Aus dem Experimental and Clinical Research Center
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

LncRNAs signature defining major subtypes of B-cell acute lymphoblastic leukemia

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Alva Rani James, Msc

aus Kozhikode (Kerala, India)

Datum der Promotion…..23.06.2019…

TABLE OF CONTENTS

TERMS AND ABBREVIATIONS

Abbreviation	Full term
lncRNAs	Long non-coding RNAs
PC	Protein-coding
BCP-ALL	B-cell precursor Acute lymphoblastic leukemia
DUX4	Double homeobox 4
Ph-like	Philadelphia-like (Ph-like)
NH-HeH	Near haploid/High hyper-diploid
BM	Bone Marrow
ID	Initial diagnosis
REL	Relapse
HSCs	Hematopoietic stem cells
AML	Acute Myleoid Leukemia
JAK-SAT	Janus kinase and Signal Transducer and Activator of Transcription
mTOR	mammalian Target of Rapamycin
PI3K-Akt	Phosphatidylinositol 3'-kinase
TGF-β	Transforming Growth Factor
RNA-Seq	RNA sequencing
CAMs	Cell adhesion molecules
FPKM	Fragments Per Kilobase Million
PVT1	Plasmacytoma variant translocation 1
LUCAT1	Lung Cancer Associated Transcript 1
TCL6	T-cell leukemia/lymphoma 6
HOTAIRM1	HOX antisense intergenic RNA myeloid 1
ANRIL	Antisense Non-coding RNA in the INK4 Locus
TERRA	Telomeric repeat-containing RNA
MIAT	Myocardial infarction associated transcript
CRNDE	Colorectal neoplasia differentially expressed
GAS5	Growth arrest-specific 5
XIST	X-inactive specific transcript

HOTTIP	HOXA transcript at the distal tip
DLEU1	Deleted Lymphocytic Leukemia 1
IKZF1	Ikaros family zinc finger protein 1
SAMD-AS2	SMAD family member one antisense 2
SMAD	SMAD family member 1
ITGA6	Integrin alpha-6
CDK6	Cyclin-dependent kinase
IL2RA	Interleukin-2 receptor alpha chain
STAR	Spliced Transcripts Alignment to a Reference
LIMMA	Linear Models for Microarray Data
GREAT	Genomic Regions Enrichment of Annotations Tool
SWAN	Subset-quantile within array Normalization
TSS	Transcription start site
DE	Differential expression
DM	Differential Methylated
GTF	Gene transfer format
BED	Browser Extensible Data
FPKM	Fragments Per Kilobase of sequence per Million mapped reads

LIST OF FIGURES

LIST OF TABLES

ZUSAMMENFASSUNG

Einführung: Die B-Vorläufer akute lymphatischen Leukämie (BCP-ALL) ist eine heterogene Krebserkrankung mit mehreren definierten Subgruppen. Neue Daten deuten darauf hin, dass lange nicht-kodierende RNAs (long noncoding RNAs - lncRNAs) eine Schlüsselrolle bei der Entwicklung und Progression der BCP-ALL spielen könnten. Daher führten wir eine Transkriptions- und DNA-Methylierungsstudie durch, um die lncRNA-Landschaft von drei BCP-ALL-Subgruppen (82 Proben) zu charakterisieren und potentielle regulative Konsequenzen zu analysieren.

Methodik: Material wurde zum Zeitpunkt der Erstdiagnose (ID) und im Rezidiv (REL) von erwachenen (n = 21) und pädiatrischen (n = 24) BCP-ALL-Patienten entnommen und unter Verwendung von RNA-Seq und DNA-Methylierungs-Array-Technologien untersucht. Die Subgruppen-spezifischen und rezidiv-spezifischen lncRNAs wurden durch differentielle Expressions (DE) Analysen mit LIMMA Voom analysiert. Durch die Analyse der Koexpression von lncRNAs mit Protein-kodierenden (PC) Genen aus allen Subgruppen schlossen wir unter Verwendung eines ‚Guilt-by-association' -Ansatzes auf potentielle Funktionen der DE lncRNAs. Zudem haben wir die Subgruppen-spezifischen lncRNAs auf einem unabhängigen Datenset von 47 BCP-ALL-Proben validiert. Die epigenetische. Die epigenetische Regulation von Subgruppen-spezifischen lncRNAs wurde durch eine differentielle Methylierungs (DM) analyse identifiziert. Die Korrelation zwischen DM und DE lncRNAs aus drei Subgruppen wurde ermittelt, um den Einfluss der epigenetischen Regulation auf die Expression von lncRNAs zu analysieren.

Ergebnisse: Wir präsentieren eine umfassende Landschaft von lncRNA-Signaturen, die drei molekulare Subtypen von BCP-ALL auf DNA-Methylierungs- und RNA-Expressionslevel klassifiziert. Die Hauptkomponentenanalyse (PCA) auf den top variablen lncRNAs auf RNA und DNA-Methylierungsniveau bestätigte eine robuste Trennung von Ph-like, DUX4 und NH-NeH BCP-ALL Subtypen. Mit integrativer bioinformatischer Analyse, zusammen 1564 subtyp-spezifische und 941 rezidiv-spezifische lncRNAs aus den drei Subtypen. Das unüberwachte hierarchische Clustering auf diesen Subtyp-spezifischen lncRNAs validierte ihre Spezifität in der unabhängigen Validierungskohorte. Unsere Studie zeigt erstmals, dass BCP-ALL-Subtyp-spezifische sowie Rezidiv-spezifische lncRNAs zur Aktivierung von Signalwegen wie TGF-β, PI3K-Akt, mTOR und Aktivierung von JAK-STAT-Signalwegen von DUX4 und Ph-like Subtypen. Endlich wurden die signifikant DM subtyp-spezifische lncRNAs profiliert. Darüber hinaus identifizierten wir 23 Subtyp-spezifische lncRNAs, die ein Hypo-

und Hypermethylierungsmuster in ihrer Promotorregion zeigen, das signifikant mit ihrer verringerten und erhöhten Expression in den jeweiligen Subtypen korreliert.

Schlussfolgerungen: Insgesamt liefert unsere Arbeit die umfassendsten Analysen für lncRNAs in BCP-ALL-Subtypen. Unsere Ergebnisse weisen auf eine Vielzahl von biologischen Funktionen im Zusammenhang mit lncRNAs und epigenetisch erleichterten lncRNAs in BCP-ALL hin und bieten eine Grundlage für funktionelle Untersuchungen, die zu neuen therapeutischen Ansätzen führen könnten.

ABSTRACT

Introduction: B-cell precursor acute lymphoblastic leukemia (BCP-ALL) is the most prevalent heterogeneous cancer in children and adults, with multiple subtypes. Emerging evidence suggests that long non-coding RNAs (lncRNAs) might play a key role in the development and progression of leukemia. Thus, we performed a transcriptional and DNA methylation survey to explore the lncRNA landscape on three BCP-ALL subtypes (82 samples) and demonstrated their functions and epigenetic profile.

Methodology: The primary BCP-ALL samples from bone marrow material were collected from diagnosis (ID) and relapse (REL) stages of adult (n = 21) and pediatric (n = 24) BCP-ALL patients, using RNA-seq and DNA methylation array technology. The subtype-specific and relapse-specific lncRNAs were analyzed by differential expression (DE) analysis method using LIMMA Voom. By analyzing the co-expression of the subtype-specific lncRNAs and protein-coding (PC) genes from all subtypes, we inferred potential functions of these lncRNAs by applying "guilt-by-association" approach. Additionally, we validated our subtype-specific lncRNAs on an independent cohort of 47 BCP-ALL samples. The epigenetic regulation of subtype-specific lncRNAs were identified using the Bumphunter package. The correlation analysis was performed between DM and DE lncRNAs from three subtypes to determine the epigenetically facilitated and silenced lncRNAs.

Results: We present a comprehensive landscape of lncRNAs signatures which classifies three molecular subtypes of BCP-ALL on DNA methylation and RNA expression levels. The principle component analysis (PCA) on most variable lncRNAs on RNA and DNA methylation level confirmed robust separation of DUX4, Ph-like and NH-HeH BCP-ALL subtypes. Using integrative bioinformatics analysis, subtype-specific and relapse-specific lncRNAs signature together determine 1564 subtype-specific and 941 relapse-specific lncRNAs from three subtypes. The unsupervised hierarchical clustering on these subtype-specific lncRNAs validated their specificity on the independent validation cohort. For the first time, our study demonstrates that BCP-ALL subtype specific as well as relapse-specific lncRNAs may contribute to the activation of key pathways including TGF-β, PI3K-Akt, mTOR and activation of JAK-STAT signaling pathways from DUX4 and Ph-like subtypes. Finally, the significantly hyper-methylated and hypo-methylated subtype-specific lncRNAs were profiled. In addition to that, we identified 23 subtypes specific lncRNAs showing hypo and hyper-methylation pattern in their promoter region that significantly correlates with their diminished and increased expression in respective subtypes.

Conclusions: Overall, our work provides the most comprehensive analyses for lncRNAs in BCP-ALL subtypes. Our findings suggest a wide range of biological functions associated with lncRNAs and epigenetically facilitated lncRNAs in BCP-ALL and provide a foundation for functional investigations that could lead to novel therapeutic approaches.

Chapter 1. Introduction

1.1 History of long non-coding RNAs (lncRNAs)

The flow of genetic information through and by messenger Ribonucleic acid (mRNA) came into light through the paper "Genetic Regulatory Mechanisms in the Synthesis of Proteins," in 1961 (Jacob & Monod, 1961). Since then, a myriad of studies discovered a large variety of RNA of different size and shape (Figure 1.2.1). Jacob & Monod postulated in their paper that lncRNAs resemble mRNA, yet they do not encode protein. Instead, lncRNAs facilitate a wide variety of mechanisms which regulate the production of gene products such as other RNAs or proteins. Today, lncRNAs have emerged as a critical layer in the genetic regulatory code. Proceeding studies and biochemical experiments were able to characterize the abundant structure and regulatory RNAs by locating their cellular localization and sequence similarity. Genetic studies identified a few lncRNAs involved in genomic imprinting and other cellular processes. For example, *XIST*, *H19* and *AIR* (Rinn & Chang, 2012). Collectively, all these classical studies identified a diverse range of RNA, but they only superficially looked on the cell surface for functions of all those identified RNAs.

1.2 Definition of lncRNAs

The new century has started with the completion of the Human genome project and discovered numerous new RNA encoding genes but no new protein-coding genes, which revealed a biological mystery about human genome: The human genome comprises only about 2% of protein-coding genes, and the rest is non-coding RNAs. The non-coding RNAs are subdivided into two types, small non-coding RNAs and long non-coding RNAs (lncRNAs). The small non-coding RNAs are microRNAs and other RNAs. The lncRNAs were defined as RNA genes >= 200 base pair (bp) in length and either no or short open reading frame (ORF). The definition is somewhat arbitrary because some small regulatory RNAs are higher than 200 nucleotides in length. Although this definition is arbitrary, the threshold separates lncRNAs from other small regulatory non-coding RNAs such as microRNAs (miRNAs) or Piwi-associated small RNAs (piRNAs) (Encode & Consortium, 2007).

The advent of full genome sequencing enabled prospecting for new "genes", which surprisingly led to the discovery of more RNAs than protein-coding genes. For instance, the number of human microRNAs (miRNAs) quickly increased from a few to nearly thousands. Transcriptome analysis by arrays and RNA sequencing (RNA-Seq) studies have demonstrated that a significant portion of the transcriptome consists of lncRNAs. However, by the discovery of next-generation technologies the scenario has been changed, and now lncRNAs are being studied widely on both molecular and genetic level because of their significant functions in a variety of disease and normal tissues/cells.

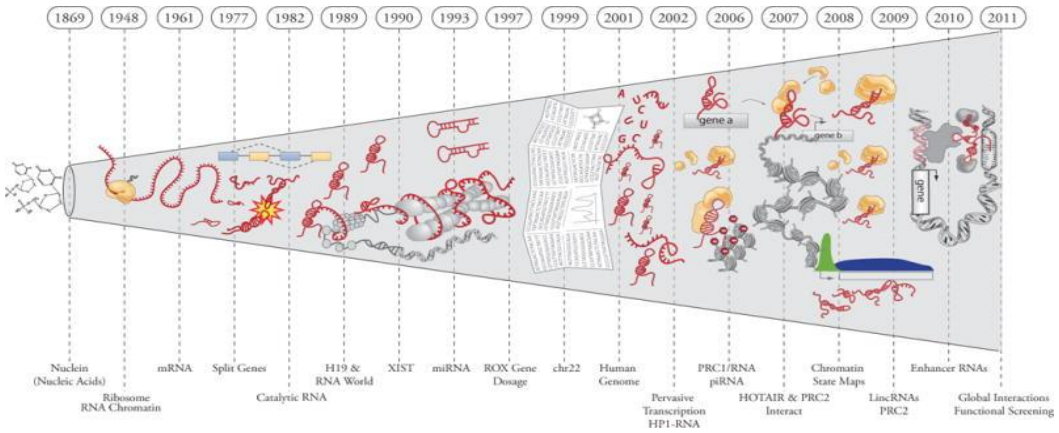

Figure 1.2.1: The time flow of the lncRNAs discovery.

The figure represents the discovery flow of lncRNAs from the time when nucleic acid was discovered until 2011. Adapted from (Rinn & Chang, 2012).

1.3 Genomic features and classification of lncRNAs

Most, but not all lncRNAs are transcribed by RNA polymerase II and are capped and polyadenylated at their 5′ and 3′ ends respectively (Rinn & Chang, 2012). LncRNAs are often defined by their location in the genome. Most of them are found near protein-coding (PC) genes, e.g. within exons of PC genes, introns of genes, and in intergenic regions. The classification of lncRNAs based on their anatomy in the genome. The biotypes of lncRNAs are, antisense, the lncRNAs that overlap PC genes in the opposite strand, sense intronic lncRNAs that are encoded within introns of PC genes, and sense overlapping lncRNAs are termed based on their transcripts overlapping PC genes. The lncRNAs located between PC genes are named long intergenic non-coding RNAs (lincRNAs) (Figure 1.3.1) (Atianand & Fitzgerald, 2014). Most of the lncRNAs have multiple exons and are subjected to alternative splicing, but they have fewer exons than PC genes.

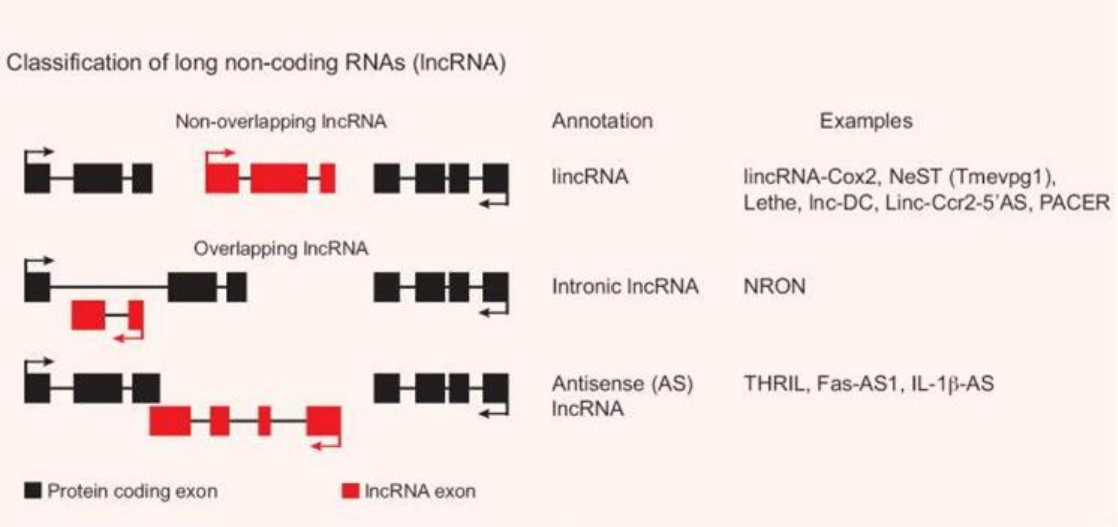

Figure 1.3.1: The classification of lncRNAs.
The anatomical definition of long non-coding RNAs (lncRNAs), based on their location within transcriptome. The diagram represents, lincRNA, intronic, Antisense and sense overlapping lncRNAs. This diagram is adapted from (Atianand & Fitzgerald, 2014).

1.4 Identification of lncRNAs

Currently, there are no standard criteria for identification of lncRNAs and most researchers use arbitrary thresholds to define lncRNAs. A widely accepted definition is based on the ORF size and was defined by the FANTOM (Functional Annotation of Mouse) project where they defined a threshold of 100 codons, to separate lncRNAs from other mRNAs genes (Kawai et al., 2001). However, the classification criteria of lncRNAs are straightforward and practical; they are subject to false positives and false negatives. For example, the XIST lncRNA in the murine cell line is approximately 15kb in size and contains 298 amino acids in ORF, which were mistaken for the protein-coding genes (Borsani et al., 1991). Various approaches can be applied to rationalize this problem.

The task of defining and annotating or separating lncRNAs from mRNAs is complex and suffers from the lack of specific defining criteria. The methods including machine learning approach and sequence conservation methods sonly provides an estimate of the likelihood that an RNA sequence is coding or non-coding. Such a dichotomous classification into mRNAs and lncRNAs might have little biological relevance as there isn't necessarily a clear distinction between the two classes. In a real-world point of view, the fact that RNAs with an exclusive coding or non-coding function are only the two extremes of a continuous process. Therefore, a definitive answer for coding and non-coding potential can only be observed by investigating the proteome experimentally in the wet lab.

1.5 Technologies used in the identification of lncRNAs

Identification of lncRNAs are based on all the transcripts obtained from the cell including coding, non-coding, and mRNAs isoforms. Advanced microarray technologies and RNA-Seq can be used for identifying lncRNAs within the cell. RNA-Seq, in contrast, is not only limited to the identification of known lncRNAs but also novel unannotated lncRNAs.

1.5.1 Microarray technologies in lncRNA identification

Conventional microarray technologies use predestined probes to find the expression level of mRNA transcripts and are not able to identify new lncRNAs. Nevertheless, it has been found that a few previously defined probe sequences are lncRNAs; therefore, microarray data analysis requires re-annotation of the probes in order to study the expression of lncRNAs. New specific probes for lncRNAs can be designed with the discovery of new and more lncRNAs. For example, some study groups designed probes matching conserved regions (intergenic and intergenic region) to identify potential non-coding RNA (nc-RNA) transcripts (Babak, Blencowe, & Hughes, 2005). However, micro-arrays are limited due to the low expression level of many lncRNAs.

1.5.2 RNA-Seq in identification of lncRNAs

The arrival of the deep sequencing technology led to the ability to sequence cDNA (derived from RNA), using the technology called RNA-seq, a high throughput and dynamic sequencing method with the unparalleled scale of data production. These approaches have been coupled to computational methods allowing the reconstruction of transcripts and their isoforms at single nucleotide resolution (Trapnell, Pachter, & Salzberg, 2009). The studies have provided an unbiased identification of non-coding transcripts across many cell types and tissues (Guttman et al., 2010). RNA-seq is widely used for discovery of novel transcripts and gene expression analysis. Advancement of RNA-seq, allowed consortia to define all the transcribed genes in the genome and to release broad catalogs. For instance, the GENCODE project released one of the complete evidence-based human reference genomes based on RNA-seq analysis on multiple cell types. The catalog consists of more than 15787 lncRNAs in the latest version (GRCh38) (Mudge & Harrow, 2015).

RNA-Seq has many advantages in studying gene expression, compared to microarray. RNA-seq more sensitive in detecting less-abundant transcripts, identifying novel alternative splicing isoforms and novel nc-RNA transcripts. Alternative splicing (AS) is a process by which exons or portions of exons or non-coding regions within a pre-mRNA transcript are differentially excluded or included, resulting in multiple RNA isoforms being encoded by a single gene on the DNA. Taking advantage of the ever-increasing

depth of sequencing and read lengths has allowed some of the first steps towards characterizing lncRNAs on a global scale. RNA sequencing has been utilized to estimate transcript abundance and to identify specific properties of distinct classes of large RNA genes in order to catalog them in a functional atlas by incorporating novel lncRNAs (Iyer et al., 2015). For example, a recent study identified 8,000 large intergenic non-coding RNAs (lincRNAs) in the human genome by integrating numerous annotation sources in combination with RNA sequencing (Arrial, Togawa, & Marcelo, 2009). This study revealed several global properties of lncRNAs, including investigating tissue-specific expression patterns, determining thousands of orthologous lincRNAs between human and mouse, and locating lncRNAs in gene deserts (the regions in the genome without any protein-coding genes) associated with the genetic trait. RNA-seq is now the gold standard method to discover lncRNAs, but a significant challenge with these data is their interpretation. Sequence reads commonly harbor multi-mapping potential, especially for lncRNAs whose DNA sequence is overall less conserved and harbors a higher degree of repetitive elements. Thus, stringent filtration and rigorous analysis are required to eliminate spurious transcripts.
Other methods to identify lncRNAs and characterize their function, are: RNA immunoprecipitation (RIP) sequencing, RIP-Seq is a protein centric approach used to find the association of specific protein with RNAs or non-coding RNAs, which uses a protein as bait to pull-down RNAs. However, the RIP-Seq approach has its limitations, for example, the task of differentiating the direct or indirect interactions between protein and RNA is difficult. In addition to that, the read length of associated RNAs are too large for identifying the actual binding sites. Finally, the assays used for RIP-Seq technology are known for having variability. Thus, multiple biological replicates are necessary.

1.6 Functions of lncRNAs

In contrast to the significant progress made in identifying and classifying lncRNAs, the functional role and mechanisms of lncRNAs remained mostly unknown. However, during the last decade, researchers investigating the role and functions of lncRNAs have exceedingly increased and made clear that lncRNAs have a broad spectrum of specific functional features in various biological processes. By now it is clear that some of these lncRNAs participate in various biological processes such as regulation of gene expression both in *cis* and *trans*, genome imprinting, X-inactivation, development, differentiation, and cell cycle regulation (Kitagawa, Kitagawa, Kotake, Niida, & Ohhata, 2013).
As of 2016, a literature-based lncRNAs database called lncRNAdb has shown 294 functionally annotated lncRNAs (Amaral, Clark, Gascoigne, Dinger, & Mattick, 2011). Below, I summaries the different types of functions carried out by lncRNAs using representative examples.

1.6.1 LncRNAs regulates the expression of their *cis* genes

LncRNAs exert their functions mainly in combination with co-expressing with their nearby (*cis*) and distant (*trans*) protein-coding genes (Guil & Esteller, 2012) (Ali et al., 2018). LncRNAs interact with genes in the same genomic loci are termed as *cis*-lncRNAs, while *trans* acting lncRNAs interact with genes on same or on different chromosomes. Recently, there were several reports of lncRNAs co-expression with its nearby protein-coding genes in several diseases and differentiation stages (Delás & Hannon, 2017). The c*is*-regulatory lncRNAs are mainly transcribed from the same promoters and enhancers of protein-coding genes, as well as from the antisense transcripts. Among these, antisense lncRNAs are epitomized due to their transcription regulatory activity at the *cis* region. Reports from FANTOM consortium suggested about 20% of transcribed PC gene has antisense lncRNAs (Kiyosawa et al., 2003). The antisense lncRNAs exert their function on their corresponding sense PC by influencing their genes expression at different levels, including transcriptional interference, and translation regulation. The following are a few examples of *cis*-regulatory lncRNAs.

Transcriptional interference: Transcriptional interference is mainly through epigenetic interaction, and through impacting PC genes. One of the best-studied examples is the antisense lncRNA *ANRIL*, which contributes to cancer initiation by reducing senescence through protein interaction contributing to the repression of tumor suppressor genes. For example, *ANRIL* is encoded by CDKN2B-AS1 which is expressed at the CDKN2B-CDKN2A gene-cluster locus. The CDKN2B-CDKN2A gene-cluster locus encodes three major tumor suppressor genes, *P14*, *P15* and *P16*, whose expression is subject to Polycomb group protein control. The antisense *ANRIL* has been shown to interact with the CBX7 protein, which is a component of the polycomb receptor factor 1 (PRC1), which can recognize H3K27me3 repressive marks on the genome. The CBX7 protein uses different regions within its domain for binding to H3K27me3 and antisense *ANRIL*; reports suggest that both interactions are important for sustained repression of the CDKN2B-CDKN2A gene-cluster locus (Qiu et al., 2016).

Translational regulation: Antisense lncRNAs exert their functions as a translational control over the sense region of PC genes. For example, the antisense lncRNA *BACE1-AS*, increases the stability of its sense PC gene BACE1 through the formation of RNA duplex in the ~100-nt region. Antisense lncRNA *BACE1-AS* acts as a positive regulator of BACE1 protein by preventing the mi RNA-induced silencing. BACE1 is a protein being present at higher levels in brains of Alzheimer's patients (Faghihi et al., 2008). On the other hand, the *trans*-acting lncRNAs may act as signals, guides or scaffolds to chromatin to regulate the expression of target genes located in the distant chromosomal domains or even at different chromosomes. The following session describes the functional properties.

1.6.2 Transcriptional regulatory functions of lncRNAs in *trans* region

The actual transcriptional regulatory functions of lncRNAs remain mostly unknown. Currently, based on the evidence and functionally characterized lncRNAs, the transcriptional regulation of lncRNAs serve mainly as a signal, decoy, guide, scaffold, and enhancer during the transcriptional process (Ma et al., 2012).

Signal: The transcription of individual lncRNAs occurs at a particular time and place to incorporate developmental evidence, interpret cellular context, or respond to diverse stimuli. Thus, the lncRNAs can serve as molecular signals at the transcription process (Figure 1.6.1 A).

Decoy: The lncRNAs are capable of acting as decoys to DNA-binding proteins such as transcription factors, chromatin modifying proteins or enhancers (Groen, Capraro, & Morris, 2014). The mode of action is mainly through the sequence homology to the target gene, such as these lncRNAs can prevent and bind their interaction with target genes by acting as bait to their specific effector proteins (Figure 1.6.1 B).

Guide: These lncRNAs guide the localization of ribonucleoproteins to specific target sites (Figure 1.6.1 C).

Scaffold: LncRNAs act as a scaffold by interacting with multiple components and activate or repress transcription. LncRNAs can bind with two or more protein partners, in which lncRNAs serve as a device to form functional protein complexes (Figure 1.6.1 D).

Enhancer: Using chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq), it has shown that gene-activating enhancers give rise to lncRNA transcripts, known as enhancer RNAs (Visel et al., 2009). In addition to that, their expression level positively correlates with that of nearby PC genes, predicting that lncRNAs are more likely to regulate mRNA synthesis. Along the same line, another Loss-of-Function study found that of 7 out of 12 lncRNA knockdowns affects the expression of their neighboring primal genes (Ørom et al., 2010). The lncRNAs also function as an activator of nearby genes via their “enhancer” function. These lncRNAs are from other genomic regions than enhancers which are called as enhancer RNA-like (eRNA) lncRNAs (Figure 1.6.1 E).

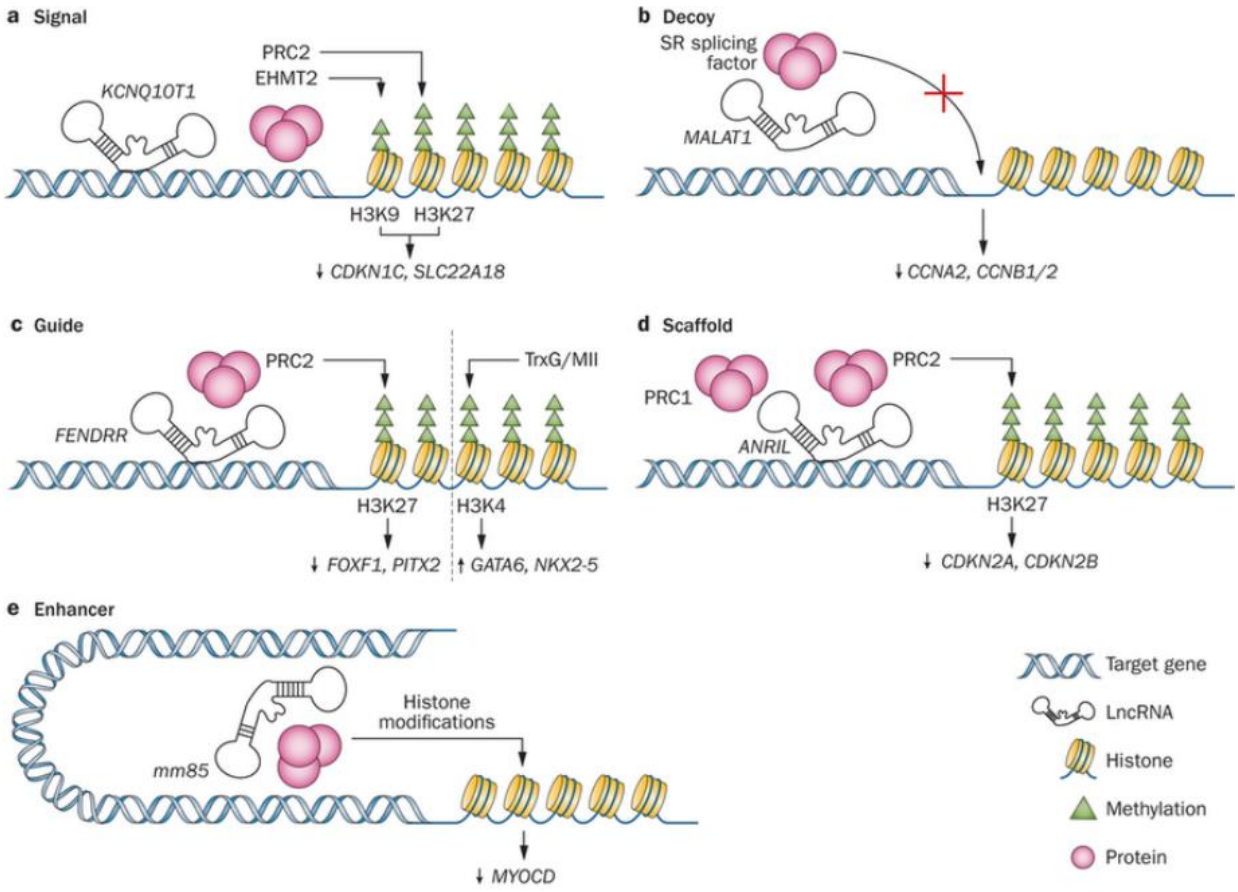

Figure 1.6.1: Molecular functions of lncRNAs
A. Signal: The figure shows the lncRNA *KCNQ1OT1* which induces transcriptional silencing by recruiting histone-lysine N-methyltransferase (EHMT2) and polycomb repressive complex 2 (PRC2) to a specific active site through chromatin methylation. B. Decoy: In the figure, alternative splicing is regulated by the lncRNA *MALAT1* by trapping the serine and arginine amino acid residues (SR, proteins involved in RNA splicing). C. Guide: *FENDRR* either silences or activate gene expression by forming a complex with PRC2 and with TrxG/All proteins respectively. D. Scaffold: Chromatin methylation is modulated by *CDKN2B-AS1* (also known as *ANRIL*) by binding to PRC1 and PRC2. E. Enhancer lncRNAs acts through chromosomal looping by an interaction between enhancer and promoter regions of genes, and it modulates target gene expression. Abbreviations: HxKy, histone (number x) lysine (number y); TrxG/Mll, trithorax-group/mixed lineage leukemia. Figure adapted from (Devaux et al., 2015).

1.7 Epigenetic gene regulation

The most studied lncRNAs expression regulation is on the epigenetic level (C. Wang et al., 2017). Epigenetic modification is heritable changes in genome leading to change in gene function without changing DNA sequences. As RNA is an integral component of chromatin, many regulatory lncRNAs can function by interacting with chromatin modifiers and re-modelers to change the epigenetic status of the target gene. Chromatin modification is one of the epigenetic processes, in which the chromatin architecture is modified. The modification is to allow access of condensed genomic DNA to the regulatory transcription machinery proteins, and thereby control gene expression. Rising information

convey that some lncRNAs 'guide' chromatin-modifying complexes (Khalil et al., 2009) as well as other nuclear proteins to specific genomic loci to utilize their effects (P. Han & Chang, 2015). Critical epigenetic regulations of lncRNAs are highlighted in the following session.

1.7.1 LncRNAs involved in chromatin-modifications

Many lncRNAs were initially characterized based on their repressive functions, including *ANRIL*, *HOTAIR*, *H19*, *KCNQ1OT1*, and *XIST* (Bhat et al., 2016). The repressive function of these lncRNAs is achieved by coupling with histone modifying or chromatin re-modeling protein complexes. The most common chromatin modifying complexes coupled with these lncRNAs are the polycomb repressive complexes 1 and 2 (PRC1 and PRC2). These complexes facilitate the chromatin compaction and heterochromatin formation in order to enact repression of gene transcription by transferring repressive post-translational modifications to specific amino acid positions on histone tail proteins. (Leeb et al., 2010).

Nearly 20% of lncRNAs are estimated to bind with PRC2 (Khalil et al., 2009). However, the biological meaning of this observation is not yet clear, it is possible for PRC2 to bind promiscuously with lncRNAs in a non-specific way. Nevertheless, if lncRNAs are predominantly functioning in the *cis*-regulatory mechanism, then the PRC2 binding is to facilitate local gene expression through the genome. Examples of these category lncRNAs include *ANRIL* and *XIST*. Likewise, PRC1 proteins, especially heterochromatin protein 1 (or CBX) proteins, have been involved in ncRNA-based biology.

1.7.2 LncRNAs in genomic imprinting and X chromosome inactivation

Genomic imprinting is an epigenetic phenomenon where epigenetic marks at specific loci are set, based on the sex of the parent of origin of the chromosome, and usually leads to expression of genes from only one chromosome. The transcription and post-transcription-based gene regulation by lncRNAs can be studied using genomic imprinting. In addition to that, reports suggest that imprinted lncRNAs may fine-tune gene expression of protein-coding genes to maintain their dosage in the cell (Kanduri, 2015). The *XIST* lncRNA is one of the classical examples in chromatin modifying lncRNAs. The lncRNA *XIST* mediates the chromatin regulation leading to the X chromosome dosage compensation in mammals. Briefly, dosage compensation refers to the process of equalizing the gene expression level of two X chromosome in the female cell to the single X in male cells (Brockdorff & Turner, 2015).

1.8 DNA Methylation and lncRNAs

DNA methylation is a fundamental form of epigenetic modification and serves multiple significant

functions, such as repression of gene transcription, maintaining genomic integrity, establishing, and repression of transposable elements (Moore, Le, & Fan, 2013). DNA methylation involves the addition of methyl group to cytosines. The genome contains CpG-rich regions, known as CpG island, which is often located at the promoter and first-exon regions. Usually, these regions are un-methylated, but when they are methylated, it blocks the transcription of related genes. LncRNAs have recently discovered as novel regulators of gene expression at the epigenetic level (Y. Zhao, Sun, & Wang, 2016). There are emerging evidence establishing the interplay between lncRNAs and DNA methylation (Y. Zhao et al., 2016). Recent studies have demonstrated several similarities in the methylation dynamics between protein-coding genes and lncRNAs, including, the TSS methylation distribution, relationship between promoter and gene expression (Li et al., 2017). One of the critical steps in epigenetic regulation during standard development programs is the establishment and maintenance of methylation patterns resulting in modulation of gene expression. Such processes are facilitated by several DNA methyltransferases (DNMTs). A recent publication from Chalei and colleagues reports one such example lncRNA which they demonstrated the lncRNA termed as *Dali*. The lncRNA *Dali* is expressed in the central nervous system. This lncRNA is essential for neural differentiation and to regulate neural gene expression partially through interacting with DNMT1 (Chalei et al., 2014). This interaction then affects DNA methylation at distal target promoters.

In addition to the functions mentioned above, another molecular mechanism of lncRNAs are, they are highly tissued specific compared to PC genes (K. C. Wang & Chang, 2011). Recently, research groups have been studying the expression of lncRNAs in the global remodeling of the epigenome and during reprogramming of somatic cells to induced pluripotent stem cells (iPSCs). The study revealed certain lncRNAs have high cell specificity regarding gene expression (Huo & Zambidis, 2013). Another study on loss-of-function of most lincRNAs expressed in mouse embryonic stem (ES) cells showed that the knockdown of lincRNAs has a major outcome on gene expression patterns, which are equal to the effects of knockdown of known ES cell regulators (Guttman et al., 2011). These studies prefigured that lncRNAs might play significant roles in regulating the developmental process. Off late, the ENCODE project analysed 31 cell types for finding the tissue specificity of lncRNAs, and they found that many lncRNAs have specific expression pattern in brain cells (Quan, Zheng, & Qing, 2017). The emerging lines of evidence suggest that any dysregulation of these lncRNAs expression can be linked to a variety of human diseases from neuron diseases to cancer or tumours (Tang et al., 2013). All these studies indicate the involvement of lncRNAs in human diseases can be more dominant than thought before.

Though considerable research development has been made since the discovery of lncRNAs, the challenge to elucidate the functions of lncRNAs remains. Unlike PC genes whose mutation would bring a drastic change in the phenotype, mutations in lncRNAs often do not cause a significant phenotype (Mattick, 2009). Also, another cause to it is that lncRNAs are more likely to function at a specific condition or specific developmental process, and so condition-specific studies of lncRNAs are necessary. With the massive amount of omics data, described lncRNAs are accumulating, and therefore for their functional predictions, computational approaches have been used to design the experimental studies and brisken the understanding of lncRNAs.

1.9 LncRNAs in cancer

Cancer is one of the leading causes of death around the world, which is about for 8.8 million (World health organization, WHO) in 2015. Understanding the underlying causes of cancer has drastically changed over the last decade. The progress in sequencing technologies has shown that cancer-associated loci cannot only be in protein-coding regions, but also in non-coding regions (Schmitt & Chang, 2016).

LncRNAs are studied widely in solid tumors, especially in breast cancer (Soudyab, Iranpour, & Ghafouri-Fard, 2016; Xu, Kong, Chen, Ping, & Pang, 2017). In breast cancer, the over expression of lncRNAs *HOTAIR* promotes the metastasis by epigenetically silencing the developmentally essential genes in the HOXD cluster (Gupta et al., 2010). LncRNAs are thus known as the functional transcripts which add on to the significant characteristics of cancer, and therefore they can be potential therapeutic targets. The comprehension of lncRNAs with the development of sequencing technologies has enabled lncRNAs in detailing their expression, function, and distribution in the human genome.

By now, we know that lncRNAs are a highly heterogeneous group of transcripts, which modulate gene expression using different mechanisms. Accordingly, some of them are found to be differentially expressed in various solid cancers, and they are directly linked to the conversion of healthy cells into tumor cells and thus represent an important factor of tumor biology.

1.10 Hallmarks of cancer

According to Hanahan and Weinberg, in their paper, “The hallmarks of cancer.” they proposed six hallmarks which collectively contribute towards the fundamental principle of malignant transformation (D Hanahan & Weinberg, 2000). These basic hallmarks are:

- Self-sustained growth signalling

- Insensitivity to growth inhibition
- Avoiding apoptosis
- Uncontrolled proliferation
- Promotion of angiogenesis
- Tissue invasion and metastasis

Two additional emerging hallmarks according to 2011, are the capability to modify or reprogram, cellular metabolism in order to most effectively support neoplastic proliferation. The second one is cancer cells to evade immunological destruction, in particular by T and B lymphocytes, macrophages, and natural killer cells (Douglas Hanahan & Weinberg, 2011).

1.11 LncRNAs in cancer hallmarks

Self-sustained growth signaling: LncRNAs promote self-sufficiency by activating/stabilizing the expression of growth factor receptors thereby enhancing signal transduction in response to the growth signals/ factors. There are multiple lncRNAs serve as receptors. For example, lncRNA *SRA*, serves as a scaffold to stabilize estrogen receptor (Lanz et al., 1999). In addition to activating signal receptors, some lncRNAs affect proliferation by regulating receptor abundance lncRNA, for example, lncRNA *PVT1* (Zhou, Chen, Feng, & Wang, 2016).

Insensitivity to growth inhibition: LncRNAs can regulate growth inhibition mostly by influencing the tumor suppressor genes that regulate cell cycles such as cyclins, CDK inhibitors, and tumor suppressor, P53 (Kitagawa et al., 2013). The process is mainly by repression of the transcription through PRC complex. Certain other lncRNAs regulate the expression of tumor suppressor gene by influencing various parts of transcription and translation. The scaffolding of transcriptional factor complexes can influence transcription initiation. Finally, the transcript stability and translation can be modulated post-transcriptionally by reducing the role of miRNAs. For example, *PTENP1* is acting as competitive endogenous RNA to inhibit miRNAs repression of *PTEN,* tumor supressor gene (L. Yang, Wang, Shen, Feng, & Jin, 2017).

Avoiding apoptosis: Apoptosis refers to the controlled cell death, one of the key pathways to control in carcinogenesis. Reports showed that some lncRNAs act on regulation of transcription of the essential apoptosis gene. LncRNA *INXS* is an example, it is expressed from the intron of B-cell lymphoma-extra large (*BCL-X*, is an anti-apoptotic protein) gene and regulates its splicing into a pro-apoptotic isoform

BCL-XS (Deocesano-Pereira et al., 2014). Another discovery is lncRNA *PRAL*, which induces apoptosis by stabilizing the complex between heat shock protein 90 (*HSP90*, assist protein to fold correctly) and *P53*. However, their mechanism of action remains unknown.

Uncontrolled proliferation: Proliferation is the potential of cancer cells for limitless replication. The maintenance of telomeres as nucleo-protein structures that stabilizes ends of chromosomes is a key factor for the proliferation of cancer cells. In the dividing cells, the telomeres shorten, so it takes a ribonucleoprotein complex telomerase to elongate the telomeric repeats through reverse transcription of an internal template RNA. The shortening of telomeres induces the production of lncRNA Telomere repeat-containing RNA (*TERRA*) (Redon, Reichenbach, & Lingner, 2010), which is transcribed from the sub-telomeric regions. Under normal conditions, *TERRA* inhbit its own expression through chromatin modifications, but recruits protein complexes for homology-directed repair of shortened or damaged telomeric sequences when activated.

Promotion of angiogenesis: Angiogenesis is the process of formation of new blood cells from existing ones. Angiogenesis can be a support for tumor cells to grow and migrate (Folkman, 1974). There are a few lncRNAs which regulate nutrient supply to tumor, mostly by regulating the expression/ function of VEGF (vascular endothelial growth factor), which is essential for the production of blood vessels. LncRNAs *MIAT* are reported to transcriptionally regulate VEGF. Knockdown of *MIAT* showed that it is required for the repression of VEGF, which resulted in microvascular dysfunction and decreased metastasis (B. Yan et al., 2015).

Tissue invasion and metastasis: Metastasis is the process by which cancer cells spread to distant parts of the body from its tissue of origin. Several reports showed that multiple lncRNAs increase the capacity of the cancer cell to invade new sites and therefore facilitate metastasis. *MALAT1* is an example lncRNA which facilitates the invasiveness of cancer cells in colorectal and nasopharyngeal carcinoma (M. H. Yang et al., 2015). Other example is, lncRNA, lincRNA-RoR which acts as a "sponge" for miR-145 which regulates ADP-ribosylation factor 6, a protein involved in the invasion of breast cancer cells (Eades et al., 2015).

1.12 Translational Implications of lncRNAs in cancer

Cancer therapy is facing the challenge of cancer cell specificity and delivering anti-cancer drugs without interfering with normal cells functions. Profiling the differential abundance of lncRNAs may assist cancer diagnosis and prognosis and furnish useful information regarding potential therapeutics (Qi &

Du, 2013). Moreover, lncRNAs are detectable from minute amounts of biological fluids like urine, blood and serum using qRT-PCR amplification making it as a diagnostic marker (Geng, Xie, Li, Ma, & Wang, 2011). For example, the highly up-regulated in liver cancer hepatocarcinoma-associated lncRNA (*HULC*) can be readily detected in the blood of HCC patients using qRT-PCR (Panzitt et al., 2007). Another example is *PCA3*, is a lncRNA that is prostate-specific and markedly over expressed in prostate cancer. Although its biological function is unclear, lncRNA PCA3 can be utilized as a biomarker in diagnostic assays for prostate cancer (Van Gils et al., 2007).

Finally, lncRNAs are an attractive therapeutic option considering their tissue-specific or cell-specific expression pattern. For example, the expression of the lncRNA, *H19* elevated in a wide range of human cancers. A plasmid, BC-819 (*DTA-H19*), has been developed to make use of this tumor-specific expression of *H19 (Smaldone & Davies, 2010)*. Intra-tumoral injections of this plasmid induce the expression of high levels of diphtheria toxin specifically in tumor, resulting in tumor size reduction in human trials. Recent studies have yielded promising results in a wide range of solid cancers including, colon, and bladder, pancreatic and ovarian cancers. Therapeutic application of lncRNAs provides an attractive treatment prospect, although still more intensive research is required. The current era of lncRNA research is giving rise to a new field within the biology of hematopoiesis and blood diseases.

1.13 Leukemia

Leukemia is mainly diagnosed based on the number of blasts typically quantified by blood tests. The exact cause of leukemia is still unknown. However, it seems to develop from a combination of genetic and environmental factors. Studies indicate both inherited, and environmental factors are involved in the formation of leukemia.

Acute leukemia is a type of leukemia occurring mostly in bone marrow characterized by the massive accumulation of immature white blood cells. These immature white blood cells are also known as blasts or leukemic cells. For instance, the risk factors are smoking, ionizing radiation, prior chemotherapy, and Down syndrome. The environmental factors including, artificial ionizing radiation, chemicals and smoking influences the genome which leads to different genetic factors leading to leukemogenesis. The genetic factors of leukemogenesis are described in the following session.

1.13.1 Leukemogenesis

The occurrence of leukemia is due to the uncontrolled proliferation of hematopoietic stem cells in the bone marrow when there is an alteration in normal cell regulatory processes (Davis, Viera, & Mead,

2014). The most common alterations in genes regulating blood cell development or homeostasis are the following:

DNA translocations: Translocations means that a part of one chromosome breaks off and becomes attached to a different part of the same chromosome or in a different chromosome altogether.

Inversions, or deletions: The deletions of the transcription factors which are essential for the normal hematopoietic development. Hematopoietic development is a normal process of immature blood cell development into all type of mature blood cells, including white blood cells, red blood cells, and platelets. For example, deletion of *IKZF1*, which is linked to crucial function in hematopoietic system its loss of function leads to lymphoid leukemias.

Mutations: The alteration of the nucleotide sequence of the genome. In leukemogenesis, two types of mutations must occur for leukemia formation one is, a mutation which improves hematopoietic cells ability to proliferate which includes *FLT3* and *KIT*. The second type is a mutation that prevents the cells from maturing including *CBFB-MYH11*.

1.13.2 Major types of Leukemia

Based on the type of bone marrow cells that are affected, leukemia can be classified into different types (Table 1.13.1). Leukemia can arise in two different types of white blood cells, myeloid and lymphoid white blood cells. When leukemia is affected in lymphoid precursor cells it is called acute lymphoblastic leukemia, and when affected in myeloid cells, it's classified as myeloid leukemia.

Table 1.13.1: The types of leukemia

Types of leukemia	Definition
Acute Myeloid Leukemia (AML)	AML arose from immature myeloid cells. Myeloid cells are the cells that make white blood cells (other than lymphocytes), red blood cells, megakaryocytes (platelet-making cells).
Acute Lymphocytic Leukemia (ALL)	ALL arises from the immature forms of lymphocytes, thus known as lymphoid or lymphoblastic leukemias. This is one of the most common leukemia in children and affects adults.
B-cell Acute Lymphoid Leukemia (BCP-ALL)	BCP-ALL is a heterogeneous disease associated with different patterns of molecular changes including protein fusions, mutations and copy number variations
T-cell precursor lineage (T-ALL)	T-ALL is biologically distinct from its counterpart, B-ALL. T-ALL shows a different dynamic form of disease response.
Chronic Myelogenous Leukemia	CML is defined by increased proliferation and differentiation of the

(CML)	granulocytic cell line. It is a myeloid proliferative disorder.
Chronic Lymphocytic Leukemia (CLL)	CLL often occurs in adults above or equal to 55 years old. In very few cases it affects young adults.

Table 1.13.1: The table contains the different types of leukemia based on their lineage and the pace of occurrence (Vardiman et al., 2009).

1.14 B-cell Acute Lymphoid Leukemia (BCP-ALL)

The present thesis is focused on B-cell based acute leukemia affecting the lymphoid cell, B-cell precursor ALL. B-cell precursor acute lymphoblastic leukemia (BCP-ALL) remains a major cause of death in pediatric patients. BCP-ALL is a heterogeneous disease associated with different patterns of molecular changes including protein fusions, mutations and copy number variations. The major chromosomal alterations are aneuploidy, the abnormal number of chromosomes, and chromosomal rearrangements, which results in oncogene deregulation or expression of chimeric fusion genes (Mullighan, 2012).

1.15 The subtypes of BCP-ALL

BCP-ALL comprise of multiple subtypes which are defined based on the structural chromosomal alterations, Somatic mutations and DNA copy number alterations that contribute to leukemogenesis. The alterations are prevalent in all age groups and so as the various subtypes (Figure 1.15.1). Identification of these subtypes is essential for diagnosis, risk classification, and, for some lesions, it enables the development of targeted therapy.

The subtypes investigated to profile their lncRNAs based molecular signature in this project are described in the following section.

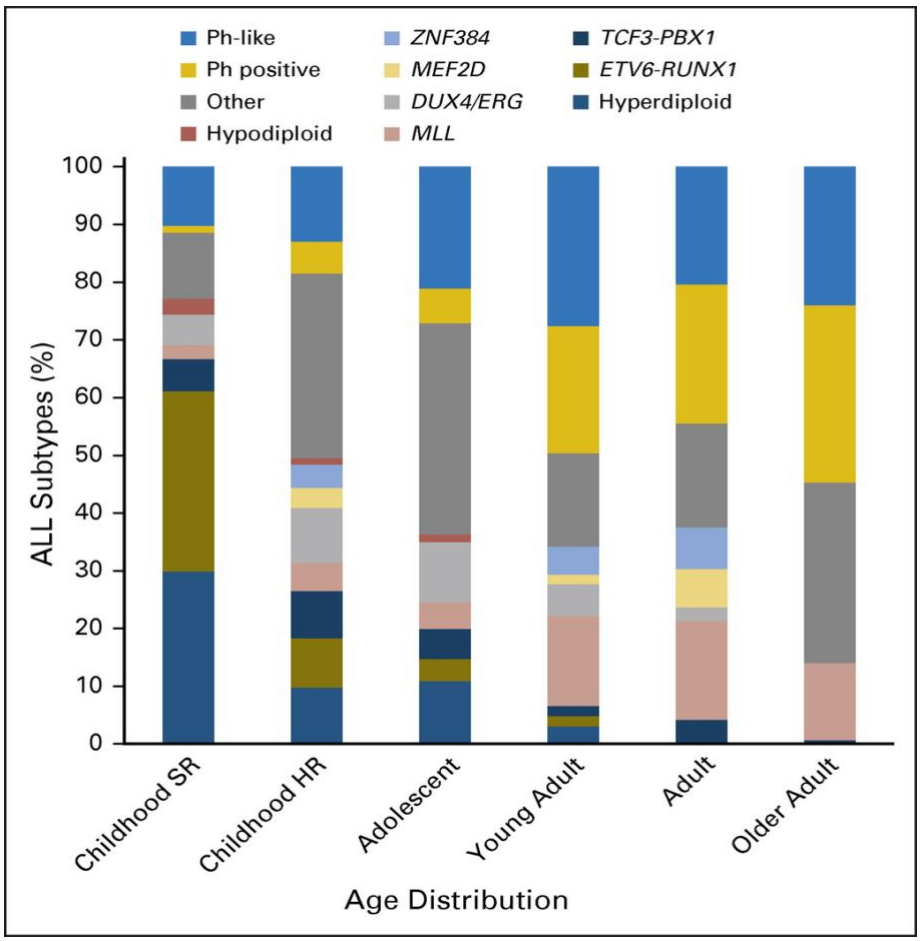

Figure 1.15.1: Subtypes in ALL across different age groups.
The figure represents different subtypes in ALL which is varying with different age groups (Iacobucci & Mullighan, 2017).

1.15.1 Double homeobox 4 (DUX4) BCP-ALL subtype

The DUX4 is a recently discovered subtype within BCP-ALL which is characterized by the IGH-DUX4 gene fusion and is prevalent in both adult and pediatric patients of BCP-ALL. The existence of DUX4 subtype was first hinted in a microarray dataset study on childhood BCP-ALL patients, where a subset of cohort displayed a unique expression profile outside the well-established subtype. The same group further performed an integrated genomic analysis on 277 ALL cases to investigate the genetic basis of this novel subtype (Yeoh et al., 2002). A recent study (Clappier et al., 2012) revealed that around 50-70% of these cases showed deletions in the intragenic region of erythroblast transformation (ETS)-specific-related gene (ERG). The ERG, a gene coding for a transcription factor in ETS family, with important functions in hematopoiesis. The genomic aberration observed was approximately non-existent in other BCP-ALL cases. Later other studies found that deletion of ERG is associated with CD2 expression and Ikaros family zinc finger protein 1 (IKZF1) deletions with a positive clinical prognosis, which is

otherwise associated with a poor prognosis (Harvey et al., 2010).

In the vast majority of cases at least one truncated copy of the DUX4 gene is usually located within subtelomeric region and is inserted (D4Z4 repeat array on chromosome 4q and 10q) into the Immunoglobulin heavy (IGH) locus. ERG-DUX4 is a less common variant involved in the insertion of DUX4 gene into an intron of the ERG gene. In both variants (IGH-DUX4 and ERG-DUX4) a 3' truncated DUX4 transcript with nucleotides added from non-coding regions of IGH or ERG is expressed, resulting in a DUX4 protein replaced with random 0-50 amino acids from non-coding partner genes in the same region. The relocation of DUX4 attributes to the truncation of C terminal of DUX4 protein and increased stability of DUX4 mRNA due to the presence of poly-A signals in the partner region (Lilljebjörn & Fioretos, 2017).

The DUX4 transcription factor is normally expressed in germinal tissues, and its expression is partially regulated by the repeat structure of D4Z4 domains, where a certain number of repeats are needed to preclude the luxated DUX4 expression. Currently, it is unclear how the expression of DUX4 fusions contributes to leukaemia development. In pediatric BCP-ALL 4-5% of the cases harbour DUX4 rearrangements, making it the sixth largest subtype of childhood BCP-ALL, slightly larger than Ph-positive subtype (Lilljebjörn & Fioretos, 2017).

Despite the common ERG deletions, DUX4-rearranged cases might also harbour other common aberrations associated with various other subtypes of BCP-ALL, such as deletions in targeting cell cycle regulator genes *CDKN2A* and *CDKN2B* and lymphoid transcription factor genes such as *IKZF1* and *PAX5*.

1.15.2 Philadelphia positive (Ph-pos) BCP-ALL subtype

The Philadelphia chromosome is a result of the molecular fusion between the ABL gene, which is located on the chromosome 9 with BCR gene located on the chromosome 22, which results in a fusion protein called BCR-ABL (Liu-Dumlao, Kantarjian, Thomas, O'Brien, & Ravandi, 2012). BCR-ABL encodes an oncogenic protein with a constitutively activated tyrosine kinase function. The prevalence of BCR-ABL positive ALL, also called Philadelphia (Ph)-positive, increases with age and occurs in up to 50% of ALL diagnosed in individuals ≥50 years old (Liu-Dumlao et al., 2012). The Ph-positive ALL is characterized by poor response to therapy, short remission duration and poor survival. The occurrence of BCR-ABL fusions is 2–5% in in pediatric ALL, and is approximately 25% in adults with ALL (El Fakih et al., 2018).

1.15.3 Philadelphia-like (Ph-like) subtype

Recently a high-risk subgroup of BCP-ALL called Philadelphia-like (*Ph-like*) has been discovered in pediatric and adult patients (Herold & Gökbuget, 2017). The Ph-like blasts harbor a similar gene expression profile as BCR-ABL1 positive ALL patients but lack the BCR-ABL1 translocation (Tran & Loh, 2016). However, instead of BCR-ABL like gene fusion, such patients harbor a wide range of genetic alterations activating tyrosine kinase signaling. Most common genomic features of these patients are deletions of *IKZF1* transcription factor and genetic alterations deregulating cytokine receptor and tyrosine kinase signaling (Tran & Loh, 2016). These include translocations and mutation of CRLF2 of approximately 50%, 12% translocations of ABL-class tyrosine kinase genes, 7% of rearrangements of JAK2 and 3-10% of the erythropoietin receptor gene (EPOR). Furthermore, 11% mutations activating JAK-STAT signaling and RAS signaling (*NRAS*, KRAS, PTPN11, and NF1, 6%) and less common kinase alterations (FLT3, NTRK3, BLNK, TYK2, and PTK2B). Kinase fusions continue to keep an intact tyrosine kinase domain and typically show a constitutive kinase activation. There is no significant difference in frequency of kinase subtypes across different age groups, apart from EPOR and JAK2 rearrangements which are increased in adult Ph-like ALL. Cytokine receptor-like factor 2 (*CRLF2*) is also known as the thymic stromal-derived lymphopoietin receptor (*TSLPR*) that forms a heterodimeric receptor with the interleukin-7 receptor a chain (*IL7Ra*) for thymic stromal lymphopoietin (TSLP). Dysregulations of CRLF2 includes its translocation into the immunoglobulin heavy chain locus (IGH-CRLF2) and less common point mutations. All these rearrangements are most common in Ph-like and Down syndrome-associated ALL and are dependent on age. For instance, with P2RY8-CRLF2 associated with young age and I-CRLF2 associated with older age and Hispanic ancestry. Flow cytometric immunophenotyping detects CRLF2 and is over expressed on the leukemic lymphoblasts. CRLF2 rearrangements are associated with poor prognosis in most studies, particularly in cases with concurrent IKZF1 alteration (Iacobucci & Mullighan, 2017).

The common therapies have shown efficacy in per-clinical models that targets JAK-STAT, PI3K/mTOR, and BCL2 signaling alone or its combinations. Another major genetic subgroup within Ph-like ALL involves ABL class of rearrangements. For example, fusions to ABL1, ABL2, CSF1R, PDGFRA or PDGFRB that are all targetable by inhibitors of ABL1, such as imatinib and dasatinib. Like Ph-positive ALL, Ph-like ALL is also associated with high-risk clinical features such as poor response to induction chemotherapy, elevated minimal residual disease (MRD) levels, and poor survival. According to world health organisation's classification of myeloid neoplasms in 2016, BCR-ABL1–like or Ph-like ALL acute leukaemia was recognized as a new leukaemia entity with clinical importance due to its association with

an unfavourable prognosis and reactivity towards Tyrosine Kinase Inhibitor (TKIs). Ph-like ALL increases with age and varies from 10% in standard-risk childhood ALL to greater than 20% in adult ALL, with a peak prevalence of 27.9% in young adults (age 21 to 39 years) (Iacobucci & Mullighan, 2017).

1.15.4 Near haploid/High hyperdiploid (NH-HeH) BCP-ALL subtype

The abnormal chromosomal number in ALL defines distinct subtypes with different response to treatment. High hyperdiploid is a subtype defined based on cytogenetic nomenclature as chromosomal count between 47 and 57; the definition criteria are universally accepted. High hyperdiploid is one of the common childhood malignancies comprising 30% of all pediatric B cell–precursor ALL. Molecularly, high hyperdiploid ALL is characterized by massive aneuploidy (abnormal number of chromosomes), authenticating a nonrandom gain of chromosomes. For example, some or all of +X, +4, +6, +10, +14, +17, +18, and +21 and other trisomies have been reported. However, the pathogenetic phenomenon of chromosomal gains remains poorly understood, but it generally is believed that gene dosage effects are of significance (Chilton et al., 2014). Genetic abnormalities like driver fusion gene is not observed in the vast majority of high hyperdiploid ALL cases. However, there is a possibility that there is yet unidentified primary aberrations present due to the low resolution of most genetic screening techniques. Previously such concealed events have been reported in aneuploid tumors, for example, the identification of structural dysregulation resulting in rearrangements of cytokine receptor-like factor 2 (*CRLF2*) in a large number of ALL patients with Down syndrome and microdeletions leading to the transmembrane protease, serin 2 (TMPRSS2)/v-its erythroblastosis virus E26 oncogene homolog (ERG) hybrid gene in prostate cancer (Mullighan et al., 2009). Profiling of a fusion gene in high hyperdiploid ALL would be of prima facie clinical importance, which may perhaps simplify the diagnostic procedures and hence provide novel treatment options. Clinical features of high hyperdiploid ALL was associated with a relatively low WBC count and a B-cell precursor immunophenotype. The prognosis of five-year overall survival rates (OS) is close to 90%.

Recent genome-wide association studies by two independent groups reported linkage to a locus in the gene AT rich interactive domain 5B (ARID5B) at the locus 10q21.2, however, it is unclear how this region affects the risk of developing high hyperdiploid childhood ALL (Studd et al., 2017) . Despite a favourable prognosis in high hyperdiploid childhood ALL, ~20% of the patients suffer a relapse, and 10% give in to the disease (Paulsson et al., 2010). The finding of extra recurrent changes could subserve in the identification of the high-risk cases and would be of great clinical significance.

In contrast, Near-haploid ALL is much rarer (<1%) ALL subtype defined based on the cytogenetic nomenclature of the 23-29 chromosome, with poor outcome (Safavi & Paulsson, 2017). The Near-haploid is mainly reported in children and adolescents. Lately, some adult cases are also reported. Because of the rarity of near-haploid ALL subtype, relatively very few studies have focused on this molecular subtype and no studies on lncRNAs are reported to this date.

In this present study, we are focusing on the three major subtypes defined above namely, DUX4, Ph-like and NH-HeH. In addition to the subtypes mentioned above, BCP-ALL has additional subtypes which are described briefly in the following session.

1.15.5 Pre-B-cell leukemia transcription factor 1 (PBX) fused

The translocation resulting in the Transcription factor 3 (TCF3) - PBX1 fusion occurs in approximately 5% of childhood to 6% of adult BCP-ALL cases. With the rise of novel therapies, it is now associated with a favorable outcome (Diakos et al., 2014).

1.15.6 Myocyte enhancer factor 2D (MEF2D) fused

MEF2D and zinc finger 384 (ZNF384) rearrangements characterize distinct B-ALL subtypes, accounting for approximately 3% to 4% and 3% of pediatric patients and approximately 6% and 7% of adult patients, respectively. The MEF2D related fusions are recently identified B-ALL subtype with relatively worse survival (Zhaohui Gu et al., 2016).

1.15.7 Mixed lineage leukemia (MLL) translocations

Mixed lineage leukemia (MLL/KMT2A) gene are common in young patients and are generally associated with poor clinical outcomes. The molecular biology of MLL fusion genes remains incompletely characterized and is complicated by the fact that more than 100 different partner genes have been identified in fusions with KMT2A gene (also known as MLL) rearrangements, particularly the t (4;11) (q21;q23) translocation, are most frequent in infants (1 year of age) and are associated with poor outcome (Winters & Bernt, 2017).

1.16 LncRNAs in leukemia

1.16.1 LncRNAs in normal hematopoiesis

Hematopoiesis is a process of formation of blood cellular components. All blood cells are derived from hematopoietic progenitor cells or hematopoietic stem cells. Hematopoietic progenitor cells are found in bone marrow which can form mature blood cells. The lncRNAs reported in normal hematopoiesis are

mainly associated with the blood cell development (B. W. Han & Chen, 2013) (Figure 1.16.1 A). Notably, two lncRNAs are associated with early blood cell development from hematopoietic progenitor cells. The first example is lncRNA HOX antisense intergenic RNA myeloid 1 (*HOTAIRM1*), in the intergenic region of the HOXA cluster and transcribed in the antisense direction. *HOTAIRM1* renders the expression of several genes that are important for myeloblasts differentiation including HOXA1 and HOXA4, which encode key transcription factors in hematopoiesis. The second example is lncRNA *EGO*, which is a conserved lncRNA transcribed antisense to ITPR1 and regulates the development of eosinophils (a type of white blood cell). Other lncRNAs which are related to the terminal differentiation of hematopoietic cells are ncRNA repressor of the nuclear factor of activated T cells (*NRON*), thymus-specific non-coding RNA (*Thy-ncR1*), and nettie Salmonella pas Theiler's (NeST), which regulate the localization, degradation, and expression of pivotal gene products, orderly. Hence, deficiency of any hematopoiesis-related lncRNAs blocks differentiation and stimulates the apoptosis of blood cell progenitors, deregulation of these lncRNAs might contribute to blood diseases, especially those associated with ineffective blood cell production, such as myelodysplastic syndrome (MDS) and anemias. Depending on the changes in lncRNA regulation, dys-regulation of these could also result in the oncogenic growth of the cells, resulting in cancer.

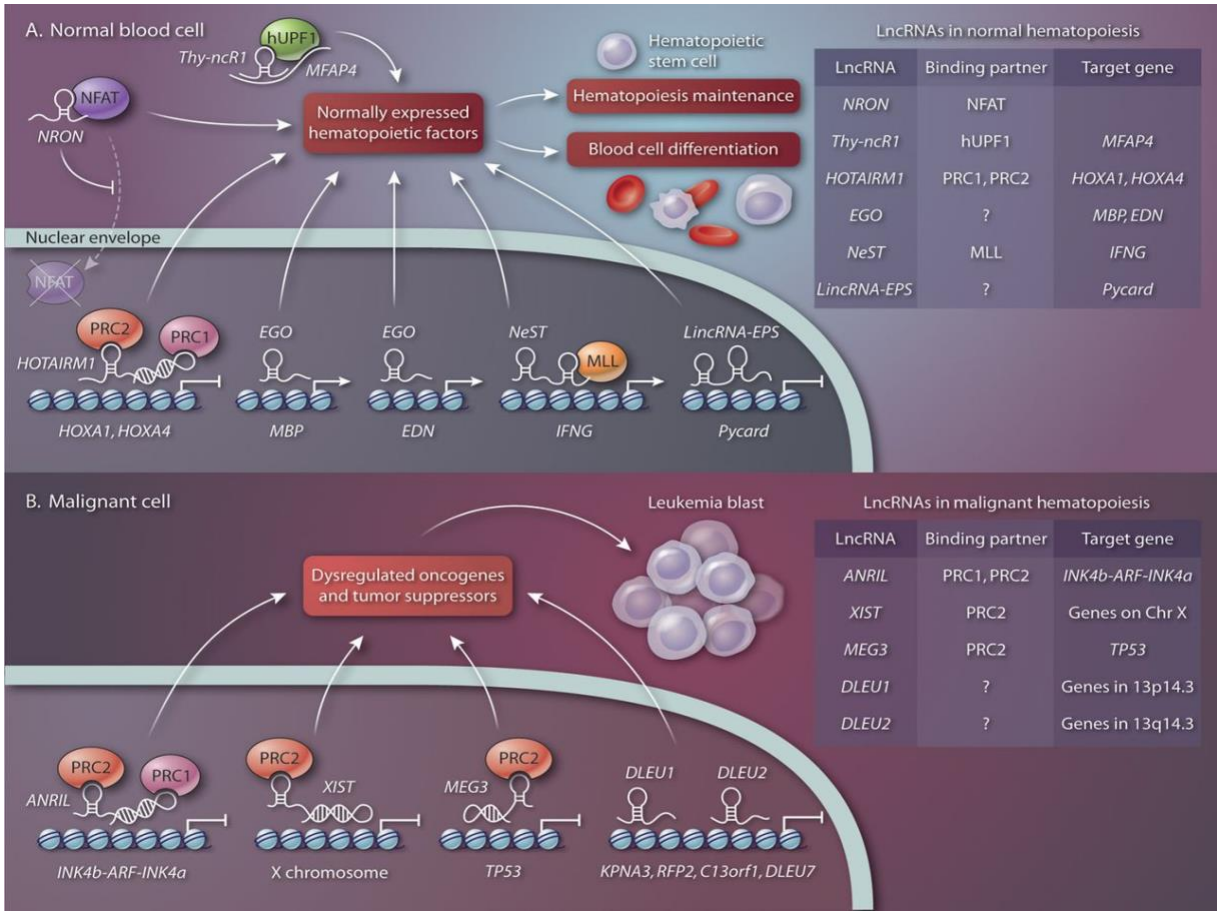

Figure 1.16.1: LncRNAs in normal and malignant leukemia
LncRNAs involved in the progression of multiple chromatin remodelling pathways and are involved in hematopoiesis and leukemogenesis. A. In healthy blood cells, lncRNAs are involved in recruiting epigenetic regulatory protein complexes, including chromatin remodeling enzymes to specific genetic target sites, by inhibiting or promoting mRNA translation or degradation, or by promoting or inhibiting the translocation of transcription factors into the nucleus, in normal hematopoiesis. B. In malignant cells, abnormally expressed lncRNAs lead to the deregulation of hematopoietic factors, resulting in an aberrant expression profile of oncogenes and tumor suppressors that leads to the pathogenesis of leukemia. Adapted from (B. W. Han & Chen, 2013).

1.16.2 LncRNAs in malignant hematopoiesis

Compared to solid cancers, little is known about lncRNAs in hematopoietic malignancies, especially in ALL subtypes. However, a few lncRNAs (Table 1.16.2, Figure 1.16.1B) have been identified in the context of leukemia and been directly involved in the tumorigenic processes of blood cell cancers (Hughes, Salvatori, Giorgi, Bozzoni, & Fatica, 2014). The before mentioned *HOTAIRM1* and *ANRIL* are two well-studied lncRNAs in malignant hematopoiesis. They might serve as potential prognostic candidates for leukemia due to their leukemic specific expression pattern. Increased *ANRIL* expression has been reported increased in many AML and ALL patients (Yu et al., 2008), whereas *MEG3* and *BIC*

expression were decreased in myeloid leukemia (Eis et al., 2005) and increased in B cell lymphoma, respectively. However, there are no lncRNAs reported as a leukemic prognostic marker.

The TP53 related lncRNAs such as *MEG3*, *lincRNA-p21*, are tumor suppressors whereas *ANRIL* is reported to function as tumor-promoting lncRNA. Several other lncRNAs have been shown to regulate key transcriptional factors that function in normal hematopoiesis, any dysregulation to these lncRNAs could contribute to malignant hematopoiesis. Some lncRNAs control other genes which participate in or regulate cancer-associated pathways. For instance, *DLEU1* and *DLEU2*, which are frequently deleted in Chronic Lymphoblastic Leukemia (CLL). In some CLL patients, these lncRNAs were shown a consistent alteration in methylation with reduced expression at transcriptional start site (TSS). In addition to that, they are correlated with transcriptional deregulation of the neighboring genes and reduced expression of genes involved in NF-kappa beta pathway (B. W. Han & Chen, 2013).

These lncRNAs associated with the tumor suppressor *TP53*, which act as regulators of the cell cycle or apoptosis and signaling pathways that are involved in leukemia hint that lncRNAs might take part in leukemogenesis either as tumor suppressors or as tumor promoters. Any dysregulation of these lncRNAs or others might furnish to the aberrant activity of leukemia-related genes, and that would further lead to malignant hematopoiesis. Specific lncRNAs having an association with particular forms of leukemia suggests that those lncRNAs may be useful for categorizing leukemia subtypes, and the possibility for therapeutic intervention may exist (B. W. Han & Chen, 2013).

By now we know that BCP-ALL is a heterogeneous blood cancer with multiple molecular subtypes, and a high relapse rate. Despite the improvements, we are still far from having a complete understanding of the rationale behind these subtypes. LncRNAs have emerged as a novel class of RNAs with diverse mechanisms in cancer progression and development. Recently, genome-wide association studies (GWAS) have unveiled that more than 80% of cancer-associated single-nucleotide polymorphisms occur in the non-coding part of the genome (Cheetham, Gruhl, Mattick, & Dinger, 2013). This suggests that a significant fraction of the genetic etiology of cancer is related to lncRNAs.

Moreover, the association of lncRNAs with various hallmarks of cancer in different cancer types shows that lncRNAs can account for cancer heterogeneity and can be used as an independent prognostic factor. However, lncRNAs defining molecular subtypes of BCP-ALL and their potential functions and epigenetic regulation are not portrayed yet. All previously reported ALL subtypes are well characterized and documented for their mRNA based molecular signature (Boer et al., 2015; Nordlund et al., 2012). Understanding the molecular signature beyond the protein-coding level that underlies BCP-ALL

heterogeneity thus remains as a significant objective in improving diagnosis and therapy. The extraordinary advancement in sequencing technology allowed the detection of low abundance transcripts on a genome-wide scale. Majority these studies explored the role of a specific single lncRNAs (Ghazavi et al., n.d.). However, these lncRNAs have not been precisely related to molecular pathways, and their functions have not been investigated. The long non-coding RNA based molecular signature behind these subtypes are less studied or characterized. Comprehensive characterization of the landscape of lncRNAs in a BCP-ALL subtype has not been achieved because most genome-wide studies have used micro-arrays, which have the disadvantage of being biased toward the inclusion of probes that map to the known protein-coding and lncRNAs transcriptome. Therefore, a comprehensive genomic delineation of lncRNAs alterations in multiple BCP-ALL subtypes not only is urgently needed but may lead to new diagnostic and therapeutic strategies for leukemia.

Table 1.16.2: LncRNAs which are reported as putatively involved in leukemia.

LncRNA	Size	Genomic location
MEG3	~1.6 Kb	Intergenic
HOTAIR	2.2 kb	Antisense between *HOXC11* and *HOXC12*
ANRIL	~3.9 kb	Antisense of *CDKN2B*
Mira	789 not	Between Hoxa6 and Hoxa7
LincRNA-p21	~3.1 kb	Upstream of *CDKN1A*
DLEU1 DLEU2	0.9 kb	Adjacent to miR-15 and miR-16 family
XIST	~19kb	Intergenic
HOTTIP	~3.8 kb	Bidirectional transcript with *HOXA13*

Table 1.16.2: The table defines the lncRNAs associated with malignant hematopeosis and their genomic target of action. Adapted from (82).

1.17 The aim of the project

The overall aim of this thesis was to profile relapse and subtype-specific lncRNAs in three significant subtypes of BCP-ALL namely, DUX4, Ph-like and NH-HeH to deepen our understanding of the functional role of lncRNAs in molecular processes in BCP-ALL. Furthermore, to investigate lncRNAs involvement in classifying the molecular subtypes of BCP-ALL. Finally, to investigate epigenetically regulated lncRNAs within the three subtypes.

In order to define lncRNAs within the BCP-ALL subtypes, we performed the integrative bioinformatics analysis on the RNA-seq and DNA methylation datasets of 82 BCP-ALL patient samples from diagnosis and relapse stages. This allowed to address the major aims of the thesis:

- Construct BCP-ALL subtype-specific and relapse-specific lncRNAs signatures
- Validate the subtype-specific lncRNAs of BCP-ALL on a independent cohort
- Define the potential functions of subtype-specific and relapse-specific lncRNAs
- Explore DNA methylation patterns of lncRNAs within the three BCP-ALL subtypes
- Unravel epigenetically altered lncRNAs within each subtype.

Overall, our data uncover the distinct mechanism of action of lncRNAs in BCP-ALL subtypes and defining how lncRNAs are involved in the pathogenesis of diseases as well as their relevance in the stratification of BCP-ALL subtypes.

Chapter 2. Materials and methods

2.1 Patient datasets

The patients used in the current study lacked routinely tested fusion genes (*BCR-ABL*, MLL translocations, *ETV6-RUNX1*) and were evenly distributed between pediatric and adult patients with early and late relapse. The sample was retrieved at initial diagnosis (ID), and relapse (REL) with the requirement of a minimal residual disease (MRD) level at complete remission below 0.01. The study group consisted of 45 patients with ID (40 samples) and matched REL (42 samples) stages of B-cell precursor ALL patients from German Multi-center Study Group ALL (GMALL) and Augmented Berlin-Frankfurt-Munster (BFM) trials. As the clinical protocol for adult and pediatric patients relapse time were different, the samples have been selected to evenly distribute into the categories of early relapse (ER; time of REL < 700 days) and late relapse (LR; time of REL => 700 days). Based on mutations, DNA methylation, translocations, and insertions or deletions these samples were further categorized into three subtypes (Table 2.1.3). These subtypes are, DUX4 (n = 23) (IGH-DUX4 fusions), Philadelphia-like (Ph-like (n = 21), and Haploid/High Hyperdiploidy (NH-HeH) (n = 16), low-hypodiploid (LH, n = 6) others (n = 16). There were 16 other samples are un-assigned because they do not belong to any of these subtypes. Patients and their clinical features are defined briefly in Table 2.1.3, which include their subtypes. All patients were treated in population-based German study trials (GMALL for adult and BFM for pediatric patients). All patients gave written informed consent to participate in these trials according to the Declaration of Helsinki. The studies were approved by the ethics board of Charité, Berlin.

Table 2.1.3: Patient clinical information and their subtypes.

Subtype	Patient	Median Age	Mean time to relapse (months)
DUX4	Adult	46	26
	Pediatric	9	83
LH	Adult	37	51
NH-HeH	Adult	22	40
	Pediatric	10	108

Ph-like	Adult	42	59
	Pediatric	14	65
Others	Adult	33	71
	Pediatric	10	25

Table 2.1.3: The table represents the patient samples (n = 82) used in this study, along with their defined subtypes and the mean of the months they were in complete remission, and their median age. Unassigned samples are hereafter refereed as others.

2.2 Major steps in RNA-Seq and DNA methylation array data analysis

2.2.1 RNA -Seq dataset preparation

RNA-Seq data preparation consists of following steps: we isolated total RNA from bone marrow mononuclear cells (MNCs) from ID, and REL conditions of BCP-ALL patients. We used Trizol reagent (Life Technologies, Grand Island, NY) for isolation and followed the manufacturer's protocol with minor modifications. Then the amount of RNA degradation is checked with gel, and capillary electrophoresis and an RNA integrity number (RIN) was assigned to all the samples. The samples with RIN greater than seven were then used for further steps.

RNA seq was performed on the Illumina HiSeq4000 platform. Paired-end reads were obtained from both ends of a fragment (paired-end sequencing). The paired-end reads were 101 base pair in length and bases with Phred +33 quality score were used further for analysis. These Phred quality scores Q are set by the Base Callers and are defined as $Q = -10\log10 (P)$, where P is the probability of the base call being incorrect. This value is an assigned quantity value to bases in DNA sequencing trace file by PHRED software. For example, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call, 1 in 1000 times, which means that the base call accuracy (that is, the probability of a correct base call) is 99.9%. These runs were performed in the high throughput sequencing core facility, German Cancer Research Center, Heidelberg, Germany. Overview of Bioinformatics analysis

We followed the alignment-counting work-flow (Figure 2.2.1) in our analysis pipeline. In order to account for the technical constraints of the tools, before finalizing each tool and analyzing method used

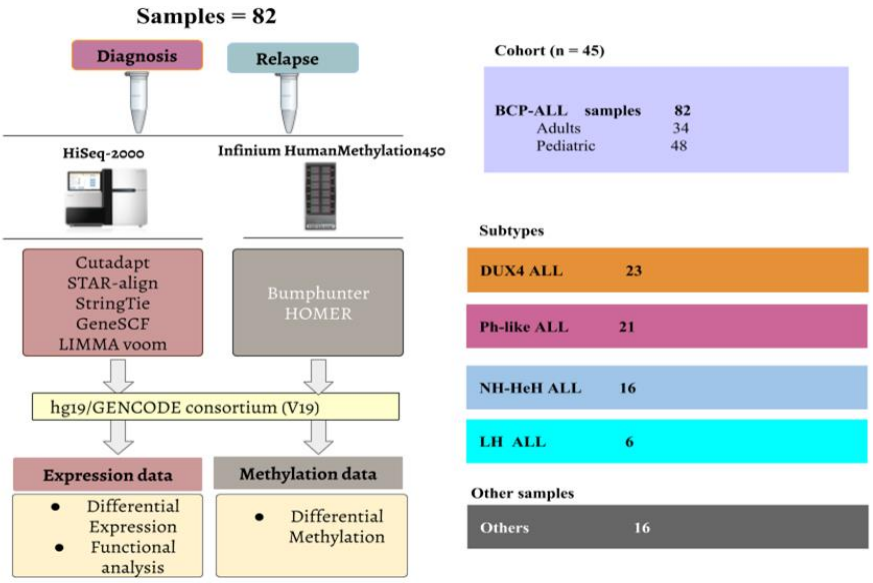

Figure 2.2.1: The global bioinformatics pipeline and the samples used in the analysis.
The diagram defines the DNA methylation and RNA-seq work-flow and important methods and tools in the pipeline. The boxes in right hand side defines the number of samples within each subtype and the total cohort used.

in this project we compared the performance and results with respect to other tools This analysis helped us to finalize the most optimized and suitable tools for our dataset and analysis. The tools used for each analysis and their versions are detailed in table 2.2.4. The reason behind the selection of each tool is defined in their respective sessions. All the analyses were performed in the UNIX environment (Ubuntu 14 LTS). The main steps involved in both dataset analysis is described in the following section.

Table 2.2.4: Bioinformatics tools and software used in analyzing RNA-Seq and DNA-methylation datasets
The table represents the tools and software and the programming languages we used in this thesis.

Tools	Description	Version
RNA-Seq data analysis		
Cutadapt	Trimming and removing low-quality reads	V1.17
RNA-Seqc	Quality metric	NA
Star-align	RNA-seq alignment	V2.4.0.1.
StrinTie/PreDE	Transcriptome assembly read quantification	v1.3.1
LIMMA Voom	Differential expression	NA

	analysis	
GREAT	*Cis* and *trans* genes located	3.0.0
GeneSCF	Functional enrichment	v. 0.1
DNA methylation analysis		
Bumphunter	Differential Methylation analysis	NA
HOMER/annotationPeaks.pl	Annotation of files	NA
Additional software and programming languages		
R-Bioconductor version (3) SHELL and BASH scripting Python 2 and 3 Pandas, Scify, numpy, matalblibplot, Seaborn		

Table 2.2.4: The table represents the bioinformatics tools used in the analysis and their versions. Additionally, the R-Bioconductor packages, and the scripting languages used in this thesis.

2.3 RNA-Seq data analysis

2.3.1 Preprocessing the Fastq files

The sequence reads from the Illumina HiSeq4000 machine came out in fastq (. fq) formatted files. The FASTQ format was first widely used in the Sanger Institute; it stores both biological sequence and its corresponding quality scores in the text-based format in the single file.

The first step of RNA-seq data analysis is to per-process the FASTQ file by measuring specific quality control metrics. The measured quality metrics are low-quality reads, followed by removing the 3' adapter sequence. The adapter sequence was from Illumina TrueSeq, we used the cutadapt tool to perform the trimming and removal of low-quality reads with the following parameters:

For removing the low-quality reads from the end of read sequences we used, *-quality-base = 33* (retains the Phred 33 quality bases) parameter with a cut-off of 5 (*--quality-cutoff = 5*). The read sequence which retained a length of >= 25 base pair, were filtered using the *-m 25* parameters. The adapter sequences were removed by adapter = [TACACTCTTTCCCTACACGACGCTCTTCCGATCT, GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT] parameter. The resulted FASTQ files were then further used for downstream analysis.

2.3.2 Read Alignment

Read alignment is the first, the most time-consuming step in RNA-Seq data analysis. The trimmed and quality controlled FASTQ files were then mapped to human reference genome version 19 (hg19) using

a mapping algorithm, STAR with options: *-twopassMode Basic –sjdbOverhang a*nd the Gencode (a project for integrated annotation of gene features) comprehensive GTF files as reference transcriptomes (gencode.v19.annotation.gtf.gz,v19) along with other default parameters. The STAR-align performed the read alignment steps in two runs, explained as follows:

In the 1st pass, STAR allows mapping to a genome with unknown junctions and extracts the novel read junctions, then insert them into the genomic index. The 2-pass alignment is a more advanced mapping strategy for a more accurate spliced alignments, where the reads are then re-mapped using the GENCODE annotation file, from the provided novel junctions reads. These novel junction reads were detected during the 1st pass alignment process. The Splice junctions are the exon-intron junctions, where the splicing takes place. The output from STRA-align tool was aligned FASTQ files in Sequence Alignment/Map (SAM) format. The read SAM format files sorted by coordinate were produced after the alignment process for each patient samples.

The SAM file is a TAB-delimited text file which consists of a header session and alignment session. However, SAM files are big and consumes more computational resource, mainly disk space, therefore they are generally stored as binary alignment mapping (BAM) files. The SAM files were then sorted and converted into BAM file using samtools sort. The BAM files store the same information as the SAM files in a smaller size due to compression and thus are more suitable for memory-efficient storage. The resulting BAMs were later used for further downstream analysis. All the runs are performed in parallel using shells scripts from 82 samples on UNIX platform.

2.3.3 Transcript assembly and read quantification

The transcript assembly and read quantification were performed using, StringTie, a fast-de-*novo* assembler. We used StringTie with default parameters, along with the *-e flag*. The *-e flag* was used to generate files which can be used to estimate the raw read counts. We performed StringTie runs input BAM file, GTF reference genome (*-G*, GENCODE annotation V19) along with other default parameters. Transcript abundance is first computed and from which the gene expression is inferred. StringTie reported read abundances in FPKM units. FPKM is commonly used for paired-end RNA-seq. In paired-end RNA-seq two reads can correspond to a single fragment. The relative abundances of transcripts are described in terms of the expected biological objects (fragments) observed from an RNA-Seq experiment. As StringTie assembles the transcripts and estimates its expression level simultaneously, the output from StringTie was many files: one contains the assembled transcript and the other contains the estimated expression in both FPKM and Transcripts Per Million (TPM) units, one assembles transcript

which in *gtf* format.

The assembled transcripts are quantified using the *PreDE* python script provided by *the StringTie* developers. Using the *PreDE* script we obtained the gene expression counts for each gene. The script summarized the FPKM values from all samples and converted it into raw read counts. In order to make a matrix of read counts of all samples, we parsed all the sample files using *os.path.join* module and then concatenated all the files based on their gene identifiers (Geneid) and expression values as a count matrix using *pandas concat* function.

This count matrix was later used for principal component analysis and differential expression analysis using the *R Bioconductor* package. The *R Bioconductor* packages for differential expression analysis takes in raw read count matrix mapped to a particular genome feature (example, *gene*) as their input. The raw read matrix was of gene identifiers (*gene* ID) as rows and their expression value of each sample as columns.

2.4 Reference genome and annotation files used

The lncRNAs were annotated using the GENCODE lncRNA annotation (V19), a manually curated and evidence-based lncRNA annotation consists of 13,860 lncRNA genes and its 23,898 transcripts. The GTF file was converted into the text file with gene identifiers, gene symbols, and gene biotype and their chromosomal position, extracted from GTF file using *awk* shell script. We then used a python script (*pandas merge*) to merge the "Geneid" between our count matrix and text file extracted from reference GTF file. The lncRNAs genes defined in GENCODE v19 version consists of 5276 antisenses with 9710 transcripts, 21 3prime overlapping non-coding RNA with 25 transcripts, 3055 lincRNAs with 3116 transcripts, 742 sense intronic with 802 transcripts, 202 sense overlapping genes with 330 transcripts and 515 processed transcript genes with 28082 transcripts, which made a total of 13,3860 lncRNAs. The same procedure was followed for identifying PC genes as well. The Gencode v19 version consists of 20,356 PC genes with 81,814 transcripts.

2.5 Unsupervised clustering using Principal Component Analysis (PCA)

We used a dimensionality reduction method PCA for unsupervised clustering on the expression and DNA-methylation of lncRNAs across all samples. We performed the principal component analysis on all 13,365 lncRNAs from all BCP-ALL RNA-seq samples. The PCA was performed using the R function *prcomp* on the FPKM values of lncRNAs. The R function *prcomp* uses the spectral decomposition approach, which examines the covariances or correlations between variables. The 3D PCA plots are

constructed using the python library *matlabplotlib* on the most variable principle components.

2.6 Identification of differentially expressed lncRNAs

The subtype-specific lncRNAs were identified from the whole cohort of 82 samples from 20 adult and 25 pediatric patients from ID and REL stages. We started off by analyzing the differential expression (DE) lncRNAs on factors including, age (Pediatric *versus* adult patients) and disease progression (diagnosis *versus* relapse) which did not result in distinct DE lncRNAs expression profile.

We next aimed to identify the DE lncRNAs or subtype-specific lncRNAs for the three subtypes, Ph-like, DUX4, and NH-HeH. For all samples, we applied a filter by removing the genes which are not expressed in at least 25% of samples to eliminate unreliability in the measurements of genes (Figure 2.7.1).

In the experimental design, we noticed that we have biologically dependent samples, for example, we had paired ID and REL samples from 45 patients, with some exceptions. Eight of 45 patients had no matching ID or REL. Moreover, the patient and subtypes were confounded. The R Bioconductor package *LIMMA Voom* can model the multi-stage scenario with its in-build duplicate correlation feature. Thus, we took advantage of it and used LIMMA Voom for the DE expression analysis.

We fed the raw read matrix as input to *LIMMA Voom,* which is then normalized using log2-counts-per-million (logCPM) approach. The *LIMMA Voom* uses *calcNormFactors* function for normalization. In the differential expression analysis and all related analysis, the raw count is rarely considered, mainly due to the varying library size, the libraries are sequenced at a greater depth and will result in higher counts. Thus, it is the norm to convert the raw counts into a scale that accounts for such library size differences. Some of the popular conversions are counts per million (CPM), log-CPM, reads per kilobase of transcript per million (RPKM), FPKM and TPM. *LIMMA Voom* started with the normalization of raw expression counts.

Normalization can significantly improve the quality of analysis and will lessen the bias across samples. Ideally, all samples are assumed to have similar distribution range of expression values. Normalization is used to ensure that the range of expression distributions of all sample are similar across the experiment. The normalization method LIMMA *Voom* employs is a trimmed mean of M-value (TMM) (Robinson & Oshlack, 2010). In *LIMMA Voom*, the normalization is performed by the *calcNormFactors* function. We used *boxplots* to visualize the difference of expression distribution of all samples' unnormalized count matrix (Figure 2.6.1 A) versus normalized count matrix (Figure 2.6.1 B). For our samples the effect of TMM-normalization is subtle, as shown in the magnitude of the scaling factors, which are all relatively

close to 1 (Figure 2.6.1 B). Normalization helps to make intuitive sense out of the data. Also, scaling enabled our data to be incorporated into the LIMMA method to conduct the DE analysis. The boxplots are constructed using *R graphical package boxplot*.

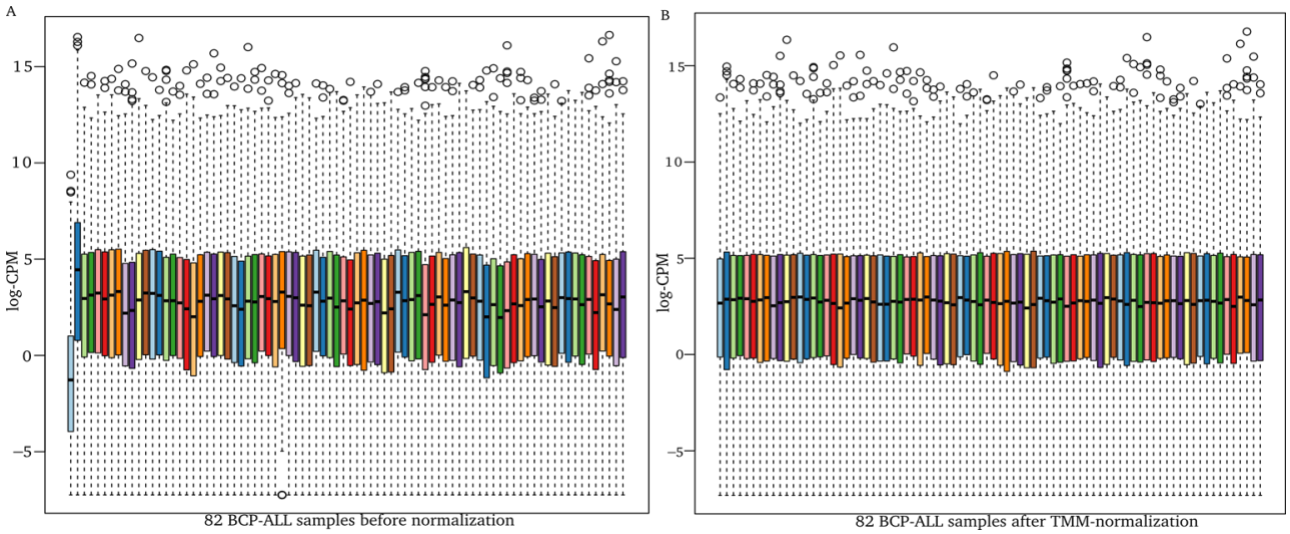

Figure 2.6.1: Box Plots of log-CPM values showing expression distributions for unnormalized data on the 82 BCP-ALL samples.
A. The unnormalized 82 BCP-ALL samples. B. The boxplot represents same dataset after TMM-normalization, showing a uniform expression distribution across all 82 samples.

2.6.1 Analysis matrix and contrasts

After normalizing the samples, the next step was to develop a design matrix. We studied three different subtypes of BCP-ALL: DUX4 (n = 23), Ph-like (n = 21), and NH-NeH (n = 16). The DE analyses on these three subtypes were performed separately where each subtype was compared to the all other samples (Figure 3.7.10). For example, when the DUX4 subtype is used for DE analysis, treatment group was formed by all DUX4 samples (n = 23) and the rest of the cohort (n = 59) as a control group. The same design was followed for the other two subtypes. For dysregulated relapse -specific lncRNAs, we used REL samples as the treatment group and ID samples as the control group within each subtype. Firstly, we started with setting up a design matrix using the model. The design matrix was formed by providing the information about the samples, including condition, control, and time (ID or REL). Contrasts for pairwise comparisons between paired samples can be set up in LIMMA using the make contrasts function. In our study design, we needed to include the following two complexities into the contrasts in order to avoid the bias including, biological dependence and time-dependence. We, therefore, needed to account for biologically dependent samples using the *duplicateCorrelation* function by

specifying patients as block argument. The *duplicateCorrelation* takes the expression matrix, the design variable, and specified block argument, that is, in our case-patient information. We then included time information also in contrast.

2.6.2 Examine DE genes from LIMMA

The differential expression analysis was done on two models. Firstly, between each subtype versus others (for subtype-specific lncRNAs) and secondly, within each subtype we looked for DE lncRNAs between ID and REL samples (for relapse-specific lncRNAs). The output from LIMMA Voom consisted of genes and its corresponding *P-value*s, which was determined using moderated t-statistic test, false discovery (FDR) value determined using Benjamini and Hochberg's method and log fold change of each gene.

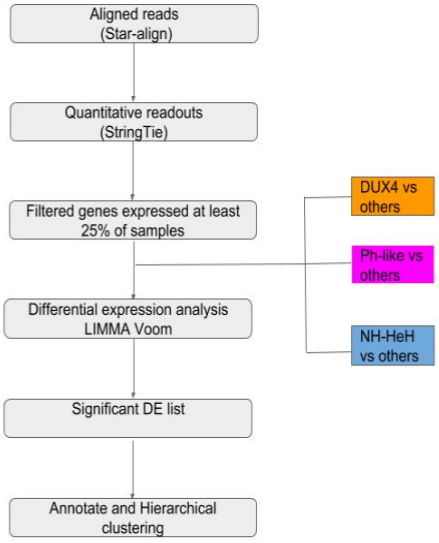

Figure 2.6.2: The DE subtype-specific lncRNAs identification workflow

The work-flow defines the steps taken to identify subtype-specific DE lncRNAs from the matrix of raw read count.

Next, the significantly differentially expressed genes were classified based on the following cutoffs: *P-value* <=0.01 and fold change ⇔+-1.5. That is, the genes above the fold change of +1.5 are up-regulated and below -1.5 are down-regulated.

In order to filter the lncRNAs from the output file using the above-mentioned cut-offs we wrote a python script using *pandas.DataFrame.query* function and for the annotation between the output matrix and GENCODE file we used *pandas merge* function. In order to find overlaps between subtype-specific lncRNAs, we used *Venn3* package from *matplotlib-venn*. The significant up and down-regulated lncRNAs are hereafter refereed to as subtype-specific and relapse-specific lncRNAs and these were then

used for the further downstream analysis.

2.7 Validation of the subtype-specific lncRNAs

An independent validation cohort of predefined BCP-ALL samples was used to validate our subtype-specific lncRNAs. The independent validation cohort consisted of 47 patients from the ID stage (age: median 32 years, range 1-80 years) (136). The samples were previously defined based on their genomic and molecular profile as DUX4 (n=17), Ph-like (n=27), and NH-HeH (n=3). We used our 1534 subtype-specific lncRNAs from our discovery cohort and performed an unsupervised clustering using complexheatmap R package using correlation-based clustering on lncRNAs expression and samples using the Spearman method. The column *barplot* represents the subtypes within the cohort, which was defined using *HeatmapAnnotation* function with *complexheatmap* R package.

2.8 Hierarchical cluster analysis

The graphical representation of high dimensional data sets is key for straightforward explanatory analysis and hypothesis generation. With genomics dataset, the most commonly used methods are heatmaps combined with hierarchical clustering. The hierarchical clustering builds a dendrogram (a tree-like structure) where the leaves are the samples or variables. The algorithm consecutively pairs together the samples showing the highest degree of similarity. These samples are then collapsed into a cluster and treated as a single object in all the following steps.

For the hierarchical cluster analysis and heatmap illustration, we used the subtype-specific and relapse-specific lncRNAs (Fold change <> +-1.5, P-value <= 0.01). The *LIMMA Voom* normalized expression of those both sets of lncRNAs were then transformed into row-based Z-scores.

$$Z - score = x - \mu \div \sigma$$

Where x is the normalized gene expression count, μ is the mean of the gene expression across the samples, and σ is the standard deviation of the gene expression. Hierarchical clustering (HC) was done on DE lncRNAs, the method is based on calculating the distance and correlation between each samples (columns) the genes (rows). The most correlated genes and samples from clusters visible in the heatmap. The distance-based correlation was performed using the Spearman method. We used the “complete” method to calculate the distances between clusters; this method uses the farthest distance between objects from the first cluster and objects from the second cluster. For the lncRNAs or rows we implemented *km =3, the* K-mean clustering split the rows for *3* clusters to define row-clusters or lncRNAs clusters. We used R function *FUNcluster* using *method = "silhouette"* method to define the optimal number of

clusters. The same approach was used for all subtypes and DNA methylation hierarchical clustering. Heatmaps provide large-scale qualitative views of the transcriptomic landscape by representing the quantitative differences in gene expression levels measured from RNA-seq or microarray technologies. We used R Bioconductor package *ComplexHeatmap* to plots our DE heatmaps and DNA methylation results (Zuguang Gu, Eils, & Schlesner, 2016).

2.9 Functional analysis by the guilt-by-association approach

LncRNAs can also positively or negatively regulate the expression of these *cis* genes located nearby, overlapping, or within protein-coding genes. We next aimed to determine functions of lncRNAs based their positive correlations with *cis* and *trans* lncRNAs. In our study, we used the "*guilt-by-association*" approach by establishing the correlations between the expression of lncRNA genes and their *cis* and *trans* PC genes. This method is most used and validated method regarding finding functions of lncRNAs. We

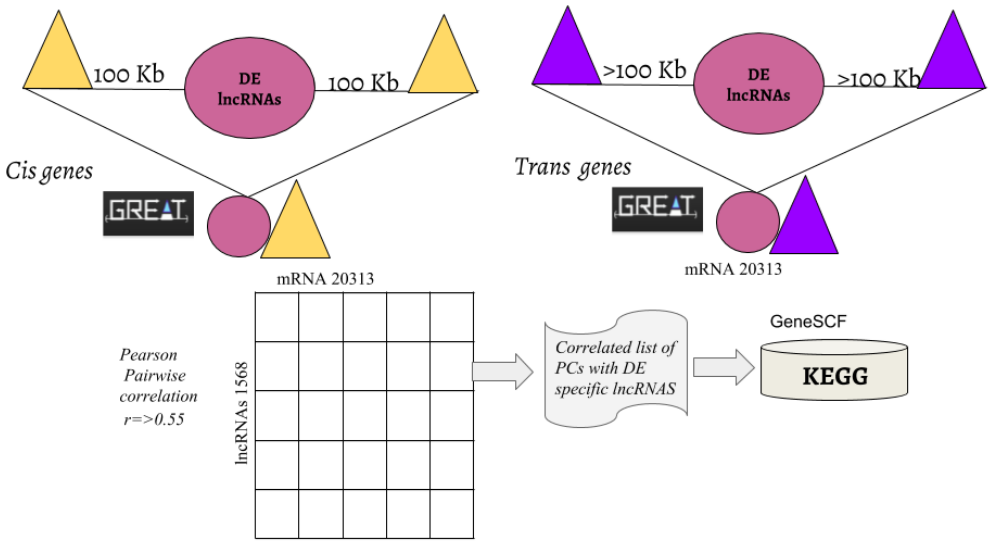

Figure 2.9.1: The work-flow used for functional predictions
Guilt-by-association approach used in functional predictions of differentially expressed lncRNAs from BCP-ALL subtypes. The circle indicates the lncRNAs and the yellow triangles are neighboring protein-coding genes. The purple triangles indicate *trans* protein coding genes which located in genomic location >100 kb distance or in another chromosome. Both *cis* and *trans* protein coding genes were located using GREAT interface. Then in the next step we calculated pairwise correlation using Pearson's correlation method, and the most significant genes were then used for functional enrichment analysis using GeneSCF tool.

located the *cis* and *trans* protein-coding genes for our subtype-specific lncRNAs, a set of 1564 lncRNAs from the three subtypes using the GREAT tool (McLean et al., 2010). GREAT is a graphical user interface (GUI), which accepts a list of background genes and test genes (Figure 2.10.1). Both the background

genes (~20,000 PC genes) and all subtype-specific (test genes) (n = 1534) lncRNAs were fed into the GREAT database in Browser Extensible Data (BED) file format. The input BED file contained the information of chromosomal position (chromosome, start, and end) then the gene symbol (Ensemble gene symbol) of lncRNAs and protein-coding (PC) genes. This BED file was constructed from the filtered subtype-specific lncRNAs text file using *sed* and *awk* one-liners on the command line. In the GREAT tool, we checked "*two nearest gene*" option with the *"Associating genomic regions with genes* session". Where we defined 100 kilobase (kb) for the *cis* PC genes, which are located within a proximity of from the lncRNAs transcription start site (TSS) site. In order to determine the *trans* PC, we defined greater >100 kb from the TSS of the lncRNAs in the checkbox. And then we submitted our query to the GREAT algorithm. The GREAT run resulted in a list of *cis* and *trans* located PC to their corresponding lncRNAs as text files with their genomic distance. Firstly, we used regular expression *[regex = r'(?P<PC>\w+.*).*\((?P<Strand>[+-])(?P<Distance>.*)\)']* and extracted all lncRNAs and their corresponding *cis* and *trans* PC genes using python *lambda extract and groupby* functions into a tab separated file. Once we had all the list of *cis* and *trans* PC genes as tab separated file, we again filtered out the *cis* PC genes from *trans* list using *query* function from *python panda's* library.

2.9.1 Co-expression analysis between subtype-specific and relapse-specific lncRNAs and their *cis* and *trans* located PC genes

Computing Pearson correlation quantifies correlations between genes. In order to test the significance of correlation we used 2-tailed test. The Pearson correlation analysis was performed using python's *scipy.stats.pearsonr* mathematical algorithm from python. We calculated a pairwise correlation on subtype-specific lncRNAs to their *cis* and *trans* protein-coding genes. For instance, for each subtype, we computed a pairwise correlation matrix between all DE lncRNA and between their *cis* and *trans* coding gene to produce two matrices of lncRNAs × cis protein coding. We obtained both positively and negatively correlated *cis* and *trans* protein-coding genes for each DE lncRNAs. We considered only positive correlations to characterize the lncRNAs of interest in our study. The significantly correlated protein-coding genes were ranked for each lncRNA by the correlation coefficient (Pearson's correlation >= 0.55, P-value <= 0.05) as co-expressed genes or positively correlated genes for each subtype. This filtering was done python script written using *query* function. The resulted gene list enabled us to generate hypotheses regarding the function of a given lncRNA based on how they are enriched in pathways. The same procedure was followed for functional characterization of dysregulated lncRNAs for relapse-specific DE lncRNAs within all subtypes.

2.9.2 Functional enrichment of significantly correlating genes using GeneSCF tool

We used Gene Set Clustering based on Functional annotation (Subhash & Kanduri, 2016) (GeneSCF version 1.0) for our functional enrichment analysis for the subtype-specific DE lncRNAs and relapse-specific DE lncRNAs. The GeneSCF is a command line tool. We had a plain text file with all our PC gene symbols which were significantly correlated with the subtype-specific lncRNAs. *gtype=sym*). The functional enrichment analysis was performed by the following parameters: defined the input text file with significantly co-expressed PC genes using *-i=input* plain text file, defined the input gene symbol using *-t=sym* and source database using *-db=KEGG*. After that, the tool outputs the gene hits along with the corresponding functional pathways as a table. The significant pathways were filtered based on P-value <=0.05, the same procedure is followed for all subtype in order to maintain consistency in the analysis process. We used *awk* to filter out the functional pathways falling within the cut-off.

2.10 DNA methylation analysis

We next, sought to comprehensively define the DNA methylation profile of BCP-ALL subtypes in patient samples. In order to define the DNA methylation profile, we isolated genomic DNA (0.5 μg) from BCP-ALL (n = 82) samples at ID and REL conditions from the same 45 patients. These samples were then hybridized onto an Illumina 450k methylation array. The beta values representing the signal density of CpG sites were obtained from DNA methylation array for all samples.

2.10.1 DNA methylation dataset preparation and normalization

A beta value refers to the measure of the degree of methylation at each measured locus. Beta values are powerful for large-scale studies as it can be transmitted and compared across samples. The obtained beta (β-values) values from each CpG sites, were then normalized using SWAN method. Normalization is used to remove the technical variation between measurements, by maintaining the true biological difference between samples and probes. SWAN normalization was used as this method improves the correlation between biological or technical replicate, while the increasing the detection of some significantly differentially methylated probes. In addition to that, we can use them with any R package further for detection of differentially methylated probes. SWAN normalization method has two parts; the first part determines the average quantile distribution using a subset of probes defined to be biologically similar based on CpG content. The second step is to then adjust the intensities of the remaining probes, mainly from Infinium II than I, by insertion onto the distribution of the subset probes. This is done for each probe type separately using linear insertion between the subset probes to define the new intensities. Gradually, while the distribution of the subset is same, the intensity distribution of Infinium I probes is

still vastly different from the distribution of Infinium II probes (Maksimovic, Gordon, & Oshlack, 2012). The obtained SWAN normalized value was saved in a matrix for further bioinformatics analysis. The data matrix consisted of CpG's identifications as rows and beta values for all 82 BCP-ALL as columns.

2.10.2 DNA methylation profile of lncRNAs across samples

The positions of each CpG's from the SWAN normalized data matrix was identified using R package *IlluminaHumanMethylation450kanno.ilmn12.hg19* using *"@data$Locations.* Which resulted in the genomic position information, including, chromosomal positions, and gene symbols for each CpG sites for our input matrix. Annotation of the CpG signals represented in SWAN normalized beta values obtained from the array resulted in the identification of 60,021 CpG probes corresponding to 7190 lncRNA genes and 120,000 CpG probes corresponding to around 15,000 PC genes. DNA methylation analysis started by looking into the level of DNA methylation profile between lncRNAs and PC genes across 82 BCP-ALL samples. The density plots were plotted using python. *plot* function, on *M-values* (β-logit2 transformed). The *logit* transformation was performed by python *scipy.special.logit* algorithm.

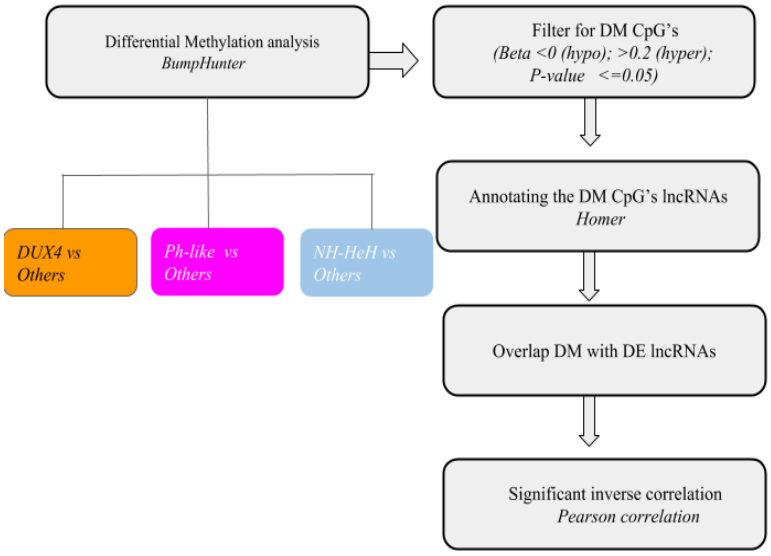

Figure 2.10.1: The DNA methylation analysis work-flow fro defining the differentially methylated subtype-specific lncRNAs
Work-flow used for DNA methylation analysis for each subtype.

2.10.3 PCA on the lncRNAs DNA methylation profile

We then used the same matrix to see how the samples are clustered based on their DNA methylation profile using PCA analysis using the R function *prcomp* on the SWAN normalized values for lncRNAs associated CpG sites (n = 60,021). The 3D PCA plots are constructed using the python library

matlabplotlib.

2.10.4 Differential methylation analysis

We performed differential methylation analysis using the R Bioconductor package, *Bumphunter* using the most variant quartile of CpG probes, searches for differentially methylated regions in an annotation-unbiased manner (Jaffe et al., 2012). We separated the ID and REL samples for each DUX4, Ph-like, and NH-HeH subtype in order to account for the biological replicate dependency. To determine differentially methylated regions (DMRs), we used R to apply 1000 permutations with the Bumphunter algorithm and considered significant regions of *P-value* < 0.05, CpGs differently methylated. Each subtype was compared with other samples for differential methylation analysis (Figure 2.12.1). In order to define statistically significant hyper-methylated genes and hypo-methylated genes we then used previously defined criteria by Bumphunter package [http://genomicsclass.github.io/book/pages/epiviz.html].The significant hyper-methylated genes were defined if the differential methylation value > 0.2 and *P-value* <= 0.05 and the significant hypo-methylated genes were defined if the differential methylation value is < 0 and *P-value* <= 0.05. The hyper-methylated genes are the ones who showed an elevated methylation rate compared to other samples, and the hypo-methylated genes are the ones which a decreased methylation rate compared to the others.

2.10.5 Association of subtype-specific DM with different genomic regions and finding subtype-specific DM lncRNAs

We associated the differentially methylated regions from three BCP-ALL subtypes using hypergeometric optimization of motif enrichment (HOMER) suite of tools [http://homer.ucsd.edu/homer/ngs/customGenomes/index.html]. We performed annotation of DM sites using 'annotationPeaks.pl.' tool using the *encode.v19.annotation.gtf* reference file. In order to get all information about the genomic regions including, the gene symbol, gene type, distance from the promoter-TSS region, and genomic regions (intron, exon, promoter-TSS, Transcription Termination site, etc.), gene type, and the distance from the promoter-TSS of each gene, we used the *-gene* parameter. The input for 'annotationPeaks.pl' tool was BED files defining the chromosomal potions of each significant DM regions obtained from *Bumphunter* and the reference file (*encode.v19.annotation.gtf*) which was converted into a tab-delimited gene data file using *awk* command line script. With these inputs '*annotatePeaks.pl*' provided us with all the essential information about the genomic region corresponding to each CpG sites for our DM genes. Using this information, we identified lncRNAs from our DM list and their genomic regions.

The genomic regions were defined as promoter-TSS and gene body. The gene body was defined if the CpGs are annotated in exonic, intronic or transcription termination site (TTS). We used the list of all lncRNAs biotype to filter the lncRNAs from the output file. The *awk* and *grep* commands were used to filter out the lncRNAs. The promoter-TSS is assigned based on the genomic window of -2000 base pairs downstream and 2000 base pair upstream to the TSS region. The regions mapped to lncRNAs were then used for analysis. These filtered DM lncRNAs were further used for remaining comparison analysis.

2.10.6 Correlation analysis between DM of lncRNAs and their expression levels

We used the results from *Bumphunter* and *LIMMA Voom* for DM and DE subtype-specific lncRNAs signatures to compare the DNA-methylation and expression. First, we overlapped the promoter-TSS methylated lncRNAs and the DE lncRNAs. Out of these, we used the overlapped promoter-TSS methylated lncRNAs for comparative analysis. Next, the reverse correlation was determined between DNA-methylation and expression level by correlating the DNA methylation values (β-logit2 transformed) with the log2 transformed FPKM values of each lncRNAs. The correlation was determined by the previously mentioned *Pearson correlation method* using python *scipy. stats. Pearson* library. The significantly correlated DM and DE promoter methylated lncRNAs are determined based on a 2-tailed *P-value*<=0.05.

Chapter 3. Results

3.1 The expression and DNA methylation profile of lncRNAs

To systematically identify subtype-specific lncRNAs from three BCP-ALL subtypes we analyzed transcriptome and DNA methylation profiles from paired ID and REL samples of 25 pediatric and 20 adult BCP-ALL patients lacking known chromosomal translocations like BCR-ABL. Based on expression signatures of PC genes, fusion genes, mutations, deletions and DNA methylation profile detected by RNA expression and DNA methylation profiles, the samples (n = 82) were classified into different molecular subtypes, namely, DUX4 (n = 23), Ph-like (n = 21), NH-HeH (n = 16), and low-hypodiploid (LH; n = 6). The DNA methylation data were extracted from the same samples using DNA methylation array platform, which accounted for 60,022 CpG's located annotated as lncRNAs (n = 7160) in the genome.

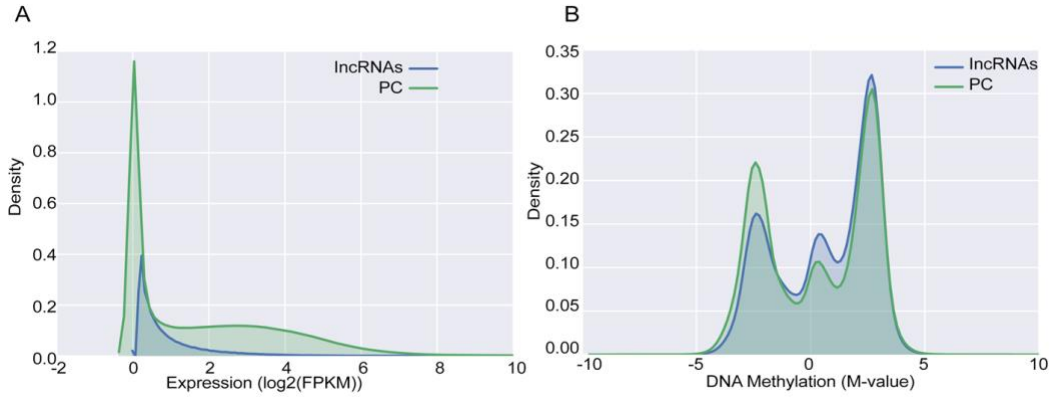

Figure 3.1.1: The expression and DNA methylation profile of lncRNAs and protein coding genes across all samples.

A. The level of distribution of expression between 13460 lncRNAs and 20,135 PC genes across 82 BCP-ALL samples. B. The level of distribution of DNA methylation rate between 60,022 CpGs probes associated with lncRNAs region and 120,000 CpGs probes associated with PC genes across 82 BCP-ALL samples. The x-axis represents the DNA methylation values, log-transformed Methylation values.

We started by comparing the distributions of expression and DNA methylation profile between lncRNAs and protein-coding (PC) across all BCP-ALL. Consistent with the previous reports (Casero et al., 2015), we observed the lncRNAs (n = 13460) were less abundantly expressed than the PC genes (n = 20,135) (Figure 3.1.1 A). Whereas, when the DNA methylation profile of CpG sites (n = 60,021) associated with 7,160 lncRNAs was compared with CpG sites associated with PC genes (n = 120,000) across all BCP-

ALL samples (Figure 3.1.1 B), we identified a similar DNA methylation profile between lncRNAs and PC genes.

3.2 Unsupervised hierarchical clustering of lncRNAs expression identified robust clusters of BCP-ALL subtypes

In order to identify the ability of lncRNAs to stratify BCP-ALL samples as the PC genes, we performed unsupervised clustering using principal component analysis (PCA) on the normalized (FPKM) expression values of 13,860 GENCODE lncRNAs. The PCA analysis revealed three distinct robust clusters of BCP-ALL subtypes, DUX4, Ph-like and NH-HeH (Figure 3.2.1 A).

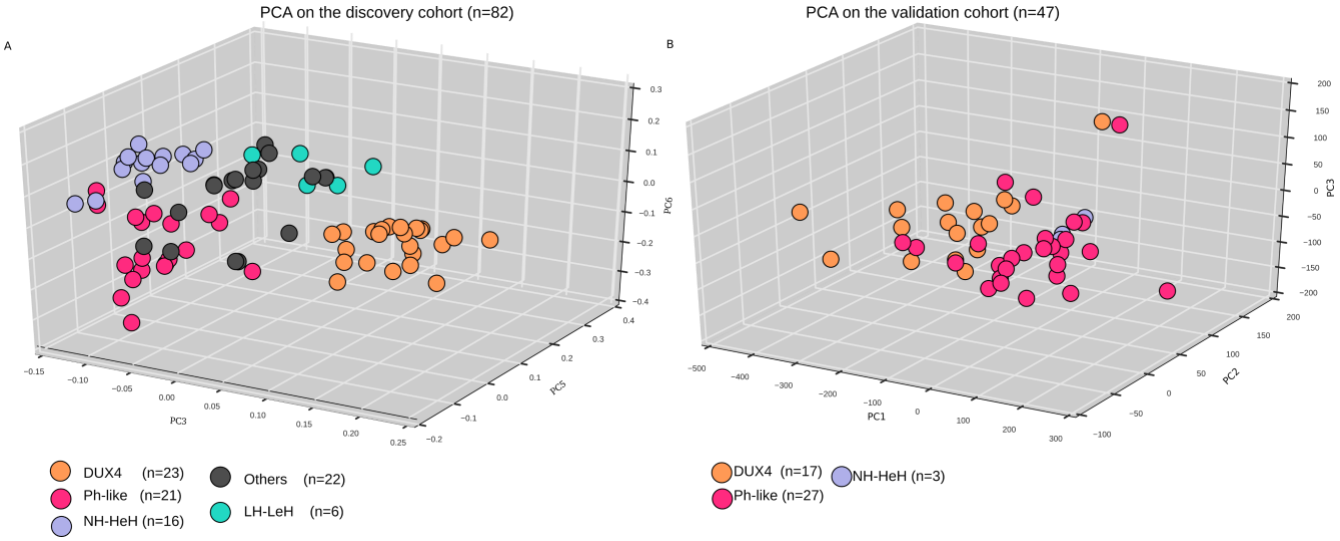

Figure 3.2.1: Unsupervised clustering of lncRNAs expression in BCP-ALL samples on the discovery and validation cohort.

A. The unsupervised clustering (PCA) on the lncRNAs expression of BCP-ALL samples of the discovery cohort (n=82), representing distinct clusters of DUX4, Ph-like and Nh-HeH subtypes. The PCA plot constructed from expression FPKM values of lncRNAs from 82 BCP-ALL samples obtained from RNA-Seq from the original discovery cohort. B. The unsupervised clustering on the independent validation cohort of 47 BCP-ALL samples, representing the distinct clusters of DUX4, Ph-like and Nh-HeH subtypes. The PCA plot constructed from expression FPKM values of lncRNAs from 47 BCP-ALL samples obtained from RNA-Seq from the validation cohort. Each point represents a BCP-ALL sample. DUX4, Ph-like, NH-HeH, LH subtype and others are represented by orange, rose, blue, green and grey respectively.

This observation was in concordance with the predefined molecular classification of BCP-ALL subtypes. In order to validate the ability of lncRNAs in distinguishing subtypes, we extended the same PCA approach on the lncRNAs expression (13860 lncRNAs FPKM value) on an independent cohort, that we termed as "independent validation cohort". The samples within the independent validation cohort were predefined based on their molecular profile into DUX4 (n = 17), Ph-like (n= 27) and NH-HeH (n = 3) subtypes. We identified a similar observation as with our discovery cohort (Figure 3.2.1 B)

in the independent cohort. We, therefore, sought to identify lncRNAs signatures which are differentially expressed within these three subtypes in our discovery cohort.

3.3 Differentially expressed lncRNAs across multiple BCP-ALL subtypes

To identify the subtype-specific differentially expressed (DE) lncRNAs signatures in the BCP-ALL samples we analyzed each subtype (DUX4, Ph-like, and NH-HeH) versus the rest of the cohort. A total of 1564 (P-value <= 0.01 and Fold change <=> -+1.5) subtype-specific lncRNAs were identified as DE from the three subtypes (Figure 3.3.1 A). By comparing subtype-specific DE lncRNAs (n = 1564) from these three subtypes, we found 59% (n = 930) of lncRNAs are specific to each subtype, the remaining lncRNAs were shared in at least two subtypes (Figure 3.3.1 B). We identified 24 lncRNAs whose expression was significantly altered in DUX4, Ph-like and NH-HeH BCP-ALL subtypes (Figure 3.3.1 B).

The large size of the DUX4 subtype signature distinguishes the DUX4 subtype as particularly perturbed at the level of lncRNAs gene expression. The subtype-specific DE lncRNAs based hierarchical clustering revealed distinct and robust clusters for each BCP-ALL subtype (Figure 3.1.1 A-C), defining lncRNAs driven molecular signatures in BCP-ALL subtypes.

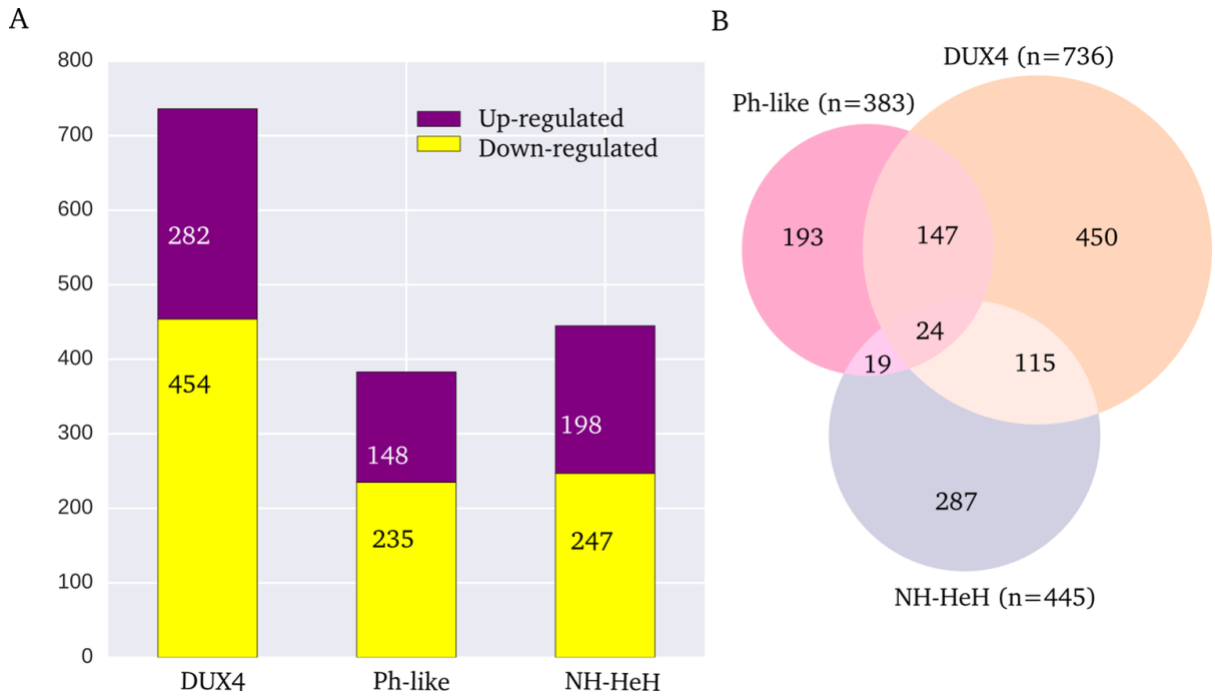

Figure 3.3.1: Number of subtype-specific lncRNAs

A. The barplots represents the number of significantly up and down regulated subtype-specific lncRNAs. B. The Venn diagram represents the overlap between subtype-specific lncRNAs.

We next compared our subtype-specific DE lncRNAs signature with different cancer types from a comprehensive genomic characterization of lncRNAs across cancers (CGC) (X. Yan et al., 2015). We found about 59% (n = 712, Figure 3.3.1 D; Hypergeometric test P-value = 9.2e-29) of our subtype-specific lncRNAs were more specific to the investigated three BCP-ALL subtypes (Figure 3.1.13 A-C). Out of the overlapped DE lncRNAs (n = 523), 23 (Appendix 1) lncRNAs were previously defined as cancer-related lncRNAs from the lnc2cancer database (Ning et al., 2018). For example, oncogenic lncRNAs *PVT1* (Tseng & Bagchi, 2015) and *GAS5* (Mazar et al., 2016) are differentially up-regulated in the DUX4 subgroup, and *CRNDE* (Huan, Xing, Lin, Xui, & Qin, 2017) is DE is down-regulated in the Ph-like subtype. Together, this demonstrates that the dysregulated expression of lncRNAs in for BCP-ALL subtypes.

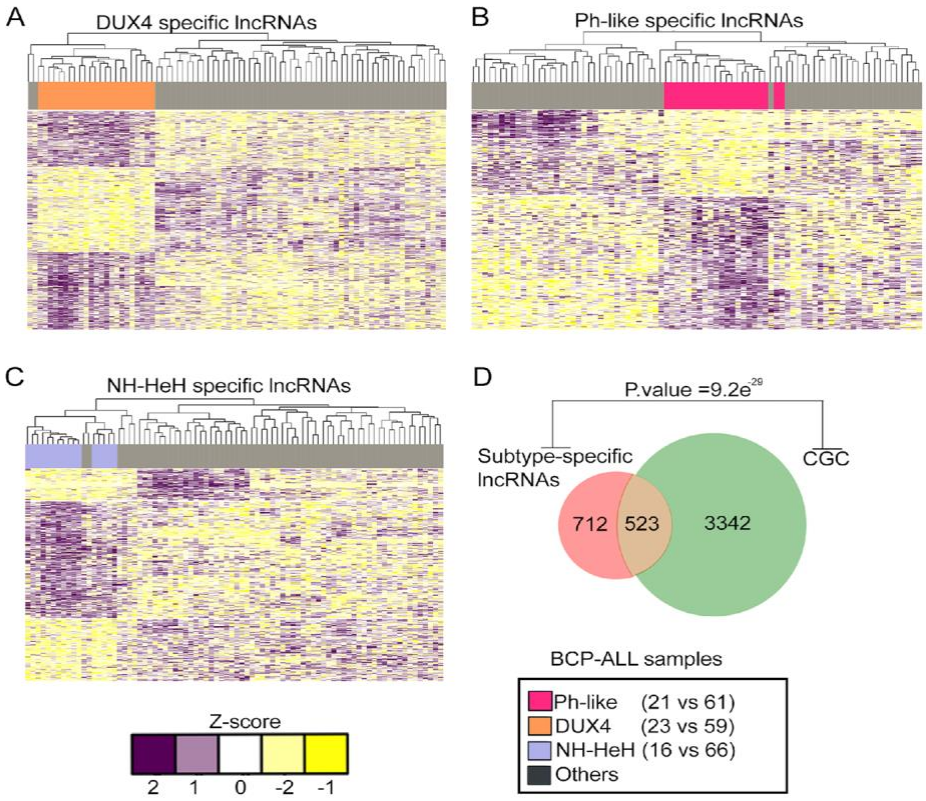

Figure 3.3.2: BCP-ALL subtype-specific differentially expressed lncRNAs.
A-C. The hierarchical clustering representing lncRNAs clustering and expression differences of the compared subtypes DUX4, Ph-like and NH-HeH; corresponding to 736, 383, and 445 subtype-specific DE lncRNAs in DUX4, Ph-like and NH-HeH subtypes, respectively. In the DUX4 subtype, 100% of samples clustered together based on the DE lncRNAs signature. The hierarchical clustering of the subtype-specific DE lncRNAs revealed that 90% (19 out of 21 samples) of Ph-like samples clustered within the predefined Ph-like subtype. For the NH-HeH subtype, 69% (11 out of 16 samples) of samples correlated and clustered together using the respective DE lncRNA signature. The BCP-ALL samples box representing the number of samples within each subtype and versus (vs) the other samples used as a control group in DE analysis. D The overlap between DE subtype specific lncRNAs from three subtypes versus a public list of dysregulated lncRNAs from 12 different cancer types comprehensive cancer genome (CGC).

3.4 Further validation of the subtype-specific lncRNAs with an independent BCP-

ALL cohort

We observed distinct clusters for three subtypes based on unsupervised HC on the 1235 subtype-specific lncRNAs (299 out of 1534 is present at least in one of the three subtypes) (Figure 3.1.14 A) in the discovery cohort (n = 82). To confirm subtype-specificity of these dysregulated lncRNAs, we made use of the previously defined independent validation cohort. Unsupervised hierarchical clustering on the expression of 1235 subtype-specific lncRNAs on the independent validation cohort identified three robust clusters confirming the (Figure 3.1.14 B) subtype-specificity of our 1235 subtype-specific lncRNAs. Taken together, the results with independent validation cohort demonstrated better and reproducible subtype-specific lncRNAs in stratifying the BCP-ALL samples.

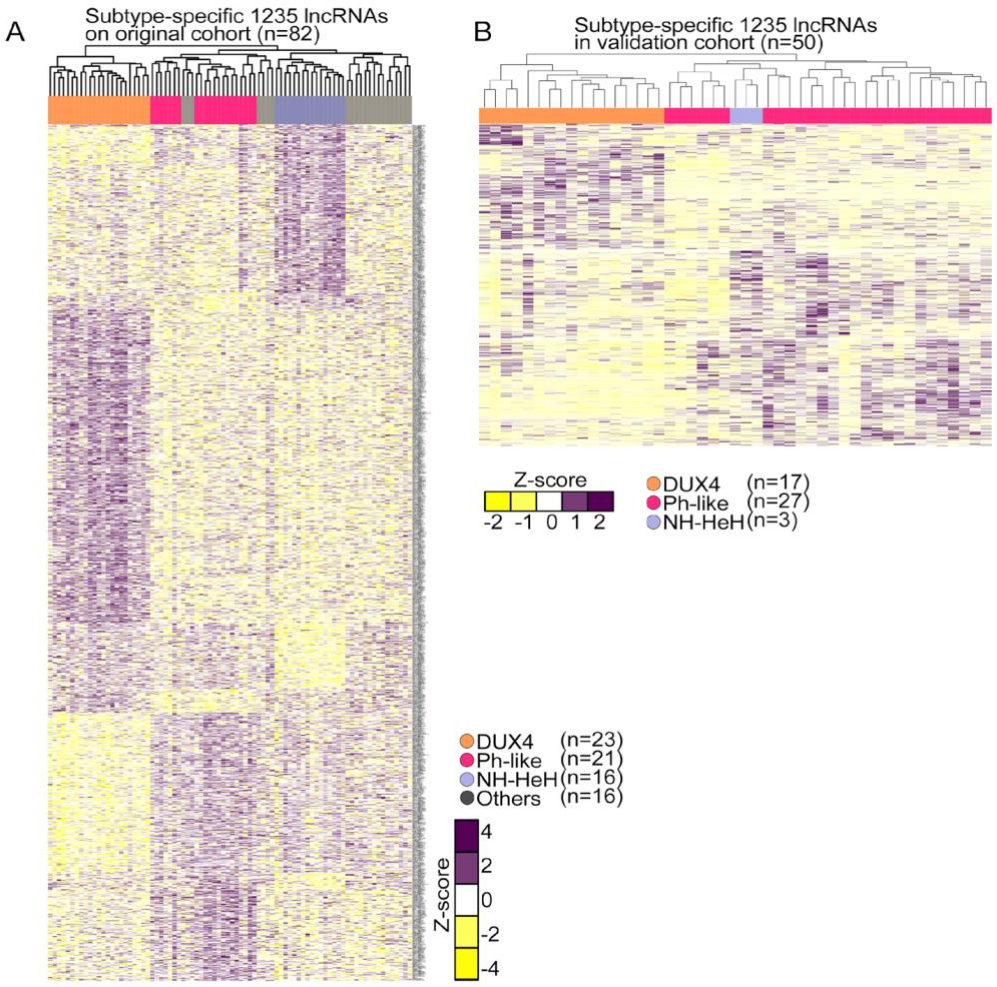

Figure 3.4.1: Validation of subtype-specific lncRNAs on independent validation cohort.
A. Heatmap illustrates hierarchical clustering (HC) DE subtype-specific lncRNAs (absolute Fold change >= +- 1.5, P-value <= 0.01) signature based on z-score transformed LIMMA normalized expression values on 1235 subtype-specific lncRNAs from DUX4 (n = 450), Ph-like (n = 193), and NH-HeH (n = 287) subtypes. The spearman correlation-based clustering is done on the lncRNAs. B. The heatmap represents the expression pattern of 1235 subtype-specific lncRNAs in the independent validation cohort. The unsupervised HC defines three distinct clusters within the independent validation cohort.

3.4.1 Identification of subtype-specific lncRNAs functions

We next asked whether these 1235 subtype-specific lncRNAs might be globally related to the alteration of different biological functions and molecular pathways. Since lncRNAs exert their actions by regulating the PC genes in *cis* and *trans* regions, we performed functional enrichment analysis using the previously defined *guilt-by-association* approach on both *cis* and *trans* PC genes (Table 4.4.1). This enabled us to generate hypotheses regarding the function(s) of a given subtype-specific lncRNAs. Functional enrichment analysis was performed based on the correlation between neighboring (*cis*) and distally (*trans*) located protein-coding (PC) genes (within ± 100 kb cis and >± 100 kb window for *trans*) of the subtype-specific lncRNAs from the subtypes. The significantly co-expressed *cis* and *trans* PC genes based on their positive correlation rate (Pearson correlation >= 0.55 and two-tailed P-value <= 0.05) were then used for functional enrichment analysis. Consistent with other reports (Casero et al., 2015), we observed the higher number of positive correlations (Pearson correlation rate >= 0.55 and two tail P-value <= 0.05) than the negative correlated *cis* and *trans* genes. The table represents the number of positive correlated PC genes (Table 3.4.5).

Table 3.4.5: Number of BCP-ALL subtype specific co-expressed lncRNAs with it's *cis* and *trans* PC genes.

Subtypes	*Cis* PC genes (n = 929)	LncRNAs co-expressed with *cis* PC genes (n = 621)	*Trans* PC genes (n = 753)	LncRNAs co-expressed with *trans* PC genes (n = 552)
DUX4	669	451 (736)	492	379 (736)
Ph-like	260	170 (383)	261	173 (383)

Table 3.4.5: The table represents the number of subtype-specific lncRNAs with *cis* (<100 Kb proximity) and *trans* (>100 Kb) protein-coding genes and the number of their co-expressions. The numbers shown within the bracket is the otal number of DE lncRNAs corresponding to the respective subtypes. The percentage of *cis* co-expression is, 68%, 44.3%, in the DUX4 and Ph-like subtypes respectively. The percentage of trans-co-expression is, 51.5%, 45.2%, in the DUX4 and Ph-like subtypes respectively.

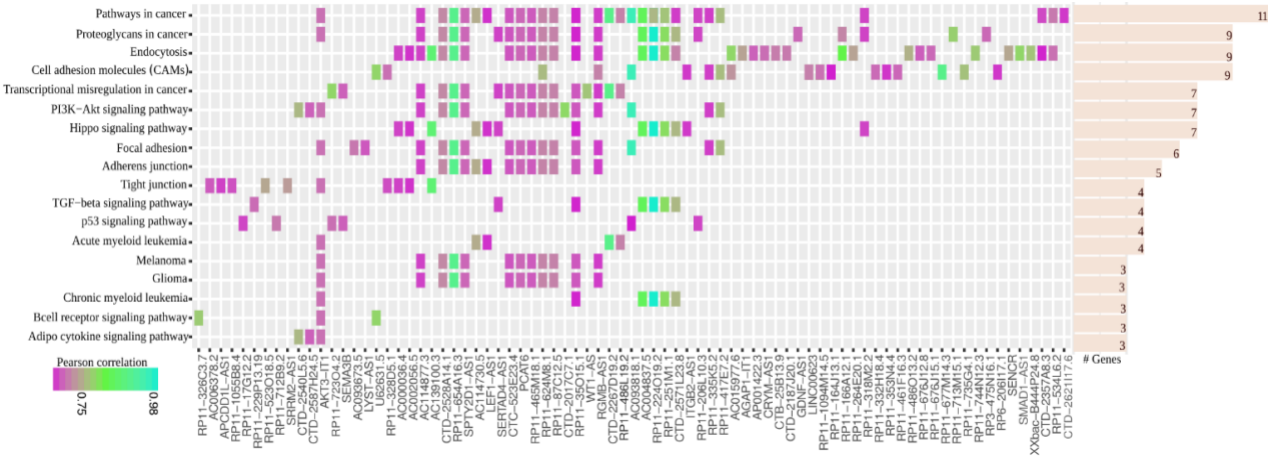

Figure 3.4.2: The molecular pathways of lncRNAs involved in the DUX4 subtype.
The plot depicts the molecular pathway analysis from the functional enrichment analysis for nearby (<= 100 kb proximity) cis protein-coding genes correlated (Pearson correlation coefficient >= 0.55 and 2-tailed P-value <= 0.05) with DE lncRNAs in the DUX4 subtype. The barplot in the right-hand side represents the number of genes involved in each pathway. The KEGG pathways or biological functions presented in the plot are with P-value <= 0.05 and > 2 genes within each pathway. The hypergeometric P-values are obtained from GeneSCF tool for the pathways.

Out of these *cis* (n = 451) and *trans* (n = 379) co-expressed lncRNAs from DUX4 subtype, we identified 58 and 127 lncRNAs significantly co-expressed with 45 *cis* and 49 *trans* PC respectively. In Ph-like subtype, we identified 24 (Appendix C) and 20 subtype-specific lncRNAs co-expressed with 25 *cis* and 37 *trans* located PC genes respectively. These genes were enriched in signaling pathways, including, Janus kinase and Signal Transducer and Activator of Transcription (JAK-STAT), cytokine-cytokine kinase receptor, and phosphoinositide 3-kinase (P13K-Akt) signaling pathways (Figure 3.4.3, Figure 3.4.4 C) based on the *cis* PC gene co-expression-based analysis.

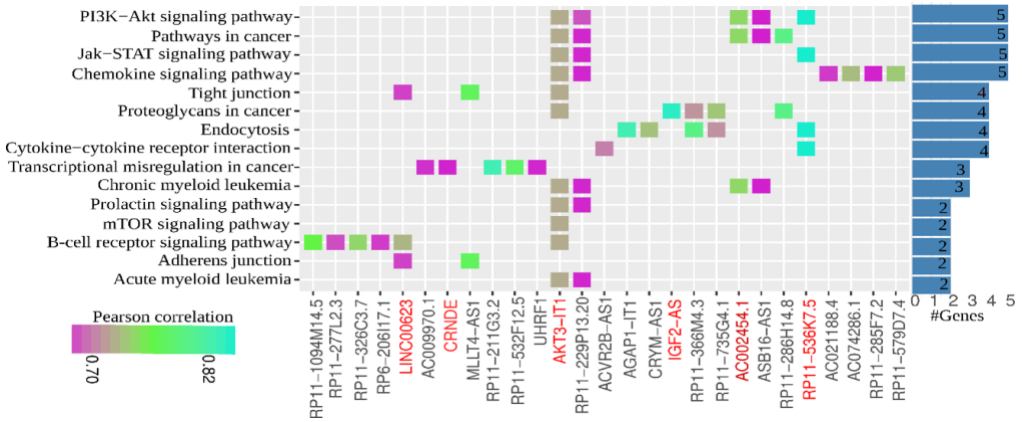

Figure 3.4.3: The molecular pathways of lncRNAs involved in the Ph-like subtype.
A. The plot depicts the molecular pathway analysis from the functional enrichment analysis for nearby (<= 100 kb proximity) cis protein-coding genes correlated (Pearson correlation coefficient >= 0.55 and 2-tailed P-value <= 0.05) with DE lncRNAs in the Ph-like subtype. The barplot on the right-hand side represents the number of genes involved in each pathway. The KEGG pathways or biological functions presented in the plot are with P-value <= 0.05 and > 2 genes within each pathway. The hypergeometric P-values are obtained from GeneSCF tool for the pathways.

These 45 *cis* and 49 *trans* located PC genes were enriched in pathways involved in proliferation, apoptosis, and differentiation in leukemia. For example, the DUX4 subtype we identified signaling pathways including, Transforming growth factor Beta (TGF-Beta), P53, Endocytosis, and hippo signaling pathway, and Proteoglycans in cancer pathways (Figure 3.4.2, Figure 3.4.4 A-B).

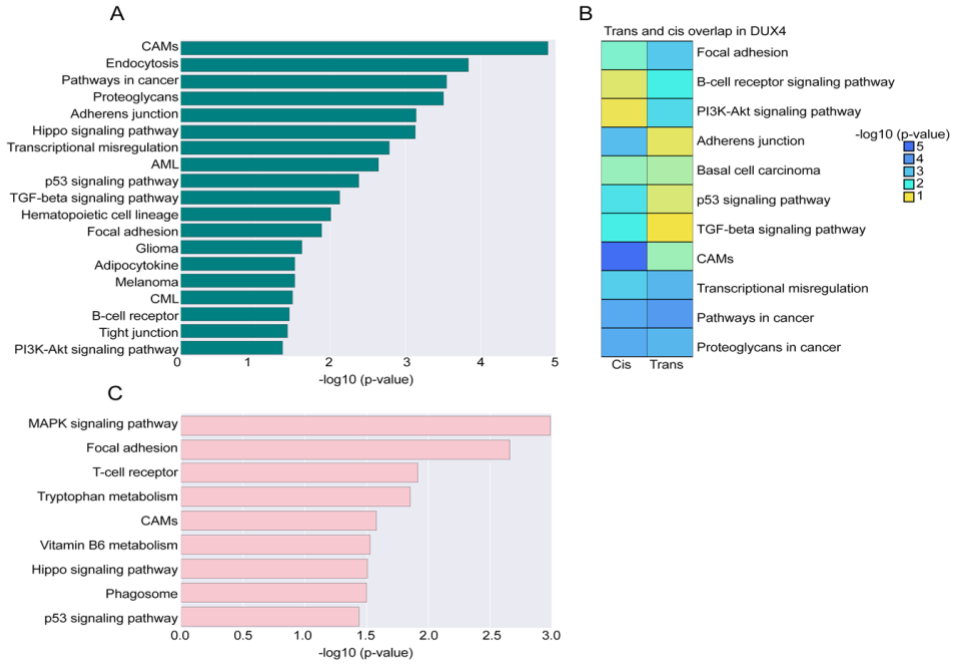

Figure 3.4.4: Comparison of molecular pathways from cis and trans based analysis on subtype-specific DE lncRNAs.

A. Molecular pathway analysis from functional enrichment analysis for distant (>100 kb) trans protein-coding genes correlated (Pearson correlation >0.55 and P-value <=0.05) with DE lncRNAs in the DUX4 subtype. B. The molecular pathways overlapped between cis (<100 kb proximity) and trans (>100 kb) based functional enrichment analysis in the DUX4 subtype. C. Molecular pathway analysis from functional enrichment analysis for distant (>100 kb) trans protein-coding genes correlated (Pearson correlation >0.55 and P-value <=0.05) with DE lncRNAs in the Ph-like subtype.

3.4.2 The lncRNAs based and mRNAs based functional enrichment analysis showed the same pathways in the subtypes

In order to validate our predictions, we then linked the functions predicted for subtype-specific lncRNAs to those predicted for respective subtype-specific mRNAs. The comparisons were done between the DE PC based and DE lncRNAs based functional pathways from DUX4 and Ph-like subtypes. Interestingly, we observed all signaling activated in DUX4 DE PC based analysis are activated or inhibited in the DE DUX4 specific lncRNAs based analysis (Figure 3.4.5 A). Whereas, in the Ph-like subtype, we identified 60% (9 out of 15 pathways) overlap between the two sets of predicted functions (Figure 3.4.5 B). For example, JAK-STAT, Cytokine-cytokine receptor and endocytosis pathways had overlapped between both analyses; yet, there were cases where subtype-specific lncRNAs appeared to be more strongly associated with a function or pathway than subtype-specific PC genes. The key signaling pathways mTOR and P13K-Akt signaling pathways were more exclusive for lncRNAs based analysis in the Ph-like subtype.

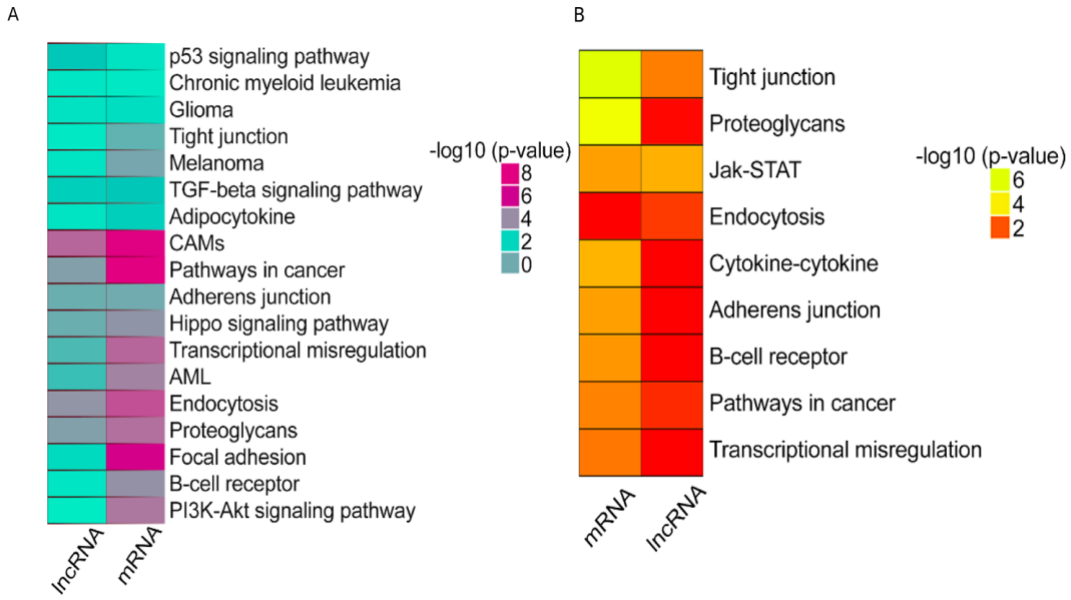

Figure 3.4.5: Subtype-specific lncRNAs and PC genes displayed enrichment of same pathways in DUX4 and Ph-like subtypes.
A. The heatmap depicts the concordance between the protein-coding and lncRNAs based predictions for DUX4 subtypes. B. The heatmap depicts the overlapping pathways from both lncRNAs and protein-coding in the Ph-like subtype. The KEGG pathways or biological functions presented in the heatmaps and barplots show with P-value <= 0.05 and > 2 genes involved in each pathway. The hypergeometric P-values are obtained from GeneSCF for the pathways. CAMs: Cell adhesion molecules, CML: Chronic myeloid leukemia, AML: Acute myeloid leukemia.

3.4.3 DUX4 Subtype-specific lncRNAs represented in functional pathways predictions

One of the key functions of lncRNAs is its ability to regulate the expression of its neighboring protein-coding genes in the genome. In order to investigate to what degree this general concept may be applied

to the BCP-ALL subtypes-specific lncRNAs, we looked into genes which are present in different predicted signaling pathways from each subtype. Firstly, we started with identifying PC genes involved in activated signaling pathways associated with cellular processes including proliferation, differentiation, migration and cell survival. Secondly, we looked for subtype-specific lncRNAs which are co-expressed and located at the same locus of these PC genes.

Among the co-expressed DUX4 specific lncRNAs (Pearson correlation > 0.55 and P-value <= 0.05, Appendix B) with *cis*-PC genes we identified 58 lncRNAs are correlated with their *cis*-PC genes that are involved in key signaling pathways. Some of these *cis*-PC genes are oncogenes involved in leukemia progression. For example, *TGFB2* gene, *SMAD1* is and *ITGA6*. Among these, the *TGFB2* gene is enriched in key signaling pathways, including, Hippo and TGF-Beta signaling pathway and endocytosis. The antisense lncRNA *RP11-224O19.2* which is encoded at *TGFB2* locus aligning at the 5' end showed a strong significant co-expression (Pearson correlation = 0.96, P-value = 2.6e-45) with the *TGFB2* (Figure 3.4.6 A). Interestingly, both *TGFB2* and its antisense lncRNA *RP11-224O19.2* are significantly up-regulated in the DUX4 samples (Figure 3.4.6 B, Table 4.4.2).

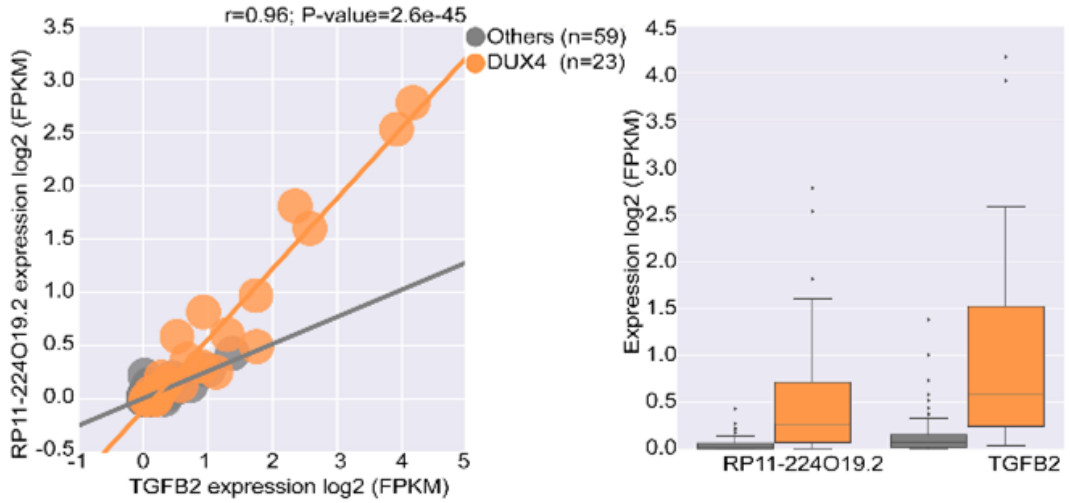

Figure 3.4.6: The subtype-specific lncRNA RP11-224O19.2 co-expressed with TGFB gene in DUX4 subtype
The lncRNA RP11-224O19.2 and its cis oncogene TGFB2 is significantly co-expressed. Antisense RP11-224O19.2 and its cis oncogene TGFB2 are encoded in the same locus (left panel). Antisense RP11-224O19.2 (absolute fold-change = 2.786, P-value = 9.74E-08) and TGFB2 (absolute fold-change = 3.84, P-value = 2.74E-10) genes are significantly up-regulated in DUX4 samples (Right panel).

The *SMAD* proteins are signal transducers and transcriptional modulators that mediate multiple signaling pathways. The *SMAD* protein mediates the signals of the bone morphogenetic proteins (BMPs), which are involved in a range of biological activities including cell growth, apoptosis, morphogenesis, development, and immune responses (Blank & Karlsson, 2011). We observed the antisense lncRNA,

SMAD1-AS2 positively correlated (Pearson correlation = 0.75; 2 tailed P-value = 2.9e-16) with its sense gene, *SMAD1*. Other example, *ITGA6* gene is enriched in the PI3K-Akt signaling pathway, cell adhesions molecule and focal adhesion. The antisense lncRNAs *AC093818.1* (Pearson correlation = 90; P-value = 2.1e-30) and *AC078883.3* (Pearson correlation = 0.68; P-value = 2.8e-12) are co-expressed with their sense PC gene, *ITGA6.* Interestingly, both *ITGA6* (absolute Fold change = 9.5; FDR = 9.1e-10) and the novel lncRNAs were co-expressed and significantly up-regulated in the DUX4 samples.

3.4.4 Ph-like Subtype-specific lncRNAs represented in functional pathways

Dissecting the pathways appeared in the Ph-like DE lncRNAs based analysis, we identified 24 Ph-like specific lncRNAs co-expressed with *CDK6* and *IL2RA* oncogenes. These oncogenes were enriched in pathways, including, PI3K-Akt and JAK-STAT signaling pathways, and Cytokine-cytokine receptor interaction and endocytosis pathways. Twenty-four (Appendix C) Ph-like lncRNAs are co-expressed with these oncogenes. In Ph-like subtype, *CDK6* gene is enriched in the PIK3-Akt pathway in the Ph-like subtype, a pathway which prevents apoptosis. The *CDK6* is a gene which appears to be frequently up-regulated in the malignant hematopoiesis (Scheicher et al., 2015) with a critical role in AML and ALL driven by mixed lineage leukemia fusion proteins (Van Der Linden et al., 2014). We then looked for lncRNAs associated with the *CDK6* protein and identified seven novel lncRNAs (Figure 3.4.7 A, Table 3.4.6), including its antisense lncRNA, *AC002454.1* (-42918 bp, Pearson correlation = 0.72, Figure 3.4.7 B). The role of antisense lncRNA *AC002454.1* had previously been reported, including the ability to regulate *CDK6* by inducing the cell cycle disorder (Y. Wang, Li, Yang, Liu, & Wang, 2015). The association of *CDK6* and *AC002454.1* are referred to as "head-to-head" association, as the 5' end, both genes are aligned. The antisense lncRNAs are widely reported to be linked with multiple functions such as they regulate protein-coding genes positively or negatively.

Table 3.4.6: Novel lncRNAs co-expressed with oncogene *CDK6*, *TGFB2*, and *IL2RA*

Subtype-specific lncRNAs	Pearson coefficient	P-value	Subtype	Oncogene
RP11-347C18.3	0.56	3.25E-008	Ph-like	CDK6
RP11-461F16.3	0.62	5.21E-010		
RP11-96H19.1	0.62	3.89E-010		
RP11-228B15.4	0.64	7.68E-011		
MME-AS1	0.56	3.68E-008		
CTB-39G8.3	0.57	1.78E-008		
AC002454.1	0.72	2.21E-014		

RP11-582J16.4	0.55	8.08E-008		
AC009970.1	0.64	6.23E-011		
RP11-229P13.20	0.66	1.44E-011		
LINC00114	0.57	3.06E-008		
CTB-118N6.3	0.61	9.70E-010		
SOCS2-AS1	0.62	4.94E-010		
CTD-2561B21.10	0.61	9.91E-010		
RP11-413E1.4	0.56	4.36E-008		
KB-1460A1.1	0.55	7.77E-008		
AC012309.5	0.59	4.10E-009		
RP11-37B2.1	0.59	4.76E-009		
ASB16-AS1	0.65	3.86E-011		
LINC00426	0.62	6.32E-010		
LINC01071	0.57	2.46E-008		
RP11-536K7.5	0.74	5.11E-15		IL2RA
RP11-224O19.2	0.98	1.08E-061	DUX4	TGFB2
AC004837.5	0.83	6.11E-023		
RP11-251M1.1	0.79	7.39E-019		
CTD-2571L23.8	0.75	2.94E-016		
RP11-35O15.1	0.65	3.36E-011		
AC139100.3	0.58	1.00E-008		
RP11-158M2.3	0.58	1.50E-008		
RP11-672A2.5	0.56	4.68E-008		
CTD-2357A8.3	0.55	7.46E-008		
RP11-677M14.3	0.55	6.68E-008		

Table 3.4.6: The subtype-specific lncRNAs from DUX4, Ph-like associated with different signalling pathways. The table represents the lncRNAs and correlation rate between subtype-specific lncRNAs and PC genes enriched in various signaling pathways.

Another example is the *IL2RA*, a gene which is involved in diverse biological functions such as cell proliferation, apoptosis, cell surface immune response, and MAPK cascade. Recently, *IL2RA* is found to be specifically up-regulated by pre-B cell receptor (pre-BCR) signaling during early B cell development (Sadras et al., 2017), and cells with activated tyrosine kinases by a manifold to pre-BCR signaling in both in Ph+ALL and in Ph-like ALL (Roberts et al., 2012). The *IL2RA* gene locus encodes some non-

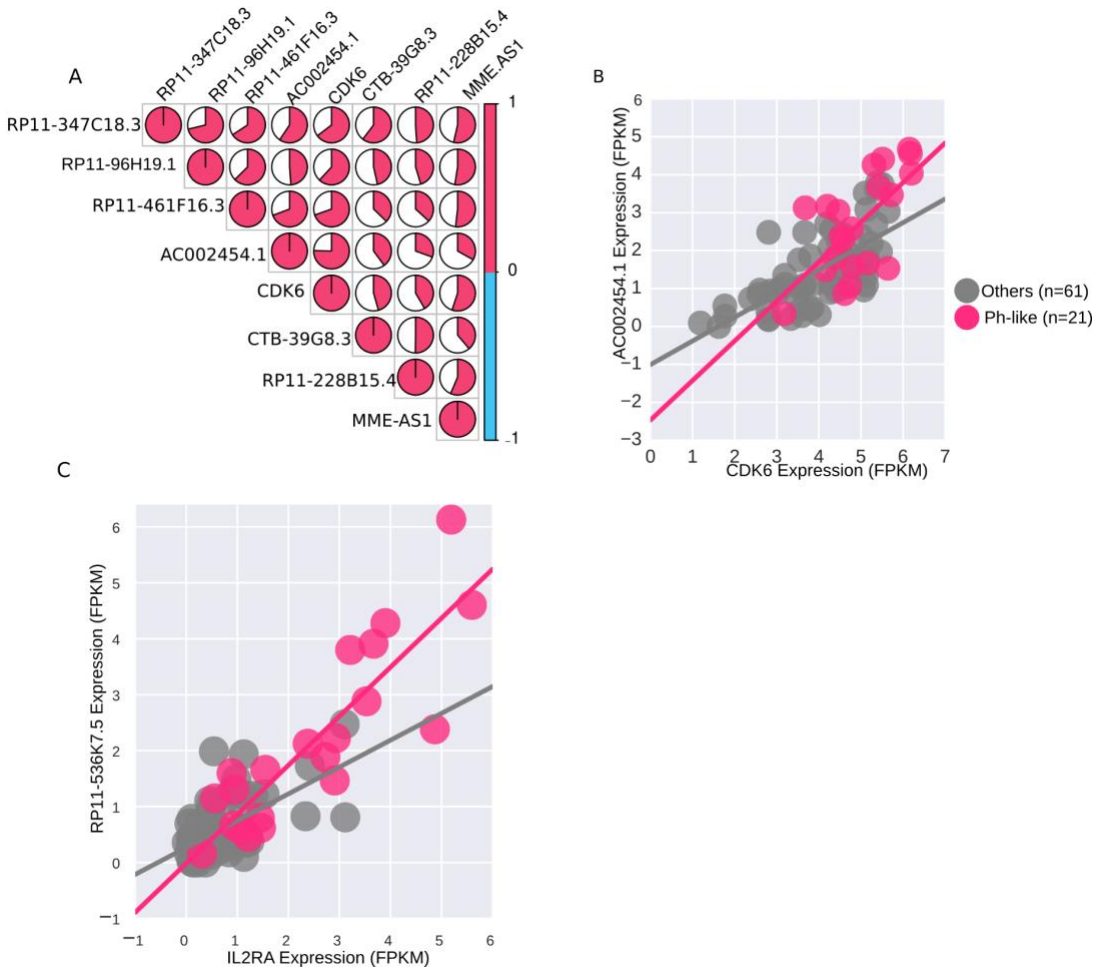

Figure 3.4.7: The subtype-specific lncRNAs co-expressed with oncogenes involved in key signaling pathways in Ph-like subtypes

A Seven novel lncRNAs co-expressed with CDK6 gene. B. The expression of cis antisense lncRNA AC002454.1 significant co-expressed with its cis oncogene CDK6 in Ph-like subtype. Both CDK6 (Absolute fold change = 1.01, P-value = 0.0005) and antisense lncRNA AC002454.1 (Absolute fold change = 1.79, P-value = 0.00015) are up-regulated in Ph-like samples. C. Expression of antisense lncRNA RP11-536K7.5 showed significant co-expression with expression of its cis oncogene IL2RA. Both RP11-536K7.5 (absolute fold-change = 2.79, P-value = 3.07E-008) and IL2RA (absolute fold-change = 3.11, P-value = 3.97e-1) are up-regulated in Ph-like samples

coding genes; antisense *RP11-536K7.5* is an example (Qureshi, Mattick, & Mehler, 2010). The antisense lncRNA *RP11-536K7.5* and *IL2RA* gene are up-regulated (absolute fold change= 2.79; *P-value* = 3.07e-08) in the Ph-like samples. Furthermore, the expression of *RP11-536k7.5* shows a significant positive correlation (Pearson correlation = 0.73, *P-value* = 5e-14) with the expression of *IL2RA* gene (Figure 3.4.7 C, Table 3.4.6). The antisense lncRNA *AC002454.1*, is up-regulated (absolute fold change = 1.7; *P-value* = 0.00015) in the Ph-like samples (Figure 3.4.7 right panel) compared to others.

Table 3.4.7: Subtype-specific novel DE lncRNAs co-expressed with oncogenes, which are associated with important molecular pathways.

LncRNAs	Cis PC	Pearson correlation coefficient	Associated pathways
RP11-224O19.2 *AC004837.5* *RP11-251M1.1* *CTD-2571L23.8*	*TGFB2*	0.98 0.83 0.79 0.75	Hippo TGF-β Endocytosis
AC093818.1 *AC078883.3*	*ITGA6*	0.95 0.68	PI3K-Akt
U62631.5	*CD22*	0.78	CAMs B cell receptor signaling pathway
CTD-2267D19.2 *RP11-486L19.2*	*RARA*	0.89 0.70	Pathways in cancer transcriptional mis-regulation in cancer pathways

Table 3.4.7: The table represents the novel subtype-specific DE lncRNAs co-expressed with its *cis* genes such as TGFB2, ITGA6, CD22, and RARA genes, which were enriched in vital molecular pathways in BCP-ALL.

In summary, global co-expression analysis and gene-expression profiling suggest an important and previously unappreciated role for lncRNAs in BCP-ALL subtypes. Our analyses highlight important putative functions for subgroups of the subtype-specific lncRNA genes whose expression correlates tightly with leukemic oncogenes.

3.5 Dysregulated relapse-specific lncRNAs as markers of BCP-ALL subtypes

LncRNAs are reported to be linked with clinical outcome of several diseases (Herrera-Solorio et al., 2017). Relapsed ALL, refers to the return of ALL in patients, and are mainly due to the poor outcome of conventional therapy. In order to investigate the relapse-specific DE lncRNAs, we performed differential expression analysis between initial diagnosis (ID) and relapse (REL) samples within each subtype. The DE analysis resulted in 941 dysregulated lncRNAs, in which, we identified 192 lncRNAs whose expression are up-regulated (Absolute fold change > +- 1.5; P-value <= 0.01) in relapsed samples compared to diagnosis samples from three subtypes.

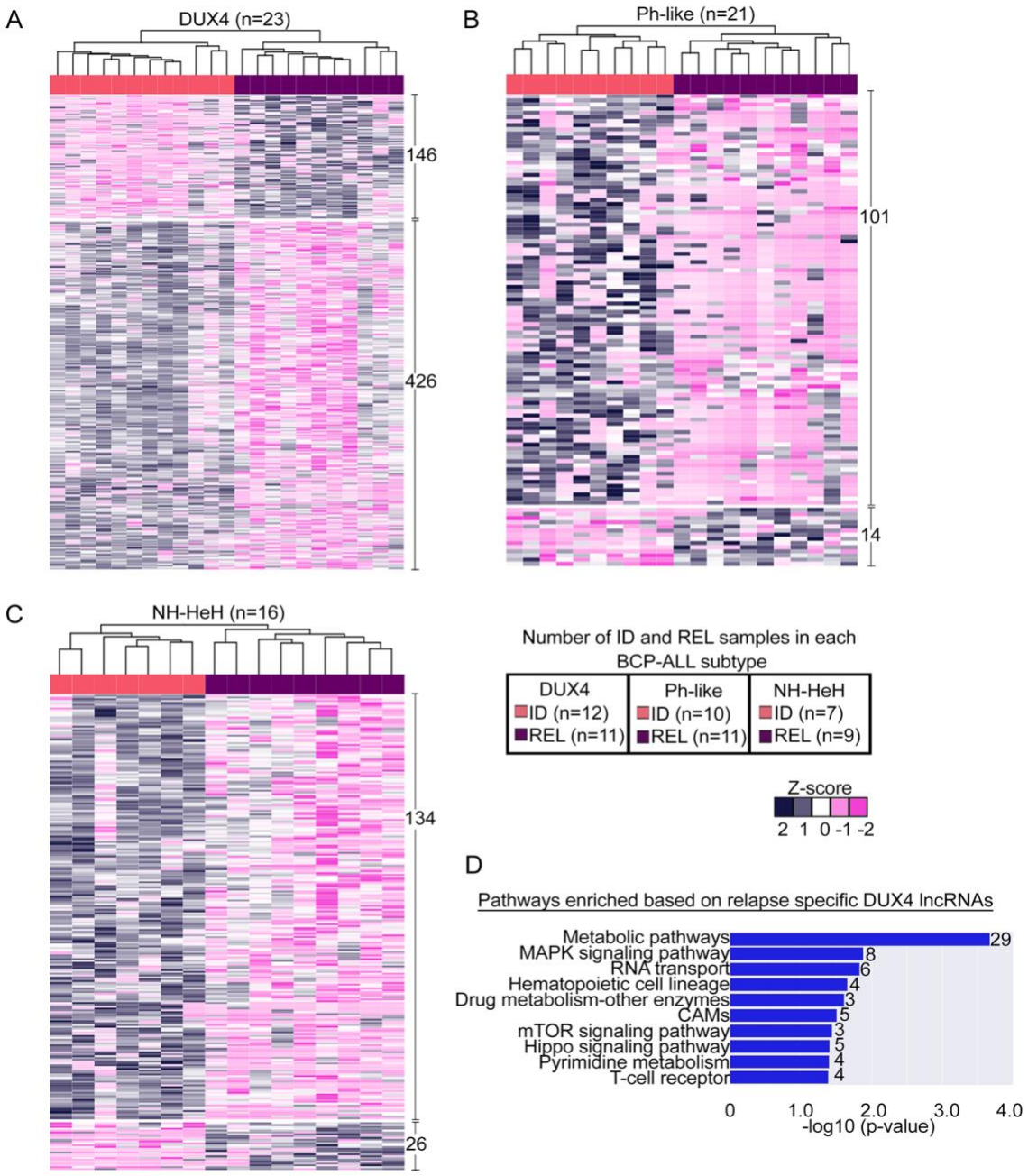

Figure 3.5.1: Relapse-specific DE lncRNAs from BCP-ALL subtypes.
A-C. Heatmap depicting the hierarchical clustering on relapse-specific DE lncRNAs signature on Z-score transformed LIMMA normalized expression values from DUX4, Ph-like and NH-HeH subtypes. Each heatmap shows the up and down-regulated lncRNAs specific to ID and REL samples. D. Molecular pathway analysis with the number of genes involved in each pathway from the enrichment analysis of the nearby (< 100 kb proximity) cis protein-coding genes correlated (Pearson correlation > 0.55 and P-value <= 0.05) with relapse-specific DE lncRNAs in the DUX4 subtype. The legend box indicates the number of ID and REL samples within each group. CAMs: Cell adhesion molecules. E. The overlap between relapse-specific and subtype-specific lncRNAs from three subtypes.

These relapse-specific lncRNAs signature consisted of 14 lncRNAs up-regulated in the Ph-like subtype and 26 DE relapse specific lncRNAs in the NH-HeH subtype, whereas in the DUX4 subtype we found 146 relapses specifically up-regulated lncRNAs (Figure 3.5.1 A-C). The majority of dysregulated relapse-specific lncRNAs observed in our analysis are novel lncRNAs. However, we found 10 lncRNAs from our dysregulated set that had been reported in another disease including a different type of cancers (Table 3.5.8, Figure 3.5.2 A-C). In addition to that, we identified 61 relapse-specific (Appendix D) lncRNAs are significantly overlapped (Hypergeometric test P-value = 2.6 x 10e-4) with a recently published prognostic markers (Ali et al., 2018) lncRNAs from various cancers.

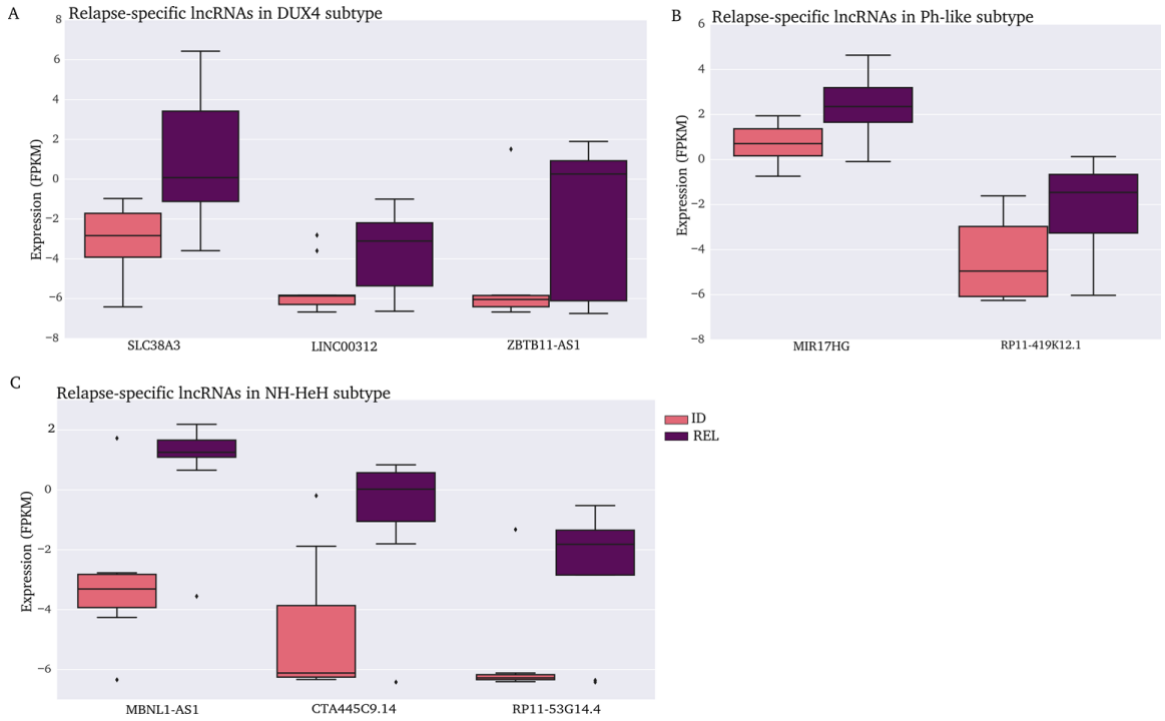

Figure 3.5.2: Relapse-specific lncRNAs markers identified in other cancers.

A. The boxplot represents, examples of relapse-specific lncRNAs, *SLC38A3 (*absolute fold change = -1.961, *P-value* = 7.14 x 10e-4*)*, *LINC00312 (*absolute fold change = -1.028, *P-value* = 1.69 x 10e-3*)* and *ZBTB11-AS1 (*absolute fold change = -1.55, *P-value* = 9.58 x 10e-3*)* which are significantly up-regulated in relapse samples compared to diagnosis samples in DUX4 subtype. B. The boxplot represents, examples of relapse-specific lncRNAs MIR17HG (*absolute fold change = -0.82*, *P-value* = 7.81 x 10e-4) and *RP11-419K12.1 (*absolute fold change = *-1.18*, *P-value* = 1.80 x 10e-3*)* which are significantly up-regulated in relapse samples compared to diagnosis samples in Ph-like subtype. C. The boxplot represents, examples of relapse-specific lncRNAs *MBNL1-AS1* absolute fold change = *-1.95*, *P-value* = 5.62 x 10e-4), *CTA-445C9.14* (absolute fold change = *-1.90*, *P-value* = 2.59 x 10e-3) and *RP1-153G14.4* (absolute fold change = *-1.40*, *P-value* = 5.34 x 10e-3) which are significantly up-regulated in relapse samples compared to diagnosis samples in NH-HeH subtype.

Table 3.5.8: Examples of previously reported lncRNAs identified as relapse-specific lncRNAs in BCP-ALL subtypes.

Relapse-specific lncRNAs	Disease association
TCL6 (DUX4)	Chromosomal translocations T-cell leukaemia/lymphoma
LINC00312 (DUX4, Ph-like, NH-HeH)	Proliferation, invasion, and migration of thyroid cancer, Nasopharyngeal carcinoma
miR-17-92a-1 (DUX4, Ph-like, NH-HeH)	Development, progression, and aggressiveness of colorectal cancer

Table 3.5.8: The differentially expressed lncRNAs between relapse (REL) and initial diagnosis (ID), from three subtypes, which were previously reported for its disease association, selected representative examples from relapse-specific lncRNAs, which are previously identified in other diseases. Representative examples from ten disease-associated lncRNAs

3.5.1 Functional analysis for relapse-specific lncRNAs as markers of BCP-ALL subtypes

We then aimed to infer the putative global molecular functions associated with these relapse-specific lncRNAs. For each subtype, we used the previously defined *guilt-by-association* approach to predict the putative functional pathway involved. Relapse-specific lncRNAs within Ph-like and NH-HeH subtypes did not show any significant correlation with activation of pathways. In contrast, we identified 56% (n = 321) relapse-specific lncRNAs within the DUX4 subtype correlated with their *cis* PC genes. These strongly correlated relapse-specific lncRNAs showed activation of PC genes involved in vital signaling pathways and metabolic pathways. For example, Hippo, mTOR, and MAPK signaling pathways, cell adhesions molecule (CAMS) and metabolic pathways (number of genes involved >= 3 and *P-value* <= 0.05) (Figure 3.5.1 D). These pathways are comparable to the identified pathways from the subtype-specific analysis. Taken together, the results indicate that relapse-specific markers from DUX4 subtype may be functionally engaged in metabolic and signaling pathways (Figure 3.5.1 D).

3.6 DNA Methylation Patterns of lncRNA genes are altered in BCP-ALL subtypes

In order to analyze the methylation status of loci located at the lncRNAs genomic position in the BCP-ALL subtypes, we used DNA methylation array data (collected from Illumina 450k methylation array) from the same patients (n = 45) included matched diagnosis (ID) and relapse (REL) samples (n = 82). Consistent with the RNA-seq dataset, the unsupervised clustering (PCA) on the DNA methylation profile of lncRNAs identified with distinct clusters for DUX4, the Ph-like and the NH-HeH subtypes (Figure 3.6.1 A).

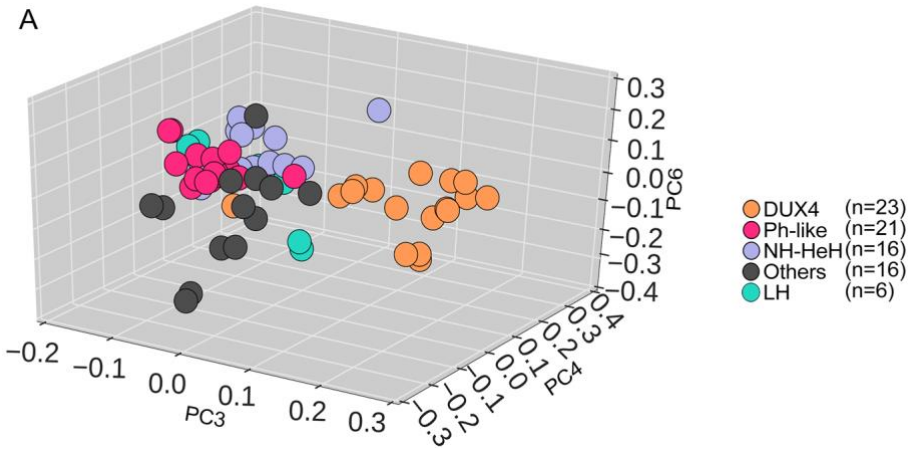

Figure 3.6.1: Hierarchical clustering of CpG's associated with DM lncRNA

A. PCA of CpG's associated with lncRNAs on DNA methylation SWAN normalized beta values. Each point represents a BCP-ALL sample. DUX4, Ph-like, near-haploid, and others are represented by orange, rose, blue and grey respectively.

Given these findings, we next looked into the differential hypo-methylated (Methylation value < 0; P-value <= 0.05) and hyper-methylated (Methylation value >0.2; P-value <= 0.05) CpGs associated with lncRNAs between each subtype (Figure 3.6.2 A-C). We identified 1118 DM lncRNAs from three subtypes, with methylation distribution of 29.25% at promoter-TSS, 46% in the intronic and intergenic region and renaming in the gene body of the genome. About 10% of promoter-TSS methylated lncRNAs displayed a significant inverse correlation with its RNA expression level. In the DUX4 and NH-HeH subtypes the number of hypo-methylated lncRNAs (differential methylation value < 0, P-value <= 0.05) were higher compared to the number of hyper-methylated lncRNAs. Whereas in Ph-like subtype the hyper-methylated (67%) lncRNAs were higher than hypo-methylated (33%) lncRNAs.

We then explored the differentially methylated regions associated with lncRNAs and annotated them based on their genomic position. The differentially methylated CpGs were located in different genomic positions. In each subtype, we identified an average of 28% of DM lncRNAs in the promoter-TSS region (defined as region between -2000 base pair to +2000 base pair within TSS) the remaining within gene body (exon, and Transcription termination site (TTS), Figure 3.6.2 A-C (right panel).

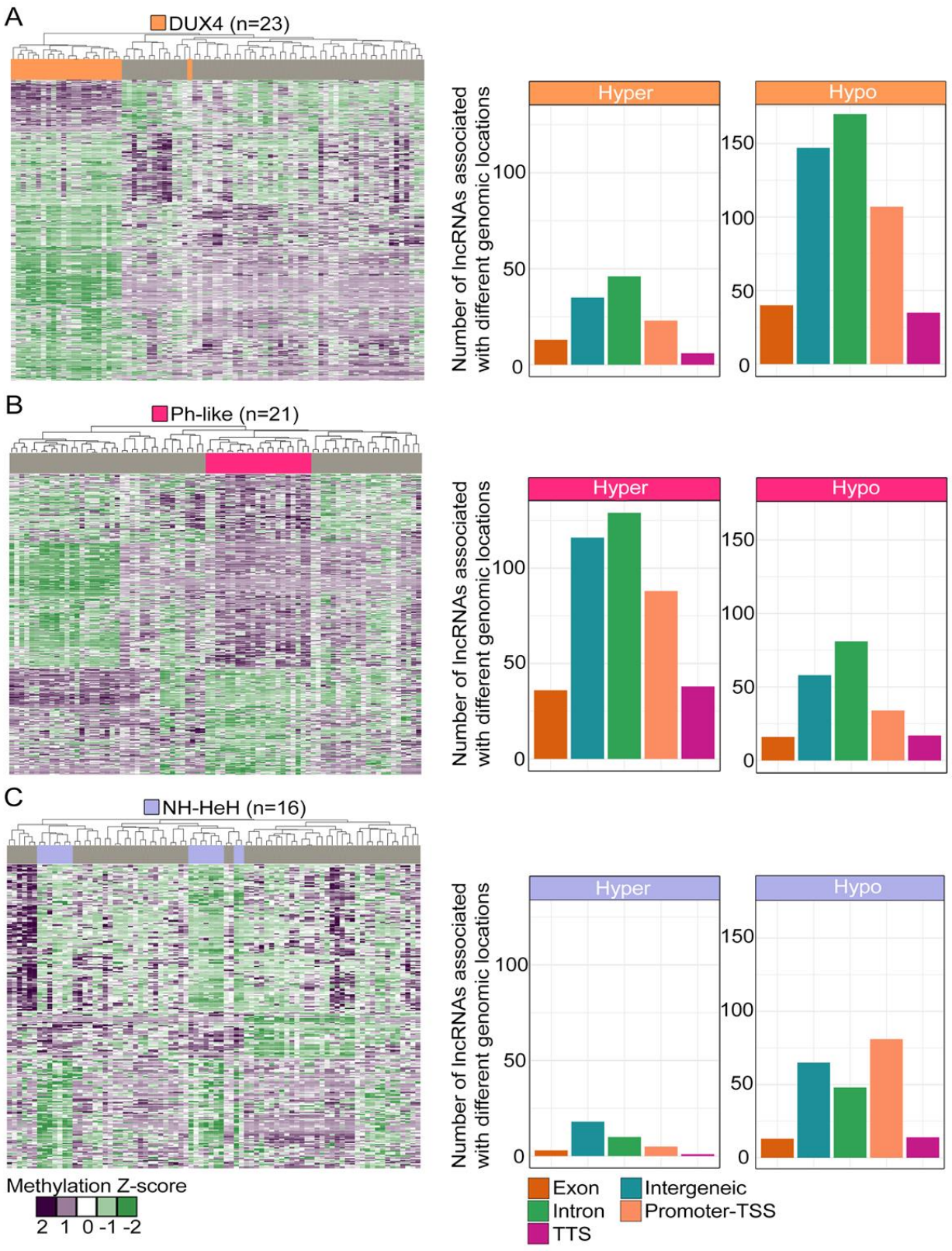

Figure 3.6.2: Hierarchical clustering of CpG's associated with DM lncRNAs from each subtype

A. The heatmap representing hierarchal clustering on 544 differentially methylated (DM) CpG's associated with 434 lncRNAs in DUX4 subtype. In the DUX4 subtype, we identified 328 (76%) differentially hypo-methylated and 106 (25%) hyper-methylated lncRNAs. B. The heatmap representing hierarchal clustering on 518 DM CpG's associated with 450 lncRNAs in the Ph-like subtype. In Ph-like subtype, we observed 302 (67%) hyper-methylated lncRNAs and 148 (33%) hypo-methylated lncRNAs. C.The heatmap representing hierarchal clustering on 295 DM CpG's associated with 234 lncRNAs in NH-HeH subtype. In the NH-HeH subtype, we identified 200 (86%) hypo-methylated and 34 (14%) hyper-methylated lncRNAs. The heatmap is plotted using SWAN normalised Methylation values. The bar plots below each heatmap represent the distribution of DM lncRNAs in the genome (Promoter-TSS and gene body) lncRNAs from each subtype. The distribution DM Promoter-TSS lncRNAs are as follows: 25%, 29% and 39% in DUX4, Ph-like, and NH-HeH subtype, respectively. The promoter methylated lncRNAs, we identified a higher degree of hypo-methylated and lower number hyper-methylated lncRNAs in DUX4 and NH-HeH subtypes. However, the Ph-like subtype has shown a higher degree of hyper-methylated DM lncRNAs than hypo-methylated DM lncRNAs.

3.6.1 Correlation between subtype-specific differentially expressed and differentially methylated lncRNAs

In order to systematically define epigenetically silenced or facilitated lncRNAs in the three subtypes, we correlated the expression FPKM and the beta value of promoter-TSS DM lncRNAs. We identified 6.7% of lncRNAs with hyper-methylated and hypo-methylated promoters with a reduced or increased RNA expression level within BCP-ALL subtypes. For instance, in the DUX4 subtype, 17% (n = 22) of DM lncRNAs at promoter region (Differential hyper-methylation > 0.2; Differential hypo-methylation < 0; P-value <= 0.05) were differentially expressed (P-value <= 0.01 and Absolute fold change = +- 1.5). Out of that, 15 lncRNAs were significantly inversely correlating with their RNA expression levels (Pearson correlation test, two-tailed P-value <= 0.05, Figure 3.6.3 A). Whereas in the Ph-like subtype, 9% (n = 11) of significantly DM lncRNAs at promoter region (Differential hyper-methylation > 0.2 ; Differential hypo-methylation < 0; P-value <= 0.05) are differentially expressed (P-value <= 0.01 and Absolute fold change =+- 1.5), and out of that we found 73% (n = 7) lncRNAs with a significant inverse correlation with their RNA expression level (P-value <= 0.05, Figure 3.6.2 D). Analogously, in the NH-HeH subtype we observed three promoters associated lncRNAs overlapping with the DE signature, where 2 out of 3 showed a significant anti-correlation to their expression level.

Thus, the DM promoter-TSS methylated lncRNAs harboring statistically significant DM at promoter regions, and the strong anti-correlation with its expression level collectively determined 23 putative epigenetically facilitated and silenced lncRNAs from our three BCP-ALL subtypes (Table 3.6.9).

Table 3.6.9: The list of significantly correlated DNA methylation and the expression for promoter methylated lncRNAs (n = 23) from BCP-ALL subtypes.

DM lncRNAs	Pearson correlation coefficient	*P-value*	Methylation	Absolute Fold change	Subtypes
AC003075.4	-0.31	0.004	1.43	-1.26	DUX4
AC099754.1	-0.32	0.002	-1.74	3.2	

AC104655.3	-0.26	0.017	-2.27	2.07	
CACNA1C-AS1	-0.45	2.03E-05	1.97	-1.62	
CTB-25B13.9	-0.26	0.016	-1.73	1.46	
IGF2-AS	-0.24	0.028	-1.33	4.95	
LINC01006	-0.39	0.001	-2.06	2.53	
PVT1	-0.40	0.001	-2.13	1.15	
RGMB-AS1	-0.26	0.0193	-1.48	5.96	
RP11-125B21.2	-0.35	0.001	-1.75	4.11	
RP11-138M12.1	-0.70	5.21E-13	-5.98	3.77	
RP11-367G6.3	-0.30	0.004	1.98	-1.63	
RP11-624M8.1	-0.50	1.34E-06	-3.34	4.13	
RP11-789C17.3	-0.36	0.001	-2.27	3.2	
SERTAD4-AS1	-0.25	0.0232	-1.98	1.79	
LINC01006	-0.38	0.0003	1.44	-1.56	Ph-like
RP11-138M12.1	-0.70	5.21E-13	2.06	-1.44	
RP11-305F18.1	-0.64	5.36E-11	1.76	-2.08	
AC099754.1	-0.33	0.002	1.21	-1.36	
ACVR2B-AS1	-0.36	0.0009	2.18	-1.75	
LINC00996	-0.39	0.0003	-1.56	2.11	
ERICH1-AS1	-0.40	0.0006	-1.82	2.21	
DIO3OS	-0.31	0.0037	-1.76	4.05	NH-HeH
U3	-0.83	1.346E-22	-2.01	2.43	

Table 3.6.9: The lncRNAs are promoter differentially methylated and correspondingly differentially expressed. DM: Differentially methylated. The significance is calculated based on Pearson correlation rate and two -tailed *P-value* <= 0.05.

Of these 23 epigenetically modulated lncRNAs, we observed novel lncRNAs to show an anti-correlation between the DNA methylation rate and expression levels. For instance, in the DUX4 subtype, lncRNA *R11-138M12.1* and *RP11-624M8.1*, showed a significant hypo-methylated at their promoter region and is transcriptionally up-regulated in the DUX4 subgroup (Pearson correlation coefficient = -0.69; *P-value* = 5.1E-13 for *R11-138M12.1*; Pearson correlation coefficient = -0.50; *P-value* = 1.36E-06 for *RP11-624M8.1*; Figure 3.6.2 B -C). The same lncRNA *R11-138M12.1* showed significant hypermethylation at the promoter region and a concordant down-regulation in the Ph-like subgroup (Figure 3.6.2 E).

Besides the novel lncRNAs, we identified certain others lncRNAs which are previously brought in the context of different cancers from our epigenetically facilitated set. The lncRNA *PVT1* is an example which we observed in the DUX4 subtype with significant anti-correlation (*P-value* <= 0.01) to its expression level. These findings suggest that epigenetic silencing of lncRNAs may be a mechanism that contributes to the dysregulation of expression of lncRNAs in the BCP-ALL subtypes.

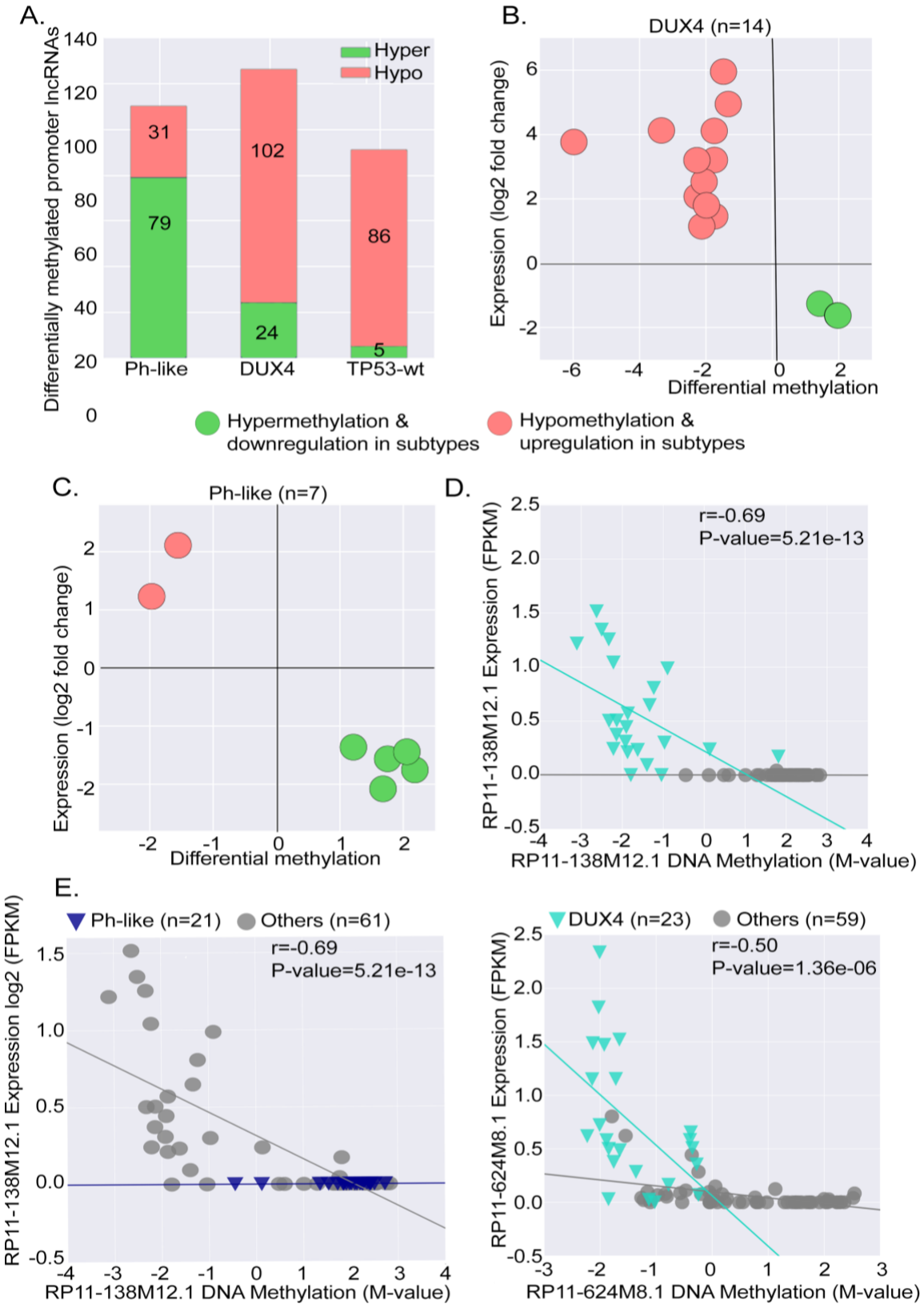

Figure 3.6.3: The epigenetically altered promoter methylated lncRNAs and their expression.

.

A. The barplot depicts the distribution of hyper and hypo-methylated LncRNAs in promoter region. B-C. The promoter-TSS DM lncRNAs with significant negative correlation with DE expression profile from the DUX4 and Ph-like subtypes. D. Two representative examples of hypo-methylated lncRNAs with increased expression profile from a DUX4 subtype with significant inverse correlation between DNA methylation and expression levels. The lncRNA *RP11-138M12.1* (Pearson correlation coefficient = -0.69, 2-tailed P-value = 5.21e-13), *RP11-624MB.1* (Pearson correlation coefficient = -0.50, P-value = 1.36e-06) are examples with hypo-methylation and up-regulated expression pattern. E. A representative example of the promoter hyper-methylated lncRNA, *RP11-138M12.1* (Pearson correlation coefficient = -0.69, 2-tailed P-value = 5.21e-13) with down-regulated expression pattern, and with inverse correlation within the Ph-like subtype.

3.6.2 Chromatin markers associated with intronic and intergenic methylated subtype-specific lncRNAs

Around 46% (n = 512) of the DM subtype-specific lncRNAs are localized in the intronic and intergenic genomic regions. We next aimed to investigate whether these lncRNAs regions have chromatin markers encoded within their genomic location. A recent human genome-wide chromatin marker study (114) has provided us with a rich resource to identify chromatin markers. Genome-wide mapping of B-lymphocyte cell line by searching for epigenetic markers within our DM subtype-specific intronic and intergenic regions revealed a significant number of lncRNAs (n = 53, Fisher extract test P-value = 2.2E-16) with enhancer and insulator markers. Out of these, lncRNAs, *RP11-134O21.1*, *RP11-398B16.2*, *RP11-689B22.2*, *CTC-458I2.2,* and *LINC00880* were DE expressed, with a significant negative correlation between DNA methylation and expression levels in the DUX4 subtype (Table 3.6.10). Together these show both intronic and intergenic DM lncRNAs associated with strong enhancer and insulator regions can accelerate its expression at the epigenetic level.

Table 3.6.10: The list of significantly correlated DNA methylation and expression for intronic and Intergenic methylated lncRNAs (n = 5) from DUX4 BCP-ALL subtypes.

DM lncRNAs	Absolute Fold change	Methylation value	Pearson correlation rate	*P-value*	Epi-markers	Biotype
RP11-134O21.1	2.54	-1.56	-0.63	1.9E-010	Enhancer	Intron
RP11-398B16.2	2.08	-1.85	-0.47	0.0007	Insulator	
RP11-689B22.2	1.52	-3.37	-0.47	0.008	Enhancer	
CTC-458I2.2	-1.16	3.38	-0.42	0.0001	Enhancer	
LINC00880	-1.45	2.23	-0.25	0.02	Enhancer	Intergenic

Table 3.6.10: The significance is calculated based on Pearson correlation rate and two -tailed *P-value* <= 0.05. The lncRNAs are promoter differentially methylated and differentially expressed in their corresponding subtypes. These lncRNAs are with enchancer and insulator epigenetic markers. DM: Differentially methylated.

We further compared our list of DMR associated lncRNAs from each subtype with published list disease-associated lncRNAs and identified 24 previously reported disease associated lncRNAs (Table 3.6.11) within our BCP-ALL subtype-specific DM lncRNAs.

Table 3.6.11: The list of DM lncRNAs which are previously reported due to there disease associations (n = 24) from BCP-ALL subtypes.

DM Subtype-specific lncRNAs	Subtypes	Annotation	Methylation value	P-value
ADAMTS9-AS2	Ph-like	intron	1.777801	0.00171869
DANCR	NH-HeH	promoter-TSS	-1.787494	0.02195348
DLEU2	NH-HeH	intron	-1.857272	0.007403689
DLX6-AS1	DUX4	Intergenic	-1.639343	0.007447799
	NH-HeH	intron	-1.368489	0.03913844
EGOT	Ph-like	intron	-1.995476	0.01631441
ERICH1-AS1	DUX4	intron	-2.788796	0.0000274
	NH-HeH	intron	-2.396816	0.0002673425
FENDRR	NH-HeH	intron	-1.742717	0.004852748
HOTAIRM1	DUX4	promoter-TSS	-1.858395	0.0002810779
	Ph-like	promoter-TSS	1.428301	0.0202788
HOXA-AS2	NH-HeH	promoter-TSS	-1.509447	0.02797619
HOXA11-AS	NH-HeH	Promoter-TSS	-1.84395	0.001730609
	Ph-like	exon	-1.934465	0.02145395
IGF2-AS	DUX4	Promoter-TSS	-1.326866	0.0322549
	NH-HeH	Promoter-TSS	-1.405105	0.03537591
KCNQ1OT1	Ph-like	exon	2.251927	0.008248937
LINC00261	Ph-like	intron	-2.112499	0.01462022
LINC00467	DUX4	promoter-TSS	2.000734	0.0018457
LINC00473	DUX4	promoter-TSS	-1.442647	0.02018019
MEG3	DUX4	intron	-2.518794	0.0001154195
NEAT1	DUX4	exon	2.230128	0.000287578
	Ph-like	exon	-1.787521	0.01741656
PVT1	DUX4	Promoter-TSS	-2.126776	0.0009309612
	Ph-like	Promoter-TSS	1.960143	0.002582014
RGMB-AS1	DUX4	Promoter-TSS	-1.471133	0.007150467
RP11-325I22.2	NH-HeH	Intergenic	-1.778581	0.02285671
TCL6	DUX4	Promoter-TSS	1.749096	0.006956922
	Ph-like	intron	1.844357	0.04590776
TP73-AS1	DUX4	Promoter-TSS	-1.34541	0.01637621

UCA1	DUX4	intron	-4.572273	0.000000778
	Ph-like	intron	2.03285	0.02043621
ZNRD1-AS1	Ph-like	Promoter-TSS	1.679822	0.01186042

Table 3.6.11: The list of previously reported lncRNAs which are associated with other disease including leukemia identified within our DM subtype-specific lncRNAs. For example, HOTAIRM1, previously reported to be associated with malignant hematopieosis.

Chapter 4. Discussion

BCP-ALL is a genetically heterogeneous disease consisting of many subtypes. Understanding the molecular signature behind these subtypes could improve its diagnosis and treatment. In the recent past, compared to protein-coding genes, there is a growing appreciation for investigating the role of lncRNAs in cancer development and progression given their surprising functions and their aberrant gene expression patterns. Now, it is apparent that lncRNAs are involved in the tumorigenesis of leukemias. Nevertheless, a comprehensive characterization of the transcriptome, DNA-methylation, and their functional contribution in distinct BCP-ALL subtypes are lacking.

Unsupervised clustering of lncRNAs expression and DNA methylation profile demonstrated the ability of lncRNAs in classifying the established BCP-ALL subtypes within our cohort. This finding was further subsequently validated in an independent validation cohort of 47 patients. In addition to that, we cataloged a comprehensive set of 1564 subtype-specific and 941 relapse-specific lncRNAs using RNA-Seq data. Finally, we present a catalog of 1118 differentially methylated lncRNAs based on the DNA methylation array data from the same patients across the three subtypes. In addition, to that we highlight 23 lncRNAs whose expression levels were epigenetically facilitated or silenced.

Interestingly, 36% of DUX4 and Ph-like specific lncRNAs (n = 229) and 62% (n = 321) of relapse-specific lncRNAs within DUX4 were found be associated with pivotal signaling and metabolic pathways relevant to the progression of ALL. We present a catalog of lncRNAs based on an integrative analysis which brought significant insight and advances over previous studies as it provides the most comprehensive dataset and their potential functions in BCP-ALL, a resource of clinically relevant relapse-specific lncRNAs signature and discloses their utility in prognosis.

The discussion part is divided into two sessions the first section discusses the methods used in the thesis and in the second part, I evaluate the major results. One of the challenges of this thesis was to determine the best practices for our RNA-Seq data analysis. Our first goal was to determine right algorithms which addresses all the major caveats of the datasets in order to draw reliable conclusions. In the following section we discuss the major advantages and disadvantages of the algorithms used.

4.1 RNA-seq for determining the subtype-specific and relapse-specific lncRNAs

RNA-Seq technology revolutionized the study of transcriptomes. RNA-Seq allows the detection and the quantification of expressed genes or transcripts in a biological sample. RNA-Seq has clear advantages over previously existing technologies, such as microarray and Sanger sequencing. For example, microarray relies upon existing knowledge about genome sequence, high background levels owing to cross-hybridization and a limited dynamic range of detection owing to both background and saturation of signals. Unlike, former technologies, the RNA-seq approach is not limited to detecting known lncRNAs, but is also able to identify novel lncRNAs that are present in the human genome. (Zhong Wang, Gerstein, & Snyder, 2009). Owing to the rapid decrease in sequencing costs, RNA-seq may soon replace microarrays completely. However, choosing the right alignment tool is always is the key to RNA-Seq data analysis for maximum detection of significant genes involved in BCP-ALL samples. With the onset of RNA-Seq technology, there are some tools and methods developed for the RNA-seq sequence alignment. The appropriate mapping and read quantification algorithm were chosen based on their capabilities to address existing constrains. This enabled extraction of reliable information out of raw RNA-seq data.

In certain instances, we observed increased misalignment of spliced reads or junction reads for lncRNAs. This was true for antisense and sense lncRNAs. For example, the read sequences mapping to the exon-exon junction of lncRNAs were borrowed from their sense of PC genes, resulting in increased misalignment. Inherent problems of all *de novo* RNA-seq aligners including, Tophat and cufflinks, are its inability to accurately detect the splicing events or junction reads that are in the short sequence on the junctions. Therefore, this problem increases the under-detection of splicing events or increases the misalignment rate.

A recently developed aligner, *STAR-align* provided solutions for the above issues combined with faster performance. *STAR-align* mitigates this problem by its 2-pass option in which it first obtains information about possible splice junction loci from annotation databases. Then it is also possible to run a second mapping pass, supplying it with novel splice junction loci found in the first mapping pass. In the second pass, STAR will not discover any new junctions but will align spliced reads with short overhangs across the previously detected junctions. The STAR-align takes the splice junction reads into accounts by its in-built parameters. Additional reasons for using STAR-align tool was due to its better mapping accuracy (sensitivity and precision) and computational resources (runtime and disk space) compared to former mapping tools (Dobin et al., 2013). We, therefore, decided to finalize STAR-align for our RNA-seq datasets owing to its advantages over former tools.

4.2 Transcriptome alignment and read quantification

The transcriptome is the complete set of the transcript in a cell for a specific developmental stage or physiological condition. Studying the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding developments that are attributable to the disease. LncRNAs have complex patterns of expression and regulation compared to PC genes. Like RNA-seq reads alignment, determining the quantities of these reads were also challenging. The major challenge was ambiguous read mapping due to close paralogs present in the genome and the generally low expression level of lncRNAs. In addition to that our data set is strand non-specific RNA-seq data composed the second challenge in the work-flow construction to determine the useful tool for read quantification or extracting the gene expression count.

When the dataset is strand non-specific the gene expression quantification of some subtypes of lncRNAs other than lincRNAs is not possible, as they may share sense and antisense sequence with PC genes with overlapping exons. When there are a sense and antisense overlapping gene across a given lncRNAs gene body, the tendency of borrowing reads from these neighboring partners would increase, and therefore an accurate estimation of the exact gene expression level is not possible. The *StringTie* (transcriptome assembler) algorithm overcame this issue with its final transcriptome assembly and read quantification solution, regardless of the non-strand specificity of our reads. The StringTie algorithm accurately quantified reads which are mapped in splice junctions or exon-exon junction of each lncRNAs without borrowing its neighboring reads from the sense protein-coding genes. To test the ability of *StringTie* with different other programs which quantify gene expression we tested *feature-count,* and *Ht-seq*. These tools resulted in biased results for the sense and antisense lncRNAs. In addition to that, the output files from StringTie can be processed by R Bioconductor differential expression analysis programs like Deseq2, edgeR, LIMMA Voom. Whereas other similar transcriptome assemblers (example Cufflinks), need one more additional step of read quantification using other tools like Ht-seq. Therefore, using *StringTie* enabled us to reduce the complexity of characterizing antisense or partially intronic transcripts from our strand unspecific RNA-Seq libraries. StringTie had fast run, which is several times faster than Cufflinks (our former tool of choice), HTSeq (Anders, Pyl, & Huber, 2014), and FeatureCount (Liao, Smyth, & Shi, 2014).

4.3 Addressing major caveats in our multi factorial design model for differential

expression analysis

Accurate detection of significant differentially expressed genes is one of the essential steps in RNA-seq work-flow. RNA-seq data is heteroscedastic in nature, meaning that, some genes have a higher variance than others. However, the ultimate rationale in the choice of the tool for DE analysis is based on ability of an algorithm which could leverage all caveats present in the datasets. In our case, we had 82 samples from 45 patients, from 2 disease stages with corresponding matching diagnosis (primary sample) and relapse from each patient. However, for 8 individual patients, we had either of ID or REL samples. And DESeq2 has no statistical function to leverage this sample imbalance and biological dependency or confounders. The differential expression analysis using R-Bioconductor, *LIMMA Voom*, accommodates the mean-variance relationship using precision weights calculated by the *Voom* function and has additional features to accommodate the co-founders within our dataset.

4.4 Functional enrichment analysis of lncRNAs

The high throughput technologies have enabled a detailed exploration of lncRNAs in different diseases. Functional association of lncRNAs is a daunting task, due to low expression rate, and cell or tissue-specificity. However, the functional role, if any, of these lncRNAs, are largely unexplored. Unlike protein-coding genes, lncRNAs have no functional databases like DAVID, reactome, or Panther. Recently, there are methods like "*guilt-by-association*" approach is used to identify potential functions of lncRNAs. By systematically applying this approach we defined the functions of our subtype-specific and relapse-specific lncRNAs. The approach is based on the hypothesis described in several earlier studies, that lncRNAs exert their functions by regulating neighboring (*cis*) and distally (*trans*) located protein-coding genes. The correlation-based "*guilt-by-association*" approach is based on the hypotheses that co-expressed genes are more likely to be co-regulated, or share the same functions. Compared to other methods, including genome-wide-clustering, which is based on hierarchical clustering, L-means and self-organizing maps (SOMs) require some form of a selection of cluster number or size. Therefore, these methods require careful selection of parameters to ensure that the clusters obtained would result in reliable conclusions. The network approach is another method similar to clustering approach, which is a computationally difficult task, even though there exist several algorithms and statistical approaches for this purpose.

Functional enrichment analysis is mostly done using the database. For example, Database for Annotation Visualization and Integrated Discovery (DAVID) (Huang, Sherman, & Lempicki, 2008), and Gene Set Enrichment Analysis (GSEA) (Y. Lee, Huang, & Zhang, 2006; Subramanian et al., 2005). However, one

drawback of such tools is since these databases are growing dramatically and it becomes more difficult for those enrichment tools to include that updated information. Therefore, the need for an enrichment tool which performs the real-time analysis is required. We used Gene Set Clustering based on Functional annotations (GeneSCF) for the functional enrichment analysis. GeneSCF is a relatively new tool which performs real-time functional enrichment analysis. GeneSCF makes use of source databases directly while enrichment analysis is performed. This tool is a more reliable compared to its predecessors mainly because of its real-time analysis nature (Subhash & Kanduri, 2016).

4.5 DNA methylation array on subtype-specific lncRNAs profiling

We used Illumina's Infinium HumanMethylation450 BeadChip for profiling lncRNAs within our BCP-ALL samples. Compared to the current methods for genome-wide DNA methylation profile such as tilling microarrays and bisulfite genomic DNA sequencing of selected regions, the Illumina's Infinium HumanMethylation450 BeadChip has its advantages. For instance, these methods (tilling micro-arrays and bisulfite genomic DNA sequencing) require large amounts of sample material and labor, making it difficult to use in large-scale studies where there is a limited number of samples are used. The Infinium HumanMethylation450 array makes it possible to assess the methylation status of >450 000 CpGs located throughout the genome, which means it covers 96% of CpG or known as CGIs (regions where cytosine nucleotide is followed by guanine nucleotide, separated by a phosphate group) islands within the genomes, offering a comprehensive coverage with high-throughput compatibility to large sample size (Dedeurwaerder et al., 2014). The signal intensity is measured using an Illumina scanner (iScan) to generate the beta values.

4.6 Unsupervised hierarchal clustering revealed lncRNAs expression and methylation pattern correlated with established molecular subtypes of BCP-ALL

Unsupervised hierarchal clustering of lncRNAs expression and DNA methylation data identified the robust classification of previously established BCP-ALL subtypes from 45 relapsed patients. Emerging evidences have shown that lncRNAs can be used as a classifier in the molecular subtypes of different solid tumors, including, ovarian cancer, glioma, and lung squamous cell carcinoma (Du et al., 2013). Of the few studies performed in leukemias, most have analyzed the association of lncRNAs in AML (Lei et al., 2017) and pediatric ALL (Gioia et al., 2017) classification there are no much comprehensive studies on BCP-ALL subtypes. Furthermore, our study framework is based on the relapsed BCP-ALL patients

from different age groups of both pediatric and adult cases. This observation was true for DNA-methylation dataset and 1534 subtype-specific and 1118 differentially methylated lncRNAs. Taken together, our analysis provides insights of lncRNA-based stratification of BCP-ALL patients on the established molecular subtypes of ALL.

4.7 Validated set of BCP-ALL subtype-specific lncRNAs

We further corroborated the ability of lncRNAs to classify the established molecular subtypes in an already processed independent validation cohort of 47 BCP-ALL patients. In line with our observation, we identified similar clusters of molecular subtypes based on lncRNAs expression profile in the independent validation cohort upon unsupervised hierarchical clustering. This result strengthens our preliminary observation that subtype-specific lncRNAs can be used as classifiers like PC genes in molecular subtypes of BCP-ALL. In-depth transcriptomic analyses using RNA-seq revealed that lncRNAs profiling could recapitulate the molecularly defined subtypes of BCP-ALL, which agrees with the findings of recent studies (W. Zhao, Luo, & Jiao, 2014).

4.8 BCP-ALL subtype-specific lncRNAs showing oncogene properties like drug resistance

Besides the subtype-specificity, we also identified previously detected/described lncRNAs within our set. A closer look at these molecular subtype–specific lncRNAs identified 23 lncRNAs previously validated and reported as onco-lncRNAs in different cancers including leukemia. The representative examples of these onco-lncRNAs in the subtypes are discussed in the following session. For example, tumor suppressor *GAS5* is up-regulated in the DUX4 subtype; it has been reported to be related to cell-cycle arrest and apoptosis properties in other cancer types (Nobili et al., 2016). Recently, an isoform of *GAS5* lncRNAs (*GAS6-AS2*) (Bester et al., 2018) is reported to be associated with chemotherapy resistance in AML. However, there are no direct reports of *GAS5* in BCP-ALL subtypes. Notably, the highly expressed *MIR155HG* lncRNA mainly associated with B-cell malignancies (CLL) and B-cell receptor signaling was down-regulated within the DUX4 subtype (Vargova et al., 2011). *MIR155HG* associated with aggressive phenotype in cytogenetically normal acute myeloid leukemia (CN-AML) was found to be down-regulated in DUX4 subtype compared to others. Another example is lncRNA *PVT1*, which is widely reported in AML is up-regulated in the DUX4 subtype. The lincRNA *MIAT* up-regulated in Ph-like subtype within our cohort, over-expression of *MIAT* is reported to be associated with CLL (Sattari et al., 2016). The lincRNA *CCDC26* (Hirano et al., 2015) associated with cell proliferation,

differentiation and apoptosis in AML is up-regulated Ph-like subtype. Besides, the known lncRNAs, we have identified novel lncRNAs as BCP-ALL subtype-specific lncRNAs. Taken together, these subtype-specific lncRNAs signature from different molecular subtypes may serve in defining the core lncRNAs that orchestrate the key oncogenic properties of BCP-ALL subtypes.

4.9 Relapse-specific lncRNAs markers in BCP-ALL subtypes

A hand full of studies consistent with our studies reported the role of lncRNAs in relapse and its importance as prognostic factor (Zhonghao Wang, Wu, Feng, Zhao, & Tao, 2017) in several cancer types (Ali et al., 2018). However, so far there are no studies reporting relapse specific lncRNAs in BCP-ALL subtypes. Like protein-coding RNAs, several lncRNAs are reported as markers of other diseases. In our analyses, we found associations between lncRNAs and relapse on three different subtypes of BCP-ALL. One of the leading causes of death in ALL patients is the disease relapse. Chemotherapy resistance of relapsed blasts compared to what is observed in diagnosis is a key hallmark of ALL relapse. Innovative strategies are urgently required due to the frequent failure of conventional salvage chemotherapy, including intensified drug schedules and stem cell transplantation, in the treatment of relapsed ALL (Bhatla et al., 2014). Thus, there is always a great interest in characterizing the molecular drivers of relapsed ALL because it has a poor outcome with conventional therapy and is increasing with age (Iacobucci & Mullighan, 2017).

We cataloged a comprehensive set of 941 relapse-specific lncRNAs which are driving the BCP-ALL progression. Initially, we looked for relapse-specific lncRNAs in adults and pediatric patients. Due to the heterogeneity of BCP-ALL, we could not infer any significant relapse-specific lncRNAs and thus investigated relapse-specific lncRNAs signature within each subtype. Within each subtype, we found that the relapse-specificity of lncRNAs was more pronounced than in the whole cohort (n = 82). Interestingly, when compared to relapse-specific mRNA signature the relapse-specific lncRNAs signature was stronger with clear separation between ID and REL samples for all three subtypes. Gene expression profiling of leukemic blasts in the matched diagnosis and relapse patient pairs in protein-coding genes has revealed a common gene signature reflective of relapse, gene markers involved in proliferation and cell cycle regulation, apoptosis, DNA repair and drug resistance (Bhatla et al., 2014). However, lncRNAs based relapse-specific markers are not studied much, especially in the context of BCP-ALL.

4.9.1 Relapse-specific onco-lncRNAs

Although the majority of the relapse-specific lncRNAs identified were novel ones, we identified a hand full of previously defined onco-lncRNAs. The examples included *MIAT*, *CCDC26*, *TCL6*, *RP11-701P16.5*, *MIR503* Host Gene (*MIR503HG*), *MIR17HG* and *GAS5*, with cell-cycle arrest and apoptosis properties in cancer (Kitagawa et al., 2013). In the DUX4 subtype, we observed lncRNAs, T-cell leukemia/lymphoma 6 (*TCL6*), which was up-regulated in diagnosis compared to relapse. *TCL6* is reported to its leukemogenesis properties in T cell leukemia (H. et al., 2017). However, its association with BCP-ALL is not documented, and its functions are unclear. The results of our "guilt-by-association" study highlights its association with the PI3K-Akt signaling pathway in DUX4 subtype.

Other examples are, the BDNF antisense RNA (BDNF-AS) and Insulin-like growth factor 2 antisense (IGF2-antisense, embryonic stem cell-related (ESRG) are up-regulated in the diagnosis and LINC00312 up-regulated in relapsed samples. Among the disease-associated lncRNAs, in Ph-like, we observed *SNHG3* (a marker for malignant melanoma), as up-regulated in relapse stages. The lncRNAs AP000688.29 was down-regulated and *MIR17HG* was up-regulated in relapse. The lncRNA *MIR17HG46* was reported to suppress apoptosis in myc-driven lymphomas (Ott, Rosenwald, & Campo, 2013) and was DE in relapse compared to diagnosis samples within the Ph-like subtype. In NH-NeH subtype we observed *AP000688*.29 and *IFNG-AS1* as down-regulated, whereas MBNL1 antisense RNA 1 (*MBNL1-AS1*), *CTA-445C9.14* and *RP1-153G14.4* were up-regulated in relapse samples. Overall, the relapse-specific lncRNAs highlights the oncogenic relevance in BCP-ALL subtypes.

4.9.2 Relapse-specific lncRNAs as prognostic markers

Besides the oncogenic properties, lncRNAs can act as prognostic markers and aid for disease diagnosis and treatment (Ali et al., 2018). We identified a significant enrichment of a subset of relapse-specific lncRNAs (n = 61) with recently identified independent prognostic markers from 14 different solid cancer types. Out of these, for example, lncRNA LUCAT1 was previously reported for its role in drug resistance in solid cancer (Z. Han & Shi, 2018). Within the DUX4 subtype, we identified up-regulated expression of LUCAT1 at relapse, providing a novel insight into treatment resistance for BCP-ALL subtypes. Together, this illustrates that the catalog of relevant lncRNAs in different subtypes of BCP-ALL serves as subtype-specific and relapse-specific markers with the potential of RNA based treatments for BCP-ALL subtypes.

4.10 Molecular functions identified using subset-specific and relapse-specific

lncRNAs

The subtype-specific and relapse-specific lncRNAs showed significant correlation between genes enriched in key pathways associated with cell proliferation, growth, survival, metabolism, and autophagy based on the correlations with their neighboring and distinct protein-coding genes. These findings indicate that BCP-ALL subtype-specific and relapse-specific lnRNAs are associated with tumorigenesis of hematopoietic cells. LncRNAs are emerging as new players in cancer, due to their potential roles in both oncogenic and tumour suppressive pathways. They are frequently dysregulated in a variety of human cancers; however, the biological functions of a vast majority of them remain unknown. Recently, evidence of lncRNAs molecular mechanisms and function has begun to accumulate, providing insight into the functional roles they may play in tumorigenesis (Serviss, Johnsson, & Grandér, 2014).

The *guilt-by-association* method designates potential or putative functions to lncRNAs based on its co-expression of characterized PC genes. The certainty of the association is based on the condition of available expression data. For instance, time-dependent data can be notably crucial because aberrant regulation of expression can be informative of the certain pathways by which lncRNAs functions (Bartonicek, Maag, & Dinger, 2016). Our study consists of time-dependent dataset from two disease stages. The appropriateness and popularity of this approach have given rise to several subtypes of analyses including the *cis* and *trans*-correlation-based association.

We identified a remarkable fraction of subtype-specific lncRNAs (621 out 1534) with significant co-expression with their *cis* and *trans* located protein-coding genes. Notably, 32% of these lncRNAs are involved in pathways associated with proliferation, apoptosis and differentiation in leukemia, including, JAK-STAT, mTOR, PIK3-AKT, TGF-beta, MAPK, P53, hippo and NF-kappa B signaling pathways from both DUX4 and Ph-like subtypes. The co-expression between the protein-coding genes and subtype-specific lncRNAs provided a possible explanation of co-regulation or co-activation of lncRNAs, with their *cis* and *trans* PC genes. We also demonstrated that several lncRNAs were co-expressed with oncogenes associated with leukemias.

4.10.1 Potential functions of DUX4 specific DE lncRNAs associated with signaling pathways

In the DUX4 subtype, we report lncRNAs signature (n = 185, both *cis* and *trans* based analysis) associated with pathways reported to play a key role in leukemogenesis, such as TGF-Beta signaling pathway, P53, Endocytosis, hippo, proteoglycans, and pathways in cancer. Considering the functional nexus between these lncRNAs and leukemia related pathways, targeting these lncRNAs provide novel insights for new therapeutic targets.

In ALL, *TGFB* has complex roles; it regulates the proliferation of the distinct myeloid stem cells (Dong & Blobe, 2006). Recently there are some lncRNAs documented to be associated with *TGFB* gene. For instance, *Lnc-ATB*, a TGFB2 induced lncRNA that could mediate TGF-β-induced epithelial-mesenchymal transition and has been reported to promote metastasis in various solid cancer such as hepatocellular carcinoma, colorectal cancer, gastric cancer, and breast cancer. However, the lncRNAs associated with BCP-ALL subtypes are not reported. We identified the antisense lncRNA, *RP11-224O19.2* and other novel lncRNAs significantly correlated with *TGFB*, and are enriched in TGF-beta pathway, indicating their functional relatedness or regulatory relationships. Interestingly, the subtype-specific lncRNAs and subtype-specific PC are globally predicted to activate or inhibit the same key signaling pathways in the DUX4 subtype.

4.10.2 Potential functions of Ph-like specific DE lncRNAs associated with signaling pathways

The Ph-like subtype is both molecularly and functionally well characterized based on mRNAs/protein expression levels, whereas non-coding genes are not much studied. We have identified a list of 24 novel dysregulated Ph-like specific lncRNAs crucial in signaling pathways associated with Ph-like subtype. The pathways controlling the cell proliferation, differentiation, and survival of hematopoietic cells were identified based on functional enrichment analysis, for example, the PI3K and mTOR signaling pathways. In addition to that, our functional predictions identified other prominent pathways which trigger chemotherapy resistance in BCP-ALL, including, JAK-STAT2, Cytokine-cytokine receptor and endocytosis pathways. The lncRNAs associated with these pathways are antisense or sense intronic to the mRNA genes with a significant co-expression pattern. Characterization of the lncRNAs involved in this pathway may be of interest in the search for new potential therapies.

Some of the functions predicted here have been validated by previous studies, suggesting that our *guilt-by-association* approach is valid. For example, lncRNA *AC002454.1* was recently reported to regulate *CDK6* to participate in cell cycle dysfunction in the endometriosis pathogenesis. LncRNA *AC002454.*1 is an antisense lncRNA of *CDK6* gene. The results of our *guilt-by-association* study highlight an association of this lncRNA with the PIK3-Akt pathway. Both the *CDK6* gene and antisense *AC002454.1* are significantly co-expressed and up-regulated in the Ph-like subtype.

Interestingly, we observed a significant co-expression between oncogene *IL2RA* and its antisense lncRNA *RP11-536K7.5*. Recently, *IL2RA* gene was found to be specifically up-regulated by pre-B cell receptor (pre-BCR) signaling during early B cell development, and cells with oncogenically activated tyrosine kinases by a manifold to pre-BCR signalling in both in Ph+ALL and in Ph-like ALL (J.-W. Lee

et al., 2015). In Ph-like subtype, we observed *IL2RA* gene enriched in the cytokine-kinase signaling pathway. Both *IL2RA* and *RP11-536K7.5* were up-regulated in Ph-like samples. The ability of *IL2RA* to stabilize oncogenic signaling strength in Ph-like ALL is important for leukemia initiation and development. Our analysis indicates the co-regulation or co-regulation of *RP11-536K7.5* with *IL2RA* gene, which provided a new context for further characterization of *RP11-536K7.5* lncRNA. We predicted the positive association between cytokine-kinase signaling pathway and *RP11-536K7.5* lncRNA, but the mechanisms involved are still poorly understood, therefore, further studies are needed to better understand *RP11-536K7.5*/ cytokine-kinase signaling transduction.

It was noteworthy that subtype-specific lncRNAs and subtype-specific PC are globally predicted to activate or inhibit the same pathways. However, some exclusivity appeared. For example, lncRNAs specific to the Ph-like subtype particularly involved in the activation of mTOR and the PI3K-Akt signaling pathway. Considering the functional nexus between Ph-like specific lncRNAs and the activation of pathways such as mTOR and PI3K signaling pathways, targeting those lncRNAs may be a promising novel therapeutic option for BCP-ALL subtypes provided a new context for further characterization of *RP11-536K7.5* lncRNA. We predicted the positive association between cytokine-kinase signaling pathway and *RP11-536K7.5* lncRNA, but the mechanisms involved are still poorly understood, therefore, further studies are needed to better understand *RP11-536K7.5*/ cytokine-kinase signaling transduction.

4.10.3 Molecular and functional association of relapse-specific lncRNAs signature

We applied the "*guilt-by-association*" approach also on the relapse-specific lncRNAs markers within the subtypes for investigating their functions. However, the relapse-specific signature from the Ph-like and the NH-HeH subtypes did not show any significant enrichment of pathways. A potential reason can be that the DUX4 subtype is particularly perturbed in both relapse-specific and subtype-specific classification, and therefore the number of dysregulated lncRNAs are high compared to the other two subtypes.

In the DUX4 subtype, a notable observation was a strong correlation between relapse-specific lncRNAs with genes involved in the activation of metabolic pathways and signaling pathways. We identified 112 relapse-specific lncRNAs co-expressed with 29 PC genes activated in metabolic pathways. Out of this 112 lncRNA, eight lncRNAs were previously reported as biomarker lncRNAs in the context of various cancers. For example, we identified oncogenic lncRNA *LUCAT1* reported to be associated with poor prognosis in lung cancer. In addition to that, eight relapse-specific lncRNAs associated with metabolic

pathways in the DUX4 subtype was previously reported due to their ability to dysregulate metabolic pathways in multiple tumor contexts. Taken together, the global co-expression analysis and gene-expression profiling suggest important and previously unappreciated roles of lncRNAs in the BCP-ALL subtypes.

4.11 Differentially methylated lncRNAs in BCP-ALL subtypes

We are demonstrating 1118 epigenetically modified novel lncRNAs and previously reported disease-associated lncRNAs within each subtype. Our work additionally underscores the importance of epigenetic alterations in modulating lncRNAs transcriptional activities. Although previous studies have demonstrated cross-talk between DNA methylation and transcriptional activities of lncRNAs, their role in the aetiology of BCP-ALL subtypes has not been investigated. DNA methylation analyses of lncRNAs revealed that DNA methylation might underlie the differential expression of BCP-ALL subtype-specific lncRNAs.

4.11.1 Epigenetically altered lncRNAs within DUX4 subtype

In the DUX4 subtype, we identified lincRNA *PVT1*, *LINC00312* and *TCL6* as differentially hypo-methylated in the promoter region. Interestedly, lincRNA *PVT1* is differentially hypo-methylated, with an up-regulated expression pattern in DUX4 samples. The lincRNA *PVT1* is a well-defined lncRNA for its oncogenic properties (Colombo, Farina, Macino, & Paci, 2015) with multiple roles in cell growth, and differentiation in chronic and T-cell leukemia and many other solid tumors. However, there are no studies in the context of BCP-ALL subtypes.

Other interesting examples are lncRNA *LINC00312*, and *TCL6* were extensively studied on expression levels, but they are not studied at the epigenetic level. We are reporting its expression and DNA methylation profile in BCP-ALL subtypes. Both *TCL6* and *LINC00312* are lowly expressed in DUX4 samples. Intriguingly, *TCL6* is differentially up-regulated in ID compared to REL condition in the DUX4 subtype. The lncRNA *TCL6* is on-lncRNA reported in CLL due to its leukemogenic properties. However, it is not much documented in BCP-ALL. The *TCL6*, on-lncRNA is hyper-methylated and negatively correlating its expression pattern in the DUX4 subtypes.

In addition to leukemia related lncRNAs, we also identified certain lncRNAs with prognostic value in other cancers. For instance, lncRNA *LINC00472* is a tumor suppressor lncRNA in breast cancer. We observed hypomethylation of *LINC00472* at DNA-methylation level and a higher expression at the

transcriptome level. Besides known ones, we have identified 15 novel lncRNAs epigenetically up and down-regulating its expression profile in the DUX4 subtype.

4.11.2 Epigenetically altered lncRNAs within Ph-like subtype

In Ph-like subtype, lncRNAs such as *SOX2-OT*, *MIR7-3HG* and *PVT1* are DM methylated at the promoter region. In Ph-like subtype, we identified *PVT1* as hyper-methylated with a corresponding lower expression level in Ph-like samples. The lncRNA *SOX2-OT* is promoter hyper-methylated in the Ph-like subtype. Shahryari A et al. hints the genomic association of *SOX2* and *SOX2-O*T resembles that of *ANRIL* and *CDKN2B*. Similarly, the lncRNA *ANRIL* resides in the intronic region of the protein-coding gene *CDKN2B*, in the antisense/opposite strand (Shahryari, Jazi, Samaei, & Mowla, 2015). However, there are no direct reports of DNA methylation activity of *SOX2-OT*.

In addition to that, we have observed seven novel lncRNAs potentially epigenetically regulating their gene expression, out of these four were differentially hyper-methylated lncRNAs with down-regulated expression profile and three were differentially hypo-methylated lncRNAs with up-regulated expression profile within Ph-like samples.

4.11.3 Epigenetically altered lncRNAs within NH-HeH subtype

In NH-HeH subtype, we identified lncRNAs including, *LINC00312*, *DANCR* and *IGF2-AS*, as promoter differentially hypo-methylated and with a corresponding low expression level. In addition to the reported lncRNAs, we identified novel lncRNAs within our subtype which significantly facilitates its expression level. This observation was true for all three subtypes. These findings suggest that epigenetic silencing of lncRNA genes may be a mechanism that contributes to the dysregulation of expression of lncRNAs in BCP-ALL subtypes.

Moreover, we identified 53 novel intronic and intergenic DM lncRNAs with super enhancer insulator chromatin markers from our subtypes, which provided a new context for further characterization. The probes for many lncRNA genes were not available in the DNA methylation microarray platform, some lncRNAs that are epigenetically regulated may not be identified in our analysis. Taken together, these results provide a valuable resource that will allow us to investigate epigenetically dysregulated lncRNAs and provided a list of subtype-specific lncRNAs whose expression is epigenetically facilitated.

CONCLUSIONS

My doctoral studies covered a number of aspects pertaining to the broad field of lncRNAs defining subtypes of BCP-ALL. The main conclusions of this studies presented in this thesis are:

- We present a catalog of validated subtype-specific novel lncRNAs through our integrative analysis demonstrating the ability of lncRNAs to classify BCP-ALL subtypes
- Subtype-specific lncRNAs and subtype-specific protein-coding genes are globally predicted to activate or inhibit the same pathways, which are involved in cell proliferation, apoptosis, differentiation in leukemia
- Relapse-specific lncRNAs markers in ALL subtypes and these lncRNAs are associated with both keys signaling and metabolic pathways
- Identified novel and known differentially methylated subtype-specific lncRNAs.
- Epigenetically facilitated dysregulated subtype-specific lncRNAs from these subtypes.

Together, these data extend the spectrum of known involvement of lncRNAs in BCP-ALL subtypes and represents BCP-ALL subtype-specific lncRNAs involved in key signaling and metabolic pathways. Additionally, we highlight key lncRNAs deregulated through epigenetic mechanisms. These findings may open promising avenues for the future studies to investigate key bio-markers and potential therapeutic targets in BCP-ALL subtypes.

REFERENCES

Ali, M. M., Akhade, V. S., Kosalai, S. T., Subhash, S., Statello, L., Meryet-Figuiere, M., Kanduri, C. (2018). PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers. *Nature Communications*, *9*(1). http://doi.org/10.1038/s41467-018-03265-1

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., & Mattick, J. S. (2011). LncRNAdb: A reference database for long noncoding RNAs. *Nucleic Acids Research*, *39*(SUPPL. 1). http://doi.org/10.1093/nar/gkq1138

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq – A Python framework to work with high-throughput sequencing data HTSeq – A Python framework to work with high-throughput sequencing data. *Bioinformatics.*, *31*(2), 0–5. http://doi.org/10.1093/bioinformatics/btu638

Arrial, R. T., Togawa, R. C., & Marcelo, M. M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: Case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics*, *10*. http://doi.org/10.1186/1471-2105-10-239

Atianand, M. K., & Fitzgerald, K. A. (2014). Long non-coding rnas and control of gene expression in the immune system. *Trends in Molecular Medicine*. http://doi.org/10.1016/j.molmed.2014.09.002

Babak, T., Blencowe, B. J., & Hughes, T. R. (2005). A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, *6*. http://doi.org/10.1186/1471-2164-6-104

Bartonicek, N., Maag, J. L. V, & Dinger, M. E. (2016). Long noncoding RNAs in cancer: Mechanisms of action and technological advancements. *Molecular Cancer*, *15*(1). http://doi.org/10.1186/s12943-016-0530-6

Bester, A. C., Lee, J. D., Chavez, A., Lee, Y. R., Nachmani, D., Vora, S., Victor, J., Sauvageau, M., Monteleone, E., Rinn, J. L., Provero, P., Church, G. M., Clohessy, J.G., Pandolfi, P. P. (2018). An Integrated Genome-wide CRISPRa Approach to functionalize lncRNAs in Drug Resistance. *Cell*, *173*(3), 649–664.e20. http://doi.org/10.1016/j.cell.2018.03.052

Bhat, S. A., Ahmad, S. M., Mumtaz, P. T., Malik, A. A., Dar, M. A., Urwat, Shah R.A., Ganai, N. A. (2016). Long non-coding RNAs: Mechanism of action and functional utility. *Non-Coding RNA Research*, *1*(1), 43–50. http://doi.org/10.1016/j.ncrna.2016.11.002

Bhatla, T., Jones, C. L., Meyer, J. A., Vitanza, N. A., Raetz, E. A., & Carroll, W. L. (2014). The biology of relapsed acute lymphoblastic leukemia: opportunities for therapeutic interventions. *Journal of Pediatric Hematology/oncology*, *36*(6), 413–8. http://doi.org/10.1097/MPH.0000000000000179

Blank, U., & Karlsson, S. (2011). The role of Smad signaling in hematopoiesis and translational hematology. *Leukemia*. http://doi.org/10.1038/leu.2011.95

Boer, J. M., Koenders, J. E., Van Der Holt, B., Exalto, C., Sanders, M. A., Cornelissen, J. J., Valk P. J., den Boer M. L., Rijneveld, A. W. (2015). Expression profiling of adult acute lymphoblastic leukemia identifies a BCR-ABL1-like subgroup characterized by high non-response and relapse rates. *Haematologica*. http://doi.org/10.3324/haematol.2014.117424

Borsani, G., Tonlorenzi, R., Simmler, M. C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Ballabio A. (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, *351*(6324), 325–329. http://doi.org/10.1038/351325a0

Brockdorff, N., & Turner, B. M. (2015). Dosage compensation in mammals. *Cold Spring Harbor Perspectives in Biology*, *7*(3). http://doi.org/10.1101/cshperspect.a019406

Casero, D., Sandoval, S., Seet, C. S., Scholes, J., Zhu, Y., Ha, V. L., Luong, A., Parekh, C., Crooks, G. M. (2015). Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nature Immunology*, *16*(12), 1282–1291. http://doi.org/10.1038/ni.3299

Chalei, V., Sansom, S. N., Kong, L., Lee, S., Montiel, J. F., Vance, K. W., & Ponting, C. P. (2014). The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. *eLife*, *3*(November), 1–24. http://doi.org/10.7554/eLife.04530

Cheetham, S. W., Gruhl, F., Mattick, J. S., & Dinger, M. E. (2013). Long noncoding RNAs and the genetics of cancer. *British Journal of Cancer*, *108*(12), 2419–2425. http://doi.org/10.1038/bjc.2013.233

Chilton, L., Hills, R. K., Harrison, C. J., Burnett, A. K., Grimwade, D., & Moorman, A. V. (2014). Hyperdiploidy with 49-65 chromosomes represents a heterogeneous cytogenetic subgroup of acute myeloid leukemia with differential outcome. *Leukemia*. http://doi.org/10.1038/leu.2013.198

Clappier, E., Baruchel, A., Rapion, J., Caye, A., Khemiri, A., Hernandez, L., Kabongo, E., Leblanc, T., Yakouben, K., Plat G., Costa, V., Ferster, A., Rossi, S., Girard, S., Dastugue, N., Bakkus, M., Suciu S., Benoit, Y., Bertrand, Y., Soulier, J., Cave, H. (2012). ERG Intragenic Deletion Characterizes a Distinct Oncogenic Subtype of B-Cell Precursor Acute Lymphoblastic Leukemia with a Favourable Outcome Despite Frequent IKZF1 Deletions. *ASH Annual Meeting Abstracts*, *120*(21), 121-.

Colombo, T., Farina, L., Macino, G., & Paci, P. (2015). PVT1: A rising star among oncogenic long noncoding RNAs. *BioMed Research International*. http://doi.org/10.1155/2015/304208

Davis, A. S., Viera, A. J., & Mead, M. D. (2014). Leukemia: An overview for primary care. *American Family Physician*, *89*(9), 731–738.

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., & Fuks, F. (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics*, *15*(6), 929–941. http://doi.org/10.1093/bib/bbt054

Delás, M. J., & Hannon, G. J. (2017). lncRNAs in development and disease: From functions to

mechanisms. *Open Biology*. http://doi.org/10.1098/rsob.170121

Deocesano-Pereira, C., Amaral, M. S., Parreira, K. S., Ayupe, A. C., Jacysyn, J. F., Amarante-Mendes, G. P., Verjovski-Almeida, S. (2014). Long non-coding RNA INXS is a critical mediator of BCL-XS induced apoptosis. *Nucleic Acids Research*, *42*(13), 8343–8355. http://doi.org/10.1093/nar/gku561

Devaux, Y., Zangrando, J., Schroen, B., Creemers, E. E., Pedrazzini, T., Chang, C. P., Dorn G.W. 2nd., Thum T., Heymans, S. (2015). Long noncoding RNAs in cardiac development and ageing. *Nature Reviews Cardiology*. http://doi.org/10.1038/nrcardio.2015.55

Diakos, C., Xiao, Y., Zheng, S., Kager, L., Dworzak, M., & Wiemels, J. L. (2014). Direct and indirect targets of the E2A-PBX1 leukemia-specific fusion protein. *PLoS ONE*, *9*(2). http://doi.org/10.1371/journal.pone.0087602

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Dong, M., & Blobe, G. C. (2006). Role of transforming growth factor-β in hematologic malignancies. *Blood*. http://doi.org/10.1182/blood-2005-10-4169

Du, Z., Fei, T., Verhaak, R. G. W., Su, Z., Zhang, Y., Brown, M., Chen, Y., Liu, X. S. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nature Structural & Molecular Biology*, *20*(7), 908–913. http://doi.org/10.1038/nsmb.2591

Eades, G., Wolfson, B., Zhang, Y., Li, Q., Yao, Y., & Zhou, Q. (2015). lincRNA-RoR and miR-145 Regulate Invasion in Triple-Negative Breast Cancer via Targeting ARF6. *Molecular Cancer Research*, *13*(2), 330–338. http://doi.org/10.1158/1541-7786.MCR-14-0251

Eis, P. S., Tam, W., Sun, L., Chadburn, A., Li, Z., Gomez, M. F., Lund, E., Dahlberg, J. E. (2005). Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(10), 3627–32. http://doi.org/10.1073/pnas.0500613102

El Fakih, R., Jabbour, E., Ravandi, F., Hassanein, M., Anjum, F., Ahmed, S., & Kantarjian, H. (2018). Current paradigms in the management of Philadelphia chromosome positive acute lymphoblastic leukemia in adults. *American Journal of Hematology*. http://doi.org/10.1002/ajh.24926

Encode, T., & Consortium, P. (2007). Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. *Nature*, *447*, 199–816. http://doi.org/10.1038/nature05874

ENCODE Project Consortium, A. I. E. of D. E. in the H. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., Finch, CE, St

Laurent, G., Kenny, P.J.,Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase. *Nature Medicine*, *14*(7), 723–730. http://doi.org/10.1038/nm1784

Folkman, J. (1974). Tumor Angiogenesis. *Advances in Cancer Research*, *19*(C), 331–358. http://doi.org/10.1016/S0065-230X(08)60058-5

Geng, Y. J., Xie, S. L., Li, Q., Ma, J., & Wang, G. Y. (2011). Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. *Journal of International Medical Research*, *39*(6), 2119–2128. http://doi.org/10.1177/147323001103900608

Ghazavi, F., Moerloose, B. De, Loocke, W. Van, Delabesse, E., Uyttebroeck, A., Nieuwerburgh, F. Van, & Deforce, D. (n.d.). Unique long non-coding RNA expression signature in ETV6 / RUNX1-driven B-cell precursor acute lymphoblastic leukemia, *7*(45).

Gioia, R., Drouin, S., Ouimet, M., Caron, M., St-onge, P., Richer, C., & Sinnett, D. (2017). LncRNAs downregulated in childhood acute lymphoblastic leukemia modulate apoptosis, cell migration, and DNA damage response, *8*(46), 80645–80650.

Groen, J. N., Capraro, D., & Morris, K. V. (2014). The emerging role of pseudogene expressed non-coding RNAs in cellular functions. *International Journal of Biochemistry and Cell Biology*. http://doi.org/10.1016/j.biocel.2014.05.008

Gu, Z., Churchman, M., Roberts, K., Li, Y., Liu, Y., Harvey, R. C., Mullighan, C. G. (2016). Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nature Communications*, *7*. http://doi.org/10.1038/ncomms13331

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. http://doi.org/10.1093/bioinformatics/btw313

Guil, S., & Esteller, M. (2012). Cis-acting noncoding RNAs: Friends and foes. *Nature Structural and Molecular Biology*. http://doi.org/10.1038/nsmb.2428

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L., Wang, Y., Brzoska, P., Kong B., Li, R., West, R.B., van de Vijver, M.J., Sukumar, S., Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, *464*(7291), 1071–1076. http://doi.org/10.1038/nature08975

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., Lander, E. S. (2011). LincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, *477*(7364), 295–300. http://doi.org/10.1038/nature10398

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., Regev, A. (2010). Ab initio reconstruction of

cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, *28*(5), 503–510. http://doi.org/10.1038/nbt.1633

H., S., T., S., H., W., G., S., H., Z., F., S.,Ye, D. (2017). Decreased TCL6 expression is associated with poor prognosis in patients with clear cell renal cell carcinoma. *Oncotarget*, *8*(4), 5789–5799. http://doi.org/http://dx.doi.org/10.18632/oncotarget.11011

Han, B. W., & Chen, Y. Q. (2013). Potential pathological and functional links between long noncoding RNAs and hematopoiesis. *Science Signaling*. http://doi.org/10.1126/scisignal.2004099

Han, P., & Chang, C. P. (2015). Long non-coding RNA and Chromatin Remodeling. *RNA Biol*, 0. http://doi.org/10.1080/15476286.2015.1063770

Han, Z., & Shi, L. (2018). Long non-coding RNA LUCAT1 modulates methotrexate resistance in osteosarcoma via miR-200c/ABCB1 axis. *Biochemical and Biophysical Research Communications*, *495*(1), 947–953. http://doi.org/10.1016/j.bbrc.2017.11.121

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. http://doi.org/10.1007/s00262-010-0968-0

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*. http://doi.org/10.1016/j.cell.2011.02.013

Harvey, R. C., Mullighan, C. G., Wang, X., Dobbin, K. K., Davidson, G. S., Bedrick, E. J., Chen, I.M., Atlas, S.R., Kang, H., Ar K., Wilson, C.S., Wharton, W., Murphy, M., Devidas, M., Carroll, A.J., Borowitz, M.J., Bowman, W.P., Downing, J.R., Relling, M., Yang, J., Bhojwani, D., Carroll, W. L., Camitta, B., Reaman, G. H., Smith, M., Hunger, S. P., Willman, C. L. (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), 4874–4884. http://doi.org/10.1182/blood-2009-08-239681

Herold, T., & Gökbuget, N. (2017). Philadelphia-Like Acute Lymphoblastic Leukemia in Adults. *Current Oncology Reports*, *19*(5). http://doi.org/10.1007/s11912-017-0589-2

Herrera-Solorio, A. M., Armas-López, L., Arrieta, O., Zúñiga, J., Piña-Sánchez, P., & Ávila-Moreno, F. (2017). Histone code and long non-coding RNAs (lncRNAs) aberrations in lung cancer: Implications in the therapy response. *Clinical Epigenetics*. http://doi.org/10.1186/s13148-017-0398-3

Hirano, T., Yoshikawa, R., Harada, H., Harada, Y., Ishida, A., & Yamazaki, T. (2015). Long noncoding RNA, CCDC26, controls myeloid leukemia cell growth through regulation of KIT expression. *Molecular Cancer*, *14*(1). http://doi.org/10.1186/s12943-015-0364-7

Huan, J., Xing, L., Lin, Q., Xui, H., & Qin, X. (2017). Long noncoding RNA CRNDE activates Wnt/β-catenin signaling pathway through acting as a molecular sponge of microRNA-136 in human

breast cancer. *American Journal of Translational Research*, *9*(4), 1977–1989.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. http://doi.org/10.1038/nprot.2008.211

Hughes, J. M., Salvatori, B., Giorgi, F. M., Bozzoni, I., & Fatica, A. (2014). CEBPA-regulated lncRNAs, new players in the study of acute myeloid leukemia. *Journal of Hematology & Oncology*, *7*(1), 69. http://doi.org/10.1186/s13045-014-0069-1

Huo, J. S., & Zambidis, E. T. (2013). Pivots of pluripotency: The roles of non-coding RNA in regulating embryonic and induced pluripotent stem cells. *Biochimica et Biophysica Acta - General Subjects*. http://doi.org/10.1016/j.bbagen.2012.10.014

Iacobucci, I., & Mullighan, C. G. (2017). Genetic basis of acute lymphoblastic leukemia. *Journal of Clinical Oncology*. http://doi.org/10.1200/JCO.2016.70.7836

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y. M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, *47*(3), 199–208. http://doi.org/10.1038/ng.3192

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*. http://doi.org/10.1016/S0022-2836(61)80072-7

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, *41*(1), 200–209. http://doi.org/10.1093/ije/dyr238

Kanduri, C. (2015). Long noncoding RNAs: Lessons from genomic imprinting. *Biochimica et Biophysica Acta*, *1859*(1), 102–111. http://doi.org/10.1016/j.bbagrm.2015.05.006

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., Adachi, J., Fukuda, S., Aizawa, K., Izawa, M., Nishi, K., Kiyosawa, H., Kondo, S., Yamanaka, I., Saito, T., Okazaki, Y., Gojobori, T., Bono, H., Kasukawa, T., Saito, R., Kadota, K., Matsuda, H., Ashburner, M., Batalov S, Casavant, T.., Fleischmann, W., Gaasterland, T., Gissi, C., King, B., Kochiwa, H., Kuehl, P., Lewis, S.., Matsuo,Y., Nikaido, I., Pesole, G., Quackenbush, J., Schriml, L. M., Staubli, F., Suzuki, R., Tomita, M., Wagner, L., Washio, T., Sakai, K., Okido, T., Furuno, M., Aono, H., Baldarelli, R., Barsh, G., Blake, J., Boffelli, D., Bojunga, N., Carninci, P., de Bonaldo, M. F., Brownstein, M. J., Bult, C., Fletcher, C., Fujita, M., Gariboldi, M., Gustincich, S., Hill D., Hofmann, M., Hume, D. A., Kamiya M., Lee N. H., Lyons P., Marchionni, L., Mashima, J., Mazzarelli, J., Mombaerts, P., Nordone, P., Ring,, B., Ringwald M., Rodriguez, I., Sakamot,o N., Sasaki, H., Sato, K., Schönbach, C., Seya, T., Shibata, Y., Storch, K. F., Suzuki, H., Toyo-oka, K., Wang, K. H., Weitz, C., Whittaker, C., Wilming, L., Wynshaw-

Boris, A.., Yoshida, K, Hasegawa, Y., Kawaji, H., Kohtsuki, S., Hayashizaki, Y. Hayashizaki, Y. (2001). Functional annotation of a full-length mouse cDNA collection. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation meeting 2. *Nature*, *409*(6821), 685–690.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, R. D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, *106*(28), 11667–11672. http://doi.org/10.1073/pnas.0904715106

Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H., & Ohhata, T. (2013). Cell cycle regulation by long non-coding RNAs. *Cellular and Molecular Life Sciences*. http://doi.org/10.1007/s00018-013-1423-0

Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S., Arakawa, T., Carninci, P., Hayashizaki, Y. RIKEN GER Group. GSL Members. (2003). Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Research*. http://doi.org/10.1101/gr.982903

Lanz, R. B., McKenna, N. J., Onate, S. A., Albrecht, U., Wong, J., Tsai, S. Y., Tsai M. J., O'Malley, B. W. (1999). A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell*, *97*(1), 17–27. http://doi.org/10.1016/S0092-8674(00)80711-4

Lee, J.-W., Chen, Z., Geng, H., Xiao, G., Park, E., Parekh, S., Kornblau, S. M., Melnick, A., Abbas, A., Paietta, E., Muschen, M. (2015). CD25 (IL2RA) Orchestrates Negative Feedback Control and Stabilizes Oncogenic Signaling Strength in Acute Lymphoblastic Leukemia. *Blood*, *126*(23), 1434–1434. Retrieved from http://www.bloodjournal.org/content/126/23/1434

Lee, Y., Huang, Y. X., & Zhang, F. H. (2006). Expression-Anchored Pathway Profiles of Individual Samples Predicts Survival, Yang X et al. *Proc Natl Acad Sci U S A PLoS Comput Biol Journal of the Royal Statistical Society Series B*, *102*(57), 85–98.

Leeb, M., Pasini, D., Novatchkova, M., Jaritz, M., Helin, K., & Wutz, A. (2010). Polycomb complexes act redundantly to repress genomic repeats and genes. *Genes and Development*, *24*(3), 265–276. http://doi.org/10.1101/gad.544410

Lei, L., Xia, S., Liu, D., Li, X., Feng, J., Zhu, Y., Guo, L., Chen, F., Cheng, H., Chen, K., Hu, H., Che,n X., Li, F., Zhong, S., Mittal, N., Yang, G., Qian, Z., Han, L., He, C. (2017). Genome-wide characterization of lncRNAs in acute myeloid leukemia. *Briefings in Bioinformatics*, (December 2016), 1–9. http://doi.org/10.1093/bib/bbx007

Li, J., Han, W., Shen, X., Han, S., Ye, H., & Huang, G. (2017). DNA methylation signature of long noncoding RNA genes during human pre-implantation embryonic development. *Oncotarget*, *8*(34), 56829–56838. http://doi.org/10.18632/oncotarget.18072

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for

assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. http://doi.org/10.1093/bioinformatics/btt656

Lilljebjörn, H., & Fioretos, T. (2017). New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood*. http://doi.org/10.1182/blood-2017-05-742643

Liu-Dumlao, T., Kantarjian, H., Thomas, D. A., O'Brien, S., & Ravandi, F. (2012). Philadelphia-positive acute lymphoblastic leukemia: Current treatment options. *Current Oncology Reports*, *14*(5), 387–394. http://doi.org/10.1007/s11912-012-0247-7

Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J., & Tian, W. (2012). Molecular mechanisms and function prediction of long noncoding RNA. *The Scientific World Journal*, *2012*(1), 541786. http://doi.org/10.1100/2012/541786

Maksimovic, J., Gordon, L., & Oshlack, A. (2012). SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology*, *13*(6). http://doi.org/10.1186/gb-2012-13-6-r44

Mattick, J. S. (2009). The genetic signatures of noncoding RNAs. *PLoS Genetics*. http://doi.org/10.1371/journal.pgen.1000459

Mazar, J., Rosado, A., Shelley, J., Marchica, J., Westmoreland, T. J. (2016). The long non-coding RNA GAS5 differentially regulates cell cycle arrest and apoptosis through activation of BRCA1 and p53 in human neuroblastoma. *Oncotarget*, *5*(0), 6589–6607. http://doi.org/10.18632/oncotarget.14244

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger A.M., Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. http://doi.org/10.1038/nbt.1630

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*. http://doi.org/10.1038/npp.2012.112

Mudge, J. M., & Harrow, J. (2015). Creating reference gene annotation for the mouse C57BL6/J genome assembly. *Mammalian Genome*. http://doi.org/10.1007/s00335-015-9583-x

Mullighan, C. G. (2012). Molecular genetics of B-precursor acute lymphoblastic leukemia. *Journal of Clinical Investigation*. http://doi.org/10.1172/JCI61203

Mullighan, C. G., Collins-Underwood, J. R., Phillips, L. A., Loudin, M. G., Liu, W., Zhang, J., Ma, J., Coustan-Smith, E., Harvey, R. C., Willman, C. L., Mikhail, F. M., Meyer J., Carroll, A. J., Williams, R. T., Cheng J., Heerema N. A., Basso, G., Pession, A., Pui C. H., Raimondi, S. C., Hunger, S. P., Downing, J. R., Carroll, W. L, Rabin, K. R. (2009). Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. *Nature Genetics*, *41*(11), 1243–6. http://doi.org/10.1038/ng.469

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Guo, M., Yue, M., Wang, L., Li, X. (2018).

Lnc2Cancer : a manually curated database of experimentally supported lncRNAs associated with various human cancers, *44*(May), 980–985. http://doi.org/10.1093/nar/gkv1094

Nobili, L., Lionetti, M., Neri, A., Nobili, L., Lionetti, M., & Neri, A. (2016). Long non-coding RNAs in normal and malignant hematopoiesis. *Oncotarget*, *7*(31), 50666–50681. http://doi.org/10.18632/oncotarget.9308

Nordlund, J., Kiialainen, A., Karlberg, O., Berglund, E. C., Göransson-Kultima, H., Sonderkær, M., Sønderkær, M., Nielsen, K. L., Gustafsson, M. G., Behrendtz, M., Forestier, E., Perkkiö, M., Söderhäll, S., Lönnerholm G., Syvänen, A. C. (2012). Digital gene expression profiling of primary acute lymphoblastic leukemia cells. *Leukemia*, *26*(6), 1218–1227. http://doi.org/10.1038/leu.2011.358

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, *143*(1), 46–58. http://doi.org/10.1016/j.cell.2010.09.001

Ott, G., Rosenwald, A., & Campo, E. (2013). Understanding MYC-driven aggressive B-cell lymphomas: pathogenesis and classification. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*. http://doi.org/10.1182/asheducation-2013.1.575

Panzitt, K., Tschernatsch, M. M. O., Guelly, C., Moustafa, T., Stradner, M., Strohmaier, H. M., Buck, C. R., Denk, H., Schroeder, R., Trauner, M., Zatloukal, K. (2007). Characterization of HULC, a Novel Gene With Striking Up-Regulation in Hepatocellular Carcinoma, as Noncoding RNA. *Gastroenterology*, *132*(1), 330–342. http://doi.org/10.1053/j.gastro.2006.08.026

Paulsson, K., Forestier, E., Lilljebjörn, H., Heldrup, J., Behrendtz, M., Young, B. D., & Johansson, B. (2010). Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21719–21724. http://doi.org/10.1073/pnas.1006981107

Qi, P., & Du, X. (2013). The long non-coding RNAs, a new cancer diagnostic and therapeutic gold mine. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *26*(2), 155–65. http://doi.org/10.1038/modpathol.2012.160

Qiu, J.-J., Wang, Y., Liu, Y.-L., Zhang, Y., Ding, J.-X., & Hua, K.-Q. (2016). The long non-coding RNA ANRIL promotes proliferation and cell cycle progression and inhibits apoptosis and senescence in epithelial ovarian cancer. *Oncotarget*, *7*(22). http://doi.org/10.18632/oncotarget.8744

Quan, Z., Zheng, D., & Qing, H. (2017). Regulatory Roles of Long Non-Coding RNAs in the Central Nervous System and Associated Neurodegenerative Diseases. *Frontiers in Cellular Neuroscience*, *11*. http://doi.org/10.3389/fncel.2017.00175

Qureshi, I., Mattick, J., & Mehler, M. (2010). Long non-coding RNAs in nervous system function and

disease. *Brain Research*, 20–35. http://doi.org/10.1016/j.brainres.2010.03.110.Long

Redon, S., Reichenbach, P., & Lingner, J. (2010). The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. *Nucleic Acids Research*, *38*(17), 5797–5806. http://doi.org/10.1093/nar/gkq296

Rinn, J. L., & Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, *81*(1), 145–166. http://doi.org/10.1146/annurev-biochem-051410-092902

Roberts, K. G., Morin, R. D., Zhang, J., Hirst, M., Zhao, Y., Su, X., Chen, S. C., Payne-Turner, D., Churchman, M. L., Harvey, R. C., Chen, X., Kasap, C., Yan, C., Becksfort, J., Finney, R. P., Teachey, D. T., Maude, S. L., Tse, K., Moore, R., Jones, S., Mungall, K., Birol, I., Edmonson, M. N., Hu, Y., Buetow, K. E., Chen, I. M., Carroll, W. L., Wei, L., Ma, J., Kleppe, M., Levine, R. L., Garcia-Manero, G., Larsen. E., Shah N. P., Devidas, M., Reaman, G., Smith, M., Paugh, S. W., Evans, W. E., Grupp, S. A., Jeha, S., Pui, C. H., Gerhard, D. S, Downing, J. R., Willman, C. L., Loh, M., Hunger, S. P., Marra, M. A., Mullighan, C. G. (2012). Genetic Alterations Activating Kinase and Cytokine Receptor Signaling in High-Risk Acute Lymphoblastic Leukemia. *Cancer Cell*, *22*(2), 153–166. http://doi.org/10.1016/j.ccr.2012.06.005

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. http://doi.org/10.1186/gb-2010-11-3-r25

Sadras, T., Heatley, S. L., Kok, C. H., Dang, P., Galbraith, K. M., McClure, B. J., Muskovic, W., Venn, N. C., Moore, S., Osborn, M., Revesz, T., Moore, A. S., Hughes, T. P., Yeung, D., Sutton, R., White, D. L. (2017). Differential expression of MUC4, GPR110 and IL2RA defines two groups of CRLF2-rearranged acute lymphoblastic leukemia patients with distinct secondary lesions. *Cancer Letters*, *408*, 92–101. http://doi.org/10.1016/j.canlet.2017.08.034

Safavi, S., & Paulsson, K. (2017). Near-haploid and low-hypodiploid acute lymphoblastic leukemia: Two distinct subtypes with consistently poor prognosis. *Blood*. http://doi.org/10.1182/blood-2016-10-743765

Sattari, A., Siddiqui, H., Moshiri, F., Ngankeu, A., Nakamura, T., Kipps, T. J., & Croce, C. M. (2016). Upregulation of long noncoding RNA MIAT in aggressive form of chronic lymphocytic leukemias. *Oncotarget*, *7*(34), 54174–54182. http://doi.org/10.18632/oncotarget.11099

Scheicher, R., Hoelbl-Kovacic, A., Bellutti, F., Tigan, A.-S., Prchal-Murphy, M., Heller, G., Schneckenleithner, C., Salazar-Roa, M., Zöchbauer-Müller, S., Zuber, J., Malumbres, M., Kollmann, K., Sexl, V. (2015). CDK6 as a key regulator of hematopoietic and leukemic stem cell activation. *Blood* (Vol. 125). http://doi.org/10.1182/blood-2014-06-584417

Schmitt, A. M., & Chang, H. Y. (2016). Long Noncoding RNAs in Cancer Pathways. *Cancer Cell*. http://doi.org/10.1016/j.ccell.2016.03.010

Serviss, J. T., Johnsson, P., & Grandér, D. (2014). An emerging role for long non-coding RNAs in cancer metastasis. *Frontiers in Genetics*. http://doi.org/10.3389/fgene.2014.00234

Shahryari, A., Jazi, M. S., Samaei, N. M., & Mowla, S. J. (2015). Long non-coding RNA SOX2OT: Expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis. *Frontiers in Genetics*. http://doi.org/10.3389/fgene.2015.00196

Smaldone, M. C., & Davies, B. J. (2010). BC-819, a plasmid comprising the H19 gene regulatory sequences and diphtheria toxin A, for the potential targeted therapy of cancers. *Current Opinion in Molecular Therapeutics*, *12*(5), 607–16.

Soudyab, M., Iranpour, M., & Ghafouri-Fard, S. (2016). The role of long non-coding RNAs in breast cancer. *Archives of Iranian Medicine*. http://doi.org/0161907/AIM.0011

Studd, J. B., Vijayakrishnan, J., Yang, M., Migliorini, G., Paulsson, K., & Houlston, R. S. (2017). Genetic and regulatory mechanism of susceptibility to high-hyperdiploid acute lymphoblastic leukaemia at 10p21.2. *Nature Communications*, *8*. http://doi.org/10.1038/ncomms14616

Subhash, S., & Kanduri, C. (2016). GeneSCF: A real-time based functional enrichment tool with support for multiple organisms. *BMC Bioinformatics*, *17*(1). http://doi.org/10.1186/s12859-016-1250-z

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. 102(43):15545-50 https://doi.org/10.1073/pnas.0506580102

Tang, J. Y., Lee, J. C., Chang, Y. T., Hou, M. F., Huang, H. W., Liaw, C. C., & Chang, H. W. (2013). Long noncoding RNAs-related diseases, cancers, and drugs. *The Scientific World Journal*. http://doi.org/10.1155/2013/943539

Tran, T. H., & Loh, M. L. (2016). Ph-like acute lymphoblastic leukemia. *ASH Education Program Book*, *2016*(1), 561–566. http://doi.org/10.1182/asheducation-2016.1.561

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. http://doi.org/10.1093/bioinformatics/btp120

Tseng, Y.-Y., & Bagchi, A. (2015). The PVT1-MYC duet in cancer. *Molecular & Cellular Oncology*, *2*(2), e974467. http://doi.org/10.4161/23723556.2014.974467

Van Der Linden, M. H., Willekes, M., Roon, E., Seslija, L., Schneider, P., Pieters, R., & Stam, R. W. (2014). MLL fusion-driven activation of CDK6 potentiates proliferation in MLL-rearranged infant ALL. *Cell Cycle*, *13*(5), 834–844. http://doi.org/10.4161/cc.27757

Van Gils, M. P. M. Q., Cornel, E. B., Hessels, D., Peelen, W. P., Witjes, J. A., Mulders, P. F. A., , Rittenhouse H. G., Schalken J. A., Schalken, J. A. (2007). Molecular PCA3 diagnostics on prostatic fluid. *Prostate*, *67*(8), 881–887. http://doi.org/10.1002/pros.20564

Vardiman, J. W., Thiele, J., Arber, D. A., Brunning, R. D., Borowitz, M. J., Porwit, A., Harris, N. L., Le Beau, M. M., Hellström-Lindberg, E., Tefferi, A., Bloomfield, C. D. (2009). The 2008 revision of

the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: Rationale and important changes. *Blood*. http://doi.org/10.1182/blood-2009-03-209262

Vargova, K., Curik, N., Burda, P., Basova, P., Kulvait, V., Pospisil, V., Savvulidi, F., Kokavec, J., Necas, E., Berkova, A., Obrtlikova, P., Karban, J., Mraz, M., Pospisilova, S., Mayer, J., Trneny, M., Zavadil, J., Stopka, T. (2011). MYB transcriptionally regulates the miR-155 host gene in chronic lymphocytic leukemia. *Blood*, *117*(14), 3816–3825. http://doi.org/10.1182/blood-2010-05-285064

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. http://doi.org/10.1038/nature07730

Wang, C., Wang, L., Ding, Y., Lu, X., Zhang, G., Yang, J., Zheng, H., Wang, H., Jiang, Y., Xu, L. (2017). LncRNA structural characteristics in epigenetic regulation. *International Journal of Molecular Sciences*. http://doi.org/10.3390/ijms18122659

Wang, K. C., & Chang, H. Y. (2011). Molecular Mechanisms of Long Noncoding RNAs. *Molecular Cell*. http://doi.org/10.1016/j.molcel.2011.08.018

Wang, Y., Li, Y., Yang, Z., Liu, K., & Wang, D. (2015). Genome-wide microarray analysis of long non-coding RNAs in Eutopic secretory endometrium with endometriosis. *Cellular Physiology and Biochemistry*, *37*(6), 2231–2245. http://doi.org/10.1159/000438579

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. http://doi.org/10.1038/nrg2484

Wang, Z., Wu, Q., Feng, S., Zhao, Y., & Tao, C. (2017). Identification of four prognostic LncRNAs for survival prediction of patients with hepatocellular carcinoma. *PeerJ*, *5*, e3575. http://doi.org/10.7717/peerj.3575

Winters, A. C., & Bernt, K. M. (2017). MLL-Rearranged Leukemias-An Update on Science and Clinical Approaches. *Frontiers in Pediatrics*, *5*. http://doi.org/10.3389/fped.2017.00004

Xu, S., Kong, D., Chen, Q., Ping, Y., & Pang, D. (2017). Oncogenic long noncoding RNA landscape in breast cancer. *Molecular Cancer*, *16*(1). http://doi.org/10.1186/s12943-017-0696-6

Yan, B., Yao, J., Liu, J. Y., Li, X. M., Wang, X. Q., Li, Y. J., Tao, Z. F., Song, Y. C. , Chen, Q., Jiang, Q. (2015). LncRNA-MIAT regulates microvascular dysfunction by functioning as a competing endogenous RNA. *Circulation Research*, *116*(7), 1143–1156. http://doi.org/10.1161/CIRCRESAHA.116.305510

Yan, X., Hu, Z., Feng, Y., Hu, X., Yuan, J., Zhao, S. D., Shan, W., He, Q., Fan, L., Kandalaft, L. E., Tanyi, J. L., Li C, Yuan, C. X., Zhang, D., Yuan, H., Hua, K., Lu, Y., Katsaros, D., Huang, Q., Montone, K., Fan, Y., Coukos, G., Boyd, J., Sood, A. K., Rebbeck, T., Mills, G. B., Dang, C. V., Zhang, L. (2015). Comprehensive Genomic Characterization of Long Non-coding RNAs across

Human Cancers. *Cancer Cell*, *28*(4), 529–540. http://doi.org/10.1016/j.ccell.2015.09.006

Yang, L., Wang, H., Shen, Q., Feng, L., & Jin, H. (2017). Long non-coding RNAs involved in autophagy regulation. *Cell Death & Disease*. http://doi.org/10.1038/cddis.2017.464

Yang, M. H., Hu, Z. Y., Xu, C., Xie, L. Y., Wang, X. Y., Chen, S. Y., & Li, Z. G. (2015). MALAT1 promotes colorectal cancer cell proliferation/migration/invasion via PRKA kinase anchor protein 9. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1852*(1), 166–174. http://doi.org/10.1016/j.bbadis.2014.11.013

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui,C. H., Evans, W. E., Naeve, C., Wong, L., Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, *1*(2), 133–143. http://doi.org/10.1016/S1535-6108(02)00032-6

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A. P., & Cui, H. (2008). Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature*, *451*(7175), 202–206. http://doi.org/10.1038/nature06468

Zhao, W., Luo, J., & Jiao, S. (2014). Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci Rep*, *4*, 6591. http://doi.org/10.1038/srep06591

Zhao, Y., Sun, H., & Wang, H. (2016). Long noncoding RNAs in DNA methylation: New players stepping into the old game. *Cell and Bioscience*. http://doi.org/10.1186/s13578-016-0109-3

Zhou, Q., Chen, J., Feng, J., & Wang, J. (2016). Long noncoding RNA PVT1 modulates thyroid cancer cell proliferation by recruiting EZH2 and regulating thyroid-stimulating hormone receptor (TSHR). *Tumor Biology*, *37*(3), 3105–3113. http://doi.org/10.1007/s13277-015-4149-9

EIDESSTATTLICHE VERSICHERUNG

„Ich, Alva Rani James, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: LncRNAs signature defining major subtypes of B-cell Acute lymphoblastic leukemia selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst“.

Datum Alva Rani James

Appendix A

Cancer associated lncRNAs from Lnc2Cancer databased lncRNAs identified within our DE BCP-ALL subtypes specific lncRNAs		
lncRNAs	**logFC**	**Subtype**
WT1-AS	4.4467075041	DUX4
IGF2-AS	4.9492924115	DUX4
LINC00617	1.6149985664	DUX4
RP11-528G1.2	1.5962063225	DUX4
RGMB-AS1	5.9567554334	DUX4
PVT1	1.1472967711	DUX4
GAS5	0.7491196605	DUX4
NAMA	2.5975131564	DUX4
RP11-385J1.2	1.2329503925	DUX4
EGOT	-1.0943326581	DUX4
CCDC26	-3.2170024233	DUX4
MIR155HG	-2.6518949391	DUX4
ERICH1-AS1	-2.7489007994	DUX4
SBF2-AS1	-1.3594259494	DUX4
RP11-473M20.11	-0.8438258236	DUX4
RP4-583P15.10	-1.5308828468	DUX4
EGOT	1.274358117	Ph-like
CCDC26	2.6225278418	Ph-like
GAS6-AS1	1.9037862297	Ph-like
SOCS2-AS1	1.0747748025	Ph-like
ZEB1-AS1	1.5226204666	Ph-like
ERICH1-AS1	2.213397887	Ph-like
RP11-473M20.11	1.1719914008	Ph-like
IGF2-AS	-2.0742530532	Ph-like

RGMB-AS1	-3.4663322292	Ph-like
CRNDE	-0.9892797054	Ph-like
MYCNOS	1.7545902012	near-haploid
ERICH1-AS1	3.06193078	near-haploid
FTX	0.793491499	near-haploid
LINC00961	2.3894010913	near-haploid
WT1-AS	-3.1921542827	near-haploid
GAS6-AS1	-1.7290376602	near-haploid
RP11-528G1.2	-1.2262908531	near-haploid
GAS5	-0.90656684	near-haploid
NAMA	-2.5361169641	near-haploid

Appendix B

58 DUX4 specific lncRNAs correlated with the cis PC genes involved in different signaling pathways which are significantly enriched within DUX4 subtype

Cis PC genes	DUX4 specific lncRNAs	Pearson correlation rate	P-value
ERG	RP11-719K4.6	0.5768319431	1.41E-08
ERG	AC135048.13	0.5769253172	0.000000014
ERG	VIPR1-AS1	0.5724957843	1.92E-08
FCGR1A	CTD-2616J11.10	0.7385737284	2.39E-15
FCGR1A	AC069363.1	0.5663844552	2.92E-08
FCGR1A	CTD-2616J11.10	0.7385737284	2.39E-15
FCGR1A	AC069363.1	0.5663844552	2.92E-08
FCGR1A	CTD-2616J11.10	0.7385737284	2.39E-15
FCGR1A	AC069363.1	0.5663844552	2.92E-08

IGF1R	CTD-2528A14.1	0.7109735395	7.23E-14
IGF1R	CTC-523E23.4	0.6726852059	4.47E-12
IGF1R	CTB-25B13.9	0.6484099706	4.51E-11
IGF1R	RP11-1157N2-B.2	0.6173232018	6.54E-10
IGF1R	RP11-314O13.1	0.6143967848	8.28E-10
IGF1R	RP11-654A16.3	0.8986194406	2.34E-30
IGF1R	RP11-35O15.1	0.6743790376	3.77E-12
IGF1R	CTD-2134A5.3	0.5981854606	0.000000003
IGF1R	LINC00637	0.5503288421	8.51E-08
IGF1R	RP11-725G5.2	0.573910225	1.73E-08
IGF1R	RP11-186F10.2	0.6320289475	1.92E-10
IGF1R	RP11-87C12.5	0.704751345	1.48E-13
IGF1R	RP11-264E20.1	0.6218023229	4.53E-10
IGF1R	AF131215.3	0.6213166158	4.71E-10
IGF1R	RP11-166A12.1	0.6229334601	4.12E-10
IGF1R	CTD-2516F10.2	0.5737725158	1.75E-08
IGF1R	CTD-2187J20.1	0.5654915282	0.000000031
IGF1R	SMAD1-AS2	0.583731427	8.62E-09
IGF1R	RP11-713M15.1	0.5534522823	6.94E-08
IGF1R	SPTY2D1-AS1	0.6839701765	1.41E-12
IGF1R	RGMB-AS1	0.6594708095	1.61E-11
IGF1R	RP11-696N14.1	0.6367590983	1.27E-10
IGF1R	RP11-744N12.3	0.5740255383	1.72E-08
IGF1R	RP11-624M8.1	0.7066372789	1.19E-13
IGF1R	RP11-528G1.2	0.614170638	8.44E-10
IGF1R	AC114877.3	0.6621082443	1.26E-11
IGF1R	AC062029.1	0.6091826178	1.26E-09
IGF1R	RP11-523O18.5	0.6478935766	4.73E-11

IGF1R	RP11-465M18.1	0.6761671632	3.15E-12
IGF1R	RP11-63K6.7	0.5629188709	0.000000037
IGF1R	RP11-298A8.2	0.5590243606	0.000000048
IGF1R	LINC00954	0.6288612965	2.51E-10
IGF1R	RP11-563N6.6	0.6119674043	0.000000001
IGF1R	PCAT6	0.6841884724	1.38E-12
IGF1R	RP11-735G4.1	0.6342321907	1.58E-10
IGF1R	RP5-1077I2.3	0.5660757169	2.98E-08
IGF1R	LINC00694	0.568809002	2.47E-08
IGF1R	AC015977.6	0.624252876	3.7E-10
IGF1R	RP1-293L8.2	0.6415439113	8.36E-11
IGF1R	RP3-395M20.2	0.5851375181	7.78E-09
IGF1R	RP11-15H20.6	0.6028228348	2.06E-09
IGF1R	SRRM2-AS1	0.6099631582	1.18E-09
IGF1R	LINC01006	0.5932008351	4.29E-09
IGF1R	IGF2-AS	0.6394800895	1E-10
IGF1R	CTD-2528A14.1	0.7109735395	7.23E-14
IGF1R	CTC-523E23.4	0.6726852059	4.47E-12
IGF1R	CTB-25B13.9	0.6484099706	4.51E-11
IGF1R	RP11-1157N2-B.2	0.6173232018	6.54E-10
IGF1R	RP11-314O13.1	0.6143967848	8.28E-10
IGF1R	RP11-654A16.3	0.8986194406	2.34E-30
IGF1R	RP11-35O15.1	0.6743790376	3.77E-12
IGF1R	CTD-2134A5.3	0.5981854606	0.000000003
IGF1R	LINC00637	0.5503288421	8.51E-08
IGF1R	RP11-725G5.2	0.573910225	1.73E-08
IGF1R	RP11-186F10.2	0.6320289475	1.92E-10
IGF1R	RP11-87C12.5	0.704751345	1.48E-13

IGF1R	RP11-264E20.1	0.6218023229	4.53E-10
IGF1R	AF131215.3	0.6213166158	4.71E-10
IGF1R	RP11-166A12.1	0.6229334601	4.12E-10
IGF1R	CTD-2516F10.2	0.5737725158	1.75E-08
IGF1R	CTD-2187J20.1	0.5654915282	0.000000031
IGF1R	SMAD1-AS2	0.583731427	8.62E-09
IGF1R	RP11-713M15.1	0.5534522823	6.94E-08
IGF1R	SPTY2D1-AS1	0.6839701765	1.41E-12
IGF1R	RGMB-AS1	0.6594708095	1.61E-11
IGF1R	RP11-696N14.1	0.6367590983	1.27E-10
IGF1R	RP11-744N12.3	0.5740255383	1.72E-08
IGF1R	RP11-624M8.1	0.7066372789	1.19E-13
IGF1R	RP11-528G1.2	0.614170638	8.44E-10
IGF1R	AC114877.3	0.6621082443	1.26E-11
IGF1R	AC062029.1	0.6091826178	1.26E-09
IGF1R	RP11-523O18.5	0.6478935766	4.73E-11
IGF1R	RP11-465M18.1	0.6761671632	3.15E-12
IGF1R	RP11-63K6.7	0.5629188709	0.000000037
IGF1R	RP11-298A8.2	0.5590243606	0.000000048
IGF1R	LINC00954	0.6288612965	2.51E-10
IGF1R	RP11-563N6.6	0.6119674043	0.000000001
IGF1R	PCAT6	0.6841884724	1.38E-12
IGF1R	RP11-735G4.1	0.6342321907	1.58E-10
IGF1R	RP5-1077I2.3	0.5660757169	2.98E-08
IGF1R	LINC00694	0.568809002	2.47E-08
IGF1R	AC015977.6	0.624252876	3.7E-10
IGF1R	RP1-293L8.2	0.6415439113	8.36E-11
IGF1R	RP3-395M20.2	0.5851375181	7.78E-09

IGF1R	RP11-15H20.6	0.6028228348	2.06E-09
IGF1R	SRRM2-AS1	0.6099631582	1.18E-09
IGF1R	LINC01006	0.5932008351	4.29E-09
IGF1R	IGF2-AS	0.6394800895	1E-10
IGF1R	CTD-2528A14.1	0.7109735395	7.23E-14
IGF1R	CTC-523E23.4	0.6726852059	4.47E-12
IGF1R	CTB-25B13.9	0.6484099706	4.51E-11
IGF1R	RP11-1157N2-B.2	0.6173232018	6.54E-10
IGF1R	RP11-314O13.1	0.6143967848	8.28E-10
IGF1R	RP11-654A16.3	0.8986194406	2.34E-30
IGF1R	RP11-35O15.1	0.6743790376	3.77E-12
IGF1R	CTD-2134A5.3	0.5981854606	0.000000003
IGF1R	LINC00637	0.5503288421	8.51E-08
IGF1R	RP11-725G5.2	0.573910225	1.73E-08
IGF1R	RP11-186F10.2	0.6320289475	1.92E-10
IGF1R	RP11-87C12.5	0.704751345	1.48E-13
IGF1R	RP11-264E20.1	0.6218023229	4.53E-10
IGF1R	AF131215.3	0.6213166158	4.71E-10
IGF1R	RP11-166A12.1	0.6229334601	4.12E-10
IGF1R	CTD-2516F10.2	0.5737725158	1.75E-08
IGF1R	CTD-2187J20.1	0.5654915282	0.000000031
IGF1R	SMAD1-AS2	0.583731427	8.62E-09
IGF1R	RP11-713M15.1	0.5534522823	6.94E-08
IGF1R	SPTY2D1-AS1	0.6839701765	1.41E-12
IGF1R	RGMB-AS1	0.6594708095	1.61E-11
IGF1R	RP11-696N14.1	0.6367590983	1.27E-10
IGF1R	RP11-744N12.3	0.5740255383	1.72E-08
IGF1R	RP11-624M8.1	0.7066372789	1.19E-13

IGF1R	RP11-528G1.2	0.614170638	8.44E-10
IGF1R	AC114877.3	0.6621082443	1.26E-11
IGF1R	AC062029.1	0.6091826178	1.26E-09
IGF1R	RP11-523O18.5	0.6478935766	4.73E-11
IGF1R	RP11-465M18.1	0.6761671632	3.15E-12
IGF1R	RP11-63K6.7	0.5629188709	0.000000037
IGF1R	RP11-298A8.2	0.5590243606	0.000000048
IGF1R	LINC00954	0.6288612965	2.51E-10
IGF1R	RP11-563N6.6	0.6119674043	0.000000001
IGF1R	PCAT6	0.6841884724	1.38E-12
IGF1R	RP11-735G4.1	0.6342321907	1.58E-10
IGF1R	RP5-1077I2.3	0.5660757169	2.98E-08
IGF1R	LINC00694	0.568809002	2.47E-08
IGF1R	AC015977.6	0.624252876	3.7E-10
IGF1R	RP1-293L8.2	0.6415439113	8.36E-11
IGF1R	RP3-395M20.2	0.5851375181	7.78E-09
IGF1R	RP11-15H20.6	0.6028228348	2.06E-09
IGF1R	SRRM2-AS1	0.6099631582	1.18E-09
IGF1R	LINC01006	0.5932008351	4.29E-09
IGF1R	IGF2-AS	0.6394800895	1E-10
IGF1R	CTD-2528A14.1	0.7109735395	7.23E-14
IGF1R	CTC-523E23.4	0.6726852059	4.47E-12
IGF1R	CTB-25B13.9	0.6484099706	4.51E-11
IGF1R	RP11-1157N2-B.2	0.6173232018	6.54E-10
IGF1R	RP11-314O13.1	0.6143967848	8.28E-10
IGF1R	RP11-654A16.3	0.8986194406	2.34E-30
IGF1R	RP11-35O15.1	0.6743790376	3.77E-12
IGF1R	CTD-2134A5.3	0.5981854606	0.000000003

IGF1R	LINC00637	0.5503288421	8.51E-08
IGF1R	RP11-725G5.2	0.573910225	1.73E-08
IGF1R	RP11-186F10.2	0.6320289475	1.92E-10
IGF1R	RP11-87C12.5	0.704751345	1.48E-13
IGF1R	RP11-264E20.1	0.6218023229	4.53E-10
IGF1R	AF131215.3	0.6213166158	4.71E-10
IGF1R	RP11-166A12.1	0.6229334601	4.12E-10
IGF1R	CTD-2516F10.2	0.5737725158	1.75E-08
IGF1R	CTD-2187J20.1	0.5654915282	0.000000031
IGF1R	SMAD1-AS2	0.583731427	8.62E-09
IGF1R	RP11-713M15.1	0.5534522823	6.94E-08
IGF1R	SPTY2D1-AS1	0.6839701765	1.41E-12
IGF1R	RGMB-AS1	0.6594708095	1.61E-11
IGF1R	RP11-696N14.1	0.6367590983	1.27E-10
IGF1R	RP11-744N12.3	0.5740255383	1.72E-08
IGF1R	RP11-624M8.1	0.7066372789	1.19E-13
IGF1R	RP11-528G1.2	0.614170638	8.44E-10
IGF1R	AC114877.3	0.6621082443	1.26E-11
IGF1R	AC062029.1	0.6091826178	1.26E-09
IGF1R	RP11-523O18.5	0.6478935766	4.73E-11
IGF1R	RP11-465M18.1	0.6761671632	3.15E-12
IGF1R	RP11-63K6.7	0.5629188709	0.000000037
IGF1R	RP11-298A8.2	0.5590243606	0.000000048
IGF1R	LINC00954	0.6288612965	2.51E-10
IGF1R	RP11-563N6.6	0.6119674043	0.000000001
IGF1R	PCAT6	0.6841884724	1.38E-12

IGF1R	RP11-735G4.1	0.6342321907	1.58E-10
IGF1R	RP5-1077I2.3	0.5660757169	2.98E-08
IGF1R	LINC00694	0.568809002	2.47E-08
IGF1R	AC015977.6	0.624252876	3.7E-10
IGF1R	RP1-293L8.2	0.6415439113	8.36E-11
IGF1R	RP3-395M20.2	0.5851375181	7.78E-09
IGF1R	RP11-15H20.6	0.6028228348	2.06E-09
IGF1R	SRRM2-AS1	0.6099631582	1.18E-09
IGF1R	LINC01006	0.5932008351	4.29E-09
IGF1R	IGF2-AS	0.6394800895	1E-10
THBS4	RP11-455F5.5	0.5834534532	8.79E-09
THBS4	AC006369.2	0.581104785	1.04E-08
THBS4	RP11-206L10.3	0.6205508833	5.02E-10
THBS4	AC009495.2	0.5729739692	1.85E-08
THBS4	LINC01001	0.5885699321	6.05E-09
THBS4	RP11-455F5.5	0.5834534532	8.79E-09
THBS4	AC006369.2	0.581104785	1.04E-08
THBS4	RP11-206L10.3	0.6205508833	5.02E-10
THBS4	AC009495.2	0.5729739692	1.85E-08
THBS4	LINC01001	0.5885699321	6.05E-09
IFNG	LA16c-380H5.2	0.6021783111	2.17E-09
IFNG	RP11-229P13.19	0.6873513366	9.92E-13
IFNG	AC006369.2	0.6132095599	9.11E-10
IFNG	RP11-206L10.3	0.5639277141	3.45E-08
IFNG	AC069363.1	0.5755674534	1.54E-08
BHLHE40	BHLHE40-AS1	0.6099960726	1.18E-09

TPO	CTD-2134A5.4	0.5924493	4.54E-09

Appendix C

24 Cis lncRNAs correlated with genes activated in signaling pathways in Ph-like subytpe			
Cis PC genes	**Ph-like specific lncRNAs**	**Pearson correlation rate**	**P-value**
ERG	AC009970.1	0.6526738535	3.05E-011
AGAP1	AC091814.3	0.5751982452	1.58E-008
AGAP1	AGAP1-IT1	0.8032790178	1.09E-019
AKT3	AKT3-IT1	0.6973830726	3.37E-013
AGAP1	CRYM-AS1	0.7090921007	9.00E-014
ERG	FLNB-AS1	0.5648719539	3.24E-008
AGAP1	IGF2-AS	0.5689062092	2.46E-008
IFNG	MIAT	0.5516374441	7.81E-008
MLLT4	MLLT4-AS1	0.7485779553	6.23E-016
AGAP1	RGMB-AS1	0.6187846799	5.80E-010
IFNG	RP11-1094M14.5	0.606262208	1.58E-009
AGAP1	RP11-125B21.2	0.6099589362	1.18E-009
ERG	RP11-228B15.4	0.6340048541	1.62E-010
ERG	RP11-229P13.20	0.5607579369	4.28E-008
AGAP1	RP11-332H18.4	0.5736201641	1.77E-008
AGAP1	RP11-366M4.3	0.7741984559	1.47E-017
AKT3	RP11-382A20.2	0.6097118707	1.20E-009
ERG	RP11-473M20.7	0.5570139637	5.49E-008

AGAP1	RP11-481J2.2	0.5984889177	2.88E-009
AGAP1	RP11-735G4.1	0.6853865908	1.22E-012
AGAP1	RP11-744N12.3	0.5923224789	4.58E-009
AGAP1	RP11-80H8.4	0.6215682327	4.62E-010
ERG	SOCS2-AS1	0.5923422688	4.58E-009
AGAP1	ZEB2-AS1	0.55993078	4.52E-008

Appendix D

61 relapse-specific lncRNAs overlapped with prognostic markers (S-phase lncRNAs) from various cancers

Hypergeometric test is done on the overlap between relapse-specific lncRNAs (864, without duplicated lncRNAs) on 634 S-phase prognostic lncRNAs from Pan-cancer paper, we got a P-value of 0.00026 on the overlap

Geneid	Log Fold change	Subtype	freq	CANCERS
ENSG00000271966	1.8985864055	Ph-like	2	KICH
ENSG00000271966	1.8985864055	Ph-like	2	HNSC
ENSG00000271797	1.8237031774	DUX4	1	LIHC
ENSG00000214145	1.7741480353	NH-HeH	1	LIHC
ENSG00000249635	1.5171090944	DUX4	1	KIRC
ENSG00000225806	1.2744862632	DUX4	2	KIRC
ENSG00000225806	1.2744862632	DUX4	1	LIHC
ENSG00000267745	1.2531138011	DUX4	1	KIRC
ENSG00000229989	1.208303717	DUX4	1	KIRC
ENSG00000258210	1.178750383	DUX4	2	COAD
ENSG00000224950	1.1622599607	DUX4	1	HNSC
ENSG00000272377	1.1518481691	NH-HeH	1	STAD
ENSG00000251141	1.0820345446	NH-HeH	1	KIRC
ENSG00000235477	1.0368402864	DUX4	1	KIRC
ENSG00000272377	1.0007919201	DUX4	1	STAD
ENSG00000225431	0.9792792873	NH-HeH	1	KIRC

ENSG00000259005	0.9754401642	Ph-like	2	HNSC
ENSG00000259005	0.9754401642	Ph-like	1	THCA
ENSG00000245904	0.9697603784	NH-HeH	1	KIRC
ENSG00000245904	0.9697603784	NH-HeH	1	HNSC
ENSG00000253854	0.9683243951	NH-HeH	1	HNSC
ENSG00000258458	0.9658526226	NH-HeH	2	KICH
ENSG00000132832	0.9542306016	NH-HeH	2	KIRC
ENSG00000132832	0.9542306016	NH-HeH	1	KIRC
ENSG00000269275	0.950853756	Ph-like	1	KIRP
ENSG00000273007	0.9293166535	DUX4	1	KICH
ENSG00000229956	0.9230710207	DUX4	3	BLCA
ENSG00000257496	0.9071366649	DUX4	1	KIRC
ENSG00000273321	0.8869802261	Ph-like	1	KIRC
ENSG00000232995	0.8461955111	DUX4	2	KICH
ENSG00000254343	0.8409647741	DUX4	1	KIRC
ENSG00000261971	0.8122099133	DUX4	1	KIRC
ENSG00000261971	0.8122099133	DUX4	1	HNSC
ENSG00000255142	0.8026421154	NH-HeH	1	KIRC
ENSG00000271270	0.7972751156	DUX4	2	KICH
ENSG00000271270	0.7972751156	DUX4	1	LIHC
ENSG00000262903	0.790551093	NH-HeH	2	HNSC
ENSG00000262903	0.790551093	NH-HeH	1	KIRC
ENSG00000203327	0.7777332889	DUX4	2	KIRC
ENSG00000203327	0.7777332889	DUX4	1	KIRC
ENSG00000249684	0.7751868547	DUX4	1	KIRC
ENSG00000203327	0.7723922873	NH-HeH	2	KIRC
ENSG00000203327	0.7723922873	NH-HeH	1	KIRC
ENSG00000224616	0.7700697066	DUX4	2	COAD

ENSG00000231154	0.7402699106	NH-HeH	1	HNSC
ENSG00000225177	0.7323191686	DUX4	2	COAD
ENSG00000225177	0.7323191686	DUX4	2	KIRC
ENSG00000245904	0.724870551	DUX4	1	KIRC
ENSG00000245904	0.724870551	DUX4	1	HNSC
ENSG00000227953	0.7215896414	DUX4	3	COAD
ENSG00000227953	0.7215896414	DUX4	2	COAD
ENSG00000231160	0.719501534	DUX4	1	HNSC
ENSG00000251141	0.7055409481	DUX4	1	KIRC
ENSG00000272142	0.7051632182	DUX4	2	BLCA
ENSG00000272142	0.7051632182	DUX4	2	BLCA
ENSG00000237489	0.7044213697	DUX4	1	HNSC
ENSG00000251432	0.6928448368	DUX4	2	KICH
ENSG00000242798	0.6656524129	DUX4	2	KIRC
ENSG00000242798	0.6656524129	DUX4	1	KIRC
ENSG00000251661	0.6521627605	DUX4	1	BRCA
ENSG00000179406	0.6341635179	DUX4	1	KIRC
ENSG00000179406	0.6341635179	DUX4	1	KIRC
ENSG00000228544	0.6248451995	DUX4	1	KIRC
ENSG00000232931	0.6190056715	DUX4	2	HNSC
ENSG00000232931	0.6190056715	DUX4	2	HNSC
ENSG00000240291	0.6188511007	DUX4	2	HNSC
ENSG00000240291	0.6188511007	DUX4	1	KIRC
ENSG00000231770	0.5830600454	DUX4	1	BRCA
ENSG00000260219	-0.6415049402	DUX4	1	HNSC
ENSG00000237471	-0.696257352	DUX4	1	KIRC
ENSG00000203999	-0.7656032616	DUX4	3	BRCA
ENSG00000203999	-0.7656032616	DUX4	2	COAD

ENSG00000237476	-0.7724465539	DUX4	2	BRCA
ENSG00000237476	-0.7724465539	DUX4	1	LUSC
ENSG00000223486	-0.7978375931	DUX4	1	KIRC
ENSG00000269959	-0.8649871927	DUX4	2	HNSC
ENSG00000269959	-0.8649871927	DUX4	2	KICH
ENSG00000261189	-0.972774636	DUX4	2	BLCA
ENSG00000262152	-1.0599678801	DUX4	2	COAD
ENSG00000262152	-1.0599678801	DUX4	2	COAD
ENSG00000272502	-1.0857976669	NH-HeH	3	BRCA
ENSG00000272502	-1.0857976669	NH-HeH	2	KIRP
ENSG00000258701	-1.1876038012	DUX4	1	BLCA
ENSG00000248323	-1.1894163242	DUX4	3	KICH
ENSG00000248323	-1.1894163242	DUX4	3	KIRC
ENSG00000234432	-1.2407728884	NH-HeH	1	HNSC
ENSG00000259498	-1.3129758934	NH-HeH	1	BLCA
ENSG00000261584	-1.5542111472	DUX4	1	KIRC
ENSG00000261584	-1.5542111472	DUX4	1	HNSC
ENSG00000229619	-1.9591715435	NH-HeH	1	THCA
ENSG00000229619	-1.9591715435	NH-HeH	1	KIRC

Curriculum Vitae

For reasons of data protection law, my curriculum vitae is not included in the electronic version of the dissertation

Publication list

1. **James, A. R,** Schroeder, M. P, Neumann, M, Bastian, L, Eckert, C, Gökbuget, N, … Baldus, C. D. (2019). Long non-coding RNAs defining major subtypes of B cell precursor acute lymphoblastic leukemia. ***Journal of Hematology and Oncology, 12*(1), 1–16. https://doi.org/10.1186/s13045-018-0692-3**
2. Schroeder, M. P, Bastian, L, Eckert, C, Gökbuget, N, **James, A. R**, Tanchez, J. O., … Baldus, C. D. (2019). Integrated analysis of relapsed B-cell precursor Acute Lymphoblastic Leukemia identifies subtype-specific cytokine and metabolic signatures. ***Scientific Reports*, *9*(1), 4188. https://doi.org/10.1038/s41598-019-40786-1**
3. Bastian, L, Schroeder, M. P, Eckert, C, Schlee, C, Ortiz, J, Kämpf, S, **James, A. R** … Brüggemann, M. (n.d.). PAX5 biallelic genomic alterations define a novel subgroup of B-cell precursor acute lymphoblastic leukemia. ***Leukemia*. https://doi.org/10.1038/s41375-019-0430-z**
4. von der Heide EK, Neumann M, Vosberg S, **James AR,** Schroeder MP, Ortiz-Tanchez J, Isaakidis K, Schlee C, Luther M, Jöhrens K, Anagnostopoulos I, Mochmann LH, Nowak D, Hofmann WK, Greif PA, Baldus CD. *Molecular alterations in bone marrow mesenchymal stromal cells derived from acute myeloid leukemia patients.* ***Leukemia. 2016;(April):1-10. doi:10.1038/leu.2016.324.***
5. Mondal T, Subhash S Vaid R, Enroth S, Uday S Reinius B Mitra S Mohammed A, James **AR,** Hoberg E, Moustakas A, Gyllensten U, Jones SJ, Gustafsson CM, Sims AH, Westerlund F, Gorab E, Kanduri C. *MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures.* ***Nat Commun. 2015; 6:7743. doi:10.1038/ncomms8743.***
6. Gawel DR, **Rani James A,** Benson M, et al. *The Allergic Airway Inflammation Repository - A user-friendly, curated resource of mRNA expression levels in studies of allergic airways.* ***Allergy Eur J AllergyClin Immunol. 2014;69(8):1115-1117. doi:10.1111/all.12432***

Oral Presentation and poster presentation

June 2017: A Specific long non-coding RNA Expression Signature Defines the Philadelphia-like B-cell Acute Lymphoblastic Leukemia Subtype. Oral presentation 2nd international Symposium on Frontiers in Molecular science, in Basel, Switzerland.

October 2017: A Specific long non-coding RNA Expression Signature Defines the Philadelphia-like B-

cell Acute Lymphoblastic Leukemia Subtype. Oral presentation 7th German cancer consortium Annual scientific retreat, Berlin, Germany

October 2017: A Specific long non-coding RNA Expression Signature Defines the Philadelphia-like B-cell Acute Lymphoblastic Leukemia Subtype. Poster presentation 6th German cancer consortium Annual scientific retreat, Heidelberg, Germany

Acknowledgements

First and foremost, I would like to express my sincere gratitude to the almighty God for all his blessings and grace showered upon me to accomplish yet another step in my academic life. Heartfelt thanks to my family to whom I express my deepest gratitude for standing with me and guiding me through all seasons of my life. Your teachings and experiences shaped me and without your unconditional love, and prayers, I could never have made it so far. I am extremely thankful for offering your wisdom and firm backing during all my tough decisions. I also gratefully acknowledge all my spiritual brethren for your unconditional love and support.

I would like to sincerely thank Prof. Dr. Claudia Baldus for giving me an opportunity to pursue my PhD thesis under her supervision. Without her constant support and guidance, the research work presented here would not have taken this shape. Your mentoring and thoughtful discussions that encouraged me to become a better scientist. I also greatly acknowledge Dr. Martin Neumann for his guidance, valuable and timely discussions which helped me to accomplish this work. I also sincerely acknowledge my senior Dr. Michael P Schroeder incredible time and patience during writing my thesis and for your great mentoring throughout project helped me with timely completion of the work. Special thanks to all members of AG Baldus group who supported and made the work environment friendly, cooperative and always supported me in various ways. A special thanks to Jutta Ortiz Tanchez for the great time we spend together at the early period of my Phd thesis. A special thanks to Santhilal Subhash, who was my mentor in my early scientific career, who really encouraged me to pursue PhD.

Most importantly, I thank my fiancé Mikhail Trishchakin, for your support, encouragement, patience and most importantly your prayers through thick and thin. I would not have been able to get through any of this without you being by my side (figuratively).

In addition, I would like to thank Deutsches Krebsforschungszentrum (DKFZ) for funding me throughout my PhD. Special thanks to Dr. Eleanor Horn and her team for helping me to get through all the bureaucratic troubles right from my application submission to BSIO. Once again, I would like to thank my parents, to whom I dedicate this thesis.