Tampere University

SUSANNA TEPPO

Genome-wide Transcriptional Characterization of the ETV6-RUNX1-positive Childhood Leukemia

Tampere University Dissertations 239

Tampere University Dissertations 239

SUSANNA TEPPO

Genome-wide Transcriptional Characterization of the ETV6-RUNX1-positive Childhood Leukemia

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the auditorium F115
of the Arvo building, Arvo Ylpön katu 34, Tampere,
on the 3rd of April 2020, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine and Health Technology
Finland

Responsible supervisor and Custos	Docent Olli Lohi Tampere University Finland	
Supervisor	PhD Keijo Viiri Tampere University Finland	
Pre-examiners	Docent Pieta Mattila University of Turku Finland	Docent Gisela Barbany Karolinska Institutet Sweden
Opponent	Professor Monique den Boer Princess Máxima Center for Pediatric Oncology, Utrecht Erasmus University Medical Center, Rotterdam The Netherlands	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-1526-9 (print)
ISBN 978-952-03-1527-6 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-1527-6

PunaMusta Oy – Yliopistopaino
Tampere 2020

ACKNOWLEDGEMENTS

This work was done in Tampere University, Faculty of Medicine and Health Technology, and at Tampere Center for Child Health Research, affiliated in the Tampere University and Tampere University Hospital. I am thoroughly grateful to Finland, its institutions, and the people who keep the stones rolling. I want to thank one of the founders and the soul of the Tampere Center for Child Health Research, Emer. Prof. Markku Mäki, as well as the current director Prof. Per Ashorn, and Prof. Kalle Kurppa, for the good working environment and for the opportunities to learn different aspects of children's health. I am also thankful for the support from science foundations - Finnish Hematology Association (SHY), Ida Montin Foundation, the Cancer Society of Finland (Syöpäsäätiö), Emil Aaltonen Foundation, the Orion Research Foundation, and the Finnish Hemopathy Foundation (Veritautien tutkimussäätiö) - which has enabled me to focus on this work and participate in world-class meetings during the past years.

I would like to warmly thank the leader of our group, my supervisor, docent, MD Olli Lohi, for the amazing time I've got to spend diving into molecular biology in pediatric hemato-oncology and beyond. I am greatly thankful for all the trust, opportunities and demands you have given me, which have enabled growth towards being an actual scientist. I have always felt proud to present our work in meetings which is fundamentally accounted for your significant research themes and integrity. I would also like to thank my other supervisor, PhD Keijo Viiri. I have been extremely fortunate to know I have an expert bioscientist to lean on if anything goes wrong in the lab.

I whole-heartedly thank Assoc. Prof. Merja Heinäniemi. This thesis would be something totally different without your contribution. I have also been privileged to take part on other research projects that you lead. You are an exceptional bright-minded scientist who have kept challenging the rest of our working group with ideas and questions. I would also like to warmly thank Prof. Matti Nykter, Prof. Ann-Christine Syvänen, PhD Jessica Nordlund, Assoc. Prof. Minna Kaikkonen-Määttä, MSc Tapio Vuorenmaa, MSc Thomas Liuksiala, and all the other co-authors for making these publications possible. I also thank Adj. Prof. Leena Latonen and MD,

docent Ilkka Junttila who offered a valuable back-up support as the members of my thesis follow-up group. The pre-examiners of this thesis, docent Pieta Mattila and docent Gisela Barbany are thanked for the excellent help in improving the text and for the words of encouragement.

I would like to thank all the past and present, wonderful HemoRes-scientists for the help, discussions, and time together, especially Kaisa Teittinen, Toni Grönroos, Saara Laukkanen, Laura Oksa, Miikka Voutilainen, Artturi Mäkinen, Atte Nikkilä, Veronika Zapilko, and Noora Hyvärinen. I would also like to thank Jorma Kulmala for keeping the cells and people happy. I also thank Mikko Oittinen and all the other great researchers in Keijo's group for sharing the office and making the workdays more delightful. Joel Johnson is also greatly acknowledged for the language review throughout this thesis. I am also happy to have been a part of an extended research team with the excellent scientists in Kuopio, especially Juha Mehtonen, Mari Lahnalampi, Maria Bouvy-Liivrand, and Petri Pölönen. In addition, I have had a great pleasure to share the lab and thoughts with the brilliant scientists in the CeliRes group, especially Laura Airaksinen, Minna Hietikko, Heidi Kontro, and Suvi Kalliokoski, as well as Anne Heimonen, Soili Peltomäki, and Kaija Laurila. You had a great positive influence on the working atmosphere in our shared floor.

I would want to thank my brother, MD Eero Teppo. You are not only one of the most intelligent but also the most thoughtful person I know. The science community is so lucky to have you in. Among the many things I've learnt from you, you have kept reminding me of the bigger pictures in science when my mind has got lost in a detailed biochemical swamp. I am also cordially thankful to Hanna, Anniina, and Annukka, and my extended family and friends, for all the support and time together. Importantly, I want to thank my parents for passing on the fundamental mindset on how to manage this or any work.

Kristian, of the countless things I appreciate in you, here I especially want to thank for your unfaltering support during the long final steps of this process.

ABSTRACT

Acute lymphoblastic leukemia (ALL) is the most common cancer affecting in childhood. It occurs typically in early B-lineage cells and is characterized by a few specific initiating genomic alterations. One of the most common alterations is the translocation resulting in the *ETV6-RUNX1* (*E/R*) fusion gene. Progression to overt ALL requires additional genetic abnormalities that are recurrently found at essential B-cell lineage identity determining genes. Besides DNA, alterations in various RNA species and proteins could also have marked unwanted effects on cell behavior. E/R functions as an aberrant transcription factor but its direct target genes have thus far remained uncertain.

We set out to study genome-wide gene regulation in childhood precursor B-ALL (preB-ALL) by studying nascent RNA transcription in cell lines and patient samples. For the examination of target sites, we generated a cell line model with an inducible E/R. We detected enhancer regions by the expression of eRNA transcripts and deciphered a possible target gene by correlating between expression level changes. Two thirds of the E/R-regulated genes were repressed by direct regulation via RUNX1 DNA binding. We further showed E/R-mediated downregulation of B-cell specific super-enhancers. Some of the regulated genes were observed to be differentially expressed among E/R patients when compared to other preB-ALL patients.

RAG and AID are enzymes that have been linked to the genesis of secondary genetic alterations in B-cell leukemia. We explored the nascent RNA transcription across B-lymphoid cells at the genomic sites that are often deleted in childhood precursor B-ALL and noticed significant association with specific transcriptional features, namely RNA polymerase II stalling and convergent transcription. These features seem to expose the DNA to double strand breaks especially by revealing RAG recombination signal sequences. We noticed high *RAG1* expression in the E/R subtype, and abnormal expression of *AICDA* among the non-classified precursor B-ALL cases.

This thesis identifies genome-wide targets of the E/R fusion and specific transcriptional features that are associated with recurrent DNA breakpoint sites in childhood precursor B-ALL.

TIIVISTELMÄ

Akuutti lymfoblastileukemia (ALL) on lasten yleisin syöpä. Useimmiten se saa alkunsa epäkypsästä B-solusta (preB), jossa tapahtuu tietty altistava geneettinen muutos. Yksi yleisimmistä muutoksista on translokaatio, joka johtaa *ETV6-RUNX1* (*E/R*) fuusiogeenin syntymiseen. Leukemian puhkeamiseen vaaditaan lisäksi muita geneettisiä muutoksia, jotka usein osuvat B-solun identiteetille tärkeisiin geeneihin. DNA-vaurioiden lisäksi solun toiminta voi häiriintyä RNA-molekyylien ja proteiinien toiminnan muutoksista. E/R on epänormaali transkriptiotekijä ja sen suorat säätelykohteet ovat vielä jääneet epäselviksi.

Tässä työssä tutkimme lasten prekursori B-ALL:ssa (preB-ALL) tapahtuvaa genominlaajuista geeniensäätelyä tarkastelemalla varhaista RNA-transkriptiota solulinjoissa ja potilasnäytteissä. E/R-fuusion kohdegeenien kartoittamista varten teimme solulinjamallin, jossa fuusion tuotantoa voidaan säädellä. Määritimme tehostaja-alueet tehostaja-RNA:iden (eRNA) ilmentymisen perusteella sekä niiden mahdolliset kohdegeenit perustuen signaalimuutosten samankaltaisuuteen. E/R-fuusion säätelemistä geeneistä kaksi kolmasosaa hiljeni suoran RUNX1-välitteisen DNA-sitoutumisen kautta. Lisäksi E/R vähensi B-solu-spesifisten tehostaja-alueiden luentaa. Osa geeneistä myös ilmentyi eri tavalla E/R-potilaiden leukemiasoluissa verrattuna muiden preB-ALL alityyppien potilaiden soluihin.

RAG ja AID entsyymit on liitetty DNA-katkosten syntymiseen B-solu-leukemiassa ja niiden toimintaan tiedetään liittyvän avoimena oleva kromatiini. Tutkimme RNA-transkriptiota B-linjan soluissa keskittyen lasten leukemiassa usein nähtäviin DNA-katkoskohtiin. Huomasimme, että katkoskohtiin assosioituvat tietyt transkriptionaaliset ominaisuudet: RNA-polymeraasin pysähtyminen sekä yhtäaikainen geenienluenta päällekkäisiltä DNA-juosteilta. Nämä piirteet näyttävät altistavan DNA:n katkoksille erityisesti paljastamalla RAG-entsyymin rekombinaatiosignaalisekvenssejä. Huomasimme myös korkean *RAG1*-geenin luennan erityisesti E/R-potilailla sekä AID-entsyymiä koodaavan geenin epätavallisen luennan osalla korkean riskin preB-ALL potilaita.

Tässä väitöskirjassa tunnistettiin E/R-fuusion genominlaajuisia säätelykohteita sekä toistuvien DNA-katkosten kohdille ominaisia transkriptionaalisia piirteitä lasten leukemiassa.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

ALL	acute lymphoblastic leukemia
AML	acute myeloid leukemia
bp, kb	base pair, kilobase
cDNA	complementary DNA
ChIP-seq	chromatin immunoprecipitation sequencing
CLL	chronic lymphocytic leukemia
CLP	common lymphoid progenitor
CML	chronic myeloid leukemia
convT	convergent transcription
DNA	deoxyribonucleic acid
DNase-seq	DNase I hypersensitive sites sequencing
DRIP-seq	DNA:RNA immunoprecipitation sequencing
E/R	ETV6-RUNX1
eRNA	enhancer RNA transcript
FANTOM	Functional annotation of the mammalian genome
FISH	fluorescence in situ hybridization
GRO-seq	global run-on sequencing
H3K27ac	acetylation of histone 3 lysine 27
H3K4me3	trimethylation of histone 3 lysine 4
HAT	histone acetyl transferase
HDAC	histone deacetylase
Hi-C	chromosome conformation capture method
IGH	immunoglobulin heavy chain
iPS	induced pluripotent stem cell
LMPP	lymphoid-primed multipotent progenitor
lncRNA	long non-coding RNA
miRNA	microRNA
MNase-seq	micrococcal nuclease sequencing
MRD	minimal residual disease, residual malignant cells

pat	promoter upstream transcript
pol II	RNA polymerase II
preB-ALL	precursor B-cell acute lymphoblastic leukemia
proB	progenitor B-cell
qPCR	quantitative real-time polymerase chain reaction
RHD	Runt homology domain
RLFS	R-loop forming sequence
rRNA	ribosomal ribonucleic acid
RSS	recombination signal sequence
sgRNA	small guide RNA
snoRNA	small nucleolar RNA
SNP	single-nucleotide polymorphism
TAD	topologically associating domain
TF	transcription factor
TSS	transcription start site
TTS	transcription termination site

ABL1	ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase
AICDA, AID	Activation-induced cytidine deaminase
APOBEC	apolipoprotein B mRNA editing enzyme
ARPP21	CAMP Regulated Phosphoprotein 21
BCR	BCR Activator Of RhoGEF And GTPase
BET	bromodomain and extraterminal domain protein family
BRD	Bromodomain Containing
BTG1	BTG Anti-Proliferation Factor 1
CBFB	Core-Binding Factor Subunit Beta
CDKN2A	Cyclin Dependent Kinase Inhibitor 2
CEBPA	CCAAT Enhancer Binding Protein Alpha
CLIC5	Chloride Intracellular Channel 5
CREBBP	CREB Binding Protein
CRLF2	Cytokine Receptor Like Factor 2
CTCF	CCCTC-Binding Factor
DUX4	Double Homeobox 4
EBF1	EBF Transcription Factor 1
EPOR	Erythropoietin receptor
ERG	ETS Transcription Factor ERG

ETS	E26 transformation-specific family
ETV6	ETS Variant Transcription Factor 6
FOXO1	Forkhead Box O1
GTF2B	general transcription factor, transcription initiation factor IIB
H3	histone 3
IGH	Immunoglobulin Heavy Locus
IGLL1	Immunoglobulin Lambda Like Polypeptide 1
IKZF1	IKAROS Family Zinc Finger 1
KMT2A	Lysine Methyltransferase 2A
MYOD1	Myogenic Differentiation 1
NR3C1	Nuclear Receptor Subfamily 3 Group C Member 1
p300	E1A Binding Protein P300
PAX5	Paired Box 5
PI3K	phosphoinositide 3-kinases
RAG	recombination-activating gene
RUNX1	RUNX Family Transcription Factor 1
Ser2P, Ser5P	Serine 2/5 phosphorylation
SOX	SRY-Box Transcription Factor
SPI1, PU.1	Spi-1 Proto-Oncogene
TAL1	TAL BHLH Transcription Factor 1
TBL1XR1	Transducin Beta Like 1 X-Linked Receptor 1
TCF3	Transcription Factor 3
VLA-4	Integrin α4β1 (Very Late Antigen-4)
VPREB1	V-Set Pre-B Cell Surrogate Light Chain 1

ORIGINAL PUBLICATIONS

I Teppo, S., Laukkanen, S., Liuksiala, T., Nordlund, J., Oittinen, M., Teittinen, K., Grönroos, T., St-Onge, P., Syvänen, AC., Nykter, M., Viiri, K., Heinäniemi M.*, & Lohi, O.* (2016). Genome-wide repression of eRNA and target gene loci by the ETV6-RUNX1 fusion in acute leukemia. *Genome Research.* 26(11): 1468–1477.

II Teppo, S., Heinäniemi, M., & Lohi, O. (2017). Deregulation of the non-coding genome in leukemia. *RNA Biology.* 14(7): 827-830.

III Heinäniemi, M., Vuorenmaa, T.*, Teppo, S.*, Kaikkonen, M. U.*, Bouvy-Liivrand, M., Mehtonen, J., Niskanen, H., Zachariadis, V., Laukkanen, S., Liuksiala, T., Teittinen, K., & Lohi, O. (2016). Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots. *eLife.* 5: e13087.

* equal contribution

1 INTRODUCTION

Acute leukemia is a type of blood cancer which is characterized by rapid proliferation and growth of abnormal cells that fill the bone marrow. Leukemia is the most common cancer in childhood and is diagnosed in approximately 4000 children each year in Europe. The incidence peak of acute lymphoblastic leukemia (ALL) is at 2-5 years of age and most cases arise in precursor B-cells. Remarkable progress has been made in the treatment of childhood ALL during the past decades, with the current cure rate of over 90%. This success is mainly brought about by the use of conventional cytotoxic chemotherapy that is also associated with major short- and long-term side-effects. Treatment is currently tailored according to risk grouping, which is partly defined by the underlying genetics and the treatment response.

The genomic diversity of childhood ALL has been investigated in several studies and the classification of leukemia subtypes has progressed rapidly during the recent years. These improvements have been made possible by the advances in genomics, including novel technologies and integration of data types such as DNA alteration, RNA expression, and epigenetic data. Relatively few secondary genomic alterations are typically found in childhood ALL but, curiously, they seem to accumulate to certain genomic sites.

The ETV6-RUNX1 subtype comprises approximately 25% of childhood B-ALL cases. The translocation between chromosomes 12 and 21, first noticed in the early 1990s, occurs *in utero* during fetal hematopoiesis at an early B-cell progenitor cell, and additional alterations accumulate before leukemia initiation during early childhood. Precursor B-ALL subtypes differ by sensitivity to chemotherapeutic drugs and by the overall gene expression profiles. This implies that the initiating alteration, including the ETV6-RUNX1 fusion, induces unique genetic and molecular features in the leukemic cell.

Many studies have aimed at revealing the role played by the ETV6-RUNX1 fusion in leukemia, but many details are yet to be elucidated. The research presented in this thesis was aimed at gaining further knowledge on the transcriptional regulation and molecular biology of this subtype of childhood ALL.

2 REVIEW OF THE LITERATURE

2.1 Cancer in children

There are 17 million new cancer cases and 9.6 million cancer deaths worldwide each year. Less than 1% of cancers occur in children. (*Cancer Statistics for the UK*). Although the 5 year survival rate of all cancers is over 80%, almost a hundred thousand children die for it every year worldwide (Sullivan *et al.*, 2013). In Finland, around 150 children (age < 15 years) are diagnosed with cancer each year, and cancer causes around 15% of childhood deaths (in 2017, 24 of the 182 deaths) (SVT, kuolleisuustilasto). The most common cancer in children is leukemia (around 35%), followed by central nervous system tumors and lymphomas (Madanat-Harjuoja *et al.*, 2014). Mutational load is significantly lower in pediatric cancers than in adult cancer types (Gröbner *et al.*, 2018). Treatment of leukemia in children is characterized as one of the major successes of chemotherapy of cancer.

2.1.1 Childhood acute lymphoblastic leukemia

Childhood acute leukemia incidence is around 40-50 per one million children per year (*Syöpä Suomessa Syöpärekisteri*; Steliarova-Foucher *et al.*, 2017). The proportion of leukemia of all cancers is the highest among children (35% among 0-9 years old) and decreases by age (15% of cancers among 15-19 of age and 3% among adults), while the proportion of epithelial cancers increases (*Cancer Statistics for the UK*; Steliarova-Foucher *et al.*, 2017). Most leukemia cases in children are acute lymphoblastic (ALL), in contrast to myeloid leukemias or chronic types. In contrast, only 10% of the adult leukemias are ALLs (most being chronic lymphocytic or acute myeloid diseases). Approximately the same number of adults and children are diagnosed with ALL each year (*Cancer Statistics for the UK*; *Syöpä Suomessa - Syöpärekisteri*).

Childhood ALL can be divided into subgroups based on cell lineage (B- or T-cells), differentiation status (early progenitors or precursors), and genetics. Ninety percent of early childhood ALL arise in B-cells (Toft *et al.*, 2018). Symptoms include fever, fatigue, hemorrhage, and paleness. Patients in the Nordic countries receive

standard chemotherapy treatment according to the contemporary NOPHO (Nordic Society for Pediatric Hematology and Oncology) protocol (Toft *et al.,* 2018). Treatments are tailored based on age, white blood cell counts, minimal residual disease after induction chemotherapy, and specific genetic subtypes. High risk patients are often defined by the age over 10 years, white blood cell count over 50 000/μl, and/or by having hypodiploid or BCR-ABL1 genetics. These features were reported to identify 12% of preB-ALL patients with less than 50% relapse-free survival. (Harvey *et al.*, 2010.) However, advances in treating BCR-ABL1-positive patients with tyrosine kinase inhibitors have increased the survival of this subgroup up to 70% (Biondi *et al.*, 2019).

Treatment protocol for adults has been adapted from pediatric protocols, however, lower doses of drugs are needed to avoid induction related deaths (Terwilliger and Abdul-Hay, 2017). By applying pediatric protocol, almost 70% of adult ALL patients achieve long-term remission (Jabbour *et al.*, 2015; Toft *et al.*, 2018). Poorer overall survival is partly due to higher proportion of poor prognostic genetic subtypes in adults: *KMT2A*-rearranged, low hypodiploid, and kinase-driven ALLs account for more than 65% of adult cases (Iacobucci and Mullighan, 2017; Gu *et al.*, 2019).

Childhood leukemia incidence has risen 15% worldwide from the 1980s to 2010 and the reasons are unknown (Steliarova-Foucher *et al.*, 2017). Ionizing radiation exposure is the clearest causal factor for childhood leukemia, especially increasing the risk of B-cell leukemias. Other factors include Down syndrome, germ-line variations in genes linked to B-cell development or DNA repair, and the use of chemotherapy agents (Saida, 2017). On the contrary, breast feeding and daycare attendance are associated as protective factors (Infante-Rivard, Fortier and Olson, 2000; Ma *et al.*, 2002; Greaves, 2018). Evidence supporting the relevance of timing of infections in early childhood has been gained from epidemiological studies and more recently from animal models (Rodríguez-Hernández *et al.*, 2017; reviewed in Greaves, 2018). It has even been suggested that a significant part of leukemias could be prevented (Greaves, 2018).

Relapse in ALL has approximately 10% of incidence and is associated with positive minimal residual disease at the end of induction (Pui and Campana, 2017; Toft *et al.*, 2018). The recurrence of ALL is the most frequent cause of premature death (1% incidence at 10 years from diagnosis), however, patients also have increased risk of non-relapse mortality compared to normal population. Survivors are at increased risk in developing growth hormone deficiency, neuropathy, hypogonadism (related to fertility), and, if treated with anthracyclines, cardiac-related

effects. Late effects in long-term survivors are dependent on treatment regimen and life-threatening effects are no longer as common, although long-term risk based follow-up is needed. (Essig *et al.*, 2014; Ford *et al.*, 2019; Mulrooney *et al.*, 2019.)

Up to 80% of pediatric leukemia cases occur in resource-limited (low- or middle income) countries and survival rates differ between countries (Sullivan *et al.*, 2013; Bonaventure *et al.*, 2017). The differences may reflect differing diagnostic characterization, risk stratification, restrictions in overtreatment, and adherence to protocols by oncologists. Because of the discrepancy, the highest impact globally will come from not only more efficient but also more cost-efficient and local options for examination and care.

2.1.2 B-cell differentiation and leukemia

Billions of blood cells are produced each day in hematopoiesis in human body through proliferation, differentiation and maturation. B-cell differentiation is characterized by specific cell surface markers and recombination of immunoglobulin genes. All blood cells originate from pluripotent hematopoietic stem cells (HSCs). HSCs develop towards lymphoid-primed multipotent cells (LMPPs) and subsequently to common lymphoid progenitor (CLP) population, which can direct differentiation toward either T- or B-cells under specific transcription factor guidance. B-lymphoid directed progenitors remain plastic until an activation loop containing TCF3, FOXO1, EBF1, and PAX5 is complete, after which progenitor-B cells are produced. (Lin *et al.*, 2010; Jacobsen and Nerlov, 2019.) Differentiation towards B-cell lineage is also characterized by the expression of certain genes (Figure 1). After successful rearrangement of immunoglobulin heavy chain (*IGH*) and preB-cell receptor formation (containing IgH; surrogate light chains VPREB1 and IGLL1; and proximal CD79A/B signaling molecules) on the surface, the cell can enter stroma-dependent proliferating large preB state (Joshi *et al.*, 2014). Differentiation is then continued towards small preB and immature B-cell state, at which point cells leave the bone marrow for maturation in secondary lymphoid organs.

Immunoglobulin heavy chain genes are recombined during pro-B states by recombination activating gene (RAG1 and RAG2) mediated cleavage activity. RAG enzymes are produced specifically in lymphoid lineage precursor cells for the crucial process called VDJ-recombination, which eventually leads to production of antibody repertoire needed in mature B-cell mediated immune response. Recombination and cleavage of DNA requires multiple interactions between proteins and DNA features.

RAG1 anchors at recombination signal sequence (RSS) nonamer site in the genome (ACAAAAACC), whereas RAG2 surveys nearby spacer and RSS-heptamer sequences (CACAGTG), binds with methylated H3K4 (Matthews *et al.*, 2007), and serves as a cofactor for RAG1. Cleaved sites are ligated by non-homologous end joining. (Reviewed in Schatz and Swanson, 2011.) RAG1 prefers to bind to single-stranded DNA and the binding affinity is influenced by conformational accessibility to RSS site and by sequence variations exhibited especially in the spacer and the nonamer sequences. Even transcription factors have been suggested to function in targeting the enzyme, such as PAX5 that binds RSS sites in heavy chain variable regions (Zhang *et al.*, 2006).

As in many cancers, cells in leukemia are immature. PreB-ALL cells display a differentiation block at pro- or preB cell state. PreB-ALL blast immunophenotype is usually CD19$^+$, TdT$^+$ (*DNTT*), CD22$^+$, CD79A$^+$ and variably CD10$^+$ and CD34$^+$. Precursor cell states are characterized by the activity of RAGs and on-going *IGH* rearrangement. RAG activity is suggested to be an important mechanism for oncogenic structural variations in ALL by illegitimate off-targeting (Aplan *et al.*, 1990; Zhang and Swanson, 2008; Papaemmanuil *et al.*, 2014).

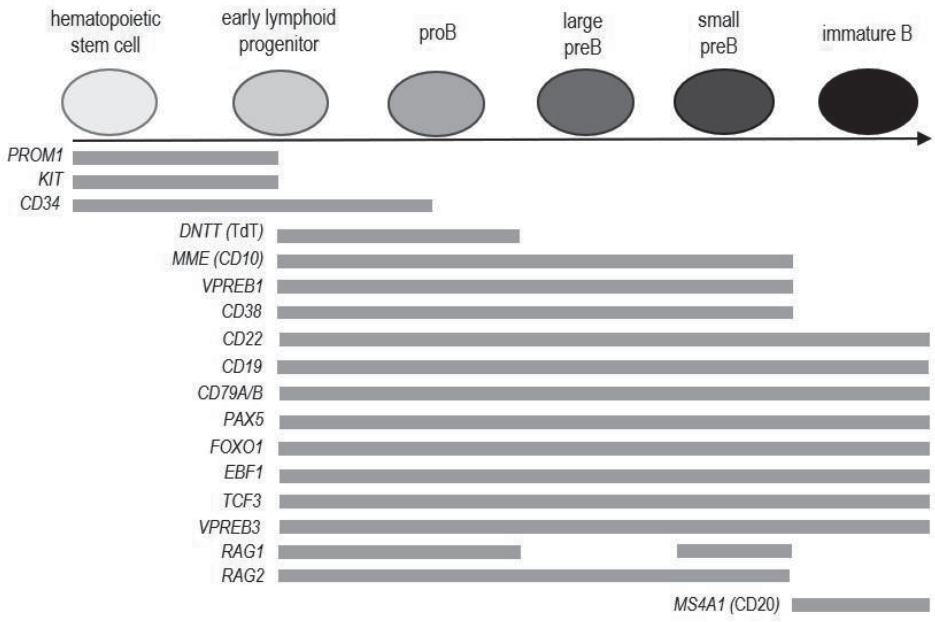

Figure 1. A diagram of the major B-cell states during differentiation in the bone marrow. The bars depict the expression of chosen genes that characterize B-cell differentiation.

2.2 Genetic subtypes of preB-ALL

Childhood B-cell leukemia can be divided in groups based on recurrent structural variations (Figure 2). These groups also differ in their overall transcriptome signal. All risk stratifying changes in chromosomal copy numbers, specific deletions, rearrangements, and fusion genes are assessed in the clinics by SNP arrays, FISH, G-banding, and/or PCR (NOPHO protocol). Studies are on-going to identify new molecules and variations for an improved outcome prediction.

2.2.1 The classical subtypes

The six classical genetic subtypes of precursor B-cell leukemia include: 1) high hyperdiploidy with 51-67 chromosomes; 2) hypodiploidy with less than 44 chromosomes; 3) t(12;21)(p13;q22) translocation encoding ETV6-RUNX1; 4) t(1;19)(q23;p13) translocation encoding TCF3-PBX1; 5) t(9;22)(q34;q11.2) translocation encoding BCR-ABL1; and 6) *KMT2A*- (previously called *MLL*) rearrangements, particularly the t(4;11)(q21;q23) (KMT2A-AF4).

High hyperdiploid subtype has a good prognosis and is present in around 25% of preB-ALL, similarly to ETV6-RUNX1 subtype. Contrary to hyperdiploidy, low or near-haploid hypodiploidy is rare (around 2%) and is presented with poor prognosis (Nachman *et al.*, 2007). Five percent of preB-ALL belong to TCF3-PBX1 group, which has a good prognosis with intensified treatment but which may be associated with increased risk of central nervous system relapse (Jeha *et al.*, 2009).

BCR-ABL1, also called Philadelphia chromosome, is present in 3% of pediatric preB-ALL, and was associated with dismal prognosis before the addition of tyrosine kinase inhibitors to treatment (Druker *et al.*, 2001; Biondi *et al.*, 2019). *KMT2A*-rearrangements are rare in children (1%), mostly occurring in infants (< 1 years of age). This subtype has a very low frequency of somatic mutations, although half of the *KMT2A*-rearranged cases carry activating mutation in a PI3K-RAS pathway component (Andersson *et al.*, 2015). KMT2A is a histone methyl transferase, and many of the KMT2A-fusion partners in infant leukemia are also known to bind with factors playing central roles in transcriptional processes (Mullighan, 2012).

KMT2A-AF4 (Gale *et al.*, 1997), *ETV6-RUNX1* (Hjalgrim *et al.*, 2002; Zuna *et al.*, 2011; Schäfer *et al.*, 2018), *TCF3-PBX1* (Hein *et al.*, 2019), and hyperdiploidy (Taub *et al.*, 2002; Maia *et al.*, 2003) alterations have been suggested to occur *in utero* during fetal hematopoiesis.

2.2.2 New subtypes

Until recently, up to 30% of pediatric B-ALL could not be classified as being any of the known subtypes. New findings based on cytogenetics and gene expression-based classification in B-ALL have identified additional groups and recurrent expressional changes.

Two new subtypes were added into the official WHO classification in 2016: BCR-ABL1-like and iAMP21 (Arber *et al.*, 2016). The BCR-ABL1-like group is characterized by alterations in *IKZF1* and in other kinase genes than *BCR* or *ABL1* (e.g. *ABL2*, *PDGFRB*, and *CSF1R*). It was identified by gene expression profiling in which the cases resembled samples that contained the BCR-ABL1 fusion (Den Boer *et al.*, 2009). The subgroup comprises of around 8% of pediatric preB-ALL patients (Iacobucci and Mullighan, 2017). Many of these cases are sensitive to tyrosine kinase inhibitor treatment. Intrachromosomal amplification of chr 21 (iAMP21) was first characterized as multiple copies of *RUNX1* gene in FISH studies (Coniat *et al.*, 2001; Soulier *et al.*, 2003). However, RUNX1 is not expected to be a driver in this abnormality, and these cases do not usually differ from other subgroups by gene expression profile (Harrison, 2009). Pediatric patients with iAMP21 are typically older and treated on intensive therapy due to initial poor survival. (Harrison, 2009, 2015). These two new classifications have had immediate benefit on prognostication and tailoring the treatment regimen.

Similar to the BCR-ABL1-like group, novel ETV6-RUNX1-like and *DUX4*-rearranged cases were identified by RNA-sequencing (Lilljebjörn *et al.*, 2016). The E/R-like group was characterized by clustering with the E/R samples and by coexisting *ETV6* and *IKZF1* alterations without the E/R translocation. They are also enriched with *ARPP21* deletions (Zaliova *et al.*, 2019). Some cases without the fusion were observed to cluster with E/R cases based on DNA methylation earlier (Nordlund *et al.*, 2015). E/R-like group has approximately 4% incidence among pediatric B-ALL. DUX4 subgroup is associated with *ERG* deletions and has approximately 5% incidence. In addition, ALL cases with *CRLF2* alterations (Russell *et al.*, 2009), and *MEF2D*-, *ZNF384*-, or *PAX5*-rearrangements, were characterized fairly recently and have been classified as their own groups. (Iacobucci and Mullighan, 2017).

Recently, expression profiling and genomic analyses from 1223 B-ALL (children and adults) patients resulted in characterization of six groups not specifically characterized before: 1) *PAX5* and *CRLF2* fusions (9%), 2) PAX5 p.P80R (2%), 3) IKZF1 p.N159Y (< 1%), 4) ZEB2 p.H1038R/IGH-CEBPE (<1%), 5) TCF3/4-

HLF (<1%), and 6) *NUTM1* fusions (2%) (Li *et al.*, 2018). The group with PAX5 and CRLF2 fusions was associated with intermediate risk. As the number of patients in other groups were small, prognosis for them were not yet analyzed. In addition, further classification of B-ALL cases into a total of 23 groups was recently performed using RNA-seq data on a group of 1988 patients of which 1140 were children (< 16 years of age) (Gu *et al.*, 2019). Like in Li *et al.*, this work describes groups for PAX5alt (7% of children), PAX5 P80R (1.3%), IKZF1 N159Y (0.4%), HLF (0.5%), and NUTM1 (0.9%), in addition to previously described subtypes. As many as 13% of cases in this cohort were classified as BCR-ABL1-like. A part of the CRLF2-altered cases was classified within the BCR-ABL1-like subgroup and another part separately (CRLF2 (non-Ph-like), 1%). Now, only five percent of the pediatric cases remained unclassified ("other" subtype). It remains to be seen whether further classification will improve risk stratification and identification of targetable vulnerabilities in each individual's leukemic genome.

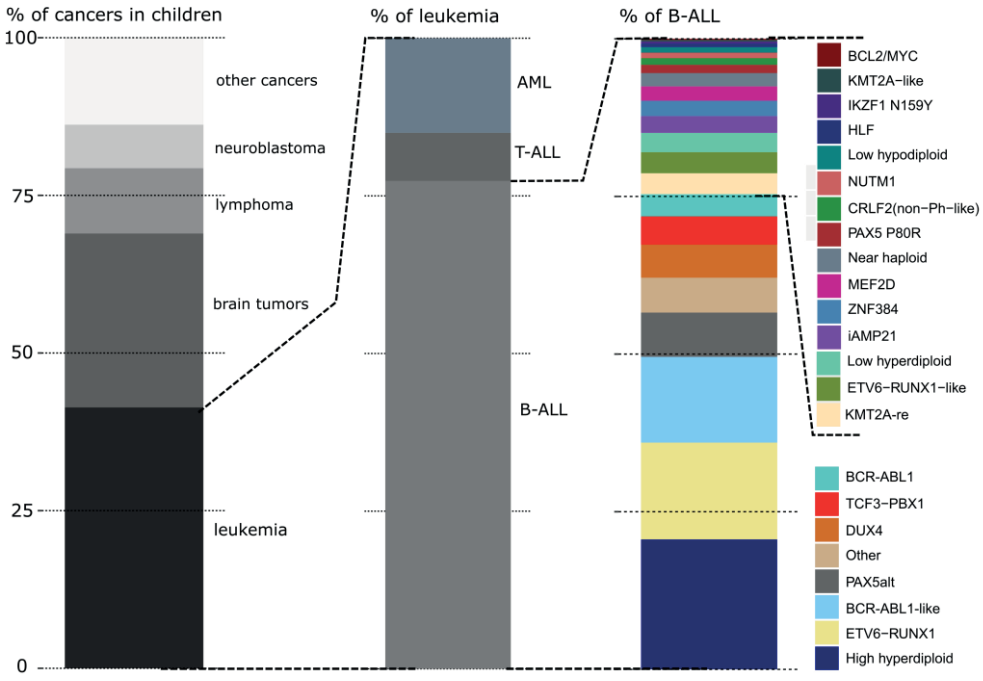

Figure 2. Cancer in childhood. B-cell acute lymphoblastic leukemia (B-ALL) subtype percentages represent patients under 16 years of age in the dataset from Gu *et al.*, 2019.

2.2.3 Secondary genetic alterations

Co-drivers in preB-ALL leukemogenesis have only been started to comprehend. DNA structural variations in ALL cluster in pathways related to transcription factors (TFs), lymphoid cell differentiation, cell cycle, RAS signaling, JAK/STAT signaling, PI3K/AKT/mTOR signaling, chromatin structure modifiers, and epigenetic regulators (Montaño *et al.*, 2018). Recurrently altered genes in preB-ALL include *PAX5*, *IKZF1*, *CDKN2A/B*, *EBF1*, *RAG1/2*, *BTG1*, *TBL1XR1, TCF3,* and *LEF1* (Mullighan *et al.*, 2007; reviewed in Sun, Chang and Zhu, 2017). Secondary alterations are usually not specific to any pediatric preB-ALL subtype and can also be found in adult cases. However, some alterations are enriched to or lack in certain subtypes. For example, alterations in histone modifiers and RAS pathway genes are missing from the ETV6-RUNX1 subtype (Alexandrov *et al.*, 2013; Papaemmanuil *et al.*, 2014). Some studies have aimed to infer the sequential order of secondary mutations in order to decipher significance of each in the clonal process (Anderson *et al.*, 2011). For example, in the E/R disease, *PAX5* and *CDKN2A/B* deletions were shown to occur early in the leukemogenic process. Deletion in the other *ETV6* allele was also shown to occur early, but did not seem to be necessary for any subsequent alterations (Lilljebjörn *et al.*, 2010).

2.3 ETV6-RUNX1

ETV6-RUNX1 (E/R) translocation is found in the cancer cells of 20-25% of child patients diagnosed with B-cell acute lymphoblastic leukemia. The peak incidence is at 2-5 years of age. E/R patients have almost excellent prognosis with the current treatment and minimal residual disease (MRD) follow-up strategies. The 10 year event-free survival was reported to be 95.3% for the E/R-positive preB-ALL patients (Piette *et al.*, 2018). However, E/R subtype is known to have a relatively high late relapse rate, with estimations ranging from 3 to 10%, which correlates with the MRD level after induction treatment (Harbott *et al.*, 1997; Forestier *et al.*, 2008; Bokemeyer *et al.*, 2014; O'Connor *et al.*, 2018). This chapter highlights known characteristics of E/R-positive cells.

2.3.1 Cell of origin

There are two types of evidence for prenatal origin of the E/R translocation. First, E/R has been found in as many as 5% of healthy newborns in cord blood studies, with reports between 0 - 0.01% (Lausten-Thomsen *et al.*, 2011) to 1 - 5% (Mori *et al.*, 2002; Zuna *et al.*, 2011; Schäfer *et al.*, 2018). The prevalence has been under debate between claims of virtually non-existence to a relatively high percentage of newborns that would carry E/R-cells. The most recent work reporting 5% incidence was obtained by studying CD19-enriched mononuclear cells from cord blood with an improved method that, unlike in all the previous reports, investigated DNA rather than the presence of the fusion RNA molecule (Fueller *et al.*, 2014). Despite precursor cells being rare in peripheral blood and cord blood (Kurzer and Weinberg, 2018), it has been possible to detect E/R even in old, dried peripheral blood spots collected from newborns (Guthrie cards) of patients that later developed leukemia (Wiemels, Cazzaniga, *et al.*, 1999; Hjalgrim *et al.*, 2002; Morak *et al.*, 2013).

The second evidence for prenatal occurrence is that if monochorionic twins both get leukemia, they usually have the same E/R breakpoint in their leukemic cells (Ford *et al.*, 1998; reviewed in Ford and Greaves, 2017). This suggests that the preleukemic clone emerges during pregnancy and transfers between the two individuals. A few preleukemic E/R cells have also been detected in samples from the healthy twin of a diseased sibling (Wiemels, Ford, *et al.*, 1999; Hong *et al.*, 2008). The concordance rate in monozygotic twins is 10% (i.e. the healthy sibling has 10% chance of also developing E/R leukemia) (Greaves *et al.*, 2003).

Evidence on cell state origin have been gained from studying immunoglobulin and TCR rearrangements. Most reports have shown similar *IGH/TCR* rearrangements in the cancer cells of twin siblings (Ford *et al.*, 1998; Alpar *et al.*, 2015). As no polyclonal rearrangements were found, it was thought to be unlikely that E/R occurred in a non-committed (RAG- and CD19-negative) cell. In addition, a cell population characterized with $CD34^{+}/CD19^{+}/CD38^{low/-}$ was suggested as the cancer-propagating cells in E/R-leukemia (Castor *et al.*, 2005; Hong *et al.*, 2008). A small fraction of these cells, interpreted as E/R preleukemic cells, was found in the blood of the healthy twin sibling of a leukemic patient, but not in other healthy individuals (Hong *et al.*, 2008).

Fetal hematopoiesis differs from adult hematopoiesis (Böiers *et al.*, 2013; Popescu *et al.*, 2019). Human fetal liver CD19-positive cells were shown to differ from cord blood (neonatal) CD19-positive cells in the expression levels of genes, especially with higher *IL7R*, *KIT*, and *LIN28B*, and far lower *DNTT* expression (Böiers *et al.*, 2018).

At a specific time, 40% of fetal bone marrow cells are proB-cells, of which one third are characterized as being $CD10^-/CD34^+/CD19^+$, named as pre-pro-B cell population, which was almost undetectable in adults (O'Byrne *et al.*, 2019). In addition, when E/R was expressed in a human induced pluripotent stem cell model, a CD19-negative, IL7R-positive cell compartment was expanded (Böiers *et al.*, 2018). These E/R expressing cells were suggested to lie upstream of the pre-proB-cell state (O'Byrne *et al.*, 2019). In conclusion, the translocation is now suggested to occur in either multipotent stem cell or very early committed progenitor during fetal hematopoiesis.

2.3.2 Structure

A reciprocal translocation t(12;21)(p13;q22), rearrangement between *ETV6* (*TEL*) and *RUNX1* (*AML1*), was found in lymphoid leukemias in the 1990s (Romana, Le Coniat and Berger, 1994; Kobayashi and Rowley, 1995). The translocation fuses almost the entire *RUNX1* and the five first exons of *ETV6* (Golub *et al.*, 1995) (Figure 3). Breakpoints cluster relatively closely between patients in introns between exons 1 and 2 of *RUNX1*, and exons 5 and 6 of *ETV6* (Thandla *et al.*, 1999; Wiemels and Greaves, 1999; Wiemels *et al.*, 2000). No specific mutational signature is yet found close to these regions, however, signs of non-homologous end joining repair have been reported (Wiemels and Greaves, 1999; Eguchi-Ishimae *et al.*, 2001; Papaemmanuil *et al.*, 2014).

Both *ETV6* and *RUNX1* are normally expressed in hematopoietic stem cells and progenitor cells. RUNX1 protein is essential for normal fetal hematopoiesis, and in adults the knockdown of *RUNX1* results in expansion of stem and progenitor cell states in addition to impaired B- and T-cell formation (reviewed in Mevel *et al.*, 2019). *RUNX1* was known to be recurrently rearranged in myeloid leukemias before it was found in preB-ALL, and mutations in it are also found in T-cell ALL (Grossmann *et al.*, 2011). ETV6 is crucial in transitioning hematopoiesis from fetal liver to bone marrow. In addition, ETV6 appears non-essential to lymphoid differentiation but important in maintaining a normal progenitor pool in the bone marrow, thus functioning more in promoting self-renewal than differentiation. (Wang *et al.*, 1998; Hock *et al.*, 2004; reviewed in Rasighaemi and Ward, 2017). ETV6 is also seen deleted and translocated in other cancers, especially in 25% of early T-cell leukemias (ETP-ALL) (Zhang *et al.*, 2012).

Protein structures in both ETV6 and RUNX1 have been associated with E/R function. The PNT domain in ETV6 (sometimes called HLH domain, for its helix-loop-helix structure) functions in protein-protein interactions, binding with e.g. another ETV6, other ETS factors, or histone deacetylases (HDACs), resulting mostly in transcriptional repression. The PNT domain was reported essential for differentiation impediment at proB cell state in a mouse E/R model (Fischer *et al.*, 2005). This domain has been found to bind repressor proteins also in the fusion format (ETV6-RUNX1) (Fenrick *et al.*, 1999). Repressive function of the PNT domain in E/R has been shown in reporter gene assays in a T-cell line with IL3 gene (Uchida *et al.*, 1999) and in fibroblasts with TCRα gene (Hiebert *et al.*, 1996). All the initially tested E/R-regulated genes were previously deciphered as RUNX1 targets (listed in Kitabayashi *et al.*, 1998). Functional relevance of the interaction with HDACs have been tested using an HDAC inhibitor (Wang and Hiebert, 2001; Starkova *et al.*, 2007). Expression of the reporter genes and some of the putative E/R-regulated genes (from Fine *et al.*, 2004) were released upon HDAC inhibition (Starkova *et al.*, 2007).

The RHD (Runt-homology domain) mediates DNA-binding in RUNX1 and in E/R. Transduction with a DNA-binding deficient, RHD mutant version (R201Q) of E/R did not lead to enhanced colony formation ability in mouse hematopoietic stem cells like the normal E/R, indicating that E/R directly disturbs RUNX1 targets and not only sequesters co-activators (Morrow *et al.*, 2007). The RHD domain is also needed for heterodimerization with core binding factor beta CBFB, which is important for RUNX1 (also called CBFα) function. For effective binding of RUNX1 to DNA, CBFB blocks inhibition mediated by an adjacent region called NRDB. (Kanno *et al.*, 1998.) Both ETV6 and RUNX1 domains in E/R are reported to bind corepressor Sin3A and to contribute to repression (Fenrick *et al.*, 1999).

The transactivation domain (also called proline, serine, and threonine rich region, PST) interacts with p300, CREBBP, and other transcriptional activators, which are likely to mediate association between multiple transcription factors. The ID domain inhibits the transactivation domain that is located next to it. In addition, the extreme C-terminal VWRPY domain mediates interaction with TLE co-repressor. (Kanno *et al.*, 1998; Kitabayashi *et al.*, 1998).

E/R lacks the ETS domain from ETV6, which, in addition to importantly mediating DNA-protein-interactions, interacts with proteins such as HLH- and Runt-domain proteins (reviewed in Sharrocks, 2001). A small part of patients lack exon 5 of ETV6 (known as the central region) in the E/R protein which was reported to be redundant for E/R mediated transcriptional regulation by a reporter

gene assay and in clinical data (Zaliova *et al.*, 2011), although it has been suggested to be essential to activate progenitor expansion in HSCs (Morrow *et al.*, 2007). NCoR corepressor and HDAC3 have been reported to bind the ETV6 central region (Wang and Hiebert, 2001).

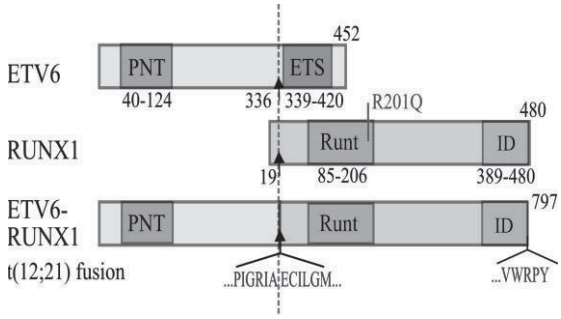

Figure 3. Schematic structure of the ETV6-RUNX1 fusion protein and its wild type partners. AML1c (NP_001745) variant of RUNX1 is visualized. Runt = RHD domain. From Teppo, Heinäniemi and Lohi, 2017.

2.3.3 Alterations in genes and pathways

E/R functions as an aberrant transcription factor. The effects on gene expression have been deduced from comparisons of profiles between the E/R and the other preB-ALL subtypes (Moos *et al.*, 2002; Ross *et al.*, 2003; Fine *et al.*, 2004; Andersson *et al.*, 2005; van Delft *et al.*, 2005; Gandemer *et al.*, 2007) or between E/R-silenced and control cell line with endogenous E/R (Starkova *et al.*, 2007; Fuka *et al.*, 2011; Zaliova *et al.*, 2011; Ghazavi *et al.*, 2016). Few of the reported genes are shared between the studies.

One recurrently reported gene is *EPOR*, which is around 7-fold more expressed in the E/R-subtype than others and its promoter is bound by E/R (Ross *et al.*, 2003; Inthal *et al.*, 2008; Torrano *et al.*, 2011). EPOR is usually restricted to myeloid lineage (Baruchel *et al.*, 1997). JAK inhibitors have been suggested for E/R patients to target EPOR downstream effectors (Chatterton *et al.*, 2014). Another example, *PIK3C3* (Vps34) belonging to a phosphoinositide 3-kinase (PI3K) family, is upregulated in E/R leukemia and there is some evidence that its function in the inhibition of autophagy could be targeted (Polak *et al.*, 2019). PI3K/AKT/mTOR signaling pathway was reported to be active in the E/R subtype and silencing of the E/R led to its inactivation (Fuka *et al.*, 2012). PI3K pathway acts downstream of many

receptors (including EPOR) and its function could be inhibited by e.g. rapamycin (Harrison, 2013). Some other efforts have been made to go beyond individual target genes and understand perturbed pathways and interactions in the E/R disease. For example, using microarray gene expression data and B-cell interactome datasets, MYC was found the most perturbed transcription factor in E/R disease, and cell adhesion genes were specifically targeted in the E/R subtype (Hajingabo *et al.*, 2014). At cell phenotype level, the silencing of E/R in REH cells led to reduced proliferation (Zaliova *et al.*, 2011). Induction of ETV6-RUNX1 has been shown to enhance the self-renewal of progenitor B cells and expand hematopoietic cells or early B-cell progenitors in a fetal mouse cell model (Morrow *et al.*, 2004) and in a human iPS cell model (Böiers *et al.*, 2018).

Recurrent secondary events in E/R-leukemia are most often copy number alterations, mostly deletions, which is characteristic of the E/R subtype (Mullighan *et al.*, 2007; Papaemmanuil *et al.*, 2014). More than 80% of the diagnosed E/R cases display additional alteration in either the non-rearranged alleles of *ETV6* (deletion, 70%) and *RUNX1* (extra copy, 20%), or the derivative chromosome der21(t12;21) (10%, duplication) (Cavé *et al.*, 1997; Stams *et al.*, 2006; Al-Shehhi *et al.*, 2013). Structural alterations comprising the *ETV6* and *RUNX1* genes are sometimes gained in relapse (Peter *et al.*, 2009; Kuster *et al.*, 2011). Loss of *NR3C1*, a gene coding for the receptor responding to glucocorticoid drugs, is present in approximately 10% of the E/R-leukemia relapses and is associated with the E/R subtype (Mullighan *et al.*, 2008; Kuster *et al.*, 2011; Bokemeyer *et al.*, 2014). In a study, all the E/R cases with an *NR3C1* aberration had positive MRD at the end of induction and went to stem cell transplantation (Bokemeyer *et al.*, 2014). Loss of *VPREB1* and *CDKN1B* in relapsed cases is also associated with inferior outcome (Kuster *et al.*, 2011; Bokemeyer *et al.*, 2014). *VPREB1* deletion prevalence was reported to be the highest in cases with E/R (Mangum *et al.*, 2014). However, the so-called NCI risk (MRD ratio, white blood cell count, age, etc.) was shown to be a better prognostic factor in long-term follow-up of E/R patients than any secondary mutation (Enshaei *et al.*, 2013).

All "driver" copy number alterations have been reported to be dissimilar between E/R twins (Bateman *et al.*, 2010). Based on expression profiles, E/R twins were reported to cluster with the other E/R-cases but not specifically with each other (Teuffel *et al.*, 2004). As evidence suggests that E/R-positive precursor cells are found in 1:100 newborn (Mori *et al.*, 2002; Zuna *et al.*, 2011; Schäfer *et al.*, 2018), and the leukemia incidence is in the order of 1:10000, secondary events leading to

leukemia occur much more rarely than the E/R translocation. This also indicates that although E/R occurs *in utero*, all the secondary variations occur after birth.

2.4 Transcription of the genome

Transcription is the process of preparing functional RNA molecules by the instructions encoded in DNA. Regulation of RNA transcription is an important process that guides cell fates during cell differentiation and maintains molecular homeostasis and function throughout cell life. Regulation occurs especially at the enhancer and promoter areas. Technological improvements have made it possible to realize the pervasively transcribed genome: 50 - 80% of the genome is thought to be transcribed in at least some cell type, while still a big part of DNA is repetitive and normally inert (Djebali *et al.*, 2012; Hangauer, Vaughn and McManus, 2013).

One aim in the functional genomics field is to assign a molecular phenotype for each genetic variation. The sequence of the human genome was solved in 2003 but annotation of the various regions is still on-going. Approximately 300 000 human enhancers (12% of the genome) with putative target genes were annotated in 2017, combining data from different sources including Ensembl and FANTOM projects (Fishilevich *et al.*, 2017). Additional works have defined a few millions of enhancers in over a hundred cell types (Gao *et al.*, 2016; Gao and Qian, 2019). Enhancer regions can be localized at several kilobases from transcription start sites. Looping of DNA brings the distal enhancers in the proximity of promoters. Insulators (CTCF) can be situated in between enhancers and promoters and are thought to act as borders of topologically associating domains (TADs) in chromosomes (Dixon *et al.*, 2012).

2.4.1 Transcription factors

Gene expression is regulated through combinatorial action of promoters and regulatory elements which are bound by transcription factors (TFs). TFs orchestrate the regulation of transcriptional networks in cells. Enhancer is a regulatory element that amplifies transcription, is typically a few hundred base pairs long, and contains binding motifs for several TFs. These motifs are 6-12 bp long DNA sequences that are recognized and favored for binding by specific TFs. (Spitz and Furlong, 2012).

TFs can cooperate in regulating the genes by directly interacting with each other or indirectly by recruiting common cofactors or different components of

multiprotein complexes. They can also help in unwinding the chromatin, e.g. as a pioneering TF, or prevent folding and thus serve as a place-holder factor. A pioneering TF, such as PAX5 in B cells (McManus *et al.*, 2011), recruits chromatin modifiers to facilitate binding of other factors, and may not cause any immediate response to gene expression by itself. Place-holder function has been suggested for e.g. SOX proteins: SOX2 is a general TF bound to many sites in embryonic stem cells, potentially keeping these sites open, but later in B-cell development some of the sites are replaced by SOX4. Different modes of actions may explain for the relatively low correlation found between binding of a TF and the expression of the nearby genes. In one study, only 4% of the genes bound by MYOD1 had changes in expression after removing the TF; however, its occupancy was associated with increased level of H3K27ac at the sites. (Spitz and Furlong, 2012.) Transcription factors can also change their influence in a manner that is dictated by chromatin landscape. This has been seen in IKZF1-deficient high-risk preB-leukemia, in which normally supportive EBF1 can be redirected to incorrect enhancer regions and promote an altered B-cell faith (Hu, Yoshida and Georgopoulos, 2017). In addition, binding of a TF with a cofactor protein may change the preference of binding motif even though the cofactor does not contain a DNA binding domain (Siggers *et al.*, 2011).

2.4.2 RNA polymerase II

RNA polymerase II (pol II) is the main enzyme reading the genome and binds to a 50- to 100 bp stretch of DNA. RNA pol II is accumulated in promoters of almost all genes, especially in many developmentally regulated and stimulus-responsive genes (Guenther *et al.*, 2007; Muse *et al.*, 2007). Upon transcription initiation, RNA pol II pre-initiation complex forms with general transcription factors (e.g. GTF2B) (Parvin and Sharp, 1993). RNA pol II releases contact with the general transcription factors at the promoter, and serine residues within the pol II are phosphorylated, resulting in early elongation and subsequently to productive elongation, which are regulated by kinases and other factors. (Nechaev and Adelman, 2011.)

RNA polymerase II that is situated at a promoter is called poised, regardless of its initiation or elongation status. Some of the poised polymerases can be stalled, meaning the elongation complex has stopped RNA synthesis. On the other hand, some of the stalled polymerases can be paused, which specifies that the stalled RNA pol II is expected to continue transcription after a temporary pause. (Nechaev and

Adelman, 2011). Stalling of RNA pol II occurs during elongation at promoter-proximal regions, and is regulated by certain protein factors and DNA signal of the transcribed gene (Nechaev *et al.*, 2010). Stalling is also considered to act as a damage check-point when the machinery encounters transcription-blocking DNA lesions (reviewed in Lans *et al.*, 2019). Pausing of pol II can also poise a gene for activation by maintaining open chromatin near its TSS. Locations of stalled RNA pol II in genomes have been detected by using permanganate which detects single-stranded thymines in DNA (Kainz and Roberts, 1992), by RNA pol II ChIP (Kim *et al.*, 2005), or by analyzing local elevations in nascent RNA signals (Core, Waterfall and Lis, 2008).

Transcriptional phase of the genome-wide RNA pol II complexes can be predicted by ChIP-seq targeting specific domain modifications in the polymerases. Serine residue 5 is phosphorylated (ser5P) during early elongation near the promoter and its abundance decreases toward productive elongation or termination, whereas ser7P and ser2P phosphorylation levels increase towards the end of the process, activating splicing and 3' end processing (Egloff, Dienstbier and Murphy, 2012). H3K36me3 can also serve as a marker for elongation (Bannister *et al.*, 2005). (Adelman and Lis, 2012).

2.4.3 Convergent transcription and DNA:RNA hybrids

Convergent transcription (convT) is defined as overlapping sense and antisense transcription and is a widespread transcriptional feature. Antisense transcripts have been detected in approximately half of the transcribed genes (Core, Waterfall and Lis, 2008). It has been suggested as a mechanism that interrupts transcription by causing collision of RNA pol II molecules moving in the opposite directions (Ward and Murray, 1979). Convergent transcription was also found to be associated with AID enzyme's off-target sites (Meng *et al.*, 2014).

Nascent RNA transcripts can anneal back to the DNA template, forming DNA:RNA hybrid, and displacing the coding strand as single-stranded DNA (Drolet *et al.*, 1995). These structures are called R-loops and they associate with convergent transcription and open chromatin regions. (Reviewed in Skourti-Stathaki and Proudfoot, 2014). R-loop formation is enriched at transcription start and termination sites (TSS and TTS) of genes. It was estimated that approximately 5% of the genome is engaged in R-loops (Lim *et al.*, 2015; Sanz *et al.*, 2016). Although R-loops are part

of the normal processes, they are also actively suppressed by topoisomerase, helicase, RNase H1 activity, and by the rapid processing of RNA.

Genome-wide R-loop distribution has been characterized by using hybrid specific antibody S9.6 based DNA:RNA immunoprecipitation (DRIP-seq) (Ginno *et al.*, 2012; Sanz *et al.*, 2016; Sanz and Chédin, 2019). In addition, inactive RNase H1 ChIP-seq has been used (Chen *et al.*, 2017). R-loops are associated with G-rich RNA (repeats such as CGG/GCC) and intra-strand structure formation. Based on the characteristic structure, R-loop forming sequences (RLFS) have been predicted genome-wide (Jenjaroenpun *et al.*, 2015).

2.4.4 Long non-coding and enhancer RNAs

RNA can act as an important housekeeping molecule (ribosomal RNA and transfer RNA); in messaging and translation (messenger RNA); in regulation (microRNAs, enhancer RNAs); or in RNA processing (small nucleolar RNAs). Long non-coding RNAs (lncRNAs) are a heterogenous group in size (two hundred to one million nucleotides) and function. It is unclear whether they will eventually be subclassified based on their mechanism of action or by active domains, or whether they have some other character to group by (de Hoon, Shin and Carninci, 2015). LncRNAs also encompass the relatively unstable RNA transcripts - enhancer RNAs (eRNAs) and promoter upstream transcripts (PROMPTs; also called promoter antisense transcript, pat, or upstream antisense RNA transcription, uaRNA) (Core, Waterfall and Lis, 2008; Preker *et al.*, 2008). The roles of these transcripts are not clear, however, in some cases the mere transcription has been found to be functionally more relevant than the transcript product (Engreitz *et al.*, 2016).

Genome-wide transcriptional features are able to be captured after invention of high-throughput strand-specific RNA-sequencing (reviewed in Levin *et al.*, 2010) and nascent RNA sequencing methods (Core, Waterfall and Lis, 2008). RNA molecules are transcribed from active enhancer and promoter sites. Enhancer RNAs may keep the chromatin in open conformation and attract components for looping (Kaikkonen *et al.*, 2013) as well as participate in bridging promoter-enhancer connections (Jensen, Jacquier and Libri, 2013; Lai *et al.*, 2013). Enhancers are especially associated with divergent transcription (transcription from both strands to different directions). Genome-wide abundance of genes with bidirectional promoters was reported as 10% in the beginning of 2000 (Trinklein *et al.*, 2004) and later suggested to occur in 80% of active gene promoters (Core, Waterfall and Lis,

2008). Enhancer and lncRNA signals have more recently been studied using single-cell sequencing methods (Kouno *et al.*, 2019). This improved resolution revealed that a subpopulation of cells may only transcribe enhancer from one strand (unidirectional), although enhancers are classically defined as being bidirectional.

Promoters and enhancers share many features like divergent transcription and both are bound by TFs (Core, Waterfall and Lis, 2008). In addition, enhancers can initiate transcription, and promoters can enhance transcription at another promoter. The two are therefore suggested to belong to the same functional unit (Core *et al.*, 2014; Andersson, Sandelin and Danko, 2015). Promoters and transcription start sites (TSSs) can be predicted using histone marker associations. In addition, promoters often have stalled or paused RNA polymerase II. Modifications of histone tails at enhancer and promoter regions reflect their transcriptional activity state. (Spitz and Furlong, 2012). Especially, H3K27ac and H3K4me1/3 correlate with active transcription sites, together with transient H3K79 methylation at enhancers (Bernstein *et al.*, 2002). DNA accessibility measurements (e.g. DNase- or MNase-seq) and the presence of coactivators (e.g. p300) can also guide in enhancer recognition.

Enhancer RNA landscape is highly cell type specific. In addition to different genes being transcribed and regulated, enhancer usage can differ for the same gene, as was shown for the gene *SPI1* (PU.1) in B cells vs. myeloid cells (Leddin *et al.*, 2011). In another example, one enhancer of the 14 detected for *CEBPA* gene was found to regulate the gene expression in myeloid cells only (Avellino *et al.*, 2016). This highly differentiated regulatory system may partly explain how genetic diseases end up being tissue specific. On the other hand, redundancy in enhancer usage has also been reported: another enhancer can replace the function of another, which makes gene regulation less dependent on individual variations (Osterwalder *et al.*, 2018). Large, highly active regions of chromatin called super-enhancers regulate genes critical to cell identity. Super-enhancers differ from typical enhancers in transcription factor density and sensitivity to perturbation (Whyte *et al.*, 2013).

Molecular mechanisms as to how lncRNAs affect transcription include signaling, guiding chromatin modifying enzymes, scaffolding multiple proteins, and acting as a decoy to trap TFs from regulatory sites (reviewed in Wang and Chang, 2011). The functions for most lncRNAs is unknown, although some have been addressed recently e.g. with the help of CRISPR genome editing technique. For example, Liu *et al.* screened 10 000 lncRNAs using sgRNA mediated silencing, and found 230 of them crucial for CML cell line survival (Liu *et al.*, 2018). The on-going FANTOM6

(functional annotation of the mammalian genome) project is focusing on lncRNAs with some preliminary data (Ramilowski *et al.*, 2019 preprint).

Antisense transcription at the coding gene loci is widespread in the genome. Perturbance in antisense transcript can alter the expression of the sense mRNA (Katayama *et al.*, 2005). Antisense transcripts initiate from promoter, terminator, or intronic sequences, and are associated with R-loops. R-loop formation is thought to promote a substantial amount of antisense lncRNA transcription (Tan-Wong, Dhir and Proudfoot, 2019). Antisense transcript expression profiles across cancers have been elucidated from strand-specific RNA-seq (Balbin *et al.*, 2015). One mechanism for the concordantly regulated antisense gene is the stabilization of the sense transcript. Some atlases for lncRNAs have been produced (Hon *et al.*, 2017). Based on a study, 20% of lncRNAs are eRNAs, although this might be underestimation as eRNAs are relatively unstable (Sigova *et al.*, 2013).

3 AIMS OF THE STUDY

ETV6-RUNX1 translocation, resulting in an aberrant transcription factor fusion protein, characterizes the second most common subtype of precursor B-cell acute lymphoblastic leukemia in children. In this thesis, we set out to study the genome-wide transcriptional regulation and features to gain insights for the underlying mechanisms in this disease.

The aims of this study were:

1) to investigate genomic targets and transcriptional regulation by the ETV6-RUNX1 fusion in preB-ALL (I-II);

2) to probe genome-wide nascent RNA profiles from the ETV6-RUNX1 preB-ALL cell lines and patient samples, as well as other preB-ALL subtypes (I-III); and

3) to explore transcriptional features at the recurrent structural variation sites in the ETV6-RUNX1 subtype (III).

4 MATERIALS AND METHODS

Detailed information can be found in the online supplemental materials of the original publications which are referred to in by Roman numerals (I-III).

4.1 Molecular cloning, virus production and transduction (I)

ETV6-RUNX1 cDNA was cloned into inducible LentiX pLVX-Tight-Puro expression vector (Clontech, Mountain View, CA, USA). Point mutation G1553A was implemented using site-directed mutagenesis PCR resulting in R518Q in ETV6-RUNX1 protein (R201Q in normal RUNX1). In addition, short hairpin RNA (shRNA) oligos targeting ETV6-RUNX1 (target sequence GAATAGCAGAATGCATACTT) were cloned into pLVX-shRNA1-vector (Clontech). Transfection grade plasmids were purified using Midiprep PureYield kit (Promega, Madison, Wisconsin, USA) and viruses were produced in HEK293T (ATCC CRL-3216) cells using HTX packaging mix and Xfect reagent (Clontech). Nalm6-cells (ACC 128) were infected with the regulatory vector TetOn Advanced and subsequently with one of the response vectors: pLVX-Tight-Puro-ETV6-RUNX1 (E/R), pLVX-Tight-Puro-ETV6-RUNX1-mutated (E/Rmut), or pLVX-Tight-Puro-LUC (luciferase control) (Clontech). REH cells (ACC 22) were co-infected with viral particles containing the pLVX-shE/R and a construct targeting N-terminus of ETV6 (clone TRCN0000003855, Sigma Aldrich, Saint Louis, MO, USA). Control cell line was produced using shRNA virus against luciferase (SCH007V, Sigma Aldrich). Stably transduced cells were selected with puromycin (0.5 μg/ml or 1 μg/ml, Clontech).

4.2 Cell culture and mononuclear cell extraction (I-III)

Nalm6-cells (ACC 128), REH cells (ACC 22), and KOPN-8 cells (ACC 552) were bought from DSMZ, Braunschweig, Germany) were cultured in RPMI 1640 (#31870074, Gibco, Thermo Fisher Scientific, Waltham, MA, USA), with 2 mM L-

glut, 100 U penicillin, 100 µg/ml streptomycin, and 10% Tet System Approved FBS (Clontech) (Nalm6-Tet-cells) or 10% FBS (Gibco) (normal Nalm-6, REH, and KOPN-8) at 37°C in 5% CO2 and split every 2-3 days. Mycoplasma was routinely measured in the cell lab and found negative (PCR Mycoplasma Test Kit I/C, PromoCell GmbH, Germany). The cell lines were authenticated by short tandem repeats (STR) genotyping (Eurofins Genomics, Ebersberg, Germany), and checked by PCR of the known fusion gene (*ETV6-RUNX1* in REH).

Induction of the E/R fusion gene or the control luciferase was done with 500 ng/ml doxycycline (Clontech) for 4, 12 and 24 hours. Two independent replicates were included in the GRO-seq study. In assessing changes in proliferation of E/R-positive cells, AlamarBlue reagent (Life technologies) was used and the reactions were measured with Tecan fluorometer (Tecan, Switzerland).

Mononuclear cells were extracted freshly using Ficoll-Paque Plus from primary bone marrow samples collected in EDTA tubes (GE Healthcare, Chicago, Illinois, US). Samples were diluted 1:1 in Hank's balanced salt solution (HBSS) (Gibco) and centrifuged through Ficoll at 720 x g, 30 min at room temperature with breaks off. Extracted mononuclear cell layer was washed with cold HBSS 2-3 times and centrifuged at +4C, 400 x g with 5 min in between. Part of the cells were used for nuclei extraction directly while the rest were frozen in 15% dimethyl sulfoxide (DMSO) in 40% FBS/RPMI (Gibco).

4.3 RNA extraction and quantitative PCR (I)

Total RNA was extracted from cells using GeneJET RNA Purification Kit (Thermo Fisher Scientific). cDNA synthesis (reverse transcription) was performed using iScript (BioRad) with 1 µg of RNA as a starting material. RT-qPCR was performed using SsoFast EvaGreen® Supermix (BioRad) and BioRad CFX96TM Real Time System (BioRad). Quantification was done using the relative $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). For ChIP-qPCR, enrichment was calculated relative to input chromatin sample.

4.4 Chromatin immunoprecipitation, western blotting and immunofluorescence staining (I-III)

Crosslinking of proteins to chromatin was performed with 1.1% final concentration of formaldehyde (J.T. Baker, Avantor, Center Valley, PA, USA) for 10-15 min. Two antibodies against ETV6 were pooled in study I and used to detect ETV6-RUNX1 binding to chromatin in cell lines: sc-166835 (Santa Cruz Biotechnology Inc., Dallas, TX, USA; RRID:AB_2101020) and HPA000264 (Atlas Antibodies, RRID:AB_611466), and DNA was sonicated using Covaris S220 (Covaris Inc., Woburn, MA, USA). In addition, antibodies against Ser2P (Abcam Cat# ab5095, RRID:AB_304749), Ser5P (Abcam Cat# ab5131, RRID:AB_449369) (Abcam, Cambridge, MA, USA), and H3K4me3 (Abcam Cat# ab8580, RRID:AB_306649) were used in study III and the chromatin was cut using MNase (cat# 88216, Thermo Fisher). Libraries were prepared for the latter and sequencing data was submitted to the Gene Expression Omnibus (GEO) database (GSE67540).

For western blotting, proteins were extracted from cells with NE-PER nuclear and cytoplasmic extraction kit (Pierce, Life Technologies, Waltham, MA USA), separated on SDS-PAGE, and transferred onto a nitrocellulose membrane. Anti-ETV6 (HPA000264, RRID:AB_611466, 0.4 μg/ml) was used to detect the fusion protein. For immunofluorescence images, cells were spun onto slides, fixed with paraformaldehyde and permeabilized using Triton-X. Anti-ETV6 (1 ug/ml, Atlas Antibodies, HPA000264, RRID:AB_611466) was used to detect E/R fusion protein.

4.5 Nuclei extraction and global run-on sequencing method (I-III)

Nuclei were isolated for run-on yielding ~5×10^6 nuclei per condition. Nuclear run-on (illustrated in Figure 4) was done in the presence of Br-UTP (sc-214314A, Santa Cruz Biotechnology, Inc., Dallas, Texas, U.S.A.). After run-on, Trizol-LS (#10296010, Thermo) was added for RNA-extraction. The run-on products were treated with DNase I (TURBO DNA-free Kit, Invitrogen, Life Technologies), base hydrolyzed (RNA fragmentation reagent, Ambion, Life Technologies), 3-ends were dephosphorylated using PNK (New England Biolabs, Ipswich, MA, USA), and the RNA was immuno-purified using Br-UTP beads (Santa Cruz, CA, USA). Samples were then 3' poly-A tailed (PolyA polymerase, New England Biolabs, Ipswich, MA,

USA) to allow first-strand cDNA synthesis with NTI223 library adaptor, containing poly-dT tail for priming. cDNA synthesis was performed (Superscript III RT) and the samples were treated with exonuclease I (20U/ul) to catalyze removal of excess index-oligos. RNaseH was used to break the RNA-strand from the DNA-RNA-hybrids, and the resulting single stranded cDNA was circularized (CircLigase 100 U/ul, EpiBio). The cDNA was amplified and then concentrated using ChIP DNA clean & concentrator kit (ZymoResearch, Tustin, CA, USA), and purified from Novex 10% TBE gel (180-300 bp). Sequencing was done with Illumina Hi-Seq2000 (GenCore, EMBL Heidelberg, Germany). (Detailed process in the supplemental materials of studies I and III.)

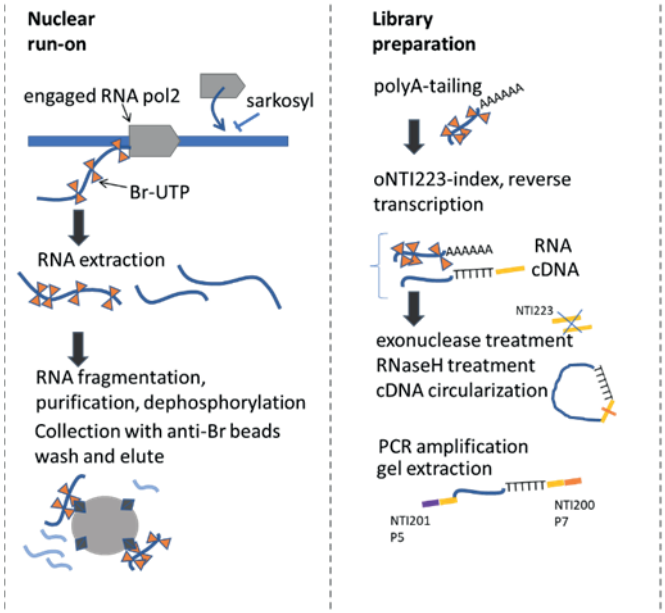

Figure 4. Schematic presentation of the run-on method for GRO sequencing.

4.6 Transcriptome data (I, III)

The microarray gene expression data sets were retrieved from the NCBI GEO database, all measured on Affymetrix GeneChip Human Genome U133 Plus 2.0 array. From a total of 1382 preB-ALL samples, 1004 with known cytogenetics (153 E/R, 153 BCR-ABL1, 151 hyperdiploid, 198 *KMT2A*-rearranged, 82 TCF3-PBX1,

and 267 others) were included in study III, and 664 samples (137 E/R, 76 TCF3-PBX1, 145 BCR-ABL1, 178 *KMT2A*-rearranged, and 128 hyperdiploid) were included in study I. The general analyses of processing the above data are presented in separate publications (Mehtonen *et al.*, 2019; Pölönen *et al.*, 2019). R-package 'mclust' (Fraley, Raftery and Murphy, 2012) was used in the differential expression analysis of microarray probe signals. Genes with log2 fold change > 0.5 or < -0.5 and with adjusted p-value (Mann-Whitney test) below 0.01 were considered differentially expressed. In study III, t-distributed stochastic neighbor embedding (t-SNE)-method was used for visualization of the expression data in two dimensions.

In study I, RNA-seq data for preB-ALL patients was obtained from the GEO database (accession GSE79373) containing nine E/R patients, seven high hyperdiploids, and one other preB-ALL subtype. Aligned reads were summarized using featureCounts (Liao, Smyth and Shi, 2014) and differential expression was calculated using the DESeq2 package (Love, Huber and Anders, 2014).

4.7 Sequencing data (I-III)

Sequencing data that were used in the analyses are summarized in Table 1. Patient samples were chosen by availability to represent good GRO-seq signal in precursor B-cells. Blacklisted sequences were discarded (containing poorly mappable, rRNA, or snoRNA sequences; study III Supplemental file 6). BedGraph and bigWig files were generated with reads normalized to a total of 10^7 mapped reads. The bigWig files were further converted to track hubs and visualized as strand-specific custom Track Hub in the UCSC Genome browser. HOMER (http://homer.ucsd.edu/homer) findPeaks program was used in *de novo* transcript detection from GRO-seq data and in peak detection from ChIP-seq data.

Table 1. Compilation of data produced and reanalyzed in studies I-III. Accession codes refer to NCBI Gene Expression Omnibus database.

Data produced in studies I-II (Teppo *et al.*, 2016; Teppo, Heinäniemi and Lohi, 2017)

Sample	Type	Method	Accession code	Replicates	Main application
Nalm6-LUC 24 h	preB-ALL cell line, control	GRO-seq	GSM1648604, -05	2	Differential expression analysis
Nalm6-LUC 0 h	preB-ALL cell line, control	GRO-seq	GSM1648606, -07	2	Signal observations
Nalm6- E/R 4 h	preB-ALL cell line, E/R induced	GRO-seq	GSM1648608, -09	2	
Nalm6- E/R 12 h		GRO-seq	GSM1648610, -11	2	
Nalm6- E/R 24 h		GRO-seq	GSM1648612, -13	2	Differential expression analysis
Nalm6-E/Rmut 24 h	preB-ALL cell line, E/Rmut induced	GRO-seq	GSM1648614, -15	2	Differential expression analysis
preB-ALL E/R+	Primary cells, bone marrow	GRO-seq	NA	3 patients	Visualization of signal
preB-ALL other	Primary cells, bone marrow	GRO-seq		2	

Other data used in studies I-II

Sample	Type	Method	Accession code	Replicates	Main application
Normal bone marrow	CD34+ cells (HSC)	ChIP-seq / RUNX1, ERG, FLI1	GSE45144 (Beck *et al.*, 2013)	1	Prediction of TF binding
SEM	preB-ALL cell line, *KMT2A*-re	ChIP-seq / RUNX1	GSE42075 (Wilkinson *et al.*, 2013)	1	Prediction of RUNX1 binding and open chromatin regions
Normal bone marrow or cord blood	CD19+, CD20+, or CD34+ cells	ChIP-seq / H3K27ac	GSM1027287, GSM1003459, GSM772870, -85, -94 (Bernstein *et al.*, 2010; Hnisz *et al.*, 2013)	1-3	Super-enhancer data in Hnisz *et al.*, 2013, prediction of open chromatin regions
GM12878, CD34+ -cells, Kasumi-1	Hematopoietic cells	DNase-seq, ChIP-seq / RUNX1, H3K4me1, H3K27ac, P300	See Supplemental methods of study I (ENCODE Project Consortium, 2012; Ptasinska *et al.*, 2012)	24 in total	Prediction of open chromatin regions
Leukemic cells	preB-ALL patients	RNA-seq	GSE79373 (Neveu *et al.*, 2016)	17 patients	Differential expression analysis of transcripts

Data produced in study III (Heinäniemi *et al.*, 2016)

Sample	Type	Method	Accession code	Replicates	Main application
REH	Cell line, ETV6-RUNX1	GRO-seq	GSM1649155 - 62	8	Defining pol II stalling and convT in B-lineage
Nalm6	Cell line, other preB-ALL	GRO-seq	GSM1649153 - 54, GSM1648604 - 05	4	
KOPN-8	Cell line, KMT2A-rearranged	GRO-seq	NA	4	Visualization of signal
ALL patient 1 (del12)	Primary cells, other preB-ALL	GRO-seq		2	
ALL patient 2 (NK)	Primary cells, other preB-ALL	GRO-seq		1	
ALL patient 3 (RUNX1 CNV)	Primary cells, other preB-ALL	GRO-seq		1	

ALL patient 4 (CRLF2+)	Primary cells, other preB-ALL	GRO-seq		1	
ALL patient 5 (Hyperd P1)	Primary cells, high hyperdiploid	GRO-seq		1	
ALL patient 6 (Hyperd P2)	Primary cells, high hyperdiploid	GRO-seq		1	
ALL patient 7 (ETV6-RUNX1)	Primary cells, t(12;21)	GRO-seq		1	
Normal bone marrow	CD19+ mononuclear cells	GRO-seq		1	
Nalm6	Cell line, other preB-ALL	ChIP-seq / phospho Ser2 RNA Pol II	GSM2144895, GSM2144898	2	Evaluating RNA pol II stalling
Nalm6		ChIP-seq / phospho Ser5 RNA Pol II	GSM2144896, GSM2144899	2	
Nalm6		ChIP-seq / input	GSM2144894, GSM2144897	2	
Nalm6		ChIP-seq / H3K4me3	GSM2166074	1	Analysis of overlap between the widest pol II stalling regions and the widest peaks
REH	Cell line, ETV6-RUNX1+ preB-ALL	ChIP-seq / phospho Ser2 RNA Pol II	GSM2144901	1	Evaluating RNA pol II stalling
REH		ChIP-seq / phospho Ser5 RNA Pol II	GSM2144902	1	
REH		ChIP-seq input	GSM2144900	1	
REH		ChIP-seq / H3K4me3	GSM2166075	1	Analysis of overlap between the widest pol II stalling regions and the widest peaks

Data reanalyzed in study III

Sample	Type	Method	Accession code	Replicates	Main application
GM12878, GM12750, GM12004	B-lymphoid cell lines	GRO-seq	GSM1480326, GSM980645, GSM980644 (Core *et al.*, 2014; I. X. Wang *et al.*, 2014)	3	Defining RNA pol II stalling and convT in B-lineage
GM12878	B-lymphoid cell line	HiC	GSM1551571, -72, -74, -75 (Rao *et al.*, 2014)	4	Defining TADs in B-lymphoid cells
H1 ESC	Embryonic stem cell line	GRO-seq	GSM1006728, GSM1006729 (Sigova *et al.*, 2013)	2	Defining RNA pol II stalling and convT in ES-cells
NT2	Embryonic stem cell line	DRIP-seq	GSM1108095 – 99 (Ginno *et al.*, 2013)	3	Testing overlap between DNA:RNA hybrids and RLFS motifs or signal features

Other data used in study III

Sample	Type	Method	Accession code	Replicates	Main application
GM12878	B-lymphoid cell line	DNase-seq	GSM736496, GSM736620, GSE29692	2	Comparison of DNase-seq peak width at sites with and without RNA pol II stalling
H1 and H7 ESC	Embryonic stem cell lines	DNase-seq	GSM736582, GSM736638, GSM736610 (Thurman *et al.*, 2012)	3	
GM12878	B-lymphoid cell line	ChIP-seq / histone modifications	GSE29611 (ENCODE Project Consortium, 2012)	1-4	Analysis of overlap between the widest RNA pol II stalling regions and the widest peaks in each ChIP-seq
Leukemic cells	ETV6-RUNX1	WGS	Papaemmanuil *et al.*, 2014	51 patients	Structural variation breakpoints
Leukemic cells	KMT2A-re	WGS	Andersson *et al.*, 2015	22	
Leukemic cells	hyperdiploid	WGS	Paulsson *et al.*, 2015	16	
Leukemic cells	hypodiploid	WGS	Holmfeldt *et al.*, 2013	20	

4.8 Analysis of GRO-seq data (I)

E/R-regulated novel transcripts were manually classified as either alternative TSSs, eRNAs (representing the most abundantly expressed eRNAs), antisense transcripts, promoter-associated transcripts (pats), or novel intergenic transcripts. Transcripts of <15 kb in length were considered as potential eRNAs. In addition to characteristic bidirectional enhancers, unidirectional enhancers were annotated but only those residing in putative open chromatin regions were accepted.

Gene coordinates were retrieved for all the RefSeq transcripts from the iGenomes database (Illumina, San Diego, USA). When applicable, only intronic regions were quantified from GRO-seq data to accurately catch primary transcription. Differential expression was analyzed using R/Bioconductor package edgeR (Robinson, McCarthy and Smyth, 2009). Significantly regulated genes were defined by adjusted p-value < 0.05 (Benjamini-Hochberg methods using p-values from moderated t-test). Only genes that lacked change in the E/Rmut sample were included. Top 5 enhancers for each coding gene (transcript-centric analysis) were based on Euclidian distance score between the fold change values in the gene and the enhancer region with correlation > 0. Detection of enriched TF binding motifs in enhancer regions close to regulated genes was performed using the HOMER program findMotifsGenome.pl with binomial test.

In the enhancer-centric approach, enhancer detection was based on GRO-seq signal at open chromatin regions (Table 1) and statistical analysis was done as for transcript-centric genes. Enhancer RNAs lacking response in the E/Rmut sample were included. Quantification of eRNA abundance in ChIP-seq mapped transcription factor binding sites was done using the HOMER annotatePeaks.pl program. Before the analysis, the TF peaks were centered by the respective TF motif.

For predicting putative function of the E/R-regulated enhancers, or the E/R-regulated coding transcripts, GREAT v.3.0 tool (McLean *et al.*, 2010) and DAVID tool (Huang, Sherman and Lempicki, 2009) were used, respectively.

4.9 Analysis of transcriptional features at structural variations (III)

Data used in the analyses are summarized in Table 1. Breakpoint coordinates for structural variations were first retrieved from whole genome sequencing data of 51 ETV6-RUNX1-positive patients (Papaemmanuil *et al.*, 2014). Transcriptional

features at the breakpoint regions were then analyzed using GRO-seq signal in the context of topologically associated domains resolved from B-lymphoid cells (Rao *et al.*, 2014). Breakpoint regions were annotated with RSS/heptamer sequence and the overlap with transcription start sites (TSS), RNA polymerase II stalling, R-loop forming sequences (RLFS), and convergent transcription (convT) was determined separately for the non-RSS- and RSS-breakpoints. For assigning breakpoint recurrence, breakpoints at 1 kb distance were stitched together. The overlap frequencies were compared to random sampling of genomic regions of similar size. Breakpoint coordinates were also retrieved from WGS data of high hyperdiploid, hypodiploid, and KMT2A-rearranged cases (Holmfeldt *et al.*, 2013; Andersson *et al.*, 2015; Paulsson *et al.*, 2015) and analyzed separately.

Signal features (convergent transcription and pol II stalling) and *de novo* enhancers were defined from GRO-sequencing data based on deeply sequenced REH, Nalm6, and GM12878 cells. GRO-seq data from replicates were pooled for each cell and sample type (Table 1; study III Supplemental file 1). Convergent transcriptions were identified as overlapping transcription from both strands for at least 100 bp. A combined bed track was then used to analyze the level of convergence at these sites. The gene annotations from UCSC were retrieved (hg19, GRCh37 Genome Reference Consortium Human Reference 37). RNA pol II stalling was quantified from GRO-seq signal across the gene regions using a genome-wide change point analysis which searches for abrupt changes in signal (R-package 'changepoint') (Killick and Eckley, 2014). Signal level at changepoint was compared against the median from across the whole gene. In addition, ChIP-seq data on RNA pol II phosphorylated on serine 2 or 5 were used to analyze stalling regions. In addition, separate analysis for signal features was carried out for embryonic stem cells.

R-loop potential was predicted from DRIP-seq data which is based on a structure-specific antibody that recognizes DNA-RNA-hybrid regions. The signal level from NT2 embryonic stem cells was quantified at transcription start site regions using HOMER. RLFS motif search was performed using the software QmRLFS-finder that uses predictions based on structural models of known sequences (Jenjaroenpun *et al.*, 2015). RLFS motif density was calculated across DNA sequences using BEDtools coverage tool.

DNase-seq and additional ChIP-seq data were used to characterize wide RNA pol II stalling events in embryonic stem cells and B-lymphoid cells. The width of DNase-seq peaks that overlapped with RNA pol II stalling events (from the same cell type) were compared with the ones that did not overlap and the 5% widest regions were identified.

4.10 Statistical tests (I, III)

Binomial test was used in assessing statistical significance for observing breakpoint events overlapping a transcriptional feature (III). For example, in testing the enrichment of breakpoints at convergently transcribed (convT) regions, success in population was defined by the number of 1kb regions containing convT divided by all the analyzed 1kb regions, samples taken were all the regions with breakpoints, and sample success was defined as the number of regions with breakpoints overlapping convT regions divided by the number of all the breakpoint regions. Hypergeometric test was used to assess whether greater overlap frequency was seen between breakpoints and annotated genomic regions (III). For example, in assessing enrichment of breakpoints inside convT-enhancers, samples taken were all the convT-positive enhancers from all the enhancers, population success was all the enhancers with breakpoints, and sample success was a convT-positive enhancer with breakpoint. Hypergeometric test was also performed in assessing overrepresentation of ChIP-peaks in the vicinity of E/R-repressed genes or E/R regulated enhancers with RUNX1-peak in super-enhancer region (study I). The nonparametric Wilcoxon rank sum test (Mann-Whitney) was applied to assess quantified signal levels between categories (e.g. DRIP-seq signal between RLFS-negative and -positive regions) (III). Random sampling to 1000-fold was used in estimating the overlap between stitched breakpoint regions with transcriptional features (III).

5 RESULTS

5.1 ETV6-RUNX1 functions mainly as a repressive transcription factor (I)

We prepared an inducible E/R cell line model to study early effects of the putative aberrant transcription factor. Induction of E/R in Nalm6 cells led to an 18-fold expression level relative to REH preB-ALL cell line (which has endogenous fusion expression) and the protein was visible in western blot after four hours (Figure 5). We collected samples at 0, 4, 12 and 24 hours and performed GRO-sequencing assay for two induction series replicates (Table 1). Indirect effects were filtered out using E/R-mutant version (R201Q) that is not able to bind DNA via RHD (Runt)-domain in RUNX1 (Li *et al.*, 2003).

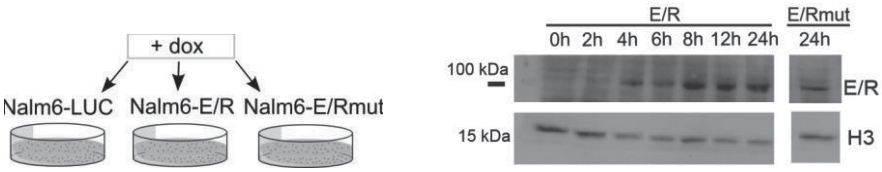

Figure 5. ETV6-RUNX1 induction in the cell model. Adapted from Teppo *et al.,* 2016 *Genome Research.*

Overall, two-thirds of the observed changes in primary transcription at coding regions were repressive. We next studied active enhancer RNA regions nearby the coding genes for enrichment of specific transcription factor motifs (transcript-centric approach; Figure 6A). ETS and RUNX1 motifs were enriched at the associated enhancers nearby the downregulated genes, whereas mainly ETS-factor motifs were enriched nearby the upregulated genes (study I, Figure 2A). In addition to the RUNX motif, we saw enrichment of RUNX1 ChIP-peaks (analyzed from SEM preB-ALL cell line) nearby the downregulated genes (Figure 6B). At least one

enhancer with evidence of RUNX1 site (motif or ChIP peak) was detected for almost all the genes (103/108) identified by the transcript-centric approach (study I Table S2). Either RUNX1 motif or ChIP-peak was identified for 315 of the 467 putative (top 5) enhancers and 244 of the enhancers contained ETS motif. We also saw enrichment of ETS-RUNX motif at RUNX1-peaks near the downregulated genes (study I Figure 2C).

We next examined GRO-seq signal at the top 1000 RUNX1 ChIP-peak regions. The nascent RNA signal was specifically downregulated in the E/R-sample and not in the E/R-mutant version compared to the LUC control (Figure 6C). The signal was also gradually downregulated in the induction time series (study I Figure 3B).

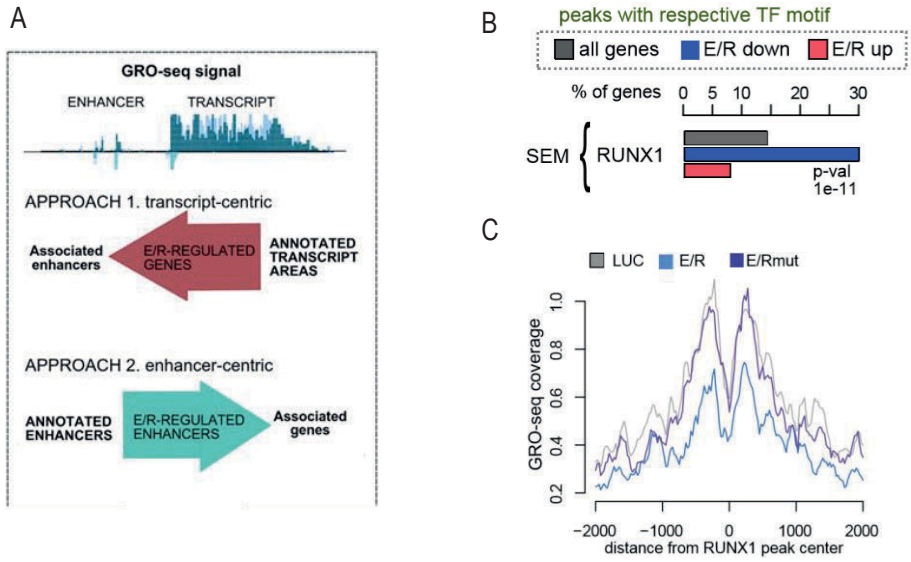

Figure 6. A) A representation of the two approaches used to define ETV6-RUNX1 regulated regions in study I. B) RUNX1-peaks from a ChIP-seq study were enriched nearby the downregulated genes. C) GRO-seq signal from the Nalm6 cell model illustrated at RUNX1-motif centered ChIP-seq peaks. Adapted from Teppo *et al.*, 2016 *Genome Research*.

5.2 Noncoding RNAs in ETV6-RUNX1 leukemia (I-II)

Our analysis revealed genomic regions at which transcription changes after induction of E/R. These changes included annotated genes, both coding and non-coding, as well as yet to be annotated putative long non-coding RNAs and enhancer RNAs. These can be intragenic (antisense transcripts read from the opposite strand of a gene) or intergenic (between annotated genes). Our analysis also revealed novel transcription start sites (TSSs), tails (RNA polymerase running longer than the annotated gene), and promoter antisense transcription (pat).

In addition to eRNAs described earlier, we found 57 novel E/R-regulated long non-coding transcripts. These were further manually classified as potential eRNAs (29, based on size and visual examination), or lncRNAs (28) (study I Table S4 online). 15 of the 57 transcripts studied here showed similar change when we compared the expression in E/R patients vs. other preB-ALL-patients in RNA-seq data. Nine of the 28 lncRNAs were antisense of annotated genes, including a transcript antisense to *IGLL1*. We also found E/R regulation for seven annotated lncRNAs, including *LOC728175* and *LOC374443*.

We compiled a list of B-cell super-enhancers using H3K27ac data accompanied with the most prominent RUNX1-bound sites in SEM and HSC cells (refer to Table 1), and quantified GRO-seq signal from the E/R induction model at these sites. This allowed us to examine regulation directly at putative enhancer sites (enhancer-centric approach, Figure 6A). Of the 534 enhancers defined by H3K27ac and found regulated by E/R in GRO-seq (study I Figure S3 online), 59 were super-enhancers (study I Table S3 online), and the expression level of 28 of them correlated with a nearby gene (study I Table S2 online). RUNX1 ChIP-peaks were found to enrich in super-enhancer regions over all H3K27ac sites (4.4-fold). Similarly, E/R-regulated enhancers that contained RUNX1-peaks were enriched to super-enhancer sites (6-fold) (study I Figure 3C). Overall, two thirds of the regulated super-enhancers were annotated from $CD19^+/CD20^+$-cells and two-thirds were downregulated (study I Table S3 online; study I Figure 3D).

5.3 ETV6-RUNX1 affects genes related to transmembrane signaling (I)

We analyzed gene ontology enrichment of the E/R-regulated enhancer regions and coding genes. Cell adhesion and transmembrane signaling as functional annotation terms were enriched in both analyses. We also tested whether the coding genes found regulated in our cell model were differentially expressed between E/R patients and other preB-ALL patients. For this we used microarray datasets, consisting of E/R, t(1;19), t(9;22), *KMT2A*-rearranged, and high hyperdiploid cases. Total of 133 of the 183 genes were detected in the dataset, and 35 of them were found to be significantly altered between the patient subgroups. Of the 35 genes, 13 had similarly differential expression in another patient cohort in RNA-seq analysis.

For example, direct regulation of *ITGA4* gene, encoding an integrin subunit and related to adhesion, was consistently downregulated in E/R samples, and we showed its regulation at mRNA level and in ChIP-seq (study I Figure 5B). Consistently downregulated genes included *IL21R*, *LAT2, LAIR1, CLEC14A, CLEC2D, CMTM7*, all of which are transmembrane proteins related to immunoregulation.

5.4 R-loops and convergent transcription co-occur with RNA polymerase II stalling (III)

For a general overview on transcriptional features in the genome (see Table 1 for the data produced and used), we first assessed the co-occurrence of R-loop forming sequences (RLFS) and convergent transcription (convT) with RNA pol II stalling sites (Table 2). 30% of the RNA pol II stalled regions contained RLFS which was a significant overlap when compared to random regions. High local RLFS motif density was associated with GRO-seq signal intensity and both elevated at TSS and TTS sites (study III Figure 3 supplement 1 online). We also saw higher incidence of RNA pol II stalling in the sense strands in regions with higher antisense transcription (study III Fig 3B). We assessed whether the presence of RLFS motif at TSSs is associated with DRIP-seq signal level (DNA-RNA hybrids) and found a 2.1-fold increase. Similarly, DRIP-seq signal was elevated at TSS sites with convergent transcription. These results showed that transcription stalling occurs at sites with convergent transcription and RLFS motifs, and is associated with R-loop formation.

Table 2. Enrichment of RLFS motifs and convergent transcription at RNA pol II stalling sites and at DNA-RNA-hybrid sites.

	RLFS MOTIF	CONVERGENT TRANSRIPTION
RNA POL II STALLING (B-LYMPHOID)	30%; $p < 0.001$ for the overlap belonging to the random distribution (16%) (study III Fig 3A)	10 - 50% of convT regions overlapped with RNA pol II stalling; overlap increased with increasing convT level quartile (study III Fig 3B)
RNA POL II STALLING (ES CELLS)	29%; $p < 0.001$ for the overlap belonging to the random distribution (12%) (study III Fig 3A)	5 - 35% of convT regions overlapped with RNA pol II stalling; overlap increased with increasing convT signal level quartile (study III Fig 3B)
DRIP-SEQ (ES CELLS)	2.1-fold elevation in DRIP-seq signal; $p < 2.2 \times 10^{-16}$ for the signal levels coming from the same population (study III Fig3C, source data 2)	1.7-fold elevation in DRIP-seq signal; $p < 2.2 \times 10^{-16}$ for the signal levels coming from the same population (study III Fig3D, source data 2)

Next we tested whether wide open chromatin regions were associated with RNA pol II stalling sites. For the interpretation of open chromatin sites, we retrieved DNase-seq data from B-lymphoid and embryonic stem (ES) cell lines and analyzed it together with the stalling sites (analyzed from GRO-seq from the respective cell types). Wider DNase-seq signal was observed at sites overlapping with RNA pol II stalling (876 bp in B-lymphoid, and 412 bp in ES cells; study III Fig 3 source data 2 online). We then compared the 5% widest RNA pol II stalling sites to the widest peaks from DNase-seq and ChIP-seq for histone markers. We found that the features with the highest odds for co-occurrence at RNA pol II stalling regions were pol II Ser5P, DNase hypersensitivity, H3K4me3, and H3K36me3 (odds ratios > 10; study III Figure 4C and source data 1 online).

5.5 Transcriptional features at genomic breakpoint regions (III)

Coordinates for structural variation (SV) breakpoints in 51 ETV6-RUNX1-positive patients were retrieved (Papaemmanuil *et al.*, 2014) and analyzed together with nascent RNA signal generated by using GRO-sequencing. For the quantification of SV regions, the genome was sorted into topologically associated domains (TADs) based on chromosome conformation data from a B-lymphoid cell line (Rao *et al.*, 2014). This resulted in 354 (of the total of 2900) TADs with at least one breakpoint among the E/R patients. 64% of the TADs contained more than one breakpoint

event among the patients (study III supplemental data 1 online). Initial observations for the specific transcription signal feature - RNA pol II stalling and convergent transcription - were gained by visually examining the most frequent breakpoint sites. Some of these sites did not contain previously annotated lncRNA although long transcripts were expressed based on the GRO-seq data. We also noticed that the breakpoints were often located a few kilobase from the TSS, with simultaneous transcription on both strands at these sites. We suspected that several of them might be intragenic enhancers. In addition, convergent transcription at intragenic breakpoint sites was associated with elevation in GRO-seq signal, that we hypothesized as being stalled RNA pol II (example of elevated signal in Figure 7A).

These observations led us to perform genome-wide analysis of convergent transcription and RNA pol II stalling at the breakpoint regions resolved from the ETV6-RUNX1-positive preB-ALL patients. Convergent transcription and pol II stalling were enriched at TADs with the most frequent breakpoints (41/73 with convT-level over background, $p = 0.00038$; 42/73 with RNA pol II stalling, $p = 0.014$) (Figure 7B).

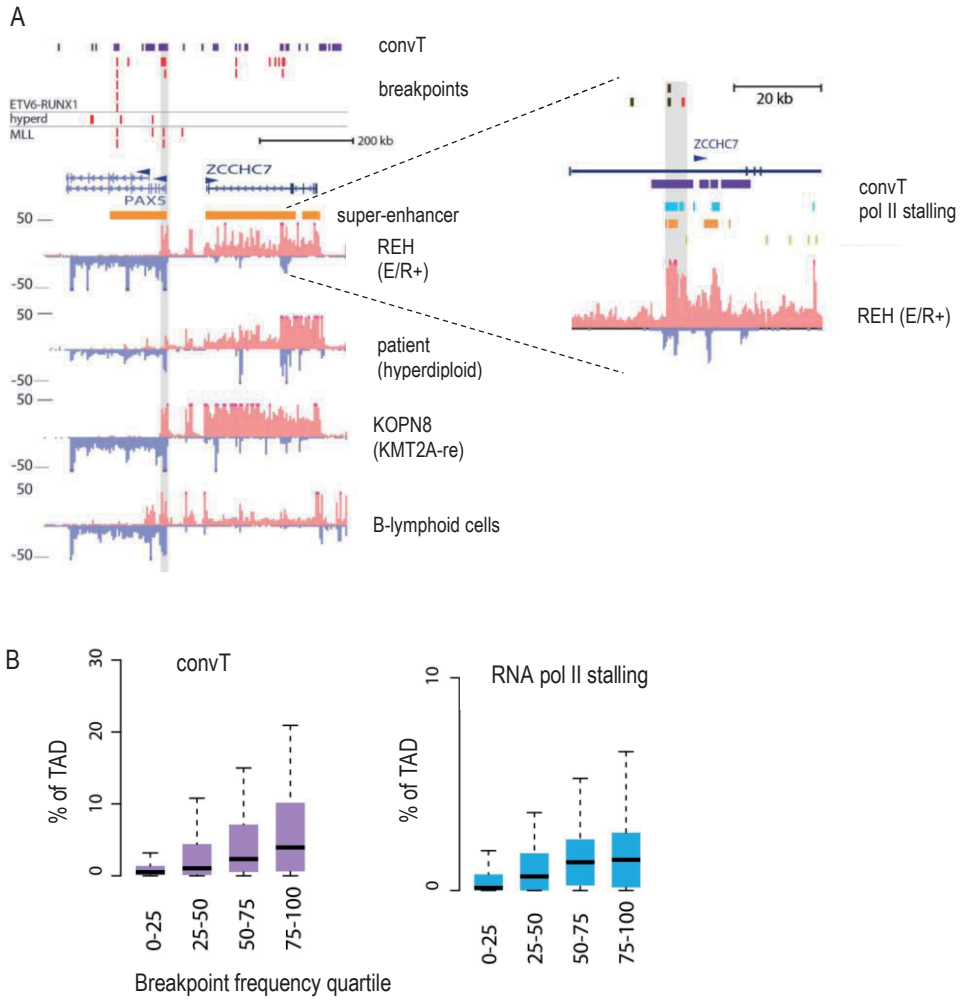

Figure 7. A) GRO-seq signal at an example region with recurrent breakpoints in ETV6-RUNX1 preB-ALL (locus with *PAX5* and *ZCCHC7* genes in chromosome 9). Zoomed view on *ZCCHC7* shows an example of local elevation in the signal with transcription on both strands. B) Topologically associated domains (TADs) with breakpoints were assigned into quartiles based on breakpoint frequency per TAD size (number of breakpoints per kilobase). Convergent transcription (left) and RNA pol II stalling (right) were enriched in TADs with frequent breakpoints. Adapted from Heinäniemi *et al.,* 2016 *eLife*.

We then took the breakpoints without recombination signal sequence under investigation to study a potential RAG-independent mechanism (416 non-RSS/NR-breakpoints, of which 124 intragenic). We found enrichment of breakpoints at regions with RNA pol II stalling and convergent transcription (Figure 8A; Table 3). Non-RSS-breakpoint recurrence, defined as a 1kb window with more than one breakpoint, increased the percentage of overlap with pol II stalling and convergent transcription (Figure 8A).

We then analyzed the breakpoints with RSS motif (335 R-breakpoints, of which 156 intragenic) for overlap with the transcriptional features. 56% of the breakpoints overlapped with pol II stalling at intragenic regions and 44% with genome-wide convergent transcription (Table 3). The overlap with RNA pol II stalling was more clear for the recurrent breakpoints (68% of the recurrent regions with stalling) (Figure 8B). Similarly, the overlap with convergent transcription was considerable in the recurrent breakpoint sites, up to 91% at regions with four or more breakpoints. The regions with RLFS motifs or annotated TSSs showed less marked co-occurrence (Figure 8B). The presence of both convT and RNA pol II stalling best characterized the recurrent intragenic breakpoints (43/48).

Table 3. The percentages of breakpoints that overlap with convergent transcription or RNA pol II stalling.

	BREAKPOINTS WITHOUT RSS	BREAKPOINTS WITH RSS
CONVT (B-LYMPHOID)	**9 %** (39 of 416) vs. 3.9% background; $p = 5.16 \times 10^{-7}$	**44%** (149 of 335) vs. 3.9% background; $p < 2.2 \times 10^{-16}$
RNA POL II STALLING (B-LYMPHOID) INTRAGENIC REGIONS	**29%** (36 of 124) vs. 12% background; $p = 4.09 \times 10^{-7}$	**56%** (87 of 156) vs. 12% background; $p < 2.2 \times 10^{-16}$

We also studied the expression levels of *RAG* and *AICDA* across a preB-ALL transcriptome dataset and noticed 10.8-fold higher median expression of *RAG1* in the ETV6-RUNX1 subtype (Figure 9A). Known classical subtypes clustered separately based on the global gene expression in an unsupervised analysis of sample similarities (study III Figure 5). *AICDA* expression was more prevalent in patients without any classic cytogenetics (study III Figure 5F and G).

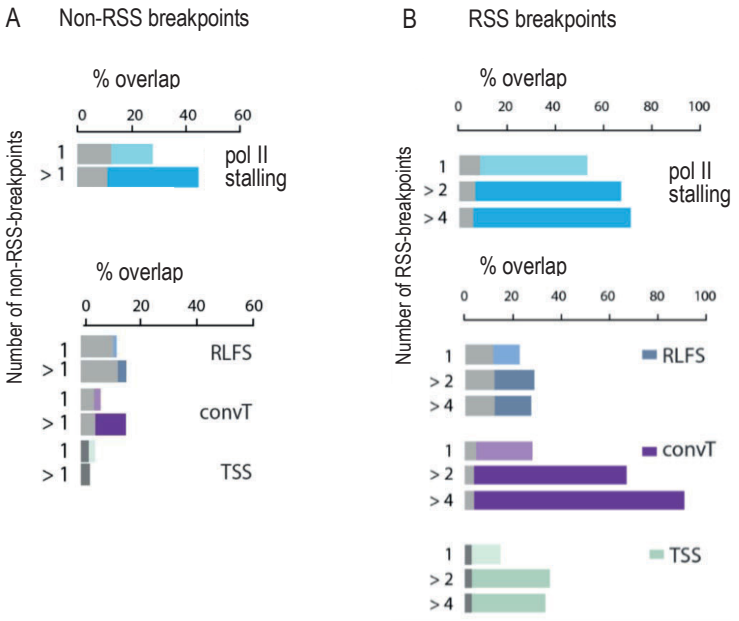

Figure 8. The percentages of breakpoints that overlap with RNA pol II stalling, convergent transcription (convT), R-loop forming sequences (RLFS), or transcription start sites (TSS) at regions with A) non-RSS-breakpoints, and B) RSS-breakpoints, resolved from ETV6-RUNX1 patients (Papaemmanuil *et al.*, 2014). RSS = recombination signal sequence. The overlap is shown separately for breakpoints binned by the recurrence in the dataset. Adapted from Heinäniemi *et al.,* 2016 *eLife*.

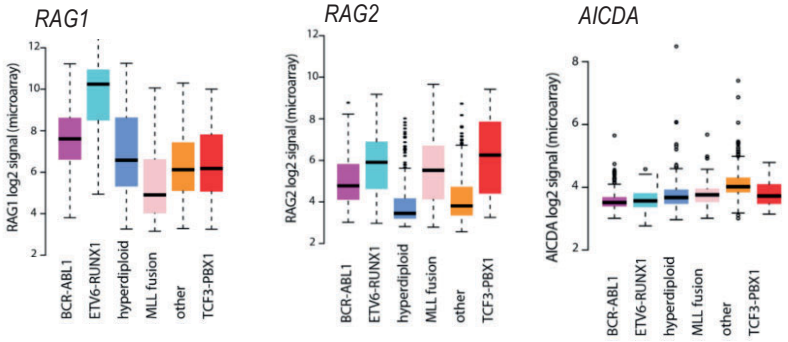

Figure 9. *RAG1*, *RAG2*, and *AICDA* expression in different preB-ALL subtypes based on the combined microarray studies with a total of 1382 patients. *MLL*-fusion = *KMT2A*-rearranged subtype. Adapted from Heinäniemi *et al.,* 2016 *eLife*.

6 DISCUSSION

6.1 ETV6-RUNX1 target genes

The second most common characterizing alteration in childhood leukemia, *ETV6-RUNX1* translocation, results in an aberrant fusion of two hematopoietic transcription factors. A few studies have therefore focused on its transcriptional consequences. However, many of the outcomes in terms of targeted genes have not been consistently reproducible across studies with different cohorts or cell models. There could be several potential reasons. First, when the putative target genes are deduced by comparing patient expression profiles, the results depend much on what the control group is comprised of. Heterogeneity of the subtypes may be one reason as to why reported gene sets discriminating between ALL and AML differed almost completely between early studies (Golub *et al.*, 1999; Moos *et al.*, 2002; van Delft *et al.*, 2005). As high hyperdiploid is the most common subtype in preB-ALL (besides E/R), they are likely overrepresented in many datasets, and thus some of the differential expression results may more describe high hyperdiploids than the E/R subgroup. In our study, we only tested a subset of genes for differential expression in an RNA-seq study and across microarray studies. In the RNA-seq cohort, the control group comprised of almost only hyperdiploid cases. However, the microarray data contained a better representation of subtypes. Cell differentiation status also differs between ALL subtypes (Andersson *et al.*, 2010; Chen *et al.*, 2016). Gene expression profiles of the E/R cases resemble most closely the normal proB cell state (Andersson *et al.*, 2010), while the *KMT2A*-rearranged cases are more immature (early lymphoid) and the TCF3-PBX1 subtype more mature preB-cells (Chen *et al.*, 2016). Thus, differential gene expression results may partly reflect the cell differentiation status and not the leukemia. Considering expression profiles during the differentiation of human B-cells (Mehtonen, Teppo, *et al.,* manuscript), many of the genes that were listed as being E/R-related in e.g. a report by Gandemer

et al., 2007 seem to define normal proB-cells. Cautious prediction on their possible role in leukemia is therefore necessary.

In addition to comparison of expression profiles between patients, cell models may generate different results depending on the genomic background of the chosen cell line. The clearest differences are between studies in human and mouse cells which are not advisable to be compared at gene level. Additional differences may stem from different overexpression or silencing level and the time of perturbation in the model. There are a few studies that have reported silencing of the E/R fusion in REH cells (a human E/R-positive preB-ALL cell line) (Fuka *et al.*, 2011; Zaliova *et al.*, 2011; Mangolini *et al.*, 2013; Ghazavi *et al.*, 2016). In Zaliova *et al.,* siRNAs were used to silence E/R down to 42% of the control level, yet no significant differentially expressed genes were found in the microarray study. This is in contrast to the report by Fuka *et al.,* where 50-80% silencing of E/R resulted in 777 genes deregulated in both REH and AT2 cell lines (with rather low correlation between the cell lines), and 52% of the genes were upregulated after knockdown. This study used an shRNA sequence that matched the wild type RUNX1 and was also shown to silence RUNX1 down to 46% (Zaliova *et al.*, 2011) which probably had some effect on the results. Ghazavi *et al.* reported approximately 50% silencing of E/R resulting in 134 lncRNAs deregulated, of which 30% were upregulated at knockdown. In our study, we managed to silence the E/R in REH down to 40% using lentivirally mediated shRNA that targeted evenly the *ETV6* and *RUNX1* sides of the fusion point. The silencing level may not have been enough for reliable estimation of changes in nascent RNA transcription genome-wide as only a few genes were found altered. However, some potential target genes were altered at mRNA level as measured by qPCR. Although a better knockdown level would have been favorable, complete knock-off of E/R may not be possible for experiments, as E/R-positive leukemic cells have been suggested to be dependent on the expression (Montano *et al.*, 2019 preprint).

Our cell model with inducible E/R fusion was based on a preB-ALL cell line Nalm6. By the experimental setup, we aimed to clarify the target genes behind the phenotypic differences as compared to other subtypes, and to reveal the genomic target sites of ETV6-RUNX1 in the genome. The use of inducible expression vector allowed a controlled study design without unwanted long-term effects. We also used a version of E/R with a mutated DNA-binding domain as a control to exclude effects not caused by RUNX1-mediated DNA binding. This approach led to exclusion of around 15% of putative target genes. ETV6-RUNX1 was present in the cells for a relatively short time (less than 24 hours) which probably decreased the

amount of indirect effects in transcription as compared to models with constitutively expressed transgenes. Nalm6 cell line resembles cells at preB differentiation state. Even though the cell line is relatively close to E/R-positive leukemic cells, the expression changes that we observed are affected by its cell state and unique genomic alterations. For example, we could not capture putative E/R-mediated downregulation at genes that are only expressed in earlier cell states. The selection of the model cell line is always a compromise and in an ideal world, one would have several different models that could help to better identify changes that are specific to background.

Direct genome-wide regulation by transcription factors can be interrogated by using chromatin immunoprecipitation (ChIP-seq). This method is widely used but is dependent on the antibody specificity and sensitivity to recognize the correct sites. It is common that only a few percent of TF binding motifs in a cell are actually bound by the TF (Spitz and Furlong, 2012). Moreover, binding of a TF at a certain site is not a direct measure of regulation of nearby genes. To address these challenges, we employed global nuclear run-on sequencing to decipher targets of E/R by measuring changes in nascent RNA transcription levels genome-wide. Nascent RNA expression at enhancer correlates with its activity and can thus be used to reveal activity of the regulatory region. The GRO-seq based method allowed more correct interpretation of the involvement of each putative enhancer site in the regulation. TF motif and ChIP-peak enrichment analyses suggested that E/R acts via RUNX1-dependent enhancer sites in the case of downregulated genes. This is in accordance with previous studies reporting mainly repression by the E/R which may be related to its protein-protein interactions with co-repressors (Fenrick *et al.*, 1999; Uchida *et al.*, 1999). In comparison, after silencing of E/R in another study, a nearby RUNX1 ChIP peak was found in similar levels for both up- and downregulated genes, and no enrichment of RUNX1 motif was seen at the gene promoters (Fuka *et al.*, 2011). This report used RUNX1 ChIP-seq data from human megakaryocytes and mouse early hematopoietic cells and may reflect differences in the cell models as well as illustrate the need to confirm findings with several data types.

In addition to the RUNX1-domain mediated downregulation of a part of the genes, we saw clear enrichment of ETS motifs at the deregulated sites after inducing the E/R. This suggests that these genes are not regulated through the RUNX1 RHD-domain but by ETV6 or other ETS-factors. Interestingly, the fusion protein does not contain the ETS-domain for DNA-binding by ETV6. It is known that ETS proteins (including ELF, ELK, ERG, and SPI) favor a similar DNA-binding motif, and ETV6 can heterodimerize with several of the members through the PNT

domain (Fenrick *et al.*, 1999). Therefore, although it is not possible to conclude with this data, other ETS proteins than ETV6 could be behind the upregulated genes in our experiment. This idea has been presented in a study where both E/R and ETV6 were needed to repress a promoter containing several TF binding sites, and ETV6 could be replaced with either ETS1 or FLI1 (Fears *et al.*, 1997). Expressing ETV6 in REH cells also led to both ETS- and RUNX1 motif enrichments at ETV6 binding sites, possibly revealing binding sites for E/R (Neveu *et al.*, 2018). The same study identified enrichment of ETS-IRF and IRF motifs at the binding sites. Similarly, RUNX1 and ETS factors have been shown to occupy a shared motif (Hollenhorst *et al.*, 2009). Interestingly, the DNA-binding deficient, RHD-domain mutated E/R version was not able to exclude these putatively non-RUNX1-mediated (ETS-factor dimerization) effects, although no defects in the putative PNT-domain mediated interactions were introduced.

The non-rearranged version of ETV6 is deleted in many but not all the cases of E/R leukemia. The deletion of ETV6 is also a common feature in E/R relapses. Thus far, ETV6 is reported to be always deleted in the ETV6-RUNX1-like subtype and has been suggested to be the main driver event (Lilljebjörn *et al.*, 2016; Zaliova *et al.*, 2017). The E/R-like cases resemble E/R leukemia in gene expression profile, which indicates that not only the E/R protein influence the profile. For example, a gene previously reported as E/R-specific (Ross *et al.*, 2003), *CLIC5*, was shown to be downregulated by ETV6 (Neveu *et al.*, 2016); therefore, its high expression in the E/R subtype seems to be due to loss or inactivation of the normal *ETV6* allele. Negative correlation between ETV6 deletion and duplication of ETV6-RUNX1 (duplication of der21(t12;21)) has been shown, suggesting that the two events are redundant. E/R seems to be more potent if it is expressed higher than ETV6, i.*e.* it is the ratio of these two that matters (McLean *et al.*, 1996; Lilljebjörn *et al.*, 2010). However, it has been suggested that even if the other ETV6 allele is not deleted, it is not expressed in the E/R disease (Patel *et al.*, 2003; Gunji *et al.*, 2004). Therefore, studies could benefit from looking at its RNA level expression in addition to the genetic alterations at the locus.

Soon after the E/R translocation was discovered, it was concluded that only one of the products, E/R, is consistently expressed in the t(12;21) cases, as the reciprocal RUNX1-ETV6 was detected in only 10/22 patients (McLean *et al.*, 1996). It was also reported that adding E/R, RUNX1-ETV6, or them together, was insufficient to induce leukemia (or factor independent growth) in mouse HSC cells (Andreasson *et al.*, 2001). However, subsequent studies have mentioned expression of the RUNX1-ETV6 in approximately 75% of the cases (Nakao *et al.*, 1996; Stams *et al.*, 2005) and

the expression level of *RUNX1-ETV6* has been reported as an independent prognostic factor in E/R leukemia (Stams *et al.*, 2005). As RUNX1-ETV6 contains the DNA binding domain ETS, and only the first exon of RUNX1, it could affect gene regulation as an aberrant ETS-factor. It was also suggested to share structural similarities for FUS-ERG fusion (in AML, Andreasson *et al.*, 2001), which was shown to bind DNA together with a complex containing other ETS factors, RUNX1 and TAL1, suggesting the fusion may affect regulation via these complexes (Sotoca *et al.*, 2016). Some E/R cases have additional RUNX1-fusions and simultaneously lack the RUNX1-ETV6, suggesting the translocations occur during the E/R translocation by a complex three-way manner (Lilljebjörn *et al.*, 2016). Off note, the most widely used E/R-positive cell line REH was reported not to express the reciprocal fusion and to have deletion in the other ETV6 allele (Uphoff *et al.*, 1997). This might be relevant in any future experiments where the overall play between the ETS-factors, RUNX1, and the fusions might be studied.

Genes that regulate B-cell differentiation are often targets of structural variations in ALL (Mullighan *et al.*, 2007). We noticed that E/R regulated 7% (56) of the super-enhancers resolved from $CD19^+$ or $CD34^+$ cells, with a preference to downregulate the expression. This exemplifies a way for an aberrant TF to interfere with gene regulation and contribute to differentiation impediment. TCF3-HLF, another fusion in preB-ALL leukemia, was shown to hijack the binding sites of HLF and rewire the enhancer landscape by targeting genes that represent features of proB- to preB transition (Huang *et al.*, 2019). Alterations in TFs and other genes that are important in B cell development may have direct consequences to treatment efficacy. This was described in glucocorticoids which were shown to suppress B-cell checkpoint genes to push cells through developmental blocks (Kruth *et al.*, 2017).

In conclusion, our analysis yielded a list of putative E/R responsive genomic sites, parts of which are seen to be similarly regulated in primary patient cells when compared to other subtypes. By considering the approach and the innate limitation and strengths in each reported study on the ETV6-RUNX1 function, it might be possible to capture the target genes and pathways most essential in the disease. The possible roles of them in the leukemogenesis or in the phenotype of E/R disease would need to be studied further in the future.

6.2 Enhancers in leukemia

During the studies, we generated nascent RNA profiles, which include all the newly transcribed RNAs in the genome, for several preB-ALL cell lines and primary patient samples. This data is especially suitable for studying transcriptional regulation and features. We noticed that regions with high number of breakpoints in preB-ALL contained previously unannotated non-coding RNA transcription. The genome-wide nascent RNA profiles may provide sites of interest for further examination. Large projects are also currently focusing on identifying pathogenic noncoding variants in many cancers (Gutierrez-Camino, Martin-Guerrero and García-Orad, 2017). That is increasingly possible with improved annotation and high-throughput methods. Genome-wide association studies have revealed a few risk loci for childhood leukemia, many of which lie in noncoding regions of the genome (Ellinghaus *et al.*, 2012; Vijayakrishnan *et al.*, 2018). For example, in leukemia, a susceptibility locus for childhood leukemia near *IKZF1* gene was shown to localize to a B-cell super-enhancer, while the enhancer was also shown to have differential activity and protein binding (Brown *et al.*, 2019). This exemplifies how a putative risk SNPs or somatic mutations in non-coding regions could be selected for further scrutinization with the help of improved enhancer annotation. In addition to mutations, individual enhancers have been reported to be altered in leukemia by translocations, amplifications, insertions, and deletions, reported especially on T-ALL, CLL, and AML (reviewed in Bhagwat, Lu and Vakoc, 2018).

Enhancers can affect gene expression both locally and from further distance by chromatin looping. When linking potential enhancers to genes after E/R induction, we only inferred local associations and did not consider the three-dimensional folding of chromatin. Defining interactions in three dimensions would be possible by using a chromosome conformation capture method such as HiC (Lieberman-Aiden *et al.*, 2009). By defining enhancers using nascent RNA transcription-based prediction, we were however able to capture more reliable pool of active regulatory regions than what would have been possible by analyzing TF/target gene co-expression or by ChIP-seq alone. Gene-enhancer distance metric have been used before to assign enhancer-gene pairs (Visel *et al.*, 2007; Fishilevich *et al.*, 2017) and combining it with eRNA/target co-expression (Andersson *et al.*, 2014) probably decreases false positive hits.

The enhancer landscape in human hematopoietic cells from blood has been mapped using RNA-seq and ATAC-seq profiles and was shown to act as a good discriminator between cell states (Corces *et al.*, 2016). Enhancer profiling based on

H3K27ac has been utilized in some leukemia studies (McKeown *et al.*, 2017; Wong *et al.*, 2017). We have noticed that different genetic preB-ALL subgroups can be clustered based on their enhancer RNA (GRO-seq) profiles (Heinäniemi *et al.,* unpublished). RNA-sequencing data has recently been utilized in identification of novel subtypes in ALL, including the so-called phenocopies (i.e. cases that have similar gene expression profile but lack a trademark genetic lesion such as a fusion gene). Epigenomic analyses seem to be able to similarly distinguish between subtypes (Nordlund and Syvänen, 2018). In fact, the so-called E/R-like subgroup was first noticed in a DNA methylation based classification (Nordlund *et al.*, 2015) before it was reported using gene expression data (Lilljebjörn *et al.*, 2016).

Studies reporting lncRNA profiles are emerging for different cell types and cancers. LncRNA profile based classifications of ALL subtypes have also been reported (Fernando *et al.*, 2015; James *et al.*, 2019). LncRNAs have been examined specifically in the E/R subtype using microarray data (Ghazavi *et al.*, 2016). Putative lncRNAs, from our study and others, could be further investigated from the GRO-seq data from patient cells. Further studies are needed to decipher any functional annotation to the putative E/R-related lncRNAs.

Enhancers could be utilized in treatment by targeting the co-factor proteins transcribing these regions. For example, BET inhibitors target a subgroup of BRD-protein bound enhancers, depending on other TFs and some still unknown factors. Also, kinase units in Mediator protein complex or general transcription factors could be inhibited, or the balance of histone deacetylases (HDAC) and acetyltransferases (HATs) could be altered for reprogramming. (Reviewed in Bhagwat, Lu and Vakoc, 2018). Targeting an enhancer or cofactor present specifically in cancer could have minimal amount of side effects, such as was proposed for TRIM33 cofactor that binds with PU.1 specifically at an upstream enhancer of the pro-apoptotic gene *BCL2L11* in preB-ALL (Wang *et al.*, 2015).

6.3 Transcriptional features at breakpoint regions

The use of next generation sequencing technologies has markedly increased the knowledge on the transcription of the genome. Transcriptional features, such as R-loops, have been linked to genomic instability (Sollier *et al.*, 2014; Hatchi *et al.*, 2015). On the other hand, secondary alterations in E/R leukemia have been suggested to predominantly arise from RAG enzyme mediated off-targeting in sites other than the immunoglobulin genes (Papaemmanuil *et al.*, 2014). *RAG1* and *RAG2* are

normally expressed during B-cell development for recombination of the immunoglobulin gene regions. They are also expressed in leukemic cells, especially in the E/R subtype.

We investigated the structural variation sites in ETV6-RUNX1-positive preB-ALL by analyzing transcriptional features (from GRO-seq data) at breakpoint regions (from WGS data). Transcription start sites (TSSs) in precursor B-cells have been considered susceptible to double strand breaks, mediated by H3K4me3 that is recognized by the RAG2 enzyme (Matthews *et al.*, 2007; Teng *et al.*, 2015). However, we noticed the breakpoints often resided a few kilobases from the TSS. Our analysis showed that the breakpoints overlap with RNA pol II stalling and convergent transcription, especially in case of recurrent breakpoint sites. We showed signals at the recurrent loci of, *PAX5*, *BTG1*, *CDKN2A/B,* and *RAG1/2* , all of which are often hit by deletions in preB-ALL (Mullighan *et al.*, 2007; Papaemmanuil *et al.*, 2014). Some of the regions may present novel enhancer RNAs or lncRNAs. We also showed genome-wide associations between elevated R-loop signal, convergent transcription, and RNA pol II stalling in normal and leukemic human cells. Although all the features are a part of the normal transcriptional process, genomic loci with recurrent breakpoints were particularly enriched with them. RNA pol II stalling was also associated with H3K4me3. Wide H3K4me3 regions have been suggested to ensure transcriptional consistency of key genes in cell identity and function (Benayoun *et al.*, 2014).

We classified the structural variation breakpoints into two groups based on the presence of an RSS motif. Both RSS- and non-RSS breakpoint sites were found to be enriched at stalled RNA pol II and convergent transcription sites. No clear difference in associations with transcriptional features were seen between the two breakpoint types, however, the breakpoints with RSS-motif had especially high concurrence of the features, indicating that there might be a higher demand for the exposure of the exact RSS motif for the RAG enzymes. This is supported by the notion that RAG-mediated cleavage occurs on unpaired and unwound DNA (Akamatsu and Oettinger, 1998). In contrast, non-RSS breakpoints could develop during transcription elongation and convergent transcription by mechanisms dependent on the exposure of the region in general.

R-loops are expected to be formed in regions where RNA pol II is stalling (Skourti-Stathaki and Proudfoot, 2014; Jenjaroenpun *et al.*, 2015). In addition to R-loops, convergent transcription has been hypothesized to cause pauses in transcription due to collisions between the crossing polymerases (Prescott and Proudfoot, 2002). Although we only saw a slight enrichment of breakpoints to R-

loop forming sequence sites (RLFS), we showed that they overlap with RNA pol II stalling events, and there was an increased signal level of DRIP-seq at RLFS-positive sites compared with RLFS-negative sites in ES cells. At the time, DRIP-seq data was not available for B-lineage cells but instead we used RLFS sequence prediction (Jenjaroenpun *et al.*, 2015). The modest enrichment of breakpoints to RLFS sites may be influenced by the wide proportion of genome, especially the TSS sites, that is prone to R-loops as predicted by the program (15646 TSSs with RLFS-motif, 8220 without) and reported by others (Ginno *et al.*, 2013; Lim *et al.*, 2015; Sanz *et al.*, 2016). Additional mechanisms seem to be needed to result in breakage. For example, factors affecting the stability of the DNA:RNA hybrids could play a role. One of these factors is topoisomerase that functions in rewinding the strands behind RNA pol II, inhibiting R-loop formation (Atkin, Raimer and Wang, 2019). Defects in RNA processing after transcription could slow down the rewinding of DNA at R-loop prone sites and thereby expose to breaks. Indeed, quite recently, mutations in genes affecting RNA splicing that results in RNA pol II accumulation at certain mis-spliced loci have been suggested in leukemogenesis in AML (Yoshimi *et al.*, 2019).

We used data from embryonic stem cells to show a general overlap of DNA:RNA-hybrids with RLFS motifs along with transcriptional features. We also showed overlap between stalling and convergent transcription using ES cell data with the idea that general properties of transcription are not dependent on cell type. When the overlap between the breakpoints in E/R leukemic cells and the transcriptional features in stem cells was analyzed, we did not see clear enrichments. This reflects cell type specific expression of the genome.

Our analysis comprised of data from B-lymphoid cell lines and preB-ALL leukemic cells (primary and cell lines). Nascent RNA signals were used to estimate the transcription occurring at the breakpoint sites before the breakage. Our study shows associations between transcriptional features, but the putative causes behind the associations and their regulation were not explored. In addition to ETV6-RUNX1, we also studied breakpoints resolved from other preB-ALL subtypes. Although only 7% of the breakpoints in *KMT2A*-rearranged cases had RSS motif evidence (Andersson *et al.*, 2015), we saw similar enrichment of convergent transcription and pol II stalling as in the case of ETV6-RUNX1. Similar transcriptional features were also seen for the breakpoints studied from the hyperdiploid and hypodiploid cases. This indicates that comparable transcriptional vulnerability may underlie the generation of the secondary alterations in these subtypes. The analysis could be extended in the future by using a model where GRO-seq signal was produced from preleukemic precursor B-cells and the structural

variation sites from the established leukemic cells were later resolved. Additional breakpoint data from more patients would also give more certainty for the recurrence of breaks inside specified regions.

Our study provides evidence on how genomic regions, especially with a nucleotide composition resembling RSS-motifs, are further selected for breakages recurrently seen at B-cell specific genomic sites. Recurrence is often used to define "driver genes" which are proposed to be particularly essential for the cancer progression. However, regarding the emerging evidence on transcriptional and epigenomic features, at least some of the recurrently altered genes could instead just be more prone to double strand breaks at that specific cell type and state and not all of them necessarily act as active players in leukemogenesis.

6.4 RAG and AID in secondary structural alterations

Current evidence pinpoints the role of RAGs especially in the E/R-positive preB-ALL. We noticed high expression of *RAG1* especially in the E/R subtype, in line with previous reports (Ross *et al.*, 2003). This could partly reflect the proB cell state of the E/R cells, although *RAG1* has also been suggested to be directly regulated by the E/R in a mouse model (Swaminathan *et al.*, 2015). E/R-positive preleukemic cells were also reported to have increased Rag1 and Rag2 levels compared to proB cells in wild type mice in another study (Rodríguez-Hernández *et al.*, 2017). Higher proportion of structural variations with RSS-motif have been reported in E/R-subtypes than in others (Zhang *et al.*, 2012; Papaemmanuil *et al.*, 2014; Andersson *et al.*, 2015), indicating that the RAG-mediated mechanism may be more prevalent in E/R-positive cells. Signs of RAG off-targeting have however been reported in individual genes also in other leukemia subtypes (Aplan *et al.*, 1990; Marculescu *et al.*, 2002; Raschke *et al.*, 2005; Mullighan *et al.*, 2008; Iacobucci *et al.*, 2009; Novara *et al.*, 2009; Waanders *et al.*, 2012).

Epidemiologic data show strong association between day-care attendance (used as a surrogate for exposure to common infections in early life) and reduced risk of acute lymphoblastic leukemia, while the risk of other childhood cancers is not affected (Urayama *et al.*, 2010). Already a hundred years ago, general infections were suspected as a cause of leukemia (reviewed in Greaves, 2018). Both RAG and AID enzymes were linked to secondary mutations in preB-cells with repeated inflammatory stimuli (Swaminathan *et al.*, 2015), which provided mechanistic evidence as to how overly strong inflammatory stimuli later in childhood may

increase the risk of secondary events. Abnormal cytokine production has also been shown to favor outgrowth of E/R-positive progenitor B cell clone (Ford *et al.*, 2009). In addition, relocating E/R-positive mice into a non-sterile room was sufficient to induce leukemic formation in an E/R mouse model (Rodríguez-Hernández *et al.*, 2017). Evidence of RAG-mediated secondary breakpoints were also seen in a human E/R iPS model (Böiers *et al.*, 2018).

AICDA (AID) is not normally expressed in bone marrow precursor cells (Gazumyan *et al.*, 2012) but its expression is induced in the presence of inflammatory agents in preB cells (Rosenberg and Papavasiliou, 2007; Swaminathan *et al.*, 2015). *AICDA* normally functions in mature germinal center B-cells during somatic hypermutation process introducing point mutations to immunoglobulin heavy chain gene, making antibodies more variable. Its function is dependent on active transcription at the site (Alt *et al.*, 2013). Intragenic enhancer RNAs have previously been linked with AID-mediated genomic instability in lymphomas (Meng *et al.*, 2014). After combining samples from independent gene expression microarray studies, we saw unusual expression of *AICDA* in a subgroup of preB-ALL patient. These patients mostly did not belong to any of the classic subgroups. *AICDA* expression has been suggested to be associated with a high-risk disease phenotype in a previous work (Swaminathan *et al.*, 2015). Given that AID is known to be targeted to stalled RNA pol II and convergent transcription sites (Yamane *et al.*, 2011; Meng *et al.*, 2014; Wang *et al.*, 2014), it may be possible that it is behind some of the oncogenic alterations among high-risk patients. PreB-ALL clones have also been shown to consistently carry somatically mutated *IGH* variable gene segments, which is indicative of AID activity (Bonaventure *et al.*, 2017). However, we did not specifically study if any of the genes deleted in the WGS dataset was a known AID off-target or whether there were sequence-specific clues of it. AID is similar to APOBEC cytidine deaminase proteins of which mutational signatures are found widely in cancers, also in leukemia (Rebhandl *et al.*, 2014; Wagener *et al.*, 2015; Li *et al.*, 2017). Therefore, this protein family may function in preB-ALL in general, and their relevance in leukemogenesis could be studied further in the future.

6.5 Do we still need more studies on ETV6-RUNX1 leukemia?

Several studies have been conducted on the ETV6-RUNX1-positive childhood leukemia and new evidence on its role as a differentiation staller has emerged. There are two main clinical aspects that would still benefit from additional knowledge on the disease: overtreatment and late relapses. E/R leukemia is suspected to be overtreated which causes unnecessary side-effects, but at the same time late relapses occur. Adjustment of the treatment regime may not be wise as long as excellent predictors for poor response or relapse are lacking. We are currently running a study where we compare E/R-positive patients with good early therapy response to those with suboptimal therapy response. We hope that this will give insights into whether and which of the genomic alterations contribute to therapy failure or increased risk of relapse. This type of data could help in optimizing therapy for the E/R patients in future: to decrease chemotherapy agents for the ones with good-prognostic genomic features and to increase therapy for the ones with poor-prognostic features.

The E/R fusion junctional region has been suggested to function as an antigen (Yotnda *et al.*, 1998; Chang *et al.*, 2017) and E/R-directed neoepitope-reactive T-cells were recently detected in ETV6-RUNX1-positive patients (Zamora *et al.*, 2019). If the E/R fusion protein itself proves to be targetable, we may not need additional knowledge on the molecular biology of the E/R leukemia to combat the E/R-positive clones. ETV6-RUNX1 has been considered a promising target for its high clonality and dependency of cells on its expression. E/R is clearly a first hit; thus, it is indeed likely to be found in every single leukemic cell and even in preleukemic cells. E/R-positive cancer cell lines such as REH are also suggested to be dependent on the fusion expression (Diakos *et al.*, 2007; Montano *et al.*, 2019), meaning, the cells would go through apoptosis if it is silenced.

Transcription factors, such as ETV6-RUNX1, could also be targeted by small molecule drugs. Most successful examples of TFs as drug targets function as receptors, such as retinoic acid receptor (RAR) in acute promyelocytic leukemia and glucocorticoid receptor (GR, *NR3C1*) in ALL. Drugs could also be designed to inhibit essential protein-protein-interactions in specific leukemic subtypes, like the interaction between the histone methyltransferase KMT2A (MLL) and its coactivators Menin and Ledgf (Uckelmann *et al.*, 2020), or between the fusion CBFB-SMMHC and RUNX1, without disturbing interaction between the normal CBFB and RUNX1 (Illendula *et al.*, 2015; Bushweller, 2019). In conclusion, studies focusing on targeting abnormal transcription factor fusions directly could have high impact on the next generation of leukemia treatment.

7 SUMMARY AND CONCLUSIONS

The ETV6-RUNX1-positive leukemia displays a characteristic gene expression profile, but the direct genome-wide transcriptional consequences have mostly remained elusive. Knowing these changes could help explain how the fusion predisposes carriers to leukemia and how it affects the leukemic cell phenotype. We aimed to uncover these issues by genome-wide analysis of nascent RNA transcription after induction of E/R in leukemic cells. We examined changes in the expression levels of both annotated and novel gene areas and linked putative enhancer regions to the differentially expressed genes. Our results suggest that the repressive function of ETV6-RUNX1 is partly mediated by RUNX1 binding sites at promoters and enhancers. We also showed repressive effect on the B-cell related enhancer sites of the genome. This may represent a way for E/R to slow down differentiation of precursor B cells that carry the fusion and thereby increase the probability to gain secondary structural variations. In addition, induction of E/R caused alterations in genes related to signaling with the microenvironment. Some of the direct effects could play a role in the features of E/R leukemia.

Many of the genetic alterations in ETV6/RUNX1-leukemia are suggested to be caused by illegitimate activity of the RAG and AID enzymes. Both enzymes function specifically during the lymphocyte development and are responsible for generating vast diversity of antibodies. RAG enzymes are abundantly expressed in the E/R subtype and a part of the breakpoint regions are marked with RAG recognition signal sequences. *AICDA* is not normally expressed in precursor B cells but is known to be induced by inflammatory stimuli and could therefore contribute to alterations in precursor leukemia. We found clear associations between certain transcriptional features, namely convergent transcription and RNA polymerase II stalling, and DNA breakpoint sites in precursor B-ALL. These sites were also associated with open chromatin and DNA:RNA hybrids (R-loops). Genes that are necessary for B-cell differentiation are often marked with abundant enhancer RNA transcription and are recurrently altered in leukemia. Our results support a hypothesis that transcriptional features leave the genome vulnerable to secondary genomic alterations in precursor B-cells. Further studies are needed to validate the findings and for the possible application in prevention and/or treatment of childhood ALL.

REFERENCES

Akamatsu, Y. and Oettinger, M. A. (1998) 'Distinct Roles of RAG1 and RAG2 in Binding the V(D)J Recombination Signal Sequences', *Molecular and Cellular Biology*. American Society for Microbiology, 18(8), pp. 4670–4678. doi: 10.1128/mcb.18.8.4670.

Al-Shehhi, H. *et al.* (2013) 'Abnormalities of the der(12)t(12;21) in *ETV6-RUNX1* acute lymphoblastic leukemia', *Genes, Chromosomes and Cancer*, 52(2), pp. 202–213. doi: 10.1002/gcc.22021.

Alexandrov, L. B. *et al.* (2013) 'Signatures of mutational processes in human cancer', *Nature*, 500(7463), pp. 415–421. doi: 10.1038/nature12477.

Alpar, D. *et al.* (2015) 'Clonal origins of ETV6-RUNX1+ acute lymphoblastic leukemia: studies in monozygotic twins', *Leukemia*. Nature Publishing Group, 29(4), pp. 839–846. doi: 10.1038/leu.2014.322.

Alt, F. W. *et al.* (2013) 'Mechanisms of programmed DNA lesions and genomic instability in the immune system.', *Cell*, 152(3), pp. 417–29. doi: 10.1016/j.cell.2013.01.007.

Anderson, K. *et al.* (2011) 'Genetic variegation of clonal architecture and propagating cells in leukaemia.', *Nature*. Nature Publishing Group, 469(7330), pp. 356–361. doi: 10.1038/nature09650.

Andersson, A. *et al.* (2005) 'Molecular signatures in childhood acute leukemia and their correlations to expression patterns in normal hematopoietic subpopulations.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(52), pp. 19069–74. doi: 10.1073/pnas.0506637102.

Andersson, A. *et al.* (2010) 'Gene expression signatures in childhood acute leukemias are largely unique and distinct from those of normal tissues and other malignancies.', *BMC medical genomics*. BioMed Central, 3, p. 6. doi: 10.1186/1755-8794-3-6.

Andersson, A. K. *et al.* (2015) 'The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias.', *Nature genetics*, 47(4), pp. 330–7. doi: 10.1038/ng.3230.

Andersson, R. *et al.* (2014) 'An atlas of active enhancers across human cell types and tissues', *Nature*. Nature Publishing Group, 507(7493), pp. 455–461. doi: 10.1038/nature12787.

Andersson, R., Sandelin, A. and Danko, C. G. (2015) 'A unified architecture of transcriptional regulatory elements', *Trends in Genetics*. Elsevier Ltd, pp. 426–433. doi: 10.1016/j.tig.2015.05.007.

Andreasson, P. *et al.* (2001) '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 genetics and cytogenetics*, 130(2), pp. 93–104. doi: 10.1016/s0165-4608(01)00518-0.

Aplan, P. D. *et al.* (1990) 'Disruption of the human SCL locus by "illegitimate"V-(D)-J recombinase activity.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 250(4986), pp. 1426–9. doi: 10.1126/science.2255914.

Arber, D. A. *et al.* (2016) 'The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia', *Blood*. American Society of Hematology, pp. 2391–2405. doi: 10.1182/blood-2016-03-643544.

Atkin, N., Raimer, H. and Wang, Y.-H. (2019) 'Broken by the Cut: A Journey into the Role of Topoisomerase II in DNA Fragility', *Genes*, 10(10), p. 791. doi: 10.3390/genes10100791.

Avellino, R. *et al.* (2016) 'An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation', *Blood*, 127(24), pp. 2991–3003. doi: 10.1182/blood-2016-01-695759.

Balbin, O. A. *et al.* (2015) 'The landscape of antisense gene expression in human cancers', *Genome Research*. Cold Spring Harbor Laboratory Press, 25(7), pp. 1068–1079. doi: 10.1101/gr.180596.114.

Bannister, A. J. *et al.* (2005) 'Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes', *Journal of Biological Chemistry*, 280(18), pp. 17732–17736. doi: 10.1074/jbc.M500796200.

Baruchel, A. *et al.* (1997) 'The majority of myeloid-antigen-positive (My+) childhood B-cell precursor acute lymphoblastic leukaemias express TEL-AML1 fusion transcripts.', *British journal of haematology*. Blackwell Publishing Ltd, 99(1), pp. 101–6. doi: 10.1046/j.1365-2141.1997.3603174.x.

Bateman, C. M. *et al.* (2010) 'Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia', *Blood*, 115(17), pp. 3553–3558. doi: 10.1182/blood-2009-10-251413.

Beck, D. *et al.* (2013) 'Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes', *Blood*. American Society of Hematology, 122(14), pp. e12–e22. doi: 10.1182/blood-2013-03-490425.

Benayoun, B. A. *et al.* (2014) 'H3K4me3 breadth is linked to cell identity and transcriptional consistency.', *Cell*, 158(3), pp. 673–88. doi: 10.1016/j.cell.2014.06.027.

Bernstein, B. E. *et al.* (2002) 'Methylation of histone H3 Lys 4 in coding regions of active genes', *Proceedings of the National Academy of Sciences of the United States of America*, 99(13), pp. 8695–8700. doi: 10.1073/pnas.082249499.

Bernstein, B. E. *et al.* (2010) 'The NIH roadmap epigenomics mapping consortium', *Nature Biotechnology*, pp. 1045–1048. doi: 10.1038/nbt1010-1045.

Bhagwat, A. S., Lu, B. and Vakoc, C. R. (2018) 'Enhancer dysfunction in leukemia', *Blood*. American Society of Hematology, pp. 1795–1804. doi: 10.1182/blood-2017-11-737379.

Biondi, A. *et al.* (2019) 'Long-term follow up of pediatric Philadelphia positive acute lymphoblastic leukemia treated with the EsPhALL2004 study: High white blood cell count at diagnosis is the strongest prognostic factor', *Haematologica*. Ferrata Storti Foundation, pp. e13–e16. doi: 10.3324/haematol.2018.199422.

Den Boer, M. L. *et al.* (2009) 'A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study', *The Lancet Oncology*, 10(2), pp. 125–134. doi: 10.1016/S1470-2045(08)70339-5.

Böiers, C. *et al.* (2013) 'Lymphomyeloid Contribution of an Immune-Restricted Progenitor Emerging Prior to Definitive Hematopoietic Stem Cells', *Cell Stem Cell*, 13(5), pp. 535–548. doi: 10.1016/j.stem.2013.08.012.

Böiers, C. *et al.* (2018) 'A Human IPS Model Implicates Embryonic B-Myeloid Fate Restriction as Developmental Susceptibility to B Acute Lymphoblastic Leukemia-

Associated ETV6-RUNX1', *Developmental Cell*, 44(3), pp. 362-377.e7. doi: 10.1016/j.devcel.2017.12.005.

Bokemeyer, A. *et al.* (2014) 'Copy number genome alterations are associated with treatment response and outcome in relapsed childhood ETV6/RUNX1-positive acute lymphoblastic leukemia.', *Haematologica*. Ferrata Storti Foundation, 99(4), pp. 706–14. doi: 10.3324/haematol.2012.072470.

Bonaventure, A. *et al.* (2017) 'Worldwide comparison of survival from childhood leukaemia for 1995–2009, by subtype, age, and sex (CONCORD-2): a population-based study of individual data for 89 828 children from 198 registries in 53 countries', *The Lancet Haematology*. Elsevier Ltd, 4(5), pp. e202–e217. doi: 10.1016/S2352-3026(17)30052-2.

Brown, A. L. *et al.* (2019) 'Inherited genetic susceptibility of acute lymphoblastic leukemia in Down syndrome', *Blood*, p. blood.2018890764. doi: 10.1182/blood.2018890764.

Bushweller, J. H. (2019) 'Targeting transcription factors in cancer — from undruggable to reality', *Nature Reviews Cancer*. Springer Science and Business Media LLC. doi: 10.1038/s41568-019-0196-7.

Cancer Statistics for the UK. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk (Accessed: 25 October 2019).

Castor, A. *et al.* (2005) 'Distinct patterns of hematopoietic stem cell involvement in acute lymphoblastic leukemia', *Nature Medicine*. Nature Publishing Group, 11(6), pp. 630–637. doi: 10.1038/nm1253.

Cavé, H. *et al.* (1997) 'ETV6 is the target of chromosome 12p deletions in t(12;21) childhood acute lymphocytic leukemia.', *Leukemia*, 11(9), pp. 1459–64. doi: 10.1038/sj.leu.2400798.

Chang, T.-C. *et al.* (2017) 'The neoepitope landscape in pediatric cancers', *Genome Medicine*, 9(1), p. 78. doi: 10.1186/s13073-017-0468-3.

Chatterton, Z. *et al.* (2014) 'Epigenetic deregulation in pediatric acute lymphoblastic leukemia', *Epigenetics*. Taylor and Francis Inc., 9(3), pp. 459–467. doi: 10.4161/epi.27585.

Chen, D. *et al.* (2016) 'The Expression Pattern of the Pre-B Cell Receptor Components Correlates with Cellular Stage and Clinical Outcome in Acute Lymphoblastic Leukemia.', *PloS one*. Public Library of Science, 11(9), p. e0162638. doi: 10.1371/journal.pone.0162638.

Chen, L. *et al.* (2017) 'R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters', *Molecular Cell*. Cell Press, 68(4), pp. 745-757.e5. doi: 10.1016/j.molcel.2017.10.008.

Coniat, M. B. Le *et al.* (2001) 'Chromosome 21 abnormalities with AML1 amplification in acute lymphoblastic leukemia', *Genes Chromosomes and Cancer*, 32(3), pp. 244–249. doi: 10.1002/gcc.1188.

Corces, M. R. *et al.* (2016) 'Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution', *Nature Genetics*. Nature Research, 48(10), pp. 1193–1203. doi: 10.1038/ng.3646.

Core, L. J. *et al.* (2014) 'Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers.', *Nature genetics*. NIH Public Access, 46(12), pp. 1311–20. doi: 10.1038/ng.3142.

Core, L. J., Waterfall, J. J. and Lis, J. T. (2008) 'Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.', *Science (New York, N.Y.)*, 322(5909), pp. 1845–8. doi: 10.1126/science.1162228.

van Delft, F. W. *et al.* (2005) 'Prospective gene expression analysis accurately subtypes acute leukaemia in children and establishes a commonality between hyperdiploidy and t(12;21) in acute lymphoblastic leukaemia.', *British journal of haematology*, 130(1), pp. 26–35. doi: 10.1111/j.1365-2141.2005.05545.x.

Diakos, C. *et al.* (2007) 'RNAi-mediated silencing of TEL/AML1 reveals a heat-shock protein– and survivin-dependent mechanism for survival', *Blood*, 109(6), pp. 2607–2610. doi: 10.1182/blood-2006-04-019612.

Dixon, J. R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485(7398), pp. 376–380. doi: 10.1038/nature11082.

Djebali, S. *et al.* (2012) 'Landscape of transcription in human cells.', *Nature*, 489(7414), pp. 101–8. doi: 10.1038/nature11233.

Drolet, M. *et al.* (1995) 'Overexpression of RNase H partially complements the growth defect of an Escherichia coli ΔtopA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 92(8), pp. 3526–3530. doi: 10.1073/pnas.92.8.3526.

Druker, B. J. *et al.* (2001) 'Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome', *New England Journal of Medicine*, 344(14), pp. 1038–1042. doi: 10.1056/NEJM200104053441402.

Egloff, S., Dienstbier, M. and Murphy, S. (2012) 'Updating the RNA polymerase CTD code: Adding gene-specific layers', *Trends in Genetics*, pp. 333–341. doi: 10.1016/j.tig.2012.03.007.

Eguchi-Ishimae, M. *et al.* (2001) 'Breakage and fusion of the TEL (ETV6) gene in immature B lymphocytes induced by apoptogenic signals.', *Blood*, 97(3), pp. 737–43. doi: 10.1182/blood.v97.3.737.

Ellinghaus, E. *et al.* (2012) 'Identification of germline susceptibility loci in ETV6-RUNX1-rearranged childhood acute lymphoblastic leukemia', *Leukemia*, 26(5), pp. 902–909. doi: 10.1038/leu.2011.302.

ENCODE Project Consortium, T. E. P. (2012) 'An integrated encyclopedia of DNA elements in the human genome.', *Nature*. NIH Public Access, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Engreitz, J. M. *et al.* (2016) 'Local regulation of gene expression by lncRNA promoters, transcription and splicing', *Nature*. Nature Research, 539(7629), pp. 452–455. doi: 10.1038/nature20149.

Enshaei, A. *et al.* (2013) 'Long-term follow-up of ETV6-RUNX1 ALL reveals that NCI risk, rather than secondary genetic abnormalities, is the key risk factor', *Leukemia*, pp. 2256–2259. doi: 10.1038/leu.2013.136.

Essig, S. *et al.* (2014) 'Risk of late effects of treatment in children newly diagnosed with standard-risk acute lymphoblastic leukaemia: a report from the Childhood Cancer Survivor Study cohort.', *The Lancet. Oncology*, 15(8), pp. 841–51. doi: 10.1016/S1470-2045(14)70265-7.

Fears, S. *et al.* (1997) 'Functional characterization of ETV6 and ETV6/CBFA2 in the regulation of the MCSFR proximal promoter.', *Proceedings of the National Academy of Sciences of the United States of America*, 94(5), pp. 1949–1954. doi: 10.1073/pnas.94.5.1949.

Fenrick, R. *et al.* (1999) 'Both TEL and AML-1 contribute repression domains to the t(12;21) fusion protein.', *Molecular and cellular biology*, 19(10), pp. 6566–6574.

Fernando, T. R. *et al.* (2015) 'LncRNA Expression Discriminates Karyotype and Predicts Survival in B-Lymphoblastic Leukemia', *Molecular Cancer Research*, 13(5), pp. 839–851. doi: 10.1158/1541-7786.MCR-15-0006-T.

Fine, B. M. *et al.* (2004) 'Gene expression patterns associated with recurrent chromosomal translocations in acute lymphoblastic leukemia', *Blood*, 103(3), pp. 1043–1049. doi: 10.1182/blood-2003-05-1518.

Fischer, M. *et al.* (2005) 'Defining the oncogenic function of the TEL/AML1 (ETV6/RUNX1) fusion protein in a mouse model.', *Oncogene*, 24(51), pp. 7579–7591. doi: 10.1038/sj.onc.1208931.

Fishilevich, S. *et al.* (2017) 'GeneHancer: genome-wide integration of enhancers and target genes in GeneCards', *Database*, 2017. doi: 10.1093/database/bax028.

Ford, A. M. *et al.* (1998) 'Fetal origins of the TEL-AML1 fusion gene in identical twins with leukemia', *Proceedings of the National Academy of Sciences*, 95(8), pp. 4584–4588. doi: 10.1073/pnas.95.8.4584.

Ford, A. M. *et al.* (2009) 'The TEL-AML1 leukemia fusion gene dysregulates the TGF-beta pathway in early B lineage progenitor cells.', *The Journal of clinical investigation*. American Society for Clinical Investigation, 119(4), pp. 826–36. doi: 10.1172/JCI36428.

Ford, A. M. and Greaves, M. (2017) 'ETV6-RUNX1 + Acute Lymphoblastic Leukaemia in Identical Twins', in *Advances in experimental medicine and biology*, pp. 217–228. doi: 10.1007/978-981-10-3233-2_14.

Ford, J. S. *et al.* (2019) 'Barriers and facilitators of risk-based health care for adult survivors of childhood cancer: A report from the Childhood Cancer Survivor Study.', *Cancer*. doi: 10.1002/cncr.32568.

Forestier, E. *et al.* (2008) 'Outcome of ETV6/RUNX1-positive childhood acute lymphoblastic leukaemia in the NOPHO-ALL-1992 protocol: frequent late relapses but good overall survival.', *British journal of haematology*, 140(6), pp. 665–72. doi: 10.1111/j.1365-2141.2008.06980.x.

Fraley, C., Raftery, A. E. and Murphy, T. B. (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.

Fueller, E. *et al.* (2014) 'Genomic inverse PCR for exploration of ligated breakpoints (GIPFEL), a new method to detect translocations in leukemia.', *PloS one*. Public Library of Science, 9(8), p. e104419. doi: 10.1371/journal.pone.0104419.

Fuka, G. *et al.* (2011) 'The leukemia-specific fusion gene ETV6/RUNX1 perturbs distinct key biological functions primarily by gene repression.', *PloS one*. Public Library of Science, 6(10), p. e26348. doi: 10.1371/journal.pone.0026348.

Fuka, G. *et al.* (2012) 'Silencing of ETV6/RUNX1 abrogates PI3K/AKT/mTOR signaling and impairs reconstitution of leukemia in xenografts', *Leukemia*, 26(5), pp. 927–933. doi: 10.1038/leu.2011.322.

Gale, K. B. *et al.* (1997) 'Backtracking leukemia to birth: Identification of clonotypic gene fusion sequences in neonatal blood spots', *Proceedings of the National Academy of Sciences of the United States of America*, 94(25), pp. 13950–13954. doi: 10.1073/pnas.94.25.13950.

Gandemer, V. *et al.* (2007) 'Five distinct biological processes and 14 differentially expressed genes characterize TEL/AML1-positive leukemia.', *BMC genomics*. BioMed Central, 8, p. 385. doi: 10.1186/1471-2164-8-385.

Gao, T. *et al.* (2016) 'EnhancerAtlas: A resource for enhancer annotation and analysis in 105 human cell/tissue types', *Bioinformatics*. Oxford University Press, 32(23), pp. 3543–3551. doi: 10.1093/bioinformatics/btw495.

Gao, T. and Qian, J. (2019) 'EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions', *PLOS Computational Biology*. Edited by I. Ioshikhes, 15(10), p. e1007436. doi: 10.1371/journal.pcbi.1007436.

Gazumyan, A. *et al.* (2012) 'Activation-Induced Cytidine Deaminase in Antibody Diversification and Chromosome Translocation', in *Advances in Cancer Research*. Academic Press Inc., pp. 167–190. doi: 10.1016/B978-0-12-394280-7.00005-1.

Ghazavi, F. *et al.* (2016) 'Unique long non-coding RNA expression signature in ETV6/RUNX1-driven B-cell precursor acute lymphoblastic leukemia', *Oncotarget*, 7(45), pp. 73769–73780. doi: 10.18632/oncotarget.12063.

Ginno, P. A. *et al.* (2012) 'R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters', *Molecular Cell*, 45(6), pp. 814–825. doi: 10.1016/j.molcel.2012.01.017.

Ginno, P. A. *et al.* (2013) 'GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination.', *Genome research*, 23(10), pp. 1590–600. doi: 10.1101/gr.158436.113.

Golub, T. R. *et al.* (1995) 'Fusion of the TEL gene on 12p13 to the AML1 gene on 21q22 in acute lymphoblastic leukemia.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(11), pp. 4917–4921. doi: 10.1073/pnas.92.11.4917.

Golub, T. R. *et al.* (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 286(5439), pp. 531–7. doi: 10.1126/science.286.5439.531.

Greaves, M. (2018) 'A causal mechanism for childhood acute lymphoblastic leukaemia', *Nature Reviews Cancer*. Nature Publishing Group, 18(8), pp. 471–484. doi: 10.1038/s41568-018-0015-6.

Greaves, M. F. *et al.* (2003) 'Leukemia in twins: lessons in natural history.', *Blood*. American Society of Hematology, 102(7), pp. 2321–33. doi: 10.1182/blood-2002-12-3817.

Gröbner, S. N. *et al.* (2018) 'The landscape of genomic alterations across childhood cancers', *Nature*. Nature Publishing Group, 555(7696), pp. 321–327. doi: 10.1038/nature25480.

Grossmann, V. *et al.* (2011) 'Prognostic relevance of RUNX1 mutations in T-cell acute lymphoblastic leukemia', *Haematologica*, 96(12), pp. 1874–1877. doi: 10.3324/haematol.2011.043919.

Gu, Z. *et al.* (2019) 'PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia', *Nature Genetics*. Nature Publishing Group, 51(2), pp. 296–307. doi: 10.1038/s41588-018-0315-5.

Guenther, M. G. *et al.* (2007) 'A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells', *Cell*, 130(1), pp. 77–88. doi: 10.1016/j.cell.2007.05.042.

Gunji, H. *et al.* (2004) 'TEL/AML1 shows dominant-negative effects over TEL as well as AML1', *Biochemical and Biophysical Research Communications*, 322(2), pp. 623–630. doi: 10.1016/j.bbrc.2004.07.169.

Gutierrez-Camino, A., Martin-Guerrero, I. and García-Orad, A. (2017) 'Genetic susceptibility in childhood acute lymphoblastic leukemia', *Medical Oncology*, 34(10), p. 179. doi: 10.1007/s12032-017-1038-7.

Hajingabo, L. J. *et al.* (2014) 'Predicting interactome network perturbations in human cancer: application to gene fusions in acute lymphoblastic leukemia.', *Molecular biology of the cell*. American Society for Cell Biology, 25(24), pp. 3973–85. doi: 10.1091/mbc.E14-06-1038.

Hangauer, M. J., Vaughn, I. W. and McManus, M. T. (2013) 'Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs.', *PLoS genetics*. Public Library of Science, 9(6), p. e1003569. doi: 10.1371/journal.pgen.1003569.

Harbott, J. *et al.* (1997) 'Incidence of TEL/AML1 fusion gene analyzed consecutively in children with acute lymphoblastic leukemia in relapse.', *Blood*, 90(12), pp. 4933–7.

Harrison, C. J. (2009) 'Cytogenetics of paediatric and adolescent acute lymphoblastic leukaemia', *British Journal of Haematology*, pp. 147–156. doi: 10.1111/j.1365-2141.2008.07417.x.

Harrison, C. J. (2013) 'Targeting signaling pathways in acute lymphoblastic leukemia: new insights', *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, pp. 118–125. doi: 10.1182/asheducation-2013.1.118.

Harrison, C. J. (2015) 'Blood Spotlight on iAMP21 acute lymphoblastic leukemia (ALL), a high-risk pediatric disease', *Blood*. American Society of Hematology, 125(9), pp. 1383–1386. doi: 10.1182/blood-2014-08-569228.

Harvey, R. C. *et al.* (2010) 'Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome', *Blood*, 116(23), pp. 4874–4884. doi: 10.1182/blood-2009-08-239681.

Hatchi, E. *et al.* (2015) 'BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair.', *Molecular cell*, 57(4), pp. 636–47. doi: 10.1016/j.molcel.2015.01.011.

Hein, D. *et al.* (2019) 'The preleukemic TCF3-PBX1 gene fusion can be generated in utero and is present in ≈0.6% of healthy newborns', *Blood*. American Society of Hematology, pp. 1355–1358. doi: 10.1182/blood.2019002215.

Heinäniemi, M. *et al.* (2016) 'Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots.', *eLife*, 5. doi: 10.7554/eLife.13087.

Hiebert, S. W. *et al.* (1996) 'The t(12;21) translocation converts AML-1B from an activator to a repressor of transcription.', *Molecular and cellular biology*, 16(4), pp. 1349–1355.

Hjalgrim, L. L. *et al.* (2002) 'Presence of clone-specific markers at birth in children with acute lymphoblastic leukaemia', *British Journal of Cancer*, 87(9), pp. 994–999. doi: 10.1038/sj.bjc.6600601.

Hnisz, D. *et al.* (2013) 'Super-Enhancers in the Control of Cell Identity and Disease', *Cell*, 155(4), pp. 934–947. doi: 10.1016/j.cell.2013.09.053.

Hock, H. *et al.* (2004) 'Tel/Etv6 is an essential and selective regulator of adult hematopoietic stem cell survival', *Genes and Development*, 18(19), pp. 2336–2341. doi: 10.1101/gad.1239604.

Hollenhorst, P. C. *et al.* (2009) 'DNA specificity determinants associate with distinct transcription factor functions', *PLoS Genetics*, 5(12). doi: 10.1371/journal.pgen.1000778.

Holmfeldt, L. *et al.* (2013) 'The genomic landscape of hypodiploid acute lymphoblastic leukemia.', *Nature genetics*, 45(3), pp. 242–52. doi: 10.1038/ng.2532.

Hon, C.-C. *et al.* (2017) 'An atlas of human long non-coding RNAs with accurate 5' ends', *Nature*. Nature Research, 543(7644), pp. 199–204. doi: 10.1038/nature21374.

Hong, D. *et al.* (2008) 'Initiating and Cancer-Propagating Cells in TEL-AML1-Associated Childhood Leukemia', *Science*, 319(5861), pp. 336–339. doi: 10.1126/science.1150648.

de Hoon, M., Shin, J. W. and Carninci, P. (2015) 'Paradigm shifts in genomics through the FANTOM projects', *Mammalian Genome*. Springer US, 26(9–10), pp. 391–402. doi: 10.1007/s00335-015-9593-8.

Hu, Y., Yoshida, T. and Georgopoulos, K. (2017) 'Transcriptional circuits in B cell transformation', *Current Opinion in Hematology*, 24(4), pp. 345–352. doi: 10.1097/MOH.0000000000000352.

Huang, D. W., Sherman, B. T. and Lempicki, R. a. (2009) 'Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research*, 37(1), pp. 1–13. doi: 10.1093/nar/gkn923.

Huang, Y. *et al.* (2019) 'The Leukemogenic TCF3-HLF Complex Rewires Enhancers Driving Cellular Identity and Self-Renewal Conferring EP300 Vulnerability', *Cancer Cell*. doi: 10.1016/j.ccell.2019.10.004.

Iacobucci, I. *et al.* (2009) 'Identification and molecular characterization of recurrent genomic deletions on 7p12 in the IKZF1 gene in a large cohort of BCR-ABL1-positive acute lymphoblastic leukemia patients: On behalf of Gruppo Italiano Malattie Ematologiche dell'Adulto Acute Leukemia Working Party (GIMEMA AL WP)', *Blood*, 114(10), pp. 2159–2167. doi: 10.1182/blood-2008-08-173963.

Iacobucci, I. and Mullighan, C. G. (2017) 'Genetic Basis of Acute Lymphoblastic Leukemia.', *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. American Society of Clinical Oncology, 35(9), pp. 975–983. doi: 10.1200/JCO.2016.70.7836.

Illendula, A. *et al.* (2015) 'Chemical biology. A small-molecule inhibitor of the aberrant transcription factor CBFβ-SMMHC delays leukemia in mice.', *Science (New York, N.Y.)*, 347(6223), pp. 779–84. doi: 10.1126/science.aaa0314.

Infante-Rivard, C., Fortier, I. and Olson, E. (2000) 'Markers of infection, breast-feeding and childhood acute lymphoblastic leukaemia', *British Journal of Cancer*. Churchill Livingstone, 83(11), pp. 1559–1564. doi: 10.1054/bjoc.2000.1495.

Inthal, A. *et al.* (2008) 'Role of the Erythropoietin Receptor in ETV6/RUNX1-Positive Acute Lymphoblastic Leukemia', *Clinical Cancer Research*, 14(22), pp. 7196–7204. doi: 10.1158/1078-0432.CCR-07-5051.

Jabbour, E. *et al.* (2015) 'New insights into the pathophysiology and therapy of adult acute lymphoblastic leukemia', *Cancer*. John Wiley and Sons Inc., pp. 2517–2528. doi: 10.1002/cncr.29383.

James, A. R. *et al.* (2019) 'Long non-coding RNAs defining major subtypes of B cell precursor acute lymphoblastic leukemia', *Journal of Hematology and Oncology*. BioMed Central Ltd., 12(1). doi: 10.1186/s13045-018-0692-3.

Jeha, S. *et al.* (2009) 'Increased risk for CNS relapse in pre-B cell leukemia with the t(1;19)/TCF3-PBX1', *Leukemia*, 23(8), pp. 1406–1409. doi: 10.1038/leu.2009.42.

Jenjaroenpun, P. *et al.* (2015) 'QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences.', *Nucleic acids research*, 43(W1), pp. W527-34. doi: 10.1093/nar/gkv344.

Jensen, T. H., Jacquier, A. and Libri, D. (2013) 'Dealing with pervasive transcription', *Molecular Cell*, pp. 473–484. doi: 10.1016/j.molcel.2013.10.032.

Joshi, I. *et al.* (2014) 'Loss of Ikaros DNA-binding function confers integrin-dependent survival on pre-B cells and progression to acute lymphoblastic leukemia.', *Nature immunology*. NIH Public Access, 15(3), pp. 294–304. doi: 10.1038/ni.2821.

Kaikkonen, M. U. *et al.* (2013) 'Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription.', *Molecular cell*, 51(3), pp. 310–25. doi: 10.1016/j.molcel.2013.07.010.

Kainz, M. and Roberts, J. (1992) 'Structure of transcription elongation complexes in vivo', *Science*, 255(5046), pp. 838–841. doi: 10.1126/science.1536008.

Kanno, T. *et al.* (1998) 'Intrinsic transcriptional activation-inhibition domains of the polyomavirus enhancer binding protein 2/core binding factor alpha subunit revealed in the presence of the beta subunit.', *Molecular and cellular biology*, 18(5), pp. 2444–54. doi: 10.1128/mcb.18.5.2444.

Katayama, S. *et al.* (2005) 'Antisense transcription in the mammalian transcriptome.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 309(5740), pp. 1564–6. doi: 10.1126/science.1112009.

Killick, R. and Eckley, I. A. (2014) 'Changepoint: An R package for changepoint analysis', *Journal of Statistical Software*. American Statistical Association, 58(3), pp. 1–19. doi: 10.18637/jss.v058.i03.

Kim, T. H. *et al.* (2005) 'A high-resolution map of active promoters in the human genome', *Nature*, 436(7052), pp. 876–880. doi: 10.1038/nature03877.

Kitabayashi, I. *et al.* (1998) 'Interaction and functional cooperation of the leukemia-associated factors AML1 and p300 in myeloid cell differentiation', *The EMBO Journal*, 17(11), pp. 2994–3004. doi: 10.1093/emboj/17.11.2994.

Kobayashi, H. and Rowley, J. D. (1995) 'Identification of cytogenetically undetected 12p13 translocations and associated deletions with fluorescence in situ hybridization.', *Genes, chromosomes & cancer*, 12(1), pp. 66–9.

Kouno, T. *et al.* (2019) 'C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution.', *Nature communications*, 10(1), p. 360. doi: 10.1038/s41467-018-08126-5.

Kruth, K. A. *et al.* (2017) 'Suppression of B-cell development genes is key to glucocorticoid efficacy in treatment of acute lymphoblastic leukemia.', *Blood*. American Society of Hematology, 129(22), pp. 3000–3008. doi: 10.1182/blood-2017-02-766204.

Kurzer, J. H. and Weinberg, O. K. (2018) 'Identification of early B cell precursors (stage 1 and 2 hematogones) in the peripheral blood', *Journal of Clinical Pathology*, 71(9), pp. 845–850. doi: 10.1136/jclinpath-2018-205172.

Kuster, L. *et al.* (2011) 'ETV6/RUNX1-positive relapses evolve from an ancestral clone and frequently acquire deletions of genes implicated in glucocorticoid signaling.', *Blood*, 117(9), pp. 2658–67. doi: 10.1182/blood-2010-03-275347.

Lai, F. *et al.* (2013) 'Activating RNAs associate with Mediator to enhance chromatin architecture and transcription', *Nature*, 494(7438), pp. 497–501. doi: 10.1038/nature11884.

Lans, H. *et al.* (2019) 'The DNA damage response to transcription stress', *Nature Reviews Molecular Cell Biology*. Springer Science and Business Media LLC. doi: 10.1038/s41580-019-0169-4.

Lausten-Thomsen, U. *et al.* (2011) 'Prevalence of t(12;21)[ETV6-RUNX1]-positive cells in healthy neonates.', *Blood*. American Society of Hematology, 117(1), pp. 186–9. doi: 10.1182/blood-2010-05-282764.

Leddin, M. *et al.* (2011) 'Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells', *Blood*, 117(10), pp. 2827–2838. doi: 10.1182/blood-2010-08-302976.

Levin, J. Z. *et al.* (2010) 'Comprehensive comparative analysis of strand-specific RNA sequencing methods', *Nature Methods*, 7(9), pp. 709–715. doi: 10.1038/nmeth.1491.

Li, J.-F. *et al.* (2018) 'Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 115(50), pp. E11711–E11720. doi: 10.1073/pnas.1814397115.

Li, Z. *et al.* (2003) 'Energetic contribution of residues in the Runx1 Runt domain to DNA binding.', *The Journal of biological chemistry*, 278(35), pp. 33088–96. doi: 10.1074/jbc.M303973200.

Li, Z. *et al.* (2017) 'APOBEC signature mutation generates an oncogenic enhancer that drives LMO1 expression in T-ALL', *Leukemia*. Nature Publishing Group, 31(10), pp. 2057–2064. doi: 10.1038/leu.2017.75.

Liao, Y., Smyth, G. K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.', *Bioinformatics (Oxford, England)*, 30(7), pp. 923–30. doi: 10.1093/bioinformatics/btt656.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science*, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.

Lilljebjörn, H. *et al.* (2010) 'The correlation pattern of acquired copy number changes in 164 ETV6/RUNX1-positive childhood acute lymphoblastic leukemias.', *Human molecular genetics*. Oxford University Press, 19(16), pp. 3150–8. doi: 10.1093/hmg/ddq224.

Lilljebjörn, H. *et al.* (2016) 'ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-ALL', *Nature Communications*, 7, p. 11790. doi: 10.1038/ncomms11790.

Lim, Y. W. *et al.* (2015) 'Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Goutières syndrome', *eLife*. eLife Sciences Publications Ltd, 4(JULY2015). doi: 10.7554/eLife.08007.

Lin, Y. C. *et al.* (2010) 'A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate.', *Nature immunology*. NIH Public Access, 11(7), pp. 635–43. doi: 10.1038/ni.1891.

Liu, Y. *et al.* (2018) 'Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites', *Nature Biotechnology*. Nature Publishing Group. doi: 10.1038/nbt.4283.

Livak, K. J. and Schmittgen, T. D. (2001) 'Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.', *Methods (San Diego, Calif.)*, 25(4), pp. 402–408. doi: 10.1006/meth.2001.1262.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.', *Genome biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Ma, X. *et al.* (2002) 'Daycare attendance and risk of childhood acute lymphoblastic leukaemia.', *British journal of cancer*, 86(9), pp. 1419–24. doi: 10.1038/sj.bjc.6600274.

Madanat-Harjuoja, L. M. *et al.* (2014) 'Childhood cancer survival in Finland (1953-2010): A nation-wide population-based study', *International Journal of Cancer*, 135(9), pp. 2129–2134. doi: 10.1002/ijc.28844.

Maia, A. T. *et al.* (2003) 'Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins.', *Leukemia*, 17(11), pp. 2202–6. doi: 10.1038/sj.leu.2403101.

Mangolini, M. *et al.* (2013) 'STAT3 mediates oncogenic addiction to TEL-AML1 in t(12;21) acute lymphoblastic leukemia.', *Blood*. American Society of Hematology, 122(4), pp. 542–9. doi: 10.1182/blood-2012-11-465252.

Mangum, D. S. *et al.* (2014) 'VPREB1 deletions occur independent of lambda light chain rearrangement in childhood acute lymphoblastic leukemia', *Leukemia*, pp. 216–220. doi: 10.1038/leu.2013.223.

Marculescu, R. *et al.* (2002) 'V(D)J-mediated Translocations in Lymphoid Neoplasms: A Functional Assessment of Genomic Instability by Cryptic Sites', *The Journal of Experimental Medicine*, 195(1), pp. 85–98. doi: 10.1084/jem.20011578.

Matthews, A. G. W. *et al.* (2007) 'RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination.', *Nature*, 450(7172), pp. 1106–10. doi: 10.1038/nature06431.

McKeown, M. R. *et al.* (2017) 'Superenhancer analysis defi nes novel epigenomic subtypes of non-APL AML, including an RARα dependency targetable by SY-1425, a potent and selective RARα agonist', *Cancer Discovery*. American Association for Cancer Research Inc., 7(10), pp. 1136–1153. doi: 10.1158/2159-8290.CD-17-0399.

McLean, C. Y. *et al.* (2010) 'GREAT improves functional interpretation of cis-regulatory regions.', *Nature biotechnology*. Nature Publishing Group, 28(5), pp. 495–501. doi: 10.1038/nbt.1630.

McLean, T. W. *et al.* (1996) 'TEL/AML-1 dimerizes and is associated with a favorable outcome in childhood acute lymphoblastic leukemia.', *Blood*, 88(11), pp. 4252–8.

McManus, S. *et al.* (2011) 'The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells', *EMBO Journal*, 30(12), pp. 2388–2404. doi: 10.1038/emboj.2011.140.

Mehtonen, J. *et al.* (2019) 'Data-driven characterization of molecular phenotypes across heterogeneous sample collections', *Nucleic Acids Research*. Oxford University Press (OUP), 47(13), pp. e76–e76. doi: 10.1093/nar/gkz281.

Meng, F.-L. *et al.* (2014) 'Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability.', *Cell*, 159(7), pp. 1538–48. doi: 10.1016/j.cell.2014.11.014.

Mevel, R. *et al.* (2019) 'RUNX transcription factors: orchestrators of development', *Development*, 146(17), p. dev148296. doi: 10.1242/dev.148296.

Montano, A. *et al.* (2019) 'ETV6/RUNX1 Fusion Gene Abrogation Decreases The Oncogenic Potencial Of Tumour Cells In A Preclinical Model Of Acute Lymphoblastic Leukaemia.', *bioRxiv*. Cold Spring Harbor Laboratory, p. 809525. doi: 10.1101/809525.

Montaño, A. *et al.* (2018) 'New challenges in targeting signaling pathways in acute lymphoblastic leukemia by NGS approaches: An update', *Cancers*. MDPI AG. doi: 10.3390/cancers10040110.

Moos, P. J. *et al.* (2002) 'Identification of Gene Expression Profiles That Segregate Patients with Childhood Leukemia', *Clinical Cancer Research*. American Association for Cancer Research, 8(10), pp. 3118–3130.

Morak, M. *et al.* (2013) 'Clone-specific secondary aberrations are not detected in neonatal blood spots of children with ETV6-RUNX1-positive leukemia.', *Haematologica*. Ferrata Storti Foundation, 98(9), pp. e108-10. doi: 10.3324/haematol.2013.090860.

Mori, H. *et al.* (2002) 'Chromosome translocations and covert leukemic clones are generated during normal fetal development.', *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), pp. 8242–8247. doi: 10.1073/pnas.112218799.

Morrow, M. *et al.* (2004) 'TEL-AML1 promotes development of specific hematopoietic lineages consistent with preleukemic activity', *Blood*, 103(10), pp. 3890–3896. doi: 10.1182/blood-2003-10-3695.

Morrow, M. *et al.* (2007) 'TEL-AML1 preleukemic activity requires the DNA binding domain of AML1 and the dimerization and corepressor binding domains of TEL.', *Oncogene*, 26(30), pp. 4404–4414. doi: 10.1038/sj.onc.1210227.

Mullighan, C. G. *et al.* (2007) 'Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia', *Nature*, 446(7137), pp. 758–764. doi: 10.1038/nature05690.

Mullighan, Charles G. *et al.* (2008) 'BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros', *Nature*. Nature Publishing Group, 453(7191), pp. 110–114. doi: 10.1038/nature06866.

Mullighan, C. G. *et al.* (2008) 'Genomic Analysis of the Clonal Origins of Relapsed Acute Lymphoblastic Leukemia', *Science*, 322(5906), pp. 1377–1380. doi: 10.1126/science.1164266.

Mullighan, C. G. (2012) 'The molecular genetic makeup of acute lymphoblastic leukemia.', *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, 2012(1), pp. 389–96. doi: 10.1182/asheducation-2012.1.389.

Mulrooney, D. A. *et al.* (2019) 'The changing burden of long-term health outcomes in survivors of childhood acute lymphoblastic leukaemia: a retrospective analysis of the St Jude Lifetime Cohort Study.', *The Lancet. Haematology*, 6(6), pp. e306–e316. doi: 10.1016/S2352-3026(19)30050-X.

Muse, G. W. *et al.* (2007) 'RNA polymerase is poised for activation across the genome', *Nature Genetics*, 39(12), pp. 1507–1511. doi: 10.1038/ng.2007.21.

Nachman, J. B. *et al.* (2007) 'Outcome of treatment in children with hypodiploid acute lymphoblastic leukemia', *Blood*, 110(4), pp. 1112–1115. doi: 10.1182/blood-2006-07-038299.

Nakao, M. *et al.* (1996) 'Detection and quantification of TEL/AML1 fusion transcripts by polymerase chain reaction in childhood acute lymphoblastic leukemia.', *Leukemia*, 10(9), pp. 1463–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8751464.

Nechaev, S. *et al.* (2010) 'Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila', *Science*, 327(5963), pp. 335–338. doi: 10.1126/science.1181421.

Nechaev, S. and Adelman, K. (2011) 'Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, pp. 34–45. doi: 10.1016/j.bbagrm.2010.11.001.

Neveu, B. *et al.* (2016) 'CLIC5: a novel ETV6 target gene in childhood acute lymphoblastic leukemia.', *Haematologica*. Ferrata Storti Foundation, 101(12), pp. 1534–1543. doi: 10.3324/haematol.2016.149740.

Neveu, B. *et al.* (2018) 'Genome wide mapping of ETV6 binding sites in pre-B leukemic cells.', *Scientific reports*. Nature Publishing Group, 8(1), p. 15526. doi: 10.1038/s41598-018-33947-1.

Nordlund, J. *et al.* (2015) 'DNA methylation-based subtype prediction for pediatric acute lymphoblastic leukemia', *Clinical Epigenetics*, 7(1), p. 11. doi: 10.1186/s13148-014-0039-z.

Nordlund, J. and Syvänen, A. C. (2018) 'Epigenetics in pediatric acute lymphoblastic leukemia', *Seminars in Cancer Biology*. Academic Press, pp. 129–138. doi: 10.1016/j.semcancer.2017.09.001.

Novara, F. *et al.* (2009) 'Different molecular mechanisms causing 9p21 deletions in acute lymphoblastic leukemia of childhood', *Human Genetics*. Springer, 126(4), pp. 511–520. doi: 10.1007/s00439-009-0689-7.

O'Byrne, S. *et al.* (2019) 'Discovery of a CD10-negative B-progenitor in human fetal life identifies unique ontogeny-related developmental programs.', *Blood*. American Society of Hematology, 134(13), pp. 1059–1071. doi: 10.1182/blood.2019001289.

O'Connor, D. *et al.* (2018) 'Genotype-Specific minimal residual disease interpretation improves stratification in pediatric acute lymphoblastic leukemia', *Journal of Clinical Oncology*. American Society of Clinical Oncology, 36(1), pp. 34–43. doi: 10.1200/JCO.2017.74.0449.

Osterwalder, M. *et al.* (2018) 'Enhancer redundancy provides phenotypic robustness in mammalian development', *Nature*. Nature Publishing Group, 554(7691), pp. 239–243. doi: 10.1038/nature25461.

Papaemmanuil, E. *et al.* (2014) 'RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia.', *Nature genetics*. Nature Publishing Group, 46(2), pp. 116–25. doi: 10.1038/ng.2874.

Parvin, J. D. and Sharp, P. A. (1993) 'DNA topology and a minimal set of basal factors for transcription by RNA polymerase II', *Cell*, 73(3), pp. 533–540. doi: 10.1016/0092-8674(93)90140-L.

Patel, N. *et al.* (2003) 'Expression profile of *wild-type ETV6* in childhood acute leukaemia', *British Journal of Haematology*. John Wiley & Sons, Ltd (10.1111), 122(1), pp. 94–98. doi: 10.1046/j.1365-2141.2003.04399.x.

Paulsson, K. *et al.* (2015) 'The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia.', *Nature genetics*, 47(6), pp. 672–6. doi: 10.1038/ng.3301.

Peter, A. *et al.* (2009) 'Interphase FISH on TEL/AML1 positive acute lymphoblastic leukemia relapses - Analysis of clinical relevance of additional TEL and AML1 copy number changes', *European Journal of Haematology*, 83(5), pp. 420–432. doi: 10.1111/j.1600-0609.2009.01315.x.

Piette, C. *et al.* (2018) 'Differential impact of drugs on the outcome of ETV6-RUNX1 positive childhood B-cell precursor acute lymphoblastic leukaemia: results of the EORTC CLG 58881 and 58951 trials', *Leukemia*. Nature Publishing Group, 32(1), pp. 244–248. doi: 10.1038/leu.2017.289.

Polak, R. *et al.* (2019) 'Autophagy inhibition as a potential future targeted therapy for ETV6-RUNX1-driven B-cell precursor acute lymphoblastic leukemia', *Haematologica*. Ferrata Storti Foundation, 104(4), pp. 738–748. doi: 10.3324/haematol.2018.193631.

Pölönen, P. *et al.* (2019) 'HEMap: An interactive online resource for characterizing molecular phenotypes across hematologic malignancies', *Cancer Research*. American Association for Cancer Research Inc., 79(10), pp. 2466–2479. doi: 10.1158/0008-5472.CAN-18-2970.

Popescu, D.-M. *et al.* (2019) 'Decoding human fetal liver haematopoiesis', *Nature*. doi: 10.1038/s41586-019-1652-y.

Preker, P. *et al.* (2008) 'RNA exosome depletion reveals transcription upstream of active human promoters', *Science*, 322(5909), pp. 1851–1854. doi: 10.1126/science.1164096.

Prescott, E. M. and Proudfoot, N. J. (2002) 'Transcriptional collision between convergent genes in budding yeast.', *Proceedings of the National Academy of Sciences of the United States of America*, 99(13), pp. 8796–801. doi: 10.1073/pnas.132270899.

Ptasinska, A. *et al.* (2012) 'Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding.', *Leukemia*, 26(8), pp. 1829–41. doi: 10.1038/leu.2012.49.

Pui, C.-H. and Campana, D. (2017) 'Minimal residual disease in pediatric ALL', *Oncotarget*. Impact Journals, LLC, 8(45), p. 78251. doi: 10.18632/ONCOTARGET.20856.

Ramilowski, J. A. *et al.* (2019) 'Functional Annotation of Human Long Non-Coding RNAs via Molecular Phenotyping', *bioRxiv*, p. 700864. doi: 10.1101/700864.

Rao, S. S. P. *et al.* (2014) 'A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping', *Cell*. Elsevier, 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.

Raschke, S. *et al.* (2005) 'Homozygous deletions of CDKN2A caused by alternative mechanisms in various human cancer cell lines.', *Genes, chromosomes & cancer*, 42(1), pp. 58–67. doi: 10.1002/gcc.20119.

Rasighaemi, P. and Ward, A. C. (2017) 'ETV6 and ETV7: Siblings in hematopoiesis and its disruption in disease', *Critical Reviews in Oncology/Hematology*. Elsevier, 116, pp. 106–115. doi: 10.1016/J.CRITREVONC.2017.05.011.

Rebhandl, S. *et al.* (2014) 'APOBEC3 signature mutations in chronic lymphocytic leukemia', *Leukemia*. Nature Publishing Group, pp. 1929–1932. doi: 10.1038/leu.2014.160.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2009) 'edgeR: A Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*. Oxford University Press, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.

Rodríguez-Hernández, G. *et al.* (2017) 'Infection Exposure Promotes ETV6-RUNX1 Precursor B-cell Leukemia via Impaired H3K4 Demethylases.', *Cancer research*. American Association for Cancer Research, 77(16), pp. 4365–4377. doi: 10.1158/0008-5472.CAN-17-0701.

Romana, S. P., Le Coniat, M. and Berger, R. (1994) 't(12;21): a new recurrent translocation in acute lymphoblastic leukemia.', *Genes, chromosomes & cancer*, 9(3), pp. 186–91. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7515661.

Rosenberg, B. R. and Papavasiliou, F. N. (2007) 'Beyond SHM and CSR: AID and Related Cytidine Deaminases in the Host Response to Viral Infection', in *Advances in immunology*, pp. 215–244. doi: 10.1016/S0065-2776(06)94007-3.

Ross, M. E. *et al.* (2003) 'Classification of pediatric acute lymphoblastic leukemia by gene expression profiling', *Blood*, 102(8), pp. 2951–2959. doi: 10.1182/blood-2003-01-0338.

Russell, L. J. *et al.* (2009) 'Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia', *Blood*, 114(13), pp. 2688–2698. doi: 10.1182/blood-2009-03-208397.

Saida, S. (2017) 'Predispositions to Leukemia in Down Syndrome and Other Hereditary Disorders', *Current Treatment Options in Oncology*. Springer New York LLC. doi: 10.1007/s11864-017-0485-x.

Sanz, L. A. *et al.* (2016) 'Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals', *Molecular Cell*. Cell Press, 63(1), pp. 167–178. doi: 10.1016/j.molcel.2016.05.032.

Sanz, L. A. and Chédin, F. (2019) 'High-resolution, strand-specific R-loop mapping via S9.6-based DNA–RNA immunoprecipitation and high-throughput sequencing', *Nature*

Protocols. Nature Publishing Group, 14(6), pp. 1734–1755. doi: 10.1038/s41596-019-0159-1.

Schäfer, D. *et al.* (2018) 'Five percent of healthy newborns have an ETV6-RUNX1 fusion as revealed by DNA-based GIPFEL screening.', *Blood*. American Society of Hematology, 131(7), pp. 821–826. doi: 10.1182/blood-2017-09-808402.

Schatz, D. G. and Swanson, P. C. (2011) 'V(D)J recombination: mechanisms of initiation.', *Annual review of genetics*, 45, pp. 167–202. doi: 10.1146/annurev-genet-110410-132552.

Sharrocks, A. D. (2001) 'The ETS-domain transcription factor family', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 2(11), pp. 827–837. doi: 10.1038/35099076.

Siggers, T. *et al.* (2011) 'Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex', *Molecular Systems Biology*, 7. doi: 10.1038/msb.2011.89.

Sigova, A. A. *et al.* (2013) 'Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 110(8), pp. 2876–81. doi: 10.1073/pnas.1221904110.

Skourti-Stathaki, K. and Proudfoot, N. J. (2014) 'A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression.', *Genes & development*, 28(13), pp. 1384–96. doi: 10.1101/gad.242990.114.

Sollier, J. *et al.* (2014) 'Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability.', *Molecular cell*, 56(6), pp. 777–85. doi: 10.1016/j.molcel.2014.10.020.

Sotoca, A. M. *et al.* (2016) 'The oncofusion protein FUS-ERG targets key hematopoietic regulators and modulates the all-trans retinoic acid signaling pathway in t(16;21) acute myeloid leukemia.', *Oncogene*. Nature Publishing Group, 35(15), pp. 1965–76. doi: 10.1038/onc.2015.261.

Soulier, J. *et al.* (2003) 'Amplification of band q22 of chromosome 21, including AML1, in older children with acute lymphoblastic leukemia: an emerging molecular cytogenetic subgroup.', *Leukemia*, 17(8), pp. 1679–82. doi: 10.1038/sj.leu.2403000.

Spitz, F. and Furlong, E. E. M. (2012) 'Transcription factors: From enhancer binding to developmental control', *Nature Reviews Genetics*, pp. 613–626. doi: 10.1038/nrg3207.

Stams, W. A. G. *et al.* (2005) 'Expression Levels of TEL, AML1, and the Fusion Products TEL-AML1 and AML1-TEL versus Drug Sensitivity and Clinical Outcome in t(12;21)-Positive Pediatric Acute Lymphoblastic Leukemia', *Clinical Cancer Research*. American Association for Cancer Research, 11(8), pp. 2974–2980. doi: 10.1158/1078-0432.CCR-04-1829.

Stams, W. A. G. *et al.* (2006) 'Incidence of additional genetic changes in the TEL and AML1 genes in DCOG and COALL-treated t(12;21)-positive pediatric ALL, and their relation with drug sensitivity and clinical outcome.', *Leukemia*, 20(3), pp. 410–6. doi: 10.1038/sj.leu.2404083.

Starkova, J. *et al.* (2007) 'The Identification of (ETV6)/RUNX1-Regulated Genes in Lymphopoiesis Using Histone Deacetylase Inhibitors in ETV6/RUNX1-Positive Lymphoid Leukemic Cells', *Clinical Cancer Research*, 13(6), pp. 1726–1735. doi: 10.1158/1078-0432.CCR-06-2569.

Steliarova-Foucher, Eva *et al.* (2017) 'International incidence of childhood cancer, 2001–10: a population-based registry study', *The Lancet Oncology*. Lancet Publishing Group, 18(6), pp. 719–731. doi: 10.1016/S1470-2045(17)30186-9.

Sullivan, R. *et al.* (2013) 'New policies to address the global burden of childhood cancers', *The Lancet Oncology*. doi: 10.1016/S1470-2045(13)70007-X.

Sun, C., Chang, L. and Zhu, X. (2017) 'Pathogenesis of ETV6/RUNX1-positive childhood acute lymphoblastic leukemia and mechanisms underlying its relapse.', *Oncotarget*. Impact Journals, LLC, 8(21), pp. 35445–35459. doi: 10.18632/oncotarget.16367.

Swaminathan, S. *et al.* (2015) 'Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia.', *Nature immunology*. Europe PMC Funders, 16(7), pp. 766–774. doi: 10.1038/ni.3160.

Syöpä Suomessa - Syöpärekisteri. Available at: https://syoparekisteri.fi/tilastot/syopa-suomessa.

Tan-Wong, S. M., Dhir, S. and Proudfoot, N. J. (2019) 'R-Loops Promote Antisense Transcription across the Mammalian Genome.', *Molecular cell*. doi: 10.1016/j.molcel.2019.10.002.

Taub, J. W. *et al.* (2002) 'High frequency of leukemic clones in newborn screening blood samples of children with B-precursor acute lymphoblastic leukemia', *Blood*, 99(8), pp. 2992–2996. doi: 10.1182/blood.V99.8.2992.

Teng, G. *et al.* (2015) 'RAG Represents a Widespread Threat to the Lymphocyte Genome.', *Cell*, 162(4), pp. 751–65. doi: 10.1016/j.cell.2015.07.009.

Teppo, S. *et al.* (2016) 'Genome-wide repression of eRNA and target gene loci by the ETV6-RUNX1 fusion in acute leukemia', *Genome Research*, 26(11), pp. 1468–1477. doi: 10.1101/gr.193649.115.

Teppo, S., Heinäniemi, M. and Lohi, O. (2017) 'Deregulation of the non-coding genome in leukemia', *RNA Biology*. Taylor & Francis, pp. 1–4. doi: 10.1080/15476286.2017.1312228.

Terwilliger, T. and Abdul-Hay, M. (2017) 'Acute lymphoblastic leukemia: a comprehensive review and 2017 update', *Blood cancer journal*, 7(6), p. e577. doi: 10.1038/bcj.2017.53.

Teuffel, O. *et al.* (2004) 'Prenatal origin of separate evolution of leukemia in identical twins', *Leukemia*. Nature Publishing Group, 18(10), pp. 1624–1629. doi: 10.1038/sj.leu.2403462.

Thandla, S. P. *et al.* (1999) 'ETV6-AML1 translocation breakpoints cluster near a purine/pyrimidine repeat region in the ETV6 gene.', *Blood*, 93(1), pp. 293–9..

Thurman, R. E. *et al.* (2012) 'The accessible chromatin landscape of the human genome', *Nature*, 489(7414), pp. 75–82. doi: 10.1038/nature11232.

Toft, N. *et al.* (2018) 'Results of NOPHO ALL2008 treatment for patients aged 1-45 years with acute lymphoblastic leukemia', *Leukemia*. Nature Publishing Group, 32(3), pp. 606–615. doi: 10.1038/leu.2017.265.

Torrano, V. *et al.* (2011) 'ETV6-RUNX1 promotes survival of early B lineage progenitor cells via a dysregulated erythropoietin receptor', *Blood*, 118(18), pp. 4910–4918. doi: 10.1182/blood-2011-05-354266.

Trinklein, N. D. *et al.* (2004) 'An abundance of bidirectional promoters in the human genome', *Genome Research*, 14(1), pp. 62–66. doi: 10.1101/gr.1982804.

Uchida, H. *et al.* (1999) 'Three distinct domains in TEL-AML1 are required for transcriptional repression of the IL-3 promoter.', *Oncogene*, 18(4), pp. 1015–22. doi: 10.1038/sj.onc.1202383.

Uckelmann, H. J. *et al.* (2020) 'Therapeutic targeting of preleukemia cells in a mouse model of NPM1 mutant acute myeloid leukemia', *Science (New York, N.Y.)*. NLM (Medline), 367(6477), pp. 586–590. doi: 10.1126/science.aax5863.

Uphoff, C. *et al.* (1997) 'Occurrence of TEL-AML1 fusion resulting from (12;21) translocation in human early B-lineage leukemia cell lines', *Leukemia*, 11(3), pp. 441–447. doi: 10.1038/sj.leu.2400571.

Urayama, K. Y. *et al.* (2010) 'A meta-analysis of the association between day-care attendance and childhood acute lymphoblastic leukaemia', *International Journal of Epidemiology*, 39(3), pp. 718–732. doi: 10.1093/ije/dyp378.

Vijayakrishnan, J. *et al.* (2018) 'Genome-wide association study identifies susceptibility loci for B-cell childhood acute lymphoblastic leukemia', *Nature Communications*, 9(1), p. 1340. doi: 10.1038/s41467-018-03178-z.

Visel, A. *et al.* (2007) 'VISTA Enhancer Browser - A database of tissue-specific human enhancers', *Nucleic Acids Research*, 35(SUPPL. 1). doi: 10.1093/nar/gkl822.

Waanders, E. *et al.* (2012) 'The origin and nature of tightly clustered BTG1 deletions in precursor B-cell acute lymphoblastic leukemia support a model of multiclonal evolution', *PLoS Genetics*. Public Library of Science, 8(2). doi: 10.1371/journal.pgen.1002533.

Wagener, R. *et al.* (2015) 'Analysis of mutational signatures in exomes from B-cell lymphoma cell lines suggest APOBEC3 family members to be involved in the pathogenesis of primary effusion lymphoma', *Leukemia*. Nature Publishing Group, pp. 1612–1615. doi: 10.1038/leu.2015.22.

Wang, E. *et al.* (2015) 'The transcriptional cofactor TRIM33 prevents apoptosis in B lymphoblastic leukemia by deactivating a single enhancer', *eLife*, 4, p. e06377. doi: 10.7554/eLife.06377.

Wang, I. X. *et al.* (2014) 'RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II.', *Cell reports*, 6(5), pp. 906–15. doi: 10.1016/j.celrep.2014.01.037.

Wang, K. C. and Chang, H. Y. (2011) 'Molecular Mechanisms of Long Noncoding RNAs', *Molecular Cell*, 43(6), pp. 904–914. doi: 10.1016/j.molcel.2011.08.018.

Wang, L. C. *et al.* (1998) 'The TEL/ETV6 gene is required specifically for hematopoiesis in the bone marrow', *Genes & Development*. Cold Spring Harbor Laboratory Press, 12(15), pp. 2392–2402. doi: 10.1101/gad.12.15.2392.

Wang, L. and Hiebert, S. W. (2001) 'TEL contacts multiple co-repressors and specifically associates with histone deacetylase-3', *Oncogene*, 20(28), pp. 3716–3725. doi: 10.1038/sj.onc.1204479.

Wang, X. *et al.* (2014) 'A source of the single-stranded DNA substrate for activation-induced deaminase during somatic hypermutation.', *Nature communications*, 5, p. 4137. doi: 10.1038/ncomms5137.

Ward, D. F. and Murray, N. E. (1979) 'Convergent transcription in bacteriophage λ: Interference with gene expression', *Journal of Molecular Biology*, 133(2), pp. 249–266. doi: 10.1016/0022-2836(79)90533-3.

Whyte, W. A. *et al.* (2013) 'Master transcription factors and mediator establish super-enhancers at key cell identity genes.', *Cell*, 153(2), pp. 307–19. doi: 10.1016/j.cell.2013.03.035.

Wiemels, J. L., Cazzaniga, G., *et al.* (1999) 'Prenatal origin of acute lymphoblastic leukaemia in children.', *Lancet (London, England)*, 354(9189), pp. 1499–503. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10551495.

Wiemels, J. L., Ford, A. M., *et al.* (1999) 'Protracted and variable latency of acute lymphoblastic leukemia after TEL-AML1 gene fusion in utero.', *Blood*, 94(3), pp. 1057–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10419898.

Wiemels, J. L. *et al.* (2000) 'Microclustering of TEL-AML1 translocation breakpoints in childhood acute lymphoblastic leukemia.', *Genes, chromosomes & cancer*, 29(3), pp. 219–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10992297.

Wiemels, J. L. and Greaves, M. (1999) 'Structure and possible mechanisms of TEL-AML1 gene fusions in childhood acute lymphoblastic leukemia.', *Cancer research*, 59(16), pp. 4075–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10463610.

Wilkinson, A. C. *et al.* (2013) 'RUNX1 Is a Key Target in t(4;11) Leukemias that Contributes to Gene Activation through an AF4-MLL Complex Interaction', *Cell Reports*, 3(1), pp. 116–127. doi: 10.1016/j.celrep.2012.12.016.

Wong, R. W. J. *et al.* (2017) 'Enhancer profiling identifies critical cancer genes and characterizes cell identity in adult T-cell leukemia', *Blood*. American Society of Hematology, 130(21), pp. 2326–2338. doi: 10.1182/blood-2017-06-792184.

Yamane, A. *et al.* (2011) 'Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes', *Nature Immunology*, 12(1), pp. 62–69. doi: 10.1038/ni.1964.

Yoshimi, A. *et al.* (2019) 'Coordinated alterations in RNA splicing and epigenetic regulation drive leukaemogenesis', *Nature*. Springer Science and Business Media LLC. doi: 10.1038/s41586-019-1618-0.

Yotnda, P. *et al.* (1998) 'Cytotoxic T cell response against the chimeric ETV6-AML1 protein in childhood acute lymphoblastic leukemia.', *Journal of Clinical Investigation*, 102(2), pp. 455–462. doi: 10.1172/JCI3126.

Zaliova, M *et al.* (2011) 'Revealing the role of TEL/AML1 for leukemic cell survival by RNAi-mediated silencing', *Leukemia*. Nature Publishing Group, 25(2), pp. 313–320. doi: 10.1038/leu.2010.277.

Zaliova, Marketa *et al.* (2011) 'TEL/AML1-positive patients lacking TEL exon 5 resemble canonical TEL/AML1 cases.', *Pediatric blood & cancer*, 56(2), pp. 217–25. doi: 10.1002/pbc.22686.

Zaliova, M. *et al.* (2017) '*ETV6/RUNX1* -like acute lymphoblastic leukemia: A novel B-cell precursor leukemia subtype associated with the CD27/CD44 immunophenotype', *Genes, Chromosomes and Cancer*, 56(8), pp. 608–616. doi: 10.1002/gcc.22464.

Zaliova, M. *et al.* (2019) 'Genomic landscape of pediatric B-other acute lymphoblastic leukemia in a consecutive European cohort.', *Haematologica*. Ferrata Storti Foundation, 104(7), pp. 1396–1406. doi: 10.3324/haematol.2018.204974.

Zamora, A. E. *et al.* (2019) Pediatric patients with acute lymphoblastic leukemia generate abundant and functional neoantigen-specific CD8 + T cell responses, *Sci. Transl. Med.* Available at: http://stm.sciencemag.org/.

Zhang, J. *et al.* (2012) 'The genetic basis of early T-cell precursor acute lymphoblastic leukaemia', *Nature*, 481(7380), pp. 157–163. doi: 10.1038/nature10725.

Zhang, M. and Swanson, P. C. (2008) 'V(D)J Recombinase Binding and Cleavage of Cryptic Recombination Signal Sequences Identified from Lymphoid Malignancies', *Journal of Biological Chemistry*, 283(11), pp. 6717–6727. doi: 10.1074/jbc.M710301200.

Zhang, Z. *et al.* (2006) 'Transcription factor Pax5 (BSAP) transactivates the RAG-mediated VH-to-DJH rearrangement of immunoglobulin genes', *Nature Immunology*, 7(6), pp. 616–624. doi: 10.1038/ni1339.

Zuna, J. *et al.* (2011) 'ETV6/RUNX1 (TEL/AML1) is a frequent prenatal first hit in childhood leukemia', *Blood*, 117(1), pp. 368–369. doi: 10.1182/blood-2010-09-309070.

PUBLICATIONS

PUBLICATION I

Genome-wide repression of eRNA and target gene loci by the ETV6-RUNX1 fusion in acute leukemia

Teppo, S., Laukkanen, S., Liuksiala, T., Nordlund, J., Oittinen, M., Teittinen, K., Grönroos, T., St-Onge, P., Syvänen, AC., Nykter, M., Viiri, K., Heinäniemi M., & Lohi, O.

Genome Research 2016, 26(11): 1468–1477
doi: 10.1101/gr.193649.115

Research

Genome-wide repression of eRNA and target gene loci by the ETV6-RUNX1 fusion in acute leukemia

Susanna Teppo,[1] Saara Laukkanen,[1] Thomas Liuksiala,[1,2] Jessica Nordlund,[3] Mikko Oittinen,[1] Kaisa Teittinen,[1] Toni Grönroos,[1] Pascal St-Onge,[4] Daniel Sinnett,[4,5] Ann-Christine Syvänen,[3] Matti Nykter,[2,6] Keijo Viiri,[1] Merja Heinäniemi,[7,8] and Olli Lohi[1,8]

[1]*Tampere Center for Child Health Research, University of Tampere and Tampere University Hospital, 33520 Tampere, Finland;* [2]*Institute of Biosciences and Medical Technology, University of Tampere, 33520 Tampere, Finland;* [3]*Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, 75105, Uppsala, Sweden;* [4]*CHU Sainte-Justine Research Center, Université de Montréal, Montréal, Quebec, H3T 1J4, Canada;* [5]*Department of Pediatrics, Faculty of Medicine, Université de Montréal, Montréal, Quebec, H3T 1J4, Canada;* [6]*Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland;* [7]*Institute of Biomedicine, School of Medicine, University of Eastern Finland, 70211 Kuopio, Finland*

Approximately 20%–25% of childhood acute lymphoblastic leukemias carry the *ETV6–RUNX1* (*E/R*) fusion gene, a fusion of two central hematopoietic transcription factors, ETV6 (TEL) and RUNX1 (AML1). Despite its prevalence, the exact genomic targets of E/R have remained elusive. We evaluated gene loci and enhancers targeted by E/R genome-wide in precursor B acute leukemia cells using global run-on sequencing (GRO-seq). We show that expression of the *E/R* fusion leads to widespread repression of RUNX1 motif–containing enhancers at its target gene loci. Moreover, multiple super-enhancers from the CD19$^+$/CD20$^+$-lineage were repressed, implicating a role in impediment of lineage commitment. In effect, the expression of several genes involved in B cell signaling and adhesion was down-regulated, and the repression depended on the wild-type DNA-binding Runt domain of RUNX1. We also identified a number of E/R-regulated annotated and de novo noncoding genes. The results provide a comprehensive genome-wide mapping between E/R-regulated key regulatory elements and genes in precursor B cell leukemia that disrupt normal B lymphopoiesis.

[Supplemental material is available for this article.]

Childhood acute lymphoblastic leukemia (ALL) is a heterogeneous disease consisting of distinct clinical subtypes characterized by specific chromosomal translocations or mutations. The most common ALL subtypes are hyperdiploid and *ETV6-RUNX1* (*E/R*) fusion gene positive pre-B-ALL, which both comprise ~20%–25% of the cases (Inaba et al. 2013). Both leukemias have been suggested to arise from a progenitor cell in utero and typically advance into overt disease after accumulating additional mutations during early childhood (Wiemels et al. 1999; Bateman et al. 2015). The *E/R* fusion can be detected also among healthy newborns with prevalence estimates ranging from 0.01% to 1% (Mori et al. 2002; Greaves et al. 2011; Lausten-Thomsen et al. 2011; Zuna et al. 2011).

After 20 years since the discovery of the *E/R* fusion (Romana et al. 1994; Golub et al. 1995), the mechanism(s) by which it contributes to the development of B-ALL remains not fully understood. E/R is a fusion of two essential hematopoietic transcription factors (TF): ETV6 (TEL) (Wang et al. 1998) and RUNX1 (AML1) (Wang et al. 1996). The translocation between Chromosomes 12 and 21, t(12;21), fuses the N terminus of *ETV6* to nearly full-length *RUNX1*, thus retaining the DNA-binding domain of RUNX1 (Runt). The E/R fusion is suggested to function as an aberrant TF: The transactivating function of RUNX1 is lost, and the fusion protein is converted into a repressor through the ETV6 moiety that recruits corepressors (SIN3A or NCOR) and epigenetic modifiers (HDACs) (Hiebert et al. 1996; Fears et al. 1997; Fenrick et al. 1999; Song et al. 1999; Uchida et al. 1999; Guidez et al. 2000; Hiebert et al. 2001; Morrow et al. 2007). During the past decade, a number of microarray studies were performed exploring the gene expression profiles of *E/R*-positive patient samples (Yeoh et al. 2002; Ross and Zhou 2003; Fine et al. 2004; Andersson et al. 2005, 2007; Gandemer et al. 2007). However, these studies generated only a limited number of common genomic targets of E/R, and the functions of E/R have remained elusive.

Global run-on assay followed by next generation sequencing (GRO-seq) is a recently described method that allows for genome-wide detection of primary transcript levels by directly measuring nascent RNA production, including transcription at regulatory elements (Core et al. 2008). At these regions, RNA polymerase II (RNAP II) generates so-called enhancer RNAs (eRNAs) that show dynamic activation patterns reflecting regional regulatory activity (Core et al. 2008; Kaikkonen et al. 2013). Enhancers contain binding sites for multiple sequence-specific TFs that can be revealed through profiling dynamical eRNA expression and de novo motif discovery. In the context of lineage determination, a unique subset

[8]**Co-senior authors.**
Corresponding author: susanna.teppo@uta.fi
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.193649.115.

26:1468–1477 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/16; www.genome.org

of enhancers termed super-enhancers (SE) has been distinguished. They consist of enhancers with clustering of binding sites for multiple TFs and high histone K27 acetylation levels, reminiscent of locus control regions characterized at beta-globin or immune gene loci (Li et al. 2002; Whyte et al. 2013; Pott and Lieb 2014; Adam et al. 2015). Recently, aberrant enhancer activity at the *TAL1* gene locus was shown to drive leukemogenesis in T-ALL (Mansour et al. 2014). However, the majority of cancer-relevant regulatory elements remain uncharacterized.

Here we applied GRO-seq on an inducible cell model to elucidate the genomic targets of the E/R fusion in precursor B-ALL.

Results

ETV6-RUNX1 is predominantly a repressive fusion TF

To study the molecular functions and targets of the E/R fusion in ALL, we generated a doxycycline-inducible human *E/R* expressing cell line using an *E/R*-negative precursor B-ALL cell line Nalm-6 (Nalm6-E/R) (Fig. 1A; Supplemental Fig. S1A–E). Induction with doxycycline for 24 h increased the *E/R* mRNA level by 18-fold compared to REH cells (in which *E/R* is expressed endogenously) (Supplemental Fig. S1A), and at protein level, E/R was detectable as early as 4–8 h after induction (Fig. 1B). Since the E/R fusion retains the DNA-binding domain of the RUNX1 protein (Runt), we generated another cell line with a targeted mutation (Nalm6-E/Rmut): Substitution of an arginine with glutamine (R201Q) in human RUNX1 reduces DNA-binding affinity of the Runt domain by 1000-fold (Li et al. 2003; Morrow et al. 2007). Moreover, a reciprocal cell line was created in which endogenous *E/R* is silenced by ~60% (REH-sh*E/R*) (Supplemental Fig. S1F).

We applied the GRO-seq assay on our inducible cell model to investigate early E/R-mediated transcriptional events. After 0, 4, 12, and 24 h of *E/R* induction, nuclei were extracted, and GRO-seq was performed with a high level of concordance ($r^2 = 0.96$) between two independent experiments (Fig. 1C). Mature transcripts

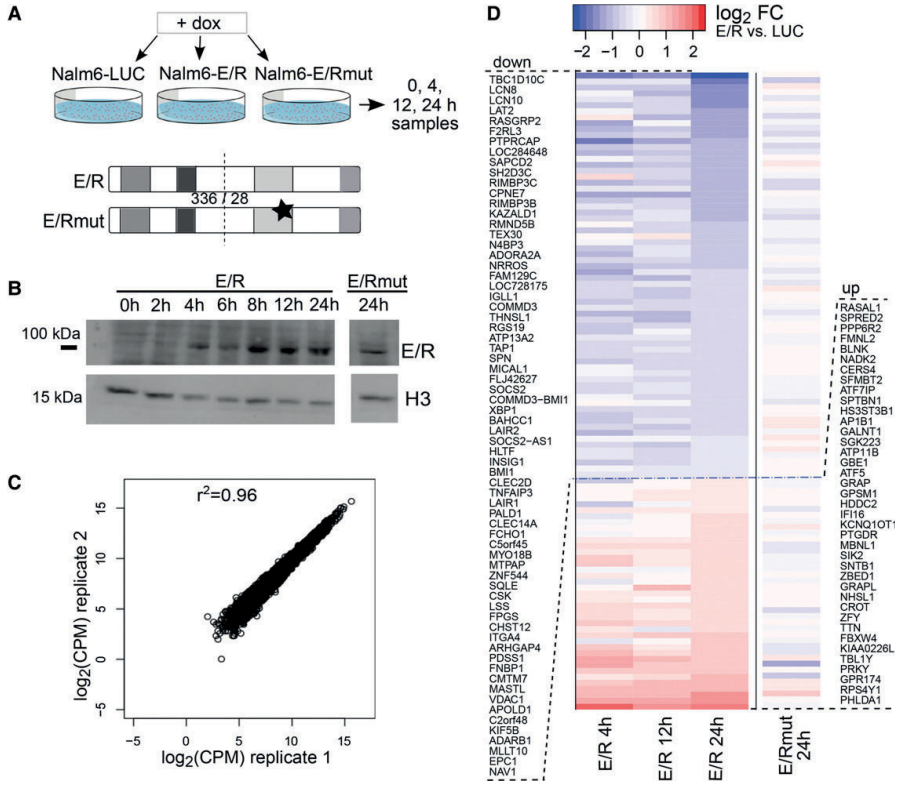

Figure 1. An inducible cell culture model uncovers early transcriptional changes downstream from the ETV6-RUNX1 fusion protein. (*A*) The doxycycline-inducible cell culture model (Nalm6-E/R) with two controls (Nalm6-E/Rmut and Nalm6-LUC) is schematically illustrated and the sampling time points indicated (see also Supplemental Fig. S1). Mutation at the DNA-binding domain of E/R is marked with a star. (*B*) Expression of E/R fusion protein after doxycycline induction at indicated time points (a representative Western blot of two replicates). The H3 antibody was used as a loading control. (*C*) Highly consistent results were obtained from biological replicates, as shown by a correlation plot depicting the GRO-seq signal from Nalm6-LUC cells. Pearson r^2 values for the biological replicate pairs of all samples were between 0.96–0.98. (*D*) A heatmap illustrating magnitude and direction of changes in the GRO-seq signal for the annotated transcripts altered in the E/R sample (and not in E/Rmut) at 24 h. Log_2 fold changes of indicated samples relative to Nalm6-LUC are shown in color with shades of blue and red indicating down-regulation and up-regulation, respectively. E/R-mediated changes were predominantly repressive (for genes, see Supplemental Table S5).

as measured by RT-qPCR showed considerable concordance with primary transcription in GRO-seq (Supplemental Table S1). Overall, two-thirds of the DNA-binding-dependent transcriptional alterations at 24 h were repressive (Fig. 1D).

ETV6-RUNX1–mediated repression occurs at enhancers carrying RUNX1 binding sites

GRO-seq allows TF binding sites to be inferred by investigating sequence-specific DNA motifs within enhancer regions that display expression of eRNAs. We explored changes of eRNA expression in the vicinity of E/R-regulated genes that, based on the mutant E/R results, depended on direct DNA-binding (for transcript-centric list, see Supplemental Table S2 for coordinates) and tested for the overrepresentation of TF motifs (see Supplemental Material, "Genomic regions used in analysis" and "TF motif enrichment analysis"). As shown in Figure 2A, enhancers containing the ETS motif were enriched at both up- and down-regulated loci, whereas the RUNX1 motif ranked highly in the repressed group. As an independent confirmation, we retrieved ChIP-seq profiles for RUNX1 in an ALL cell line (SEM) and for RUNX1, ERG, and FLI1 (the latter two representing ETS factors) in hematopoietic stem cells (HSC) (GSE42075, Wilkinson et al. 2013; GSE45144, Beck et al. 2013). There was significant enrichment of RUNX1 peaks in the ±400 kb vicinity of regulated genes, as shown in Figure 2B, agreeing with the motif enrichment analysis (Fig. 2A). Again, RUNX1 peaks were associated with E/R-mediated repression (Fig. 2B). The result remained consistent when less stringent peak cutoffs were tested (Supplemental Fig. S2). Among the enhancers associated with E/R-regulated genes at ±400 kb from the altered transcription start site (TSS) of a transcript, 67% (315 of 467) contained either a RUNX1 motif or a RUNX1 binding site (ChIP peak in HSC or SEM cells). Notably, we found evidence of a nearby E/R-regulated enhancer for 96% of genes identified by the transcript-centric approach (104/108) and for all but one, at least one enhancer region with evidence of RUNX1 binding or motif (Supplemental Table S2).

A previous work indicated that RUNX1 and ETS factors can also occupy a shared ETS-RUNX motif (Hollenhorst et al. 2009). We observed marked enrichment of repressed RUNX1 binding sites harboring ETS-RUNX motif in both SEM and HSC cells (five-fold, Fig. 2C). ERG ChIP peaks centered by either ERG motif (Fig. 2B) or ETS-RUNX motif (Fig. 2C) showed enrichment in the E/R-repressed category, agreeing with the eRNA-based motif analysis. In contrast, the up-regulated genes lacked enrichment of the studied ChIP peaks (RUNX and ETS factors). The TF motif and ChIP peak enrichment results suggest that E/R acts mainly through binding to a RUNX1 motif and that a subset of genes may be coregulated with ETS factors or indirectly by other TFs.

A

E/R downregulated regions	p-value	Motif match to known TFs	% of regions	% of backgr.
AGGAAGTG	1e-20	ETS motif (0.9)	50	31
CAAACCACAAAC	1e-16	RUNX motif (0.8)	9	2
GTTTTCTG	1e-12	EWS-fusion motif (0.7)	30	18
E/R upregulated regions	**p-value**	**Motif match to known TFs**	**% of regions**	**% of backgr.**
GGCAGGAAGCAG	1e-15	ETS motif (0.7)	21	5
ACTCCCCTCTTCTTA	1e-14	PU.1 motif (0.6)	9	1

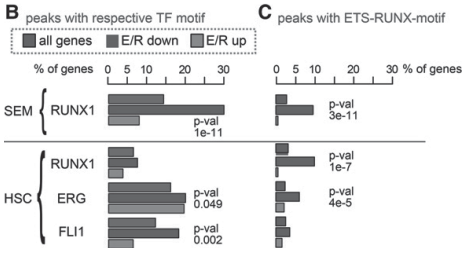

Figure 2. Enrichment of TF motifs at enhancers in the vicinity of E/R-regulated genes. (*A*) Enriched TF motifs at putative enhancer regions within 400 kb of the transcription start site (TSS) of E/R-regulated genes are shown (binomial test with FDR < 0.01 and enrichment in >5% of regions). *P*-value, best-known matching TF motif, and the percentage of regions containing the motif are indicated. The similarity score to known TF motifs is shown in parentheses. (*B*,*C*) Enrichment of ChIP peaks nearby (max 400 kb from TSS) all expressed genes (gray) or those that were either down-regulated (blue) or up-regulated (red) by E/R. Horizontal bars show the percentage of genes associated with a ChIP peak containing either the indicated TF motif (in *B*) or the ETS-RUNX motif (in *C*). *P*-value of the hypergeometric test is shown for statistically significant enrichment in the repressed group ($P < 0.05$). The most prominent ChIP peaks from each data set were used for the analysis. See Supplemental Figure S2 for additional details.

ETV6-RUNX1 targets super-enhancers associated with CD19$^+$/CD20$^+$-cell identity

To further scrutinize the motif and ChIP peak findings (Fig. 2), we explored the eRNA signal distribution upon *E/R* induction from top RUNX1 peaks in SEM cells. As shown in Figure 3, A and B, repression of eRNA signal centered to RUNX1 motif occurred specifically upon wild-type *E/R* expression, reaching the maximal effect at the 24-h time point.

Given the role of RUNX1 in HSC differentiation (for review, see Lutterbach and Hiebert 2000), we next compiled annotated B cell and HSC enhancers based on H3K27ac data (distinguishing between "regular" enhancers and super-enhancers as in Hnisz et al. 2013), and included the most prominent RUNX1-bound sites in SEM and HSC cells (Supplemental Fig. S3A). The GRO-seq signal was then quantified at these sites. This "enhancer-centric" approach allowed us to pinpoint regulation by E/R that directly considers alterations of eRNA (Supplemental Fig. S3B,C). We next tested whether the RUNX1 peaks in SEM cells were generally enriched at SE regions and found enrichment over all H3K27ac-positive regions (4.4-fold, *P*-value 5×10^{-44}) (Fig. 3C). Importantly, E/R-regulated sites (based on the eRNA level analysis) with RUNX1 peaks were also enriched at SE regions (6.0-fold, *P*-value 1.6×10^{-10}). Motivated by this observation, we compared the proportion of SEs from CD34$^+$ cells and CD19$^+$/CD20$^+$ cells, which represent HSC and later stage of B cell differentiation, respectively. The amount of affected SE regions in CD19$^+$/CD20$^+$ cells was 1.5-fold higher compared to CD34$^+$ enhancers, and 65% of the SEs were repressed (Fig. 3D; Supplemental Table S3). These findings indicate the role for the E/R fusion as an impediment for B cell differentiation.

In total, among the 534 E/R-regulated H3K27ac regions revealed by the enhancer-centric analysis, 59 were SEs (for the SE-coordinates, see Supplemental Table S3), and 28 of them (47%) correlated with an alteration at a nearby gene (for the list, see Supplemental Table S2). Intriguing examples of repressed SEs with multiple RUNX1 peaks are the two loci shown in Figure 3E,

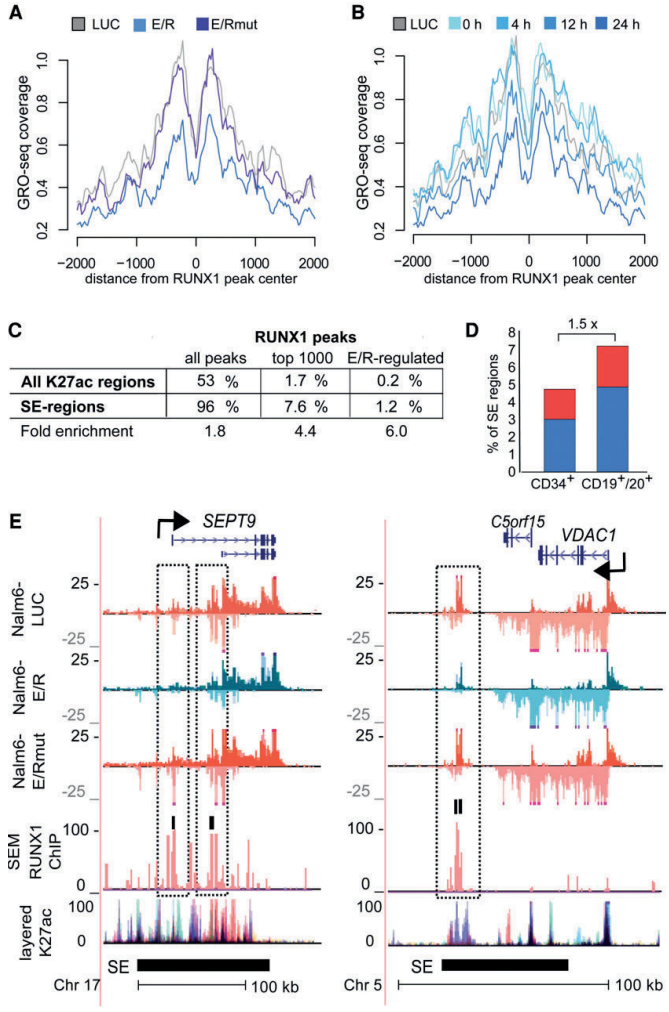

Figure 3. eRNA transcription profiles reveal E/R targets and repression at SE regions. Histograms of GRO-seq coverage are shown at RUNX1-bound genomic regions comparing control (LUC), E/R mutant (E/Rmut), and wild-type E/R samples (*A*) and across the time series of *E/R* induction (*B*). Signal profile represents the most prominent intergenic RUNX1 ChIP peaks from SEM cells centered by the RUNX1 motif. Early repression of eRNAs is evident at the center of RUNX1-bound sites upon expression of *E/R* but not *E/Rmut*. (*C*) Enrichment of RUNX1 ChIP peaks in SEM cells at SE regions over regular H3K27ac regions is shown. The fold enrichment at SE regions is indicated for all peaks (*P*-value $<10^{-50}$), top 1000 peaks (*P*-value 5.2×10^{-44}), and E/R-regulated RUNX1 peaks (*P*-value 1.6×10^{-10}). (*D*) Proportion of E/R-regulated SE regions in $CD34^+$ or $CD19^+/CD20^+$ cells is shown as bar plots. Down-regulation and up-regulation are indicated in blue and red, respectively. In total, 37 of 500 SEs in $CD19^+/CD20^+$ were regulated by E/R, a 1.5-fold excess compared to $CD34^+$ cells (22 of 452 SEs in $CD34^+$). (*E*) GRO-seq signal tracks at representative E/R-regulated SEs are shown. Repression of eRNA signal at prominent RUNX1 ChIP peaks (highlighted in the figure) is observed in the vicinity of *SEPT9* and *VDAC1* genes. Transcript variants 10 and 11 of *SEPT9* are shown in the RefSeq track. Two biological replicates of each GRO-seq sample (Nalm6-LUC, Nalm6-E/R, Nalm6-E/Rmut) are shown with different shades of color, and signals *above* and *below* the axis indicate plus and minus strands, respectively. RUNX1 ChIP peaks in SEM cells are shown in light red overlaid with the input control in shades of blue and purple. SE track is based on $CD19^+/20^+$ cell data from Hnisz et al. 2013. Layered H3K27ac track indicates active enhancers and is shown as an overlaid signal from seven cell lines retrieved from ENCODE (The ENCODE Project Consortium 2012). Color key: GM12878, red; H1-hESC, yellow; HSMM, green; HUVEC, light blue; K562, blue; NHEK, purple; NHLF, pink.

containing cancer-related genes *SEPT9* and *VDAC1* (Peterson and Petty 2010; Brahimi-Horn et al. 2015).

GRO-seq establishes noncoding transcript targets of ETV6-RUNX1

Recent findings have indicated a role for misexpression of long noncoding RNAs (lncRNAs) in various steps of tumorigenesis (Yang et al. 2014). In addition to the eRNAs, our GRO-seq profiling revealed 28 novel E/R-regulated lncRNA transcripts, which were manually classified as either antisense RNAs (14) or intergenic RNAs (14) to depict their putative function (Supplemental Table S4). One of the most robustly E/R down-regulated and annotated lncRNA was *LOC728175* (uncharacterized transcript), which is associated with a SE region in $CD19^+/CD20^+$ cells (Fig. 4A). Down-regulation of *LOC728175* was confirmed by RT-qPCR at mature RNA level after *E/R* induction (Fig. 4B). In evaluating the clinical relevance of this finding, we analyzed its expression in patient samples. RNA-seq measurement revealed that *LOC728175* expression was lower among *E/R*-positive patients ($n = 9$), compared to other precursor B-ALL subtypes ($n = 8$) (Fig. 4C). Similarly, an E/R-regulated noncoding transcript near the SE region of *CLEC2D* locus (Fig. 4D) was repressed among *E/R*-positive patients (Fig. 4E,F). Overall, among the de novo transcripts regulated by E/R in GRO-seq (57 transcripts), 15 showed concordant change in expression (adjusted $P < 0.05$) (Supplemental Table S4).

ETV6-RUNX1 regulates cell adhesion and transmembrane signaling pathways

As SEs are essential in differentiation, their enrichment suggested significant consequences for cellular pathways. The enhancer-centric approach in parallel to transcript-centric approach jointly implicated 183 coding and 13 noncoding annotated transcripts as E/R targets in a Runt DNA-binding domain-dependent manner (Supplemental Table S5). To unveil the pathways implicated, we performed gene ontology analyses for the transcript-centric gene list using the DAVID software (Huang et al. 2009) and for the enhancer-centric regulatory region list by using the GREAT software (see Supplemental Material; McLean et al. 2010). Interestingly, genes related to cell adhesion and transmembrane or

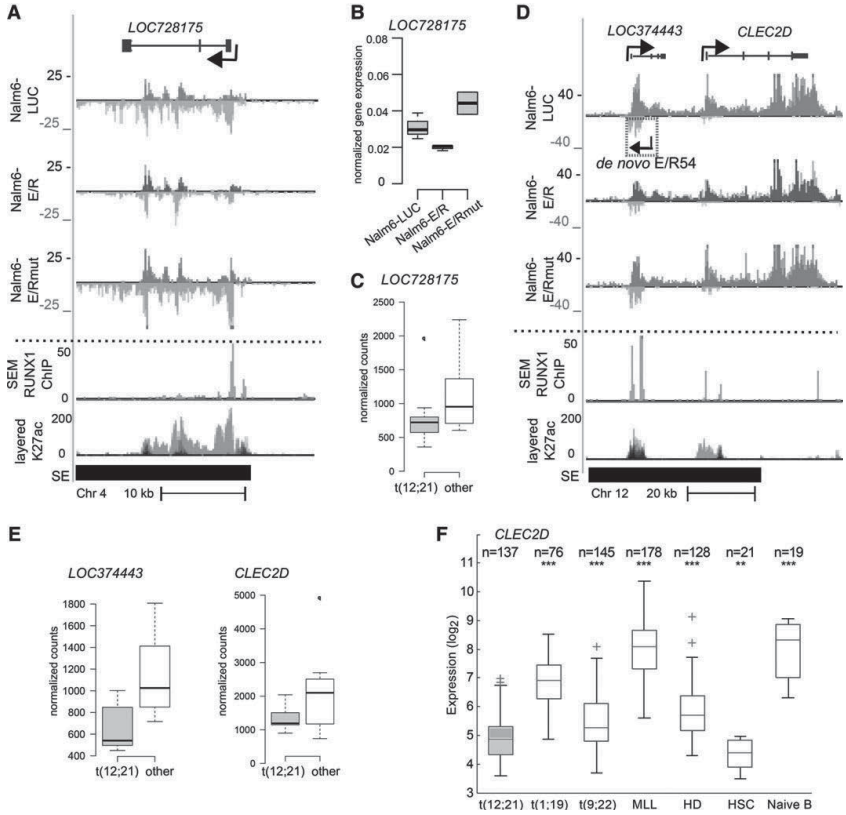

Figure 4. E/R-regulated noncoding genes retain altered expression in patient samples. (*A*) GRO-seq and ChIP-seq signals (as in Fig. 3E) indicate down-regulation of *LOC728175* transcription by E/R, likely through a RUNX1 binding site located at TSS. (*B*) E/R-mediated down-regulation of *LOC728175* in Nalm6-E/R cells as measured by RT-qPCR after 24 h of *E/R* induction. Expression was normalized to the housekeeping gene *GAPDH*. A representative experiment is shown with technical variation (lowest and highest datum within 1.5 × IQR). (*C*) RNA-seq normalized count values for *LOC728175* among *E/R*-positive (t12;21) and *E/R*-negative (other) patients indicate lower expression in the *E/R*-positive group (adjusted *P*-value 0.37; E/R, $n = 9$; other, $n = 8$). (*D*) GRO-seq and ChIP-seq signals shown at a locus containing three repressed genes (*LOC374443*, *CLEC2D*, and a de novo transcript *E/R54*) and colocalizing with RUNX1-binding sites. Tracks are as in Figure 3E. (*E*) RNA-seq normalized count values for *LOC374443* and *CLEC2D* show that decreased expression is maintained at the diseased state (*LOC374443* adjusted *P*-value 0.020; *CLEC2D* adjusted *P*-value 0.027). No reads mapped to *E/R54*. (*F*) Combined microarray data indicate repression of *CLEC2D* among *E/R*-positive patients. In each comparison, statistical significance (Mann-Whitney *U* test) was tested against *E/R*-positive subtype: (***) $P < 0.001$; (**) $P < 0.01$. Tukey whiskers are shown for each box plot (1.5 × IQR).

intracellular signaling were highly ranked among functional annotation groups in both analyses (Fig. 5A; for detailed results, see Supplemental Table S6).

This prompted us to examine the genes related to cell adhesion and signaling more closely. Knockout of *ITGA4*, a gene that belongs to the enriched integrin signaling pathway (Fig. 5A) and encodes a transmembrane subunit of the VLA4 receptor, results in a differentiation block prior to pro-B cell stage in mouse (Arroyo et al. 1996). A locus containing *ITGA4* is among the E/R-regulated SE regions that contain a ChIP peak for RUNX1 (Fig. 5B), suggesting direct binding. As shown in Figure 5C, expression of the mature *ITGA4* transcript was repressed by E/R containing the wild-type Runt domain and up-regulated in REH sh*E/R*-knockdown cells. We confirmed direct binding of wild-type E/R at *ITGA4* promoter, coinciding with SE, with a ChIP experiment using ETV6 antibodies (Fig. 5D). The direct effect was further supported by the lack of ChIP-enrichment of the E/Rmut sample.

We also investigated E/R targets between different RUNX1 fusions and observed *LAT2* (linker for activation of T-cells family member 2), a signaling transmembrane protein, as a shared target for E/R and the RUNX1-RUNX1T1 fusion (Brdicka et al. 2002; Janssen et al. 2003; Duque-Afonso et al. 2011a). In acute myeloid leukemia, binding of the RUNX1-RUNX1T1 fusion results in widespread alterations throughout the epigenome (Ptasinska et al. 2012, 2014). A binding site for RUNX1-RUNX1T1 (Duque-Afonso et al. 2011b) at intronic region 3 of *LAT2*, which contains prominent peaks in RUNX1-ChIP assay, was markedly repressed in GRO-seq by the E/R fusion (Fig. 5E). Repression of mature *LAT2* transcript was evident after *E/R* induction, whereas up-regulation was detected in REH sh*E/R*-knockdown cells (Fig. 5F). Both

ITGA4 and *LAT2* were consistently repressed in a large patient data set among *E/R*-positive patients (Fig. 6A).

ALL patient data sets discern ETV6-RUNX1 targets

In examining the persistence of observed changes in overt disease, we compiled a large single-platform gene expression data set from the GEO microarray repository, including 664 precursor B-ALL samples of all ages and with known genetics (Supplemental Material; Heinäniemi et al. 2013; Liuksiala et al. 2014). This was also motivated by the limited overlap between previous microarray studies in patients (Supplemental Fig. S3D; Supplemental Material) and cell culture models (Fuka et al. 2011; Linka et al. 2013). With the unified sample set, a quarter (35/133) of the E/R fusion targets identified by the GRO-seq and measured in the microarray data set were found differentially expressed between *E/R*-positive and *E/R*-negative ALL samples (Supplemental Table S7). The correlation between the $\log_2$ fold changes of the 35 differentially expressed genes was significant (Pearson's correlation coefficient 0.529 with a *P*-value of 0.0011) (Fig. 6B). In a majority of the genes (23/29), concordant alterations were observed in RNA-seq analysis of a cohort of pediatric ALL patients consisting of nine *E/R*-positive and eight other precursor B-ALL patients (13/29 with adjusted $P < 0.05$) (Supplemental Table S7).

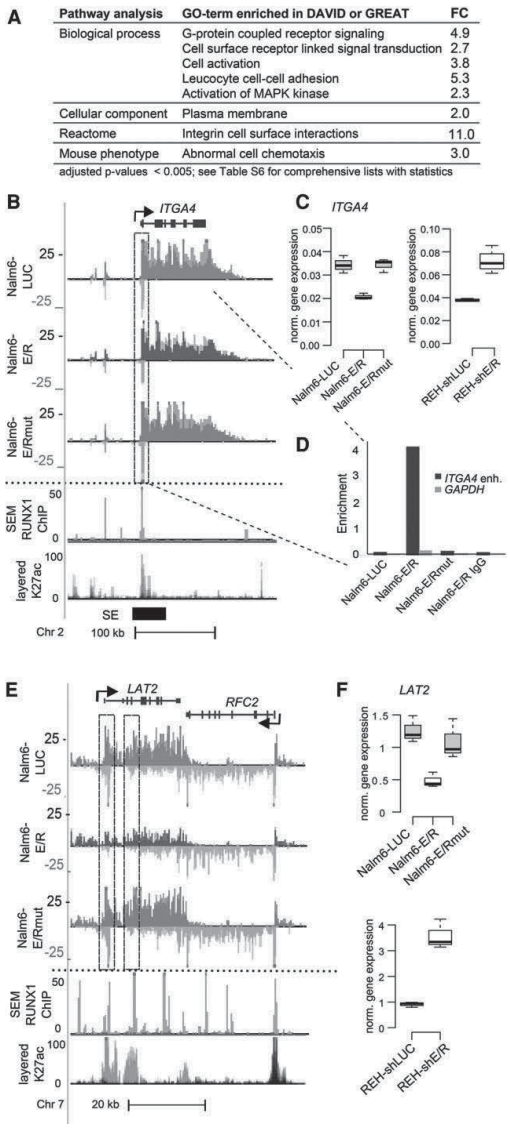

Figure 5. Transmembrane signaling is affected by E/R. (*A*) Excerpt of the most significantly enriched gene ontology and pathway terms for E/R-regulated transcript-centric and enhancer-centric analyses (see also Supplemental Table S6). (*B*) GRO-seq and ChIP-seq signals at the *ITGA4* locus are shown as in Figure 3E. The annotated SE region present in both $CD19^+/CD20^+$ cells and $CD34^+$ cells (Hnisz et al. 2013) harbors a prominent RUNX1 peak downstream from the TSS. This RUNX1 peak is referred to in *D*. The primary transcript is repressed by E/R. (*C*) *ITGA4* mRNA expression level as measured by RT-qPCR after 24 h of *E/R* induction (Nalm6-E/R) or after silencing of endogenous *E/R* in REH cells (REH-sh*E/R*). A representative experiment with technical variation (1.5 × IQR) is shown. Expression is normalized to the housekeeping gene *GAPDH*. (*D*) ETV6 ChIP assay with qPCR (primer sites are in the *middle* of the RUNX1 peak highlighted in *B*) validates the binding of E/R after 24 h induction, while E/Rmut, LUC, or nonspecific IgG antibody shows no enrichment in comparison to the control region at the *GAPDH* promoter area. A representative figure of two independent ChIP experiments is shown. (*E*) *LAT2* transcript is repressed in Nalm6-E/R cells (refer to Fig. 3E for tracks). Dashed boxes indicate the TSS and intron 3 of *LAT2*. (*F*) Expression of *LAT2* mRNA relative to housekeeping gene *HMBS* as measured by RT-qPCR after induction of *E/R* for 24 h or after silencing of endogenous *E/R* in REH cells. A representative experiment is shown with technical variation (1.5 × IQR).

Discussion

Enhancers represent key control switches that regulate initiation of transcription at nearby gene loci by integrating signals from multiple TFs (Arner et al. 2015). Transcription of eRNA is considered the most precise mark of functional looping between an activated enhancer and its regulated gene promoter, and the dynamic changes observed in eRNA levels allow elucidation of key TFs upon cell differentiation and environmental stimuli (Wang et al. 2011; Kaikkonen et al. 2013). We present here the first application of eRNA quantification to elucidate aberrant transcriptional activity downstream from a fusion TF. With our controllable cell model coupled to GRO-seq, we were able to identify E/R targets at both regulatory regions and at protein-coding and noncoding genes. To establish a link between regulatory region activity and gene transcription downstream from E/R, a correlation-based approach was applied to link enhancers to target genes and combined with motif annotation and ChIP-seq data. Further, we used the mutated E/R to exclude effects that did not involve direct DNA binding. Complementary experimental validation using an ETV6 ChIP assay showed that a promoter region of *ITGA4* was bound by the E/R fusion and not by its mutated form. This demonstrates how the initial regulatory map based on GRO-seq signal correlation can be further validated and refined.

Directly measuring nascent RNA production, GRO-seq bypasses issues of antibody specificity and cross-linking efficiency

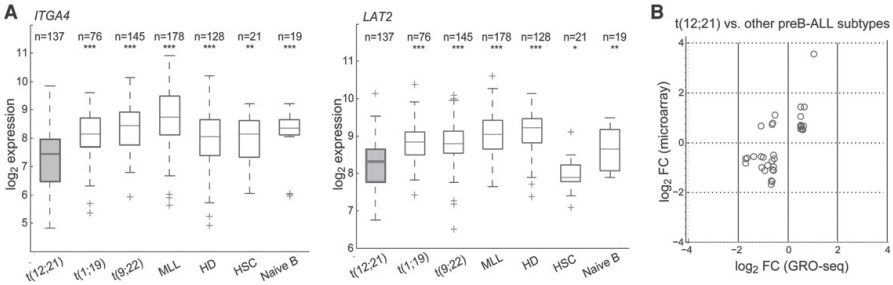

Figure 6. Patient data indicate that E/R-perpetuated changes persist at diagnostic samples. (*A*) Expression of *ITGA4* and *LAT2* in microarrays shows repression among *E/R*-positive patients compared to other subtypes: (MLL) *KMT2A* (*MLL*) rearranged; (HD) hyperdiploid; (HSC) hematopoietic stem cells. In each comparison, statistical significance (Mann-Whitney *U* test) was tested against the *E/R*-positive subtype: (***) $P < 0.001$; (**) $P < 0.01$; (*) $P < 0.05$. (*B*) Concordance between differentially expressed genes on GRO-seq and combined microarray (Pearson's correlation coefficient 0.529, $P = 0.0011$) (Supplemental Table S7).

associated with ChIP assays (Slattery et al. 2014). Moreover, mapping between an enhancer and gene transcription is achieved more directly than by integrating mature transcript levels. By detecting de novo eRNAs in the vicinity of annotated E/R-regulated genes, and in an alternative approach by quantifying eRNA levels at previously characterized enhancers, we observed a rapid repression of prominent RUNX1 binding sites, and RUNX1 and ETS motifs, as underlying the global changes in eRNA levels and gene transcription upon *E/R* induction. Typically, the gene loci that were concomitantly repressed harbored multiple E/R-regulated enhancers with prominent RUNX1 peaks. Significant regulatory effects at enhancers translated into significant alterations of transcription at nearby genes in one-third of genomic loci. Lack of more complete concordance may reflect complex regulatory mechanisms or imply a role for E/R in epigenetic remodeling that impact the region with a delay.

Recently, a subset of T-ALL patients was found to harbor mutations near the oncogenic *TAL1* locus that leads to establishment of a highly active SE via MYB binding (Mansour et al. 2014). We observed enrichment of E/R-regulated RUNX1 binding sites in SEs. Moreover, several affected binding sites localized to gene loci with an established role in cancer, including *SEPT9* and *VDAC1*. Intriguingly, *SEPT9* is a leukemic fusion partner for *KMT2A* (*MLL*) and associated with oncogenic signaling pathways in many cancers (e.g., by interacting with HIF-1-alpha) (Peterson and Petty 2010), and the knockout of *VDAC1* in mouse induces tumor growth (Brahimi-Horn et al. 2015). Accompanying these potential oncogenic changes, repression of super-enhancers associated with $CD19^+/CD20^+$ cells could underlie the differentiation arrest at the pro-/pre-B cell stage in *E/R*-positive patients and provide further evidence implicating regulatory regions in cancer development (Hnisz et al. 2015).

Our main finding that approximately two-thirds of E/R targets were repressed is in accordance with early reports that suggested a repressive role for E/R (Hiebert et al. 1996; Fenrick et al. 1999; Guidez et al. 2000). Although the majority of the E/R targets were repressed, one-third of them were up-regulated, possibly through other TFs such as ETS factors as suggested by motif analysis, and a few of the genes were regulated independently of DNA binding through the Runt domain, suggesting indirect effects (data not shown). One potential caveat in our experimental setup is the ectopic expression of *E/R*, which may allow nonspecific DNA binding. To moderate possible cell-line–specific effects and to aid in distinction of clinically relevant targets, we integrated data from patient samples. We were able to pinpoint a considerable number of genes targeted by E/R within hours of induction that were also present in primary ALL patients carrying the t(12;21) translocation. In future, it would be informative to evaluate E/R targets during various stages of early B lymphopoiesis to better understand the dynamics of disease initiation.

In conclusion, we demonstrate that E/R functions as a repressive TF at genomic sites containing the RUNX1 motif. Based on our results and literature (Schindler et al. 2009; van Delft et al. 2011), we suggest that E/R impairs B cell differentiation through transcriptional reprogramming and renders pre-leukemic cells susceptible to additional genetic hits for an extended period of time. The results pave the way for further characterization of the TF network that mediates commitment to B cell fate and the specific contribution of leukemic TF fusions derailing this process.

Methods

Cloning, cell culture, and chromatin immunoprecipitation

pLVX-Tight-Puro-*ETV6-RUNX1* (*E/R*) construct was generated by cloning *ETV6-RUNX1* cDNA (a kind gift from Professor Renate Panzer-Grümayer) into the inducible expression vector (Clontech). Point mutation G1553A was introduced by site-directed mutagenesis resulting in the *E/R*mut construct. The sh*E/R* construct was generated by cloning short hairpin RNA (shRNA) oligos targeting *ETV6-RUNX1* (target sequence GAATAGCAGAATGC ATACTT) into pLVX-shRNA1 vector (Clontech). Nalm6-cells (ACC 128, DSMZ) were infected with the regulatory vector TetOn and subsequently with one of the pLVX response vectors: *E/R*, *E/Rmut*, or *LUC* (luciferase control) (Clontech). The expression of *E/R* in Nalm6-cells was induced with 500 ng/mL doxycycline (Clontech) and confirmed by RT-qPCR and Western blotting. RT-qPCR was performed using iScpript (BioRad) and SsoFast EvaGreen Supermix (BioRad), and the relative $2^{-\Delta\Delta C_T}$ method (Livak and Schmittgen 2001) was used for quantification. Western blot to detect E/R was performed using anti-ETV6 (HPA000264, RRID:AB_611466, Atlas Antibodies). For chromatin immunoprecipitation (ChIP), 2×10^7 cells per immunoprecipitation were harvested after 24 h induction and cross-linked using ethylene glycol bis[succinimidylsuccinate] (EGS) (Thermo Fisher Scientific) and formaldehyde (J.T. Baker, Avantor). Two antibodies against ETV6 were pooled for IP (sc-166835, RRID:AB_2101020

[Santa Cruz Biotechnology] and HPA000264, RRID:AB_611466 [Atlas Antibodies]).

Further details and all primer sequences are listed in Supplemental Material.

GRO-seq assay and processing of GRO-seq and ChIP sequencing reads

The nuclei isolation (yielding ~5×10^6 nuclei per condition), the nuclear run-on reaction, and library preparation were performed as previously described (Wang et al. 2011, Kaikkonen et al. 2013; Supplemental Material). The ChIP-seq data from human HSC and SEM cells (GSE45144, Beck et al. 2013; GSE42075, Wilkinson et al. 2013; originally mapped to hg18) were reanalyzed starting from raw reads. See Supplemental Material for details of processing and visualizing the GRO and ChIP sequencing reads.

Genomic regions used in analysis

ChIP-seq: Peak detection was performed using HOMER program findPeaks style factor (http://homer.salk.edu/homer/ngs/peaks.html) against the respective control (IgG or input). Due to the different number of peaks called from each experiment, the 1000 peaks with highest enrichment were used to represent prominent binding of the TF in question (HOMER program getTopPeaks.pl) (Supplemental Fig. S2). Hypergeometric *P*-values for RUNX1-peak enrichments were calculated for greater or equal difference than observed.

GRO-seq: The HOMER program findPeaks.pl was used to identify de novo transcripts from GRO-seq data using a pooled read library and allowing gaps at nonmappable regions. Transcript categories identified and further details are described in Supplemental Material.

Differential expression analysis of GRO-seq transcripts

Transcript quantification is described in detail in the Supplemental Material.

For the transcript-centric analysis, transcripts expressed at a level RPKM > 1 (reads per kilobase per million mapped reads) in at least two samples and with at least 20 reads within the quantified region in any sample were used for statistical analysis, excluding snoRNAs. Differentially expressed transcripts were identified using the R/Bioconductor package edgeR (Robinson et al. 2010). A linear model was fitted to the RLE-normalized data using a group-mean design matrix. Contrasts between the different conditions were performed, and transcripts with at least 1.5-fold change in expression level and adjusted *P*-value < 0.05 (Benjamini-Hochberg method using *P*-values from moderated *t*-test) were defined as significantly regulated (Benjamini and Hochberg 1995). The correlation and Euclidean distance between the $\log_2$ fold change values observed for the transcript and dynamically regulated eRNAs (as defined above) ±400 kb from TSS was used to assign candidate regulatory elements for each transcript. The transcript-centric analysis was coupled with TF motif enrichment analysis ±400 kb from significantly regulated genes (for further details, see Supplemental Material).

For the enhancer-centric analysis, the eRNAs passing the expression level-based cutoffs (RPKM 2.5 and at least 10 reads) at annotated B cell, HSC, and RUNX1-bound enhancers based on previously reported H3K27ac ChIP-seq data (Hnisz et al. 2013), and the top ChIP-seq peaks in HSC (GSE45144) and SEM cells (GSE42075), were included in the statistical analysis, performed as above. The eRNAs with adjusted *P*-value < 0.1 and lacking a response in the E/Rmut sample were associated with nearby genes based on $\log_2$ fold change values of eRNA and gene body transcription, similarly as above. Further, only gene transcripts with significant change (two-tailed *t*-test) between E/R 24 h and LUC were reported. Notice that this approach performs multiple testing corrections on eRNA changes and prioritizes E/R regulation at enhancers, with no fold change cutoff for associated gene transcripts applied, compared to the transcript-centric analysis. Consistent with the gene transcription profiles, multidimensional scaling of the eRNA profiles grouped biologically similar samples together (Supplemental Fig. S3C), validating the enhancer-centric approach as a complementary way to improve detection of E/R targets.

Microarray data set and RNA sequencing

The microarray data were retrieved from the NCBI GEO database representing healthy and malignant hematological samples hybridized to Affymetrix GeneChip Human Genome U133 Plus 2.0 array (for GSE accession numbers, references, and further details of data processing, see Supplemental Material). Differential expression was defined by a minimum absolute $\log_2$ fold change of 0.5 between group medians and maximum *Q*-value of 0.01.

Publicly available RNA-seq data for 17 primary BCP-ALL patients were obtained from the Gene Expression Omnibus data set GSE79373. The data were available from nine BCP-ALL patients harboring the *ETV6-RUNX1* fusion gene and eight BCP-ALL patients without *ETV6-RUNX1* (hyperdiploid $n = 7$, other $n = 1$). Aligned reads were summarized using featureCounts (Liao et al. 2014). Differential expression was performed using the DESeq2 package in R (Love et al. 2014). Expression analysis was only run on the regions selected based on GRO-seq: 35 E/R-regulated protein-coding genes (intersect of GRO-seq data and the combined microarray analysis) and 57 de novo regions. See Supplemental Tables S4 and S7 for normalized count values used in RNA-seq box plots.

Data access

GRO sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE67519.

Acknowledgments

We thank Minna Kaikkonen for advice in GRO-seq methodology. We thank the Sequencing Service GeneCore Sequencing Facility (EMBL, http://genecore3.genecore.embl.de/genecore3/index.cfm) for DNA sequencing. This work was supported by grants from the Academy of Finland (project number 277816 to O.L.; 276634 to M.H.; 265575 to K.V.); The Foundation for Pediatric Research (O.L.); Jane and Aatos Erkko Foundation (O.L.); Sohlberg Foundation (K.V.); Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital (grant numbers 9R029 and 9S032 to O.L.); the Swedish Cancer Society (CAN2010/592 to A.C.S.); the Swedish Research Council for Science and Technology (VR-NT 90559401 to A.C.S.); and The Finnish Cultural Foundation (Interdisciplinary Science Workshops to M.H.).

References

Adam RC, Yang H, Rockowitz S, Larsen SB, Nikolova M, Oristian DS, Polak L, Kadaja M, Asare A, Zheng D, et al. 2015. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521:** 366–370.

Andersson A, Edén P, Lindgren D, Nilsson J, Lassen C, Heldrup J, Fontes M, Borg A, Mitelman F, Johansson B, et al. 2005. Gene expression profiling

of leukemic cell lines reveals conserved molecular signatures among subtypes with specific genetic aberrations. *Leukemia* **19:** 1042–1050.
Andersson A, Ritz C, Lindgren D, Edén P, Lassen C, Heldrup J, Olofsson T, Råde J, Fontes M, Porwit-MacDonald A, et al. 2007. Microarray-based classification of a consecutive series of 121 childhood acute leukemias: prediction of leukemic and genetic subtype as well as of minimal residual disease status. *Leukemia* **21:** 1198–1203.
Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drablos F, Lennartsson A, Ronnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347:** 1010–1014.
Arroyo AG, Yang JT, Rayburn H, Hynes RO. 1996. Differential requirements for α4 integrins during fetal and adult hematopoiesis. *Cell* **85:** 997–1008.
Bateman CM, Alpar D, Ford AM, Colman SM, Wren D, Morgan M, Kearney L, Greaves M. 2015. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. *Leukemia* **29:** 58–65.
Beck D, Thoms JA, Perera D, Schütte J, Unnikrishnan A, Knezevic K, Kinston SJ, Wilson NK, O'Brien TA, Göttgens B, et al. 2013. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood* **122:** e12–e22.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J R Stat Soc B* **57:** 289–300.
Brahimi-Horn MC, Giuliano S, Saland E, Lacas-Gervais S, Sheiko T, Pelletier J, Bourget I, Bost F, Féral C, Boulter E, et al. 2015. Knockout of *Vdac1* activates hypoxia-inducible factor through reactive oxygen species generation and induces tumor growth by promoting metabolic reprogramming and inflammation. *Cancer Metab* **3:** 8.
Brdicka T, Imrich M, Angelisová P, Brdicková N, Horváth O, Spicka J, Hilgert I, Lusková P, Dráber P, Novák P, et al. 2002. Non-T cell activation linker (NTAL): a transmembrane adaptor protein involved in immunoreceptor signaling. *J Exp Med* **196:** 1617–1626.
Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.
Duque-Afonso J, Solari L, Essig A, Berg T, Pahl HL, Lübbert M. 2011a. Regulation of the adaptor molecule LAT2, an *in vivo* target gene of AML1/ETO (*RUNX1/RUNX1T1*), during myeloid differentiation. *Br J Haematol* **153:** 612–622.
Duque-Afonso J, Yalcin A, Berg T, Abdelkarim M, Heidenreich O, Lübbert M. 2011b. The HDAC class I-specific inhibitor entinostat (MS-275) effectively relieves epigenetic silencing of the *LAT2* gene mediated by AML1/ETO. *Oncogene* **30:** 3062–3072.
The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.
Fears S, Gavin M, Zhang DE, Hetherington C, Ben-David Y, Rowley JD, Nucifora G. 1997. Functional characterization of *ETV6* and *ETV6/CBFA2* in the regulation of the *MCSFR* proximal promoter. *Proc Natl Acad Sci* **94:** 1949–1954.
Fenrick R, Amann JM, Lutterbach B, Wang L, Westendorf JJ, Downing JR, Hiebert SW. 1999. Both TEL and AML-1 contribute repression domains to the t(12;21) fusion protein. *Mol Cell Biol* **19:** 6566–6574.
Fine BM, Stanulla M, Schrappe M, Ho M, Viehmann S, Harbott J, Boxer LM. 2004. Gene expression patterns associated with recurrent chromosomal translocations in acute lymphoblastic leukemia. *Blood* **103:** 1043–1049.
Fuka G, Kauer M, Kofler R, Haas OA, Panzer-Grümayer R. 2011. The leukemia-specific fusion gene *ETV6/RUNX1* perturbs distinct key biological functions primarily by gene repression. *PLoS One* **6:** e26348.
Gandemer V, Rio AG, de Tayrac M, Sibut V, Mottier S, Ly Sunnaram B, Henry C, Monnier A, Berthou C, Le Gall E, et al. 2007. Five distinct biological processes and 14 differentially expressed genes characterize *TEL/AML1*-positive leukemia. *BMC Genomics* **8:** 385.
Golub TR, Barker GF, Bohlander SK, Hiebert SW, Ward DC, Bray-Ward P, Morgan E, Raimondi SC, Rowley JD, Gilliland DG. 1995. Fusion of the *TEL* gene on 12p13 to the *AML1* gene on 21q22 in acute lymphoblastic leukemia. *Proc Natl Acad Sci* **92:** 4917–4921.
Greaves M, Colman SM, Kearney L, Ford AM. 2011. Fusion genes in cord blood. *Blood* **117:** 369–370.
Guidez F, Petrie K, Ford AM, Lu H, Bennett CA, MacGregor A, Hannemann J, Ito Y, Ghysdael J, Greaves M, et al. 2000. Recruitment of the nuclear receptor corepressor N-CoR by the TEL moiety of the childhood leukemia–associated TEL-AML1 oncoprotein. *Blood* **96:** 2557–2561.
Heinäniemi M, Nykter M, Kramer R, Wienecke-Baldacchino A, Sinkkonen L, Zhou JX, Kreisberg R, Kauffman SA, Huang S, Shmulevich I. 2013. Gene-pair expression signatures reveal lineage control. *Nat Methods* **10:** 577–583.
Hiebert SW, Sun W, Davis JN, Golub T, Shurtleff S, Buijs A, Downing JR, Grosveld G, Roussell MF, Gilliland DG, et al. 1996. The t(12;21) translocation converts AML-1B from an activator to a repressor of transcription. *Mol Cell Biol* **16:** 1349–1355.
Hiebert SW, Lutterbach B, Amann J. 2001. Role of co-repressors in transcriptional repression mediated by the t(8;21), t(16;21), t(12;21), and inv(16) fusion proteins. *Curr Opin Hematol* **8:** 197–200.
Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155:** 934–947.
Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. 2015. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* **58:** 362–370.
Hollenhorst PC, Chandler KJ, Poulsen RL, Johnson WE, Speck NA, Graves BJ. 2009. DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet* **5:** e1000778.
Huang DW, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37:** 1–13.
Inaba H, Greaves M, Mullighan CG. 2013. Acute lymphoblastic leukaemia. *Lancet* **381:** 1943–1955.
Janssen E, Zhu M, Zhang W, Koonpaew S, Zhang W. 2003. LAB: a new membrane-associated adaptor molecule in B cell activation. *Nat Immunol* **4:** 117–123.
Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, et al. 2013. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell* **51:** 310–325.
Lausten-Thomsen U, Madsen HO, Vestergaard TR, Hjalgrim H, Nersting J, Schmiegelow K. 2011. Prevalence of t(12;21)[*ETV6-RUNX1*]-positive cells in healthy neonates. *Blood* **117:** 186–189.
Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. 2002. Locus control regions. *Blood* **100:** 3077–3086.
Li Z, Yan J, Matheny CJ, Corpora T, Bravo J, Warren AJ, Bushweller JH, Speck NA. 2003. Energetic contribution of residues in the Runx1 Runt domain to DNA binding. *J Biol Chem* **278:** 33088–33096.
Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930.
Linka Y, Ginzel S, Krüger M, Novosel A, Gombert M, Kremmer E, Harbott J, Thiele R, Borkhardt A, Landgraf P. 2013. The impact of TEL-AML1 (ETV6-RUNX1) expression in precursor B cells and implications for leukaemia using three different genome-wide screening methods. *Blood Cancer J* **3:** e151.
Liuksiala T, Teittinen KJ, Granberg K, Heinäniemi M, Annala M, Mäki M, Nykter M, Lohi O. 2014. Overexpression of SNORD114–3 marks acute promyelocytic leukemia. *Leukemia* **28:** 233–236.
Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C}_T$ method. *Methods* **25:** 402–408.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550.
Lutterbach B, Hiebert SW. 2000. Role of the transcription factor AML-1 in acute leukemia and hematopoietic differentiation. *Gene* **245:** 223–235.
Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, et al. 2014. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346:** 1373–1377.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501.
Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, Hows JM, Navarrete C, Greaves M. 2002. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci* **99:** 8242–8247.
Morrow M, Samanta A, Kioussis D, Brady HJM, Williams O. 2007. TEL-AML1 preleukemic activity requires the DNA binding domain of AML1 and the dimerization and corepressor binding domains of TEL. *Oncogene* **26:** 4404–4414.
Peterson EA, Petty EM. 2010. Conquering the complex world of human septins: implications for health and disease. *Clin Genet* **77:** 511–524.
Pott S, Lieb JD. 2014. What are super-enhancers? *Nat Genet* **47:** 8–12.
Ptasinska A, Assi SA, Mannari D, James SR, Williamson D, Dunne J, Hoogenkamp M, Wu M, Care M, McNeill H, et al. 2012. Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia* **26:** 1829–1841.
Ptasinska A, Assi SA, Martinez-Soria N, Imperato MR, Piper J, Cauchy P, Pickin A, James SR, Hoogenkamp M, Williamson D, et al. 2014. Identification of a dynamic transcriptional network in t(8;21) AML that regulates differentiation block and self-renewal. *Cell Rep* **8:** 1974–1988.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.

Romana SP, Le Coniat M, Berger R. 1994. t(12;21): a new recurrent translocation in acute lymphoblastic leukemia. *Genes Chromosomes Cancer* **9:** 186–191.

Ross M, Zhou X. 2003. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **102:** 2951–2959.

Schindler JW, Van Buren D, Foudi A, Krejci O, Qin J, Orkin SH, Hock H. 2009. TEL-AML1 corrupts hematopoietic stem cells to persist in the bone marrow and initiate leukemia. *Cell Stem Cell* **5:** 43–53.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. 2014. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39:** 381–399.

Song H, Kim JH, Rho JK, Park SY, Kim CG, Choe SY. 1999. Functional characterization of TEL/AML1 fusion protein in the regulation of human CR1 gene promoter. *Mol Cells* **9:** 560–563.

Uchida H, Downing JR, Miyazaki Y, Frank R, Zhang J, Nimer SD. 1999. Three distinct domains in TEL-AML1 are required for transcriptional repression of the IL-3 promoter. *Oncogene* **18:** 1015–1022.

van Delft FW, Horsley S, Colman S, Anderson K, Bateman C, Kempski H, Zuna J, Eckert C, Saha V, Kearney L, et al. 2011. Clonal origins of relapse in *ETV6-RUNX1* acute lymphoblastic leukemia. *Blood* **117:** 6247–6254.

Wang Q, Stacy T, Binder M, Marin-Padilla M, Sharpe AH, Speck NA. 1996. Disruption of the *Cbfa2* gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. *Proc Natl Acad Sci* **93:** 3444–3449.

Wang LC, Swat W, Fujiwara Y, Davidson L, Visvader J, Kuo F, Alt FW, Gilliland DG, Golub TR, Orkin SH. 1998. The *TEL/ETV6* gene is required specifically for hematopoiesis in the bone marrow. *Genes Dev* **12:** 2392–2402.

Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, et al. 2011. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474:** 390–394.

Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153:** 307–319.

Wiemels JL, Ford AM, Van Wering ER, Postma A, Greaves M. 1999. Protracted and variable latency of acute lymphoblastic leukemia after *TEL-AML1* gene fusion in utero. *Blood* **94:** 1057–1062.

Wilkinson AC, Ballabio E, Geng H, North P, Tapia M, Kerry J, Biswas D, Roeder RG, Allis CD, Melnick A, et al. 2013. *RUNX1* is a key target in t(4;11) leukemias that contributes to gene activation through an AF4-MLL complex interaction. *Cell Rep* **3:** 116–127.

Yang L, Froberg JE, Lee JT. 2014. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci* **39:** 35–43.

Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, et al. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1:** 133–143.

Zuna J, Madzo J, Krejci O, Zemanova Z, Kalinova M, Muzikova K, Zapotocky M, Starkova J, Hrusak O, Horak J, et al. 2011. *ETV6/RUNX1* (*TEL/AML1*) is a frequent prenatal first hit in childhood leukemia. *Blood* **117:** 368–369.

Received April 26, 2015; accepted in revised form September 12, 2016.

PUBLICATION II

Deregulation of the non-coding genome in leukemia

Teppo, S., Heinäniemi M., & Lohi, O.

RNA Biology 2017, 14(7): 827-830
doi: 10.1080/15476286.2017.1312228

RNA BIOLOGY
2017, VOL. 14, NO. 7, 827–830
https://doi.org/10.1080/15476286.2017.1312228

Taylor & Francis
Taylor & Francis Group

POINT OF VIEW

Deregulation of the non-coding genome in leukemia

Susanna Teppo [a], Merja Heinäniemi [b], and Olli Lohi [a]

[a]Tampere Center for Child Health Research, Faculty of Medicine and Life Sciences, University of Tampere and Tampere University Hospital, Tampere, Finland; [b]Institute of Biomedicine, School of Medicine, University of Eastern Finland, Kuopio, Finland

ABSTRACT
Methodological advances that allow deeper characterization of non-coding elements in the genome have started to reveal the full spectrum of deregulation in cancer. We generated an inducible cell model to track transcriptional changes after induction of a well-known leukemia-inducing fusion gene, ETV6-RUNX1. Our data revealed widespread transcriptional alterations outside coding elements in the genome. This adds to the growing list of various alterations in the non-coding genome in cancer and pinpoints their role in diseased cellular state.

ARTICLE HISTORY
Received 7 February 2017
Revised 21 March 2017
Accepted 24 March 2017

KEYWORDS
eRNA; GRO-seq; leukemia; nascent RNA; transcriptional regulation

Approximately 80 % of the genome is transcribed into RNA species in at least some cell type or at some stage of development.[1,2] Non-coding regulatory (non-housekeeping) RNAs are currently defined by their size, genomic location or presumptive function. Enhancer RNAs (eRNA), which have a length span from 0.1 to 10 kb, mainly fall into the category of long non-coding RNAs (lncRNAs) although they are better defined by their transcriptional regulatory function. Larger clusters of enhancers with multiple transcription factor (TF) binding sites and open chromatin marks are termed super-enhancers and they define cell identity.[3,4] Locations of enhancer elements are often deduced from certain histone marks (H3K4me1, H3K27ac), transcription factor binding profiles (p300), or open chromatin states (eg. DNAse- and ATAC-seq). The development of global nascent RNA sequencing techniques, such as global run-on sequencing (GRO-seq),[5] has revealed that transcription of eRNAs is highly correlated with marks such as H3K27ac (for review see ref. 6) and to transcription at nearby gene promoters,[7,8] and is considered the most reliable mark of an active enhancer.[7,9] The functions of eRNAs are yet unclear: they can be passive byproducts of transcription or function actively in recruitment of transcription factors (reviewed in ref. 10), like in the case of Yin-Yang (YY)1.[11]

Misregulation of ncRNAs is common in cancer although recurrent structural variations have been challenging to find. For example, in a study with whole-genome sequencing of 150 tumor/normal pairs of chronic lymphocytic leukemia, only one recurrent non-coding mutation cluster was found at a potential regulatory element.[12] However, this may also reflect the lacking annotations. We recently analyzed whole genome sequencing data from precursor B-cell acute lymphoblastic leukemia (pre-B-ALL) in the context of chromatin architecture and found that the topologically associated domains with the highest number of breakpoints contained unannotated ncRNAs.[13] Functional studies manipulating lncRNA production in leukemia have shown diverse roles in cancer-related pathways.[14-16] In addition, functional studies on enhancers have highlighted their overall role in cancer, as reviewed in ref. 17. In leukemia, somatic mutation of a non-coding element generated a MYB binding site upstream of oncogenic TAL1 locus, and a deletion of the mutated (but not wild type allele) super-enhancer in a T-ALL cell line decreased expression of TAL1 and impaired cell survival.[18] Altered transcription at enhancers may also result from structural or quantitative changes in both enhancer elements and their regulating proteins. Duplication of NOTCH1-driven MYC enhancer was observed in T-ALL and its relevance demonstrated in a mouse knockout model.[19] Moreover, aberrations in chromatin structure and especially in insulator regions induce abnormal gene expression, as exemplified by activation of TAL1 due to a deletion of upstream insulator element.[20] Misregulated transcription during delicate differentiation processes in haematopoietic precursors may also cause cancer by predisposing to secondary mutations. Convergent transcription and RNA polymerase II stalling strongly correlate with structural variation clusters and seem to provide vulnerable regions for RAG and AID mediated double strand breaks in lymphoma and leukemia.[13,21] Although ncRNA expression profiles using microarray or RNA-seq have been published (eg. refs. 22-26), many nascent transcripts have remained unnoticed because of rapid degradation of several ncRNA species. New methods to address this challenge have emerged, such as GRO-seq, PRO-seq or TT-seq that enable monitoring various nascent transcripts and engaged RNA polymerase II in leukemia.[27,28]

CONTACT Susanna Teppo susanna.teppo@uta.fi Tampere Center for Child Health Research, Faculty of Medicine and Life Sciences, University of Tampere, Laakarinkatu 1, Arvo, Tampere 33520, Finland.

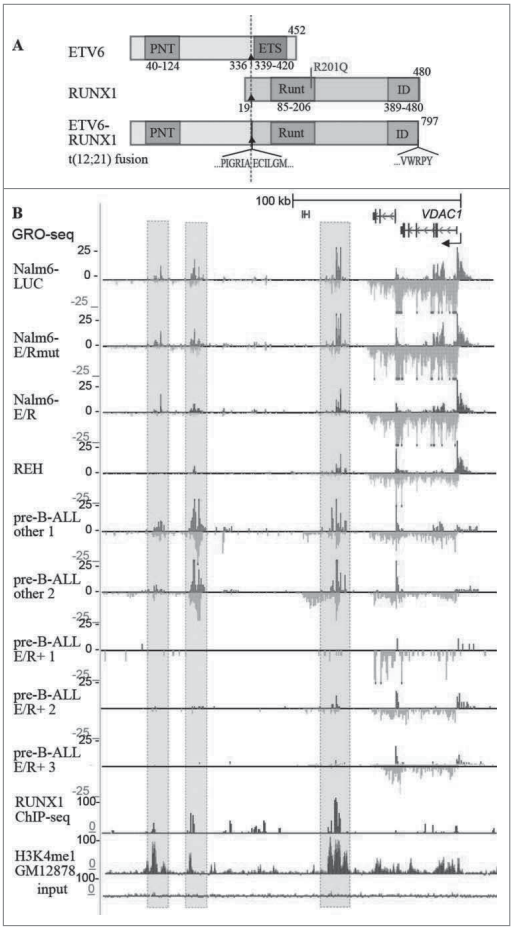

Figure 1. (A) A schematic representation of the ETV6-RUNX1 (E/R, TEL-AML1) fusion protein resulting from a recurrent t(12;21) translocation in pediatric pre-B acute lymphoblastic leukemia. ETV6-RUNX1 includes the pointed (PNT) domain of ETS variant 6 (ETV6) but lacks the ETS domain that is involved in DNA binding of the normal TF protein. The 480 aa long RUNX1 variant 1 (AML-1c, NP_001745) is illustrated with the point mutation R201Q in the Runt domain which impedes its DNA binding capability (this was used to generate E/Rmut in ref. 30). ID = Runx inhibitory domain. (B) GRO-seq signal (nascent RNA transcription) is shown for E/R-negative as red and E/R-positive samples as blue tracks at an example genomic region. Signals above and below the axis indicate plus and minus strands, respectively. RUNX1 ChIP peaks in SEM cells (GSE42075, ref. 42) and an enhancer marker H3K4me1 ChIP-seq in B-cells (GM12878, ref. 2) are shown and coincide with the GRO-seq signal. Three enhancer regions that are downregulated by E/R via RUNX1-mediated binding are highlighted. Nalm6-E/R = 24h expression of E/R in a pre-B-ALL cell line; REH = E/R-positive cell line; pre-B-ALL other = E/R-negative patient; pre-B-ALL E/R+ = E/R-positive patient.

We addressed this issue in the ETV6-RUNX1 (E/R, TEL-AML1) fusion positive leukemia,[30] which represents 25 % of pediatric acute lymphoblastic leukemias, and causes alterations in gene expression that predispose to leukemia.[29] With the help of an inducible E/R cell model and GRO-seq, we explored dynamics of gene expression and the activity of their regulatory elements simultaneously, exposing the transcriptional circuitry downstream of the E/R fusion (Fig. 1).[30] We analyzed enhancers based on eRNA correlation with GRO-seq signal change at differentially expressed genes (transcript-centric approach). Secondly, we generated an enhancer-centric approach that directly applied the statistical framework on eRNA levels to identify significantly regulated enhancers (enhancer annotation was based on H3K27ac and RUNX1 ChIP-seq data) and correlated these changes to that of nearby transcripts. We found at least one similarly altered putative enhancer element within +/−400 kb for almost all the deregulated coding transcripts using transcript-centric approach. E/R regulated approximately 20% of transcribed regions with RUNX1 ChIP peaks, and 5% of CD19/20 (B-cell)-related enhancers. Interestingly, CD19/20 specific super-enhancers were mostly downregulated, implying a way for E/R to arrest cell differentiation.

It has been proposed that any transcription may possess regulatory activity. A recent study showed that half of the studied transcribed gene loci (12 lncRNA and 6 mRNA) regulated a nearby gene in *cis* independently of whether the locus was a coding or non-coding one.[31] As the non-coding genome is only weakly conserved,[1,32] most non-coding regions may function in a way which is not dependent on the sequence of transcript itself but rather the sequence of its promoter or its location in the genome. In the case of E/R leukemia, we classified 57 deregulated novel lncRNAs (over 5 kb long) as either potential eRNAs or lncRNAs based on the GRO-seq signal. One fourth of the novel and 3 of 7 annotated transcripts were concordantly differentially expressed in RNA-seq data with 8 E/R-positive and 9 other subtype pre-B-ALL patients.[30] For example, KCNQ1OT1, which acts in epigenetic regulation,[33-35] was upregulated in our E/R cell model GRO-seq and the patient RNA-seq data. Signal changes at ZEB1 and ZEB1-AS1 serve as an example of a simultaneous downregulation of gene and its promoter-associated RNA, with ZEB also being linked to cancer[36,37] and late B cell differentiation.[38] Functional roles of the novel transcripts in E/R leukemia remains to be explored in future. Nascent RNA profiles of diagnostic patient samples of distinct ALL subtypes will give further insights into the derailed transcriptional network downstream of the oncogenic TF fusions.

Already, thousands of regulatory lncRNA transcripts[39] and hundreds of thousands of enhancer regions have been found. It is now known that ncRNAs are widely specific to a certain cell type and developmental stage. For example, most lncRNAs that are expressed at various stages of mouse B cell development are not expressed in a closely related T-cell lineage.[40] A recent study noted that distal regulatory elements varied across distinct haematopoietic lineages so that they are better discriminators of cell identity than mRNA levels.[41] This was also reflected in our work, where we noticed that sample separation based on quantification of global eRNA transcription was equally good as that based on quantification of transcription at protein coding regions.[30] We can assume that the increasing knowledge of the interplay between various elements of genome and their transcriptional products will significantly contribute to our understanding of the diverse types of leukemia and cancer in near future.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

ORCID

Susanna Teppo http://orcid.org/0000-0003-2569-8030
Merja Heinäniemi http://orcid.org/0000-0001-6190-3439
Olli Lohi http://orcid.org/0000-0001-9195-0797

References

1. ENCODE Project Consortium TEP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799-816; PMID:17571346; https://doi.org/10.1038/nature05874
2. ENCODE Project Consortium TEP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012; 489:57-74; PMID:22955616; https://doi.org/10.1038/nature11247
3. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 2013; 153:307-19; PMID:23582322; https://doi.org/10.1016/j.cell.2013.03.035
4. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. Cell 2013; 155:934-47; PMID:24119843; https://doi.org/10.1016/j.cell.2013.09.053
5. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 2008; 322:1845-8; PMID:19056941; https://doi.org/10.1126/science.1162228
6. Lam MTY, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci 2014; 39:170-82; PMID:24674738; https://doi.org/10.1016/j.tibs.2014.02.007
7. Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Mol Cell 2013; 51:310-25; PMID:23932714; https://doi.org/10.1016/j.molcel.2013.07.010
8. Kim T-K, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 2010; 465:182-7; PMID:20393465; https://doi.org/10.1038/nature09033
9. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 2011; 474:390-4; PMID:21572438; https://doi.org/10.1038/nature10006
10. Takemata N, Ohta K. Role of non-coding RNA transcription around gene regulatory elements in transcription factor recruitment. RNA Biol 2017; 14:1-5; PMID:27763805; https://doi.org/10.1080/15476286.2016.1248020
11. Sigova AA, Abraham BJ, Ji X, Molinie B, Hannett NM, Guo YE, Jangi M, Giallourakis CC, Sharp PA, Young RA. Transcription factor trapping by RNA in gene regulatory elements. Science (80−) 2015; 350:978-81; PMID:26516199; https://doi.org/10.1126/science.aad3346
12. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, Munar M, Rubio-Pérez C, Jares P, Aymerich M, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 2015; 526:519-24; PMID:26200345; https://doi.org/10.1038/nature14666
13. Heinäniemi M, Vuorenmaa T, Teppo S, Kaikkonen MU, Bouvy-Liivrand M, Mehtonen J, Niskanen H, Zachariadis V, Laukkanen S, Liuksiala T, et al. Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots. Elife 2016; 5:e12068; PMID:26896675; https://doi.org/10.7554/eLife.13087
14. Blume CJ, Hotz-Wagenblatt A, Hüllein J, Sellner L, Jethwa A, Stolz T, Slabicki M, Lee K, Sharathchandra A, Benner A, et al. p53-dependent non-coding RNA networks in chronic lymphocytic leukemia. Leukemia 2015; 29:2015-23; PMID:25971364; https://doi.org/10.1038/leu.2015.119
15. Guo G, Kang Q, Zhu X, Chen Q, Wang X, Chen Y, Ouyang J, Zhang L, Tan H, Chen R, et al. A long noncoding RNA critically regulates Bcr-Abl-mediated cellular transformation by acting as a competitive endogenous RNA. Oncogene 2015; 34:1768-79; PMID:24837367; https://doi.org/10.1038/onc.2014.131
16. Lu Y, Li Y, Chai X, Kang Q, Zhao P, Xiong J, Wang J. Long non-coding RNA HULC promotes cell proliferation by regulating PI3K/AKT signaling pathway in chronic myeloid leukemia. Gene 2017; 607:41-6; PMID:28069548; https://doi.org/10.1016/j.gene.2017.01.004
17. Sur I, Taipale J. The role of enhancers in cancer. Nat Rev Cancer 2016; 16:483-93; PMID:27364481; https://doi.org/10.1038/nrc.2016.62
18. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science (80-) 2014; 346:1373-7; PMID:25394790; https://doi.org/10.1126/science.1259037
19. Herranz D, Ambesi-Impiombato A, Palomero T, Schnell SA, Belver L, Wendorff AA, Xu L, Castillo-Martin M, Llobet-Navás D, Cordon-Cardo C, et al. A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. Nat Med 2014; 20:1130-7; PMID:25194570; https://doi.org/10.1038/nm.3665
20. Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science (80-) 2016; 351:1454-8; PMID:26940867; https://doi.org/10.1126/science.aad9024
21. Meng F-L, Du Z, Federation A, Hu J, Wang Q, Kieffer-Kwon K-R, Meyers RM, Amor C, Wasserman CR, Neuberg D, et al. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell 2014; 159:1538-48; PMID:25483776; https://doi.org/10.1016/j.cell.2014.11.014
22. Fernando TR, Rodriguez-Malave NI, Waters EV, Yan W, Casero D, Basso G, Pigazzi M, Rao DS. LncRNA expression discriminates karyotype and predicts survival in B-lymphoblastic leukemia. Mol Cancer Res 2015; 13:839-51; PMID:25681502; https://doi.org/10.1158/1541-7786.MCR-15-0006-T
23. Nordlund J, Kiialainen A, Karlberg O, Berglund EC, Göransson-Kultima H, Sønderkær M, Nielsen KL, Gustafsson MG, Behrendtz M, Forestier E, et al. Digital gene expression profiling of primary acute lymphoblastic leukemia cells. Leukemia 2012; 26:1218-27; PMID:22173241; https://doi.org/10.1038/leu.2011.358
24. Ghazavi F, De Moerloose B, Van Loocke W, Wallaert A, Helsmoortel HH, Ferster A, Bakkus M, Plat G, Delabesse E, Uyttebroeck A, et al. Unique long non-coding RNA expression signature in ETV6/RUNX1-driven B-cell precursor acute lymphoblastic leukemia. Oncotarget 2016; 7:73769-80; PMID:27650541; https://doi.org/10.18632/oncotarget.12063
25. Teittinen KJ, Laiho A, Uusimäki A, Pursiheimo J-P, Gyenesei A, Lohi O. Expression of small nucleolar RNAs in leukemic cells. Cell Oncol 2013; 36:55-63; PMID:23229394; https://doi.org/10.1007/s13402-012-0113-5
26. Ronchetti D, Manzoni M, Agnelli L, Vinci C, Fabris S, Cutrona G, Matis S, Colombo M, Galletti S, Taiana E, et al. lncRNA profiling in early-stage chronic lymphocytic leukemia identifies transcriptional fingerprints with relevance in clinical outcome. Blood Cancer J 2016; 6:e468; PMID:27611921; https://doi.org/10.1038/bcj.2016.77
27. Zhao Y, Liu Q, Acharya P, Stengel KR, Sheng Q, Zhou X, Kwak H, Fischer MA, Bradner JE, Strickland SA, et al. High-resolution mapping of RNA polymerases identifies mechanisms of sensitivity and

resistance to BET inhibitors in t(8;21) AML. Cell Rep 2016; 16:2003-16; PMID:27498870; https://doi.org/10.1016/j.celrep.2016.07.032
28. Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P. TT-seq maps the human transient transcriptome. Science (80-) 2016; 352:1225-8; PMID:27257258; https://doi.org/10.1126/science.aad9841
29. Inaba H, Greaves M, Mullighan CG. Acute lymphoblastic leukaemia. Lancet 2013; 381:1943-55; PMID:23523389; https://doi.org/10.1016/S0140-6736(12)62187-4
30. Teppo S, Laukkanen S, Liuksiala T, Nordlund J, Oittinen M, Teittinen K, Grönroos T, St-Onge P, Sinnett D, Syvänen A-C, et al. Genome-wide repression of eRNA and target gene loci by the ETV6-RUNX1 fusion in acute leukemia. Genome Res 2016; 26:1468-77; PMID:27620872; https://doi.org/10.1101/gr.193649.115
31. Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature 2016; 539:452-5; PMID:27783602; https://doi.org/10.1038/nature20149
32. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 2007; 17:556-65; PMID:17387145; https://doi.org/10.1101/gr.6036807
33. Sunamura N, Ohira T, Kataoka M, Inaoka D, Tanabe H, Nakayama Y, Oshimura M, Kugoh H. Regulation of functional KCNQ1OT1 lncRNA by β-catenin. Sci Rep 2016; 6:20690; PMID:26868975; https://doi.org/10.1038/srep20690
34. Ribarska T, Goering W, Droop J, Bastian K-M, Ingenwerth M, Schulz WA. Deregulation of an imprinted gene network in prostate cancer. Epigenetics 2014; 9:704-17; PMID:24513574; https://doi.org/10.4161/epi.28006
35. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, Komorowski J, Nagano T, Mancini-DiNardo D, Kanduri C. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 2008; 32:232-46; PMID:18951091; https://doi.org/10.1016/j.molcel.2008.08.022
36. Lehmann W, Mossmann D, Kleemann J, Mock K, Meisinger C, Brummer T, Herr R, Brabletz S, Stemmler MP, Brabletz T. ZEB1 turns into a transcriptional activator by interacting with YAP1 in aggressive cancer types. Nat Commun 2016; 7:10498; PMID:26876920; https://doi.org/10.1038/ncomms10498
37. Zhang P, Sun Y, Ma L. ZEB1: At the crossroads of epithelial-mesenchymal transition, metastasis and therapy resistance. Cell Cycle 2015; 14:481-7; PMID:25607528; https://doi.org/10.1080/15384101.2015.1006048
38. Malpeli G, Barbi S, Zupo S, Tosadori G, Scardoni G, Bertolaso A, Sartoris S, Ugel S, Vicentini C, Fassan M, et al. Identification of microRNAs implicated in the late differentiation stages of normal B cells suggests a central role for miRNA targets ZEB1 and TP53. Oncotarget 2017; 8:11809–26; PMID:28107180; https://doi.org/10.18632/oncotarget.14683
39. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature 2017; 543:199-204; PMID:28241135; https://doi.org/10.1038/nature21374
40. Brazão TF, Johnson JS, Müller J, Heger A, Ponting CP, Tybulewicz VLJ. Long noncoding RNAs in B-cell development and activation. Blood 2016; 128:e10-9; PMID:27381906; https://doi.org/10.1182/blood-2015-11-680843
41. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet 2016; 48:1193-203; PMID:27526324; https://doi.org/10.1038/ng.3646
42. Wilkinson AC, Ballabio E, Geng H, North P, Tapia M, Kerry J, Biswas D, Roeder RG, Allis CD, Melnick A, et al. RUNX1 is a key target in t(4;11) leukemias that contributes to gene activation through an AF4-MLL complex interaction. Cell Rep 2013; 3(1):116-27; PMID:23352661; https://doi.org/10.1016/j.celrep.2012.12.016

PUBLICATION III

Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots

Heinäniemi, M., Vuorenmaa, T., Teppo, S., Kaikkonen, M. U., Bouvy-Liivrand, M., Mehtonen, J., Niskanen, H., Zachariadis, V., Laukkanen, S., Liuksiala, T., Teittinen, K., & Lohi, O.

eLife 2016, 5: e13087
doi: 10.7554/eLife.13087

eLIFE elifesciences.org

RESEARCH ARTICLE

Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots

Merja Heinäniemi[1]*, Tapio Vuorenmaa[1,2†], Susanna Teppo[3†], Minna U Kaikkonen[2†], Maria Bouvy-Liivrand[1], Juha Mehtonen[1], Henri Niskanen[2], Vasilios Zachariadis[4], Saara Laukkanen[3], Thomas Liuksiala[3], Kaisa Teittinen[3], Olli Lohi[3,5]*

[1]School of Medicine, University of Eastern Finland, Kuopio, Finland; [2]A. I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland; [3]School of Medicine, University of Tampere, Tampere, Finland; [4]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden; [5]Tampere University Hospital, Tampere, Finland

Abstract Progression of malignancy to overt disease requires multiple genetic hits. Activation-induced deaminase (AID) can drive lymphomagenesis by generating off-target DNA breaks at loci that harbor highly active enhancers and display convergent transcription. The first active transcriptional profiles from acute lymphoblastic leukemia (ALL) patients acquired here reveal striking similarity at structural variation (SV) sites. Specific transcriptional features, namely convergent transcription and Pol2 stalling, were detected at breakpoints. The overlap was most prominent at SV with recognition motifs for the recombination activating genes (RAG). We present signal feature analysis to detect vulnerable regions and quantified from human cells how convergent transcription contributes to R-loop generation and RNA polymerase stalling. Wide stalling regions were characterized by high DNAse hypersensitivity and unusually broad H3K4me3 signal. Based on 1382 pre-B-ALL patients, the ETV6-RUNX1 fusion positive patients had over ten-fold elevation in RAG1 while high expression of AID marked pre-B-ALL lacking common cytogenetic changes.
DOI: 10.7554/eLife.13087.001

***For correspondence:** merja.heinaniemi@uef.fi (MH); olli.lohi@staff.uta.fi (OL)

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 21

Received: 16 November 2015
Accepted: 09 June 2016
Published: 19 July 2016

Reviewing editor: Scott A Armstrong, Memorial Sloan Kettering Cancer Center, United States

Introduction

In precursor lymphoblastic leukemia, primary genetic lesions often arise *in utero* (***Wiemels et al., 1999***; ***Mori et al., 2002***; ***Maia et al., 2003***, ***Bateman et al., 2015***), while the onset of overt disease requires additional genetic alterations. Whole-genome sequencing (WGS) of ETV6-RUNX1 (also known as TEL-AML1) positive acute leukemias suggested that the secondary lesions are predominantly caused by off-target activity of the RAG complex (***Papaemmanuil et al., 2014***). In a similar fashion, the expression of the AID complex in more mature B cells is implicated in genomic instability and development of lymphomas (***Meng et al., 2014***; ***Qian et al., 2014***; ***Robbiani et al. 2015***). To date, WGS in leukemia have been reported from several pre-B-ALL subtypes (***Andersson et al., 2015***; ***Holmfeld et al., 2013***; ***Paulsson et al., 2015***; ***Zhang et al., 2012***), resulting in a comprehensive characterization of the underlying genetic alterations. Therefore, the research focus on leukemia genetics is moving into characterization of the mechanisms by which these lesions occur and the consequences of the resulting clonal heterogeneity.

Antigen receptor genes are assembled from discrete gene segments by RAG-mediated V(D)J recombination at sites of recombination signal sequences (RSS) during early lymphocyte

eLife digest Some of the most common cancers found in children are called precursor leukemias, which may start to develop before birth. Cancerous cells often contain alterations to the genetic information in their DNA. In precursor leukemias, the most common genetic changes involve deleting, adding or rearranging segments of the DNA sequence.

Several researchers have sequenced the entire DNA of childhood leukemia cells, with the result that almost all of the genetic alterations linked to these conditions have been catalogued. These efforts have shown that certain DNA regions are particularly affected by mutations, but no one knows why errors occur so frequently in these regions.

Recent evidence also suggests that transcription – the process of producing useful molecules from a stretch of DNA – can play a role in generating genetic alterations. Heinäniemi et al. have now used a technique called global run-on sequencing to measure the extent of transcription in many different types of leukemia cells. This revealed that in the error-prone DNA regions, two processes – called convergent transcription and transcriptional stalling – interfere with transcription. Both processes temporarily leave the normally double-stranded DNA unzipped as two single strands and free of nucleosomes, which makes DNA more vulnerable to breaking. This would explain how pieces of DNA might be lost, added, or moved to cause the genetic errors that lead to leukemia.

Further investigation revealed that two protein complexes called RAG and AID, which rearrange segments of DNA in immune cells, are likely to cause the errors in the vulnerable DNA regions. Different amounts of RAG and AID were present in different subtypes of leukemia cells, and these amounts also varied with the risk classification of the disease. Further studies are now needed to investigate the exact roles of these protein complexes. This could eventually help scientists devise strategies to protect the DNA of people with leukemia from these errors, which could reduce the risk of the cancer reoccurring.

DOI: 10.7554/eLife.13087.002

development (***Gellert 2002***; ***Schatz and Swanson, 2011***). Cells incorporate multiple strategies to control the action of the RAG complex to appropriate genomic loci: the expression of *RAG1* and *RAG2* is limited to precursor stages of lymphocytes, the activity of the complex is attenuated during S-phase of cell cycle, and RAG cleavage is directed towards RSS pair containing sequences (***Schatz and Swanson, 2011***). The engagement of RAG2 is further limited by the histone modification H3K4me3, which is typically found at transcription start sites (TSS) (***Matthews et al., 2007***; ***Teng et al., 2015***). However, RSS and RSS-like motifs are found only at around 7–40% of breakpoints at SV (genomic imbalance, translocation or inversion) sites (***Andersson et al., 2015***; ***Papaemmanuil et al., 2014***). Furthermore, the RSS motifs and H3K4me3 occur frequently in the genome suggesting that additional features, possibly even additional complexes including AID (***Swaminathan et al., 2015***), are relevant for the genetic instability underlying leukemia SV.

In lymphomas, AID off-target effects localize to intragenic super-enhancer (SE) and promoter areas characterized by transcription from both strands, i.e. convergent transcription (convT) (***Meng et al., 2014***). Notably, VH gene segment recombination by RAG at the IgH locus coincides with sense- and antisense transcription (***Bolland et al., 2004***), which could be relevant also at off-target sites. Secondly, stalled polymerases, which are found at exons, R-loops and actively paused at TSS regions (***Jonkers and Lis, 2015***), expose single stranded DNA, recruiting AID via Spt5 binding (***Pavri et al., 2010***). Furthermore, the polymerase complex displaces nucleosomes completely or partially (the H2A/H2B moiety), which in vitro promotes cleavage by RAGs (***Bevington and Boyes, 2013***). Despite these intriguing findings, the relevance of transcription-coupled processes has not been systematically characterized, and the clinical relevance of RAG and AID expression in the different leukemia subtypes remains unclear. RNA polymerases engaged into primary transcription across the genome can be measured using Global-Run-On sequencing (GRO-seq) (***Kaikkonen et al., 2013***). Therefore, this method is ideally suited to distinguish features of transcription at SV sites, including convT and RNA polymerase stalling. To this end, we acquired the first patient profiles of nascent transcriptional activity in leukemic blasts representing seven cytogenetic subgroups and performed integrative analysis of various genome-wide profiles and patient transcriptomes.

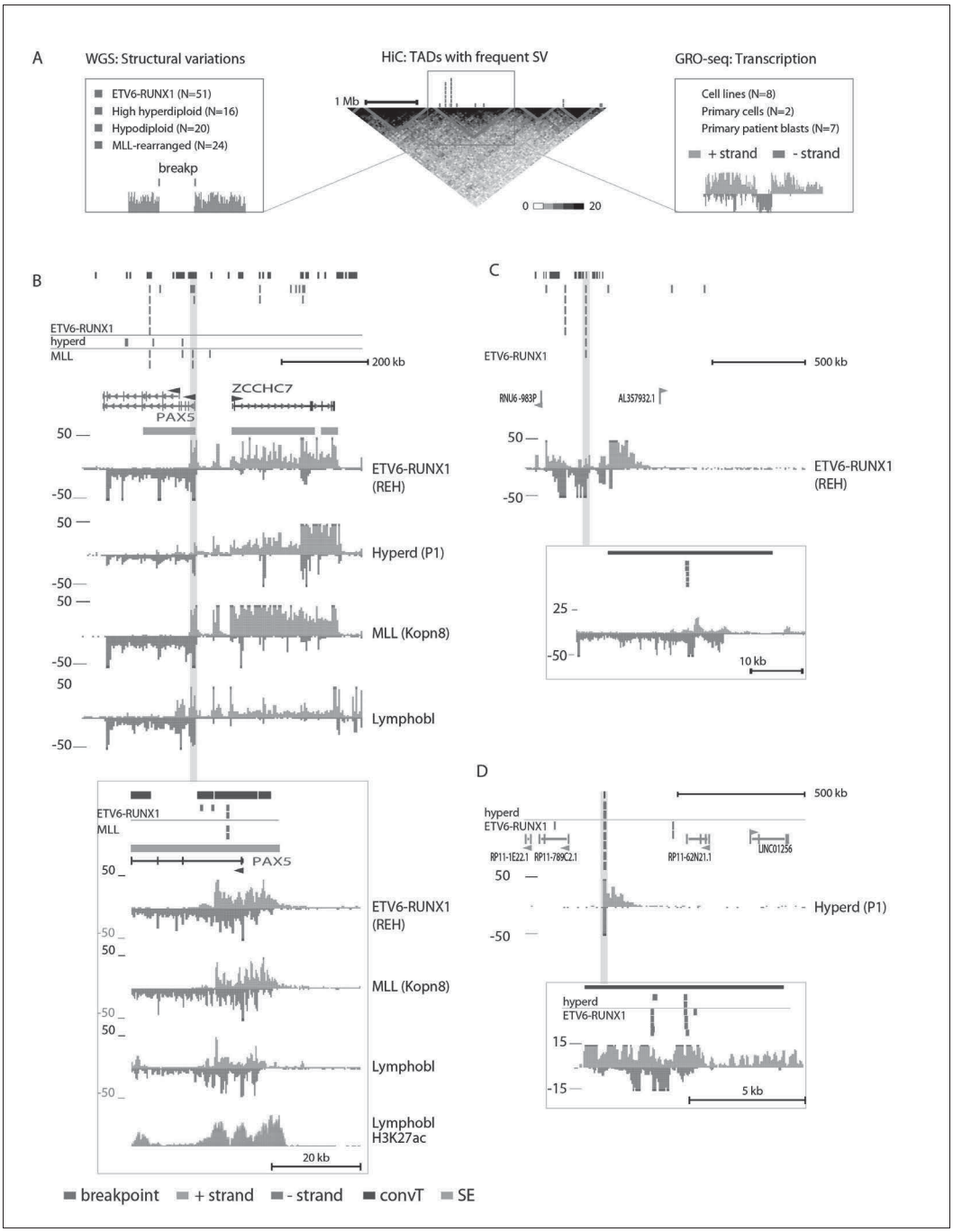

Figure 1. Integrative analysis of transcription and high-recurrence SV sites highlights novel transcribed regions. (**A**) WGS data from the ETV6-RUNX1 (51 cases; ***Papaemmanuil et al., 2014***), high hyperdiploid (16 cases; ***Paulsson et al., 2015***), hypodiploid (20 cases; ***Holmfeldt et al., 2013***) and MLL-

Figure 1 continued on next page

Figure 1 continued

rearranged (22 cases; ***Andersson et al., 2015***) subtypes of precursor B-ALL was integrated with profiles of transcriptional activity assayed using GRO-seq from ALL patient and cell line samples (see also ***Figure 1—figure supplement 1*** and ***Supplementary file 1***). HiC data from B-lymphoid cells (***Rao et al., 2014***) was used to define TADs based on the HiC interaction frequency, shown as grey scale heatmap, in order to distinguish TADs with highest frequency of SV. (**B**) The *PAX5* and *ZCCHC7* loci are located in the TAD shown that has high SV frequency in hyperdiploid, ETV6-RUNX1- and MLL-fusion positive patients (4, 20 and 6 breakpoints, respectively, ***Figure 1—source data 1***). The GRO-seq signal profiles from three pre-B-ALL cytogenetic subtypes and normal B-lymphoblastoid cells are displayed as indicated in the figure (see also ***Figure 1—figure supplement 4*** and ***Figure 2—figure supplement 2***). The y-axis shows the normalized read density (plus strand in red, minus strand in blue). convT regions regions are indicated in purple and leukemia breakpoints in red. The TSS region of *PAX5* overlaps convT that co-localized with an intragenic SE (B-lymphoblastoid H3K27ac track is shown at the bottom). (**C**) A TAD with the same number of breakpoints (20) in ETV6-RUNX1 patients is shown with signal from REH cells (see also ***Figure 1—figure supplement 4***). Genomic annotations include the location of GENCODE transcripts (in green). A strong transcription signal is visible that spans approximately 500 kb near the TAD boundary, lacking annotated transcripts. A zoom-in panel shows the most recurrent SV site. (**D**) The TAD visualized represents a genomic region that harbors most SV in HeH (see ***Figure 1—figure supplement 5*** for the hypodiploid SV hotspot). The GRO-seq signal (track from patient 1) indicates a novel locus with abundant transcription in leukemic samples (refer to ***Figure 1—figure supplement 4*** for all GRO-seq profiles). The highest recurrence of SV occurs at the convT overlapping mid-region (zoom-in panel), which has also two ETV6-RUNX1 breakpoints.

DOI: 10.7554/eLife.13087.003

The following source data and figure supplements are available for figure 1:

Source data 1. Identified topologically associated domains.

DOI: 10.7554/eLife.13087.004

Figure supplement 1. Transcriptional activity in leukemic cells from patients, cell lines and primary healthy B-lineage cells is captured in GRO-seq signals.

DOI: 10.7554/eLife.13087.005

Figure supplement 2. Summary of data used in the integrative analysis.

DOI: 10.7554/eLife.13087.006

Figure supplement 3. Transcriptional activity in TADs binned by breakpoint frequency.

DOI: 10.7554/eLife.13087.007

Figure supplement 4. Data from all signal tracks for regions displayed in ***Figure 1***.

DOI: 10.7554/eLife.13087.008

Figure supplement 5. TAD with frequent SV in hypodiploid patients.

DOI: 10.7554/eLife.13087.009

Results

Integrative analysis of transcription and genomic instability in leukemic cells

Transcriptional activity from ALL cells representing seven different pre-B-ALL cytogenetic subtypes was assayed using GRO-seq (both primary patient and cell line samples, see ***Supplementary file 1*** and Materials and methods), and jointly analyzed with WGS data from the ETV6-RUNX1 (51 cases; ***Papaemmanuil et al., 2014***), high hyperdiploid (HeH, 16 cases; ***Paulsson et al., 2015***), hypodiploid (20 cases; ***Holmfeldt et al., 2013***) and MLL-rearranged (22 cases at diagnosis and 2 relapses; ***Andersson et al., 2015***) subtypes of precursor B-ALL. GRO-seq signals and breakpoint data are shown in ***Figure 1—figure supplement 1*** at the *CDKN2A* locus, a significant SV site in childhood ALL (***Sulong et al., 2009***).

To systematically identify regions with high frequency of SV across the genome, topologically-associated domains (TADs) were retrieved based on HiC data from B-lymphoid lineage cells (***Rao et al., 2014***). TADs reflect the three-dimensional structure of chromatin. These natural boundaries to transcriptional activity were used to divide the chromosomes into subregions for analysis (see ***Figure 1—source data 1*** and Materials and methods). To link typical transcriptional activity patterns and hotspots of genomic instability, we related the breakpoint frequency with chromatin domains, as illustrated in ***Figure 1A*** (see also ***Figure 1—figure supplement 2***).

The most frequent SV regions encompass novel transcribed regions

An increasing trend of transcriptional activity was observed when TADs were compared based on breakpoint frequency quartiles (see Materials and methods, ***Figure 1—figure supplement 3***). TADs with highest SV count are shown in ***Figure 1*** (see also ***Figure 1—source data 1*** and ***Figure 1—***

figure supplement 4). The *PAX5* and *ZCCHC7* genes are located within a TAD region with 20 breakpoints in the ETV6-RUNX1, 4 in HeH and 6 in MLL subtype (excluding the MLL-fusion itself) (*Figure 1B*). Frequent SV were also found in TADs with no annotated coding genes (*Figure 1C*, 20 breakp in ETV6-RUNX1; *Figure 1D*, 4 breakp in HeH), yet GRO-seq exhibited transcription signal spanning several hundred thousand base pairs in both regions, typical of long non-coding transcripts (*Sun et al., 2015*). There was evidence of non-coding transcripts, based on Refseq and GENCODE, but none matched the same location (refer to *Supplementary file 2* for all genomic coordinates shown; a TAD with frequent SV in hypodiploid subtype is shown in *Figure 1—figure supplement 5*). The nascent ALL transcriptomes thus reveal novel transcribed regions as recurrent SV-associated hotspots in the two most common ALL subtypes.

Convergent transcription and RNA polymerase stalling are prevalent at genomic regions with frequent breakpoint events

The prevailing notion is that active transcription start sites (TSS) in pre-B cells are susceptible to RAG off-targeting due to the H3K4me3 chromatin mark (*Matthews et al., 2007*; *Teng et al., 2015*). However, we noticed that the recurrent breakpoints often lied several kb downstream of TSS, as highlighted in *Figure 1B and D* (see inserts), and coincided with simultaneous transcription on both strands, ie. convT spanning a minimum of 100 bp. In closer examination of the signal data from leukemia SV hotspots, many of these regions likely correspond to transcription from intragenic enhancers that generate enhancer RNAs (eRNA) that are typically a few kb in size (*Kaikkonen et al. 2013*). In agreement, a significant enrichment of breakpoints in enhancers overlapping with convT was observed (hypergeometric test P=0.00012 for intergenic and P=4.6e-08 for all enhancers identified based on eRNA signal, see Materials and methods and *Figure 2—source data 1*). An overlapping eRNA transcript at the TSS region of *PAX5*, confirmed by the active enhancer chromatin marker H3K27ac, led to convT extending nearly 20 kb, with SV sites located between 3.7–9.7 kb downstream of the TSS (*Figure 1B*, see insert).

Secondly, convT in the vicinity of intragenic breakpoints was often associated with localized elevation in the GRO-seq signal, as exemplified at the *ZCCHC7* and *RAG* loci (*Figure 2A*, see also *Figure 2—figure supplement 1*). The observed signal features were highly reproducible between biological replicates and shared among a subset of cytogenetic groups (*Figure 2—figure supplement 2*). We hypothesized that they represent RNA polymerase II (Pol2) stalling events. Previous analyses of Pol2 stalling have focused on promoter proximal regions (*Adelman and Lis, 2012*). To examine such events genome-wide and across gene bodies, we developed a general analysis approach that identifies change points within gene regions and reports those with high elevation in the signal level (see Materials and methods and *Figure 2—source data 1* for the identified regions) (*Killick et al., 2012*). As additional confirmation, we analyzed stalling from Pol2 ChIP-seq in the REH and Nalm6 cell lines (*Figure 2A*). To distinguish between different Pol2 complexes (*Zhou et al., 2012*), antibodies against the serine 2 or serine 5 phosphorylated Pol2 were used (see Materials and methods).

Genome-wide analysis of convT and Pol2 stalling (see Materials and methods and *Figure 2—source data 1*) substantiated the relevance of these observations: considering the breakpoint frequency per TAD size, the top ranked TADs in each ALL subtype represented genomic regions with abundant convT and Pol2 stalling (*Figure 2B*). Significant enrichment was confirmed for the upper quartiles (hypergeometric test P=0.00038 in ETV6-RUNX1, P=0.00018 in hyperdiploid, P=0.028 in hypodiploid and P=0.00004 in MLL-rearranged). The increased overlap was found for breakpoints with and without RSS motifs (denoted as R-breakp and NR-breakp, see *Figure 2—figure supplement 3* and Materials and methods) and it was preserved when total transcriptional activity was considered (*Figure 2—figure supplement 4*). Furthermore, the distinct transcriptional profile of embryonic stem cells (ES) had lower overlap (*Figure 2—figure supplement 5*).

For comparison, chromatin segmentation of B-lymphoid cells was similarly analyzed (see *Figure 1—source data 1* and *Figure 2—source data 1*). TADs with high number of breakpoints consistently had significant overlap with chromatin segments representing active transcription (refer to *Figure 2—source data 1*), supporting a transcription-coupled mechanism for the observed genetic instability. We then distinguished regions with overlap to the transcriptional features defined here within active promoters and enhancers. Comparing these against the TAD SV frequency quartiles

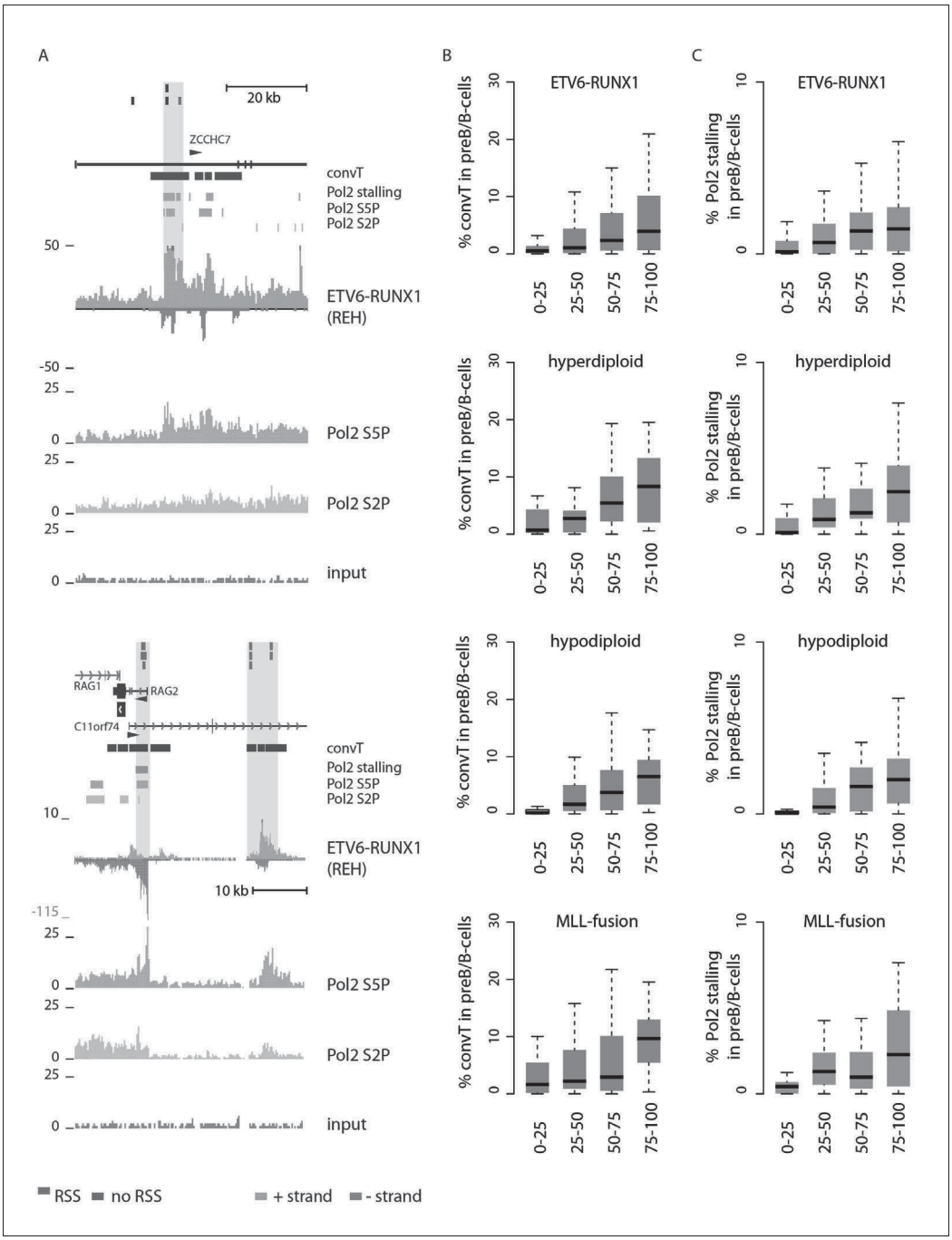

Figure 2. Convergent transcription and Pol2 stalling characterize genomic regions with high number of breakpoint events. (**A**) The GRO-seq signal in the ETV6-RUNX1 positive REH cell line is shown to exemplify the co-occurrence of convT (in purple) and local elevation in GRO-seq signal (Pol2 stalling,

Figure 2 continued on next page

Figure 2 continued

in light blue) at both R- and NR-breakp (in red and brown, respectively) that reside within intronic (*ZCCHC7*), TSS (*RAG2*) or putative enhancer regions (*RAG2*). The elevated signal is also visible in Pol2 ChIP-seq signal (Pol2 S2P in green, Pol2 S5P in orange, input in grey). See also ***Figure 2—figure supplement 1***. The percentage of TAD spanned by convT (in **B**) or Pol2 stalling (in **C**) in pre-B/B-lymphoid cells is summarized as boxplots from TADs divided into quartiles based on number of breakpoints per bp (see also ***Figure 1—figure supplement 3***, ***Figure 2—figure supplement 3–6***). The quartile ranges are for exclusive lower and inclusive upper value in the range, as indicated. Refer to ***Figure 2—source data 1*** for statistical analysis.

DOI: 10.7554/eLife.13087.010

The following source data and figure supplements are available for figure 2:

Source data 1. Identified convT and Pol2 stalling regions.

DOI: 10.7554/eLife.13087.011

Figure supplement 1. Data from all signal tracks for regions displayed in ***Figure 2***.

DOI: 10.7554/eLife.13087.012

Figure supplement 2. The GRO-seq signal from replicate samples generated from ALL cells displayed at the *PAX5/ZCCHC7* locus.

DOI: 10.7554/eLife.13087.013

Figure supplement 3. Signal feature span for TADs ordered separately by R-breakp or NR-breakp frequency.

DOI: 10.7554/eLife.13087.014

Figure supplement 4. Signal feature span normalized by total transcribed area for TADs sorted by breakpoint frequency.

DOI: 10.7554/eLife.13087.015

Figure supplement 5. Overlap of TADs with convT in ES cells.

DOI: 10.7554/eLife.13087.016

Figure supplement 6. TAD analysis using promoter and enhancer chromatin segments stratified by convT and Pol2 stalling.

DOI: 10.7554/eLife.13087.017

(***Figure 2—figure supplement 6***), as before, revealed the most pronounced enrichment in convT/Pol2 stall overlapping regions.

Next, we set out to define what may link convT and Pol2 stalling regions with AID and RAG recruitment. The signal feature detection for convT (as in ***Meng et al., 2014***) and Pol2 stalling (as defined here) enables this on a genome-wide level.

R-loop formation and convergent transcription co-occur with Pol2 stalling

RNA polymerases are expected to stall at regions harboring R-loop forming sequences (RLFS) (***Skourti-Stathaki et al., 2014a***; ***Jenjaroenpun et al., 2015***). The sensitivity of DNA sequence to form R-loops can be computationally predicted (***Jenjaroenpun et al., 2015***) (see Materials and methods). These RLFS motif containing regions exhibited a significantly higher overlap with Pol2 stalling sites when compared to random intragenic regions (***Figure 3B***, empirical P<0.001 in B-lineage and ES cells). A highly concordant local RLFS motif density and GRO-seq signal profile was observed across gene regions (***Figure 3—figure supplement 1A and B***). The profiles peaked near TSS, where the presence of RLFS motifs led to a significant elevation in the median GRO-seq signal level (***Figure 3—figure supplement 1***, 2.1-fold increase in B-lineage cells, Wilcoxon rank sum test P<2.2e-16, 95% CI 2.1–2.3). As a second mechanism, collisions due to convT may halt transcription (***Prescott and Proudfoot, 2002***) in a dynamic and cell-specific manner. Accordingly, higher antisense signal at convT regions (see Materials and methods) increased the overlap with Pol2 stalling sites on the sense strand (***Figure 3B***), intriguingly exceeding that observed for RLFS motifs (***Figure 3A***).

As an additional experimental validation of R-loops, we used DNA-RNA-immunoprecipitation sequencing (DRIP-seq) results from ES cells (see Materials and methods) that correspond to detection of DNA-RNA hybrids (***Ginno et al., 2013***). The 2.1-fold elevation in median DRIP-seq signal confirmed that RLFS motifs favor DNA-RNA hybrid formation (***Figure 3C***, Wilcoxon rank sum test P<2.2e-16, 95% CI 2.0–2.1, see ***Figure 3—source data 2*** for each replicate). Moreover, DRIP-seq quantification showed 1.7-fold higher median signal at convT-positive TSS regions (***Figure 3D***, Wilcoxon rank sum test P<2.2e-16, 95% CI 1.6–1.7). These results demonstrate that transcription stalling occurs at RLFS and convT regions in mammalian cells that associates with R-loop formation based on evidence from ES cells.

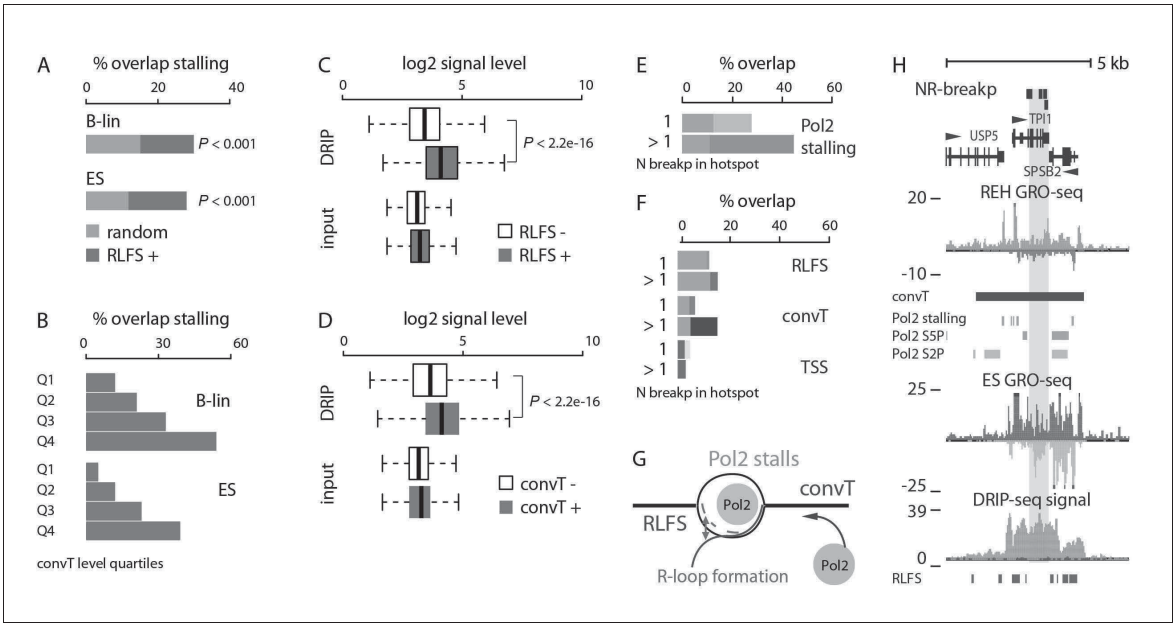

Figure 3. Indication of transcription-coupled genetic instability at leukemia SV hotspots lacking RSS motifs. (**A**) Overlap between RLFS motif harboring intragenic regions and detected Pol2 stalling sites in B-lineage and ES cells. The high overlap of RLFS-positive regions is statistically significant compared to random regions (empirical P is indicated for 30% and 28% overlaps, respectively). (**B**) Overlap of detected Pol2 stalling sites also increases based on the strength of antisense signal level for B-lineage and ES cell convT regions divided into quartiles. (**C**) The influence of RLFS at TSS on ES cell DRIP-seq signal level is shown (Wilcoxon rank sum test P is indicated). Input signal levels are shown as control. (**D**) ES cell DRIP-seq signal is plotted similarly as in **C**, from convT-positive and -negative TSS regions. The DRIP-signal is higher in convT-positive TSS (Wilcoxon rank sum test P is indicated, TSS with convT N = 11774, TSS without convT N = 12092, refer to ***Figure 3—source data 2*** for statistical analysis based on separate DRIP-seq replicates). (**E**) The percentages of breakpoint regions with no RSS motifs overlapping intragenic Pol2 stalling sites found in B-lineage cells are shown as barplots. The mean overlap observed in random sampling is indicated in grey bars (further statistical analysis is presented in ***Supplementary file 3***). Categories with increasing cut-off for recurrence (1: non-recurrent in dim color, >1 and above: recurrent in darker color) were tested. (**F**) Overlap with RLFS, convT and annotated TSS is shown, as in **E**, for ETV6-RUNX1 NR-breakp (see also ***Supplementary file 3***). (**G**) A schematic model illustrating how transcription from both strands (convT) or RLFS can locally arrest the Pol2 complex leading to recruitment of DNA damage-sensing complexes to R-loops, such as AID or BRCA (***Alt et al., 2013***, ***Hatchi et al., 2015***), in an RSS-independent manner. (**H**) NR-breakp hotspot with the highest recurrence (*TPI1* locus) is shown. DRIP-seq signal (shown in tones of red overlaid with input control signal in blue), and RLFS motifs indicated as a magenta bar track represent two levels of independent data that were integrated with GRO-seq data (signal from REH and ES cells is shown) to characterize properties of convT and Pol2 stalling regions. The breakpoint data (NR-breakp in brown) and detected convT (in purple) and Pol2 stalling in B-lineage cells (in blue) are shown. At the the recurrent breakpoint sites antisense transcription of neighboring gene (*SPSB2* primary transcript) leads to a broad convT region, as indicated in the figure. Elevated DRIP-signal indicates formation of DNA-RNA hybrids (see also ***Figure 3—figure supplement 3***).

DOI: 10.7554/eLife.13087.018

The following source data and figure supplements are available for figure 3:

Source data 1. Breakpoint clustering to regions.

DOI: 10.7554/eLife.13087.019

Source data 2. Statistical analysis of separate DRIP-seq and DNAse-seq replicates.

DOI: 10.7554/eLife.13087.020

Figure supplement 1. GRO-seq, RLFS and DRIP-seq signal profiles across genes.

DOI: 10.7554/eLife.13087.021

Figure supplement 2. Venn diagrams comparing SV within Pol2 stalling regions based on GRO- and ChIP-seq profiles.

DOI: 10.7554/eLife.13087.022

Figure supplement 3. Data from all signal tracks for regions displayed in ***Figure 3***.

DOI: 10.7554/eLife.13087.023

Transcriptional-coupled instability at RSS-independent SV hotspots

A mechanistic link between R-loops and AID off-targeting has been established in lymphomas (***Alt et al., 2013***). With this in mind, we investigated regions where off-targeting could occur via R--loops by focusing on breakpoints without RSS-motifs (data shown in figures represents the 416 ETV6-RUNX1 NR-breakp, refer to ***Figure 3—source data 1*** and ***Supplementary file 3*** for all statistical results). We observed significant genome-wide enrichment of breakpoints with the investigated transcriptional features (***Figure 3E*** and F, 29% overlap with Pol2 stalling within gene regions, binomial test P=4.088e-07; 9% genome-wide overlap with convT, P=5.16e-07). This enrichment of breakpoints to convT and Pol2 stalling regions was significant across a wide range of transcriptional activity (refer to ***Supplementary file 3***). Co-occurrence of breakpoints within a 1-kb window was used to distinguish non-recurrent (one breakpoint) and recurrent (more than one breakpoint) events (***Figure 3—source data 1***). Breakpoint recurrence was found to increase the overlap with both Pol2 stalling (***Figure 3E***)and convT (***Figure 3F***). The mean overlap observed in 1000-fold random sampling (grey bars) confirmed the specificity of the overlap (note that Pol2 stalling is analyzed from intragenic regions only). The breakpoints in Pol2 stalling sites were concordant with analysis using Pol2 ChIP-seq (by 78%) and they co-localized with both Ser2 and Ser5 phosphorylated forms of Pol2 complex (***Figure 3—figure supplement 2***). A schematic model summarizing the possible underlying mechanisms based on these results is shown in ***Figure 3G***. The distinct integrated genomic profiles are collectively depicted at the *TPI1* loci, representing an SV hotspot with the highest number of NR-breakp in ETV6-RUNX1 cases (***Figure 3H***, see also ***Figure 3—figure supplement 3*** and ***Figure 2A***). At the breakpoint region, both RLFS and convT are visible and overlap the elevated DRIP-seq signal measured from ES cells.

Access to RAG cleavage sites increases at Pol2 stalling regions

Next, we focused on deciphering whether the transcriptional features associate with RAG off-targeting. We hypothesized that locally depleted nucleosomes around the Pol2 complex (***Bevington and Boyes, 2013***) may enhance access to RSS/RSS-like sequences. To this end, we retrieved DNAse hypersensitivity data from ENCODE (***The ENCODE Project Consortium, 2012***; see Materials and methods). DNAse-seq signal peaks were significantly wider when overlapping with Pol2 stalling sites (***Figure 4A***). A 876 bp (95% CI, 855–896) increase was observed in B-lymphoblastoid cells and 412 bp (95% CI, 395–429) in ES cells (Wilcoxon rank sum test P<2.2e-16 in both cell types, see also ***Figure 3—source data 2***). This was reproducibly observed using peaks located within gene TSS, body or end regions (***Figure 4A***). We selected TSS regions with RSS motifs for closer examination and found that Pol2 stalling sites at these TSS were significantly wider than at other TSS (***Figure 4B***), with a difference of 259 bp (95% CI, 79–475 bp, Wilcoxon rank sum test P=0.0024). Thus, wide Pol2 stalling increases the likelihood of RSS motif occurrence in accessible chromatin. The width of stalling did not correlate positively (Pearson's correlation −0.11; 95% CI, −0.09 to −0.13) with the transcription level of the corresponding gene, indicating that stalling events, and not just active transcription, are important. We further analyzed the top 5% of widest Pol2 stalling regions by comparing them to widest peaks from DNAse hypersensitivity and ChIP for histone marks (see Materials and methods). The odds ratios for the overlap are visualized as a heatmap (see ***Figure 4C***, OR>10 is shown in darkest color tone, refer to ***Figure 4—source data 1*** for more statistics). In addition to DNAse-seq and Pol2 ChIP peaks, the H3K4me3 was found among the top category, confirmed also by ChIP-seq data acquired from REH and Nalm6 cells (***Figure 4—source data 1***).

Next, the ETV6-RUNX1 R-breakp (335; 156 intragenic) were analysed for the genome-wide overlap with the transcriptional features. A 66% overlap was found with Pol2 stalling at intragenic regions (binomial test P<2.2e-16) and a 44% genome-wide overlap with convT (binomial test P<2.2e-16, see also ***Figure 3—source data 1*** for joint analysis across pre-B-ALL subtypes). The overlap with Pol2 stalling had high agreement between GRO-seq and ChIP-seq (***Figure 3—figure supplement 2***) and it increased at recurrent R-breakp (***Figure 4D***). In addition, overlap with convT (***Figure 4E***) was considerable (91%) at regions with 4 or more breakpoints. In comparison, regions with RLFS motifs or annotated TSSs showed less marked enrichment (up to 36%) (***Figure 4E***). Similar, as for NR-breakp, the significant overlap with transcriptional features was preserved at a wide range of expression levels (***Supplementary file 3***). A schematic model that links the obtained results with vulnerability to RAG cleavage is shown in ***Figure 4F***. As in ***Figure 3I***, the different profiles are depicted at the SV

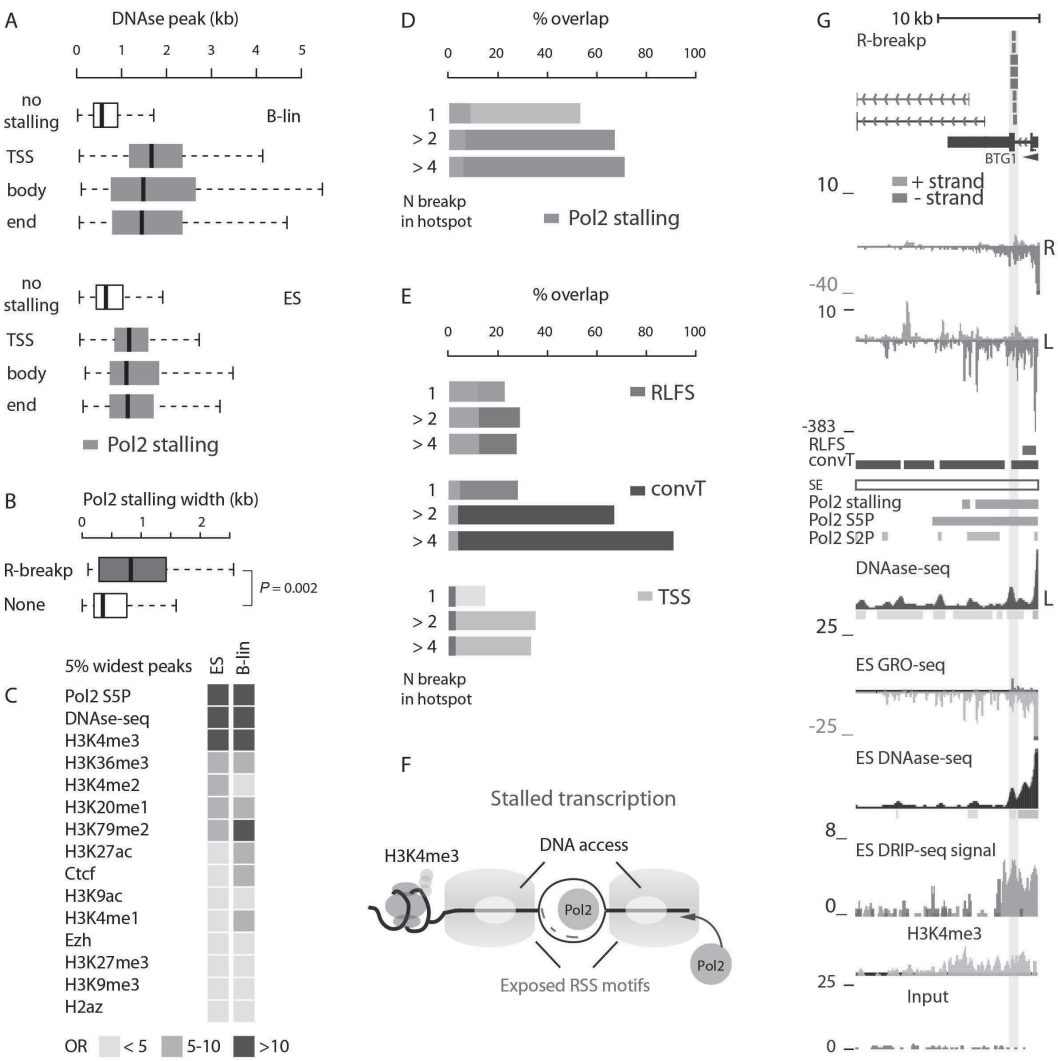

Figure 4. SV with RSS motifs localize to Pol2 stalling regions with broad open chromatin regions. (**A**) DNA access based on DNAse-seq peak width (GM12878 or H1 ES from ENCODE) is compared between regions with no Pol2 stalling (no color) and overlapping Pol2 stalling (light blue, cell-specific Pol2 stalling coordinates are listed in ***Figure 2—source data 1***) at TSS, body and end region of transcripts (refer to ***Figure 3—source data 2*** for statistical analysis based on separate DNAse-seq replicates). (**B**) The TSS stalling width is compared between TSS harboring R-breakp and TSS with no breakpoints (Wilcoxon rank sum test P is indicated, TSS with R-breakp N = 38, TSS without breakpoints N = 11957, 95% CI for size difference 67–491 bp). (**C**) The 5% widest Pol2 stalling regions were overlapped with similarly defined widest peaks in different ChIP- and DNAse-seq data (refer to ***Figure 4—source data 1*** for details and all statistics). The odds-ratio (OR) for the overlap is visualized in color from discrete categories (<5; 5–10; >10, with darker color tones indicating higher OR). Pol2 S5P, DNAse-seq and H3K4me3 peaks had highest OR based on both B-lineage and ES cell data. **D** and **E**: The percentages of R-breakp overlapping Pol2 stalling (as in ***Figure 3E***) or RLFS, convT and annotated TSS (as in ***Figure 3F***) are shown as barplots, respectively. Overall, the recurrence was higher compared to NR-breakp and therefore two categories for recurrent R-breakp are shown (>2; >4). The overlap with convT reaches 91% at highly recurrent R-breakp hotspots (source data can be found in ***Figure 2—source data 1***, S6 and statistics

Figure 4 continued on next page

Figure 4 continued

for genes binned by their transcription level in ***Supplementary file 3***). (**F**) A schematic model illustrating how the transcriptional features may lead to the recruitment of RAG1 and RAG2 based on RSS-motif recognition and chromatin. Pol2 stalling associated with DNA accessibility and wide deposition of the H3K4me3 mark. (**G**) R-breakp hotspot with the highest recurrence (*BTG1* locus) is shown. B-lymphoblastoid and ES cell tracks from DNAse-seq and H3K4me3 from pre-B-ALL cells (Nalm6) represent signals with highest overlap to wide Pol2 stalling (other tracks as in ***Figure 3H***, see also ***Figure 4—figure supplement 1***).

DOI: 10.7554/eLife.13087.024

The following source data and figure supplements are available for figure 4:

Source data 1. Overlap of wide Pol2 stalling regions with unusually wide peaks representing other chromatin features.

DOI: 10.7554/eLife.13087.025

Figure supplement 1. Data from all signal tracks for regions displayed in ***Figure 4***.

DOI: 10.7554/eLife.13087.026

Figure supplement 2. GRO-seq signal profile at multiple clustered deletion regions.

DOI: 10.7554/eLife.13087.027

hotspot with the highest number of R-breakp (***Figure 4G*** *BTG1* locus, see also ***Figure 4—figure supplement 1***). Further examples in ***Figure 4—figure supplement 2*** show RSS-dependent clustered deletions as defined in (***Papaemmanuil et al., 2014***). Overall, the presence of both convT and Pol2 stalling best characterized the recurrent ETV6-RUNX1 breakpoints with RSS motifs (101/148; compared to 20/70 without motif), with 90% (43/48, empirical P=0.002) co-occurrence at intragenic sites (see also ***Supplementary file 4***).

AID expression marks pre-B-ALL lacking common cytogenetic changes

To elucidate the potential for RAG and AID mediated genetic instability in leukemia blasts, we compared the expression of the genes *RAG1, RAG2* and *AICDA* across a transcriptome data set with 1382 pre-B-ALL patients (***Figure 5—source data 1***, ***Figure 5***). Among samples with annotation of cytogenetic subtype (N = 1008), the ETV6-RUNX1 cases (N = 153) exhibited 10.8-fold higher median level of *RAG1* expression relative to other cases with annotated cytogenetic type (Wilcoxon rank sum test P<2.2e-16, 95% CI 8.6–13.6-fold, ***Figure 5A***) and also high *RAG2* expression (***Figure 5B***). Moreover, *AICDA* expression was also detected in a specific subset of patients. It was highest in the 'other' group (N = 267) that does not carry recurrent fusion genes or karyotypic changes (***Figure 5C***, no statistical evaluation was performed as majority of signal values were below detection level of 4.2 in log2 scale). As comparison, we carried out unsupervised analysis of sample similarities based on the global gene expression profiles. To visualize these molecular subtypes in two dimensions, we utilized the t-Distributed Stochastic Neighbor Embedding (t-SNE) method (***van der Maaten and Hinton, 2008***) (see Materials and methods, refer to ***Figure 5—source data 1*** for coordinates). The t-SNE map places highly similar samples in close proximity. The discrete expression states (high; low; not detected) of *RAG1, RAG2* and *AICDA* were evident in distinct groups (***Figure 5D–F***, respectively, the annotated ALL subtypes are colored in ***Figure 5G***). Upon further examination, high levels of *AICDA* expression were particularly prevalent in sample clusters that corresponded to high risk cases from two independent ALL datasets (hypergeometric test P=7.19e-47, ***Figure 5H***, see ***Supplementary file 5*** for patient characteristics). The highest level of *AICDA* expression was presented by a relapsed ALL case, and the *RAG1* and *RAG2* expression levels were 3.09- and 1.93-fold increased at relapse, respectively. Based on the integrated patient profiles, the expression of AID and RAG is distinct in leukemia subtypes and clinical prognosis groups.

Discussion

Next generation sequencing technologies have enabled the elucidation of mechanisms regulating transcription and the analysis of genetic alterations across different cancer genomes. Precursor leukemias are unique in that they often harbor SV and have relatively few mutations (***Roberts and Mullighan, 2015***). Recently, a functional role of transcription in genomic instability has begun to emerge (***Hatchi et al., 2015***; ***Sollier et al., 2014***). The maturing lymphoid cells are vulnerable to off-target effects downstream of RAG and AID activity that is required for immune gene rearrangement (***Meng et al., 2014***; ***Qian et al., 2014***, ***Papaemmanuil et al., 2014***, ***Swaminathan et al., 2015***). The

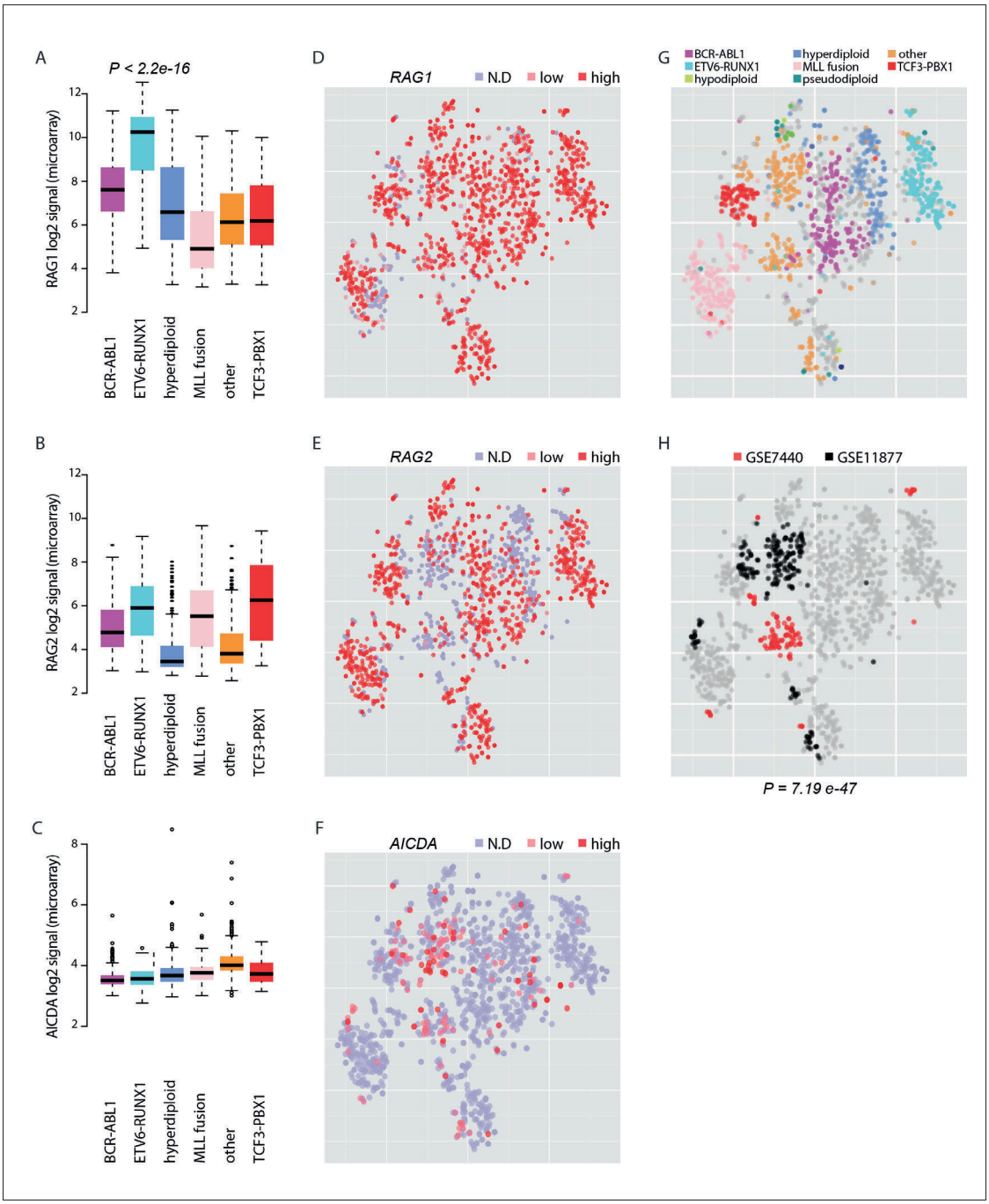

Figure 5. Expression of AID and RAG across molecular subtypes of leukemia. The log2 expression signal is summarized as boxplots for (**A**) *RAG1* (**B**) *RAG2* and (**C**) *AICDA* across the pre-B-ALL subtypes (*N* = 153 BCR-ABL1, *N* = 153 ETV6-RUNX1, *N* = 151 hyperdiploid, *N* = 198 MLL rearrangement,

Figure 5 continued on next page

Figure 5 continued

N = 267 other, *N* = 82 TCF3-PBX1). Wilcoxon rank sum test p-value is indicated for differential *RAG1* expression in the ETV6-RUNX1 subtype (N = 153, patients with cytogenetic subtype information N = 1008) (in **A**). (**D–F**) Alternative representation of discrete expression states for *RAG1*, *RAG2*, and *AICDA*, respectively (red: high, pink: low, grey: not detected). The data points shown as a t-SNE map correspond to the full set of pre-B-ALL patient samples (*N* = 1382) (see also ***Figure 5—source data 1***). Their relative positions are defined by the transcriptome similarity. The sample groups can be compared to annotated cytogenetic types, as colored on the same map in (**G**, **H**). The location of high-risk samples (*N*=295) from two independent studies is indicated in color on the same map (COG studies GSE7740 in red and GSE11877 in black, see also ***Supplementary file 5***). Hypergeometric test p-value is indicated for enrichment of detected *AICDA* expression in the high risk studies (*N* = 112, refer to ***Supplementary file 5*** for population statistics).

DOI: 10.7554/eLife.13087.028

The following source data is available for figure 5:

Source data 1. pre-B-ALL transcriptome samples.

DOI: 10.7554/eLife.13087.029

present study represents a systematic investigation of SVs detected in acute pre-B-cell leukemia using WGS in the context of global transcriptional activity in leukemic cells. We identified specific transcriptional features, namely convergence of transcription and Pol2 stalling, as key factors underlying secondary genetic lesions frequently seen in precursor B leukemias.

Pol2 stalling and convT strongly associate with recurrent breakpoint sites across the genome and at gene loci implicated in leukemia such as *CDKN2A* and *PAX5* (***Sulong et al., 2009***). While protein-coding secondary hits required in disease progression have been recognized for some time, our integrative analysis identified several putative long non-coding RNAs and eRNAs, which merit further investigation. Earlier work has linked eRNAs generated from intragenic superenhancers with AID-mediated instability in lymphomas (***Meng et al., 2014***), proposing that convT leads to arrested transcription, in agreement with experimental evidence from yeast cells (***Prescott and Proudfoot, 2002***). Similarly, it has been shown that Pol2 stalling and R-loops expose ssDNA for AID targeting (***Huang et al., 2007***, ***Pavri et al., 2010***, ***Alt et al., 2013***). We show for the first time that leukemia breakpoints similarly display significant enrichment to enhancers overlapping convT. We further demonstrate a link between convT and elevated R-loop levels and Pol2 stalling on a genome-wide level, with evidence from normal and leukemic human cells. These mechanisms of transcription-coupled genetic instability, earlier implicated in lymphomas (***Pavri et al., 2010***; ***Meng et al., 2014***; ***Pefanis et al., 2014***) and breast cancer (***Hatchi et al., 2015***), therefore have relevance in multiple different cancer types.

Breakpoints carrying RSS-like recognition motifs for RAG1 showed high overlap with the vulnerable regions as defined by convT and Pol2 stalling. Therefore, we propose that also RAG1 access to its target sites is related to the fidelity of elongation. Previous studies investigating motif recognition and genome-wide binding profiles of RAGs have shed light on the mechanisms how this complex is recruited to DNA (***Bevington and Boyes, 2013***, ***Teng et al., 2015***); however these studies have been carried out using normal cells or mouse models that limit their integration with patient WGS data. The chromatin mark H3K4me3 typically found at active promoters serves as a docking site for RAG2 (***Matthews et al., 2007***, ***Teng et al., 2015***). RAG-mediated cleavage further requires recognition of RSS motifs by RAG1 (***Schatz and Swanson, 2011***). Our results revealed that TSS that carry breakpoints with an RSS motif differ from unaffected TSS by the presence of unusually wide Pol2 stalling. We show that Pol2 stalling sites, in general, have increased DNA accessibility. Further, the top 5% of widest stalling regions are characterized by unusually broad DNAse hypersensitive regions and H3K4me3 signal. Unique regulation of Pol2 pausing and elongation has been recognized to be related to broad H3K4me3 domains across a wide variety of cell types (***Benayoun et al., 2014***, ***Scheidegger and Nechaev, 2016***). Together, these properties of Pol2 stalling sites may favor both the recognition and cleavage by the RAG complex.

In this study, we developed a genome-wide approach to capture Pol2 stalling events across gene bodies using change points analysis. This extends previous approaches to detect promoter-proximal pausing events (reviewed in ***Adelman and Lis, 2012***) to analysis of slowing down of Pol2 within the full transcribed region. The feasibility of our approach was confirmed by high overlap of detected regions with RLFS rich regions that represent known structural obstacles to the progression of transcription (***Skourti-Stathaki et al., 2014a***; ***Skourti-Stathaki et al., 2014b***). Furthermore, analysis of Pol2 stalling from Pol2 ChIP-seq profiles acquired in pre-B-ALL cells had high agreement with the

GRO-seq profiles. The slowing of Pol2 upon transition from initiation to elongation, measured by the Pol2 Ser5 phosphorylation, occurs at AID hypermutation sites within the IgH-V region (***Wang et al., 2014a***). We show that this type of Pol2 stalling had high overlap with leukemia breakpoints.

While RAG has a well-established role in pre-B cells, expression of AID represents a recently discovered threat for lymphoid precursor genome integrity. (***Swaminathan et al., 2015***) showed that infection-triggered attenuation of IL-7 receptor signaling led to strong AID expression, thus exposing pre-leukemic cells to additional off-targeting events. Moreover, a negative effect on patient survival and increased relapse frequency were observed in high *AICDA* expressing leukemia patients (***Swaminathan et al., 2015***). We found that high expression of *RAG1/2* or *AICDA* is markedly distinct between different subtypes of pre-B-ALL at the leukemia state. Prevalent *AICDA* expression was a distinguishing feature of high risk pre-B-ALL cases, in line with the previous data (***Swaminathan et al., 2015***). Furthermore, the molecular profiles of patients belonging to the cytogenetic subtype designated as 'other', had high similarity, placing them in close proximity on the t-SNE map. This genetically heterogeneous category of rare cytogenetic types had a distinct elevation in *AICDA* expression. Further investigation of the WGS profiles focusing on this patient category may shed light on whether *AICDA* expression could serve as a putative underlying factor that may spur the diversity of DNA lesions in these patients. Similarly, the over ten-fold higher *RAG1* expression could also be relevant for the prevalent development of leukemia carrying the ETV6-RUNX1 initiating fusion. The *RAG* locus is under complex regulation of local chromatin looping by SATB1 (***Hao et al., 2015***) that controls silencing and activating regulatory elements and was shown to directly control the elevated *RAG1* expression in mice. The enhancer activity in patient blast cells, as captured here in the nascent transcriptomes, will help understanding the regulation of such key loci in detail.

As more data on SV becomes available across cancers, further efforts should be made to elucidate the contribution of different complexes in transcription-coupled genomic instability and to develop strategies for dampening their levels and activity. Translation of these measures into clinical practice could impact treatment efficacy by decreasing clonal heterogeneity and relapse risk.

Materials and methods

GRO-seq samples

Primary bone marrow or blood samples from pediatric precursor B-ALL patients that represented different cytogenetic subtypes were used for GRO-seq assay (refer to ***Supplementary file 1*** for cytogenetic and blast count data). The study was approved by the Regional Ethics Committee in Pirkanmaa, Tampere, Finland (#R13109). The study was conducted according to the guidelines of the Declaration of Helsinki, and a written informed consent was received by the patient and/or guardians. In addition, three ALL cell lines (REH, Kopn-8 and Nalm-6) representing different genetic subtypes (ETV6-RUNX1 fusion, MLL rearrangement and 'other') were included to complement the dataset. REH (ACC-20), Nalm6 (ACC-128), and Kopn8 (ACC-552) cell lines were obtained from the Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (Germany). Mycoplasma status was defined negative by PCR (PCR Mycoplasma Test Kit I/C, PromoCell GmbH, Germany) for all cells. The cell lines were authenticated by PCR of known fusion genes: ETV6-RUNX1 in REH, MLL-MLLT1 in Kopn8, and the lack of recurrent fusions in Nalm6. In addition, the generated genome-wide results can be used in verification of cell line specific markers. We reasoned that a collection of samples that represent both primary blasts and cell lines of different cytogenetics types (and genetic complexity) would be ideal to capture the patterns of transcriptional activity in the lymphoid lineage and leukemic cells. Furthermore, 4–8 replicates were collected from a subset of samples to ensure reproducibility of the results (***Figure 2—figure supplement 2***). In cell culture studies, same cell lines with similar conditions are defined as biological replicates, as nuclei were extracted from temporally independent experiments. For nuclei extractions in co-culture experiments, same cell lines with different culture conditions but with same time points were processed simultaneously. For example, total of eight extractions from REH cells were performed ('Sample name' column), in six slightly different culture conditions ('Cell culture type' and 'Time point (h)' columns), and with two replicate samples collected for two conditions in independent experiments ('Biological replicate' column). There were no technical replicates in the sense that multiple nuclei extractions would have

been made from the same biological replicate. Patient samples (N = 7) collected represent different cytogenetic subtypes and were used as additional confirmation at individual gene loci: for most subtypes N = 1, except hyperdiploid N = 2; ETV6-RUNX1 is represented also by the REH cell line; and replicate samples were generated for Patient 1 that correspond to cultured and freshly isolated cells. (Re-analyzed GRO-seq data: lymphoblastoid data is from three donors; ESC data is from two independent experiments where several technical replicates were pooled). For cell culture conditions and further details, see ***Supplementary file 1***. The nuclei isolation was performed as previously described (***Kaikkonen et al., 2013***), yielding ~1–5×10^6 nuclei per condition. The REH, Nalm6, lymphoblastoid and ES cell samples that represent very deeply sequenced data, were used in the genome-wide analysis (***Supplementary file 1***, GEO accession numbers for deposited pre-B-ALL data: GSE67519 and GSE67540).

GRO-seq assay

Cells were suspended in 10 ml of swelling buffer (10 mM Tris-HCl, 2 mM MgCl2, 3 mM CaCl2 and 2 U/ml SUPERase Inhibitor [Thermofisher, Carlsbad, CA, USA] RNAse inhibitor) and let swell for 5 min. The cells were pelleted for 10 min at 400 × g and resuspended in 500 µl of swelling buffer supplemented with 10% glycerol. Subsequently, 500 µl of swelling buffer supplemented with 10% glycerol and 1% Igepal was added drop by drop to the cells while being vortexed gently. Nuclei were washed twice with 10 ml of swelling buffer supplemented with 0.5% Igepal and 10% glycerol, and once with 1 ml of freezing buffer containing 50 mM Tris-HCl pH 8.3, 40% glycerol, 5 mM MgCl2 and 0.1 mM EDTA. Nuclei were counted and centrifuged at 900 × g for 6 min and suspended to a concentration of 5 million nuclei per 100 µl of freezing buffer, snap-frozen and stored -80˚C until run-on reactions. The nuclear run-on reaction buffer (NRO-RB; 496 mM KCl, 16.5 mM Tris-HCl, 8.25 mM MgCl2 and 1.65% Sarkosyl (Sigma-Aldrich, Steinheim, Germany) was pre-heated to 30˚C. Then each ml of the NRO-RB was supplemented with 1.5 mM DTT, 750 µM ATP, 750 µM GTP, 4.5 µM CTP, 750 µM Br-UTP (Santa Cruz Biotechnology, Inc., Dallas, Texas, USA) and 33 µl of SUPERase Inhibitor (Thermofisher, Carlsbad, CA, USA). 50 µl of the supplemented NRO-RB was added to 100 µl of nuclei samples, thoroughly mixed and incubated for 5 min at 30˚C. GRO-Seq libraries were subsequently prepared as previously described (***Kaikkonen et al., 2013***). Briefly, the run-on products were treated with DNAse I according to the manufacturer's instructions (TURBO DNA-free Kit, Thermofisher, Carlsbad, CA, USA), base hydrolysed (RNA fragmentation reagent, Thermofisher, Carlsbad, CA, USA), end-repaired and then immuno-purified using Br-UTP beads (Santa Cruz Biotechnology, Inc., Dallas, Texas, USA). Subsequently, a poly-A tailing reaction (PolyA polymerase, New England Biolabs, Ipswich, MA, USA) was performed according to manufacturer's instructions, followed by circularization and re-linearization. The cDNA templates were PCR amplified (Illumina barcoding) for 11–14 cycles and size selected to 180–300 bp length. The ready libraries were quantified (Qubit dsDNA HS Assay Kit on a Qubit fluorometer, Thermofisher, Carlsbad, CA, USA) and pooled for 50 bp single-end sequencing with Illumina Hi-Seq2000 (GeneCore, EMBL Heidelberg, Germany). GRO-Seq reads were trimmed using the HOMER v4.3 (http://homer.salk.edu/homer) software (homerTools trim) to remove A-stretches originating from the library preparation. From the resulting sequences, those shorter than 25 bp were discarded.

ChIP-seq assay

ChIP-seq was performed using antibodies against the Ser2 and Ser5 phosphorylated forms of Pol2 and against the histone mark H3K4me3 in REH (N = 1) and Nalm6 cells (N = 2). Ser5 phosphorylation is present before the Pol2 is released to active elongation and it diminishes within the gene body and is greatly reduced downstream of the poly(A) site, where Ser2 phosphorylation is predominantly found (***Zhou et al., 2012***). For ChIP, 40 million cells were crosslinked with 1% formaldehyde for 10–15 mins. The reactions were quenched by adding glycine to a final concentration of 125 mM, and the cells were centrifuged and washed twice with ice-cold PBS. For ChIP 5 or 10 million cells were used (Pol2 or H3K4me3, respectively). Nuclei were extracted by washing cell pellet twice with 1 ml of MNase buffer (10 mM Tris ph 7.4, 10 mM NaCl, 5 mM MgCl2, 0.5% IGEPAL CA-630 [Sigma-Aldrich, Steinheim, Germany], 1x protease inhibitor cocktail [PIC, Roche, Basel, Switzerland], 1 mM PMSF [Thermofisher, Carlsbad, CA, USA]). Nuclei were spun down (1500 × g, +4˚C, 5 min) and suspended into 90 µl of MNase buffer supplemented with 5 mM CaCl2 and 0.1% Triton-X. Different

amounts of MNase (0.5–20 U; #88216, Thermofisher, Carlsbad, CA, USA) was added to the nuclei in 10 µl volume and incubated at 37˚C for 10 mins. To stop the reaction, 100 µl of 2x Lysis buffer was added to the reaction (1% SDS, 40 mM EDTA, 100 mM Tris-HCl pH 8.1) and samples were sonicated using Bioruptor (Diagenode) for 5 cycles (30 s - 30 s) to break the nuclei. The lysate was cleared by centrifugation and supernatant was diluted with RIPA buffer (for Pol2 antibodies, 1X PBS, 1% NP-40, 0.5% Sodium deoxycholate, 0.1% SDS, PIC) or dilution buffer (for H3K4me3; 20 mM Trix-HCl pH 7.4, 100 mM NaCl, 2 mM EDTA, 0.5% TritonX, PIC). The diluted lysate was pre-cleared by rotating for 2 h at 4˚C with 60 µl 80% CL-4B sepharose slurry (GE Healthcare, UK). Before use, sepharose was washed twice with TE buffer, blocked for 1 hr min at room temperature with 0.5% BSA and 20 µg/ml glycogen in 1 ml TE buffer, washed twice with TE and brought up to the original volume with TE. The beads were discarded, and 1% of the supernatant were kept as ChIP input. The protein of interest was immunoprecipitated by rotating the supernatant with 3–5 µg antibody overnight at 4˚C. Antibodies against Ser2P (cat# ab5095, RRID:AB_304749) and Ser5P (cat# ab5131, RRID:AB_449369) were purchased from Abcam (Cambridge, MA, USA). The Ab was captured using 25 µl blocked Protein G Sepharose 4 Fast Flow (GE Healthcare, UK) and rotating the sample for 2 hr at 4˚C. Sepharose was blocked as CL-4B above, except that it was rotated overnight at 4˚C. The beads were pelleted (1 min, 1000×g, 4˚C) and the supernatant discarded. The beads used to bind Ser2P/5P Ab were washed five times with 5X LiCl IP wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% Sodium deoxycholate) and twice with TE in 0.45 µm filter cartridges (Ultrafree MC, Millipore, Bedford, MA, USA). The beads used to pull down H3K4me3 Ab were washed three times with wash buffer I (20 mM Tris/HCl pH 7.4, 150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA), twice with buffer II (20 mM Tris/HCl pH 7.4, 500 mM NaCl, 1% Triton X-100, 2 mM EDTA) and buffer III (10 mM Tris/HCl pH 7.4, 250 mM LiCl, 1% IGEPAL CA-630, 1% Na-deoxycholate, 1 mM EDTA), once with TE + 0.2% TritonX and twice with TE. Immunoprecipitated chromatin was eluted twice with 100 µl elution buffer (TE, 1% SDS). The NaCl concentration was adjusted to 300 mM with 5 M NaCl and crosslinks were reversed overnight at 65˚C. The samples were sequentially incubated at 37˚C for 2 h each with 0.33 mg/ml RNase A and 0.5 mg/ml proteinase K (both from Thermofisher, Carlsbad, CA, USA). The DNA was isolated using the ChIP DNA Clean & Concentrator (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions. Sequencing libraries were prepared from collected DNA by blunting, A-tailing, adaptor ligation as previously described (***Heinz et al., 2010***) using barcoded adapters (NextFlex, Bioo Scientific, Austin, TX, USA). Between the reactions, the DNA was purified using Sera-Mag SpeedBeads (Thermofisher, Carlsbad, CA, USA). Libraries were PCR-amplified for 15–16 cycles, size selected for 230–350bp fragments by gel extraction and single-end sequenced on a Hi-Seq 2000 (Illumina) for 50 cycles.

Processing of GRO-seq, ChIP-seq, DRIP-seq and HiC sequencing reads

The GRO-seq data from lymphoblastoid cells (GSE39878, ***Wang et al., 2014b***; GSE60456, ***Core et al., 2014***), ES cells (GSE41009, ***Sigova et al., 2013***), DRIP-seq data from ES cells (GSE45530, ***Ginno et al., 2013***) and HiC data from human lymphoblastoid GM12878 cells (GSM1551571, GSM1551572, GSM1551574, GSM1551575; ***Rao et al., 2014***) were downloaded from SRA (raw reads) and processed similarly as the new samples: reads were quality controlled and subsequently aligned to the human hg19 reference genome version. Specifically, the quality of raw sequencing reads was confirmed using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and subsequently bases with poor quality scores were trimmed (requiring a minimum 97% of all bases in one read to have a min phred quality score of 10) using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Samples sequenced on multiple lanes were pooled after quality control. Read stacks were collapsed from ChIP-seq files using fastx (collapse). The Bowtie software (bowtie-0.12.9v0.1.x) (***Langmead et al., 2009***) was used for aligning the GRO-seq, ChIP-seq and DRIP-seq reads to the human genome (version hg19). Up to two mismatches and up to three locations were accepted and the best alignment was reported for each read. For the GRO-seq reads this step was preceded by removing reads mapping to rRNA regions (AbundantSequences as annotated by iGenomes) and discarding reads overlapping with so-called blacklisted regions that represent unusual low or high mappability as defined by ENCODE, ribosomal and small nucleolar RNA (snoRNA) loci from ENCODE and further manually curated for the human genome (bed file with sequences is provided in ***Supplementary file 6***).

HiC

Reads from paired-end sequencing were separately filtered and aligned to the genome using bowtie. The reads were checked for MboI restriction sites before doing the alignments and the sequences after GATC sites were trimmed out to improve mappability. The HOMER v4.3 (http://homer.salk.edu/homer) software was used in further processing of HiC-data. Paired-end reads were connected and read pairs with exact same ends were only considered once and read pairs were removed if they were separated by less than 1.5× the estimated sequencing insert length to remove likely continuous genomic fragments or re-ligation events. Paired-end reads originating from regions of unusually high tag density were left out by removing reads from 10 kb regions that contain more than five times the average number of reads. Background model for normalization of HiC-data was generated with 50 kb resolution. The topological domains were identified using the HOMER command' findHiCDomains.pl' using a resolution of 50 kb. This analysis is based on a statistic referred to as the 'directionality index', which describes the tendency of a given position to interact with either the chromatin upstream or downstream from its current position.

GRO-seq

Combined tagDirectories from GRO-seq samples were made by pooling the sequencing data for each cell and sample type with fragment length set to 75. The findPeaks.pl program in the The HOMER v4.3 software (http://homer.salk.edu/homer) was used to identify *de novo* transcripts from GRO-seq data using pooled sequencing reads per sample type. Deeply sequenced REH, Nalm6 and lymphoblastoid cells were used to define signal features in B-cell lineage and separate analysis was carried out for ES cells (see ***Supplementary file 1***). Gaps were allowed at non-mappable regions (-style groseq -uniqmap).

ChIP-seq

Peaks were identified using findPeaks (-style histone –size 1000).

Signal tracks

BedGraph and bigWig files were generated with reads in each sequencing experiment normalized to a total of 10^7 mapped reads. The bigWig files were further converted to track hubs and visualized as strand-specific, overlaid MultiTracks as a custom Track Hub in the UCSC Genome browser.

Genomic regions used in analyses

The hg19 genome version from UCSC (available from iGenomes) was used to specify chromosome lengths in the analysis. The gene annotations from Refseq and UCSC known gene tables were retrieved using the UCSC Table Browser (hg19, GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1)). Unique transcript coordinates were used in analysis, that is, any transcripts sharing the same start and end coordinate were considered together. The TSS regions were defined as +/- 1 kb regions around the annotated start coordinate. Only transcripts mapping to canonical chromosomes were kept, also those on chrM were removed.

Enhancers

Super-enhancer coordinates from CD19+, CD20+ and HSC cells were obtained (***Hnisz et al., 2013***) and merged for visualization of tracks. De novo enhancer detection was performed from the deeply sequenced REH, Nalm6 and lymphoblastoid cells based on the transcript identification result. Transcripts with length <15 kb and the characteristic bidirectionality or co-localization with enhancer locations defined using DNAse and chromatin marker data were used to distinguish eRNAs (see ***Figure 2—source data 1*** for data).

Analysis of SV in context of chromosome subregions

TADs reflect the three dimensional structure of chromatin, forming natural boundaries that divide the chromosomes into sub-regions. To identify TADs with highest frequency of breakpoints, HiC-data analysis was performed using HOMER 4.3. As our goal is to generate a natural division of the genome into sub-regions that are relevant in context of transcriptional regulation, this approach is superior to arbitrarily assigning sub-regions based on fixed windows. The pre-B-ALL breakpoints and

annotation data (***Andersson et al., 2015***; ***Holmfeldt et al., 2013***; ***Papaemmanuil et al., 2014***; ***Paulsson et al., 2015***) were analyzed in context of TADs. Specifically, TADs were overlapped with subtype-specific breakpoints (***Figure 1—source data 1*** presents TADs sorted based on the count of breakpoints). Subsequently, TADs with breakpoints were divided into quartiles based on breakpoint frequency per bp to analyze enrichment of feature overlap that exceeds the genomic background level. To obtain the total transcribed area width within each TAD, the TAD coordinates were overlapped with the detected GRO-seq transcripts (bedtools intersect -wao). The combined SV data represents in total 1680 breakpoints and is the most comprehensive collection of pre-B-ALL SV that we are aware of.

Chromatin segmentation data

BroadChromHMM chromatin segmentations were obtained from GM12878 and H1 ES cells including the following segment types: 1_Active_Promoter, 2_Weak_Promoter, 3_Poised_Promoter, 4_Strong_Enhancer, 5_Strong_Enhancer, 6_Weak_Enhancer, 7_Weak_Enhancer, 8_Insulator, 9_Txn_Transition, 10_Txn_Elongation, 11_Weak_Txn, 12_Repressed", 13_Heterochrom/lo, 14_Repetitive/CNV, 15_Repetitive/CNV. The sizes of the segments of each type were used to calculate the total span from the genome. Each segment type was then overlapped with a combined bed file specifying convT and Pol2 stalling regions. Overlapping and non-overlapping pieces were returned and analyzed separately (bedtools intersect, followed by bedtools subtract with the overlapping pieces given as parameter b).

Distinguishing breakpoints based on RSS-like motifs or recurrence

Two types of breakpoints were distinguished based on RSS motif annotation to result in the following region assignment: regions containing a consensus RSS/heptamer sequence motif were used to categorize co-localized breakpoints as putative RSS-dependent breakpoints (R-breakp: 447 in total, 335 in the ETV6-RUNX1 subtype), while regions devoid of recognition sequence were used to classify RSS-independent lesions (NR-breakp: 938 in total, 416 in the ETV6-RUNX1 subtype). Regions harboring unresolved breakpoints were left out from majority of analysis performed (285 regions harboring 295 breakpoints in the ETV6-RUNX1 subtype that were mainly isolated and non-recurrent). The RSS assignment for other breakpoints was obtained in the following way: the resolved breakpoints were extended to both sides by 10 bp, resulting in a 21 bp region. The MEME analysis in ***Papaemmanuil et al. 2014*** for 708 resolved breakpoints from ETV6-RUNX1 patients was replicated and comparable sequence logos to that reported previously were obtained and used to annotate RSS status. A p-value cut-off of 0.003 was chosen for the MEME motif scanning based on FIMO analysis of the ETV6-RUNX1 data.

To evaluate recurrence, the breakpoint ends at 1 kb distance were stitched together to form regions (each with at least one breakpoint, see ***Figure 3—source data 1***), annotating the number of breakpoints inside (BEDTools mergeBed –d 1000 –n). Overlap of breakpoint regions with RLFS, TSS, convT and Pol2 stalling regions were obtained using BEDTools with 1 kb window. The overlap frequencies were compared to random sampling of similarly sized genomic regions. Further comparisons were performed separating recurrent (>1 breakpoint per stitched region) and non-recurrent regions, and with increasing the cut-off for the number of breakpoint events per stitched region. The same was repeated for breakpoints within genes binned into four categories based on their transcription level. The transcript regions were quantified using data from REH, Nalm6, and lymphoblastoid cells, and normalized by RPKM. The maximum expression value was to divide transcripts into quartiles based on the expression level.

Signal feature analysis

The visual examination of SV sites served as the first step to define transcriptional features of potential relevance. This motivated the analysis of regions with overlapping transcription from both strands (convT) and local elevations in the signal (Pol2 stalling), with detailed definitions given below. For genome-wide analysis of signal feature overlap with SV, feature tracks from several samples were combined. This approach was deemed most appropriate to address the dynamic nature of transcriptional activity and to avoid missing regions that due to high recurrence of SV may be deleted in subset of leukemic cells studied.

ConvT

ConvT regions were identified as transcripts that overlap on opposite strands by at least 100 bp (as in ***Meng et al., 2014***). Subsequently, a combined bed track was created for the leukemic and lymphoblastoid samples using bedTools mergeBed command (-d 0). The data from ES cells (GSE41009) served as an independent control. The level of convT was quantified using the HOMER program analyzeRNA.pl from both strands separately and normalized by region size. The minimum value obtained per region (comparing + and - strands) was assigned as convT level.

Pol2 stalling

Change-point detection is the mathematical problem of finding abrupt changes in a signal, typically applied in context of time series (***Killick et al., 2012***). Both approximate and exact methods exist for estimating the point at which the statistical properties of a sequence of observations change. The analysis of changepoints in the signal mean was carried out using functions implemented in the R package 'changepoint' (***Killick and Eckley, 2014***). An exact method with favorable computational cost was recently introduced in context of time-series data (***Killick et al., 2012***). This method called PELT was selected to detect the changepoints from scaled (zero mean, equal variance) signal profiles calculated at 50-bp resolution (generated bedGraph files are available under the GEO accession GSE67540). The analysis was performed separately for each gene, using Bayesian Information Criterion as a penalty term with the changepoint counted as a parameter (function cpt.mean with parameters penalty = 'BIC1', method ='PELT'). The input dataset representing primary transcription activity at gene loci was generated by overlapping the GRO-seq signal file strand-specifically with transcript coordinates from UCSC and Refseq (see genomic regions used). The analysis only considered regions with annotation match (in minimum 5% of identified transcript covered by annotation; a minimum of 50% overlap with the identified transcript; annotated and detected starts do not differ more than 10 kb). In order to define Pol2 stalling sites, the signal level between changepoints were compared to the median across the whole gene, and intervals above 90% quantile were reported as stalled. For ChIP-seq, this cut-off was relaxed to 80% due to higher background signal. Notice also that there is no strand information based on ChIP-seq. The analysis was carried out separately for each of the deeply sequenced (REH, Nalm6 and lymphoblastoid) GRO-seq datasets and ChIP-seq replicates and subsequently merged to one bed file specifying stalled region coordinates (bedTools merge –d 100). The ES GRO-seq dataset GSE41009 was processed similarly and used as an independent control. To study whether there was a relationship between the stalled region size overlapping TSS regions and R-breakp frequency, the following intersects were calculated using BEDTools (intersectBed -wa | uniq): (i) overlap of Pol2 stalling sites and TSS regions harboring R-breakp and (ii) overlap of Pol2 stalling sites and TSS regions not harboring R- or NR-breakp. Subsequently, the sizes of Pol2 stalling sites in each coordinate file were calculated and the Wilcoxon rank sum test used for evaluating statistical significance for the difference in Pol2 stalling width. Secondly, top 5% widest Pol2 stalling sites were identified and compared to top 5% widest peaks from ChIP-seq and DNAse-seq profiles (see below).

Signal comparison at gene regions

The HOMER command annotatePeaks.pl was used to create a transcriptional profile of active genes (RPKM > 0.5) in ES, lymphoblastoid and REH cells by scaling the histogram to each region (i.e 0–100%) using a bin size of 100. RLFS motif density was calculated across genes with the BEDtools coverage tool. A density plot representing RLFS frequency across gene regions was then obtained as above.

DRIP-seq and RLFS motif data for R-loop detection

Data from replicate DRIP-seq experiments with two different restriction enzyme digestions (GSE45530) were used in the analysis. Log2 signal levels were quantified using HOMER at TSS regions. Statistical significance was estimated separately for the two different restriction enzyme digestions. RLFS motif search was performed using the software QmRLFS-finder that predicts R-loop forming sequences based on structural models of known sequences (***Jenjaroenpun et al., 2015***). The fasta input file was generated by extracting DNA sequences based on the hg19 genome version.

DNAse-seq data and additional ChIP-seq data to characterize wide Pol2 stalling events

The DNAse-seq peaks were first overlapped with the Pol2 stalling regions detected based on the GRO-seq signal. Only peaks with a score above 15 were considered. The width of the overlapping peaks was then compared to the width of peaks with no overlap using the Wilcoxon rank sum test. Next, top 5% widest peaks were obtained from the DNAse-seq and ChIP-seq data (refer to ***Figure 4—source data 1***). The overlap with 5% widest Pol2 stalling regions was subsequently evaluated using the BEDTools fisher tool.

Transcriptome data

Gene expression data from pre-B-ALL studies was combined from microarray datasets retrieved from the NCBI GEO database as part of a data collection representing both healthy and malignant samples hybridized to hgu133Plus2 genome-wide microarrays (in preparation for submission). In total, 1382 pre-B-ALL samples were included. Probe-level quality control was performed to exclude samples with very high difference in data location or distribution as measured by median and inter-quartile range of raw probe intensities. Samples that passed this filtering were processed using the RMA probe summarization algorithm with probe mapping to Entrez Gene IDs (from BrainArray version 18.0.0, ENTREZG), followed by bias correction using the R package 'bias'. The Barnes-Hut-SNE algorithm (computationally faster approximation of t-SNE) implementation from the R package 'Rtsne' (***Krijthe, 2015***) was used to discover near-optimal representation of sample distances in two dimensions (using parameter values perplexity 30 and theta 0.5) using 15% genes with highest variance. The t-SNE method belongs to dimensionality reduction methods that include also traditional methods such as Principal Component Analysis. The main objective of the method is to accurately place highly similar samples (here based on the high-dimensional gene expression profile) to close proximities in lower dimensions. The result can be visualized in two-dimensions as a scatter plot that allows observing sample groups based on the molecular profiles. According to our experience, this method provides better separation between sample groups compared to more traditional methods for large heterogeneous sample collections. To identify whether a given gene was expressed or unexpressed in a sample, a Gaussian finite mixture model (testing equal and variable variance models, best fit chosen by BIC) was fitted by expectation-maximization algorithm to the probe signals (R package 'mclust', version 4.3, ***Fraley and Raftery, 2002***).

Statistical tests

Statistical significance was estimated using several tests to ensure reliability, including tests that rely on assumptions about data distributions and empirical tests that rely on randomization of data points. The statistical tests used, exact values of N, definitions of center and dispersion and precision measures are indicated in Results, in the respective supplementary tables or figure legends.

Binomial test

Test for independent random trials with binary (success/failure) outcome, with replacement. This test was used to assess the statistical significance of observing breakpoint events overlapping a transcriptional feature (Pol2 stalling or convT). Success in population was defined using 1 kb windows across the genome. The windows overlapping the studied feature was divided by the total number of 1 kb windows analyzed. E.g. in the Pol2 stalling analysis, the total number of windows overlapping Pol2 stalling regions divided by this number of 1 kb windows within gene coordinates (included to the input for the change point analysis), define probability of success.

Hypergeometric test

Test for independent random trials with binary (success/failure) outcome, without replacement. This test was used to assess the statistical significance of observing greater than or equal overlap frequency between breakpoints and an annotated set of genomic regions. E.g. to test for enrichment of breakpoints inside convT-positive enhancers, convT-positive enhancers with breakpoints define sample success; all enhancers with breakpoints population success; and sample taken is all convT-positive enhancers (from the population of all enhancers). The related Fisher's test (implemented in BEDTools fisher) was used to obtain similar statistics with odds ratios.

Wilcoxon rank sum test (Mann-Whitney test)

A nonparametric two-sided Wilcox test was performed to estimate whether two samples (continuous values, unknown distribution) come from the same population (R function wilcox.test). This test was applied to quantified signal levels compared between categories.

Random sampling

This test can be used to obtain an empirical estimate of random overlap frequencies. The sampling was performed 1000-fold within the same genomic context as used in the analysis. To estimate the significance of overlap between stitched breakpoint regions with e.g. convT regions, the stitched regions were allocated random genomic coordinates, thus preserving the size distribution and breakpoint event frequencies within stitched regions. The observed random region overlap was used as the empirical p-value estimate. Further, the z-test was used to evaluate whether there was evidence to reject the null hypothesis that the observed feature overlap value would belong to the empirical distribution obtained.

Acknowledgements

We would like to thank Ville Hautamäki for comments on signal analysis methods and the EMBL Gene Core sequencing team for the sequencing service provided. The work was supported by grants from the Emil Aaltonen Foundation, Jane and Aatos Erkko Foundation, Finnish Cancer Foundation, Academy of Finland, Sigrid Juselius Foundation,Finnish Cultural Foundation, Paulo Foundation, Foundation for Pediatric Research, the Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital, University of Tampere and University of Eastern Finland.

Additional information

Funding

Funder	Grant reference number	Author
Suomen Kulttuurirahasto	00150214	Merja Heinäniemi Olli Lohi
Itä-Suomen Yliopisto		Merja Heinäniemi
The Finnish Cancer Foundation		Merja Heinäniemi
Emil Aaltosen Säätiö		Merja Heinäniemi
Suomen Akatemia	276634	Merja Heinäniemi
Tampereen Yliopisto		Susanna Teppo Saara Laukkanen Thomas Liuksiala Olli Lohi
Sigrid Juselius Foundation		Minna U Kaikkonen
Suomen Akatemia	277816	Olli Lohi
Jane ja Aatos Erkon Säätiö		Olli Lohi
Paulo Foundation		Olli Lohi
Lastentautien Tutkimussäätiö		Olli Lohi
Competitive State Research Financing of the Expert Responsibility area of Tampere University Hospital		Olli Lohi

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

MH, ST, MUK, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; TV, OL, Conception and design, Analysis and interpretation of data, Drafting or revising the article; MB-L, JM, HN, TL, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; VZ, Analysis and interpretation of data, Drafting or revising the article; SL, KT, Acquisition of data, Drafting or revising the article

Author ORCIDs

Merja Heinäniemi, http://orcid.org/0000-0001-6190-3439
Susanna Teppo, http://orcid.org/0000-0003-2569-8030
Olli Lohi, http://orcid.org/0000-0001-9195-0797

Ethics

Human subjects: The study was approved by the Regional Ethics Committee in Pirkanmaa, Tampere, Finland (#R13109). The study was conducted according to the guidelines of the Declaration of Helsinki, and a written informed consent was received by the patient and/or guardians.

Additional files

Supplementary files

• Supplementary file 1. GRO-seq sample summary. Description of the patient and cell line GRO-seq samples used in the analysis, including the cell culture conditions, replicate information and the total number of pooled sequencing reads obtained after quality filtering and alignment. A more detailed table for cultured samples with replicate information and accession codes is provided at the bottom. Sample accession codes for already published and re-analyzed GRO-seq data, and additional GRO-seq data displayed in ***Figure 1—figure supplement 1*** are listed in worksheet 2.
DOI: 10.7554/eLife.13087.030

• Supplementary file 2. Genomic coordinates for regions displayed. The coordinates of example gene regions displayed in the main and supplementary figures are listed (hg19 human genome version).
DOI: 10.7554/eLife.13087.031

• Supplementary file 3. Breakpoint hotspot analysis for genes binned by the transcription level. Hypergeometric test statistics for genes stratified by expression level. Breakpoint overlap with transcriptional features was tested within the binned intragenic regions. Data for ETV6-RUNX1 subtype and all pre-B-ALL subtypes are shown as separate worksheets. Related to ***Figures 3*** and ***4***.
DOI: 10.7554/eLife.13087.032

• Supplementary file 4. Intragenic recurrent SV in ETV6-RUNX1 patients with overlap to vulnerable regions. The patient and region identifiers for recurrent intragenic SV in ETV6-RUNX1 patients are listed, reporting separately those co-localized with Pol2 stalling or convT regions.
DOI: 10.7554/eLife.13087.033

• Supplementary file 5. Clinical data for patients with high *AICDA* expression. Study description, sample identifier, cytogenetic group, age and dataset identifier are listed for the patients within high *AICDA* expression level. Statistical analysis testing enrichment of detected AICDA expression in high risk studies is summarized in worksheet 2.
DOI: 10.7554/eLife.13087.034

• Supplementary file 6. Custom blacklisted genomic regions. Blacklisted regions discarded from the analysis that were deemed to represent low-mappability, rRNA and snoRNA loci based on GRO-seq signal. Coordinates refer to the hg19 human genome version.
DOI: 10.7554/eLife.13087.035

Major datasets

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Heinäniemi M, Teppo S, Kaikkonen MU, Bouvy-Liivrand M, Lohi O	2015	ALL cells	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67540	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE67540)
Heinäniemi M, Teppo S, Lohi O	2015	Genome-wide mapping of TEL-AML1 targets in acute leukemia	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67519	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE67519)

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG	2014	RNA-DNA DIFFERENCES IN NASCENT RNA	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39878	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE39878)
Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT	2014	Analysis of transcription start sites from nascent RNA identifies a unified architecture of initiation at mammalian promoters and enhancers	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60456	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE60456)
Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA	2013	Divergent transcription of lncRNA/mRNA gene pairs in embryonic stem cells	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41009	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE41009)
Ginno PA, Lim YW, Lott PL, Korf I, Chédin F	2013	DNA-RNA Immunoprecipitation sequencing (DRIP-seq) of human NT2 cells	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45530	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE45530)
Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, Geeting KP, Gnirke A, Melnikov A, McKenna D, Stamenova EK, Lander ES, Aiden EL	2014	A three-dimensional map of the human genome at kilobase resolution reveals prinicples of chromatin looping	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE63525)
Sandstrom R	2011	DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29692	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE29692)
Shoresh N	2011	Histone Modifications by ChIP-seq from ENCODE/Broad Institute	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29611	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE29611)

Sandstrom R	2011	CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30263	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE30263)
Myers R, Pauli F	2011	Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32465	Publicly available at NCBI Gene Expression Omnibus (accession no: GSE32465)

References

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews. Genetics* **13**:720–731. doi: 10.1038/nrg3293

Alt FW, Zhang Y, Meng FL, Guo C, Schwer B. 2013. Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell* **152**:417–429. doi: 10.1016/j.cell.2013.01.007

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, Wu G, Li Y, Mulder H, Raimondi S, Pounds S, Kang G, Shi L, et al. 2015. The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nature Genetics* **47**:330–337. doi: 10.1038/ng.3230

Bateman CM, Alpar D, Ford AM, Colman SM, Wren D, Morgan M, Kearney L, Greaves M. 2015. Evolutionary trajectories of hyperdiploid ALL in monozygotic twins. *Leukemia* **29**:58–65. doi: 10.1038/leu.2014.177

Benayoun BA, Pollina EA, Ucar D, Mahmoudi S, Karra K, Wong ED, Devarajan K, Daugherty AC, Kundaje AB, Mancini E, Hitz BC, Gupta R, Rando TA, Baker JC, Snyder MP, Cherry JM, Brunet A. 2014. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**:673–688. doi: 10.1016/j.cell.2014.06.027

Bevington S, Boyes J. 2013. Transcription-coupled eviction of histones H2A/H2B governs V(D)J recombination. *The European Molecular Biology Organization Journal* **32**:1381–1392. doi: 10.1038/emboj.2013.42

Bolland DJ, Wood AL, Johnston CM, Bunting SF, Morgan G, Chakalova L, Fraser PJ, Corcoran AE. 2004. Antisense intergenic transcription in V(D)J recombination.. *Nature Immunology* **5**:630–637. doi: 10.1038/ni1068

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics* **46**:1311–1320. doi: 10.1038/ng.3142

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**:611–631. doi: 10.1198/016214502760047131

Gellert M. 2002. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annual Review of Biochemistry* **71**:101–132. doi: 10.1146/annurev.biochem.71.090501.150203

Ginno PA, Lim YW, Lott PL, Korf I, Chédin F. 2013. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Research* **23**:1590–1600. doi: 10.1101/gr.158436.113

Hao B, Naik AK, Watanabe A, Tanaka H, Chen L, Richards HW, Kondo M, Taniuchi I, Kohwi Y, Kohwi-Shigematsu T, Krangel MS. 2015. An anti-silencer- and SATB1-dependent chromatin hub regulates Rag1 and Rag2 gene expression during thymocyte development. *The Journal of Experimental Medicine* **212**:809–824. doi: 10.1084/jem.20142207

Hatchi E, Skourti-Stathaki K, Ventz S, Pinello L, Yen A, Kamieniarz-Gdula K, Dimitrov S, Pathania S, McKinney KM, Eaton ML, Kellis M, Hill SJ, Parmigiani G, Proudfoot NJ, Livingston DM. 2015. BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Molecular Cell* **57**:636–647. doi: 10.1016/j.molcel.2015.01.011

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* **38**:576–589. doi: 10.1016/j.molcel.2010.05.004

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**:934–947. doi: 10.1016/j.cell.2013.09.053

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, et al. 2013. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nature Genetics* **45**:242–252. doi: 10.1038/ng.2532

Huang FT, Yu K, Balter BB, Selsing E, Oruc Z, Khamlichi AA, Hsieh CL, Lieber MR. 2007. Sequence dependence of chromosomal R-loops at the immunoglobulin heavy-chain Smu class switch region. *Molecular and Cellular Biology* **27**:5921–5932. doi: 10.1128/MCB.00702-07

Jenjaroenpun P, Wongsurawat T, Yenamandra SP, Kuznetsov VA. 2015. QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Research* **43**:W527–W534. doi: 10.1093/nar/gkv344

Jonkers I, Lis JT. 2015. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology* **16**:167–177. doi: 10.1038/nrm3953

Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, Chun HB, Tough DF, Prinjha RK, Benner C, Glass CK. 2013. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell* **51**:310–325. doi: 10.1016/j.molcel.2013.07.010

Killick R, Eckley IA. 2014. changepoint : An R Package for Changepoint Analysis. *Journal of Statistical Software* **58**:1–19 . doi: 10.18637/jss.v058.i03

Killick R, Fearnhead P, Eckley IA. 2012. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association* **107**:1590–1598. doi: 10.1080/01621459.2012.737745

Krijthe J. 2015. *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.10. https://CRAN.R-project.org/package=Rtsne.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25. doi: 10.1186/gb-2009-10-3-r25

Maia AT, van der Velden VH, Harrison CJ, Szczepanski T, Williams MD, Griffiths MJ, van Dongen JJ, Greaves MF. 2003. Prenatal origin of hyperdiploid acute lymphoblastic leukemia in identical twins. *Leukemia* **17**:2202–2206. doi: 10.1038/sj.leu.2403101

Matthews AG, Kuo AJ, Ramón-Maiques S, Han S, Champagne KS, Ivanov D, Gallardo M, Carney D, Cheung P, Ciccone DN, Walter KL, Utz PJ, Shi Y, Kutateladze TG, Yang W, Gozani O, Oettinger MA. 2007. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature* **450**:1106–1110. doi: 10.1038/nature06431

Meng FL, Du Z, Federation A, Hu J, Wang Q, Kieffer-Kwon KR, Meyers RM, Amor C, Wasserman CR, Neuberg D, Casellas R, Nussenzweig MC, Bradner JE, Liu XS, Alt FW. 2014. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159**:1538–1548. doi: 10.1016/j.cell.2014.11.014

Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, Hows JM, Navarrete C, Greaves M. 2002. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proceedings of the National Academy of Sciences of the United States of America* **99**:8242–8247. doi: 10.1073/pnas.112218799

Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, Alexandrov LB, Van Loo P, Cooke SL, Marshall J, Martincorena I, Hinton J, Gundem G, van Delft FW, Nik-Zainal S, Jones DR, Ramakrishna M, Titley I, Stebbings L, Leroy C, et al. 2014. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature Genetics* **46**:116–125. doi: 10.1038/ng.2874

Paulsson K, Lilljebjörn H, Biloglav A, Olsson L, Rissler M, Castor A, Barbany G, Fogelstrand L, Nordgren A, Sjögren H, Fioretos T, Johansson B. 2015. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Nature Genetics* **47**:672–676. doi: 10.1038/ng.3301

Pavri R, Gazumyan A, Jankovic M, Di Virgilio M, Klein I, Ansarah-Sobrinho C, Resch W, Yamane A, Reina San-Martin B, Barreto V, Nieland TJ, Root DE, Casellas R, Nussenzweig MC. 2010. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**:122–133. doi: 10.1016/j.cell.2010.09.017

Pefanis E, Wang J, Rothschild G, Lim J, Chao J, Rabadan R, Economides AN, Basu U. 2014. Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. *Nature* **514**:389–393. doi: 10.1038/nature13580

Prescott EM, Proudfoot NJ. 2002. Transcriptional collision between convergent genes in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America* **99**:8796–8801. doi: 10.1073/pnas.132270899

Qian J, Wang Q, Dose M, Pruett N, Kieffer-Kwon KR, Resch W, Liang G, Tang Z, Mathé E, Benner C, Dubois W, Nelson S, Vian L, Oliveira TY, Jankovic M, Hakim O, Gazumyan A, Pavri R, Awasthi P, Song B, et al. 2014. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**:1524–1537. doi: 10.1016/j.cell.2014.11.013

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:1665–1680. doi: 10.1016/j.cell.2014.11.021

Robbiani DF, Deroubaix S, Feldhahn N, Oliveira TY, Callen E, Wang Q, Jankovic M, Silva IT, Rommel PC, Bosque D, Eisenreich T, Nussenzweig A, Nussenzweig MC. 2015. Plasmodium infection promotes genomic instability and AID-dependent B cell lymphoma. *Cell* **162**:727–737. doi: 10.1016/j.cell.2015.07.019

Roberts KG, Mullighan CG. 2015. Genomics in acute lymphoblastic leukaemia: insights and treatment implications. *Nature Reviews. Clinical Oncology* **12**:344–357. doi: 10.1038/nrclinonc.2015.38

Schatz DG, Swanson PC. 2011. V(D)J recombination: mechanisms of initiation. *Annual Review of Genetics* **45**:167–202. doi: 10.1146/annurev-genet-110410-132552

Scheidegger A, Nechaev S. 2016. RNA polymerase II pausing as a context-dependent reader of the genome. *Biochemistry and Cell Biology = Biochimie Et Biologie Cellulaire* **94**:82–92. doi: 10.1139/bcb-2015-0045

Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **110**:2876–2881. doi: 10.1073/pnas.1221904110

Skourti-Stathaki K, Proudfoot NJ. 2014a. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes & Development* **28**:1384–1396. doi: 10.1101/gad.242990.114

Skourti-Stathaki K, Kamieniarz-Gdula K, Proudfoot NJ. 2014b. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516**:436–439. doi: 10.1038/nature13787

Sollier J, Stork CT, García-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA. 2014. Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Molecular Cell* **56**:777–785. doi: 10.1016/j.molcel.2014.10.020

Sulong S, Moorman AV, Irving JA, Strefford JC, Konn ZJ, Case MC, Minto L, Barber KE, Parker H, Wright SL, Stewart AR, Bailey S, Bown NP, Hall AG, Harrison CJ. 2009. A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* **113**:100–107. doi: 10.1182/blood-2008-07-166801

Sun M, Gadad SS, Kim DS, Kraus WL. 2015. Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. *Molecular Cell* **59**:698–711. doi: 10.1016/j.molcel.2015.06.023

Swaminathan S, Klemm L, Park E, Papaemmanuil E, Ford A, Kweon SM, Trageser D, Hasselfeld B, Henke N, Mooster J, Geng H, Schwarz K, Kogan SC, Casellas R, Schatz DG, Lieber MR, Greaves MF, Müschen M. 2015. Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia. *Nature Immunology* **16**:766–774. doi: 10.1038/ni.3160

Teng G, Maman Y, Resch W, Kim M, Yamane A, Qian J, Kieffer-Kwon KR, Mandal M, Ji Y, Meffre E, Clark MR, Cowell LG, Casellas R, Schatz DG. 2015. RAG Represents a Widespread Threat to the Lymphocyte Genome. *Cell* **162**:751–765. doi: 10.1016/j.cell.2015.07.009

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74. doi: 10.1038/nature11247

van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**:2579–2605.

Wang X, Fan M, Kalis S, Wei L, Scharff MD. 2014a. A source of the single-stranded DNA substrate for activation-induced deaminase during somatic hypermutation. *Nature Communications* **5**:4137. doi: 10.1038/ncomms5137

Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG. 2014b. RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. *Cell Reports* **6**:906–915. doi: 10.1016/j.celrep.2014.01.037

Wiemels JL, Ford AM, Van Wering ER, Postma A, Greaves M. 1999. Protracted and variable latency of acute lymphoblastic leukemia after TEL-AML1 gene fusion in utero. *Blood* **94**:1057–1062.

Zhang J, 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, Collins-Underwood JR, Ma J, Roberts KG, Pounds SB, Ulyanov A, Becksfort J, Gupta P, et al. 2012. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**:157–163. doi: 10.1038/nature10725

Zhou Q, Li T, Price DH. 2012. RNA polymerase II elongation control. *Annual Review of Biochemistry* **81**:119–143. doi: 10.1146/annurev-biochem-052610-095910