Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Terapia Pediatrica e Farmacologia dello Sviluppo

Ciclo XXVI

Identificazione e analisi funzionale di fattori regolatori dei geni globinici

Settore scientifico disciplinare di afferenza

[MED/03]

Presentata da:	Dott.ssa Alessia Desogus
Coordinatore Dottorato:	Prof. Paolo Moi
Tutor/ Relatore:	Prof. Paolo Moi
Correlatrici:	Dott.ssa Manuela Uda
	Dott.ssa Maria Giuseppina Marini

Esame finale anno accademico 2012 – 2013

UNIONE EUROPEA
Fondo sociale europeo

REGIONE AUTONOMA DELLA SARDEGNA

FSE 2007-2013
POR SARDEGNA
obiettivo competitività regionale e occupazione

La presente tesi è stata prodotta durante la frequenza del corso di Dottorato in Terapia Pediatrica e Farmacologia dello Sviluppo dell'Università degli Studi di Cagliari, a.a. 2012/2013 - XXVI ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività l.3.1 "Finanziamento di corsi di dottorato finalizzati alla formazione di capitale umano altamente specializzato, in particolare per i settori dell'ICT, delle nanotecnologie e delle biotecnologie, dell'energia e dello sviluppo sostenibile, dell'agroalimentare e dei materiali tradizionali.

INDICE

1. ABSTRACT

Genome wide association studies have identified two quantitative trait loci outside of the β-globin cluster associated with fetal hemoglobin (HbF) levels, number of F cell and β-thalassemia severity: the *HBS1L-MYB* intergenic region and the *BCL11A* gene. In order to understand the functional role of the associated variants at these loci we applied "Genome Wide Chromosome Conformation Capture" (Hi-C), followed by a novel technique for a selective enrichment at these target regions, to characterize whether they are involved in long range physical interactions able to modulate *HBS1L-MYB* and *BCL11A* expression. As a first step we optimized a Hi-C protocol in the K562 fetal erythroid cell line and we set up the conditions for the new method based on the selective enrichment of target regions. We were able to validate the new capture system analyzing the β-globin locus, as a control model, detecting all the interactions found by other "3C-like" technologies with a higher resolution. We then analyzed the data at the *HBS1L-MYB* and *BCL11A* loci; the most significant detected interactions involved four *HBS1L-MYB* intergenic regions, three of which contain the GWAS SNPs, the *HBS1L* and *MYB* genes. We hypothesized a contact model where the associated variants could exert their regulatory role likely altering transcription factors binding sites and thus DNA long-range interactions resulting in different levels of *MYB* expression. Indeed, although we can not exclude the implication of *HBS1L* gene in this mechanism, *MYB* represents the best candidate in modulating HbF levels given its role in the erythropoiesis kinetics. Our results highlighted the power of the new capture system, able to identify chromatin interactions with very high resolution simultaneously at different loci. Finally, we discovered new chromatin interactions that support the transcription factor MYB as a potential good candidate to develop new targeted therapeutic strategies to treat β-thalassemia patients.

2.INTRODUZIONE

2.1 Emoglobina e geni globinici

La componente principale dei globluli rossi è l'emoglobina (Hb), una cromoproteina tetramerica globulare scoperta da Hunefeld nel 1840, che trasporta l'ossigeno dai polmoni ai tessuti attraverso il circolo sanguigno.

E' costitutita da quattro subunità, ciascuna delle quali contiene un gruppo eme circondato e protetto da una catena polipeptidica globinica (Figura 1).

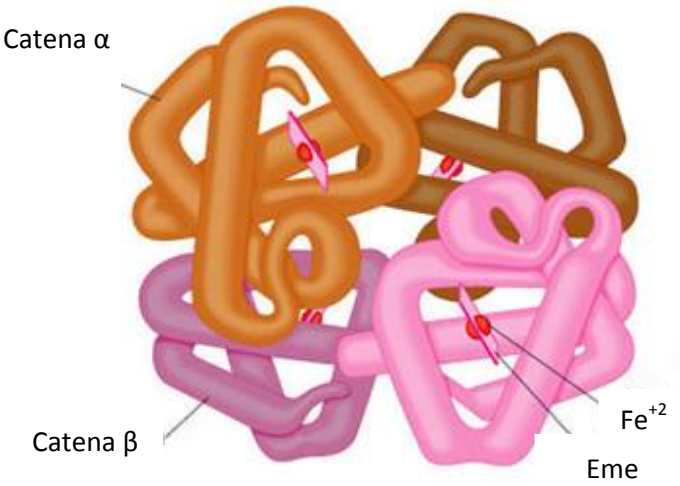

Figura 1: Struttra molecolare dell'emoglobina.

Il gruppo eme conferisce la colorazione rossa ai globuli rossi e si trova all'interno di una tasca idrofobica costituita da venti amminoacidi; i suoi due principali costituenti sono la protoporfirina IX, la parte organica, ed un atomo di ferro inorganico che si trova al suo interno e con cui interagisce mediante quattro legami. Il ferro lega reversibilmente l'ossigeno, e tale legame innesca un cambiamento conformazionale producendo il cosiddetto effetto di cooperazione positiva che induce una maggiore affinità per l'ossigeno stesso da parte degli altri gruppi eme. Il ferro deve mantenere uno stato di ossidazione +2, in quanto un eventuale passaggio allo stato ferrico +3 comprometterebbe il corretto trasporto dell'ossigeno con conseguente anossia. Lo stato ferroso (Fe+2) viene garantito grazie alle proprietà proteiche e strutturali della molecola globulare, tali per cui gli amminoacidi idrofilici sono rivolti all'esterno, mentre quelli idrofobici all'interno, creando le condizioni ottimali per il corretto posizionamento del ferro.

La porzione proteica è costituita da quattro catene polipeptidiche: due di tipo *α* (ζ e α), di 141 residui amminoacidici e due di tipo *β* (ε, γ, δ e β), di 146 aminoacidi (Bianco, Silvestroni. 1998). Le quattro subunità

si assemblano spontaneamente mediante interazioni elettrostatiche ed idrofobiche (Ranney, Sharma. 1991).

I geni per le catene globiniche di tipo α e di tipo β sono organizzati in due raggruppamenti chiamati *clusters* e sono disposti nello stesso ordine con il quale vengono trascritti ed espressi durante lo sviluppo. Tutti i geni globinici sono costituiti da tre esoni e due introni e questa similarità strutturale induce a pensare che si siano generati da un unico progenitore comune.

Il *cluster* α-globinico è situato nell'estremità distale del braccio corto del cromosoma 16, in corrispondenza della banda 16p13.3 e si estende per circa 30 Kb. E' costituito dal gene embrionale ζ, dai geni fetali/adulti α_1 e α_2, dai tre pseudogeni ψζ, ψα_1 e ψα_2 e dal gene θ, identificato più recentemente (Marks et al. 1986) (Figura 2). I geni α_1 e α_2 differiscono strutturalmente a livello del II introne e del III esone, ma i trascritti sono identici e variano solo nella quantità; infatti α_2 produce un trascritto 2,6 volte più abbondante rispetto a α_1 (Liebhaber et al. 1986), e tale evento è riconducibile ad una lieve diversità nella sequenza promotrice.

Il *cluster* α possiede un'alta densità di GC (54%), localizzate in regioni chiamate HRV (*High Variability Region*), oltre che sequenze ripetute appartenenti alla famiglia AluI che favoriscono processi di ricombinazione.

A monte del gene ζ, ad una distanza di circa 40 Kb, sono presenti una serie di siti ipersensibili alla DNase I che complessivamente prendono il nome di HS-40. Nonostante tali regioni non siano assimilabili alla *locus control region* del *cluster* β-globinico, è stato tuttavia dimostrato il loro coinvolgimento nella regolazione dei geni α-globinici. L'esempio più eclatante è dato dalla presenza di delezioni naturali all'interno di questi siti che riducono l'espressione dei geni a valle, condizione tipica dei portatori di α-talassemia.

Il *cluster* β-globinico, di circa 70 Kb, si trova nella regione distale del braccio corto del cromosoma 11, nella banda 11p15.4 e contiene cinque geni strutturali: uno embrionale ε, due fetali $^{G}\gamma$ e $^{A}\gamma$, e due adulti δ e β; è presente inoltre lo pseudogene ψβ1 (*HBBP1*) che mappa tra i geni $^{A}\gamma$ e δ, ma non è funzionale (Figura 2).

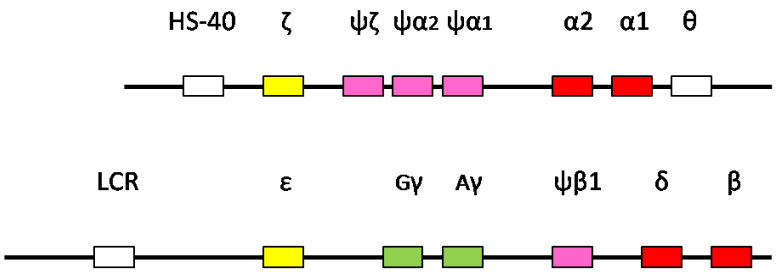

Figura 2: Rappresentazione schematica dei *cluster*s α-globinico in alto e β-globinico in basso.

Il processo di *switching* globinico rappresenta uno dei meccanismi di regolazione spazio-temporale dell'espressione genica più studiato, che determina la progressiva e sequenziale variazione dell'espressione dei geni embrionali e fetali, fino all'attivazione dei geni β-globinici di tipo adulto, parallelamente alla variazione nella sede dell'ematopoiesi (Figura 3).

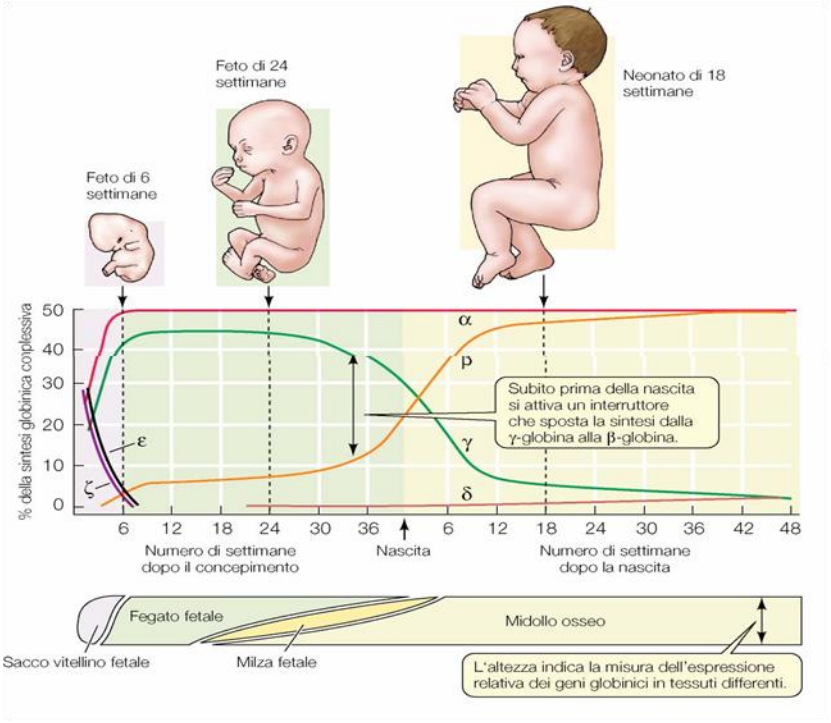

Figura 3: Espressione genica delle diverse catene globiniche durante lo sviluppo. (L'immagine proviene da: "biologia l'informazione e l'eredità". Savada, Orians, Heller; Zanichelli 2005).

Durante il periodo embrionale nel sacco vitellino sono attivi i geni responsabili della produzione delle Hb di Gower I ($\zeta_2\varepsilon_2$), Gower II ($\alpha_2\varepsilon_2$) e Portland ($\zeta_2\gamma_2$). Dopo le prime due settimane di gestazione l'espressione del gene ζ diminuisce, mentre aumenta quella dei geni per le globine α. Dopo circa sei settimane dal concepimento le globine ε sono sostituite dalle globine γ; in questa fase la sede eritropoietica si sposta nel fegato e l'Hb risultante, denominata emoglobina fetale (HbF), rappresenta circa il 90% dell'emoglobina totale. L'HbF ha una maggiore affinità per l'ossigeno rispetto all'Hb adulta (HbA) permettendo un trasferimento più efficiente dell'ossigeno dal sangue materno a quello fetale. Dopo la nascita, quando l'ematopoiesi si sposta definitivamente nel midollo osseo, si ha il secondo e più importante *switching* che si completa entro i primi due anni di vita, con una progressiva riduzione nella sintesi di γ globine a favore delle β globine che porta al quadro adulto in cui le diverse quote di emoglobina sono così rappresentate: HbA ($\alpha_2\beta_2$) ~96%, HbA_2 ($\alpha_2\delta_2$) ~3% e HbF ($\alpha_2\gamma_2$) ~1% dell'emoglobina totale contenuta nei globuli rossi.

2.2 Regolazione dei geni globinici

L'espressione dei geni β-globinici è regolata dalla *Locus Control Region* (*LCR*), una regione di circa 25 Kb localizzata in 5' rispetto al gene ε. La *LCR* è costituita da siti ipersensibili alla DNase I (HS), il cui numero varia tra specie; nell'uomo e nel topo sono cinque. Gli HS1-4 sono presenti solo in cellule eritroidi, mentre l'HS5 è ubiquitario, ma non costitutivo (Li et al. 1999). Il ruolo centrale della *LCR* nell'attivazione della trascrizione dei geni globinici è confermata da numerosi studi, in cui è stato osservato che la delezione dell'intera regione compromette l'espressione di tutti i geni del *cluster*, inibendola e causando la β-talassemia (Curtin et al. 1985).

Il processo che coordina finemente l'espressione delle globine durante le varie fasi dello sviluppo avviene mediante l'intervento sia di elementi *cis-acting* che di fattori *trans-acting*. Tra i primi troviamo sequenze specifiche vicine o all'interno dei geni stessi e della *LCR*, che agiscono come *enhancer*, *silencer* o *insulator*. Tali regioni legano a loro volta elementi *trans-acting* che comprendono fattori di trascrizione e fattori implicati nel rimodellamento della cromatina (Vakoc et al. 2005). Inoltre la *LCR* stessa influenza la struttura della cromatina costituendo una regione cromosomica aperta, cioè più accessibile ai fattori di trascrizione ed interagisce con i singoli promotori dei geni globinici e con l'apparato di trascrizione dell'RNA polimerasi per regolare l'espressione stadio specifica dei singoli geni del *cluster* (Qiliang et al. 2002).

Malgrado i numerosi studi sul ruolo svolto dalla *LCR* sul controllo dell'espressione dei geni globinici il meccanismo d'azione non è ancora chiaro. Nel corso degli anni sono stati proposti quattro modelli:

- Il *looping model* prevede l'interazione della *LCR* con i promotori dei geni interessati mediante la formazione di *loop* di DNA (Wijgerde et al. 1995).
- Nel *tracking model* il legame di fattori e cofattori di trascrizione con la *LCR* porta alla formazione di un complesso che migra lungo il DNA, fino ad incontrare l'apparato di trascrizione sul promotore del gene da esprimere attivando la trascrizione.
- Il *tracking model* facilitato incorpora entrambi i modelli precedenti: la *LCR* si ripiega e viene in contatto con i promotori dei geni a valle portando alla liberazione di un complesso che si muove lungo in DNA, fino ad incontrare il promotore del gene da esprimere.
- Il *linking model* ipotizza che fattori di trascrizione legati alla cromatina possano definire i domini da trascrivere e mediare il legame sequenziale stadio specifico dei vari fattori.

Tutti i modelli direttamente o indirettamente implicano un'alterazione nella struttura della cromatina.

Alcuni studi condotti più recentemente suggeriscono che la *LCR* formi un olocomplesso (*Active Chromatin Hub*; ACH) interagendo con i geni globinici in relazione alla fase di sviluppo (De Laat, Grosveld. 2003). Tale interazione sembra essere mediata dalla formazione di *loop* di DNA, in cui complessi DNA-proteina giocano un ruolo importante nel guidare il contatto che si stabilisce tra i vari elementi del *locus*.

2.3 La β-Talassemia

Le talassemie sono un gruppo eterogeneo di anemie a carattere ereditario autosomico recessivo, causate da mutazioni a carico dei geni globinici che risultano in un difetto nella sintesi dell'emoglobina.

La forma di talassemia più diffusa in Italia è la β-talassemia caratterizzata dalla riduzione (β^+) o dal deficit (β^0) della sintesi di catene β-globiniche.

I soggetti β-talassemici sono caratterizzati da una severa anemia che si manifesta attraverso due meccanismi: l'eritropoiesi inefficace da morte intramidollare degli eritroblasti e iperemolisi periferica.

La ridotta o assente produzione di catene β determina uno sbilanciamento del rapporto esistente tra catene α e β prodotte a livello di precursori eritroidi, con eccesso relativo di catene α ($\alpha/\beta > 1$). Queste si aggregano e tendono a precipitare formando dei corpi insolubili che determinano un danno ossidativo a carico delle membrane, già allo stadio di precursori eritroidi, i quali vanno incontro a prematura apoptosi e quindi ad eritropoiesi inefficace. Solo pochi proeritroblasti sono in grado di arrivare a maturazione completa ed essere quindi immessi nel torrente circolatorio come eritrociti. Queste cellule, che per via della maturazione anomala presentano all'interno numerose inclusioni citoplasmatiche, vengono prematuramente rimosse dal torrente circolatorio dalle cellule reticoloendoteliali della milza, del fegato e del midollo osseo, determinando una condizione di anemia emolitica cronica. Lo stato anemico stimola la sintesi di eritropoietina nel tentativo di indurre una iperplasia eritroide compensatoria. La notevole espansione midollare è in grado di produrre solo un numero limitato di nuovi elementi cellulari maturi, mentre causa un notevole aumento dei progenitori eritroidi immaturi che vanno incontro ad arresto maturativo a causa dell'accumulo di α-globine. La massiva espansione del midollo osseo ha notevoli effetti negativi sulla crescita, lo sviluppo e la funzione di numerosi organi e sistemi. Inoltre come conseguenza dell'anemia emolitica si osserva una notevole splenomegalia e talvolta può svilupparsi un'insufficienza cardiaca congestizia.

2.3.1 Distribuzione geografica della β-talassemia

La prevalenza della β-talassemia non è nota, ma l'incidenza alla nascita della forma grave è stimata in 100.000/anno. La malattia è stata inizialmente descritta nel bacino mediterraneo, ma le forme gravi sono presenti anche nell'Asia Centro e Sud-Orientale, in India e in Cina. In particolare la più alta incidenza dei portatori sani si osserva nel Sud Est Asiatico (16%), Cipro (~ 14%) ed in Sardegna (Italia ~ 12%) (Galanello, Cao. 1998; Weatherall et al. 2002). Gli individui portatori della malattia sono più resistenti alla grave infezione da *Plasmodium Falciparum* rispetto ai soggetti sani, e questo ha portato ad un'alta frequenza di mutazioni β-talassemiche in zone con presente o passata endemicità malarica, nonostante la condizione patogenetica degli omozigoti (selezione bilanciante) (Flint et al. 1998). Tuttavia l'epidemiologia della malattia sta cambiando, soprattutto a causa dei flussi migratori e dei matrimoni misti tra i diversi gruppi

etnici diventando un importante problema di assistenza sanitaria, economica e sociale anche in Nord Europa, Nord e Sud America, Caraibi e Australia.

2.3.2 Clinica e genetica della β-talassemia

Nelle β-talassemie si distinguono tre quadri clinici ed ematologici di gravità crescente: la talassemia *major* o morbo di Cooley, una grave anemia incompatibile con la vita in assenza di trasfusioni regolari; la talassemia intermedia, un'anemia microcitica di gravità variabile, molto eterogenea geneticamente e fenotipicamente, in cui i pazienti non necessitano di trasfusioni regolari per la sopravvivenza; il portatore sano, clinicamente asintomatico risultante dallo stato eterozigote per la β-talassemia.

I pazienti omozigoti o composti eterozigoti per mutazioni β-globiniche possono sviluppare sia talassemia *major* (TM) che talassemia intermedia (TI).

Individui con TM di solito mostrano sintomi di anemia grave entro i primi 2 anni di vita e richiedono regolari trasfusioni di sangue per sopravvivere. La malattia si manifesta con senso di affaticamento, modesti rialzi termici improvvisi oppure associati a lievi episodi influenzali o a disturbi della dentizione, pallore molto accentuato, ittero e crescita inferiore alla norma. Altri segni importanti sono le deformazioni ossee, l'ingrossamento di fegato e milza, la comparsa di calcoli biliari e l'accumulo di ferro che può portare allo scompenso cardiaco. Il quadro ematologico si caratterizza per livelli di Hb di 4-6 g/dl, globuli rossi 2.000.000 mmc, eritroblastemia, anisopoichilocitosi, resistenze globulari aumentate, iperplasia eritroblastica midollare, HbF 70-90% e HbA_2 > 3.5%. Fino all'età di 10-11 anni con un regime trasfusionale regolare che mantenga una concentrazione minima di Hb di 9.5-10.5 g/dl la crescita e lo sviluppo del paziente risulteranno nella norma. Più tardivamente se non viene attuata un'adeguata terapia con chelanti del ferro, si manifestano complicazioni a carico del cuore (miocardiopatia dilatativa e pericardite), del fegato (epatite cronica, fibrosi e cirrosi), delle ghiandole endocrine (con conseguente diabete mellito, ipotiroidismo, ipoparatiroidismo), della ghiandola pituitaria e, meno comunemente, delle ghiandole surrenali correlate al sovraccarico di ferro post-trasfusionale in questi organi.

La TI è definita come un quadro clinico ad ampio spettro di variabilità applicato a quei pazienti con fenotipo β-talassemico lieve-moderato, in grado di mantenere spontaneamente livelli di Hb uguali o superiori a 7 gr/dl, senza regolare fabbisogno trasfusionale. L'esordio della malattia è più tardivo rispetto ai soggetti con TM e lo sviluppo psichico e fisico è normale. Le deformazioni ossee sono scarse o assenti mentre si ha costante epato-splenomegalia e frequente calcolosi biliare. Il quadro ematologico è qualitativamente analogo a quello della TM, tuttavia l'anemia microcitica ipocromica è meno spiccata ed è quasi sempre associata a reticolocitosi. La sopravvivenza di questi pazienti è variabile, data la variabilità del quadro clinico ed ematologico. Dalla terza decade di vita possono rendersi necessarie le trasfusioni, anche se per definizione la terapia trasfusionale non dovrebbe essere indicata. Tuttavia questo schematismo non ha

riscontro nella pratica e l'appartenenza di un individuo alla classe *major* o intermedia è basata essenzialmente su osservazioni cliniche e spesso affidata al giudizio del medico. Infatti da un lato vi sono clinici che optano per un regime trasfusionale regolare, in modo tale da prevenire gli effetti collaterali causati da un'eccessiva attività del midollo osseo, viceversa altri preferiscono evitare le trasfusioni per non rendere il paziente dipendente ed esposto agli effetti collaterali che ne derivano. Quindi vi è una crescente consapevolezza della necessità di stabilire criteri più obiettivi per la diagnosi al fine di ottenere una gestione ottimale del paziente e ridurre i problemi dovuti ad un mancato o eccessivo trattamento terapeutico.

La significativa variabilità di fenotipi clinici nella β-talassemia riflette prima di tutto l'elevata eterogeneità osservata a livello molecolare, con oltre 200 mutazioni diverse identificate (Cao, Galanello. 2010), per lo più costituite da sostituzioni/delezioni/inserzioni di singoli nucleotidi che alterano il processo di trascrizione, di maturazione di mRNA, di traduzione o di stabilità della molecola messaggero. In particolare delezioni o inserzioni puntiformi nelle regioni esoniche determinano uno scivolamento del messaggio con l'alterazione di tutti i codoni a valle della mutazione, producendo una proteina non funzionante con un suo arresto più o meno precoce (*Frameshift*). Anche le mutazioni nonsenso, rappresentate da variazioni puntiformi che creano una tripletta di stop nella regione codificante, determinano un'interruzione prematura della traduzione. Tali mutazioni sono associate ad un fenotipo β°-talassemico. I difetti di trascrizione interessano soprattutto mutazioni nel promotore determinando una riduzione della quantità di mRNA e producendo forme di β^+-talassemia più o meno lievi in relazione alla regione del promotore colpita. Gli introni sono interessati da mutazioni in grado di alterare un sito di *splicing*, sia distruggendo il sito canonico che creando un sito di *splicing* criptico. Tali mutazioni producono una talassemia β° o β^+ più o meno grave, a seconda che interessino direttamente i siti di *splicing* o sequenze interne agli introni. Più raramente le β-talassemie sono dovute a delezione limitate e intrageniche o a delezioni più estese che coinvolgono in modo variabile anche i geni δ, $^{G}\gamma$ e $^{A}\gamma$ (β-talassemie complesse, comprendenti le δβ- e le γδβ-talassemie). Eccezionalmente sono in gioco le delezioni della *LCR* che silenziano, pur lasciandoli strutturalmente intatti, tutti i geni β-globinici, e che costituiscono il primo esempio di difetto funzionale di un gene per alterazione di sequenze lontane con funzione regolatrice.

Inoltre l'estrema eterogeneità fenotipica riscontrata nelle sindromi β-talassemiche è direttamente correlata al grado di squilibrio tra le catene globiniche α e β e/o γ. Pertanto qualunque elemento o meccanismo capace di ridurre questo squilibrio, sia riducendo la sintesi di catene α-globiniche che aumentando quella delle catene β e/o γ, può essere in grado di migliorare il fenotipo clinico della malattia (Wheatherall, Clegg. 2001).

Esistono diversi meccanismi che portano all'insorgenza della TI (Galanello, Cao. 1998).

Il meccanismo clinicamente più importante è il risultato dell'eredità di alleli β^+ in omozigosi o di composti eterozigoti β^+/β^+, che permettono la presenza di una consistente quantità residua di β-globina. Viceversa, composti eterozigoti β^+/β^0 risultano in un ampio spettro di fenotipi, che vanno dalle forme più lievi a quelle più severe, rendendo il risultato clinico difficile da prevedere.

Il secondo meccanismo consiste nella coeredità di β e α-talassemia. In questo caso la severità del fenotipo clinico correla con un miglioramento dello sbilanciamento delle catene β e α, come conseguenza di una ridotta produzione di entrambe le globine. Una delezione di uno dei quattro geni α-globinici è sufficiente per il miglioramento della β^+ talassemia, mentre invece è necessaria la delezione di due dei quattro geni α-globinici, o la presenza di una mutazione inattivante a questi geni, per il miglioramento clinico della β^0-talassemia (Wheatherall, Clegg. 2001).

Il terzo meccanismo è la coeredità di mutazioni β^0 e di fattori genetici capaci di mantenere una continua produzione di catene γ anche in età adulta, riducendo così la severità dello sbilanciamento tra catene α/non α. A questo riguardo l'HbF rappresenta uno dei maggiori e meglio studiati modificatori della gravità della patologia in pazienti affetti da β-talassemia. La sintesi di HbF avviene solo nel 5%-8% di una sottopopolazione di cellule eritroidi chiamate F cellule. Nel contesto di una deficienza severa della sintesi delle catene β-globiniche, anche bassi livelli di γ-globina nelle F cellule sono in grado di ridurre lo sbilanciamento delle catene α/non α. Questo avviene per esempio nella $\delta\beta^0$-talassemia, causata da mutazioni di varia entità che coinvolgono diverse regioni/geni del cluster β, in caso di piccole delezioni che coinvolgono solo la regione in 5' del promotore β-globinico, e infine nelle forme di HPFH (persistenza ereditaria di emoglobina fetale nell'adulto) dovute a mutazioni puntiformi nei promotori dei geni $^A\gamma$ o $^G\gamma$ (tra cui i polimorfismi -196 C->T; -XmnI-$^G\gamma$) (Tasiopoulou et al. 2008; Thein. 2005).

Pertanto negli ultimi decenni la regolazione dello *switching* da emoglobina fetale ad adulta e la possibilità di indurre la sintesi di HbF sono stati oggetto di intensi studi.

2.3.3 Approcci terapeutici per la cura della β-talassemia

Attualmente le strategie terapeutiche utilizzate per la cura della β-talassemia consistono in regolari trasfusioni di sangue, per tutta la durata della vita del paziente, con somministrazione di una terapia ferrochelante. La trasfusione di sangue infatti permette al paziente di avere la quantità di emoglobina necessaria per il trasporto dell'ossigeno ai tessuti, con riduzione nell'attività della milza e di manifestazioni cliniche come le alterazioni ossee. Questo continuo apporto di sangue crea però un accumulo di ferro su organi vitali come cuore, fegato e ghiandole endocrine. Tale condizione di "sovraccarico di ferro" può portare alla morte se i pazienti non vengono trattati con un'opportuna terapia ferrochelante.

La Deferoxamina (DFO), ad uso parenterale, è stato il primo farmaco utilizzato per sequestrare il ferro in eccesso. Successivamente sono stati introdotti due chelanti orali, il Deferasirox ed il Deferiprone, che oltre

a sostituire la lunga e talvolta dolorosa infusione sottocutanea, garantiscono rispettivamente un profilo farmacologico più sicuro ed una ridotta probabilità di sviluppare problemi cardiaci. La disponibilità dei chelanti orali ha notevolmente migliorato la qualità della vita e la longevità dei malati. Tuttavia il sovraccarico di ferro rimane a tutt'oggi un problema significativo.

Il trapianto di midollo osseo fornisce l'unica cura definitiva per la talassemia (Gaziev, Lucarelli. 2003), ma è limitato dalla disponibilità di donatori HLA compatibili ed è attualmente accessibile solo ad un 30-40% dei malati; inoltre richiede una precisa valutazione del quadro clinico di ciascun paziente e delle reali possibilità di successo. Infatti il trapianto di midollo osseo comporta notevoli rischi tra cui infezioni, insufficienza cardiaca, danni agli organi con sovraccarico di ferro ed una mortalità di circa il 10%.

Terapie alternative, come la terapia genica con le cellule staminali o l'induzione della produzione di HbF con composti farmacologici, sono ancora in fase sperimentale sebbene ci siano stati importanti progressi sia da un punto di vista scientifico che clinico (Persons. 2009; Sankaran, Orkin. 2013).

La terapia genica per esempio offre una valida alternativa, in quanto porterebbe alla sostituzione del gene difettoso con il gene sano, opportunamente veicolato nelle cellule staminali ematopoietiche del paziente stesso. Questa pratica comporta però delle limitazioni, prima fra tutte lo sviluppo di un adeguato vettore capace di penetrare efficacemente nelle cellule e garantire l'espressione stabile ed elevata del gene corretto. I retrovirus utilizzati inizialmente si sono rivelati troppo instabili, in quanto il virus non sempre viene completamente inattivato. Successivamente sono stati considerati gli adenovirus ed infine nel 2000 sono stati introdotti i vettori lentivirali (May et al. 2000); questi ultimi hanno l'abilità di infettare cellule come quelle staminali ematopoietiche che non si replicano o che si replicano con una bassa efficienza, ed inoltre hanno la capacità di veicolare ampi frammenti di DNA (fino a 10 Kb), consentendo una migliore espressione del gene terapeutico.

Diverse molecole farmacologiche tra cui agenti ipometilanti il DNA, come la 5-azacitidina e la decitabina, gli inibitori delle istone deacetilasi e i derivati del butirrato sono stati utilizzati per il trattamento di soggetti con β-talassemia allo scopo di diminuire lo sbilanciamento delle catene globiniche aumentando la sintesi di quelle γ. I risultati ottenuti nei *trials* clinici sono stati però deludenti (Pace, Zein. 2006). Questi agenti inducono la sintesi di HbF con diversi meccanismi che non sono del tutto noti. Inoltre presentano effetti collaterali importanti come la creazione di mutazioni, modificazioni epigenetiche non specifiche, cambiamenti nell'espressione genica ed il rischio di sviluppare tumori o altre patologie. Un altro agente terapeutico in grado di incrementare i livelli di HbF è l'idrossiurea. Nel caso della β-talassemia il possibile ruolo di questo farmaco è ancora poco chiaro, così come la sua efficacia, soprattutto se comparato agli studi effettuati su pazienti affetti da anemia falciforme. Le evidenze cliniche suggeriscono che pazienti con forme severe di β-talassemia trasfusione dipendente raramente rispondono con un incremento sufficiente della concentrazione di emoglobina. In particolare l'aumento di HbF derivato dal trattamento con

idrossiurea non è costante in tutti i pazienti; alcuni non hanno nessun incremento, altri invece mostrano un modesto aumento, ma comunque non sufficiente a garantire un'efficacia clinica. Inoltre in diversi pazienti è stata osservata una diminuzione della risposta dopo trattamento a lungo termine.

Purtroppo lo sviluppo di terapie mirate alla modulazione di target specifici è stata complicata e limitata dalle scarse conoscenze relative al preciso meccanismo molecolare che regola l'espressione dei geni globinici. Tuttavia le recenti scoperte di nuovi *loci* associati ai livelli di HbF e al miglioramento del quadro clinico β-talassemico offrono l'opportunità di sviluppare nuove strategie terapeutiche per la cura della malattia.

2.4 Studi GWAS identificano tre *loci* associati ai livelli di emoglobina fetale

Negli ultimi anni studi di associazione estesi all'intero genoma (GWAS) condotti in differenti gruppi etnici hanno identificato due *loci* principali al di fuori del *cluster* β-globinico che regolano i livelli di HbF in età adulta, spiegandone circa il 50% della variabilità: la regione intergenica tra il fattore di elongazione GTP-binding e l'oncogene *MYB* (*HBS1L-MYB*) ed il gene *BCL11A* sui cromosomi 6q23.3 e 2p16 rispettivamente (Menzel et al. 2007; Thein et al. 2007; Lettre et al. 2008; Uda et al. 2008). Varianti a questi *loci* sono state anche associate alla gravità del fenotipo clinico dell'anemia falciforme e della β-talassemia, ed insieme ai difetti delle α-globine rappresentano circa il 75% delle differenze fenotipiche tra talassemia *major* ed intermedia (Galanello et al. 2009; Danjou et al. 2012).

Le varianti in associazione tra i geni *HBS1L* e *MYB* risiedono in una regione di 126 kb caratterizzata dalla presenza di tre blocchi di *linkage disequilibrium* (LD): HMIP-1, HMIP-2 e HMIP-3. Gli SNPs che mostrano l'associazione più significativa con l'HbF (~9 SNPs più una delezione di 3-bp) coprono circa 24Kb della regione intergenica all'interno del blocco di LD HMIP-2, e presumibilmente comprendono la/le variante/i causative funzionalmente coinvolte nella regolazione dell'HbF.

Nonostante le ampie evidenze genetiche, e sebbene i due geni fiancheggianti (*HBS1L* e *MYB*) rappresentino entrambi i potenziali target, il meccanismo biologico in grado di spiegare l'associazione rimane ancora da chiarire (Thein et al. 2007; Uda et al. 2008).

L'*HBS1L* codifica per una proteina con attività GTPasica coinvolta in una serie di processi cellulari (Wallrapp et al. 1998), ma il suo ruolo nello sviluppo delle cellule eritroidi è ancora ignoto.

Al contrario il *MYB,* che codifica per il fattore di trascrizione c-MYB, è considerato un regolatore chiave dell'eritropoiesi e dell'ematopoiesi (Oh, Reddy. 1999; Vegiopoulos et al. 2006).

L'importanza funzionale della regione intergenica è stata dimostrata da diversi studi. La prima osservazione è risultata dall'inserzione di un transgene all'interno della regione intergenica *Hbs1l-Myb* murina in grado di abolire completamente la trascrizione di *Myb* causando un quadro di grave anemia (Mukay et al. 2006). Un successivo studio in cui l'inserzione del transgene mappava nella regione murina ortologa all'HMIP-2 ha

mostrato elevati livelli di espressione dei geni globinici embrionali in cellule eritroidi isolate dalla milza, confermando quindi nel topo il ruolo cruciale di questa regione nella regolazione dei livelli delle globine (Suzuki et al. 2013). Studi su progenitori eritroidi isolati da individui con elevati livelli di HbF hanno evidenziato una riduzione nell'espressione di entrambi i geni *MYB* e *HBS1L* (Jiang et al. 2006), e l'inattivazione specifica del *MYB* nelle stesso tipo di cellule ha determinato l'incremento nella produzione di catene γ (Sankaran et al. 2011). Esperimenti condotti in cellule K562 viceversa hanno mostrato che l'iperespressione di *MYB* reprime l'espressione dei geni γ, mentre quella dell'*HBS1L* non sortisce alcun effetto (Jiang et al. 2006). E' quindi evidente che i risultati dei vari studi non siano univoci, tuttavia la maggior parte supporta l'ipotesi che il *MYB* sia il gene coinvolto nella regolazione dell'HbF. Il meccanismo attraverso il quale quest'ultimo agirebbe regolandone i livelli non è definito; gli alti livelli di espressione di *MYB* in cellule ematopoietiche immature in proliferazione ed i bassi livelli al termine della differenziazione suggerirebbero un ruolo nella transizione e nel mantenimento di un equilibrio tra i processi di proliferazione e differenziazione (Gonda, Metcalf. 1984; Emambokus et al. 2003). Oltre a questo però non si può escludere anche un effetto diretto sul locus β-globinico.

Esperimenti di *Chromosome Conformation Capture* (3C) condotti con l'obiettivo di comprendere il ruolo delle varianti associate al *locus*, sia su cellule eritroidi murine che umane, hanno evidenziato nella regione intergenica la presenza di numerosi elementi regolatori con caratteristiche *enhancer*, che agirebbero a distanza sulla trascrizione del *MYB* interagendo fisicamente con esso attraverso la formazione di *loop* di cromatina (Stadhouders et al. 2012; Stadhouders et al. 2014). Le varianti associate altererebbero il sito di legame per fattori di trascrizione influenzando in questo modo le interazioni a distanza e quindi l'espressione del *MYB*.

Il gene *BCL11A*, localizzato nel braccio corto del cromosoma 2, codifica per un fattore di trascrizione *zinc-finger* caratterizzato da diverse isoforme che condividono un dominio N-terminale ma si differenziano per il numero di *zinc-finger* carbossiterminali. BCL11A lega motivi GC-*rich* delle regioni regolatrici dei suoi geni bersaglio agendo come repressore trascrizionale; è espresso nei precursori dei globuli rossi ed è stato precedentemente implicato nei tumori ematopoietici (Liu et al. 2003; Satterwhite et al. 2001). Nel fegato fetale e nei progenitori eritroidi il trascritto e i livelli proteici di BCL11A mostrano un'espressione stadio-specifica, suggerendo che il prodotto proteico di questo gene possa agire come repressore del gene γ globinico (Sankaran et al. 2008). Il silenziamento del gene mediante siRNA nelle cellule progenitrici eritroidi dell'adulto è capace di indurre l'espressione della γ globina con un grado direttamente correlato all'estensione del silenziamento e senza alterare il processo dell'eritropoiesi (Wilber et al. 2011). In cellule eritroidi primarie BCL11A interagisce direttamente con la cromatina nel *cluster* β-globinico umano, creando un complesso con i fattori di trascrizione eritroidi quali GATA-1, FOG1 e SOX6 ed il complesso repressore di

rimodellamento della cromatina NuRD (nucleosome remodelling and histone deacetylase complex) (Xu et al. 2010).

Inoltre, è stato stabilito che anche in modelli murini Bcl11a svolge un ruolo cruciale nello *switching* e nel silenziamento dei geni globinici. Topi transgenici per il *locus* β-globinico umano esprimono la γ-globina in modo molto simile alle globine embrionali endogene (Sankaran et al. 2009). I topi in cui il gene *Bcl11a* murino è stato silenziato (*knockout*) mostrano un'eritropoiesi normale, ma sono incapaci di silenziare i geni globinici embrionali negli eritrociti maturi, consentendo così un'espressione persistente della γ-globina quando è presente il *locus* β-globinico umano intatto (Xu et al. 2011) e supportando il ruolo del BCL11A nel controllo dello *switching* globinico nei mammiferi.

Le varianti geniche associate mediante studi GWAS mappano all'interno di una regione di circa 15Kb nell'introne 2 del *BCL11A*. In particolare sono stati rilevati due segnali indipendenti con gli rs1427407 e rs7606173, in forte LD con altre varianti che mostrano valori di *p-value* statisticamente distintinguibili, e che insieme spiegano l'associazione al *locus* (Lettre et al. 2008).

Questa regione è caratterizzata dalla presenza di marcatori della cromatina conservati, tipici di regioni *enhancer*, compresa la presenza degli istoni H3K4me1, H3K27ac e siti ipersensibili alla DNase I (DHS) eritroidi specifici (Bauer et al. 2013); inoltre sono presenti siti di legame per importanti fattori di trascrizione eritroidi GATA1 e TAL1 che si sovrappongono ai DHS. Gli SNPs nella regione associata correlano con la variazione di espressione del *BCL11A* in precursori eritroidi e le loro sequenze circostanti funzionano *in vivo* come *enhancer* eritroidi e stadio-specifici necessari per l'espressione del gene.

Complessivamente questi dati hanno definito MYB e BCL11A come principali regolatori dell'eritropoiesi e dello *switching* globinico, giocando un ruolo cruciale nel silenziamento dell'espressione delle γ globine e rappresentando due possibili target terapeutici per la cura della β-talassemia.

2.5 Elementi regolatori nel genoma umano

Il corretto funzionamento dei processi biologici richiede un preciso controllo sull'espressione spazio-temporale dei geni. Sebbene le cellule contengano le stesse informazioni genetiche, i geni sono accesi o spenti in relazione ad uno specifico stadio di sviluppo e alla loro funzione in un determinato tessuto/organo.

Per capire i meccanismi coinvolti in questo fine processo è fondamentale identificare quali sono gli elementi regolatori associati ad uno specifico gene target. Tale traguardo sarebbe fondamentale per comprendere come l'espressione di un gene venga alterata in condizioni patologiche. Così, una delle maggiori sfide della ricerca genomica è quella di identificare tutti gli elementi funzionali del DNA, che possono esplicare sia un'azione stimolatoria che repressoria sul processo trascrizionale.

A tale classe appartengono i promotori, gli *enhancers*, i *silencers* e gli *insulators.* (Figura 4); questi elementi contengono delle sequenze specifiche per il legame di fattori *cis* e *trans acting*, come i fattori di trascrizione, i coattivatori, gli elementi di rimodellamento della cromatina ed altri elementi proteici.

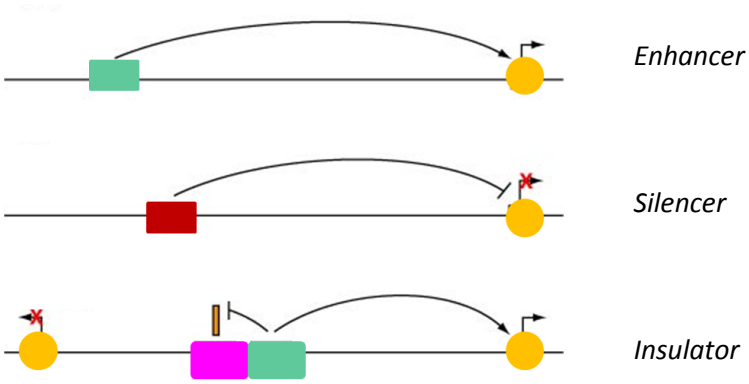

Figura 4: Elementi regolatori della trascrizione.

Il promotore è una sequenza di DNA localizzata al 5' del gene che lega l'enzima RNA polimerasi II e tutti i fattori proteici della macchina trascrizionale.

Gli *enhancers* sono brevi sequenze di DNA che legano fattori di trascrizione in grado di attivare la trascrizione di un gene bersaglio. Essi agiscono in maniera indipendente sia in funzione della posizione che dell'orientamento rispetto al promotore, infatti possono trovarsi anche alcune centinaia di kilobasi a monte o a valle dal gene, o in una sequenza intronica.

I *silencers* sono elementi con caratteristiche simili agli *enhancers* quali la modalità con cui esplicano la loro azione, indipendente sia dalla posizione che dall'orientamento, ma il loro effetto è opposto: reprimono l'attività trascrizionale.

Gli *insulators* bloccano un gene target dalla possibilità di reclutare il macchinario della trascrizione che opera in un gene adiacente bloccando la comunicazione tra un *enhancer* ed un promotore, o prevenendo la diffusione dell'eterocromatina (Gaszner, Felsenfeld. 2006). Solitamente la loro funzione è posizione dipendente ma orientamento indipendente rispetto al promotore.

L'importanza degli elementi regolatori è dovuta anche al fatto che mutazioni nella loro sequenza, tra cui sostituzioni/delezioni/inserzioni, possono essere associate a numerose malattie o correlate alla variazione di tratti quantitativi.

2.6 Nuove metodiche per lo studio della struttura della cromatina e delle interazioni tra regioni di DNA

I processi biologici di ogni organismo sono strettamente correlati alla funzione svolta dai vari elementi regolatori del DNA (promotori, *enhancers*, *insulators*, *silencers*). Tali fattori collaborano tra di loro per governare finemente quelli che sono i *pattern* di espressione spazio-temporale di ciascun gene, tessuto e organismo. Negli ultimi anni i passi in avanti nel campo della genomica hanno permesso di identificare un numero altissimo di questi elementi, mentre sono ancora da definire i target ed il loro meccanismo d'azione. Questo interrogativo è legato al fatto che questi elementi sono distanti anche centinaia di Kb dai loro geni bersaglio e talvolta localizzati anche in un diverso cromosoma (Spitz et al. 2003; Pennacchio et al. 2006; Marinic et al. 2013; Sagai et al. 2005). In quest'ultimo decennio un crescente numero di evidenze ha portato alla luce un nuovo concetto, quello della prossimità spaziale; infatti regioni anche molto distanti nel genoma possono essere portate in stretta vicinanza tra loro grazie allo stabilirsi di *loop* di DNA, cioè vere e proprie interazioni cromatiniche. Tali osservazioni suggeriscono come l'architettura del genoma giochi un ruolo chiave nel controllo dell'espressione genica (Fraser, Bickmore. 2007). Le interazioni possono essere sia intracromosomiche o in *cis,* cioè tra regioni del DNA situate sullo stesso cromosoma, che intercromosomiche o in *trans*, quindi su cromosomi diversi; esse non solo sono cruciali per eventi quali trascrizione ed espressione genica (Osborne et al.2004; Spilianakis et al. 2005; Osborne et al. 2007;), ma possono influenzare tanti altri processi nucleari come la ricombinazione (Skok et al. 2007) ed il silenziamento genico (Bantignies et al. 2003).

Nel 2002 Dekker e colleghi hanno sviluppato una nuova strategia chiamata "*Chromosome Conformation Capture*" (3C; Dekker et al. 2002), che ha dimostrato *in vivo* l'esistenza dei *loops* di cromatina tra elementi regolatori del DNA ed i loro geni target. Tale metodica ha permesso di analizzare la frequenza dei contatti tra regioni del genoma di interesse in una determinata popolazione cellulare ed in uno specifico stadio di sviluppo (De Wit, de Laat. 2012). Negli anni successivi si sono sviluppate altre tecniche correlate al 3C che hanno permesso lo studio della struttura del DNA superando quelli che erano i limiti della microscopia ottica e del 3C stesso (Figura 5).

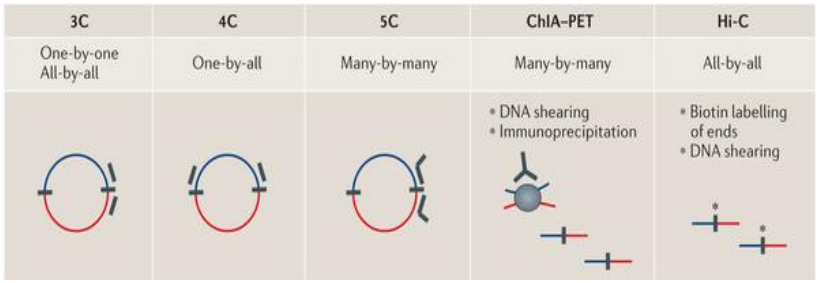

Figura 5: Rappresentazione schematica delle metodiche "3C": "3C", "4C", "5C", "ChIA-PET" e "Hi-C". Da: Dekker et al. Nat Rev Genet 390-403 (2013).

Le tecnologie “3C” nel loro complesso prevedono cinque passaggi comuni:

1) Il “*cross-linking*” del DNA con la formaldeide, un potente agente fissante che cattura le interazioni DNA-DNA *in vivo*.
2) La digestione enzimatica della cromatina con un enzima di restrizione opportunamente scelto.
3) La ligazione dei frammenti di DNA *cross-linkati* tra di loro.
4) La reversione del *cross-linking*.
5) Infine l’analisi delle frequenze di interazione che prevede strategie differenti a seconda dell’approccio utilizzato.

Tra le metodiche successive al 3C, il 4C (*Circular Chromosome Conformation Capture*; Zhao et al. 2006) per esempio mira ad aumentare la risoluzione dell’esperimento in quanto analizza le interazioni che intercorrono tra una specifica regione ed il resto del genoma (*one by all)*; l’Hi-C (*Genome-wide Chromosome Conformation Capture*; Lieberman et al. 2009) offre una visione *genome-wide* in quanto costruisce una matrice delle interazioni tra tutti i frammenti di restrizione lungo l’intero genoma (*all by all*); il 5C (*Chromosome Conformation Capture Carbon Copy*; Dostie et al. 2006) permette di mappare su larga scala anche centinaia di interazioni contemporaneamente (*many by many*); il Chia-Pet (*Chromatin Interaction Analysis by Paired-End Tag Sequencing*; Fullwood, Ruan. 2009) associa al 3C uno step di immunoprecipitazione per analizzare i contatti tra una proteina e le regioni del genoma che lega.

Nel complesso le tecnologie “3C” offrono importanti informazioni sulla struttura 3D del DNA e sulle interazioni cromatiniche, mirando non solo alla conferma della relazione tra specifiche regioni del DNA che altri studi avevano messo in luce, ma anche all’identificazione di nuovi elementi regolatori e target mai considerati essere correlati ad una determinata patologia/tratto quantitativo. I dati risultanti potranno poi essere validati attraverso studi funzionali indipendenti per mettere in luce quelli che sono i meccanismi d’azione alla base dei normali processi biologici, ma anche di quelli patologici.

3. SCOPO DELLA TESI

Malattie ereditarie dell'emoglobina come la β-talassemia e l'anemia falciforme pregiudicano la salute di innumerevoli persone. L'ultimo rapporto OMS sottolinea come ogni anno in tutto il mondo, per mancanza di trasfusioni e per i problemi associati al trattamento ferrochelante a lungo termine, muoiono circa 27.000 individui. La β-talassemia è stata definita dall'Unione Europea come una malattia rara, perché non colpisce più dello 0,05% della popolazione totale. Per questo motivo l'industria farmaceutica e le istituzioni pubbliche sono poco interessate allo sviluppo di una cura definitiva per la malattia. Viceversa l'Italia, con la presenza di quasi 6000 pazienti, è uno dei Paesi in cui la patologia presenta la più alta incidenza e dove, nonostante il miglioramento delle terapie tradizionali, una cura definitiva è fortemente necessaria e richiesta.

Un grande passo avanti in questa direzione è stato compiuto grazie agli studi GWAS, che partendo dalla consapevolezza che l'HbF rappresenta il principale modificatore della β-talassemia e anemia falciforme, hanno identificato due principali QTLs (*Quantitative Trait Loci*) implicati nella regolazione dei livelli di emoglobina fetale che mappano al di fuori del *cluster* β-globinico: la regione intergenica *HBS1L-MYB* nel cromosoma 6 e l'introne 2 del gene *BCL11A* nel cromosoma 2 (Menzel et al. 2007; Thein et al. 2007; Lettre et al. 2008; Uda et al. 2008).

Tuttavia i meccanismi molecolari alla base di queste associazioni geniche sono stati chiariti solo parzialmente, e la comprensione delle implicazioni funzionali e del ruolo che svolgono rimane ancora sconosciuta.

Sebbene gli studi GWAS abbiano permesso l'identificazione di geni implicati nella regolazione di numerosi tratti quantitativi/malattie complesse, spesso le varianti associate non sono quelle causative, ma marcatori in LD (*Linkage Disequilibrium*) con esse, presentano effetti di piccole dimensioni, sono difficili da valutare con test funzionali, e sono per la maggior parte localizzate in regioni del genoma non codificanti, complicando la comprensione del loro ruolo funzionale (Maurano et al. 2012).

Dati di letteratura evidenziano un arricchimento significativo di tali varianti in regioni del DNA che mostrano un'alta densità di marcatori cromatinici, suggerendo la presenza di tipici elementi genomici funzionali con una potenziale attività regolatrice (*enhancers,* promotori*, insulators* ecc) (Cookson et al. 2009).

Tra le varie ipotesi si ritiene che gli SNPs associati possano alterare elementi regolatori (per esempio modificando siti di legame per fattori di trascrizione) che agiscono a lunga distanza, *in cis* o *in trans,* modulando l'espressione dei geni bersaglio. L'avvento delle nuove tecnologie di "*Chromosome Conformation Capture*, 3C" ha reso possibile investigare *in vivo* l'organizzazione spaziale della cromatina e

rilevare la frequenza delle interazioni fisiche lungo il genoma tra elementi genetici con un ruolo chiave nella regolazione genica.

In tale contesto si è definito il mio progetto di ricerca, con l'obiettivo di verificare se le regioni contenenti gli SNPs associati alla regolazione dei livelli di HbF e al miglioramento della gravità clinica della β-talassemia, instaurino interazioni a lunga distanza *in cis* in grado di influenzare l'espressione di *HBS1L/MYB* e del *BCL11A*, mediante l'applicazione di tecniche "3C", in una linea cellulare con caratteristiche eritroidi (K562).

Uno dei laboratori più all'avanguardia nello studio delle tecniche "3C" e della struttura cromatinica è il laboratorio di "Nuclear Dynamics" del Babraham Insitute di Cambridge, diretto dal Dottor Peter Fraser, in cui ho trascorso parte del mio percorso di Dottorato. Durante la permanenza nel suo laboratorio la mia attività di ricerca si è articolata come segue:

1) Ottimizzazione del protocollo Hi-C. La metodica dell'Hi-C (Lieberman et al. 2009) è in grado di rivelare le interazioni che intercorrono tra regioni geniche lungo tutto il genoma senza ipotesi a priori. Se da un lato tale approccio permette una visione *genome wide,* che riflette la potenza della metodica, dall'altra però si evidenzia il limite dovuto alla bassa risoluzione (1 Mb). Per superare tale limite la prima parte del mio progetto ha previsto la messa a punto di alcune modifiche al protocollo seguito nel laboratorio Nuclear Dynamics.
2) Applicazione di una nuova metodica, sviluppata nel laboratorio del Dott.Fraser, denominata SCRiBL (sequence capture of regions interacting with bait loci), che si basa su una reazione di ibridazione in cui si ha l'arricchimento selettivo delle regioni di interesse, utilizzando come substrato la libreria derivata dall'Hi-C modificato. Questo approccio permette di superare nettamente la risoluzione dell'Hi-C risolvendo fino a poche Kb.
3) Validazione dello SCRiBL. La scelta di eseguire una nuova metodica ha previsto necessariamente la sua validazione. Il *cluster* β-globinico è quello che meglio si prestava a questo scopo. Infatti, oltre ad essere la regione genica maggiormente studiata è anche quella più frequentemente utilizzata nelle tecniche "3C", rappresentando quindi il controllo sperimentale ideale.
4) Analisi dei *loci HBS1L-MYB* e *BCL11A* mediante il programma SeqMonk . Lo scopo è stato quello di valutare se i frammenti contenenti le varianti geniche associate ai due *loci* in esame potessero essere coinvolti in interazioni a distanza regolando *in cis* l'espressione genica.

Questo progetto, che ha previsto per la prima volta l'utilizzo della nuova metodica SCRiBL, ci ha permesso con un unico esperimento di valutare simultaneamente le interazioni cromatiniche a lunga distanza lungo più *loci,* con una risoluzione maggiore di altre tecniche "3C" e senza conoscere a priori i siti di interazione.

Vista la rilevanza clinica della persistenza di emoglobina fetale nei soggetti affetti da emoglobinopatie, i risultati ottenuti contribuirebbero ad una maggiore comprensione del ruolo delle varianti geniche ai *loci HBS1L-MYB* e *BCL11A* e a chiarire il meccanismo d'azione alla base della persistente espressione dei geni γ

globinici nell'adulto. A lungo termine queste nuove conoscenze potrebbero fornire gli strumenti per la ricerca di strategie terapeutiche innovative ed alternative per la cura della β-talassemia.

4. MATERIALI E METODI

4.1 Coltura della linea cellulare K562

In questo progetto sono state utilizzate le cellule K562 (*Human erythromyeloblastoid leukemia cell line*), una linea di leucemia mieloide cronica umana in crisi blastica terminale; la coltura è avvenuta con terreno RPMI 1640 (Sigma), addizionato con 10% di Siero Fetale Bovino (FBS, Life Technologies), 1% di penicillina (100 U/ml), 1% di streptomicina (100 µg/ml) ed è stata mantenuta alla temperatura di 37°C in un incubatore (Hereaus HERA CELL150) ad atmosfera umidificata al 5% di CO_2.

4.2 Preparazione della libreria Hi-C

La metodica Hi-C che ho appreso nel laboratorio Nuclear Dynamics segue il protocollo "Takashi Nagano/Stefan Schoefender", una versione modificata del metodo originariamente descritto da Lieberman (Lieberman et al. 2009; Figura 6), con alcune successive modifiche apportate da Mayra Furlan Magaril.

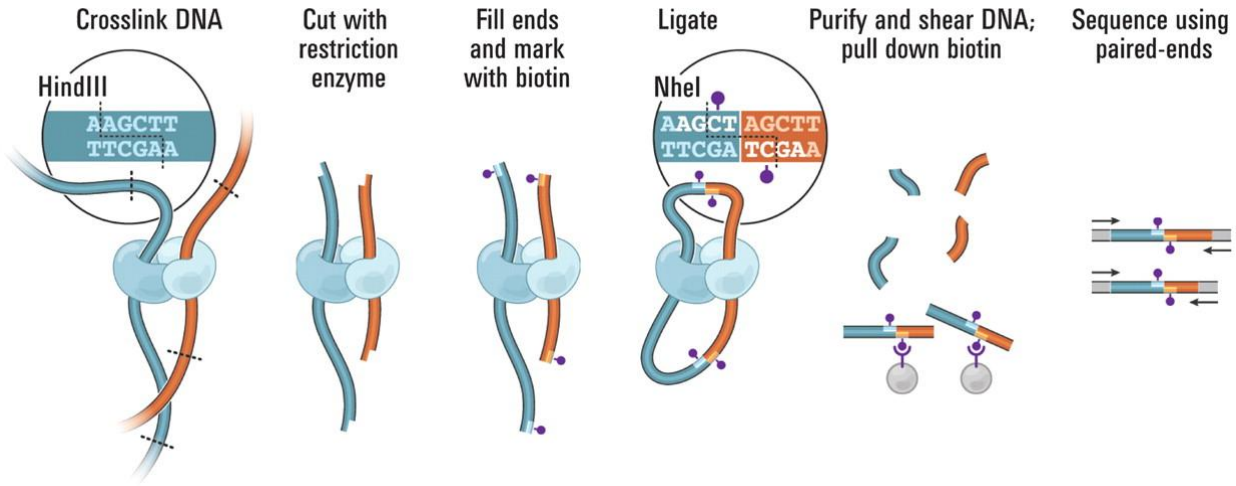

Figura 6: Rappresentazione dei passaggi cruciali per la preparazione della libreria Hi-C. Tratto da: "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome"; Lieberman et al.2009.

4.2.1 Fissaggio delle cellule

Le cellule K562 sono state contate con la camera di Burker (Knittelglaser 0.100 mm, Tiefe Depth 0.0025 mm^2) e suddivise in aliquote contenenti 30 milioni di cellule ciascuna.

Dopo centrifugazione il *pellet* è stato risospeso in 40 ml di terreno Dulbecco's modified eagle medium (DMEM, Life Technologies) supplementato con 10% di Siero Fetale Bovino (FBS, Life Technologies).

Le cellule sono state fissate con formaldeide al 2% (esente da acidi, Merck) e incubate 10 min a temperatura ambiente. La formaldeide è stata neutralizzata con l'aggiunta di 6 ml di Glicina fredda 1M (Life

Technologies) e incubazione di 5 min a temperatura ambiente. Le cellule sono state lavate con Phosphate-Buffered Saline freddo (PBS, Life Technologies), congelate in azoto liquido e conservate a -80°C.

4.2.2 Lisi cellulare e digestione enzimatica

Sono stati condotti due esperimenti ciascuno dei quali con un'aliquota di 30 milioni di cellule K562: il primo per testare una nuova condizione, la doppia digestione enzimatica, mentre il secondo, successivamente al primo, per eseguire l'esperimento Hi-C completo; il protocollo seguito è lo stesso, tranne in alcuni passaggi in cui è stato espressamente dichiarato.

Un *pellet* di 30 milioni di cellule K562 precedentemente fissate è stato risospeso in 50 ml di buffer di lisi (10 mM Tris-HCl pH8, 10 mM NaCl, 0,2% Igepal CA-630 (Sigma), 1 compressa di inibitore delle proteinasi (Complete EDTA-free, Roche)) e incubato in ghiaccio per 30 min. Successivamente la cromatina è stata centrifugata, risospesa in NEB buffer 2 1,25x e suddivisa in 5 aliquote. Per eliminare completamente i residui del buffer di lisi i tubi sono stati centrifugati ulteriormente e ciascun *pellet* è stato risospeso nel buffer specifico o di maggiore compatibilità dell'enzima/enzimi di restrizione da utilizzare nella fase seguente. Per rimuovere le proteine che non sono state fissate al DNA è stato addizionato sodio dodici solfato ad una concentrazione finale dello 0.3% (SDS, Invitrogen), i campioni sono stati incubati a 37°C per 60 min e l'azione dell'SDS è stata interrotta con l'aggiunta di Triton X-100 per una concentrazione finale del 2% (Sigma). Per il test di controllo della doppia digestione enzimatica sono state valutate 5 condizioni:

1- Digestione con HindIII (100 U/µl, New England Biolabs).
2- Digestione con BglII (50 U/µl, New England Biolabs).
3- Digestione con entrambi gli enzimi addizionati contemporaneamente.
4- Digestione con entrambi gli enzimi addizionati in due tempi, prima HindIII e dopo 9 ore BglII.
5- Nessuna digestione (controllo negativo o controllo non digerito).

Nell'Hi-C completo invece tutte le aliquote sono state digerite con HindIII e BglII contemporaneamente. In tutte le condizioni sono state usate 1500 unità di ciascun enzima e l'incubazione è avvenuta a 37°C overnight a 950 rpm.

4.2.3 Biotinilazione delle estremita' di DNA e ligazione

Le estremità dei frammenti ottenuti sono protrudenti (*sticky)* sia in 3' che in 5', per cui per renderle *blunt* è stato eseguito un "*fill-in*" utilizzando oltre ai deossinucleotidi standard dATP biotinilata (Biotin-14-dATP, Invitrogen) e 50 unità dell'enzima Klenow (DNA polymerase I large fragment, New England Biolabs) incubando a 37°C per 75 min.

Per la ligazione invece che procedere secondo il protocollo originale e portare la cromatina in soluzione, è stata attuata la ligazione "*in nuclei*", evitando la rottura della membrana nucleare con l'SDS e procedendo

direttamente alla reazione con 50 unità di enzima DNA Ligase T4 (1 U/µl, Invitrogen), incubando overnight a 16°C. Il giorno successivo i campioni sono stati trattati overnight con la Proteinase K (Roche) per revertire il *cross-linking*.

4.2.4 Purificazione del DNA

Successivamente al trattamento con RNase A (Roche) è stata effettuata un'estrazione con fenolo pH 8 1:1 (Sigma) per rimuovere i lipidi e le proteine dai prodotti di ligazione. La fase superiore contenente il DNA è stata poi estratta con fenolo:cloroformio (Sigma) 1:1. Per precipitare il DNA sono stati aggiunti 0,1x di sodio acetato pH 5,2 3M (Sigma) e 2,5x di etanolo assoluto (VWR Chemicals).
I *pellets* di DNA, risospesi in TE (10 mM Tris-HCl, 0.1 mM EDTA), sono stati estratti ancora con fenolo:cloroformio (Sigma), il DNA è stato precipitato come descritto sopra, lavato con etanolo al 70%, risospeso in TE e purificato con le colonne Amicon (Amicon Ultra 0,5 Centrifugal Filter Devices 30 K, Millipore) secondo le istruzioni fornite dal produttore.

4.2.5 Quantificazione col PicoGreen Assay

Le librerie di DNA non possono essere accuratamente quantificate tramite misurazione dell'assorbanza a 260 nm, in quanto la presenza di elevate quantità di ATP e DTT nel buffer utilizzato per la ligazione potrebbero portare alla co-precipitazione con il DNA. Per una quantificazione più precisa è stato usato il Quant-it PicoGreen ds DNA Assay kit (Invitrogen); il sistema utilizza un fluoroforo che si lega al DNA ad alta affinità rilasciando fluorescenza, la cui intensità aumenta con l'aumentare del legame al DNA, quindi della quantità di DNA presente. Il kit fornisce un DNA standard (fago λ) dal quale vengono fatte delle diluizioni seriali da 2 pg/µl a 2ng/µl col buffer TE 1x fornito dal kit. Anche il PicoGreen stesso viene diluito 1:200 in TE 1x prima di essere miscelato con i campioni da misurare (sia la curva standard che i campioni stessi) in un rapporto di 1:1. La fluorescenza è stata misurata nel lettore multipiastra Cytofluor II (PerSeptive Biosystem) con un'eccitazione di lunghezza d'onda pari a 485 nm ed un'emissione di 530 nm, e così è stata evinta la quantità di DNA.

4.2.6 Controlli di efficienza dell'Hi-C

Prima di poter procedere col resto dell'esperimento sono stati eseguiti dei controlli per valutare la qualità della libreria in quanto i passaggi iniziali di fissaggio, digestione, *fill-in* e ligazione sono molto critici.
Per prima cosa sono state eseguite delle PCR per rilevare le cosiddette "*short*" e "*long range interactions*", interazioni che si stabiliscono tra frammenti di restrizione adiacenti e ad una distanza maggiore di 20 kb, rispettivamente; per quanto riguarda le "*long range*" sono stati scelti due frammenti che distano 1,5 Mb e

che mappano ciascuno all'interno dei due *clusters* istonici nel cromosoma 6, le cui interazioni sono state riscontrate precedentemente in laboratorio e riportate in letteratura (Rastegar et al. 2004).
Avendo eseguito una doppia digestione enzimatica si avranno sia frammenti tagliati con HindIII che con BglII, per cui è necessario testare tutte le possibili combinazioni di prodotti: ligazione tra frammenti di restrizione HindIII/HindIII, BglII/BglII ed infine HindIII/BglII. I *primers* utilizzati sono orientati verso il sito di taglio e sono elencati nella Tabella 1.

ID PRIMER	SEQUENZA PRIMER	INTERAZIONE	DIMENSIONE AMPLICONE
HIST1-HIND-1F HIST1-HIND-2F	ttaagccaacccagttgtcc aagcaggaaaaggcatagca	SHORT RANGE HINDIII/HINDIII	376
HIST1-BGL-1F HIST1-BGL-2F	tcaagccaacaccttgacac tcttcgggaggaaaatcctt	SHORT RANGE BGLII/BGLII	96
HIST1-BGL-2F HIST1-HIND-1F	tcttcgggaggaaaatcctt ttaagccaacccagttgtcc	SHORT RANGE HINDIII/BGLII	145
HIST1-BGL-1R HIST1-BGL-4R	acgcatcaacatctcagcag atcatttgctgcttgggatg	SHORT RANGE BGLII/BGLII	364
HIST1-HIND-2F HIST2-HIND-2F	aagcaggaaaaggcatagca tcttgggttgtgggactttc	LONG RANGE HINDIII/HINDIII	395
HIST1-BGL-1R HIST2-BGL-3F	acgcatcaacatctcagcag gtgcaaggcagtgttgaaga	LONG RANGE BGLII/BGLII	283
HIST1-HIND-2F HIST2-BGL-3F	aagcaggaaaaggcatagca gtgcaaggcagtgttgaaga	LONG RANGE HINDIII/BGLII	397

Tabella 1: Elenco dei *primers* utilizzati per i controlli di interazione a breve distanza (*short range*) e a lunga distanza (*long range*) per le varie combinazioni di frammenti: HindIII/HindIII, BglII/BglII e HindIII/BglII.

Per ogni reazione abbiamo addizionato 250 ng di DNA alla miscela finale utilizzando il kit Hot Start Taq Polymerase (Qiagen) ed i parametri seguiti per le PCR sono stati i seguenti:

95°C	15 min	
60°C	1 min	36 cicli
72°C	1 min	
94°C	30 sec	
60°C	2 min	
72°C	10 min	

Nell'ultimo test di controllo gli ampliconi derivanti dalle PCR per le interazioni a breve distanza dei prodotti di ligazione HindIII/HindIII e BglII/BglII sono stati digeriti con HindIII e/o NheI, e con BglII e/o ClaI, rispettivamente, per valutare la proporzione della libreria che è stata resa *blunt* e biotinilata ed avere una prima indicazione anche sulla digestione enzimatica (Tabelle 2 e 3). Gli enzimi NheI e ClaI vengono utilizzati in quanto la ligazione tra due frammenti HindIII/HindIII e BglII/BglII rispettivamente porta alla costituzione del sito di taglio per tali enzimi nel punto di giunzione. Le condizioni sono riportate più in dettaglio in Tabella 2 nel caso dei prodotti HindIII/HindIII e in Tabella 3 per BglII/BglII.

	HINDIII	**NHEI**	**HINDIII/NHEI**	**NON DIGERITO**
DNA (µl)	5,5	5,5	5,5	5,5
NEB Buffer (µl)	1,5	1,5	1,5	1,5
Enzima/i (µl)	1	1	0,75 per ciascun enzima	-
BSA 10X (µl)	1,5	1,5	1,5	1,5
Acqua (µl)	5,5	5,5	6	6,5

Tabella 2: Digestioni enzimatiche sul prodotto dell'interazione a breve distanza HindIII/HindIII per il controllo di ligazione e biotinilazione della libreria Hi-C. I valori sono espressi in µl.

	BGLII	**CLAI**	**BGLII/CLAI**	**NON DIGERITO**
DNA (µl)	5,5	5,5	5,5	5,5
NEB Buffer (µl)	1,5	1,5	1,5	1,5
Enzima/i (µl)	1	1	0,75 per ciascun enzima	-
BSA 10X (µl)	1,5	1,5	1,5	1,5
Acqua (µl)	5,5	5,5	6	6,5

Tabella 3: Digestioni enzimatiche sul prodotto dell'interazione a breve distanza BglII/BglII per il controllo di ligazione e biotinilazione della libreria Hi-C.I valori sono espressi in µl.

4.2.7 Calcolo della percentuale di digestione

Sia nell'esperimento di controllo della doppia digestione che nell'Hi-C completo è stata effettuata una q-PCR col fluorescente SYBR Green sull'Abi Prism 7000 SD System (Applied Biosystems) per calcolare la percentuale di efficienza delle varie digestioni enzimatiche. I parametri di reazione sono stati i seguenti:

50°C	2 min	
95°C	10 min	
95°C	30 sec	40 cicli
55°C	30 sec	
62°C	1 min	

Sono state utilizzate tre diverse coppie di *primers*, la prima orientata su un sito di restrizione HindIII, la seconda su un sito BglII e la terza non conteneva alcun sito di restrizione (Tabella 4).

NOME PRIMER	**SEQUENZA**	**CARATTERISTICA**
HS-CDKN1A-HINDIII-A-F	tcacagcatttcctcactgc	Orientati verso un sito HindIII
HS-CDKN1A-HINDIII-A-R	agccattgtggaggacaagt	
HS-CDKN1A-BGLII-A-F	ggaggatgcctctcttcaaa	Orientati verso un sito BglII
HS-CDKN1A-BGLII-A-R	tgttggtcatcacacctgct	
HS-CDKN1A-CONTROL-A-F	ttaaggcttaggtctggaatgg	Non orientati verso siti HindII/BglII
HS-CDKN1A-CONTROL-A-R	tttgtagttgcctccccttg	

Tabella 4: Elenco dei *primers* utilizzati per la q-PCR di controllo della percentuale di digestione degli enzimi HindIII e BglIII sulla libreria Hi-C.

4.2.8 Rimozione della biotina dalle estremità dei frammenti non ligati

L'esperimento per il controllo di efficienza della doppia digestione è terminato con la q-PCR, mentre da questo punto in poi è stato portato avanti solo l'Hi-C completo. Non tutti i frammenti biotinilati sono stati ligati, per cui per evitare di selezionare in seguito questo materiale non specifico, la dATP biotinilata presente è stata rimossa sfruttando l'attività esonucleasica della DNA Polimerasi T4. L'attività dell'enzima è stata arrestata con l'aggiunta di EDTA 0,5 M pH 8 (Invitrogen). I campioni sono stati successivamente purificati mediante estrazione fenolo:cloroformio seguita da precipitazione con sodio acetato ed etanolo.

4.2.9 Frammentazione del DNA e riparazione delle estremità

Il DNA è stato frammentato ulteriormente mediante sonicazione (Covaris E220) per ottenere dei frammenti aventi come picco 400 bp. I parametri impostati sono stati i seguenti:

Duty Factor	10%
Peak Incident Power	140
Cycles per burst	200
Time (sec)	55

La sonicazione ha portato ad una rottura delle molecole di DNA in punti casuali generando delle estremità protrudenti, che sono state rese *blunt* sfruttando l'attività esonucleasica 3'-5' della T4 DNA Polimerasi (che rimuove la protrusione in 3') e l'attività polimerasica della DNA Polimerasi I Klenow (che riempie la protrusione in 5'); contemporaneamente è stato addizionato anche l'enzima T4 polinucleotide chinase per addizionare un gruppo fosfato all'estremità in 5'. I campioni sono stati purificati (QiaQuick PCR purification kit, Qiagen) ed eluiti in TE. Per permettere la ligazione degli adattatori Illumina più avanti nel protocollo, è stata sfruttata l'attività esonucleasica dell'enzima Klenow per adenilare le estremità 3' dei frammenti sonicati e "riparati".

4.2.10 Selezione dei frammenti in base alle dimensioni con le biglie magnetiche

Per raggiungere una misura dei frammenti più uniforme e compatibile con il sequenziamento (dalle 200 alle 700 paia di basi) mediante piattaforme di nuova generazione (NGS) Illumina, è stata effettuata una selezione in base alle dimensioni mediante il sistema Agencourt AMPure XP beads (Beckman Coulter), sfruttando la tecnologia delle biglie paramagnetiche in fase solida (SPRI), seguendo il protocollo standard (Figura 7).

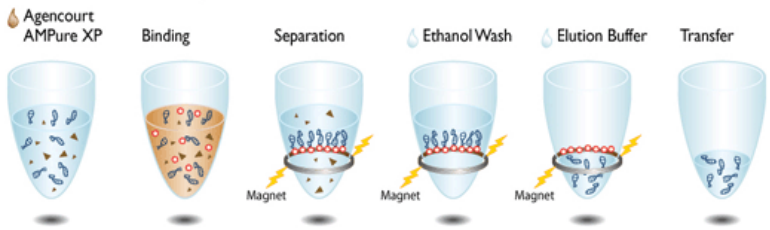

Figura 7: Schema riassuntivo dei vari passaggi effettuati con le Agencourt AMPure XP beads per la selezione dei prodotti di ligazione in base alle dimensioni.

Più in dettaglio la selezione è avvenuta in due passaggi utilizzando diverse concentrazioni di biglie per rimuovere frammenti di diverso peso molecolare; nel primo passaggio con 0,6 volumi di biglie sono stati rimossi i frammenti più grandi, mentre nel secondo con 0,95 volumi sono stati eliminati quelli di più basso peso molecolare.

La concentrazione della libreria è stata misurata col Picogreen Assay (citato precedentemente).

4.2.11 "Pull-down" dei frammenti biotinilati e ligazione degli adattatori alle estremità

Le molecole di DNA biotinilate sono state ibridate alle Dynabeads MyOne Streptavidin C1 (Invitrogen), biglie magnetiche contenenti la streptavidina per consentire così la "cattura" dei prodotti di ligazione.

Le biglie sono state prima lavate e poi combinate con la libreria Hi-C. Il DNA che aderisce alle biglie è stato lavato con No Tween Buffer 1X (NTD, Tris-HCl 5 mM, EDTA 0,5 mM, NaCl 1M), con T4 DNA Ligase Reaction Buffer 1x (New England Biolabs) ed infine risospeso in quest'ultimo.

La libreria Hi-C è ancora legata alle biglie e tale condizione migliora l'efficienza di ligazione degli adattatori decrementando la mobilità dei frammenti di DNA e facilitando la rimozione degli oligonucleotidi non legati. Gli adattatori "Illumina TruSeq DNA Adapters" conferiscono un'identità specifica alla libreria, una sorta di codice a barre unico capace di discriminare il campione nel caso in cui venisse sequenziato con altri nella stessa *lane*. La reazione di ligazione è avvenuta addizionando l'adattatore numero 6 e l'enzima T4 DNA Ligase (New England biolabs) alla libreria Hi-C, incubando 2h a temperatura ambiente seguendo le istruzioni del produttore.

4.2.12 Amplificazione della libreria Hi-C

Al termine della ligazione il campione è stato lavato ed infine risospeso in NEB buffer 2 1x (New England Biolabs).

L'amplificazione della libreria è fondamentale per generare un quantitativo di molecole di DNA tale da permettere il sequenziamento e l'esecuzione dello SCRiBL; è altrettanto importante utilizzare il minor numero di cicli possibile per assicurare un'amplificazione lineare senza creare artefatti tipici della PCR. Perciò sono state eseguite delle amplificazioni di controllo per riuscire ad identificare l'esatto numero di cicli necessario utilizzando i *primers* TruSeq PCR 1.0 e TruSeq PCR 2.0 (Tabella 5) poiché compatibili con le estremità delle molecole di DNA della libreria a cui sono stati ligati gli adattatori.

NOME PRIMER	SEQUENZA
TruSeq PCR Primer 1.0 (100µM)	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3'
TruSeq PCR Primer 2.0 (100 µM)	5'-CAAGCAGAAGACGGCATACGAGAT'-3'

Tabella 5: Elenco dei due *primers* Illumina TruSeq utilizzati per amplificare la libreria Hi-C.

Il programma seguito è stato il seguente:

98°C	30 sec	
98°C	10 sec	x N° cicli
65°C	30 sec	
72°C	30 sec	
72°C	7 min	

I cicli testati sono stati il 7, l'8, il 9 e il 12. Dopo aver determinato la condizione ottimale, l'intera libreria è stata amplificata in reazioni multiple e purificata per rimuovere i reagenti residui della PCR con 1,8x di AMPure XP beads. Dopo un lavaggio in etanolo (VWR Chhemicals) al 70% le biglie sono state risospese in TE ed è stato infine possibile recuperare con estrema attenzione il surnatante pulito ovvero la libreria Hi-C finita.

4.2.13 Profilo del Bioanalyzer e quantificazione della libreria Hi-C col kit Kapa SYBR

Mediante saggio elettroforetico sul Bioanalyzer 2100 (Agilent Technologies) è stata determinata sia la concentrazione della libreria che l'esatta distribuzione dei frammenti in base alle dimensioni. E' stata necessaria una validazione anche mediante PCR quantitativa (qPCR) con il kit Kapa SYBR fast qPCR (Kapa Biosystem) seguendo le istruzioni indicate dal produttore.

4.2.14 Sequenziamento della libreria Hi-C

Per il sequenziamento genico sono state utilizzate le tecnologie di nuova generazione (Next generation Sequencing, NGS) con la piattaforma Illumina che prevede l'ibridazione della libreria su una superficie chiamata *flow cell* grazie alla presenza di specifici adattatori alle estremità di ciascuna molecola di DNA della libreria. Quest'ultima viene amplificata mediante *bridge amplification* e formazione di *clusters* contenenti milioni di frammenti di DNA a doppio filamento.

La libreria Hi-C è stata sequenziata condividendo con un altro campione una lane della *flow cell* mediante la piattaforma HiSeq 2000 Illumina e corsa *paired-end* da 50 paia di basi. Il termine *paired-end* indica l'abilità del sequenziatore di leggere entrambe i filamenti delle molecole di DNA producendo una sequenza corta di DNA, 50 bp in questo caso, chiamata *read*.

4.3 Preparazione delle bait ad RNA per l'arricchimento della libreria Hi-C

Per poter individuare le interazioni all'interno di regioni di DNA al di sotto di un megabase, l'Hi-C non è la metodica ideale; per superare tale limitazione è stato ideato nel laboratorio Nuclear Dynamics un sistema di arricchimento della libreria Hi-C non ancora pubblicato, chiamato SCRiBL, che abbiamo eseguito per il

mio progetto di tesi con l'obiettivo di aumentare la risoluzione per il rilevamento delle interazioni cromatiniche attraverso l'ibridazione con piccoli frammenti ad RNA (bait). Il protocollo utilizzato per la preparazione di queste ultime è stato creato al Babraham Institute e discusso di seguito.

4.3.1 Preparazione del DNA dei cloni BAC

Il substrato utilizzato per la costruzione delle bait ad RNA, necessarie per l'esecuzione dell'esperimento SCRiBL, è rappresentato dal DNA genomico veicolato in cloni BAC (Bacterial artificial chromosomes) corrispondente alle regioni di nostro interesse: il *locus* β-globinico, il gene *BCL11A* e la porzione intergenica *HBS1L-MYB*. Oltre a questi ne sono stati inseriti altri per meglio valutare l'efficienza della nuova metodica SCRiBL. I BAC forniti dall'Invitrogen sono elencati nella tabella seguente con le relative coordinate dei frammenti del genoma che veicolano.

NOME BAC	INIZIO	FINE	CROMOSOMA
RP11-104D9 *Hbs1l-Myb*	**135366522**	**135547346**	**6**
RP11-65A9 *Bcl11a*	**60612928**	**60789817**	**2**
CTD-2063A20 *Locus* β-globinico	**5243054**	**5329081**	**11**
RP11-135I7	60789970	60959188	2
CTC-215I11	104266	263351	16
RP11-1013N13	26082711	26261136	6
CTD-2536K9	27146920	27354198	7
CTD-3054H22	27025000	27146421	7
CTD-2545C21	51770709	51961472	2

Tabella 6: Elenco dei cloni BAC utilizzati per la preparazione delle *bait* ad RNA. In grossetto sono evidenziati i *loci* di interesse: β-globinico, *BCL11A* e *HBS1L-MYB.*

Il DNA veicolato dai cloni BAC è stato estratto, in seguito a coltura batterica, con il kit NucleoBond BAC100 (Macherey Nagel) seguendo le istruzioni indicate dal produttore. Per confermare la corretta identità dei cloni, sono state eseguite delle PCR utilizzando dei *primers* disegnati alle estremità di ciascuna sequenza (Tabella 7).

IDENTITA' BAC	PRIMERS 1° ESTREMITA'	PRIMERS 2° ESTREMITA'
RP11-104D9	ttctccttttagaccaccctca gtggatggtattgcatgtgg	ccagggagaacctcagatca ccgaatggaaacaaatcctg
RP11-65A9	ctccaggcacatgaacaaaa gctgcctctggaagattcac	ggcactgggtagaattcaagtt tcatagctggaagttggcaat
CTD-2063A20	taactgcagagccagaagca gtctgctttggaaggactgg	tgtgacactcgcagcaagtta ccaagccgtcacttcttagc
RP11-135I7	tggtaaccacgtgcctgata gttagcccaggctgctattg	atcatgccactcacactcca agccaaggtcactccacttc
CTC-215I11	atggactacggcacttccac ttgggtgtactcttccagca	tctactgccctctcccttca tctttgtgctggtgctgtct
RP11-1013N13	atgaaggctgtggaaggtgt gggaaagtagcttcgcaatg	cttccagacactccgttcct cctaccaaattgcagcttcc
CTD-2536K9	tgaccagcgaatgcatagag agatggccaatctgctgaac	gacaccattcagcaccttcc ctcctggtgaaaggacagga
CTD-3054H22	cacagcactctgctggactc gcatatagtggcagcagctc	ccagttcagtcctccgtttg aatcggactcctggtctcct
CTD-2545C21	ttaagcctgcagcacttttt caaccaagcaactggaggtt	accaccctactgctgtgagc tcatgctgcctttatgttgtg

Tabella 7: Elenco delle sequenze dei *primers* utilizzati per il controllo di entrambe le estremità delle sequenze contenute nei cloni BAC.

4.3.2 Digestione enzimatica dei DNA dei cloni BAC

Il DNA dei vari BAC è stato collezionato in quantità equimolari in tre tubi, ciascuno contenente 20 µg in totale. Nel primo tubo il DNA è stato digerito con 400 unità di HindIII (100 U/µl, New England Biolabs), mentre nel secondo con 400 unità di BglII (50 U/µl, New England Biolabs), e nell'ultimo, rappresentando il controllo negativo, non è stato addizionato alcun enzima. La digestione è stata incubata a 37°C overnight.
Una piccola aliquota di ciascuna miscela è stata controllata su gel d'agarosio allo 0,8% per avere una conferma sull'avvenuta digestione. Il DNA è stato estratto mediante l'utilizzo delle colonne Phase Lock Gel (5prime) con un uguale volume di fenolo:cloroformio (Sigma), precipitato con 0,1 volumi di sodio acetato 3M pH 5,2 (Invitrogen) e 2,5 volumi di etanolo al 100%. I *pellets* sono stati lavati con etanolo (VWR Chemicals) al 70%, risospesi in Tris-HCl 10 mM pH 7,5 (Sigma) ed il DNA quantificato al Nanodrop.

4.3.3 Ligazione degli adattatori alle estremità dei frammenti

Alla miscela di sequenze di DNA ottenute dalla digestione sono stati legati due adattatori contenenti la sequenza universale del T7 *promoter,* preparati precedentemente mediante addizione di due oligonucleotidi 100 µM, di cui uno universale ed uno specifico per l'enzima di restrizione (Tabella 8) e il buffer di "*annealing*" (10mM Tris, pH8.0; 50mM NaCl; 1mM EDTA), per ottenere una concentrazione finale

di adattatori di 20 µM. Questi ultimi differiscono per le estremità, una recante il sito di taglio per l'enzima HindIII, che quindi si legherà ai frammenti digeriti con tale enzima, e l'altra presenta quello per BglII.

ADAPTER PRIMER	SEQUENZA
LuoT7 HindIII/BglII	tctagtcgacggccagtgaattgtaatacgactcactatagggcga
LuoT7HindIII rc	[Phos]agcttcgccctatagtgagtcgtattacaattcactggccgtcgactaga[SpcC3]
LuoT7 BglII rc	[Phos]gatctcgccctatagtgagtcgtattacaattcactggccgtcgactaga[SpcC3]

Tabella 8: Elenco delle sequenze utilizzate per la costituzione degli adattatori T7 HindIII e BglII. Essi sono HPLC purificati e presentano l'estremità in 5' col gruppo fosfato libero in corrispondenza del sito di restrizione HindIII (seconda riga) e BglII (terza riga) sottolineati.

La reazione di ligazione agli adattatori è avvenuta utilizzando 10 µg di DNA, un eccesso molare degli adattatori, 800 unità dell'enzima T4 DNA Ligase (New England Biolabs) e incubando i campioni a 16°C per 16h.

4.3.4 Sonicazione del DNA e riparazione delle estremità

I campioni sono stati frammentati ulteriormente mediante sonicazione (Covaris E220) per ottenere una dimensione dei frammenti con un picco di 200 bp, impostando i seguenti parametri:

Duty Factor	10%
Peak Incident Power	175
Cycles per burst	200
Time (sec)	180
Temperature	7°C

Prima e dopo la sonicazione sono state prelevate delle aliquote e valutate mediante corsa su gel d'agarosio al 2% per confermare o meno l'avvenuta frammentazione.
Per riparare le estremità dei frammenti è stata sfruttata l'attività di tre enzimi: la T4 DNA Polimerase (New England Biolabs), la T4 DNA Polinucleotide chinase (New Englan Biolabs) ed il Klenow (DNA Polimerase I large fragment, New England Biolabs).

4.3.5 Purificazione e selezione dei frammenti in base alle dimensioni

I frammenti sono stati successivamente selezionati in base alle dimensioni con le Agencourt AMPure XP beads (Beckman Coulter) in due passaggi successivi, come è avvenuto per la libreria Hi-C.

4.3.6 Trascrizione *in vitro* con UTP biotinilato e purificazione

Per la trascrizione *in vitro*, che converte il DNA in RNA, è stato utilizzato il kit MEGAscript T7 (Ambion), secondo le istruzioni del produttore, con l'aggiunta di UTP biotinilato (Biotin-16-UTP, Roche) che permette di marcare le piccole sequenze con la biotina. Il kit utilizza l'RNA polimerasi T7 per poter sintetizzare molecole di RNA a singolo filamento sfruttando la capacità dell'enzima di riconoscere la sequenza consenso del promotore T7 incorporata negli adattatori inseriti alle estremità dei frammenti di DNA. Il giorno successivo per eliminare ogni residuo di DNA alla reazione sono state addizionate 2 unità di Turbo DNase (Ambion) e incubata a 37°C per 15 min. Per inattivare quest'ultimo enzima è stata utilizzata la stessa quantità di EDTA 0,5 M (Invitrogen). Le bait ad RNA sono state quindi purificate da tutti i reagenti residui dalla trascrizione *in vitro* con il kit MEGAclear (Ambion) e quantificate mediante Nanodrop.

Per valutare l'incorporazione della biotina è stato incubato il fluoroforo "Alexa Fluor-647 streptavidin" (Life Technologies) coniugato con la streptavidina con un'aliquota di baits HindIII e separatamente con un'aliquota di bait BglII. I campioni sono stati caricati su gel d'agarosio all'1,5% e visualizzati al ChemiDoc XRS System sia per confermare la corretta dimensione delle bait che l'incorporazione della biotina.

4.4 SCRiBL (Sequence Capture of Regions interacting with Bait Loci)

L'esperimento SCRiBL prevede l'ibridazione ("capture") tra libreria Hi-C e le baits ad RNA a singolo filamento biotinilate, per arricchire selettivamente il *cluster* β-globinico, la regione intergenica *HBS1L-MYB* e il *locus BCL11A* di nostro interesse. Questo sistema è stato sviluppato per incrementare l'abilità nel rilevare interazioni significative, focalizzando l'attenzione su un gruppo di *loci* piuttosto che sull'intero genoma.

4.4.1 Ibridazione

La preparazione della reazione di ibridazione, considerata la parte centrale dell'esperimento, si suddivide in tre fasi:

1) Preparazione della libreria Hi-C.
2) Preparazione del buffer di ibridazione.
3) Preparazione delle bait HindIII-BglII.

Durante la fase 1 500 ng della libreria Hi-C sono stati concentrati mediante pompa a vuoto SpeedVac ed il *pellet* è stato risospeso in una miscela di reagenti costituita da: salmon spearm DNA (Ambion), Cot-1 Human DNA (Invitrogen) e una miscela composta da TruSeq *blocker index* 6 e TruSeq *universal blocker* (Tabella 9)*;* i bloccanti TruSeq sono corti oligonucleotidi a doppia elica capaci di appaiarsi alle estremità di DNA dove sono presenti gli adattatori per impedire il legame con sequenze aspecifiche.

BLOCKER ID	**SEQUENZA**
TruSeq universal blocker	aatgatacggcgaccaccgagatctacactctttccctacacgacgctcttccgatct
TruSeq blocker index6	caagcagaagacggcatacgagatattggcgtgactggagttcagacgtgtgctcttccgatc

Tabella 9: Sequenze utilizzate per bloccare le estremità dei frammenti di DNA della libreria Hi-C per evitare di legare molecole *off-target* aspecifiche.

Nel secondo passaggio è stato preparato il buffer di ibridazione addizionando i seguenti reagenti: 11,15x di SSPE (Gibco), 11,15x di Denhardt's (Invitrogen), 11,15 mM di EDTA (Gibco) e 0,223% di SDS (Promega).
Infine le bait finali sono state preparate miscelando le baits HindIII e BglII per ottenere una concentrazione finale di 500 ng. Per stabilizzarle è stata addizionata SUPERase-In (30 U; Ambion).
Per la reazione di ibridazione è stata impostato il termociclatore (PTC-200, MJ Research) col seguente programma:

STEP 1	95°C	5 min
STEP 2	65°C	24 h

La procedura è stata eseguita molto rapidamente in modo da tenere il coperchio dello strumento aperto il meno possibile ed evitare un'evaporazione eccessiva che potrebbe compromettere la resa dell'esperimento.
Per prima cosa è stato trasferito il tubo contenente il DNA nella parte centrale della griglia dell'amplificatore (Figura 8; quadrato rosso) ed il programma è iniziato. Quando la temperatura ha raggiunto i 65°C è stato inserito il buffer di ibridazione nella parte superiore (Figura 8 quadrato blu); dopo 5 min è stato trasferito anche l'RNA nella parte più in basso (Figura 8; quadrato verde) e lasciato per 2 min (Figura 8).

Figura 8: Schema rappresentante la posizione di DNA, bait ad RNA e buffer di Ibridazione nell'amplificatore per la reazione di ibridazione dello SCRiBL.

La reazione è avvenuta nel modo seguente: il buffer di ibridazione è stato combinato alle *bait* ad RNA (Figura 9; blu nel verde, step 1); immediatamente la libreria Hi-C è stata addizionata alla miscela buffer di ibridazione-RNA bait (Figura 9; rosso nel verde, step 2) ed il programma è proseguito a 65°C per 24h.

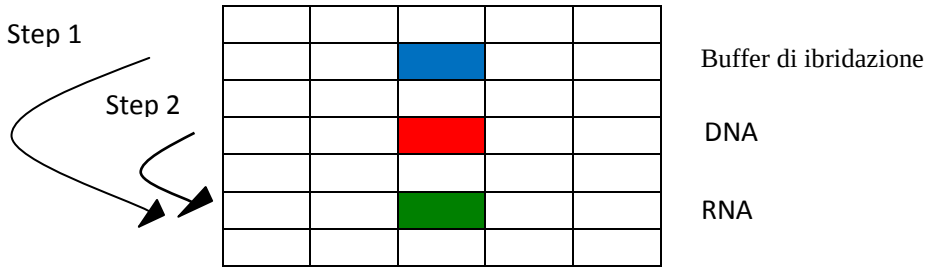

Figura 9: Schema rappresentante i passaggi della reazione di ibridazione dello SCRiBL.

4.4.2 "Pull-down" dei frammenti biotinilati

Gli ibridi DNA-RNA biotinilato che si sono generati sono stati recuperati ("*pull-down*") sfruttando la capacità delle biglie magnetiche Dynabeads MyOne Streptavidin T1 (Life Technologies) rivestite con la streptavidina di legarsi alla biotina incorporata alle bait. Le biglie con il DNA legato sono state lavate con Wash Buffer 1 (SSC 1x, SDS 0,1%) e Wash Buffer 2 (SSC 0,1x, SDS 0,1%). Infine la libreria SCRiBL è stata recuperata dalle biglie e risospesa in NEB buffer 2 1x (New England Biolabs).

4.4.3 Amplificazione e purificazione dei prodotti di ibridazione

Come nell'Hi-C la libreria SCRiBL deve essere amplificata per produrre materiale sufficiente per il sequenziamento ma utilizzando il minor numero di cicli per evitare di creare molti artefatti durante la PCR; sono stati testati i cicli 6, 7, 8 e 12. Una volta definita la condizione ottimale di PCR l'intera libreria "arricchita" è stata amplificata in reazioni multiple. I prodotti sono stati collezionati in unico tubo, purificati con 1,8x di Agencourt AMPure XP beads (Beckman Coulter) e infine abbiamo così ottenuto la libreria finale SCRiBL.

4.4.4 Profilo del Bioanalyzer e quantificazione della libreria SCRiBL col kit Kapa SYBR

La validazione della libreria SCRiBL è avvenuta allo stesso modo di quella Hi-C mediante saggio elettroforetico sul Bioanalyzer 2100 e q-PCR con il Kapa SYBR kit (Kapa Biosystem). Da un lato la libreria è stata valutata quantitativamente per ottenere con massima precisione la concentrazione, e dall'altra qualitativamente, in quanto il sequenziamento richiede un media di dimensione dei frammenti tra le 200 e 700 bp.

4.4.5 Sequenziamento della libreria SCRiBL

La libreria SCRiBL è stata inserita come unico campione in una *lane* della *flow-cell* e sequenziata con la piattaforma MiSeq Illumina in una corsa *paired-end* da 100 paia di basi in accordo con le istruzioni del produttore.

4.5 Analisi dei dati

4.5.1 L'algoritmo HiCUP

I dati ottenuti dal sequenziamento delle librerie Hi-C e SCRiBL sono stati sottoposti ad un controllo di qualità FastQC, per verificare la presenza di eventuali contaminazioni.

Successivamente sono stati processati mediante l'utilizzo di un algoritmo chiamato Hi-C User Pipeline V0.3.0 o più semplicemente HiCUP (www.bioinformatics.babraham.ac.uk/projects/hicup/) creato al Babraham Institute. Questa pipeline riceve i dati FASTQ di sequenziamento e mediante il programma Botwie allinea le sequenze al genoma di riferimento (GRCh37/hg19). HICUP processa i dati grezzi per produrre un valido set di coppie di *reads* (*di-tag)* da utilizzare per l'analisi dei risultati. HICUP produce tre grafici che riassumono le caratteristiche della libreria: il primo raffigura la distribuzione delle coppie di *reads* valide e non valide; il secondo rappresenta la suddivisione dei *di-tag* tra i diversi tipi di interazione, *in trans* o *in cis*; l'ultimo riguarda la rimozione dei duplicati, prodotti di ligazione con esattamente le stesse dimensioni e le stesse *reads,* che vengono generati dalla reazione di PCR e alterano i dati reali.

4.5.2 Il software SeqMonk

I dati di sequenziamento vengono visualizzati e analizzati su SeqMonk (www.bioinformatics.bbsrc.ac.uk/projects/seqmonk), un programma sviluppato al Babraham Institute da Simon Andrews.

SeqMonk si interfaccia all'operatore mediante un pannello (Figura 10) in cui vengono visualizzate diverse *tracks*; la prima (A) rappresenta la parte del genoma che si intende analizzare, per esempio singoli geni o

porzioni intergeniche; quella successiva (B) presenta la regione di DNA coperta dalle bait ad RNA, quindi le regioni oggetto di studio che intendiamo arricchire, che corrispondono al colore grigio.

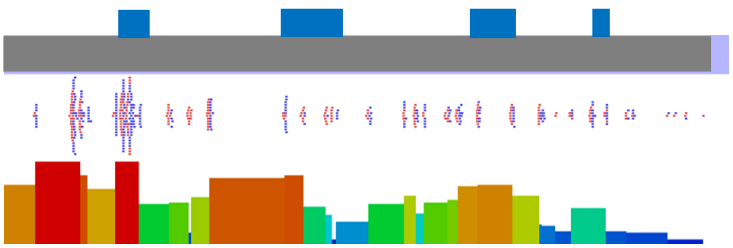

Figura 10: Rappresentazione dell'interfaccia di SeqMonk con l'operatore.

Più in basso si possono osservare le *reads* (C) ottenute dal sequenziamento dei prodotti di ligazione ed il cui colore varia a seconda dell'orientamento, blu per il *forward* e rosso per il *reverse*. Infine l'ultima parte è quella che rappresenta i risultati sottoforma di colonne (D), ciascuna delle quali corrisponde ad un frammento di restrizione, in cui altezza e colore dipendono dall'intensità del segnale: dal più debole e basso (blu) al più forte e alto (rosso), con dei valori ed altezze intermedie di colore giallo-verde- celeste (Figura 11). In questo modo a livello visivo si ha un immediato riscontro delle informazioni che vogliamo rilevare.

Figura 11: Rappresentazione della scala di colori utilizzata dal programma SeqMonk per indicare l'intensità dei segnali in funzione del colore; ai due estremi il rosso, indice di un segnale molto forte, e il blu che rappresenta invece un segnale molto debole.

L'analisi più adatta per i nostri dati di Hi-C e di SCRiBL prevede una quantificazione delle interazioni cromatiniche sulle regioni di interesse e si suddivide in due punti:

1) Definizione della lista di probes, cioè le regioni del genoma che vogliamo considerare come unità di misura.
2) Calcolo di un valore per ciascuna *probe*, dato dal anumero delle reads.

Nell'esperimento Hi-C la risoluzione non è altissima per cui per poter avere dei dati significativi le *probes* che vengono costruite corrispondono a più frammenti di restrizione uniti insieme, per esempio 50 frammenti; mentre con lo SCRiBL il nostro obiettivo è quello di poter di aumentare la risoluzione fino a poter riscontrare significativi *loop* di DNA partendo da singoli frammenti.

Per la quantificazione dei dati vera e propria è stata utilizzata l'opzione "*read count quantitation*" che applica un conteggio delle *reads* che si sovrappongono a ciacuna *probe*. Il sistema utilizzato per guidare

l'analisi dei dati è chiamato "Hi-C *other ends*". Una volta definite le *probes* si sceglie una regione di riferimento (*anchor region*) che può variare di dimensioni, da un singolo frammento di restrizione, che rappresenta la più piccola unità di misura, a sequenze più ampie che possono contenere anche un intero gene o *locus*. Il comando richiesto al programma è quello di identificare e calcolare le "*other ends*", cioè le *reads* che mappano insieme all'*anchor region* considerata, in quanto i prodotti di ligazione sono generati appunto dal legame tra due frammenti del genoma. A seguito della quantificazione si identificano le interazioni con la regione di interesse e più il contatto è frequente più il segnale è forte. L'intensità come già spiegato sopra viene indicata in base all'altezza e al colore della colonna, che riflette il numero di *reads.*

5. RISULTATI

Il nostro obiettivo è stato quello di indagare il potenziale ruolo di regolazione delle regioni contenenti gli SNPs associati alla modulazione dei livelli di HbF ed al miglioramento del fenotipo β-talassemico. Oltre al *locus* β-globinico, che è stato utilizzato come controllo di validazione sperimentale, sono stati investigati i *loci HBS1L-MYB* e *BCL11A*. Per valutare l'ipotesi di partenza, che tali regioni includano elementi regolatori capaci di agire in *cis* creando interazioni a lunga distanza per influenzare l'espressione spazio-temporale dei geni *target*, abbiamo scelto due approcci sperimentali per la caratterizzazione della struttura cromatinica nei *loci* di interesse.

Inizialmente è stato eseguito l'Hi-C (*Genome-wide Chromosome Conformation Capture*), una metodica all'avanguardia che rispetto alle altre tecnologie "3C" è in grado di offrire una visione *genome-wide* delle interazioni cromosomiche, sebbene non con un elevato potere di risoluzione.

Successivamente, considerando che il nostro interesse era rivolto all'analisi di specifici *loci*, abbiamo deciso di applicare e validare una nuova metodica chiamata SCRiBL (*Sequence Capture of regions interacting with Bait Loci*), al momento in via di pubblicazione, basata sull'arricchimento selettivo delle regioni bersaglio, per migliorare il potere risolutivo dell'Hi-C. Lo SCRiBL infatti è uno strumento potente in grado di utilizzare singoli frammenti di restrizione come unità di misura per riscontrare interazioni cromatiniche significative (vedere materiali e metodi).

5.1 Hi-C

5.1.1 Ottimizzazione dell'esperimento

Per poter avviare l'esperimento è stato necessario come primo passaggio selezionare la linea cellulare. Le colture di progenitori eritroidi isolati da sangue periferico sono ritenute il modello più adatto per tali studi, in quanto mimano le varie fasi dell'eritropoiesi. Tuttavia, tali cellule possiedono una limitata capacità proliferativa, si differenziano in maniera asincrona, per cui è difficile asserire con precisione in quale stadio di maturazione si trovano senza un attento monitoraggio della cinetica, ed infine sono difficili da reperire. Pertanto, dovendo mettere a punto una nuova metodica, la nostra scelta è ricaduta su una linea cellulare eritroide caratterizzata da una maggiore versatilità.

Tra le linee cellulari eritroidi in grado di esprimere i geni γ globinici fetali, le cellule K562 (*Human erythromyeloblastoid leukemia cell line*), Hel (Human erythroleukemia *cell line*) e KU812, la scelta è ricaduta sulle prime, una linea di leucemia mieloide cronica (CML) in crisi blastica terminale. Le K562 infatti, oltre ad esprimere in condizioni standard di coltura i geni γ globinici, esprimono i fattori di trascrizione *MYB* ad alti livelli e *BCL11A* a bassi livelli. Presentano inoltre altri vantaggi: non sono economicamente dispendiose,

possiedono un'alta capacità proliferativa, sono facilmente coltivabili *in vitro* e sono già state impiegate nello studio della struttura cromatinica ("3C" Dekker et al. 2002; "5C" Dostie et al. 2006). Complessivamente quindi le K562 possono essere considerate un ottimo modello sperimentale, capace di rappresentare in maniera affidabile un fenotipo eritroide fetale molto utile nello studio di tali globine.
Il secondo passo è stato quello di definire le condizioni sperimentali per l'Hi-C.
Abbiamo infatti deciso di apportare delle modifiche al protocollo seguito nel laboratorio Nuclear Dynamics dal gruppo di ricerca diretto dal Dottor Peter Fraser, con lo scopo di ottimizzare la metodica e aumentarne la risoluzione attraverso due passaggi fondamentali: la digestione enzimatica e la ligazione.

5.1.1.1 Doppia digestione enzimatica

I primi passaggi dell'Hi-C consistono nella fissazione delle cellule con la formaldeide, nella lisi cellulare e nella digestione enzimatica. Per l'analisi di *loci* superiori alle 10 Kb viene utilizzato un enzima di restrizione 6 *cutter* (solitamente HindIII o EcoRI). La dimensione dei frammenti di restrizione e di conseguenza la scelta dell'enzima, rappresenta il primo fattore che influenza la risoluzione con la quale vengono mappate le interazioni. Allo scopo di aumentare la frequenza di taglio su tutto il genoma e produrre un più alto numero di frammenti di dimensioni inferiori abbiamo testato una nuova condizione mai riportata in letteratura: la doppia digestione enzimatica. Sono stati scelti opportunamente due enzimi 6 *cutter*, HindIII e BglII in quanto presentano un'alta efficienza di digestione, assenza di sensibilità alla metilazione nei dinucleotidi CpG e tagliano i *loci* di interesse con un'alta frequenza. Il nostro obiettivo è stato quello di ridurre il più possibile l'unità di riferimento su cui individuare con maggiore precisione le sequenze coinvolte nei contatti cromatinici.

5.1.1.2 Ligazione "in nuclei"

La seconda modifica apportata alla metodica riguarda l'evento di ligazione.
Nei protocolli "3C" di Dekker (Dekker et al. 2002) e "Hi-C" di Liebermann (Liebermann et al. 2009) dopo la digestione ed il *fill-in* la cromatina viene portata in soluzione. L'aggiunta di SDS (sodio dodecil solfato) e l'incubazione ad alte temperature favoriscono la lisi della membrana nucleare portando al rilascio delle molecole di DNA nel citoplasma. Immediatamente si procede alla ligazione in condizioni di alta diluizione per favorire la giunzione intramolecolare di frammenti *cross-linkati* tra loro piuttosto che favorire il contatto casuale.
Nel laboratorio Nuclear Dynamics sono stati condotti esperimenti di ligazione "in nuclei", cioè all'interno del nucleo. Paradossalmente è stata riscontrata una maggiore efficienza e specificità evitando la permeabilizzazione della membrana nucleare con l'SDS e procedendo direttamente alla ligazione nel comparto nucleare. Abbiamo così deciso di testare tale condizione nell'esperimento.

5.1.1.3 Test sull'efficienza di digestione

Una prima aliquota di 30 milioni di cellule K562 è stata dedicata alla valutazione della nuova condizione della doppia digestione enzimatica. Dopo la ligazione, la cromatina è stata purificata e precipitata portando così alla costituzione della libreria "grezza" Hi-C utilizzata per analizzare le varie digestioni singolarmente. Come già descritto in maniera più approfondita nei materiali e metodi sono state valutate cinque condizioni sperimentali: 1) digestione con HindIII (H), 2) digestione con BglII (B), 3) doppia digestione con entrambi gli enzimi addizionati nella miscela di reazione contemporaneamente (H/B i), 4) doppia digestione con entrambi gli enzimi addizionati in sequenza a 9 ore di distanza l'uno dall'altro (H/B s), 5) controllo negativo, cellule non digerite (ND).

I prodotti di ligazione sono stati quantificati mediante q-PCR con tre coppie di *primers* che amplificano frammenti tra le 100 bp e le 177 bp in una regione a monte del gene *CDKN1A*. Come già accennato nei materiali e metodi la prima coppia amplifica una sequenza contenente un sito di taglio per l'enzima HindIII, la seconda per BglII e la terza coppia amplifica una regione di controllo che non contiene nessun sito di restrizione.

Il metodo di calcolo utilizzato è quello dei Ct (*Cycle Threshold*) ed è stata applicata la formula seguente: "100-100/2^((CtR-CtC)D-(CtR-CtC)UND)".

In Figura 12 sono rappresentate le percentuali di digestione relative all'enzima HindIII per ognuna delle cinque condizioni.

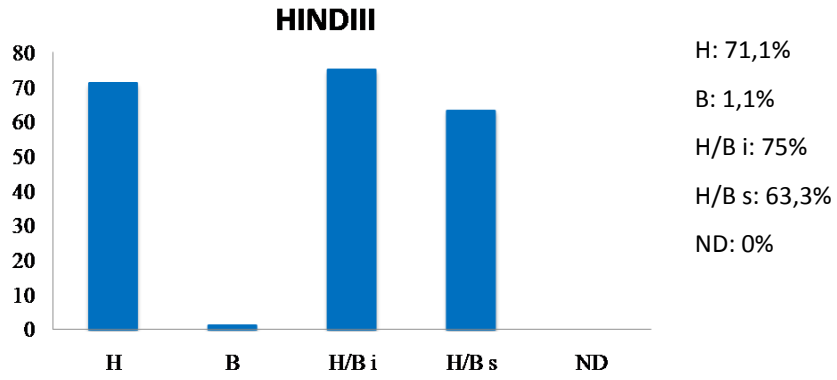

Figura 12: Grafico delle percentuali di digestione relative all'enzima HindIII in cinque diverse condizioni per saggiare la doppia digestione enzimatica. H: campione digerito con HindIII; B: campione digerito con BglII; H/B i: campione digerito con HindIII e BglII addizionati insieme; H/B s: campione digerito con HindIII e BglII addizionati in maniera sequenziale; ND: campione non digerito.

HindIII ha digerito la cromatina con le seguenti percentuali (Figura 12): 71,1% di efficienza nella digestione singola HindIII (H); 75% e 63,3% nella doppia digestione con entrambe gli enzimi insieme (H/B i) e nella

digestione sequenziale (H/B s) rispettivamente. Come atteso invece non si è ottenuta nessuna digestione nel campione digerito con BglII (B), così come in quello non digerito (ND).

Per quanto riguarda la digestione del secondo enzima, BglII, possiamo osservare i risultati in Figura 13.

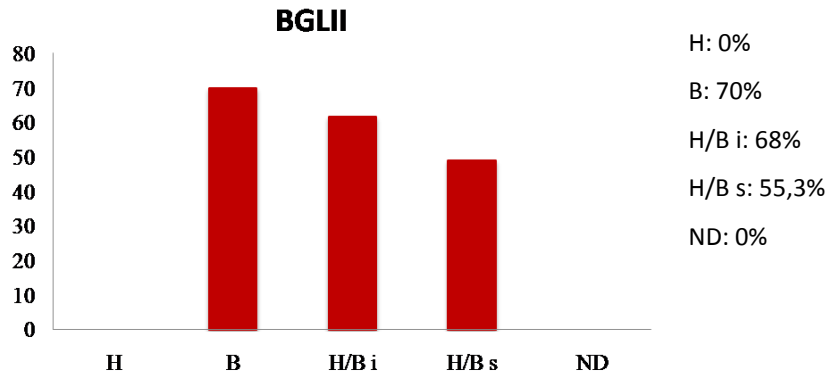

Figura 13: Grafico delle percentuali di digestione relative all'enzima BglII in cinque diverse condizioni per saggiare la doppia digestione enzimatica. H: campione digerito con HindIII; B: campione digerito con BglII; H/B i: campione digerito con HindIII e BglII addizionati insieme; H/B s: campione digerito con HindIII e BglII addizionati in maniera sequenziale; ND: campione non digerito.

L'efficienza di digestione per l'enzima BglII espressa in percentuale è la seguente: 0% per la digestione singola con HindIII (H) e nel campione non digerito (ND), 70% per la digestione singola con BglII (B), 68% e 55,3% per la doppia digestione con entrambi gli enzimi addizionati insieme (H/B i) ed in maniera sequenziale (H/B s), rispettivamente.

Anche BglII, come HindIII, ha digerito la cromatina efficientemente, per cui la doppia digestione è avvenuta con successo.

Per poter ottenere una libreria di alta qualità uno dei requisiti principali è l'alta percentuale di digestione enzimatica. Idealmente il 100% del DNA dovrebbe essere digerito, ma in realtà questo non avviene, per cui si ritiene che un'efficienza di digestione > del 60% sia un valore sufficiente per procedere con le fasi successive dell'esperimento.

In questo caso la nuova condizione di doppia digestione è stata altamente efficiente (circa il 70%) quindi ideale per la costruzione della libreria Hi-C.

5.1.2 Controlli di efficienza

Una volta definita la condizione ottimale per la doppia digestione enzimatica, questa è stata applicata su una seconda aliquota di 30 milioni di cellule K562 per eseguire l'intero esperimento Hi-C. Trattandosi di una metodica lunga, complessa e dispendiosa, è stato necessario effettuare dei controlli di efficienza e qualità

in più punti del protocollo. Le prime fasi dell'Hi-C sono le più critiche ed influenzano l'intero esperimento: il *cross-linking*, la lisi cellulare, la digestione enzimatica, il *fill-in*, la biotinilazione e la ligazione.

Una volta costituita la libreria Hi-C "grezza" sono stati eseguiti i seguenti test di qualità:

1) Rilevamento delle interazioni a breve distanza (*short range interactions*).

2) Rilevamento delle interazioni a lunga distanza (*long range interactions*).

3) Test di digestione enzimatica.

4) q-PCR.

1) Rilevamento delle interazioni "*short range*"

Il riscontro delle *short range interactions* costituisce la prima prova che i frammenti siano stati digeriti, resi *blunt* e ligati correttamente. Si esegue una PCR utilizzando diverse coppie di *primers* per il rilevamento di tutti i prodotti di ligazione possibili: frammenti HindIII/HindIII, frammenti BglII/BglII e frammenti HindIII/BglII. I *primers* sono stati disegnati tra i geni *HIST1H3D* e *HIST1H4E*, che appartengono ad uno dei *clusters* istonici del cromosoma 6, nello stesso orientamento in modo da essere entrambi *forward* o *reverse* e da trovarsi in due frammenti di restrizione adiacenti. Tale condizione rappresenta la combinazione di ligazione che avviene con più frequenza, in quanto più i frammenti si trovano vicini nel genoma più è probabile che essi vengano ligati insieme. I prodotti attesi di amplificazione sono stati ottenuti in tutte le PCR eseguite (Figura 14).

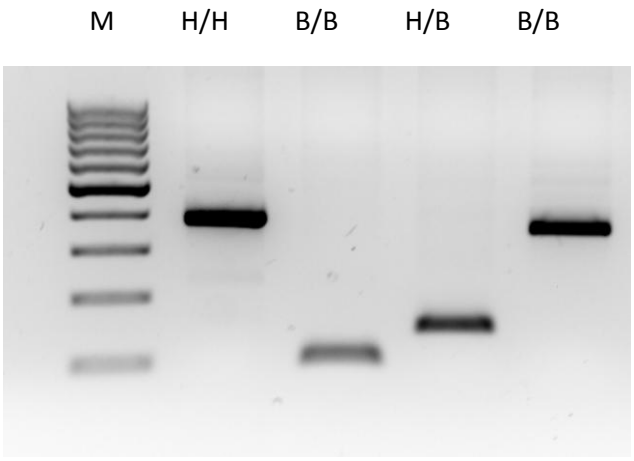

Figura 14: Frammenti di PCR delle *short range interactions* per i prodotti di ligazione HindIII/HindIII (H/H, ~380 bp), BglII/BglII (B/B, ~100 bp), HindIII/BglII (H/B, 145 bp) e BglII/BglII (B/B, 364 bp); M: marcatore di peso molecolare noto.

2) Rilevamento delle interazioni "*long range*"

Il secondo controllo di qualità testa la presenza delle *long range interactions*; esse non solo ci indicano che la digestione, il *fill-in* e la ligazione siano avvenuti con successo, ma ci permettono di valutare anche la qualità del *cross-linking*. Se infatti le cellule non sono state fissate correttamente con la formaldeide le interazioni tra regioni geniche che si trovano a grande distanza linearmente nel genoma, ma in prossimità spaziale, non verrebbero riscontrate.

Anche in questo caso sono state valutate le diverse combinazioni di ligazione (Figura 15 A, B). I *primers* di ciascuna coppia possono avere entrambi gli orientamenti, *reverse* e *forward,* e sono stati disegnati affinché si trovino nei due diversi *clusters* istonici localizzati nel cromosoma 6 ad una distanza di 1,5 Mb.

La corretta amplificazione di tutti i prodotti attesi attesta l'alta qualità della libreria.

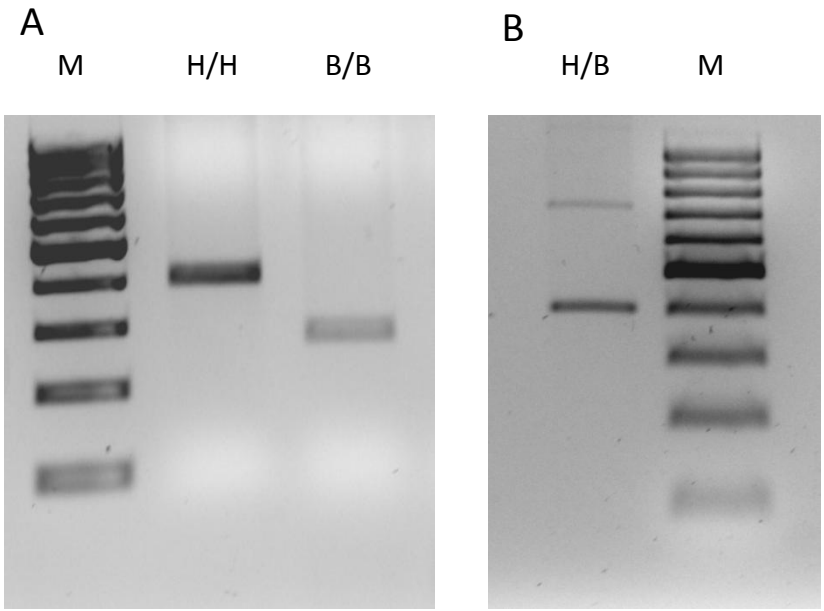

Figura 15: Frammenti di PCR delle *long range interactions*; A: Prodotti di ligazione HindIII/HindIII (H/H, 408 bp) e BglII/BglII (B/B, 288 bp); B: Prodotto di ligazione HindIII/BglII (H/B, 410 bp); M: Marcatore di peso molecolare noto.

3) Test di digestione enzimatica

Gli ampliconi derivati dalle PCR di rilevamento delle *short range interactions* per i prodotti di ligazione HindIII/HindIII e BglII/BglII sono stati digeriti enzimaticamente rispettando le seguenti condizioni: a) HindIII; b) NheI; c) HindIII/NheI; d) nessuna digestione, per i frammenti HindIII/HindIII; mentre e) BglII; f) ClaI; g) BglII/ClaI; h) nessuna digestione, per i frammenti BglII/BglII. Questo test diretto di digestione enzimatica oltre che indicare la bontà della digestione stessa, fornisce un'indicazione anche sull'efficienza del *fill-in* e quindi della ligazione. Le figure 16 e 17 riportano i risultati ottenuti.

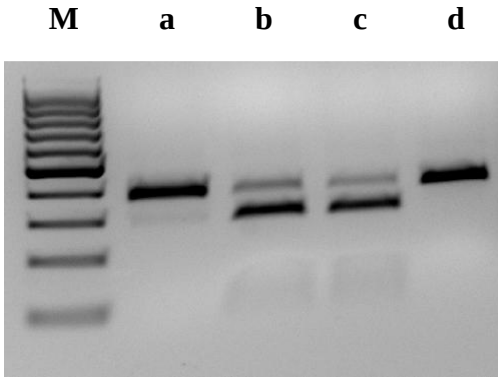

Figura 16: Digestione enzimatica di controllo sul prodotto della *short range interaction* HindIII/HindIII. (a): digestione con HindIII; (b): digestione con NheI; (c): digestione con HindIII e NheI; (d): nessun digestione. M: marcatore di peso molecolare noto.

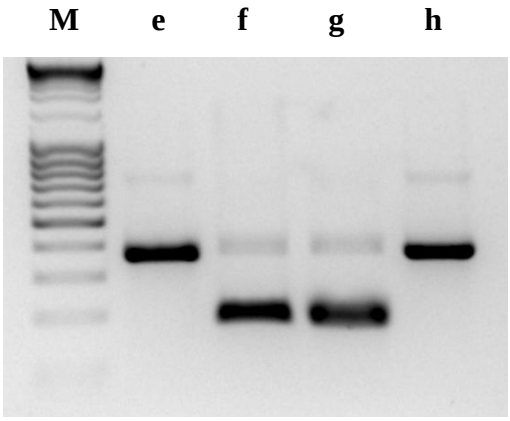

Figura 17: Digestione enzimatica di controllo sul prodotto della *short range interaction* BglII/BglII. (e): digestione con BglII; (f): digestione con ClaI; (g): digestione con BglII e ClaI; (h): nessuna digestione. M: marcatore di peso molecolare noto.

I *primers* utilizzati per le interazioni a breve distanza non riconoscono il DNA genomico, poiché sono stati disegnati per amplificare prodotti di ligazione tra due frammenti HindIII (~380 bp) e BglII (364 bp) rispettivamente, di cui uno è invertito rispetto all'altro. Se la ligazione dei frammenti resi *blunt* con il *fill-in* è avvenuta con successo si avrà la formazione di un nuovo sito di restrizione per gli enzimi NheI, nel caso dei prodotti di ligazione HindIII/HindIII e ClaI nel caso dei prodotti BglII/BglII. NheI e ClaI digeriscono infatti i relativi frammenti, e le bande risultanti avranno una dimensione inferiore (Figura 16 b; Figura 17 f). Sebbene una piccola frazione di DNA sembrerebbe non digerita da NheI o ClaI nella digestione singola e/o combinata (Figura 16 b, c; Figura 17 f, g) risulta però evidente che la maggior parte del DNA sia stato tagliato e quindi che l'efficienza di digestione, *fill-in* e ligazione rientri nei parametri richiesti per la prosecuzione dell'esperimento.

4) q-PCR

Un ulteriore test, mirato esclusivamente al controllo della qualità della digestione enzimatica, è stato condotto mediante q-PCR; sono stati così quantificati i prodotti di amplificazione delle tre stesse coppie di *primers* utilizzati per testare l'efficienza della doppia digestione enzimatica: la prima coppia orientata verso

un sito di restrizione HindIII, la seconda verso un sito BglII e la terza su una sequenza che non presenta siti di restrizione.

Anche in questo caso viene utilizzato per l'analisi il metodo dei Ct che ha permesso di calcolare l'efficienza di digestione dell'esperimento Hi-C (Figura 18).

Entrambi gli enzimi hanno digerito la cromatina con efficienza al di sopra del 70%, definendo così l'alta qualità della libreria.

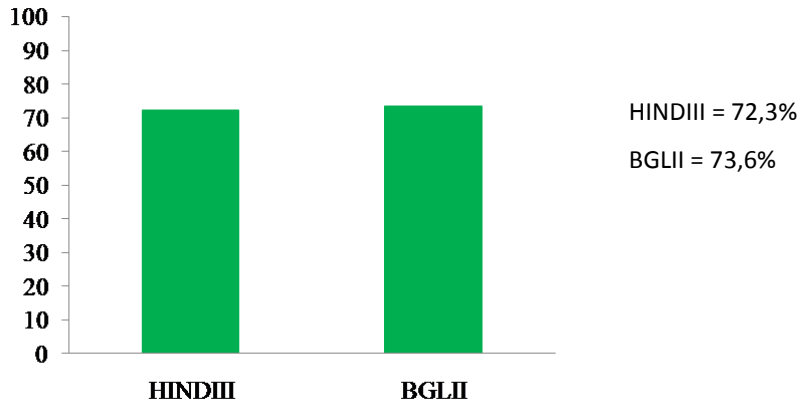

Figura 18: Efficienza della digestione enzimatica per gli enzimi HindIII e BglII nell'esperimento Hi-C completo.

Complessivamente tutti i controlli effettuati hanno avuto un riscontro positivo confermando che la doppia digestione, il *fill-in* e la ligazione sono avvenuti con alta efficienza e la libreria rispecchia le condizioni ottimali dettate dall'esperimento.

5.1.3 Amplificazione della libreria e controlli di qualità

La molecole di DNA della libreria sono state ulteriormente frammentate mediante sonicazione e selezionate in base alle dimensioni mediante l'utilizzo delle biglie magnetiche Agencourt AMPure XP. Il "*pull-down*" ha permesso la raccolta dei prodotti di ligazione sfruttando l'affinità della biotina per la streptavidina di cui sono rivestite le biglie magnetiche Dynabeads MyOne Streptavidin C1. In seguito al legame con gli adattatori Illumina, che rappresentano il substrato per l'attacco delle molecole alla *flow-cell*, la libreria è stata amplificata per produrre un quantitativo di molecole sufficiente per il sequenziamento. Prima di procedere sono stati effettuati dei test di amplificazione in modo da valutare il ciclo ideale, che fornisca DNA sufficiente per il sequenziamento e produca meno artefatti durante la reazione di PCR. Sono stati testati i cicli 7, 8, 9 e 12 (Figura 19).

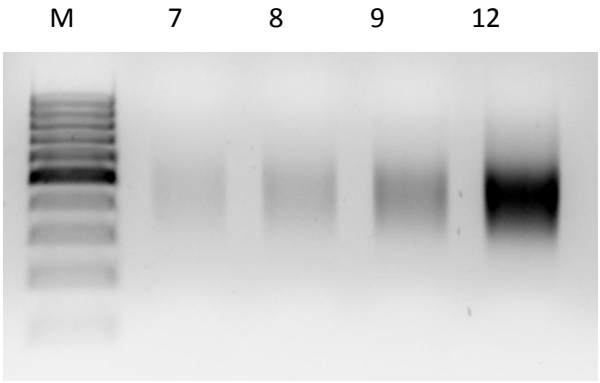

Figura 19: Test di amplificazione della libreria Hi-C; i cicli valutati sono stati il 7, l'8, il 9 ed il 12. M: marcatore di peso molecolare noto.

Come mostrato in Figura 19 un prodotto di PCR è visibile in tutti i pozzetti. E' stato appurato mediante esperimenti precedenti che l'amplificazione ideale corrisponde a 3 cicli inferiori rispetto ad un buon prodotto ben visibile su gel (ciclo 9). In questo caso la condizione migliore è rappresentata da 6 cicli. Il successo delle PCR di prova inoltre ci fornisce un controllo indiretto sulla libreria; i *primers* utilizzati Illumina TruSeq 1.0 e 2.0 sono specificamente compatibili con gli adattatori Illumina appena legati alle estremità delle molecole di DNA, per cui l'amplificazione riflette la corretta incorporazione di questi ultimi.

5.1.4 Controlli qualitativi e quantificazione della libreria Hi-C

La libreria Hi-C "finita" per essere ritenuta idonea al sequenziamento deve superare due controlli: un saggio elettroforetico sul Bioanalyzer 2100 e una q-PCR eseguita con il KAPA SYBR kit.

La prima prova fornisce anche una valutazione qualitativa, oltre al calcolo della concentrazione viene infatti analizzata la dimensione dei frammenti e riportata sotto forma di ferogramma (Figura 20):

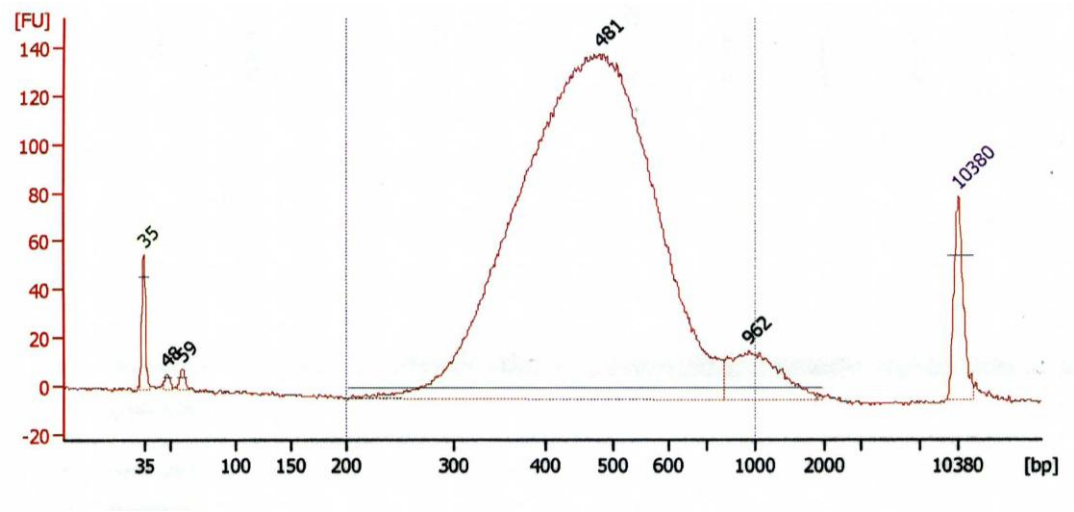

Figura 20: Elettroferogramma risultatante dal saggio elettroforetico sul Bioanalyzer 2100 della libreria Hi-C; nell'asse delle ordinate è rappresentata l'unità di fluorescenza e nell'asse delle ascisse la dimensione delle molecole di DNA espressa in bp.

L'asse delle ordinate riporta la concentrazione delle molecole misurate in base alla fluorescenza emessa durante il saggio, mentre le ascisse indicano la dimensione dei frammenti in bp. Per effettuare la misurazione è stato utilizzato un marcatore i cui estremi sono i picchi a 35 bp e 10380 bp rispettivamente.
La dimensione media dei frammenti della libreria Hi-C è di 481 bp, un valore perfettamente compatibile con i parametri richiesti per il sequenziamento. La presenza di un piccolo segnale a 962 bp è probabilmente dovuta ad una selezione dei frammenti ad alto peso molecolare mediante biglie magnetiche non del tutto efficiente.
La concentrazione della libreria calcolata dal Bioanalyzer equivale a 18,18 ng/µl per un totale di 60 µl (1090 ng complessivamente).
Il secondo test è stato eseguito utilizzando il kit Kapa SYBR fast qPCR, un sistema molto accurato ideato proprio per la quantificazione delle librerie di DNA. La concentrazione risultante è di 16,2 ng/µl. Solitamente i valori ottenuti dai 2 controlli discordano leggermente, per cui viene utilizzato come valore ultimo la media tra i due.
Per il sequenziamento è fondamentale conoscere con massima precisione la concentrazione della libreria Hi-C per poter caricare la giusta quantità di campione nella *flow cell*; un eccesso o una scarsità di DNA potrebbero causare perdita nella complessità della libreria e quindi nei dati finali.

5.1.5 Elaborazione dei dati della libreria Hi-C con l'algoritmo HiCUP

Il sequenziamento, eseguito con la piattaforma HiSeq 2000 Illumina, produce coppie di *reads* (*di-tag*) per ciascuna molecola di DNA che non si trovano sulla stessa sequenza, ma su frammenti provenienti da due diversi punti, anche in diversi cromosomi. La libreria Hi-C è infatti molto complessa poiché rappresenta la struttura 3D del genoma.
La *pipeline* HiCUP applica una serie di filtri che eliminano gli artefatti sperimentali per ottenere solamente le coppie di *reads* valide (per dettagli vedere materiali e metodi).
Il primo grafico (Figura 21) illustra la distribuzione dei *di-tag* informativi (87,3%) e non informativi (12,7%). I *di-tag* rappresentano le coppie di *reads* localizzate nei prodotti di ligazione e vengono definiti non validi quando riflettono gli artefatti sperimentali che si generano durante la costruzione della libreria. Sono state rilevate le seguenti percentuali di *di-tag* non validi: 6,7% in frammenti di dimensioni non compatibili con l'analisi, 0,4% separati da pochi altri frammenti, risultato di una digestione incompleta, 3% in frammenti adiacenti che si sono ligati nello stesso orientamento, 1,1% nello stesso frammento, 0,4% nello stesso frammento di restrizione, di cui una singola *read* mappa sul sito di taglio, infine 1,1% appartengono allo stesso frammento di restrizione circolarizzato.
Complessivamente questi valori danno un'ulteriore indicazione sull'alta qualità della libreria, in quanto il numero di *di-tag* validi è più alto rispetto ai valori delle librerie generate precedentemente nel laboratorio

Nuclear Dynamics. Tale dato ci induce a pensare che l'introduzione della doppia ligazione rappresenta effettivamente un sistema più efficiente.

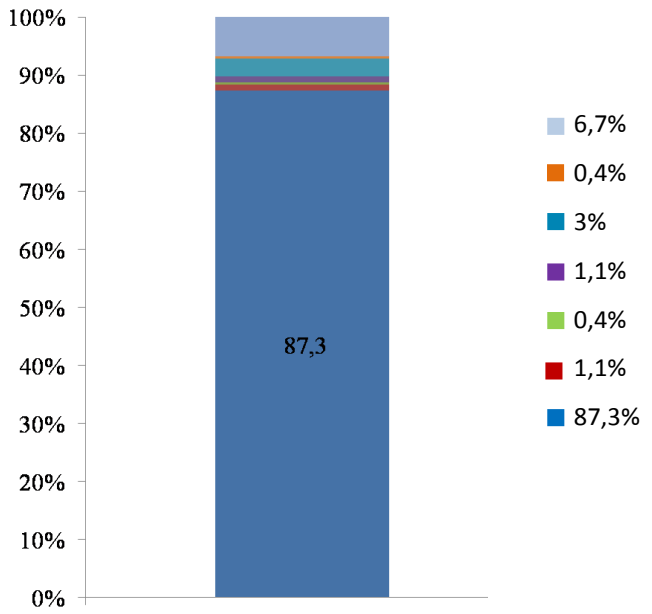

Figura 21: Distribuzione delle coppie di *reads* dell'esperimento Hi-C.
di-tag di dimensione errata;
di-tag separati da pochi frammenti;
di-tag in frammenti adiacenti che si sono ligati nello stesso orientamento;
di-tag sullo stesso frammento;
di-tag sullo stesso frammento di cui una *read* mappa sul sito di taglio;
di-tag sullo stesso frammento ricircolarizzato;
di-tag validi.

Il secondo grafico (Figura 22) riporta la distribuzione delle interazioni *in trans* (9,4%), *in cis* ad una distanza < a 10 Kb (16,6%) e > di 10 Kb (74,1%). Come è stato già discusso nei materiali e metodi una percentuale bassa di interazioni intercromosomiche indica un basso rumore di fondo.

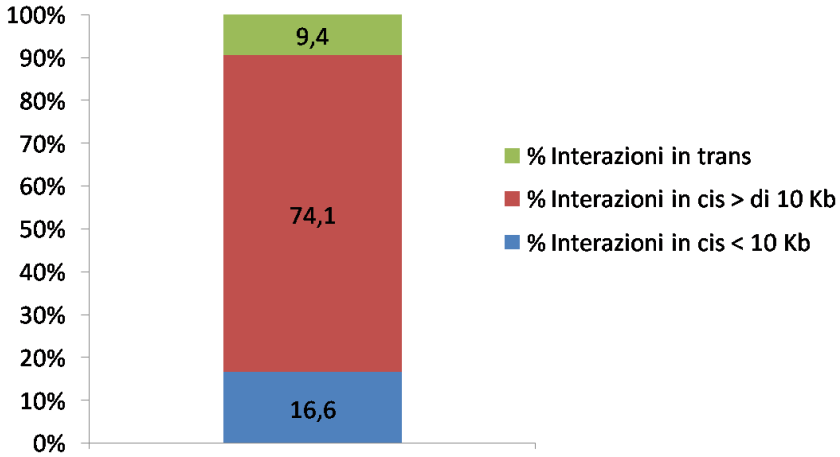

Figura 22: Distribuzione delle coppie di *reads* in base al tipo di interazione sotto forma di percentuale; in verde abbiamo le interazioni in *trans*; in rosso le interazioni in *cis* ad una distanza maggiore di 10 Kb; in blu le interazioni in *cis* ad una distanza inferiore a 10 Kb.

L'ultimo filtro applicato da HiCUP elimina i duplicati generati dalla PCR. Come possiamo notare in Figura 23 i *di-tag* validi sono 44.810.216 e tra questi 44.592.450 sono privi di duplicati, equivalenti al 99,5% del totale.

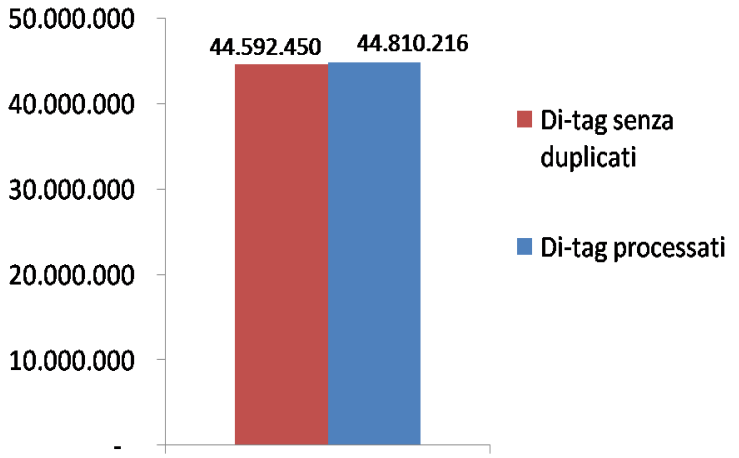

Figura 23: Raffigurazione delle coppie di *reads* processate valide in blu e di quelle che non comprendono i duplicati in rosso.

5.2 Preparazione delle bait ad RNA

La SCRiBL arricchisce specifiche regioni di DNA di interesse attraverso l'ibridazione tra la libreria Hi-C e piccole sequenze ad RNA biotinilate. Queste ultime sono state costruite utilizzando come substrato il DNA dei *loci* di interesse veicolato da cloni BAC (*Bacterial Artificial Chromosome*): il *cluster* β-globinico, la regione intergenica *HBS1L-MYB* e il *BCL11A*. Altri loci sono stati utilizzati per testare la nuova metodica (Tabella 6) e non saranno oggetto di discussione in questa tesi.

Il protocollo eseguito è stato ideato al Babraham Institute. Il primo passaggio ha previsto la digestione enzimatica di quantità equimolari di DNA proveniente da tutti i cloni BAC (Tabella 6) con i due enzimi utilizzati nell'Hi-C, HindIII e BglII, in due tubi separati. L'avvenuta digestione è stata valutata mediante corsa elettroforetica su gel di agarosio di una piccola aliquota di ciascuna reazione, confrontandole con una frazione di DNA non digerito (Figura 24).

La digestione enzimatica ha originato una serie di frammenti di dimensioni differenti creando uno *smear* (Figura 24 H; B); mentre il DNA non digerito è rappresentato da una sola banda ad alto peso molecolare (Figura 24 Co). La miscela di DNA digerita con ciascuno dei due enzimi è stata tenuta in due tubi separati fino alla fine del protocollo per generare due set di bait differenti.

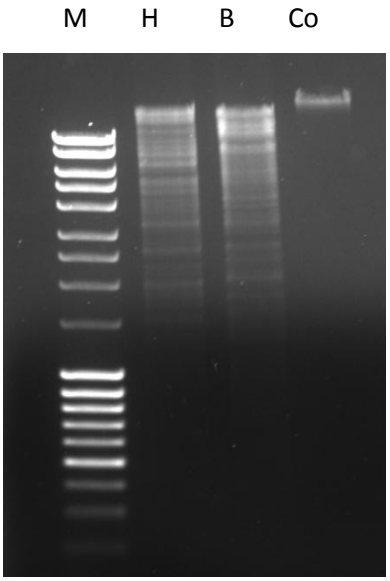

Figura 24: Digestione enzimatica dei cloni BAC;
H: Miscela dei cloni BAC digeriti da HindIII;
B: Miscela dei cloni BAC digeriti da BglII;
Co: Miscela dei cloni BAC non digeriti (controllo).
M: Marcatore di peso molecolare noto.

Successivamente i frammenti di DNA sono stati ligati a specifici adattatori che presentano le sequenze universali del promotore del fago T7 e poi ulteriormente frammentati mediante sonicazione. Sono state recuperate due aliquote prima e dopo la sonicazione e visualizzate su gel (Figura 25).

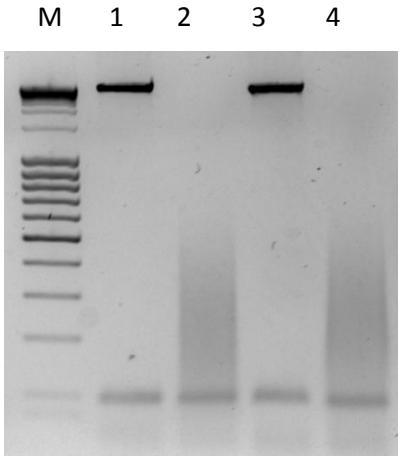

Figura 25: Controllo delle aliquote prima e dopo sonicazione rispettivamente per il campione digerito con HindIII e con BglII. 1: campione digerito con HindIII pre-sonicazione; 2: campione digerito con HindIII post-sonicazione; 3: campione digerito con BglII pre-sonicazione; 4: campione digerito con BglII post sonicazione; M: marcatore di peso molecolare noto.

Come possiamo osservare in Figura 25 prima della frammentazione le molecole di DNA hanno alto peso molecolare con una banda meno intensa all'altezza di 100 bp che rappresenta gli adattatori in eccesso residui dalla reazione di ligazione (pozzetti 1 e 3). Dopo l'uso del Covaris i frammenti sono stati tagliati in maniera casuale per ottenere un picco di 200 bp, ed il risultato è confermato da uno *smear* tra le 500 bp e 100 bp per entrambi i campioni (pozzetti 2 e 4).

La purificazione con le Agencourt AMPure XP beads ha permesso di selezionare ulteriormente i frammenti in base alle dimensioni eliminando i pesi molecolari più alti e più bassi; le molecole di DNA eluite infatti presentano una media tra le 200 bp e le 400 bp (Figura 26).

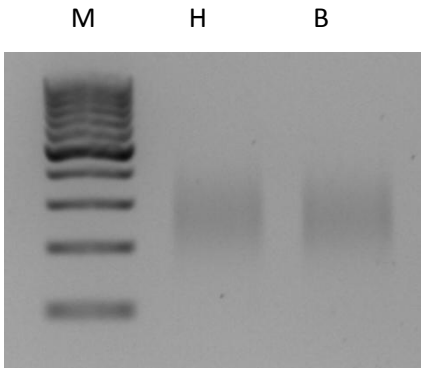

Figura 26: Controllo dei campioni dopo purificazione e selezione con le XP beads.
H: miscela dei cloni BAC digerita con HindIII;
B: miscela dei cloni BAC digerita con BglII;
M: marcatore di peso molecolare noto.

La *trascrizione in vitro* ha convertito il DNA in molecole ad RNA a singolo filamento (bait) che sono state biotilinate grazie all'inserimento nella reazione di UTP biotinilato. La presenza della biotina è essenziale per la reazione di ibridazione e la successiva selezione dei frammenti di DNA arricchiti nelle regioni di interesse. Per poter testare l'incorporazione della biotina è stata ibridata un'aliquota delle bait digerite con HindIII inizialmente (bait Hind) e un'aliquota delle bait digerite con BglII (bait Bgl) con un fluoroforo coniugato con la streptavidina. L'affinità tra quest'ultimo e la biotina determinerà il legame del fluoroforo all'RNA (Figura 27 A-B-C).

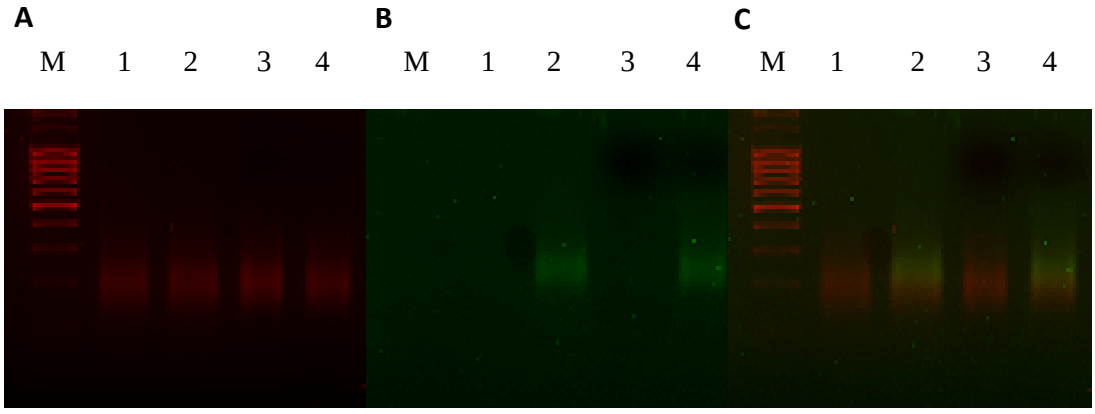

Figura 27: Test eseguito per il controllo dell'incorporazione della biotina mediante rilevamento dell'emissione di fluorescenza da parte del fluoroforo Alexa Fluor-647. A: visualizzazione dell'RNA in rosso e del marker mediante rilevamento del segnale proveniente dall'etidio bromuro; B: visualizzazione della biotina in verde mediante rilevamento del segnale del fluoroforo; C: visualizzazione combinata dell'RNA in rosso e della biotina in verde. 1: bait HindIII non incubate col fluoroforo. 2: bait HindIII incubate col fluoroforo. 3: bait BglII non incubate col fluoroforo. 4: bait BglII incubata col fluoroforo.

Nella Figura 27 A il colore rosso presente in tutti i pozzetti corrisponde all'RNA, e DNA nel caso del marcatore. La Figura B invece mostra la lettura della sola biotina nei quattro campioni. Infine la Figura C combina le prime due, permettendo di visualizzare sia l'RNA che la biotina contemporaneamente. Infatti

come atteso i pozzetti 2 e 4, che contengono i campioni incubati col fluoroforo, confermano la presenza della biotina.

Possiamo affermare che la preparazione delle bait è avvenuta con successo: le dimensioni dei frammenti ad RNA rispettano i parametri richiesti per l'ibridazione ed inoltre la biotina è stata incorporata correttamente.

5.3 SCRiBL (Sequence Capture of Regions Interacting with Bait Loci)

Lo SCRiBL prevede l'arricchimento selettivo delle regioni di interesse mediante ibridazione tra la libreria Hi-C e le bait ad RNA.

Successivamente le regioni target così arricchite vengono "recuperate" sfruttando l'affinità tra la biotina incorporata nelle bait e la streptavidina coniugata a delle biglie magnetiche.

5.3.1 Amplificazione della libreria SCRiBL, quantificazione e controllo di qualità

La libreria SCRiBL risultante, composta da ibridi DNA:RNA, anche in questo caso deve essere amplificata per produrre materiale sufficiente per il sequenziamento cercando allo stesso tempo però di non generare artefatti nella reazione di PCR.

Sono stati così eseguiti dei test di amplificazione a 9, 12 e 15 cicli per valutare la condizione ottimale (Figura 28), ed è stato scelto il ciclo 7.

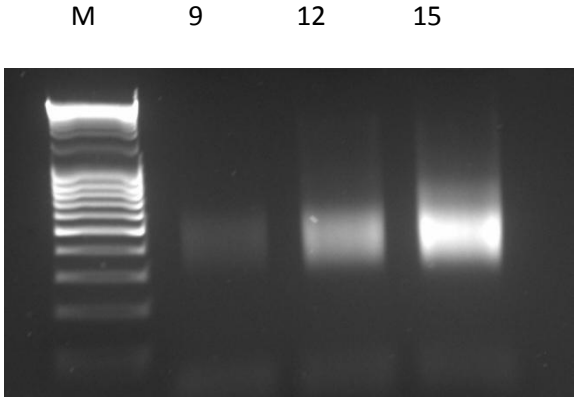

Figura 28: Test di amplificazione della libreria SCRiBL. I cicli raffigurati sono il 9, il 12 e il 15. M: Marcatore di peso molecolare noto.

Dopo la purificazione dei prodotti finali di PCR sono stati eseguiti gli ultimi test di qualità.

Come per la libreria Hi-C è stato eseguito il saggio elettroforetico sul Bioanalyzer 2100 (Figura 29) e la quantificazione mediante q-PCR col KAPA SYBR kit.

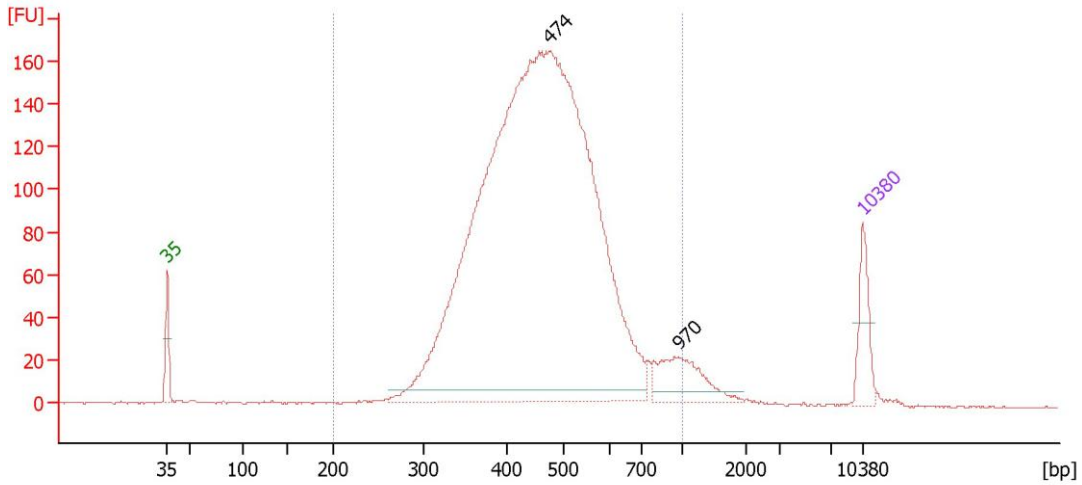

Figura 29: Elettroferogramma risultante dalla corsa elettroforetica sul Bioanalyzer 2100 per la libreria SCRiBL. Nell'asse delle ordinate è presente l'unità di fluorescenza che riflette la concentrazione del campione e nelle ascisse la dimensione delle molecole di DNA espressa in bp.

L'elettroferogramma (Figura 29) mostra una dimensione media di frammenti di 474 bp, che rispetta i parametri richiesti per il sequenziamento. Come nella libreria Hi-C è presente una piccola irregolarità all'altezza di 970 bp, probabilmente dovuta alla di selezione con le AMPure XP beads; la reazione potrebbe non essere stata efficiente al 100% lasciando un piccolissimo residuo di frammenti ad alto peso molecolare. Tuttavia essendo un quantitativo minimo la libreria è risultata comunque idonea al sequenziamento.

La concentrazione ottenuta equivale a 4.23 ng/µl, mentre il kit KAPA indica 3.63 ng/µl. Anche in questo caso si è tenuto conto della media matematica tra i due campioni come concentrazione finale della libreria.

5.3.2 Elaborazione dei dati della libreria SCRiBL con l'algoritmo HiCUP

I dati di sequenziamento della libreria SCRiBL, avvenuto con la piattaforma MiSeq Illumina, sono stati processati da HiCUP con le stesse modalità utilizzate per la libreria Hi-C. Inoltre sono stati prodotti altri indicatori specifici per il protocollo SCRiBL per valutarne qualità ed efficienza.

Le coppie di *reads* valide equivalgono al 92,3%, delle totali (Figura 30), mentre quelle non informative sono complessivamente il 7,7% e si suddividono in: 3% in frammenti di dimensione incorretta, 0,4% separati da pochi frammenti, 2,6% in frammenti adiacenti ligati nello stesso orientamento, 0,3% e 0,5% che mappano sullo stesso frammento e nel primo caso una *read* si sovrappone al sito di restrizione, infine 0,9% che si trovano sullo stesso frammento che si è circolarizzato.

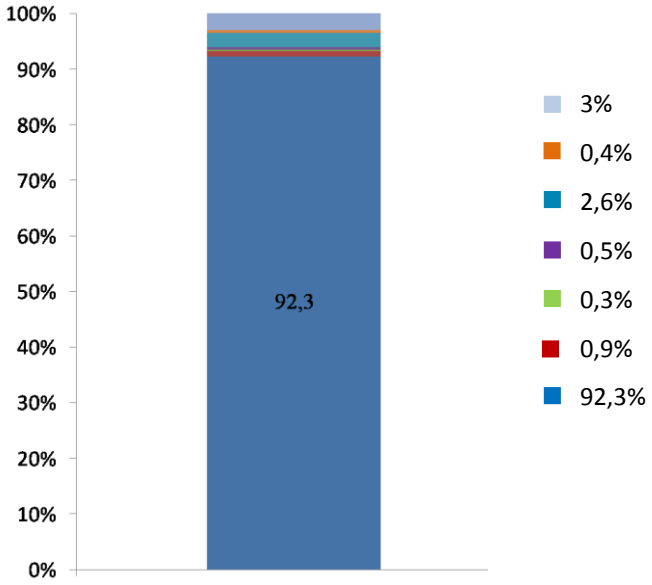

Figura 30: Distribuzione delle coppie di *reads* dell'esperimento Hi-C.
- *di-tag* di dimensione errata;
- *di-tag* separati da pochi frammenti;
- *di-tag* in frammenti adiacenti che si sono ligati nello stesso orientamento;
- *di-tag* sullo stesso frammento;
- *di-tag* sullo stesso frammento di cui una *read* mappa sul sito di taglio;
- *di-tag* sullo stesso frammento ricircolarizzato;
- *di-tag* validi.

Il secondo grafico (Figura 31) rappresenta le distribuzione delle interazioni riscontrate: 9% per le *trans*, 16,18% per le *cis* < a 10 Kb ed il 75% per le *cis* > a 10 Kb.

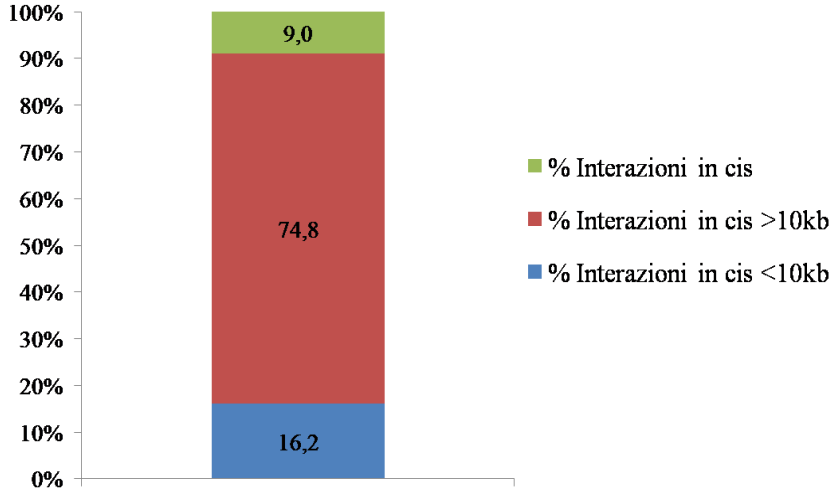

Figura 31: Distribuzione delle coppie di *reads* in base al tipo di interazione sotto forma di percentuale; in verde abbiamo le interazioni in *trans*; in rosso le interazioni in *cis* ad una distanza maggiore di 10 Kb; in blu le interazioni in *cis* ad una distanza inferiore a 10 Kb.

Infine l'ultimo passaggio rimuove i duplicati generati dalla PCR: le *reads* senza duplicati risultano 15.506.340, il 99,6% di quelle valide (Figura 32).

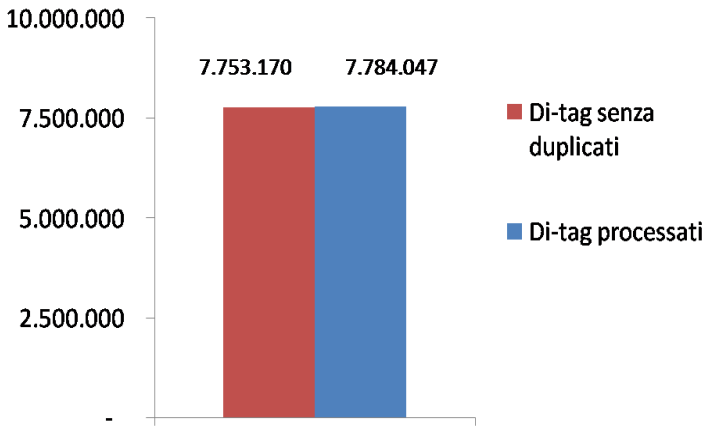

Figura 32: Raffigurazione delle coppie di *reads* valide processate in blu e senza duplicati in rosso.

Anche la libreria SCRiBL possiede un'altissima qualità e tutti i parametri rispettano i valori richiesti.

Lo SCRiBL rappresenta una nuova metodica per cui per valutare il grado di efficienza dell'esperimento HiCUP ha effettuato ulteriori quantificazioni.

Per prima cosa sono state calcolate le *reads* che allineano nelle regioni target (reads "catturate"), che riflettono la qualità e specificità dell'ibridazione (Tabella 10). Come si può osservare nella Tabella seguente si tratta di 1.044.690 su 15.506.340 *reads* valide totali, e la percentuale è di 6,73%.

***READS* VALIDE SENZA DUPLICATI**	15.506.340
***READS* "CATTURATE"**	1.044.690
***READS* "NON CATTURATE"**	14.461.650
PERCENTUALE DI *READS* CATTURATE (Efficienza SCRiBL)	6,73%

Tabella 10: Quantificazione di HiCUP delle *reads* totali valide per l'esperimento SCRiBL, di quelle "catturate" cioè provenienti dalle regioni di interesse coperte dalle bait, di quelle "non catturate" quindi fuori dalle regioni target e infine la *reads* catturate espresse in percentuale rispetto alle *reads* totali.

La specificità dell'ibridazione può essere stata influenzata dalle proprietà delle molecole di DNA della libreria, le cui estremità terminano con delle sequenze che sono parzialmente complementari tra loro ("adesive"), determinando il legame tra frammenti *target* e non *target* ("*off-target*"), creando delle lunghe molecole aspecifiche. Gli *off-target* possono permanere durante l'intero esperimento, fino ad essere sequenziati insieme ai *target*, generando così l'alto livello di *background*. Sono state utilizzate preventivamente delle sequenze bloccanti (*blockers*) ad alta concentrazione molare nella preparazione del DNA per la reazione di ibridazione. Ma nonostante l'addizione dei *blockers* persiste un alto livello di aspecificità, per cui probabilmente queste sequenze non bloccano efficientemente il legame con i

frammenti *off-target,* o ancora vengono inserite in un punto della procedura sperimentale in cui sono già costituite le catene aspecifiche.

Nonostante le *reads* "catturate" siano una piccolissima percentuale rispetto a quelle processate occorre tener presente che esse non sono distribuite lungo tutto il genoma, ma solo sulle porzioni geniche selezionate con le bait; per cui distribuendo 1.044.690 *reads* "catturate" su tali regioni, risultano comunque essere un numero molto elevato che garantisce l'identificazione di interazioni significative.

Una seconda quantificazione che HiCUP fornisce all'operatore è la percentuale di arricchimento dello SCRiBL per ciascuna regione utilizzata, rispetto all'Hi-C (Tabella 11). L'arricchimento è dato dal rapporto tra le *reads* "catturate" dallo SCRiBL in ogni regione selezionata e le *reads* provenienti dalle stesse regioni nell'esperimento Hi-C.

ID BAC	CROMOSOMA	% READS PER SINGOLA BAIT NELLO SCRiBL	% READS PER SINGOLA BAIT NELL'Hi-C	ARRICCHIMENTO (%)
RP11-104D9 (*Hbs1l-Myb*)	**6**	**0,75**	**0,00797**	**94,2**
RP11-65A9 (*Bcl11a*)	**2**	**0,1**	**0,00649**	**14,9**
CTD-2063A20 (*Locus* β)	**11**	**0,39**	**0,00492**	**80,2**
RP11-135I7 Regione intergenica	2	0,59	0,00702	83,5
CTC-215I11 (*Locus* α)	16	1,02	0,01122	90,6
RP11-1013N13 (*Cluster* istonici)	6	0,48	0,00529	90,8
CTD-2536K9 (*Cluster* HOX-A)	7	0,8	0,01149	69,4
CTD-3054H22 (*Cluster* HOX-A)	7	0,24	0,00289	83,7
CTD-2545C21 Regione ENCODE	2	0,09	0,00231	37,0

Tabella 11: Lista dei cloni BAC utilizzati per la costruzione delle bait. Evidenziati in grossetto vi sono i tre BAC che coprono i *loci* di interesse, mentre gli altri sono stati utilizzati solo per testare la nuova metodica. Calcolo generato da HiCUP sulla percentuale di arricchimento data dal rapporto tra le reads dallo SCRiBL per ciascuna delle regioni utilizzate e quelle invece che provengono dall'Hi-C.

Questi dati mettono a confronto i due approcci evidenziando la potenza dello SCRiBL, che nonostante l'alto *background* è capace di arricchire in maniera consistente le regioni di interesse. Tuttavia abbiamo

osservato che la regione al *locus BCL11A* risulta avere una percentuale di arricchimento nettamente inferiore sia a quella attesa che rispetto a quella ottenuta agli altri *loci*. Questo risultato ci ha indotto ad eseguire ulteriori controlli per chiarire le cause di tale anomalia (vedi di seguito paragrafo "Analisi del *locus BCL11A*").

5.4 Analisi delle interazioni con il software SeqMonk

SeqMonk è il programma utilizzato per l'analisi dei dati di sequenziamento derivanti dagli esperimenti di Hi-C e SCRiBL, permettendone la visualizzazione sul genoma di riferimento e fornendo gli strumenti per filtrare e quantificare i dati. L'analisi viene eseguita mediante due passaggi: la scelta della regione di riferimento o "*anchor region*" e l'identificazione e calcolo delle *reads* provenienti dai prodotti di ligazione ad essa corrispondenti. L'interpretazione dei dati a livello visivo è in funzione dell'altezza e del colore dei singoli frammenti di restrizione che vengono visualizzati come colonne (vedere materiali e metodi); più queste sono alte più il segnale è forte.

5.4.1 Analisi del *locus* β-globinico

Il *locus* β-globinico rappresenta una delle regioni maggiormente studiate a livello genetico e molecolare, caratterizzata anche dal punto di vista della struttura cromatinica. L'avvento delle tecnologie "3C" ha per la prima volta confermato *in vivo* la formazione di *loop* nel DNA, ed in particolare il *locus* β-globinico murino è stato tra i primi oggetto di tale studio (Tolhuis et al. 2002). E' stato dimostrato che la *LCR* interagisce fisicamente con i singoli geni del *cluster* in funzione dello stadio di sviluppo (Tolhuis et al. 2002).

Per cui ho considerato il *locus* β-globinico umano come modello ideale per poter validare la metodica dello SCRiBL, mai utilizzata finora, confrontando i nostri risultati con quelli riportati in letteratura.

In particolare, abbiamo considerato il manoscritto di Dostie e colleghi pubblicato su "Genome Research" nel 2006, che riporta un'analisi delle interazioni lungo il *cluster* β-globinico in cellule K562, mediante *Choromosome Conformation Capture Carbon Copy* (5C; Dostie et al. 2006).

Utilizzando come riferimento (*anchor region*) l'insieme dei siti HS2-HS3-HS4 della *LCR* gli autori evidenziano un'elevata frequenza di interazioni tra tale regione e i geni γ globinici (Dostie et al. 2006).

Analogamente l'esperimento di SCRiBL, condotto sempre in K562 (mantenendo le medesime condizioni di crescita riportate da Dostie, in modo da contenere quanto più possibile le variabili relative alla coltura cellulare), è stato analizzato mediante il programma SeqMonk considerando lo stesso riferimento, ovvero l'insieme dei siti HS2-HS3-HS4 (chr11:5301442-5310696) (Figura 33).

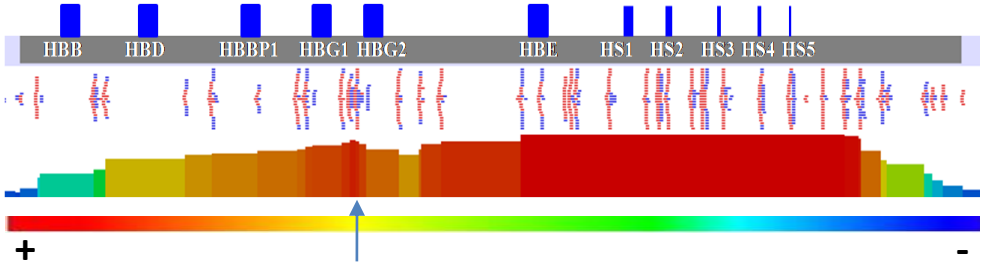

Figura 33: Analisi SeqMonk mendiante "Hi-C *other ends*" e opzione "*Smoothing*" con i siti HS2, HS3 e HS4 della *LCR* considerati come regione *anchor*.

I nostri dati confermano la stessa forte interazione evidenziata nell'esperimento 5C di Dostie tra la *LCR* ed i geni fetali γ (Figura 33; blocco rosso indicato con la freccia blu), permettendo la validazione della nostra metodica sperimentale.

Le interazioni tra regioni adiacenti nel genoma come per esempio la *LCR* e il gene *HBE*, non possono essere valutate in quanto la loro stessa vicinanza potrebbe causare un'alta percentuale di collisioni casuali non funzionali tra i frammenti cromatinici. Per cui i segnali ad alta intensità all'interno del grande blocco rosso in Figura 33, che comprendono numerosi frammenti di restrizione, potrebbero essere il risultato di un artefatto dovuto appunto alla prossimità dei frammenti nel genoma, piuttosto che rappresentare delle reali interazioni.

Il contatto tra la *LCR* e i geni *HBG1/2* è più evidente se viene applicato ai miei dati il comando "*smoothing*", che permette di raggruppare frammenti di restrizione in funzione dell'intensità del loro segnale, creando vere e proprie aree in cui i segnali risultano più omogenei e quindi più visibili. Quest'analisi ha così messo in risalto l'interazione che si instaura nel *locus* con più frequenza (Figura 33).

Oltre che con i geni *HBG1/2*, abbiamo potuto osservare che anche una regione situata tra le globine γ e δ ed una regione a monte del gene *HBB* interagiscono con la nostra *anchor region* (Figura 34). La prima contiene lo pseudogene *HBBP1* (chr11:5260613-5264777; 4,1 Kb), mentre la seconda mappa 1,5 Kb a monte del gene *HBB* (chr11:5249858-5250845; 990bp). Tali interazioni sono più evidenti se all'analisi non si applica il comando *smoothing*, ma si valuta in maniera indipendente ogni singolo segnale proveniente da ciascun frammento di restrizione (Figura 34; blocco giallo scuro in corrispondenza della freccia nera per *HBBP1* e piccolo frammento verde indicato con la freccia rosa per la piccola regione in 5' al gene β globinico).

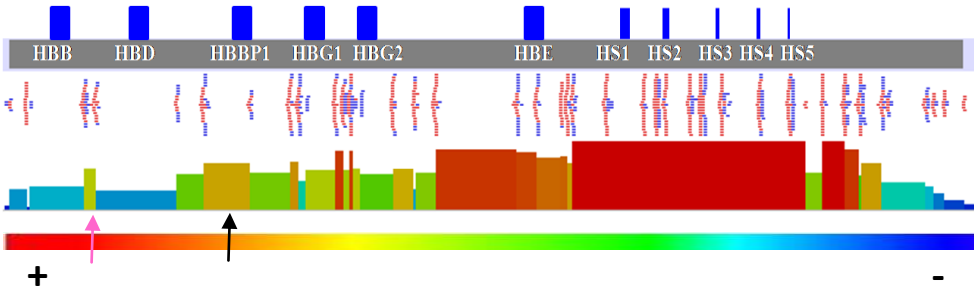

Figura 34: Analisi SeqMonk mediante "Hi-C *other ends*" con l'insieme dei siti HS2, HS3 e HS4 della *LCR* come regione *anchor*.

Kim e colleghi nel 2011 avevano evidenziato un legame tra la *LCR* ed un frammento HindIII che copre oltre il gene β-globinico anche delle regioni a monte ed a valle comprendenti il frammento da noi evidenziato in quest'analisi (990 bp) per un totale di 7,8 Kb (chr11:5243046-5250850).

Mentre, per quanto riguarda l'interazione tra la *LCR* ed una regione contenente lo psuedogene, Dostie e colleghi sono stati i primi ad identificarla su un frammento di 7 Kb (chr11:5259736-5266756).

Nel complesso i nostri risultati già descritti da Dostie e Kim (Dostie et al. 2006; Kim et al. 2011), non solo convalidano lo SCRiBL, in quanto comparabili e sovrapponibili ai dati della letteratura, ma dimostrano che questa metodica presenta una risoluzione e sensibilità maggiore, in quanto permette di rilevare anche le interazioni che avvengono con frequenza minore (come quella con la regione a monte del gene *HBB*) più difficili da rilevare globalmente con le altre tecniche "3C".

Una volta validata la nostra metodica, abbiamo voluto investigare in dettaglio il quadro di interazioni lungo l'intero *locus*. Per questo abbiamo condotto un'analisi indipendente scegliendo come *anchor region* proprio il frammento contenente lo pseudogene *HBBP1* (chr11:5260613-5264777) (Figura 35).

Le interazioni più forti sono state osservate (in rosso) in corrispondenza di frammenti che includono la regione intergenica e i geni *HBG1/2* (Figura 35 frecce verdi; chr11:5272426-5273149 e chr11:5273725-5274717) e quelli che comprendono il sito HS3 e la regione tra l'HS2 e l'HS3 della *LCR* (Figura 35 freccia blu; chr11:5304496-5307382).

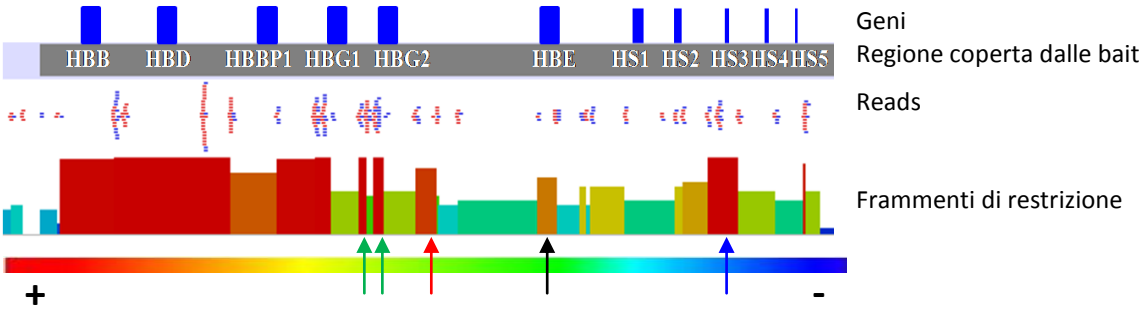

Figura 35: Analisi SeqMonk mediante "Hi-C *other ends*" eseguita con il frammento contenente lo pseudogene *HBBP1* come regione di riferimento.

All'interno del *locus* vi sono altri frammenti che mostrano una frequenza di interazione minore e mappano nella regione intergenica tra i geni *HBG2 e HBE* (Figura 35 freccia rossa; chr11:5277666-5279523), e sullo stesso *HBE* (Figura 35 freccia nera; chr11:5288771-5290592). Al momento il ruolo di tali regioni nella regolazione del *cluster* β-globinico non è nota e studi funzionali mirati saranno necessari per valutare la loro funzione.

Ho effettuato due ulteriori analisi scegliendo come *anchor* i geni globinici *HBD* e *HBB*, per valutare anche il quadro di interazioni associato ai geni globinici adulti (Figure 36 e 37).

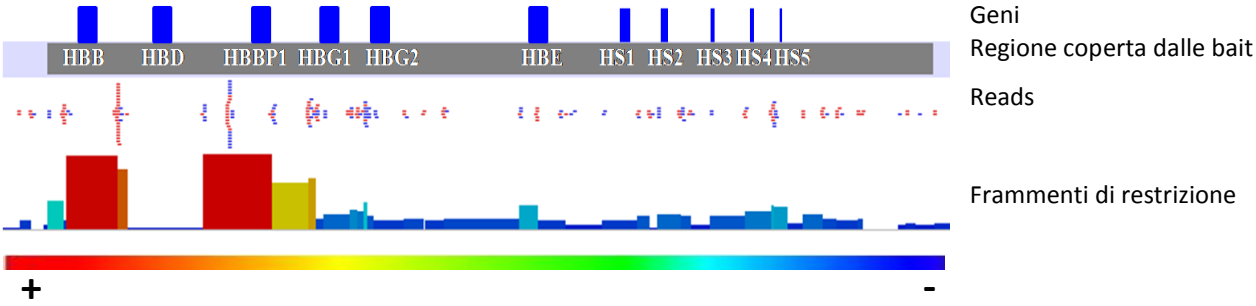

Figura 36: Analisi SeqMonk mediante "Hi-C *other ends*" eseguita con il frammento contenente il gene *HBD* come riferimento.

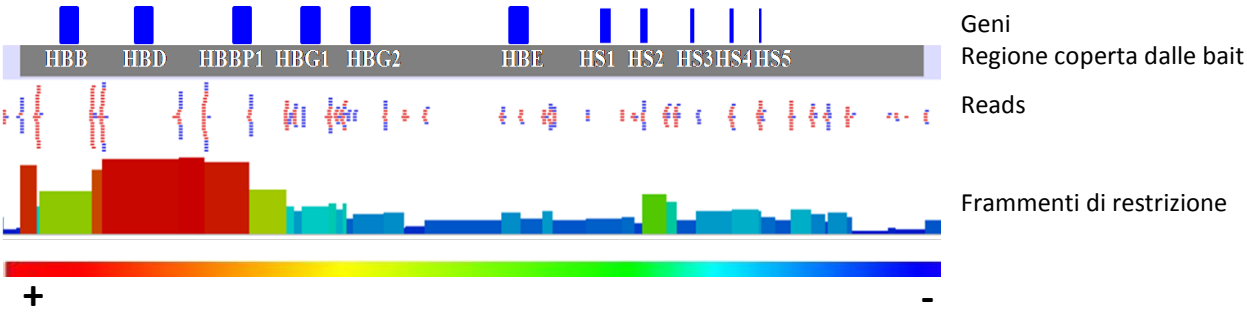

Figura 37: Analisi SeqMonk mediante "Hi-C *other ends*" eseguita con il frammento contenente il gene *HBB* come riferimento.

In entrambi i casi i segnali più alti sono circoscritti ai frammenti immediatamente adiacenti ai geni utilizzati come riferimento, probabilmente perché vicini linearmente nel genoma. Nel resto del *cluster* non vi è la presenza di altri picchi con la stessa intensità. Vi è però una debolissima interazione con alcuni siti della *LCR*: *HBD* interagisce con l'HS5, mentre *HBB* con l'HS2 e con meno intensità con l'HS4.

I dati derivanti dall'Hi-C e dallo SCRiBL sono stati comparati utilizzando il programma SeqMonk per confermare e quantificare la copertura maggiore dello SCRiBL rispetto all'Hi-C.

Considerando il *locus* β-globinico come regione di controllo, sono state quantificate le *reads* in entrambe le librerie in due condizioni differenti: senza *anchor region,* solamente mediante "*read count quantification"* (Figura 38; vedere materiali e metodi) in cui si effettua un calcolo delle *reads* totali per ciascun frammento di restrizione, e usando la *LCR* come riferimento mediante "Hi-C *other ends*" (Figura 39; vedere materiali e metodi). In entrambe le condizioni analizzate è evidente come con lo SCRiBL si ottenga una copertura nettamente superiore della regione target, in termini di numero di *reads,* da cui deriva la maggiore intensità del segnale (Figura 39 B e Figura 40 B).

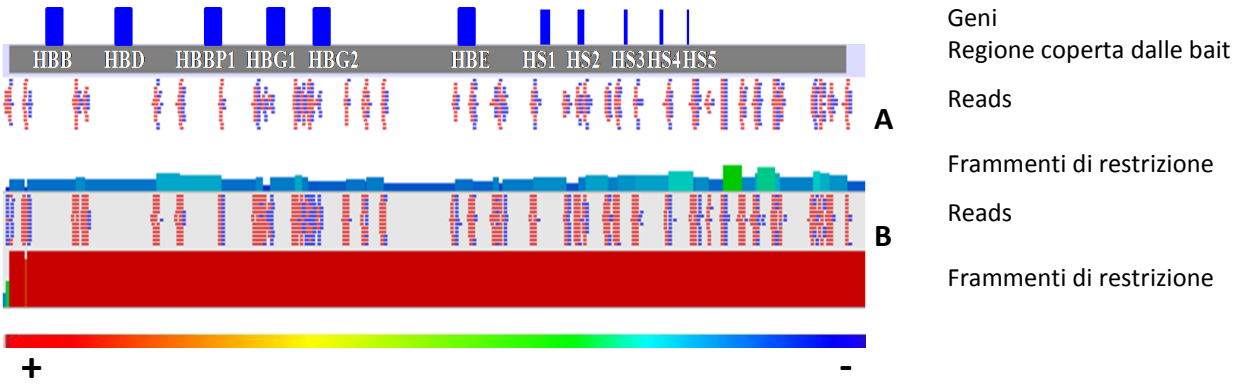

Figura 38: Quantificazione delle *reads* senza utilizzare *anchor region* mediante l'opzione "*read count quantitation*". Il pannello A mostra i dati dell'esperiemnto Hi-C, il B dello SCRiBL.

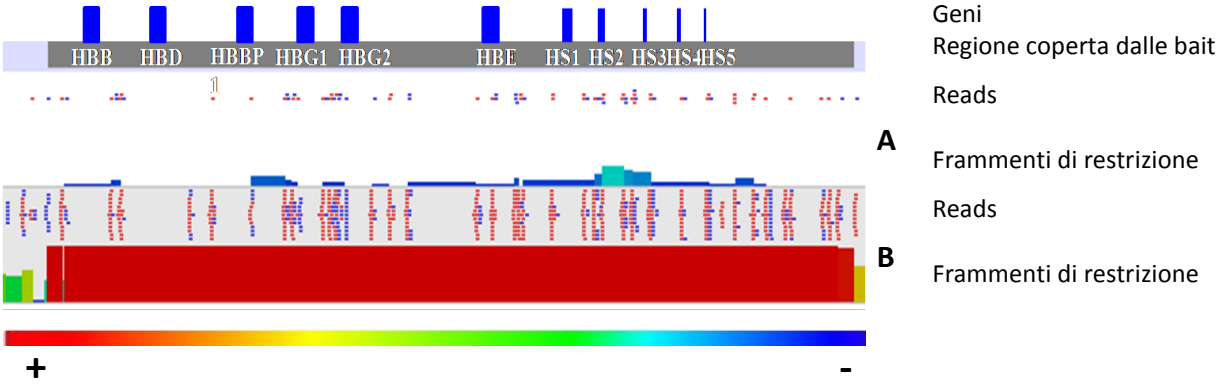

Figura 39: Quantificazione delle *reads* utilizzando la *LCR* come *regione anchor mediante "*Hi-C *other ends"*. Il pannello A mostra i dati dell'Hi-C, mentre il B dello SCRiBL.

Abbiamo ottenuto tali dati nonostante le due librerie Hi-C e SCRiBL siano state sequenziate con modalità differenti, condizione che influenza il numero totale delle *reads* risultanti da una singola *lane* di *flow-cell*. La libreria Hi-C è stata sequenziata con HiSeq 2000 condividendo, per motivi legati ai costi, una *lane* con una differente libreria. La libreria SCRiBL invece è stata sequenziata con il MiSeq in una singola *lane* riducendo notevolmente costi e tempi legati al sequenziamento. Tuttavia, sebbene il MiSeq generi un numero minore di *reads* rispetto all'HiSeq 2000, le caratteristiche stesse dell'esperimento SCRiBL giustificano la nostra

scelta. Infatti mentre le *reads* prodotte dall'Hi-Seq per la libreria Hi-C sono distribuite lungo l'intero genoma, ci si aspetta che quelle generate dal MiSeq per la libreria SCRiBL si concentrino nei *loci* arricchiti con le bait. Questo permette di raggiungere in tali regioni un numero di *reads* comunque nettamente maggiore con la libreria SCRiBL sequenziata con il MiSeq rispetto alla libreria Hi-C sequenziata con l'HiSeq, tale consentire l'individuazione di interazioni significative ad alta risoluzione.
Queste osservazioni dimostrano, come atteso, che tra le due metodiche lo SCRiBL offre una risoluzione maggiore rivelandosi un approccio più robusto e potente.

5.4.2 Selezione degli SNPs

Nel 2008 un gruppo di ricerca all'interno dell'IRGB con cui ho collaborato durante il mio corso di Dottorato ha condotto uno studio GWA su otto tratti quantitativi correlati a parametri ematologici, tra cui i livelli di HbF, in una coorte di 6.148 individui sardi reclutati nell'ambito del progetto ProgeNIA (Pilia et al. 2006; Uda et al. 2008). L'obiettivo era quello di identificare fattori che regolano i livelli di HbF e ne determinano la sua persistenza. Varianti a tre *loci* sono state chiaramente associate ai livelli di HbF: nel *cluster* β-globinico, all'interno dell'introne 2 del gene *BCL11A* e nella regione intergenica *HBS1L-MYB*.
In studi successivi, gli stessi autori hanno collegato direttamente queste varianti al miglioramento della gravità del fenotipo clinico della β-talassemia e dell'anemia falciforme. Altri gruppi hanno condotto studi GWA analoghi sia sulla popolazione generale che su pazienti affetti da β-talassemia e/o anemia falciforme, appartenenti a diverse etnie, confermando l'associazione nelle medesime regioni ai *loci HBS1L-MYB* e *BCL11A* (Menzel et al. 2007; Thein et al. 2007; Lettre et al. 2008).
Non avendo nessuna chiara evidenza circa la funzione biologica degli SNPs associati a questi *loci*, si ipotizza che le varianti in associazione mappino in elementi di DNA regolatori capaci di alterare l'espressione del gene *target,* rappresentando quindi dei fattori chiave che guidano le differenze fenotipiche. In particolare si ritiene che queste sequenze regolatrici includano elementi che possono agire a lunga distanza *in cis* per influenzare l'espressione spazio-temporale di *HBS1L/MYB* e del *BCL11A*. Con il mio progetto di Dottorato ho voluto investigare la presenza di interazioni *in cis* tra le regioni in cui mappano le varianti identificate con gli studi GWAS e il promotore o altre regioni regolatrici a questi *loci*, mediante la nuova metodica SCRiBL. In particolare 12 e 8 SNPs, che mappano rispettivamente nei *loci BCL11A* e *HBS1L-MYB,* sono stati selezionati in base ai dati dei GWAS disponibili in letteratura (Thein et al. 2007; Menzel et al. 2007; Galarneau et al. 2010; Lettre et al. 2008; Uda et al. 2008; Badens et al. 2011; He et al. 2011; Sedgewick et al. 2008; Nuinoon et al. 2010) in associazione con i livelli di l'HbF/numero di cellule F e miglioramento della gravità del fenotipo talassemico e dell'anemia falciforme. La Tabella 12 riporta il numero dell'rs, la posizione sul *build* 37/hg19, il *locus* di appartenenza ed il frammento di restrizione che li contiene.

ID SNP (RS)	POSIZIONE	*LOCUS*	FRAMMENTO DI RESTRIZIONE
rs28384513	135376209 T>G	*HBS1L-MYB*	Chr6:135374498-135376661 "A"
rs7776054	135418916 A>G	*HBS1L-MYB*	Chr6:135417836-135419821 "B"
rs9399137	135419018 T>C	*HBS1L-MYB*	Chr6:135417836-135419821 "B"
rs9389268	135419631 A>G	*HBS1L-MYB*	Chr6:135417836-135419821 "B"
rs9402685	135419688T>C	*HBS1L-MYB*	Chr6:135417836-135419821 "B"
rs4895441	135426573A>G	*HBS1L-MYB*	Chr6:135424365-135427991 "C"
rs9402686	135427817G>A	*HBS1L-MYB*	Chr6:135424365-135427991 "C"
rs4895440	135426558A>T	*HBS1L-MYB*	Chr6:135424365-135427991 "C"
rs6732518	60708597 C>T	*BCL11A*	Chr2:60706711-60711096 "D"
rs10189857	60713235 A>G	*BCL11A*	Chr2:60711097-60716672 "E"
rs6545816	60714861 A>C	*BCL11A*	Chr2:60711097-60716672 "E"
rs11886868	60720246 C>T	*BCL11A*	Chr2:60717201-60720944 "F"
rs7599488	60718347 C>T	*BCL11A*	Chr2:60717201-60720944 "F"
rs1427407	60718043 T>G	*BCL11A*	Chr2:60717201-60720944 "F"
rs766432	60719970 C>A	*BCL11A*	Chr2:60717201-60720944 "F"
rs4671393	60720951 A>G	*BCL11A*	Chr2:60720945-60722183 "G"
rs7557939	60721347 G>A	*BCL11A*	Chr2:60720945-60722183 "G"
rs6706648	60722040 C>T	*BCL11A*	Chr2:60720945-60722183 "G"
rs10184550	60729294 G>A	*BCL11A*	Chr2:60725013-60731839 "H"
rs7606173	60725451 G>C	*BCL11A*	Chr2:60725013-60731839 "H"

Tabella 12: Elenco degli SNPs emersi dagli studi GWAS considerati nel mio studio correlati con i livelli di HbF/ numero di cellule F e miglioramento del fenotipo β-talassemico. I colonna: Identità dello SNP (rs); II colonna: posizione nel genoma e sostituzione nucleotidica; III colonna: *Locus* di appartenenza; IV colonna: Frammento di restrizione che contiene lo SNP.

5.4.3 Analisi della regione intergenica *HBS1L-MYB*

Numerose evidenze hanno dimostrato che la regione intergenica *HBS1L-MYB,* in cui mappano gli SNPs associati dai GWAS, contenga importanti elementi di regolazione per i livelli di HbF. L'esatto meccanismo d'azione non è stato ancora definito, ma l'ipotesi più avvalorata ritiene che tali elementi agiscano *in cis* con interazioni a lunga distanza modulando l'espressione dei geni fiancheggianti *HBS1L* e/o *MYB*, tra cui quest'ultimo rappresenta il migliore candidato.

Anche per questo *locus* in una prima fase abbiamo identificato i frammenti di restrizione contenenti gli SNPs (Tabella 12).

In particolare abbiamo considerato i seguenti 3 frammenti come riferimento: "A" contentente lo SNP rs28384513, "B" con gli rs 7776054, 9399137, 9389268 e 9402685, e "C" che presenta gli rs 4895441, 9402686 e 4895440 (Tabella 12) ed avviato con il programma SeqMonk la ricerca delle interazioni *in cis* lungo il *locus* mediante l'analisi dei dati ottenuti con lo SCRiBL. L'opzione "Hi-C *others ends*" utilizza tali frammenti come *anchor region*, ed individua così le *reads* corrispondenti ad essi provenienti dai prodotti di ligazione sequenziati.

In Figura 40 sono presenti tre pannelli 1, 2, 3 che riportano le analisi effettuate con le 3 *anchor region* rispettivamente.

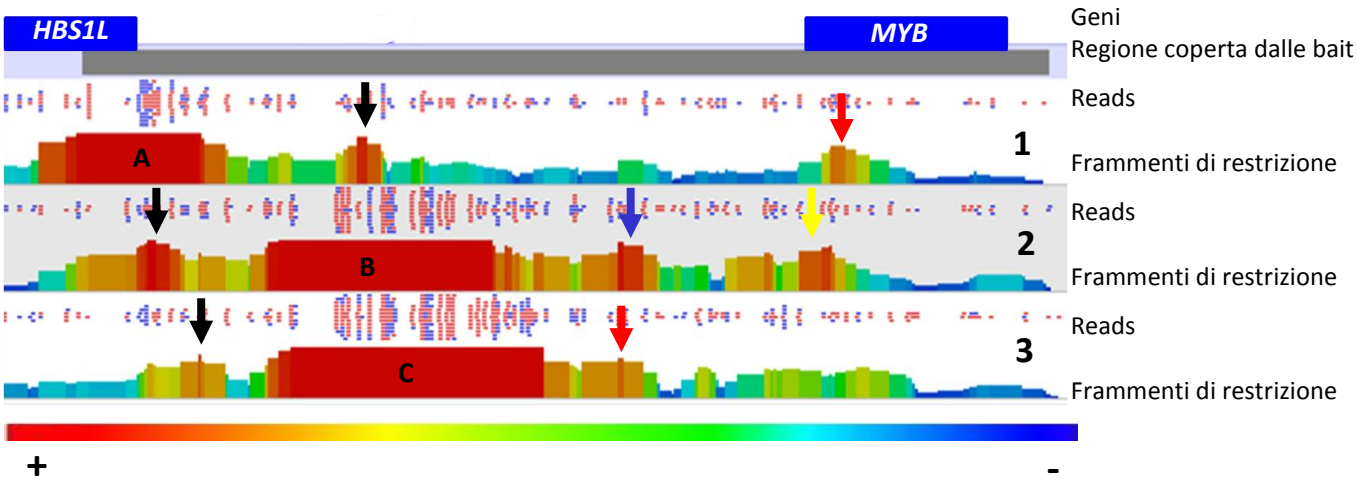

Figura 40: Analisi della regione intergenica *HBS1L-MYB* mediante il programma SeqMonk. Pannello 1: presenta il frammento A come *anchor region*; Pannello 2: presenta il frammento B come *anchor region*; Pannello 3: presenta il frammento C come *anchor region*.

L'*anchor* A (chr6:135374498-135376661) mostra interazioni circoscritte a due regioni (Figura 40 pannello 1): la prima, che possiede un segnale più intenso (chr6:135415956-135422368; indicata con la freccia nera), è localizzata nella porzione intergenica a -84 Kb dal sito di inizio trascrizione di *MYB* che si sovrappone all'*anchor region* B, e la seconda mappa all'interno del gene *MYB* in una regione che comprende gli esoni 2, 3 e 4 (chr6:135506158-135511182; freccia rossa).

Il pannello 2 raffigura l'analisi effettuata con l'*anchor region* B; i segnali più intensi sono circoscritti a 3 regioni: la prima include la regione dell'*anchor* A immediatamente a monte del promotore del gene *HBS1L* (chr6:135374498-135382830; freccia nera), la seconda si trova nella porzione intergenica a -36 Kb dal sito di inizio trascrizione di *MYB* (chr6:135466461-135471244; freccia blu), e la terza si estende dalla regione promotrice di *MYB* fino all'esone 2 (chr6:135500361-135507267; freccia gialla). Mentre quest'ultima è già

stata riportata in altri studi (Stadhouders et al. 2014), le prime due non erano mai state identificate finora utilizzando come riferimento un segmento a -84 Kb.

Infine considerando l'*anchor region* C (pannello 3) l'analisi rivela due piccolissimi picchi di segnale in corrispondenza di due regioni, una a monte del gene *HBS1L* (chr6:135388165-135388461; freccia nera) e la seconda in una porzione intergenica in corrispondenza della regione a -36 kb emersa con l'*anchor region* B (chr6:135466461-135467729; freccia rossa).

Non possiamo tuttavia escludere altre eventuali interazioni con le regioni immediatamente adiacenti alle *anchors* poiché, come precedentemente citato a proposito dell'analisi sul *cluster* β, la stessa vicinanza rende impossibile discriminare la presenza di un'interazione reale, perché gli alti segnali riscontrati sono più probabilmente riconducibili a delle collisioni casuali non funzionali dovute esclusivamente alla prossimità spaziale.

Complessivamente tali risultati indicano la presenza di un contatto fisico tra le regioni contenenti gli SNPs, seppur con differenze nell'intensità del segnale, e le regioni promotrici dei geni *HBS1L-MYB*, guidandoci all'ipotesi di un possibile ruolo regolatorio di tali regioni nell'espressione dei geni adiacenti.

5.4.4 Analisi del *locus BCL11A*

Le varianti associate con i GWAS ai livelli di HbF e al miglioramento del fenotipo clinico della β-talassemia mappano in una regione di circa 15 Kb dell'introne 2 del *BCL11A,* che mostra marcatori cromatinici compatibili con la presenza di elementi genici funzionali con una potenziale attività regolatoria.

Il nostro disegno di studio ha previsto di utilizzare lo stesso approccio applicato allo studio del *locus* β-globinico per identificare la presenza di eventuali interazioni *in cis* con i frammenti in cui mappano le varianti al *locus BCL11A* (Tabella12).

Come accennato nel paragrafo "Elaborazione dei dati SCRiBL con l'algoritmo HiCUP" con l'algoritmo HICUP è stata osservata a questo *locus* una bassa percentuale di arricchimento (14% circa) delle *reads* derivate dallo SCRiBL. Era quindi evidente che durante l'esecuzione della procedura sperimentale, successiva all'Hi-C, doveva essersi verificato un problema tecnico, causa di questo risultato. Mentre la percentuale delle *reads* provenienti dall'Hi-C per il locus *BCL11A* sono paragonabili a quelle ottenute agli altri *loci*.

Inizialmente abbiamo eseguito un'analisi specifica su SeqMonK che prevede la quantificazione delle *reads* validate con HiCUP e la loro precisa localizzazione nel *locus*. Con sorpresa questa analisi ha rivelato che se si esclude il normale rumore di fondo dell'esperimento, un'ampia area che includeva l'intera regione genomica del *BCL11A* di circa 100 Kb, il promotore e tutta la regione 3'UTR, era completamente priva di segnale indicando l'assenza di arricchimento (Figura 41) risultante dalla reazione di ibridazione con le bait.

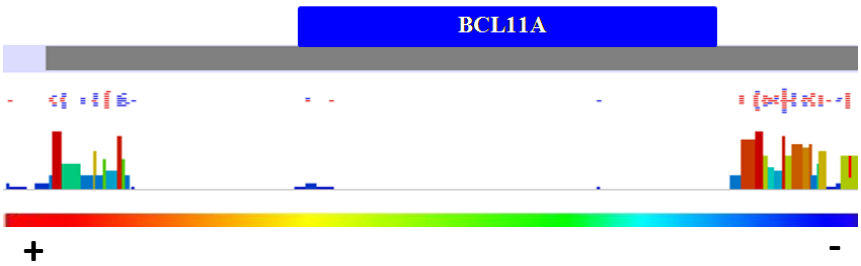

Figura 41: Analisi SeqMonk mediante "*read count quantitation*", ovvero calcolo delle *reads* totali, per ciascun frammento di restrizione.

E' evidente che tale l'arricchimento non è avvenuto con successo ed in maniera uniforme: solo piccole regioni a monte e a valle del gene, all'interno dell'area coperta dalle bait, presentano un segnale.

Per chiarire la causa del problema abbiamo vagliato delle ipotesi, tra cui la possibilità che quella regione non sia stata "catturata" come le altre perché le bait ad RNA potevano presentare dei difetti. Le bait sono state prodotte mediante *trascrizione in vitro* condotta sul DNA genomico estratto dai cloni BAC che coprivano la regione di interesse e di cui erano state verificate solo le estremità 5' e 3' (vedere materiali e metodi).

I cloni BAC, come i plasmidi, vengono prodotti in alte quantità attraverso coltura batterica, sfruttando l'alta capacità proliferativa dei batteri. Sebbene non avvenga di frequente è possibile che durante la coltura si sia verificato un evento di ricombinazione che ha determinato la delezione di un'ampia porzione di DNA nel BAC. Abbiamo quindi valutato la bontà del BAC mediante PCR con cinque coppie di *primers*: due coppie già utilizzate nel controllo iniziale che amplificano ciascuna estremità della sequenza contenuta nel BAC RP11-65A9; 3 coppie invece che amplificano regioni interne al gene *BCL11A* clonato nel BAC. Tra queste ultime, due amplificano sequenze all'interno dell'introne 2, dove sono presenti gli SNPs associati, ed una in un'altra area scelta casualmente all'interno del gene. Per ciascuna amplificazione ho utilizzato un DNA genomico umano come controllo positivo ed ho ottenuto i seguenti prodotti di PCR (Figura 42):

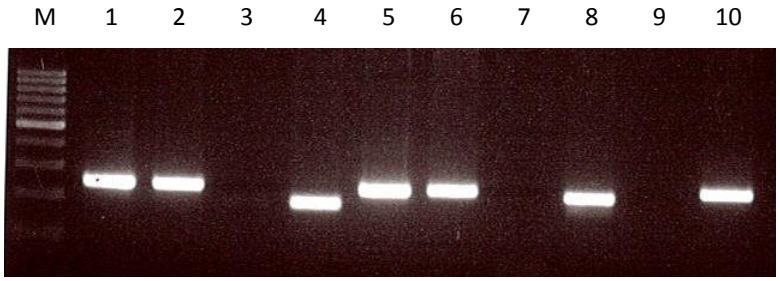

Figura 42: Test di amplicazione per il BAC RP11-65A9 contenente *BCL11A*; sono state testate cinque coppie di *primers* che amplificano regioni diverse lungo la sequenza contenuta nel BAC. 1-3-5-7-9: DNA estratto dal BAC; 2-4-6-8-10: DNA genomico (controllo positivo). 1-2: *primers* nell'estremità 5' del gene; 3-4: *primers* all'interno del gene; 5-6:

primers nell'estremità 3' del gene; 7-8: *primers* nella regioni contenente gli SNP; 9-10: *primers* nella regione contenente gli SNPs; M: Marcatore di peso molecolare noto.

Come mostrato in Figura 42, per il DNA genomico estratto dal BAC abbiamo ottenuto un prodotto di PCR solo con le coppie di *primers* alle estremità della sequenza del clone (pozzetti 2 e 6). I tre prodotti all'interno del gene sono presenti solo nel controllo positivo (pozzetti 4, 8, 10) ed assenti nel BAC (pozzetti 3, 7, 9).

Questo conferma che il DNA derivante dal clone BAC RP11-65A9 non presenta l'intera sequenza ed è privo di un ampia regione che comprende il *BCL11A*. L'ipotesi della delezione in seguito ad un evento di ricombinazione avvenuto durante la crescita batterica è quindi risultata fondata, rendendo di conseguenza impossibile l'analisi delle interazioni nel *locus*.

L'introne 2 che comprende gli SNPs oggetto di studio non può essere esplorato perché all'interno della regione priva di segnale. Questo problema tecnico ha compromesso l'esecuzione di un'analisi affidabile e riflette gli inconvenienti tecnici che spesso si presentano quando si sviluppa una nuova metodica.

6. DISCUSSIONE

Le tecniche di "*Chromosome Conformation Capture"*, superando i limiti della microscopia ottica, hanno rivoluzionato lo studio dell'architettura della cromatina. Sequenze di DNA in prossimità spaziale nel nucleo o impegnate in interazioni fisiche possono essere valutate quantitativamente per fornire una misura che riflette potenzialmente la loro frequenza di associazione e/o prossimità. Grazie all'applicazione di queste metodiche si è potuto osservare *in vivo* come il genoma sia organizzato in una complessa rete di contatti tra geni ed elementi regolatori sullo stesso cromosoma (intracromosomiche, *in cis*) o tra diversi cromosomi (intercromosomiche, *in trans*), mettendo in luce la relazione funzionale tra organizzazione nucleare e trascrizione genica. Queste evidenze hanno rappresentato una svolta importante nella ricerca dei meccanismi molecolari alla base del gran numero di associazioni geniche identificate dai GWAS, che spesso mappano in regioni del DNA non codificanti. Infatti queste, seppur senza un chiaro ruolo biologico, sembrerebbero implicate nella regolazione dell'espressione di geni bersaglio spesso in maniera spazio e stadio specifica.

Con il mio progetto di tesi mi sono proposta di studiare il quadro di interazioni cromatiniche ai *loci HBS1L-MYB* e *BCL11A* identificati dai GWAS come regioni chiave associate alla modulazione dei livelli di HbF e al miglioramento del fenotipo clinico della β-talassemia. Per raggiungere tale scopo abbiamo ottimizzato un protocollo per l'Hi-C in cellule K562 con caratteristiche eritroidi fetali, e applicato una nuova metodica "3C" ad alta risoluzione denominata SCRiBL, validata mediante l'analisi del *cluster* β-globinico.

6.1 Hi-C e SCRiBL

La tecnica Hi-C "cattura" i contatti cromatinici che si instaurano in una determinata popolazione cellulare mediante l'utilizzo della formaldeide, un potente agente fissante, a cui fa seguito una digestione enzimatica, lisi, ligazione, massivo sequenziamento della libreria che ne deriva con le piattaforme di nuova generazione (NGS), e allineamento delle sequenze sul genoma. Sebbene l'Hi-C sia una metodica ad alta risoluzione se comparata al 3C, in grado di rivelare il complesso intreccio di contatti dinamici per la prima volta lungo tutto il genoma, risulta meno sensibile quando si intende analizzare il quadro di interazioni all'interno di specifici *loci* (de Wit, de Laat. 2012; Dekker. 2006), risolvendo infatti non meno di 1 Mb. Inoltre, la sua applicazione è limitata anche da altri fattori tra cui i costi elevati, principalmente legati all'alta copertura richiesta nel sequenziamento, e la complessità di gestione e analisi bioinformatica dei dati prodotti.

L'ottimizzazione del protocollo Hi-C con l'introduzione di una doppia digestione enzimatica e della ligazione "in nuclei", ci ha permesso da un lato di sfruttare la potenza dell'Hi-C e allo stesso tempo di migliorarne la risoluzione.

Per la prima volta infatti, per la creazione della libreria Hi-C, il genoma è stato frammentato da due endonucleasi di restrizione 6 *cutter*, HindIII e BglII. Questo approccio ha permesso di generare un numero maggiore di frammenti di restrizione, che rappresentano l'unità di misura utilizzata per il rilevamento delle interazioni, e di conseguenza di ridurne le dimensioni rispetto ai protocolli standard.

I punti critici della prima fase dell'esperimento, come la digestione, il *fill-in*, la biotinilazione e l'evento di ligazione, hanno ampiamente superato i numerosi e stringenti controlli di qualità (vedere materiali e metodi e risultati). I prodotti di ligazione così generati sono stati sottoposti ad ulteriori passaggi sperimentali e controlli qualitativi e quantitativi che ci hanno permesso di ottenere la libreria finale Hi-C idonea per le applicazioni successive: sequenziamento con la piattaforma HiSeq 2000 Illumina e SCRiBL. Il sequenziamento *paired-end* della libreria Hi-C è stato eseguito sia per verificare direttamente la qualità dello stesso esperimento Hi-C, che parallelamente per confrontare i due approcci Hi-C e SCRiBL in termini di qualità e risoluzione.

La nuova metodica SCRiBL, sviluppata per arricchire specifiche regioni di interesse mediante selezione per ibridazione su fase liquida, ha dimostrato di essere rispetto l'Hi-C più flessibile, efficiente e di avere una maggiore sensibilità in termini di numero di nucleotidi risolti.

I *loci* β-globinico, *HBS1L-MYB* e *BCL11A* oggetto del nostro studio sono stati selettivamente arricchiti mediante ibridazione della libreria Hi-C con le bait ad RNA biotinilate, originate per trascrizione *in vitro* del relativo DNA genomico veicolato all'interno di cloni BAC. La libreria SCRiBL risultante ha superato tutti i controlli quantitativi e qualitativi in termini di concentrazione e di dimensione media dei frammenti, rispettando quindi i parametri richiesti per il sequenziamento *paired-end* mediante MiSeq Illumina.

Per verificare l'efficienza della procedura sperimentale adottata per la preparazione delle due librerie, i dati risultanti dal sequenziamento di entrambe dopo essere stati allineati al genoma di riferimento sono stati filtrati ed elaborati con la pipeline HiCUP, che seleziona solamente le coppie di *reads* (*di-tag*) informative eliminando quelle non valide dovute a *bias* sperimentali. Questo primo filtro fornisce un'ulteriore indicazione sulla qualità della libreria, infatti maggiore sarà il numero dei *di-tag* informativi maggiore sarà la qualità. In entrambi i casi la percentuale delle coppie di *reads* valide riscontrata è molto elevata: 87,3% per l'Hi-C e 92,3% per lo SCRiBL.

HiCUP inoltre ha consentito di valutare da un punto di vista qualitativo e tecnico la nuova metodica SCRiBL, stabilendo l'efficienza dell'esperimento e la percentuale di arricchimento per ogni *locus* rispetto all'Hi-C. Questi test hanno evidenziato un *background* maggiore rispetto a quello atteso che si riflette in un 6,7% di *reads* che coprono le regioni bersaglio, dovuto al sequenziamento di molte molecole di DNA *target-off target*, ovvero frammenti di DNA che mappano anche al di fuori dalle regioni di arricchimento. Queste vengono generate probabilmente dall'ibridazione delle sequenze terminali delle molecole della libreria che diventano "adesive", in quanto presentano le stesse estremità in seguito all'incorporazione degli adattatori

a tutti i prodotti di ligazione. Tale evento era atteso, sebbene in misura minore. Infatti prima dell'ibridazione sono stati addizionati alla libreria Hi-C sia dei bloccanti, oligonucleotidi che riconoscono tali estremità, che reagenti quali "Salmon Sperm" e "Human Cot-1 DNA" che hanno la proprietà di bloccare ibridazioni aspecifiche. Visto però il risultato non possiamo escludere che il legame con le sequenze *off-target* si sia verificato prima dell'ibridazione, o che i bloccanti utilizzati non siano ideali in termini di sequenza, dimensione e/o concentrazione per esplicare tale funzione.

Questo risultato potrebbe sollevare dei dubbi riguardo la maggiore efficienza e robustezza dello SCRiBL rispetto all'Hi-C. Tuttavia il confronto fornito da HiCUP tra le due metodiche, identificato come percentuale di arricchimento per tutte le regioni target, derivata dal rapporto tra le *reads* provenienti dallo SCRiBL e da quelle dell'Hi-C, è molto alta, superando in alcuni casi il 90%. Il *locus BCL11A* è l'unico che presenta un valore intorno al 15%, dovuto ad un problema tecnico avvenuto durante la procedura sperimentale di cui si parlerà più avanti in dettaglio (vedere paragrafo "Analisi del *locus BCL11A*"). Quindi complessivamente possiamo affermare che, nonostante si possa migliorare il protocollo per diminuire il *background* ed aumentare l'efficienza di arricchimento, i dati forniti da HiCUP sono tali da ritenere lo SCRiBL un metodica ideale per indagare la presenza di interazioni cromatiniche simultaneamente su specifici *loci* e con una risoluzione maggiore dell'Hi-C.

6.2 Validazione dello SCRiBL mediante analisi del *locus* β-globinico

Nonostante gli incoraggianti risultati ottenuti dall'analisi con HiCUP, si è reso necessario validare l'esperimento SCRiBL. Le innumerevoli conoscenze acquisite sui meccanismi molecolari che governano l'espressione dei geni globinici e le interazioni cromatiniche precedentemente definite con altre metodiche "3C" in cellule K562, rendono il *cluster* β-globinico il modello ideale di validazione (Dostie et al. 2006; Kim et al. 2011).

Mediante l'analisi dei dati derivati dallo SCRiBL con il software SeqMonk abbiamo confermato le più importanti interazioni all'interno del *locus* β-globinico descritte in letteratura, prima fra tutte quella più forte tra la *LCR* e i geni γ globinici, espressi nelle K562.

Inoltre lo SCRiBL ha replicato l'interazione tra la *LCR* e la regione intergenica tra i geni γ e δ globinici (contenente lo pseudogene *HBBP1*) riportata per la prima volta nel 2006 da Dostie e colleghi in un esperimento di 5C. Grazie all'utilizzo di due enzimi di restrizione per la costruzione della libreria è stato possibile ridurre le dimensioni della regione coinvolta nel legame fino a circa 4 Kb (chr11:5260613-5264777) rispetto le 7 Kb del frammento EcoRI di Dostie. L'analisi delle interazioni considerando *HBBP1* come *anchor* ha riportato come atteso il contatto con la *LCR*, ma anche con i geni γ, che rappresentano le due aree del *cluster* che instaurano contatti con la più alta frequenza. Questi risultati suggeriscono che *HBBP1* possa avere un ruolo funzionale agendo per esempio come *anchor* per guidare i *loop* di cromatina

LCR-dipendenti, meccanismo cruciale per coordinare l'espressione spazio-temporale dei geni globinici (Holwerda, de Laat. 2012). Altre evidenze supportano tale ipotesi, come la localizzazione all'interno del frammento di 4 Kb ed esattamente nell'introne 2 di *HBBP1* di due SNPs (rs10128556 e rs2071348) correlati alla modulazione dei livelli di HbF e al miglioramento del fenotipo β-talassemico (Sherva et al. 2010; Nuinoon et al. 2010). Inoltre, come riportato nei dati ENCODE (genome.ucsc.edu/ENCODE), disponibili su UCSC (genome.ucsc.edu), ed Ensembl (www.ensembl.org/index.html), il frammento contenente lo pseudogene mostra in varie linee cellulari tra cui le K562, la presenza di marcatori che indicano una regione di cromatina trascrizionalmente attiva, come siti ipersensibili alla DNase I, presenza di istoni H3K27Ac, H3K9Ac, H3K4Me1 e H3K4M3 considerati marcatori di regioni con potenziale attività di promotore*/enhancer,* e di siti di legame per importanti fattori di trascrizione (GATA, MYC, CEBPB).

Seppur con minore intensità è emerso un segnale di interazione molto interessante tra la *LCR* ed un frammento di circa 990 bp che mappa a 1,5 Kb a monte del gene β-globinico (chr11:5249858-5250845). Kim e colleghi nel 2011 (Kim et al. 2011) avevano descritto la stessa interazione ma con un frammento di maggiori dimensioni, circa 7,8 Kb (chr11:5243046-5250850), che include il nostro frammento di 990 bp, tutto il gene *HBB* più una regione a valle. A differenza di questi autori, che hanno utilizzato per l'esperimento di 3C HindIII come unico enzima di restrizione per frammentare il genoma, la nostra scelta di eseguire per la prima volta una doppia digestione si è dimostrata di successo, permettendo infatti di ottenere una risoluzione nettamente superiore rispetto agli altri approcci (7,8 Kb Kim et al. versus 990 bp nel nostro caso). Nonostante questa interazione, così come le altre, debba essere ulteriormente validata con approcci indipendenti, possiamo speculare che all'interno delle 990 bp vi sia un elemento regolatore, a cui corrisponde un sito ipersensibile alla DNase I (ENCODE: genome.ucsc.edu/ENCODE; UCSC: genome.ucsc.genome), che interagendo con la *LCR* sarebbe in grado di silenziare in K562 il gene β-globinico, che infatti risulta represso in queste cellule.

Infine l'ultima considerazione riguarda il confronto tra i due approcci, l'Hi-C e lo SCRiBL, utilizzando il *locus* β-globinico come modello. SeqMonk ci ha permesso infatti di avere un riscontro visivo delle informazioni fornite dalla *pipeline* HiCUP, evidenziando infatti come sia il numero complessivo di *reads* che l'intensità dei segnali siano nettamente superiori nello SCRiBL.

Pertanto possiamo affermare di aver validato una nuova metodica capace di riscontrare interazioni cromatiniche a lunga distanza, in maniera sistematica, obiettiva e con un'altissima risoluzione.

6.3 Analisi delle interazioni nella regione intergenica *HBS1L-MYB*

La regione intergenica *HBS1L-MYB,* di circa 126 Kb, mappa diverse varianti geniche identificate negli studi GWAS come associate ai livelli di HbF/numero di cellule F e al miglioramento clinico di β-talassemia e anemia falciforme. Con l'obiettivo di fornire una spiegazione funzionale per queste associazioni, alcuni studi

hanno evidenziato come le varianti geniche alterino il legame di fattori di trascrizione chiave, talvolta eritroidi specifici, in elementi regolatori che interagiscono a lunga distanza *in cis* con il gene *MYB,* regolandone l'espressione (Stadhouders et al. 2014).

Con questa attività ricerca abbiamo indagato le interazioni cromatiniche lungo il *locus HBS1L-MYB* applicando la nuova metodica SCRiBL che, avendo un potere risolutivo maggiore di altre metodiche "3C", poteva fornire nuove informazioni sulle modalità d'azione delle varianti geniche identificate con i GWAS.

L'analisi dei dati è stata condotta con il programma SeqMonk utilizzando tre *anchor region* (A, B e C) corrispondenti ai tre frammenti di restrizione contenenti ciascuno almeno uno degli SNPs associati (Tabella 12).

L'*anchor* A, comprendente l'rs 28384513 localizzato nella regione promotrice del gene *HBS1L* e che si estende fino a coprirne il primo esone e parte del primo introne, interagisce con due regioni: una che si sovrappone all' *anchor* B (Chr6:135415956-135422368), contenente gli rs 7776054, 9399137, 9389268, 9402685 e l'rs 66650371 che identifica una delezione di 3 bp, e la seconda di 5 Kb che comprende gli esoni 2, 3 e 4 del *MYB*.

Utilizzando come *anchor* la regione B, si conferma l'interazione con l'*anchor* A (vista precedentemente) e si evidenziano altri due intensi segnali: uno comprende il promotore del gene *MYB* fino all'esone 2 (chr6:135500361-135507267) ed il secondo copre una regione di circa 4,7 Kb a -36 Kb dal sito di inizio della trascrizione di *MYB*.

Infine l'ultima regione di riferimento, C, che include gli rs 4895441, 9402686 e 4895440 risulta avere segnali più deboli rispetto ai precedenti che interessano due piccole regioni, una a -114 Kb e l'altra a -36 Kb dal sito di inizio della trascrizione di *MYB.* Quest'ultima corrisponde ad una porzione dell'area già emersa nell'analisi precedente effettuata con l'*anchor* B, confermando il potenziale ruolo nel definire la conformazione cromatinica del *locus* e quindi dei meccanismi che regolano i processi funzionali.

Nel loro complesso tutte le interazioni riscontrate lungo il *locus* supporterebbero l'ipotesi secondo cui le regioni identificate dagli SNPs agiscano modulando i livelli di HbF attraverso la regolazione dei geni fiancheggianti, *HBS1L* e *MYB*. Tuttavia, sebbene sia stata osservata una riduzione contemporanea dell'espressione di entrambi i geni in individui con elevati livelli di HbF, la maggior parte degli studi funzionali condotti sia sull'uomo che su modelli murini indicano il *MYB* quale migliore candidato nella regolazione dei geni γ globinici e quindi dei livelli di HbF (Jiang et al. 2006; Wahlberg et al. 2009). Pertanto tutte le interazioni che coinvolgono il frammento 135374498-135376661, che comprende il 5' dell'*HBS1L* dove mappa lo SNP associato all'HbF dai GWAS, sono di difficile interpretazione. Pur non potendo escludere un ruolo diretto dell'*HBS1L* nella regolazione dei geni γ globinici, i risultati finora disponibili indirizzano piuttosto verso un meccanismo di co-regolazione dei geni a questo *locus*. Solo studi funzionali mirati potranno meglio definire tale quadro.

Tra le interazioni più interessanti identificate con la nostra analisi emerge quella tra i frammenti *anchor* B e C con una regione intergenica a circa -36 Kb dall'inizio della trascrizione di *MYB* (chr6:135466461-135471244).

Dati di ENCODE (genome.ucsc.edu/ENCODE) pubblicamente disponibili su UCSC Genome Browser (genome.ucsc.edu), indicano in K562 la presenza in tale regione di marcatori istonici quali H3K4me1 e H3K27ac, siti ipersensibili alla DNase I e numerosi siti di legame per fattori di trascrizione, tra cui oltre a proteine eritroidi specifiche (GATA1 e GATA2) sono stati rilevati diversi fattori legati al rimodellamento della cromatina (CTCF, p300 e BRG1 che fa parte del complesso SWI/SNF), che definiscono questa, come un'area di cromatina attiva con potenziali funzioni *enhancer*.

Considerando i risultati nel loro insieme possiamo postulare il modello in Figura 43, in cui si evince la presenza di un preciso e sottile meccanismo di regolazione, in cui le regioni emerse dall'analisi instaurano interazioni tra loro formando una struttura a fiore, simile a quella formulata per gli stessi *loci* murino e umano in altri lavori di letteratura (Stadhouders et al. 2012; Stadhouders et al. 2014), sebbene le regioni interessate non siano completamente sovrapponibili.

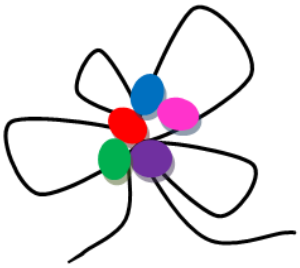

Figura 43: Modello rappresentante le interazioni cromatiniche identificate dall'analisi con SeqMonk, implicate nella regolazione del gene *MYB* e potenzialmente dell'*HBS1L*.
Verde: regione *anchor* A (comprendente anche il promotore del gene *HBS1L)*;
Blu: regione intergenica 135466461-135471244 a -36 Kb dal sito di inizio della trascrizione di *MYB*;
Rosso: regione *anchor* B;
Viola: regione promotrice di *MYB* (comprendente la prima parte del gene fino all'esone 2);
Rosa: regione *anchor* C.

La presenza dei polimorfismi in queste regioni potrebbe alterare la sequenza di legame ed il reclutamento di fattori di trascrizione, oppure influenzare la struttura della cromatina nei frammenti adiacenti, alterando il normale meccanismo di regolazione, presumibilmente, dell'espressione di *MYB*. Considerando i dati di letteratura infatti è noto il suo ruolo chiave nel processo di eritropoiesi. Nonostante il suo meccanismo d'azione sulla regolazione dei livelli di HbF non sia ancora stato definito chiaramente, sono state avanzate delle ipotesi. Quella più accreditata riguarda il ruolo giocato dal MYB nella cinetica dell'eritropoiesi: bassi livelli porterebbero ad una progressione più lenta nel ciclo cellulare ed un'accelerazione nel processo di differenziazione (Jiang et al. 2006; Sankaran et al. 2011; Bianchi et al. 2010); così la terminazione prematura del ciclo cellulare durante l'eritropoiesi adulta produrrebbe più cellule eritroidi che sintetizzano HbF rispetto ad HbA (Stamatoyannopoulous. 2005). Un secondo meccanismo d'azione è stato avanzato in un lavoro recentissimo di Stadhouders (Stadhouders et al. 2014), che ha mostrato l'occupazione di MYB lungo

il *cluster* β-globinico e i geni *KLF1* e *BCL11A*, postulando che l'oncogene MYB agisca direttamente attivando tali regolatori negativi. Altri dati di letteratura hanno inoltre osservato che una repressione di *MYB* condurrebbe ad un'importante riduzione di tali legami, portando alla mancata attivazione dei repressori della produzione delle catene γ globiniche, la cui espressione quindi non verrebbe "spenta" (Bianchi et al. 2010; Sankaran et al. 2011; Suzuki et al. 2013).

I risultati dello SCRiBL caratterizzano ulteriormente la regione intergenica *HBS1L-MYB* mettendo in luce sia le regioni contenenti le varianti geniche, già ampiamente studiate in ricerche precedenti, che la nuova regione (a -36 Kb) che viene coinvolta insieme alle altre nella formazione di *loop* potenzialmente in grado di regolare i processi di trascrizione al *locus*.

I risultati ottenuti portano ad una maggiore comprensione dei reali meccanismi alla base delle varianti geniche associate, fornendo informazioni significative utili per la manipolazione di questo processo di regolazione, che emerge come un promettente target terapeutico per lo sviluppo di cure alternative per le β-emoglobinopatie.

6.4 Problemi tecnici emersi nell'analisi del *locus BCL11A*

Come per i *loci* precedentemente descritti anche il *BCL11A* è stato coinvolto nella regolazione dei livelli di HbF e nel miglioramento clinico della β-talassemia da studi GWAS. Questi hanno identificato varianti geniche associate circoscritte in una regione di circa 15 Kb nell'introne 2 del gene. Studi funzionali hanno evidenziato che tale regione svolge un ruolo di *enhancer* eritroide specifico regolando l'espressione del *BCL11A* (Bauer et al. 2013).

Lo SCRiBL a questo *locus* è stato disegnato per verificare la presenza di interazioni a lunga distanza tra questa regione *enahcer* ed il resto del *locus BCL11A* che spieghi il ruolo biologico alla base dell'associazione genetica.

L'analisi prima con HiCUP e poi con SeqMonk dei risultati dello SCRiBL ha portato alla luce un serio problema tecnico che purtroppo ha compromesso l'esperimento e quindi il rilevamento di interazioni al *locus*. La percentuale di arricchimento di questa regione, secondo le statistiche fornite da HiCUP, è risultata estremamente bassa (14%) rispetto agli altri loci (80-90%), senza una ragione apparente. Una più approfondita ricerca ha rivelato però la delezione dell'intero gene e del suo promotore nella sequenza genica veicolata dal clone BAC, di cui erano state verificate solamente le estremità 5' e 3'. Di conseguenza le bait ad RNA non contenendo tale regione non hanno potuto arricchirla. Anche la quantificazione delle *reads* su SeqMonk ha confermato una povertà di segnali rendendo impossibile un'analisi efficiente e affidabile dei contatti cromatinici.

Sicuramente questo risultato ci ha indotto a valutare criticamente tutti i reagenti utilizzati nelle varie fasi del protocollo e ad inserire ulteriori punti di controllo, primo fra tutti la verifica della corretta identità della regione genica contenuta nel BAC.

6.5 Prospettive per il futuro

Il progetto della mia tesi di Dottorato ha previsto l'applicazione per la prima volta della nuova metodica "3C" SCRiBL in cellule eritroidi K562, per l'identificazione di interazioni a lunga distanza su specifici *loci* simultaneamente.

I risultati raggiunti dimostrano che lo SCRiBL è un approccio potente, efficiente e con una risoluzione maggiore di altre metodiche "3C", tra cui l'Hi-C con il quale è stato comparato in questa ricerca.

Nonostante questo, come per tutte le metodiche utilizzate per la prima volta, anche lo SCRiBL presenta notevoli margini di miglioramento che possono essere raggiunti agendo sia sul protocollo sperimentale che sull'analisi dei dati prodotti.

In particolare nel prossimo futuro per ottimizzare lo SCRiBL intendiamo eseguire quanto segue:

1) Costruire un nuovo set di bait ad RNA che comprenda l'intera sequenza genomica del *BCL11A.*
2) Migliorare la specificità dell'evento di ibridazione tra la libreria Hi-C e le bait ad RNA, modificando opportunamente le loro concentrazioni; utilizzando nuove sequenze bloccanti e testando in quali punti del protocollo è più indicata la loro incorporazione; intervenendo sulle temperature di ibridazione per aumentare la stringenza.
3) Sviluppare una *pipeline* specifica per l'analisi dei dati SCRiBL. Infatti i risultati sono stati finora analizzati con un algoritmo messo a punto per esperimenti Hi-C e quindi non completamente adatto all'elaborazione dei dati SCRiBL. Questo rappresenta forse il punto cruciale su cui intervenire per: estrapolare da librerie così complesse il numero maggiore di informazioni possibili; discriminare in maniera efficiente i reali prodotti di ligazione tra i frammenti *cross-linkati* contro le interazioni *random*; risolvere le interazioni tra frammenti adiacenti o separati da poche Kb, le cui intensità di segnale risultano così alte spesso solo in funzione della loro prossimità spaziale; valutare le interazioni *in trans* di più difficile individuazione con gli strumenti a disposizione al momento; normalizzare i dati che possiedono livelli differenti di arricchimento derivati da prodotti di ligazione di cui una delle *reads* può mappare sia al di fuori che dentro le bait. Per concludere, una nuova *pipeline* ci permetterebbe quindi di produrre dati ancora più affidabili e sensibili rendendo lo SCRiBL più potente ed efficiente.
4) Eseguire l'esperimento in progenitori eritroidi derivati da sangue periferico di individui sani e pazienti affetti da talassemia *major* ed intermedia con aplotipi estremi associati ai livelli di HbF per i loci *HBS1L-MYB* e *BCL11A*. Ormai è risaputo che le interazioni a lunga distanza sono implicate nella

regolazione dell'espressione genica di elementi target, e che la presenza di varianti geniche all'interno di regioni ad azione regolatoria potrebbe alterare il normale processo di trascrizione attraverso un effetto diretto o indiretto sulla cromatina e sul reclutamento dei vari fattori di trascrizione. L'obiettivo è quello di esplorare la conformazione cromatinica dei *loci* presi in esame per individuare interazioni non solo eritroidi, ma anche aplotipo specifiche. Tali informazioni produrrebbero nuove valutazioni per la comprensione dei meccanismi biologici alla base dell'azione svolta dalle varianti genetiche e guiderebbero la scelta di nuovi target terapeutici alternativi per l'induzione di HbF nell'adulto, portando al miglioramento fenotipico della β-talassemia.

5) Validare le interazioni identificate con lo SCRiBL. Le regioni di interazione individuate saranno successivamente confermate con metodi indipendenti tra cui 3C e FISH ad alta risoluzione (3D-FISH o crio-FISH) che utilizzano metodiche di fissazione che conservano bene l'ultrastruttura nucleare. Infine le regioni validate saranno ulteriormente studiate mediante saggi funzionali tra cui quelli luceferasici per evidenziare una potenziale azione *enhancer/silencer/promoter*.

6.6 Conclusioni

Gli studi di associazione estesi all'intero genoma hanno permesso di identificare varianti geniche associate ad un numero molto elevato di malattie complesse e tratti quantitativi, la maggior parte delle quali mappa in sequenze non codificanti del DNA e pertanto di difficile interpretazione da un punto di vista funzionale. Grazie all'avvento delle tecniche sullo studio dell'architettura della cromatina è stato rivelato che le varianti geniche, localizzate all'interno di elementi di regolazione del DNA, agiscono attraverso interazioni a lunga distanza dirette su altri elementi regolatori o geni target. (Dekker. 2006; de Laat, Grosveld. 2003; Chambeyron, Bickmore. 2004). Questo influenza processi biologici come la trascrizione genica riflettendo inoltre il complesso intreccio di interazioni cromatiniche tra geni ed elementi funzionali del genoma. Le tecnologie "3C" sfruttano il potente potere fissativo della formaldeide per "scattare un'istantanea" delle interazioni tra segmenti di cromatina lungo tutto il genoma, catturandole *in vivo* in quel preciso momento. La frequenza con la quale i frammenti di restrizione verranno poi ligati misura la frequenza con la quale interagiscono nel nucleo (Dekker et al. 2002).

Col nostro progetto di ricerca partendo dalla libreria Hi-C abbiamo applicato e validato la nuova metodica SCRiBL per il riscontro delle interazioni ad altissima risoluzione sulla regione intergenica *HBS1L-MYB* e sul *locus* contenente *BCL11A,* e analizzato il *cluster* β-globinico, nella linea cellulare K562.

I risultati ottenuti mostrano contatti importanti tra frammenti situati nella regione intergenica *HBS1L-MYB*, alcuni contenenti gli SNPs associati ai livelli di HbF, e le regioni promotrici di *MYB* e *HBS1L*, suggerendo quindi un'azione diretta su tali geni che a loro volta andrebbero a modulare l'espressione delle catene γ e

quindi di HbF. Ciononostante non è chiaro il coinvolgimento di *HBS1L,* mentre invece l'implicazione di *MYB* è supportata dal suo ruolo nell'eritropoiesi.

Sono stati postulati due meccanismi che potrebbero essere responsabili dell'azione di MYB sui livelli di catene γ globiniche: il primo riguarda la partecipazione alle cinetiche di eritropoiesi attraverso un bilanciamento dei processi di proliferazione e differenziazione dei progenitori eritroidi, mentre il secondo si basa sull'attivazione diretta dei regolatori negativi BCL11A e KLF1.

I risultati ottenuti sottolineano il ruolo di MYB come potenziale target terapeutico per la riattivazione della produzione di HbF nei pazienti affetti da β-talassemia.

Nel complesso questa ricerca evidenzia l'efficienza e la potenza delle metodiche "3C" Hi-C e SCRiBL nel delucidare meccanismi d'azione e i processi di regolazione che coinvolgono vari elementi del genoma come geni, *enhancers, silencers, insulators*, e le varianti geniche, a cui spesso non è facile attribuire un preciso ruolo funzionale o definire il bersaglio su cui agiscono.

7. BIBLIOGRAFIA

Badens C, Joly P, Agouti I, Thuret I, Gonnet K, Fattoum S, Francina A, Simeoni MC, Loundou A, Pissard S. **Variants in genetic modifiers of β-thalassemia can help to predict the major or intermedia type of the disease.** Haematologica 2011; 96(11):1712–1714.

Bantignies F, Grimaud C, Lavrov S, Gabut M, Cavalli G. **Inheritance of Polycomb-dependent chromosomal interactions in Drosophila.** Genes Dev. 2003 Oct 1; 17(19):2406-20.

Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, Shao Z, Canver MC, Smith EC, Pinello L, Sabo PJ, Vierstra J, Voit RA, Yuan GC, Porteus MH,Stamatoyannopoulos JA, Lettre G, Orkin SH. **An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level.** Science. 2013 Oct 11; 342(6155):253-7.

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. **Hi-C: a comprehensive technique to capture the conformation of genomes.** Methods. 2012 Nov; 58(3):268-76.

Bianchi E, Zini R, Salati S, Tenedini E, Norfo R, Tagliafico E, Manfredini R, Ferrari S. Blood. **c-myb supports erythropoiesis through the transactivation of KLF1 and LMO2 expression.** 2010 Nov 25; 116(22):e99-110.

Bianco, Silvestroni I. **Le talassemie. Un problema medico-sociale: ieri e oggi**. Istituto Italiano di Medicina Sociale, Roma, 1998.

Cao A, Galanello R. **β-thalassemia**. 2010. Genetics in Medicine. 12: 61-76.

Chambeyron S, Bickmore WA. **Does looping and clustering in the nucleus regulate gene expression?** Curr Opin Cell Biol. 2004 Jun; 16(3):256-62.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. **Mapping complex disease traits with global gene expression**. Nat. Rev. Genet. 10, 184 (2009).

Curtin P, Pirastu M, Kan YW, Gobert-Jones JA, Stephens AD, Lehmann H. **A distant gene deletion affects β-globin gene function in an atypical γδβ-thalassemia.** J. Clinic. Invest, 76, 1554, 1985.

Danjou F, Anni F, Perseu L, Satta S, Dessì C, Lai ME, Fortina P, Devoto M, Galanello R. **Genetic modifiers of β-thalassemia and clinical severity as assessed by age at first transfusion**. Haematologica. 2012. 97: 989-93.

de Laat W, Grosveld F. **Spatial organization of gene expression: the active chromatin hub**. Chromosome Res. 2003; 11(5):447-59.

de Wit E, de Laat W. **A decade of 3C technologies: insights into nuclear organization.** Genes Dev. 2012 Jan 1; 26(1):11-24.

Dekker J, Rippe K, Dekker M, Kleckner N. **Capturing chromosome conformation.** Science. 2002 Feb 15; 295(5558):1306-11.

Dekker J. **The three 'C' s of chromosome conformation capture: controls, controls, controls.** Nat Methods. 2006 Jan; 3(1):17-21.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. **Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.** Genome Res. 2006 Oct; 16(10):1299-309.

Efstratiadis A, Posakony JW, Maniatis T, Lawn RM, O'Connell C, Spritz RA, DeRiel JK, Forget BG, Weissman SM, Slightom JL, Blechl AE, Smithies O, Baralle FE, Shoulders CC, Proudfoot NJ. **The structure and evolution of the human β-globin gene family**. Cell, 21,653-668, 1980.

Emambokus N, Vegiopoulos A, Harman B, Jenkinson E, Anderson G, Frampton J. **Progression through key stages of haemopoiesis is dependent on distinct threshold levels of c-Myb.** EMBO J. 2003 Sep 1; 22(17):4478-88.

Fattoum S. **Evolution of hemoglobinopathy prevention in Africa: results, problems and prospect**. Mediterr J Hematol Infect Dis. 2009 Nov 10;1(1).

Flint J, Harding RM, Boyce AJ, Clegg JB. **The population genetics of the hemoglobinopathies**. Bailliere's Clinical Hematology. 1998; 11: 1-50.

Fraser P, Bickmore W. **Nuclear organization of the genome and the potential for gene regulation.** Nature. 2007 May 24; 447(7143):413-7.

Fullwood MJ, Ruan Y. **ChIP-based methods for the identification of long-range chromatin interactions.** J Cell Biochem. 2009 May 1; 107(1):30-9.

Galanello R, Cao A. **Relationship between genotype and phenotype. Thalassemia intermedia**. 1998; Ann NY Acad Sci. 850: 325-333.

Galanello R, Sanna S, Perseu L, Sollaino MC, Satta S, Lai ME, Barella S, Uda M, Usala G, Abecasis GR, Cao A. **Amelioration of Sardinian beta0 thalassemia by genetic modifiers.** 2009. Blood. 114: 3935-7.

Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G.. **Fine-mapping at three loci known to affect fetal haemoglobin levels explains additional genetic variation**. Nat Genet. 2010; 42: 1049-51.

Gaszner M, Felsenfeld G. **Insulators: exploiting transcriptional and epigenetic mechanisms.** Nat Rev Genet. 2006 Sep; 7(9):703-13.

Gaziev J, Lucarelli G. **Stem cell transplantation for hemoglobinopathies**. Curr Opin Pediatr. 2003; 15: 24-31.

Gonda TJ, Metcalf D. **Expression of myb, myc and fos proto-oncogenes during the differentiation of a murine myeloid leukaemia.** Nature. 1984 Jul 19-25; 310(5974):249-51.

He Y, Lin W, Luo J. **Influences of genetic variation on fetal hemoglobin** .Pediatr Hematol Oncol 2011; 28(8):708–717.

Holwerda S, de Laat W. **Chromatin loops, gene positioning, and gene expression.** Front Genet. 2012 Oct 17; 3:217

Jiang J, Best S, Menzel S, Silver N, Lai MI, Surdulescu GL, Spector TD, Thein SL. **cMYB is involved in the regulation of fetal hemoglobin production in adults**. Blood. 2006 Aug 1; 108(3):1077-83.

Kim Y, Kim S, Geun Kim C, Kim A. **The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal γ-globin genes. Nucleic.** Acids Res. 2011 Sep 1;39(16):6944-55.

Lettre G, Sankaran VG, Bezerra MA, Araújo AS, Uda M, Sanna S, Cao A, Schlessinger D, Costa FF, Hirschhorn JN, Orkin SH. **DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease.** Proc Natl Acad Sci USA. 2008; 105: 11869–11874.

Lieberman-Aiden─Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny Eric S. Lander Job Dekker. **Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.** Science 9 October 2009: Vol. 326 no. 5950 pp. 289-293.

Liebhaber SA, Cash FE, Ballas SK. **Human α-globin gene expression the dominant role of the a 2-locus in mRNA and protein synthesis.** Journal of Biological Chemistry; (1986); 261:15327-15333.

Li Q, Zhang M, Duan Z, Stamatoyannopoulos G**. Structural analysis and mapping of DNAse I hypersensitivity of HS5 of the β-globin locus control region**. Genomics. 1999; 61:183-193.

Liu P, Keller JR, Ortiz M, Tessarollo L, Rachel RA, Nakamura T, Jenkins NA, Copeland NG. **Bcl11a is essential for normal lymphoid development**. Nat Immunol 2003; 4: 525–532.

Loudianos G, Cao A, Ristaldi MS, Pirastu M, Tzeti M, Kannavakis E, Kattamis C. **Molecular basis of δβ-thalassemia with normal fetal hemoglobin level**. Blood. 1990; 75: 526-8.

Marinić M, Aktas T, Ruf S, Spitz F. **An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape.** Dev Cell. 2013 Mar 11; 24(5):530-42.

Marks J, Shaw JP, Shen CK. **Sequence organization and genomic complexity of primate theta 1 globin gene, a novel alpha-globin-like gene**. Nature. 1986; 321: 785-788.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. **Systematic localization of common disease-associated variation in regulatory DNA**. Science.2012 Sep 7; 337(6099):1190-5.

May C, Rivella S, Callegari J, Heller G, Gaensler KM, Luzzatto L, et al. **Therapeutic haemoglobin synthesis in β-thalassaemic mice expressing lentivirus-encoded human β-globin.** Nature 2000 Jul 6; 406:82-6.

Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, Foglio M, Zelenika D, Boland A, Rooks H, Best S, Spector TD, Farrall M, Lathrop M, Thein SL. **A QTL influencing F cell production maps to a gene encoding a zincfinger protein on chromosome 2p15**. Nat Genet. 2007; 39: 1197-9.

Mukai HY, Motohashi H, Ohneda O, Suzuki N, Nagano M, Yamamoto M..**Transgene insertion in proximity to the c-myb gene disrupts erythroid-megakaryocytic lineage bifurcation.** Mol Cell Biol. 2006 Nov;26(21):7953-65.

Nuinoon M, Makarasara W, Mushiroda T, Setianingsih I, Wahidiyat PA, Sripichai O, Kumasaka N, Takahashi A, Svasti S, Munkongdee T**.** Mahasirimongkol S,Peerapittayamongkol C, Viprakasit V, Kamatani N, Winichagoon P, Kubo M, Nakamura Y, Fucharoen S. **A genome-wide association identified the common genetic variants influence disease severity in β^0-thalassemia/hemoglobin E**. Hum Genet 2010, 127(3):303–314.

Oh IH, Reddy EP. **The myb gene family in cell growth, differentiation and apoptosis**. Oncogene.1999; 18(19):3017-3033.

Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W, Fraser P. **Active genes dynamically colocalize to shared sites of ongoing transcription**. Nat Genet. 2004 Oct; 36(10):1065-71.

Osborne CS, Chakalova L, Mitchell JA, Horton A, Wood AL, Bolland DJ, Corcoran AE, Fraser P. **Myc dynamically and preferentially relocates to a transcription factory occupied by Igh**. PLoS Biol. 2007 Aug; 5(8):e192.

Pace BS, Zein S. **Understanding mechanisms of gamma-globin gene regulation to develop strategies for pharmacological fetal hemoglobin induction**. Dev Dyn. 2006. 235: 1727-37.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. **In vivo enhancer analysis of human conserved non-coding sequences.** Nature. 2006 Nov 23; 444(7118):499-502.

Persons DA**. Hematopoietic stem cell gene transfer for the treatment of hemoglobin disorders**. 2009. Hematology Am Soc Hematol Educ Program. 690-697. 42.

Phillips JE, Corces VG. **CTCF: master weaver of the genome**. Cell. 2009 Jun 26; 137(7):1194-211.

Pilia G, Chen WM, Scuteri A, Orrú M, Albai G, Dei M, Lai S, Usala G, Lai M, Loi P, Mameli C, Vacca L, Deiana M, Olla N, Masala M, Cao A, Najjar SS, Terracciano A, Nedorezov T, Sharov A, Zonderman AB, Abecasis GR, Costa P, Lakatta E, Schlessinger D. **Heritability of cardiovascular and personality traits in 6,148 Sardinians.** PLoS Genet. 2006 Aug 25; 2(8):e132.

Qiliang Li, Peterson KR, Xiangdong Fang, Stamatoyannopoulos G. **Locus control regions**. Blood, 100 (9), 3077-3086, 2002.

Ranney HM, Sharma V. **Struttura e funzione dell'emoglobina**. In: Williams WJ, Beutler E, Erslev AJ, Lichtman MA. Ematologia. Mc Graw Hill Libri Italia srl, 1991.

Rastegar M, Kobrossy L, Kovacs EN, Rambaldi I, Featherstone M. **Sequential histone modifications at Hoxd4 regulatory regions distinguish anterior from posterior embryonic compartments.** Mol Cell Biol. 2004 Sep; 24(18):8090-103.

Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T. **Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb.** Development**.** 2005 Feb; 132(4):797-803.

Sankaran VG, Menne TF, Xu J, Akie TE, Lettre G, Van Handel B, Mikkola HK, Hirschhorn JN, Cantor AB, Orkin SH. **Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A.** Science. 2008 Dec 19;322(5909):1839-42.

Sankaran VG, Xu J, Ragoczy T, Ippolito GC, Walkley CR, Maika SD, Fujiwara Y, Ito M, Groudine M, Bender MA, Tucker PW, Orkin SH. **Developmental and species-divergent globin switching are driven by BCL11A**. 2009. Nature. 460: 1093-7.

Sankaran VG, Menne TF, Šćepanović D, Vergilio JA, Ji P, Kim J, Thiru P, Orkin SH, Lander ES, Lodish HF. **MicroRNA-15a and -16-1 act via MYB to elevate fetal hemoglobin expression in human trisomy 13.** Proc Natl Acad Sci U S A. 2011 Jan 25;108(4):1519-24.

Sankaran VG, Orkin SH. **The switch from fetal to adult hemoglobin**. Cold Spring Harb Perspect Med. 2013 Jan 1; 3(1):a011643.

Satterwhite E, Sonoki T, Willis TG, Harder L, Nowak R, Arriola EL, Liu H, Price HP, Gesk S, Steinemann D, Schlegelberger B, Oscier DG, Siebert R, Tucker PW, Dyer MJ.. **The BCL11A gene family: involvement of BCL11A in lymphoid malignancies**. Blood. 2001; 98: 3413–3420.

Sedgewick AE, Timofeev N, Sebastiani P, So JC, Ma ES, Chan LC, Fucharoen G, Fucharoen S, Barbosa CG, Vardarajan BN, Farrer LA, Baldwin CT, Steinberg MH, Chui DH. **BCL11A is a major HbF quantitative trait locus in three different populations with β hemoglobinopathies.** Blood Cells Mol Dis. 2008 Nov-Dec; 41(3):255-8.

Sherva R, Sripichai O, Abel K, Ma Q, Whitacre J, Angkachatchai V, Makarasara W, Winichagoon P, Svasti S, Fucharoen S, Braun A, Farrer LA. **Genetic modifiers of Hb E/beta0 thalassemia identified by a two-stage genome-wide association study.** BMC Med Genet. 2010 Mar 30; 11:51.

Skok JA, Gisler R, Novatchkova M, Farmer D, de Laat W, Busslinger M. **Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes.** Nat Immunol. 2007 Apr; 8(4):378-87.

Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA. **Interchromosomal associations between alternatively expressed loci**. Nature. 2005 Jun 2; 435(7042):637-45.

Spitz F, Gonzalez F, Duboule D. **A global control region defines a chromosomal regulatory landscape containing the HoxD cluster.** Cell. 2003 May 2; 113(3):405-17.

Stadhouders R, Aktuna S, Thongjuea S, Aghajanirefah A, Pourfarzad F, van Ijcken W, Lenhard B, Rooks H, Best S, Menzel S, Grosveld F, Thein SL, Soler E. **HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers.** J Clin Invest. 2014 Apr 1; 124(4):1699-710.

Stadhouders R, Thongjuea S, Andrieu-Soler C, Palstra RJ, Bryne JC, van den Heuvel A, Stevens M, de Boer E, Kockx C, van der Sloot A, van den Hout M, van Ijcken W, Eick D, Lenhard B, Grosveld F, Soler E. **Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development.** EMBO J. 2012 Feb 15; 31(4):986-99.

Stamatoyannopoulos G. **Control of globin gene expression during development and erythroid differentiation.** Exp Hematol. 2005 Mar; 33(3):259-71.

Suzuki M, Yamazaki H, Mukai HY, Motohashi H, Shi L, Tanabe O, Engel JD, Yamamoto M. **Disruption of the Hbs1l-Myb locus causes hereditary persistence of fetal hemoglobin in a mouse model**. Mol Cell Biol. 2013; 33(8):1687–1695.

Tasiopoulou M, Boussiou M, Sinopoulou K, Moraitis G, Loutradi- Anagnostou A, Karababa P. **G gamma-196 C-->T, A gamma-201 C-->T: two novel mutations in the promoter region of the gamma-globin genes associated with nondeletional hereditary persistence of fetal hemoglobin in Greece.** 2008. Blood Cells Mol Dis. 40: 320-2.

Thein SL. **Genetic modifiers of β-thalassemia**. Haematologica. 2005; 90: 649-60.

Thein SL, Menzel S, Peng X, Best S, Jiang J, Close J, Silver N, Gerovasilli A, Ping C, Yamaguchi M, Wahlberg K, Ulug P, Spector TD, Garner C, Matsuda F, Farrall M, Lathrop M. **Intergenic variants of HBS1L-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults.** Proc Natl Acad Sci USA. 2007 Jul 3; 104(27):11346-51.

Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. **Looping and interaction between hypersensitive sites in the active β-globin locus.** Mol Cell. 2002 Dec; 10(6):1453-65.

Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, Usala G, Busonero F, Maschio A, Albai G, Piras MG, Sestu N, Lai S, Dei M, Mulas A, Crisponi L, Naitza S, Asunis I, Deiana M, Nagaraja R, Perseu L, Satta S, Cipollina MD, Sollaino C, Moi P, Hirschhorn JN, Orkin SH, Abecasis GR, Schlessinger D, Cao A. **Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia**. Proc Natl Acad Sci USA. 2008; 105: 1620-5. 16.

Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA. Proximity **among distant regulatory elements at the β-globin locus requires GATA-1 and FOG-1**. Molecular Cell, 17: 453-462. 2005.

Vegiopoulos A, Garcia P, Emambokus N, Frampton J. **Coordination of erythropoiesis by the transcription factor c-Myb.** Blood. 2006; 107(12):4703-4710.

Vegiopoulos A, Garcia P, Emambokus N, Frampton J. **Coordination of erythropoiesis by the transcription factor c-Myb.** Blood. 2006;107(12): 4703-4710.

Wahlberg K, Jiang J, Rooks H, Jawaid K, Matsuda F, Yamaguchi M, Lathrop M, Thein SL, Best S.**The HBS1L-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells.** Blood. 2009 Aug 6; 114(6):1254-62.

Wallrapp C, Verrier SB, Zhouravleva G, Philippe H, Philippe M, Gress TM, Jean-Jean O. **The product of the mammalian orthologue of the Saccharomyces cerevisiae HBS1 gene is phylogenetically related to eukaryotic release factor 3 (eRF3) but does not carry eRF3-like activity.** FEBS Lett. 1998; 440(3):387-392.

Wheatherall DJ, Clegg JB. **The thalassemia syndromes**. 4th ed. Oxford, England 2001; Blackwell Science Ltd.

Weatherall DJ, Clegg JB, Higgs DR, Wood WG. **The hemoglobinopathies**. In: Scriver CR, Beaudet Al, Sly WS, Valle D, Vogelstein B, editors. The metabolic and molecular bases of inherited disease (OMMBID). 2002. Chapter 101. New York, NY: McGraw-Hill.

Wilber A, Hargrove PW, Kim YS, Riberdy JM, Sankaran VG, Papanikolaou E, Georgomanoli M, Anagnou NP, Orkin SH, Nienhuis AW, Persons DA**. Therapeutic levels of fetal hemoglobin in erythroid progeny of β-thalassemic CD34+ cells after lentiviral vector-mediated gene transfer**. Blood. 2011; 117: 2817-2826.

Wijgerde M, Grosveld F, Fraser P. **Transcription complex stability and chromatin dynamics in vivo**. Nature. 1995; 377:209-213.

Woon Kim Y, Kim S, Geun Kim C, Kim A. **The distinctive roles of erythroid specific activator GATA-1 and NF-E2 in transcription of the human fetal γ-globin genes.** Nucleic Acids Res. 2011 Sep 1; 39(16):6944-55.

Xu J, Peng C, Sankaran VG, Shao Z, Esrick EB, Chong BG, Ippolito GC, Fujiwara Y, Ebert BL, Tucker PW, Orkin SH. **Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing.** 2011; Science. 334: 993-6.

Xu J, Sankaran VG, Ni M, Menne TF, Puram RV, Kim W, Orkin SH. **Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6**. Genes Dev. 2010; 24: 783-798.

Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R. **Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.** Nat Genet. 2006 Nov; 38(11):1341-7.

E’ un piacere per me poter ringraziare tutte le persone che hanno partecipato al raggiungimento di questo importante traguardo.

Primo tra tutti il mio tutor e relatore Prof. Paolo Moi per avermi dato la possibilità di intraprendere questo percorso e crescere non solo da un punto di vista scientifico e lavorativo; la Dottoressa Maria Giuseppina Marini, per la costante disponibilità, la pazienza, la fiducia, l’aiuto fornito in questi tre anni e per essere un costante punto di riferimento; la Dottoressa Manuela Uda, per il fondamentale e valido supporto, i preziosi suggerimenti e tutto il tempo che mi ha dedicato durante il periodo di stesura della Tesi; la Dottoressa Maria Serafina Ristaldi per la disponibilità, i consigli sia a livello professionale che umano e per avermi dato la possibilità di intraprendere il periodo di ricerca a Cambridge.

I’d like to say a big thank you to all the people in the Nuclear Dynamics laboratory who have been much more than just colleagues: Stefan, Mayra, Biola, Louise, Andrew, Joerg and Sara. Thank you for the patience, the useful discussions and the big support you gave to me every day with my project. Thank you for all the “children” jokes, for having listened “radio Italia”, for having introduced me to “George”, for all the “going out” and for having taught me more than 20 words that mean walking slowly. Thanks to my Italian labmate Sara, you are a great friend. Finally thank you to my “Boss”, Peter Fraser, for having accepted me in his research group, for having believed in me and for having allowed me to learn so much and to make the experiments for my thesis. You all have made the “Cambridge experience” amazing and unforgettable.

Ancora il mio pensiero è dedicato a TUTTA la mia famiglia, mio marito, i miei genitori, mia nonna e tutte le persone, amici, colleghi e parenti che mi hanno accompagnato e “sopportato” durante questo percorso e mi sono stati vicini soprattutto nei momenti più duri ed impegnativi, è anche grazie a voi che ho potuto raggiungere questo traguardo.

Infine vorrei ringraziare me stessa per tutti i sacrifici e non aver mai mollato. ☺