UNIVERSITA' DEGLI STUDI DI ROMA TOR VERGATA

MACROAREA DI SCIENZE MATEMATICHE, FISICHE E NATURALI

CORSO DI LAUREA MAGISTRALE IN BIOINFORMATICA

TESI DI LAUREA

La caratterizzazione del reguloma della Leucemia Linfoblastica Acuta pediatrica rivela meccanismi di evoluzione tumorale

Relatore:
Manuela Helmer Citterich

Candidato:
Stefano Di Giovenale

Correlatore:
Gerardo Pepe

Relatore esterno:
Giacomo Corleone

Anno Accademico 2020/2021

Ringraziamenti

Vorrei utilizzare questo spazio per ringraziare tutte le persone che mi hanno stimolato al raggiungimento di questo obiettivo.

CI tengo a ringraziare Maurizio per la fiducia e i mezzi messimi a disposizione, e Giacomo che mi ha introdotto al mondo della ricerca e per avermi spinto a fare cose che non credevo avrei mai potuto fare. Con loro ringrazio anche tutto il Laboratorio Fanciulli e l'unità di Bioinformatica che mi hanno accolto in un ambiente amichevole in cui poter lavorare serenamente.

Un ringraziamento speciale a Elena, Domenico, Laura e Daniele per avermi sempre ascoltato e supportato in questo percorso.

Un posto speciale è senza dubbio riservato a Simone ed Andrea per la fraterno legame che ho con loro.

Non potrebbero mai mancare in questo spazio anche Federico, Valerio, Gianluca ed Alessandro perché nessuno è in grado di raggiungere traguardi se al suo fianco non ha amici veri.

Un ringraziamento speciale a Marta per avermi supportato e sopportato in questo percorso, ma soprattutto per avermi fatto apprezzare tutti i traguardi raggiunti, anche quando non ne comprendevo l'importanza.

Indice

1. Summary

B-cell Acute Lymphoblastic Leukemia (B-ALL) is the most common malignancy in children. Indeed, B-ALL accounts for 25% of pediatric cancer and represents the primary diagnosis of leukemia in the pediatric population (Stanulla and Schrappe, 2009). Despite the great efforts in identifying the genetic alteration that prime B-ALL phenotype and the development of treatments targeting those alterations, 15%-20% of patients experience an aggressive cancer relapse (Hunger and Mullighan, 2015). Moreover, patients affected by relapse are likely drug-resistant to therapy. This condition commonly leads to a severe prognosis with a low survival rate which accounts from 15% to 50%. Despite evidence that non-genetic mechanisms such as chromatin accessibility variation, and nucleosome remodelling may drive the development of drug resistance (Marine et al., 2020) in cancer, a deeper understanding of the epigenetic mechanisms underlying B-ALL progression is lacking.

cis-Regulatory Regions (cRRs) are *loci* of non-coding DNA that regulate the transcription of target genes. cRRs are identified through next generation sequencing approaches such as ATAC-seq. Enhancers are the most common cRRs that contain numerous binding sites for sequence-specific transcription factors (TFs). Once TFs are recruited, enhancers accomplish their duty through a three-dimensional spatial rearrangement, named loop, that allows the physical contact with the target gene's promoter. In this way, an enhancer regulates the gene expression of the target gene.

Moreover, enhancers produce a class of non-coding RNA named enhancer RNA (eRNA). The eRNA perform various tasks, for example, they can interact with RNA polymerase II influencing its stability on the promoter. Furthermore, Zangh et al. revealed that the alteration in eRNAs transcription is cancer-specific (Zhang et al., 2019).

Here I present a study that aims to describe the landscape of DNA accessibility in a large longitudinal cohort of paediatric B-ALL and to classify at high resolution the *cis*-regulatory DNA regions sustaining B-ALL relapse. We profiled 32 longitudinal samples from B-ALL to create the most extensive map of the accessibility landscape of 32 Pediatric B-cell Acute Lymphoblastic Leukemia (B-ALL) to date. The integration of Next Generation Sequencing (NGS) techniques and computational pipelines allowed the identification of novel non-genetic targets of B-ALL progression.

The large cohort of 32 patients' samples were provided by Ospedale Pediatrico Bambino Gesù and include 4 B-ALL stages respectively: 8 healthy, 11 onset, 7 remissions, and 8 relapses. The ATAC-seq sample preparation and analyses were accomplished in Fanciulli's lab at National Cancer Institute IFO-IRE in Rome. Samples were profiled with ATAC-seq, a technique used to investigate chromatin accessibility by using hyperactive Tn5 transposase. Changes in genomic accessibility are controlled by the combined work of TFs, RNA polymerases, chromatin remodelers. During the development, chromatin remodelling establishes and maintains cell and tissue type, development transition and response to *stimuli*. For these reasons, alterations in chromatin accessibility have a great interest in cancer progression.

The computational workflow used in this study allows the identification and classification of ATAC-seq profiles. Profiles visually identified as peaks, which represent genomic regions characterized by a high accumulation of reads. Each peak depicts a nucleosome accessible regions. Peak size and shape vary according to the clonality of the peak within the sample and the penetrance within the population. This work aims to classify each peak using 2 indexes: Sharing Index (SI) and Ranking Index (RI), as described by Patten and Corleone et al 2018. SI is a metric able to summarize the penetrance of each peak evaluating the peak's sharing between patients at the same timepoint. RI is a percentile score that describes sample peaks'

clonality. Our analysis revealed that highly shared peaks are generally clonal. Conversely, less shared peaks are subclonal. Thus, SI is a proxy of clonality . This method allowed to stratify the accessibility landscape of B-ALL samples and to prioritize the investigation towards 11k peaks that potentially drive B-ALL progression. Further data stratification based on unsupervised clustering methods allowed the detection of 6k peaks that arise in onset, shrink after treatment and re-expand in relapse.

Furthermore, the integration of data stored in the public repository The Cancer eRNA Atlas (TCeA) (Chen and Liang, 2020) showed that 117 selected enhancer are actively transcribed. Finally, we characterized each enhancer through a multi-layer analysis which includes data such as ENCODE (Davis et al., 2018), HeRA (Zhang et al., 2021), TARGET, and ChIP-ATLAS (Oki et al., 2018) to ultimately assigns function to the selected enhancers. Among the selected 117 enhancers, three actively regulate the expression of DCTD, BCL2, and MYB genes. The RNA interference experiment targeted the eRNA transcribed from the DCTD's enhancer shown a reduction in cell proliferation around 30%.

Taken together, our analyses revealed that regulatory regions support B-ALL progression. Moreover, we highlight the alteration in chromatin accessibility during cancer progression. Thus, we detected the DCTD's enhancer as a putative therapeutic target specific in onset and relapse. Those results shed the light on the involvement of the accessibility alteration in the B-ALL progression. The identification of enhancers B-ALL-specific could be used to identify novel prognostic and therapeutic markers.

1.Introduzione

1.1 Cancer Hallmarks

Il genoma umano contiene tutte le informazioni che consentono ad una cellula totipotente di potersi differenziare in ciascuno dei molteplici tipi cellulari necessari allo sviluppo di un organismo. La normale maturazione di una cellula è però costantemente minata da alterazioni nella sequenza dell'acido desossiribonucleico (*DeoxyriboNucleic* Acid, DNA) denominate mutazioni. Tali variazioni genetiche possono essere dovute ad eventi spontanei nel corso della replicazione del DNA all'interno del nucleo, oppure possono essere causate da stimoli chimici sia endogeni che esogeni. Le cellule hanno però maturato dei sistemi di riparo che, bilanciando le alterazioni genetiche che si manifestano, ne permettono il normale sviluppo (Sniegowski, 1997). Quando viene meno l'equilibrio tra mutazioni e riparo la cellula può andare incontro ad un destino diverso da quello per cui era programmata. Proprio nei cambiamenti dinamici del genoma è stata identificata una delle cause principali per l'insorgenza dei tumori.

Il termine tumorigenesi si riferisce al complesso processo *multistep* che comporta l'acquisizione di un vantaggio proliferativo da parte di cellule pre-cancerose rispetto alle cellule normali localizzate nel medesimo tessuto. In ciascuno *step* la cellula acquisisce uno dei dieci possibili *hallmak* tumorali (Fig. 1) che, nel loro insieme, sono fondamentali sia per l'insorgenza che per la progressione tumorale.

I primi sei *hallmarks* identificati sono stati: autosufficienza nei segnali di crescita, insensibilità a segnali inibitori della crescita, capacità replicativa priva di limiti, elevata angiogenesi, invasione dei tessuti e metastasi (Hanahan and Weinberg, 2000). Negli ultimi dieci anni, grazie all'aumento della disponibilità di dati derivanti da sequenziamenti di ultima generazione (*Next Generation Sequencing*, NGS*)*, sono stati identificati l'instabilità genomica, la stimolazione infiammatoria, la deregolazione energetica e la

capacità di evadere il sistema immunitario come ulteriori *hallmarks*. I primi due sono definiti *enabling characteristics*, i restanti sono invece classificati come *emerging hallmarks* (Hanahan and Weinberg, 2011).

L'instabilità genomica e la stimolazione infiammatoria fanno parte delle *enabling characteristics* dal momento che favoriscono la cellula all'acquisizione di altri *hallmarks* e la conseguente trasformazione cancerosa. Un esempio di instabilità genomica deriva dalle mutazioni in geni che codificano proteine per la detezione ed il riparo di danni al DNA. Infatti se questi sistemi di riparazione vengono meno può verificarsi un incremento del *mutation rate*, aumentando quindi la probabilità di incorrere in mutazioni in geni codificanti per oncosoppressori o oncogeni (Cheng et al., 2008).
La deregolazione energetica e la capacità di evadere il sistema immunitario fanno invece parte degli *emerging hallmarks* poiché donano un vantaggio alla sopravvivenza cellulare (Hanahan and Weinberg, 2011).
Sebbene i tumori siano comunemente considerati come malattie genetiche lo stato cromatinico e le modifiche epigenetiche sembrano svolgere un ruolo fondamentale nell'insorgenza, nello sviluppo e infine nella progressione tumorale (Flavahan et al., 2017). Diversi esempi possono essere annoverati tra le alterazioni che comportano deregolazioni del panorama epigenetico cellulare. Per esempio le alterazioni delle metilazioni e demetilazioni a carico delle regioni ricche in C e G a monte dei promotori (isole CpG) possono indurre silenziamento o attivazione di oncogeni e oncosoppressori (Cheng et al., 2008). Un ulteriore esempio proviene da alterazioni delle modifiche post-traduzionali subite dalle code istoniche (Portela and Esteller, 2010).

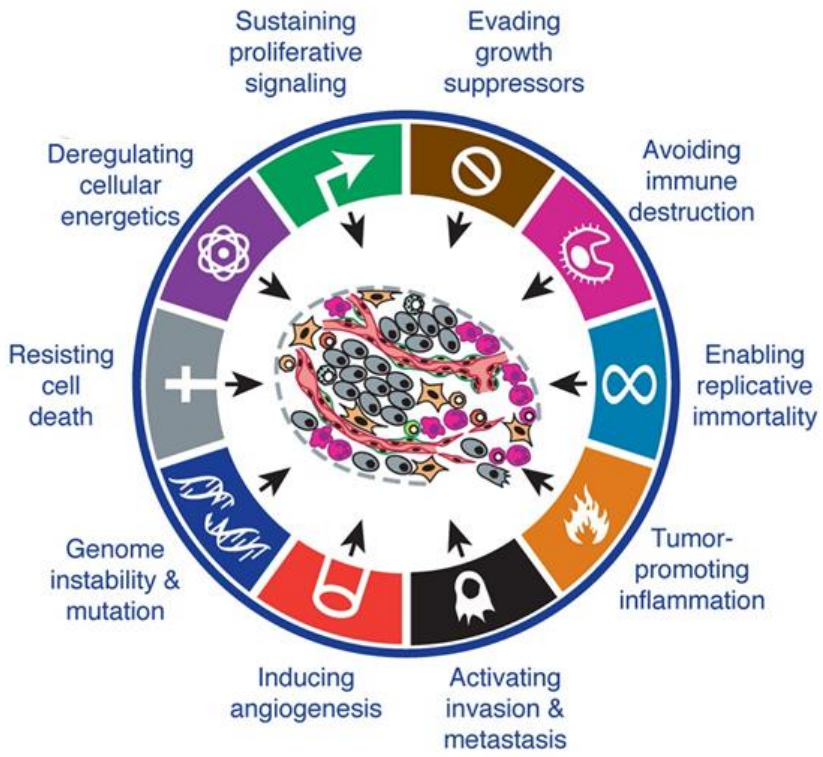


Fig.1: Rappresentazione grafica dei Cancer Hallmarks. Adattata da (Hanahan and Weinberg, 2011)

1.2 Leucemia Linfoblastica Acuta pediatrica di tipo B

La classificazione dei tumori è compiuta tenendo in considerazione diversi parametri tra cui: il sito di insorgenza, il tipo cellulare coinvolto, lo stadio della malattia, la diagnosi, l'aggressività e la presenza o meno di metastasi. Più in generale si possono dividere i tumori solidi da quelli liquidi. I primi sono definiti come una massa di tessuto non contenente porzioni liquide che va incontro a una divisione sregolata; i secondi racchiudono tutti i tumori che affliggono le cellule del sangue. In particolare, i tumori che colpiscono la porzione corpuscolata del sangue prendono il nome di leucemie o linfomi, in base al tipo cellulare in cui avviene la trasformazione maligna.

Nel caso della Leucemia Linfoblastica Acuta (*Acute Lymphoblastic Leukemia*, ALL) il midollo osseo produce un numero elevato di globuli bianchi immaturi denominati linfoblasti. Normalmente i linfoblasti

completerebbero il loro normale sviluppo differenziandosi in linfociti maturi, ma nel caso di questa patologia rimangono indifferenziati, continuano a replicarsi nel midollo osseo, interferendo così con la produzione delle altre cellule della linea ematopoietica.

1.2.1 Epidemiologia

L'ALL è caratterizzata dalla trasformazione maligna, con conseguente proliferazione incontrollata, delle cellule linfoidi nei primi stati del differenziamento nel sangue, midollo osseo e siti extra-midollari. L'ALL è la forma tumorale più comune in età pediatrica, da sola è infatti responsabile del 25% dei tumori pediatrici e circa dell'80% delle leucemie giovanili (Stanulla and Schrappe, 2009). Sebbene esistano due tipologie di ALL distinte in base al tipo cellulare, linfociti T o linfociti B, che incorre nella trasformazione maligna l'ALL di tipo B (B-ALL) è la forma che si manifesta maggiormente. Negli Stati Uniti la maggior parte dei casi di B-ALL si manifesta tra il primo e il quarto anno di vita con un'incidenza di 8 casi ogni 100.000, con un'incidenza maggiore negli individui di sesso maschile rispetto a quelli di sesso femminile (Howlader et al., 2021).

1.2.2 Midollo osseo

Il tessuto maggiormente colpito dall'alterata proliferazione dei linfoblasti è il midollo osseo. Il midollo osseo (*bone marrow, BM*) è costituito dal tessuto spugnoso all'interno delle cavità ossee in cui risiedono le cellule precursori pluripotenti della linea ematopoietica. È, infatti, la fonte principale delle cellule staminali ematopoietiche che hanno la funzione di rinnovamento della porzione corpuscolare del sangue (Cabrita et al., 2003).
Negli adulti il BM è localizzato nella porzione medio-diafisale delle ossa periferiche ed è composto da due porzioni differenti: il midollo osseo rosso o ematopoietico, ed il midollo osseo giallo, il cui colore è determinato dalla presenza di un gran numero di cellule adipose. Nelle ossa dello scheletro

assile (cranio, colonna vertebrale, coste e sterno) il tessuto adiposo coesiste con quello ematopoietico in proporzioni variabili. La variazione dei rapporti tra i due tessuti è guidata dalle richieste dell'organismo. Per esempio se è presente un'elevata richiesta di cellule ematopoietiche si assiste ad una diminuzione del midollo giallo in favore di quello rosso in modo da soddisfare tale richiesta (Gimble et al., 1996).

Il midollo osseo rosso è composto dallo stroma, dai cordoni ematopoietici e dai capillari sinusoidi. La componente fondamentale dello stroma sono le cellule reticolari che hanno il compito di sostenere lo sviluppo cellulare tramite il rilascio di fattori di crescita indispensabili al differenziamento. Una volta maturate le cellule del sangue migrano dai vasi che irrorano il BM, lasciando spazio per la maturazione di ulteriori cellule staminali ematopoietiche (Tavassoli and Yoffey, 1984).

Il BM, però, non è responsabile solo della produzione di nuove cellule ematiche, ma è anche il luogo in cui le cellule staminali mesenchimali si differenziano in osteoblasti, condrociti, miociti, adipociti e cellule neuronali (Braccini et al., 2005).

1.2.3 B-Linfopoiesi

L'attuale modello di sviluppo ematopoietico ipotizza che tutte le cellule mature del sangue derivano da una divisione gerarchica delle cellule staminali ematopoietiche (*hematopoietic stem cells*, HSC) (Shizuru et al., 2005). Le HSC devono mantenere un rigido equilibrio tra il proprio rinnovamento e il differenziamento nelle linee progenitrici mieloidi e linfoidi (Singer et al., 2006).

Nel corso del differenziamento le cellule HSC, multipotenti, perdono progressivamente la loro plasticità differenziandosi in cellule funzionali, modificando anche il programma di espressione genica della linea prescelta. In particolare, le cellule progenitrici linfoidi per differenziarsi in cellule B mature hanno bisogno dell'espressione di tre fattori trascrizionali (TF) fondamentali come E2A, EBF1 e Pax5, oltre al recettore per

l’interluchina-7. E2A e EBF1 svolgono la loro attività di TF agendo da *pioneer factor* inducendo il rimodellamento cromatinico necessario al differenziamento (Maier et al., 2004). Pax5 invece ha un duplice ruolo: da un lato promuove l’espressione di geni indispensabili per lo sviluppo dei linfociti B maturi, dall’altro invece reprime la trascrizione di geni che non permetterebbero un corretto sviluppo del linfocita (Cobaleda and Busslinger, 2008).

Nel corso del differenziamento i linfociti, oltre ad un rimodellamento cromatinico e ad un cambiamento dell’espressione genica, acquisiscono una serie di mutazioni indispensabili per uno sviluppo corretto. È stato infatti evidenziato come nei linfociti maturi sia attivo un elevato processo mutazionale che agisce in ogni passo del differenziamento che permette alle HSP di completare il processo maturativo. Tale plasticità cellulare è indispensabile per permettere ai linfociti di poter riconoscere i molteplici antigeni che potrebbe incontrare. Sebbene l’acquisizione di mutazioni sia indispensabile ad un corretto funzionamento dei linfociti maturi, è stato evidenziato come, nei linfociti in cui avviene una trasformazione maligna, le alterazioni genetiche siano generate dai medesimi sistemi mutazionali presenti nei linfociti maturi (Machado et al., 2021).

Durante il processo di maturazione si osserva, inoltre, il compattamento della cromatina, la diminuzione della visibilità dei nucleoli e la notevole diminuzione delle dimensioni cellulari quando il progenitore linfoide si avvia al differenziamento. Il primo progenitore identificabile delle cellule linfoidi è il linfoblasto, una cellula di dimensione cospicue, che ha la capacità di dividersi due o tre volte per generare i prolinfociti. Questo tipo cellulare seguirà ulteriori processi di maturazione per poi entrare nel torrente circolatorio tramite i capillari sinusoidi del midollo osseo e, infine, stanziare negli organi linfoidi come linfocita B maturo.

1.2.4 Istologia ALL

Dall'analisi delle biopsie di BM ottenute da pazienti affetti da ALL è evidenziata una predominanza dei linfoblasti caratterizzati da un alto rapporto nucleo citoplasma, cromatina nucleare finemente dispersa e presenza dei nucleoli. Inoltre, si evidenzia una diffusa sostituzione degli elementi normali del midollo osseo con fogli uniformi di linfoblasti ovali aventi nuclei frastagliati (Fig. 2).

La diagnosi è effettuata considerando la percentuale di blasti riscontrati nella biopsia di BM. Una percentuale prossima al 25% è clinicamente sufficiente per poter diagnosticare un caso di ALL.

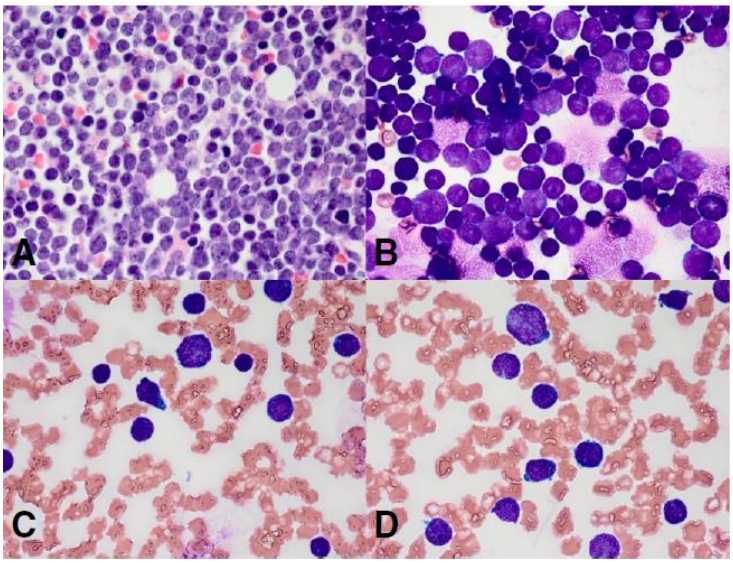


Fig.2: istologia del midollo osseo in pazienti di B-ALL. A. Biopsia di midollo osseo (100x) in cui è evidenziata una sostituzione dei normali elementi del midollo osseo con uno strato di linfoblasti con margini ovali e nuclei convoluti. B. Sono rappresentati i linfoblasti con un rapporti nucleo:citoplasma elevato, cromatina nucleare finemente dispersa e nucleoli prominenti. C-D. Striscio di sangue periferico (100x) mostra linfoblasti con rapporto nucleo:citoplasma elevato e pseudopodi citoplasmatici.

Figura adattata da (Woo et al., 2014)

1.2.5 Alterazioni Genetiche

Analisi genomiche hanno evidenziato che circa il 75% delle ALL giovanili sono affette da aberrazioni genetiche ricorrenti incluse anaeuplodia e riarrangiamenti cromosomali su larga scala (Woo et al., 2014). Il continuo flusso di dati derivanti da sequenziamenti NGS di questi anni ha permesso nel 2016 la ridefinizione delle classi di ALL da parte della *World Health Organization*. La nuova classificazione consiste nell'aggiunta, rispetto alla precedente datata 2008, di due nuove aberrazioni ricorrenti in ALL, passando da 8 a 10 sottoclassi di ALL (Arber et al., 2016).

- B-ALL con t(9;22)(q34.1;q11.2);BCR-ABL1
- B-ALL con t(v;11q23.3);KMT2A rearranged
- B-ALL con t(12;21)(p13.2;q22.1); ETV6-RUNX1
- B-ALL con iperdiploidia
- B-ALL con ipodiploidia
- B-ALL con t(5;14)(q31.1;q32.3) IL3-IGH
- B-ALL con t(1;19)(q23;p13.3);TCF3-PBX1
- B-ALL, *not otherwise specified*, NOS
- *B-ALL, BCR-ABL1–like*
- *B-ALL con iAMP21*

Le sempre più raffinate analisi NGS hanno permesso di identificare, oltre alle macro alterazioni cromosomiche, anche quelle mutazioni submicroscopiche del DNA in oncosoppressori, geni coinvolti nell'ematopoiesi, nella regolazione del ciclo cellulare o coinvolti nella regolazione del ciclo cellulare, come *IKZF1*, *CRLF2*, *PAX5* e *FLT3* (Woo et al., 2014)

Sebbene siano stati individuati numerosi *target* e le terapie farmacologiche siano largamente efficienti con una remissione completa nell'80-90% dei pazienti, la grande sfida è quella di identificare dei possibili *target* per le recidive. Infatti nel 15-20% dei pazienti viene riscontrata una recidiva

tumorale con un tasso di sopravvivenza estremamente basso (Hunger and Mullighan, 2015).
Dato il fallimento delle terapie convenzionali risulta necessaria una più approfondita comprensione del coinvolgimento dell'epigenetica nella progressione tumorale, fino ad oggi ancora largamente inesplorato.

1.3 Epigenomica

Per poter essere contenuto all'interno del nucleo, con diametro di circa 6 micrometri, i 2 metri di DNA devono essere finemente condensati. Il genoma eucariotico, pertanto, è organizzato in una struttura ordinata che prende il nome di cromatina. Le unità fondamentali della cromatina sono i nucleosomi, ciascuno dei quali è composto da circa 147 paia di basi (bp) di DNA che compie circa due giri intorno all'ottamero istonico.
Gli ottameri istonici sono composti da un eterotetramero centrale costituito dagli istoni H3 e H4, fiancheggiati da due eterodimeri formati dagli istoni H2A e H2B. Ogni nucleosoma è separato da circa 10-60 bp di DNA chiamato *linker*. Ciascun istone possiede inoltre una coda ammino-terminale (N-terminale) di 20-35 residui ricchi di amminoacidi basici che protrudono verso l'esterno della superficie del nucleosoma (Peterson and Laniel, 2004).
Con il termine epigenetica si intendono tutte le informazioni nucleari ereditabili durante la divisione cellulare che controllano lo sviluppo, il differenziamento del tessuto e la responsività cellulare. Le modifiche epigenetiche consentono di avere tipi cellulari differenti sebbene la sequenza di DNA contenuta nel nucleo sia sempre la stessa, non si deve però fare l'errore di definire il panorama epigenetico come un qualcosa di immutabile. Infatti l'epigenoma di una cellula può variare per cause dovute all'ambiente, inoltre tali cambiamenti vengono mantenuti durante le divisioni cellulari e vanno a determinare dei cambiamenti nell'espressione genica cellulare (Feinberg, 2018).

Le informazioni epigenetiche agiscono in tre diverse forme, ciascuna delle quali interviene ad un livello di organizzazione cromatinico differente. La prima opera direttamente sulle basi azotate metilando le citosine nella sequenza del DNA. La seconda forma d'informazione epigenetica risiede nelle modifiche post-traduzionali delle code istoniche dei nucleosomi. Infine il terzo livello comprende l'organizzazione ordinata della cromatina facilitando le interazioni tra regioni regolatorie e promotori per guidare l'espressione genetica (Feinberg, 2018).

1.3.1 Metilazioni del DNA

Nei mammiferi, incluso l'uomo, l'informazione contenuta nel DNA è composta da due livelli: la sequenza nucleotidica e il sistema di modifiche chimiche delle basi azotate che guidino la cellula nella corretta interpretazione della sequenza. In questo scenario le metilazioni del DNA rappresentano un sofisticato meccanismo molecolare per l'annotazione dell'informazione genetica.

La metilazione del DNA è una modifica covalente concentrata prevalentemente all'interno di regioni ricche in C e G (isole CpG) a monte dei promotori. Un gran numero di proteine con compiti differenti rientrano nei meccanismi di regolazione delle metilazioni. Un esempio sono le proteine *DNMT3A* e *DNMT3B* responsabili delle metilazioni *de-novo*; mentre le proteine *TET1, TET2* e TET3 sono responsabili della rimozione dei gruppi metilici; infine le proteine *DNMT1 e UHRF1* hanno il ruolo di mantenimento delle metilazioni nel corso della replicazione del materiale genetico (Dor and Cedar, 2018).

Se metilate, le isole CpG a monte di promotori o *enhancer*, ne reprimono l'espressione. Tale repressione è operata da proteine in grado di riconoscere i gruppi metilici e di reclutare fattori deputati alla chiusura della cromatina, rendendo così la regione meno accessibile da parte di fattori trascrizionali e dell'RNA polimerasi (Domcke et al., 2015).

Le metilazioni delle isole CpG sono altamente dinamiche nel corso della vita di un individuo e della cellula. I principali fattori che causano cambiamenti nella metilazione delle isole CpG, e quindi dell'espressione genica, sono l'età (Maegawa et al., 2010), l'ambiente (Waterland and Jirtle, 2003) o l'insorgenza di malattie.
Nel caso dei tumori le alterazioni epigenetiche dovute ai cambiamenti nelle metilazioni del DNA sono responsabili in larga parte dei cambiamenti dell'espressione genica (Esteller, 2011).

1.3.2 Modifiche post-traduzionali delle code istoniche

Tutte le code N-terminali degli istoni sono soggette a modifiche post-traduzionali come: acetilazione, metilazione, fosforilazione, ubiquitinazione SUMOilazioni e ADP-ribosilazioni (Kouzarides, 2007). In base alle modifiche presenti sulle code N-terminali possiamo dividere la cromatina in due stati differenti: eucromatina accessibile e attivamente trascritta; eterocromatina meno accessibile e non trascritta. L'eucromatina è caratterizzata da alti livelli di acetilazione e metilazioni nelle posizioni H3K27, H3K26 e H3K79. Contrariamente l'eterocromatina è caratterizzata dalle metilazione nelle posizioni H3K9, H3K27 e H4K20 (Li et al., 2007).
All'interno delle regioni di eucromatina è possibile identificare una stretta relazione tra modifiche post-traduzionali ed espressione genica. Infatti i geni espressi sono caratterizzati da alti livelli di H3K4me3, H3K27ac, H2BK5ac e H4K20me1 sulle code di istoni fiancheggianti il promotore, mentre i marker H3K79me1 and H4K20me1 sono caratteristici degli istoni contenenti il corpo del gene (Karlić et al., 2010). Anche sugli istoni fiancheggianti gli *enhancer* è possibile identificare dei *pattern* specifici di modifiche post-traduzionali. Gli *enhancer* attivi sono infatti identificati dalla mono-metilazione H3K4Me1 e l'acetilazione H3K27Ac (Heinz et al., 2015).
Come già accennato il *core* istonico del nucleosoma può avere molteplici modifiche post-trascrizionali simultaneamente, questo permette alle modifiche covalenti di dialogare tra di loro, definendo una sorta di codice

istonico. Non è quindi una sola modifica delle code istoniche a determinare il destino di una porzione di DNA, ma la combinazione di molteplici modifiche concorrenti sullo stesso nucleosoma.

Per essere più precisi non è possibile definire l'insieme delle modifiche istoniche come un vero e proprio codice. Questo perché in alcuni casi le modifiche istoniche possono essere associate a effetti diversi. Un esempio è quello delle metilazioni H3K9 associati frequentemente all'inibizione della trascrizione di un gene, ma che in determinati contesi può essere un *marker* di attivazione trascrizionale (Peterson and Laniel, 2004).

Anche le modifiche istoniche sono soggette a plasticità e all'interno del nucleo di una cellula normale viene mantenuto un equilibrio tra le proteine *eraser*, che hanno il compito di rimuovere le modifiche istoniche, e le *writer*, che invece sono responsabili dell'aggiunta. Numerosi lavori hanno evidenziato come nell'insorgenza delle patologie questo equilibrio venga meno, alterando così sia la normale struttura cromatinica che il corretto programma di espressione genica (Zhao and Shilatifard, 2019). Molteplici studi si sono poi concentrati sul cambiamento delle acetilazioni e metilazioni a carico degli istoni nei tumori, identificandone una diminuzione. Le prime spesso associate ad un aumento dell'enzima HDAC responsabile delle deacetilazioni, le seconda causate dall'espressione aberrante di metil-trasferasi e demetilasi con conseguente perdita o aggiunta di metilazioni in maniera da determinare l'attivazione o l'inibizione a seconda della posizione e del contesto (Portela and Esteller, 2010).

1.3.3 *Packaging* della cromatina

Il genoma presenta una fine organizzazione tridimensionale. Tale organizzazione permette di identificare tre strutture principali: compartimenti, domini associati topologicamente (*topological associating domains*, TAD) e isolatori.

I compartimenti consistono nell'organizzazione su larga scala del nucleo in due differenti regioni genomiche, divise in compartimenti attivi e inattivi

(Schoenfelder and Fraser, 2019). Queste sono strutture altamente plastiche in grado di riorganizzarsi nel corso del differenziamento cellulare (Dixon et al., 2015).

I TAD sono invece delle organizzazioni superiori composti da diverse megabasi di DNA che permettono di compartimentalizzare e facilitare le interazioni tra regioni genomiche al proprio interno (Dixon et al., 2012). Un modo di vedere i confini dei TAD è quello di immaginarli come degli isolatori che permettono di facilitare le interazioni *enahncer*-promotore all'interno del medesimo TAD, scoraggiando possibili cross-interazioni tra TAD differenti e regolando in questo modo l'espressione genica (Symmons et al., 2014). Quindi I TAD possono essere identificati come dei microambienti che promuovono l'interazione spaziale di regioni regolatorie con i relativi promotori *target* anche se questi si trovano a distanze considerevoli.

Sebbene il promotore, insieme ai fattori trascrizionali e alla polimerasi, sia fondamentale per l'espressione genica, diverse regioni regolatorie possono modificare questo processo in maniera considerevole. Infatti, la regolazione trascrizionale di un gene può essere compiuta da fattori trascrizionali che si legano a distanze considerevoli dal promotore di quest'ultimo, anche ad 1 megabase. È proprio grazie alla conformazione tridimensionale della cromatina che permette alle regioni regolatorie di trovarsi in prossimità del promotore del gene e poterne così influenzarne la trascrizione (Andersson and Sandelin, 2020).

Gli *enhancer* insieme ai promotori sono le maggiori regioni funzionali agenti in *cis* e in grado di alterare e regolare l'espressione genica. Gli *enhancer* sono regioni genomiche legate da molteplici fattori trascrizionali e possono essere situate in posizioni diverse rispetto al gene *target*, infatti sono stati evidenziati *enhancer* che regolano geni a monte, a valle e possono trovarsi anche all'interno di regioni introniche del gene *target* o di altri geni (Plank and Dean, 2014).

L'attività di un *enhancer* può essere dedotta a partire dalle modifiche istoniche che concorrono sugli istoni fiancheggianti. Per permettere lo

spostamento dei nuclesomi e l'apertura della cromatina, esponendo così un *enhancer* al legame dei fattori trascrizionali, è necessaria la presenza delle modifiche istoniche H3K4me1 e H3K27ac (Murakawa et al., 2016).
Questa serie di modifiche istoniche permettono il rimodellamento dei nucleosomi interessati con l'esposizione del DNA regolatorio, ed il successivo legame di TFs, RNA polimerasi II (Li et al., 2016) con la conseguente produzione di non-coding RNA chiamati *enhancer* RNA (eRNA).
L'integrazione dei dati di ChIP-seq del H3K27ac, ChIP-seq del H3K4me ed RNA *sequencing* (RNA-seq) hanno permesso di individuare regioni genomiche intergeniche di dimensione media di 20kb che evidenziano un elevato segnale trascrizionale. Queste regioni prendono il nome di *super-enhancer* (SE) dal momento che condividono le stesse funzionalità e caratteristiche degli *enhancer* canonici (EC) ma con alcune differenze. Come già accennato i SE, rispetto agli EC, hanno dimensioni maggiori, maggiore attività trascrizionale, maggiore densità di fattori trascrizionali e, infine, una maggiore attivazione e regolazione dell'espressione del gene *target* (Tang et al., 2020).

1.3.4 Regioni regolatorie nei tumori

A causa della loro importanza nella regolazione dell'espressione genica nel corso delle diverse fasi dello sviluppo, l'attività degli *enhancer* è coinvolta nell'insorgenza e nello sviluppo dei tumori. Due evidenze fondamentali vengono condotte a supporto di questa tesi. La prima si basa sugli studi di associazione genomica su larga scala (*genome wide association studies*, GWAS) che hanno dimostrato la presenza di numerose varianti geniche predisponenti l'insorgenza dei tumori e che non mappano all'interno di regioni codificanti, ma in regioni identificate come *enhancer* putativi. Secondo, lo sviluppo di tecniche di sequenziamento *high-throughput* ha permesso l'identificazione di *enhancer* putativi e le loro differenti attivazioni tra cellule tumorali e normali (Sur and Taipale, 2016).

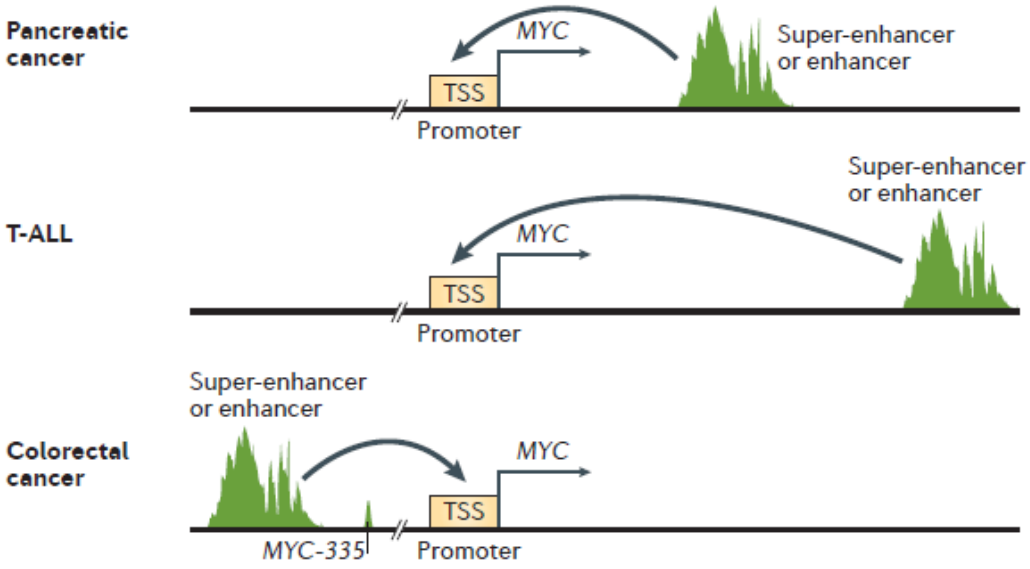


Fig. 3: Raffigurazione di come uno stesso gene, in questo caso MYC, possa essere regolato da enhancer o super-enhancer differenti in tumori differenti. Adattato da (Kaikkonen et al., 2013)

L'attività di un *enhancer* può essere alterata in molteplici modi, ma principalmente è possibile identificare due tipi di alterazioni: *cis* e *trans*. Le alterazioni *cis* possono portare a tre diverse modifiche dell'attività dell' *enhancer*: l'aumento del suo numero di copie può comportarne un aumento dell'attività; riarrangiamenti strutturali possono portare al cambiamento del gene *target*; mutazioni di singolo nucleotide possono alterare siti di legame a fattori trascrizionali e creare degli *enahncer de-novo*.

Le più comuni *trans* attivazioni o repressioni di un *enhancer* sono dovute a mutazioni in fattori trascrizionali e alterazioni nelle proteine coinvolte nelle modifiche istoniche (Sur and Taipale, 2016).

Le alterazioni identificate negli *enhancer* sono generalmente paziente-specifiche. Infatti nelle cellule tumorali diversi geni importanti per la crescita cellulare sono regolati da *enhancer* acquisiti *de-novo* nel corso della tumorigenesi, questi *enhancer* sono specifici per ciascun tipo tumorale, anche se regolano il medesimo gene (Fig. 3) (Hnisz et al., 2013).

1.3.5 *Enhancer* RNA

Solamente circa 1-2% del trascrittoma provvede alla sintesi delle proteine, il restante 98-99% è deputato, invece, alla produzione di tutti gli RNA non codificanti, di cui fanno parte gli RNA *transfer*, RNA ribosomiali, piccoli RNA nucleari, micro RNA e *long-non coding* RNA (Sartorelli and Lauberth, 2020). Studi recenti hanno però identificato una nuova tipologia di RNA prodotti dagli *enhancer* attivi. Questi RNA, della dimensione di circa 150 nucleotidi. prendono il nome di enhancer RNA (eRNA) e sono trascritti in maniera tessuto specifica.

Ancora non è stata compiuta una caratterizzazione completa degli RNA di questo tipo in tessuti e linee cellulari. Infatti, la loro trascrizione può avvenire in maniera bidirezionale, producendo trascritti non poliadenilati, instabili e provi di siti di *splicing*, oppure unidirezionale, sintetizzando RNA stabili e processati sottoposti a *splicing* (Sartorelli and Lauberth, 2020).

Gli eRNA ricoprono un numero elevato di compiti nella regolazione dell'espressione genica, nella definizione dello stato cromatinico e nella stabilizzazione del *loop enhancer*-promotore. Infatti diversi studi dimostrano il coinvolgimento di questi RNA nel mantenimento della cromatina in uno stato aperto in modo da rendere possibile il legame di fattori trascrizionali (Kaikkonen et al., 2013).

Principalmente questi compiti sono svolti tramite l'interazione degli eRNA con specifiche proteine come l'istone acetiltransferasi CBP e la proteina Ying Yang 1 (YY1). In maniera simile gli eRNA sono in grado di interagire con l'RNA polimerasi II aumentando il legame nelle regioni codificanti proteine e, simultaneamente, creando le condizioni favorevoli per l'inizio della trascrizione. Infine, è stato dimostrato come l'interazione degli eRNA con le coesine e la proteina CBP sia in grado di aumentare l'attività degli *enhnacer* mediante la regolazione della formazione del *loop* e la deposizione di gruppi acetile sulle code istoniche (Sartorelli and Lauberth, 2020).

Anche gli eRNA, come i relativi *enhancer*, mostrano una forte componente tumore-specifica. È infatti stata dimostrata tramite un'analisi multi-omica la forte espressione tumore-specifica degli eRNA, suggerendo un possibile utilizzo di queste molecole come possibili *marker* diagnostici (Zhang et al., 2019).

1.4 Cancer Evolution

La tumorigenesi è un processo *multistep* in cui possono essere individuati meccanismi evoluzionistici e, in quanto tale, è sottoposta ad una pressione selettiva effettuata dall'ambiente circostante (Nowell, 1976). Il processo evolutivo è quindi responsabile della comparsa di un popolazione non omogenea di cellule all'interno della massa tumorale. Non è possibile, pertanto, individuare un'unica popolazione cellulare all'interno del tessuto tumorale, ma numerose sottopopolazioni che, nel loro insieme, definiscono l'eterogeneità intra-tumorale (ITH). Le differenze all'interno di queste popolazioni cellulari possono essere dovute a diverse alterazioni come: mutazioni puntiformi, variazione del numero di copie, cambiamenti epigenetici e trascrittomici che alterano l'espressione genica, e la risposta immunitaria anti-tumorale (Black and McGranahan, 2021).

La progressione tumorale ha luogo con l'insorgenza di una prima alterazione cellulare che, assicurando un vantaggio selettivo rispetto alle altre cellule normali del tessuto, produce un'espansione clonale. Secondo questo modello, sebbene siano presenti differenti popolazioni cellulari all'interno di un tumore, per tutte le popolazioni sarà possibile individuare un progenitore comune (Navin et al., 2010). Bisogna tuttavia considerare che nel corso evolutivo si possono generare alterazioni che non sono necessariamente riconducibili all'attribuzione di un vantaggio selettivo, questo tipo di alterazioni prendono il nome di neutrali (Williams et al., 2016). La comprensione del processo di evoluzione cellulare è fondamentale per definire i principi alla base dell'insorgenza della resistenza alle terapie antineoplastiche.

Nonostante le alterazioni genetiche abbiano un ruolo fondamentale nella resistenza ai farmaci, non rappresentano necessariamente l'unico evento. Infatti è possibile riscontrate l'insorgenza di resistenza ai farmaci in un tumore anche senza alterazioni del panorama mutazionale, suggerendo un coinvolgimento di meccanismi non-genetici alla base del fenomeno (Marine et al., 2020).

Diverse strategie di campionamento tumorale sono state sviluppate per comprende e caratterizzare l'ITH di un tumore. Questi approcci possono essere divisi in due metodologie di campionamento differente: *Geographical sampling* e *Longitudinal sampling*. Mentre la prima ha lo scopo di confrontare campionamenti derivanti da aree diverse appartenenti ad aree diverse della stessa massa tumorale, la seconda ha lo scopo di confrontare lo stesso sito tumorale ma in diversi *timepoint* della malattia, principalmente esordio, remissione e recidiva (Yates and Campbell, 2012). Come già ipotizzato da Nowell, un ruolo fondamentale nel definire quali sottopopolazioni hanno la capacità di propagarsi è svolto dall'ambiente in cui le cellule si trovano (Gatenby et al., 2007). Una prova del ruolo svolto dal microambiente tumorale è stato identificata primariamente nell'insorgenza di rami organo-specifici all'interno di alberi filogenetici in studi che correlano le metastasi ai tumori primari (Yachida et al., 2010)

2. Scopo del lavoro

Il lavoro descritto in questa tesi ha come obiettivo l'identificazione di regioni regolatorie altamente coinvolte nella plasticità cellulare in pazienti affetti da B-ALL pediatrica e in grado di guidare l'insorgenza di recidiva tumorale e la resistenza ai trattamenti. A questo proposito, la nostra analisi è stata compiuta a partire da 32 campioni longitudinali di *Assay for Transposase-Accessible Chromatin using sequencing* (ATACseq) ottenuti da pazienti in cura presso l'Ospedale Bambino Gesù di Roma e sequenziati presso l'Istituto Tumori Regina Elena nel laboratorio del Dottor Maurizio Fanciulli.
I 32 campioni analizzati erano così suddivisi: 6 Sani, 11 Esordi, 7 Remissioni e 8 Recidive.
Il primo passo dell'analisi è stato quello di assegnare a ciascuna regione rilevata due *score* denominati *Ranking Index* (RI) e *Sharing Index* (SI), che indicano rispettivamente la clonalità del picco nel campione e la condivisione del picco tra tutti i campioni dello stesso stato tumorale. Una volta identificate le possibili regioni *target*, ci siamo focalizzati sui geni putativamente controllati da queste regioni. Successivamente gli sforzi si sono concentrati sulla descrizione del comportamento di queste regioni a più livelli e in linee cellulari differenti. Queste regioni sono state ulteriormente stratificate individuando quali fossero precedentemente individuate come *enhancer*. Infine, sono stati individuati *enhancer* coinvolti nella progressione tumorale e selettivamente attivi in esordi e recidive.
Questo lavoro ha permesso l'identificazione di una pletora di regioni regolatorie coinvolte nella progressione tumorale, con particolare attenzione alla loro presenza nelle recidive, producendo finora la più grande mappa di alterazioni nell'accessibilità cromatinica di B-ALL.

3.Materiali e Metodi

3.1 Raccolta dei dati

I dati utilizzati in questo progetto derivano da pazienti in cura presso l'Ospedale Bambino Gesù di Roma e sequenziati presso l'Istituto Tumori Regina Elena nel laboratorio del dottor Maurizio Fanciulli. In aggiunta, profili di: RNA-seq phs000464 da TARGET (https://ocg.cancer.gov/programs/target), ChIP-seq da ChIP-ATLAS(Oki et al., 2018), ChIP-seq di H3K27ac ottenuti da ENCODE (Davis et al., 2018).

3.2 ATACseq

L' *Assay for Transposase-Accessible Chromatin using sequencing (*ATACseq*)* (Minnoye et al., 2021) è una tecnica biotecnologica in grado di determinare regioni di cromatina accessibili o la posizione dei nucleosomi. Questo metodo consiste nella preparazione di una libreria di *next-generation sequencing* (NGS) utilizzando la trasposasi Tn5 iperattiva. Le librerie vengono sequenziate tramite tecnologia NGS, nel nostro caso *paired-end* di dimensioni 150 bp, l'intera libreria.

3.2.1 Analisi computazionale di ATACseq

Le *reads paired-end* di lunghezza 150 bp ottenute dal sequenziamento con NextSeq 500 sono state allineate al genoma Hg19 di riferimento tramite l'utilizzo di Bowtie (Langmead and Salzberg, 2012) v 2.3.5.1 con i parametri di *default*. Da questo processo vengono generati i file SAM che poi sono stati convertiti in BAM per poi essere ordinati e indicizzati tramite il *tool* samtools v 1.7. I file BAM così ottenuti sono poi stati deduplicati, ovvero sono state eliminate le *reads* che hanno esattamente la stessa posizione di inizio e fine in quanto considerate artefatti di PCR, tramite *Genome Analysis Toolkit* (GATK) (McKenna et al., 2010) v 4.1.9.0 con i parametri 'MarkDuplicates -REMOVE_DUPLICATES true'. Infine è stata eseguita la

chiamata dei picchi tramite il *tool* macs2 v 2.2.6 con i parametri 'callpeak --format AUTO --nomodel --shift -100 --extsize 200 -B --SPMR --call-summits -q 0.01 -g hs'. Infine sono state eliminate le regioni considerate *blacklisted* dai file (https://www.encodeproject.org/files/ENCFF001TDO) con estensione summits.bed ottenuti da macs2. Le *blacklisted regions sono siti* del DNA che hanno un segnale anomalo negli esperimenti NGS (Amemiya et al., 2019), tramite l'utilizzo del tool bedtools (Quinlan and Hall, 2010) v 2.29.2 con parametri 'intersect -v'.

3.3 Generazione Masterlist

La *masterlist* è un file contenente la lista di tutti i picchi identificati in almeno un campione. Una volta identificati i picchi in ciascun campione tramite MACS2 sono stati concatenati i file con estensione *NarrowPeak*. Utilizzando *bedtools merge* con parametri *standard* sono stati fusi i picchi che mostravano una sovrapposizione spaziale. In questo modo è stato ottenuto un file contenente i picchi unici identificati nei campioni.

3.4 *Ranking Index*

Il *Ranking Index* (RI) è un metodo di normalizzazione quantilica usato per determinare la clonalità dei picchi identificati in ciascun campione. Questo indice permette di rappresentare la clonalità di ciascun picco all'interno del campione sequenziato. Tale considerazione è possibile poiché il segnale di un picco di ATACseq in campione è sovrapponibile alla somma delle intensità dei singoli picchi identificati in un ATAC *single cell* (Buenrostro et al., 2015).

Il procedimento prevede il conteggiato per ciascun picco di ogni campione del numero di *reads* che mappano in quella regione utilizzando i file bam e i narrowPeak ottenuti da MACS2. Il conteggio è ottenuto tramite l'utilizzo di bedtools multicov con parametri standard. Infine è stato utilizzato uno script R per normalizzare il conteggio delle *reads* tramite la formula :

Nscore=((conteggio delle *reads* / dimensione del picco)*10^6))*10^3 / numero totale delle reads (FPKM) (Patten and Corleone et al., 2018).
Per poi assegnare un valore da 1 (Nscore alto) a 100 (Nscore basso) tramite la funzione *ntile* del pacchetto dplyr in R.

3.5 *Sharing Index*

Lo Sharing Index (SI) permette di definire la condivisione di ciascun picco tra i pazienti in ciascuna condizione. Sono state applicate due diverse strategie di assegnazione dello SI. In entrambe le strategie è utilizzato il *file masterlist* precedentemente descritto.

1. Tramite *bedtools intersect* con i parametri -wa -wb -loj -f 0.4 è stato identificato per ciascun file *NarrowPeak* la presenza di un determinato picco. Infine tramite uno script python è stato contato il numero di campioni che presentano il picco. In questo caso lo SI va da 1 a 35.
2. Per prima cosa sono stati divisi i campioni nei rispettivi *timepoint* della malattia: sano (MUD), esordio (ES), remissione (REM), recidiva (REC). Lo SI è stato ottenuto usando *bedtools intersect* con i parametri '-wa -wb -loj -f 0.4' utilizzando il file *masterlist* e i file *NarrowPeak* divisi per *status*. Infine, tramite uno script python è stato contato il numero di campioni che presentano il picco. In *output* sono stati ottenuti quattro *file*, uno per ciascun *timepoint*, contenenti tutti i picchi identificati in almeno un campione analizzato, ma con SI che va da 1 al numero di campioni in ciascun *timepoint*.

3.5 Selezione dei picchi di interesse

Per la selezione dei picchi coinvolti nella progressione della patologia sono state utilizzati tre principi diversi che coinvolgono le alterazioni dell'SI e RI.

3.5.1 Selezione dei picchi esordio specifici

Per identificare i picchi presenti solo negli esordi rispetto ai campioni sani, primariamente sono stati selezionati tutti i picchi con SI mediano compreso tra 8 e 11 per gli esordi e tra 4 e 8 nei sani. L'identificazione delle *threshold* è stata compiuta tramite molteplici prove empiriche utilizzando diverse combinazioni. Come criterio di selezione delle soglie è stato usato il tool web GREAT (McLean et al., 2010) v 3.0.0, che permette di assegnare un significato biologico a gruppi di regioni non codificanti analizzando le annotazioni dei geni più vicini, con l'utilizzo dei parametri di *default*. Le soglie selezionate sono quelle più stringenti, ovvero considerano un numero di pazienti elevato che deve condividere lo stesso picco, che allo stesso tempo permettevano di ottenere annotazioni delle regioni selezionate legate alla ALL o a malattie leucemiche. Considerando la maggior eterogeneità degli esordi rispetto ai campioni sani, è stato deciso di considerare un intervallo più conservativo per gli esordi rispetto ai sani. Successivamente si sono trovati i picchi unici degli esordi tramite l'utilizzo di bedtools con parametri intersect -v. In questo modo sono stati selezionati 5253 picchi.

3.5.2 Selezione dei picchi clonali negli esordi

Per l'identificazione dei picchi che evidenziano un'espansione clonale tra ES e MUD si è valutato l'incremento del RI. A partire dai picchi scartati nel passaggio precedente sono stati identificati i picchi che evidenziavano un incremento di RI tra ES e MUD maggiore del 20% della clonalità. In questo modo sono stati selezionati 2469 picchi.

3.5.3 Selezione dei picchi con aumento di clonalità tra remissione e recidiva

Per la selezione dei picchi che evidenziano espansione clonale tra remissioni e recidive inizialmente sono stati selezionati i picchi comuni alle due condizioni tramite *bedtools intersect* con parametri *standard*. Infine è

stata compiuta una selezione per mantenere solo i picchi che evidenziavano un incremento maggiore di 10 tra la medie degli RI dei REM e la media degli RI delle REC. In questo modo sono stati selezionati 3382 picchi.

3.6 *Heatmap* conta delle *reads*

A partire dagli 11.000 picchi selezionati è stata prodotta un'*heatmap* per determinare la clusterizzazione dei campioni e dei picchi. La clusterizzazione è stata compiuta utilizzando la conta delle *reads* sequenziate per ciascuno dei picchi in ciascun campione. Per determinare il numero di *reads* per ogni campione è stato usato *bedtools multicov* con parametri standard utilizzando i file BAM e la lista degli 11.000 siti selezionati. Successivamente i *file* di *output* sono stati uniti in un unico *dataframe* e normalizzati tramite edgeR. Per la clusterizzazione per righe e colonne è stato usato il metodo "euclidean" per il calcolo delle distanze e il metodo "ward D2" per la formazione del dendogramma.

3.7 Analisi dei motivi

I fattori trascrizionali (TF) sono proteine in grado di legare il DNA coinvolti nella regolazione della trascrizione. Il legame dei TF avviene a sequenze di DNA specifiche che prendono il nome di motivi di legame. L'analisi dei motivi di legame è stata eseguita utilizzando findMotifsGenome.pl con parametri standard facente parte del *tool* HOMER (Heinz et al., 2010) v4.11 con parametri *standard*.

3.8 Annotazione dei picchi

Per identificare il gene più vicino a ciascun picco e la relativa distanza in numero di basi è stato utilizzato annotatePeaks.pl di HOMER v4.11 con l'utilizzo di parametri *standard*.

3.9 Caratterizzazione degli eRNA

3.9.1 Identificazione degli enhancer attivi

Per identificare gli *enhancer* attivamente trascritti sono stati utilizzati i risultati derivanti dall'analisi su più livelli disponibili nel portale *The Cancer eRNA Atlas* (TCeA) (Chen and Liang, 2020). Integrando le regioni identificate nella linea cellulare LAL-B, nei *cluster* C1 e C2, ottenuti a partire dai dati di ATAC-seq di paziente, e nel TCeA tramite *bedtools intersect* con parametri standard sono stati identificati 117 regioni codificanti eRNA.

3.9.2 Identificazione dei geni *target* degli *enhancer*

Per ciascuno dei 117 *enhancer* identificati è stato assegnato un solo gene come possibile *target*. Per questa caratterizzazione sono state usate diverse metodiche:

1. Utilizzo dei geni *target* per le regioni in cui era precedentemente stato caratterizzato in TCeA. Inizialmente è stato scaricato il *file* con i geni *target* identificati in TCeA. Tramite *bedtools intersect* con parametri *standard* sono stati identificati i geni *target*, quando disponibili, per i 117 eRNA identificati.
2. Definizione del *Trascription Start Site* (TSS) più vicino (Nasser et al., 2021) tramite l'utilizzo di annotatePeaks.pl di HOMER v4.11 con parametri standard

3.9.3 Identificazione degli eRNA identificati in campioni di paziente

Per poter valutare l'espressione dei 117 eRNA identificati precedentemente è stato prima modificato il file GTF inserendo gli eRNA. Il GTF così prodotto è stato poi utilizzato per allineare i fastq dei campioni tramite STAR con parametri '-quantMode TranscriptomeSAM --outSAMtype BAM SortedByCoordinate' ed infine è stata valutata l'espressione tramite RSEM.

3.10 Analisi dell'espressione

I dati utilizzati derivano dal database pubblico del progetto TARGET (https://ocg.cancer.gov/programs/target), phs000464. I dati usati per questa analisi sono disponibili presso https://portal.gdc.cancer.gov/projects. I dati di trascrittomica in RPKM sono stati prima filtrati selezionando solo i campioni esordio-recidiva *matched*. Successivamente sono stati filtrati solo i geni *target* delle regioni genomiche facenti parte dei *cluster* C1-C2. L'*heatmap* è stata prodotta utilizzando lo z-*score*, ed il metodo Ward D2 per la clusterizzazione.

3.11 Analisi del cistroma dei 117 enahncer

L'analisi è stata condotta sui *file* BAM di linee cellulari primarie e linee cellulari tumorali ottenuti da ENCODE. Successivamente i file BAM così ottenuti sono poi stati deduplicati, tramite l'utilizzo di GATK v 4.1.9.0 come precedentemente descritto per l'analisi delle ATAC-seq (Metodi 3.2.1). Per ciascun *file* è stata eseguita la chiamata dei picchi tramite il *tool* macs2 v 2.2.6 con i parametri 'callpeak --format AUTO --nomodel --shift -100 --extsize 200 -B --SPMR --call-summits -q 0.01 -g hs'. I picchi rilevati da ciascun esperimento di ChIP-seq analizzato sono poi stati intersecati con i 117 eRNA, precedentemente selezionati, tramite l'utilizzo di bedtools *intersect* con parametri *standard*. I risultati ottenuti dagli esperimenti di ChIP-seq del medesimo TF sono infine stati concatenati e poi uniti tramite bedtools *merge*.

3.12 Identificazione degli eRNA in campioni di paziente

Per identificare gli eRNA è stato usata la pipeline STAR-RSEM. Per prima cosa è stato modificato il file gtf di hg19 in modo tale da aggiungere le coordinate genomiche degli *enhancer* altrimenti assenti poiché il file GTF

include solo le informazioni relative ai geni . Successivamente è stato prodotto il genoma di riferimento utilizzando il *tool* STAR v2.7.8° con i seguenti parametri *"—runMode genomeGenerate*". L'allineamento è stato possibile tramite l'utilizzo di STAR con i seguenti parametri "*--readFilesCommand 'zcat' --quantMode TranscriptomeSAM --outSAMtype BAM SortedByCoordinate*".

Per la quantificazione è stato usato RSEM v 1.3.3. Il genoma di riferimento è stato prodotto utilizzando *rsem-prepare-reference* con parametri standard. La quantificazione è stata compiuta con *rsem-calculate-expression* con paramtri "*--bam --paired-end*"

3.13 Analisi *promoter-capture*

Gli esperimenti di *promoter-capture* permettono di identificare le interazioni che intercorrono tra promotori e potenziali regioni regolatorie distali. A differenza degli esprimenti di *Hi-C*, il *pormoter-capture* permette di diminuire la complessità delle librerie di *Hi-C*, in cui vengono indagate tutte le interazioni cromatiniche, tramite un preliminare l'arricchimento dei promotori presenti nel campione analizzato prima del sequenziamento (Schoenfelder et al., 2018) e permette così .

Il sequenziamento, prodotto presso l'Istituto Tumori Regina Elena nel laboratorio del dottor Maurizio Fanciulli, è stato condotto su un replicato di LAL-B.

I file FASTQ ottenuti sono poi stati analizzati con l'utilizzo del tool *juicer* (Durand et al., 2016) con parametri "*-S postproc*". Il risultato di questa analisi produce un *file* che permette la visualizzazione delle interazioni rilevate.

3.14 Analisi statistiche

Per le analisi statistiche sono stati utilizzati due test statistici. Il primo è il test di Kruskal-Wallis è un metodo non parametrico che utilizza le mediane di diversi gruppi e permette di verificare se tali gruppi provengono dalla stessa popolazione o da popolazioni con le mediane uguali. Il secondo è il Dunn-test è un test non parametrico *post hoc*, eseguito successivamente al test di Kruskal-Wallis, che permette compiere confronti multipli tra i vari gruppi presi in analisi.

4. Risultati

Per poter comprendere e definire le cause alla base dell'insorgenza tumorale e della comparsa di recidive è necessario compiere degli studi che tengano in considerazione la progressione e l'evoluzione tumorale.

L'unico modo per poter affrontare questo tipo di problematica biologica è tramite l'analisi di un elevato numero di campioni, possibilmente ottenuti dallo stesso paziente in diversi *timepoint* della patologia. La tecnica di campionamento precedentemente descritta prende il nome di campionamento longitudinale. Sebbene numerosi studi abbiano tentato, con ottimi risultati, di comprendere le alterazioni genetiche osservate nella B-ALL infantile, un numero esiguo di lavori ha posto l'accento sulla progressione tumorale dal punto di vista epigenetico.

Nonostante siano numerose le componenti dell'epigenetica, una delle maggiori problematiche è l'identificazione degli *enhancer*. Questo è principalmente dovuto alla grande mole di dati da integrare per poter distinguere un *enhancer* dalle altre regioni non codificanti presenti nel genoma.

L'analisi multi-*step* utilizzata in questo lavoro ha come obiettivo l'identificazione di nuovi *target* terapeutici della progressione della patologia, con principale interesse nelle regioni non codificanti coinvolte sia nell'esordio che nella recidiva tumorale. I dati sono stati ottenuti dai profili di accessibilità di pazienti di B-ALL in vari *timepoint* della patologia. L'analisi di questi profili ha consentito di identificare quale fosse la dinamica di apertura e chiusura della cromatina nel corso della progressione tumorale.

La forza dello studio risiede nella larga coorte di pazienti e nell'integrazione di dati ottenuti da *database* pubblici (HeRA (Zhang et al., 2021), TCeA (Chen and Liang, 2020), TARGET ed ENCODE (Davis et al., 2018)) per descrivere ad alta precisione e risoluzoine le regioni genomiche identificate.

4.1 Descrizione dei dataset a disposizione

Lo studio è stato condotto a partire dai profili di ATAC-seq di 32 campioni di B-ALL infantile nei diversi stadi della patologia, così divisi: 6 sani, 11 esordi, 7 remissioni e 8 recidive. I picchi identificati mediante l'analisi delle ATACseq di ciascun campione mostrano una distribuzione con media e deviazione *standard* rispettivamente pari a 57.082,84 e 11.682.78 (Fig. 4). Solo un campione di recidiva mostra un numero di picchi estremamente basso, 24.162, rispetto alla media dei picchi identificati, 57082. (50-60 milioni)

L'estrazione dei linfociti B compiuta sui pazienti è stata ottenuta mediante un arricchimento delle popolazioni cellulari positive alla proteina di membrana CD19+, *marker* di membrana dei linfociti B espresso in tutti gli stadi dei linfociti B, comprese le plasmacellule

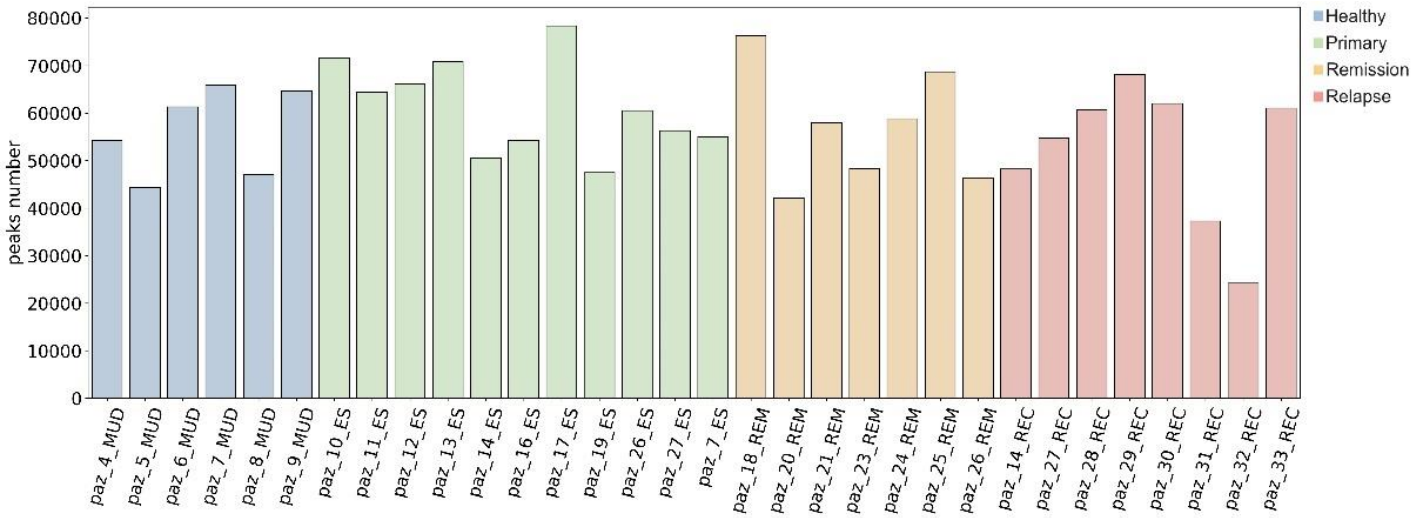


Figura 4 : Il barplot rappresenta il numero di picchi identificato in ciascun campione, per tanto sull'asse delle X sono disposti i singoli campioni colorati in base al timepoint di appartenenza (blu:sano, verde:esordio, giallo:remissione, rosso:recidiva), mentre l'asse delle Y mostra il numero di picchi identificati.

4.2 *Ranking index* e *Sharing index*

Una volta processati i dati grezzi delle ATACseq (Metodi 3.2.1), i risultati sono stati processati tramite l'assegnazione di due nuovi indici: lo *Sharing index* (SI) e il *Ranking index* (RI). La necessità di assegnare due nuovi indici a ciascun picco deriva dalla mancanza di informazioni relative alla clonalità e alla penetranza tramite l'utilizzo delle *pipeline* bioinformatiche standard e dall'assenza di una metodica computazione in grado di analizzare in maniera sistematica i profili di ATAC-seq. Inoltre, la forma e l'altezza di un picco sono strettamente legate alla clonalità del picco nel campione e alla penetranza nella popolazione (Patten and Corleone et al., 2018).

L'assegnazione del RI a ciascun picco permette di identificare la clonalità di un picco all'interno di un campione ed è descritta nei Metodi. I valori che possono essere assegnati vanno da 1 a 100, identificando rispettivamente un picco estremamente clonale (1) e un picco estremamente sottoclonale (100). La definizione di picco clonale o sottoclonale è effettuata tramite il conteggio delle *read* derivanti da quella porzione genomica, infatti un picco con poche *read* sarà definito sottoclonale, mentre un picco con molte *read* è considerato clonale.

Il valore di SI assegnato è invece in grado di identificare la penetranza di un picco nei campioni. L'assegnazione di questo indice mira ad identificare quanti campioni condividano il medesimo picco. Per definire la penetranza sono state utilizzate due metodiche differenti, entrambe descritte nei metodi (Metodi 3.5).

Successivamente è stata valutata la relazione che intercorre tra SI (Fig. 5, asse delle X), qui calcolato come la condivisione di un picco in tutti e 32 i campioni analizzati, e RI (Fig. 5, asse delle Y). Questa analisi ha permesso di evidenziare come ci sia una stretta relazione tra penetranza e clonalità. Infatti, è possibile vedere come regioni ad alta clonalità (basso RI) siano anche quelle che mostrano un'elevata penetranza (alto SI), mentre le regioni sottoclonali (alto RI) siano più paziente specifiche, e questo è riscontrato in tutti i *timepoint*. Il risultato ottenuto è in linea con le aspettative,

dal momento che i picchi clonali si ipotizzano essere indispensabili per la vitalità cellulare e quindi saranno anche largamente condivisi in tutti i campioni.

Questo risultato evidenzia anche come sia possibile utilizzare lo SI e il RI per definire la clonalità di un picco in ciascun campione, considerando che, con uno spettro di valori maggiore, il RI permette di discriminare in maniera più fine la clonalità di un picco rispetto allo SI. Quest'ultimo è in grado, tuttavia, di identificare rapidamente quali siano i picchi maggiormente condivisi in ciascun *timepoint*, come descritto nella seconda metodica di definizione dello SI.

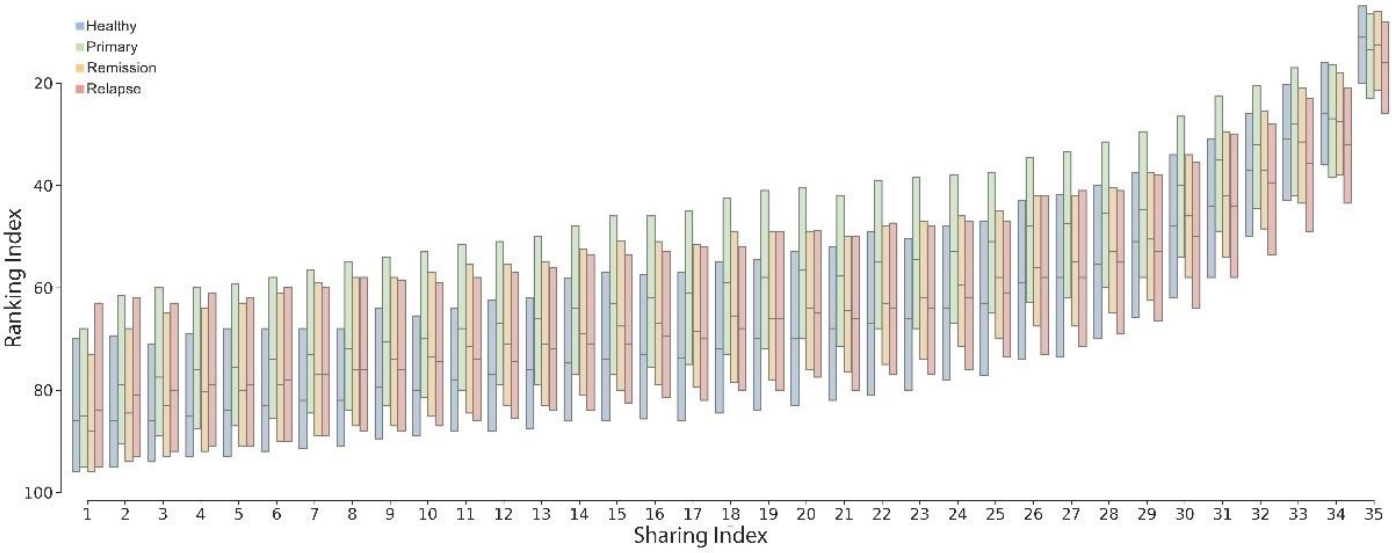


Figura 5 : Il boxplot evidenzia un andamento concorde tra SI e RI. Infatti al crescere dell'SI (asse delle x) diminuisce RI, ovvero aumenta la clonalità, (asse delle y). Qui per ogni SI sono stati raggruppati gli RI nei vari timepoint (blu:sano, verde:esordio, giallo:remissione, rosso:recidiva)

Una volta definita la correlazione tra SI e RI è stato necessario definire la percentuale di picchi dei vari *timepoint* che confluiscono in ciascun valore dello SI (Fig. 6). Questo risultato ha permesso di definire chiaramente il maggior contributo degli esordi nelle basse penetranze. Ciò denota una maggiore eterogeneità degli esordi rispetto agli altri *timepoint*, e potrebbe

essere attribuito ad un processo simile all'evoluzione in atto negli esordi, dove le cellule tumorali attivano numerosi sistemi di adattamento al microambiente.

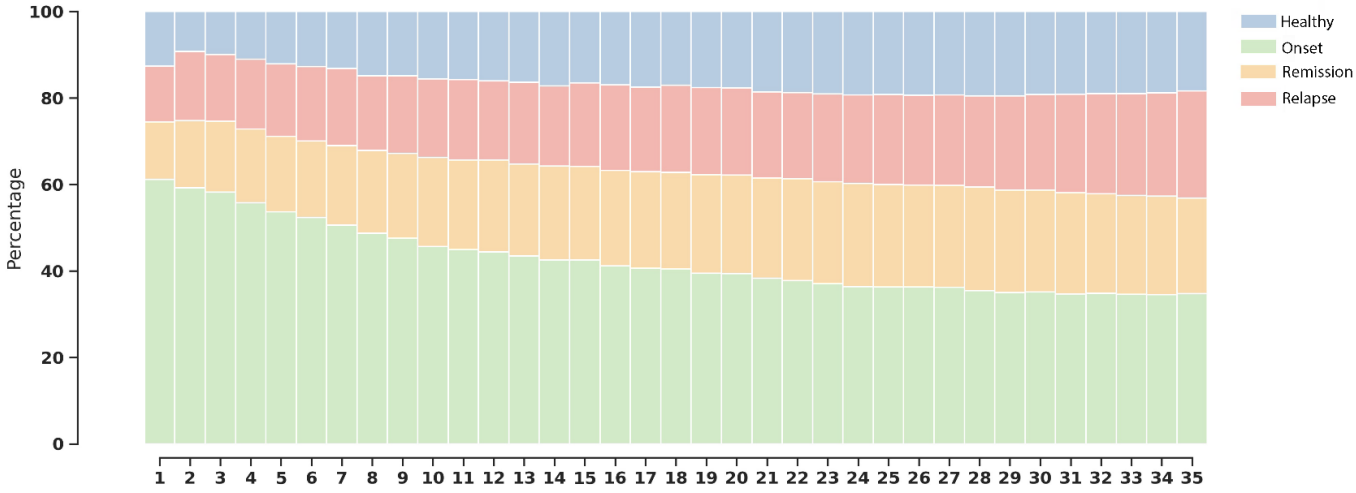


Figura 6 : L'asse delle X rappresenta lo Sharing Index da 1 a 35 mentre sull'asse delle Y sono presenti le percentuali dei picchi. È rappresentata una grande omogeneità in sani, remissioni e recidive, a differenza degli esordi che mostrano una grande eterogeneità tra i campioni.

4.3 Distribuzione nel genoma dei picchi

Successivamente è stata valutata la distanza dei picchi dal sito di inizio della trascrizione (*Trascription Start Site,* TSS) più prossimo (Fig. 7). L'analisi compiuta su tutti i picchi identificati da almeno un campione in ogni *status* mira ad identificare la distanza assoluta dal TSS più vicino. Analizzando l'andamento è stato possibile notare come le remissioni e le recidive mostrino un numero di picchi costante nelle varie distanze dal TSS. Al contrario esordi e sani mostrano un andamento opposto al crescere della lontananza dal TSS. Infatti, i campioni sani mostrano un maggiore arricchimento in prossimità del TSS e un contributo minore a distanze maggiori di 100 kilobasi, contrariamente gli esordi mostrano un numero di picchi maggiore a distanze elevate piuttosto che in prossimità del TSS.

Poiché un *enhancer* più trovarsi ad una distanza dal TSS che va da 5 kb a 100 kb è possibile considerare ciascun picco compreso in questi intervalli di distanza come un potenziale *enhancer*.

Risulta quindi evidente come gli esordi mostrino un maggiore contributo, rispetto agli altri *timepoint*, in ciascun intervallo di distanza, questo potrebbe però essere dovuto sia alla maggiore eterogeneità sia al maggior numero di campioni disponibili per gli esordi.

Risulta però inevitabile osservare come a distanze maggiori dal TSS gli esordi mostrino un maggior livello di picchi. Questo, unito al risultato ottenuto nel precedente paragrafo, potrebbe definire una maggiore eterogeneità degli esordi per quanto concerne le regioni con funzione regolatoria, quindi la plasticità cellulare.

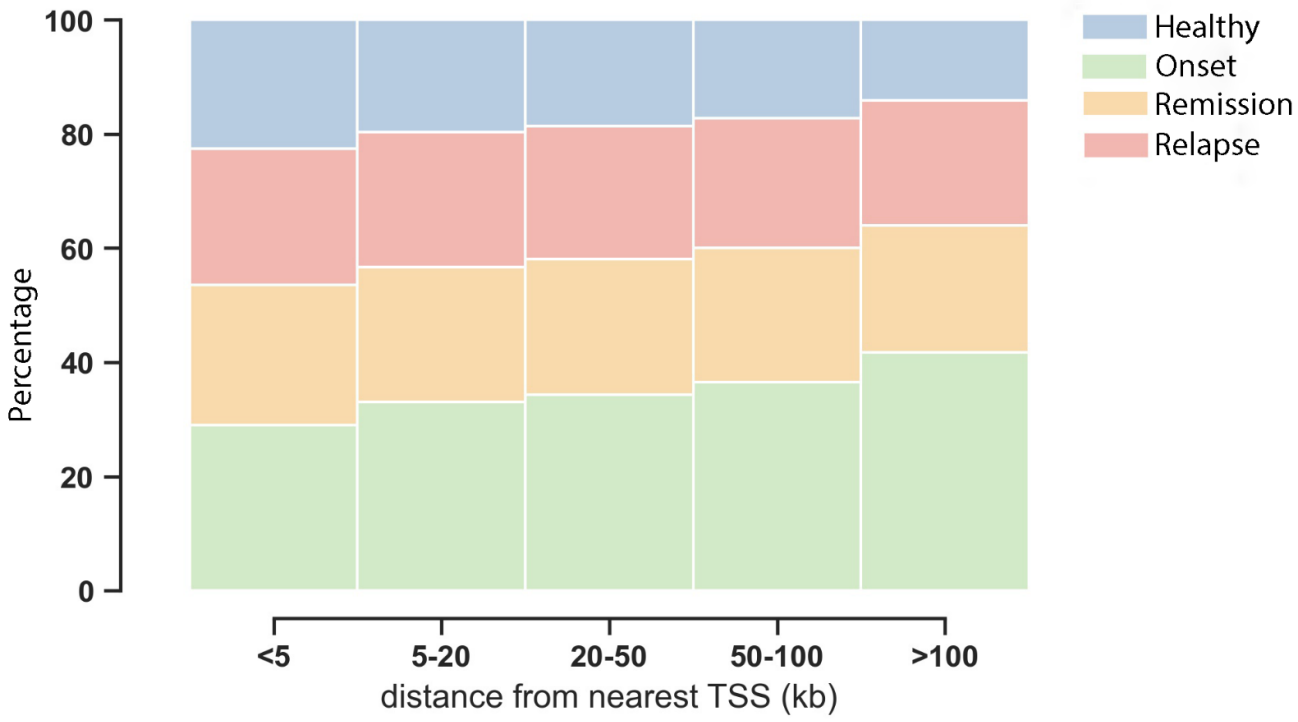


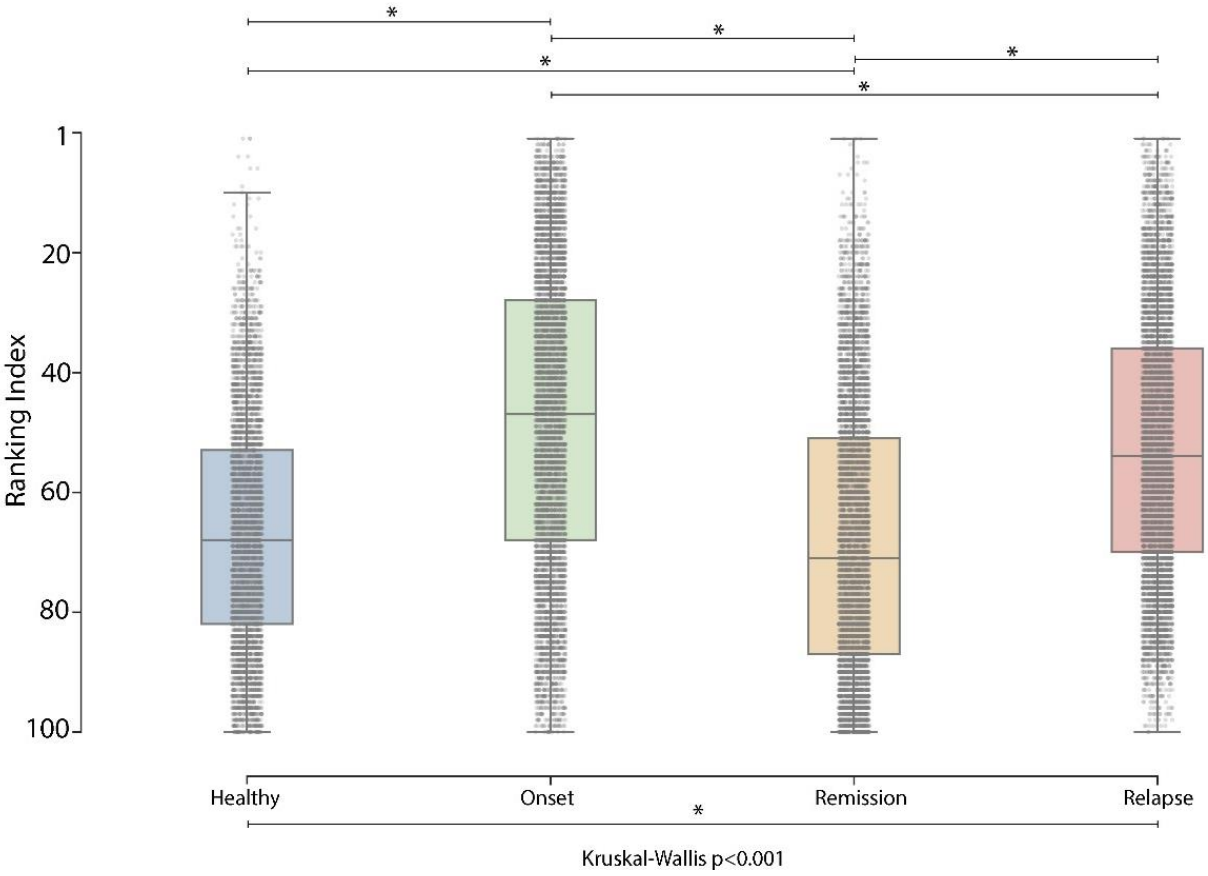
Figura 7 : L'asse delle X rappresentata la distanza in kilobasi dal TSS più vicino mentre sull'asse delle Y sono presenti le percentuali dei picchi. Gli esordi mostrano un maggiore numero di picchi con potenziali funzioni regolatorie, data la loro distanza dal TSS.

4.4 Identificazione delle regioni genomiche coinvolte nella progressione tumorale

Oltre a descrivere clonalità e penetranza lo SI e il RI sono stati utilizzati per poter identificare tutti quei picchi coinvolti nella progressione tumorale.
Tramite l'utilizzo di *threshold* e l'utilizzo di SI e RI si è resa possibile la selezione di picchi che presentano cambiamenti di clonalità e penetranza nei i vari *timepoint* (Metodi 3.5). Per l'individuazione dei picchi di interesse la selezione è avvenuta in base a tre principi biologici differenti. Tale processo di stratificazione è necessario per concentrare le analisi su un numero di picchi limitato rispetto alle circa 150.000 regioni identificate nella *masterlist*, composta da tutte le regioni di cromatina aperta identificate in almeno un campione sequenziato. Inizialmente sono stati selezionati i picchi che mostrano un'elevata condivisione tra gli esordi rispetto sani basandosi sull'SI, questo ha consentito di individuare 6.000 regioni aperte all'interno degli esordi e completamente chiuse nei sani, utilizzando tale criterio di selezione vengono anche escluse tutte le regioni del DNA aperte ma che mostrano elevata condivisione sia nei sani che negli esordi. L'esclusione di queste regioni è necessaria dal momento che se una regione è stata rilevata aperta in tutti i campioni sia di esordio che normali si ipotizza il suo coinvolgimento nelle attività necessarie alla sopravvivenza cellulare, quindi non direttamente legate all'esordio o alla progressione della patologia. Tra i picchi esclusi dalla prima selezione sono stati selezionati 1700 picchi che mostrano una riduzione di RI tra esordio e sano, questi sono quindi i picchi che mostrano un incremento di clonalità nel passaggio tra le due condizioni. Infine, sono stati aggiunti 3300 picchi, che rappresentano tutti quei picchi con un incremento di clonalità tra remissioni ed recidive. Analizzando gli RI degli 11.000 picchi selezionati (Fig.8) tra le varie condizioni si evidenzia come ci sia un andamento della clonalità di interesse. Infatti, i picchi mostrano una bassa clonalità dei pazienti sani, per poi incrementare nel momento di insorgenza della malattia, per poi

diminuire post-trattamento e tornare a livelli dei sani nelle remissioni ed, infine, aumentare nuovamente quando si manifesta la recidiva.

Figura 8 : Il boxplot rappresenta l'andamento della clonalità degli 11.000 picchi selezionati. Ciascun boxplot identifica un differente timepoint, mentre sull'asse delle Y è rappresentato il Ranking Index.



La vicinanza della mediana tra sani, 68, e remissioni, 71, (*p-value*<0.02) è indicativa della somiglianza fenotipica tra queste due condizioni. Questo è vero anche per esordi e recidive anche se le mediane mostrano una minor somiglianza (*p-value*<0.00001), probabilmente dovuta al momento in cui vengono diagnosticate le recidive. Infatti, i pazienti, dopo una remissione, sono soggetti a monitoraggi che permettono la diagnosi precoce di recidiva, questo significa che nelle remissioni troveremo una minore clonalità rispetto agli esordi.

Per rendere più robusto questo risultato è stata compiuta un'analisi sul numero di *reads* sequenziate per ciascun picco nei diversi campioni. Successivamente è stata effettuata una clusterizzazione degli 11.000 pichi. Questa analisi ha prodotto un'*heatmap* con i picchi sulle righe, i pazienti sulle colonne e lo z-score che ne identifica l'intensità (Fig. 9).

Di interesse è come vengono raggruppati i picchi in base al *clustering* non supervisionato applicato sia sulle righe che sulle colonne. Infatti, i pazienti riescono ad essere raggruppati in maniera molto netta mostrando due gruppi ben distinti di esordio-recidiva e sano-remissione. Il comportamento dei picchi non è però uniforme. Infatti, sono evidenziabili quattro differenti *cluster* di picchi identificati con C1, C2, C3 e C4. I primi due *cluster* mostrano una decisiva apertura di queste regioni genomiche nel passaggio dal fenotipo sano a quello malato; il *cluster* C3 ha invece un comportamento inverso, identificando regioni soggette a chiusura tra sano e malato; il C4 che mostra, invece, un'elevata variabilità tra le quattro condizioni.

Questo risultato mette in luce come le selezioni e le valutazioni compiute basandoci su SI e RI siano effettivamente rispecchiate dal numero di *reads* normalizzate identificate (Metodi 3.6) da ciascun picco a diversi *timepoint*.

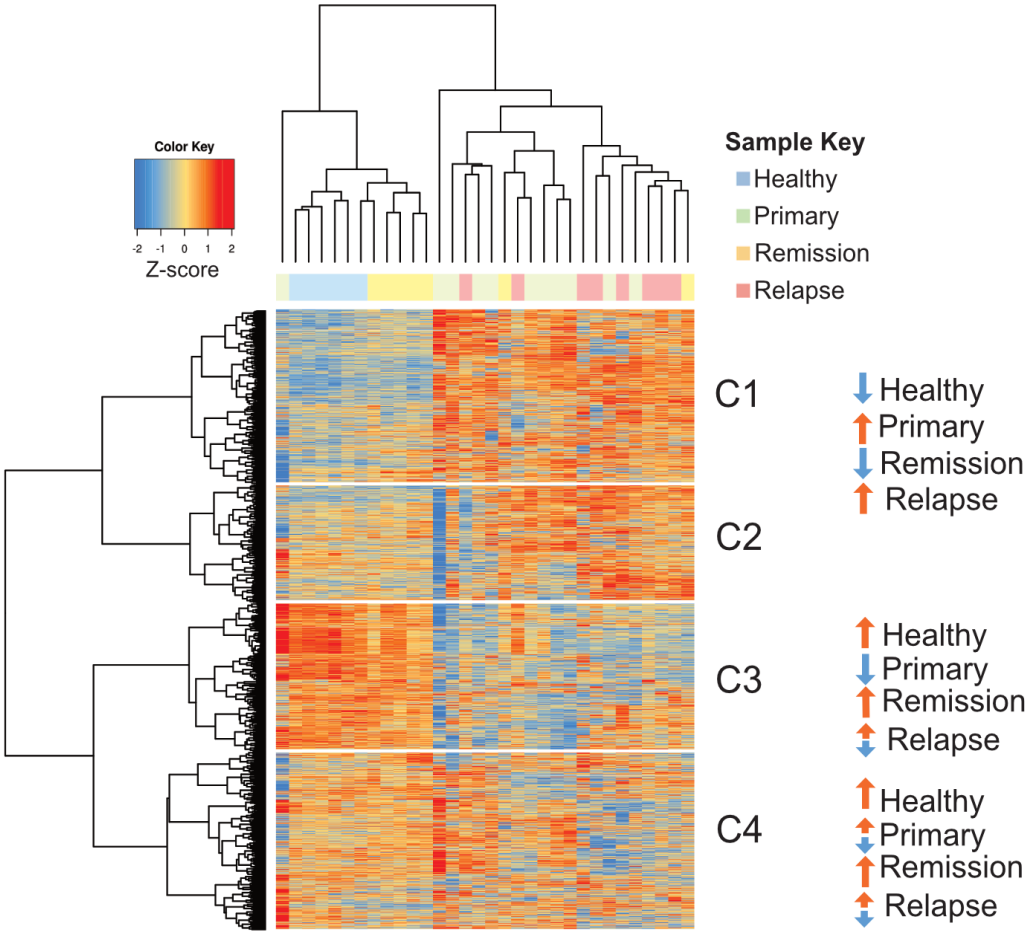


Figura 9 : L'heatmap mostra i campioni sulle colonne, gli 11.000 picchi sulle righe e lo Z-score del numero di reads sequenziate ne definsice l'intensità.

4.5 Analisi del cistroma attivo nei *cluster*

I fattori trascrizionali (TF) sono delle proteine coinvolte in un elevato numero di compiti all'interno del nucleo, principalmente sono responsabili del controllo dell'espressione genica e del conseguente destino cellulare (Stadhouders et al., 2019). Per poter espletare la loro funzione i TF hanno bisogno di legarsi a determinate sequenze TF-specifiche del DNA.

Pertanto, una volta identificate delle regioni genomiche che alterano la loro accessibilità durante il decorso della malattia, è indispensabile individuare quali TF riconoscono le sequenze di legame contenute in queste regioni. Tramite l'utilizzo di HOMER (Heinz et al., 2010) è stato possibile identificare

quali potessero essere i TF in grado di legare le regioni appartenenti a ciascun *cluster*. Principalmente si è cercato di individuare i TF con la maggiore affinità per le regioni dei *cluster* C1 e C2, rispetto a C3 e C4, ovvero quei TF coinvolti direttamente nella progressione della B-ALL.

Una volta effettuata l'analisi con HOMER, è stato graficato il rapporto tra la percentuale delle regioni che contengono la sequenza TF-specifica e la percentuale delle regioni che contengono la medesima sequenza ma in un campione casuale di regioni genomiche (Fig. 10).

Nel *circular barplot* sono stati selezionati i 10 TF che mostravano un rapporto maggiore nei *cluster* C1 e C2 rispetto ai *cluster* C3 e C4. Sebbene tutti e 10 i TF siano di notevole interesse in quanto coinvolti nella progressione tumorale di B-ALL (Tejedor et al., 2021), l'attenzione si è focalizzata su ERG ed EBF.

La selezione è stata compiuta considerando due aspetti principali: una maggiore percentuale di sequenze attese e l'osservazione dei dati di proliferazione cellulare successiva al *knockout* dei fattori trascrizionali presenti sul portale DepMap.

ERG è un TF facente parte della famiglia dei *TF erythroblast transformation-specific* (ETS), tutti i membri sono coinvolti in: apoptosi, proliferazione cellulare, angiogenesi, infiammazione e sviluppo embrionale. Sebbene il rapporto tra segnale del promotore e segnale dell'*enhancer* (REP) mostri un incremento minore dei *cluster* C1, 2.66, e C2, 3.00, paragonati a C3, 2.00, e C4, 2.25, ciò che risulta interessante è la percentuale di frammenti contenenti sequenze di legame attese che ha una media prossima al 16% nei diversi *clusters*. Questo sta ad indicare che le sequenze nei cluster C1 e C2 che permettono il legame di ERG siano un numero elevato. Inoltre i dati di proliferazione cellulare pubblicati su DepMap successivamente al *knockout* di ERG in molteplici linee cellulari, evidenziano .

EBF1 è parte della famiglia degli *early B-cell factors* (EBFs) composta da 4 TF coinvolti nel differenziamento e maturazione di molteplici linee cellulari inclusi i linfoblasti progenitori dei linfociti B. Qui l'interesse è stato attratto

dall'elevato REP evidenziato in C1 e un rapporto nettamente più basso in C3, *cluster* contenente regioni che si chiudono nella progressione della patologia e aperte nei pazienti sani e nelle remissioni. Dato il ruolo svolto da EBF1 (dati derivanti da DepMap) mostrano una grande specificità per linfociti B, nelle linee ematopoietiche e linfoidi ed ALL.

Al fine di comprendere quali delle regioni individuate nei 4 cluster siano legate da EBF1 e ERG sono stati disegnati esperimenti di ChIP-seq dei due TF, ma i risultati non sono ancora disponibili a causa della poca efficienza degli anticorpi fi'ora utilizzati.

Comprendere quali regioni siano legate dai fattori trascrizionali identificati consentirebbe di migliorare la comprensione delle dinamiche che comportano le alterazioni nell'accessibilità cromatinica. Per approfondire il funzionamento molecolare si potrebbero condurre esperimenti di siRNA tramite la tecnologia dell'RNA *interference* per determianre se il silenziamento di ERG ed EBF1 produca la chiusura delle regioni facenti parte dei *cluster* C1 e C2. In questo modo si potrebbe anche definire se, per tali regioni, questi TF svolgano il ruolo di *pioneering factor*, alterando quindi il profilo di accessibilità. Inoltre, integrando esperimenti di Hi-C o *Promoter-Capture* si potrebbe definire in maniera più chiara se il legame dei fattori trascrizionali possa essere coinvolto nella formazione di *loop* cromatinici in grado di regolare l'espressione genica.

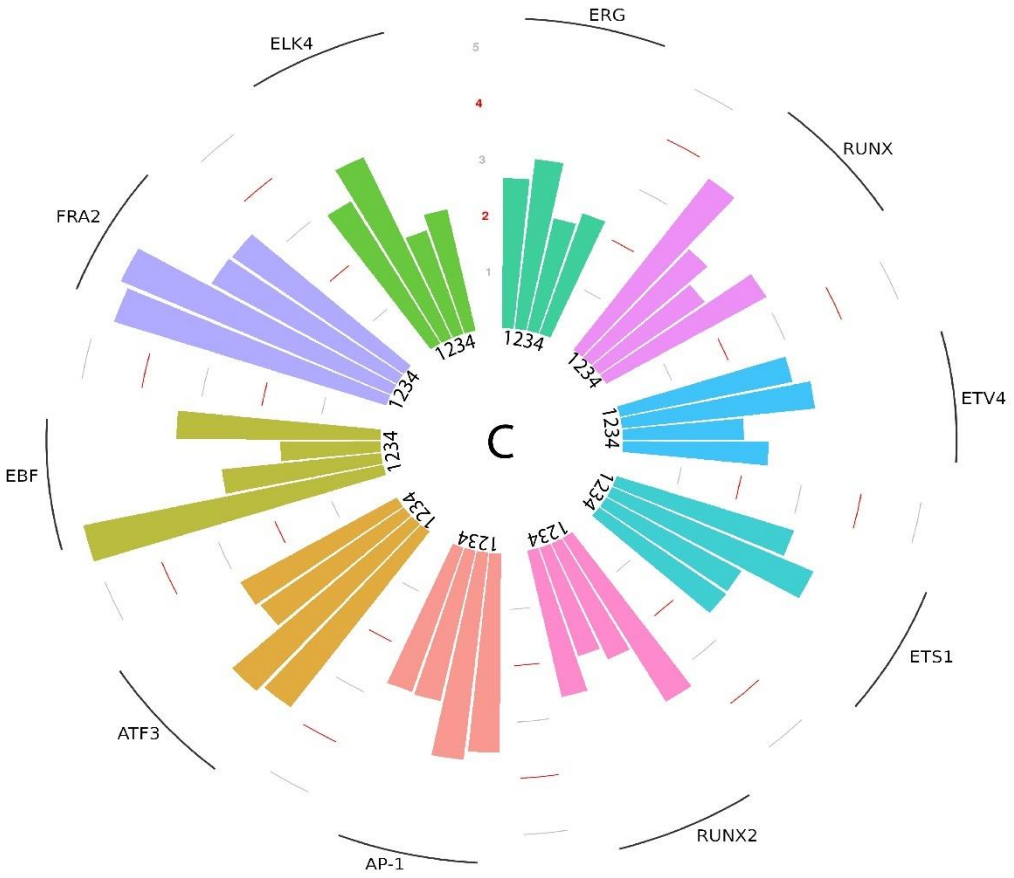


Figura 10 : Il circular barplot mostra 10 fattori trascrizionali e per ciascuno sono rappresentate 4 barre. Ognuna delle barre rappresenta il rapporto tra percentuale delle sequenze che contengono la sequenza di binding del fattore trascrizionale atteso e la percentuale delle sequenze di binding osservate per ciascun cluster (C1, C2, C3 e C4). Tutti i fattori trascrizionali mostrati hanno un rapporto maggiore nei cluster C1 e C2 rispetto a C3 e C4.

4.6 Identificazione degli *enhancer* attivamente trascritti

Le regioni dei *cluster* C1C2 comprendono tutti i siti che mostrano una maggiore apertura nei campioni di esordio e recidiva rispetto ai campioni di sani e remissioni. Solo una piccola parte di questi sono già stati caratterizzati come *enhancer* veri e propri. Per definire quali sono gli

enhancer tra le circa 6.000 regioni dei due cluster sono stati usati i dati di espressione degli eRNA contenuti dal portale *The Cancer eRNA Atlas* (TCeA). Il TCeA contiene più di 300.000 *canonical-enhancer* e *super-enhancer* annotati e quantificati in più di 20.000 campioni ottenuti da *The Cancer Genome Atlas*, *Genotype-Tissue Expression Portal* e *Cancer Cell Line Encyclopedia*. Per l'identificazione degli *enhancer* che si trovano in C1C2 è stato aggiunto un ulteriore parametro di stringenza. Infatti, oltre alla presenza in TCeA, è richiesto che il picco identifficato sia presente anche all'interno delle ATACseq della linea cellulare LAL-B, i cui profili contengono circa 150.000 picchi. Questa ulteriore condizione si è resa necessaria per identificare gli *enhancer* coinvolti nella progressione tumorale ma che potessero anche essere successivamente saggiati in laboratorio. Sono così stati identificati 117 *enhancer* presenti in tutti e tre i *dataset* presi in considerazione. Pertanto, ogni *enhancer* individuato è presente sia nei pazienti sia nelle LAL-B, ma è anche stato precedentemente caratterizzato come un *enhancer* in grado di produrre eRNA in TCeA.

Le cellule LAL-B sono cellule primarie mononucleari isolate dal midollo osseo di paziente affetto da BCP-ALL. Il confronto tra i picchi di ATACseq ottenuti dalle LAL-B e l'insieme dei picchi ottenuti dai campioni di esordio e recidiva conferma che questa linea cellulare è in grado di ricapitolare efficientemente entrambi i fenotipi, anche se c'è una maggior corrispondenza tra LAL-B ed esordio (Fig. 11). Il grande numero di picchi di esordio che non hanno nessuna sovrapposizione con gli altri stadi può essere giustificato dall'alta variabilità presente nei pazienti che mostrano spesso dei picchi paziente specifici. Inoltre, bisogna considerare che una linea cellulare non riesce a ricapitolare la complessità e la variabilità paziente specifica della patologia.

Possedere un modello sperimentale in grado di rappresentare in maniera efficiente il fenotipo della malattia è una condizione indispensabile per poter condurre studi sulla malattia e indagare i meccanismi molecolari che ne sono alla base.

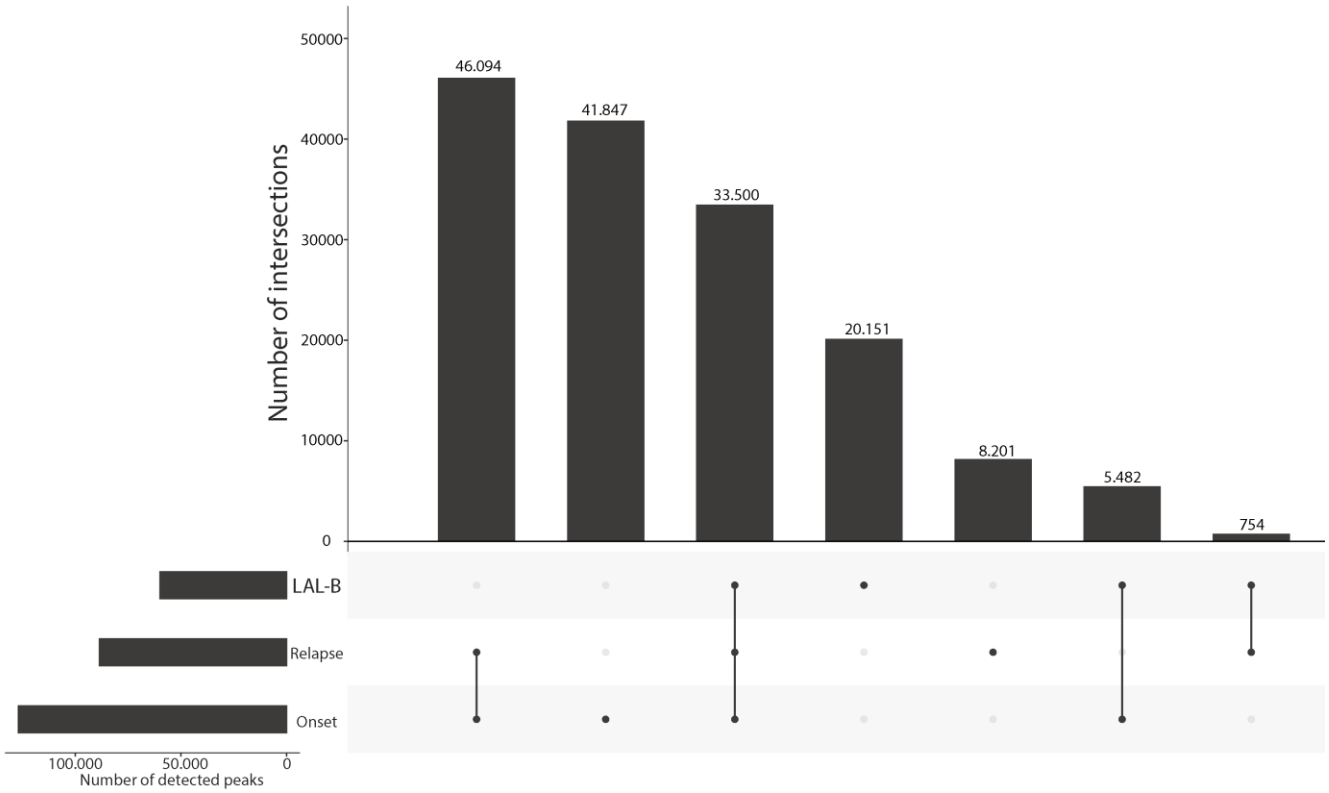


Figura 11 :. L'upsetplot è stato ottenuto analizzando le intersezioni dei picchi di ATAC-seq rilevati negli esordi, nelle recidive e nella linea cellulare LAL-B, in modo da poter identificare quale condizione (esordio o recidiva) contenesse il numero di siti maggiori con le LAL-B. In basso a sinistra sono rappresentate le tre condizioni (LAL-B, Onset e Relapse) con il numero di picchi identificati. Le barre sulla destra, invece, identificano il numero di picchi trovati in comune tra le condizioni unite dalla barra sottostante.

4.7 Caratterizzazione degli eRNA in altri tessuti

Una volta identificati i 117 *enhancer* putativamente coinvolti nella progressione della B-ALL, è stata compiuta una caratterizzazione del grado di acetilazione sull'istone H3 in posizione K27 (H3K27ac) di tessuti e linee cellulari sia sani che malati. La caratterizzazione delle H3K27ac risulta fondamentale dal momento che è stata identificata una correlazione tra livelli di acetilazione ed attività di un *enhancer* (Sungalee et al., 2021).

La prima analisi è stata compiuta sui dati di H3K27Ac di linee cellulari disponibili su ENCODE, per un totale di 161 linee analizzate. Il risultato dell'analisi è riassunto dall'*heatmap* (Fig. 12) in cui è mostrato il RI di ciascuno dei 117 *enhancer* nelle varie linee cellulari. Questa analisi rivela che la maggior parte degli *enhancer* osservati dell'analisi sia non attivo nella gran parte delle linee cellulari presenti su ENCODE. Solo un piccolo gruppo di regioni mostra alta clonalità in alcune linee cellulari, nessuna delle quali è però connessa al sistema immunitario o a tipologie di leucemia. Questo suggerisce che gli *enhancer* identificati siano per la maggior parte B-ALL specifici. Inoltre, i siti indagati mostrano bassi livelli di acetilazione nelle linee linfocitarie come i linfociti B e linfociti T, questo è una conferma delle osservazioni compiute sui *cluster* C1 e C2. Una delle limitazioni dell'analisi delle linee cellulari presenti su ENCODE è la assenza di linee cellulari di B-ALL. Tuttavia, è stato comunque possibile ottenere una panoramica in grado di descrivere l'attivazione degli *enhancer* individuati in un numero elevato (n=161) di linee cellulari provenienti da molteplici tessuti.

Un'ulteriore caratterizzazione è stata compiuta utilizzando i risultati pubblicati nel portale HeRA che contiene tutti gli eRNA attivamente trascritti nei tessuti contenuti all'interno di GTeX (Lonsdale et al., 2013). GTeX è un portale in cui sono contenuti dati di espressione genica derivanti da 54 tessuti sani. Questo tipo di analisi è in grado di descrivere, oltre l'attivazione degli *enhancer* dedotta dai segnali di H3K27Ac, i livelli di produzione degli eRNA nei vari tessuti.

L'analisi dei dati di HeRA evidenzia come le regioni appartenenti a ciascun *cluster* siano presenti anche in altri tessuti, seppur con livelli molto bassi, sempre prossimi allo 0 e in un numero limitato. In tutti i *cluster* il tessuto denominato *Cells Leukemia* mostra mediane maggiori di RPKM maggiori rispetto agli altri tessuti, sebbene non sia specificato che tipologie di leucemia siano rappresentate. Caratteristica per il *cluster* C2 (Fig. 12B) è la presenza di eRNA identificati in *Whole Blood* con la mediana minore alla maggior parte dei tessuti rappresentati. È inoltre evidente come il numero

di eRNA identificati nei vari tessuti sia minore per i *cluster* C1 e C2 (Fig. 13 A-B) mentre il numero di eRNA identificati in ciascun tessuto aumenta per il cluster C3 (Fig. 13 C) e C4 (Fig. 13 D). Quest'ultimo mostra anche una maggiore variabilità di espressione inter-tissutale ed intra-tissutale. Questi risultati confermano quanto osservato precedentemente tramite ENCODE in quanto i livelli degli eRNA sono estremamente bassi e spesso non rilevati in tutti i tessuti. Il maggior risultato dell'analisi è quello di aver definito come varia la trascrizione dei siti di ciascun *cluster* nei diversi tessuti presenti su HeRA.

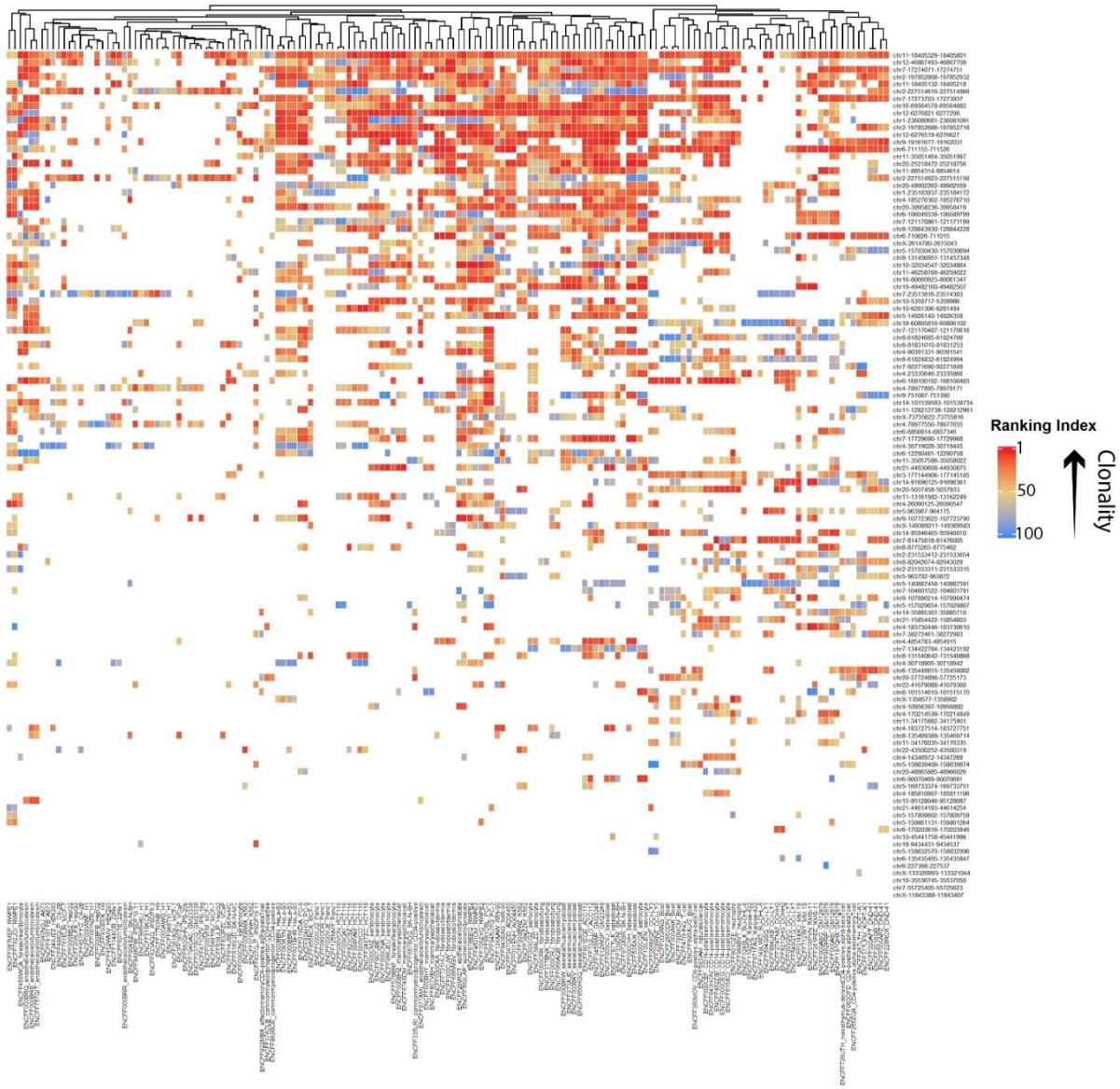


Figura 12 L'heatmap è il risultato dell'analisi dei segnali di acetilazione di 161 linee cellulari (asse delle x). Per ogni enhancer (asse delle Y) è stato valutato il Ranking Index assegnato a quel sito nell'analisi dei segnali di ChIPseq H3K27Ac in ognuna delle 161 linee cellulari ottenute da ENCODE. Gli enhancer sono stati ordinati in maniera decrescente in base al numero di linee cellulari in cui è stato rilevato il segnale in quella regione. Le linee cellulari sono state clusterizzate tramite il metodo Ward D2.

C1
RPM
Prostate
Kidney
Cells_fibroblast
Uterus
Esophagus_gast
Artery_cor
Skin_sun
Skin_no_sun
Artery_aor
Vagina
Lung
Heart_Atr
Breast
Adipose_sub
Bladder
Brain_Spi
Minor_salivary_gland
Small_intestine
Whole_blood
Brain_Hippo
Brain_Hypo
Brain_Nuc
Brain_Fro
Nerve
Brain_Cortex
Adipose_vis
Brain_Put
Thyroid
Testis
Esophagus_muc
Brain_Sub
Brain_Cerebellum
Artery_tib
Cervix_endo
Esophagus_mus
Spleen
Brain_Amy
Colon_sigmoid
Heart_Left
Ovary
Brain_Cer
Colon_transverse
Liver
Muscle_skeletal
Brain_Ant
Cells_ebv
Pituitary
Pancreas
Stomach
Brain_Cau
Adrenal
Cells_leukemia
Cervix_ecto
Fallopian_tube
C2
RPM
Whole_blood
Stomach
Pituitary
Brain_Spi
Cells_fibroblast
Adipose_vis
Prostate
Small_intestine
Muscle_skeletal
Artery_cor
Minor_salivary_gland
Heart_Atr
Adrenal
Skin_no_sun
Colon_transverse
Lung
Artery_aor
Thyroid
Brain_Cerebellum
Breast
Esophagus_gast
Vagina
Liver
Uterus
Colon_sigmoid
Fallopian_tube
Kidney
Nerve
Skin_sun
Esophagus_mus
Cervix_endo
Artery_tib
Testis
Bladder
Esophagus_muc
Pancreas
Ovary
Brain_Cer
Spleen
Adipose_sub
Brain_Cortex
Cervix_ecto
Brain_Sub
Brain_Put
Brain_Nuc
Brain_Ant
Brain_Cau
Cells_ebv
Heart_Left
Brain_Fro
Brain_Amy
Brain_Hypo
Brain_Hippo
Cells_leukemia
C3
RPM
Kidney
Stomach
Heart_Left
Lung
Small_intestine
Cervix_ecto
Artery_cor
Bladder
Esophagus_mus
Heart_Atr
Colon_sigmoid
Minor_salivary_gland
Esophagus_muc
Cells_fibroblast
Thyroid
Skin_no_sun
Muscle_skeletal
Vagina
Esophagus_gast
Whole_blood
Breast
Colon_transverse
Adipose_vis
Skin_sun
Cervix_endo
Brain_Cerebellum
Liver
Artery_aor
Brain_Fro
Brain_Put
Artery_tib
Ovary
Adrenal
Brain_Cortex
Prostate
Brain_Amy
Uterus
Fallopian_tube
Brain_Sub
Brain_Hippo
Adipose_sub
Brain_Hypo
Pancreas
Brain_Ant
Testis
Nerve
Brain_Spi
Cells_ebv
Spleen
Pituitary
Cells_leukemia
Brain_Cau
Brain_Nuc
Brain_Cer
C4
RPM
Whole_blood
Heart_Left
Cervix_ecto
Adipose_vis
Lung
Adrenal
Liver
Small_intestine
Adipose_sub
Cervix_endo
Kidney
Uterus
Brain_Hippo
Heart_Atr
Brain_Hypo
Skin_no_sun
Skin_sun
Cells_fibroblast
Colon_sigmoid
Esophagus_gast
Stomach
Brain_Cerebellum
Breast
Brain_Spi
Muscle_skeletal
Testis
Artery_cor
Fallopian_tube
Prostate
Bladder
Spleen
Brain_Amy
Esophagus_mus
Minor_salivary_gland
Ovary
Vagina
Brain_Cortex
Thyroid
Brain_Put
Pituitary
Artery_aor
Brain_Fro
Esophagus_muc
Colon_transverse
Brain_Nuc
Brain_Sub
Brain_Cer
Nerve
Brain_Cau
Artery_tib
Cells_ebv
Brain_Ant
Pancreas
Cells_leukemia

Figura 13 : Valutazione dell'espressione in RPM degli eRNA presenti nei 4 cluster e descritti nel database HeRA. Per ciascun cluster sono stati graficati il numero di eRNA identificati in ciascun tessuto e i livelli di RPKM di questi. I tessuti in ogni cluster (C1, C2, C3 e C4) sono ordinati con le mediane crescenti da sinistra a destra. I boxplot C1, C2, C3 e C4si riferiscono rispettivamente al Cluster 1, Cluster 2, Cluster 3 e Cluster 4. I numeri sotto a ciascun boxplot rappresentano il numero di eRNA trovati per ogni cluster in ciascun tessuto, label sull'asse delle x.

4.8 Analisi del cistroma dei 117 *enhancer*

Per ottenere una completa caratterizzazione degli *enhancer* coinvolti nella progressione e per comprenderne i meccanismi molecolari di azione è necessario individuare i TF che si legano in questi siti. In questo caso, dato il numero limitato di siti da indagare, sono state utilizzate le ChIPseq pubblicate disponibili sul portale ChIP-Atlas (Oki et al., 2018) dei 10 fattori trascrizionali (AP1, ATF3, EBF1, ERG, ETS1, ETV4, FRA2, RUNX1, RUNX2, ELK4) (Fig. 10) coinvolti nel legame dei siti necessari per la progressione tumorale. ChIP-Atlas è un portale *web* in cui sono raccolti tutti i collegamenti ai dati grezzi di ChIPseq pubblicati.

Nella selezione dei dati da utilizzare si è tenuto in considerazione: l'anno di pubblicazione, la qualità del sequenziamento e la tipologia di linea cellulare in cui è stato compiuto il sequenziamento, e sono state scelte preferenzialmente linee cellulari leucemiche qualora disponibili. Sono stati nuovamente analizzati 32 esperimenti di ChIPseq così suddivisi: 8 di AP-1, 4 di ATF3, 2 di EBF1, 4 di ELK4, 1 di ERG, 2 di ETS1, 3 di ETV4, 2 di FRA2, 5 di RUNX1 e 2 di RUNX2.

Nonostante l'analisi delle ChIP-seq, ottenute come precedentemente descritto, consenta di identificare, per ciascuno dei 117 enhancer, il legame di almeno un fattore trascrizionale (Fig. 14), alcuni presentano legami di più fattori trascrizionali. Il TF ERG, rilevato in 77 dei 117 *enhancer* (n=77), mostra un grande coinvolgimento nel legame delle sequenze degli *enhancer*, infatti il numero di siti che lo contengono è secondo solamente a RUNX2 (n=96), di cui si conosce già il coinvolgimento in altri tipi di leucemia

e pertanto potrebbe essere aggiunto ai potenziali *target* coinvolti nella progressione e, in questo caso, nella regolazione dell'espressione dei geni *target*. Oltre al numero di regioni legate dai singoli TF è interessante notare come il segnale di molteplici fattori trascrizionali sia stato rilevato nella stessa regione genomica. La presenza di più fattori trascrizionali legati alla medesima regione genomica potrebbe essere ulteriormente indagata tramite esperimenti di ChIPseq per ciascun fattore trascrizionale per definirne la reale presenza nella B-ALL. Inoltre, si potrebbero condurre esperimenti di RNA *interference* per ciascuno dei fattori trascrizionali per determinare se uno dei 9 TF svolga il ruolo di *pioneer factor*, ovvero un fattore trascrizionale in grado di legare la cromatina condensata e permettere l'apertura dei .

L'unico fattore trascrizionale non rappresentato è ELK4 di cui non sono stati trovati segnali di legame all'interno delle regioni selezionate.

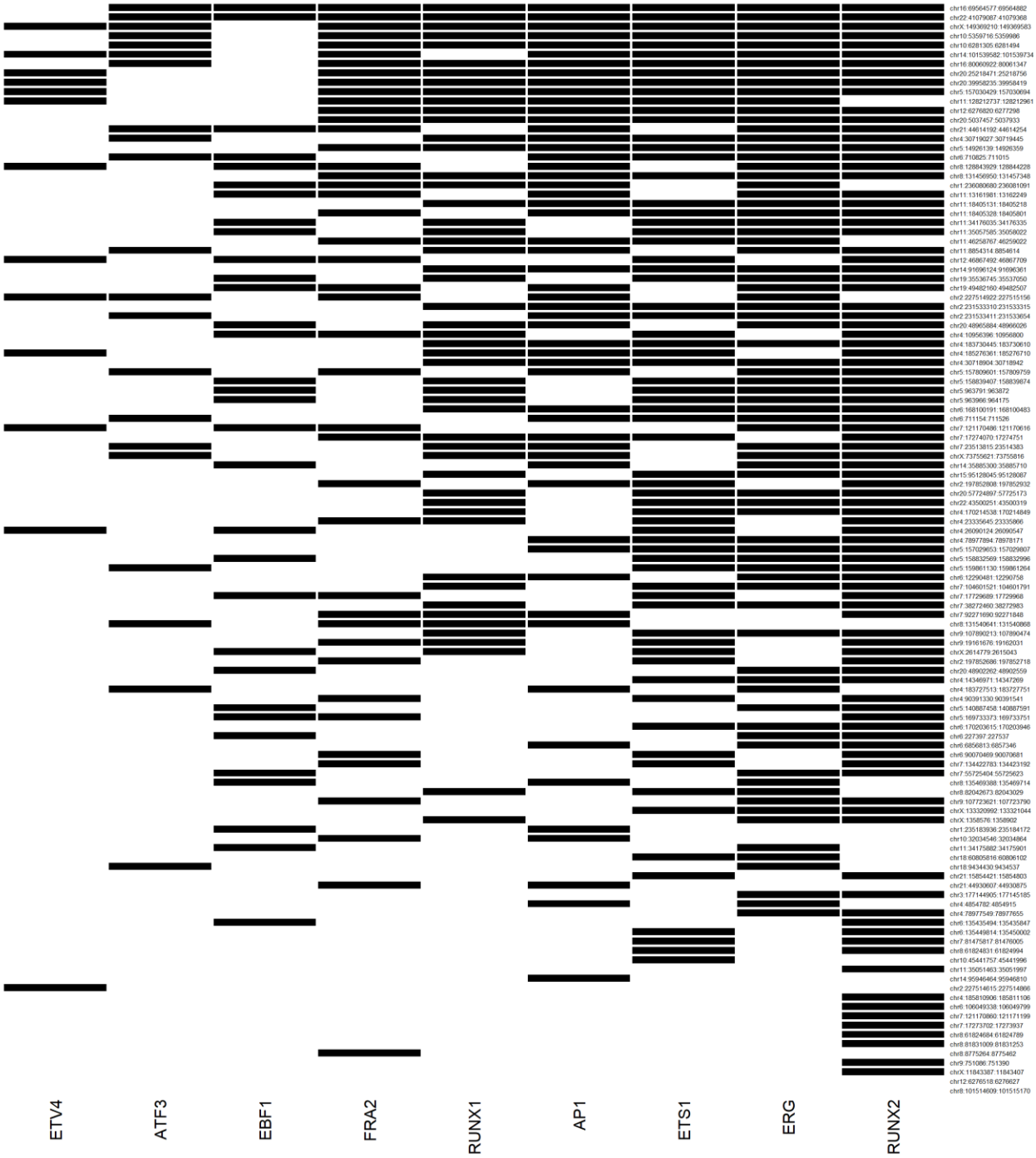


Figura 14: Sono mostrati i 117 enhancer identificati, sulle righe, e i fattori trascrizionali, sulle colonne La presenza della barra nera indica il rilevamento del segnale del fattore trascrizionale all'interno di quella regione. I fattori trascrizionali riportati sono 9, poiché ELK4 non ha rilevato segnali di legame con nessuno dei 117 enhancer.

4.9 Identificazione dei geni *target*

Una volta identificati e caratterizzati i 117 *enhancer* è stato fondamentale identificare i geni *target* di ciascuno di essi. Questa procedura è stata possibile tramite l'utilizzo di due metodiche complementari. Innanzitutto, sono stati assegnati i geni *target* testati all'interno di TCeA. Per tutti gli *enhancer*, il cui gene *target* non era presente in TCeA, è stata utilizzata la metodica del *nearest gene*. Questa metodica risulta tutt'ora come una delle più precise nell'assegnare un *enhancer* al proprio gene *target*.(Nasser et al., 2021). Per ogni *enhancer* è stato deciso di assegnare 1 solo gene alla luce delle più recenti pubblicazioni che identificano in una mediana di 1.5 il numero di geni regolati da un singolo *enhancer* (Boix et al., 2021).

L'assegnazione del gene *target* è poi proseguita attraverso un ulteriore processo di validazione. Infatti, a partire dai dati di *Promoter-Capture* prodotti in laboratorio è stata compiuto un controllo sui geni assegnati a ciascun *enhancer*.

L'unione delle varie metodiche ha consentito, infine, l'assegnazione di tutti i geni *target* ai 117 *enhancer* selezionati, in alcuni casi mostrando anche l'associazione di più *enhancer* allo stesso gene , per un totale di 100 geni assegnati ai 117 enhancer.

4.10 Analisi dell'espressione dei geni *target*

Una volta identificati i geni *target* per i 117 *enhancer* è stato interessante indagare quanto questi geni sono espressi in pazienti affetti da B-ALL pediatrica. I dati di RNAseq sono stati ottenuti dallo studio Therapeutically Applicable Research to Generate Effective Treatments (TARGET). Lo studio TARGET rappresenta la più ampia collezione di trascrittomi di B-ALL Questo *dataset* presenza di campioni longitudinali dello stesso paziente. Relativamente ai dati di RNAseq utilizzati non sono presenti campioni in tutte le fasi della patologia ma solo relativi ad esordio e recidiva. I dati analizzati riguardano l'espressione di 85 dei 100 geni *target*,

precedentemente individuati, in 96 campioni ottenuti da 48 pazienti. L'integrazione dei 100 geni *target* con dati ottenuti da TARGET non include geni di cui non è stata rilevata l'espressione nei dati scaricati.

L'*heatmap* prodotta con i dati *matched* di esordi e recidive, mostra come non ci sia alterazione tra i due stadi. Questo risultato si aggiunge alla nostra precedente osservazione in cui gli *enhancer* non presentano alterazione nell'accessibilità cromatinica tra gli stadi di esordio e recidiva. Sebbene siano presenti delle regioni dell'*heatmap* con arricchimento significativamente più elevato, queste non mostrano consistente *pattern* di alterazione dell'espressione. Nella maggior parte dei casi i campioni che mostrano un gruppo di geni più espresso rispetto al resto della popolazione sono sequenziamenti derivanti dal medesimo paziente nei due stadi della patologia. Pertanto è possibile ricondurre le alterazioni a delle caratteristiche più paziente specifiche che dovute a cambiamenti dell'espressione di geni tra i due stati della patologia. Risulta comunque evidente come tra esordio e recidive la maggior parte dei geni non mostri un cambiamento netto di espressione, ma tutti si attestano intono ad uno Z-score pari a 0. Una delle limitazioni riscontrate è l'assenza dei dati di RNA-seq derivanti da remissioni. Osservare l'espressione anche in un campione fenotipicamente sano permetterebbe di dare maggiore robustezza all'analisi e, inoltre, conferirebbe un'indicazione ulteriore sulla regolazione che un *enhancer* svolge sul suo gene *target*.

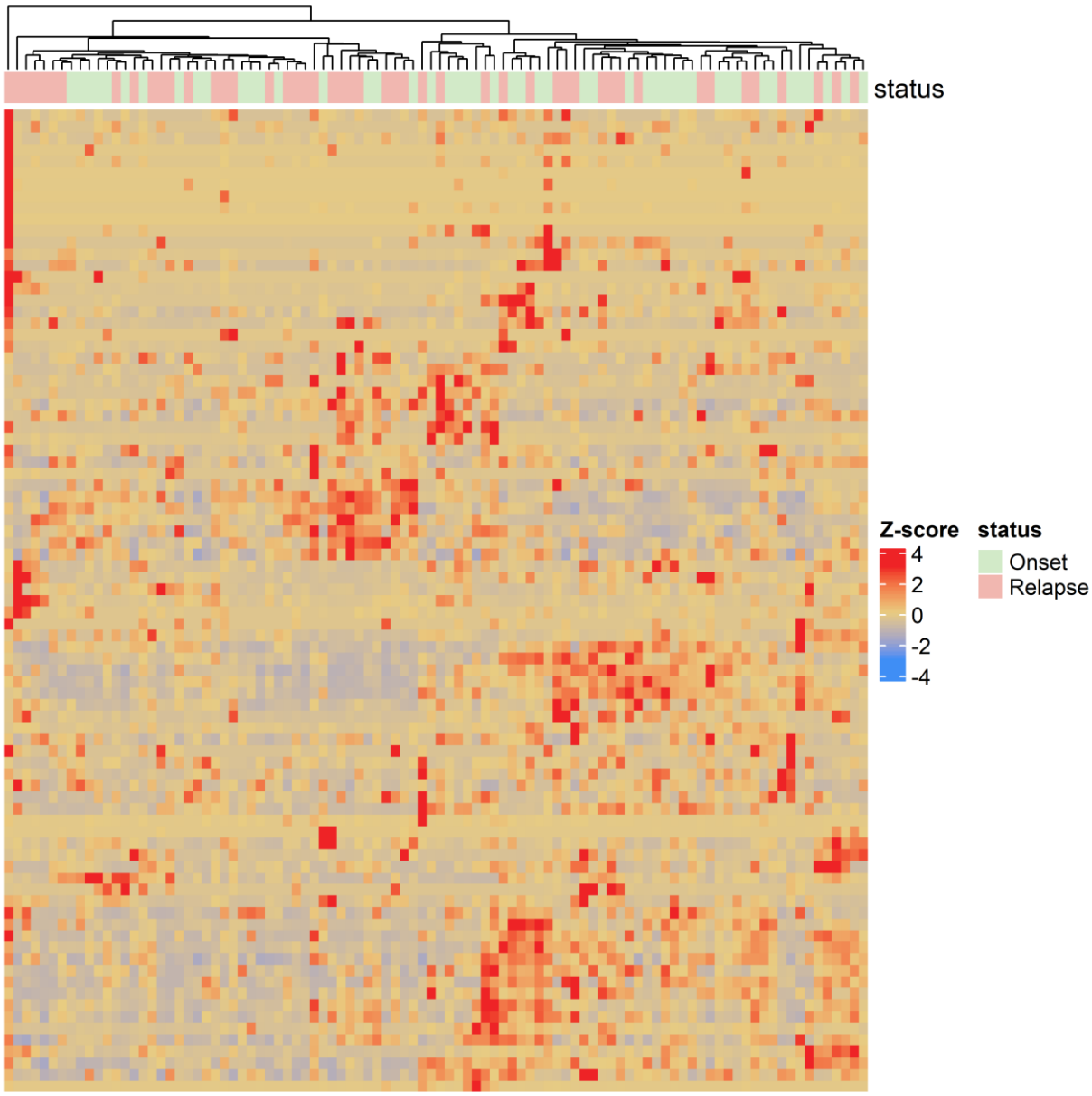


Figura 14:. L'heatmap mostra lo Z-score dei dati di espressione ottenuti da TARGET riguardati 96 campioni ottenuti da 48 pazienti.

4.11 Identificazione degli eRNA nei pazienti

L'osservazione dell'attivazione degli *enhancer* è stata compiuta, come precedentemente descritto integrando i profili ATAC-seq della nostra coorte di pazienti, i dati TCeA ed i profili di H3K27ac ottenuti da ENCODE. In aggiunta, abbiamo valutato la della produzione di eRNA in una selezione dei campioni di paziente analizzati di cui avevamo a disposizione l'RNA. Questo tipo di analisi è di importanza rilevante per definire effettivamente quali siano gli *enhancer* produttivi nei campioni, ma allo stesso tempo più complicata di una RNA-seq canonica in quanto i livelli di espressione degli eRNA sono sensibilmente inferiori agli mRNA prodotti dai geni. Questo tipo di analisi è stata eseguita su sequenziamenti RNA-seq totale di 2 pazienti sani, 1 paziente afferente alle remissioni e 3 campioni di esordio. I risultati ottenuti mostrano una sovrapposizione con le osservazioni precedentemente effettuate (Fig. 15). Non sono stati rilevati eRNA prodotti da tutti i 117 *enhancer* coinvolti nella progressione, probabilmente dovuta alle basse quantità di eRNA prodotto da queste regioni e dal rapido *tournover* degli eRNA (Larsson et al., 2019).

Ciò che risulta evidente è la netta clusterizzazione dei campioni, infatti i 3 esordi mostrano una grande somiglianza nell'espressione come i campioni normali e la remissione che ancora una volta mostrano estrema similarità.

Analizzando i *pattern* di espressione degli eRNA si possono evidenziare 3 diversi comportamenti: *enhancer* che mantengono livelli bassi di espressione in tutte le condizioni, *enhancer* che aumentano l'espressione negli esordi e infine quelli che hanno un'espressione variabile. Il fatto di non aver individuato *enhancer* che riducono i livelli di espressione negli esordi rispetto a normali e remissioni evidenzia il ruolo determinante della nostra metodica di selezione incentrata sulla caratterizzazione degli eventi di sviluppo della malattia.

Si evidenzia, un campione normale che mostra un' attivazione di un gruppo di *enhancer* non concorde agli altri campioni fenotipicamente sani. Sebbene

non si sia certi del motivo, potrebbe essere dovuto ad una differente malattia che affligge il paziente, che risulta però negativo alla diagnosi di B-ALL.

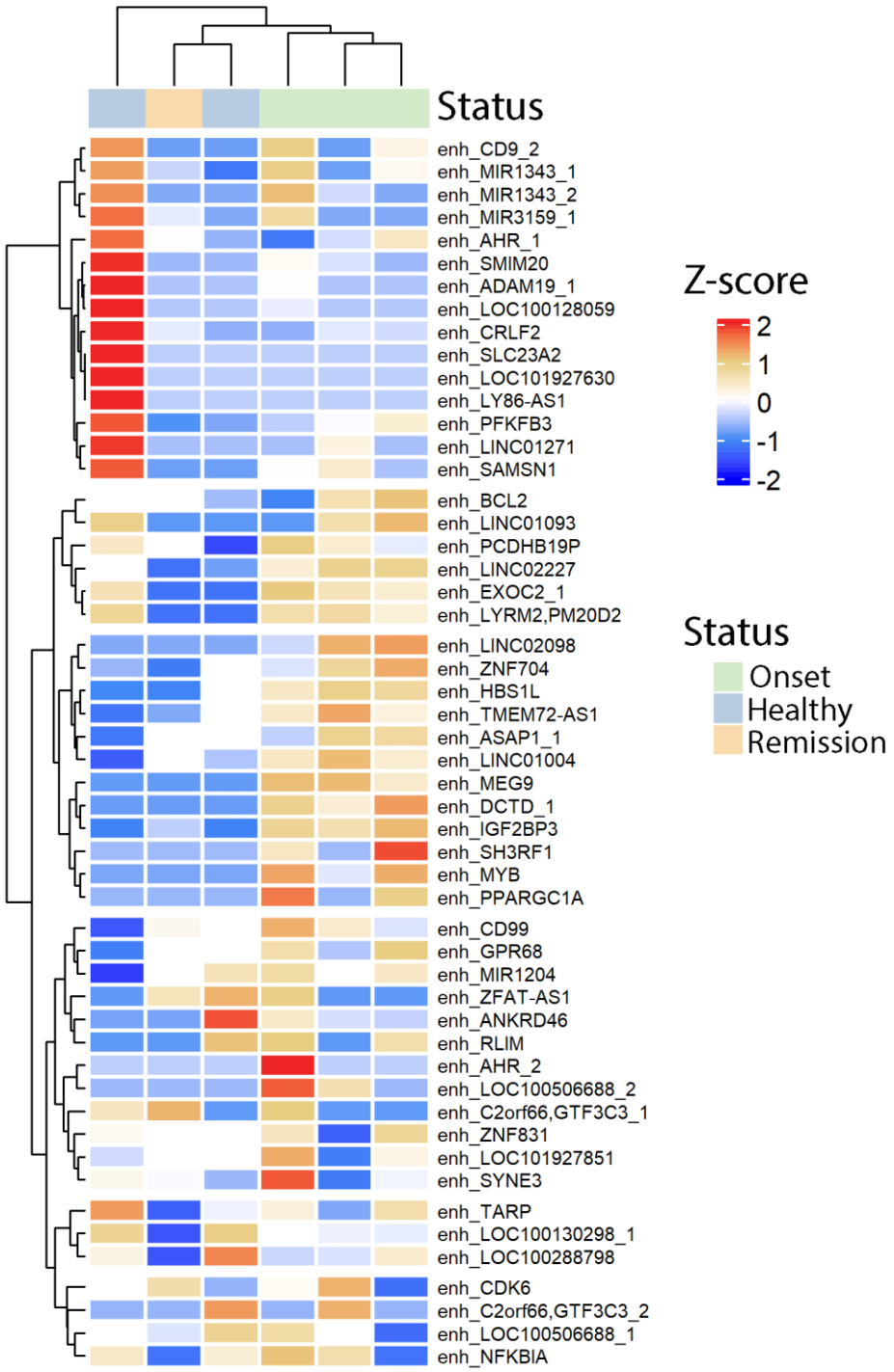


Figura 15: In 2 campioni normali, 1 remissione e 3 esordi è stato analizzato l'mRNA totale. L'heatmap mostra i risultati dell'RNA-seq totale condotto su 2 campioni normlai, 1 remissione e 3 esordi. I valori di espressione sono relativi allo Z-score dei TMM degli enahncer identificati. Non è stata identificata l'espressione di tutti per tutti i 117 enhancer, per questo è rappresentato un numero minore di righe. Il nome di ciascuna riga identifica l'enhancer in base al gene target.

4.12 Identificazione di potenziali *target* della patologia

Il passaggio finale di questo studio è stato volto ad identificare regioni regolatorie come possibili *target* terapeutici della patologia.
L'identificazione dei potenziali *target* di progressione di B-ALL è stata compiuta integrando i dati finora analizzati e considerando quali *enahncer* target mostrassero un maggiore interesse nei dati dei pazienti, nei database pubblici e infine nella linea cellulare di LAL-B, che rappresentano il modello sperimentale prescelto per valutare l'inibizione del *target* individuato. I dati utilizzati per l'identificazione dei potenziali target terapeutici sono stati stati ottenuti da*repository* pubbliche come: TARGET, HeRA, DepMap ed ENCODE, oltre quelli prodotti in laboratorio di RNAseq, ATACseq, ChiPseq pubblicate e Hi-C *promoter-capture*. Quest'analisi integrata ha permesso di identificare gli *enhancer* di: DCTD, MYB e BCL2 come potenziali *target* della progressione della patologia.

4.12.1 DCTD

La proteina codificata dal gene DCTD è un enzima responsabile della catalisi della deaminazione da dCMP a dUMP, quest'ultimo substrato per l'enzima timidilato sintasi, coinvolto nella reazione di sintesi del dTMP.
L'analisi del portale DepMap (Meyers et al., 2017) evidenzia univocamente una specificità dell'inibizione della vitalità cellulare in linee di ALL se si effettua un *knockout* del gene DCTD tramite la tecnologia CRISPR-Cas

(Fig. 16). Questo importante dato ha pertanto indirizzato la scelta di DCTD come potenziale *target* terapeutico.

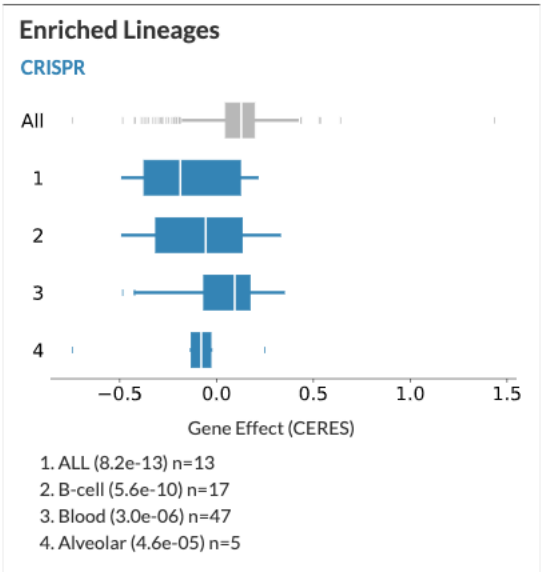


Figura 16: Immagine dal portale DepMap in cui è rappresentata la sensibilità delle linee cellulari di ALL al knockout di DCTD.

L'obiettivo non è stato tuttavia identificare un gene *target* ma l'*enhancer* coinvolto nella regolazione della sua espressione. Per questo motivo è stato affiancato all'identificazione del gene *target* la conferma di formazione del *loop* tra la regione identificata e il promotore di DCTD tramite l'analisi dei dati di Hi-C *promoter-capture* (Metodi 3.13) ottenuti dal sequenziamento di un replicato della linea cellulare LAL-B., identificando un contatto diretto tra le due regioni (Fig. 17).

Inoltre, come è possibile osservare dall'analisi del cistroma dei 117 *enhancer* condotta nel paragrafo 4.7 (Fig. 14), la regione identificata come *enhancer* di DCTD, con coordinate chr4: 183730445-183730610, contiene sequenze di *binding* di RUNX2, ERG, AP-1 e RUNX2 identificati in linee leucemiche (Capitolo 4.7), sottolineando l'attivazione di questa regione anche in altri tipi cellulari oltre che in B-ALL.

Un'ulteriore prova del coinvolgimento dell'*enahncer* di DCTD nella progressione della patologia proviene dal *boxplot* del rapporto tra segnale ottenuto dall'*enhancer* e quello del promotore di DCTD (Fig. 18).

Il rapporto tra segnale del promotore e segnale dell'enhancer può essere considerato come un rapporto tra una costante e una variabile. Infatti, il DCTD è un gene *housekeeping*, pertanto presente in tutte le cellule, anche quelle sane. È però possibile identificare una netta diminuzione del rapporto negli esordi e nelle recidive, mostrando una maggiore apertura dell'*enhancer* in queste due condizioni. Tale risultato mette in evidenza come l'*enhancer* di DCTD sia un fattore importante sia nell'insorgenza che nella progressione tumorale, candidando la regione *cis*-regolatoria come possibile *target* cellulare. Inoltre, dal momento che l'*enhancer* mostra una maggiore apertura solo in esordi e recidive può essere considerato un target specifico per le cellule cancerose.

Figura 17. L'immagine mostra il risutalto dell'analisi del promotore-capture della regione chr4:183,705,346-183,841,329. Il loop, arco nero, mostra l'interazione della regione di sinistra, enhancer di DCTD, con il promotore del gene DCTD. Sono inoltre mostrati 3 profili ATAC-seq di pazienti a cui è stato diagnosticata B-ALL. Le analisi di promoter-capture sono state eseguite sulla linea cellulare LAL-B.

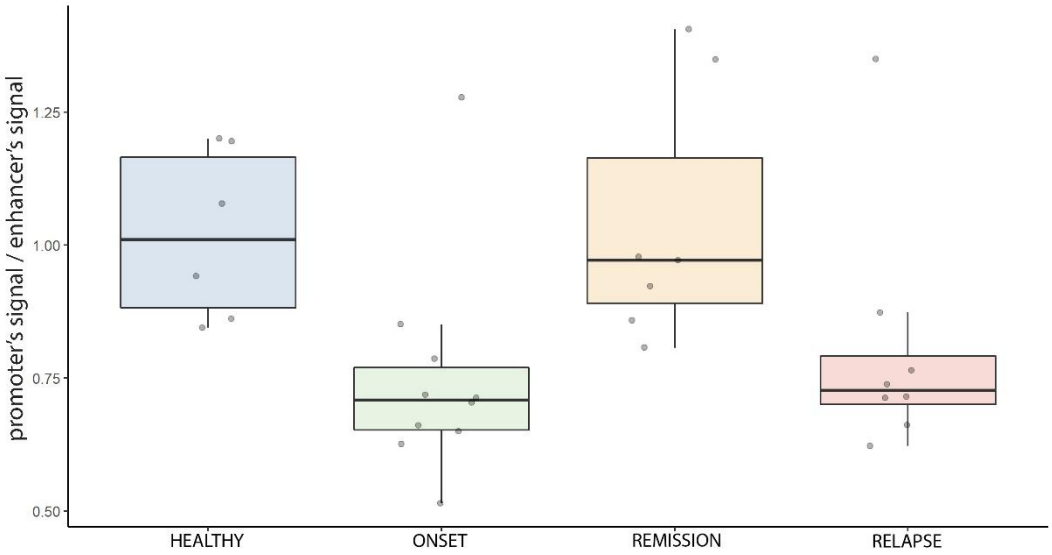


Figura 18: Il boxplot mostra, per ogni status, il rapporto tra il segnale del promotore e il segnale dell'enhancer ottenuti dalle analisi delle ATAC-seq dei pazienti, ogni punto raprresenta un paziente.

4.12.2 BCL2

La selezione per BCL2 è avvenuta principalmente considerando lo stretto legame tra questo gene e i tumori ematologici come la leucemie mieloide acuta e il mieloma multiplo.

Il gene BCL2 codifica per una proteina della membrana esterna mitocondriale che ha il compito di bloccare la morte cellulare per apoptosi, questo è uno dei motivi per cui la proteina risulta *overespressa* in un grande numero di tumori. Sebbene sia ben noto il ruolo svolto dalla proteina, l'identificazione degli *enhancer* in grado di regolarne l'espressione ancora è incompleta. La regione identificata nel lavoro, con coordinate *chr18: 60805817-60806103*, presenta molteplici informazioni che suggeriscono un suo diretto coinvolgimento nella regolazione dei livelli di BCL2 all'interno della cellula. In primo luogo, come mostrato precedentemente (Fig. 15), anche per questa regione è stata individuala la produzione di eRNA nei campioni ottenuti da paziente, con un incremento negli esordi rispetto a sani

e recidive. Inoltre in un campione sano e uno di remissione non è stato possibile individuare neanche un livello basale di espressione. Inoltre le analisi di Promoter-Capture evidenziano una connessione tra la regione identificata nelle ATACseq di pazienti e il promotore di BCL2 (Fig. 19).
Una volta saggiato il contatto tra il sito identificato ed il promotore di BCL2 le analisi si sono concentrate sull'andamento dell'apertura durante il decorso della patologia. Tramite l'utilizzo del RI è stato valutato l'andamento della clonalità che segue l'andamento desiderato, ovvero è possibile individuare un incremento della clonalità nei campioni di esordio e recidiva, mentre nei normali e nelle remissioni la clonalità è più bassa. Un'ulteriore valutazione si potrebbe compiere sul numero di normali e remissioni che presentano il picco aperto, infatti nei normali è stato osservato solo in 4 campioni su 6 mentre nelle remissioni in 6 su 7, questo incremento potrebbe suggerire, in un numero ristretto di paziente, l'apertura del sito a livelli di clonalità alti e poi mantenuto durante la remissione (Fig. 20).

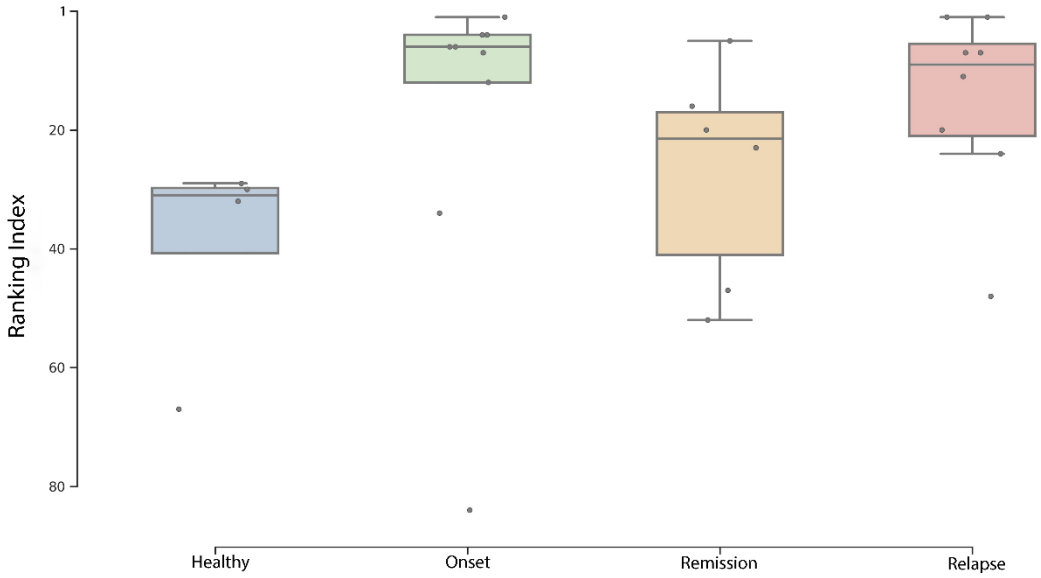


Figura 20: Sono graficati i Ranking Index assegnati a ciascun paziente per l'enhancer di BCL2 diviso per status della patologia.

A differenza di DCTD l'andamento del rapporto tra segnale dell'*enhancer* e del promotore è in costante diminuzione, suggerendo una continua apertura

dell'*enhancer* durante la progressione tumorale (Fig.20 A). Analizzando però i segnali del promotore e dell'*enhancer* nei diversi stadi (Fig. 20 B) è possibile osservare come ci sia un aumento del segnale dell'*enhancer* negli esordi rispetto ai campioni sani, mentre il promotore rimane pressoché costante, come ipotizzato già per DCTD. È particolare invece il comportamento che ha il segnale del promotore nelle remissioni e infine nelle recidive. Infatti l'*enhancer* rimane costate, ma questa volta è il segnale del promotore che diminuisce. Questo potrebbe suggerire il coinvolgimento di nuove regioni che hanno effetto contrario rispetto all'enhancer individuato, che comunque mostra un notevole coinvolgimento nell'esordio della patologia.

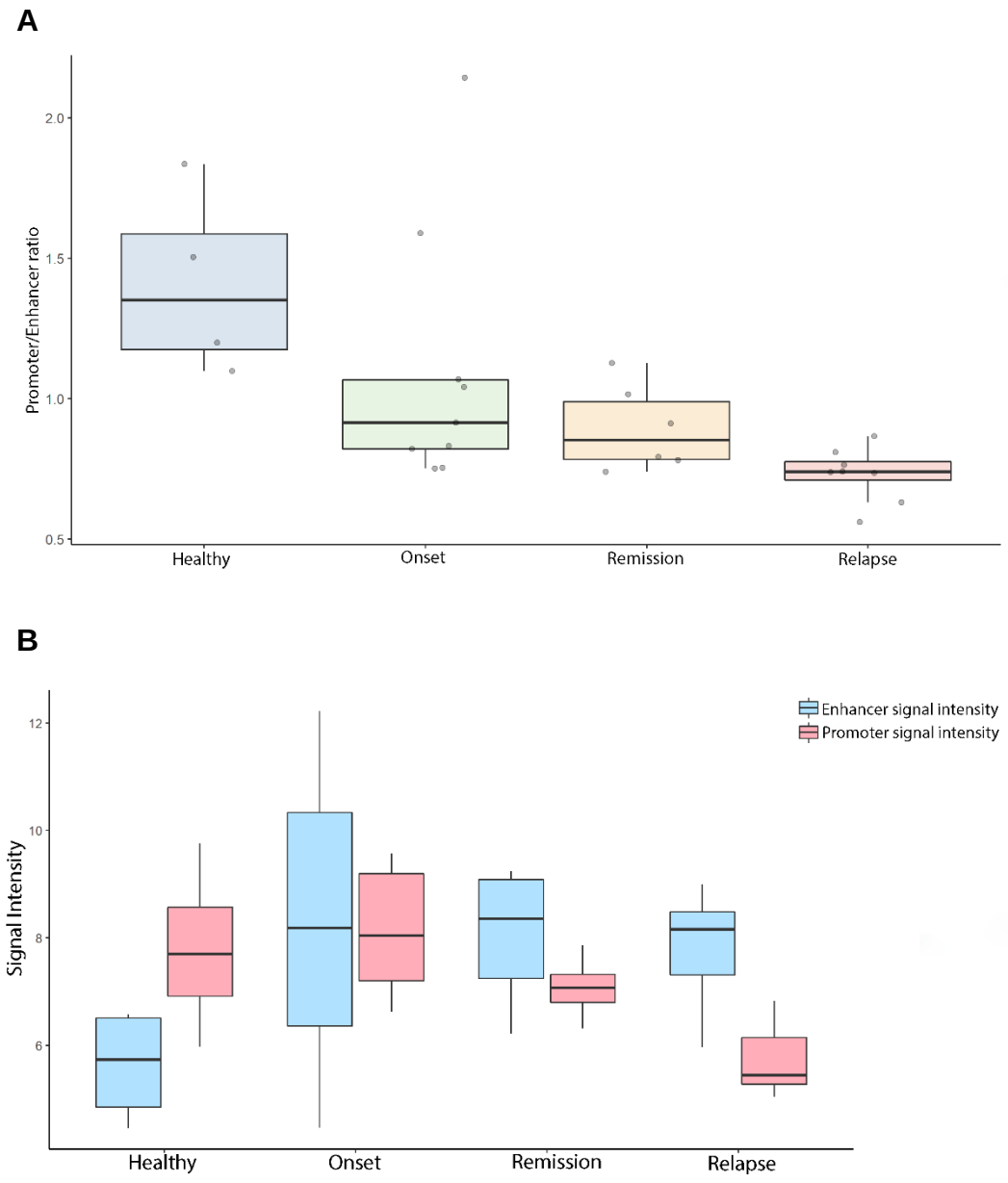


Figura 20: 20A per ogni status è rapresentato il rapporto tra segnale del promotore e segnale dell'enhancer di BCL2 ottenuti dalle analisi delle ATAC-seq dei pazienti, ogni punto raprresenta un paziente per ogni paziente. 20B sono rappresentati per ciascuno status i segnali separati di promotore ed enhancer.

4.12.3 MYB

Il gene MYB codifica per una proteina in grado di legare il DNA con funzioni di fattore trascrizionale, ha un ruolo principale nella regolazione dell'ematopoiesi. Dato il largo coinvolgimento di MYB nell'insorgenza e progressione dei tumori ematologici, l'identificazione di un *enhancer* in grado di regolare la sua espressione potrebbe migliorare le terapie farmacologiche. L'*enhancer* identificato, con coordinate *chr18-:60805817-60806103*, fa parte dei *cluster* che mostrano una modulazione positiva dell'apertura durante l'insorgenza e la recidiva. L'individuazione dell'*enhancer* di MYB come possibile *target* prende sempre in considerazione i risultati ottenuti dalle molteplici analisi. Infatti l'analisi del H3K27Ac di ENCODE (Fig. 12) mostra come l'*enhancer* non sia attivo in molte linee cellulari tumorali, questo potrebbe però essere dovuto alla non espressione di questo specifico *enhancer* in tutti i tessuti maligni ed identificandolo come specifico per la B-ALL. A questa osservazione si aggiunge anche la presenza di un *loop* evidente tra la regione individuata ed il promotore di MYB. Anche se i risultati di *Promoter-Capture* sembrerebbero non confermare l'assegnazione del gene *target* secondo la strategia del gene più vicino, possiamo ipotizzare che in questa regione ci sia un rimodellamento cromatinico che porta all'avvicinamento degli *enahncer* ai promotore di H1SBL e MYB. Pertanto è possibile evidenziare come tutta la regione possa essere determinante nella determinazione della struttura tridimensionale dell'intero tratto genomico. Inoltre anche per l'*enhancer* di MYB è evidenziata una maggiore produzione di eRNA negli esordi rispetto ai sani e alla remissione (Fig. 15) dimostrando che il picco di apertura cromatinica osservato con ATACseq è corrisposto ad una maggiore attività trascrizionale. Anche per MYB, come per BCL2, il comportamento del rapporto tra segnale del promotore e segnale dell'*enhancer* non rispecchia l'andamento di DCTD desiderato (Fig. 21A). Infatti sembrerebbe quasi avere un andamento inverso a quello atteso, con una maggiore apertura dell'*enhancer* in sani e remissioni. Graficando in

maniera separata i segnali del promotore e dell'*enhancer* è però evidente come ci sia un aumento del segnale dell'*enahncer* negli esordi rispetto ai sani, con una riduzione nella remissione che comporta anche una diminuzione cospicua del promotore. Più particolare è il comportamento delle recidive, in cui l'abbassamento del segnale dell'*enhancer* non comporta una riduzione del segnale del promotore.

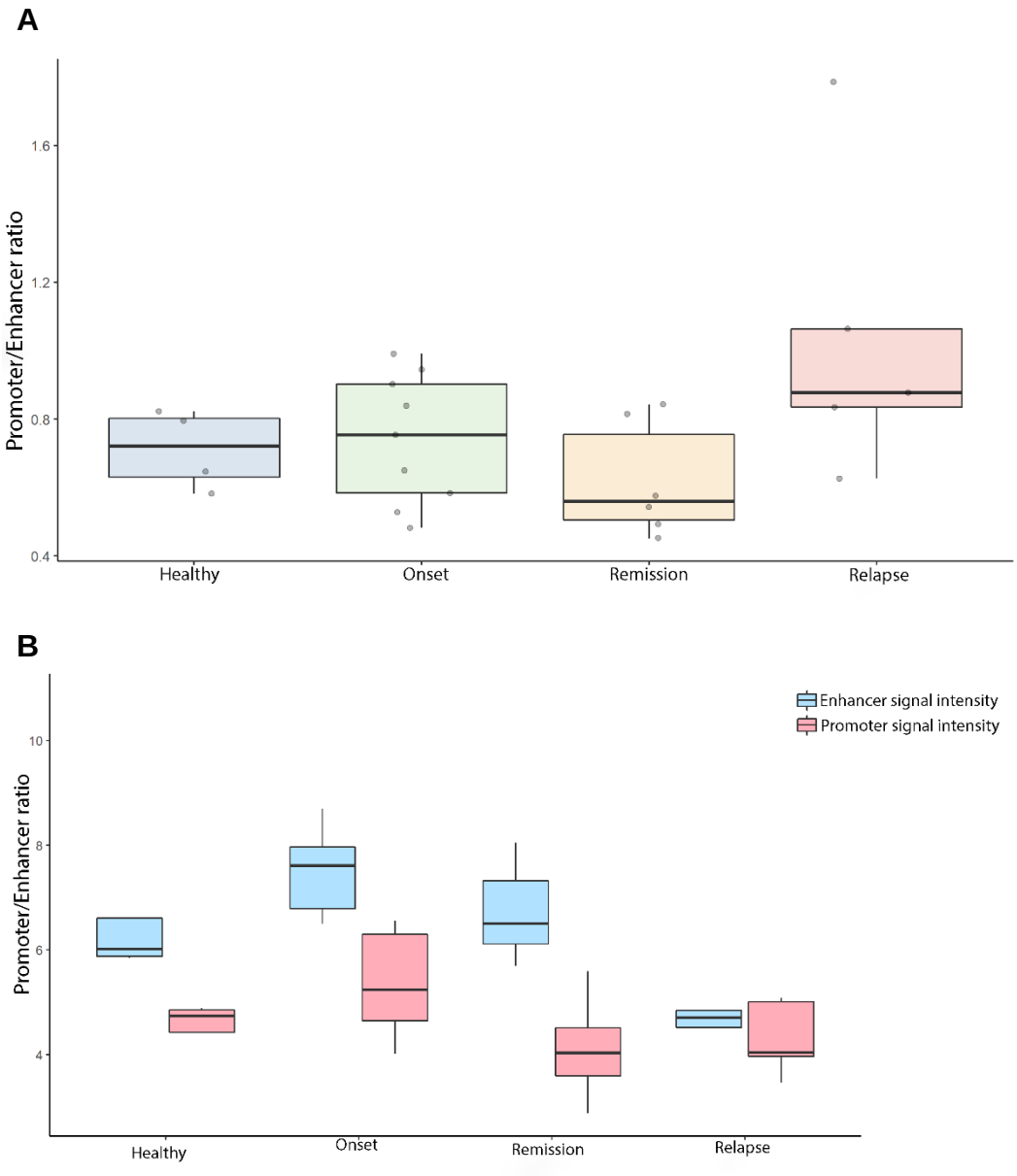


Figura 20: 20A per ogni status è rapresentato il rapporto tra segnale del promotore e segnale dell'enhancer di MYB ottenuti dalle analisi delle ATAC-seq dei pazienti, ogni punto raprresenta un paziente per ogni paziente. 20B sono rappresentati per ciascuno status i segnali separati di promotore ed enhancer,

4.13 RNA *interference* dell'*enhancer* di DCTD

Successivamente alla selezione e caratterizzazione computazionale dell'*enhancer* di DCTD è indispensabile validare tale risultato con esperimenti compiuti su linee cellulari di LAL-B e NALM-6. Le NALM-6 sono una linea cellulare derivante da linfociti di un paziente affetto da ALL recidivante.

Ciò che si vuole valutare è il ruolo svolto dall'eRNA dell'*enhancer* di DCTD nella regolazione del suo gene *target*, valutandone i livelli di: proliferazione cellulare, mRNA di DCTD e livello della proteina DCTD.

Per poter eseguire questa tipologia di esperimento è stata utilizzata la tecnologia dell'RNA *interference* (RNAi), tramite il disegno e la sintesi di frammenti di RNA complementari (siRNA) alla regione identificata. Una volta assorbiti dalla cellula, i siRNA hanno affinità per il trascritto *target*, in questo caso l'eRNA, prodotto. L'appaiamento tra il siRNA e l'RNA target darà vita ad una doppia elica di RNA che sarà riconosciuta da sistemi cellulari che ne guideranno la degradazione. Secondo questo procedimento si dovrebbe osservare una netta diminuzione degli eRNA prodotti dall'*enhancer* e successivamente sarà possibile valutarne le implicazioni.

I risultati dell'esperimento mostrano una riduzione di circa il 30% della proliferazione cellulare tramite l'utilizzo dell'eRNA di DCTD come target del siRNA. Si osserva anche una riduzione anche nella produzione dell' mRNA di DCTD sempre intorno al 30% rispetto al controllo (Fig. 21 A), confermando il ruolo dell'enhancer individuato come regolatore per la produzione della proteina.

I risultati dell'interferito dell'eRNA sono anche confermati dall'interferito dell'eRNA del gene di DCTD (Fig. 21 C) che evidenziano anche qui una diminuzione nella replicazione cellulare intorno al 30%. Per consolidare questi risultati è stato condotto anche un esperimento di RNAi diretto però contro l'mRNA di DCTD. I risultati suggeriscono ancora di più il coinvolgimento dell'*enhancer* nella regolazione di DCTD dal momento che

è, anche qui, osservata una riduzione della proliferazione cellulari del 20% sia il LAL-B che in NALM-6 (Fig. 22).

Per un'ulteriore conferma è stata anche valutata la quantità di proteina DCTD prodotta dalla cellula, anche per i livelli proteici si osserva una riduzione dopo il trattamento con il sieRNA.

Alla luce di ciò che è stato appena descritto si può ipotizzare il ruolo svolto dall'*enhancer* di DCTD nella regolazione dell'espressione genica di DCTD. Sebbene siano molte le possibili funzioni svolte da un eRNA, dati i risultati ottenuti, è plausibile ipotizzare che il ruolo del trascritto prodotto dall'*enhancer* sia implicato nella formazione del *loop* tra *enhancer* e promotore (Fitz et al., 2020). Per determinare in maniera più approfondita il meccanismo molecolare in cui è coinvolto l'eRNA sono però necessari ulteriori approfondimenti, come definire se, in alternativa, il ruolo sia svolto a livello di regolazione della polimerasi modulandone il passaggio da processività a pausa.

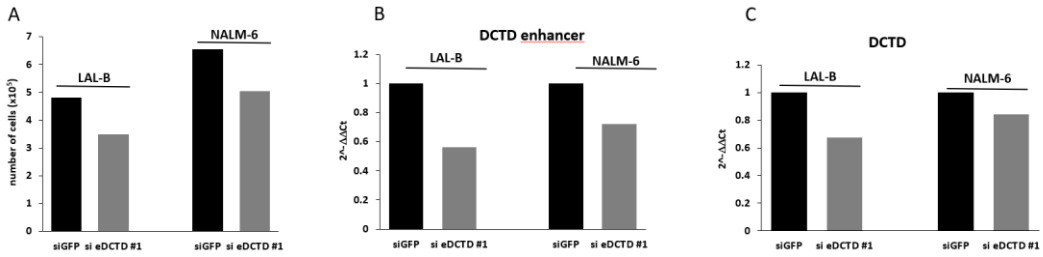


Figura 21: 21A Il sieRNA mostra una riduzione di circa il 30% in LAL-B e NALM-6 rispetto sl controllo (siGFP). 21B Mostra il contrllo della riduzione dell'eRNA bersaglio. 21C Evidenzia la diminuzione, anche in questo caso di circa il 30%, dei livelli di mRNA di DCTD.

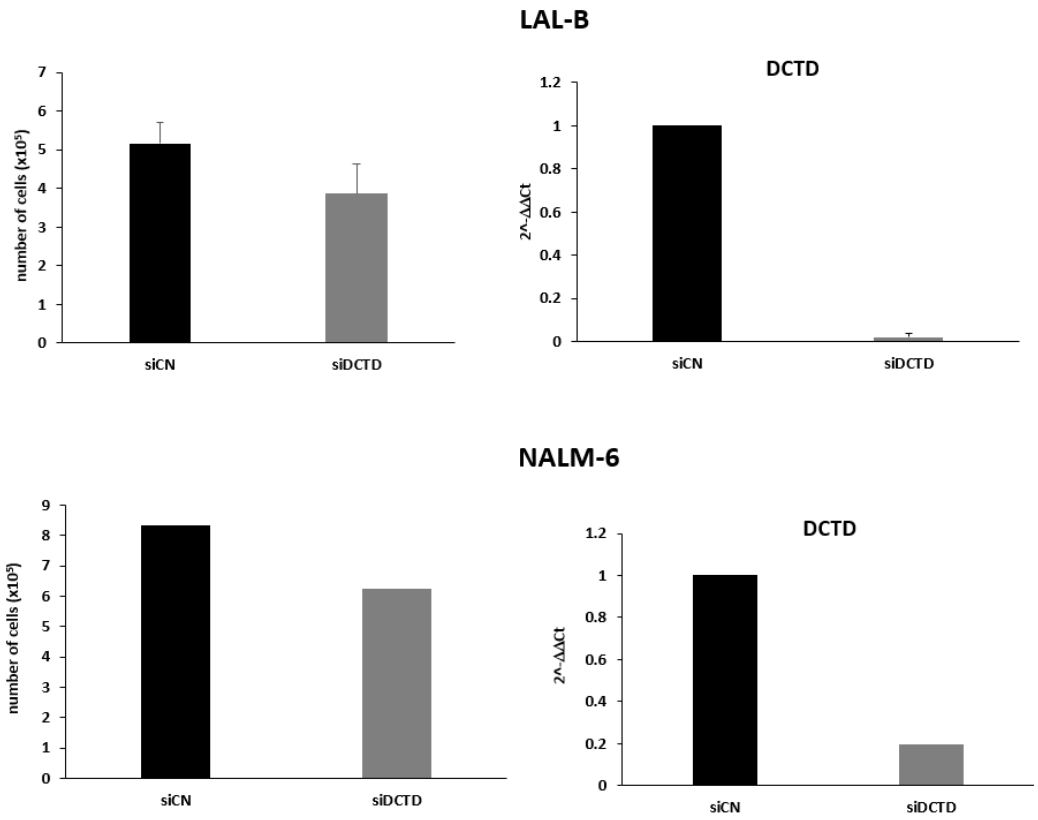


Figura 22: Risultati degli esperimenti di RNAi con target l'mRNA di DCTD, sulla sinistra sono riportati gli effetti sul numero di cellule, emntre sulla destra è evidenziata la riduzione sui livelli dell'mRNA di DCTD.

4.14 Discussione

La Leucemia Linfoblastica Acuta di tipo B (B-ALL) rappresenta il tumore più comune in età pediatrica, è infatti responsabile di circa il 25% dei tumori pediatrici e circa dell'80% delle leucemie pediatriche (Stanulla and Schrappe, 2009). Data l'elevata incidenza, la ricerca si è fortemente concentrata sull'identificazione di nuovi possibili *target* terapeutici. Numerosi studi hanno portato ad una fine caratterizzazione delle alterazioni genetiche alla base della patologia. Gli avanzamenti nella definizione delle alterazioni genetiche ha permesso l'identificazione di molteplici terapie con un alto tasso di remissione completa (80%-90%). Ciò che rimane da definire sono le alterazioni che comportano la resistenza ai farmaci ed il conseguente sviluppo di recidive in circa il 15-20% dei pazienti. La principale problematica è la definizione di un piano terapeutico in grado di

aumentare la sopravvivenza in caso di recidiva, che attualmente è compreso tra il 15% e il 50% (Hunger and Mullighan, 2015).

Nonostante siano stati compiuti grandi passi in avanti nella comprensione delle alterazioni genetiche coinvolte nella progressione della B-ALL, non è presente, ad oggi, una risposta farmacologica ai casi di recidiva tumorale. A differenza delle alterazioni genetiche, il coinvolgimento dell'epigenetica nel decorso della B-ALL non è stato oggetto di approfondimento. A tal proposito, il progetto di tesi mira all'identificazione di regioni regolatorie fondamentali per la progressione tumorale. L'analisi è stata basata su 35 campioni longitudinali di pazienti pediatrici in trattamento presso l'Ospedale Bambino Gesù di Roma. I profili di accessibilità ottenuti tramite ATACseq sono poi stati sottoposti a successive analisi bioinformatiche tramite l'utilizzo di *pipeline standard* combinate a nuove metodiche di analisi.

Una volta ottenuti i picchi di accessibilità da ciascun campione sono stati assegnati a ciascun picco due indici: *Sharing index* e *Ranking index*, seguendo la strategia applicata nel lavoro di Patten and Corleone et al. (Patten and Corleone et al., 2018). Questi due indici hanno permesso di compiere una stratificazione dei 140.000 picchi iniziali tenendo in considerazione la clonalità di un picco all'interno del campione sequenziato (*Ranking Index*) e la penetranza di un picco tra i pazienti di uno status (*Sharing Index*). La stratificazione ha permesso l'individuazione di 11.000 picchi che rispecchiano tre condizioni biologiche fondamentali nella progressione tumorali, che sono: picchi con alta penetranza negli esordi e bassa penetranza nei sani, picchi che mostrano aumento di clonalità negli esordi rispetto ai sani, picchi che mostrano un aumento di clonalità nelle recidive rispetto alle remissioni. I picchi così selezionati sono sottoclonali negli esordi, la clonalità poi aumenta con l'insorgenza della malattia per poi diminuire post-trattamento e infine aumentare nuovamente nelle recidive (Fig. 7). La *clusterizzazione* (Fig. 9), compiuta con lo *z-score* della conta grezza delle *reads* per ciascuno degli 11.000 picchi, ha evidenziato come solo i cluster C1 e C2, contenenti 6.000 picchi, seguano l'andamento di

interesse. L'analisi dei motivi delle regioni appartenenti ai quattro cluster ha poi permesso di identificare i 10 fattori trascrizionali (AP1, ATF3, EBF1, ERG, ETS1, ETV4, FRA2, RUNX1, RUNX2, ELK4) che hanno il rapporto atteso-osservato, calcolato sui risultati di HOMER, maggiore nei cluster C1 e C2 (Fig. 10). A supporto della nostra osservazione, il lavoro recentemente pubblicato da Tajedor et al. (Tejedor et al., 2021) identifica i medesimi TF determinanti nella progressione della B-ALL. Andrà comunque approfondito se i TF ERG ed EBF svolgano un ruolo nelle alterazioni dell'apertura cromatinica nei siti appartenenti ai *cluster* C1 e C2.

Il lavoro è poi continuato per determinare quali di queste 6.000 regioni potessero essere identificate univocamente come *enhancer*. Per fare ciò è stato utilizzato il portale HeRA (Zhang et al., 2021) che contiene gli *enhancer* individuati tramite l'integrazione di molteplici dati ottenuti con tecniche NGS, e relativi geni *target*. L'analisi è stata completata tramite la valutazione della presenza delle regioni di C1 e C2 in HeRA e all'interno della linea cellulare LAL-B. Queste cellule primarie di ALL sono in grado di ricapitolare in maniera efficiente il fenotipo degli esordi come mostrato precedentemente (Fig. 11). Gli *enhancer* così identificati ammontano a 117. Per confermare questa osservazione è stata anche valutata la presenza dei fattori trascrizionali precedentemente individuati, e per fare ciò sono stati utilizzati i dati di ChIPseq pubblici di linee cellulari prevalentemente leucemiche. Come mostrato (Fig. 14) è stata riscontrato il legame di molteplici fattori agli *enhancer*, confermando quanto osservato. Questo rafforza il concetto secondo cui le regioni identificate siano necessarie nella progressione della B-ALL, inoltre l'aumento di queste regioni in pazienti recidivanti suggerisce un coinvolgimento di queste regioni nella resistenza ai farmaci.

Ulteriori caratterizzazioni dei 117 *enhancer* hanno portato all'identificazione di possibili target della progressione di B-ALL. Infatti, tramite l'integrazione delle informazioni ottenute da DepMap, ENCODE e le analisi del *Ranking Index* e di Promoter-Capture prodotte nel nostro laboratorio è stato possibile

individuare gli *enhancer* che hanno come geni target DCTD, MYB e BCL2 come possibili nuovi target terapeutici. Infatti è stato possibile individuare un incremento nella produzione di eRNA da parte di questi *enahncer* mediante un'RNAseq totale condotta su campione di pazienti (Fig. 15).

L'interesse è poi stato concentrato prevalentemente su DCTD dato il suo coinvolgimento nella produzione di basi azotate e quindi indispensabile in tutte le cellule in attiva proliferazione. È stato pertanto realizzato un esperimento di RNA *interference* diretto contro l'eRNA dell'*enhancer* di DCTD. I risultati (Fig. 21) mostrano una riduzione della proliferazione cellulare di circa il 30% nelle LAL-B e nelle NALM-6, oltre che una diminuzione del 30% della produzione di mRNA del gene DCTD.

Nonostante questi risultati siano di grande interesse andranno condotti ulteriori sforzi per la caratterizzazione di questi *enahncer* per determinare se la riduzione di trascritto e della proliferazione non siano dovuti ad interazioni indesiderate con *target* aspecifici. Per fare ciò sono state identificate due linee cellulari K562 e OCI-LY3 come possibili controlli poiché non mostrano apertura nell'*enhancer* di DCTD individuato. Sarebbe, inoltre, di interesse valutare il comportamento delle LAL-B successivamente al *knock-out* dell'*enhancer* tramite la tecnologia CRISPR-Cas.

In conclusione, questo lavoro ha permesso di caratterizzare in maniera efficiente il comportamento del panorama di accessibilità nel corso della progressione tumorale. L'identificazione di regioni non codificanti come fattori preponderanti nel decorso della patologia pone l'accento sulla necessità di compiere questo tipo di analisi e studi a livelli sempre più approfonditi.

Un possibile miglioramento di questo lavoro è individuabile nella produzione di una metodica computazionale in grado di dissezionare in maniera sistematica i profili di accessibilità dei pazienti e poterne determinare una classificazione basata sulla variazione dell'apertura di ciascun picco nei vari stadi della patologia. In questo modo si potrebbero perciò superare la più grade limitazioni riscontrata, ovvero la selezione dei picchi tramite il

confronto di soli due *status* alla volta, senza tenere in considerazione le alterazioni dell'accessibilità di una regione genomica durante la progressione tumorale.

5. Bibliografia

Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports *9*, 9354.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nature Reviews Genetics *21*, 71–87.

Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–2405.

Black, J.R.M., and McGranahan, N. (2021). Genetic and non-genetic clonal diversity in cancer evolution. Nature Reviews Cancer 1–14.

Boix, C.A., James, B.T., Park, Y.P., Meuleman, W., and Kellis, M. (2021). Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature *590*, 300–307.

Braccini, A., Wendt, D., Jaquiery, C., Jakob, M., Heberer, M., Kenins, L., Wodnar-Filipowicz, A., Quarto, R., and Martin, I. (2005). Three-Dimensional Perfusion Culture of Human Bone Marrow Cells and Generation of Osteoinductive Grafts. STEM CELLS *23*, 1066–1072.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Cabrita, G.J.M., Ferreira, B.S., da Silva, C.L., Gonçalves, R., Almeida-Porada, G., and Cabral, J.M.S. (2003). Hematopoietic stem cells: from the bone to the bioreactor. Trends in Biotechnology *21*, 233–240.

Chen, H., and Liang, H. (2020). A High-Resolution Map of Human Enhancer RNA Loci Characterizes Super-enhancer Activities in Cancer. Cancer Cell *38*, 701-715.e5.

Cheng, Y.-W., Pincas, H., Bacolod, M.D., Schemmann, G., Giardina, S.F., Huang, J., Barral, S., Idrees, K., Khan, S.A., Zeng, Z., et al. (2008). CpG Island Methylator Phenotype Associates with Low-Degree Chromosomal Abnormalities in Colorectal Cancer. Clin Cancer Res *14*, 6005–6013.

Cobaleda, C., and Busslinger, M. (2008). Developmental plasticity of lymphocytes. Current Opinion in Immunology *20*, 139–148.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794–D801.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. The Lancet *392*, 777–786.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cels *3*, 95–98.

Esteller, M. (2011). Epigenetic changes in cancer. F1000 Biol Rep *3*.

Feinberg, A.P. (2018). The Key Role of Epigenetics in Human Disease Prevention and Mitigation. New England Journal of Medicine.

Fitz, J., Neumann, T., Steininger, M., Wiedemann, E.-M., Garcia, A.C., Athanasiadis, A., Schoeberl, U.E., and Pavri, R. (2020). Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. Nat Genet *52*, 505–515.

Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science *357*.

Gatenby, R.A., Smallbone, K., Maini, P.K., Rose, F., Averill, J., Nagle, R.B., Worrall, L., and Gillies, R.J. (2007). Cellular adaptations to hypoxia and acidosis during somatic evolution of breast cancer. British Journal of Cancer *97*, 646–653.

Gimble, J.M., Robinson, C.E., Wu, X., and Kelly, K.A. (1996). The function of adipocytes in the bone marrow stroma: an update. Bone *19*, 421–428.

Hanahan, D., and Weinberg, R.A. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576–589.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nature Reviews Molecular Cell Biology *16*, 144–154.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. Cell *155*, 934–947.

Howlader, N., Noone, A.M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D.R., Chen, H.S., et al. (2021). SEER Cancer Statistics Review, 1975-2018, National Cancer Institute. (Bethesda, MD).

Hunger, S.P., and Mullighan, C.G. (2015). Acute Lymphoblastic Leukemia in Children (Massachusetts Medical Society).

Kaikkonen, M.U., Spann, N.J., Heinz, S., Romanoski, C.E., Allison, K.A., Stender, J.D., Chun, H.B., Tough, D.F., Prinjha, R.K., Benner, C., et al. (2013). Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. Molecular Cell *51*, 310–325.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahoviček, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. PNAS *107*, 2926–2931.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature *565*, 251–254.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Li, W., Notani, D., and Rosenfeld, M.G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nature Reviews Genetics *17*, 207–223.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580–585.

Machado, H.E., Mitchell, E., Øbro, N.F., Kübler, K., Davies, M., Maura, F., Leongamornlert, D., Sanders, M.A., Cagan, A., McDonald, C., et al. (2021). Genome-wide mutational signatures of immunological diversification in normal lymphocytes. BioRxiv 2021.04.29.441939.

Maegawa, S., Hinkal, G., Kim, H.S., Shen, L., Zhang, L., Zhang, J., Zhang, N., Liang, S., Donehower, L.A., and Issa, J.-P.J. (2010). Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. *20*, 332–340.

Maier, H., Ostraat, R., Gao, H., Fields, S., Shinton, S.A., Medina, K.L., Ikawa, T., Murre, C., Singh, H., Hardy, R.R., et al. (2004). Early B cell factor cooperates with Runx1 and mediates epigenetic changes associated with mb-1 transcription. Nat Immunol *5*, 1069–1077.

Marine, J.-C., Dawson, S.-J., and Dawson, M.A. (2020). Non-genetic mechanisms of therapeutic resistance in cancer. Nature Reviews Cancer *20*, 743–756.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis -regulatory regions. Nat Biotechnol *28*, 495–501.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet *49*, 1779–1784.

Minnoye, L., Marinov, G.K., Krausgruber, T., Pan, L., Marand, A.P., Secchia, S., Greenleaf, W.J., Furlong, E.E.M., Zhao, K., Schmitz, R.J., et al. (2021). Chromatin accessibility profiling methods. Nat Rev Methods Primers *1*, 1–24.

Murakawa, Y., Yoshihara, M., Kawaji, H., Nishikawa, M., Zayed, H., Suzuki, H., Fantom Consortium, null, and Hayashizaki, Y. (2016). Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. Trends Genet *32*, 76–88.

Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature 1–6.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., et al. (2010). Inferring tumor progression from genomic heterogeneity. Genome Res. *20*, 68–80.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., and Meno, C. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Reports *19*, e46255.

Patten, D.K., Corleone, G., Győrffy, B., Perone, Y., Slaven, N., Barozzi, I., Erdős, E., Saiakhova, A., Goddard, K., Vingiani, A., et al. (2018). Enhancer mapping uncovers phenotypic heterogeneity and evolution in patients with luminal breast cancer. Nature Medicine *24*, 1469–1480.

Peterson, C.L., and Laniel, M.-A. (2004). Histones and histone modifications. Current Biology *14*, R546–R551.

Plank, J.L., and Dean, A. (2014). Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. Molecular Cell *55*, 5–14.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nat Biotechnol *28*, 1057–1068.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Sartorelli, V., and Lauberth, S.M. (2020). Enhancer RNAs are an important regulatory layer of the epigenome. Nature Structural & Molecular Biology *27*, 521–528.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics *20*, 437–455.

Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S.W., and Fraser, P. (2018). Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. JoVE (Journal of Visualized Experiments) e57320.

Shizuru, J.A., Negrin, R.S., and Weissman, I.L. (2005). Hematopoietic Stem and Progenitor Cells: Clinical and Preclinical Regeneration of the Hematolymphoid System. Annual Review of Medicine *56*, 509–538.

Singer, R.A.J., Montecino-Rodriguez, E., and Dorshkind, K. (2006). Aging, B lymphopoiesis, and patterns of leukemogenesis. Exp Gerontol *42*, 391–395.

Sniegowski, P. (1997). Evolution: Setting the mutation rate. Current Biology *7*, R487–R488.

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345–354.

Stanulla, M., and Schrappe, M. (2009). Treatment of Childhood Acute Lymphoblastic Leukemia. Seminars in Hematology *46*, 52–63.

Sungalee, S., Liu, Y., Lambuta, R.A., Katanayeva, N., Donaldson Collier, M., Tavernari, D., Roulland, S., Ciriello, G., and Oricchio, E. (2021). Histone acetylation dynamics modulates chromatin conformation and allele-specific interactions at oncogenic loci. Nature Genetics *53*, 650–662.

Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. Nature Reviews Cancer *16*, 483–493.

Symmons, O., Uslu, V.V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. Genome Res. *24*, 390–400.

Tang, F., Yang, Z., Tan, Y., and Li, Y. (2020). Super-enhancer function and its application in cancer targeted therapy. Npj Precision Oncology *4*, 1–7.

Tavassoli, M., and Yoffey, J. (1984). Bone marrow: structure and function. Scandinavian Journal of Haematology *32*, 335–335.

Tejedor, J.R., Bueno, C., Vinyoles, M., Petazzi, P., Agraz-Doblas, A., Cobo, I., Torres-Ruiz, R., Bayón, G.F., Pérez, R.F., López-Tamargo, S., et al. (2021). Integrative methylome-transcriptome analysis unravels cancer cell vulnerabilities in infant MLL-rearranged B-cell acute lymphoblastic leukemia. J Clin Invest.

Waterland, R.A., and Jirtle, R.L. (2003). Transposable Elements: Targets for Early Nutritional Effects on Epigenetic Gene Regulation. Mol Cell Biol *23*, 5293–5300.

Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. Nature Genetics *48*, 238–244.

Woo, J.S., Alberti, M.O., and Tirado, C.A. (2014). Childhood B-acute lymphoblastic leukemia: a genetic update. Experimental Hematology & Oncology *3*, 16.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Yates, L.R., and Campbell, P.J. (2012). Evolution of the cancer genome. Nature Reviews Genetics *13*, 795–806.

Zhang, Z., Lee, J.-H., Ruan, H., Ye, Y., Krakowiak, J., Hu, Q., Xiang, Y., Gong, J., Zhou, B., Wang, L., et al. (2019). Transcriptional landscape and

clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. Nature Communications *10*, 4562.

Zhang, Z., Hong, W., Ruan, H., Jing, Y., Li, S., Liu, Y., Wang, J., Li, W., Diao, L., and Han, L. (2021). HeRA: an atlas of enhancer RNAs across human tissues. Nucleic Acids Research *49*, D932–D938.

Zhao, Z., and Shilatifard, A. (2019). Epigenetic modifications of histones in cancer. Genome Biology *20*, 245.