are divided into 6 panels depending on their survival or refractory status and on the delay before relapse events. Blue circles stand for outliers. PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse.



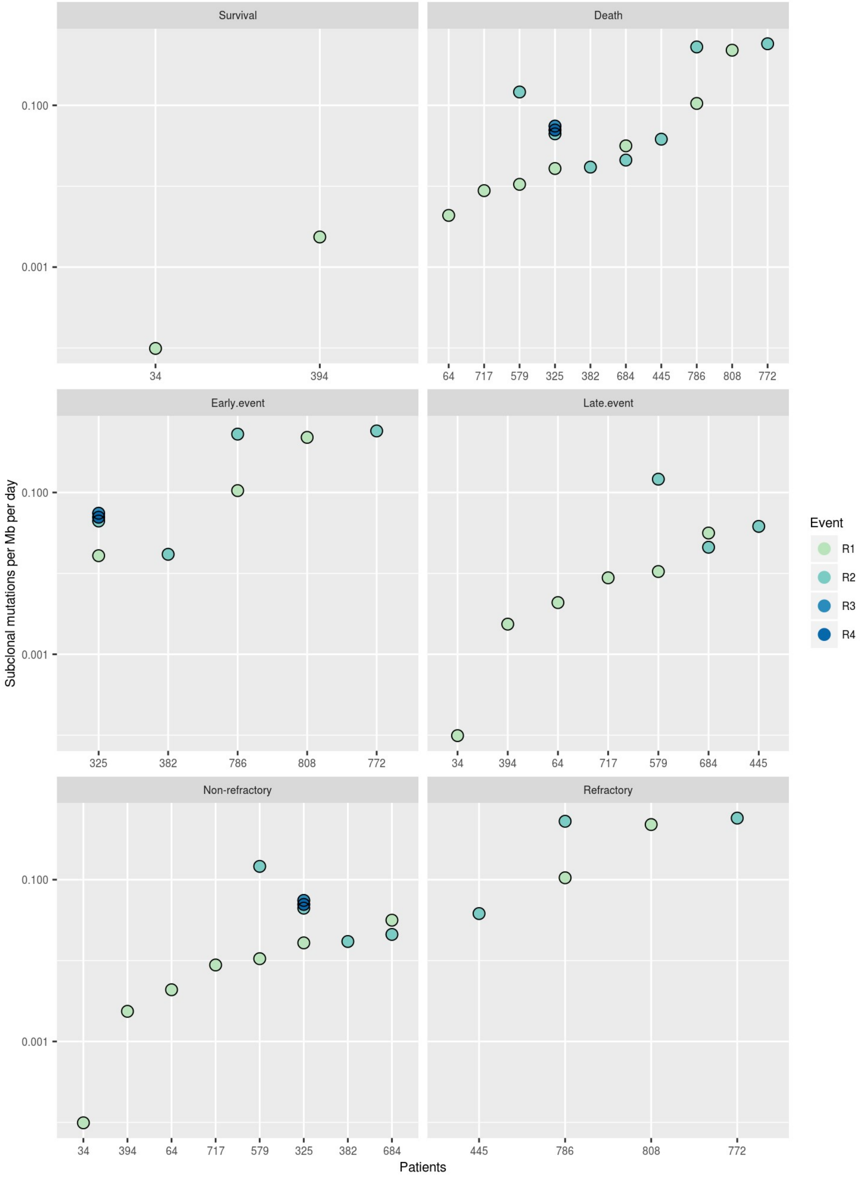
S4 Fig. Mutation spectra of 5 hypermutatable cases. PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse.



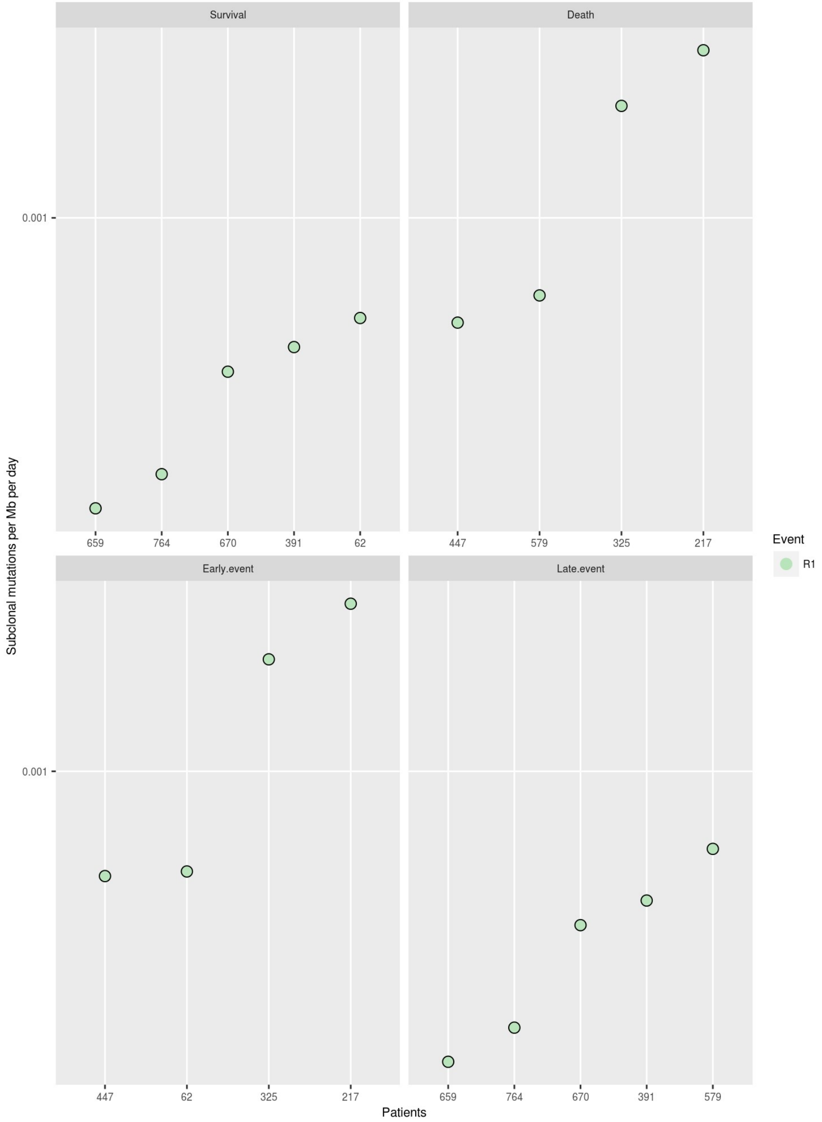
S5 Fig. Overview of putative drivers (somatic SNVs, small indels and CNVs) identified among 19 childhood cALL patients and grouped into relevant treatment pathways. Genes (top) are ordered according to the number of events identified in the cohort (descending order from left to right). Pathways are ordered according to the number of genes identified as mutated (descending order from left to right). The color scale indicates the VAFs of SNVs and indels (from light green - VAF ≤0.1 to dark blue - VAF =1). CNVs are indicated in cyan. Numbers in cells indicate the number of multiclonal mutations co-occurring in the same gene. PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse; SNV: single nucleotide variant; CNV: copy number variant; VAF: variant allele frequency; CTX: cyclophosphamide; L-ASP: asparaginase.



S6 Fig. Early and late cALL relapses show different clonal dynamics (WGS). The "dynamic mutation rate" (number of dynamic mutation per Mb per day) was calculated from WGS data by considering mutations showing a significant VAF shift only (Methods). Each value represents the clonal dynamics of a tumor between PT and R1. Patients are divided into 4 panels depending on their survival status and on the delay before relapse events. R1 stands for the analysis of PT vs. R1. The dynamic mutation rate was significantly lower in late relapses compared to early events (p =0.0079, Mann–Whitney U test) and for patients who survived compared to those who did not (p =0.0159, Mann–Whitney U test). PT: primary tumor; R1: first relapse; WES: whole-exome sequencing; WGS: whole-genome sequencing.



S7 Fig. Early and late cALL relapses show different subclonal expansion (WES). The "subclonal mutation rate" (number of newly acquired subclonal mutation per Mb per day) was calculated from WES data (Methods). Each value represents the subclonal expansion of a tumor between two events (either PT vs. R1 or R(n) vs. R(n+1)). Patients are divided into 6 panels depending on their survival or refractory status and on the delay before relapse events. R1 stands for the analysis of PT vs. R1, R2 for R1 vs. R2, R3 for R2 vs. R3 and R4 for R3 vs. R4. The subclonal expansion was significantly lower in late relapses compared to early events ($p = 0.0053$, Mann–Whitney U test) and in non-refractory compared to refractory events ($p = 0.0022$, Mann–Whitney U test). PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse; WES: whole-exome sequencing.



S8 Fig. Early and late cALL relapses show different subclonal expansion (WGS). The “subclonal mutation rate” (number of newly acquired subclonal mutation per Mb per day) was calculated from WGS data (Methods). Each value represents the subclonal expansion of a tumor between PT and R1. Patients are divided into 4 panels depending on their survival status and on the delay before relapse events. R1 stands for the analysis of PT vs. R1. The subclonal expansion was significantly lower in late relapses compared to early events (p =0.0317, Mann–Whitney U test) and in patients who survived compared to those who did not (p =0.0159, Mann–Whitney U test). PT: primary tumor; R1: first relapse.

5.9. Tables

Table 1. Childhood B-ALL patient clinical information.

Patient ID	Gender	Diagnosis age (mo.)	Year of diagnosis	Risk group	WBC (.10^9/l)	Platelet (.10^9/l)	Treatment protocol	Nb of relapse event	Sequenced events	Tumor purity (%)	Type	Event free survival (days)	Death
34	M	129	1995	High	2.8	323.0	DFCI 91-01	1	PT|R1	NA|94	Relapse	2147	N
62	M	29	1996	Standard	22.3	85.0	DFCI 91-01	1	PT|R1	95|97	Relapse	1025	N
64	F	67	1991	Standard	3.8	30.0	DFCI 91-01	2	PT|R1	NA|95	Relapse	1600	Y
217	F	23	1997	High	99.2	16.0	DFCI 95-01	1	PT|R1	94|87	Relapse	259	Y
325	M	160	1999	High	7.7	12.0	DFCI 95-01	5	PT|R1|R2|R3|R4	98|94|95|81|99	Relapse	306|106|309|41	Y
382	M	60	2000	NA	4.7	NA	DFCI 95-01	3	R1|R2	96|70	Relapse	534|105	Y
391	M	35	2000	Standard	10.0	47.0	DFCI 95-01	1	PT|R1	98|92	Relapse	1143	N
394	F	92	2000	Standard	10.5	11.0	DFCI 95-01	1	PT|R1	99|95	Relapse	1189	N
445	F	45	1997	NA	NA	NA	DFCI 2000-01	2	R1|R2	97|93	Refractory	2227|106	Y
447	F	129	2002	High	61.1	94.0	DFCI 2000-01	1	PT|R1	98|95	Relapse	982	Y
579	M	75	2002	High	183.1	29.0	DFCI 2000-01	2	PT|R1|R2	96|90|91	Relapse	1240|496	Y
659	M	35	2004	Standard	29.91	24.0	DFCI 2000-01	1	PT|R1	100|97	Relapse	1420	N
670	M	33	2005	Standard	5.48	20.0	DFCI 2005-01	1	PT|R1	93|94	Relapse	1199	N
684	F	74	2005	NA	8.49	317.0	DFCI 2005-01	2	PT|R1|R2	84|42|48	Relapse	1153|588	Y
717	F	73	2006	Standard	5.55	24.0	DFCI 2005-01	1	PT|R1	100|65	Relapse	1380	Y
764	M	55	2008	High	59.2	21.0	DFCI 2005-01	1	PT|R1	96|73	Relapse	1107	N
772	M	39	2008	Standard	29.72	4.0	DFCI 2005-01	2	R1|R2	93|45	Refractory	736|228	Y
786	M	206	2009	High	120.9	57.0	DFCI 2005-01	2*	PT|R1|R2	97|98|95	Refractory	35|22	Y
808	M	15	2010	High	265.7	47.0	DFCI 2005-01	1	PT|R1	93|38	Refractory	199	Y

*2nd sample of the same event

Patient ID	Karyotype (PT)	Karyotype (R events)
34	46,XY	46,XY
62	54,XY,+4,+6,+8,+10,+14,+17,+18,+21	55,XXY,+4,+6,+8,+10,+14,+der(17)del(17)(p12),+18,+21[12]/46,XY[7]
64	46,XX	46,XX,t(7;11)(q32;q14),+der(11)t(1;7;11)(q21;q32;q14),-21[17]/46,XX,idem,t(9;22)(p24;q22.1)[8]
217	46,XX,inv(4)(p14q27),del(9)(p22)	46,XX,inv(4)(p14q28),del(9)(p22),t(12;14)(q12;q11)[8]/46,XX,inv(4)(p14q28),del(9)(p22)[3]
325	46,XY	NA | NA | NA | NA
382	47,XY,t(9;14)(p13;q32),+mar	NA | 47,XY,del(2)(p22?),add(3)(p25),t(9;14)(p13;q32),t(16;Y)(q24;q11.2),+mar
391	52,XY,+3,+6,+11,+17,+21,+22	52,XY,+X,+6,+14,+17,+21,+21
394	46,XX,-12,-21,+mar1,+mar2	46,XX,-12,-21,+mar1,+mar2[1]/46,XX,der(12),-21,+mar3[5]
445	NA	47~48,-X,der(1),del(1)x2,-6,der(8),der(17),+21,+21,+mar | NA
447	47,XX,add(6)(q27),der(15)t(?13;15)(p11;p11),add(20)(?q11.2),+21	47,XX,add(1)(q32),add(6)(q27),der(15)(p11),+21
579	46,XY	46,XY | 46,XY
659	55,XXY,+2,+?3,+4,+6,+10,+14,+18,+21	54,XY,+4,+6,del(9)(p21),+10,del(12)(p13),+17,+21
670	51~56,XY,+X,+2,+4,+6,+9,+10,+14,+17,+21	53,XY,+4,+6,+8,+10,+14,+17,+21
684	46,XX	46,XX | 46,XX,dup(1)(q21q23)
717	46,XX,hsr(21)(q22.1)	46,XX,hsr(21)(q22.1)
764	56,XY,+X,+4,+5,+6,del(9p),+10,+14,+17,+18,+21,+21	NA
772	23~39,X,Y,+3,+5,+14,+16,+17,+20,[9]/52,XY,+X,+Y,+14,+20,+21,+21[13]	26~27,XY,+14,+18,+21
786	NA	NA | NA
808	46,XY,t(11;19)(q23;p13.1)	46,XY,t(11;19)(q23;p13.1)

Information concerning the different relapse events are separated by a bar in the cells. WBC: white blood cell; NA: not applicable; N: no; Y: yes; PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse.

S1 Table. SNVs and indels identified among 19 childhood B-ALL patients.

Patient	Hugo Symbol	Chr	Start position	End position	Ref allele	Alt allele	Type	AA change	VAF PT	VAF R1	VAF R2	VAF R3	VAF R4	dbSNP RS	Protein Change	1000Genome AF	ESP6500 EA AF
34	KRAS	12	25378647	25378647	T	A	MS	K117N	3.070	5.155	NA	NA	NA		p.K117N	0	0
34	KMT2D	12	49435963	49435963	C	T	SG	W2006*	31.250	55.573	NA	NA	NA		p.W2006*	0	0
34	NRAS	1	115258747	115258747	C	T	MS	G12D	40.991	41.015	NA	NA	NA	rs121913237	p.G12D	0	0.000116
62	CREBBP	16	3801726	3801726	C	T	SS	_1260_	10.354	0.000	NA	NA	NA			0	0
62	KRAS	12	25398284	25398284	C	T	MS	G12D	39.664	54.983	NA	NA	NA	rs121913529	p.G12D	0	0
64	CREBBP	16	3788606	3788606	A	C	MS	Y1450D	0.000	36.184	NA	NA	NA		p.Y1450D	0	0
64	POLB	8	42196150	42196150	A	G	MS	K3R	0.000	16.487	NA	NA	NA		p.K3R	0	0
217	KRAS	12	25398284	25398284	C	T	MS	G12D	14.184	0.000	NA	NA	NA	rs121913529	p.G12D	0	0
325	A2M	12	9265979	9265979	C	T	MS	V83I	0.000	0.000	0.000	13.717	12.219		p.V83I	0.000199681	0
325	TP53	17	7577120	7577120	C	T	MS	R273H	0.000	0.000	0.000	0.216	2.817	rs28934576	p.R273H	0.000199681	0
325	ABCC3	17	48753407	48753407	C	T	MS	T1008I	0.000	0.000	0.354	0.962	2.666		p.T1008I	0	0
325	CBS	21	44485543	44485543	G	A	MS	A207V	0.197	0.000	0.186	43.366	45.228		p.A207V	0	0
325	CREBBP	16	3786704	3786704	A	G	MS	Y1503H	0.000	0.473	0.196	46.451	40.445		p.Y1503H	0	0
325	CSF1R	5	149457757	149457757	C	T	MS	R216Q	0.586	0.000	0.000	3.685	0.526		p.R216Q	0	0
325	ELN	7	73471760	73471760	G	A	MS	G469D	0.000	0.000	0.000	50.705	52.801		p.G469D	0	0
325	FOXL2	3	138665530	138665530	G	A	MS	A12V	0.000	0.000	0.000	12.840	9.671		p.A12V	0	0
325	FPGS	9	130566979	130566979	G	T	MS	S129I	0.000	0.000	0.430	44.932	44.577		p.S129I	0	0
325	HRAS	11	532740	532740	A	G	MS	F156L	0.000	0.347	0.000	4.669	1.759		p.F156L	0	0
325	IKZF1	7	50450333	50450333	T	C	MS	F173L	0.000	0.000	0.000	36.895	40.404		p.F173L	0	0
325	KRAS	12	25378647	25378647	T	G	MS	K117N	15.806	0.262	0.000	0.257	0.000		p.K117N	0	0
325	MSH6	2	48026839	48026839	T	C	MS	F573L	0.000	0.000	0.000	5.746	2.876		p.F573L	0	0
325	NDUFS3	11	47600860	47600860	G	A	MS	R36Q	0.000	0.000	0.000	5.316	0.540		p.R36Q	0	0
325	PBRM1	3	52643776	52643776	T	C	MS	E707G	0.000	0.000	0.000	7.629	0.673		p.E707G	0	0
325	PMS2	7	6042221	6042221	G	A	SG	R134*	0.000	0.000	0.000	76.678	84.718	rs63750871	p.R134*	0	0
325	POLD1	19	50909737	50909737	A	G	MS	K486R	0.000	0.000	0.000	5.930	4.644		p.K486R	0	0
325	PTGS2	1	186648531	186648531	A	G	MS	V31A	0.000	0.000	0.000	32.846	35.354		p.V31A	0	0
325	PTPN11	12	112888199	112888199	C	A	MS	A72D	23.112	49.374	51.936	51.571	47.138	rs121918454	p.A72D	0	0
325	SETD2	3	47164760	47164760	G	A	SG	R456*	23.644	0.000	0.000	0.000	0.000		p.R456*	0	0
325	SMARCB1	22	24176338	24176338	C	T	MS	R377C	0.000	0.000	0.444	46.116	54.293		p.R377C	0	0
325	SRSF2	17	74733209	74733209	C	T	MS	G12S	0.000	0.000	0.000	53.677	38.340		p.G12S	0	0
325	TP53	17	7577121	7577121	G	A	MS	R273C	0.000	0.000	0.000	1.098	3.636	rs121913343	p.R273C	0	0
325	TP53	17	7577574	7577574	T	C	MS	Y236C	0.000	0.000	0.000	2.736	9.214		p.Y236C	0	0
325	USH2A	1	216040396	216040396	G	A	MS	A2933V	0.352	0.000	0.000	8.563	14.908		p.A2933V	0	0
325	WHSC1	4	1920309	1920309	G	A	MS	A457T	0.000	0.000	0.000	1.723	4.591		p.A457T	0	0
325	NR3C1	5	142662145	142662145	A	C	MS	D724E	0.000	9.251	4.733	53.115	40.104		p.D723E	0	0
325	NR3C1	5	142678252	142678252	G	A	MS	P626S	0.000	22.411	42.008	0.000	0.000		p.P625S	0	0
325	NR3C1	5	142675130	142675130	A	G	MS	Y641H	0.266	0.000	0.000	49.978	44.865		p.Y640H	0	0
325	KMT2C	7	151874713	151874713	G	A	SG	R2609*	0.000	0.000	0.563	54.148	45.785		p.R2609*	0	.|.
382	CYP2C9	10	96702006	96702006	C	T	MS	T130M	NA	3.516	37.706	NA	NA	rs200965026	p.T130M	0.000399361	0
382	POLQ	3	121186462	121186462	C	T	MS	V2291M	NA	26.552	46.005	NA	NA	rs371518241	p.V2291M	0.000998403	0.000116
382	APC	5	112173648	112173648	G	A	MS	R786H	NA	24.038	54.007	NA	NA		p.R786H	0	0
382	BAX	19	49459056	49459056	G	A	MS	G67R	NA	32.051	75.802	NA	NA	rs398122513	p.G67R	0	0
382	CHD2	15	93567691	93567691	G	T	MS	R1748L	NA	50.637	74.675	NA	NA		p.R1748L	0	0
382	EGFR	7	55229299	55229299	G	A	MS	V536M	NA	52.797	69.296	NA	NA		p.V536M	0	0
382	EZH2	7	148526829	148526829	C	T	MS	G159R	NA	43.620	57.304	NA	NA		p.G159R	0	0
382	FGG	4	155532958	155532958	G	A	SG	R134*	NA	23.600	58.480	NA	NA		p.R134*	0	0
382	FLT3	13	28608240	28608240	G	A	MS	P606S	NA	10.810	29.762	NA	NA		p.P606S	0	0
382	KIT	4	55599341	55599341	T	C	MS	Y823H	NA	51.440	68.758	NA	NA		p.Y823H	0	0
382	MSH2	2	47693856	47693856	C	T	MS	R524C	NA	87.560	100.000	NA	NA		p.R524C	0	0
382	NOTCH2	1	120468210	120468210	C	T	MS	R1410H	NA	10.081	4.926	NA	NA	rs202022988	p.R1410H	0	0
382	SLC19A1	21	46951822	46951822	C	T	MS	V144M	NA	68.598	100.000	NA	NA		p.V144M	0	0
382	SLC1A3	5	36684051	36684051	G	A	MS	G459S	NA	32.529	48.410	NA	NA		p.G459S	0	0
382	KIT	4	55593437	55593437	G	A	MS	V532I	NA	24.194	57.844	NA	NA	rs55792975	p.V532I	0	0.000233
382	BRCA1	17	41256182	41256182	C	T	MS	R133H	NA	47.822	72.516	NA	NA	rs80357357	p.R133H	0	.|.
382	CDKN2A	9	21971186	21971186	G	A	SG	R58*	NA	84.821	100.000	NA	NA	rs121913387	p.R58*	0	.|.
382	KDM6A	X	44922694	44922694	C	T	SG	R519*	NA	5.928	0.000	NA	NA	rs397514628	p.R519*	0	.|.
382	KDM6A	X	44969369	44969369	C	T	SG	R1351*	NA	49.603	100.000	NA	NA		p.R1351*	0	.|.
382	POLL	10	103343407	103343407	C	T	MS	R308Q	NA	8.523	48.780	NA	NA		p.R308Q	0	.|.

Patient	Hugo Symbol	Chr	Start position	End position	Ref allele	Alt allele	Type	AA change	VAF PT	VAF R1	VAF R2	VAF R3	VAF R4	dbSNP RS	Protein Change	1000Genome AF	ESP6500 EA AF
391	CBL	11	119148875	119148875	G	C	SS	_366_	12.294	0.000	NA	NA	NA	rs397517077		0	0
391	CREBBP	16	3786796	3786796	C	G	MS	W1472S	0.000	63.012	NA	NA	NA		p.W1472S	0	0
391	KRAS	12	25398285	25398285	C	A	MS	G12C	20.140	59.289	NA	NA	NA	rs121913530	p.G12C	0	0
391	KMT2C	7	151945051	151945051	A	G	MS	I823T	0.000	4.208	NA	NA	NA	rs2838171	p.I823T	0	0
394	NRAS	1	115258744	115258744	C	T	MS	G13D	0.412	46.840	NA	NA	NA	rs121434596	p.G13D	0	0
394	SH2B3	12	111884825	111884826	-	GGGGT	FI	_305_	2.464	42.715	NA	NA	NA		p.CT305fs	0	
445	BAP1	3	52438516	52438516	A	C	SG	Y401*	NA	4.457	2.430	NA	NA	rs200156887	p.Y401*	0.000199681	0.000116
445	KRAS	12	25398285	25398285	C	T	MS	G12S	NA	46.170	45.617	NA	NA	rs121913530	p.G12S	0	0
445	MTHFR	1	11854822	11854822	C	T	MS	R377H	NA	1.718	3.694	NA	NA		p.R377H	0	0
447	NRAS	1	115258744	115258744	C	T	MS	G13D	4.664	0.000	NA	NA	NA	rs121434596	p.G13D	0	0
447	NRAS	1	115258748	115258748	C	T	MS	G12S	7.465	0.000	NA	NA	NA	rs121913250	p.G12S	0	0
447	KMT2C	7	151927025	151927025	A	G	MS	Y987H	5.000	0.000	NA	NA	NA	rs183684706	p.Y987H	0	.|.
447	ATM	11	108202234	108202235	-	T	FI	_2527_	70.281	55.739	NA	NA	NA		p.M2527fs	0	
579	ATIC	2	216214353	216214353	C	T	MS	T585M	0.364	0.000	3.829	NA	NA	rs145243313	p.T585M	0	0
579	FLT3	13	28592642	28592642	C	A	MS	D835Y	45.706	33.206	0.000	NA	NA	rs121913486	p.D835Y	0	0
579	NRAS	1	115258744	115258744	C	T	MS	G13D	0.000	17.284	48.546	NA	NA	rs121434596	p.G13D	0	0
659	NRAS	1	115258747	115258747	C	G	MS	G12A	41.975	16.718	NA	NA	NA	rs121913237	p.G12A	0	0
659	NRAS	1	115258747	115258747	C	T	MS	G12D	41.975	16.718	NA	NA	NA	rs121913237	p.G12D	0	0.000116
670	KRAS	12	25398284	25398284	C	T	MS	G12D	58.244	43.579	NA	NA	NA	rs121913529	p.G12D	0	0
684	BRCA2	13	32914147	32914147	C	A	SG	C1885*	4.960	0.000	0.000	NA	NA	rs397507794	p.C1885*	0	0
684	EP300	22	41573234	41573234	T	C	MS	V1840A	0.000	0.000	74.705	NA	NA		p.V1840A	0	0
684	KDM6A	X	44911048	44911048	G	T	SS	_250_	9.862	0.000	0.000	NA	NA			0	0
684	MYCN	2	16082455	16082455	G	T	MS	S90I	0.000	0.000	73.529	NA	NA		p.S90I	0	0
717	BRCA2	13	32945172	32945172	A	C	MS	E2856A	0.000	44.678	NA	NA	NA	rs11571747	p.E2856A	0.000199681	0.001977
717	ARID1A	1	27105880	27105880	C	G	MS	L1831V	48.872	23.835	NA	NA	NA	rs150534917	p.L1831V	0.000199681	3.49E-4|3.49E-4
717	CBR1	21	37444216	37444216	C	A	SG	C167*	0.465	4.846	NA	NA	NA			0	0
717	TNFAIP3	6	138197255	138197255	G	A	MS	D253N	0.000	25.641	NA	NA	NA		p.D253N	0	0
717	USH2A	1	215960026	215960026	G	A	MS	T3458M	0.000	23.909	NA	NA	NA	rs149485593	p.T3458M	0	0
717	APC	5	112176819	112176819	C	T	MS	P1843L	0.000	45.138	NA	NA	NA	rs368080169	p.P1843L	0	0.000116
717	PYGL	14	51378885	51378885	G	A	MS	T586M	0.250	50.499	NA	NA	NA	rs144989341	p.T586M	0	0.000465
717	PAX5	9	36846902	36846902	-	CC	FI	_346_	37.500	0.000	NA	NA	NA		p.Y346fs	0	
717	USP8	15	50786479	50786480	-	AAGA	SS	_886_	0.000	43.337	NA	NA	NA	rs535741597		0	
764	CHGA	14	93397703	93397703	C	A	MS	P155Q	25.385	45.037	NA	NA	NA		p.P155Q	0	0
764	NRAS	1	115256530	115256530	G	T	MS	Q61K	23.014	0.000	NA	NA	NA	rs121913254	p.Q61K	0	0
764	PTPN11	12	112888210	112888210	G	A	MS	E76K	0.000	50.881	NA	NA	NA	rs121918464	p.E76K	0	0
772	SETD2	3	47163029	47163029	T	C	MS	T1033A	NA	0.000	17.567	NA	NA	rs145759179	p.T1033A	0.000998403	0.00093
772	ABCC2	10	101571298	101571298	G	T	MS	A636S	NA	0.000	20.344	NA	NA		p.A636S	0	0
772	BRCA1	17	41244831	41244831	T	C	MS	K906R	NA	0.000	100.000	NA	NA		p.K906R	0	0
772	CDKN2A	9	21971134	21971134	G	A	MS	P75L	NA	5.376	100.000	NA	NA		p.P75L	0	0
772	CDKN2A	9	21971179	21971179	G	T	MS	A60E	NA	0.000	100.000	NA	NA	rs36204594	p.A60E	0	0
772	PTCH1	9	98232134	98232134	C	T	MS	R603H	NA	0.000	20.817	NA	NA	rs199523893	p.R603H	0	0
772	TP53	17	7577121	7577121	G	A	MS	R273C	NA	0.000	100.000	NA	NA	rs121913343	p.R273C	0	0
772	SETD2	3	47161933	47161933	A	G	MS	I1398T	NA	0.000	21.433	NA	NA	rs145732065	p.I1398T	0	0.000581
772	ATM	11	108196896	108196896	C	T	MS	L2307F	NA	0.000	19.704	NA	NA	rs56009889	p.L2307F	0	0.002676
772	NF1	17	29553477	29553478	-	C	FI	_676_	NA	0.000	100.000	NA	NA	rs587781807	p.T676fs	0	
786	ASXL1	20	31023882	31023882	C	A	MS	P1123T	0.237	4.918	0.000	NA	NA		p.P1123T	0	0
808	NOTCH1	9	139401234	139401234	G	A	MS	R1279C	0.000	25.626	NA	NA	NA	rs182330532	p.R1279C	0.000399361	0.000944
808	XDH	2	31564244	31564244	A	G	MS	I1179T	0.000	25.304	NA	NA	NA	rs139515054	p.I1179T	0.000399361	0.002911
808	CYP2B6	19	41510282	41510282	A	G	MS	K139E	0.000	38.244	NA	NA	NA	rs12721655	p.K139E	0.000798722	0.003605
808	ASNS	7	97498378	97498378	C	T	MS	A31T	0.000	21.522	NA	NA	NA		p.A31T	0	0
808	ASNS	7	97498404	97498404	A	G	MS	M22T	0.830	23.256	NA	NA	NA		p.M22T	0	0
808	KRAS	12	25398284	25398284	C	T	MS	G12D	39.683	52.632	NA	NA	NA	rs121913529	p.G12D	0	0
808	POLD1	19	50919896	50919896	C	T	MS	L995F	0.000	33.452	NA	NA	NA		p.L995F	0	0
808	SCNN1A	12	6465035	6465035	T	C	MS	H355R	0.000	36.153	NA	NA	NA		p.H296R	0	0
808	KIT	4	55570011	55570011	A	G	MS	N293S	0.000	24.518	NA	NA	NA	rs137909416	p.N293S	0	0.000116
808	MSH2	2	47702191	47702191	A	G	MS	N596S	0.000	25.744	NA	NA	NA	rs63749831	p.N596S	0	0.000233
808	DHFR	5	79949873	79949873	A	C	MS	N30K	0.000	20.243	NA	NA	NA	rs201745474	p.N30K	0	0.000723

Patient	Hugo Symbol	Chr	Start position	End position	COSMIC n overlapping mutations	MutationTaster converted rankscore	MutationTaster pred	MutationTaster score	Polyphen2 pred	Polyphen2 score	SIFT pred	SIFT score
34	KRAS	12	25378647	25378647	9	0.80722	D	1	D;D	1.0;1.0	D	0.01
34	KMT2D	12	49435963	49435963	0	0.80722	A	1	0	0	D	0
34	NRAS	1	115258747	115258747	481	0.80722	D	1	B	0.372	D	0
62	CREBBP	16	3801726	3801726	2	0.80722	D	1	0	0	0	0
62	KRAS	12	25398284	25398284	15813	0.80722	D	1	B;P	0.385;0.517	D	0
64	CREBBP	16	3788606	3788606	1	0.80722	D	1	D;D	1.0;1.0	D	0
64	POLB	8	42196150	42196150	0	0.80722	D	1	D;D	1.0;1.0	D	0.04
217	KRAS	12	25398284	25398284	15813	0.80722	D	1	B;P	0.385;0.517	D	0
325	A2M	12	9265979	9265979	0	0.0931	N	1	B	0.013	T	0.37
325	TP53	17	7577120	7577120	669	0.80722	A	1	P;D;P;P	0.631;1.0;0.831;0.48	D	0.01
325	ABCC3	17	48753407	48753407	0	0.52501	D	0.999956	B	0.131	T	0.12
325	CBS	21	44485543	44485543	0	0.80722	D	1	D;D	0.999;1.0	D	0
325	CREBBP	16	3786704	3786704	3	0.80722	D	1	D;D	1.0;1.0	D	0
325	CSF1R	5	149457757	149457757	0	0.0931	N	0.999998	B;B;B;B	0.025;0.009;0.009;0.005	T	0.66
325	ELN	7	73471760	73471760	0	0.0931	N	1	0	0	T	0.62
325	FOXL2	3	138665530	138665530	0	0.34617	D	0.718219	P	0.839	T	0.19
325	FPGS	9	130566979	130566979	0	0.80722	D	1	D;D	0.997;1.0	D	0
325	HRAS	11	532740	532740	0	0.80722	D	1	P	0.955	D	0
325	IKZF1	7	50450333	50450333	131	0.80722	D	1	P;D;D	0.504;0.999;0.99	D	0.03
325	KRAS	12	25378647	25378647	9	0.80722	D	1	D;D	1.0;1.0	D	0.01
325	MSH6	2	48026839	48026839	2	0.80722	D	1	D;D;D	0.998;0.998;1.0	D	0
325	NDUFS3	11	47600860	47600860	0	0.19575	N	0.999967	P;B	0.577;0.038	T	0.24
325	PBRM1	3	52643776	52643776	0	0.80722	D	1	D;D;D;D;D;D;D;D;P;D;D	1.0;0.996;1.0;0.983;0.981;1.0;1.0;1.0;0.63;1.0;1.0	D	0.02
325	PMS2	7	6042221	6042221	0	0.80722	A	1	0	0	T	1
325	POLD1	19	50909737	50909737	0	0.80722	D	1	D;D	1.0;1.0	D	0
325	PTGS2	1	186648531	186648531	0	0.52099	D	0.99994	B	0.005	D	0.03
325	PTPN11	12	112888199	112888199	38	0.80722	D	1	D;D	0.997;0.999	D	0.03
325	SETD2	3	47164760	47164760	0	0.80722	A	1	0	0	T	1
325	SMARCB1	22	24176338	24176338	3	0.80722	D	1	D;D;D	1.0;0.999;1.0	D	0
325	SRSF2	17	74733209	74733209	0	0.80722	D	1	B;B	0.235;0.235	D	0.01
325	TP53	17	7577121	7577121	533	0.80722	D	1	D;D;D;D	1.0;1.0;1.0;1.0	D	0
325	TP53	17	7577574	7577574	94	0.4457	D	0.997534	D;D;D;D;D;D	0.998;0.982;1.0;0.998;0.999;1.0	D	0
325	USH2A	1	216040396	216040396	0	0.0931	N	1	B	0.019	T	1
325	WHSC1	4	1920309	1920309	0	0.80722	D	1	P;B;P;P	0.688;0.222;0.554;0.554	T	0.59
325	NR3C1	5	142662145	142662145	0	0.58432	D	0.999993	D;D;D	1.0;1.0;1.0	D	0
325	NR3C1	5	142678252	142678252	0	0.80722	D	1	D;D;D	1.0;1.0;1.0	D	0
325	NR3C1	5	142675130	142675130	0	0.34437	D	0.699203	D;D;D	0.992;0.997;0.99	T	0.08
325	KMT2C	7	151874713	151874713	1	0.41672|0.41672	D|D	0.989244|0.989244	.|.	.|.	.|T	.|1
382	CYP2C9	10	96702006	96702006	1	0.35167	D	0.771229	D;D;D	1.0;1.0;1.0	D	0
382	POLQ	3	121186462	121186462	1	0.0931	N	1	B;B	0.012;0.005	T	0.18
382	APC	5	112173648	112173648	1	0.80722	D	1	D;D	0.997;0.998	T	0.06
382	BAX	19	49459056	49459056	0	0.80722	D	1	D;D;D	0.997;0.979;0.995	D	0.02
382	CHD2	15	93567691	93567691	0	0.80722	D	1	D	0.997	T	0.11
382	EGFR	7	55229299	55229299	0	0.58432	D	0.999999	D;D;D;D	0.999;0.998;0.994;0.989	D	0
382	EZH2	7	148526829	148526829	2	0.80722	D	1	D;D;D;D;D;D	1.0;1.0;1.0;1.0;1.0;1.0	D	0
382	FGG	4	155532958	155532958	0	0.80722	A	1	0	0	T	1
382	FLT3	13	28608240	28608240	3	0.80722	D	1	D;P	1.0;0.901	D	0
382	KIT	4	55599341	55599341	10	0.80722	D	1	D;D	1.0;1.0	D	0
382	MSH2	2	47693856	47693856	4	0.80722	D	1	D;D;D	1.0;1.0;1.0	D	0
382	NOTCH2	1	120468210	120468210	1	0.27362	N	0.955934	B	0	T	0.6
382	SLC19A1	21	46951822	46951822	0	0.80722	D	1	D;D;D;D	1.0;1.0;1.0;1.0	D	0
382	SLC1A3	5	36684051	36684051	0	0.80722	D	1	D	1	D	0.04
382	KIT	4	55593437	55593437	1	0.193	N	0.999973	B;B	0.075;0.089	T	0.51
382	BRCA1	17	41256182	41256182	0	0.80722|0.80722	D|D	1|1	D;D;D;D;D;D;D;D;D;D;D|.	1.0;1.0;0.999;1.0;1.0;1.0;1.0;1.0;1.0;0.998;1.0|.	D|.	0|.
382	CDKN2A	9	21971186	21971186	1459	0.09310|0.09310	N|N	1|1	P|.	0.824|.	D|T	0.01|0.59
382	KDM6A	X	44922694	44922694	9	0.80722|0.80722	A|A	1|1	.|.	.|.	T|.	1|.
382	KDM6A	X	44969369	44969369	6	0.80722|0.80722	A|A	1|1	.|.	.|.	T|.	0.99|.
382	POLL	10	103343407	103343407	0	0.27238|0.27238	N|N	0.959223|0.959223	P|P;D;D;D;D	0.733|0.555;0.988;0.959;0.999;0.97	D|D	0|0.01

Patient	Hugo Symbol	Chr	Start position	End position	COSMIC n overlapping mutations	MutationTaster converted rankscore	MutationTaster pred	MutationTaster score	Polyphen2 pred	Polyphen2 score	SIFT pred	SIFT score
391	CBL	11	119148875	119148875	3	0.80722	D	1	0	0	0	0
391	CREBBP	16	3786796	3786796	0	0.80722	D	1	D;D	1.0;1.0	D	0
391	KRAS	12	25398285	25398285	5144	0.80722	D	1	D;D	1.0;1.0	D	0.01
391	KMT2C	7	151945051	151945051	1	0.80722	D	1	D	0.999	T	0.41
394	NRAS	1	115258744	115258744	262	0.80722	D	1	B	0.434	D	0.03
394	SH2B3	12	111884825	111884826	0							
445	BAP1	3	52438516	52438516	0	0.80722	A	1	0	0	T	0.37
445	KRAS	12	25398285	25398285	5144	0.80722	D	1	P;P	0.557;0.682	D	0.01
445	MTHFR	1	11854822	11854822	0	0.80722	D	1	D;D	1.0;1.0	D	0
447	NRAS	1	115258744	115258744	262	0.80722	D	1	B	0.434	D	0.03
447	NRAS	1	115258748	115258748	236	0.80722	D	1	B	0.329	D	0.04
447	KMT2C	7	151927025	151927025	2	0.80722|0.80722	D|D	1|1	.|D;D	.|1.0;0.999	.|D	.|0
447	ATM	11	108202234	108202235	0							
579	ATIC	2	216214353	216214353	0	0.80722	D	1	D;D	1.0;1.0	D	0
579	FLT3	13	28592642	28592642	252	0.80722	D	1	D	1	D	0
579	NRAS	1	115258744	115258744	262	0.80722	D	1	B	0.434	D	0.03
659	NRAS	1	115258747	115258747	481	0.80722	D	1	P	0.775	D	0
659	NRAS	1	115258747	115258747	481	0.80722	D	1	B	0.372	D	0
670	KRAS	12	25398284	25398284	15813	0.80722	D	1	B;P	0.385;0.517	D	0
684	BRCA2	13	32914147	32914147	0	0.80722	A	1	0	0	T	0.54
684	EP300	22	41573234	41573234	0	0.3159	N	0.653951	B	0.001	T	0.82
684	KDM6A	X	44911048	44911048	10	0.80722	D	1	0	0	0	0
684	MYCN	2	16082455	16082455	0	0.27277	N	0.958193	B	0.202	D	0.01
717	BRCA2	13	32945172	32945172	0	0.35693	D	0.814991	D	0.981	T	0.15
717	ARID1A	1	27105880	27105880	0	0.42266|0.42266	D|D	0.992031|0.992031	.|D;D;D	.|0.983;0.985;0.992	.|T	.|0.24
717	CBR1	21	37444216	37444216	0	0.0931	N	0.999998	0	0	T	0.16
717	TNFAIP3	6	138197255	138197255	26	0.52501	D	0.999957	D	1	T	0.12
717	USH2A	1	215960026	215960026	0	0.0931	N	1	D	0.977	T	0.1
717	APC	5	112176819	112176819	2	0.80722	D	1	D;D	1.0;1.0	0	0
717	PYGL	14	51378885	51378885	0	0.46774	D	0.999194	D;D;D	0.995;0.998;1.0	D	0.01
717	PAX5	9	36846902	36846902	12							
717	USP8	15	50786479	50786480	0							
764	CHGA	14	93397703	93397703	0	0.0931	N	0.999999	B	0.221	T	0.34
764	NRAS	1	115256530	115256530	610	0.80722	D	1	P	0.948	D	0.03
764	PTPN11	12	112888210	112888210	88	0.80722	D	1	D;D	0.998;0.995	D	0
772	SETD2	3	47163029	47163029	0	0.2514	N	0.990508	B;B	0.01;0.01	T	0.13
772	ABCC2	10	101571298	101571298	0	0.80722	D	1	P	0.902	T	0.44
772	BRCA1	17	41244831	41244831	0	0.2305	N	0.998373	B;B;B;P;D;P	0.108;0.108;0.394;0.541;0.965;0.778	D	0.05
772	CDKN2A	9	21971134	21971134	1368	0.80722	D	1	D	1	T	0.3
772	CDKN2A	9	21971179	21971179	1375	0.80722	D	1	D	0.997	D	0.01
772	PTCH1	9	98232134	98232134	0	0.80722	D	1	D;D;D;D	0.996;0.998;0.992;0.998	T	0.07
772	TP53	17	7577121	7577121	533	0.80722	D	1	D;D;D;D	1.0;1.0;1.0;1.0	D	0
772	SETD2	3	47161933	47161933	0	0.27466	N	0.953008	B;B	0.039;0.039	T	0.1
772	ATM	11	108196896	108196896	0	0.80722	D	1	D	0.999	T	0.07
772	NF1	17	29553477	29553478	13							
786	ASXL1	20	31023882	31023882	0	0.37973	D	0.933348	P;P	0.926;0.926	D	0
808	NOTCH1	9	139401234	139401234	2	0.80722	D	1	D	0.996	T	0.06
808	XDH	2	31564244	31564244	0	0.80722	D	1	D	0.998	D	0
808	CYP2B6	19	41510282	41510282	0	0.0931	N	1	D;D	0.994;0.995	D	0
808	ASNS	7	97498378	97498378	0	0.80722	D	1	P	0.786	D	0.05
808	ASNS	7	97498404	97498404	0	0.80722	D	1	B	0.034	T	0.07
808	KRAS	12	25398284	25398284	15813	0.80722	D	1	B;P	0.385;0.517	D	0
808	POLD1	19	50919896	50919896	0	0.45634	D	0.998568	P;P	0.886;0.789	D	0.05
808	SCNN1A	12	6465035	6465035	0	0.29938	N	0.821302	B;B;B	0.02;0.019;0.01	T	0.35
808	KIT	4	55570011	55570011	0	0.31723	N	0.6381	B;B	0.004;0.023	T	0.82
808	MSH2	2	47702191	47702191	4	0.80722	D	1	B;B;B	0.016;0.088;0.012	T	0.41
808	DHFR	5	79949873	79949873	0	0.80722	D	1	B;B;B	0.128;0.031;0.031	0	0

To be considered as damaging (D), mutations had to present a MutationTaster score >0.9, a Polyphen2 score >0.909 or a Sift score ≤0.05. Mutations presenting a Polyphen2 score between 0.447 and 0.909 were considered as possibly damaging (P). For each algorithm, mutations that did not meet the respective filtering criteria were annotated as tolerated (T). VAFs were adjusted for tumor purity. Chr: chromosome; Ref: reference; Alt: alternative; AA: amino acid; MS: missense; SS: splice site; SG: stop gain; FI: frameshift indel; AF: allele frequency; VAF: variant allele frequency; EA: EuropeanAmerican; pred: prediction; NA: not applicable; PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse.

S2 Table. CNVs identified among 19 childhood B-ALL patients.

Sample	Hugo Symbol	Chromosome	Start position	End position	Type
217_PT	CDKN2A	9	21902343	21997597	LOSS
217_PT	PAX5	9	36997417	37000687	GAIN
217_PT	NFKB1	4	100772603	103932556	LOSS
217_R1	CDKN2A	9	21902354	21997597	LOSS
217_R1	PAX5	9	36997417	37000687	GAIN
217_R1	NFKB1	4	100772603	103933567	LOSS
325_PT	CDKN2A	9	21965073	22012566	LOSS
325_R1	CDKN2A	9	21965073	22012566	LOSS
325_R3	NOTCH2	1	120548111	120548258	GAIN
391_PT	NOS3	7	150670943	150873566	GAIN
391_PT	POLE	12	132250360	133378852	GAIN
391_PT	SLC19A1	21	45774410	48095982	GAIN
391_R1	CBR3	21	37490655	37512565	GAIN
391_R1	POLE	12	132259946	133378852	GAIN
391_R1	SLC19A1	21	45774410	48095982	GAIN
394_PT	NF2	22	30005024	30005224	LOSS
394_PT	RUNX1	21	35514745	37092998	GAIN
394_PT	U2AF1	21	44514819	44524447	GAIN
394_R1	RUNX1	21	36170347	37092998	GAIN
394_R1	U2AF1	21	44514817	44524447	GAIN
447_PT	IKZF1	7	50417522	50462935	LOSS
447_PT	SETD2	3	47011359	47204518	LOSS
447_R1	IKZF1	7	50413334	50462935	LOSS
447_R1	SETD2	3	47010116	47204518	LOSS
579_PT	IKZF1	7	50413334	50462138	LOSS
579_PT	ABCC5	3	183280319	184526415	LOSS
579_PT	CRHR2	7	30689733	31163196	LOSS
579_PT	ELN	7	73072839	76048314	LOSS
579_PT	IMPDH1	7	127669008	128465450	LOSS
579_PT	MAT1A	10	80853508	82043576	LOSS
579_PT	NT5C2	10	102556902	105616713	LOSS
579_PT	POLB	8	40698506	42446238	LOSS
579_PT	POLE	12	131456449	133779375	LOSS
579_PT	POLN	4	809787	7771634	LOSS
579_PT	POR	7	73072839	76048314	LOSS
579_PT	RAC1	7	4911336	6495466	LOSS
579_R1	IKZF1	7	50413334	50462138	LOSS
579_R1	SMARCA4	19	10054253	11710309	LOSS
579_R1	ABCC4	13	95844281	96152308	LOSS
579_R1	ABCC5	3	183145833	184526415	LOSS
579_R1	CRHR2	7	30689733	31163196	LOSS
579_R1	ELN	7	73072839	76048314	LOSS
579_R1	NT5C2	10	102556902	105587304	LOSS
579_R1	POLN	4	18058	4238315	LOSS
579_R1	POR	7	73072839	76048314	LOSS
579_R1	RAC1	7	4914358	6495466	LOSS
579_R1	SCNN1A	12	5577867	7318631	LOSS
62_PT	NOS3	7	150639871	151123518	GAIN
62_PT	POLE	12	132304059	133779375	GAIN
64_PT	NF2	22	30005015	30005215	LOSS
64_R1	RB1	13	48985764	49064267	LOSS
659_R1	CDKN2A	9	21701942	22176961	LOSS
659_R1	PAX5	9	36882141	37070398	LOSS
659_R1	ETV6	12	NA	NA	LOSS
670_PT	POLN	4	18058	3436588	GAIN
670_R1	POLN	4	18058	3436588	GAIN
684_R1	NF2	22	30005083	30005183	LOSS
717_PT	RUNX1	21	NA	NA	GAIN
717_R1	ATRX	X	76937501	76944248	GAIN
717_R1	RUNX1	21	NA	NA	GAIN
764_PT	CDKN2A	9	NA	NA	LOSS
764_R1	CDKN2A	9	NA	NA	LOSS
772_R1	KDM6A	X	44820527	44833938	LOSS
772_R1	MED12	X	70329118	70344029	GAIN
772_R2	KDM6A	X	44820427	44833938	LOSS
786_R2	NF2	22	30005021	30005221	LOSS
808_PT	MDM4	1	204498229	204498329	LOSS

NA: not applicable; PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse.

S3 Table. Number of SNVs and indels shared between consecutive events.

Samples	Total T1	Total sub. T1	sub. T1 (%)	Total T2	Total sub. T2	sub. T2 (%)	Shared T1-T2	Persisted	T2-spec	Ratio T2/T1
34.TvsR	26	10	38	26	9	35	17	0.65	9	1.00
62.TvsR	17	9	53	23	14	61	8	0.47	15	1.35
64.TvsR	25	25	100	67	37	55	2	0.08	65	2.68
217.TvsR	17	14	82	18	13	72	3	0.18	15	1.06
325.TvsR	36	27	75	48	29	60	15	0.42	33	1.33
325.R1vsR2	48	29	60	60	36	60	27	0.56	33	1.25
325.R2vsR3	60	36	60	911	570	63	19	0.32	892	15.18
325.R3vsR4	911	570	63	963	603	63	734	0.81	229	1.06
382.R1vsR2	782	365	47	713	89	12	633	0.81	80	0.91
391.TvsR	13	9	69	8	5	63	1	0.08	7	0.62
394.TvsR	32	19	59	47	32	68	6	0.19	41	1.47
445.R1vsR2	579	520	90	1386	1328	96	379	0.65	1007	2.39
579.TvsR	26	17	65	29	20	69	9	0.35	20	1.12
579.R1vsR2	29	20	69	2185	2171	99	7	0.24	2178	75.34
447.TvsR	17	14	82	20	8	40	2	0.12	18	1.18
659.TvsR	5	4	80	17	6	35	0	0.00	17	3.40
670.TvsR	16	8	50	15	7	47	3	0.19	12	0.94
684.TvsR	30	28	93	31	27	87	5	0.17	26	1.03
684.R1vsR2	31	27	87	3823	97	3	8	0.26	3815	123.32
717.TvsR	1852	161	9	5119	417	8	973	0.53	4146	2.76
764.TvsR	17	11	65	53	10	19	7	0.41	46	3.12
772.R1vsR2	34	18	53	5282	3975	75	11	0.32	5271	155.35
786.TvsR	180	173	96	28	19	68	10	0.06	18	0.16
786.R1vsR2	28	19	68	19	10	53	11	0.39	8	0.68
808.TvsR	17	12	71	9985	1756	18	5	0.29	9980	587.35

Coding somatic mutations including non-synonymous SNVs, in-frame and frameshift indels as well as mutations located in splicing sites were considered here. Persisted values correspond to the ratio of the number of variants shared between T1 and T2 over the total number of variants identified at T1. T1: event 1; T2: event 2; sub.: subclonal; PT: primary tumor; R1, R2, R3, R4: first, second, third and fourth relapse.

5.10. Acknowledgments

The authors are indebted to the patients and their parents for participating in this study. Patient tissue samples were provided by the CHU Sainte-Justine Pediatric Cancer biobank. WES was performed at the Integrated Clinical Genomics Centre In Pediatrics, CHU Sainte-Justine; WGS was performed by Illumina, Inc. (San Diego, CA, US); WGG was performed at the McGill University and Genome Quebec Innovation Centre; Computations were made on the supercomputer Briarée from Université de Montréal, managed by Calcul Québec and Compute Canada.

5.11. Competing interests

The authors (JFS, CR, PC, MO, JH and DS) declare no conflict of interest.

5.12. References

1. Institute of Medicine (US) and National Research Council (US) National Cancer Policy Board. Childhood Cancer Survivorship: Improving Care and Quality of Life.; Hewitt M, Weiner SL, Simone JV, editors. Washington (DC): National Academies Press (US). 2003.
2. Siegel DA, King J, Tai E, Buchanan N, Ajani UA, Li J. Cancer incidence rates and trends among children and adolescents in the United States, 2001-2009. Pediatrics. 2014;134(4):e945-55. doi: 10.1542/peds.2013-3926.
3. Hunger SP, Mullighan CG. Acute Lymphoblastic Leukemia in Children. N Engl J Med. 2015;373(16):1541-52.
4. Smith MA, Seibel NL, Altekruse SF, et al. Outcomes for children and adolescents with cancer: challenges for the twenty-first century. J Clin Oncol. 2010;28:2625-34.
5. Linabery AM, Ross JA. Trends in childhood cancer incidence in the U.S. (1992-2004). Cancer. 2008;112:416-32.
6. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381): 306-13. doi: 10.1038/nature10762.
7. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152:714-26.
8. Irving J, Matheson E, Minto L, Blair H, Case M, Halsey C, Swidenbank I, Ponthan F, Kirschner-Schwabe R, Groeneveld-Krentz S, Hof J, Allan J, Harrison C, Vormoor J, von Stackelberg A, Eckert C. Ras pathway mutations are prevalent in relapsed childhood acute lymphoblastic leukemia and confer sensitivity to MEK inhibition. Blood. 2014;124(23):3420-30.
9. Inthal A, Zeitlhofer P, Zeginigg M, Morak M, Grausenburger R, Fronkova E, Fahrner B, Mann G, Haas OA, Panzer-Grümayer R. CREBBP HAT domain mutations prevail in relapse cases of high hyperdiploid childhood acute lymphoblastic leukemia. Leukemia. 2012;26(8):1797-803.
10. Mullighan CG, Zhang J, Kasper LH, Lerach S, Payne-Turner D, Phillips LA, Heatley SL, Holmfeldt L, Collins-Underwood JR, Ma J, Buetow KH, Pui CH, Baker SD, Brindle PK, Downing JR. CREBBP mutations in relapsed acute lymphoblastic leukaemia. Nature. 2011;471(7337):235-9.
11. Meyer JA, Wang J, Hogan LE, Yang JJ, Dandekar S, Patel JP, Tang Z, Zumbo P, Li S, Zavadil J, Levine RL, Cardozo T, Hunger SP, Raetz EA, Evans WE, Morrison DJ, Mason CE, Carroll WL. Relapse specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat Genet. 2013;45(3):290-4.
12. Tzoneva G, Perez-Garcia A, Carpenter Z, Khiabanian H, Tosello V, Allegretta M, Paietta E, Racevskis J, Rowe JM, Tallman MS, Paganin M, Basso G, Hof J, Kirschner-Schwabe R, Palomero T, Rabadan R, Ferrando A. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nat Med. 2013;19(3):368-71.
13. Ma X, Edmonson M, Yergeau D, Muzny DM, Hampton OA, Rusch M, et al. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat Commun. 2015;6:6604.
14. Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. Leukemia. 2014;28(1):34-43.
15. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, Cibulskis K, Tesar B, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ. Evolution and impact of subclonal mutations in chronic

lymphocytic leukemia. Cell. 2013;152(4):714-26. doi: 10.1016/j.cell.2013.01.019.

16. Green MR, Gentles AJ, Nair RV, Irish JM, Kihira S, Liu CL, et al. Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. Blood. 2013;121:1604-11.

17. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012, 150:264-78.

18. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, Downing JR. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. Science. 2008;322(5906):1377-80. doi: 10.1126/science.1164266.

19. Miyaki M, Konishi M, Tanaka K, Kikuchi-Yanoshita R, Muraoka M, Yasuno M, et al. Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. Nat Genet. 1997;17:271-2.

20. Kolodner RD, Tytell JD, Schmeits JL, Kane MF, Gupta RD, Weger J, et al. Germline MSH6 mutations in colorectal cancer families. Cancer Res. 1999;59:5068-74.

21. Wu Y, Berends MJ,Mensink RG, Kempinga C,Sijmons RH, van Der Zee AG, et al. Association of hereditary nonpolyposis colorectal cancer-related tumors displaying low microsatellite instability with MSH6 germline mutations. Am J Hum Genet. 1999;65:1291-8.

22. Li Z, Kong L, Yu L, Huang J, Wang K, Chen S, Yu M, Wei S. Association between MSH6 G39E polymorphism and cancer susceptibility: a meta-analysis of 7,046 cases and 34,554 controls. Tumour Biol. 2014;35(6):6029-37. doi: 10.1007/s13277-014-1798-z.

23. Marinovic-Terzic I, Yoshioka-Yamashita A, Shimodaira H, Avdievich E, Hunton IC, Kolodner RD, Edelmann W, Wang JY. Apoptotic function of human PMS2 compromised by the nonsynonymous single-nucleotide polymorphic variant R20Q. Proc Natl Acad Sci U S A. 2008;105(37):13993-8. doi: 10.1073/pnas.0806435105.

24. Barrio S, Shanafelt TD, Ojha J, Chaffee KG, Secreto C, Kortüm KM, Pathangey S, Van-Dyke DL, Slager SL, Fonseca R, Kay NE, Braggio E. Genomic characterization of high-count MBL cases indicates that early detection of driver mutations and subclonal expansion are predictors of adverse clinical outcome. Leukemia. 2016. doi: 10.1038/leu.2016.172.

25. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10(8):e1003665. doi: 10.1371/journal.pcbi.1003665. ECollection 2014.

26. El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics - Special Issue: Proceedings of ISMB. 2015;31(12):i62-i70.

27. Elamin Y, Toomey S, Carr A, Gately K, Rafee S, Grogan W, Morris PG, Breathnach OS, Crown J, O'Byrne K, Hennessy B. Protein tyrosine phosphatase non receptor 11 (PTPN11/Shp2) as a driver oncogene and a novel therapeutic target in non-small cell lung cancer (NSCLC). J Clin Oncol. 2015;33 (suppl; abstr 11077).

28. Bledsoe RK, Montana VG, Stanley TB, Delves CJ, Apolito CJ, McKee DD, Consler TG, Parks DJ, Stewart EL, Willson TM, Lambert MH, Moore JT, Pearce KH, Xu HE. Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. Cell. 2002;110(1):93-105.

29. Tzoneva G, Perez-Garcia A, Carpenter Z, Khiabanian H, Tosello V, Allegretta M, et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. Nature medicine. 2013;19(3):368-71.

30. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, Lindblom A, Lagerstedt K, Thibodeau SN, Lindor NM, Young J, Winship I, Dowty JG, White DM, Hopper JL, Baglietto L, Jenkins MA, de la Chapelle A. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. Gastroenterology. 2008;135(2):419-28. doi: 10.1053/j.gastro.2008.04.026.

31. Pui CH, Pei D, Coustan-Smith E, Jeha S, Cheng C, Bowman WP, Sandlund JT, Ribeiro RC, Rubnitz JE, Inaba H, Bhojwani D, Gruber TA, Leung WH, Downing JR, Evans WE, Relling MV, Campana D. Clinical utility of sequential minimal residual disease measurements in the context of risk-based therapy in childhood acute lymphoblastic leukaemia: a prospective study. Lancet Oncol. 2015;16(4):465-74. doi: 10.1016/S1470-2045(15)70082-3.

32. Ford AM, Mansur MB, Furness CL, van Delft FW, Okamura J, Suzuki T, Kobayashi H, Kaneko Y, Greaves M. Protracted dormancy of pre-leukemic stem cells. Leukemia. 2015;29(11):2202-7. doi: 10.1038/leu.2015.132.

33. Ford AM, Fasching K, Panzer-Grümayer ER, Koenig M, Haas OA, Greaves MF. Origins of "late" relapse in childhood acute lymphoblastic leukemia with TEL-AML1 fusion genes. Blood. 2001;98(3):558-64.

34. Greaves MF. Aetiology of acute leukaemia. Lancet. 1997;349: 344-349.

35. Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood. 2007;109(2):683-92.

36. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. Leuk Res. 1997;21(9):817-23.

37. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9: 357-9.

38. Picard. Broadinstitute. 2016. [cited 19 July 2016]. Available: http://broadinstitute.github.io/picard/.

39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

40. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811-7.

41. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-76.

42. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, Perou CM, Børresen-Dale AL, Kristensen VN. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107(39):16910-5.

43. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

44. Ramos AH, Lichtenstein L, Gupta M, Lawrence MS, Pugh TJ, Saksena G, Meyerson M, Getz G. Oncotator: cancer variant annotation tool. Hum Mutat. 2015;36(4):E2423-9.

45. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

46. Exome Variant Server. NHLBI Exome Sequencing Project (ESP). 2016. [cited 19 July 2016]. Available:http://evs.gs.washington.edu/EVS/.

47. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339(6127):1546-58.

48. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.

49. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

50. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361-2.
51. Spinella JF, Cassart P, Richer C, Saillour V, Ouimet M, Langlois S, St-Onge P, Sontag T, Healy J, Minden MD, Sinnett D. Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations. Oncotarget. 2016. doi: 10.18632/oncotarget.11796.
52. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics Knowledge for Personalized Medicine. Clinical Pharmacology & Therapeutics. 2012;92(4):414-7.

CHAPITRE 6

Résumé, Discussion et Conclusion

6.1. Résumé des résultats principaux

Les travaux menés au cours de cette thèse ont tout d'abord permis de développer et de rendre disponible un outil d'analyse (SNooPer) qui facilitera l'étude des données somatiques par la communauté de la même manière qu'il nous a aidé au cours de ces travaux.

Ces travaux ont également permis de caractériser une série d'évènements *drivers* rares ou récurrents impliqués dans le processus leucémogène. Par le biais d'une analyse orientée, outre l'identification d'acteurs rares et fonctionnels impliqués dans l'initiation ou la progression de la leucémie pédiatrique, nous avons été parmi les premiers à démontrer l'importance et le besoin d'une analyse prenant en compte l'ensemble des évènements présentant un potentiel *driver*, y compris les moins fréquents. Cela semble indispensable à la mise en évidence de l'ensemble des fonctions cellulaires qui par leur dérégulation peuvent participer au développement de la maladie.

Nos projets nous ont également permis de mettre en évidence de nouveaux évènements récurrents chez des patients atteints de LAL-T. Nous avons par exemple identifié une mutation au niveau d'*U2AF1* qui est présente chez 10% des patients de notre cohorte. Nous avons démontré l'influence de cette mutation sur l'épissage alternatif de gènes clés et ainsi confirmé l'importance du dysfonctionnement de l'épissage dans le développement de la leucémie. Nous avons également identifié *MED12* et *USP9X* comme nouveaux *drivers* portés par le chromosome X. Ces gènes n'avaient jamais été associés à la LAL-T et représentent un intérêt particulier dans le cadre du débalancement de l'incidence de la maladie en fonction du sexe (ratio garçon:fille =1.22). De plus, la majorité des nouvelles mutations identifiées l'ont été chez des patients présentant des lymphoblastes bloqués aux stades les plus précoces de différenciation, phénotype généralement associé au plus mauvais pronostic. Cette observation souligne la diversité génomique de cette leucémie qui explique la complexité de prise en charge des patients et démontre le besoin d'un effort de caractérisation approfondie.

Enfin, grâce à l’étude longitudinale de patients LAL-B ayant subi une ou plusieurs rechutes, nous avons mis en évidence des modèles variables de dynamique clonale qui distinguent les rechutes précoces des tardives et suggèrent deux mécanismes de résistance au traitement. Un premier que l’on pourrait qualifier d’“actif”, impliquant l'acquisition rapide de mutations de résistance, et un deuxième, “passif”, suggérant la réactivation de cellules leucémiques en quiescence/dormance qui dans certains cas correspondaient aux cellules composant la population majoritaire au moment du diagnostic.

6.2. Discussion

Malgré les avancées majeures suivant l'avènement du NGS il y a 10 ans, la caractérisation génomique exhaustive des tumeurs représente toujours un défi important tant au niveau de la collecte du matériel biologique et du design expérimental que du point de vu de l'analyse. Parmi les défis auxquels sont confrontés la communauté scientifique, 4 points en particuliers sont abordés au sein de ma thèse et seront discutés plus en détails: i) identification de mutations somatiques au sein d'un séquençage de faible couverture (*low-pass*); ii) caractérisation de *drivers* rares; iii) intégration de données multiples; iv) étude de la dynamique clonale des populations de cellules leucémiques.

6.2.1. L'analyse du séquençage de faible couverture

Bien qu'étant devenu de plus en plus routinière au sein des laboratoires analysant des données de séquençage de cellules tumorales, l'identification de mutations somatiques ponctuelles (SNVs ou indels) demeure une étape clé dans la caractérisation du paysage génomique des cancers. Le choix d'un outil adapté aux besoins et aux données est donc crucial. Certains outils comme Strelka [116], Varscan2 [117] ou MuTect [118] sont devenus des standards, mais leur efficacité est souvent dépendante du contexte. De plus, la nature complexe des génomes de cancer combinée à une faible couverture des régions séquencées *(low-pas*s) ainsi qu'aux problèmes techniques de séquençage et d'alignement peuvent affecter la spécificité et/ou la sensibilité de la plupart de ces outils. Le faible taux de concordance entre ces algorithmes [119,120,147] en est une preuve directe. La majorité des articles effectuant des études comparatives les testent en utilisant des données présentant des couvertures importantes ou, dans le cas contraire, notent généralement une forte dégradation des performances [119,120,147,148]. Pourtant, avec le développement de projets de recherche portant sur des cohortes de plus en plus grandes ainsi que l'entrée progressive du NGS en clinique en tant qu'outil diagnostic, les approches de séquençage systématique de génomes imposent un compromis entre la qualité, la rapidité et le coût de génération des données qui se fait souvent au détriment de la profondeur de couverture. Le défi réside donc dans l'identification de mutations somatiques au sein de données de qualité variable.

En développant SNooPer (chapitre 2: SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing), notre but était de contribuer à la recherche en génomique du cancer en mettant à la disposition de la communauté scientifique un nouvel outil s'adaptant aux données. Cet outil offre une alternative efficace aux algorithmes traditionnels et montre des performances intéressantes y compris dans les conditions les plus difficiles. La flexibilité de la forêt d'arbres décisionnels qu'il exploite (*Random Forest classifier*), couplée à l'entrainement utilisant des données issues d'une technologie indépendante, permettent à SNooPer de se départir des biais techniques et des erreurs systématiques. Le développement d'un modèle basé sur l'entrainement préalable est intéressant puisqu'il permet une analyse qui adapte ses paramètres aux données plutôt que de reposer sur des paramètres définis par l'utilisateur et qui se limitent en général à une petite série de filtres catégoriels ou numériques, dans le meilleur des cas. En rendant SNooPer disponible, nous comblons un vide dans le panel d'outils disponibles en proposant pour la première fois une analyse spécialisée capable de prendre en charge des données sous-optimales et de retenir des mutations qui présentent des caractéristiques causant leur rejet par les autres méthodes. SNooPer a été utilisé pour l'identification des mutations caractérisées dans les papiers 2, 3 et 4 (Whole-exome sequencing of a rare case of familial childhood acute lymphoblastic leukemia reveals putative predisposing mutations in Fanconi anemia genes; A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells; Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations) constituant les chapitres 3 et 4. SNooPer a notamment permis une exploration approfondie des données des études 3 et 4 conduisant à l'identification de mutations somatiques filtrées par les méthodes traditionnelles et pourtant validées par la suite. Ces différences de performance sont reportées dans le papier accompagnant l'outil puisque les jeux de données tests utilisés sont en partie communs aux 3 projets (papiers 1, 3 et 4).

Paradoxalement, les avantages proposés par SNooPer sont directement liés à la particularité d'un entrainement préalable qui représente également sa limitation majeure. L'accès à un jeu de données validées sur une plateforme indépendante et assez large pour être utilisé dans ce contexte peut représenter une complication importante. Afin de réduire l'impact de ce problème, dans le cas où des sous-ensembles de validation ne seraient pas facilement disponibles, nous avons développé une série de modèles de classification représentatifs des designs expérimentaux les plus courants (disponibles sur https://sourceforge.net/projects/snooper/). Ces modèles peuvent être utilisés pour l'analyse de n'importe quel jeu de données comparable à celui utilisé pour l'entrainement. Toutefois, bien que des différences limitées entre les deux jeux de données ne devraient avoir qu'une influence mineure sur l'analyse, des résultats optimaux ne seront obtenus qu'en utilisant des données séquencées grâce à la même chimie, sur la même plateforme ainsi que préparées selon le même processus, incluant les étapes d'alignement et de nettoyage préalable.

Une autre limitation du programme est l'utilisation du format mpileup de SAMtools (http://samtools.sourceforge.net/) comme entrée ce qui borne le nombre de variables informatives utilisables par l'algorithme lors de la classification. Afin de corriger ce biais, une version exploitant directement le format BAM, offrant par exemple plus d'informations sur les spécificités d'alignement, est en cours de développement.

6.2.2. Mutations somatiques rares, les évènements *drivers* oubliés

Récemment, de nombreuses études ont souligné l'importance de prendre en compte l'étendue de l'hétérogénéité mutationnelle tumorale en montrant que chaque type de tumeur ne présente que quelques gènes mutés de manière récurrente contre plusieurs gènes rarement mutés [149-151]. Ce phénomène de "*long-tail*" des gènes peu fréquemment mutés complique la distinction *driver* vs. *passenger*, puisqu'ils présentent des taux mutationnels comparables au sein d'un type tumoral, et rend inefficaces la plupart des méthodes d'identification courantes [133]. Une des limitations récurrentes des travaux publiés réside dans l'identification des *drivers* circonscrits aux évènements redondants au sein d'un type tumoral. Par exemple, une des méthodes populaires basée sur une définition simpliste de ce qu'est un "*driver*" se concentre sur la détection de "*hills*" (pour collines) mutationnelles au sein des génomes étudiés, se limitant ainsi à la redondance entre tumeurs et se dégageant des plaines correspondant au taux mutationnel de base du génome. Le principe même de la méthode néglige la possibilité d'événements sélectionnés dans un nombre très limité de cas étant donné qu'ils n'apportent d'avantages sélectifs qu'au sein de contextes génétiques spécifiques ou d'environnements particuliers [9]. Elle est donc généralement inefficace à l'identification de ces *drivers* contextuels [127] dont l'analyse requerrait une puissance statistique uniquement obtenu par le biais de cohortes d'une taille conséquente [152]. Ainsi, même les études pan-cancers les plus récentes exploitant les données TCGA (The Cancer Genome Atlas) [152-157] ont montré des limites analytiques les empêchant de compléter le catalogue de mutations somatiques [133].

Pourtant, comme nous l’avons montré au niveau des différents travaux qui composent cette thèse, il est possible d’identifier de nouveaux *drivers* rares, y compris au sein de types tumoraux déjà bien caractérisés. Dans le chapitre 3, l'analyse de l'exome de 3 enfants atteints de LAL-B (A novel somatic mutation in ACD induces telomere lengthening and apoptosis resistance in leukemia cells), a mené à l’identification de mutations rares ou privées au potentiel *driver* (*DOT1L* p.V114F, *HCFC1* p.Y103H, *ACD* p.G223V) au sein de groupes moléculaires pourtant courants et largement étudiés (hyperdiploïde et t(12;21)). La validation fonctionnelle de la mutation *ACD* p.G223V a permis de démontrer son impact sur la résistance à l'apoptose et la régulation de la taille des télomères au sein de cellules LAL de type pré-B. Cette étude est d'ailleurs la première à décrire l'implication d'une mutation somatique au niveau d'*ACD* dans le développement de la leucémie. Depuis, le régulateur épigénétique DOT1L a également confirmé son rôle au niveau du processus leucémogène chez les patients présentant une haute hyperdiploïdie [158]. Un processus d'analyse équivalent utilisé dans l'étude de notre cohorte de leucémie LAL-T présenté au chapitre 4 (Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations) a permis d’identifier des mutations uniques comme *USP9X* p.Q117*. Outre l'intérêt de l'identification de nouveaux gènes ou voies de signalisation *drivers*, ces études représentent une preuve de principe démontrant la complexité génétique sous-jacente des cancers. Ainsi, nos données suggèrent que chaque tumeur est composée de *drivers* communs, apportant un avantage sélectif lié à la forte pénétrance des mutations, et de *drivers* rares, voire privés, dont l'effet est dépendant du contexte mutationnel ou physiologique. La variabilité interindividuelle démontrée par ces résultats, y compris au niveau de cancers très bien caractérisés comme la LAL-B, est d'une importance particulière puisqu'elle permet d'expliquer en partie les divergences de réponses au traitement ou d'évolution de la maladie observées à l'intérieur même d'un sous-type de LAL pourtant homogène. Elle justifie le développement de nouvelles méthodes d'analyse permettant une caractérisation systématique des *drivers* communs et rares qui composent les tumeurs étudiées.

Malgré tout, les méthodes d’analyse utilisées présentent néanmoins certaines limitations. Les outils de prédiction qu’elles exploitent entraînent la génération de nombreux faux positifs et négatifs [159]. Une étape complémentaire de réduction de la liste de candidats sans *a priori*, comme le criblage par perte de fonction présenté également au niveau du chapitre 3, permet de limiter les validations fonctionnelles aux gènes présentant un impact significatif lorsque dérégulés. Toutefois, ces cribles basés sur une compétition interclonale, réalisés au sein d’un contexte génétique uniforme, ont tendance à faire ressortir les acteurs majeurs de voies cellulaires clés et donc à sous-estimer l’importance de candidats contextuels.

À l’heure de la génomique translationnelle (ou “*precision medicine*”), de nouvelles méthodes systématiques d’identification des *drivers* rares sont nécessaires. Dans ce cadre, des approches d’analyse par réseau montrent des résultats prometteurs [125,133]. Elles intègrent l’information mutationnelle des gènes à la topologie locale d'interactions entre protéines en utilisant de larges bases de données comme TCGA. HotNet2 [133] a notamment permis de mettre en évidence plusieurs sous-réseaux d’intérêt, comme celui des complexes cohésine et condensine, contenant de nombreux gènes mutés peu fréquemment (~1% des cas) et délaissés par les autres méthodes d’analyses mais présentant toutes les caractéristiques de *drivers* potentiels.

Outre une analyse *in silico* innovante, la caractérisation de nouveaux *drivers* rares nécessite également la mise en place de validations fonctionnelles adaptées. Aujourd'hui, la technologie CRISPR-Cas9, couplée à des xénogreffes cellulaires, offre une possibilité envisageable permettant de se rapprocher de la complexité du contexte tumoral d'origine et ainsi favoriser la mise en évidence du rôle contextuel de ces mutations rares. Dans le cas où des cellules tumorales du patients ne seraient pas disponibles ou exploitables, il est possible d'induire une combinaison de mutations par CRISPR-Cas9 [160] au sein de cellules humaines (pré-leucémiques dans notre cas) afin de se rapprocher du contexte génétique observé, en incluant ou non la mutation d'intérêt. Réaliser ensuite des xénogreffes des deux types cellulaires au sein de modèles murins immunodéficients permettrait d'observer si la combinaison génétique incluant la mutation rare conduit à un phénotype différent confirmant son impact (croissance de la tumeur, invasion, survie des individus). Dans l'éventualité où les cellules primaires du patients seraient disponibles, une xénogreffe au sein de modèles murins immunodéficients d'une lignée cellulaire derivée de ces cellules portant la mutation d'intérêt, et en parallèle de la lignée après correction de la mutation par CRISPR-Cas9, devrait de la même façon permettre de mesurer l'effet de la présence de la mutation sur le développement et l'évolution de la tumeur induite. Cette expérience, réalisée récemment par Antal *et al.* au niveau du gène de la proteine kinase C dans un contexte de cancer du colon [161], conduit à démontrer l'avantage sélectif éventuel apporté par la mutation dans un contexte tumoral spécifique.

6.2.3. Le paysage génomique de la LAL-T de l'enfant: le défi de l’intégration de données

Bien que les mécanismes moléculaires impliqués dans la pathogenèse de la LAL-T ont été largement étudiés, certains restent encore mal compris. De nombreuses études de séquençage génomique à grande échelle portent sur la LAL-T adulte uniquement ou combinent des données adultes et pédiatriques [63,162-165], limitant la caractérisation d'évènements spécifiques aux cohortes pédiatriques [62].

Dans l'étude présenté au chapitre 4 (Genomic characterization of pediatric T-cell acute lymphoblastic leukemia reveals novel recurrent driver mutations), nous avons utilisé une approche intégrative combinant des données immunophénotypiques, génomiques et transcriptomiques, ainsi que des données de génotypage afin de caractériser 30 LAL-T pédiatriques (diagnostic, rémission et rechute lorsque disponibles). Ces données nous ont permis de classer nos patients en fonction du degré de maturation des cellules tumorales. Cette étape, particulièrement critique lorsqu'elle concerne les tumeurs les plus immatures comme celle du groupe ETP-ALL qui présentent des profils immunophénotypique et transcriptomique comparables à des progéniteurs myéloïdes immatures et à des cellules souches hématopoïétiques, conduit fréquemment à un diagnostic erroné et contribue à la faible caractérisation des groupes ambigus [62,164,166]. Nous avons ensuite déterminé les combinaisons de mutations somatiques connues ou nouvelles (incluant SNVs, indels, CNVs et réarrangements chromosomiques) présentes dans chaque tumeur et spécifiques à chaque groupe de maturation. Nous avons identifié de nouveaux *drivers* associés à LAL-T pour la première fois (*U2AF1*, *MED12* et *USP9X*). À l'aide d'analyses fonctionnelles *in vitro*, nous avons démontré l'épissage aberrant provoqué par la mutation récurrente (p.R35L) identifiée au niveau d'U2AF1, un des membres du spliceosome. Nous avons également démontré que la perte de fonction de MED12, membre du complexe Mediator, ainsi que de la déubiquitinase USP9X, codé par deux gènes suppresseurs de tumeurs liés à l'X, conféraient une résistance à l'apoptose après traitements chimiothérapiques. Soixante pourcent des mutations nouvellement identifiées l'ont été au niveau de LAL immatures, en particulier chez les 2 ETP-ALLs de la cohorte, soulignant la complexité sous-jacente du paysage génomique de LAL-T et la nécessité d'études intégratives à grande échelle permettant de comprendre les mécanismes biologiques de cette leucémie et de ses divers sous-types. L'immaturité des cellules du groupe ETP-ALL représente un phénotype à haut risque associé à une mauvaise réponse à la chimiothérapie entraînant un risque élevé d'échec de l'induction et de rechute. L'identification de nouveaux gènes impliqués dans la pathogenèse de ce groupe fournit des pistes de solution pour optimiser le pronostic et proposer de nouvelles stratégies de traitement basées sur le profil mutationnel spécifique de ce sous-type.

L'identification de *MED12* comme un nouveau *driver* est particulièrement intéressante. Ce résultat fait écho aux travaux récents d'Aifantis et de son équipe qui, à travers la caractérisation du lncRNome de la LAL-T, a identifié LUNAR1 (LeUkemia-induced Noncoding Activator RNA-1). Ce long ARN non codant régulé par la voie Notch contribue au développement de la leucémie en modulant la liaison du complexe Mediator en agissant sur MED1 et MED12 [167]. Bien que des études complémentaires soient nécessaires pour comprendre la contribution de la perte de fonction de *MED12* à la leucémogenèse, conjointement ou non aux mutations activant *NOTCH1*, notre étude appuie l'importance du rôle du complexe Mediator dans le développement de la LAL-T pédiatrique. En outre, bien que l'étude n'ait pas encore été publiée, des mutations au niveau de *MED12* et *USP9X* ont été récemment rapportées chez 3% des patients LAL-T pédiatriques d'une cohorte analysée par l'équipe de Charles Mullighan (http://www.bloodjournal.org/content/126/23/691?sso-checked=true), confirmant le potentiel *driver* de ces gènes.

De plus, *MED12* et *USP9X* représentent un intérêt particulier dans le contexte du débalancement de l'incidence de la maladie en fonction du sexe pour lequel une contribution génétique semble de plus en plus évidente [145]. En effet, l'accumulation de mutations perte de fonction au niveau de gènes suppresseurs de tumeurs liés au chromosome X pourrait influencer la distribution en fonction du sexe. Jusqu'à maintenant, *KDM6A* (*UTX*) qui présente des mutations inactivatrices spécifiques aux patients masculins [165] et échappe à l'inactivation du X contrairement à *PHF6* [162], représentait le candidat le plus solide pour influencer ce ratio. Toutefois, le taux mutationnel de *KDM6A* au niveau de la LAL-T ne permet pas à lui seul d'expliquer l'effet observé. Dans le chapitre 4, outre la démonstration de l'effet fonctionnel des pertes de fonction de *MED12* et *USP9X*, nous montrons la possibilité d'un échappement à l'inactivation du X pour le gène *MED12*, celui d'*USP9X* ayant déjà été démontré [168]. Ces résultats font de *MED12* et *USP9X* des candidats supplémentaires pouvant participer au débalancement.

Bien que notre étude aille dans le sens d'une contribution génétique, nous ne pouvons pas exclure la possibilité que d'autres mécanismes sous-jacents, tels que les différences hormonales, puissent également contribuer à l'écart sexuel observé dans l'incidence de la LAL-T. En effet, le fait que le biais sexuel ait été identifié comme plus prononcé chez les femmes en âge de procréer sur la base de l'analyse de 5,202 patients atteints de la LAL [169] pourrait en effet suggérer un rôle hormonal. En fait, de nombreux facteurs sous-jacents pourraient contribuer à la différence de genre observée, y compris les hormones sexuelles, les différences génétiques et épigénétiques ainsi que les causes environnementales [170-177]. Il a été démontré que les facteurs hormonaux agissent sur l'activité des lymphocytes B et T par le biais de récepteurs liés à la membrane qui influencent des voies de signalisation cellulaire situées en aval comme NF-kB et JAK-STAT [178]. D'ailleurs, des associations genre-spécifiques impliquant l'allèle rs12203592 de l'interféron-4 (IRF4) couplé à l'absence d'inhibition de la voie NF-kB par les œstrogènes, régulant elle-même IRF4, ont été identifiées [179]. Toutefois, les niveaux hormonaux à l'âge des patients de l'étude (âge moyen au diagnostic: 11.8 ans) sont très similaires quelque soit le sexe [176], ce qui suggère une influence limitée. Une programmation prénatale de la régulation de l'expression des gènes par les œstrogènes a été suggérée [180], mais des recherches plus poussées sont nécessaires pour expliquer les mécanismes impliqués et y voir une association avec le biais observé au niveau de la LAL-T pédiatrique.

Une des limitations de ce travail est la taille réduite de la cohorte. Les associations identifiées, l'influence éventuelle des pertes de fonctions au niveau de gènes suppresseurs de tumeurs sur le ratio garçon:fille, ainsi que les taux mutationnels des gènes mis en exergue devront être validés au sein de cohortes plus importantes. On peut d'ailleurs considérer que cette limitation concerne l'ensemble des travaux composant cette thèse.

Une autre limitation réside dans le mode d'exploitation des données multiples. En raison de la quantité/qualité limitée de l'ARN pour certains de nos patients, le séquençage ne pouvait être réalisé que pour 11 des 30 cas de l'étude. Ces données ont néanmoins été utilisées, en conjonction avec des données génomiques, cytogénétiques et immunophénotypiques, pour la classification de nos patients LAL-T, l'identification d'anomalies chromosomiques cryptiques (t(10;11)(p12;q14)) ainsi que l'analyse de l'expression de gènes d'intérêts dans lesquels de nouvelles mutations ont été identifiées (*MED12*, *USP9X* et *U2AF1*). Bien que ces données n'aient été que partielles, notre objectif était d'utiliser et d'intégrer toutes les sources de données disponibles pour améliorer l'interprétation globale de nos résultats. Toutefois, comme le rappelait très récemment Elaine Mardis [115], le défi de l'intégration ne réside pas exclusivement dans l’analyse indépendante de plusieurs types de données comme ici, ou de tests de corrélations un à un *a posteriori*, mais plutôt dans une analyse méta-dimensionnelle combinant simultanément l'ensemble des informations [181,182]. Ainsi, le but est d’intégrer les différentes dimensions génomiques, transcriptomiques, épigénomiques, protéomiques et métabolomiques pouvant contribuer aux phénotypes observés. Afin de mieux caractériser les génomes de cancers, cette intégration de données doit maintenant devenir un des objectifs principaux de la communauté. Bien que certains algorithmes comme DriverNet [183] et OncoIMPACT [184] permettent une priorisation des gènes mutés en prenant en compte leurs connections aux gènes dérégulés au sein des données d’expression correspondantes, l’intégration reste limitée à deux dimensions de données. Les progrès modestes réalisés ces dernières années dans le déchiffrage de l’architecture “multi-omiques” des phénotypes complexes, en particulier au niveau des cancers, démontrent le besoin d’exploiter des méthodes plus complètes.

6.2.4. La dynamique des populations cellulaires à l'origine des rechutes

Caractériser la dynamique clonale des cellules à l'origine des rechutes ainsi que les évènements génomiques contribuant aux mécanismes de résistance est essentiel à la mise en place de stratégies thérapeutiques les contournant. Au niveau du chapitre 5 (Mutational dynamics of early and late relapsed childhood B-cell acute lymphoblastic leukemia: rapid clonal expansion and long-term dormancy), à travers l'analyse des données de génotypage et de séquençage d'exome/génome d'échantillons sériels de tumeurs primaires et de rechutes uniques ou multiples, nous avons déchiffré l'évolution spatio-temporelle des clones composants les tumeurs de 19 patients LAL-B. Notre objectif était d'évaluer la dynamique de deux types de rechutes qui se distinguent par le délai d'apparition de l'évènement (avant ou après 36 mois suivant le diagnostic). Nous avons distingué deux modèles spécifiques à chaque groupe: un modèle hautement dynamique/adaptatif spécifique aux rechutes précoces qui illustre l'émergence rapide des clones les plus adaptés aux nouvelles conditions; et un modèle d'évolution lent, voire quasi-inerte dans certains cas, qui suggère la réactivation d'une population cellulaire en dormance conduisant à l'apparition d'événements tardifs. Ces deux modes ont été capturés que ce soit par le biais d'une quantification de la dynamique clonale générale ou par la mesure de l'apparition de nouvelles sous-clonalités dans la tumeur.

Nos travaux se distinguent de la majorité des projets portant sur l'étude de l'impact du traitement sur les populations cellulaires tumorales [76,185,186] en proposant, lorsque possible, un suivi longitudinal permettant de saisir la complexité de l'évolution temporelle. Outre la rareté de ce *design* expérimental liée à la complexité de la collecte de tels échantillons, à notre connaissance, cette étude est la première à capturer les deux modes de dynamique régissant les rechutes tardives et précoces par le biais de l'étude de la fluctuation de la fréquence allélique des mutations somatiques. Ce décryptage séquentiel est pourtant essentiel à l'identification de l'ensemble des déterminants de l'échec thérapeutique [187]. De plus, l'accès au matériel tumoral prélevé au moment du diagnostic et jusque 6 ans plus tard, au moment de la rechute, permet d'interroger l'origine des cellules initiant les rechutes à long terme, mécanisme peu documenté faute de *design* adéquat. Jusqu'à récemment, les données attribuant les récurrences tardives à l'émergence de clones dérivés de la population originale au moment du diagnostic étaient très limitées et se cantonnaient généralement à l'utilisation des réarrangements physiologiques des gènes d'immunoglobuline (IGH/IGK) comme marqueurs spécifiques des clones [188-190]. Ici, l'analyse des données de séquençage nous a permis de montrer que dans certains cas une majorité de mutations identifées au niveau de la tumeur primaire sont partagées avec la rechute tardive. L'identité génotypique identifiée ici entre les populations clonales de la tumeur primaire et de la rechute fournit sans ambiguité la preuve d'une ré-émergence possible de la population prédominante au moment du diagnostic. Ces données sont contradictoires avec les résultats obtenus récemment par l'équipe de Mel Greaves qui suggéraient la réactivation de cellules souches pre-leucémiques distinctes ne partageant aucune mutation avec la population de la tumeur primaire mise à part les évènements les plus ancestraux (ex: fusion *BCR-ABL1*) [18]. Cette divergence est particulièrement intéressante. L'absence d'évolution clonale à la suite d'une rechute implique que les cellules leucémiques n'ont subi aucun changement biologique pendant leur dormance prolongée. Si tel est le cas, il est probable que la réapparition de la leucémie ait été provoquée par un relachement de la surveillance immunitaire de l'hôte contrôlant jusque-là le clone [188], plutôt qu'à une évolution du clone permettant la mise en place d'un mécanisme d'échappement à cette surveillance, comme pourraient le suggèrer les travaux de Mel Greaves [18]. Si l'absence de changement au niveau moléculaire est confirmé dans la majorité des rechutes tardives, cela représente un atout clinique puisque la leucémie conserve sa chimiosensibilité. Dans ce cadre, le traitement permet d'aboutir rapidement à une nouvelle rémission comme nous l'avons observé ici.

Le faible nombre de cas étudiés (8 rechutes précoces et 11 tardives) représente ici aussi une des limitations majeures de ce projet, en particulier dans le cas de résultats contradictoires avec certains travaux publiés. Par conséquent, bien que les profils obtenus pour les deux groupes soient significativement différents malgré la taille de notre échantillonnage, une validation utilisant les données d'une cohorte étendue renforcerait les conclusions de l'étude.

De plus, en dépit d'une couverture de séquençage relativement profonde (couverture exonique moyenne =200X), elle reste limitante et ne permet pas une caractérisation exhaustive de l'architecture sous-clonale. En effet, le processus courant de séquençage se fait à partir d'une masse cellulaire. Cette méthode, utilisée dans les différents projets qui composent cette thèse, entraîne plusieurs limitations qui ne laissent entrevoir qu'une image partielle et simplifiée de la réalité. L'utilisation de telles données afin d'étudier la structure de la population cellulaire se fait en extrapolant le fait que la fréquence allélique des mutations somatiques identifiées est représentative de la proportion de clones mutés dans la masse de cellules séquencées. Le premier problème est inhérent à ce principe: bien que compensé par un séquençage de plus en plus “profond” [191], la limite de détection des sous-clonalités atteint rarement des valeurs inférieures à 0.01 de fréquence allélique pour 1,000X de profondeur [192,193]. Cette limite laisse inexplorées de multiples sous-clonalités qui, même en représentant moins de 1% de la masse tumorale totale, diversifient le phénotype de la tumeur [194]. La déconvolution de ces données pose un autre problème. Regrouper des mutations identifiées au sein d’une tumeur, dont la proportion de contamination en cellules normales est souvent inconnue, pour en déduire un nombre indéterminé de sous-clones est un défi de taille lorsque la technologie utilisée produit des séquences courtes rendant particulièrement complexe la détermination du “*phasing*” des variations génétiques [195]. De nouvelles technologies de séquençage sont néanmoins disponibles et devraient permettre à l’avenir de contourner ces difficultés. Le séquençage de cellules uniques, par exemple, a le potentiel de déchiffrer la complexité tumorale à sa plus petite échelle et donc de résoudre le problème de sensibilité. Toutefois, la technologie est encore émergente et le biais imposé par la pré-amplification nécessaire au séquençage d’une molécule unique entraîne la génération de nombreuses erreurs rendant les données difficiles à interpréter [196-197]. De plus, le coût de leur acquisition reste prohibitif. Les technologies produisant de longues séquences (>1 kilobase) peuvent quant à elles résoudre le problème de *phasing*.

Enfin, l'identification de deux phénotypes de dynamique clonale soulève la question de l'identification des éventuels déterminants génétiques les causant, question abordée seulement partiellement dans cette étude. Dans ce contexte, le problème se pose sans doute différemment en fonction du phénotype. En ce qui concerne celui des rechutes précoces, l'hypothèse d'une dynamique forte alimentée à l'origine par l'apparition de mutations somatiques dans un ou plusieurs gènes d'une des voies de réparation des dommages à l'ADN est une hypothèse séduisante. D'ailleurs, l'identification d'une surreprésentation de cas du groupe précoce avec des mutations au niveau de ces gènes (62.5% contre 18.2% des patients avec rechutes tardives) va dans ce sens. Pour vérifier cet effet, non significatif à l'échelle de notre cohorte, il est indispensable d'augmenter notre pouvoir statistique. Dans ce cas, l'étude du matériel de rechute d'une cohorte étendue incluant les deux phénotypes est indispensable. Pour ce qui est des rechutes tardives, la solution est probablement plus complexe. Si la forte plasticité observée pour le premier groupe présente sans doute une origine somatique, celle du phénotype de dormance pourrait être germinale, somatique ou issue d'une interaction entre les deux “génomes”. Divers mécanismes ont déjà été proposés pour tenter d'expliquer les phénomènes de rechutes tardives, incluant une prolifération contrôlée, une activité angiogénique limitée ainsi qu'une interaction avec le micro-environnement de la niche stromale maintenant les cellules tumorales hors de portée de la surveillance immunitaire, éventuellement par le biais d'une signalisation chimique [198-200]. Si des polymorphismes participent au phénotype observé, une étude d'association basée sur des données de séquençage [201] (“*sequencing-based association study*”) de cas présentant des phénotypes marqués est envisageable. Toutefois, le pouvoir statistique limité dû à la rareté des cas de rechutes tardives risquent de la compromettre. Si le phénotype est lié à des modifications somatiques, une étude similaire à celle proposée pour le premier phénotype devrait permettre de les identifier. Toutefois, la diversité des mécanismes pouvant expliquer le phénotype imposent une analyse sans *a priori*, intégrant les mutations somatiques les plus rares et tenant compte des réseaux d'interaction des protéines concernées. Enfin, une analyse intégrant une mesure de l'interaction entre polymorphismes et mutations somatiques pourrait également être intéressante dans ce contexte.

6.3. Conclusion

De manière générale, cette thèse a eu pour objectif de contribuer à l'identification et à la caractérisation des déterminants et processus génomiques qui constituent la variabilité inter- et intra-tumorale des leucémies pédiatriques. Elle permet non seulement d'améliorer l'état des connaissances spécifiques de la biologie de la LAL pédiatrique en mettant en évidence de nouveaux acteurs moléculaires de la maladie, mais interroge également sur des principes plus larges qui pourraient être extrapolés à d'autres types de tumeurs tels que l'implication des mutations contextuelles dans le développement de cancer ou la dynamique évolutive conduisant aux rechutes.

De manière générale, nous espérons qu'elle puisse à terme constituer l'une des pierres de l'édifice que représente le prochain grand défi de notre communauté scientifique: une caractérisation exhaustive, intégrative et systématique des tumeurs et de leurs rechutes. Ultimement, ces connaissances devraient permettre une meilleure stratification des cas, promouvoir la recherche translationnelle et aboutir à une prise en charge personnalisée des patients.

RÉFÉRENCES

1. Mitra D, Shaw AK, Hutchings K. Trends in incidence of childhood cancer in Canada, 1992-2006. Chronic Dis Inj Can. 2012;32(3):131-9.

2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7-30.

3. Greaves M. Leukaemia firsts in cancer research and treatment. Nat Rev Cancer. 2016;16(3):163-72.

4. Downing JR, Wilson RK, Zhang J, Mardis ER, Pui CH, Ding L, et al. The Pediatric Cancer Genome Project. Nat Genet. 2012;44(6):619-22.

5. Orkin SH, Zon LI. SnapShot: hematopoiesis. Cell. 2008;132(4):712. doi:10.1016/j.cell.2008.02.013.

6. Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. Cell. 2008;132(4):631-44. doi: 10.1016/j.cell.2008.01.025.

7. Larsson J, Karlsson S. The role of Smad signaling in hematopoiesis. Oncogene. 2005;24(37):5676-92. Review.

8. Kennedy JA, Barabé F. Investigating human leukemogenesis: from cell lines to in vivo models of human leukemia. Leukemia. 2008;22(11):2029-40. doi: 10.1038/leu.2008.206.

9. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al., 2007. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446(7132):153-8.

10. Colby-Graham MF, Chordas C. The childhood leukemias. J Pediatr Nurs. 2003;18(2):87-95.

11. Passegué E, Jamieson CH, Ailles LE, Weissman IL. Normal and leukemic hematopoiesis: are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? Proc Natl Acad Sci U S A. 2003;100 Suppl 1:11842-9.

12. Kreso A, Dick JE. Evolution of the cancer stem cell model. Cell Stem Cell. 2014;14(3):275-91. doi: 10.1016/j.stem.2014.02.006.

13. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13. doi: 10.1038/nature10762.

14. Quail DF1, Taylor MJ, Postovit LM. Microenvironmental regulation of cancer stem cell phenotypes. Curr Stem Cell Res Ther. 2012;7(3):197-216.

15. Chaffer CL, Weinberg RA. How does multistep tumorigenesis really proceed? Cancer Discov. 2015;5(1):22-4. doi: 10.1158/2159-8290.CD-14-0788.

16. Plaks V, Kong N, Werb Z. The Cancer Stem Cell Niche: How Essential Is the Niche in Regulating Stemness of Tumor Cells? Cell Stem Cell. 2015;16(3):225-38. doi: 10.1016/j.stem.2015.02.015.

17. Greaves M. Cancer stem cells renew their impact. Nat Med. 2011;17(9):1046-8. doi: 10.1038/nm.2458.

18. Ford AM, Mansur MB, Furness CL, van Delft FW, Okamura J, Suzuki T, et al. Protracted dormancy of pre-leukemic stem cells. Leukemia. 2015;29(11):2202-7. doi: 10.1038/leu.2015.132.

19. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group. Ann Intern Med. 1985;103(4):620-5.

20. Nemazee D. Receptor editing in lymphocyte development and central tolerance. Nat Rev Immunol. 2006;6(10):728-40. Review.

21. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet. 2013;381(9881):1943-55. doi: 10.1016/S0140-6736(12)62187-4. Review.

22. Ries LAG, Smith MA, Gurney JG, Linet M, Tamra T, Young JL, et al. Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995, National Cancer Institute, SEER Program. Leuk Res. 2009;33(3):355-62. doi: 10.1016/j.leukres.2008.08.022.

23. Lim JY, Bhatia S, Robison LL, Yang JJ. Genomics of racial and ethnic disparities in

childhood acute lymphoblastic leukemia. Cancer 2014;120:955-62.

24. Hunger SP, Mullighan CG. Acute Lymphoblastic Leukemia in Children. N Engl J Med. 2015;373(16):1541-52. doi: 10.1056/NEJMra1400972. Review.

25. Graux C. Biology of acute lymphoblastic leukemia (ALL): clinical and therapeutic relevance. Transfus Apher Sci. 2011;44(2):183-9. doi: 10.1016/j.transci.2011.01.009. Review.

26. Gale KB, Ford AM, Repp R, Borkhardt A, Keller C, Eden OB, Greaves MF. Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. Proc Natl Acad Sci U S A. 1997;94(25):13950-4.

27. Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, et al. Prenatal origin of acute lymphoblastic leukaemia in children. Lancet. 1999;354(9189):1499-503.

28. Greaves MF, Maia AT, Wiemels JL, Ford AM. Leukemia in twins: lessons in natural history. Blood. 2003;102(7):2321-33.

29. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. Nat Rev Cancer. 2003;3(9):639-49.

30. Yuan Y, Zhou L, Miyamoto T, Iwasaki H, Harakawa N, Hetherington CJ, et al. AML1–ETO expression is directly involved in the development of acute myeloid leukemia in the presence of additional mutations. Proc Natl Acad Sci U S A. 2001;98(18):10398-403.

31. Higuchi M, O'Brien D, Kumaravelu P, Lenny N, Yeoh EJ, Downing JR. Expression of a conditional AML1–ETO oncogene bypasses embryonic lethality and establishes a murine model of human t(8;21) acute myeloid leukemia. Cancer Cell. 2002;1(1):63-74.

32. Daley, G. Q., Van Etten, R. A. & Baltimore, D. Induction of chronic myelogenous leukemia in mice by the P210 bcr/abl gene of the Philadelphia chromosome. Science. 1990;247(4944):824-30.

33. Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proc Natl Acad Sci U S A. 2002;99(12):8242-7.

34. Greaves M. Infection immune responses and the aetiology of childhood leukaemia. Nat Rev Cancer. 2006;6(3):193-203.

35. Greaves M. When one mutation is all it takes. Cancer Cell. 2015;27(4):433-4. doi: 10.1016/j.ccell.2015.03.016.

36. Horwitz M. The genetics of familial leukemia. Leukemia. 1997;11(8):1347-59

37. Hunger SP, Mullighan CG. Redefining ALL classification: toward detecting high-risk ALL and implementing precision medicine. Blood. 2015;125(26):3977-87. doi: 10.1182/blood-2015-02-580043.

38. Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. Blood. 2007;109(2):683-92.

39. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet. 2009;41(9):1006-10.

40. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet. 2009; 41(9):1001-5.

41. Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Koehler R, Greaves M, et al. Verification of the susceptibility loci on 7p12.2, 10q21.2, and 14q11.2 in precursor B-cell acute lymphoblastic leukemia of childhood. Blood. 2010;115(9):1765-7.

42. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. J Natl Cancer Inst. 2013;105(10):733-42.

43. Migliorini G, Fiege B, Hosking FJ, Ma Y, Kumar R, Sherborne AL, et al. Variation at

10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. Blood. 2013;122(19):3298-307.
44. Perez-Andreu V, Roberts KG, Harvey RC, Yang W, Cheng C, Pei D, et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. Nat Genet. 2013;45(12):1494-8.
45. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. Haematologica. 2010;95(9):1608-11. doi: 10.3324/haematol.2010.022459.
46. Holmfeldt L, Wei L, Diaz-Flores E, Walsh M, Zhang J, Ding L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45(3):242-52.
47. Shah S, Schrader KA, Waanders E, Timms AE, Vijai J, Miething C, et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. Nat Genet. 2013;45(10):1226-31.
48. Noetzli L, Lo RW, Lee-Sherick AB, Callaghan M, Noris P, Savoia A, et al. Germline mutations in ETV6 are associated with thrombocytopenia, red cell macrocytosis and predisposition to lymphoblastic leukemia. Nat Genet. 2015;47(5):535-8. doi: 10.1038/ng.3253.
49. Kim AS, Eastmond DA, Preston RJ. Childhood acute lymphocytic leukemia and perspectives on risk assessment of early-life stage exposures. Mutat Res. 2006;613(2-3):138-60.
50. Preston DL, Kusumi S, Tomonaga M, Izumi S, Ron E, Kuramoto A, et al. Cancer incidence in atomic bomb survivors. Part III. Leukemia, lymphoma and multiple myeloma, 1950-1987. Radiat Res. 1994;137(2 Suppl):S68-97.
51. Doll R, Wakeford R. Risk of childhood cancer from fetal irradiation. Br J Radiol. 1997;70:130-9.
52. Ward, G. The infective theory of acute leukaemia. Br. J. Child. Dis. 1917;14:10-20.
53. Kinlen L. Evidence for an infective cause of childhood leukaemia: comparison of a Scottish New Town with nuclear reprocessing sites in Britain. Lancet. 1988;2(8624):1323-7.
54. Greaves MF. Speculations on the cause of childhood acute lymphoblastic leukemia. Leukemia. 1988; 2:120-25.
55. Neveu B, Spinella JF, Richer C, Lagacé K, Cassart P, Lajoie M, et al. CLIC5: a novel ETV6 target gene in childhood acute lymphoblastic leukemia. Haematologica. 2016;101(12):1534-43.
56. Ford AM, Bennett CA, Price CM, Bruin MC, Van Wering ER, Greaves M. Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia. Proc Natl Acad Sci U S A. 1998;95(8):4584-8.
57. Swaminathan S, Klemm L, Park E, Papaemmanuil E, Ford A, Kweon SM, et al. Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia. Nat Immunol. 2015;16(7):766-74. doi: 10.1038/ni.3160.
58. Pui CH, Carroll WL, Meshinchi S, Arceci RJ. Biology, Risk Stratification, and Therapy of Pediatric Acute Leukemias: An Update. J Clin Oncol. 2011;29(5):551-65. doi: 10.1200/JCO.2010.30.7405.
59. Tsuzuki S, Seto M, Greaves M, Enver T. Modeling first-hit functions of the t(12;21) TEL-AML1 translocation in mice. Proc Natl Acad Sci U S A. 2004 ;101(22):8443-8.
60. Fischer M, Schwieger M, Horn S, Niebuhr B, Ford A, Roscher S, et al. Defining the oncogenic function of the TEL/AML1 (ETV6/RUNX1) fusion protein in a mouse model. Oncogene. 2005;24(51):7579-91.
61. Salesse S, Verfaillie CM. BCR-ABL: from molecular mechanisms of leukemia induction to treatment of chronic myelogenous leukemia. Oncogene. 2002;21(56):8547-59.
62. Zhang J, Ding L, Holmfeldt L, Wu G, Heatley SL, Payne-Turner D, et al.The genetic basis

of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012;481(7380):157-63. doi: 10.1038/nature10725.

63. De Keersmaecker K, Atak ZK, Li N, Vicente C, Patchett S, Girardi T, et al. Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. Nat Genet. 2013; 45(2):186-90.

64. Roberts KG, Li Y, Payne-Turner D, Harvey RC, Yang YL, Pei D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. N Engl J Med. 2014;371(11):1005-15.

65. Andersson AK, Ma J, Wang J, Chen X, Gedman AL, Dang J, et al; St. Jude Children's Research Hospital-Washington University Pediatric Cancer Genome Project. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. Nat Genet. 2015;47(4):330-37.

66. Virely C, Moulin S, Cobaleda C, Lasgi C, Alberdi A, Soulier J, et al. Haploinsufficiency of the IKZF1 (IKAROS) tumor suppressor gene cooperates with BCR-ABL in a transgenic model of acute lymphoblastic leukemia. Leukemia. 2010;24(6):1200-4.

67. Joshi I, Yoshida T, Jena N, Qi X, Zhang J, Van Etten RA, et al. Loss of Ikaros DNA-binding function confers integrin-dependent survival on pre-B cells and progression to acute lymphoblastic leukemia. Nat Immunol. 2014;15(3):294-304. doi: 10.1038/ni.2821.

68. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature. 2007;446(7137):758-64.

69. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LA, Miller CB, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med. 2009;360(5):470-80. doi: 10.1056/NEJMoa0808253.

70. Zhang J, Mullighan CG, Harvey RC, Wu G, Chen X, Edmonson M, et al. Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood. 2011;118(11):3080-7.

71. Hebert J, Cayuela JM, Berkeley J, Sigaux F. Candidate tumor-suppressor genes MTS1 (p16INK4A) and MTS2 (p15INK4B) display frequent homozygous deletions in primary cells from T- but not from B-cell lineage acute lymphoblastic leukemias. Blood. 1994;84(12):4038-44.

72. Weng AP, Ferrando AA, Lee W, Morris JP 4th, Silverman LB, Sanchez-Irizarry C, et al. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. Science. 2004;306(5694):269-71.

73. Van Vlierberghe P, Ferrando A. The molecular basis of T cell acute lymphoblastic leukemia. J Clin Invest. 2012;122(10):3398-406.

74. Silverman LB, Stevenson KE, O'Brien JE, Asselin BL, Barr RD, Clavell L, et al. Long-term results of Dana-Farber Cancer Institute ALL Consortium protocols for children with newly diagnosed acute lymphoblastic leukemia (1985–2000). Leukemia. 2010;24(2):320-34. doi: 10.1038/leu.2009.253.

75. Hunger SP, Lu X, Devidas M, Camitta BM, Gaynon PS, Winick NJ, et al. Improved survival for children and adolescents with acute lymphoblastic leukemia between 1990 and 2005: a report from the children's oncology group. J Clin Oncol. 2012;30(14):1663-9. doi: 10.1200/JCO.2011.37.8018.

76. Ma X, Edmonson M, Yergeau D, Muzny DM, Hampton OA, Rusch M, et al. Rise and fall of subclones from diagnosis to relapse in pediatric B all. Nat Commun. 2015;6:6604. doi: 10.1038/ncomms7604.

77. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012;481(7382):506-10. doi: 10.1038/nature10738.

78. Walter MJ, Shen D, Ding L, Shao J, Koboldt DC, Chen K, et al. Clonal architecture of secondary acute myeloid leukemia. N Engl J Med. 2012;366(12):1090-8. doi: 10.1056/NEJMoa1106968.
79. Mullighan CG, Phillips LA, Su X, Ma J, Miller CB, Shurtleff SA, et al. Genomic anlysis of the clonal origins of relapsed all. Science. 2008;322(5906):1377-80. doi: 10.1126/science.1164266.
80. Nowell PC. The Clonal Evolution of Tumor Cell Populations. Science. 1976;194(4260):23-8.
81. Greaves M. Darwin and evolutionary tales in leukemia. The Ham-Wasserman Lecture. Hematology Am Soc Hematol Educ Program. 2009:3-12. doi: 10.1182/asheducation-2009.1.3.
82. Merlo LMF, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nat Rev Cancer. 2006;6:924-35.
83. Zhang J, Niu C, Ye L, Huang H, He X, Tong WG. Identification of the haematopoietic stem cell niche and control of the niche size. Nature. 2003;425(6960):836-41.
84. Scadden DT. The stem cell niche in health and leukemic disease. Best Pract Res Clin Haematol. 2007;20:19-27.
85. Gatenby RA, Gillies RJ. A microenvironmental model of carcinogenesis. Nat Rev Cancer. 2008;8:56-61.
86. Lathia JD, Heddleston JM, Venere M, Rich JN. Deadly teamwork: neural cancer stem cells and the tumor microenvironment. Cell Stem Cell. 2011;8:482-5.
87. Anderson AR, Weaver AM, Cummings PT, Quaranta V. Tumor morphology and phenotypic evolution driven by selective pressure from the microenvironment. Cell. 2006; 127:905-15. S0092-8674(06)01348-1 [pii]. 10.1016/j. Cell.2006.09.042
88. Stratton MR. Exploring the genomes of cancer cells: progress and promise. Science. 2011; 331:1553-8.10.1126/science.1204040.
89. Klein CA. Parallel progression of primary tumours and metastases. Nature reviews. Cancer. 2009; 9:302-12.10.1038/nrc2627.
90. Oosterhuis JW, Looijenga LH. Testicular germ-cell tumours in a broader perspective. Nature reviews. Cancer. 2005; 5:210-22.10.1038/nrc1568.
91. Fleming HE, Janzen V, Lo Celso C, Guo J, Leahy KM, Kronenberg HM, et al. Wnt signaling in the niche enforces hematopoietic stem cell quiescence and is necessary to preserve self-renewal in vivo. Cell Stem Cell. 2008;2:274-83.
92. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467:1114-7. nature09515 [pii]. 10.1038/nature09515.
93. Bateman CM, Colman SM, Chaplin T, Young BD, Eden TO, Bhakta M, et al. Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. Blood. 2010; 115:3553-8.
94. Mullighan CG, Zhang J, Kasper LH, Lerach S, Payne-Turner D, Phillips LA, et al. CREBBP mutations in relapsed acute lymphoblastic leukaemia. Nature. 2011;471(7337):235-9. doi: 10.1038/nature09727.
95. Inthal A, Zeitlhofer P, Zeginigg M, Morak M, Grausenburger R, Fronkova E, et al. CREBBP HAT domain mutations prevail in relapse cases of high hyperdiploid childhood acute lymphoblastic leukemia. Leukemia. 2012;26(8):1797-803. doi: 10.1038/leu.2012.60.
96. Meyer JA, Wang J, Hogan LE, Yang JJ, Dandekar S, Patel JP, et al. Relapse specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. Nat Genet. 2013;45(3):290-4. doi: 10.1038/ng.2558.
97. Tzoneva G, Perez-Garcia A, Carpenter Z, Khiabanian H, Tosello V, Allegretta M, et al. Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in

relapsed ALL. Nat Med. 2013;19(3):368-71.

98. Mardis ER. The translation of cancer genomics: time for a revolution in clinical cancer care. Genome Med. 2014;6(3):22. doi: 10.1186/gm539. eCollection 2014.

99. Bonilla X, Parmentier L, King B, Bezrukov F, Kaya G, Zoete V, et al. Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. Nat Genet. 2016;48(4):398-406.

100. Krauthammer M, Kong Y, Bacchiocchi A, Evans P, Pornputtapong N, Wu C, et al. Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. Nat Genet. 2015;47(9):996-1002.

101. Al-Ahmadie HA, Iyer G, Lee BH, Scott SN, Mehra R, Bagrodia A, et al. Frequent somatic CDH1 loss-of-function mutations in plasmacytoid variant bladder cancer. Nat Genet. 2016;48(4):356-8.

102. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44(6):685-9.

103. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012;487(7406):239-43.

104. Giannakis M, Hodis E, Jasmine Mu X, Yamauchi M, Rosenbluh J, Cibulskis K, et al. RNF43 is frequently mutated in colorectal and endometrial cancers. Nat Genet. 2014;46(12):1264-6.

105. Tan J, Ong CK, Lim WK, Ng CC, Thike AA, Ng LM, et al. Genomic landscapes of breast fibroepithelial tumors. Nat Genet. 2015 Nov;47(11):1341-5.

106. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486(7403):395-9.

107. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012;486(7403):353-60. doi: 10.1038/nature11143.

108. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400-4.

109. Rausch T, Jones DT, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012;148(1-2):59-71.

110. Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga J, et al. Integrated molecular analysis of adult T cell leukemia/lymphoma. Nat Genet. 2015;47(11):1304-15.

111. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. PLoS One. 2012;7(10):e46688. doi: 10.1371/journal.pone.0046688.

112. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. 2011;44(1):47-52.

113. Nordlund J, Bäcklin CL, Wahlberg P, Busche S, Berglund EC, Eloranta ML, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biol. 2013;14(9):r105. doi: 10.1186/gb-2013-14-9-r105.

114. Busche S, Ge B, Vidal R, Spinella JF, Saillour V, Richer C, et al. Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. Cancer Res. 2013;73(14):4323-36. doi: 10.1158/0008-5472.CAN-12-4367.

115. Mardis ER. The challenges of big data. Dis Model Mech. 2016;9(5):483-5. doi: 10.1242/dmm.025585.

116. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811-7. doi: 10.1093/bioinformatics/bts271.

117. Koboldt DC1, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2 Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-76. doi: 10.1101/gr.129684.111.

118. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Mar;31(3):213-9. doi: 10.1038/nbt.2514.

119. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15:244. doi: 10.1186/1471-2164-15-244.

120. Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One. 2016;11(3):e0151664. doi: 10.1371/journal.pone.0151664. eCollection 2016.

121. Bareke E, Saillour V, Spinella JF, Vidal R, Healy J, Sinnett D, et al. Joint genotype inference with germline and somatic mutations. BMC Bioinformatics. 2013;14 Suppl 5:S3. doi: 10.1186/1471-2105-14-S5-S3.

122. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. 2014;6(1):5. doi: 10.1186/gm524. eCollection 2014.

123. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8. doi: 10.1038/nature12213.

124. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC,
et al. MuSiC:
identifying mutational significance in cancer genomes. Genome Res. 2012;22(8):1589-98. doi: 10.1101/gr.134635.111.

125. Cho A, Shim JE, Kim E, Supek F, Lehner B, Lee I. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. Genome Biol. 2016;17(1):129. doi: 10.1186/s13059-016-0989-x.

126. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69(16):6660-7. doi: 10.1158/0008-5472.CAN-09-1133.

127. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108-13.

128. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 2012;40(Web Server issue):W452-7. doi: 10.1093/nar/gks539.

129. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688. doi: 10.1371/journal.pone.0046688.

130. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9. doi: 10.1038/nmeth0410-248.

131. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations:

application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118. doi: 10.1093/nar/gkr407.
132. Forbes SA, Tang G, Bindal N, Bamford S, Dawson E, Cole C, et al. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. 2010;38(Database issue):D652-7. doi: 10.1093/nar/gkp995.
133. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106-14. doi: 10.1038/ng.3168.
134. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22(2):398-406. doi: 10.1101/gr.125567.111.
135. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14 Suppl 3:S7. doi: 10.1186/1471-2164-14-S3-S7.
136. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet. 2014;46:1160-5.
137. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet. 2014;46:1-7.
138. Smith KS, Yadav VK, Pedersen BS, Shaknovich R, Geraci MW, Pollard KS, et al. Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. Nucleic Acids Res. 2015;43:5307-17.
139. Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: An integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res. 2015;43:8123-34.
140. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016;17(1):128. doi: 10.1186/s13059-016-0994-0.
141. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 2014;15(10):480.
142. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol. 2012;8(12):e1002822. doi: 10.1371/journal.pcbi.1002822.
143. Bassik MC, Kampmann M, Lebbink RJ, Wang S, Hein MY, Poser I, et al. A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. Cell. 2013;152(4):909-22. doi: 10.1016/j.cell.2013.01.030.
144. Kampmann M, Bassik MC, Weissman JS. Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proc Natl Acad Sci U S A. 2013;110(25):E2317-26. doi: 10.1073/pnas.1307002110.
145. Goldberg JM, Silverman LB, Levy DE, Dalton VK, Gelber RD, Lehmann L, et al. Childhood T-cell acute lymphoblastic leukemia: the Dana-Farber Cancer Institute acute lymphoblastic leukemia consortium experience. Nat Protoc. 2015;10(4):643. doi: 10.1038/nprot0415-644d.
146. Haydu JE, Ferrando AA. Early T-cell Precursor Acute Lymphoblastic Leukemia (ETP T-ALL). Curr Opin Hematol. 2013;20(4):369-73. doi: 10.1097/MOH.0b013e3283623c61.
147. Roberts ND, Kortschak RD, Parker WT, Schreiber AW, Branford S, Scott HS, et al. A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics. 2013;29(18):2223-30. doi: 10.1093/bioinformatics/btt375.
148. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data_a comparison of mutation callers. Genome Med. 2013;5(10):91. doi: 10.1186/gm495. eCollection 2013.
149. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature.

2009;458(7239):719-24. doi: 10.1038/nature07943. Review.

150. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339(6127):1546-58. doi: 10.1126/science.1235122.

151. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153(1):17-37. doi: 10.1016/j.cell.2013.03.002. Review.

152. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495-501. doi: 10.1038/nature12912.

153. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333-9. doi: 10.1038/nature12634.

154. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45(10):1134-40. doi: 10.1038/ng.2760.

155. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20. doi: 10.1038/ng.2764.

156. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74. doi: 10.1016/j.cell.2011.02.013.

157. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18(3):507-22. doi: 10.1089/cmb.2010.0265.

158. de Smith AJ, Ojha J, Francis SS, Sanders E, Endicott AA, Hansen HM, et al. Clonal and microclonal mutational heterogeneity in high hyperdiploid acute lymphoblastic leukemia. Oncotarget. 2016. doi: 10.18632/oncotarget.12238.

159. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10(11):1081-2. doi: 10.1038/nmeth.2642.

160. Drost J, van Jaarsveld RH, Ponsioen B, Zimberlin C, van Boxtel R, Buijs A, et al. Sequential cancer mutations in cultured human intestinal stem cells. Nature. 2015;521(7550):43-7. doi: 10.1038/nature14415.

161. Antal CE, Hudson AM, Kang E, Zanca C, Wirth C, Stephenson NL, et al. Cancer-associated protein kinase C mutations reveal kinase's role as tumor suppressor. Cell. 2015;160(3):489-502. doi: 10.1016/j.cell.2015.01.001.

162. Van der Meulen J, Sanghvi V, Mavrakis K, Durinck K, Fang F, Matthijssens F, et al. The H3K27me3 demethylase UTX is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. Blood. 2015;125(1):13-21.

163. Vicente C, Schwab C, Broux M, Geerdens E, Degryse S, Demeyer S, et al. Targeted sequencing identifies association between IL7R-JAK mutations and epigenetic modulators in T-cell acute lymphoblastic leukemia. Haematologica. 2015;100(10):1301-10.

164. Neumann M, Heesch S, Gökbuget N, Schwartz S, Schlee C, Benlasfer O, et al. Clinical and molecular characterization of early T-cell precursor leukemia_a high-risk subgroup in adult T-ALL with a high frequency of FLT3 mutations. Blood Cancer J. 2012;2(1):e55. doi: 10.1038/bcj.2011.49.

165. Van Vlierberghe P, Palomero T, Khiabanian H, Van der Meulen J, Castillo M, Van Roy N, et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. Nat Genet. 2010;42(4):338-42.

166. Coustan-Smith E, Mullighan CG, Onciu M, Behm FG, Raimondi SC, Pei D, et al. Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. Lancet Oncol. 2009;10(2):147-56.

167. Trimarchi T, Bilal E, Ntziachristos P, Fabbri G, Dalla-Favera R, Tsirigos A, et al.

Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. Cell. 2014;158(3):593-606. doi: 10.1016/j.cell.2014.05.049.

168. Zhang Y, Castillo-Morales A, Jiang M, Zhu Y, Hu L, Urrutia AO, et al. Genes that escape X-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving. Mol Biol Evol. 2013;30(12):2588-601.

169. Chiaretti S, Vitale A, Cazzaniga G, Orlando SM, Silvestri D, Fazi P, et al. Clinico-biological features of 5202 patients with acute lymphoblastic leukemia enrolled in the Italian AIEOP and GIMEMA protocols and stratified in age cohorts. Haematologica. 2013;98(11):1702-10. doi: 10.3324/haematol.2012.080432.

170. Zahm SH, J F Fraumeni JF Jr. Racial, ethnic, and gender variations in cancer risk: considerations for future epidemiologic research.Environ Health Perspect. 1995;103(Suppl 8): 283-6.

171. Klein SL. The effects of hormones on sex differences in infection: from genes to behavior. Neurosci Biobehav Rev. 2000;24(6):627-38.

172. Cook MB, Dawsey SM, Freedman ND, Inskip PD, Wichner SM, Quraishi SM, et al. Sex disparities in cancer incidence by period and age. Cancer Epidemiol Biomarkers Prev. 2009;18(4):1174-82. doi: 10.1158/1055-9965.EPI-08-1118.

173. Giacomoni PU, Mammone T, Teri M. Gender-linked differences in human skin. J Dermatol Sci. 2009;55(3):144-9. doi: 10.1016/j.jdermsci.2009.06.001. Review.

174. Kaminsky Z, Wang SC, Petronis A. Complex disease, gender and epigenetics. Ann Med. 2006;38(8):530-44.

175. Gabory A, Attig L, Junien C. Sexual dimorphism in environmental epigenetic programming. Mol Cell Endocrinol. 2009 May;304(1-2):8-18. doi: 10.1016/j.mce.2009.02.015.

176. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. Nat Rev Genet. 2008;9(12):911-22. doi: 10.1038/nrg2415.

177. Ghazeeri G, Abdullah L, Abbas O. Immunological differences in women compared with men: overview and contributing factors. Am J Reprod Immunol. 2011;66(3):163-9. doi: 10.1111/j.1600-0897.2011.01052.x.

178. Dorak MT, Karpuzoglu E. Gender Differences in Cancer Susceptibility: An Inadequately Addressed Issue. Front Genet. 2012;3:268. doi: 10.3389/fgene.2012.00268.

179. Do TN, Ucisik-Akkaya E, Davis CF, Morrison BA, Dorak MT. An intronic polymorphism of IRF4 gene influences gene transcription in vitro and shows a risk association with childhood acute lymphoblastic leukemia in males. Biochim Biophys Acta. 2010;1802(2):292-300. doi: 10.1016/j.bbadis.2009.10.015.

180. Bond GL, Hirshfield KM, Kirchhoff T, Alexe G, Bond EE, Robins H, et al. MDM2 SNP309 accelerates tumor formation in a gender-specific and hormone-dependent manner. Cancer Res. 2006;66(10):5104-10.

181. Holzinger ER, Ritchie MD. Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. Pharmacogenomics. 2012;13(2):213-22. doi: 10.2217/pgs.11.145.

182. Kim J, Gao L, Tan K. Multi-analyte network markers for tumor prognosis. PLoS One. 2012;7(12):e52973. doi: 10.1371/journal.pone.0052973.

183. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13(12):R124. doi: 10.1186/gb-2012-13-12-r124.

184. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia Y, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles.Nucleic Acids Res. 2015;43(7):e44. doi: 10.1093/nar/gku1393.

185. Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, et al. Mutations driving CLL and their evolution in progression and relapse. Nature. 2015;526(7574):525-30.

doi: 10.1038/nature15395.
186. Madan V, Shyamsunder P, Han L, Mayakonda A, Nagata Y, Sundaresan J, et al. Comprehensive mutational analysis of primary and relapse acute promyelocytic leukemia. Leukemia. 2016. doi: 10.1038/leu.2016.237.
187. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13. doi: 10.1038/nature10762.
188. Frost L, Goodeve A, Wilson G, Peake I, Barker H, Vora A. Clonal stabilityin late-relapsing childhood lymphoblastic leukaemia. Br J Haematol. 1997;98(4):992-4.
189. Vora A, Frost L, Goodeve A, Wilson G, Ireland RM, Lilleyman J et al. Late relapsing childhood lymphoblastic leukemia. Blood 1998;92:2334-7.
190. Levasseur M, Maung ZT, Jackson GH, Kernahan J, Proctor SJ, Middleton PG. Relapse of acute lymphoblastic leukaemia 14 years after presentation: use of molecular techniques to confirm true re-emergency. Br J Haematol 1994;87:437-8.
191. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, Moch H, et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nat Commun. 2012;3:811. doi: 10.1038/ncomms1814.
192. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6:10001. doi: 10.1038/ncomms10001.
193. Stead LF, Sutton KM, Taylor GR, Quirke P, Rabbitts P. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. Hum Mutat. 2013;34(10):1432-8. doi: 10.1002/humu.22365.
194. Schmitt MW, Prindle MJ, Loeb LA. Implications of genetic heterogeneity in cancer. Ann N Y Acad Sci. 2012;1267:110-6. doi: 10.1111/j.1749-6632.2012.06590.x.
195. Beerenwinkel N, Greenman CD, Lagergren J. Computational Cancer Biology: An Evolutionary Perspective. PLoS Comput Biol. 2016;12(2):e1004717. doi: 10.1371/journal.pcbi.1004717. eCollection 2016.
196. Yuan K, Sakoparnig T, Markowetz F, Beerenwinkel N. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biol. 2015;16:36. doi: 10.1186/s13059-015-0592-6.
197. Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics. 2014 Jan 24;15:27. doi: 10.1186/1471-2105-15-27.
198. Giancotti FG. Mechanisms governing metastatic dormancy and reactivation. Cell. 2013;155(4):750-64. doi: 10.1016/j.cell.2013.10.029.
199. Sosa MS, Bragado P, Aguirre-Ghiso JA. Mechanisms of disseminated cancer cell dormancy: an awakening field. Nat Rev Cancer. 2014;14(9):611-22. doi: 10.1038/nrc3793.
200. Sneddon JB1, Werb Z. Location, location, location: the cancer stem cell niche. Cell Stem Cell. 2007;1(6):607-11. doi: 10.1016/j.stem.2007.11.009.
201. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5-23. doi: 10.1016/j.ajhg.2014.06.009.

ANNEXE

CLIC5: A novel ETV6 target gene in childhood ALL

CLIC5: a novel ETV6 target gene in childhood acute lymphoblastic leukemia

Benjamin Neveu,[1,2] Jean-François Spinella,[1,3] Chantal Richer,[1] Karine Lagacé,[1,2] Pauline Cassart,[1] Mathieu Lajoie,[1] Silvana Jananji,[1] Simon Drouin,[1] Jasmine Healy,[1] Gilles R.X. Hickson[1,4] and Daniel Sinnett[1,2,5]

[1]CHU Sainte-Justine Research Center, Montreal; [2]Department of Biochemistry and Molecular Medicine, Faculty of Medicine, University of Montreal; [3]Molecular biology program, Faculty of Medicine, University of Montreal; [4]Department of Pathology and Cellular Biology, Faculty of Medicine, University of Montreal and [5]Department of Pediatrics, Faculty of Medicine, University of Montreal, Montreal, Canada

Haematologica 2016
Volume 101(12):1534-1543

Correspondence:

daniel.sinnett@umontreal.ca

Received: May 19, 2016.
Accepted: August 11, 2016.
Pre-published: August 18, 2016.

doi:10.3324/haematol.2016.149740

Check the online version for the most updated information on this article, online supplements, and information on authorship & disclosures: www.haematologica.org/content/101/12/1534

ABSTRACT

The most common rearrangement in childhood precursor B-cell acute lymphoblastic leukemia is the t(12;21)(p13;q22) translocation resul-ting in the *ETV6-AML1* fusion gene. A frequent concomitant event is the loss of the residual *ETV6* allele suggesting a critical role for the *ETV6* transcriptional repressor in the etiology of this cancer. However, the precise mechanism through which loss of functional *ETV6* contributes to disease pathogenesis is still unclear. To investigate the impact of *ETV6* loss on the transcriptional network and to identify new transcriptional targets of *ETV6*, we used whole transcriptome analysis of both pre-B leukemic cell lines and patients combined with chromatin immunoprecipitation. Using this integrative approach, we identified 4 novel direct *ETV6* target genes: *CLIC5, BIRC7, ANGPTL2* and *WBP1L*. To further evaluate the role of chloride intracellular channel protein CLIC5 in leukemogenesis, we generated cell lines overexpressing *CLIC5* and demonstrated an increased resistance to hydrogen peroxide-induced apoptosis. We further described the implications of *CLIC5*'s ion channel activity in lysosomal-mediated cell death, possibly by modulating the function of the transferrin receptor with which it colocal-izes intracellularly. For the first time, we showed that loss of *ETV6* leads to significant overexpression of *CLIC5*, which in turn leads to decreased lysosome-mediated apoptosis. Our data suggest that heightened *CLIC5* activity could promote a permissive environment for oxidative stress-induced DNA damage accumulation, and thereby contribute to leukemogenesis.

Introduction

ETV6 is a known transcriptional repressor[1] involved in hematopoiesis.[2] *ETV6* rearrangements are frequently observed in multiple hematological diseases, including precursor B-cell acute lymphoblastic leukemia (pre-B ALL), the most common pediatric cancer.[3] In fact, the t(12;21)(p13;q22) translocation, which generates an in-frame *ETV6-AML1* fusion product,[4] is the most frequent chromosomal abnormality in childhood pre-B ALL, pres-ent in 20% of cases.[5] The expression of *ETV6-AML1* is systematically observed in t(12;21)-positive pre-B ALL,[6,7] indicating a possible function for this chimeric protein in pre-B ALL etiology. However, it was shown that the frequency of the t(12;21) translocation is 100 times greater than that of pre-B ALL,[8] suggesting that its presence alone is insufficient to induce leukemia. It has been demonstrated that the second non-rearranged *ETV6* allele is frequently deleted or inactivated in t(12;21)-positive pre-B ALL,[7,9-11] and recent studies have reported germline *ETV6* loss-of-function mutations that were shown to be associated with familial hematological disorders, including ALL.[12] Although these data suggest that *ETV6* plays a key tumor suppressor role and that its complete inactivation may be required for leukemogenesis,[13] little is known about the function of *ETV6* in normal hematopoiesis and leukemic transformation.

Given its role in transcriptional repression, we postulated that loss of *ETV6* could result in deregulated expression of downstream target genes and perturb key cellular processes and pathways leading to oncogenesis. Only two transcriptional targets of *ETV6* have been identified to date: the MMP3 matrix metallopeptidase[14] and the anti-apoptotic protein BcL-xL.[15] To comprehensively identify novel *ETV6* target genes, we combined whole transcriptome analysis and chromatin immunoprecipitation (ChIP) assays. Using an *in vitro* cell-based system combined with data from childhood pre-B ALL patient tumors, we identified 4 genes (*CLIC5, BIRC7, ANGPTL2* and *WBP1L*) whose expression was directly regulated by *ETV6*. Functional interrogation of *CLIC5* revealed its implication in lysosome-mediated cell death, possibly by regulating iron homeostasis through the transferrin receptor with which it colocalizes intracellularly. In this study, we provide the first evidence of a role for *CLIC5*-mediated resistance to oxidative stress that may contribute to leukemogenesis.

Methods

Complete methods can be found in the *Online Supplementary Methods* section.

Expression profiling by RNA-sequencing

Total RNA from two different Reh clones (generated in methylcellulose media) each stably expressing ETV6-His and ETV6ΔETS_NLS-His (and pLENTI control) were processed through the TruSeq Stranded Total RNA protocol and sequenced on the HiSeq 2500 system (Illumina). Reads for each sample were mapped to the hg19 reference genome using STAR with default settings,[16] and read counts per genes were determined using HTSeq-count.[17] To identify differentially expressed genes (DEGs), we used the R bioconductor package edgeR[18] with Benjamini-Hochberg *P*-value adjustment. The two clones were considered as biological replicates.

The patient cohort used for RNA sequencing was composed of 9 hyperdiploid and 9 t(12;21) patients. Total RNA was extracted from leukemic bone marrow samples of all patients and from control pre-B cells ($CD19^+CD10^+$) isolated from healthy cord blood samples. cDNA libraries were prepared using the SOLiD Total RNA-seq kit and sequenced on the SOLiD 4/5500 System (Life Technologies). Reads were aligned to the hg19 reference genome and read counts per gene obtained using LifeScope Genomic Analysis Software with default parameters.

Quantitative real-time PCR

350ng of total RNA were retro-transcribed with M-MLV reverse transcriptase (Life Technologies). cDNA was then subjected to quantitative real-time PCR using the primer sets listed in the *Online Supplementary Table S1*. Relative expression was determined by the 2-($^{\Delta\Delta}$Ct) comparative method[19] using *GAPDH* as the reference gene.

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was performed on $10x10^6$ transduced Reh cells cross-linked directly in cell medium for 10min with 1% methanol-free formaldehyde (Polysciences, Inc.). Immunoprecipitation of sheared chromatin was carried out using anti-HA magnetic beads (Thermo Fisher Scientific). Beads were eluted twice with HA peptides (Thermo Fisher Scientific) before reverse cross-linking. DNA was purified twice by standard phenol/chloroform/isoamyl alcohol (Sigma-Aldrich) extraction prior to qRT-PCR analysis (primers are listed in the *Online Supplementary Table S2*).

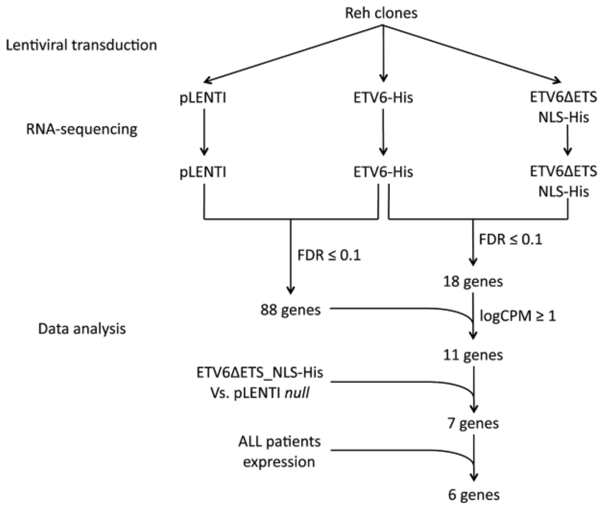


Figure 1. Schematic representation of the transcriptome-based design to detect putative direct *ETV6* target genes. To identify direct targets of *ETV6*, we first designed an *in vitro* RNA-seq experiment using $ETV6^{-/-}$ Reh-derived clones. Cells were transduced with lentiviral constructs to express ETV6-His and ETV6ΔETS_NLS-His. Total RNA was extracted from stable cell populations and RNA-seq libraries were sequenced. Expression profiles were analyzed using EdgeR. Gene expression profiles in ETV6-His cells were first compared with ETV6ΔETS_NLS-His and pLENTI cells to identify repressed genes (FDR ≤ 0.1). We then included data from the ETV6ΔETS_NLS-His *vs.* pLENTI comparison and further considered genes whose expression remains constant (*P*-value≥0.05 or logFC≥-0.5) which are more likely to be direct *ETV6* targets. Finally, only genes that showed a specific overexpression in t(12;21)-positive childhood pre-B ALL (pre-B acute lymphoblastic leukemia) patients were considered. ALL: acute lymphoblastic leukemia; FDR; false discovery rate; logCPM: log counts per million reads.

Table 1. Expression status of the 6 putative direct *ETV6* target genes.

			Gene Symbol					
			CLIC5	*BIRC7*	*ANGPTL2*	*WBP1L*	*LRRC4*	*SLC51A*
	ETV6-His	logFC	-3.23	-1.93	-1.31	-0.79	-1.02	-1.31
	vs.	logCPM	4.66	4.02	5.00	5.78	3.90	3.11
	ETV6ΔETS_NLS-His	PValue	7.00E-52	3.44E-11	4.24E-11	8.04E-06	9.18E-05	1.14E-04
		FDR	9.59E-48	1.45E-07	1.45E-07	9.18E-03	6.29E-02	6.79E-02
	ETV6-His	logFC	-3.20	-1.39	-1.68	-1.10	-1.16	-1.50
In vitro	*vs.*	logCPM	3.75	3.02	4.37	5.18	3.23	2.43
	pLENTI	PValue	6.40E-39	2.30E-05	3.34E-14	1.06E-09	3.99E-06	1.45E-05
		FDR	8.09E-35	4.76E-03	6.41E-11	6.71E-07	1.20E-03	3.28E-03
	ETV6ΔETS_NLS-His	logFC	0.03	0.53	-0.38	-0.32	-0.15	-0.20
	vs.	logCPM	4.62	3.85	4.81	5.49	3.63	2.90
	pLENTI	PValue	0.88	0.03	0.02	0.06	0.55	0.50
		FDR	1	1	1	1	1	1
ALL patients		logFC	6.36	6.79	5.82	1.92	4.25	4.27
t(12;21) *vs.* B-cells		logCPM	10.55	9.74	9.42	10.68	9.30	8.26
		PValue	1.27E-30	1.57E-13	6.39E-08	4.87E-07	7.67E-12	1.91E-06
		FDR	6.91E-29	1.55E-12	3.11E-07	2.17E-06	6.25E-11	7.59E-06

ALL: acute lymhoblastic leukemia; logFC: log fold change; logCPM: log counts per million reads; FDR: false discovery rate.

Apoptosis assays

Apoptosis was induced by treating cells for 20h with hydrogen peroxide (PRXD; Sigma-Aldrich), camptothecin (CPT; Tocris Bioscience) or doxorubicin (DOXO; Sigma-Aldrich) and assayed by Alexa Fluor 488-coupled Annexin V and propidium iodide (PI) double staining. 1x10^4 stained cells were analyzed by flow cytometry. Total apoptosis includes Annexin V+/ PI - (early apoptotic), Annexin V+/PI + (late apoptotic) and Annexin V-/PI+ (necrotic) cells.

Immunofluorescence microscopy

Reh cells were seeded at 2x10^6 cells/mL in a 96 well glass plate (Whatman) and fixed in a 3.7% formaldehyde solution. Immunostaining was performed overnight with CLIC5A antibody (ab191102 dil. 1:1000; Abcam) and transferrin receptor antibody (ab84036 dil. 1:200; Abcam). Goat anti-Mouse Alexa Fluor 488 (dil. 1:500; Thermo Fisher Scientific) was used to detect anti-CLIC5A, and Goat anti-Rabbit Alexa Fluor 546 (dil. 1:500; Thermo Fisher Scientific) was used to detect anti-transferrin receptor. Hoechst 33258 DNA stain (dil. 1:500; Thermo Fisher Scientific) was included to stain nuclei.

Statistical tests

The significance of observations was assessed using one or two-tailed Fisher's exact test or Mann-Whitney U test when appropriate.

Ethics statement

The CHU Sainte-Justine Research Ethics Board approved the protocol. Informed consent was obtained from the parents of the patients to participate in this study and for publication of this report and any accompanying images.

Results

ETV6 represses the expression of 6 genes in pre-B ALL cell lines and patient samples

To identify novel direct *ETV6*-regulated genes, we carried out a transcriptome analysis in both cell lines and patient tumor samples (Figure 1). Based on the expression profiles of transduced t(12;21)-positive pre-B ALL Reh clones (*Online Supplementary Figure S1*), we found 331 genes repressed in His-tagged *ETV6* (ETV6-His) cells compared to control cells (pLENTI empty vector; *P*-value≤0.05), of which 88 remained significant after multiple testing corrections (FDR≤0.1; *Online Supplementary Table S3*). 18 genes were significantly repressed by ETV6-His compared to its DNA-binding deficient mutant ETV6ΔETS_NLS-His (FDR≤0.1; *Online Supplementary Table S4)*, of which 11 were both confidently expressed (logCPM≥1) and also present in the above-mentioned list of 88 genes. These genes are thus more likely to be direct targets of *ETV6* since their repression depends on *ETV6*'s DNA-binding domain. However, only 7 of these genes absolutely required the DNA-binding domain for repression (ETV6ΔETS_NLS-His *vs.* pLENTI; *P*-value≥0.05 or logFC≥-0.5), further confirming their direct regulation by *ETV6*: *CLIC5, BIRC7, DDIT4L, ANGPTL2, WBP1L, LRRC4*, and *SLC51A*. We then assessed whether the *ETV6*-dependent transcriptional repression observed *in vitro* translated to childhood pre-B ALL patient tumor samples. Expression of the 331 genes repressed by *ETV6 in vitro* was evaluated in transcriptome data from 9 t(12;21)-positive samples (*ETV6* negative) and compared to 9 hyperdiploid cases as well as to 3 normal pre-B cell (CD19$^+$/CD10$^+$) samples (*ETV6* positive).

We identified 45 genes that were downregulated *in vitro* (13.6%) and that were also specifically overexpressed in t(12;21)-positive patients (Figure 2). Interestingly, these include 6 of the 7 genes identified as putative direct *ETV6* targets *in vitro* (*CLIC5, BIRC7, ANGPTL2, WBP1L, LRRC4* and *SLC51A*, but not *DDIT4L*), further supporting a role for *ETV6* in their regulation.

CLIC5, BIRC7, ANGPTL2 *and* WBP1L *are direct targets of* ETV6

To validate *ETV6*-dependent expression of these 6 genes (Table 1), we used quantitative real-time PCR (qRT-PCR) in both Reh clones and the original Reh pop-

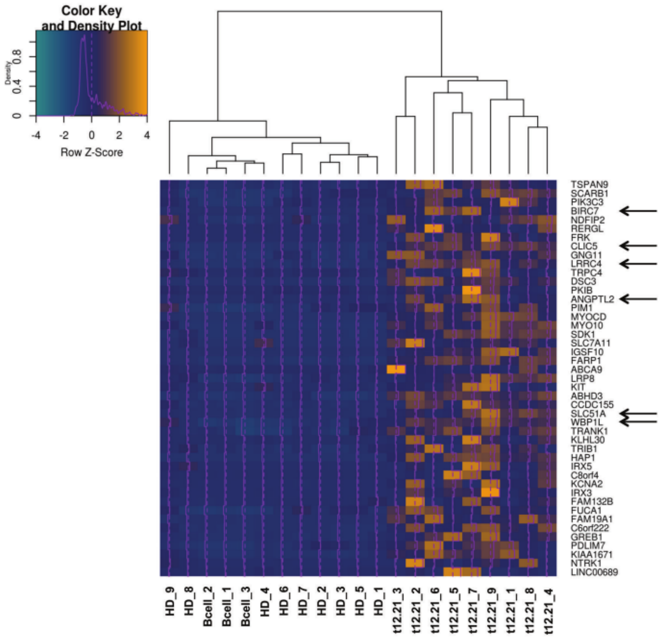


Figure 2. Expression of putative *ETV6* targets in pediatric pre-B ALL (pre-B acute lymphoblastic leukemia) patients. An unsupervised clustering heatmap was generated with the 45 genes specifically overexpressed (yellow) in t(12;21)-positive ALL (acute lymhoblastic leukemia) patients (t12.21, n=9) compared to hyperdiploid ALL patients (HD, n=9) and normal B-cells controls (B-cell, n=3). Among the 7 genes identified as putative direct *ETV6* targets in our cell line model, *CLIC5, BIRC7, ANGPTL2, WBP1L, LRRC4* and *SLC51A* were specifically overexpressed in t(12;21)-positive pediatric pre-B ALL patient tumor samples (Arrows).

ulation overexpressing ETV6 WT, ETV6-HA or GFP as a control (*Online Supplementary Figure S2*). Both ETV6 WT and ETV6-HA efficiently repressed expression of these genes except for *LRRC4* (Figure 3A). To assess the physical interaction between *ETV6* and the proximal promoters of these 5 putative *ETV6* targets we performed ChIP experiments in Reh cells overexpressing ETV6-HA or ETV6 WT as a negative control. Importantly, ETV6-HA behaves similarly to ETV6 WT in our qRT-PCR experiments (Figure 3A), indicating that the epitope tag does not negatively interfere with normal *ETV6* repressor function. As shown in Figure 3B, we successfully enriched the proximal promoters of *CLIC5, BIRC7, ANGPTL2* and *WBP1L*, but not *SLC51A*, further confirming that these 4 genes are indeed direct targets of *ETV6*.

CLIC5A *reduces hydrogen peroxide-induced apoptosis*

We pursued functional interrogation of our strongest candidate, the chloride intracellular channel *CLIC5*, to investigate its cellular function and potential contribution to childhood pre-B ALL. The *CLIC5* locus encodes two major isoforms, *CLIC5A* and *CLIC5B*[20] (*Online Supplementary Figure S3A*), transcribed by two alternative promoters and differing only from their first exon. Using ChIP experiments (as above), we showed that the *CLIC5A* promoter was specifically enriched, whereas the *CLIC5B* promoter showed no significant enrichment compared to the negative control region (*Online Supplementary Figure S3B*). *ETV6* overexpression was also shown to lead to a marked decrease of the CLIC5A protein, whereas CLIC5B levels remained constant (*Online Supplementary Figure S3C*). Together, these results confirm specific *ETV6*-mediated repression of *CLIC5A*.

In light of these results, we overexpressed the *CLIC5A* isoform in Reh cells (Figure 4A) in order to investigate its role on B-lymphoblast function. Importantly, the overexpression of *CLIC5A* in our Reh cells is highly similar to that observed in a validation cohort of t(12;21) ALL patients (*Online Supplementary Figure S4*). Given that changes in migration were observed upon silencing of *CLIC5A*,[21] we first evaluated this phenotype. We found no particular difference in migration toward stromal cell-derived factor 1 (*SDF-1*, also known as *CXCL12*) in a classic transwell experiment between control and *CLIC5A* overexpressing cells (*Online Supplementary Figure S5*). Although *CLIC5A* had never been shown to be associated with apoptosis, suppression of its closely related family member *CLIC4* had previously been shown to enhance hydrogen peroxide-induced apoptosis.[22] To test this hypothesis in our cell model, we treated *CLIC5A* overexpressing cells with hydrogen peroxide, camptothecin, or doxorubicin, and evaluated apoptosis. Of note, peroxide induces an apoptotic cell death in these conditions (Figure 4B), rather than necrosis. We observed a modest yet consistent reduction in hydrogen peroxide-induced apoptosis compared to control cells (Figure 4C), suggesting a potential role in the intracellular response to free radicals. A similar reduction in apoptosis following peroxide treatment was observed with *CLIC5A* overexpression in the IM9 B-lymphoblastoid cell line (Figure 4D-F) endogenously expressing wild-type *ETV6*, further confirming that *CLIC5A* overexpression specifically reduces hydrogen peroxide-induced apoptosis across cellular backgrounds.

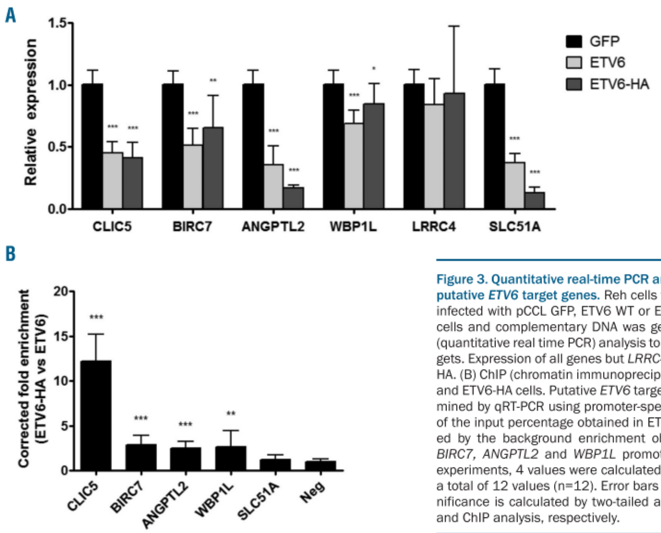


Figure 3. Quantitative real-time PCR and chromatin immunoprecipitation validation of putative *ETV6* target genes. Reh cells together with the two Reh derivated clones were infected with pCCL GFP, ETV6 WT or ETV6-HA. (A) Total RNA was extracted from these cells and complementary DNA was generated. This cDNA was submitted to qRT-PCR (quantitative real time PCR) analysis to quantify relative expression of putative *ETV6* targets. Expression of all genes but *LRRC4* is repressed by ETV6 WT (wild-type) and ETV6-HA. (B) ChIP (chromatin immunoprecipitation) experiments were performed in ETV6-WT and ETV6-HA cells. Putative *ETV6* target gene proximal promoter enrichment was determined by qRT-PCR using promoter-specific primers. Results are presented as the ratio of the input percentage obtained in ETV6-HA cells compared to ETV6 WT cells, corrected by the background enrichment obtained with an unbound region (Neg). *CLIC5, BIRC7, ANGPTL2* and *WBP1L* promoters are enriched. For both qRT-PCR and ChIP experiments, 4 values were calculated for each of the 3 cell lines and were merged for a total of 12 values (n=12). Error bars represent the standard deviation. Statistical significance is calculated by two-tailed and one-tailed Mann-Whitney U test for qRT-PCR and ChIP analysis, respectively.

CLIC5A is an endosomal ionic channel involved in lysosome-mediated apoptosis

Given that hydrogen peroxide is known to trigger lysosomal membrane permeabilization (LMP) and initiate the lysosomal-mediated apoptosis pathway (Figure 5A),[23,24] we hypothesized that *CLIC5A*'s role in protecting cells against apoptosis may function through the modulation of LMP. Since LMP has a direct impact on mitochondrial outer membrane permeabilization (MOMP) that can be assessed through mitochondrial membrane potential (MMP), we evaluated MMP in our Reh cellular model treated with hydrogen peroxide. We observed a significant reduction of MMP loss correlated with *CLIC5A* overexpression compared to control, indicating that substantially more mitochondria remained intact following peroxide exposure when *CLIC5A* was overexpressed (*Online Supplementary Figure S6*). This result supports a role for *CLIC5A* in protecting cells against peroxide-induced apoptosis and suggests that it functions upstream of MOMP, which indeed corroborates a possible implication of *CLIC5A* in LMP regulation.

Deleterious effects of hydrogen peroxide on lysosome membranes are strongly dependent on lysosomal Fe^{2+} availability since it dictates its conversion to the highly reactive hydroxyl radicals.[25] We thus investigated *CLIC5A*'s impact on lysosome-mediated cell death by modulating lysosomal Fe^{2+} availability prior to hydrogen peroxide exposure by pre-treating cells with the iron chelator deferoxamine mesylate salt (DFO).[26,27] As shown in Figure 5B, cells treated with DFO prior to peroxide showed a drastic reduction in apoptosis. The residual apoptotic activity observed in DFO-treated cells could be driven by DNA damage,[28] which appears to be *CLIC5A*-independent given the results obtained with the two DNA damaging agents camptothecin and doxorubicin (Figure 4C).

Inversely, we used ferric ammonium citrate (FAC) to positively modulate Fe^{2+} concentration in lysosomes.[29] Increased lysosomal Fe^{2+} concentration favors hydroxyl radical production from hydrogen peroxide and hence is expected to increase LMP and apoptosis. Accordingly, cells pre-treated with FAC showed increased peroxide-induced apoptosis and the protective effect of *CLIC5A* overexpression was completely lost (Figure 5C), indicating that *CLIC5A* plays a role upstream of LMP. This specific role further corroborates *CLIC5A*'s inability to protect cells against DNA damage.

To evaluate *CLIC5A*'s ion channel activity[30] in this context we pre-treated *CLIC5A* overexpressing Reh cells with indanyloxyacetic acid 94 (IAA-94), a known CLIC-specific ion channel inhibitor.[30,31] IAA-94 treatment increased peroxide-induced apoptosis and completely prevented *CLIC5A*-mediated protection (Figure 5D), similar to that which was observed in the presence of increased Fe^{2+} availability following FAC treatment (Figure 5C). Endogenous *CLIC5A* inhibition contributes to this phenotype and explains the increased apoptosis level of control cells. Together, these data confirm that *CLIC5A*'s ion channel activity is required to protect cells against peroxide-induced apoptosis, perhaps by limiting Fe^{2+} availability in the lysosomal pathway.

To further investigate the functional implications of *CLIC5A* in lysosomal apoptosis, we examined the intracellular localization of *CLIC5A* using immunofluorescence. Although we did not observe colocalization of *CLIC5A* with lysosomes (*Online Supplementary Figure S7*), we did show positive colocalization with transferrin receptor (Figure 6), which is in line with *CLIC5A*'s postulated role in modulating lysosomal Fe^{2+} availability. Transferrin receptors (*TFRC* gene) are responsible for cel-

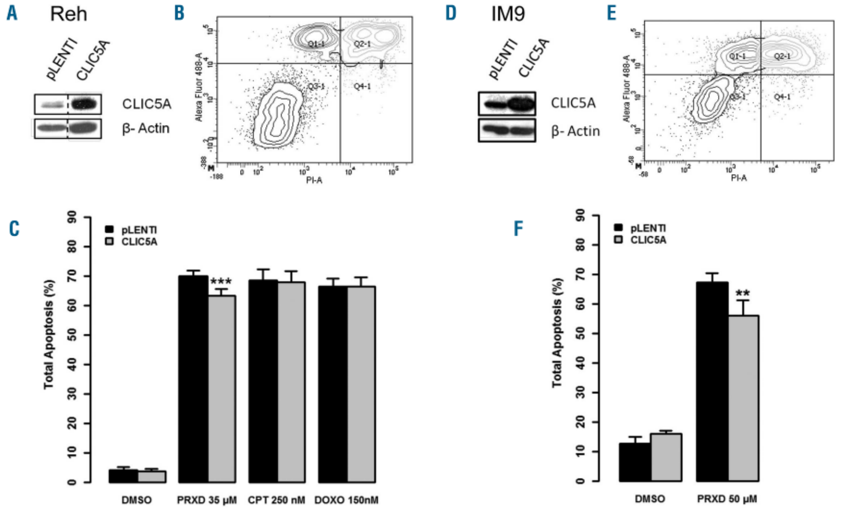


Figure 4. *CLIC5A* protects cells from hydrogen peroxide-induced apoptosis. (A) pLENTI control and *CLIC5A* stably infected Reh cells were challenged for 20h with 35μM hydrogen peroxide (PRXD), 250nM camptothecin (CPT), or 150nM doxorubicin (DOXO) and analyzed by flow cytometry with Alexa Fluor 488-coupled Annexin V and propidium iodide (PI) staining. (B) A representative example of staining following PRXD treatment is presented with PI on the x axis and Alexa Fluor 488 on the y axis. (C) Total apoptosis is calculated for each sample. *CLIC5A* overexpressing cells displayed reduced apoptosis compared to control only when treated with PRXD. Each experiment was performed 3 times in triplicates (n=9). (D) IM9 cells transduced with *CLIC5A* or pLENTI empty vector control were challenged with 50μM PRXD and processed similarly to assess apoptosis. (E) A representative example of staining and (F) total apoptosis is shown for the IM9 cell line. Again, *CLIC5A* overexpression leads to reduced apoptosis. Two independent experiments carried out in triplicate and an additional single test were performed (n=7). Error bars in (C) and (F) represent the standard deviation. Statistical significance is calculated by two-tailed Mann-Whitney U test. For (A) and (D) adjustments of brightness and contrast were applied to the whole image. DMSO: dimethyl sulfoxide.

lular iron intake through the binding and internalization of its iron-bound ligand transferrin (*TF* gene).[32,33] *CLIC5A* could perturb this process and thereby impact cellular iron concentrations and modulate lysosome sensitivity to hydrogen peroxide. Interestingly, double positive staining appears to be particularly present in distinct vesicle-like structures that are likely transferrin receptor-containing recycling endosomes, further suggesting that *CLIC5A* may interfere with normal iron homeostasis, thus reducing lysosome sensitivity to oxidative stress.

Taken together, our data strongly support a role for the newly identified *ETV6* transcriptional target *CLIC5A* in modulating lysosome-mediated apoptosis. We propose that *CLIC5A* overexpression in t(12;21)-positive, *ETV6* depleted pre-B cells could contribute to increased resistance to oxidative stress and therefore promote cell survival.

Discussion

Approximately 20% of childhood pre-B ALL patients harbor the t(12;21) translocation, yielding an ETV6-AML1 fusion protein that is, however, insufficient to initiate leukemia.[8,34,35] This process often requires further loss of the remaining wild-type *ETV6* allele,[7,9-11,13] suggesting that deregulation of the *ETV6* transcriptional machinery could play an important role in leukemogenesis. Unfortunately, very few *ETV6*-regulated transcriptional targets are known and the mechanisms through which they are involved in leukemogenesis remain elusive. Herein, we combined both *in vitro* (cell lines) and *ex vivo* (pre-B ALL patients) transcriptome data to identify candidate *ETV6* target genes, and through additional cellular assays identified 4 novel direct *ETV6* target genes: *CLIC5, BIRC7, ANGPTL2* and *WBP1L*.

The *CLIC5* gene was previously associated with the t(12;21)-positive ALL molecular signature,[36] and our pediatric pre-B ALL expression data corroborated this result with strong *CLIC5* overexpression shown to be specific to the t(12;21)-positive subgroup within our cohort. Thus *CLIC5* overexpression in these patients is likely due to *ETV6* loss. Unfortunately, very little is known about *CLIC5*'s potential contribution to leukemogenesis. Using engineered cell lines, we demonstrated an increased resistance to hydrogen peroxide-induced apoptosis following overexpression of *CLIC5A*. Notably, Reh cells already express the endogenous *CLIC5A* isoform, which could explain the modest effect of the overexpression of *CLIC5A* (mean=7.46%, *P*-value=8.32×10^{-11}, merged n=36). Given that normal B-cells show no expression of *CLIC5* (mean FPKM<0.1 over our 3 normal $CD19^{+}/CD10^{+}$ pre-B cell samples isolated from human cord blood), an eventual stronger impact of *CLIC5A* re-expression in a pre-B cell upon *ETV6* depletion can be expected.

We further linked this phenotype to *CLIC5A*'s ion chan-

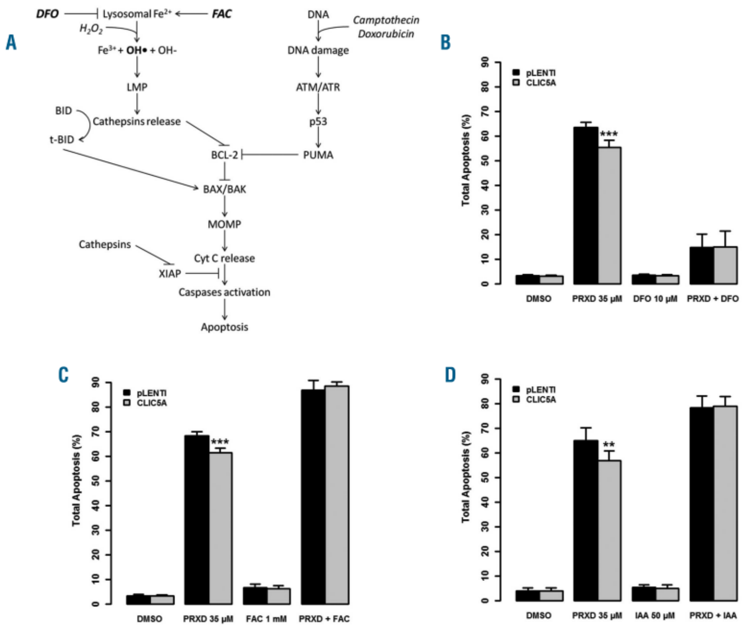


Figure 5. *CLIC5A* is implicated in lysosome-mediated cell death. (A) Schematic illustration of the lysosomal apoptosis pathway and DNA damage pathway. DFO: deferoxamine mesylate salt; FAC: ferric ammonium citrate; LMP: Lysosomal membrane permeabilization; MOMP: Mitochondrial outer membrane permeabilization. (B) pLENTI control and *CLIC5A* overexpressing Reh cells were pre-treated with 10μM DFO to chelate lysosomal ferrous iron and apoptosis was induced using 35μM PRXD for 20h followed by flow cytometry quantification. PRXD-induced apoptosis was greatly reduced with DFO treatment. (C) Similarly, 1mM FAC was used to increase ferrous iron concentration which led to higher PRXD-induced apoptosis with no protective effect of *CLIC5A* overexpression. (D) Cells were pre-treated with 50μM IAA-94 CLIC-specific ion channel inhibitor and apoptosis was assayed after a subsequent PRXD treatment. In these conditions, *CLIC5A* overexpression did not reduce apoptosis. Each experiment was carried out 3 times in triplicate (n=9). Error bars represent the standard deviation. Statistical significance is calculated by two-tailed Mann-Whitney U test.

nel activity in lysosomal-mediated cell death. The presence of *CLIC5A* on transferrin receptor-containing endosomes potentially negatively impacts lysosomal iron availability, thus reducing lysosome sensitivity to oxidative stress. However, the exact mechanism by which *CLIC5A* modulates lysosomal iron to prevent peroxide-induced apoptosis remains unclear. It has been demonstrated that changes in the concentration of several ions can modulate endosomal pH through ion dependent H^+ pumps.[37] With *CLIC5A* being a poorly selective ion channel,[38] we can hypothesize a somewhat similar function: slight differences in endosome acidification could modulate transferrin iron release or trafficking and therefore impact lysosomal iron concentration leading to differential peroxide sensitivity.

Additional experiments should be performed to further dissect *CLIC5A*'s role in lysosome-mediated apoptosis.

Nonetheless, the observed effects of *CLIC5A* in peroxide resistance could play a key role in ALL initiation. A recent study highlighted the high oxidative stress levels of leukemic blasts in the bone marrow niche induced by bone marrow stromal cell signaling.[39] Furthermore, it has been shown that increased levels of ROS in t(12;21)-positive cells was associated with DNA damage accumulation.[40] High oxidative stress levels are thought to trigger apoptotic cell death through the lysosomal pathway, thus preventing DNA damage accumulation. However, in t(12;21)-positive ALL, we have shown that loss of *ETV6* expression leads to significant overexpression of *CLIC5A*. This overexpression leads to decreased lysosome-media-ted apoptosis, which in turn can promote a more permissive environment to heighten ROS levels (Figure 7). Cells evading apoptosis can thus accumulate ROS-induced mutations at a greater rate. Moreover, this mutational process of t(12;21)-positive pre-B cells could be facilitated not only by *CLIC5A* overexpression, but together with several *ETV6* deregulated targets such as the inhibitor of apoptosis *BIRC7*. Although the overall mutational burden of ALL is low compared to other cancers,[41] the difference in ROS-induced DNA damage accumulation can contribute, over time, to promote leukemic transformation when impacting key cancer genes.

It remains challenging, however, to evaluate the contribution of this pathway to the total amount of DNA damage found in pre-B ALL patients. Although ROS-mediated alterations of nucleotides are well characterized, their signature in sequencing data remains unclear.[42,43] Interrogation of our patient sequencing data did not reveal

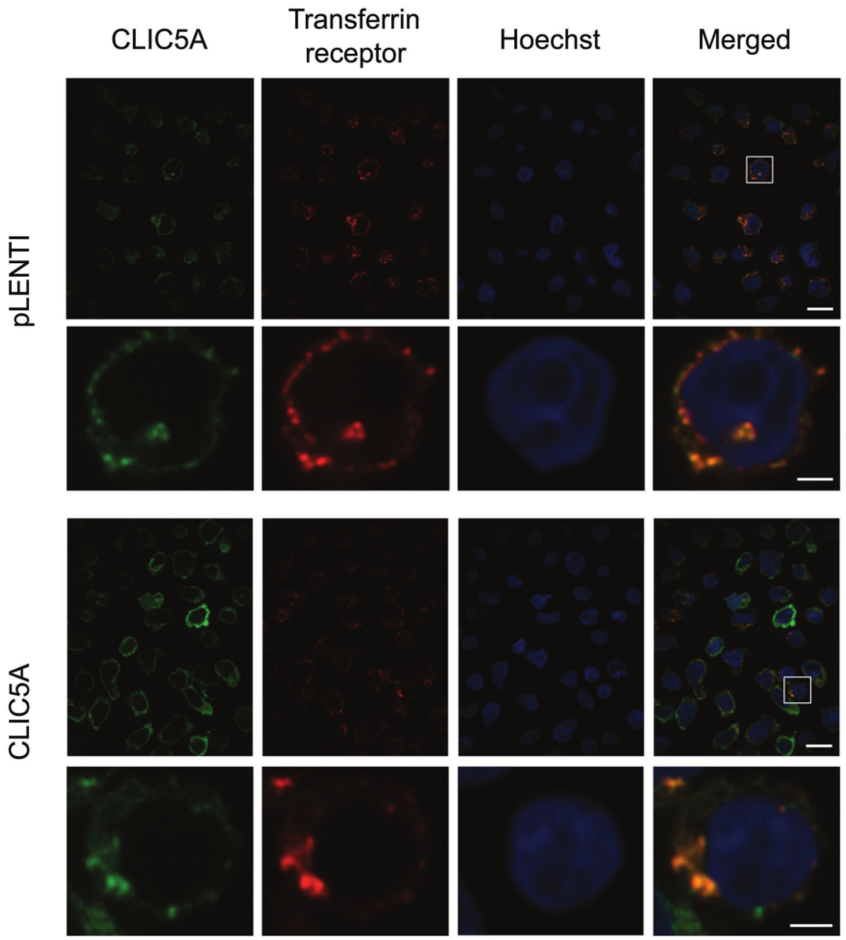


Figure 6. Colocalization of *CLIC5A* and transferrin receptors. pLENTI control and *CLIC5A* overexpressing Reh cells were used for co-localization studies. Immunostaining of both *CLIC5A* and transferrin receptor were performed in the same cells simultaneously with Hoechst DNA staining. Results were obtained at 100X magnification (upper panels; scale bar = 10μm). A strong colocalization was observed between *CLIC5A* and transferrin receptors. Additional enlargement for the marked region of the initial image is presented in lower panels (scale bar = 2μm). A merged image was generated (right panels).

any particular mutational profile that could undoubtedly be attributed to ROS. However, increased ROS-mediated DNA double strand breaks have been observed in t(12;21)-positive B-cells when assessed by comet assays.[40] These alterations do often lead to deletions or rearrangements due to aberrant repair, that are frequent events in leukemia, thus supporting a role for ROS-induced DNA damage in ALL.

With a similar *CLIC5A*-mediated protective effect on peroxide-induced apoptosis obtained in both Reh and IM9 cell lines, we demonstrated a phenotype that is independent from a particular genetic background. Interestingly, *CLIC5* expression has been associated with poor prognosis in breast cancer,[44] and expression arrays of a broad range of normal and tumoral tissues obtained through the GENT database[45] showed an overexpression of *CLIC5* in ovarian cancers (*Online Supplementary Figure S8*). Based on our results, and given the importance of oxidative stress resistance in these solid tumors that require angiogenesis to promote growth, it suggests that *CLIC5A* protection against oxidative stress may not be limited to pre-B cell leukemia. The unfavorable prognosis associated with *CLIC5* deregulation may be due to increased oxidative stress resistance, thus reducing the necessity of angiogenesis and fostering an environment prone to ROS-induced DNA damage. Although *ETV6* alterations have not been

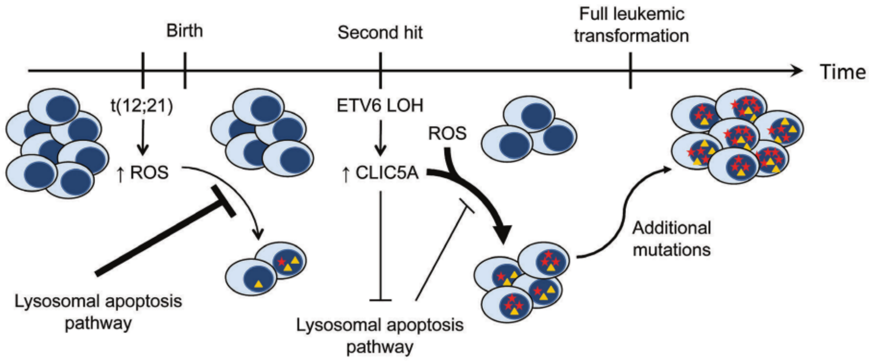

Birth
Second hit
Full leukemic transformation
Time
t(12;21)
↑ ROS
ETV6 LOH
↑ CLIC5A
ROS
Additional mutations
Lysosomal apoptosis pathway
Lysosomal apoptosis pathway

Figure 7. Proposed mechanism of *CLIC5A* involvement in *ETV6*-associated childhood pre-B ALL. The t(12;21) translocation occurs early and contributes to an increased level of reactive oxygen species (ROS). Lysosome-mediated apoptosis is sensitive to this excess of ROS and therefore prevents the accumulation of ROS-mediated DNA damage. With the subsequent deletion of the residual *ETV6* allele (LOH), *CLIC5A* expression is drastically upregulated. This overexpression of *CLIC5A* has a negative impact on the lysosomal apoptosis pathway and thus creates a permissive environment for the accumulation of mutations driven by high oxidative stress. Over time, some of these mutations may impact key cellular biological processes and pathways, which will ultimately lead to full leukemic transformation and development of childhood pre-B acute lymphoblastic leukemia (pre-B ALL).

reported in these solid tumors, other mechanisms could drive *CLIC5* overexpression. Interestingly, ETS factor binding sites (EBS) were found to be highly enriched for small non-coding mutations across a wide variety of cancers.[46] These mutations disrupt the consensus binding site of ETS factors, such as *ETV6*, and thus may prevent it's binding and the repression of its targets. Despite the presence of wild-type *ETV6* in these cases, some of its key targets (*CLIC5*, *BIRC7* or *ANGPTL2*) may be overexpressed, thus leading to a selective advantage.

While we focused on *CLIC5*, the other identified targets were also of interest. Notably, the caspases inhibitor *BIRC7* has previously been shown to be part of a t(12;21)-positive pre-B ALL molecular signature.[36] Overexpression of *BIRC7* was also observed in a variety of tumors, and contributes to oncogenesis through the inhibition of apoptosis,[47] suggesting a similar involvement for *BIRC7* in t(12;21)-positive childhood pre-B ALL. The *ANGPTL2* gene encodes a secreted protein with, among others, pro-angiogenic and anti-apoptotic properties. Although its expression has been linked to a broad range of diseases, including cancers, it has been shown to increase survival and expansion of hematopoietic stem cells.[48] *ANGPTL2*'s contribution to leukemogenesis remains unknown. Lastly, *WBP1L*, also known as *OPAL1* (Outcome Predictor in Acute Leukemia 1), is associated with the t(12;21) favorable outcome in ALL.[49] It is still unclear whether *WBP1L* plays a role in leukemia, since its molecular function is yet to be characterized.

In conclusion, for the first time, we describe a role for *CLIC5*-mediated resistance to oxidative stress that could promote cell survival and contribute to leukemogenesis. We propose a mechanism in which complete loss of wild-type *ETV6* expression in a pre-leukemic blast leads to *CLIC5A* overexpression, thus creating a permissive environment for the accumulation of mutations driven by high oxidative stress, eventually giving rise to full leukemic transformation.

Acknowledgments

The authors would like to thank the patients and their parents for participating in this study. We are grateful to Dr. Christian Beauséjour who kindly provided lentiviral-related material.

Funding

This study was supported by research funds provided by the Terry Fox Research Institute and the Canadian Institutes of Health Research. BN is the recipient of a Cole Foundation scholarship. JFS is the recipient of a Réseau de médecine génétique appliquée (RMGA) scholarship. DS holds the François-Karl Viau Research Chair in Pediatric Oncogenomics. Whole transcriptome sequencing was performed at the Integrated Clinical Genomic Center in Pediatrics at the CHU Sainte-Justine Research Center. Computations were performed on the Briarée supercomputer at the Université de Montréal, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), NanoQuébec, RMGA and the Fonds de recherche du Québec - Nature et technologies (FRQNT). GH thanks the CFI, the Cole Foundation and the Fonds de recherche du Québec-Santé for infrastructure, transition award and salary support, respectively.

References

1. Lopez RG, Carron C, Oury C, Gardellin P, Bernard O, Ghysdael J. TEL is a sequence-specific transcriptional repressor. J Biol Chem. 1999;274(42):30132-30138.
2. Wang LC, Swat W, Fujiwara Y, et al. The TEL/ETV6 gene is required specifically for hematopoiesis in the bone marrow. Genes Dev. 1998;12(15):2392-2402.
3. Bohlander SK. ETV6: a versatile player in leukemogenesis. Semin Cancer Biol. 2005;15(3):162-174.
4. Golub TR, Barker GF, Bohlander SK, et al. Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia. Proc Natl Acad Sci USA. 1995; 92(11):4917-4921.
5. Tasian SK, Loh ML, Hunger SP. Childhood acute lymphoblastic leukemia: Integrating

genomics into therapy. Cancer. 2015; 121(20):3577-3590.

6. Agape P, Gerard B, Cave H, et al. Analysis of ETV6 and ETV6-AML1 proteins in acute lymphoblastic leukaemia. Br J Haematol. 1997;98(1):234-239.
7. Poirel H, Lacronique V, Mauchauffe M, et al. Analysis of TEL proteins in human leukemias. Oncogene. 1998;16(22):2895-2903.
8. Mori H, Colman SM, Xiao Z, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proc Natl Acad Sci USA. 2002;99(12):8242-8247.
9. Patel N, Goff LK, Clark T, et al. Expression profile of wild-type ETV6 in childhood acute leukaemia. Br J Haematol. 2003; 122(1):94-98.
10. Lilljebjorn H, Soneson C, Andersson A, et al. The correlation pattern of acquired copy number changes in 164 ETV6/RUNX1-positive childhood acute lymphoblastic leukemias. Hum Mol Genet. 2010; 19(16):3150-3158.
11. Montpetit A, Larose J, Boily G, Langlois S, Trudel N, Sinnett D. Mutational and expression analysis of the chromosome 12p candidate tumor suppressor genes in pre-B acute lymphoblastic leukemia. Leukemia. 2004;18(9):1499-1504.
12. Moriyama T, Metzger ML, Wu G, et al. Germline genetic variation in ETV6 and risk of childhood acute lymphoblastic leukaemia: a systematic genetic study. Lancet Oncol. 2015.
13. Anderson K, Lutz C, van Delft FW, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. Nature. 2011;469(7330):356-361.
14. Fenrick R, Wang L, Nip J, et al. TEL, a putative tumor suppressor, modulates cell growth and cell morphology of ras-transformed cells while repressing the transcription of stromelysin-1. Mol Cell Biol. 2000;20(16):5828-5839.
15. Irvin BJ, Wood LD, Wang L, et al. TEL, a putative tumor suppressor, induces apoptosis and represses transcription of Bcl-XL. J Biol Chem. 2003;278(47):46378-46386.
16. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
17. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26 (1):139-140.
19. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25(4):402-408.
20. Shanks RA, Larocca MC, Berryman M, et al. AKAP350 at the Golgi apparatus. II. Association of AKAP350 with a novel chloride intracellular channel (CLIC) family member. J Biol Chem. 2002;277(43):40973-40980.
21. Flores-Tellez TN, Lopez TV, Vasquez Garzon VR, Villa-Trevino S. Co-Expression of Ezrin-CLIC5-Podocalyxin Is Associated with Migration and Invasiveness in Hepatocellular Carcinoma. PLoS One. 2015;10(7):e0131605.
22. Xu Y, Kang J, Yuan Z, et al. Suppression of CLIC4/mtCLIC enhances hydrogen peroxide-induced apoptosis in C6 glioma cells. Oncol Rep. 2013;29(4):1483-1491.
23. Turk B, Turk V. Lysosomes as "suicide bags" in cell death: myth or reality? J Biol Chem. 2009;284(33):21783-21787.
24. Aits S, Jaattela M. Lysosomal cell death at a glance. J Cell Sci. 2013;126(Pt 9):1905-1912.
25. Kruszewski M. Labile iron pool: the main determinant of cellular response to oxidative stress. Mutat Res. 2003;531(1-2):81-92.
26. Kurz T, Leake A, Von Zglinicki T, Brunk UT. Relocalized redox-active lysosomal iron is an important mediator of oxidative-stress-induced DNA damage. Biochem J. 2004;378(Pt 3):1039-1045.
27. Boya P, Kroemer G. Lysosomal membrane permeabilization in cell death. Oncogene. 2008;27(50):6434-6451.
28. Oller AR, Thilly WG. Mutational spectra in human B-cells. Spontaneous, oxygen and hydrogen peroxide-induced mutations at the hprt gene. J Mol Biol. 1992;228(3):813-826.
29. Repnik U, Cesen MH, Turk B. The endolysosomal system in cell death and survival. Cold Spring Harb Perspect Biol. 2013;5(1):a008755.
30. Berryman M, Bruno J, Price J, Edwards JC. CLIC-5A functions as a chloride channel in vitro and associates with the cortical actin cytoskeleton in vitro and in vivo. J Biol Chem. 2004;279(33):34794-34801.
31. Landry DW, Akabas MH, Redhead C, Edelman A, Cragoe EJ, Jr., Al-Awqati Q. Purification and reconstitution of chloride channels from kidney and trachea. Science. 1989;244(4911):1469-1472.
32. Bresgen N, Eckl PM. Oxidative stress and the homeodynamics of iron metabolism. Biomolecules. 2015;5(2):808-847.
33. Maxfield FR, McGraw TE. Endocytic recycling. Nat Rev Mol Cell Biol. 2004;5(2):121-132.
34. Andreasson P, Schwaller J, Anastasiadou E, Aster J, Gilliland DG. The expression of ETV6/CBFA2 (TEL/AML1) is not sufficient for the transformation of hematopoietic cell lines in vitro or the induction of hematologic disease in vivo. Cancer Genet Cytogenet. 2001;130(2):93-104.
35. van der Weyden L, Giotopoulos G, Rust AG, et al. Modeling the evolution of ETV6-RUNX1-induced B-cell precursor acute lymphoblastic leukemia in mice. Blood. 2011;118(4):1041-1051.
36. Ross ME, Zhou X, Song G, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood. 2003;102(8):2951-2959.
37. Scott CC, Gruenberg J. Ion flux and the function of endosomes and lysosomes: pH is just the start: the flux of ions across endosomal membranes influences endosome function not only through regulation of the luminal pH. Bioessays. 2011;33(2): 103-110.
38. Singh H, Cousin MA, Ashley RH. Functional reconstitution of mammalian 'chloride intracellular channels' CLIC1, CLIC4 and CLIC5 reveals differential regulation by cytoskeletal actin. FEBS J. 2007;274(24):6306-6316.
39. Liu J, Masurekar A, Johnson S, et al. Stromal cell-mediated mitochondrial redox adaptation regulates drug resistance in childhood acute lymphoblastic leukemia. Oncotarget. 2015.
40. Kantner HP, Warsch W, Delogu A, et al. ETV6/RUNX1 induces reactive oxygen species and drives the accumulation of DNA damage in B cells. Neoplasia. 2013;15(11):1292-1300.
41. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-421.
42. Cooke MS, Evans MD, Dizdaroglu M, Lunec J. Oxidative DNA damage: mechanisms, mutation, and disease. FASEB J. 2003;17(10):1195-1214.
43. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585-598.
44. Yau C, Sninsky J, Kwok S, et al. An optimized five-gene multi-platform predictor of hormone receptor negative and triple negative breast cancer metastatic risk. Breast Cancer Res. 2013;15(5):R103.
45. Shin G, Kang TW, Yang S, Baek SJ, Jeong YS, Kim SY. GENT: gene expression database of normal and tumor tissues. Cancer Inform. 2011;10:149-157.
46. Araya CL, Cenik C, Reuter JA, et al. Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nat Genet. 2015.
47. Wang L, Zhang Q, Liu B, Han M, Shan B. Challenge and promise: roles for Livin in progression and therapy of cancer. Mol Cancer Ther. 2008;7(12):3661-3669.
48. Thorin-Trescases N, Thorin E. Angiopoietin-like-2: a multifaceted protein with physiological and pathophysiological properties. Expert Rev Mol Med. 2014; 16:e17.
49. Holleman A, den Boer ML, Cheok MH, et al. Expression of the outcome predictor in acute leukemia 1 (OPAL1) gene is not an independent prognostic factor in patients treated according to COALL or St Jude protocols. Blood. 2006;108(6):1984-1990.