UNIVERSITA' DEGLI STUDI DI PARMA

Dottorato di Ricerca in Scienze della Prevenzione

XXVII ciclo

Studio dell'espressione dei retrotrasposoni umani *Alu* e LINE-1 e della sua possibile modulazione in risposta a fattori ambientali

Anastasia Conti

Dipartimento di Medicina Clinica e Sperimentale

Coordinatore:
Prof. Antonio Mutti

Tutor:
Prof. Giorgio Dieci

2012-2014

Indice

Riassunto **1**

Introduzione **5**

1. Gli elementi trasponibili: alla scoperta dell' *"hidden genome"* 7

I retrotrasposoni umani LINE-1, *Alu* e SVA: gli unici attivi nel genoma 8

Gli elementi LINE-1 9

Gli elementi *Alu* 12

Gli elementi SVA 16

2. Ruoli degli elementi trasponibili nell'evoluzione dei genomi 17

Classificazione evolutiva delle famiglie di elementi LINE-1, *Alu* e SVA 18

3. Elementi trasponibili come fonte di variabilità genetica 20

4. La trascrizione pervasiva: l'*"hidden genome"* è trascritto e attivo 21

I non-coding RNA: classificazione e ruoli emergenti 23

5. *Cross- talk* tra elementi ripetuti e ncRNA 31

6. Meccanismi epigenetici e controllo dell'espressione dei TEs 32

Metilazione e demetilazione del DNA 32

Modificazioni istoniche come marcatori della regolazione trascrizionale 36

7. Influenza dell'ambiente sui meccanismi epigenetici 39

I metalli 40

L'arsenico 40

Il piombo 43

Il mercurio 44

Il cadmio 45

Gli inquinanti atmosferici 47

Il benzene: metabolismo, tossicità e carcinogenesi 49

Alterazioni genetiche ed epigenetiche associate all'esposizione al benzene 51

Scopo della ricerca **55**

Parte I: Identificazione di elementi *Alu* trascritti dalla RNA polimerasi III mediante analisi di dati di RNA-seq **57**

Risultati e discussione 60

Conclusioni 85

Materiali e metodi 87

Appendice I 95

Parte II: Analisi delle alterazioni del trascrittoma cellulare causate dall'espressione di *Alu* RNA **99**

Risultati e discussione 102

Conclusioni 132

Materiali e metodi 134

Appendice II 141

Parte III: Analisi dell'espressione di elementi *Alu* e LINE-1 in risposta a trattamenti con idrochinone **145**

Risultati e discussione 148

Conclusioni 163

Materiali e metodi 167

Appendice III 173

Progetti collaterali: Identificazione dei geni trascritti dalla RNA polimerasi III negli eucarioti **175**

Risultati e discussione 178

Conclusioni 181

Materiali e metodi 183

Conclusioni **187**

Bibliografia **189**

Riassunto

Gli elementi trasponibili costituiscono circa il 45% del genoma umano e solo la classe dei retrotrasposoni non-LTR risulta attiva oggigiorno dal punto di vista della retrotrasposizione. Fanno parte di questa classe, gli elementi LINE-1 ed *Alu*, i quali si inseriscono in un locus genomico diverso da quello di origine, duplicandosi attraverso la generazione di un intermedio ad RNA, aumentando così notevolmente il loro numero di copie genomiche (circa 1 milione per gli elementi *Alu* e 500 mila per gli elementi L1). Inizialmente considerati DNA "spazzatura", ovvero una porzione di DNA priva di qualsiasi funzione, attualmente essi sono stati progressivamente rivalutati, innanzitutto in quanto in grado di contribuire alla stabilità e all'evoluzione dei genomi in molteplici modi, in secondo luogo anche in quanto in grado di influenzare numerosi processi cellulari in modi ancora da esplorare. Il fatto che tali elementi si integrino in nuovi siti genomici, fa sì che essi possano generare instabilità sia mediante mutagenesi inserzionale sia attraverso eventi di ricombinazione omologa non allelica. Inoltre gli RNA stessi, prodotti dalla trascrizione dei retrotrasposoni, possono anche interferire con i normali processi cellulari. Analizzare gli elementi ripetuti, in quanto parte del genoma non ancora ben esplorata ma con ruoli emergenti sempre più associati allo sviluppo delle malattie, è quindi di rilievo per capire in dettaglio come sia regolata la loro espressione e le conseguenze di quest'ultima nel contesto cellulare. Inoltre, dal momento che vi sono alcune evidenze di una derepressione trascrizionale degli elementi *Alu* correlata alla loro ipometilazione e che tale ipometilazione è presente nelle cellule tumorali, si può supporre che uno dei meccanismi alla base della attivazione sia proprio una variazione nel loro stato di metilazione (normalmente gli elementi *Alu* e L1 sono ipermetilati e silenti) e che la loro espressione sia da considerarsi come un importante, ma ancora inesplorato, marcatore dello stato tumorale (sono riportate alcune evidenze anche della ipometilazione di elementi L1 in cellule tumorali).

Ciò risulta di particolare rilievo se si considera che l'esposizione prolungata a differenti fattori ambientali, noti per essere legati alla comparsa di patologie, comporta anche una alterata metilazione (in genere ipometilazione) di elementi *Alu* e L1: questo avvalora l'ipotesi di una loro attivazione mediante abrogazione del silenziamento genico, e di un loro conseguente possibile coinvolgimento nella patogenesi. La derepressione di retrotrasposoni *Alu* e L1 in risposta a fattori ambientali rappresenta, pertanto, un fenomeno degno della massima attenzione.

Nell'ambito della presente tesi di dottorato si è elaborata una strategia informatica per l'identificazione di elementi *Alu* trascritti autonomamente dalla RNA polimerasi III a partire da dati grezzi di RNA-seq. Si sono analizzati i dati relativi ai trascrittomi di 7 linee cellulari (tumorali e non

tumorali) ottenuti all'interno del progetto ENCODE e si è individuato un sottogruppo di *Alu* RNA tipo cellulare-specifico che ha permesso di ipotizzare che gli elementi espressi possono avere una particolare sensibilità alla metilazione del DNA e al contesto cromatinico, i quali sono soggetti a variazioni da un tipo cellulare all'altro. Pertanto, l'analisi dei profili di espressione di elementi *Alu*, resa possibile dal sequenziamento dei trascrittomi cellulari, potrebbe rappresentare un nuovo modo, estremamente fine e sensibile, di monitorare le alterazione dell'epigenoma che accompagnano gli stati fisiologici e patologici, nonchè le esposizioni ambientali. Infatti, nonostante la notevole influenza degli elementi *Alu* sull'evoluzione del genoma e gli sporadici esempi del loro coinvolgimento in alcune forme tumorali, l'attività trascrizionale dei loci *Alu* nelle cellule normali e nelle cellule tumorali non è mai stata indagata a fondo. Detto ciò, l'identificazione di profili di *Alu* RNA potrebbe essere utile per considerarli come bio-marcatori del cancro, pertanto le nostre evidenze hanno aperto la strada ad indagini future sul possibile coinvolgimento degli *Alu* RNA nei tumori, come riportato per altri RNA non codificanti (come miRNA e lncRNA).

La sovraespressione di due *Alu* RNA in cellule HeLa ha inoltre permesso di mettere in luce un possibile meccanismo di azione degli *Alu* RNA espressi nelle cellule, in quanto si è osservata la down-regolazione di mRNA contenenti almeno 3 elementi *Alu* in antisenso nella loro 3'UTR e di lncRNA con almeno 2 elementi *Alu* senso nella forma matura (post-splicing). Tali evidenze suggeriscono fortemente un meccanismo senso-antisenso (gli *Alu* RNA sovraespressi sono in orientamento senso) alla base del decremento di espressione dei trascritti codificanti per proteine. Inoltre, fra i geni per proteine deregolati dalla sovraespressione degli elementi *Alu* si sono trovati gruppi di geni coinvolti nel ciclo cellulare e nella maturazione dell'RNA, e ciò risulta importante in quanto uno dei primi esempi in tale ambito. Infine, si è rivolta l'attenzione all'influenza che i fattori ambientali, ai quali l'uomo è quotidianamente esposto, possono esercitare sull'espressione degli elementi ripetuti *Alu* e LINE-1, incuriositi dal fatto che l'esposizione ad alcuni agenti è stata associata ad una ipometilazione di tali elementi normalmente silenti, che quindi potrebbe essere interpretata come potenziale causa di una loro derepressione. Trattamenti di cellule THP-1 e staminali ematopoietiche da cordone ombelicale con concentrazioni crescenti di idrochinone, metabolita del benzene scelto come paradigma dell'esposizione ambientale, hanno permesso di osservare un apprezzabile aumento di espressione di *Alu* e L1 RNA che potrebbe essere correlato con una loro ipometilazione genomica. Tali variazioni di espressione e metilazione si sono rivelati tipo cellulare-indipendenti, facendo supporre che il derivato del benzene eserciti effetti paragonabili in differenti contesti cellulari e che quindi, in ultima analisi, la derepressione degli elementi *Alu* e L1 possa essere un suo effetto peculiare. Lo studio condotto con l'idrochinone apre

la strada all'analisi degli effetti che altri fattori ambientali, generalmente ipometilanti, possono avere sull'espressione di elementi *Alu* e L1, e aggiungono un tassello verso la conoscenza di un possibile meccanismo alla base dell'insorgenza delle patologie indotte dai tali fattori.

Introduzione

1. Gli elementi trasponibili: alla scoperta dell' "*hidden genome*"

Gli elementi trasponibili (TEs, Transposable Elements) sono sequenze di DNA in grado di "muoversi" da un sito all'altro del genoma, creando o revertendo mutazioni ed alterando così il genotipo cellulare. Tali elementi furono scoperti per la prima volta da Barbara McClintock in *Zea mays* negli anni '40[1,2], in un'epoca in cui dominava l'idea di stabilità genetica necessaria per la trasmissione ai discendenti e per il mantenimento delle caratteristiche della specie. Cosa sorprendente è che da allora i TEs sono stati identificati in una notevole varietà di organismi eucariotici, e si è scoperto che costituiscono circa il 45% del genoma dei mammiferi[3,4] e fino al 90% del genoma di alcune specie di piante[5]. I TEs si distinguono dagli altri elementi genomici in quanto, essendo mobili, sono in grado sia di cambiare il loro intorno genomico sia quello del locus entro il quale si integrano. Inoltre, dal momento che hanno la capacità di replicarsi durante il processo di trasposizione, essi sono inevitabilmente presenti in un notevole numero di copie nel genoma e si possono considerare sia parte del genoma che entità indipendenti che vivono al suo interno. Tali elementi, a lungo considerati DNA "spazzatura", ovvero parti di quello che si potrebbe definire "hidden genome" ritenuto privo di funzione, sono stati, e sono tuttora, oggetto di approfonditi studi volti a definirne le caratteristiche strutturali e funzionali e a valutare l'impatto che la loro presenza ha sull'intero genoma.

Prendendo in considerazione il genoma umano, costituito per circa il 45% da elementi trasponibili (**Figura 1**), innanzitutto è necessario ricordare che essi possono essere suddivisi in due categorie in base alla loro organizzazione e al meccanismo di trasposizione.

I trasposoni a DNA costituiscono circa il 3% del genoma umano e si traspongono come frammenti di DNA che sono excisi e inseriti in nuovi siti genomici. Sebbene essi non siano attualmente in grado di trasporsi, sono stati attivi agli inizi dell'evoluzione dei Primati, fino a 37 millioni di anni fa[6].

I retrotrasposoni, al contrario, si traspongono in un nuovo locus genomico duplicandosi attraverso la generazione di un intermedio a RNA[6]. Essi possono dunque incrementare il proprio numero più rapidamente dei trasposoni a DNA (molti dei quali non si replicano durante il processo di trasposizione) e questo ha consentito loro di espandersi maggiormente all'interno del genoma.

I retrotrasposoni, a loro volta, sono suddivisi in due gruppi in base alla presenza o assenza di lunghe ripetizioni terminali (LTRs, Long Terminal Repeats): i retrotrasposoni LTR e i retrotrasposoni non-LTR[7].

Nell'uomo i primi sono definiti HERVs (Human Endogenous Retrovirus), ovvero elementi di tipo retrovirale (o "*virus-like")*, e costituiscono circa l'8% del genoma. Essi si sono inseriti nel DNA più di

25 milioni di anni fa e oggi la loro attività è molto limitata, se non assente[3,8]. Le loro lunghe ripetizioni terminali, sequenze regolative importanti per la mobilità, fiancheggiano una regione centrale codificante per tutte le proteine generalmente prodotte dai genomi retrovirali, fatta eccezione per quelle costituenti il capside virale.

I retrotrasposoni non-LTR comprendono le classi degli elementi LINEs (Long Interspersed Elements) e SINEs (Short Interspersed Elements), nonché gli elementi SVA (SINE/VNTR/Alu). Essi rappresentano complessivamente un terzo del genoma umano[3] e sono i soli elementi trasponibili attualmente attivi nell'uomo, come dimostrano più di 60 casi di inserzione *de novo* associate a malattie genetiche[9,10].

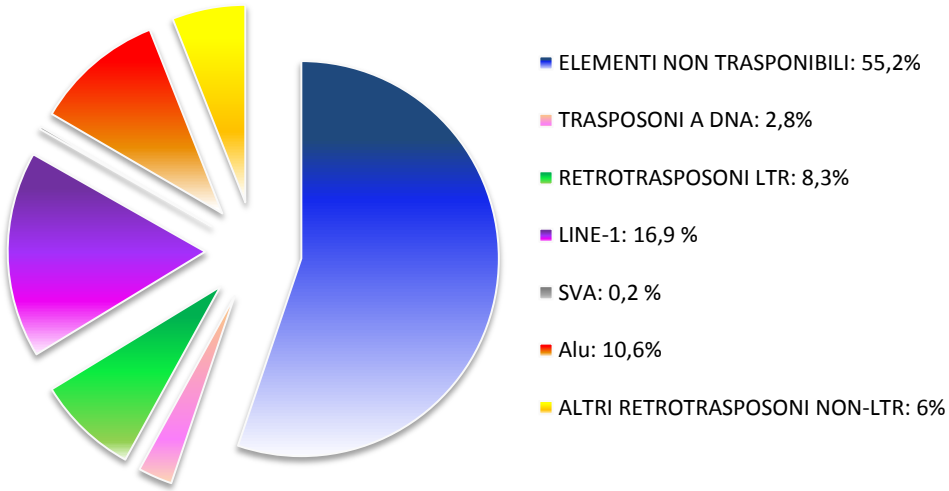


Figura 1. Elementi trasponibili contenuti nel genoma umano. Circa il 45% della sequenza del genoma deriva da elementi trasponibili, la maggioranza dei quali sono elementi L1, *Alu* e SVA.

I retrotrasposoni umani LINE-1, *Alu* e SVA: gli unici attivi nel genoma

Le evidenze di una mobilità degli elementi ripetuti umani sono limitate al sottogruppo dei retrotrasposoni non-LTR, che comprende, come accennato in precedenza, gli elementi LINEs, SINEs e SVA. In particolare, all'interno delle classi LINEs e SINEs, gli elementi attualmente attivi sono appartenenti alle sottofamiglie LINE-1 (o L1) ed *Alu* rispettivamente, che insieme costituiscono circa il 30% del genoma umano. Recenti studi hanno dimostrato che la frequenza di nuove inserzioni di tali elementi è molto maggiore di quanto fosse stato stimato in precedenza e contribuisce in modo rilevante alla diversità genetica[11-13], portando a non considerarli più segmenti di DNA privi di qualsiasi influenza ma importanti per l'evoluzione dei genomi e per conferire loro un carattere dinamico.

Si prenderanno ora in esame tali retroelementi non-LTR, molto abbondanti ed attivi nel genoma umano, e si descriveranno le loro caratteristiche di sequenza, il loro meccanismo di trascrizione e di trasposizione, per far luce su come essi possano "muoversi" nel genoma ed avere quindi un effetto su di esso a vari livelli.

Gli elementi LINE-1

Come accennato in precedenza, gli elementi LINE-1 (o L1) appartengono alla classe delle LINEs (Long Interspersed Elements) a sua volta facente parte della famiglia dei retrotrasposoni non-LTR. Il genoma umano ne contiene più di mezzo milione di copie, derivanti dalla loro retrotrasposizione nel corso dei passati 150 milioni di anni[3], e ciò contribuisce a far sì che essi costituiscano circa il 17% del genoma stesso, e che siano, quindi, una delle classi di elementi ripetuti più abbondanti.

L'elemento L1 canonico completo è lungo circa 6000 paia di basi (bp) ed è composto da una regione non tradotta in 5' (detta 5'UTR, 5'UnTranslated Region) che contiene un promotore interno atipico riconosciuto dalla RNA polimerasi II (Pol II), l'enzima che trascrive tale elemento[14]. Seguono poi due regioni codificanti (Open Reading Frame) per le proteine ORF1p ed ORF2p, una regione non tradotta in 3' (detta 3'UTR) contenente un segnale di poliadenilazione, ed una coda ricca in A di lunghezza variabile[15,16] (**Figura 2**).

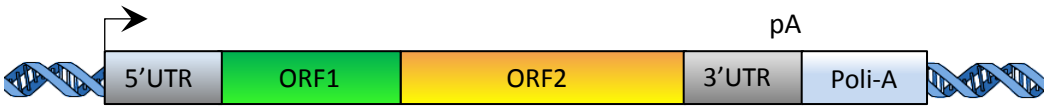


Figura 2. Rappresentazione di un elemento LINE-1 canonico: si distinguono la 5'UTR, le due sequenze codificanti per le proteine ORF1p e ORF2p, la 3'UTR contenente un segnale di poliadenilazione e la lunga coda poli-A.

ORF1 codifica per una proteina di legame all'RNA di circa 40 kDa (chiamata ORF1p) che si pensa agisca come un chaperone di acidi nucleici[17] (funzione simile alla proteina retrovirale Gag), mentre ORF2 codifica per una proteina di circa 150 kDa (chiamata ORF2p) con attività sia di endonucleasi che di trascrittasi inversa[15,18]. Entrambe le proteine sono essenziali per la retrotrasposizione di L1[19] e mostrano una tendenza al legame preferenziale del proprio RNA (preferenza *in cis*) formando con esso una particella ribonucleoproteica (RNP, RiboNucleoprotein Particle) citoplasmatica proposta come intermedio di retrotrasposizione[20-23]. Dal momento che gli elementi L1 codificano per proteine necessarie per la propria amplificazione, essi vengono definiti elementi autonomi; sono, in effetti, gli unici elementi ripetuti autonomi attivi nel genoma umano.

Tuttavia non tutte le copie L1 sono in grado di retrotrasporsi, dal momento che durante la loro amplificazione possono essere inattivate perché subiscono la rimozione dell'estremità 5'[3,24] (L1 tronche), o delezioni e altre mutazioni interne[3,25]. Pertanto tra le più di mezzo milione di copie di L1 inserite nel genoma, solo meno di 100 copie sono intatte e quindi ritenute attive dal punto di vista della retrotrasposizione[26,27].

La trascrizione degli elementi L1 avviene grazie alla presenza di un promotore riconosciuto dall'RNA polimerasi II (Pol II) privo di una sequenza TATA interno alla 5'UTR e si è visto che, affinché si abbia trascrizione, è sufficiente una regione di 100 bp, detta promotore minimo[14]. Inoltre la 5'UTR presenta anche siti di legame per fattori di trascrizione agenti *in cis*[28-33]. Recenti studi hanno dimostrato che la regione 5' terminale di L1 contiene anche un promotore forte antisenso (ASP, AntiSense Promoter) che porta alla sintesi di trascritti chimerici antisenso, corrispondenti ad una porzione della 5'UTR dell'elemento L1 e sequenze genomiche fiancheggianti l'estremità 5' dell'elemento stesso[34,35]. È inoltre stato ipotizzato che questi trascritti chimerici possano avere una funzione nella regolazione genica o promuovere la formazione di trascritti L1 a doppio filamento (dsL1 RNA, double stranded L1 RNA) che regolano la retrotrasposizione degli elementi L1 mediante un meccanismo di RNA interference[36,37]. È stato anche visto che i trascritti chimerici si sono rivelati utili per identificare elementi L1 espressi nelle cellule staminali embrionali umane, nelle linee cellulari embrionali di carcinoma e nei tessuti somatici umani[35,38,39].

Il ciclo di retrotrasposizione degli elementi L1 inizia con la trascrizione del locus genomico dell'elemento integrato da parte della Pol II, e ciò sottolinea l'importanza dell'espressione del retroelemento affinché possa successivamente trasporsi. In seguito il trascritto L1 neosintetizzato (L1 RNA) viene trasportato nel citoplasma e tradotto per produrre le proteine ORF1p ed ORF2p necessarie per la retrotrasposizione dell'elemento stesso[19]. Come detto in precedenza, ORF1p possiede attività di chaperone di acidi nucleici[40,41], una proprietà essenziale per la retrotrasposizione[41,42], e il L1 RNA mostra una forte preferenza *in cis* per le sue proteine tradotte durante la retrotrasposizione stessa[20,23]. Tale preferenza porta a pensare che probabilmente si abbiano interazioni RNA-proteina già vicino al ribosoma durante la traduzione e che siano importanti per formare la particella RNP[22]. Ciò spiegherebbe, inoltre, come mai il numero di copie di retropseudogeni nel genoma sia molto inferiore rispetto a quello degli elementi L1, in quanto la preferenza *in cis* può ridurre la possibilità che altri RNA messaggeri (mRNA) possano legarsi ad ORF1p ed ORF2p. Dal momento che più molecole di ORF1p interagiscono con il trascritto L1[42] e sono presenti in un numero maggiore rispetto a ORF2p[43], è ragionevole supporre che la proteina ORF1p sia più abbondante di ORF2p nella particella RNP, e che possa aiutare il "distacco" del

trascritto dal ribosoma e proteggerlo dalla degradazione. Inoltre, il rivestimento del L1 RNA da parte di ORF1p potrebbe impedire nuovi riassemblaggi dei ribosomi e far sì che il trascritto, nel complesso ribonucleoproteico, si diriga verso il nucleo.

Ci sono pochi dati su come, dopo l'assemblaggio del trascritto L1 con le proteine ORF1p e ORF2p, la ribonucleoproteina raggiunga il nucleo per il processo di inserzione, ma si sa che per tale processo gli elementi L1 si avvalgono di un meccanismo chiamato trascrizione inversa innescata dal sito bersaglio (Target-Primed Reverse Transcription, TPRT, **Figura 3**)[18,44,45]. Tale meccanismo è stato inizialmente descritto per l'amplificazione del retrotrasposone non-LTR R2 del baco da seta *Bombyx mori*[46], ed è stato adattato per descrivere il ciclo di retrotrasposizione di altri elementi che usano un meccanismo di inserzione non-LTR.

Dal momento che ORF2p è dotata di attività endonucleasica, l'endonucleasi inizia la reazione di integrazione introducendo un taglio nel DNA cromosomico[18]; siti di taglio preferenziali sono le sequenze 5'-TTTT/AA-3'. La presenza di tali T nel sito di taglio permette al DNA di appaiarsi con la coda poli-A dell'elemento a RNA, la quale, pertanto, gioca un ruolo molto importante nella retrotrasposizione[47,48]. L'analisi dei siti di inserzione ha dimostrato che l'estremità 3'-OH, che si forma nella reazione di incisione del DNA, funge da innesco per la retrotrascrizione dell'elemento a RNA da parte di ORF2p, in quanto perfettamente complementare con la sequenza di RNA stampo. Studi *in vitro* hanno dimostrato che le sequenze bersaglio non sono determinate solo dalla preferenza di taglio dell'endonucleasi ORF2p[49], ma riflettono anche la capacità dell'RNA di interagire con il DNA tagliato per promuovere la sua retrotrascrizione.

I passaggi finali della trasposizione, dopo la completa retrotrascrizione da parte di ORF2p del primo filamento di cDNA, non sono ancora stati ben compresi.

Essi includono il taglio del secondo filamento di DNA bersaglio, la sintesi del secondo filamento di cDNA, la riparazione della rottura nel sito di inserzione ed una reazione di ligazione utile a saldare i filamenti di DNA.

Tuttavia elementi caratteristici del processo di integrazione includono frequenti troncature delle estremità 5', la presenza di una coda poli-A all'estremità 3' e corte duplicazioni del sito bersaglio (2-20 bp) fiancheggianti la nuova copia dell'elemento[3,25].

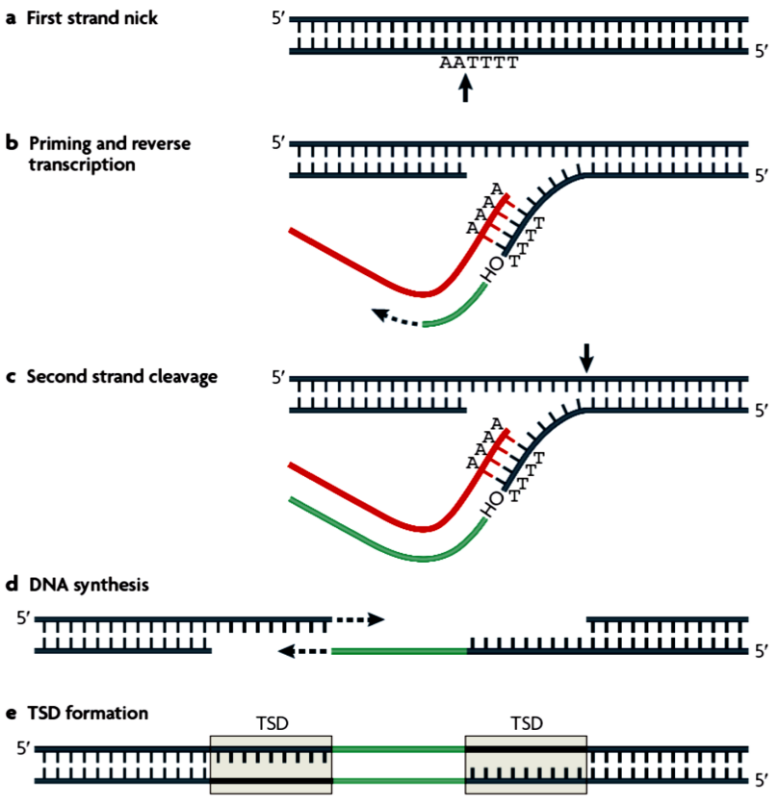


Figura 3. Rappresentazione schematica del meccanismo di Target-Primed Reverse Transcription (Adattato da Cordeaux and Batzer, 2009[50]).

È stato visto che gli elementi L1 si inseriscono preferenzialmente in regioni del genoma ricche in AT, in accordo con il loro meccanismo di retrotrasposizione, e tali regioni sono generalmente prive di geni[3].

Gli elementi *Alu*

Gli elementi *Alu* appartengono alla classe delle SINEs (Short Interspersed Elements) e sono diffusi unicamente nei Primati. Devono il loro nome alla presenza all'interno della propria sequenza di un sito di taglio per l'enzima *Alu* I e sono gli elementi ripetuti più abbondanti nel genoma umano, presenti in oltre 1 milione di copie[3]. Tali elementi contribuiscono all'11% della sequenza genomica grazie alla loro continua trasposizione avvenuta nel corso di circa 65 milioni di anni[51] e, sebbene siano stati inizialmente considerati come DNA "spazzatura", è ora divenuto evidente come possano al contrario influenzare in modo significativo la stabilità e l'espressione del genoma umano.

Il corpo del tipico elemento *Alu* è lungo circa 280 nucleotidi ed è formato da due monomeri in tandem, che derivano dal gene codificante per l'RNA 7SL (l'RNA della particella di riconoscimento del segnale SRP), separati da una corta regione ricca in A. Il monomero sinistro, detto anche braccio sinistro dell'elemento *Alu*, contiene gli elementi promotoriali A box e B box, i quali sono riconosciuti dal macchinario trascrizionale dell'RNA polimerasi III (Pol III), l'enzima che trascrive tali sequenze.

Il monomero destro è seguito da una lunga regione poli-A all'estremità 3' dell'elemento, la quale si è rivelata molto importante nel meccanismo di amplificazione dell'elemento stesso[47]. L'unità trascrizionale dell'elemento *Alu* include inoltre la sequenza genomica che si estende dalla lunga regione poli-A fino al terminatore della trascrizione da parte della Pol III, costituito nella maggior parte dei casi da almeno quattro residui di timina (TTTT) nella regione genomica fiancheggiante[52] (si parla in questo caso di terminatore canonico), nonché talvolta anche da varianti di tale sequenza (si parla di terminatore non canonico[53]).

Molti elementi *Alu* nel genoma umano, trascritti dalla Pol III, sono unici in sequenza in termini di mutazioni accumulate nel corpo dell'*Alu*, lunghezza ed eterogeneità della regione poli-A all'estremità 3', ed alla presenza di una regione di lunghezza e sequenza variabile, definita *3'trailer*, che si estende dalla coda poli-A al terminatore della trascrizione[54] (**Figura 4**).

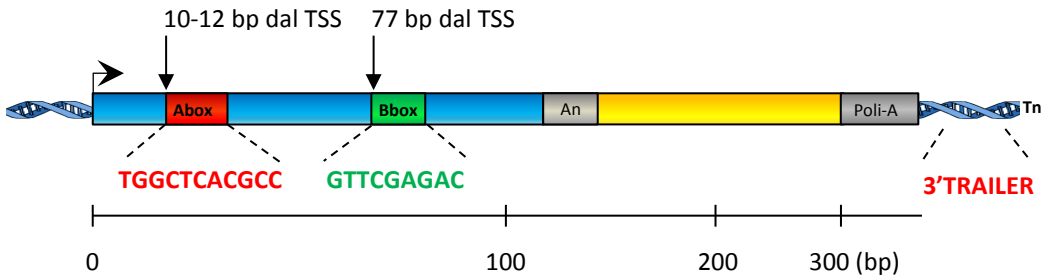


Figura 4. Rappresentazione di un elemento *Alu* canonico: si distinguono il braccio sinistro (blu) nel quale sono presenti gli elementi promotoriali A box e B box (dei quali è riportata la sequenza *consensus* e la distanza dal sito di inizio della trascrizione, TSS), la corta sequenza poli-A che collega il braccio sinistro al braccio destro (An), il braccio destro (in giallo), la lunga regione poli-A e il *3'trailer* nella sequenza genomica fiancheggiante l'elemento che si estende fino al terminatore della trascrizione di lunghezza variabile (Tn).

La storia evolutiva e la conseguente struttura degli elementi *Alu* è intimamente connessa alle modalità della loro trascrizione e trasposizione.

Come accennato in precedenza, gli elementi *Alu* sono trascritti dall'RNA polimerasi III e presentano due regioni promotrici interne, chiamate A box e B box, tipicamente presenti anche nei promotori dei geni per gli RNA transfer (tRNA)[52,55,56]. È stato dimostrato che la presenza della B box è essenziale per la loro trascrizione, mentre la A box sembra importante per definire il sito di inizio della trascrizione dell'elemento (TSS, Transcriptional Start Site), il quale si trova a circa 10-12 nucleotidi a monte della sequenza stessa[52,55-59]. Questi due elementi permettono la trascrizione *in vitro* dei retrotrasposoni *Alu*[56,60], mentre, al contrario, la loro espressione *in vivo* richiede sequenze genomiche fiancheggianti l'estremità 5' dell'elemento[61-63], le quali, molto probabilmente, contribuiscono ad attivare la trascrizione soprattutto fornendo elementi promotoriali addizionali a monte del TSS[63].

Affinché la Pol III possa trascrivere le sequenze *Alu*, è necessario che il fattore di trascrizione TFIIIC riconosca e si leghi al promotore interno (formato da A box e B box, situate rispettivamente a 10-12 e 77 basi dal TSS, **Figura 4**), e la sua subunità prossimale recluti direttamente un secondo fattore, TFIIIB, il quale si lega circa 50 bp a monte del sito di inizio della trascrizione, e permette il corretto posizionamento della Pol III a livello del TSS stesso[55,64] (**Figura 5**). TFIIIB rimane stabilmente legato al DNA per l'intera durata del processo permettendo un efficiente re-inizio della trascrizione sul medesimo templato[65]. Nonostante il legame di TFIIIB al DNA sia generalmente sequenza-indipendente, sono state riportate evidenze dell'influenza delle regioni fiancheggianti in 5' sulla trascrizione degli elementi *Alu* sia *in vitro* che in cellule trasfettate[61,62].

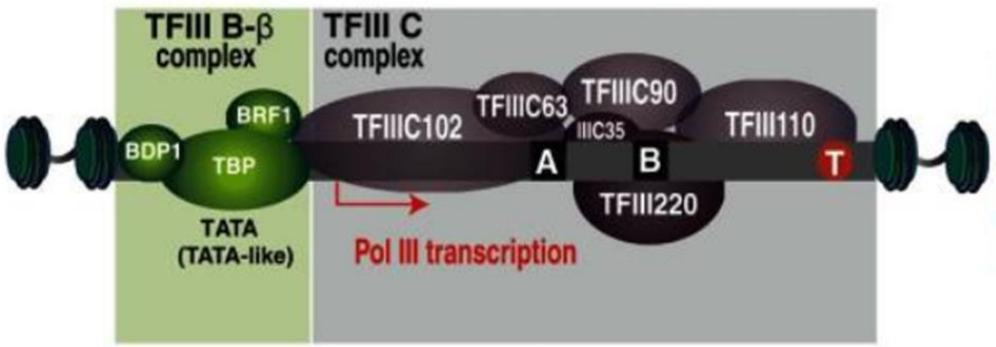


Figura 5. Rappresentazione schematica dei fattori trascrizionali che permettono la sintesi degli elementi *Alu* ad opera dell'enzima RNA polimerasi III (Adattato da Lunyak and Atallah, 2011[66]).

La maggior parte degli elementi *Alu* mancano di un terminatore necessario affinché cessi la sintesi del trascritto da parte della Pol III appena dopo la fine del corpo dell'elemento stesso[58,60]. La trascrizione si conclude, dunque, al primo terminatore canonico (TTTT) o non canonico incontrato nella sequenza genomica fiancheggiante la lunga coda poli-A, generando così trascritti lunghi da 300 a 600 nucleotidi e con sequenze variabili al 3', sia in termini di lunghezza che di composizione nucleotidica[67-73]. Quasi ogni elemento *Alu* risulta, quindi, unico per sequenza, non soltanto a seguito di mutazioni accumulate ma anche per la regione 3'*trailer*, la cui sequenza dipende dal locus genomico in cui l'elemento è localizzato (**Figura 4**).

È ormai noto come gli elementi *Alu*, al contrario di ciò che accade per gli elementi L1, tendano a retrotrasporsi maggiormente in regioni ricche di geni[3], ovvero regioni ricche in CG che possono trovarsi o in prossimità dei geni o in particolari regioni, dette isole CpG (CpG *islands*, CGI), generalmente nel promotore dei geni stessi[3].

In seguito alla scoperta delle SINEs e all'evidenza della loro incapacità di trasporsi in modo autonomo, venne proposto che questi elementi potessero amplificarsi nei genomi per retrotrasposizione[74] sfruttando il macchinario degli elementi L1, grazie alla presenza di sequenze nucleotidiche simili alle estremità 3' di SINEs e LINEs[75,76].

La retrotrasposizione di L1, trattata nel paragrafo precedente, coinvolge le due proteine codificate dall'elemento stesso, ORF1p e ORF2p, pertanto si parla di L1 come di un elemento autonomo dal punto di vista della trasposizione. Al contrario, gli elementi *Alu*, che ricorrono al macchinario di trasposizione di L1, vengono definiti elementi non autonomi e spesso in letteratura si parla di loro come di "parassiti di parassiti", ovvero, secondo questa visione, sia L1 che *Alu* sarebbero parassiti del genoma umano e a loro volta gli elementi *Alu* sarebbero parassiti degli elementi L1.

Evidenze sperimentali hanno dimostrato che la proteina ORF1p non è probabilmente essenziale nel processo di retrotrasposizione degli elementi *Alu* nonostante sembri tuttavia migliorarne l'efficienza[77]. La proteina ORF2p, al contrario, è necessaria e media l'integrazione dell'elemento *Alu* attraverso il meccanismo chiamato trascrizione inversa innescata dal sito bersaglio (Target-Primed Reverse Transcription, TPRT), già descritto per gli elementi L1 (**Figura 3**).

Tale processo inizia con il legame del trascritto *Alu* (*Alu* RNA) a due proteine della SRP (Signal Recognition Particle), che formano l'eterodimero SRP9/14[78], alla proteina PAPB (Poly A Binding Protein) che lega la sequenza poli-A [79,80] e ad almeno un'altra proteina, non ancora identificata, che lega l'RNA. L'eterodimero SRP9/14 e PABP si pensa potrebbero aiutare l'*Alu* RNA ad associarsi al ribosoma, a livello del quale potrebbe interagire con la proteina ORF2p in traduzione a partire da un trascritto L1[48,81,82]. Il complesso formato dall'*Alu* RNA e dalla proteina ORF2p verrebbe

trasportato quindi nel nucleo dove l'*Alu* RNA, tramite la regione ricca in A, può guidare l'attività endonucleasica di ORF2p che introduce, anche in questo caso, due tagli sfalsati a livello di una sequenza bersaglio ricca in T, con la quale è in grado di interagire.

L'estremità 3'OH generata dal taglio funge da innesco, e viene quindi estesa grazie all'attività retrotrascrittasica di ORF2p che utilizza come stampo l' *Alu* RNA. La proteina ORF2p prosegue la sintesi del nuovo filamento utilizzando come stampo il filamento superiore della doppia elica generatosi a causa del taglio sfalsato, e nel contempo l'*Alu* RNA viene degradato e sostituito da un nuovo tratto di DNA. Ciò genera le ripetizioni dirette ai lati della nuova inserzione (lunghe 5-15 bp a seconda della distanza tra la prima e la seconda incisione generata da ORF2p) che affiancano tipicamente sia gli elementi L1 sia gli elementi *Alu*.

Nonostante i retrotrasposoni *Alu* non presentino sempre una coda poli-A alla loro estremità 3', si è visto che la loro retrotrasposizione può avvenire ugualmente, sebbene in modo molto inefficiente[81].

Infine è importante sottolineare che tali elementi non codificano, a differenza degli elementi L1, per alcun prodotto proteico.

Gli elementi SVA

Come riportato in precedenza, oltre agli elementi L1 e *Alu*, nel genoma umano vi sono circa 2700 copie di elementi SVA (SINE/VNTR/Alu)[83], risultanti da una continua attività di retrotrasposizione nel corso di 25 milioni di anni di evoluzione[83,84]. Il tipico elemento SVA completo è lungo circa 2000 bp ed è composto da una regione all'estremità 5' formata dalla sequenza esamerica ripetuta CCCTCT, da una sequenza simile ad un elemento *Alu* (*Alu-like*), da una regione con un numero variabile di ripetizioni in tandem (VNTR, Variable Number of Tandem Repeats), da una sequenza HERV-K10-like (regione simile allo Human Endogenous Retrovirus-K10) e da un segnale di poliadenilazione che termina con una regione ricca in A di lunghezza variabile (**Figura 6**). Vi sono molte evidenze del fatto che tali elementi siano trascritti dalla RNA polimerasi II (come gli elementi L1) nonostante essi non mostrino apparentemente un promotore interno, e si pensa potrebbero avvalersi, almeno in parte, dell'attività di un promotore nelle regioni genomiche fiancheggianti[83,84]. Inoltre si ritiene che tali elementi non codifichino per alcun prodotto proteico (come gli elementi *Alu*) e, nonostante le ridotte conoscenze circa il loro meccanismo di amplificazione, si ipotizza che utilizzino il macchinario di retrotrasposizione degli elementi L1, proprio come gli elementi *Alu*. Per questo motivo, essi sono considerati elementi ripetuti non autonomi appartenenti alla famiglia dei retrotrasposoni non-LTR, ma non si possono considerare membri della famiglia delle SINEs, in

quanto di lunghezza molto maggiore rispetto agli elementi *Alu* e trascritti non dalla Pol III (vi sono, infatti, molti terminatori Pol III all'interno della loro sequenza) ma dalla Pol II.

Figura 6. Rappresentazione di un elemento SVA canonico: si distinguono le regioni $(CCCTCT)_n$, *Alu*-like, VNTR, HERV-K10-like, il segnale di poliadenilazione e la coda poli-A.

2. Ruoli degli elementi trasponibili nell'evoluzione dei genomi

Dal momento che gli elementi trasponibili sono in grado di replicarsi più velocemente del genoma, essi possono essere visti come parassiti che conferiscono svantaggi al genoma che abitano, una sorta di "cancro" del genoma stesso[85]. Per questo motivo, come accennato all'inizio della trattazione, sono stati a lungo considerati DNA "spazzatura", ovvero una porzione del genoma che, benché fosse consistente, non mostrasse alcuna funzione, una sorta di "fardello genetico" per il genoma ospite.

Oggigiorno questa visione semplicistica è stata notevolmente rivista, infatti si è osservato che i TEs e i genomi hanno co-abitato per molto tempo, e quindi che hanno interagito in molti modi differenti, e inoltre è stato riportato che tali elementi potrebbero avere importanti ruoli sia regolativi che strutturali, nonostante siano stati a lungo considerati privi di alcuna funzione apparente. Per tali motivi gli elementi trasponibili non sono più visti come parassiti quando si considera l'evoluzione dei genomi[86], ma come elementi interni al genoma stesso che ne hanno guidato l'evoluzione in molteplici modi.

I retrotrasposoni non-LTR comprendono una estesa porzione del genoma (come sopra), specialmente in piante e mammiferi, e l'effetto del loro aumento è stato tollerato durante l'evoluzione. In letteratura sono riportati molti esempi di quanto gli elementi ripetuti siano utili strumenti per studiare l'evoluzione del genoma e la funzione dei geni[87]. Un'analisi comparativa, per esempio, ha messo in luce che molte sequenze umane di elementi trasponibili sono identiche a quelle degli altri vertebrati e, inoltre, che l'inserzione di TEs può regolare l'espressione genica, aumentare la ricombinazione e il crossover ineguale. Si è infatti osservato che l'inserzione di TEs nel genoma umano ha causato molti cambiamenti nello stesso e, alcuni di tali cambiamenti, sono stati responsabili della sua evoluzione[88].

Lo sviluppo di metodologie molecolari innovative, come i saggi di retrotrasposizione nelle colture cellulari[19,81] e le tecniche computazionali di genomica comparativa, insieme alla possibilità di accedere alle sequenze genomiche di molti Primati (uomo[3], scimpanzé[44] e macaco[89]), hanno aumentato l'interesse sull'attività dei TEs, vista di profondo impatto sull'evoluzione dei genomi. Per esempio, è stato osservato nell'uomo che la ricombinazione tra TEs può causare delezioni genomiche responsabili di molte malattie genetiche[90]. Tuttavia, solo recenti confronti tra il genoma umano e quello degli altri Primati hanno permesso di determinare l'importanza delle delezioni, dovute alla ricombinazione degli elementi trasponibili, nella loro evoluzione [91,92].

Prendendo in considerazione i retrotrasposoni umani non-LTR, si può affermare che il loro impatto sul genoma è in gran parte determinato dal loro elevatissimo numero di copie (per esempio, si osserva la presenza di un elemento *Alu* in media ogni 3 kb di DNA) e dalla loro continua attività nell'arco di più di decine di milioni di anni[3].

I tassi di amplificazione di tali retroelementi non sono uniformi (la massima attività si è registrata intorno a 40 milioni di anni fa per gli elementi *Alu*, a 12-40 milioni di anni fa per gli elementi L1 e a 6 milioni di anni fa per gli elementi SVA) e gli elementi *Alu* sono i più studiati per quanto concerne l'evoluzione, in quanto possono causare diversità genomica in molteplici modi e in quanto facenti parte della famiglia più estesa di elementi ripetuti nel genoma umano. È stato stimato che migliaia di elementi *Alu* si siano integrati nel genoma umano dalla divergenza degli uomini dalle scimmie africane[93-96] e, sebbene molte inserzioni di *Alu* abbiano causato mutazioni dannose, molte altre hanno contribuito invece alla diversità genetica[90]. È inoltre importante sottolineare che le ripetizioni *Alu* hanno anche influenzato l'accumulo di polimorfismi a singolo nucleotide nel genoma[14,96,97], ed è evidente che molte delle inserzioni recenti di elementi *Alu* sono la causa di variazioni genetiche che sono state utili per studiare sia le relazioni tra popolazioni ed evoluzione, sia l'organizzazione delle famiglie di geni. Inoltre è stato suggerito che l'inserzione di un elemento *Alu*, in relazione alla variazione genetica, potrebbe anche essere utile come marcatore genetico specie-specifico[51].

Classificazione evolutiva delle famiglie di elementi LINE-1, *Alu* e SVA

Differenti periodi della storia evolutiva hanno dato origine a diverse famiglie di elementi ripetuti LINE-1, *Alu* e SVA.

La famiglia LINE-1 è comparsa circa 120 milioni di anni fa[98,99] e si sta ancora espandendo attivamente nell'uomo, come dimostrano l'esistenza di elementi L1 altamente polimorfici nelle

popolazioni umane[100-104], e le inserzioni *de novo* di elementi L1 responsabili di malattie genetiche[105]. L'analisi di centinaia di inserzioni L1 specie-specifiche nei genomi di uomo e scimpanzé supporta ulteriormente la recente mobilitazione delle famiglie di retrotrasposoni[44,106,107]. Al contrario dei retrotrasposoni non autonomi *Alu*, in cui differenti famiglie sono in grado di espandersi contemporaneamente[51,108,109], tutte le famiglie L1 nel genoma umano derivano da una singola linea ancestrale dalla quale si sono originate in successione l'una dall'altra[99].

Tali famiglie si possono distinguere per la presenza di alcune sostituzioni diagnostiche che presentano tutti i membri di una data famiglia e, a tal proposito, si pensa che 5 famiglie si siano amplificate nei Primati ominoidi (uomo e scimmie) intorno a 25 milioni di anni fa, e sono state chiamate, in ordine di comparsa, L1PA1, L1PA2, L1PA3, L1PA4 e L1 PA5[3,98,99,110]. Inoltre è stata anche identificata la famiglia evolutivamente più recente L1Hs (Homo sapiens) e si è visto che una delle sue sottofamiglie, L1Ta, è comparsa circa 4 milioni di anni fa ed è la più attiva oggigiorno dal punto di vista della retrotrasposizione nel genoma umano. Tutte le famiglie L1 sono state ben caratterizzate[101,103,110-112], e il completamento del sequenziamento del genoma di scimpanzé, ha permesso di affermare che gli elementi L1 hanno subito differenti destini evolutivi nell'uomo e nello scimpanzé, intorno a circa 6 milioni di anni fa[113]. Inoltre è stato stimato che in media il genoma umano presenta 80-100 copie di elementi L1 attivi dal punto di vista della retrotrasposizione, 6 dei quali, detti elementi L1 "caldi" (hot L1), sono probabilmente i maggiori responsabili del grande quantitativo di retrotrasposoni L1 nel genoma[26,114,115]. Questi 6 elementi appartengono tutti alla sottofamiglia L1Ta e ciò implica che si possa parlare di geni sorgente iperattivi all'interno di una stessa sottofamiglia[26]. La nozione di geni sorgente iperattivi deriva da uno dei primi modelli ideati per spiegare i moti di amplificazione alla base dell'attuale configurazione del nostro genoma. Secondo tale modello (detto "*The master gene model*"[116]), solo un certo numero di geni sorgente iperattivi, chiamati *master genes*, rimangono attivi dal punto di vista della retrotrasposizione in determinati periodi di tempo, e sono in grado di amplificarsi per generare un aumento significativo del proprio numero di copie genomiche. Per gli elementi L1, i 6 geni iperattivi sorgente sono quindi tutti appartenenti alla sottofamiglia L1Ta.

Al contrario degli elementi L1, per i quali si possono distinguere numerose famiglie, le prime analisi di conservazione delle sequenze *Alu* genomiche hanno permesso di classificare gli elementi *Alu* in solo 3 famiglie: *Alu*J, a cui appartengono gli elementi più antichi e meno attivi, *Alu*S (le sottofamiglie più attive sono *Alu*Sx, *Alu*Sq ed *Alu*Sc) a cui appartengono gli elementi intermedi[117,118], ed *Alu*Y, la famiglia evolutivamente più recente e tuttora più attiva, con le

sottofamiglie *Alu*Ya5 e *Alu*Yb8 dominanti nell'uomo[119]. Inoltre, vi sono evidenze del fatto che i membri della famiglia *Alu*Y siano emersi, in quanto famiglia a sé stante, dai 18 ai 25 milioni di anni fa e, successivamente, dopo circa 20 milioni di anni di quiescenza retrotrasposizionale, si siano espansi maggiormente nel genoma umano. Ciò spiegherebbe la grande differenza esistente tra il numero di copie di *Alu*Yb appartenenti a umani e primati non umani e, anche in questo caso, come nel caso degli elementi L1, fa identificare nei membri della famiglia *Alu*Y i geni sorgente iperattivi responsabili della loro amplificazione e del loro attuale contenuto nel genoma umano.

Infine, per quanto concerne gli elementi SVA, esistono solo sei famiglie (denominate SVA_A, SVA_B, SVA_C, SVA_D, SVA_E e SVA_F) e l'analisi dell'età evolutiva di tali famiglie, mediante lo studio della divergenza in sequenze, ha rivelato che 4 famiglie (SVA_A, SVA_B, SVA_C e SVA_D) si sono espanse prima della divergenza tra uomo, scimpanzé e gorilla, mentre le rimanenti famiglie SVA_E e SVA_F sono presenti solo nell'uomo. Inoltre, si è messo in luce che gli elementi SVA rappresentano la famiglia di retrotrasposoni più giovane nei Primati[83], e si hanno meno informazioni rispetto agli elementi ripetuti più studiati LINE-1 ed *Alu*.

3. Elementi trasponibili come fonte di variabilità genetica

Dalle mutazioni ai polimorfismi, le varietà genetiche riflettono la diversità all'interno di una popolazione, e alterazioni o cambiamenti del DNA sono alla base dell'evoluzione. Prendendo in considerazione il DNA, due meccanismi molecolari sono responsabili dell'insorgenza di diversità: le mutazioni e la ricombinazione genetica. Le mutazioni, ovvero cambiamenti nella sequenza nucleotidica, compaiono o a causa di errori durante la replicazione del DNA, che non vengono corretti, o a seguito di rotture del DNA, mentre la ricombinazione è un normale processo che ha luogo durante la meiosi cellulare. Tuttavia, entrambi i processi possono anche risultare dall'attività o dalla semplice presenza di elementi trasponibili. La trasposizione non è il risultato di errori casuali durante la replicazione o durante la riparazione di danni al DNA, ma può essere considerata come un processo mutagenico attivo, che porta a mutazioni che sono differenti dai polimorfismi a singolo nucleotide (l'integrazione infatti, generalmente non avviene in regioni non funzionali, ma all'interno dei geni codificanti per proteine, alterando la struttura e la funzione delle proteine stesse[9] con ripercussioni sia sul loro funzionamento sia sui processi in cui sono coinvolte). Al contrario, la ricombinazione omologa non allelica indotta dagli elementi TEs può essere vista come un processo casuale (sebbene possa avere luogo con facilità) a differenza del normale crossing-over meiotico, ma il suo impatto sull'instabilità genomica può risultare anche maggiore di quello

generato da eventi di mutagenesi inserzionale, portando per esempio a delezioni cromosomiche, duplicazioni, traslocazioni e inversioni.

Tali alterazioni del DNA possono interessare la funzione dei geni e la struttura dei genomi, e possono avere come peggior risultato o la morte della cellula o la sua incapacità a completare la meiosi. Tuttavia si possono anche osservare casi di inserzioni che portano ad effetti moderatamente dannosi o nulli, e tali inserzioni possono permettere la sopravvivenza cellulare e contribuire alla variabilità genetica del genoma ospite.

I TEs possono alterare il genoma in molti modi differenti (si può passare da piccole modificazioni di sequenza ad estesi riarrangiamenti cromosomici) e, dal momento che la frequenza di tali eventi non è trascurabile, essi possono essere visti come i maggiori "attori" nella generazione della diversità genetica[120,121].

Dato che, come detto in precedenza, i retrotrasposoni non-LTR L1, *Alu* e SVA sono gli unici attivi oggigiorno nel genoma umano, sono questi a determinare instabilità genomica, riarrangiamenti cromosomici ed anche, qualora i cambiamenti introdotti vengano selezionati positivamente, innovazione genetica. Tuttavia, solo raramente i nuovi assetti vengono selezionati positivamente e generalmente la mobilità degli elementi ripetuti contribuisce alla generazione di nuove patologie (per esempio, una patologia ogni mille è generata da eventi di mutagenesi inserzionale[90]).

4. La trascrizione pervasiva: l'"hidden genome" è trascritto e attivo

Ciò che emerge dalla trattazione precedente è che gli elementi ripetuti sono stati, e sono tuttora, fonte di instabilità genomica ma anche di variabilità genetica utile all'evoluzione.

Inoltre, essendo gli elementi ripetuti L1, *Alu* e SVA gli unici attivi attualmente nel genoma umano, sono stati condotti su di loro numerosi studi, che li hanno esaminati come unità trascrizionali indipendenti in grado di essere espresse e successivamente di retrotrasporsi (la loro trascrizione è condizione essenziale per la loro trasposizione), "colonizzando" così nuovi loci all'interno del genoma stesso.

Tali retrotrasposoni, eccezion fatta per gli elementi L1 che codificano per le proteine ORF1p ed ORF2p, rappresentano una porzione consistente di quello che viene definito genoma non codificante, messo in luce inizialmente grazie al primo lavoro di sequenziamento del genoma umano[3]. In seguito, con l'avvento di innovative tecnologie di sequenziamento massivo (NGS, Next Generation Sequencing), quali per esempio il sequenziamento del DNA, il sequenziamento del trascrittoma (ovvero l'insieme dei trascritti cellulari, mediante la tecnica RNA-sequencing (RNA-

seq[122])) e il sequenziamento dei frammenti genomici legati a specifiche proteine immunoprecipitate (mediante ChIP-sequencing (ChIP-seq[123])), il genoma umano è stato rapidamente conosciuto in ogni sua parte ed è stata rivelata la presenza di numerosi trascritti fino ad allora sconosciuti, di regioni con cromatina aperta e chiusa, di diverse modificazioni istoniche e nucleotidiche, nonché la presenza di diversi fattori trascrizionali di cui si poteva determinare la localizzazione[124].

Grazie a questi studi è diventato chiaro che il processo di trascrizione è molto più complesso di quanto si pensava inizialmente e che esistono elementi funzionali nelle regioni intergeniche fino ad allora trascurate (si è scoperto che circa il 75%-85% del genoma umano è, infatti, trascritto in RNA primari e processati[125]). Studi condotti sul genoma su larga scala (*genome-wide*) hanno messo in evidenza che circa metà delle regioni genomiche coinvolte nelle malattie sono intergeniche[126], le stesse regioni intergeniche che si pensava fossero silenti. Nonostante alcune di queste regioni possano funzionare solamente come elementi a DNA, oggi si è a conoscenza del fatto che le regioni intergeniche possono essere trascritte[127,128] e che al loro interno vi sono numerosi geni, non codificanti per proteine, i quali vengono trascritti in RNA che svolgono importanti funzioni[129]. Inoltre è stato osservato che non si ha trascrizione solo nelle regioni intergeniche, ma anche in regioni intrageniche ed in regioni ripetute e che, anche in questo caso, si genera una miriade di trascritti non codificanti con importanti ruoli regolativi.

Questo imprevisto livello di complessità della trascrizione (non solo umana ma più genericamente eucariotica, con esempi ben definiti anche nel lievito *Saccharomyces cerevisiae*[130,131]) ha portato alla nozione di trascrizione pervasiva, ovvero a considerare la trascrizione non ristretta solo ai geni con caratteristiche funzionali ben definite, ma estesa ad un ampio gruppo di RNA, distinti da quelli codificanti per proteine (mRNA) e da quelli non codificanti ma che presentano funzioni ben definite (come rRNA, tRNA, snRNA, snoRNA). Ciò suggerisce che tali trascritti siano "indefiniti" e rappresentino una sorta di "rumore di fondo" della trascrizione all'interno del quale risulta difficile isolare specifiche entità. Tuttavia, molti trascritti "pervasivi" sono stati caratterizzati, nonostante spesso siano privi di una funzione conosciuta.

Il prodotto della trascrizione pervasiva è quindi un ampio gruppo di RNA non codificanti per proteine (detti ncRNA, non-coding RNA)[132] all'interno del quale si possono distinguere due classi: i ncRNA di piccole dimensioni (20-200 nucleotidi, detti sncRNA, small non-coding RNA) e i ncRNA di grandi dimensioni (più di 200 nucleotidi, detti lncRNA, long non-coding RNA). All'interno di tali classi si possono identificare vari sottogruppi in base alle loro origine, alla funzione, al meccanismo di azione ed a caratteristiche peculiari. Seguirà una trattazione di tale categoria di trascritti non

codificanti, al fine di mettere in luce i ruoli regolativi della porzione di genoma considerata "spazzatura" e sottolineare l'importanza del suo studio nell'ambito del presente lavoro di tesi, dal momento che gli elementi *Alu*, vista la loro abbondanza nel genoma umano, sono la possibile fonte della famiglia più numerosa di RNA non codificanti.

I non-coding RNA: classificazione e ruoli emergenti

Come riportato sopra, i ncRNA si suddividono, in base alle dimensioni, in piccoli (sncRNA) e grandi ncRNA (lncRNA) e studi emergenti stanno sottolineando sempre più il loro coinvolgimento in molteplici processi cellulari. Ogni classe presenta delle caratteristiche peculiari ed esplica specifiche funzioni biologiche, pertanto si procederà alla trattazione dei ruoli di tali trascritti nell'ambito della loro classificazione generale.

Fanno parte della classe degli sncRNA, gli small-interfering RNA (siRNA) e gli small non-coding RNA (sRNA) e tali trascritti hanno dimensioni comprese tra i 20 e i 200 nucleotidi.

La famiglia degli siRNA include gli short-interfering RNA (siRNA), i Piwi-interacting RNA (piRNA) ed i micro RNA (miRNA)[133] e vi sono evidenze del fatto che tali piccoli ncRNA siano in grado di regolare l'espressione di singoli geni o di un gruppo di geni correlati, di silenziare alcune regioni genomiche e di controllare la risposta antivirale[134]. La regolazione dell'espressione genica, operata da tali trascritti, può avere luogo sia durante che dopo la trascrizione e, nel primo caso, si esplica attraverso silenziamento genico trascrizionale, modificazioni dello stato della cromatina (da aperta a chiusa), modificazioni istoniche e metilazione del DNA[135]. Di rilevante importanza sono i piRNA, in quanto è stato dimostrato che controllano la quantità di trasposoni e retrotrasposoni nelle cellule germinali umane, suggerendo una interazione tra ncRNA ed elementi ripetuti a fini regolativi.

La famiglia degli sRNA è invece composta da RNA non codificanti che si pensa abbiano un ruolo fondamentale nel regolare l'attività dei promotori e dei terminatori della trascrizione dei geni codificanti per proteine, e nella formazione dei cosiddetti *"gene-loop"*[136,137] (gli sRNA che formano dei *clusters* in corrispondenza dei promotori sono chiamati PASR (Promoter-Associated sRNA), mentre quelli che sono raggruppati all'estremità 3' dei geni sono chiamati TASR (Terminator-Associated sRNA)). Oltre a questa prima classificazione, recenti studi hanno messo in luce anche l'esistenza di una categoria di PASR, detta PROMPT (PROMoter uPstream Transcript), la cui presenza è correlata positivamente con l'attività dei promotori situati a valle[138], ed RNA trascritti a partire da sequenze fiancheggianti il sito di inizio della trascrizione (TSS) dei promotori attivi. Tali RNA possono essere espressi nello stesso orientamento dei geni o in orientamento opposto e

vengono detti TSSa-RNA (TSS-associated RNA) o tiRNA (transcription-initiation RNA). Infine, negli ultimi anni si è scoperta l'esistenza di spliRNA (splice-site RNA), che mappano in corrispondenza del sito di splicing donatore di esoni dei geni per proteine[136,139], e di piccoli RNA derivati dai retrotrasposoni. La distribuzione, le dimensioni e l'instabilità di questi RNA suggerisce che almeno alcuni possano essere un prodotto del fenomeno definito come RNA polimerasi II "in pausa", infatti sono documentati numerosi casi dell'enzima fermo sul DNA a 20-50 nucleotidi a valle del sito di inizio della trascrizione che mantiene potenziale attività di allungamento[140]. Tali trascritti potrebbero quindi derivare dalla dissociazione dell'enzima "in pausa" dal DNA, o dalla terminazione prematura della sua trascrizione, e il fatto che possano essere trascritti in orientamento senso o antisenso, rispetto alla trascrizione del gene codificante, suggerisce che la RNA polimerasi II possa iniziare la trascrizione in entrambe le direzioni.

È importante sottolineare inoltre, che non tutti gli sncRNA sono trascritti dalla Pol II, infatti vi sono numerosi esempi di trascritti ad opera della RNA polimerasi III, tra i quali è importante annoverare tre nuovi ncRNA che svolgono importanti ruoli regolativi: il ncRNA 17A, lo snRNA NMD29 ed il ncRNA 38A.

Il ncRNA 17A è sintetizzato della Pol III a seguito di stimoli infiammatori e promuove la produzione della variante B del recettore GABAB R2, con conseguente alterazione delle vie dipendenti dal neurotrasmettitore GABA nelle cellule cerebrali, ed un aumento della secrezione del peptide β amiloide, tipicamente presente nelle cellule cerebrali di individui affetti da Alzheimer (AD, Alzheimer's Disease)[141].

Lo snRNA NMD29 provoca la riduzione dello sviluppo del neuroblastoma (tumore infantile che si sviluppa nei bambini di età inferiore ai 3 anni), per aumentata sensibilità delle cellule tumorali ai trattamenti farmacologici. Il suo studio ha permesso di porre le basi per una nuova cura da affiancare ai trattamenti già consolidati[142].

Da ultimo, ma non per importanza, è da menzionare il ncRNA 38A, il quale è sintetizzato a seguito di uno stimolo infiammatorio, e regola lo splicing alternativo del trascritto primario del gene codificante per la proteina che interagisce con i canali potassio di tipo A (KCNIP4), con conseguente alterazione della conduttanza ionica attraverso tali canali e l'aumento della secrezione del peptide β amiloide. Il ncRNA 38A risulta sovraespresso nelle cellule della corteccia cerebrale di pazienti affetti da AD, e ciò conferma il suo ruolo, come nel caso del ncRNA 17A, nei processi neurodegenerativi[143]. Un recente studio ha anche messo in evidenza l'esistenza del trascritto murino omologo a tale ncRNA[144]: il gene per il ncRNA 38B di topo presenta la medesima collocazione genomica (nell'introne del gene KCNIP4), e il trascritto la medesima funzione (regola

anch'esso lo splicing alternativo dell'mRNA ospite) del ncRNA umano, e la sua scoperta ha permesso di aprire la strada all'identificazione di numerosi ncRNA murini omologhi a quelli umani. Tale evidenza permette di poter estendere la stesse analisi su molte unità trascrizionali Pol III identificate in topo, al fine di poter utilizzare topi transgenici per lo studio del ruolo di molti ncRNA, e di poter trasferire le stesse analisi sui rispettivi omologhi funzionali umani. Il fatto che tali ncRNA siano coinvolti nei processi di splicing alternativo aventi luogo nel cervello, con conseguenti effetti patologici rilevanti[144], potrebbe porre le basi per l'utilizzo del silenziamento genico come forma di terapia da affiancare alle terapie farmacologiche classiche.

Inoltre indagando i meccanismi alla base della regolazione epigenetica di tali ncRNA si potrebbe valutare se, a seguito di particolari stimoli, si abbia una loro derepressione trascrizionale, come ipotizzato per gli *Alu* RNA e LINE-1 RNA, che porti al loro accumulo a livello cellulare con conseguenti effetti tossici indagabili in topi transgenici e riportabili, in un secondo momento, all'uomo.

I lncRNA, al contrario dei piccoli ncRNA, rappresentano la classe più diffusa a livello cellulare e che comprende elementi in grado di coprire molteplici differenti funzioni[145-147]. Essi sono RNA non codificanti di dimensioni maggiori di 200 nucleotidi e vi sono evidenze della loro espressione in numerosi organismi comprendenti gli animali[148,149], le piante[150], il lievito[151], i procarioti[152] e persino i virus[153]. Nonostante ciò, tali trascritti non sono molto conservati tra le diverse specie se paragonati agli RNA maggiormente studiati (mRNA, miRNA e snoRNA)[154-156], e sono espressi generalmente a bassi livelli, facendoli ritenere inizialmente una sorta di "rumore di fondo" trascrizionale[154,157,158].

Numerosi studi condotti negli ultimi decenni, tuttavia, hanno permesso di scoprire che i lncRNA rivestono ruoli importanti in una grande varietà di processi biologici[159,160], che includono la trascrizione[161,162], lo splicing[163,164], la traduzione[80,165], la localizzazione delle proteine[166], l'integrità della struttura cellulare[167], l'imprinting[167], il ciclo cellulare[168], l'apoptosi[153,168], la riprogrammazione e la pluripotenza[169,170] delle cellule staminali e la risposta agli shock termici[171,172]. I lncRNA possono inoltre regolare la progressione tumorale[173] e lo sviluppo di molte altre malattie umane[174] e molti presentano una estremità 3' poliadenilata, un *cap* al 5' e più esoni,[154] ed una attivazione della trascrizione simile a quella degli mRNA (i trascritti di geni codificanti per proteine).

Tutte queste funzioni e caratteristiche biologiche li rendono molto interessanti e oggigiorno oggetto di numerose ricerche. Per comprendere al meglio questa classe di RNA e le loro funzioni, è utile classificarli sulla base della collocazione genomica, degli effetti esercitati sul DNA, del meccanismo di funzionamento e dei loro bersagli.

Sulla base della collocazione genomica i lncRNA possono essere suddivisi in intergenici, intronici, senso ed antisenso (**Figura 7**). I lncRNA intergenici (lincRNA[175], **Figura 7A**) sono trascritti, come suggerisce il termine, a partire da regioni intergeniche da entrambi i filamenti di DNA, e si è visto che possono regolare la trascrizione, lo splicing ed altri processi post-trascrizionali, nonché controllare la traduzione. Essi sono espressi a livelli paragonabili agli mRNA e, a testimonianza di ciò, sono stati trovati associati alla modificazione istonica H3K36me3, tipica dei geni attivamente trascritti[176]. Tali RNA sono più conservati dei trascritti intronici ed antisenso, hanno un'espressione tessuto-specifica maggiore rispetto agli mRNA e sono più stabili dei lncRNA intronici[177]. Questi trascritti sono di notevole interesse dal momento che è stato osservato un loro coinvolgimento nella proliferazione cellulare e nella progressione tumorale[178,179]. Al contrario dei precedenti, i lncRNA intronici (**Figura 7B**) sono poco espressi e si è esplorata la funzione solo di una loro piccola frazione; essi si trovano all'interno di introni di geni codificanti per proteine e la distinzione in senso ed antisenso è da riferirsi alla direzione della loro trascrizione rispetto a quella dei geni codificanti stessi. Infatti, i lncRNA senso (**Figura 7C**) sono trascritti a partire dallo stesso filamento da cui sono trascritti i geni codificanti per proteine, e possono contenere esoni o essere parzialmente o totalmente sovrapposti con il gene codificante, mentre i lncRNA antisenso (**Figura 7D**) sono trascritti a partire dal filamento opposto rispetto ai geni codificanti per proteine, e possono sovrapporsi parzialmente con esoni o introni del gene codificante o completamente all'introne di una sequenza codificante per proteine.

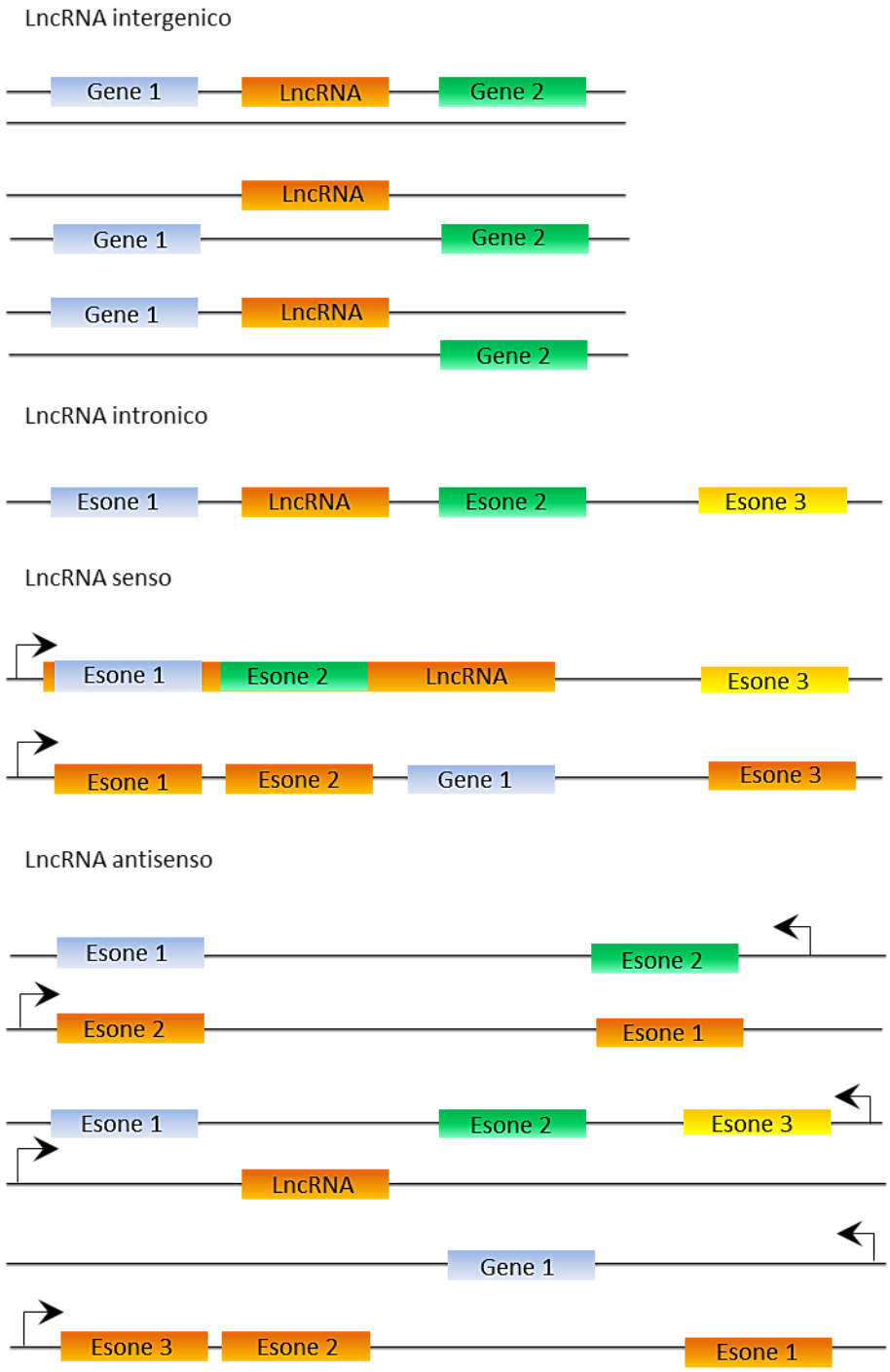


Figura 7. Rappresentazione schematica delle categorie di lncRNA sulla base della loro collocazione genomica. È possibile distinguere i lncRNA intergenici **(A)**, intronici **(B)**, senso **(C)** e antisenso **(D)**. I lncRNA sono rappresentati in arancione.

Come accennato all'inizio del paragrafo, i lncRNA possono essere classificati non solo in base alla collocazione genomica, ma anche sulla base degli effetti esercitati sulla sequenza di DNA: in tal caso si distinguono in lncRNA agenti in *cis* (*cis*-lncRNA) e lncRNA agenti in *trans* (*trans*-lncRNA), i cui ruoli saranno riportati di seguito.

Le funzioni principali dei *cis*-lncRNA, comprendono la loro capacità di interferire con la trascrizione (impedendo, per esempio, la formazione del complesso di pre-inizio[161] legandosi al promotore o interagendo con i fattori trascrizionali[161]) e di mediare modificazioni cromatiniche, reclutando, appunto, i complessi di modificazione della cromatina. Il miglior esempio a tal proposito è rappresentato dal reclutamento, da parte dell'RNA *Xist*, del complesso PRC2 (Polycomb Repressive Complex 2)[180] per indurre la trimetilazione del residuo di lisina 27 dell'istone H3 (H3K27me3) con conseguente silenziamento dei geni sul cromosoma X[181]. Al contrario, vi sono anche esempi di interazioni tra lncRNA e complessi di modificazione della cromatina che promuovono la trascrizione[182]. Come suggerisce il termine *cis*, tali trascritti regolano l'espressione di geni posti nelle loro vicinanze.

I *trans*-lncRNA, al contrario, sono in grado di agire su geni distanti dal luogo della loro sintesi e possono esplicare la propria funzione indipendentemente dalla complementarietà di sequenza con il gene bersaglio (necessaria invece per i cis-lncRNA). Oltre ad influenzare anch'essi l'espressione genica reclutando complessi di modificazione della cromatina[183,184], tali RNA possono legare i fattori di allungamento della trascrizione[185] (come nel caso dell'RNA 7SK di 330 nucleotidi) e le RNA polimerasi[171] (come il lncRNA B2 SINE che lega stabilmente il macchinario trascrizionale della Pol II inibendo la sua attività a seguito di shock termico), suggerendo un diverso meccanismo in grado di influenzare la trascrizione genica. Anche per gli *Alu* RNA è stato riportato un meccanismo di azione, a seguito di heat shock sia *in vitro* che nelle linee cellulari umane, analogo a quanto osservato per il lncRNA B2[186]: tali trascritti, infatti, sono in grado di reprimere la trascrizione degli mRNA legando direttamente e stabilmente la RNA polimerasi II e rimanendo associati al promotore insieme all'enzima. Essi impediscono alla Pol II di legare il promotore (sia a monte che a valle della TATA box) durante la formazione del complesso chiuso, e ciò comporta la repressione trascrizionale dei geni a causa della formazione di complessi trascrizionalmente inerti[187] (**Figura 8**).

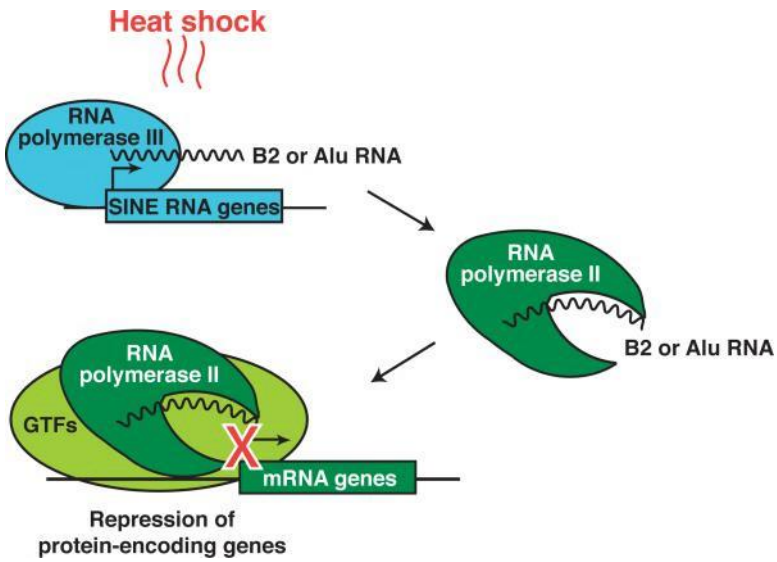


Figura 8. Rappresentazione schematica della repressione trascrizionale dei ncRNA B2 e *Alu* sui geni codificanti per protene. A seguito di shock termico, si ha un aumento della sintesi di tali trascritti, i quali interagiscono con la Pol II e con i fattori generali di trascrizione (GTFs, General Transcription Factors) impedendo la formazione del complesso chiuso e, in ultima analisi, la trascrizione genica (Adattato da Goodrich and Kugel, 2010[188]).

Un'altra classificazione emerge prendendo in considerazione il meccanismo di funzionamento dei lncRNA: si possono distinguere, pertanto, in RNA che regolano la trascrizione, RNA che regolano i processi post-trascrizionali e trascritti che presiedono ad altri meccanismi cellulari.

Per quanto riguarda la prima categoria, come detto in precedenza, essa comprende RNA che regolano l'espressione genica interferendo con la trascrizione o rimodellando la cromatina. Fanno parte di questo gruppo i lncRNA trascritti a partire da enhancer (eRNA) che, al contrario degli altri membri di questo gruppo, sono in grado di regolare positivamente la trascrizione dei geni[189]. Tra i trascritti non codificanti che regolano i processi post-trascrizionali, si distinguono quelli in grado di influenzare la maturazione degli mRNA (splicing) legando[190] o modulando[164] i fattori di splicing, oppure legandosi direttamente alle sequenze di mRNA impedendone la maturazione[163,165]. Fanno parte di questa categoria anche i trascritti che possono regolare la traduzione legandosi a fattori di traduzione[80] o al ribosoma, ed un esempio di legame al ribosoma è rappresentato dal lncRNA citoplasmatico *snaR* (small NF90-associated RNA), codificato da geni derivati da elementi *Alu* e trascritto dalla Pol III, il quale lega il ribosoma e può influenzare probabilmente la traduzione dell'mRNA NF90[52,191]. Oltre alla regolazione dello splicing e della traduzione, recenti studi hanno

anche messo in evidenza esempi di lncRNA che agiscono come i siRNA[192,193] o interagiscono con i miRNA[194].

Per quanto concerne il primo caso, essi possono agire come inibitori antisenso e promuovere la degradazione dell'mRNA, e ciò è stato visto con 1/2sbsRNA1 (half-STAU1-binding site RNA1), un lncRNA di 700 nucleotidi che lega la 3'UTR di un mRNA attraverso un elemento *Alu*[193]. Nel lncRNA è presente una sequenza *Alu* che si lega ad una sequenza *Alu* nella 3'UTR dell'mRNA bersaglio, mediante un appaiamento non perfettamente complementare. Tale appaiamento imperfetto genera una struttura secondaria intermolecolare con un elemento strutturale riconosciuto e legato dalla proteina STAU1, che a sua volta recluta la proteina UPF1, richiesta per il processo di degradazione dell'mRNA (mRNA decay). Esistono poi, come accennato in precedenza, lncRNA in grado di interagire direttamente o indirettamente con i miRNA per stabilizzare gli mRNA bersaglio[194]: tali RNA sono chiamati ceRNA (competing endogenous RNA) e fa parte di questa categoria il linc-MD1 (lincRNA associato al differenziamento muscolare), un lncRNA intergenico di 500 nucleotidi che agisce come una "spugna" per il miRNA 133 e il miRNA 135 e regola l'espressione di due fattori di trascrizione, MAML1 e MAF2C, che attivano l'espressione di geni muscolo-specifici[195].

Infine, vi sono numerosi lncRNA coinvolti in altri processi cellulari, come la localizzazione proteica[166], la replicazione dei telomeri[196], l'*RNA interference*[197], oltre alla regolazione della trascrizione e della traduzione, di cui sopra.

Per concludere, si possono distinguere lncRNA che mostrano un'espressione cellula-specifica e rispondono a diversi stimoli (*Xist*), trascritti che legano e sequestrano una proteina bersaglio ma non esercitano altre funzioni addizionali (DFHR[161]), RNA che legano le proteine e poi guidano il complesso ribonucleoproteico verso specifici bersagli (*Xist*), e trascritti che agiscono come "scaffold", ovvero come una sorta di "piattaforma centrale" che tiene unite più proteine per formare i complessi ribonucleoproteici (HOTAIR[183]).

È importante sottolineare che i lncRNA possono anche essere raggruppati, in base al tipo di interazioni con i loro bersagli, in trascritti che sono coinvolti in interazioni RNA-RNA, in ibridi RNA-DNA, interazioni mediate da strutture ad RNA, e RNA che legano le proteine[198].

L'esame dettagliato delle varie categorie di ncRNA e delle loro specifiche funzioni, ha permesso di sottolineare l'importanza di tali trascritti non codificanti nel contesto cellulare. Nell'ambito del presente lavoro di tesi, va sottolineato che gli elementi *Alu* possono essere considerati come geni codificanti per una sottofamiglia molto numerosa di lncRNA, il cui studio, ancora agli albori, potrebbe rivelare un loro coinvolgimento nella regolazione di processi cellulari essenziali. Essendo

il trascritto primario di un elemento *Alu* lungo circa 300 nt ed essendo gli elementi *Alu* molto abbondanti nel genoma, essi rappresentano potenzialmente la classe più abbondante di lncRNA. Nonostante ciò, dal momento che vi sono evidenze di un processamento di *Alu* RNA in trascritti simili agli siRNA, tali trascritti potrebbero, in ultima analisi, essere attribuiti anche alla sottofamiglia di sncRNA[199]. Detto ciò, si potrebbero ipotizzare per gli *Alu* RNA alcuni meccanismi regolativi peculiari della loro famiglia di ncRNA di appartenenza, e sono riportate alcune evidenze di un loro coinvolgimento nella regolazione della trascrizione e della traduzione, portando a concludere che non solo siano in grado perturbare le cellule generando instabilità genomica come elementi a DNA, ma anche come trascritti[200].

5. *Cross-talk* tra elementi ripetuti e ncRNA

Ogni categoria di lncRNA esaminata, come anche quella di sncRNA e di elementi ripetuti, ha permesso di far luce su una consistente porzione del genoma, ritenuta per molto tempo priva di funzione.

Vi sono infatti numerose evidenze di ncRNA che hanno origine da elementi ripetuti (per esempio un sottogruppo di miRNA non deriva da sequenze genomiche intergeniche, ma da elementi ripetuti[201] quali LINEs e SINEs, retrotrasposoni LTR e trasposoni a DNA), nonché delle interazioni tra ncRNA e TEs che sono alla base di una vasta gamma di funzioni regolative[202].

Si è scoperto infatti che alcuni lncRNA possono contenere elementi *Alu* che permettono loro di legarsi agli mRNA bersaglio e condurli alla degradazione[193], così come vi sono piRNA organizzati in *clusters* costituiti da retrotrasposoni o parti di essi. Tali RNA possono avere sequenze complementari ai trasposoni e possono legarvisi mediante appaiamenti perfetti o imperfetti e, una volta appaiati, possono interagire con il complesso delle proteine Piwi e proteggere la cellula dall'"invasione" dei retrotrasposoni stessi[203]. Tale "protezione" avviene mediante degradazione del TE RNA ad opera del complesso piRNA-proteine Piwi. Inoltre è stato riportato che i piRNA possono indurre la metilazione degli elementi L1 nel genoma umano, impedendone la trascrizione e mantenendoli in uno stato di silenziamento[204], ed anche che l'espressione di L1 RNA provoca una riduzione della sintesi di snRNA, suggerendo una possibile correlazione tra le espressioni delle due classi di trascritti[205].

Tutti questi *cross-talk* tra elementi ripetuti e trascritti non codificanti, pongono ulteriormente l'accento sul fatto che la porzione più consistente del genoma non solo non è priva di funzione, ma

è in grado di regolare, in modo molto fine, molteplici processi essenziali per la sopravvivenza e per il corretto funzionamento cellulare.

6. Meccanismi epigenetici e controllo dell'espressione dei TEs

Dalla trattazione precedente emerge che grazie agli studi condotti negli ultimi decenni, resi possibili dall'avvento di tecniche sofisticate in grado di permettere il sequenziamento sia del genoma che del trascrittoma[3,206], si è fatta luce sugli elementi trasponibili attivi (LINE-1 e *Alu*). Ormai non più DNA "spazzatura", tale porzione del genoma è stata notevolmente rivalutata ed è stato dimostrato che è implicata in numerose importanti funzioni biologiche. Rivolgendo l'attenzione all'espressione di tale porzione genomica, si è osservato che questa risulta finemente regolata attraverso diversi meccanismi epigenetici, i quali prevedono sia alterazioni nello stato del DNA sia della cromatina (ovvero il complesso di DNA e proteine in cui sono organizzati i cromosomi) e che non implicano, tuttavia, cambiamenti nella sequenza del DNA stesso[207].

Metilazione e demetilazione del DNA

La metilazione del DNA è la modificazione epigenetica più studiata in quanto cruciale per lo sviluppo degli organismi superiori. Tale modificazione, infatti, è coinvolta in numerosi processi cellulari essenziali (come la regolazione della trascrizione, il differenziamento cellulare, il mantenimento dell'identità cellulare, l'inattivazione del cromosoma X, l'imprinting genetico e la risposta cellulare ai cambiamenti ambientali[208]), risulta finemente regolata, e la sua caratteristica reversibilità è alla base di repentini cambiamenti locali nello stato di metilazione del DNA, che contribuiscono alla differente espressione genica di un determinato tipo cellulare rispetto agli altri. Tale processo biochimico prevede l'attacco covalente di un gruppo metilico (CH_3) al carbonio in posizione 5 (C_5) di un residuo di citosina (C), per formare la 5-metilcitosina (5MeC)[209] (**Figura 9A**). Il gruppo CH_3 è trasferito da un residuo di S-adenosil-L-metionina (SAM) al residuo di C dagli enzimi della famiglia delle DNA metiltrasferasi (DNMTs), che nei mammiferi sono: DNMT1, DNMT2, DNMT3A, DNMT3B e DNMT3L[210]. La DNMT1 è l'enzima responsabile della metilazione delle citosine emimetilate nei dinucleotidi CpG durante la replicazione del DNA[211] e viene detta DNMT di mantenimento, mentre le DNMT3A e DNMT3B prediligono i dinucleotidi CpG non metilati e operano la metilazione *de novo* durante lo sviluppo precoce[212]. Nonostante questa iniziale classificazione, oggigiorno si sa che le DNMT1 possono metilare il DNA genomico *de novo*[213] e,

viceversa, le DNMT3A e DNMT3B possono mantenere la metilazione del DNA durante la replicazione cellulare[214]. La DNMT2 ha una debole attività di DNA metiltrasferasi ma agisce soprattutto come RNA metiltrasferasi[215], ed infine la DNMT3L sembra aumenti la capacità delle omologhe DNMT3A e DNMT3B di legare il gruppo metilico *in vivo*[210].

Circa il 2-8% delle citosine del genoma dei mammiferi è metilato, soprattutto nei dinucleotidi CpG[216]. Nel genoma umano tali dinucleotidi sono distribuiti in modo non uniforme: si possono infatti distinguere regioni ricche e regioni povere in CG, e non tutti i siti CpG sono metilati. Esistono delle regioni particolarmente ricche in siti CpG, dette isole CpG (CpG *islands*, CGI), la cui metilazione è ridotta in corrispondenza dei promotori dei geni (circa il 50% dei geni umani presenta CGI vicino al sito di inizio delle trascrizione) e più elevata all'interno della sequenza genica o in regioni intergeniche. I promotori ricchi in CGI sono per lo più non metilati, a causa della notevole quantità di siti di legame, ricchi in CG, per i fattori di trascrizione[217], e la loro metilazione spesso è correlata con la repressione della trascrizione genica. Tuttavia, Niesen e collaboratori hanno evidenziato l'esistenza delle proteine RFX (Regulatory Factor X) in grado di legare i promotori dei geni di classe II del complesso maggiore di istocompatibilità (MHC, Major Histocompatibility Complex), nonostante tali promotori siano metilati (il legame avviene in corrispondenza dell'elemento promotoriale X box). Ciò ha permesso di svelare che non sempre la metilazione del DNA impedisce la trascrizione. Nello studio, è stato ipotizzato che il legame delle proteine RFX al promotore metilato dei geni MHC II, o impedisca alle proteine MBD (Methyl-DNA Binding Domain) di legare le citosine metilate o rimuova le proteine MBD già legate, ostacolando il loro successivo reclutamento del complesso delle istone deacetilasi e del complesso di rimodellamento della cromatina, con conseguente attivazione della trascrizione dei geni, nonostante la metilazione del loro promotore[218].

Circa 1/3 dei siti CpG del genoma umano è presente negli elementi *Alu*[219,220] (**Figura 9B**) e, nonostante nelle prime fasi dell'embriogenesi la loro espressione venga consentita, in un individuo adulto tale espressione è quasi completamente inibita. Le citosine dei siti CpG degli elementi *Alu*, infatti, risultano completamente metilate, e dal momento che ciò si verifica nelle cellule in condizioni normali, si può supporre che la ipermetilazione degli elementi ripetuti venga utilizzata dalle cellule come meccanismo di difesa[220]. Una riduzione della mobilità dei retroelementi, con conseguente "protezione" del DNA genomico, mediante ipermetilazione dei siti CpG, è stato osservato anche per gli elementi L1[221], pertanto si può concludere che la ipermetilazione di tali elementi può portare al loro silenziamento trascrizionale, proprio come avviene per i geni codificanti per proteine. Un'inibizione della loro sintesi fa sì che essi non possano retrotrasporsi e,

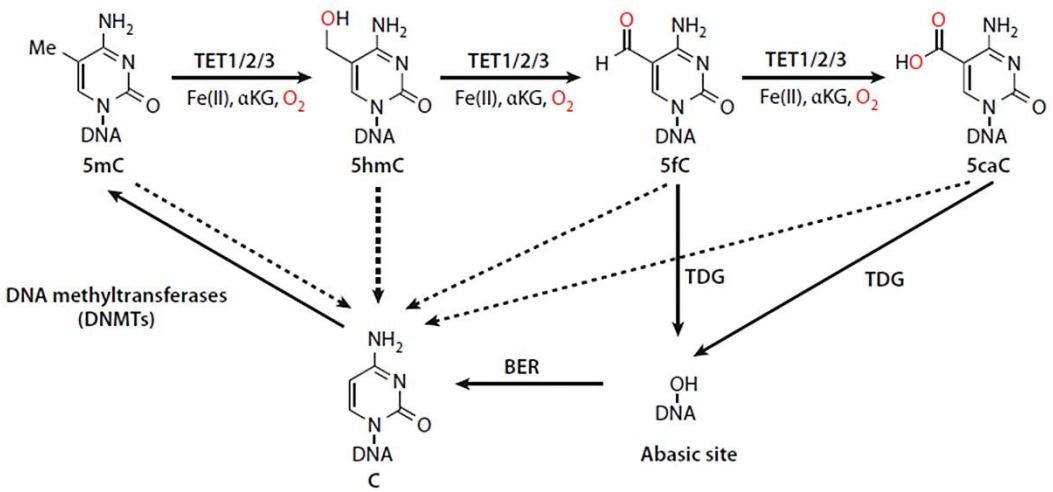
quindi, inserirsi in nuovi loci generando instabilità genomica[222] e le conseguenze riportate in precedenza[223]. L'effetto della metilazione, e quindi del silenziamento trascrizionale, dei retrotrasposoni L1 e *Alu* risulta anche in accordo con il fatto che i retroelementi effettivamente espressi nelle cellule normali sono solo una frazione molto ridotta del loro numero di copie nel genoma: circa 80-100 trascritti L1 e circa 100-1000 trascritti *Alu* nelle cellule HeLa[69,224]. L'esistenza di strategie di silenziamento avvalora l'idea che i retrotrasposoni siano la maggior fonte di instabilità genomica (come elementi mobili a DNA) e di interferenza con i normali processi cellulari (come RNA in grado di influenzare sia la trascrizione che la traduzione[200]), e permette di correlare il loro stato di metilazione con il loro silenziamento. Al contrario, uno stato di ipometilazione (ridotta metilazione) del DNA è generalmente connesso con l'attivazione trascrizionale[225] (la maggior parte dei geni essenziali per la cellula presenta CGI non metilate in corrispondenza dei TSS).

Inizialmente si pensava che la metilazione del DNA fosse un marcatore stabile, persistente ed ereditabile, e si era propensi a concludere che i gruppi metilici addizionati alle citosine non venissero più rimossi. Recenti evidenze hanno invece dimostrato che tale processo è reversibile e che i fattori di trascrizione (e proteine ad essi connesse) non solo proteggono le sequenze dalla metilazione, ma contribuiscono alla demetilazione del DNA[226]. Tale demetilazione, infatti, è stata sempre più osservata in contesti non fisiologici, ed in particolare la ipometilazione degli elementi L1 ed *Alu* è stata riscontrata in sporadici tumori e patologie[227-230], ed è stata associata a conseguente ridotta stabilità cromosomica[217,231,232] a seguito della loro retrotrasposizione.

Come la metilazione del DNA, anche la sua demetilazione è un processo enzimatico, il quale porta alla formazione della 5-idrossimetilcitosina (5hMeC) come prodotto principale, la quale è vista come la sesta base del genoma degli organismi superiori[233]. La demetilazione del DNA può essere o un processo passivo, che prevede la perdita della 5MeC a causa della inibizione del macchinario di mantenimento della metilazione, durante cicli successivi di replicazione cellulare, o un processo attivo, che può avvenire o per ossidazione o per deamminazione. Nel primo caso si ha l'ossidazione della 5MeC grazie all'azione delle deossigenasi della famiglia di enzimi TET (Ten-Eleven Translocation)[234,235]: TET1, TET2 e TET3 (**Figura 9A**). Tali enzimi sono in grado non solo di convertire la 5meC in 5hMeC, ma anche di ossidarla successivamente per sintetizzare la 5-formilcitosina (5fC) e la 5-carbossilcitosina (5caC)[236,237]. Nel secondo caso, invece, la deamminazione avviene grazie ai membri della famiglia AID/APOBEC (Activation Induced Deamminase/Apolipoprotein B mRNA-editing Enzyme Complex)[238], i quali deamminano la 5MeC o la 5hMeC con formazione di timina o 5-idrossimetiluracile (**Figura 9A**). Tali prodotti sono

successivamente convertiti in citosina durante il processo di riparazione del DNA (BER, Base Excision Repair) mediato dalla famiglia delle Uracil-DNA-Glicosilasi (UDG).

A



B

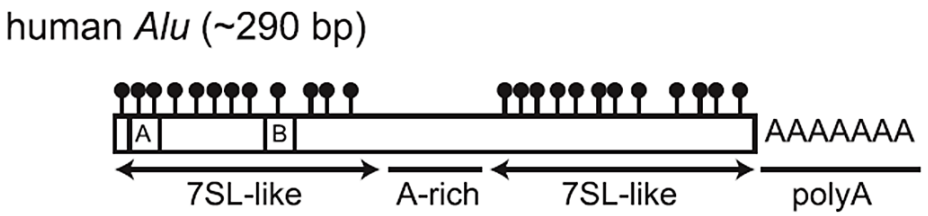


Figura 9. **(A)** Rappresentazione dei processi biochimici alla base della metilazione e demetilazione del DNA (Adattato da Shen et al., 2014[237]). **(B)** Distribuzione dei dinucleotidi CpG in un elemento *Alu* canonico di circa 290 bp. Esso presenta più di 25 siti CpG, sia all'interno che all'esterno del promotore per l'RNA Pol III (A box e B box). Questi siti, soggetti a metilazione, sono responsabili di circa 1/3 della metilazione presente nel genoma umano (Adattato da Ichiyanagi et al., 2013[220]).

La metilazione e la demetilazione del DNA sono pertanto due modificazioni epigenetiche finemente regolate e che permettono, a loro volta, una regolazione fine dell'espressione genica, sia di geni codificanti per proteine (di cui fa parte anche la classe degli elementi ripetuti L1) che non codificanti per alcun prodotto proteico (di cui fanno parte gli elementi *Alu*). Vedere i due processi enzimatici come indicativi di uno stato cellulare, ha aperto la strada verso la determinazione di numerosi ruoli funzionali dei geni che subivano una variazione in espressione dal tipo cellulare "normale" a quello "malato". Vi sono numerose evidenze, infatti, di come, regolando la trascrizione genica, la metilazione e la demetilazione del DNA siano implicate nello

sviluppo e nell'insorgenza delle malattie (per esempio dell'Alzheimer, del diabete, dell'ateroscleosi, dell'immunodeficienza, dell'artrite reumatoide, della sclerosi multipla, dell'atassia di Friedreich e del lupus eritematoso sistemico[239]), nonché siano associate a differenti forme tumorali e, in alcuni casi, si siano rilevati validi biomarcatori delle malattie. Per esempio, la ipermetilazione del gene per la glutatione S-transferasi pi 1 (GSTP1) è stata riscontrata nel cancro alla prostata[240] ed è stata utilizzata come indicatore diagnostico, mentre la ipermetilazione del gene per la O 6-metilguanina-DNA metiltrasferasi (MGMT) aiuta nella scelta della strategia terapeutica da utilizzare per il glioma[241].

Modificazioni istoniche come marcatori della regolazione trascrizionale

Oltre alla metilazione dei siti CpG, le cellule sono in grado di reprimere la trascrizione degli elementi trasponibili mediante modificazioni dello stato della cromatina, in particolare mediante modificazioni istoniche "repressive". Innanzitutto è bene ricordare che il DNA nelle cellule umane è impaccato in nucleosomi, strutture composte da un *core* di otto proteine istoniche (H2A, H2B, H3 e H4, ciascuno in duplice copia) attorno al quale è avvolto. Il DNA tra ciascun nucleosoma è detto DNA *linker* (di lunghezza variabile, tra 20 e 80 bp) e può venir legato da "istoni-*linker*" tra i quali il più diffuso è H1, mentre il DNA più strettamente legato al nucleosoma è detto DNA *core* (147 bp). La cromatina è organizzata in nucleosomi legati tra loro da istoni H1 e nel genoma si possono trovare regioni in cui la cromatina è altamente condensata (eterocromatina) e regioni in cui ha una struttura relativamente aperta (eucromatina). La differenza tra eterocromatina ed eucromatina risiede nell'organizzazione dei nucleosomi, che nel primo caso sono assemblati in strutture di ordine superiore che costituiscono un ostacolo per l'espressione genica, mentre nel secondo caso si trovano in strutture molto meno organizzate ed è possibile il legame dei fattori trascrizionali[242]. Gli istoni sono proteine basiche (contengono un gran numero di amminoacidi carichi positivamente, con almeno il 20% rappresentato da lisine ed ariginine) formate da un dominio chiamato *histone-fold* e da una coda N-terminale che sporge all'esterno del nucleosoma e può subire differenti modificazioni post-traduzionali (fosforilazioni, metilazioni e deacetilazioni sono le più frequenti) che alterano le loro interazioni con il DNA e con altre proteine nucleari[243]. Generalmente le lisine sono metilate o acetilate nelle code N-terminali degli istoni H3 ed H4, ed è riportato che un aumento nella loro acetilazione induce attivazione trascrizionale, mentre una riduzione è correlata con la repressione trascrizionale[244]. Al contrario, la metilazione è associata sia ad attivazione che a repressione della trascrizione in base al residuo di lisina modificato (per esempio, la trimetilazione delle lisine 4[245] e 36 dell'istone H3 è associata ad espressione genica,

mentre la trimetilazione delle lisine 9[246] e 27, del medesimo istone, è associata a repressione trascrizionale[247,248], **Figura 10**).

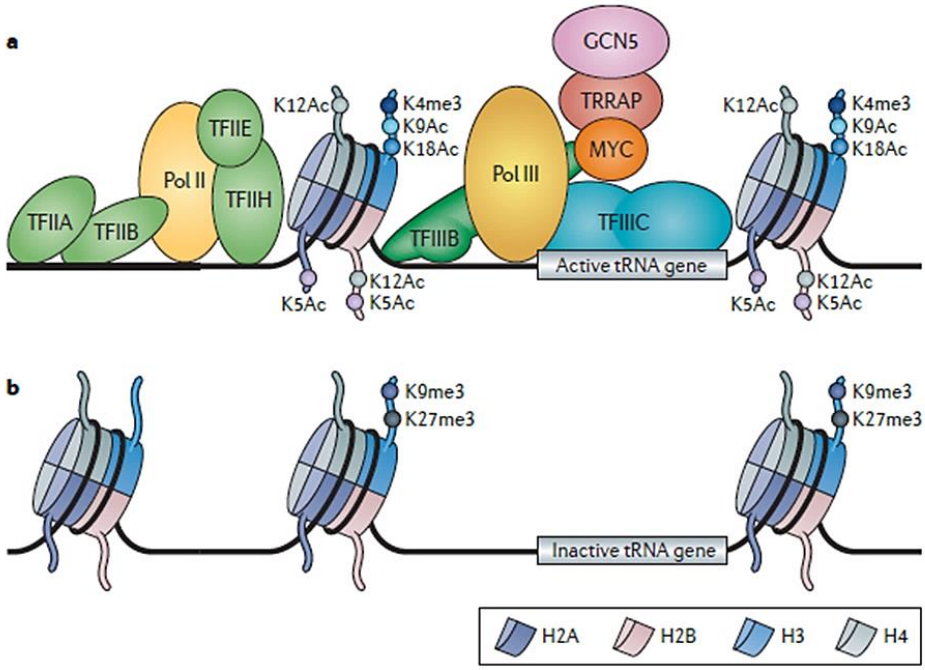


Figura 10. Rappresentazione delle modificazioni istoniche associate ad attivazione **(a)** e repressione **(b)** trascrizionale (Adattato da White et al., 2011[247]).

Recenti studi genetici e biochimici hanno permesso di determinare una correlazione tra la metilazione del DNA e le modificazioni istoniche, in particolare la metilazione delle lisine 9 e 27 dell'istone H3[249]. Nonostante siano modificazioni che coinvolgono enzimi differenti e differenti reazioni enzimatiche, i due processi sono dipendenti l'uno dall'altro, e spesso si possono trovare modificazioni associate ad uno stato trascrizionale attivo o repressivo sia a livello di DNA che di istoni. Un esempio si può trovare ancora una volta negli elementi ripetuti, in particolare negli elementi *Alu*: come detto nel paragrafo precedente, in condizioni normali essi sono repressi mediante ipermetilazione dei siti CpG all'interno della loro sequenza genomica, ed è stato visto che si trovano anche associati ad istoni H3 trimetilati a livello della lisina 9 (H3K9me3)[250], un tipo di modificazione istonica connessa a repressione trascrizionale[251] (**Figura 10**). Tali evidenze portano quindi a concludere che l'espressione degli elementi *Alu* non è solo regolata dalla metilazione del DNA, ma anche dalla metilazione di particolari code istoniche, e quindi che due meccanismi epigenetici sono responsabili della regolazione della loro trascrizione. A conferma di ciò, si è visto

che con un trattamento ipometilante (con l'agente ipometilante noto 5-aza-2'-deossicitidina) diminuiscono i livelli di H3K9me3, suggerendo che tale modificazione istonica possa essere influenzata dal grado di metilazione del DNA e possa contribuire alla repressione degli elementi *Alu*[250].

Questo permette di sottolineare che le modificazioni istoniche non solo giocano un ruolo importante nella regolazione della trascrizione di geni codificanti, ma attraverso questo processo può essere regolata anche l'attivazione o la repressione degli elementi ripetuti. Come accennato inizialmente, gli elementi L1 e *Alu* sono associati a modificazioni istoniche repressive (H2K9me2, H3K27me3) che inducono la formazione di eterocromatina e il loro silenziamento, ma allo stesso tempo è bene ricordare che tali modificazioni repressive associate ai retrotrasposoni potrebbero influenzare l'espressione dei geni ad essi vicini[252] (gli elementi ripetuti in prossimità dei geni presentano più modificazioni istoniche di quelli più lontani dai geni, e ciò suggerisce che alcune modificazioni epigenetiche dei retroelementi potrebbero essere necessarie per regolare l'espressione genica).

Inoltre è importante sottolineare che come gli elementi ripetuti, anche i ncRNA sono regolatori dei cambiamenti epigenetici, infatti influenzano la struttura della cromatina, le modificazioni istoniche (in particolare regolano la metilazione istonica) ed il reclutamento di complessi attivatori e repressori della trascrizione [253](v. trattazione precedente in cui sono sottolineati i ruoli dei membri delle differenti classi di ncRNA, come modulatori di modificazioni epigenetiche). Per esempio, i lncRNA alterano la metilazione degli istoni dei loro geni bersaglio direttamente o indirettamente, ed una classe di tali trascritti, i lincRNA, tendono a legare specifici complessi di modificazione della cromatina, pertanto la modificazione della cromatina associata a lincRNA può rappresentare uno specifico meccanismo di regolazione[254].

Infine è importante sottolineare che anche le modificazioni istoniche, come le alterazioni della metilazione del DNA, sono potenzialmente in grado di avere un impatto sull'espressione genica e sulla conseguente espressione proteica. Esse infatti sono state osservate in alcune malattie (per esempio nella malattia di Parkinson e nella malattia di Hungtington) e in diverse forme tumorali, nonché nelle stesse malattie in cui si è riscontrata un'alterata metilazione del DNA[239] (v. paragrafo precedente).

7. Influenza dell'ambiente sui meccanismi epigenetici

Numerosi studi condotti negli ultimi decenni hanno permesso di approfondire gli effetti dell'esposizione a differenti fattori ambientali sulla salute umana. Inizialmente, infatti, l'attenzione è stata concentrata su agenti, generalmente chimici, che potessero mutare la sequenza del DNA (generando cambiamenti genetici) e si consideravano tali mutazioni fondamentali per valutare il rischio di esposizione e per poterlo prevenire[255].

Al contrario, recenti scoperte hanno invece portato a considerare sempre più i cambiamenti epigenetici come possibili mediatori di specifici meccanismi di tossicità indotti dai fattori ambientali[256] e, sebbene per molti agenti tali meccanismi siano perfettamente compresi, per altri non sono stati ancora completamente delucidati.

Dal momento che i cambiamenti epigenetici, trattati nella sezione precedente, sono generalmente di modesta entità, potenzialmente cumulativi e si possono sviluppare nel tempo, non è sempre facile determinare un rapporto di causa-effetto tra l'esposizione ambientale, i cambiamenti epigenetici e le malattie.

I fattori ambientali, in grado di influenzare l'espressione del genoma umano, possono essere di varia natura e sono in grado di modificare reversibilmente sia il DNA che le code istoniche[257]. Inoltre, tali fattori sono anche potenzialmente in grado di mediare gli stati patologici perturbando i processi cellulari fondamentali, ovvero alterando l'espressione genica e, in ultima analsi, avendo un notevole impatto sull'espressione e sull'attività proteica. Negli ultimi decenni, molti studi hanno permesso di constatare che i cambiamenti epigenetici giocano un ruolo fondamentale nei processi trascrizionali alla base della regolazione dell'espressione genica, e sono sempre maggiori le associazioni tra l'esposizione a diversi fattori ambientali e tali cambiamenti. Per questo motivo, ha acquistato notevole importanza il campo della tossico-epigenomica, ovvero lo studio del rapporto tra le modificazioni epigenetiche e le alterazioni nel "normale" stato cellulare per esposizione ad agenti tossici, nell'ambito della ricerca volta a preservare la salute umana dall'esposizione ambientale. A questo proposito, infatti, è possibile che le alterazioni epigenetiche, associate all'esposizione ad agenti tossici, possano essere un fattore chiave nell'eziologia delle malattie connesse proprio all'esposizione ambientale[258-260], ma, nonostante ciò, il ruolo e la potenziale utilità dei dati epigenetici, nel processo di valutazione del rischio delle malattie, devono ancora essere definiti con precisione. Dato che la regolazione epigenetica dell'espressione genica, in risposta a determinati fattori ambientali, è sempre più documentata, si sta facendo largo la possibilità di utilizzare le modificazioni epigenetiche come biomarcatori degli effetti

dell'esposizione a tali fattori tossici, nella speranza di potere utilizzare tali marcatori come una sorta di "indicatori degli stati patologici", qualora essi siano associati ad un'espressione genica differenziale pre e post-esposizione.

Inoltre, dal momento che tali alterazioni potrebbero essere potenzialmente ereditabili, e quindi stabili per la loro rilevazione, si potrebbero anche utilizzare per predire o mettere in luce possibili cambiamenti ereditabili nell'espressione genica, a seguito dell'esposizione delle donne in gravidanza agli agenti tossici[261], prevenendo, in ultima analisi, possibili disturbi della salute del feto.

Seguirà la trattazione dei fattori ambientali per i quali sono riportate le maggiori evidenze dell'influenza sui meccanismi epigenetici *in vitro*, *in vivo* (studi sui modelli animali) e negli studi condotti sull'uomo, nonché il loro rapporto con l'espressione genica e le malattie.

I metalli

Tra i metalli maggiormente correlati con il rischio di insorgenza di patologie, è possibile annoverare l'arsenico, il piombo, il mercurio e il cadmio, i quali sono ad oggi tra i primi sette metalli nella lista delle sostanze pericolose per la salute umana, compilata dalla ATSDR (Agency for Toxic Substances & Disease Registry). Nonostante non siano ancora stati completamente compresi i meccanismi attraverso cui tali metalli esercitano la loro influenza, di notevole rilievo sono le testimonianze di come possano aumentare la produzione delle specie reattive dell'ossigeno (ROS, Reactive Oxygen Species), inducendo, così, stress ossidativo[262]. Inoltre, l'aumento dei ROS, e il conseguente danno ossidativo al DNA, si è visto che possono interferire con le interazioni tra le DNA metiltransferasi (DNMTs) e il DNA stesso[263,264], portando ad un'alterazione dello stato di metilazione genomica (ovvero un'alterata metilazione delle C nei siti CpG)[264], e tali evidenze hanno permesso di ipotizzare una possibile correlazione tra la produzione di ROS, da parte dei metalli, e le alterazioni dell'epigenoma cellulare. Detto ciò, seguirà la trattazione degli effetti epigenetici associati all'esposizione ai metalli menzionati sopra.

L'arsenico

L'arsenico è un metalloide (d'ora in poi ci si riferirà, comunque, all'arsenico come ad un metallo) tossico ubiquitario nell'ambiente, il quale contamina le acque potabili che ognuno di noi beve e, numerose evidenze, mettono in luce il fatto che centinaia di milioni di individui sono esposti in maniera cronica all'acqua contaminata e che in molti casi si supera il limite di esposizione, fissato per l'arsenico, di 10 ppb[265]. Inoltre, nonostante vi sia stata una notevole riduzione dell'esposizione

cronica a tale metallo negli ultimi decenni, si stima che un milione di individui continui oggigiorno ad essere esposto solo in Bangladesh, dove più di 9000 morti sono state attribuite proprio all'esposizione a tale metallo[266]. L'esposizione cronica all'arsenico inorganico (Ai), il primo metallo nella lista delle sostanze pericolose per la salute umana (ATSDR 2013), si è visto che è correlata con l'insorgenza di disordini cardiovascolari, respiratori, neurologici, gastrointestinali, riproduttivi e diabetici[267,268], ed è anche stata riportata l'insorgenza di encefalopatie e di neuropatie periferiche per esposizione acuta a tale metallo[269].

L'Ai è stato dichiarato agente cancerogeno per la salute umana di Gruppo 1 dalla IARC (International Agency for Research on Cancer), ed stato dimostrato che, oltre ai disordini menzionati sopra, sia in grado di provocare il tumore ai reni, alla pelle, al fegato, ai polmoni e alla vescica[270]. Gli studi condotti, al fine di valutare i possibili meccanismi di azione associati alla carcinogenesi indotta dal metallo, hanno portato a proporre diverse soluzioni, tra cui alterazioni delle vie di segnalazione cellulari, perturbazioni nel sistema di risposta al danno e nei processi di riparazione del DNA, aberrazioni cromosomiche e stress ossidativo[271,272], e sta diventando sempre più chiaro come le modificazioni epigenetiche giochino un ruolo fondamentale nella tossicità e nelle malattie indotte dall'arsenico. Si è osservato, infatti, che l'esposizione all'Ai è associata a cambiamenti epigenetici, quali la metilazione globale e gene-specifica del DNA, l'acetilazione, la metilazione e la fosforilazione istonica, e la variazione nell'espressione dei miRNA (piccoli RNA che sono stati notevolmente rivalutati come possibili marcatori dell'epigenoma cellulare) e, che, tali alterazioni potrebbero contribuire all'insorgenza dei tumori e di altre patologie[273]. Nonostante tali associazioni, tuttavia, sono necessari studi ulteriori per poter definire con precisione il ruolo dei cambiamenti epigenetici nella tossicità dell'arsenico e nelle malattie causate da tale metallo.

La metilazione del DNA è l'alterazione epigenetica, indotta dall'arsenico inorganico, più comunemente studiata. Innanzitutto, è bene ricordare che l'arsenico stesso viene attivato, a seguito dell'aggiunta di un gruppo metilico da un residuo di S-adenosil metionina (SAM), in una reazione catalizzata dall'enzima Arsenico Metiltrasferasi (AS3MT)[274]. Dal momento che le DNMTs utilizzano SAM come donatore di gruppi metile (v. paragrafo sulla metilazione del DNA), ciò ha suggerito un possibile ruolo della metilazione del DNA nella carcinogenesi dell'arsenico[275], anche a fronte del fatto che l'attività delle DNMTs diminuisce per esposizione all'Ai (nonostante non si conosca, ad oggi, il meccanismo con precisione)[276]. La maggior parte degli studi, infatti, vede l'esposizione all'Ai correlata alla ipometilazione globale del DNA, e un esempio rilevante è rappresentato da aumento del rischio di insorgenza di lesioni della pelle, indotte da arsenico, associato alla ipometilazione del DNA, interessente in quanto la ipometilazione del DNA è una

modificazione generalmente osservata nelle forme tumorali[277]. Tuttavia, vi sono anche evidenze di un aumento di metilazione globale indotta dal metallo[278,279] e, a tal proposito, è stata osservata una ipermetilazione globale del DNA in leucociti da sangue periferico di individui esposti all'arsencio in India e in Bangladesh[280,281]. Oltre alla metilazione globale del genoma, menzionata sopra, sono riportati anche numerosi casi di alterazione della metilazione gene-specifica, tra i quali si possono annoverare la metilazione del promotore del gene CDKN2A/p16 (soppressore tumorale, CDKN2A, Cyclin-Dependent Kinase Inhibitor 2A) in soggetti esposti all'arsenico in India[282] e in Cina[283], e la metilizione dei promotori dei geni PRSS3 (soppressore tumorale, Pretease Serine3), RASSF1A (Ras Association Domain Family member 1)[284] e DAPK (Death-Associated Protein Kinase1)[285], riscontrata nei tumori uroteliali e correlata con l'esposizone all'arsenico degli abitanti del New Hampshire e Taiwan. Queste e altre osservazioni confermano che l'arsenico altera la metilazione di geni coinvolti nella tumorigenesi quali: il soppressore tumorale CDKN2A/p16 e il gene TP53[279], nonché i geni VHL (soppressore tumorale, Von-Hippel Lindau)[286], RASSF1A[287], DAPK e RECK (Reversion-Inducing-Cysteine-rich protein with Kafal motifs)[288] e gli oncogeni CCND1 (Ciclyn D1), ER-α (Estrogen Receptor α)[289] e i geni per le proteine delle famiglie HRAS e KRAS (piccole proteine G appartenenti alla famiglia Ras)[290]. Inoltre, in aggiunta agli studi condotti direttamente sull'uomo, si hanno evidenze di un'alterata metilazione gene-specifica anche in numerosi studi condotti *in vitro*, e ciò conferma un ruolo rilevante dell'arsenico nella modulazione della metilazione del DNA. Per esempio si sono osservati cali di ipermetilazine nei promotori dei geni RASSF1A e p16 in linee cellulari tumorali trattate con triossido di arsenico[291], con conseguente aumento dei livelli dei loro trascritti.

Nonostante le numerose evidenze degli effetti dell'arsenico sulla metilazione del DNA, si è anche dimostrato che in soggetti esposti all'Ai in Bangladesh, la concentrazione del metallo nelle urine è inversamente proporzionale alla acetilazione globale del residuo di lisina 9 dell'istone H3 (H3K9ac) e, al contrario, positivamente correlato con la dimetilazione dello stesso residuo (H3K9me2)[292]. Inoltre, in tale studio sono state identificate anche diverse risposte a seconda del sesso, in quanto le donne mostravano una correlazione positiva tra l'esposizione all'arsenico e le modificazioni globali H3K4me3 e H3K27me3, mentre gli uomini, al contrario, una correlezione negativa[292]. Nonostante tali evidenze rilevate conducendo studi sull'uomo, la maggior parte degli effetti dell'arsenico sulle modificazioni istoniche è stata determinata mediante esperimenti *in vitro*. Huang e collaboratori, infatti, hanno dimostrato che il metallo induce l'espressione dose-dipendente del gene antiossidante ferritina attraverso la metilazione dell'arginina 3 dell'istone H4 (H4R3) e dell'arginina 17 dell'istone H3 (H4R17) nei cheratinociti umani[293]. Al contrario, sono state

riportate anche evidenze di un aumento della fosfo-acetilazione dell'istone H3 dei proto-oncogeni c-Jun e c-Fos[294] nei fibroblasti. Pertanto, tale metallo non è in grado solo di indurre una variazione nella metilazione del DNA, ma anche modificazioni a livello delle code istoniche e, dal momento che, i geni associati alle patologie indotte da tale metallo, presentano un'espressione differenziale nei soggetti patologici rispetto a quelli sani, si potrebbe pensare di utilizzare tali geni come biomarcatori delle patologie indotte dall'arsenico inorganico.

Il piombo

Il piombo è un elemento largamente presente in natura noto per i suoi innumerevoli effetti sulla salute umana: è stato stimato, infatti, che tale metallo contribuisce a circa lo 0.6% delle malattie, classificandosi in seconda posizione nella lista delle sostanze pericolose per la salute umana (ATSDR 2013). L'esposizione a bassi livelli di tale metallo, infatti, provoca problemi al sistema nervoso, al sistema immunitario[295], al sistema cardiovascolare[296] e al sistema renale[297], nonché una riduzione della fertilità maschile[298], problemi comportamentali[299] e predisposizione alla violenza[300]. Tutti i disturbi menzionati sono causati sia per esposizione cronica che acuta e, inoltre, si è osservato che tali effetti sono connessi all'esposizione già nell'utero o nella prima infanzia[301]. I meccanismi alla base degli effetti neurologici che si verificano per esposizione proungata al piombo, non sono ancora stati compeltamente delucidati, ma si pensa che la produzione di ROS, e le variazioni nell'espressione genica, possano giocare un ruolo nella patogenesi delle malattie associate a tale metallo[302].

Inoltre, è stato osservato che anche il piombo, come l'arsenico, genera una variazione nella metilazione del DNA gene-specifica riportata, per esempio, per il gene onco-soppressore CDKN2A/p16: tale gene è, infatti, risultato ipermetilato in individui con elevati livelli di piombo nel sangue, mentre solo parzialmente metilato in soggetti esposti a bassi livelli di tale metallo[303]. Nonostante generalmente le variazioni di metilazione siano indagate in geni specifici, vi sono sporadici casi riportati di come il piombo possa indurre la ipometilazione degli elementi ripetuti: si è infatti osservato che i livelli del metallo nella tibia materna sono inversamente proporzionali alla metilazione globale del DNA (misurata sugli elementi rupetuti LINE-1 e *Alu*) nei neonati, nonostante tale modificazione del DNA non è stata riscontrata in un altro studio, condotto sempre sulla tibia e indipendentemente dal primo. Tuttavia, questo secondo studio ha messo in luce una correlazione inversa tra la metilazione degli elementi LINE-1 (ma non degli elementi *Alu*) e i livelli di piombo nella rotula, portando a concludere che tale cambiamento epigenetico (negli elementi ripetuti e non solo nei geni codificanti per proteine in singola copia) potrebbe essere alla base del

meccanismo di azione del piombo e quindi delle malattie ad esso associate[304]. Infine, a differenza dell'arsenico e di altri metalli, non vi sono ad oggi casi riportati di associazione tra l'esposizione al piombo e alterazione delle modificazioni istoniche.

Il mercurio

Il mercurio è un elemento ubiquitario rilasciato nell'aria a seguito di fenomeni geologici, quali le eruzioni vulcaniche e l'evaporazione delle acque degli oceani, nonché presente in quantità elevate nella crosta terreste. Tale metallo risulta al terzo posto nella lista delle sostanze pericolose per la salute umana, in quanto l'uomo ne è esposto quotidianamente a causa delle numerose occupazioni che ne prevedono l'utilizzo: l'estrazione dell'oro, la fusione dei metalli, la combustione del carbone, la produzione del legno e le lavorazioni nelle industrie elettriche. Inoltre, il mercurio è anche presente nel pesce (in particolare nei frutti di mare), che normalmente viene consumato con la dieta, e nell'aria a seguito della combustione dei combustibili fossili e per evaporazione dei rifiuti contenenti mercurio[305].

Tale metallo si può trovare sia nella forma inorganica (mercurio inorganico) sia sottoforma di sali organici, di cui il metilmercurio (MeHg) è un esempio. Questo sale, che può essere presente ad alti livelli nel pesce, è stato considerato un potenziale agente neurotossico, in quanto in grado di causare alterazioni persistenti nell'apprendimento dei topi a seguito di esposizione perinatale. Il MeHq, infatti, attraversa facilmente la placenta e l'esposizione *in utero* è stata associata a numerosi disturbi, tra cui sono da annoverare i disturbi nell'apprendimento, la riduzione della funzione cognitiva e l'immunosoppressione[306]. Inoltre, studi condotti sull' uomo, hanno evidenziato che nei bambini e negli adulti il mercurio è stato associato a problemi nei sistemi cardiovascolare, respiratorio, neurologico ed anche ad alterazioni della normale funzionalità della tiroide, del fegato, dei reni e sistema immunitario[307], nonostante l'esposizione a bassi livelli di metallo. L'esposizione al mercurio, pertanto, non è solo causa di numerose patologie per esposizione prenatale o neonatale, ma l'esposizione cronica o acuta sembra essere un fenomeno ampiamente diffuso, che contribuisce in maniera rilevante allo sviluppo delle malattie.

Sono documentati soltanto un paio di lavori in cui vengono analizzati i cambiamenti nella metilazione del DNA nell'uomo a seguito di esposizione al mercurio. In uno studio, condotto su un campione di donne con elevati livelli di mercurio nel sangue, è stata osservata una ipermetilazione del promotore del gene GSTM1 (Glutathione S-transferase Mu 1), ma, tuttavia, non sono state trovate correlazioni statisticamente significative e non è stata neanche valutata la variazione nell'espressione del gene[308]. Un secondo studio, invece, ha rivelato una correlazione tra la

ipometilazione del promotore del gene SEPP1 (SEleno protein P Plasma 1) e un aumento dei livelli di mercurio, analizzando l'esposizione a tale metallo, mediante l'utilizzo di capelli provenienti da un campione di soggetti maschili[309]. I geni che subiscono una variazione della metilazione per esposizione al mercurio nei due studi, sono coinvolti nel sistema cellulare della risposta antiossidante, il quale riduce lo stress ossidativo e, dal momento che non è stata valutata l'espressione genica, si può solo supporre che l'alterata metilazione dei promotori dei due geni possa potenzialmente influenzare la loro espressione e alterare la risposta allo stress ossidativo.
Infine, per quanto concerne le relazioni tra l'esposizione al mercurio e le variazioni delle modificazioni istoniche, non vi sono studi su soggetti umani, ad oggi riportati. Nonostante ciò, studi condotti sui topi hanno dimostrato che il metilmercurio è associato ad un aumento della trimetilazione del residuo di lisina 27 dell'istone H3 (H3K27me3) ed a una diminuzione dell'acetilazione dell'istone H3 a livello del promotore del gene BDNF (Brain Derived Neurotrophic Factor). Tali modificazioni istoniche, per esposizione anche a bassi livelli di metilmercurio durante lo sviluppo murino, inducono la soppressione epigenetica dell'espressione del gene BDNF nell'ippocampo e ciò provoca la depressione nel topo esposto[310]. Un altro studio, condotto sulle cellule staminali embrionali murine esposte a cloruro di mercurio ($HgCl_2$), ha invece sottolineato una diminuzione della monometilazione del residuo di lisina 27 dell'istone H3 (H3K27) con conseguente decremento dei livelli della proteina THP (Total Histone Protein)[311]. Nonostante non siano ancora stati riportati effetti sulle modificazioni istoniche per esposizione dell'uomo al mercurio, o ai suoi derivati, la notevole conservazione delle risposte cellulari tra uomo e topo potrebbe permettere di ipotizzare un coinvolgimento del metallo nell'induzione di modificazioni istoniche anche nell'uomo.

Il cadmio

Il cadmio è un contaminante ambientale associato ad un notevole numero di malattie e, l'esposizione a tale metallo, avviene soprattutto attraverso il fumo delle sigarette ed ingestione, in quanto è presente in quasi tutti i cibi, seppur in quantitativi differenti[312]. Ad esempio, alte concentrazioni di cadmio si possono trovare nei crostacei, nei molluschi, nelle interiora di animali (come nel fegato e nei reni) e anche nella verdura verde e nei tuberi[312], cibi normalmente presenti nell'alimentazione quotidiana. Una particolarità di tale metallo è il possedere un'emivita molto lunga nell'uomo, corrispondente a 10-30 anni nel fegato, il principale organo bersaglio della sua tossicità a seguito di una dieta ricca di cadmio (si parla di esposizione cronica[313]). Inoltre, esso

esercita anche la sua tossicità nelle ossa, portando all'insorgenza dell'osteoporosi e della osteomalacia. Al contrario di altri metalli tossici, il cadmio è in grado di accumularsi nella placenta ed è stato visto che i livelli del metallo, nelle urine delle gestanti, sono inversamente correlati con il peso del neonato alla nascita, la circonferenza della sua testa e del torace[314]. Tale metallo è stato classificato come agente cancerogeno di Gruppo 1 (IARC), in quanto vi sono numerose evidenze della sua connessione con il tumore al fegato, alle ossa, ai reni e al pancreas, nonché con l'aumento del rischio di insorgenza delle malattie cardiovascolari[312,315].

Inoltre è stato osservato che l'esposizione al cadmio porta all'inibizione del sistema di riparo del DNA, alla generazione delle specie reattive dell'ossigeno (ROS) e all'alterazione del ciclo cellulare o dell'apoptosi[316] e, nonostante sia stato studiato l'effetto che tale metallo esercita sull'espressione genica, non si sa molto circa l'impatto che può avere sull'epigenoma. Nonostante ciò, l'esposizione prenatale al metallo è stata recentemente associata con un'alterazione della metilazione del DNA dei globuli bianchi[317], corripondente ad un arricchimento dei profili di metilazione in corrispondenza dei siti di legame per specifici fattori trascrizionali, che gli autori ipotizzano possano rappresentare una sorta di "*footprint* ambientale" nel DNA associato all'esposizione al cadmio. Inoltre, molti studi *in vitro* correlano le caratteristiche tipiche del processo tumorale per esposizione al cadmio ai cambiamenti nella metilazione del DNA. A tal proposito, infatti, sono stati osservati, nelle linee cellulari immortalizzate tratatte con cadmio (per esposizione cronica), i profili di metilazione del DNA differenziali tipici delle linee tumorali così come anche dei tessuti tumorali primari[318]. Tali risultati suggeriscono che i tumori, associati all'esposizione al metallo, potrebbero condividere gli stessi profili di metilazione e che questi potrebbero essere utilizzati come biomarcatori nel processo di valutazione del rischio di esposizione al cadmio. Ulteriori studi hanno messo in luce come il metallo possa indurre anche la proliferazione delle cellule K562 e al contempo una ipometilazione globale del DNA: trattando le cellule con la metionina prima del trattamento con il cadmio, inoltre, si è osservato che non solo viene abrogata la ipometilazione del DNA ma anche la proliferazione cellulare, portando a concludere che il cadmio potrebbe indurre la proliferazione cellulare diminuendo i livelli di metilazione del DNA[319].

Numerosi, inoltre, sono i casi riportati di variazioni della metilazione del DNA indotta dal cadmio, e in particolare è possibile annoverare la ipermetilazione dei geni soppressori tumorali p16 e RASSF1A, con conseguente loro silenziamento in cellule epiteliali della prostata (in cui il cadmio ha indotto la trasformazione a cellule tumorali) e la sovraespressione della DNMT3B[320], ed anche la ipermetilazione e diminuzione dell'espressione dei geni MSH2 (MutS Homolog 2, un gene coinvolto nella riparazione del DNA), ERCC1 (Excision Repair Cross-Complementation group 1),

XRCC1 (X-ray Repair Cross Complementing Protein 1) e OGG1 (8-Oxoguanine DNA Glycosylase)[321], associate con la sovraespressione di DNMT1 e DNMT3A. Per di più, un recente studio pone anche l'accento sulla ipometilazione degli elementi LINE-1 a seguito di esposizione a cadmio, aprendo la strada verso l'analisi degli effetti che tale inquinante ambientale può avere sugli elementi ripetuti, e non solo sui geni codificanti per proteine in singola copia nel genoma[322]. Gli studi riportati, pertanto, mettono in luce che il cadmio silenzia i geni soppressori tumorali e i geni coinvolti nella riparazione del DNA, attraverso la ipermetilazione dei loro promotori, mediata da un aumento dell'espressione delle DNA metiltrasferasi e, inoltre, si è visto che il metallo è anche in grado di aumentare l'espressione del gene HSD11B2 (Hydroxysteroid (11-Beta) Dehydrogenase 2) nei trofoblasti immortalizzati, e tale aumento di espressione è associato con la ipometilazione di tale gene, suggerendo ancora una volta che un meccanismo epigenetico è alla base della sua tossicità[323]. Infine, oltre alle modificazioni dello stato di metilazione del DNA, studi *in vitro* hanno dimostrato che il cadmio è anche in grado di aumentare i livelli delle modificazioni istoniche H3K4me3, H3K27me3 e H3K9me3 in corrispondenza del promotore del gene MT3 (MetallonThionein 3) nelle cellule uroteliali, suggerendo che l'esposizione cronica al cadmio possa alterare l'espressione genica anche attraverso una alterazione delle modificazioni istoniche. Il gene MT3, infatti, è normalmente silente e viene attivato per esposizione al metallo[324]. Infine, dal momento che le modificazioni epigenetiche indotte dal cadmio sono differenti a seconda del tipo di esposizione (ipermetilazione del DNA per esposizione cronica e ipometilazione per esposizione acuta), potrebbero essere utilizzate per predire le patologie sulla base del diverso tempo di esposizione al metallo.

Gli inquinanti atmosferici

Nonostante l'esposizione ai metalli, riportati in precedenza, provochi numerose patologie e siano sempre maggiori le evidenze di una correlazione tra tale esposizione e le modificazioni epigenetiche, stanno emergendo, tuttavia, anche evidenze circa gli effetti che gli inquinanti atmosferici possono avere sulla salute umana. Se si prende in considerazione, per esempio, l'esposizione alle polveri dell'aria (PM, Particulate Matter), si è visto che questa è associata ad un incremento della mortalità per malattie cardiovascolari ed un aumento del rischio di tumore ai polmoni[325,326]. Uno studio condotto su un gruppo di lavoratori in una acciaieria esposti al PM_{10} (polveri con diametro <10 μm), ha messo in luce che l'esposizione a lungo termine a tale inquinante porta a una variazione nello stato di metilazione del genoma ed in particolare, si è

messa in luce una ipometilazione degli elementi ripetuti LINE-1 ed *Alu*[327,328], ed anche una ipometilazione del gene iNOS (inducible Nitric Oxide Syntase). Ciò ha permesso di definire, pertanto, che il PM è in grado di alterare la metilazione del DNA (in particolare lo studio mette in evidenza una ipometilazione), e ha portato gli autori a concludere che tale effetto potrebbe essere alla base degli alterati profili di espressione genica riscontrati a seguito dell'esposizione alle polveri dell'aria, dato che non sono ancora ben noti i meccanismi attraverso cui il PM provoca tali alterazioni.

Oltre alla ipometilazione degli elementi ripetuti riportata da Tarantini e collaboratori, un ulteriore recente studio ha messo in luce che anche l'esposizione al BC (*Black Carbon*), un marcatore dell'inquinamento del traffico, porta alla ipometilazione degli elementi L1[329], misurata sul DNA di 1097 campioni di sangue provenienti da misurazioni ripetute condotte su un gruppo di uomini anziani, nell'area della città di Boston.

Pertanto, dato che la ipometilazione globale del DNA è stata riscontrata in pazienti affetti dal cancro[330] o malattie cardiovascolari[331], i cambiamenti osservati potrebbero riprodurre i processi connessi allo sviluppo delle patologie e rappresentare i maccanismi attraverso cui le polveri dell'aria possono influenzare la salute umana[326].

Le evidenze riportate sono un importante esempio di esposizione ambientale che potrebbe dereprimere la parte più abbondante del genoma. Ricordando, infatti, che tali elementi sono normalmente silenziati per ipermetilazione delle loro sequenze, tale riduzione nello stato di metilazione, potrebbe essere alla base della loro attivazione e dell'instabilità genomica conseguente, instabilità generalmente riscontrata in alcune forme patologiche. La ipometilazione degli elementi ripetuti, pertanto, potrebbe rappresentare uno dei meccanismi epigenetici alla base degli effetti nocivi dell'esposizione alle polveri dell'aria (compresi gli ioni metallici che ne fanno parte), aggiungendo così un ulteriore livello al quale l'inquinamento atmosferico è in grado di influenzare la salute umana.

Oltre agli effetti registrati per esposizione ai metalli e alle polveri dell'aria, di rilievo sono anche gli effetti dell'esposizione al benzene, un inquinante atmosferico molto diffuso, al quale verrà dedicata la trattazione successiva.

Il benzene: metabolismo, tossicità e carcinogenesi

Il benzene è un inquinante atmosferico ampiamente utilizzato nelle lavorazioni chimiche di sintesi, nei laboratori di analisi, nella preparazione e distribuzione dei distillati petroliferi. Esso viene definito inquinante ubiquitario in quanto non solo risultano esposti gli individui che esercitano determinate professioni, ma, dal momento che si trova nei gas di scarico delle autovetture, nei carburanti in concentrazioni pari all'1% , una ridotta frazione nel fumo di sigarette, e viene anche liberato dalla combustione della legna e di altri materiali organici (per esempio dalle eruzioni vulcaniche), tutti noi ne siamo più o meno esposti quotidianamente[332]. Risalgono agli anni '20 del secolo scorso i primi casi di associazione del benzene a varie forme tumorali, nel 1964 sono stati riportati casi di aumentato rischio di insorgenza di leucemia in lavoratori in industrie dedicate alla produzione di scarpe in cuoio, e negli anni '80 sono diventati sempre più numerosi i casi di insorgenza di leucemia tra gli operai di industrie metallurgiche. Tale idrocarburo è stato dichiarato cancerogeno per la salute umana (gruppo 1) dalla IARC (International Agency for Research on Cancer) nel 1982 e, numerosi casi riportati negli ultimi decenni, fanno sempre più associare l'esposizione al benzene con un aumento del rischio di insorgenza di una particolare forma di leucemia: la leucemia mieloide acuta (AML, Acute Myeloid Leukemia)[333]. Nonostante numerosi studi condotti sia su animali modello sia sull'uomo (in particolare quelli condotti negli ambienti di lavoro), non si sono, però, ancora definiti con precisione i meccanismi molecolari che permettono di mettere in relazione il benzene con tale patologia. Per la sua volatilità, il benzene è rapidamente assorbito dall'organismo per inalazione, e solo una piccola frazione viene eliminata con l'esalazione e con le urine, mentre la restante parte viene rapidamente distribuita ed è stata trovata in numerosi organi incluso il midollo osseo. Il benzene è trasformato nel fegato, dal sistema delle monoossigenasi del citocromo microsomiale P450-2E1 (CYP2E1), in ossido di benzene[334], il quale può subire riarrangiamenti non enzimatici e dare origine al fenolo. Il fenolo può essere, a sua volta, idrossilato (sempre dal CYP2E1) per formare catecolo, idrochinone e 1,2,4-triidrossibenzene. In seguito, idrochinone e catecolo sono trasportati dal fegato al midollo osseo, attraverso il flusso sanguigno, dove vengono ossidati dalle mieloperossidasi MPO (presenti ad alti livelli nel midollo osseo) negli intermedi reattivi orto-benzochinone e para-benzochinone, i quali sembrano essere i metaboliti maggiormente genotossici e leuchemogenici del benzene[335]. L'idrochinone e il catecolo possono anche essere sintetizzati a partire da chinoni derivanti dal benzene, in reazioni catalizzate dagli enzimi NAD(P)H chinone ossidoreduttasi (NQO1)[336]. Il metabolismo del benzene prevede anche la produzione di molti altri metaboliti (**Figura 11**), ma al fine della presente trattazione, i metaboliti più importanti sono idrochinone, catecolo e

benzochinone, i quali si pensa che più probabilmente siano responsabili della tossicità del benzene stesso e quindi fattore che contribuisce all'insorgenza della AML[332]. Il meccanismo attraverso il quale tale idrocarburo è in grado di indurre questa particolare forma di leucemia, come accennato sopra, sembra coinvolgere tali metaboliti fenolici ma, sebbene evidenze sperimentali indichino che questi sono necessari per la sua cancerogenicità, i meccanismi alla base degli effetti dell'esposizione al benzene non sono ancora stati compresi completamente[332,335].

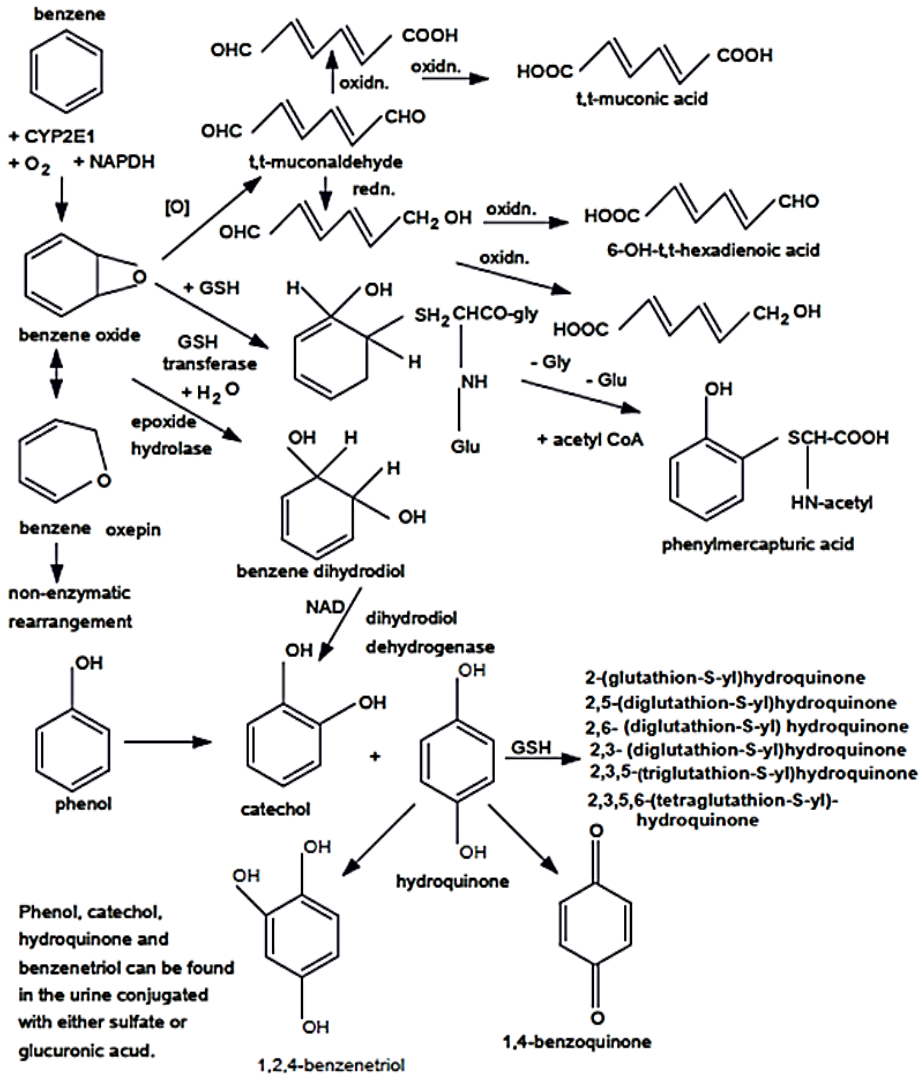


Figura 11. Rappresentazione schematica del metabolismo del benzene (Adattato da Snyder et al., 2012[335]).

Alterazioni genetiche ed epigenetiche associate all'esposizione al benzene

Come riportato per altri fattori ambientali (ioni metallici, PM_{10} e BC), studi *in vitro* ed *in vivo* hanno permesso di mettere in luce che anche il benzene è in grado di indurre danno ossidativo, nonché alterazioni nel normale svolgimento del ciclo cellulare, apoptosi, danno al DNA e rotture cromosomiche, ed è necessario che venga metabolizzato affinché possa esplicare la sua tossicità.

I metaboliti del benzene non sono mutagenici di per sé, ma sono in grado di generare specie reattive dell'ossigeno (ROS) che potrebbero creare danno al DNA sia attraverso meccanismi genetici che epigenetici. A tal proposito, vi sono numerose evidenze che tali derivati possano formare addotti covalenti con il DNA e indurre delezioni, traslocazioni, riarrangiamenti cromosomici (per esempio è stata riportata per l'idrochinone la aneusomia dei cromosomi 7 e 8 nelle cellule progenitrici ematopoietiche CD 34+[337]), e rotture a singolo e doppio filamento nel genoma stesso. Rivolgendo l'attenzione proprio all'idrochinone, il metabolita ritenuto maggiormente coinvolto nel meccanismo che porta alla citotossicità e alla genotossicità del benzene[332] (dal momento che la sua ossidazione porta a benzochinone) nonché il metabolita più abbondante[338], recenti studi *in vitro* hanno dimostrato come tale derivato sia in grado di indurre effetti citotossici quali aberrazioni cromosomiche, induzione di micronuclei, scambio di cromatidi fratelli, rotture nei filamenti di DNA e inibizione delle topoisomerasi II[332]. Tali alterazioni si pensa che potrebbero essere alla base della ematotossicità del benzene, realizzata mediante meccanismi genetici, ma negli ultimi decenni si sta sempre più esplorando la via dei cambiamenti epigenetici alla base delle leucemie e di molte altre patologie. Tali studi si basano sull'ipotesi, sempre più documentata, di un'importante correlazione tra l'esposizione a un fattore ambientale e cambiamenti epigenetici indipendenti dalle alterazioni nella sequenza del DNA.

Come riportato nella trattazione sui metalli e sugli inquinanti atmosferici, i cambiamenti epigenetici maggiormente studiati sono le alterazioni della metilazione del DNA e delle code istoniche[272,273]. Per quanto concerne le alterazioni dello stato di metilazione del DNA indotte dal benzene, un recente studio ha messo in luce che l'esposizione a bassi livelli di benzene è associata alla ipermetilazione del gene CDKN2B/p15 (Cyclin Dependent Kinase Inhibitor 2B) e alla ipometilazione del gene MAGE-1 (Melanoma Associated AntiGEn 1). Tali evidenze si sono ottenute mediante pirosequenziamento del DNA estratto dal sangue periferico di un gruppo di benzinai e vigili urbani, esposti quotidianamente al benzene[339], e hanno messo in luce che l'inquinante ubiquitario è in grado di indurre le medesime variazioni, nella metilazione dei promotori dei due geni, riscontrate nella leucemia mieloide acuta, avvalorando così un suo possibile coinvolgimento

nel rischio di insorgenza di tale patologia. Inoltre, Bollati e collaboratori hanno anche rivelato una ipometilazione dei retrotrasposoni umani L1 ed *Alu*[339] (tuttavia, maggiormente significativa per gli elementi L1) nei soggetti esposti, che ha portato gli autori a concludere che l'esposizione a bassi livelli di benzene potrebbe indurre alterazioni nella metilazione del DNA che sono stati riscontrati in diverse condizioni patologiche, compresi i tumori, soprattutto negli stadi avanzati[230,340,341]. Un ulteriore studio, ha portato a considerazioni analoghe circa la ipometilazione indotta dal benzene: anche in questo caso l'analisi del DNA, estratto dal sangue periferico di un gruppo di benzinai, mediante pirosequenziamento, ha permesso di rilevare la ipometilazione degli elementi ripetuti L1 appartenenti alla famiglia L1Pa2, ma, tuttavia non si sono ottenute variazioni significative nelle altre famiglie di elementi L1 e nelle tre famiglie di elementi *Alu* analizzate (*Alu*Yb8, *Alu*Sx e *Alu*d6)[342]. Tale studio aggiunge un tassello verso la conoscenza delle alterazioni dello stato di metilazione degli elementi ripetuti, in quanto sottolinea che si hanno differenti effetti a seconda della famiglia di provenienza di tali elementi. Gli autori, infatti, concludono che non tutte le famiglie di retrotrasposoni rispondono in modo concorde all'esposizione al benzene, e che ciò può essere determinante in quanto potrebbe aiutare a capire il ruolo degli elementi ripetuti nella risposta a fattori di rischio ambientale connessi con la salute umana e le malattie.

Infine, due studi condotti nell'ultimo anno, hanno messo anche in luce particolari variazioni della metilazione del DNA gene-specifiche. Nel primo, sono stati eseguiti trattamenti di ratti e della linea cellulare linfoblastoide F32 con diverse concentrazioni di benzene (ridotte, medie ed alte), che hanno portato a rilevare un progressivo calo dell'espressione del gene PTEN (Phosphatase and TENsin homolog deleted on chromosome 10), associato ad un aumento della metilazione del suo promotore. Tali evidenze hanno fatto concludere che il benzene potrebbe indurre un silenziamento epigenetico di tale gene, e che ciò potrebbe rappresentare uno dei meccanismi che causano la sua repressione. Inoltre, tali dati sono stati confermati dal trattamento della linea cellulare F32 con decitabina e TSA (l'inibitore delle HDAC), separatamente e insieme, e comunque dopo il trattamento con benzene, in quanto tali agenti hanno riattivato il gene PTEN mediante abrogazione della ipermetilazione indotta dal benzene. Essendo PTEN un soppressore tumorale, ed essendo ipermetilato dal benzene, gli autori concludono che le alterazioni epigenetiche potrebbero giocare un ruolo rilevante nella carcinogenesi del benzene (è il primo caso di down-regolazione del gene soppressore tumorale PTEN, nel contesto dell'ematotossicità del benzene, per variazione nel suo stato di metilazione[343]).

Infine, il secondo lavoro rivela che l'esposizione cronica al benzene provoca una ipermetilazione dei geni PRKG1 (Protein Kinase cGMP-dependent, type 1), PARD3 (Par-3 Partitioning Defective 3

homolog) e EPHA8 (EPH receptor A8) e conseguente diminuzione dei loro livelli di espressione, nonché una ipometilazione dei geni STAT3 (Signal Transducer and activator of transcription 3) e IFNGR1 (Interferon Gamma Receptor 1) con conseguente aumento dei loro trascritti, nel sangue periferico estratto da cellule monocellulari. Concentrando particolarmente l'attenzione sul gene STAT3, essendo coinvolto nel mantenimento delle caratteristiche dei tumori maligni[344] e costitutivamente attivato nella leucemia linfocitica cronica (CLL)[345], Yang e collaboratori concludono che la sua aberrante ipometilazione possa essere un potenziale biomarcatore dell'esposizione cronica al benzene[346].

Dal momento che sono stati osservati cambiamenti nello stato di metilazione dei geni codificanti per proteine ed elementi ripetuti L1, sia per esposizione a diversi fattori ambientali, sia per esposizione al benzene, nonché un calo della metilazione globale del DNA per esposizione all'idrochinone *in vitro*[347], seppur ad alti livelli di esposizione (non paragonabili con l'esposizione umana), si pensa che i cambiamenti epigenetici siano una risposta all'influenza ambientale degna della massima considerazione.

Dal momento che tali variazioni sono state riscontrate in pazienti affetti da AML, e che sono state osservate anche a seguito di esposizione al benzene, si potrebbe ipotizzare che il benzene possa produrre nel midollo osseo (il suo tessuto bersaglio) non solo alterazioni genetiche ma anche epigenetiche in grado di deregolare il sistema ematopoietico (come osservato per il gene PTEN). Tale deregolazione potrebbe spiegare l'alterata espressione di numerosi geni associati con gli effetti ematotossici nella AML[348], e con le sempre maggiori evidenze di leucemia causata dall'esposizione al benzene[349,350]. Per questi motivi la leucemia, conseguente all'esposizione prolungata al benzene, potrebbe essere un'ottima "malattia-modello" per indagare le interazioni tra geni e ambiente e i cambiamenti molecolari che promuovono o conducono allo sviluppo dei tumori nell'uomo[351].

Scopo della ricerca

Gli elementi trasponibili costituiscono circa il 45% del genoma umano e solo la classe dei retrotrasposoni non-LTR risulta attiva oggigiorno dal punto di vista della retrotrasposizione. Fanno parte di questa classe, gli elementi LINE-1 ed *Alu*, i quali si inseriscono in un locus genomico diverso da quello di origine, duplicandosi attraverso la generazione di un intermedio ad RNA[6], aumentando così notevolmente il loro numero di copie genomiche (circa 1 milione per gli elementi *Alu* e 500 mila per gli elementi L1). Inizialmente considerati DNA "spazzatura", ovvero una porzione di DNA priva di qualsiasi funzione, attualmente essi sono stati progressivamente rivalutati, innanzitutto in quanto in grado di contribuire all'evoluzione dei genomi in molteplici modi, in secondo luogo anche in quanto in grado di influenzare numerosi processi cellulari in modi ancora da esplorare. Il fatto che tali elementi si integrino in nuovi siti genomici, fa sì che essi possano generare instabilità sia mediante mutagenesi inserzionale sia attraverso eventi di ricombinazione omologa non allelica. Inoltre gli RNA stessi, prodotti dalla trascrizione dei retrotrasposoni, possono anche interferire con i normali processi cellulari[200]. Analizzare gli elementi ripetuti, in quanto parte del genoma non ancora ben esplorata ma con ruoli emergenti sempre più associati allo sviluppo delle malattie[340], è quindi di rilievo per capire in dettaglio come sia regolata la loro espressione e le conseguenze di quest'ultima nel contesto cellulare. Inoltre, dal momento che vi sono alcune evidenze di una derepressione trascrizionale degli elementi *Alu* correlata alla loro ipometilazione[69,352] e che tale ipometilazione è presente nelle cellule tumorali[200], si può supporre che uno dei meccanismi alla base della attivazione sia proprio una variazione nel loro stato di metilazione (si ricorda che normalmente gli elementi *Alu* e L1 sono ipermetilati e silenti) e che la loro espressione sia da considerarsi come un importante, ma ancora inesplorato, marcatore dello stato tumorale (sono riportate alcune evidenze anche della ipometilazione di elementi L1 in cellule tumorali[230,353]).

Ciò risulta di particolare rilievo se si considera che l'esposizione prolungata a differenti fattori ambientali, noti per essere legati alla comparsa di patologie, comporta anche una alterata metilazione (in genere ipometilazione) di elementi *Alu* e L1[322,328,339]: questo avvalora l'ipotesi di una loro attivazione mediante abrogazione del silenziamento genico, e di un loro conseguente possibile coinvolgimento nella patogenesi.

La derepressione di retrotrasposoni *Alu* e L1 in risposta a fattori ambientali rappresenta, pertanto, un fenomeno degno della massima attenzione.

Nell'ambito della presente tesi di dottorato si è innanzitutto concentrata l'attenzione sull'identificazione di elementi *Alu* trascritti autonomamente dalla RNA polimerasi III mediante analisi dei dati di RNA-sequencing ottenuti all'interno del progetto ENCODE, il più ambizioso progetto di sequenziamento dei trascrittomi di molte linee cellulari umane. Si voleva, infatti, determinare l'esistenza di un sottogruppo di elementi espressi in un contesto cellulare rispetto agli altri, in modo da poterlo utilizzare come possibile indicatore di un determinato stato cellulare e, in ultima analisi, essendo l'espressione genica associata ad uno specifico assetto epigenetico, delle alterazioni dell'epigenoma cellulare. Una volta identificato un potenziale sottogruppo di retroelementi espressi, si sono indagati i meccanismi molecolari che possono determinare la sua attivazione, nonchè gli effetti che tale attivazione può avere sul trascrittoma cellulare, dal momento che stanno emergendo evidenze di come l'espressione, e il conseguente accumulo, di *Alu* RNA può interferire con i processi cellulari fisiologici. Inoltre, si è rivolto lo sguardo verso l'influenza che i fattori ambientali, ai quali l'uomo è quotidianamente esposto, possono esercitare sull'espressione degli elementi ripetuti *Alu* e LINE-1, incuriositi dal fatto che l'esposizione ad alcuni agenti è stata associata ad una ipometilazione di tali elementi normalmente silenti (e quindi metilati), che quindi potrebbe essere interpretata come potenziale causa di una loro derepressione. L'analisi degli effetti che i fattori ambientali possono avere sulla modulazione dell'espressione dei retrotrasposoni si è ritenuta necessaria per provare a definire un possibile meccanismo alla base della loro tossicità e, in ultima analisi, all'insorgenza delle patologie, con la speranza di poterlo prevenire mediante l'impiego di opportuni accorgimenti. Infine, nell'ambito di un progetto collaterale, si è eseguita un'indagine volta a mettere in luce possibili correlazioni funzionali tra RNA non codificanti (ncRNA) trascritti dalla Pol III umani e murini, al fine di poter creare modelli sperimentali di topo in cui studiare le unità trascrizionali della RNA polimerasi III.

Parte I

Identificazione di elementi *Alu* trascritti dalla RNA polimerasi III mediante analisi di dati di RNA-seq

Come descritto ampiamente nell'introduzione, gli elementi *Alu* sono molto simili in sequenza, molto abbondanti a livello genomico, ma poco rappresentati come trascritti, ed è importante sottolineare che esistono sia trascritti *Alu* autonomi (sintetizzati dall'enzima RNA polimerasi III) sia interni a trascritti più lunghi (sintetizzati dall'enzima RNA polimerasi II). Il fatto che gli *Alu* RNA possano essere sintetizzati sia a partire da unità trascrizionali indipendenti sia come parte di trascritti più lunghi di unità trascrizionali ospiti, e che le due categorie di trascritti siano molto simili in sequenza, ha reso difficile finora l'individuazione dei singoli loci *Alu* autonomamente trascritti.

Tuttavia, grazie all'avvento di nuove tecnologie di sequenziamento, in particolare dell'RNA-sequencing (RNA-seq), si sono aperte nuove possibilità per il superamento di questo problema, essendo diventata possibile l'identificazione di trascritti *Alu* autonomi, gli unici attivi dal punto di vista della retrotrasposizione (si ricorda che gli *Alu* RNA, parte di trascritti più lunghi, non hanno alcun potenziale retrotraspositivo[54,354]).

Gli *Alu* RNA sono non codificanti per proteine e in grado di regolare processi cellulari, come la trascrizione e la traduzione[200], e tali evidenze hanno portato a rivalutare la loro importanza, e a non considerare più gli elementi genetici per essi codificanti come DNA "spazzatura", come è stato per molto tempo.

Inoltre, grazie al loro carattere non codificante, alla loro lunghezza e al loro numero di copie nel genoma, gli *Alu* RNA possono essere considerati una delle classi più numerose di long non-coding RNA (lncRNA). Negli ultimi anni si stanno studiando molti membri dell'eterogenea famiglia dei lncRNA, e si è visto che essi possono influenzare l'espressione genica a vari livelli (dalle modificazioni della cromatina al processamento post-trascrizionale); in particolare, si pensa che i lncRNA possano contribuire ai diversi stadi della progressione del cancro e, quindi, il loro studio risulta potenzialmente rilevante per il riconoscimento e trattamento di diverse forme tumorali[355]. Il fatto che tali ncRNA siano visti come possibili marcatori di riconoscimento del cancro e che gli *Alu* RNA siano RNA non codificanti, ci ha permesso di ipotizzare che tali trascritti possano potenzialmente fungere da marcatori di diversi stati cellulari, con la possibilità di identificare profili di espressione di determinati *Alu* RNA associati a specifiche patologie.

Risultati e discussione

Analisi delle sequenze di *Alu* RNA trascritti dalla RNA polimerasi III

Come accennato nella parte introduttiva della sezione, gli *Alu* RNA sono molto simili in sequenza e possono sia essere trascritti come elementi indipendenti dalla Pol III, sia come parte di trascritti più lunghi dalla Pol II. Dal momento che solo la prima categoria di trascritti è attiva dal punto di vista della retrotrasposizione (portando a instabilità genomica mediante mutagenesi inserzionale o ricombinazione omologa non allelica), sono stati condotti numerosi studi volti a definire con precisione i trascritti autonomi e i corrispondenti loci genomici di provenienza e, per far ciò, inizialmente ci si è avvalsi di tecniche quali il Northern blot e la RT-PCR, ma con scarsi risultati. In seguito, grazie alle tecniche di frazionamento da gel, primer extension e 3'-RACE (Rapid Amplification of cDNA Ends)[63,224], si sono cominciate ad avere le prime informazioni sulle sequenze di elementi *Alu* autonomi, ma solo con l'avvento delle nuove tecnologie di sequenziamento (NGS, Next Generation Sequencing), si è aperta la strada verso l'identificazione di singoli loci *Alu* trascrizionalmente attivi, ed in particolare, attraverso gli approcci di ChIP-seq (immunoprecipitazione cromatinica seguita da sequenziamento massivo), è stato possibile rilevare l'associazione *in vivo* del macchinario trascrizionale della Pol III ad elementi *Alu*, un'indicazione ragionevole del fatto che potessero essere trascritti[356-358]. Analizzando i dati di ChIP-seq, infatti, si è confermato che probabilmente solo una piccola frazione di loci *Alu* è effettivamente espressa (in accordo con le precedenti evidenze dei bassissimi livelli di espressione di *Alu* RNA nelle cellule[200]) e che i profili di espressione possono variare a seconda del tipo cellulare, dello stadio di trasformazione cellulare, nonché della risposta a piccole variazioni nelle condizioni di crescita[359] (è stato riportato recentemente, infatti, che i trascritti che hanno origine da regioni ripetute (compresi gli elementi LINEs e SINEs), presentano un'espressione tipo cellulare-specifica [125]). Oltre ai dati di ChIP-seq menzionati, anche i sequenziamenti del trascrittoma cellulare, mediante RNA-seq, potrebbero permettere di far luce sulla porzione di elementi *Alu* trascritti autonomamente, in quanto sono in grado di ovviare al problema dei bassi livelli di elementi espressi e della impossibilità di discriminare i trascritti autonomi dai non autonomi, nonché entrambe le categorie dal 7SL RNA da cui derivano (tutti i trascritti sono molto simili in sequenza e difficili da distinguere con tecniche sperimentali meno sofisticate).

È stata quindi intrapresa un'analisi volta ad identificare elementi *Alu* autonomamente espressi tenendo innanzitutto in considerazione che gli *Alu* RNA, sintetizzati dalla Pol III, nonostante siano molto simili, presentano anche alcune caratteristiche distintive nelle loro sequenze. Tali

caratteristiche distintive comprendono le differenti mutazioni accumulate nella sequenza codificante, la lunghezza ed eterogeneità della coda poli-A terminale, e la sequenza unica all'estremità 3' (detta *3'trailer*). Tale sequenza corrisponde alla regione genomica, fiancheggiante il retrotrasposone, che si estende dall'estremità 3' del corpo dell'elemento *Alu* al terminatore della trascrizione (canonico o non canonico[53]) e che può avere, in base alla collocazione di tale terminatore, una diversa lunghezza (v. **Figura 4** nell'Introduzione), portando alla sintesi di trascritti *Alu* autonomi con lunghezza compresa fra 300 nt e 600 nt circa[72]. Inoltre, in base al sito genomico di inserzione, la sequenza 3'trailer risulta specifica per un determinato elemento *Alu* e permette di distinguerlo in modo univoco dagli altri. A tal proposito, si può quindi affermare che, nonostante la generale somiglianza dei trascritti *Alu*, tenendo in considerazione le caratteristiche di sequenza peculiari di ciascun elemento, si può in linea di principio risalire al suo locus genomico di provenienza ed essere in grado di analizzare solo *Alu* trascritti dalla Pol III. Dal momento che, quindi, per distinguere gli elementi *Alu* autonomi è necessario identificare queste piccole differenze in sequenza, l'unico modo possibile è sembrato essere l'analisi dei dati di RNA-seq, la tecnica che permette di avere l'intera sequenza del trascrittoma cellulare con una risoluzione a singolo nucleotide. Seguirà la trattazione della messa a punto di una strategia finalizzata proprio all'identificazione di elementi *Alu* autonomi espressi.

Strategia bioinformatica per l'identificazione di elementi *Alu*, trascritti dalla RNA polimerasi III, mediante analisi di dati di RNA-seq

Le considerazioni fatte nel paragrafo precedente, fanno facilmente intuire che per distinguere gli *Alu* RNA autonomi da quelli ospiti di trascritti più lunghi, è necessario ricorrere a una tecnica sofisticata in grado di permettere l'analisi delle sequenze dei trascritti. Tale tecnica, detta RNA-sequencing, fa parte delle metodiche di sequenziamento di nuova generazione (NGS) e fornisce le informazioni sulla sequenza di tutti i trascritti cellulari con una risoluzione a singolo nucleotide. Per identificare, quindi, loci *Alu* singoli e trascrizionalmente attivi, si è deciso di analizzare i dati grezzi di RNA-seq ottenuti all'interno del progetto ENCODE, il più ambizioso progetto di sequenziamento dei trascrittomi di molte linee cellulari umane[125].

È stata condotta, pertanto, un'analisi *in silico* che ha previsto l'utilizzo di tali dati grezzi di RNA-seq, in particolare i dati del Cold Spring Harbor Laboratory (CSHL) relativi a sette differenti tipi cellulari, sia normali che tumorali: Gm12878 (cellule linfoblastoidi normali), H1-hESC (cellule staminali embrioniche umane), HeLa S3 (cellule da adenocarcinoma della cervice), HepG2 (cellule da carcinoma epatocellulare), HUVEC (cellule endoteliali da cordone ombelicale), K562 (cellule da

leucemia mieloide cronica), NHEK (cheratinociti epidermici normali). Le analisi sono state condotte sui dati corrispondenti ai trascritti dell'intera cellula divisi in poli-A+ e poli-A- (i trascritti sono nel formato di sequenze di 76 nucleotidi dette reads) per entrambi i replicati dell'esperimento (per ogni linea cellulare si avevano sequenziamenti da due replicati biologici indipendenti)

Per ognuno dei dataset di trascritti ottenuti si è eseguito un allineamento con il genoma umano di riferimento (hg19), in modo da determinare i loci di provenienza dei vari RNA, e si è scelto di considerare come espressi solo gli elementi *Alu* che fossero coperti da un numero di reads, mappati in modo univoco sul genoma, maggiore o uguale a 10. Così facendo si sono ottenute le coordinate dei loci corrispondenti ad elementi *Alu* espressi, i quali sono stati suddivisi in base alla loro localizzazione genomica in: (i) *Alu* intergenici/antisenso, ovvero presenti in regioni tra geni codificanti per proteine o per lincRNA (long intergenic non-coding RNA), oppure in introni/esoni di geni ma trascritti in orientamento opposto; (ii) *Alu* intronici senso, presenti in introni di geni per proteine o lincRNA trascritti nello stesso orientamento; (iii) *Alu* presenti nelle 5'UTR/3'UTR di geni per proteine o lincRNA trascritti nello stesso orientamento (5'UTR/3'UTR senso); e infine (iv) *Alu* presenti in esoni di geni per proteine o lincRNA (interamente o parzialmente sovrapposte a tali esoni) trascritti nello stesso orientamento (**Figura 12**).

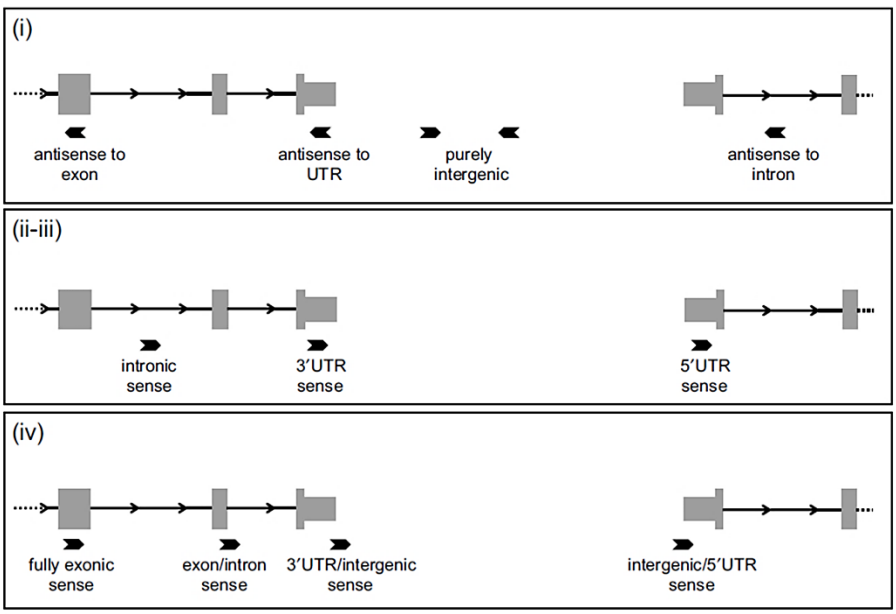


Figura 12. Rappresentazione delle differenti localizzazioni genomiche degli elementi *Alu* espressi identificati secondo la nostra strategia, a partire dai dati grezzi di RNA-seq depositati nella banca dati online di ENCODE. Le localizzazioni sono intese rispetto ad altre unità trascrizionali (geni codificanti per proteine o per lincRNA).

Dal momento che i trascritti degli elementi appartenenti ai gruppi (ii), (iii) e (iv) possono essere sintetizzati dalla Pol II (mentre per gli elementi del gruppo (i) è probabile una trascrizione da parte della Pol III, come unità trascrizionali indipendenti), si è eseguita la ricostruzione dei profili di espressione di singoli elementi *Alu* mediante risoluzione a singolo nulceotide utilizzando reads che mappano una sola volta nel genoma, e si sono scartati i trascritti con un numero rilevante di reads associati a monte e a valle degli elementi in esame (evidenze di trascrizione a monte e a valle degli elementi *Alu* sono indice del fatto che facciano parte di trascritti di maggiori dimensioni, sintetizzati dalla Pol II). Così facendo, si è riusciti a identificare solo i trascritti indipendenti e a distinguerli da trascritti ospiti di unità trascrizionali più lunghe.

La strategia bioinformatica messa a punto ha permesso, quindi, di rilevare per la prima volta trascritti *Alu* autonomi a partire dai dati grezzi di RNA-seq (per ognuna delle sette linee cellulari analizzate), di risalire alla loro localizzazione genomica, e quindi, in ultima analisi, di identificare con estrema precisione i trascrittomi di elementi *Alu* di numerose linee cellulari. Pertanto, si è messo in luce il fatto che un'appropriata analisi dei dati di RNA-seq, che contengono tutte le informazioni delle sequenze dei trascritti cellulari (anche quelli meno abbondanti, come gli *Alu* RNA), dovrebbe permettere di superare le difficoltà nell'identificazione dei trascritti *Alu*, da sempre stata ostica per la somiglianza degli RNA sia in termini di sequenza che di lunghezza. La strategia riportata ha previsto l'applicazione di filtri stringenti per identificare singoli elementi *Alu* trascritti, non presenti in molteplici copie.

Caratteristiche generali che emergono dall'analisi degli *Alu* RNA espressi individuati

La strategia bioinformatica messa a punto ha permesso di osservare che la maggior parte degli *Alu* RNA identificati come autonomi e trascrizionalmente attivi, è presente nella frazione poli-A- dell'RNA cellulare, rispetto alla frazione poli-A+ (968 *Alu* RNA individuati nel poli-A- e 394 *Alu* RNA nel poli-A+, rispettivamente), e che i trascritti sintetizzati a partire da 67 loci genomici sono presenti in entrambe le frazioni di RNA cellulare. Tali evidenze mettono in luce il fatto che, nonostante i trascritti *Alu* abbiano una regione ricca in A al loro interno (che separa il braccio sinistro dal braccio destro) e una lunga coda poli-A all'estremità 3', nella maggior parte dei casi esse non siano sufficientemente lunghe da permetterne l'inclusione nell'RNA poli-A+. Inoltre, si è osservato che in ogni linea cellulare, delle sette analizzate, il numero di elementi *Alu* espressi è molto basso (da un minimo di 149 nella linea cellulare Gm12878 a un massimo di 425 nella linea cellulare HepG2) (**Tabella 1**) e, dal momento che la strategia si è rivelata particolarmente attendibile nell'identificare con estrema precisione elementi *Alu* intergenici/antisenso (ovvero

puramente intergenici o che si sovrappongono ai geni annotati (in loro introni o esoni), ma in orientamento antisenso), trascritti autonomamente dalla Pol III, si è deciso di concentrare su tali RNA le successive analisi.

L'indagine più approfondita condotta su tali elementi, che rappresentano circa il 30% dei loci *Alu* risultati attivamente trascritti (del set di 1295 elementi *Alu* risultati espressi con la nostra strategia), ha evidenziato che circa il 22% di questi è situato negli introni di geni trascritti dalla Pol II (codificanti per proteine e per lincRNA) in orientamento antisenso, che il 20% è trascritto in più di un tipo cellulare, e che gli elementi espressi in una linea cellulare, lo sono anche in almeno un'altra (osservato per circa il 40% degli elementi espressi) (**Tabella 1**). Le osservazioni riportate permettono di mettere in luce il fatto che, nonostante l'elevato numero di copie di elementi *Alu* nel genoma, solo una piccola frazione viene effettivamente espressa e corrisponde a circa lo 0.1%, di tutti i loci *Alu* annotati, per ognuno dei tipi cellulari analizzati. Inoltre, dal momento che nelle diverse linee cellulari si ha l'espressione di un piccolo gruppo di elementi *Alu* intergenici/antisenso comune, viene richiamata la nozione di unità trascrizionali Pol III "sorgente" in grado di contribuire all'espansione degli elementi *Alu* mediante la retrotrasposizione[54] (v. Introduzione). Nonostante la presenza di tale gruppo di elementi *Alu* comune alle diverse linee cellulari, si è anche sottolineato che per ogni tipo cellulare è possibile distinguere un gruppo caratteristico di elementi *Alu* espressi, suggerendo che gli elementi *Alu*, espressi in un contesto cellulare, sono il prodotto della trascrizione sia di unità comunemente espresse che tipo cellulare-specifiche. Il fatto che esista anche un sottoinsieme di elementi linea cellulare-specifico potrebbe essere rilevante per la definizione dei profili di *Alu* RNA nelle più diverse condizioni e in diversi tipi cellulari umani, e potrebbe essere dovuto ad una particolare sensibilità di questi elementi alla metilazione del DNA ed al contesto cromatinico, che sono senz'altro soggetti a variazioni da un tipo cellulare all'altro. Pertanto determinare i profili di *Alu* RNA, attraverso l'analisi dei dati di RNA-seq, potrebbe rappresentare un nuovo modo, estremamente fine e sensibile, di monitorare le alterazioni dell'epigenoma che accompagnano stati fisiologici e patologici.

Linea cellulare	*Alu* totali	Intergenici/antisenso	Intergenici/antisenso in comune	Antisenso negli introni
Gm12878	149	48	21	13
H1-hESC	257	92	28	12
HeLa S3	276	44	32	7
HepG2	425	88	31	19
HUVEC	326	36	18	4
K562	154	71	33	20
NHEK	231	130	38	34
ALL	1295	386	78	87

Tabella 1. Nella tabella sono riassunti gli elementi *Alu* individuati come espressi secondo la nostra strategia bioinformatica. Nella colonna 1 sono riportate le linee cellulari analizzate, nella colonna 2 è messo in evidenza il ridotto numero di trascritti espressi per linea cellulare (per ogni linea cellulare è riportato il numero di elementi *Alu* risultati espressi autonomamente in entrambi i replicati degli RNA-seq analizzati). Nella colonna 3 è riportato il numero di elementi *Alu* intergenici/antisenso espressi per linea cellulare. Nella colonna 4 sono riportati gli elementi intergenici/antisenso espressi in uno o più tipi cellulari differenti. Nella colonna 5 sono riportati gli elementi *Alu* espressi, con localizzazione in orientamento antisenso negli introni di geni codificanti per proteine e geni per lincRNA. ALL è da intendersi come elementi *Alu* espressi autonomamente in una o più linee cellulari.

Infine, un'analisi volta a valutare se gli elementi *Alu*, espressi in un determinato tipo cellulare, appartenessero a specifiche famiglie (si ricorda che gli elementi *Alu* possono appartenere a 3 famiglie: *Alu*J, *Alu*S ed *Alu*Y), ha messo in luce che non vi è un'espressione preferenziale degli elementi di una famiglia rispetto alle altre (**Tabella 2**).

Tabella 2. Per ogni famiglia di elementi *Alu* sono riportati il numero assoluto di elementi e la loro percentuale (tra parentesi), la quale è stata calcolata dividendo il numero di copie di ogni famiglia per il numero totale di copie di elementi *Alu* all'interno dei differenti set (dalla seconda alla quarta colonna: set di elementi *Alu* totali nel genoma; elementi nel genoma identificati mediante l'allineamento "unico" risultati espressi in un o più tipi cellulari; set degli elementi totali intergenici/antisenso; elementi intergenici/antisenso risultati espressi mediante l'allineamento "unico" in uno o più tipi cellulari). La categoria "intergenici/antisenso espressi (read count)" si riferisce al dataset di elementi *Alu* intergenici/antisenso risultati espressi mediante l'impiego di un filtro più permissivo alla nostra strategia bioinformatica, che ha previsto che TopHat aligner distribuisse i reads che mappavano in più siti genimici, in modo casuale a uno di tali siti.

Famiglia *Alu*	Elementi totali nel genoma	Elementi espressi nel genoma	intergenici/ antisenso totali	intergenici/ antisenso espressi (numero di copie)	intergenici/ antisenso espressi (read count)
***Alu*S**	675428 (60%)	735 (57%)	513048 (60%)	219 (57%)	62%
***Alu*J**	307612 (27%)	479 (37%)	225907 (27%)	112 (29%)	27%
***Alu*Y**	140707 (13%)	81 (6%)	107922 (13%)	55 (14%)	11%
Totale	1123747	1295	846877	386	100%

Analisi degli elementi *Alu i*ntergenici espressi

L'analisi dei soli trascritti *Alu* intergenici, identificati con la strategia bioinformatica messa a punto, ha permesso di sottolineare che nelle linee cellulari esaminate, non sono solo sintetizzati RNA completi (full-length), ma anche numerosi trascritti, per così dire, tronchi, ovvero corrispondenti al solo braccio sinistro o destro di un normale elemento *Alu* dimerico.

Per quanto concerne gli *Alu* RNA full-length, essi sono stati messi in evidenza analizzando i loro profili di espressione, i quali presentavano segnali corrispondenti all'intera lunghezza del trascritto. In **Figura 13** è riportato un esempio di *Alu* RNA completo: oltre al profilo di espressione, l'analisi della sua sequenza ha rivelato la presenza di tutti gli elementi caratteristici di loci *Alu* trascritti dalla Pol III (si rimanda all'Introduzione, **Figura 4**), portando a concludere che si tratta effettivamente di un RNA autonomo completo (**Figura 13**). Inoltre si è osservata l'espressione di tale trascritto nelle cellule K562 e NHEK ma non nelle cellule HUVEC: ciò mette in luce l'espressione differenziale di *Alu* RNA da un tipo cellulare all'altro, nonché la presenza di alcuni trascritti in comune a più tipi cellulari.

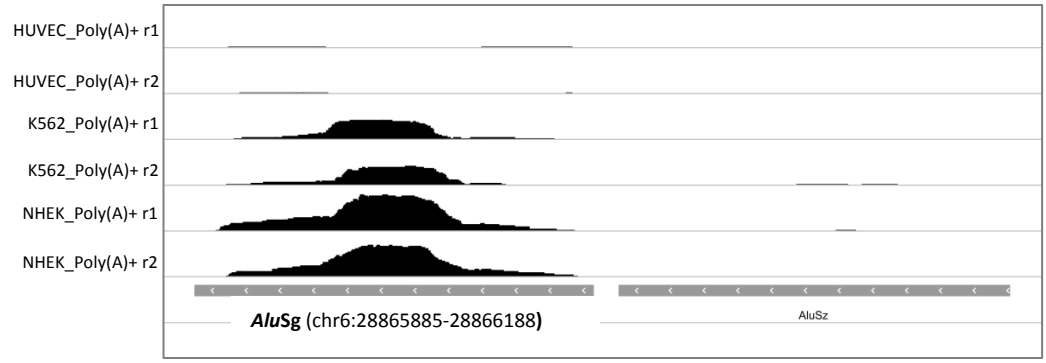


Figura 13. Profilo di espressione di un elemento *Alu*Sg intergenico risultato espresso secondo la nostra strategia bioinformatica, a partire dai dati grezzi di RNA-seq depositati nella banca dati online di ENCODE. Tale elemento è presente nella frazione di RNA poli-A+ (Poly(A)+ in figura)in tre tipi cellulari (HUVEC, K562 e NHEK) e presenta la copertura di reads per l'intera sequenza.

Al contrario, l'osservazione di altri trascritti di elementi intergenici, ha messo in luce l'esistenza nelle cellule di numerosi RNA tronchi, formati o dal solo braccio sinistro (**Figura 14A**) o dal solo braccio destro (**Figura 14B**), confermata anche dall'analisi delle loro sequenze.

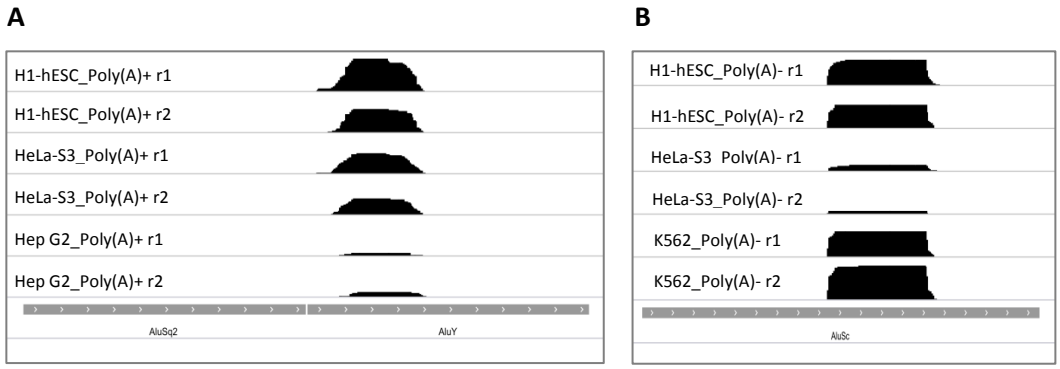


Figura 14. Profilo di espressione di un elemento *Alu*Y intergenico formato dal solo braccio sinistro **(A)** e di un elemento *Alu*Sc formato dal solo braccio destro **(B).** Tali elementi sono presente in tre tipi cellulari ciascuno e in differenti frazioni di RNA cellulare.

Tali trascritti, di cui sono riportati due esempi in **Figura 14**, possono essere interpretati come il prodotto del processamento di *Alu* RNA completi. In particolare il trascritto *Alu*Y in **Figura 14A** può essere considerato un esempio di sc*Alu*, ovvero il prodotto di trascritti *Alu* primari processati (presenti nel citoplasma cellulare e formati dal solo braccio sinistro) di cui sono riportati numerosi esempi in letteratura[360], mentre il trascritto *Alu*Sc in **Figura 14B** presenta solo il braccio destro ed è di più difficile interpretazione, in quanto non vi sono, ad oggi, evidenze di trascritti *Alu* formati dal

solo braccio destro. Tuttavia, la mancanza di RNA completo, potrebbe comunque suggerire che sia anch'esso il prodotto di un processamento dell'RNA primario e, nel complesso, l'identificazione di questi trascritti mette in luce un'intensa modificazione post-trascrizionale dei trascritti degli elementi *Alu* all'interno delle cellule.

Proseguendo l'indagine sui trascritti intergenici espressi individuati, si è anche messo in evidenza che, nella maggior parte dei casi, i profili di espressione non sono continui, ma si ha una struttura "a due picchi" separati, di cui è riportato l'esempio relativo all'elemento *Alu*Y in orientamento antisenso nel secondo introne del gene COL4A1 (**Figura 15**). L'analisi della sequenza di tale elemento, ha permesso di attribuire i due picchi al braccio sinistro e al braccio destro. La mancanza di segnali nella parte centrale potrebbe essere dovuta alla procedura che è stata eseguita per produrre i dati di RNA-seq. Infatti, dal momento che un elemento *Alu* canonico è lungo circa 300 nt e che dopo la frammentazione dell'RNA sono selezionati solo frammenti di 200 bp, per la preparazione della libreria di cDNA che viene successivamente sequenziata, tale selezione potrebbe portare ad un arricchimento in frammenti corrispondenti o alle estremità 5' o 3' di cDNA *Alu*. Il sequenziamento di tali estremità potrebbe quindi generare, in ultima analisi, una porzione centrale scarsamente rappresentata con conseguente mancanza del corrispondente segnale nel profilo di espressione.

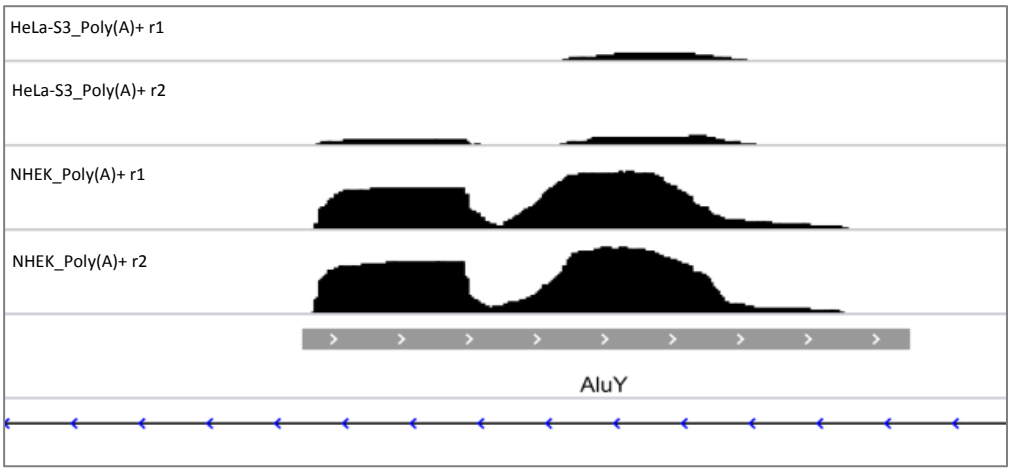


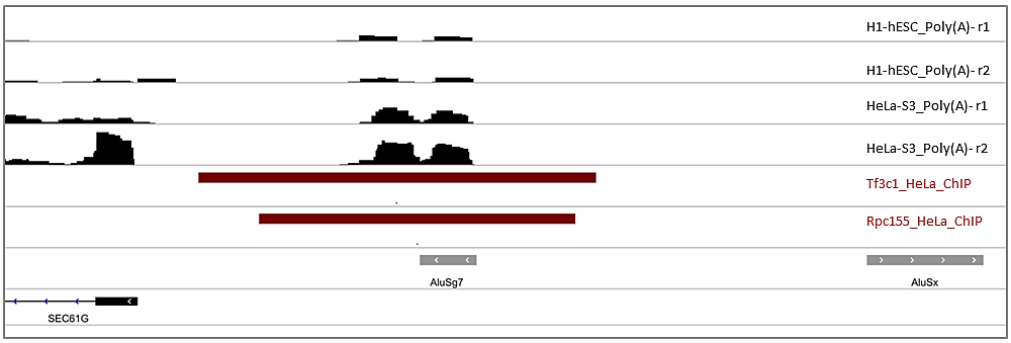
Figura 15. Profilo di espressione "a due picchi" separati di un elemento *Alu*Y intergenico.

Infine, da aggiungere alle tipologie di RNA intergenici menzionati finora, vi sono i trascritti formati o dal braccio sinistro di un elemento *Alu* e dalla regione a valle di quest'ultimo (**Figura 16A**) oppure dal braccio destro e dalla regione a monte (**Figura 16B**). Tali evidenze sono molto interessanti ed in accordo col fatto che, nel primo caso, il promotore contenuto nel braccio sinistro permette la sua trascrizione e questa potrebbe continuare a valle dell'elemento, mentre nel secondo caso il braccio destro, privo di un promotore Pol III, potrebbe essere incorporato in un trascritto che inizia a monte. Per avvalorare questa ipotesi, si sono consultati i dati di ChIP-seq (sempre ottenuti nell'ambito del progetto ENCODE) e si è confermata l'associazione del fattore TFIIIC e della Pol III proprio in corrispondenza dell'elemento *Alu*Sg7 riportato in **Figura 17A** (un esempio dei trascritti *Alu* identificati, formati dal braccio sinistro e da una regione non-*Alu* a valle), confermando che esso può essere considerato a tutti gli effetti un'unità trascrizionale della Pol III.

La presenza di numerosi monomeri di elementi *Alu* espressi in differenti linee cellulari analizzate, ha permesso di constatare che probabilmente tale espressione è più comune di quanto si potesse immaginare ed, inoltre, è in accordo con recenti evidenze di inserzioni di monomeri *Alu*, rese possibili dalla retrotrasposizione dei corrispondenti

trascritti[360] (si ricorda che affinché si abbia la retrotrasposizione degli elementi *Alu*, è necessaria la loro trascrizione).

A



B

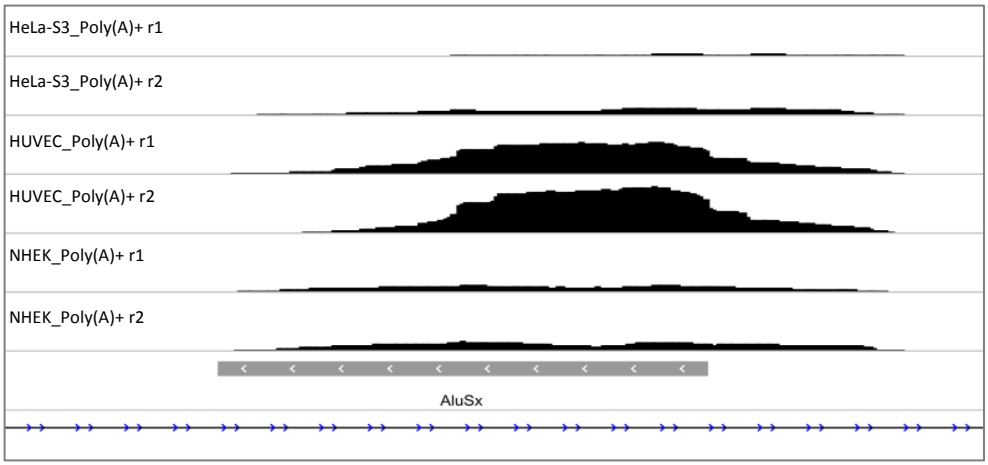


Figura 16. Profilo di espressione di un elemento *Alu* intergenico formato dal braccio sinistro e da una regione trascritta a valle **(A)** e dal braccio destro e da una regione trascritta a monte **(B)**.

Evidenze di elementi *Alu*, in unità trascrizionali più lunghe, trascritti autonomamente

Nonostante, una volta individuati i loci *Alu* autonomamente trascritti, l'attenzione sia stata rivolta maggiormente verso gli intergenici (per i quali non vi sono dubbi circa la trascrizione ad opera della Pol III), un'analisi degli elementi in orientamento senso in unità trascrizionali più lunghe (per lo più nei loro introni), ha permesso di constatare che, anche all'interno di questa categoria, alcuni potrebbero essere riconosciuti e trascritti come unità trascrizionali indipendenti. Tra questi è possibile annoverare l'*Alu*Sx1 in orientamento senso nel primo introne del gene SRGAP2, coinvolto nello sviluppo e nell'evoluzione del cervello umano[361] (**Figura 17**). L'analisi della sua sequenza ha evidenziato sia la presenza del promotore riconosciuto dalla Pol III, sia di un terminatore nella sequenza genomica fiancheggiante, il quale si trova a valle di un secondo elemento *Alu* (*Alu*Sp) sul quale sono mappati alcuni reads. Tali evidenze suggeriscono che i due elementi potrebbero essere trascritti in un unico RNA primario dimerico, e potrebbe essere stata identificata una nuova tipologia di trascritto *Alu*, finora sconosciuta.

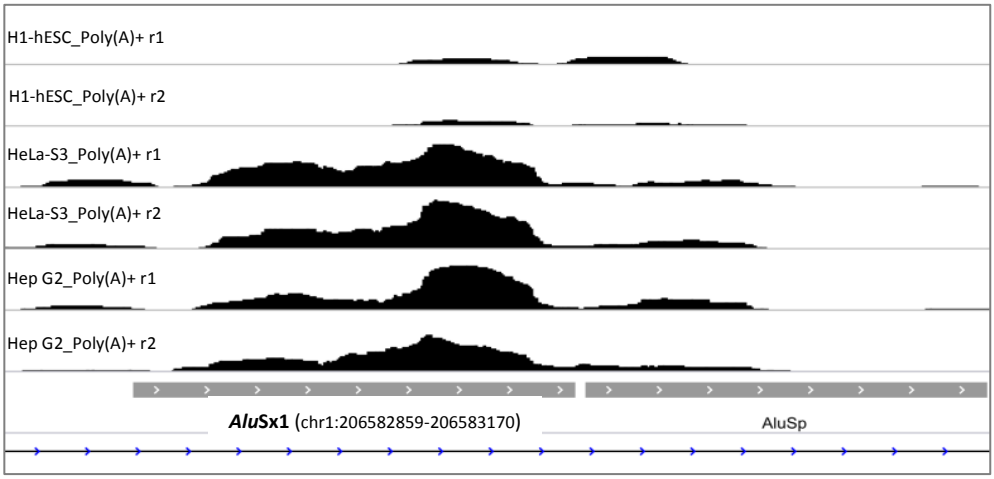


Figura 17. Profilo di espressione dell'elemento *Alu*Sx1 nel primo introne del gene SRAGP2 in orientamento senso e dell'elemento *Alu*Sp a valle.

Inoltre, è risultato di particolare interesse anche l'elemento *Alu*Jb presente nell'introne del gene USP34, che separa gli esoni 35 e 36. Tale trascritto è un esempio di *Alu* RNA intronico in grado di accumularsi nelle cellule in modo indipendente dall'espressione del gene ospite, in quanto si è osservato che l'espressione del suo braccio sinistro (e della regione di 70 nt trascritta a valle) sembra essere inversamente proporzionale all'espressione dell'esone 37 del gene, in differenti tipi cellulari (**Figura 18**).

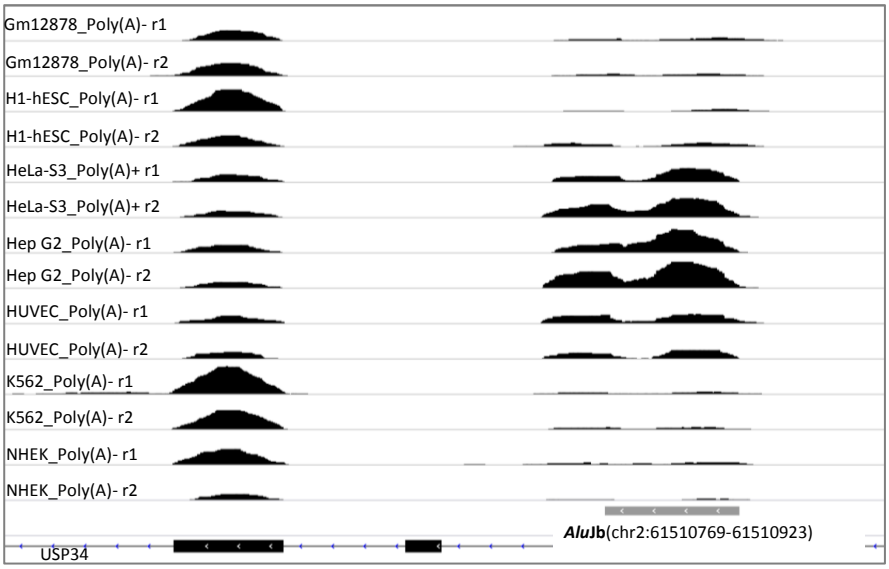


Figura 18. Profilo di espressione dell'elemento *Alu*Jb nell'introne del gene USP34 tra gli esoni 35 e 36 in orientamento senso: è messa in evidenza l'espressione dell'elemento e dell'esone 37 del gene.

Finora si sono considerati elementi *Alu* intronici senso in geni codificanti per proteine espressi autonomamente, ma l'analisi di tale categoria di trascritti, ha messo in luce anche l'espressione di loci *Alu,* negli introni di geni per lincRNA, come unità trascrizionali indipendenti. Un esempio è rappresentato dall'*Alu*Y ospite nell'introne tra gli esoni 4 e 5 del gene per il lincRNA TCONS_I2_00015350 sul cromosoma 2, in orientamento senso (**Figura 19**). L'associazione della Pol III e del fattore TFIIIC a tale locus *Alu* avvalora l'ipotesi che si tratti di un'unità Pol III indipendente.

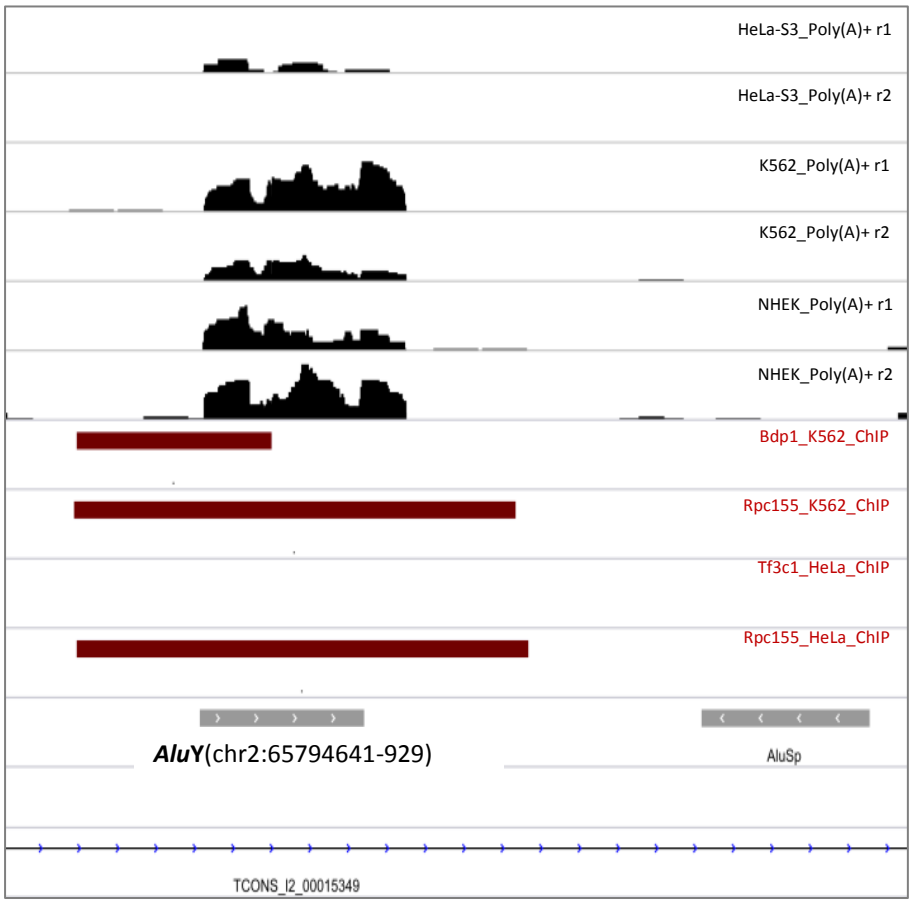


Figura 19. Profilo di espressione dell'elemento *Alu*Y nell'introne del gene per il lincRNA TCONS_I2_00015350 in orientamento senso: è messa in evidenza l'associazione dei componenti del macchinario trascrizionale della Pol III a monte del locus in esame, avvalorando l'ipotesi che si tratti di una unità trascrizionale Pol III autonoma.

Infine, si sono identificati anche elementi *Alu* nelle 3'UTR dei geni in orientamento senso, e si è osservato, per la prima volta, che i trascritti di tali elementi si accumulano nelle cellule indipendentemente dai corrispondenti mRNA (**Figura 20**).

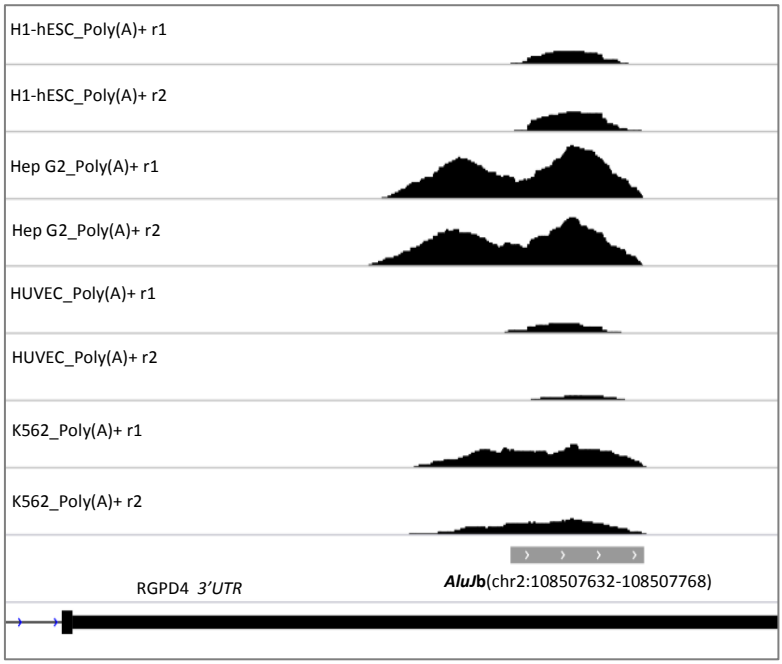


Figura 20. Profilo di espressione dell'elemento *Alu*Jb nella 3'UTR del gene RGPD4 in orientamento senso: è messa in evidenza la sua espressione in 4 tipi cellulari.

Associazione del macchinario trascrizionale della Pol III ai loci *Alu* risultati espressi

Per avere un'ulteriore conferma della trascrizione dei loci *Alu* identificati con la nostra strategia bioinformatica, si è deciso di confrontarli con quelli trovati associati al macchinario trascrizionale della Pol III (macchinario Pol III) in precedenti esperimenti di ChIP-seq[52,358], assumendo che l'associazione del macchinario trascrizionale potesse essere un'evidenza della trascrizione dei loci associati. Per far ciò si sono utilizzati i dati di ChIP-seq di ENCODE per le cellule K562 e HeLa-S3 relativi all'associazione di Bdp1 e Brf1 (componenti di TFIIIB) e di Rpc155 e TFIII110 (subunità rispettivamente della Pol III e di TFIIIC). Circa il 6% dei loci *Alu* identificati come espressi, sono stati trovati associati ad uno o più componenti del macchinario Pol III nelle cellule HeLa, e tale percentuale è aumentata restringendo l'analisi ai 44 elementi *Alu* intergenici dei quali è stata rilevata l'espressione nelle cellule HeLa stesse (13 elementi *Alu*, ovvero il 29%). I P-value di associazione di tali loci intergenici espressi (confrontato con il totale di elementi intergenici) a Bdp1, TFIIIC110 e Rpc155 sono tutti inferiori a 10^{-14}. Analogamente, molti elementi *Alu* risultati

espressi con la nostra strategia, sono stati trovati associati al macchinario Pol III nelle cellule K562 e, restringendo le analisi ai 71 elementi intergenici espressi in queste cellule, il 44% risulta associato ad almeno un componente del macchinario. Anche in questo caso i P-value per l'associazione di tali loci intergenici espressi (confrontato con il totale di elementi intergenici) a Bdp1, TFIIIC110 e Rpc155 sono tutti inferiori a 10^{-15}, sottolineando la significatività dei dati ottenuti. Tra gli elementi *Alu* riportati in precedenza, l'elemento in **Figura 15** è risultato associato alla Pol III, mentre gli elementi in **Figura 13 e 16A**, non solo alla Pol III ma anche a TFIIIC.

Per quanto concerne gli elementi *Alu* intronici senso identificati come unità trascrizionali indipendenti, l'analisi dei dati di ChIP-seq ha evidenziato anche una loro associazione al macchinario Pol III, e l'elemento in **Figura 19** ne è un esempio.

Queste evidenze confermano come la strategia bioinformatica messa a punto sia in grado di identificare loci *Alu* effettivamente espressi, in particolare intergenici/antisenso, ma anche intronici senso. Il fatto che una significativa percentuale di loci *Alu* intergenici/antisenso, identificati come espressi con la nostra strategia, risulti anche associata al macchinario Pol III sia in cellule HeLa-S3 che K562 in esperimenti indipendenti di ChIP-seq, rafforza la conclusione che si tratti di autentiche unità trascrizionali trascritte in modo autonomo dalla Pol III.

Infine, va notato come solo 4 elementi siano stati trovati espressi e associati al macchinario Pol III sia nelle cellule HeLa che K562 (chr1: 61523296-61523586; chr10: 5895538-5895651; chr1: 28672563-28672802; chr8: 144536572-144536880), suggerendo la notevole "plasticità" del trascrittoma di elementi *Alu*.

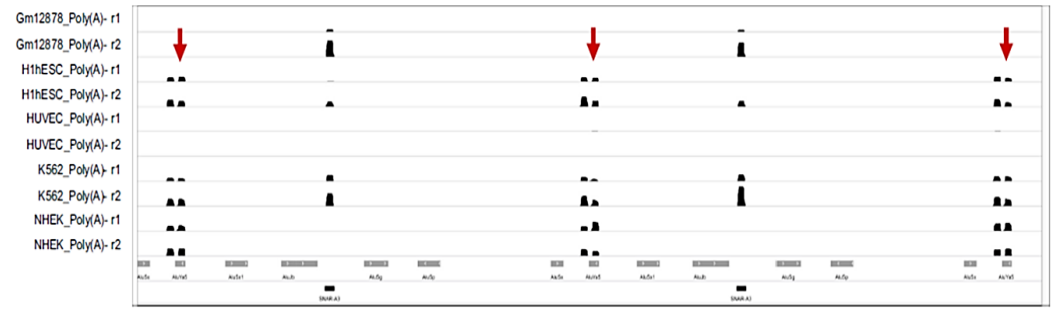
Identificazione di un nuovo gene "*Alu*Ya5-like" trascritto dalla RNA polimerasi III

Come detto finora, la strategia bioinformatica messa a punto ha previsto l'applicazione di filtri stringenti, affinché si potessero identificare solo elementi *Alu* singoli espressi come unità indipendenti. Nonostante ciò, si è voluta estendere l'analisi anche agli elementi *Alu* presenti in più copie identiche nel genoma, e, per farlo, si è applicato un filtro meno stringente rispetto al precedente, e solo limitatamente ai loci intergenici/antisenso. Tale "correzione" ha permesso, come atteso, di identificare più elementi espressi (705 invece dei precedenti 386) ed, esaminando gli elementi *Alu* espressi in almeno tre tipi cellulari, si sono messe in evidenza circa 20 copie pressoché identiche al monomero sinistro dell'elemento *Alu*Ya5, all'interno dei clusters snaR A/C e snaR A/B/D, nel braccio q del cromosoma 19[191]. I profili di espressione, con risoluzione a singolo nucleotide, di tali elementi annotati come *Alu*Ya5, hanno suggerito che in realtà si tratti di

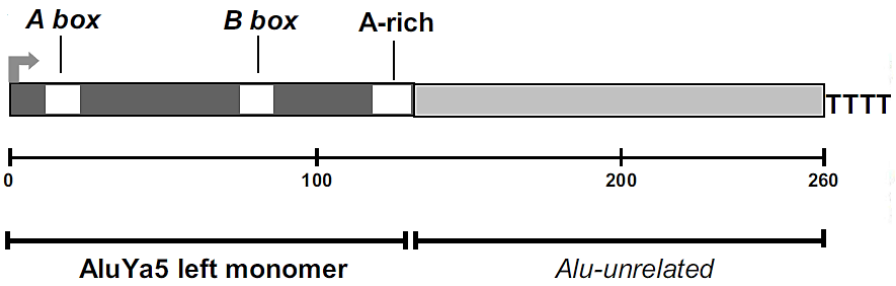
elementi incompleti, in quanto la trascrizione ha inizio in corrispondenza del braccio sinistro dell'elemento *Alu* ma continua a valle nella sua regione fiancheggiante in 3', la cui sequenza è diversa dal braccio destro di *Alu*Ya5 (**Figura 21A**). Inoltre, le loro analisi di sequenza hanno messo in evidenza che tali elementi sono formati, come accennato in precedenza, da un monomero sinistro, molto simile a quello dell'elemento *Alu*Ya5 (contenente gli elementi promotore A box e B box e la corta regione poli-A) e dalla regione a valle che, al contrario, non corrisponde al braccio destro di un elemento *Alu* canonico (detta, pertanto, *Alu*-unrelated). Inoltre, il terminatore della trascrizione incontrato nella regione genomica fiancheggiante gli elementi, è composto da 4 T, inizia alla base +260, e corrisponde all'incirca alla fine della copertura dei reads (**Figura 21B e 21C**). Le nuove unità trascrizionali identificate sono state nominate Ya5-lm (Ya5 left monomer) per la presenza del monomero sinistro simile a quello dell'elemento *Alu*Ya5.

La sequenza di tali unità ricorda quella del gene per l'RNA BC200, un elemento trascrizionalmente attivo, formato da una ripetizione monomerica *Alu* a monte, seguita da un dominio non ripetuto a valle[362].

A



B



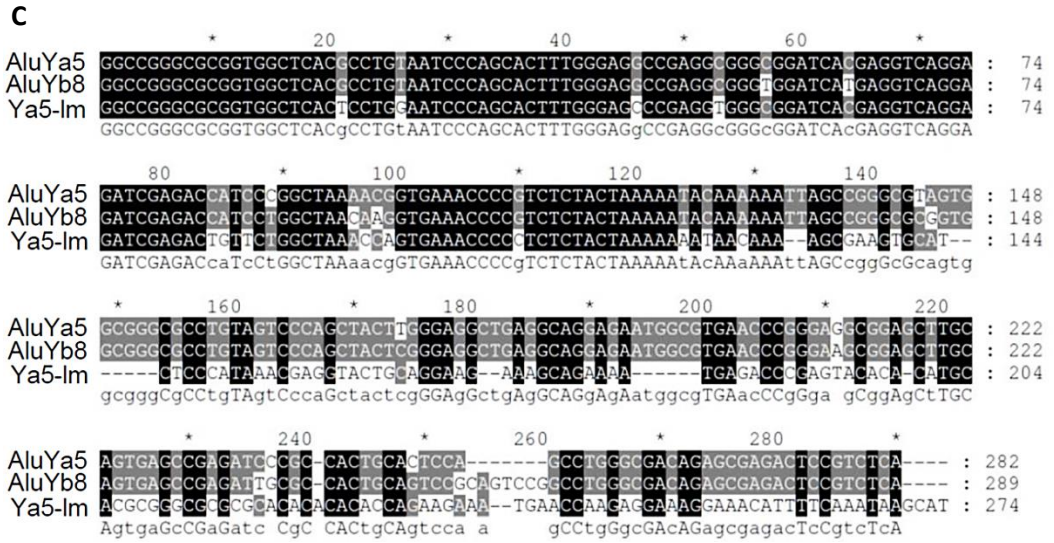


Figura 21. Profilo di espressione delle copie, pressoché identiche al braccio sinistro dell'elemento *Alu*Ya5 (indicate dalle frecce rosse), nel cluster snaR A/C/D del cromosoma 19[363], identificate in 5 diverse linee cellulari **(A).** Le nuove unità trascrizionali identificate sono costituite dal Ya5-lm e da una regione a valle *Alu*-unrelated **(B)** e ciò è confermato dall'allineamento della loro sequenza con la sequenza consensus di *Alu*Ya5 e *Alu*Yb8 (allineamento realizzato mediante l'utilizzo di Clustal X e visualizzato attraverso l'interfaccia grafica di GeneDoc) **(C)**.

L'analisi della regione genomica individuata sul cromosoma 19, ha messo in luce che l'espressione degli elementi Ya5-lm è correlata con quella degli elementi SNAR-A3 nelle vicinanze, trascritti dalla Pol III (**Figura 21A**).

I geni snaR sono disposti sul cromosoma 19 in due estese regioni invertite di ripetizioni in tandem, e i due clusters (A/C e A/B/D) sono separati da una regione di 2Mb. La nostra analisi ha messo in luce che i due clusters contengono rispettivamente 11 e 10 copie delle unità Ya5-lm, ed in entrambi i clusters tutti gli Ya5-lm, separati da circa 5300 bp, sono nello stesso orientamento dei geni snaR, e ognuno di loro è 1800 bp dal gene snaR a monte e 3300 bp dal gene snaR a valle. Infine, dal momento che le copie di Ya5-lm sono pressoché identiche, risulta difficile attribuire i reads a uno o più di essi, ma nonostante ciò, dal momento che la sequenza a valle del monomero, simile a quello dell'*Alu*Ya5, non è presente in nessun altro sito del genoma, non vi sono dubbi che uno o più geni siano trascritti per produrre un nuovo tipo di RNA derivante da elementi *Alu*. Per confermare tale conclusione si sono considerati i dati di ChIP-seq di ENCODE, relativi alle linee cellulari K562 e HeLa S3, e si è individuata in entrambe le linee l'associazione della subunità

Rpc155 in corrispondenza dell'*Alu*Ya5 di coordinate chr19: 50640453-50640584 (**Figura 22**). Ciò ha permesso di confermare che molto probabilmente solo uno o un piccolo sottoinsieme di elementi Ya5-lm sia trascrizionalmente attivo.

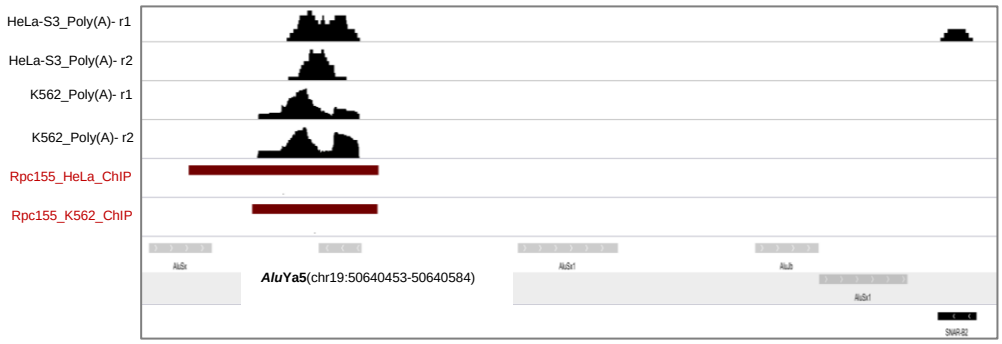


Figura 22. Profilo di espressione dell'elemento *Alu*Ya5 trovato associato alla subunità Rpc155 analizzando i dati di ChIP-seq di ENCODE realizzati su cellule HeLa S3 e K562.

Il fatto che i geni Ya5-lm e snaR siano vicini sullo stesso cromosoma, suggerisce che i primi siano stati inclusi nella stessa porzione di genoma duplicata, dalla quale si pensa che si siano espansi i secondi. Pertanto, i cluster di geni snaR potrebbero presentare un contesto cromatinico favorevole alla trascrizione di differenti ncRNA derivanti da elementi *Alu*, i quali potrebbero, una volta espressi, regolare la traduzione cellulare[191].

Trascrizione *in vitro* di elementi *Alu* con differenti profili di espressione *in vivo*

Una volta identificati elementi *Alu* con diversi livelli di espressione *in vivo*, ci si è chiesti se tali elementi potessero avere anche un diverso comportamento *in vitro*. A tal proposito, si è preso in esame un piccolo sottoinsieme di loci *Alu*, rappresentativi di diversi tipi di profili di espressione, costituito da nove elementi. Uno di questi, *Alu*Sq2_chr1: 61523296-61523586, è risultato espresso in cinque delle sette linee cellulari analizzate (H1-hESC, HeLa-S3, Hep G2, K562, NHEK) ed è stato trovato associato al macchinario della Pol III in cellule HeLa S3 e K562, pertanto è stato scelto come rappresentativo di elementi *Alu* espressi in vari contesti cellulari. Tale elemento ci ha anche incuriosito in quanto manca del motivo interno A_5TACA_6 (sostituito da A_3G) che normalmente hanno tutti gli elementi *Alu* (vedi Appendice I per la sua sequenza).

Altri due elementi (*Alu*Y_chr10b:69524852-69525156 e *Alu*Sp_chr17:4295121-429543) sono stati trovati espressi in solo un tipo cellulare dei sette esaminati (NHEK il primo e K562 il secondo) se si considera la nostra soglia d'espressione, ma anche in altri tipi cellulari, seppure a livelli inferiori.

*Alu*Y_chr10b è anche sembrato interessante in quanto non solo è stato trovato espresso con la nostra strategia bioinformatica, ma è anche stato annoverato tra gli elementi sorgente attivi per la retrotrasposizione in un recente studio[359].

Tre loci (*Alu*Sx1_chr3:139109300-130109588, *Alu*Y_chr7: 73761603-73731897, *Alu*Y_chr10a:103929453-103929749) sono stati trovati espressi in modo sporadico (in non più di 2 linee cellulari e in 1 solo replicato) ma, tuttavia, dai dati di ChIP-seq risultano associati a uno o più componenti del macchinario della Pol III.

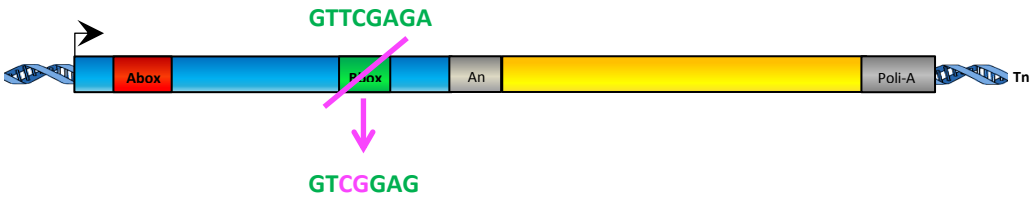
Infine, gli ultimi tre loci (*Alu*Sx_chr1:235531222-235531520, *Alu*Sx_chr10:12236879-12237173 e *Alu*Y_chr22: 41932115-41932411) non sono stati trovati espressi nella nostra analisi anche se l'elemento *Alu*Sx sul cromosoma 10 è trovato associato alla Pol III nei dati di ChIP-seq.

I nove loci in esame sono stati scelti in modo da considerare le categorie di elementi intergenici e intronici antisenso: 5 elementi *Alu* sono, pertanto, intergenici e 4 intronici antisenso. Tra gli elementi intronici antisenso, *Alu*Y_chr22: 41932115-41932411 è stato scelto in quanto in orientamento antisenso all'interno dell'introne 2 del gene POLR3H, il quale codifica per la subunità RPC8 dell'RNA polimerasi III. Tale collocazione intronica antisenso potrebbe interferire con la normale espressione del gene, come proposto per il MIR in orientamento antisenso nell'introne 1 dei geni umano e murino per la subunità RPC5 della Pol III[356]. Gli altri tre elementi *Alu* intronici antisenso si trovano nel primo introne del gene TBCE (TuBulin folding Cofactor E) (*Alu*Sx_chr1:235531222-235531520), nel terzo introne del gene CLIP2 (CAP-GLY domain containing LInker Protein 2) (*Alu*Y_chr7: 73761603-73731897) e nel primo introne del gene NUDT5 (NUcleoside Diphosphate linked moiety X-Type motif 5) (*Alu*Sx_chr10:12236879-12237173).

I loci presi in esame sono stati saggiati mediante trascrizione *in vitro* per valutare se effettivamente possono essere riconosciuti dalla RNA polimerasi III. A tal fine, essi sono stati isolati da DNA genomico umano, clonati nel vettore pGEM-T Easy (Promega) e saggiati *in vitro* in estratto nucleare da cellule HeLa. I saggi di trascrizione sono stati condotti in presenza di α-amanitina alla concentrazione 2 μg/ml (nota per inibire la Pol II) e sono state eseguite, in parallelo alle trascrizioni degli elementi wild type (wt), le trascrizioni degli elementi mutati nella B box (in cui le basi TC sono state sostituite con le basi CG nella sequenza *consensus* GTTCGAGAC, **Figura23A**), l'elemento promotore essenziale affinché avvenga la trascrizione. L'analisi dei trascritti, le cui taglie sono state predette sulla base delle analisi di sequenza, ha rivelato che ogni elemento è in grado di produrre uno specifico profilo di trascrizione *in vitro* e inoltre, a parte nel caso di *Alu*Sx_chr1:235531222-235531520 (che genera più trascritti e a bassi livelli), le efficienze di trascrizione *in vitro* sono comparabili, indicando che la differente espressione *in vivo* dipende da altri fattori (per esempio le

sequenze fiancheggianti e soprattutto il forte controllo epigenetico) ma non dalle caratteristiche del promotore interno alle unità trascrizionali. Inoltre, la mutazione della B box fa notevolmente diminuire i livelli di espressione, confermando l'importanza di tale elemento per il riconoscimento delle unità trascrizionali da parte dell'RNA polimerasi III e la loro successiva trascrizione[55] (**Figura 23B**). In parallelo alle reazioni di trascrizione degli elementi *Alu* wt e mutati nella B box (B mut) sono state anche condotte una reazione con il plasmide pGEM-T Easy vuoto (per monitorare i livelli di trascrizione aspecifica), una reazione con un elemento *Alu*Sx precedentemente caratterizzato (di taglia nota, corrispondente a 372 nt; A. Orioli e G. Dieci, dati non pubblicati), una reazione con un gene per il $tRNA^{Val}$ (AAC) wt (TRNAV18, chr6; elemento sicuramente trascritto dalla Pol III, usato quindi come controllo positivo delle reazioni. La trascrizione di tale gene genera tre trascritti per la presenza di tre terminatori della trascrizione consecutivi nella sua sequenza[53]).

A



B

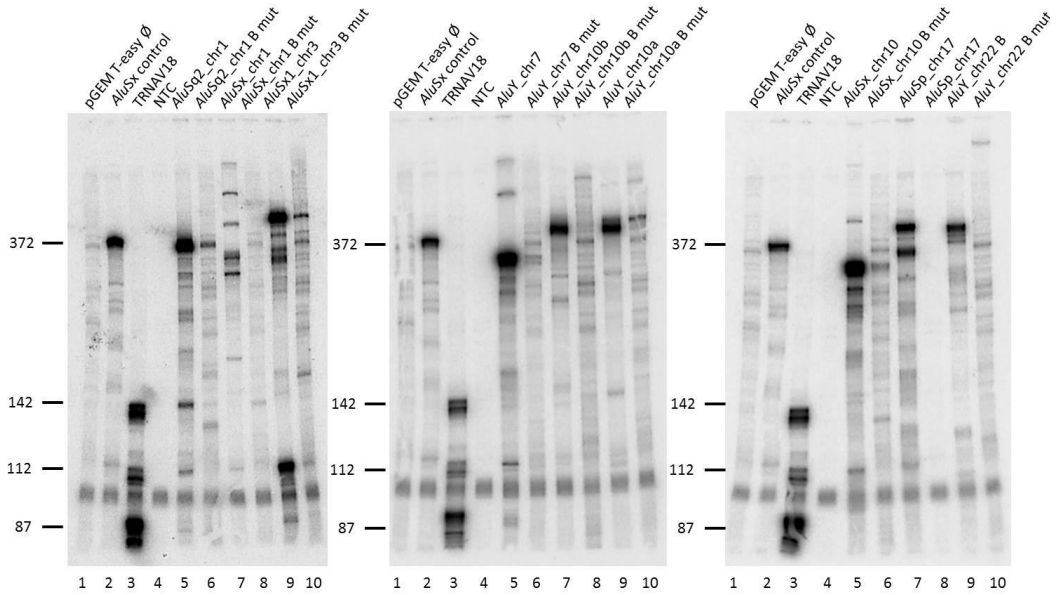


Figura 23. Rappresentazione schematica della mutazione introdotta nelle B box degli elementi *Alu* in esame (in verde la sequenza *consensus* della B box (GTTCGAGAC), in rosa le basi che sono state sostituite rispetto alla basi nella sequenza originale). **(A)**. Saggi di trascrizione *in vitro* condotti in estratto nucleare da cellule HeLa **(B)**. Le reazioni sono state eseguite con 0.5 µg di DNA plasmidico contenente le unità trascrizionali *Alu* wt (campioni 5,7,9 di ogni gel) e le corrispondenti unità trascrizionali mutate nella B box (campioni 6,8,10 in ogni gel). Per ogni gel sono state eseguite anche reazioni di controllo con il vettore vuoto (campione 1), lo stesso vettore con una sequenza *Alu* caratterizzata in precedenza come trascrizionalmente attiva (campione 2), un gene per tRNA (TRNAV18) che genera tre trascritti (87, 112 e 142 nt, campione 3) e in assenza di templato (controllo negativo, campione 4). I saggi sono stati condotti indipendentemente due volte.

Ruolo delle regioni fiancheggianti a monte delle unità trascrizionali *Alu*

Dal momento che sono state riportate in passato evidenze dell'importanza delle sequenze fiancheggianti a monte degli elementi *Alu* per una loro efficiente trascrizione *in vitro* e in cellule trasfettate[61,62], si è voluto anche analizzare in maggior dettaglio il ruolo di tali regioni, costruendo mutanti privi dalla sequenza fiancheggiante l'estremità 5' dei 9 elementi *Alu* isolati, e confrontando il loro livello di trascrizione con quello dei rispettivi costrutti wt. Per ognuno degli elementi è stata rimossa la sequenza a monte fino alla base -12 (o -15 per *Alu*Sx_chr10: 12236879-12237173) e, sia le sequenze wt che le delete in 5', sono state inserite nel vettore con lo stesso orientamento, per ridurre le differenze nell'efficienza di trascrizione dovute alla sequenza di vettore fiancheggiante. I saggi di trascrizione *in vitro* (in estratto nucleare da cellule HeLa) hanno permesso di osservare un generale calo della trascrizione nei deleti in 5' per tutti e 9 gli elementi *Alu*, nonostante il grado di inattivazione della trascrizione fosse considerevolmente diverso a seconda dell'elemento *Alu* considerato. Infatti, nel caso degli elementi *Alu*Sx_chr 1, *Alu*Sx_chr10 e *Alu*Sp_chr17 si è rilevata una riduzione di circa 4-5 volte, nel caso di *Alu*Sq2_chr1 e *Alu*Sx_chr3 di circa 1.5 volte e non apprezzabile nel caso di *Alu*Y_chr10a:103929453-103929749 (**Figura 24**). Tali dati confermano e consolidano l'ipotesi che la sequenza in 5' influenza fortemente l'efficienza di trascrizione degli elementi *Alu*, ma non rivela nessuna ovvia correlazione tra la dipendenza dalla sequenza a monte ed i profili di espressione *in vivo*.

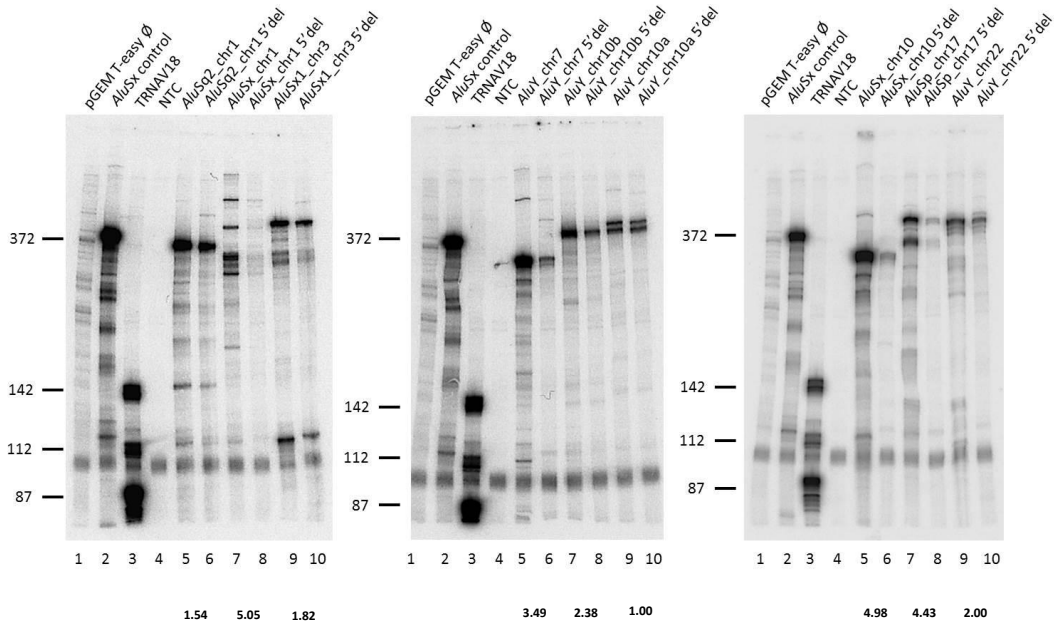


Figura 24. Saggi di trascrizione *in vitro* condotti in estratto nucleare da cellule HeLa. Le reazioni sono state eseguite con 0.5 µg di DNA plasmidico contenente le unità trascrizionali *Alu* wt (campioni 5,7,9 di ogni gel) e le corrispondenti unità trascrizionali delete delle sequenza in 5' (campioni 6,8,10 in ogni gel). Per ogni gel sono state eseguite anche reazioni di controllo con il vettore vuoto (campione 1), lo stesso vettore con un elemento *Alu* caratterizzato in precedenza come trascrizionalmente attivo *in vitro* (campione 2), un gene per tRNA (TRNAV18) che genera tre trascritti (87, 112 e 142 nt, campione 3) e in assenza di tempalto (controllo negativo, campione 4). Sono riportati anche i valori corrispondenti alla riduzione della trascrizione degli elementi a seguito della delezione della regione in 5'. I saggi sono stati condotti indipendentemente due volte.

I saggi di trascrizione *in vitro* hanno permesso di confermare ed estendere le conoscenze in merito al ruolo della regione a monte del TSS degli elementi *Alu* (la sua presenza è necessaria affinché gli elementi vengano attivamente trascritti), al ruolo dell'elemento promotore B box interno (la sua assenza non permette l'assemblaggio del macchinario Pol III e quindi la sintesi dell'*Alu* RNA) e al forte controllo epigenetico sull'espressione degli elementi *Alu in vivo* (confermato dal fatto che loci *Alu* con diversi profili di espressione *in vivo*, hanno simili efficienze di trascrizione *in vitro*).

Per quanto concerne il ruolo della regione a monte degli elementi *Alu*, vi sono evidenze del fatto che possa influenzare la trascrizione ad opera della Pol III anche in presenza di un promotore interno non mutato. Analoghe evidenze sono riportate nel caso dei geni per i tRNA, la cui struttura del promotore interno è strettamente correlata con quella dei loci *Alu*. Tali geni presentano, infatti, un notevole grado di conservazione delle regioni fiancheggianti l'estremità 5' negli eucarioti, e sono riportati esempi del fatto che la regione a monte del loro TSS influenzi la loro

trascrizione sia *in vitro* che *in vivo*[364]. I nostri dati confermano l'importanza di tale regione anche per gli elementi *Alu* e, infatti, nella maggior parte dei costrutti esaminati, si ha una rilevante riduzione della sintesi di *Alu* RNA in assenza di quest'ultima (**Figura 24**).

Infine, il fatto che i profili di espressione degli elementi *Alu in vitro* siano molto simili nonostante le loro differenti espressioni *in vivo*, conferma l'idea che l'espressione dei loci *Alu in vivo* dipenda da altri fattori, quali le sequenze genomiche fiancheggianti gli elementi e il contesto cromatinico della regione genomica in cui si trovano. È quindi messo in evidenza il forte controllo epigenetico sull'espressione *in vivo* degli elementi *Alu*, il cui silenziamento normalmente risulta associato ad un aumento della metilazione nelle loro sequenze e alla formazione di un contesto eterocromatico (vi sono evidenze della presenza della modificazione istonica H3K9me3 in corrispondenza di numerosi loci *Alu* silenti[250]), così come la loro derepressione risulta associata a modificazioni istoniche eucromatiche[365].

Evidenze dell'associazione agli elementi *Alu* espressi di fattori di trascrizione della RNA polimerasi II

La trascrizione degli elementi *Alu* è influenzata dalla regione a monte degli stessi, e questo potrebbe essere dovuto all'azione di specifici fattori di trascrizione (TFs, Transcription Factors), normalmente coinvolti nella regolazione trascrizionale della RNA polimerasi II, in grado di interagire con tale regione. A questo proposito, ci si è chiesti se i loci *Alu* identificati come espressi, mediante l'applicazione della nostra strategia bioinformatica ai dati di RNA-seq, fossero associati ad uno o più fattori del macchinario trascrizionale della RNA polimerasi II, oltre al macchinario della Pol III. Si sono analizzati, a questo fine, i dataset di ChIP-seq per molti TFs (sempre realizzati nell'ambito del progetto ENCODE) nelle 7 linee cellulari utilizzate per le analisi di espressione, e si è osservata un'elevata variabilità tra linee cellulari diverse, per quanto concerne l'associazione dei TFs, sia in termini di numero totale di loci *Alu* associati ai TFs (che oscilla tra 17 e 79), sia in termini di TFs associati ad ogni locus. In particolare, si è osservato un arricchimento di TFs nella regione a monte degli elementi *Alu* intergenici/antisenso, se confrontati con gli elementi *Alu* all'interno di unità trascrizionali ospiti. Nella maggior parte delle linee cellulari, i fattori di trascrizione trovati più frequentemente associati ai loci *Alu* espressi (escludendo i componenti del macchinario Pol III e la TBP) sono stati il fattore che lega il motivo CCCTC (CTCF, un regolatore della trascrizione e organizzatore del genoma), l'RNA polimerasi II (rilevata mediante l'associazione della sua subunità Rpb1) e il fattore di trascrizione JunD. Recenti evidenze hanno dimostrato che tali fattori sono associati ai loci attivamente trascritti dalla Pol III, ed in particolare ai geni per i tRNA, e che CTCF

potrebbe contribuire alla loro funzione nell'organizzazione nucleare a nella loro funzione di isolamento di specifiche regioni genomiche[247,366]. L'associazione dei fattori CTCF, Pol II e JunD a monte dei loci *Alu* risultati espressi rafforza, quindi, l'idea che essi siano simili agli altri geni trascritti dalla Pol III non solo per le loro proprietà trascrizionali, ma anche per la loro funzione nell'organizzazione del genoma. Inoltre, le nostre analisi hanno rivelato un'associazione linea cellulare-specifica degli elementi *Alu* espressi con il fattore CEBPB (CCAAT/Enhancer Binding Protein β), particolarmente accentuata nelle cellule HeLa S3 (P-value= 10^{-5}), HepG2 (P-value= 10^{-10}) e K562 (P-value= 10^{-6}) (v. Materiali e metodi). Dal momento che non vi sono evidenze riportate dell'associazione di tale fattore in corrispondenza dei geni trascritti dalla Pol III, la nostra analisi potrebbe aver identificato un nuovo fattore di trascrizione specifico per gli elementi *Alu*.
L'evidenza dell'associazione delle due polimerasi, Pol II e Pol III, in corrispondenza dei loci *Alu* risultati attivamente trascritti, è in accordo con numerosi casi riportati di associazione dei fattori del macchinario Pol II a monte dei geni trascritti dalla Pol III[247]. Inoltre, il fatto che i loci *Alu* intronici senso in unità trascrizionali ospiti sembrino avere una minor tendenza all'associazione del macchinario Pol III rispetto agli intergenici, potrebbe far concludere che la loro espressione avvenga o grazie al rilascio di *Alu* RNA da parte di trascritti Pol II annotati (analogamente a quanto riportato per i miRNA e gli snoRNA derivati da introni[367,368]) o attraverso la produzione e il processamento, non ancora caratterizzati, di trascritti Pol II non codificanti contenenti elementi *Alu*. Ciò potrebbe permettere di spiegare anche gli elementi *Alu* intergenici/antisenso identificati come espressi, ma non associati alla Pol III.

Modificazioni istoniche associate agli elementi *Alu* espressi

Dal momento che dalle analisi di ChIP-seq dei dati di ENCODE è possibile ricavare, oltre alle informazioni menzionate nel paragrafo precedente, anche informazioni circa le modificazioni istoniche tipiche di geni trascritti dalla Pol III (nonché condivise da questi e dai geni trascritti dalla Pol II[247]), si sono voluti altresì ricostruire i profili di modificazioni istoniche presenti in corrispondenza dei loci *Alu* risultati espressi, per avvalorare l'attendibilità della nostra strategia bioinformatica. Le analisi condotte sugli elementi *Alu* intergenici espressi non hanno prodotto risultati di facile interpretazione, probabilmente per il ridotto numero di loci analizzati, ma nonostante ciò, l'acetiltrasferasi P300 (EP300) è risultato entro i primi 10 componenti della cromatina associati agli elementi *Alu* espressi nelle cellule HepG2, HeLa S3 e K562 (l'associazione di EP300 ai loci *Alu* espressi, rispetto ai loci *Alu* totali, è risultata significativa (P-value <10^{-12}) sia

nelle cellule HeLa S3 che K562; v. Materiali e metodi), suggerendo che tali elementi potrebbero essere caratterizzati dalla presenza di elevati livelli di istoni acetilati (si ricorda che l'acetilazione istonica è connessa con l'attivazione trascrizionale). Tale considerazione è in accordo con un recente studio in cui è messo in evidenza un arricchimento di H3K27ac e P300 in corrispondenza di elementi *Alu* simili ad enhancer[365] (ovvero elementi *Alu* per i quali si è osservata non solo l'associazione con le modificazioni istoniche tipiche degli enhancer, ma anche la capacità di interagire con promotori di geni distanti nel genoma rispetto alla loro posizione, caratteristica funzionale tipica degli enhancer) e rafforza ulteriormente la veridicità degli elementi *Alu* autonomi risultati espressi nelle linee cellulari esaminate.

I risultati delle trascrizioni *in vitro*, riportati in precedenza, hanno messo in evidenza un forte controllo epigenetico sull'espressione *in vivo* degli elementi *Alu.* Nonostante, infatti, tali elementi siano normalmente silenti e associati a modificazioni tipiche della eterocromatina[250] (il che spiega i loro bassissimi livelli cellulari come trascritti), la loro espressione è risultata anche correlata con la presenza di modificazioni istoniche tipiche della cromatina aperta e degli enhancer[365]. Le evidenze del fatto che i loci *Alu*, risultati espressi con il nostro algoritmo, siano associati all'acetiltrasferasi P300, e che le loro regioni a monte siano ricche dei fattori trascrizionali Pol II, contribuiscono a porre un primo tassello verso la conoscenza dei meccanismi epigenetici che controllano la loro espressione, e potrebbero indurre a ipotizzare un coinvolgimento dei TFs della Pol II nella modulazione della trascrizione di tali elementi da parte della Pol III, nonché una loro possibile trascrizione da parte della Pol II, finora inesplorata.

Conclusioni

Dal momento che gli elementi *Alu* possono essere trascritti come unità indipendenti dalla Pol III ma anche essere ospiti di trascritti più lunghi, sintetizzati dalla Pol II, e che entrambe le categorie di trascritti sono molto simili in sequenza, è da sempre stata ardua l'identificazione dei singoli loci *Alu* trascritti autonomamente. I primi studi condotti al fine di identificare *Alu* RNA ad espressione autonoma, hanno previsto l'ibridazione di un oligonucleotide a tali trascritti (Northern blot) ma non hanno permesso di raggiungere lo scopo dal momento che, come detto in precedenza, gli *Alu* RNA sono molto simili in sequenza tra loro ed anche simili all'RNA da cui derivano (il 7SL RNA), che peraltro, ha la stessa lunghezza dei trascritti *Alu* autonomi (circa 300 nucleotidi)[54,67]. Pertanto non è stato possibile direzionare le ibridazioni solo sui bersagli di interesse e si è reso necessario l'impiego di tecniche più fini. A tal proposito, sono documentati in letteratura studi che hanno previsto l'impiego di un approccio alternativo con una tappa di amplificazione, per superare il problema dei bassissimi livelli di *Alu* RNA nelle cellule (RT-PCR), ma neanche mediante l'impiego di tale approccio è stato possibile isolare *Alu* RNA sintetizzati autonomamente, in quanto non in grado di discernere tra i trascritti da parte della Pol III e quelli ospiti di RNA più lunghi, sintetizzati dalla Pol II[54]. Solo con l'avvento delle nuove tecnologie di sequenziamento si è aperta la strada verso l'identificazione di singoli loci *Alu* trascrizionalmente attivi ed in particolare, attraverso gli approcci di ChIP-seq (immunoprecipitazione cromatinica seguita da sequenziamento massivo), è stato possibile rilevare l'associazione *in vivo* del macchinario trascrizionale della Pol III ad elementi *Alu*, un'indicazione ragionevole del fatto che potessero essere trascritti. In questo contesto, il nostro lavoro risulta di rilievo dal momento che per la prima volta sottolinea che un'analisi appropriata dei dati di RNA-seq, permette di superare le difficoltà nell'identificare i trascritti *Alu* sulla base della loro sequenza e della loro lunghezza. In questa prima parte del programma si è elaborata una strategia in grado di identificare singoli elementi *Alu* (trascritti dalla RNA polimerasi III) espressi, a seguito di sequenziamento del trascrittoma, e tale abilità ha permesso di svelare, con estrema precisione ed accuratezza, i profili di espressione di *Alu* RNA in diverse linee cellulari umane.

La possibilità di determinare i profili di espressione degli elementi *Alu* con una risoluzione a singolo locus, rappresenta pertanto un passaggio chiave per capire il loro controllo trascrizionale, problema non ancora esplorato nonostante la sua notevole rilevanza per la stabilità del genoma umano. In particolare l'algoritmo elaborato è risultato attendibile soprattutto per identificare elementi *Alu* intergenici/antisenso espressi, e la loro analisi ci ha permesso di sottolineare che

nelle cellule umane gli *Alu* RNA espressi sono solo una piccola frazione di quelli presenti nel genoma, che un sottoinsieme di trascritti è condiviso da più linee cellulari (permettendo di supporre l'importanza evolutiva e/o funzionale di tali elementi), e che vi è anche un gruppo di trascritti caratteristico di un determinato tipo cellulare. Questo gruppo di *Alu* RNA tipo cellulare-specifico ha permesso di ipotizzare che gli elementi espressi possono avere una particolare sensibilità alla metilazione del DNA e al contesto cromatinico, i quali sono soggetti a variazioni da un tipo cellulare all'altro. Pertanto determinare i profili di *Alu* RNA, attraverso RNA-seq, potrebbe rappresentare un modo nuovo, estremamente fine e sensibile, di monitorare le alterazioni dell'epigenoma che accompagnano stati fisiologici e patologici e che, cosa molto importante, probabilmente insorgono in risposta all'esposizione a determinati fattori ambientali. Identificando un sottoinsieme di *Alu* RNA specifici di una linea cellulare, si potrebbe probabilmente essere in grado di risalire alle modificazioni epigenetiche specifiche di tale tipo cellulare e valutare una loro variazione a seguito di particolari esposizioni ambientali. La nostra strategia mette quindi in luce la possibilità di analizzare le variazioni epigenetiche in tutti i contesti cellulari compatibili con il sequenziamento del trascrittoma e, dal momento che permette di risalire all'espressione differenziale di specifici loci *Alu*, permetterà anche di analizzare i profili di espressione associati allo sviluppo, alla trasformazione cellulare verso la condizione tumorale, alle alterazioni cellulari in differenti patologie e alle differenze interindividuali nell'espressione genica.

Infine, il fatto che si possano identificare profili di *Alu* RNA tipo cellulare-specifici, permetterà anche di analizzare a fondo il controllo trascrizionale degli elementi *Alu*, cosa molto importante in quanto stanno emergendo evidenze del fatto che una loro alterata espressione potrebbe essere connessa con diverse patologie. I profili di espressione degli *Alu* RNA potrebbero rivelarsi molto interessanti come marcatori di particolari stati cellulari, ambito finora inesplorato ma di notevole potenzialità date le recenti scoperte dei ruoli degli RNA non codificanti in numerosi processi cellulari (trascrizione, processi post-trascrizionali e traduzione).

Materiali e metodi

Strategia bioinformatica per l'identificazione di elementi *Alu* autonomamente espressi

Per identificare loci *Alu* singoli e trascrizionalmente attivi, si è deciso di analizzare i dati grezzi di RNA-seq ottenuti all'interno del progetto ENCODE, in particolare i dati del Cold Spring Harbor Laboratory (CSHL) relativi a sette differenti tipi cellulari, sia normali che tumorali: Gm12878, H1-hESC, HeLa S3, HepG2, HUVEC, K562, NHEK. Le analisi sono state condotte sui dati corrispondenti ai trascritti dell'intera cellula divisi in poli-A+ e poli-A- per entrambi i replicati dell'esperimento (per ogni linea cellulare si avevano sequenziamenti da due replicati biologici indipendenti), per un totale di 28 dataset.
(v.http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaseq).
Ogni dataset è risultato composto da *reads paired-end*, ovvero da sequenze di 76 nucleotidi ottenuti sequenziando entrambi i filamenti di cDNA sintetizzato per retrotrascrizione dell'RNA totale. I reads di ogni dataset sono stati allineati con il genoma di riferimento (hg19) mediante TopHat aligner (versione 2.0.11)[369], applicando sia una strategia di allineamento unico (che considera validi solo i reads che si appaiano a un solo sito nel genoma), sia una strategia di allineamento cosiddetta "best match", che considera validi anche reads che si appaiano con ugual punteggio a più di un locus genomico, attribuendoli casualmente a uno solo dei loci con ugual punteggio[370]. I reads mappati univocamente sul genoma sono stati utilizzati per la maggior parte delle analisi riportate, e sono state mantenute le impostazioni predefinite di TopHat aligner, in modo che per ogni coppia di reads (i membri di una coppia di reads sono definiti *mate*) fossero restituiti tutti gli allineamenti, con lo stesso punteggio, corrispondenti a fino 20 siti genomici di appiamento. TopHat aligner, inoltre, permette di rilevare i *reads mate*, che sono concordi (ovvero sullo stesso filamento di DNA e vicini), discordi (ovvero su filamenti opposti di DNA e lontani), o i reads per i quali solo un membro della coppia è mappato sul genoma. Si è deciso di considerare sia reads concordi che discordi (per ogni coppia di *reads mate*) in quanto la natura ripetuta e tendenzialmente polimorfica degli elementi *Alu*, poteva essere la causa del fatto che l'*aligner* attribuisse per sbaglio a uno dei due *reads mate* l'appaiamento ad una regione genomica più lontana, a causa di un allineamento leggermente peggiore nella posizione corretta (dovuto alla presenza, per esempio, di un polimorfismo in tale regione). Detto ciò, i reads allineati in modo univoco (secondo i criteri riportati sopra) sono stati contati mediante l'uso dello script HTSeq count del pacchetto HTSeq Python[371] per ognuno dei 28 dataset: sono stati tenuti solo i *reads mate* che mappavano in modo univoco sugli elementi *Alu* annotati (*unique alignment*), definendo anche

una soglia di espressione, ovvero considerando come espressi solo elementi *Alu* con più di 10 reads mappati. Innanzitutto, per poter rilevare i reads mappati sugli elementi *Alu* genomici, si è eseguito un allineamento tra i reads mappati univocamente sul genoma di riferimento e il dataset di elementi *Alu* annotati, scaricato da UCSC Table Browser Repeatmasck track.

Il dataset di *Alu* include, oltre agli elementi completi (full-length), anche gli elementi *Alu* tronchi (in 5' e/o in 3'), ma non gli elementi FLAM, FRAM o FAM (monomeri da cui gli elementi *Alu* si sono originati[200]). Inoltre, il dataset è stato modificato in modo da riportare anche il numero del cromosoma e risalire, quindi, alla localizzazione genomica degli elementi *Alu* ai quali i reads mappavano in modo univoco. Inoltre, per verificare che gli allineamenti con TopHat e la soglia impostata, fossero affidabili, si è anche utilizzato STAR aligner[372] (sviluppato indiendentemente da TopHat) per l'analisi di 2 dei 28 dataset (NHEK poli-A+, entrambi i replicati biologici) e si è trovata una corrispondenza, con i risultati di TopHat, di circa il 96%, per quanto concerne le *Alu* con più di 10 reads mappati univocamente nel genoma. Una volta identificati gli elementi *Alu* con più di 10 reads mappati sul genoma, le coordinate di tali elementi (nel formato BED file) sono state fornite allo script sitepro del Cis-regulatory Element Annotation System (http://liulab.dfci.harvard.edu/CEAS/) insieme ai profili corrispondenti al loro segnale di RNA-seq (in formato wiggle). Sitepro è stato utilizzato, nonostante sia stato sviluppato principalmente per i dati di ChIP-seq, per calcolare i profili del segnale di RNA-seq in un range di ± 500 nt dal centro del corpo di un elemento *Alu*, con una risoluzione a 50 nt. In questo modo si sono ottenute tre regioni, di 350 nt l'una, chiamate rispettivamente segmento in 5', corpo dell'*Alu* e segmento in 3', e si è potuto così ovviare al problema di *Alu* RNA ospiti di unità Pol II più lunghe applicando un filtro, per esludere falsi positivi, sulla base dei segnali di espressione spuri a monte e a valle. Per fare ciò, si è considerato che il TSS di un elemento *Alu* è localizzato circa 12 nt a monte della A box (v. **Figura 4** dell'Introduzione), pertanto, per evitare di considerare trascritti con una regione a monte più estesa, si è impostato un filtro che consentisse di attribuire un valore al segmento in 5' che fosse meno di 1/7 del valore del corpo dell'*Alu* (per includere la possibilità di TSS leggermente a monte rispetto alla posizione predetta). Analogamente, dal momento che la trascrizione di un elemento *Alu* può continuare fino al primo terminatore incontrato nella sequenza genomica fiancheggiante (dando origine a trascritti anche circa 500 nt a valle del TSS), si è scelto di imporre che il segmento in 3' avesse un valore che non fosse più della metà del valore attribuito al corpo dell'*Alu*.

Solo gli *Alu* RNA che hanno passato questo filtro, in entrambi i replicati degli RNA-seq di ENCODE (per ogni tipo cellulare), sono stati considerati loci *Alu* espressi come unità trascrizionali autonome. L'elenco completo di tali elementi *Alu* è riportato nella Supplementary Table S1 dell'articolo Conti,

A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014)[373].

Tutte le analisi sono state condotte utilizzando come genoma di riferimento GRCh37/hg19 e, per confermare l'attendibilità dei risultati, si è applicata la strategia bioinformatica messa a punto anche sul genoma di riferimento GRCh38 (nel quale il numero totale di basi è superiore solo del 2% rispetto a GRCH37). L'analisi è stata condotta sui dataset corrispondenti agli RNA-seq delle cellule NHEK (poli-A+, entrambi i replicati) e si è osservato che il 92-95% degli elementi *Alu* identificati come espressi è in comune ad entrambi i genomi di riferimento, confermando l'affidabilità della strategia elaborata per identificare elementi *Alu* autonomi trascrizionalmente attivi a partire da dati grezzi di RNA-seq.

Per i file contenenti gli allineamenti, con *reads paired-end* mappati univocamente, ottenuti utilizzando TopHat aligner per i 28 dataset di ENCODE esaminati, e STAR aligner per i dataset di cellule NHEK (replicato 1 e replicato 2), nonché la strategia bioinformatica messa a punto (con tutte le indicazioni informatiche utilizzate) si rimanda all'articolo Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014)[373].

Suddivisione degli elementi *Alu* espressi sulla base della loro localizzazione genomica

Una volta identificati gli elementi *Alu* espressi nelle 7 linee cellulari esaminate, sono state realizzate delle categorie sulla base della loro collocazione genomica (come riportato in **Figura 12**, Risultati e discussione). Le categorie realizzate sono: (i) *Alu* presenti in regioni tra geni codificanti per proteine o per lincRNA, oppure in introni/esoni di geni ma trascritti in orientamento opposto (intergenici/antisenso, chiamati intergenic/AS), (ii) *Alu* presenti in introni di geni per proteine trascritti nello stesso orientamento (chiamati intronic Refseq in quanto ospiti di geni Refseq annotati) o lincRNA trascritti sempre nello stesso orientamento (chiamati intronic lincRNA in quanto ospiti dei trascritti lincRNA annotati in UCSC), (iii) *Alu* presenti nelle 5'UTR/3'UTR di geni per proteine o lincRNA trascritti nello stesso orientamento (5'UTR/3'UTR senso in geni Refseq) e infine *Alu* presenti in esoni o introni di geni per proteine o lincRNA (interamente o parzialmente sovrapposte a tali esoni), che, non essendo riconducibili alla prime tre categorie, sono stati inseriti nella categoria "other" (iv). Le liste complete di tali elementi *Alu* sono presenti nella Supplementary Table S1 dell'articolo Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014)[373].

Dal momento che gli *Alu* RNA sono stati trovati espressi sia nei dataset corrispondenti all'RNA poli-A+ che poli-A-, nella Supplementary Table S1 è riportata anche una lista di tutti gli elementi *Alu* espressi, sia quelli espressi nella frazione poli-A+ che nella frazione poli-A- dell'RNA cellulare (vedi foglio Excel nominato "All non-redundant"[373]).

Analisi dei dati di ChIP-seq per verificare l'assoziazione del macchinario Pol III

Per avere un'ulteriore prova del fatto che le unità trascrizionali identificate con la nostra strategia bioinformatica fossero effettivamente trascritte, si sono incrociati i dati di ChIP-seq, relativi all'associazione dei componenti del macchinario trascrizionale Pol III (TFIIIC-110, RPC155, BRF1, BRF2, BDP1) nelle cellule HeLa S3 e K562, con le coordinate degli elementi *Alu* risultati espressi (estese a monte di 200 bp). I picchi di associazione dei componenti del macchinario Pol III sono stati estratti dai dati di ChIp-seq ENCODE/Stanford/Yale/USC/Harvard (v. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/).
Le liste degli elementi *Alu* risultati espressi secondo la nostra strategia, che risultano anche associati ai componenti del macchinario trascrizionale Pol III sono riportati nella Supplementary Table S2 dell'articolo Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014)[373].
Per determinare le associazioni significative dei componenti del macchinario della Pol III agli elementi *Alu* intergenici espressi si è applicato il Fisher's exact test .

Identificazione di un nuovo gene "AluYa5-*like*"

Per estendere l'analisi anche agli elementi *Alu* presenti in più copie identiche nel genoma, si è utilizzata l'opzione "-g 1" di TopHat aligner, che consente un allineamento più permissivo dei reads al genoma di riferimento[370]. Tale opzione, infatti, prevede che se un read mappa più di una volta nel genoma (con allineamento perfetto), questo sarà attribuito casualmente ad uno dei siti genomici rilevati. Detto ciò, è necessario tener presente che i reads che mappano all'interno di un elemento *Alu* potrebbero in realtà provenire da un altro locus genomico, ma allo stesso tempo non si escludono dall'analisi gli elementi *Alu* presenti in più copie identiche ed espressi. Tale strategia ha permesso di identificare trascritti simili ad *Alu* RNA (annotati come *Alu* Ya5 ma che in realtà presentano un braccio sinistro simile a quello dell'elemento *Alu*Ya5 e un dominio *Alu*-unrelated al posto del normale braccio destro), che potrebbero rappresentare una nuova categoria di trascritti derivati da elementi *Alu*, non ancora scoperta.

Associazione di fattori di trascrizione Pol II agli elementi *Alu* espressi

Dopo aver identificato *Alu* RNA autonomi, e verificato la loro associazione al macchinario Pol III, si è deciso di valutare l'eventuale associazione di altri fattori di trascrizione ai loci *Alu* identificati come espressi. Per far ciò, per ogni linea cellualre si sono incrociate le 500 bp a monte degli elementi *Alu* con le coordinate dei siti di legame dei fattori di trascrizione accessibili interrogando i dati di ChIp-seq di ENCODE.

(v. http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz). La lista e le analisi statistiche per determinare la significatività delle interazioni tra i fattori di trascrizione e gli elementi *Alu*, sono riportate nella Supplementary Table S3 dell'articolo Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014)[373]. Per determinare le associazioni significative dei fattori di trascrizione della Pol II agli elementi *Alu* espressi si è applicato il Fisher's exact test .

Realizzazione dei costrutti ricombinanti

Per testare l'efficienza di trascrizione *in vitro* di loci *Alu* con differenti profili di espressione *in vivo*, si sono scelti 9 elementi *Alu*, appartenenti alle categorie intergenici/intronici antisenso, e si è proceduto al loro clonaggio nel vettore pGEM-T Easy. I nomi dei 9 loci selezionati, le linee cellulari in cui sono stati trovati espressi (mediante analisi dei dati di RNA-seq) e le taglie predette dei loro trascritti primati (in nucleotidi) sono riportate nella tabella alla pagina seguente.

Nome elemento *Alu*	Linee cellulari in cui è espresso	Taglia predetta dei trascritti primari
AluSq2_chr1	H1-hESC, HeLa S3, HepG2, K562, NHEK	355 (T_4); 361 (T_{10})
AluSx_chr1	/	328 (TAT_3); 338 (TAT_3); 431 (T_4)
AluSx1_chr3	H1-hESC, Gm12878 (espressione sporadica)	304 (T_3GT); 311 (TCT_3); 437 (TAT_3); 443 (T_{17})
AluY_chr7	K562 (espressione sporadica)	322 (T_5)
AluY_chr10a	H1-hESC (espressione sporadica)	370 (TCT_3); 376 (T_4); 397 (T_6); 406 (T_3GT_2)
AluY_chr10b	NHEK	397 (T_5)
AluSx_chr10	/	320 (T_4); 456 (T_6)
AluSp_chr17	K562	387 (T_3CT); 424 (TAT_3); 430 (T_6)
AluY_chr22	/	378 (TGT_3); 409 (T_4); 590 (T_3CT)

Le localizzazioni genomiche dei 9 loci selezionati sono riportate nella tabella sottostante.

Nome elemento *Alu*	Localizzazione genomica
AluSq2_chr1	chr1: 61523296-61523586
AluSx_chr1	chr1: 235531222-235531520
AluSx1_chr3	chr3: 139109300-139109588
AluY_chr7	chr7: 73761603-73761897
AluY_chr10a	chr10: 103929441-103929803
AluY_chr10b	chr10: 69524852-69525156
AluSx_chr10	chr10: 12236879-12237173
AluSp_chr17	chr17: 4295121-4295437
AluY_chr22	chr22: 41932115-41932411

Tali loci sono stati amplificati mediante PCR utilizzando come templato DNA genomico estratto da saliva umana. Le reazioni di PCR sono state condotte con l'enzima GoTaq®DNA Polymerase

(Promega) e con coppie di primers che non amplificassero strettamente le unità trascrizionali *Alu*, ma che si appaiassero circa 200 bp a monte dell'ipotetico TSS e circa 100 bp a valle del terminatore della trascrizione nella sequenza genomica fiancheggiante, in modo da individuare regioni genomiche non ripetute ed essere certi di isolare solo le sequenze di interesse. I primers utilizzati per l'isolamento dei loci *Alu* wild type (wt) sono riportati in Appendice I. In seguito, si è proceduto al clonaggio dei 9 loci wt amplificati nel vettore pGEM-T Easy, utilizzando il kit pGEM®-T Easy Vector System (Promega) e si è confermata la realizzazione dei costrutti wt, mediante sequenziamento dei plasmidi ricombinanti. In seguito, si è proceduto mutando l'elemento promotore B box di tutti i 9 loci in esame, mediante PCR mutagenica condotta sui costrutti wt, che ha previsto la fusione di un frammento upstream e di un frammento downstream (rispetto alla regione da mutare) sovrapposti nella regione mutata, secondo quanto riportato in precedenza[374]. La mutazione della B box è stata progettata in modo che le basi CG, nella sequenza *consensus* GTTCGAGAC, fossero sostituite con le basi TC (v. **Figura 23A**, Risultati e discussione) e si impedisse, pertanto, il legame del fattore TFIIIC, con conseguente perdita della trascrizione da parte del macchinario Pol III. I primers per la realizzazione dei mutanti nella B box (Bmut) sono riportati in Appendice I. Gli ampliconi così ottenuti sono stati clonati nel vettore pGEM-T Easy. Anche in questo caso si è confermata l'avvenuta realizzazione dei costrutti mutati nella B box mediante sequenziamento.

Infine, oltre ai mutanti nella B box, si sono realizzati anche i mutanti privi dalla sequenza fiancheggiante l'estremità 5' dei 9 elementi *Alu* isolati (v. **Figura 24**, Risultati e discussione) progettando i primers forward in modo tale che permettessero di realizzare ampliconi con solo 12 basi a monte del TSS (15 basi nel caso di *Alu*Sx_chr10). I primers forward per la realizzazione dei mutanti deleti nella regione in 5' (5' del) sono riportati in Appendice I.

Gli ampliconi ottenuti sono stati clonati nel vettore pGEM-T Easy e sono stati selezionati i costrutti con inserti nello stesso orientamento dei rispettivi inserti wt, per evitare possibili differenze nell'efficienza di trascrizione dovute alla sequenza di vettore fiancheggiante. Anche in questo caso il sequenziamento ha confermato l'avvenuta realizzazione dei costrutti deleti in 5'.

Saggi di trascrizione *in vitro*

Tutti i plasmidi sono stati purificati con il kit Qiagen Plasmid Mini Kit (Qiagen) prima dell'esecuzione delle reazioni di trascrizione *in vitro*. Queste sono state eseguite in un volume finale di 25 µl con 500 ng di templato plasmidico, 70 mM KCl, 5mM $MgCl_2$, 1mM DTT, 2.5% glicerolo, 20 mM Tris-HCl pH 7.9, 5mM fosfocreatina, 2 µg/ml α-amanitina, 0.4 U/ml SUPERase-In RNase Inhibitor (Ambion), 40 µg di estratto nucleare da cellule HeLa[375], 0.5 mM ATP, CTP e GTP, 0.025 mM UTP e 10 µCi di [α^{32} P] UTP (PerkinElmer). Le miscele di reazione sono state incubate per 60 minuti a 30°C e in seguito interrotte con 75 µl di nuclease-free water e 100 µl di fenolo:cloroformio pH 4.2 (1:1). Gli RNA marcati radioattivamente così purificati, sono stati caricati su gel di poliacrilammide denaturante al 6% (urea 7M) e sono stati visualizzati e quantificati mediante l'utilizzo del Cyclone Phosphor Imager (PerkinElmer) e del software Quantity One (Bio-Rad).

Appendice I

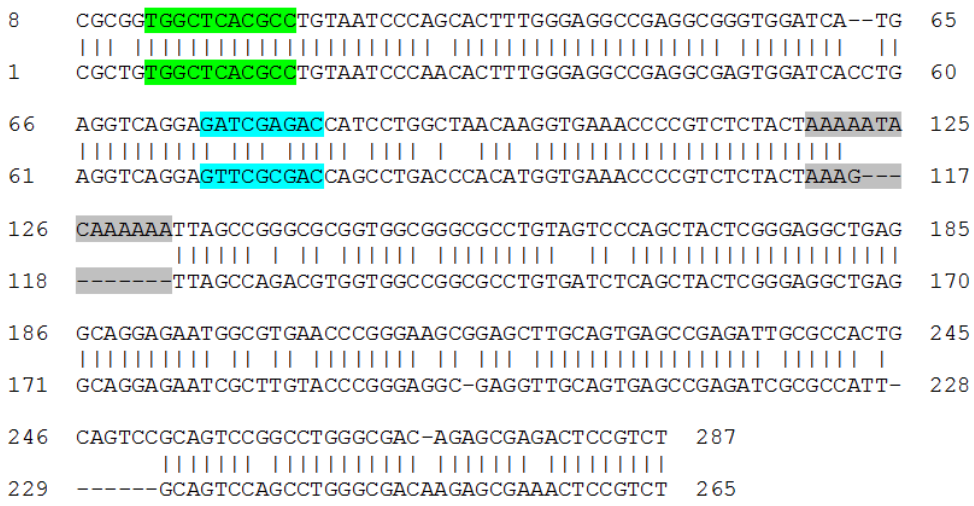
• Sono riportati i primers utilizzati nello studio:

Primers per l'amplificazione di elementi *Alu* wt	**Sequenze dei primers (5'→ 3')**
***Alu*Sq2_chr1_fw**	GCCCCAGGTGATCTCTACC
***Alu*Sq2_chr1_rev**	GTCCTCGGAGCCGCTAATTT
***Alu*Sx_chr1_fw**	CCTGCCAGGACTGCTTCTCTG
***Alu*Sx_chr1_rev**	TTAGGAGTAGCCGCACGTGATC
***Alu*Sx1_chr3_fw**	GAGACGGCAACTACCCAACC
***Alu*Sx1_chr3_rev**	GTCACATTAGCTCAAGGCTCCTAC
***Alu*Y_chr7_fw**	TCATAGTCACATATTCAGTGGCGTT
***Alu*Y_chr7_rev**	CCGTTCCTGCCTTGCCTGATGACTA
***Alu*Y_chr10a_fw**	TCGGGCTATCACAGCCTGAGCAAGC
***Alu*Y_chr10a_rev**	TAAAAATCTTCCCAAATCACACCAC
***Alu*Y_chr10b_fw**	GTTACCGATGGGGCCAGTAG
***Alu*Y_chr10b_rev**	CCAGTACAGACCAGAGACCAG
***Alu*Sx_chr10_fw**	TAGCCACTTAGCCAGTAACA
***Alu*Sx_chr10_rev**	GAGCGCTGAATTCGTTTAGT
***Alu*Sp_chr17_fw**	CCCTACGTGTTTTTAAATCACCAA
***Alu*Sp_chr17_rev**	CTCAAGTGGGTGAAGGTGTCTG
***Alu*Y_chr22_fw**	TTAGCTGAGCAGGCCGGGCG
***Alu*Y_chr22_rev**	AGTTGGTGCAGAGACACCCCCCACC

Primers per la mutagenesi della B box	Sequenze dei primers (5'→ 3')
***Alu*Sq2_chr1_Bmut_fw**	AGGTCAGGAGTCGGAGACCAGCCT
***Alu*Sq2_chr1_Bmut_rev**	GGCTGGTCTCCGACTCCTGACC
***Alu*Sx_chr1_Bmut_fw**	GAGATCAGGAGTCGGAGAGCAGCCT
***Alu*Sx_chr1_Bmut_rev**	AGGCTGCTCTCCGACTCCTGATCTC
***Alu*Sx1_chr3_Bmut_fw**	GGTCAGGAGTCGGAGACCAGCCT
***Alu*Sx1_chr3_Bmut_rev**	AGGCTGGTCTCCGACTCCTGACC
***Alu*Y_chr7_Bmut_fw**	CGAGGTCAGGAGACGGAGACCAGCCTG
***Alu*Y_chr7_Bmut_rev**	AGGCTGGTCTCCGTCTCCTGACCT
***Alu*Y_chr10a_Bmut_fw**	GAGGTCATGAGACGGAGACCATCCTG
***Alu*Y_chr10a_Bmut_rev**	CCAGGATGGTCTCCGTCTCATGACCTCG
***Alu*Y_chr10b_Bmut_fw**	GGTCAGCCGACGGAGACCAACC
***Alu*Y_chr10b_Bmut_rev**	GGTTGGTCTCCGTCGGCTGACC
***Alu*Sx_chr10_Bmut_fw**	GGTTAGGAGTCGGAGACCAGCC
***Alu*Sx_chr10_Bmut_rev**	GGCTGGTCTCCGACTCCTAACC
***Alu*Sp_chr17_Bmut_fw**	GGTCAGGAGTCGGAGACCAGCCT
***Alu*Sp_chr17_Bmut_rev**	AGGCTGGTCTCCGACTCCTGACC
***Alu*Y_chr22_Bmut_fw**	AGGTCAGGAGACGGAGACCATCCT
***Alu*Y_chr22_Bmut_rev**	AGGATGGTCTCCGTCTCCTGACCT

Primers per la delezione della regione a monte	Sequenze dei primers (5'→ 3')
***Alu*Sq2_chr1_del(-12)_fw**	AGTGTCACTTCCGGCCAGGC
***Alu*Sx_chr1_del(-12)_fw**	ATAGTTACTTTAGGCCGGGCG
***Alu*Sx1_chr3_del(-12)_fw**	ACCTCAAAGTTGGGCCGG
***Alu*Y_chr7_del(-12)_fw**	TACAGCGCGTTTGGCCGGAC
***Alu*Y_chr10-a_del(-12)_fw**	AATGATGTTAATGGCCGGGC
***Alu*Y_chr10-b_del(-12)_fw**	CAAACTTGAAGAGCCGGG
***Alu*Sx_chr10_del(-15)_fw**	AGCGGCTGGTTGGTAGG
***Alu*Sx_chr10_del(-15)_fw**	TACAATGCTGCAGCCGGG
***Alu*Y_chr22_del(-12)_fw**	GTTCTGGAGGTGGGCCGGGCG

• È riportata la sequenza di *Alu*Sq2 (1-265 nt in figura) allineata alla sequenza *consensus* di *Alu*Yb8 (8-287 nt in figura):



In verde è indicata la A box, in azzurro la B box e in grigio il motivo interno A_5TACA_6 presente in *Alu*Yb8 (e che normalmente hanno tutti gli elementi *Alu*) il quale è sostituito da A_3G nell'elemento in esame.

Il lavoro descritto in questa parte della tesi di dottorato ha portato alla pubblicazione dell'articolo: Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014)[373].

Parte II

Analisi delle alterazioni del trascrittoma cellulare causate dall'espressione di *Alu* RNA

Come accennato nella prima sezione della presente tesi di dottorato, la strategia messa a punto per identificare *Alu* RNA espressi in diversi contesti cellulari, permetterà di analizzare a fondo il controllo trascrizionale degli elementi *Alu*, cosa molto importante in quanto stanno emergendo evidenze del fatto che la loro espressione potrebbe essere connessa con diverse patologie. Poter quindi identificare un sottoinsieme di elementi differenzialmente espressi tra una condizione tumorale e la corrispondente condizione normale, potrà far concentrare l'attenzione sui membri di tale sottogruppo come marcatori di un determinato stato cellulare (se, per esempio, sono sovraespressi nelle cellule tumorali rispetto alle "normali", essi potrebbero essere coinvolti nell'insorgenza e progressione del cancro).

Numerosi studi indicano che i ncRNA possono regolare l'espressione genica attraverso il silenziamento genico trascrizionale, e modificazioni dello stato della cromatina (da aperta a chiusa) dipendenti da variazioni di modificazioni istoniche e metilazione del DNA[135]. Si è visto che alcuni lncRNA (una classe di ncRNA i cui trascritti sono più lunghi di 200 nt, v. Introduzione) sono in grado di riorganizzare la cromatina[180,182,184], altri influenzano la stabilità dei loro RNA bersaglio mediante un appaiamento senso-antisenso tra elementi *Alu* al loro interno[193], ed altri ancora possono agire da "spugne" per i miRNA esercitando un livello addizionale di regolazione post-trascrizionale[195].

Dal momento che gli elementi *Alu* possono interferire con i normali processi cellulari e che vi sono sporadici esempi del fatto che tali elementi sono espressi in condizioni patologiche[69,376], è ragionevole supporre che la loro espressione possa indurre cambiamenti significativi nel trascrittoma cellulare e che tali cambiamenti possano essere diversi a seconda del sottoinsieme di elementi *Alu* espressi.

Lo studio della deregolazione del trascrittoma cellulare, in risposta alla sovraespressione di elementi *Alu*, si presenta come una via non esplorata ma molto interessante per capire come possono agire tali elementi, e quindi definire i meccanismi molecolari della loro azione, anche in relazione alla possibilità che essi contribuiscano allo sviluppo di differenti patologie.

Risultati e discussione

Scelta degli elementi *Alu* da sovraesprimere e loro trasfezione transiente in cellule HeLa

Si è deciso di sovraesprimere due diversi elementi *Alu* in cellule HeLa (le stesse cellule in cui si è eseguita la loro trascrizione *in vitro*, v. sezione precedente), al fine di valutare gli effetti di una loro sovraespressione nel contesto cellulare e poter individuare i meccanismi di azione alla base di tali effetti. Innanzitutto è bene precisare che si è scelto di trasfettare due retroelementi *Alu* caratterizzati da due tipi di localizzazione maggiormente studiati nella prima parte della presente tesi, ovvero elementi *Alu* intergenici ed intronici antisenso.

I due elementi selezionati, sulla base delle analisi bioinformatiche riportate nella sezione precedente, sono l'elemento intergenico *Alu*Sq2_ chr1: 61523296-61523586, il quale è risultato espresso in 5 diverse linee cellulari delle 7 esaminate (H1-hESC, HeLa S3, Hep G2, K562, NHEK) e che manca del motivo conservato interno A_5TACA_6 (sostituito da A_3G), e *Alu*Y_chr22:41932115-41932411, il quale non è stato trovato espresso in nessuna delle linee cellulari analizzate ma che risulta interessante in quanto posto all'interno dell'introne 2 del gene per la subunità RPC8 dell'enzima RNA polimerasi III, in orientamento antisenso (importante in quanto potrebbe influenzare l'espressione di tale gene, dal momento che gli elementi *Alu* sono in grado di interferire con la normale maturazione dei pre-mRNA, essendo fonti di siti di splicing alternativo[377]). I due elementi *Alu* sono stati amplificati da DNA genomico, subclonati nel vettore pGEM-T Easy e successivamente clonati nel vettore di espressione pSUPER.basic, il quale presenta un promotore forte riconosciuto dalla Pol III (il promotore del gene H1) e consente la loro espressione ad alti livelli *in vivo*, non resa possibile dal loro debole promotore interno, efficiente solo *in vitro*. All'interno del vettore è stato clonato strettamente il corpo degli elementi, dal presunto TSS al primo terminatore forte T_4 a valle. I costrutti sono stati trasfettati in cellule HeLa per 48 ore e la scelta del tempo si è basata sulla letteratura corrente[199,378-380], nonostante non vi siano numerosi esempi di trasfezione transiente di elementi *Alu*. Normalmente, infatti, vengono sovraespressi geni codificanti per proteine o per piccoli RNA non codificanti, ma, nonostante ciò, ci si è fatti guidare da casi riportati di trasfezioni o con lo stesso vettore o nello stesso tipo cellulare. In parallelo alle trasfezioni dei due costrutti di elementi *Alu*, è stato anche realizzato un controllo negativo della trasfezione con il plamide pSUPER.basic vuoto, per poter monitorare gli effetti della sola trasfezione sul trascrittoma cellulare (già la sola introduzione di un plasmide perturba il contesto cellulare, in quanto fenomeno non fisiologico). L'esperimento è stato condotto indipendentemente tre volte e si è anche realizzato un controllo non trasfettato, per valutare

eventuali effetti della trasfezione sulla vitalità cellulare. Al termine delle 48 ore, infatti, si è eseguita una misura della vitalità cellulare necessaria per essere certi che gli effetti osservati, successivamente sul trascrittoma, fossero imputabili alla sovraespressione e non al fatto che le cellule fossero, in qualche modo, perturbate. La percentuale di mortalità osservata (intorno al 10%-25%) permette di concludere che l'inserimento dei plasmidi non determina perturbazioni nelle cellule, in quanto tale percentuale corrisponde a quella che si osserva normalmente in seguito alla trasfezione con lipofectamine (**Tabella 3**). Pertanto, si è potuto procedere con l'analisi dell'avvenuta sovraespressione dei due *Alu* RNA.

NOME CAMPIONE	**CELLULE VIVE**	**% CELLULE MORTE**
Ctrl non trasfettato	18.75 milioni	2%
pSUPER Ø (1)	5.4 milioni	16%
*Alu*Sq2_pSUPER (1)	1.7 milioni	18%
*Alu*Y_pSUPER (1)	3.69 milioni	8%
pSUPER Ø (2)	3.15 milioni	24%
*Alu*Sq2_pSUPER (2)	1.7 milioni	13%
*Alu*Y_pSUPER (2)	1.89 milioni	12%
pSUPER Ø (3)	4.05 milioni	16%
*Alu*Sq2_pSUPER (3)	1.08 milioni	10%
*Alu*Y_pSUPER (3)	2.47 milioni	16%

Tabella 3. Tabella in cui vengono riportati i campioni trasfettati (tra parentesi è indicato il numero corrispondente al primo, secondo e terzo replicato biologico), il numero di cellule vive dopo la trasfezione e la percentuale di cellule morte. La misurazione è stata condotta utilizzando il colorante Trypan Blue.

Analisi dell'avvenuta sovraespressione di *Alu* RNA mediante RT-PCR semiquantitativa

Una volta verificato che la maggior parte delle cellule fosse vitale (le percentuali di cellule morte, riportate in **Tabella 3**, sono in accordo con la tossicità indotta dalle lipofectamine utilizzate per le tresfezioni), si è estratto l'RNA totale dalle stesse e si è retrotrascritto utilizzando il kit *iScript cDNA Synthesis Kit* (Bio-Rad). Per ogni campione, alle reazioni di trascrizione inversa, è stata affiancata una reazione in assenza di retrotrascrittasi (controllo no RT) in modo da poter evidenziare, nelle successive reazioni di PCR, eventuali contaminazioni di DNA genomico, che avrebbero potuto falsare l'attendibilità del dato. Si sono eseguite, pertanto, delle RT-PCR semiquantitative con

primers per ognuno dei due elementi *Alu* trasfettati (progettati sia sulla loro sequenza distintiva, detta 3'trailer, sia nel loro corpo centrale; v. Appendice II) e, per verificare che la sovraespressione fosse avvenuta con successo, si è eseguita la stessa analisi anche sul gene endogeno per la β2 microbulina (B2M) che, essendo un gene costitutivo, è espresso ad alti livelli. Avere trascritti *Alu* singoli, normalmente espressi a bassissimi livelli[54], con un grado di espressione pari a quello di un gene housekeeping, è parso un buon modo per confermare l'avvenuta sovraespressione.

Per quanto concerne il trascritto *Alu*Sq2, si sono ottenute evidenze della sua sovraespressione (rispetto alla normale espressione basale) sia con la coppia di primers sul 3'trailer (**Figura 25A**) sia sul corpo centrale (**Figura 25B**), e confrontando anche la sua amplificazione con quella del gene B2M (**Figura 25C**).

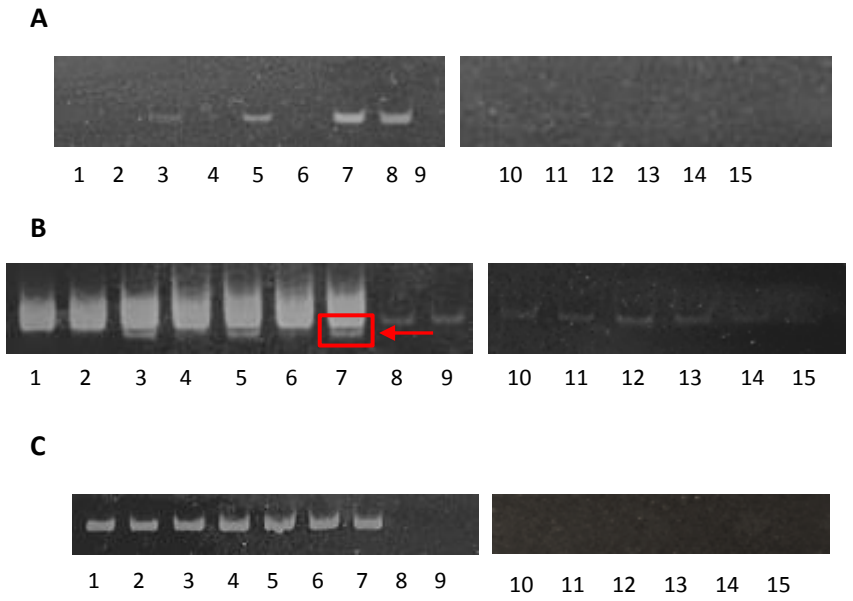


Figura 25. RT-PCR semiquantitative eseguite per valutare la sovraespressione del costrutto *Alu*Sq2_pSUPER. Sono riportate le amplificazioni realizzate mediante l'utilizzo di primers progettati sul 3'trailer di *Alu*Sq2 (specifici per l'identificazione del trascritto trasfettato) **(A)**, sul corpo dell'elemento (tali primers si appaiano anche a tutti gli elementi *Alu* cellulari autonomi e ospiti di trascritti Pol II) **(B)** e sul gene housekeeping B2M **(C)**. L'ordine di caricamento da sinistra per il primo gel (corsie 1-9, di ognuno dei tre gel riportati) è: controllo non trasfettato (Ctrl), trasfettato con pSUPERØ (1), trasfettato con *Alu*Sq2_pSUPER (1), trasfettato con pSUPERØ (2), trasfettato con *Alu*Sq2_pSUPER (2), trasfettato con pSUPERØ (3), trasfettato con *Alu*Sq2_pSUPER (3), no RT Ctrl non trasfettato, no RT pSUPERØ (1). L'ordine di caricamento da sinistra per il secondo gel (corsie 10-15, di ognuno dei tre gel riportati) è: no RT trasfettato con *Alu*Sq2_pSUPER (1), no RT trasfettato con pSUPERØ (2), no RT trasfettato con *Alu*Sq2_pSUPER (2), no RT trasfettato con pSUPERØ (3), no RT trasfettato con *Alu*Sq2_pSUPER (3), reazione in assenza di templato (controllo negativo della RT-PCR). L'amplificazione aspecifica del campione 8 potrebbe essere spiegata da una cross-contaminazione della singola reazione di RT-PCR, dal momento che le altre reazioni di controllo (no RT) non hanno portato, come atteso, ad alcuna amplificazione. Tra parentesi è indicato il numero corrispondente al primo, secondo e terzo replicato biologico dell'esperimento.

Come si può notare in **Figura 25A**, infatti, si ha un notevole incremento dell'espressione di *Alu*Sq2 (l'amplicone corrispondente è indicato in rosso in figura) se si confrontano le amplificazioni nei campioni trasfettati col costrutto in esame (3,5,7) rispetto al non trasfettato (1) o al trasfettato con plasmide vuoto (2,4,6), per ognuno dei tre replicati biologici. Lo stesso si verifica conducendo amplificazioni sul corpo dell'elemento (**Figura 25B**) e, dal momento che gli elementi *Alu* sono molto simili in sequenza soprattutto nella regione centrale (la quale è notevolmente conservata), è possibile anche osservare l'amplificazione di tutti gli *Alu* RNA cellulari (rappresentata da uno *smear* in **Figura 25B**). Ciò significa che, mediante l'utilizzo della tecnica di RT-PCR, non è possibile distinguere trascritti *Alu* autonomi da quelli ospiti di unità trascrizionali Pol II più lunghe (come riportato nella sezione precedente), ma, nonostante ciò, mancando *Alu*Sq2 del caratteristico motivo interno A_5TACA_6 (presente, invece, in tutti gli altri *Alu* RNA), è possibile rilevare il suo aumento di espressione nei campioni trasfettati, rispetto a tutti gli altri trascritti *Alu* cellulari, già confermato anche dalle RT-PCR in grado di identificare solo tale trascritto (corsie 3,5,7 in **Figura 25B**: oltre a tutti gli *Alu* RNA cellulari, si amplifica anche il trascritto in esame che, mancando di un elemento di sequenza, dà origine ad un amplicone di minori dimensioni (indicato in rosso in figura per il terzo replicato biologico, ma visibile per anche gli altri due), facilmente distinguibile su gel.

Un'ulteriore conferma dell'avvenuta sovraespressione dell'elemento *Alu*Sq2, a seguito della sua trasfezione per 48 ore in cellule HeLa, si è ottenuta confrontando la sua amplificazione (**Figura 25A**) con quella del gene housekeeping B2M (**Figura 25C**), come riportato sopra. Confrontando infatti le due amplificazioni, nonché le amplificazioni dell'*Alu*Sq2 RNA nel campione trasfettato rispetto al non trasfettato e trasfettato con plamide vuoto (**Figura 25A**), si osserva un incremento della sua espressione nelle cellule (maggiormente evidente nel terzo replicato biologico, corsia 7 in **Figura 25A**, rispetto al primo e al secondo, corsie 3 e 5 rispettivamente). Infine, essendo l'elemento *Alu*Sq2 intergenico, è possibile valutare i livelli della sua espressione basale *in vivo* (ovvero nelle cellule HeLa utilizzate), confrontando le amplificazioni ottenute nel campione trasfettato con quelle del controllo non trasfettato, utilizzando la coppia di primers sul 3'trailer (per ognuno dei tre replicati biologici dell'esperimento, **Figura 25A**). Tale confronto fa emergere quanto riportato ampiamente nell'Introduzione e confermato anche dalle nostre analisi[373], ovvero che gli *Alu* RNA in condizioni normali sono poco abbondanti: non si riesce, pertanto, a rilevare la loro amplificazione nelle condizioni sperimentali utilizzate, le stesse che, invece, mettono in evidenza un aumento della loro espressione, una volta trasfettati.

Analoghe osservazioni si possono fare per la trasfezione del costrutto *Alu*Y_pSUPER: anche in questo caso si è verificata la sovraespressione dell'*Alu*Y RNA mediante RT-PCR con coppie di primers sul suo caratteristico 3'trailer (**Figura 26A**), sul corpo centrale (**Figura 26B**) e confrontando la sua amplificazione con quella del gene endogeno B2M (**Figura 26C**). Per quanto concerne la sua amplificazione specifica, si osserva un buon aumento di espressione in tutti e tre i replicati biologici dell'esperimento, se si confrontano i campioni trasfettati (3,5,7) con i rispettivi trasfettati con il plasmide vuoto (2,4,6) e il controllo non trasfettato (1) (**Figura 26A**). Al contrario, le amplificazioni con la coppia di primers sul corpo centrale dell'elemento *Alu* non permettono di distinguerlo dagli altri *Alu* RNA cellulari (tale sequenza è completa e non può essere distinta dagli altri *Alu* RNA per la sua lunghezza, come invece nel caso di *Alu*Sq2), ma, nonostante ciò, il confronto tra la sua amplificazione (**Figura 26A**) e quella del gene endogeno (**Figura 26C**), rivela che la sua sovraespressione è avvenuta con successo. Confrontando, infine, i due elementi *Alu* trasfettati (**Figura 25A** vs **Figura 26A**), sembra che, a parità di tempo di trasfezione, si sia maggiormente espresso nelle cellule HeLa proprio *Alu*Y (le intensità della bande sono maggiori se si confrontano, a parità di condizioni di RT-PCR, i due campioni).

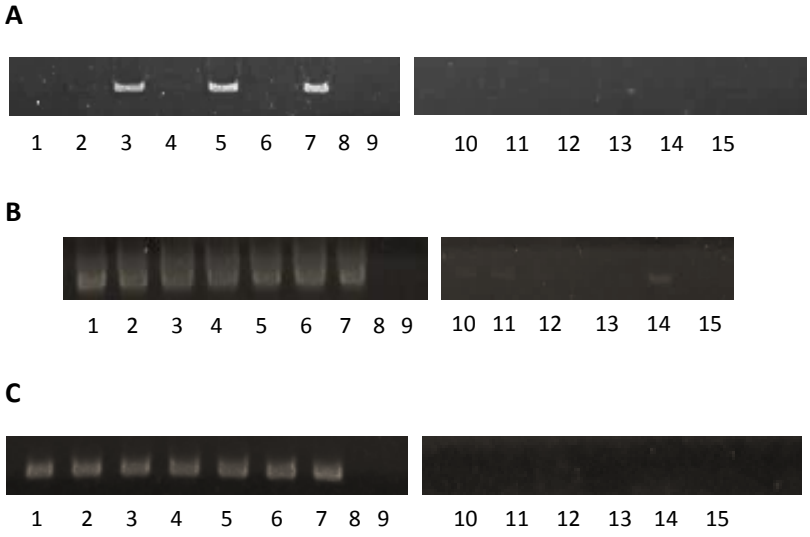


Figura 26. RT-PCR eseguite per valutare la sovraespressione del costrutto *Alu*Y_pSUPER. Sono riportate le amplificazioni realizzate mediante l'utilizzo di primers progettati sul 3'trailer di *Alu*Y (specifici per l'identificazione del trascritto trasfettato) **(A)**, sul corpo dell'elemento (tali primers si appaiano anche a tutti gli elementi *Alu* cellulari autonomi e ospiti di trascritti Pol II) **(B)** e sul gene housekeeping B2M **(C)**. L'ordine di caricamento da sinistra per il primo gel (corsie 1-9, di ognuno dei tre gel riportati) è: controllo non trasfettato (Ctrl), trasfettato con pSUPERØ (1), trasfettato con *Alu*Y_pSUPER (1), trasfettato con pSUPERØ (2), trasfettato con *Alu*Y_pSUPER (2), trasfettato con pSUPERØ (3), trasfettato con *Alu*Y_pSUPER (3), no RT Ctrl non trasfettato, no RT pSUPERØ (1).

L'ordine di caricamento da sinistra per il secondo gel (corsie 10-15, di ognuno dei tre gel riportati) è: no RT trasfettato con *Alu*Y_pSUPER (1), no RT trasfettato con pSUPERØ (2), no RT trasfettato con *Alu*Y_pSUPER (2), no RT trasfettato con pSUPERØ (3), no RT trasfettato con *Alu*Y_pSUPER (3), reazione in assenza di templato (controllo negativo della RT-PCR). Tra parentesi è indicato il numero corrispondente al primo, secondo e terzo replicato biologico dell'esperimento.

Infine, dal momento che *Alu*Y è un elemento intronico antisenso nel gene POLR3H, non è possibile rilevare i suoi livelli di espressione basale *in vivo*, come invece riportato per *Alu*Sq2. Negli esperimenti, infatti, si è utilizzato un kit di retrotrascrizione che si serve di primers oligo dT e *random examers* come innesco per la reazione, e ciò fa ottenere un pool di trascritti sia Pol II che Pol III. È stato sintetizzato, quindi, anche il pre-mRNA Pol II contenente tale elemento *Alu*Y intronico antisenso, e, nella successiva reazione di PCR, è stato amplificato, non permettendo di distinguerlo dall'amplificazione derivante dal vero e proprio *Alu*Y RNA autonomo, trascritto dall'RNA Pol III.
Mediante l'utilizzo di tale tecnica (RT-PCR) si è stati in grado, dunque, di rilevare, oltre all'avvenuta sovraespressione di entrambi gli elementi *Alu* trasfettati, anche l'espressione basale *in vivo* dell'elemento *Alu*Sq2 intergenico (espressione appena rilevabile nelle condizioni sperimentali in cui si rileva ampiamente la sua sovraespressione).
Dal momento che le RT-PCR hanno confermato l'avvenuta trasfezione di entrambi gli elementi *Alu*, si è proceduto con l'analisi degli effetti che la loro sovraespressione può avere sul trascrittoma sia codificante che non codificante (ovvero sugli mRNA e sui lncRNA).

Ibridazione dell'RNA totale ad uno Human LncRNA Expression Microarray

Al fine di valutare gli effetti della sovraespressione dei due *Alu* RNA sul trascrittoma cellulare, gli RNA totali estratti dalle cellule trasfettate (con plasmide vuoto e con i due elementi *Alu* separatamente), sono stati analizzati attraverso l'ibridazione allo Human LncRNA Expression Microarray V3.0 (Arraystar Inc.), il quale permette di tracciare un profilo di espressione di 26109 trascritti codificanti per proteine e di 30586 lncRNA. Dagli RNA totali sono stati rimossi gli RNA ribosomiali (rRNA, che costituiscono circa il 95% dei trascritti totali di una cellula e che, pertanto, possono alterare la qualità dell'ibridazione successiva), mediante l'utilizzo del kit mRNA-ONLY™ Eukaryotic mRNA Isolation Kit (Epicentre), in seguito tali RNA depleti sono stati retrotrascritti, e ogni cDNA è stato marcato con un fluoroforo per tutta la sua lunghezza. I cDNA marcati sono stati ibridati allo Human LncRNA Array v3.0 (8 x 60K, Arraystar) e si sono acquisite le immagini degli array, per le successive analisi di espressione genica. L'esperimento ha previsto l'impiego di nove

ibridazioni a micro-arrays, corrispondenti ad ognuno dei campioni in esame (trasfettato con plasmide pSUPER vuoto, trasfettato con *Alu*Sq2_pSUPER, trasfettato con *Alu*Y_pSUPER, per tre replicati biologici) e, una volta acquisite le immagini dei differenti arrays, è stata eseguita una normalizzazione quantilica dei dati per la loro successiva analisi.

Innanzitutto si è deciso di considerare come "*detected*" (rilevati) i trascritti che fossero presenti in almeno due dei tre replicati biologici. Detto ciò, si è osservato che circa il 70% dei trascritti, rispetto al totale di RNA presenti sugli array (56695), è stato rilevato nei campioni esaminati (corrispondente a 39978 trascritti). Inoltre, l'analisi condotta su tali 39978 trascritti individuati, ha messo in evidenza che 20537 corrispondono ai trascritti codificanti per proteine, mentre 19441 ai trascritti non codificanti di dimensioni maggiori di 200 nucleotidi, ovvero ai lncRNA.

Seguirà la trattazione delle analisi eseguite separatamente sulle due categorie di trascritti.

Analisi condotte sui trascritti di geni codificanti per proteine

Dopo aver effettuato la normalizzazione quantilica dei dati estratti dalle immagini degli array, si è condotta un'analisi sui trascritti codificanti per proteine, volta a mettere in luce gli effetti della sovraespressione dei due *Alu* RNA sulla loro espressione.

Tale indagine si è ritenuta di notevole interesse dal momento che vi sono numerose recenti evidenze, riportate ampiamente nell'introduzione, di come RNA non codificanti possano alterare l'espressione genica in molteplici modi[135,182,199]. Dato che i geni potenzialmente codificanti per gli *Alu* RNA sono più numerosi di qualsiasi altra famiglia di geni per RNA non codificanti, ci si è chiesti se anch'essi potessero influenzare l'espressione genica e, in caso di esito positivo, attraverso quali meccanismi molecolari.

Come riportato nel paragrafo precedente, dei 39978 trascritti identificati nei campioni in esame, 20537 sono mRNA, ovvero il 51% dei trascritti rilevati è codificante per proteine. In primo luogo si è risaliti ai geni corrispondenti a tali trascritti e si è osservato che, come atteso, dal momento che un gene può codificare per più isoforme dello stesso trascritto, il numero di geni corrispondenti ai 20537 mRNA è 15404. In seguito si è verificato se gli effetti provocati dalle due trasfezioni, sull'espressione di tali geni codificanti, fossero o meno simili. In un primo momento sono stati identificati i geni significativamente deregolati mediante un test LIMMA (Linear Models for Microarray data[381]) che ha messo in evidenza 342 e 593 trascritti down-regolati, nel caso della trasfezione con l'elemento *Alu*Sq2 e *Alu*Y, rispettivamente. Di questi, 151 sono risultati in comune tra i due esperimenti (ovvero i trascritti down-regolati in comune sono il 44% dei trascritti che

subiscono un decremento di espressione sovraesprimendo l'*Alu*Sq2 RNA). I geni up-regolati, al contrario, sono risultati essere 188 (54%) e 337 (57%) (nel caso della trasfezione con l'elemento *Alu*Sq2 e *Alu*Y rispettivamente), con una sovrapposizione del 22%. I due esperimenti sono stati quindi confrontati mediante analisi GSEA (Gene Set Enrichment Analysis, http://www.broadinstitute.org/gsea) utilizzando i geni up- e down-regolati dell'esperimento di trasfezione con *Alu*Sq2 come set di geni (gene sets). Tali gene sets sono stati confrontati con il profilo completo di tutti i 15404 geni della trasfezione con l'elemento *Alu*Y, e si sono messi in evidenza i geni deregolati in entrambi gli esperimenti. Le nostre analisi hanno messo in luce che gli insiemi dei geni significativamente deregolati sono statisticamente molto simili (P-value <0.01 e FDR < 0.01%) in entrambe le trasfezioni con gli elementi *Alu* (**Figura 27**): questa osservazione è importante in quanto *Alu*Sq2 manca di un breve tratto di sequenza presente in tutti gli altri elementi *Alu* (il motivo di collegamento tra il braccio sinistro e il braccio destro) ed esisteva la possibilità che questa variazione nella sua struttura secondaria potesse influenzare le proprietà funzionali di questo *Alu* RNA. Inoltre, si è anche osservato che il numero di geni down-regolati è quasi il doppio di quello dei geni up-regolati, in entrambe le sovraespressioni (**Figura 27**), suggerendo che l'effetto preponderante di un aumento degli elementi *Alu* nelle cellule, sia di interferenza "negativa" sull'espressione del trascrittoma codificante.

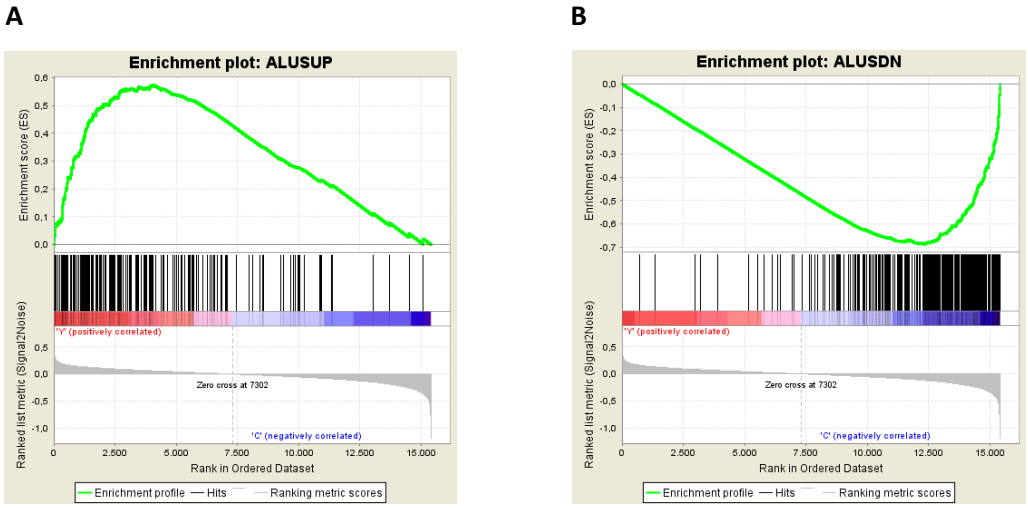


Figura 27. Grafici di arricchimento (*Enrichment plots*)che visualizzano in chiave comparativa la deregolazione (up- o down-) dei geni codificanti per proteine in risposta alla sovraespressione dei due diversi elementi *Alu*. Come si può osservare i geni sono deregolati in modo concorde (o aumenta o diminuisce la loro espressione) e lo dimostra il fatto che quando sono up-regolati nell'esperimento della trasfezione con *Alu*Sq2 (barrette nere) lo sono anche nell'esperimento della trasfezione con *Alu*Y (regione rossa) **(A)** e viceversa quando sono down-regolati **(B)**. Inoltre i geni down-regolati sono decisamente molti di più dei geni up-regolati (lo si nota bene confrontando la densità di barrette nere nei due grafici di arricchimento).

Creazione di gene sets di geni codificanti per proteine e loro analisi

L'osservazione riportata nel paragrafo precedente, ovvero la generale tendenza alla repressione dei geni codificanti per proteine in entrambi gli esperimenti di trasfezione, rimanda ad un possibile silenziamento genico ad opera di *Alu* RNA abbondanti nelle cellule. Per valutare se tale inibizione dell'espressione dei geni codificanti per proteine fosse imputabile alla presenza di elementi *Alu* nelle loro regioni non tradotte, si sono creati dei gene sets di tali geni, suddivisi in modo da poter analizzare separatamente quelli con almeno un elemento *Alu* senso o antisenso nella propria 5'UTR e, analogamente, quelli con almeno un elemento *Alu* senso o antisenso nella propria 3'UTR (v. Materiali e metodi per la creazione dei gene sets di geni; senso e antisenso sono da riferirsi rispetto alla direzione della trascrizione del gene Pol II ospite degli elementi *Alu*).

I gene sets di geni con almeno un elemento *Alu* senso o antisenso nella 5'UTR mostrano che il numero di trascritti, corrispondenti ai requisiti richiesti, è molto basso (solo 43 dei 20537 trascritti, ovvero circa lo 0.2%, presentano almeno un elemento *Alu* senso o antisenso nella propria 5'UTR).

Il 90,7% dei trascritti (ovvero 39 trascritti di 43 totali) ha solo un elemento *Alu* nella 5'UTR, ed in particolare 20 lo presentano in orientamento antisenso e 19 in orientamento senso. Dei rimanenti 4 trascritti, invece, 3 mostrano due elementi *Alu* e uno 3 elementi *Alu*, in entrambi i casi in orientamento antisenso.

Per quanto concerne, invece, i geni corrispondenti a mRNA con almeno un elemento *Alu* senso o antisenso nella 3'UTR, si è osservato che essi corrispondono a 2867 trascritti (corrispondenti al 14% degli mRNA totali identificati), ed in particolare il numero totale di elementi *Alu* presenti nella loro 3'UTR è 4855, di cui 2429 antisenso e 2426 senso. È importante sottolineare che i nostri risultati sono in accordo con quanto finora riportato, ovvero gli elementi *Alu* risiedono normalmente negli introni dei geni codificanti per proteine e, del 5% presente nelle regioni codificanti, l'82% è presente nelle 3'UTR, il 14% nelle 5'UTR e il 4% nelle regioni codificanti tradotte[382,383]. Ciò spiega il maggior numero di trascritti con almeno un elemento *Alu* nella 3'UTR rispetto alla 5'UTR.

Classificazione degli mRNA in base al numero di elementi *Alu* nella 3'UTR

I trascritti contenenti elementi *Alu* nella 3'UTR sono stati ulteriormente classificati in gene sets che potessero distinguere non solo quelli con un elemento *Alu* da quelli con più elementi *Alu*, ma anche quelli con elementi *Alu* in entrambi gli orientamenti (riassunti nel gene set chiamato genericamente "3'_ALU"), quelli con solo elementi *Alu* antisenso (3'_ANTIALU) e quelli con solo

elementi *Alu* senso (3'_SENSEALU) (**Figura 28**). La prima categoria di trascritti, 3'_ALU, può contenere anche gli mRNA della seconda e della terza categoria (un trascritto può, infatti, contenere nella 3'UTR un elemento *Alu* senso e un elemento *Alu* antisenso[384]).

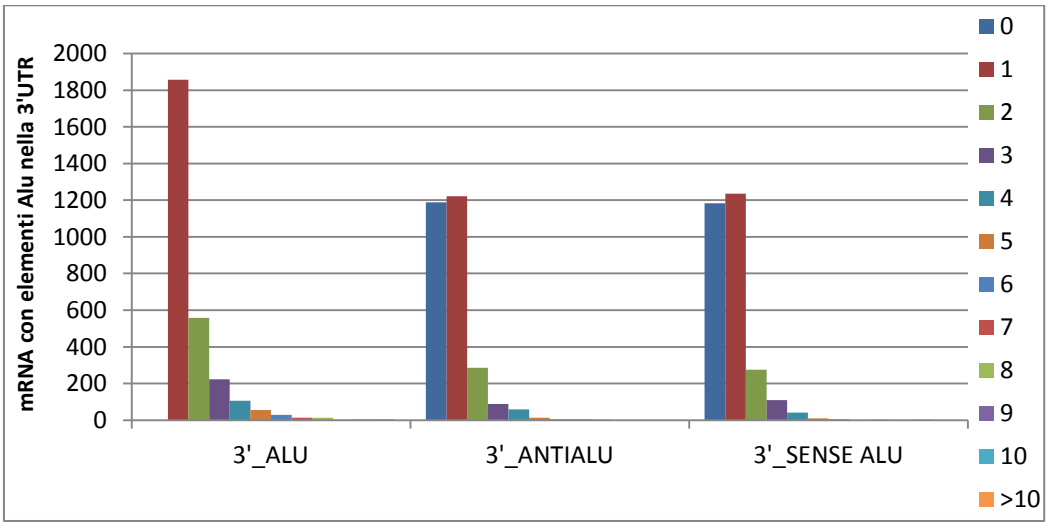


Figura 28. Rappresentazione grafica dei trascritti con elementi *Alu* nella loro 3'UTR. Tali RNA sono suddivisi in base al numero di elementi *Alu* senso, antisenso o entrambi (riassunti nel gene set chiamato "3'_ALU"), solo antisenso (3'_ANTIALU) e solo senso (3'_SENSEALU).

Come si può osservare in **Figura 28**, tutti i trascritti analizzati presentano almeno un elemento *Alu* nella propria 3'UTR (sottolineato, nel gene set "3'_ALU", dall'assenza di trascritti privi di una sequenza *Alu*), come riportato sopra, e, ciò ha reso possibile la realizzazione di sottoinsiemi all'interno dei gene sets riportati (3'_ALU, 3'_ANTIALU, 3'_SENSEALU, **Figura 28**), per poter analizzare in maggior dettaglio gli effetti degli *Alu* RNA sovraespressi sul trascrittoma codificante.

Pertanto, si sono suddividi i trascritti in base alla presenza di più di un elemento *Alu* nella loro 3'UTR ed, in particolare, si è concentrata l'attenzione su mRNA che presentassero almeno 3 o almeno 4 elementi *Alu* (senso o antisenso in diverse combinazioni, riportati nelle categorie 3'_ALU_H4, 3'_ANTIALU_H3, 3'_ANTIALU_H4, 3'_SENSEALU_H3, 3'_SENSEALU_H4, in **Tabella 4**). L'analisi è stata poi ulteriormente ristretta ai trascritti con solo almeno 3 o 4 elementi *Alu* in orientamento antisenso (riporati nelle categorie 3'_ANTIALUONLY_H3 e 3'_ANTIALUONLY_H4, in **Tabella 4**), nell'ottica di un appaiamento tra trascritti complementari (ovvero tra l'*Alu* RNA trasfettato senso e gli elementi *Alu* nelle 3'UTR in antisenso). Risalendo, poi, ai geni corrispondenti ai trascritti così classificati, si sono creati i gene sets sui quali poter condurre l'analisi di

arricchimento (GSEA; Gene Set Enrichment Analysis, http://www.broadinstitute.org/gsea), che verrà trattata nel paragrafo successivo.

	numero trascritti	numero geni
3'_ALU	2867	2258
3'_ALU_H4	228	99
3'_ANTIALU	1678	1331
3'_ANTIALU_H3	171	133
3'_ANTIALU_H4	82	60
3'_ANTIALU_ONLY	1183	933
3'_ANTIALUONLY_H3	82	64
3'_ANTIALUONLY_H4	37	24
3'_SENSEALU	1684	1331
3'_SENSEALU_H3	172	140
3'_SENSEALU_H4	63	57
5'_ALU	45	37
5'_ANTIALU	25	22
5'_SENSEALU	20	15

Tabella 4. Tabella degli mRNA con elementi *Alu* nella propria 3'UTR. I triangoli neri mettono in evidenza i gene sets corrispondenti ai trascritti con almeno 3 o 4 elementi *Alu* nella 3'UTR: dei gene sets "ALU", "SENSEALU" o "ANTIALU" fanno parte elementi sia senso che antisenso (per esempio nella categoria 3'_ALU_H4 vi sono i trascritti con un gruppo di almeno 4 *Alu*, le quali possono essere senso o antisenso secondo diverse combinazioni, mentre nella categoria 3'_ANTIALU_H3 vi sono i trascritti con un insieme di almeno 3 *Alu*, di cui almeno una in antisenso). Le categorie indicate con "ONLY" (messe in evidenza dai trangoli gialli) , al contrario, sono formate solo da elementi antisenso (sono riportati gli mRNA con solo elementi *Alu* antisenso su cui si è concentrata l'analisi). Ad esempio, nella categoria 3'_ANTIALUONLY_H3 sono raggruppati gli mRNA con solo elementi *Alu* antisenso in un gruppo di almeno 3 *Alu*, mentre nella categoria 3'_ANTIALUONLY_H4, sono presenti i trascritti con solo elementi *Alu* antisenso in un insieme di almeno 4 elementi. Nella tabella sono riportati anche i gene sets di mRNA con elementi *Alu* nella propria 5'UTR, per i quali si è eseguita la GSEA (triangoli verdi).

La scelta di suddividere i trascritti in base alla presenza di almeno 3 *Alu* nella 3'UTR è motivata dal fatto che si voleva mettere in luce, qualora presente, una correlazione diretta tra la deregolazione del trascrittoma codificante e il numero di elementi *Alu* nella 3'UTR, nell'ottica di un maggior effetto dovuto alla presenza di un maggior numero di tali retroelementi.

GSEA (Gene Set Enrichment Analysis) sui geni codificanti per proteine

Per valutare la possibilità che la presenza di più elementi *Alu* nelle regioni non tradotte in 3' possa influenzare l'espressione dei trascritti, è stata eseguita un'analisi GSEA (Gene Set Enrichment Analysis, http://www.broadinstitute.org/gsea) sui geni corrispondenti ai trascritti appartenenti ai gene sets riportati nel paragrafo precedente (**Tabella 4**), confrontando ogni esperimento di sovraespressione (con *Alu*Sq2 o *Alu*Y) con il controllo corrispondente alla trasfezione con il vettore vuoto. L'analisi ha rivelato che la presenza di elementi *Alu* nella 5'UTR non influenza significativamente il trascrittoma codificante, mentre gli mRNA con almeno 3 o 4 elementi *Alu* nella 3'UTR (di cui almeno uno è in antisenso) sono significativamente down-regolati (P-value < 0.01 e FDR< 5%) in entrambi gli esperimenti di trasfezione (v. i gene sets 3'_ANTIALU_H3 e 3'_ANTIALU_H4 in **Tabella 4** e in **Tabella 5**). Inoltre, rivolgendo l'attenzione agli mRNA aventi solo elementi *Alu* antisenso nelle loro regioni non tradotte in 3', si è osservato che i trascritti con solo elementi *Alu* in antisenso sono significativamente down-regolati quando sono presenti 3 o 4 elementi (**Tabella 5**, gene set 3'_ANTIALUONLY_H3 e 3'_ANTIALUONLY_H4), cosa che non si verifica se gli elementi sono in orientamento senso (**Tabella 5,** gene sets 3'_SENSEALU_H3, 3'_SENSEALU e 3'_ALU), eccezion fatta per i trascritti appartenenti al gene set 3'ALU_H4 (**Tabella 5**). Tali mRNA contenenti almeno 4 elementi *Alu* (di cui, però, nessun elemento potrebbe essere antisenso), infatti, risultano significativamente down-regolati nonostante possano avere 4 elementi *Alu* senso nella loro 3'UTR.

Tale down-regolazione risulta statisticamente significativa sia nelle cellule HeLa trasfettate con *Alu*Sq2 (**Tabella 5A**) che con *Alu*Y (**Tabella 5B**), portando ancora una volta ad osservare effetti simili nei due esperimenti. I geni down-regolati a seguito delle sovraespressioni dei due elementi *Alu*, pertanto, nonostante la loro differenza in sequenza, sono per la maggior parte comuni (**Figura 27**) e subiscono un decremento significativo qualora siano presenti più siti di legame per gli elementi *Alu* sovraespressi, nelle 3'UTR degli mRNA bersaglio (**Tabella 5**).

A

NAME	NES	P-value	FDR
3'_ANTIALU_H4	-1,57531	0,005252101	1,9
3'_ANTIALU_H3	-1,50539	0	2,9
3'_ANTIALUONLY_H3	-1,70806	0	0,4
3'_ALU_H4	-1,48599	0	2,7
3'_ANTIALUONLY_H4	-1,72933	0,001153403	0,5
3'_SENSEALU_H3	-1,28085	0,048241206	13,7
3'_SENSEALU	-1,23394	0	16,7
3'_ALU	-1,13868	0,002	24,2

B

NAME	NES	P-value	FDR
3'_ANTIALU_H4	-1,5409336	0,010964912	2,5
3'_ANTIALU_H3	-1,5584453	0	3,0
3'_ANTIALUONLY_H3	-1,5881761	0,00440044	3,0
3'_ALU_H4	-1,4526526	0	4,7
3'_ANTIALUONLY_H4	-1,609045	0,00969697	4,9
3'_SENSEALU_H3	-1,0965668	0,29856852	39,2
3'_SENSEALU	-1,1103501	0,038	40,8
3'_ALU	-1,0436839	0,194	51,5

Tabella 5. Tabelle in cui sono riassunti i risultati della GSEA sui geni codificanti per proteine nell'esperimento di sovraespressione di *Alu*Sq2 **(A)** e di *Alu*Y **(B)**, classificati secondo quanto riportato in **Tabella 4**. La linea tratteggiata separa i risultati statisticamente significativi (P-value < 0.01 e FDR < 5%) da quelli non statisticamente significativi: al di sopra della soglia si hanno i trascritti significativamente down-regolati sia nell'esperimento di trasfezione con *Alu*Sq2 **(A)** che con *Alu*Y **(B)** e, come si può notare, contengono tutti almeno 3 elementi *Alu* (la categoria 3'ANTIALU_H4 sta ad indicare un cluster di 4 elementi *Alu* nella 3'UTR dei trascritti, nel quale almeno un elemento *Alu* è in antisenso; le categorie contrassegnate con "ONLY" indicano che vi sono solo elementi *Alu* in antisenso in un insieme di almeno 3 o 4 elementi).

In seguito, è stata eseguita un'ulteriore analisi volta a identificare eventuali correlazioni tra la presenza, nella 3'UTR degli mRNA, di elementi di una determinata famiglia *Alu*, e la famiglia di appartenenza dei trascritti sovraespressi. In altri termini si voleva mettere in luce se i trascritti contenenti elementi *Alu* della famiglia *Alu*S fossero maggiormente down-regolati nell'esperimento di sovraespressione dell'*Alu*Sq2 RNA, e similmente per i trascritti contenenti *Alu* della famiglia Y (nel caso della trasfezione con il costrutto pSUPER_*Alu*Y). Innanzitutto si è proceduto attribuendo

la famiglia di appartenenza agli elementi *Alu* nella 3'UTR dei 2867 mRNA con almeno un elemento *Alu* nella 3'UTR. In seguito tali trascritti sono stati suddivisi in base non solo alla famiglia di provenienza e all'orientamento degli elementi *Alu*, ma anche in base alla presenza di almeno 2 elementi (senso, antisenso o entrambi) nella loro 3'UTR (gene sets indicati dai triangoli neri in **Tabella 6A**) realizzando un'analisi GSEA sui gene sets così ottenuti, addizionati ai gene sets precedenti (ovvero i gene sets in **Tabella 4**, riportati anche in **Tabella 6A,** a sinistra).

I risultati delle analisi hanno evidenziato che non vi sono effetti maggiori sui trascritti contenenti elementi *Alu* della stessa famiglia dell'elemento trasfettato (**Tabella 6B**).

A

	numero trascritti
3'_ALU	2867
3'_ALU_H4	228
3'_ANTIALU	1678
3'_ANTIALU_H3	171
3'_ANTIALU_H4	82
3'_ANTIALU_ONLY	1183
3'_ANTIALUONLY_H3	82
3'_ANTIALUONLY_H4	37
3'_SENSEALU	1684
3'_SENSEALU_H3	172
3'_SENSEALU_H4	63

	numero trascritti
ALUJ	1004
ALUJ_ANTI	516
ALUJ_ANTI_H3	29
ALUJ_H3	46
ALUS	2127
ALUS_ANTI	1220
ALUS_ANTI_H3	73
ALUS_ANTI_ONLY	879
ALUS_H3	223
ALUS_ONLY_ANTI_ONLY	736
ALUS_ONLY_ANTI_ONLY_H2	98
ALUS_ONLY_ANTI_ONLY_H3	26
ALUY	466
ALUY_ANTI	227

B

Sovraespressione AluSq2

NAME	NES	P-value	FDR
3'_ANTIALUONLY_H3	-1,7080582	0	0,7
3'_ANTIALUONLY_H4	-1,7293273	0,001153403	1,0
3'_ANTIALU_H4	-1,5753103	0,005252101	3,4
3'_ALU_H4	-1,485988	0	3,6
ALUS_ONLY_ANTI_ONLY_H3	-1,4964877	0,041966427	3,8
3'_ANTIALU_H3	-1,5053928	0	4,2
ALUS_ANTI_H3	-1,5144811	0,007399577	4,9
ALUS_H3	-1,4330319	0,001003009	5,2
3'_SENSEALU_H3	-1,2808484	0,048241206	16,7
3'_SENSEALU	-1,2339422	0	21,1

Sovraespressione AluY

NAME	NES	P-value	FDR
3'_ANTIALU_H4	-1,5409336	0,010964912	4,5
3'_ANTIALU_H3	-1,5584453	0	5,1
3'_ANTIALUONLY_H3	-1,5881761	0,00440044	5,4
3'_ALU_H4	-1,4526526	0	8,2
3'_ANTIALUONLY_H4	-1,609045	0,00969697	8,3
ALUS_ANTI_H3	-1,4065008	0,027654868	9,7
ALUS_ONLY_ANTI_ONLY_H3	-1,3667998	0,08734177	11,4
ALUS_H3	-1,2628552	0,05263158	22,8

Tabella 6. (A) Tabelle in cui sono riassunti i gene sets utilizzati per l'analisi GSEA: l'analisi è stata eseguita sia sui geni riportati in Tabella 4 (**riportati anche in Tabella 6A, a sinistra**) che sui geni in cui sono state evidenziate le famiglie di appartenenza degli elementi *Alu* nella 3'UTR (**Tabella 6A, a destra**). **(B)** Tabelle in cui sono riassunti i risultati della GSEA sui geni codificanti per proteine nell'esperimento di sovraespressione di *Alu*Sq2 (**Tabella 6B, a sinistra**) e di *Alu*Y (**Tabella 6B, a destra**), classificati secondo quanto riportato Tabella 6A. La linea tratteggiata separa i risultati statisticamente significativi (P-value < 0.01 e FDR < 10%) da quelli non statisticamente significativi: al di sopra della soglia si hanno i trascritti significativamente down-regolati

sia nell'esperimento di trasfezione con *Alu*Sq2 (**Tabella 6B, a sinistra**) che con *Alu*Y (**Tabella 6B, a destra**) e, come si può notare, contengono tutti almeno 3 elementi *Alu* (la categoria 3'ANTIALU_H4 indica un insieme di almeno 4 elementi *Alu* nella 3'UTR dei trascritti, nel quale almeno un elemento *Alu* è in antisenso; le categorie contrassegnate con "ONLY" indicano che vi sono solo elementi *Alu* in antisenso). Non vi sono effetti maggiori sui trascritti contenenti elementi *Alu* della stessa famiglia dell'elemento trasfettato (nel caso della trasfezione con *Alu*Y non è significativamente down-regolato nessun trascritto contenente elementi *Alu*Y, mentre lo sono alcuni RNA contenenti almeno 3 *Alu*S di cui almeno uno in antisenso: vedi categoria ALUS_ANTI_H3, triangolo giallo).

Inoltre tali analisi per famiglie di elementi *Alu*, hanno anche rivelato che gli elementi della sottofamiglia di *Alu*S sono più frequentemente presenti nelle 3'UTR di tutti i trascritti codificanti (quando presenti, gli elementi *Alu*S rappresentano l'86% di *Alu* senso nella 3'UTR di un trascritto e il 90% di *Alu* antisenso), e ciò potrebbe rispecchiare la loro maggiore frequenza nel genoma, dal momento che gli elementi *Alu*S sono anche a livello genomico le più abbondanti (**Figura 29**).

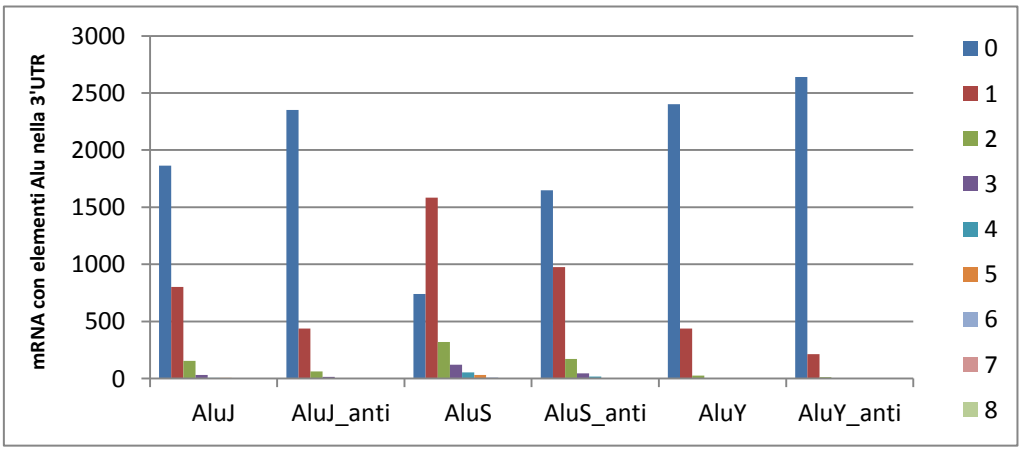


Figura 29. Rappresentazione grafica dei trascritti con elementi *Alu* nella loro 3'UTR. Gli mRNA sono suddivisi in base alla famiglia di appartenenza degli elementi *Alu* (*Alu*J, *Alu*S e *Alu*Y), all'orientamento di tali elementi (senso: AluJ, AluS, AluY; antisenso: antiAluJ, antiAluS, anti AluY) e al numero di elementi presenti (da 0 a 8, vedi legenda in figura).

Le osservazioni sopra riportate, sono rilevanti in quanto suggeriscono fortemente l'esistenza di un meccanismo antisenso alla base della diminuzione dei livelli dei trascritti, e quindi un nuovo livello di regolazione genica post-trascrizionale dipendente dall'appaiamento di *Alu* RNA alla 3'UTR di mRNA bersaglio. Ciò è in accordo con quanto riportato in un recente lavoro[199] nel quale la sovraespressione di un sottogruppo di elementi *Alu* (DR2 *Alu*, il quale è formato dai retroelementi che ospitano le caratteristiche ripetizioni dirette esanucleotidiche, tipiche dei siti di legame per i recettori nucleari dell'acido retinoico) porta ad un decremento dei livelli di mRNA che presentano un elemento *Alu* nella loro 3'UTR, con i quali il DR2 *Alu* RNA si appaia. Inoltre in tale lavoro è messo

anche in risalto il fatto che gli *Alu* RNA, derivanti dalla trascrizione dei corrispondenti elementi DR2 *Alu* trasfettati, possono essere processati e loro porzioni più piccole (dette riRNA, repeat-induced RNA) possono comunque portare ad una riduzione dei loro mRNA bersaglio, con sequenze complementari nella loro 3'UTR. Pertanto si può ipotizzare che anche nel nostro contesto sperimentale, gli elementi *Alu* sovraespressi possano subire un processamento e ridurre l'espressione degli mRNA bersaglio come trascritti di minori dimensioni, anche se tale ipotesi non è ancora stata ad oggi verificata. Nel medesimo lavoro, inoltre, viene anche osservato che la trascrizione del 2-3% delle 100 mila-200 mila ripetizioni DR2 *Alu*, poste in prossimità di unità trascrizionali Pol II attive, viene innescata dal legame dei recettori dell'acido retinoico (RAR) per produrre RNA Pol III-dipendenti. Tali trascritti vengono in seguito processati nel citoplasma in sequenze lunghe 28-65 nt mediante un processo, non ancora completamente delineato, che richiede l'intervento di DICER (una endoribonucleasi coinvolta nella biogenesi dei miRNA) e, successivamente questi piccoli prodotti, collettivamente chiamati riRNA, sono stabilizzati dal legame di AGO3 (una proteina Argonauta responsabile, in quanto tale, dell'azione catalitica del complesso RISC - RNA-Induced Silencing Complex) e guidano il complesso verso sequenze bersaglio complementari nella 3'UTR di mRNA, con conseguente loro degradazione. Tale degradazione sembra dipendere dalla proteina EDC4 (Enhancer of mRNA-DeCapping protein 4), la quale appartiene al complesso di decapping, si trova spesso in associazione ad AGO3, e porta alla degradazione XRN1-dipendente dell'mRNA bersaglio. Sembra dunque che gli eventi di regolazione che coinvolgono la trascrizione Pol III-dipendente degli elementi DR2 *Alu* possano complementare i programmi di regolazione Pol II-dipendenti. Infatti, non solo i RAR inducono uno specifico programma di trascrizione dipendente dall'RNA Pol II (determinando l'espressione di proteine importanti per il processo di differenziamento, nel contesto cellulare esaminato), ma attivano anche la trascrizione Pol III-dipendente di numerosi elementi DR2*Alu* che fiancheggiano tali unità trascrizionali Pol II. Queste evidenze hanno iniziato quindi a rivelare aspetti inattesi della regolazione dei programmi di sviluppo indotti da ligando (in questo caso dall'acido retinoico) in cellule staminali embrionali. Lo studio riportato mette in luce come i DR2 *Alu* RNA full-length, nonché i riRNA, possono essere sovraespressi in presenza di acido retinoico e possono agire a seguito della trascrizione (processo post-trascrizionale) portando alla degradazione di mRNA che presentano sequenze complementari nella propria 3'UTR. Inoltre, sono riportate evidenze della degradazione dei trascritti a seguito dell'intervento del sistema di *RNA interference* (di cui DICER e AGO3 sono componenenti fondamentali). Pertanto, avendo osservato anche noi una down-regolazione di mRNA con elementi *Alu* nella 3'UTR complementari agli *Alu* RNA sovraespressi, ciò

suggerisce concretamente che il meccanismo alla base dei nostri effetti possa essere il medesimo, e che quindi possano intervenire DICER e AGO3 nel processo che conduce alla degradazione dei trascritti codificanti per proteine. La degradazione di *Alu* RNA da parte di DICER è stata anche documentata da Kaneko e collaboratori[385], i quali hanno messo in luce che DICER1 è in grado di idrolizzare gli *Alu* RNA *in vitro*, e che ciò è in accordo con l'accumulo di trascritti *Alu* in cellule epiteliali del pigmento della retina, in cui vi è un deficit nell'espressione di DICER1. L'ipotesi meccanicistica formulata è comunque concorde con quanto riportato anche per altri RNA non codificanti, come i miRNA, i quali, appaiandosi al loro mRNA bersaglio secondo una logica senso-antisenso, ne promuovono la degradazione (oltre ad inibirne la traduzione).

Inoltre, sono riportate anche evidenze di lncRNA in grado di interferire con la normale maturazione degli mRNA portando alla loro degradazione, mediante un processo che inizia proprio con l'appaiamento senso-antisenso tra elementi *Alu*[193]. Nel lncRNA è presente una sequenza *Alu* che si associa ad una sequenza *Alu* nella 3'UTR dell'mRNA bersaglio, mediante un appaiamento non perfettamente complementare. Tale appaiamento imperfetto genera una struttura secondaria intermolecolare con una protrusione riconosciuta e legata dalla proteina STAU1, che a sua volta recluta la proteina UPF1, richiesta per il processo di degradazione dell'mRNA (mRNA decay).

Tale meccanismo potrebbe contribuire al decremento degli mRNA da noi osservato, in quanto la struttura a doppio filamento di RNA, formatasi a seguito dell'appaiamento tra l'*Alu* RNA trasfettato e l'elemento *Alu* nella 3'UTR degli mRNA down-regolati, potrebbe essere riconosciuta da STAU1 e potrebbe essere innescato il meccanismo di degradazione degli mRNA riportato da Gong e collaboratori[193]. Inoltre, l'ipotesi che STAU1 riconosca strutture a doppio filamento di RNA (dsRNA) è stata ulteriormente documentata dagli stessi autori, i quali hanno messo in luce la degradazione di mRNA che si appaiano mediante elementi *Alu* nelle loro 3'UTR[386], e avvalora l'ipotesi che gli elementi *Alu* possano portare alla degradazione degli mRNA appaiandosi ad elementi complementari nella loro 3'UTR.

Le nostre evidenze sperimentali, inoltre, aggiungono anche un importante tassello per la conoscenza dei meccanismi di azione degli *Alu* RNA nel contesto cellulare. Vi sono, infatti, solo alcune sporadiche evidenze del fatto che i trascritti con due elementi *Alu* invertiti nelle loro 3'UTR risultano down-regolati[382,387,388], e che gli mRNA possono essere degradati mediante appaiamenti o a lncRNA o ad altri mRNA che coinvolgono elementi *Alu* (vedi sopra). Recentemente, infine, uno studio ha messo in luce gli effetti sull'espressione del trascrittoma dovuti alla trasfezione di un *Alu* RNA in cellule da adenocarcinoma al senso (linea cellulare MCF7): gli autori riportano una diminuzione della vitalità cellulare e un'alterazione nell'espressione di alcuni geni coinvolti nella

risposta allo stress del reticolo endoplasmatico, ma non si soffermano su un possibile meccanismo di azione dell'*Alu* RNA sovraespresso correlabile con la deregolazione degli mRNA osservata[389]. La nostra analisi ha messo in luce che il decremento dei trascritti con almeno 3 elementi *Alu* antisenso nella loro 3'UTR, è statisticamente significativo rispetto ai trascritti con elementi *Alu* senso, e ciò è interessante dal momento che nel nostro sistema sperimentale è stata riprodotta la condizione di sovraespressione di *Alu* RNA normalmente riscontrata a seguito di diversi stimoli (stress[68], shock termico, infezione virale, azione di agenti ipometilanti) e in alcuni tumori umani[69,200,340,376,385].

Il fatto che siano maggiormente colpiti dalle nostre trasfezioni tali trascritti, e che una condizione di sovraespressione di elementi *Alu* sia stata riportata in sporadiche forme tumorali, potrebbe far ipotizzare che i bersagli che subiscono un decremento nel nostro contesto sperimentale, siano anche i primi ad essere influenzati in un contesto patologico.

Analisi condotte sui trascritti non codificanti per proteine (lncRNA)

Oltre a quanto riportato finora per gli mRNA, si sono condotte anche analisi sui lncRNA, volte a definire possibili correlazioni o differenze, nella loro risposta alle sovraespressioni di elementi *Alu*, rispetto agli effetti osservati sui trascritti codificanti per proteine. Analogamente a quanto riportato per gli mRNA, si è deciso di considerare trascritti in cui fosse presente almeno un elemento *Alu* senso o antisenso contenuto nella forma matura (post-splicing) dei lncRNA. Ciò ha portato a escludere dall'indagine circa 2028 lncRNA per i quali non erano note con precisione le posizioni dei loro esoni e per i quali, quindi, poteva essere dubbia la localizzazione di elementi *Alu*. L'analisi dei rimanenti 17413 lncRNA, ha permesso di definire il sottoinsieme di trascritti di interesse, corrispondente a 1648 trascritti su 17413 (ovvero il 9.5%) con almeno un elemento *Alu* senso o antisenso nei propri esoni. Anche in questo caso, essendo consistente il numero di trascritti, si è potuto suddividere l'insieme in sottoinsiemi in base alla localizzazione degli elementi *Alu*. In particolare, si sono distinti i lncRNA contenenti almeno una sequenza *Alu* (senso o antisenso, o entrambi, riassunti nel gene set "exons_Alus"), almeno un elemento *Alu* antisenso (exons-antiAlus) e almeno un elemento *Alu* (exons_senseAlus) (**Figura 30**).

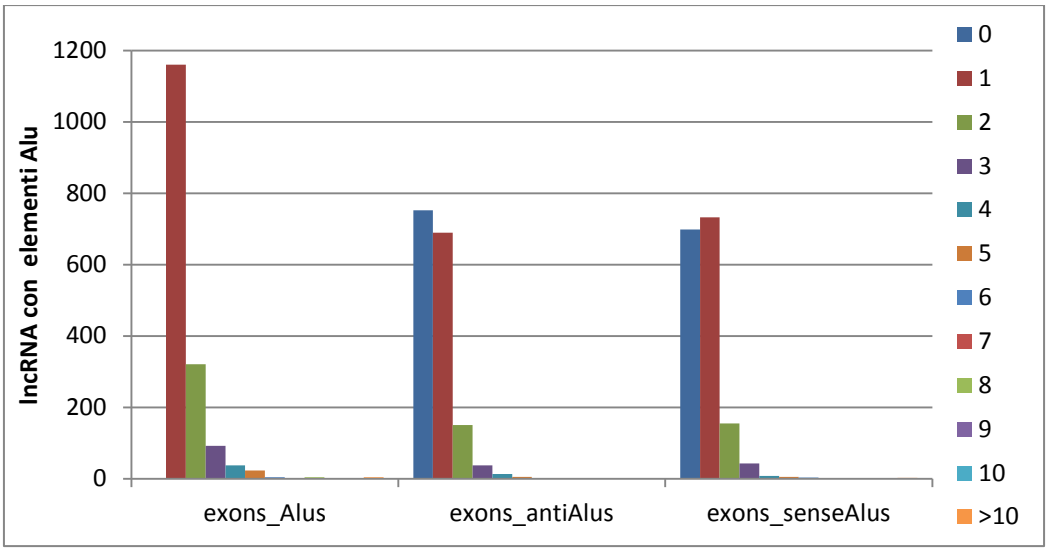


Figura 30. Rappresentazione grafica dei lncRNA contenenti almeno un elemento *Alu* nella loro forma matura: senso, antisenso o indiferentemente l'uno o l'altro (riassunti nel gene set exons_Alus), solo in orientamento antisenso (exons_antiAlus), solo in orientamento senso (exons_senseAlus).

Come si può notare in **Figura 30**, sia all'interno del gene set "exons_antiAlus" che del gene set "exons_senseAlus", la maggior parte dei lncRNA contiene un solo elemento *Alu* (689 lncRNA su 1648 trascritti totali in antisenso, e 732 su 1648 senso) e i trascritti con 1, 2 e 3 elementi *Alu* presentano una distribuzione simile fra loro, indipendentemente dall'orientamento delle sequenze *Alu* (in altri termini il numero di trascritti con 1,2,3 elementi *Alu* è molto simile se si considerano sia i lncRNA con elementi antisenso che senso). Sono stati creati, quindi, anche gene sets più specifici, realizzati considerando i trascritti contenenti almeno 2, 3, 4 *Alu* (in orientamento senso o antisenso o indifferentemente l'uno o l'altro)(**Tabella 7**). Tali classificazioni sono state eseguite tenendo presenti le evidenze ottenute analizzando gli mRNA, per i quali si era osservata una down-regolazione dei trascritti contenenti nella loro 3'UTR un insieme di elementi *Alu* antisenso, piuttosto che elementi *Alu* senso.

	numero trascritti
EXONS_ALUS	1648
EXONS_ALUS_H4	75
EXONS_ANTIALUS	896
EXONS_ANTIALUS_H3	57
EXONS_ANTIALUS_H4	20
EXONS_ANTIALUSONLY	698
EXONS_ANTIALUSONLY_H3	20
EXONS_SENSEALUS	950
EXONS_SENSEALUS_H2	218
EXONS_SENSEALUS_H3	63
EXONS_SENSEALUS_H4	20
EXONS_SENSEALUSONLY	752
EXONS_SENSEALUSONLY_H2	149
EXONS_SENSEALUSONLY_H3	29

Tabella 7. Tabella dei lncRNA con elementi *Alu* nella loro forma matura (post-splicing). I triangoli neri mettono in evidenza i gene sets corrispondenti ai trascritti con almeno 2, 3 o 4 elementi *Alu*: dei gene sets "ALU" o "ANTIALU" fanno parte elementi sia senso che antisenso (per esempio nella categoria EXONS_ALUS_H4 vi sono i trascritti con almeno 4 elementi *Alu*, i quali possono essere senso o antisenso secondo diverse combinazioni, nel gene set EXONS_ANTIALUS_H3 vi sono i trascritti con almeno 3 *Alu* antisenso e che possono inoltre contenere altri elementi senso, mentre nel gene set EXONS_SENSEALUS_H2 vi sono i trascritti con almeno 2 *Alu* senso e che possono inoltre contenere altri elementi antisenso). I gene sets indicati con "ONLY" (messi in evidenza dai triangoli gialli), al contrario, sono formati solo da elementi antisenso o solo da elementi senso (sono riportati i lncRNA con solo elementi *Alu* antisenso o solo elementi *Alu* senso). Pertanto, per esempio nel gene set EXONS_ANTIALUSONLY_H3 vi sono i lncRNA con solo elementi *Alu* antisenso in numero di almeno 3 *Alu*, mentre nella categoria EXONS_SENSEALUSALUONLY_H4, sono presenti i trascritti con solo elementi *Alu* senso in numero di almeno 4 elementi.

GSEA condotta sui trascritti non codificanti per proteine

Una volta realizzati i gene sets di lncRNA si è condotta, analogamente a quanto riportato sopra, l'analisi GSEA che ha permesso di mettere in evidenza che lncRNA, contenenti almeno 2 o 3 elementi *Alu* in orientamento senso, sono risultati significativamente down-regolati in entrambi gli esperimenti di sovraespressione (v. gene sets indicati con triangoli neri in **Tabella 8**). Inoltre l'analisi, ha anche messo in luce che i trascritti con solo elementi *Alu* senso in un gruppo di almeno due elementi (gene set EXONS_SENSEALUSONLY_H2, indicato da un triangolo giallo in **Tabella 8**) sono significativamente down-regolati (nel caso della sovraespressione dell'*Alu*Y RNA si ha una down-regolazione significativa anche per i lncRNA con solo elementi senso in un gruppo di almeno 3 sequenze *Alu*, gene set EXONS_SENSEALUSONLY_H3 indicato da un triangolo giallo in **Tabella 8**) e che, al contrario di quanto osservato per gli RNA codificanti per proteine, i lncRNA contenenti

elementi *Alu* antisenso, non sono significativamente down-regolati (FDR > 5%) e, pertanto, non si possono considerare tali trascritti per ipotizzare un possibile meccanismo di azione degli *Alu* RNA sovraespressi sulla modulazione dell'espressione del trascrittoma non codificante (**Tabella 8**). Dal momento che non si può ipotizzare un meccanismo di appaiamento diretto senso-antisenso con i trascritti degli elementi *Alu* trasfettati, il decremento di espressione di tali RNA non codificanti potrebbe essere frutto di un altro tipo di meccanismo.

A

NAME	SIZE	NES	P-value	FDR
EXONS_SENSEALUS_H2	213	-1,70287	0	3,4
EXONS_SENSEALUSONLY_H2	146	-1,58203	0	6,1
EXONS_ALUS_H4	72	-1,95263	0	2,6
EXONS_SENSEALUS_H3	62	-1,71271	0,006969	4,7
EXONS_SENSEALUSONLY_H3	29	-1,32459	0,09697	12,7
EXONS_SENSEALUS	934	-1,4452	0	9,4
EXONS_ALUS	1598	-1,45631	0	10,8
EXONS_ANTIALUS	860	-1,37294	0	13,0
EXONS_SENSEALUSONLY	738	-1,36925	0	11,5
EXONS_ANTIALUSONLY	664	-1,27574	0	14,1
EXONS_ANTIALUSONLY_H3	20	-1,29317	0,159292	14,1
EXONS_ANTIALUS_H3	56	-1,03383	0,351351	38,1
EXONS_ANTIALUS_H4	19	-1,10006	0,307018	30,3
EXONS_SENSEALUS_H4	19	-0,73974	0,836735	89,2

B

NAME	NES	P-value	FDR
EXONS_SENSEALUS_H2	-1,77183	0	1,6
EXONS_SENSEALUSONLY_H2	-1,73672	0	1,6
EXONS_ALUS_H4	-1,82491	0	2,2
EXONS_SENSEALUS_H3	-1,89585	0	3,7
EXONS_SENSEALUSONLY_H3	-1,5852	0,026596	3,7
EXONS_SENSEALUS	-1,54148	0	3,9
EXONS_ALUS	-1,5455	0	4,3
EXONS_ANTIALUS	-1,45558	0	5,0
EXONS_SENSEALUSONLY	-1,45831	0	5,6
EXONS_ANTIALUSONLY	-1,38091	0	7,4
EXONS_ANTIALUSONLY_H3	-1,23541	0,183463	14,4
EXONS_ANTIALUS_H3	-1,15286	0,220779	20,4
EXONS_ANTIALUS_H4	-0,81854	0,707775	81,3
EXONS_SENSEALUS_H4	0,728493	0,851735	94,6

Tabella 8. Tabelle in cui sono riassunti i risultati della GSEA condotta sui geni codificanti per lncRNA, suddivisi secondo quanto riportato in Tabella 7. La linee tratteggiate separano i risultati statisticamente significativi (P-value < 0.01 e FDR < 5%) da quelli non statisticamente significativi: al di sopra della soglia (rappresentata dalla

linea nera tratteggiata) si hanno i trascritti significativamente down-regolati sia nell'esperimento di trasfezione con *Alu*Sq2 **(A)** sia nell'esperimento di trasfezione con *Alu*Y **(B)**. Come si può notare, i lncRNA contengono tutti almeno 2 elementi *Alu* (triangoli neri;il gene set EXONS_ALUS_H4 indica un gruppo di 4 elementi *Alu*; le categorie contrassegnate con "ONLY" (triangoli gialli) indicano che vi sono solo gruppi di elementi *Alu* senso o antisenso).

Per spiegare il decremento di espressione di lncRNA contententi elementi *Alu* senso, è necessario tener presente che vi sono alcune evidenze, seppur controverse, del fatto che gli elementi *Alu* possano interagire con i miRNA come loro possibili bersagli. Analisi *in silico* hanno, infatti, rivelato l'esistenza di miRNA contenenti una regione seme in 5' complementare a sequenze *Alu* localizzate nella 3'UTR di mRNA[390], e avvalorano la possibilità di un'interazione funzionale e regolativa tra i due tipi di trascritti non codificanti. Inoltre, a sostegno di tale ipotesi, è stato osservato che le differenti sequenze target dei miRNA sono conservate anche tra elementi *Alu* di differenti sottofamiglie.

Interessante è anche quanto riportato da Lehnert e collaboratori[391], ovvero che gli mRNA target di miRNA sono più ricchi di sequenze *Alu* in orientamento senso che antisenso, suggerendo che questo meccanismo di regolazione possa essere specifico per elementi *Alu* senso. Il legame di complessi SRP9/14 a sequenze *Alu* in orientamento senso potrebbe ostacolare la degradazione miRNA-dipendente dei trascritti contenenti tali elementi[392] (**Figura 31A**). In questo contesto un aumento dei livelli degli *Alu* RNA potrebbe al contempo determinare il sequestro dei complessi SRP9/14 e rendere accessibili i siti target dei miRNA nei trascritti contenenti elementi *Alu*, come i lncRNA contenenti *Alu* senso in esame, i quali, attraverso i miRNA, potrebbero risultare destabilizzati (**Figura 31B**). Tale meccanismo, schematizzato in **Figura 31**, potrebbe spiegare l'osservato decremento di espressione di lncRNA con almeno un elemento *Alu* senso nella regione matura (post splicing).

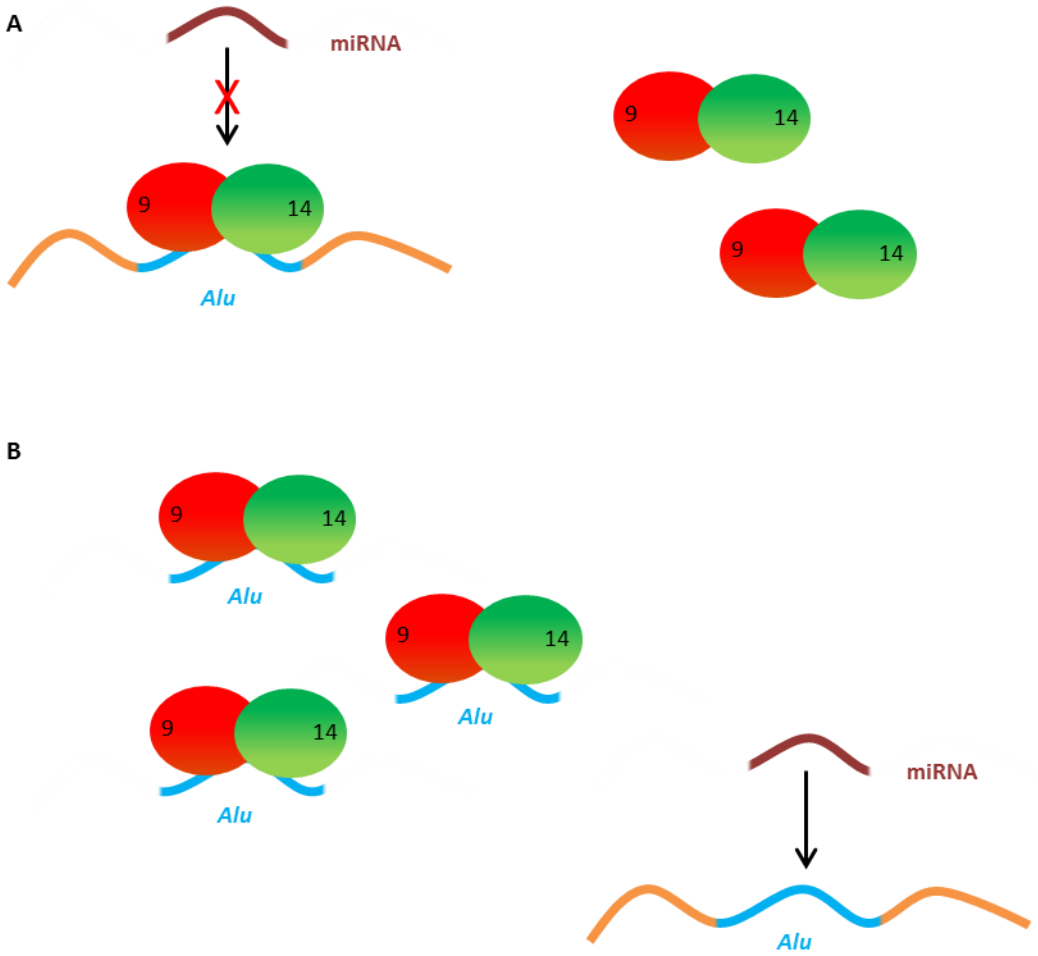


Figura 31. Rappresentazione del meccanismo di down-regolazione ipotizzato per i lncRNA. In condizioni normali il complesso SRP9/14 potrebbe legare l'elemento *Alu* senso nel lncRNA con conseguente impedimento della degradazione miRNA-dipendente del trascritto non codificante **(A)**. Al contrario, gli *Alu* RNA sovraespressi potrebbero sequestrare i complessi SRP9/14 e permettere la destabilizzazione miRNA-dipendente del lncRNA, con conseguente down-regolazione **(B)**.

Analisi dei processi cellulari in cui sono coinvolti gli mRNA e i lncRNA regolati negativamente in risposta alla sovraespressione di *Alu* RNA

Date le osservazioni riportate finora, per valutare in quali processi biologici fossero implicati gli mRNA e i lncRNA maggiormente down-regolati, si è eseguita un'analisi delle loro caratteristiche funzionali, volta a far maggior chiarezza in merito agli effetti della sovraespressione di *Alu* RNA nelle cellule Hela utilizzate. Tale analisi si è ritenuta necessaria anche a fronte del lavoro di Baryakin e collaboratori in cui viene riportato che la sovraespressione di un *Alu* RNA provoca

un'alterazione nell'espressione di alcuni geni coinvolti nella risposta allo stress del reticolo endoplasmatico in cellule MCF7[389].

L'analisi GSEA è stata utilizzata anche per individuare il significato regolativo della risposta alle sovraespressioni di *Alu*, in particolare per valutare, sulla base di misure di alterazione dell'espressione genica, quali processi cellulari potrebbero essere significativamente alterati da tali perturbazioni. Per facilitarne la lettura, i risultati di GSEA sono stati visualizzati mediante l'utilizzo dell'interfaccia grafica di Cytoscape, utilizzando il programma Enrichment Map[393,394]. In questo tipo di visualizzazione, ogni nodo del network rappresenta un processo biologico (pathway), la cui dimensione è proporzionale al numero di geni che lo compongono, il colore rappresenta il valore di "Normalized Enrichment Score"(NES) della GSEA (pathway up-regolati in rosso e down-regolati in blu), e le connessioni rappresentano i geni in comune tra due pathway. Lo spessore delle connessioni è proporzionale al numero di geni in comune. Con questa visualizzazione è anche possibile confrontare i risultati di GSEA provenienti da due esperimenti (nelle nostre analisi sono stati confrontati gli esperimenti di sovraespressione dei due elementi *Alu*). In questo caso il nodo ha due colori (un colore per il bordo e uno per il centro), che rappresentano, appunto, i due esperimenti. È inoltre possibile aggiungere nodi che contengano geni di intessere (in giallo), per poter visualizzare la loro connessione con i singoli pathway (in magenta).

Nelle nostre analisi sono stati utilizzati anche gli insiemi di geni codificanti per proteine e per lncRNA contenenti elementi *Alu* (in giallo in **Figura 32**), descritti in precedenza, significativi all'analisi GSEA (**Tabella 5 e Tabella 8**). In particolare, per poter valutare i processi biologici che risentono per primi della sovraespressione degli elementi *Alu* e della down-regolazione dei lncRNA, si è eseguita un'analisi GSEA indiretta. Ciò si è reso necessario dal momento che non esistono pathway catalogati per i trascritti non codificanti, pertanto i gene sets per la visualizzazione delle connessioni con i pathway, sono stati creati dai gene sets originali degli mRNA "sovrapposti" con i lncRNA. Per mRNA "sovrapposti" con i lncRNA si intendono tutti i trascritti non codificanti per proteine all'interno degli mRNA, che vengono trascritti e riportati nelle annotazioni come fossero "isoforme" degli mRNA ospite. In **Figura 32** sono mostrati i risultati della GSEA condotta su tutti i trascritti codificanti per proteine utilizzando come database di gene sets i pathway (CP, *canonical pathways* disponibile nel Molecular Signature Database, MSigDB, http://www.broadinstitute.org/gsea/msigdb); tale confronto, ha permesso di mettere in evidenza la funzione biologica dei geni maggiormente deregolati a seguito delle sovraespressioni dei due elementi *Alu* selezionati (in **Figura 32** sono riportati soltanto i pathway con FDR < 1% in GSEA).

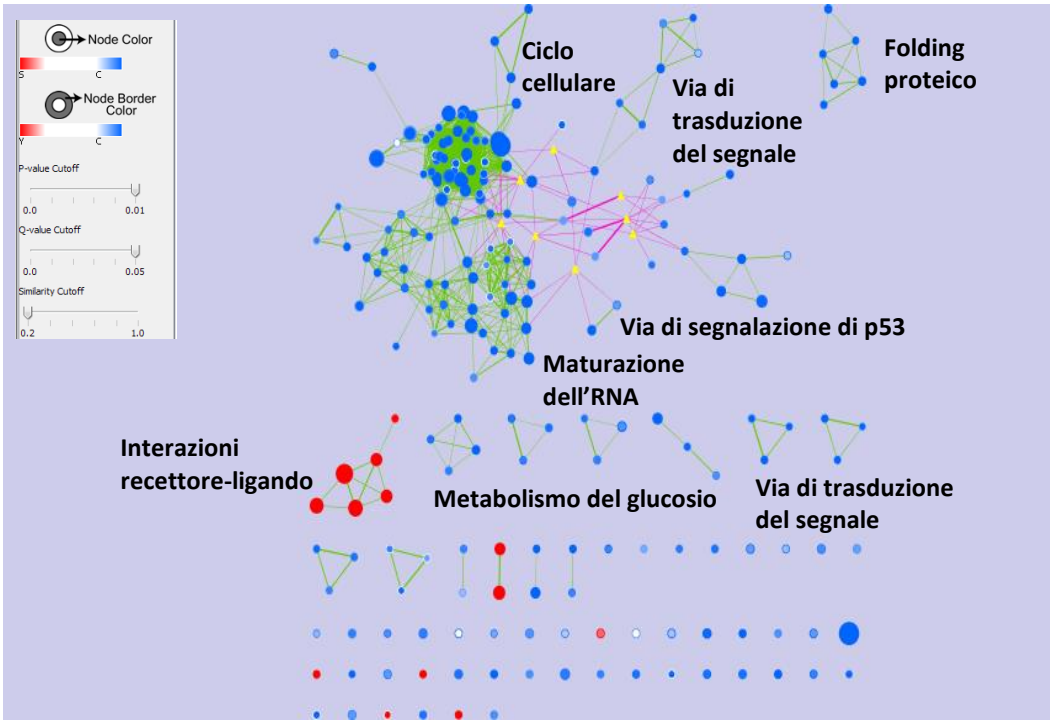


Figura 32. Rappresentazione dei geni codificanti per proteine up (rosso) e down-regolati (blu) nei due esperimenti di sovraepressione mediante l'interfaccia grafica di Cytoscape. La maggior parte dei geni risulta down-regolata e gli effetti sono concordi nei due esperimenti, infatti non si ha una differenza cromatica apprezzabile tra il centro e il bordo di ogni nodo (il centro di ogni nodo corrisponde al pathway nella trasfezione con *Alu*Sq2, mentre il bordo corrisponde al pathway nella trasfezione con *Alu*Y, vedi legenda a lato della figura). In giallo sono riportati gli insiemi di geni i cui mRNA contengono elementi *Alu* nella propria 3'UTR o perché contengono un lncRNA contenente, a sua volta, elementi *Alu*, che sono risultati significativi in GSEA (v. **Tabelle 5 e 8**) e le connessioni in magenta mettono in evidenza quali pathway vengono probabilmente influenzati per primi dalle sovraepressioni per la presenza di elementi *Alu* nei loro geni. Le connessioni in magenta rappresentano la significatività al Fisher's exact test (con soglia di significatività impostata al 5%).

L'analisi dei pathway mette in luce la down-regolazione prevalente rispetto alla up-regolazione. Tale down-regolazione interessa i trascritti codificanti per proteine appartenenti agli stessi processi biologici (rappresentato dai nodi blu sia nel bordo (trasfezione con *Alu*Y) che nel centro (trasfezione con *Alu*Sq2)), portando a concludere che le due sovraespressioni generano effetti molto simili (come confermato anche dalla analisi riassunta in **Figura 27**). Il confronto tra i gene sets dei pathway e i gene sets dei trascritti contenenti elementi *Alu* (rappresentati dai triangoli gialli in **Figura 32**) ha messo in luce che questi ultimi sono direttamente connessi (linee magenta in **Figura 32**) con processi cellulari fondamentali come il ciclo cellulare, la maturazione dei trascritti

sia codificanti che non codificanti e la via di segnalazione di p53 (portando alla down-regolazione dei suoi interattori MDM2, PPMD1) (**Figura 33**). Le connessioni riportate potrebbero suggerire che le categorie funzionali che risentono per prime dell'influenza della sovraespressione di *Alu* RNA siano quelle coinvolte nei processi cellulari, riportati sopra (**Figura 33**). Tali categorie, infatti, condividono dei geni con i gene sets da noi creati (le connessioni sono rappresentate dalle linee magenta) e, pertanto, presentano geni contenenti almeno un elemento *Alu* nella 3'UTR o nella regione matura di un lncRNA.

L'aumento di trascritti *Alu* nelle cellule HeLa potrebbe, quindi, portare *in primis* ad una down-regolazione dei trascritti contenti elementi *Alu*, ed in particolare, dei trascritti implicati nel ciclo cellulare e nella maturazione degli mRNA e dei lncRNA, processi alla base del funzionamento cellulare fisiologico. Ciò è in accordo con l'entità della perturbazione e con il fatto che tali processi siano fondamentali per il corretto funzionamento cellulare.

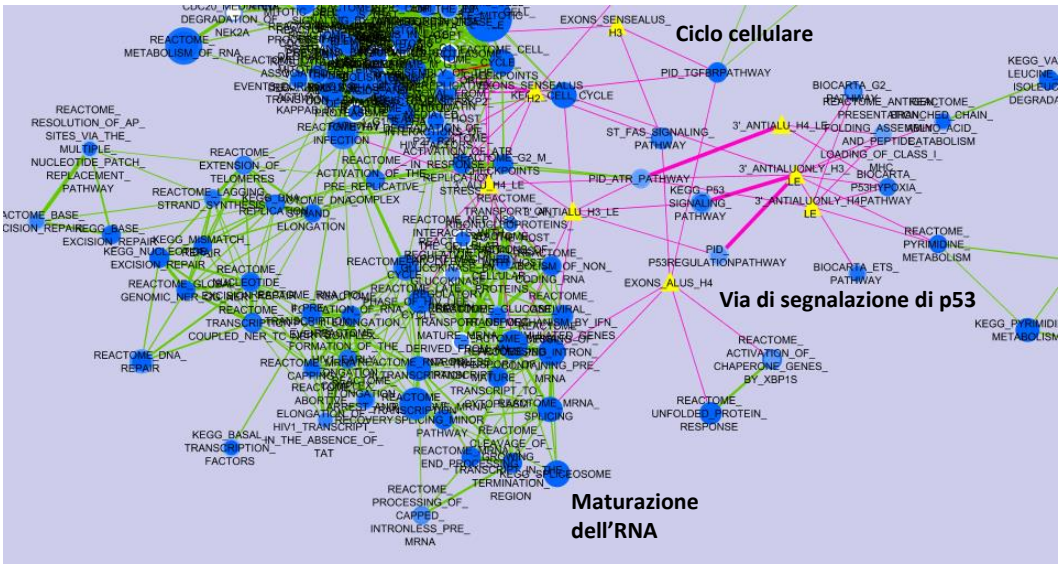


Figura 33. Rappresentazione dei pathway down-regolati nei due esperimenti di sovraepressione mediante l'interfaccia grafica di Cytoscape. Sono esplicitati i pathway che condividono geni con gli insiemi di geni da noi realizzati, ovvero i gene sets corrispondenti a mRNA e i lncRNA (o meglio agli mRNA "intercettati" dai lncRNA) con almeno un elemento *Alu* (senso, antisento, o in entrambi gli orientamenti). Le connessioni in magenta tra i nostri gene sets (in giallo) e i geni del ciclo cellulare, della maturazione dell'RNA e della via di segnalazione di p53 mettono in luce che potrebbero essere proprio tali processi a risentire per primi dalle nostre sovraespressioni.

Una volta identificati, mediante le analisi GSEA riportate sopra, i processi cellulari maggiormente influenzati dalle sovraespressioni dei due *Alu* RNA selezionati, si sono analizzati i geni effettivamente coinvolti in tali processi (**Figura 34**).

A

REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA
REACTOME_MRNA_PROCESSING
REACTOME_METABOLISM_OF_NON_CODING_RNA
REACTOME_CELL_CYCLE_MITOTIC
REACTOME_CELL_CYCLE
REACTOME_DNA_REPLICATION
KEGG_CELL_CYCLE
REACTOME_P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE
PID_P53REGULATIONPATHWAY
KEGG_P53_SIGNALING_PATHWAY
BIOCARTA_P53HYPOXIA_PATHWAY
REACTOME_METABOLISM_OF_NON_CODING_RNA

B

ALDH6A1	aldehyde dehydrogenase 6 family, member A1
CLSPN	claspin
DBT	dihydrolipoamide branched chain transacylase E2
FUS	fused in sarcoma
GINS1	GINS complex subunit 1 (Psf1 homolog)
GTF2H3	general transcription factor IIH, polypeptide 3, 34kDa
HDAC2	histone deacetylase 2
MDM2	Mdm2, p53 E3 ubiquitin protein ligase homolog (mouse)
NMT2	N-myristoyltransferase 2
NUP43	nucleoporin 43kDa
PPM1D	protein phosphatase, Mg2+/Mn2+ dependent, 1D
PSMB2	proteasome (prosome, macropain) subunit, beta type, 2
SAR1B	SAR1 homolog B (S. cerevisiae)
SF3B3	splicing factor 3b, subunit 3, 130kDa
TFDP2	transcription factor Dp-2 (E2F dimerization partner 2)
UMPS	uridine monophosphate synthetase

Figura 34. (A) Elenco dei processi biologici maggiormente colpiti dalla sovraespressione di entrambi gli elementi *Alu*. REACTOME, che compare prevalentemente, è un database di pathway, così come KEGG, BIOCARTA e PID. **(B)** Elenco dei geni down-regolati (ovvero meno espressi nel caso delle sovraespressioni di *Alu* RNA rispetto al controllo trasfettato con plamide vuoto).

Per quanto concerne il ciclo cellulare, i geni down-regolati, contenenti almeno 3 elementi *Alu* nella 3'UTR sono:

- il gene GINS1 (GINS complex subunit 1), il cui prodotto proteico risulta essenziale per l'inizio della replicazione del DNA;

- il gene HDAC2 (Histone Deacetylase 2), che codifica per l'enzima istone deacetilasi 2, importante nella regolazione della trascrizione, nella progressione del ciclo cellulare, e nello sviluppo. Vi sono recenti evidenze che la down-regolazione di tale gene porti ad una inibizione della proliferazione cellulare e all'arresto del ciclo cellulare. Hua e collaboratori, infatti, dimostrano che trasfettando le cellule HeLa con un siRNA progettato in modo da appaiarsi al trascritto HDAC2, si ha un decremento nell'espressione del trascritto e conseguente arresto delle cellule tra le fasi G0 e G1 del ciclo cellulare[395];

- il gene TFDP2(Transcription Factor Dp-2), che codifica per un fattore trascrizionale che normalmente attiva i geni del ciclo cellulare.

Dal momento che i geni menzionati potrebbero essere i bersagli diretti degli *Alu* RNA (per la presenza nella loro 3'UTR di almeno 3 elmenti *Alu* antisenso), si potrebbe ipotizzare che tali geni siano silenziati mediante destabilizzazione dei loro messaggeri attraverso un meccanismo di *RNA interference*, ampiamente riportato per piccoli RNA non codificanti (miRNA *in primis*) e anche per gli *Alu* RNA della sottofamiglia DR2 (v. trattazione precedente[199]), con conseguente inibizione della progressione del ciclo cellulare.

I trascritti down-regolati, con almeno 3 elementi *Alu* nella 3'UTR, coinvolti nella maturazione dell'RNA (codificante e non codificante), contenenti almeno 3 elementi *Alu* nella 3'UTR, corrispondono ai geni:

- FUS (FUS RNA binding protein), che codifica per un componente del complesso hnRNP coinvolto nello splicing dei pre-mRNA e nell'esportazione degli mRNA maturi nel citoplasma.

- SF3B3 (Splicing Factor 3B subunit 3), che codifica per la subunità 3 del fattore di splicing 3b, il quale insieme al fattore 3a e all'RNA 12S forma il complesso U2 snRNP, coinvolto nel processo di maturazione dell'mRNA.

L'alterazione della maturazione dei trascritti conferma ed estende il sospetto di un coinvolgimento degli *Alu* RNA nei processi post-trascrizionali come riportato per altri ncRNA[164,190] (come ampiamente illustrato nell'Introduzione).

Infine, il decremento dell'espressione di due geni che contengono almeno 3 elementi *Alu* nella propria 3'UTR, ovvero di MDM2 (gene per la ubiquitina ligasi E3 che in cellule normali porta alla degradazione di p53[396]) e di un altro interattore di p53 che inibisce la sua attività (PPM1D[397]), è in

accordo con il fatto che un aumento dei livelli di *Alu* RNA sia indicativo di una condizione di stress cellulare e inneschi un meccanismo di difesa. Se la p53 non è più degradata o inibita, ciò potrebbe portare all'arresto del ciclo cellulare, o indurre l'apoptosi cellulare.

Le evidenze riportate mettono in luce che la sovraespressione dell'*Alu*Sq2 RNA e dell'*Alu*Y RNA in cellule HeLa ha come effetto preponderante la down-regolazione dei trascritti cellulari (sia codificanti che non codificanti per proteine) e che, per alcuni di tali trascritti, si potrebbe ipotizzare un meccanismo di silenziamento. La down-regolazione osservata, infatti, ha suggerito fortemente l'esistenza di un meccanismo (senso-antisenso per gli mRNA, e che prevede il coinvolgimento di un terzo RNA per i lncRNA) reso possibile dalla presenza di un gruppo di elementi *Alu* nella sequenza degli RNA down-regolati (nella 3'UTR degli mRNA e nella forma matura dei lncRNA). La possibilità di una down-regolazione dei trascritti codificanti per proteine mediante appaiamento senso-antisenso tra *Alu* RNA è stata documentata anche da un recente lavoro condotto nelle cellule HeLa sottoposte a shock termico[398]. Gli autori hanno rivelato la presenza di un putativo sito di legame per il fattore di shock termico (HSF, Heat Shock Factor) in elementi *Alu* (sia in orientamento senso che antisenso) posti in regioni circa 5 kb a monte e a valle del sito di inizio della trascrizione dei geni down-regolati a seguito di shock termico. Inoltre, una analisi più approfondita ha permesso di osservare che la maggior parte dei siti di legame per il HSF era presente negli elementi *Alu* posti entro le 2kb a monte del TSS dei geni e in antisenso rispetto alla direzione della loro trascrizione. In altri termini, i trascritti che subivano un maggior decremento di espressione a seguito di shock termico, erano quelli con un sito di legame per il HSF posto in un elemento *Alu* in antisenso nella regione a monte dei geni codificanti per tali RNA. Lo studio, ha permesso di concludere che tale decremento, conseguente a stress termico, potrebbe essere dovuto alla trascrizione degli elementi *Alu* in antisenso o da parte della Pol III oppure da parte della Pol II richiamata dal legame del HSF a tali elementi. Inoltre, gli autori ipotizzano che i trascritti *Alu* in antisenso, potrebbero in seguito contribuire alla diminuzione dei rispettivi trascritti senso mediante interferenza trascrizionale o *RNA interference* resa possibile mediante un appaiamento senso-antisenso tra l'*Alu* RNA e l'mRNA bersaglio.

Oltre all'ipotesi meccanicistica alla base del decremento dei trascritti, l'analisi dei processi cellulari ha rivelato che i pathway biologici che potrebbero essere influenzati per primi dalla sovrespressione di *Alu* RNA nelle cellule HeLa sono il ciclo cellulare, lo splicing e la via di segnalazione di p53, interpretata come attivazione della p53 e conseguente possibile induzione dell'apoptosi cellulare. Tali evidenze sono confrontabili con quanto riportato da Pandey e collaboratori[398] a seguito di shock termico nel medesimo tipo cellulare: anche in tale contesto

sperimentale, si è osservato che il ciclo cellulare, la regolazione della trascrizione, la sintesi del DNA e dell'RNA e la regolazione dell'apoptosi sono processi down-regolati. Pertanto, dato che a seguito di shock termico si hanno numerosi esempi di una sovraespressione di *Alu* RNA, e dato che le nostre analisi sono state condotte a seguito di una sovraespressione di *Alu* RNA, i nostri dati confermano che i primi processi colpiti potrebbero essere quelli menzionati sopra, e sottolineano il plausibile coinvolgimento di meccanismi senso-antisenso nella down-regolazione dei trascritti coinvolti in tali pathway. Inoltre, lo studio condotto da Baryakin e collaboratori[389] ha messo in luce che le cellule MCF7 trasfettate con un *Alu* RNA subiscono una riduzione della vitalità, un aumento dei trascritti coinvolti nella risposta allo stress del reticolo endoplasmatico e l'induzione di cambiamenti proaoptotici. Tali evidenze, addizionate alle nostre e a quelle di Pandey e collaboratori[398], potrebbero permettere di concludere che la risposta all'accumulo di *Alu* RNA è in parte tipo cellulare-specifica (nel nostro contesto sperimentale non si osserva una riduzione della vitalità cellulare né un aumento degli RNA coinvolti nella risposta allo stress del reticolo endoplasmatico) e in parte comune alle differenti linee cellulari, in quanto si osserva sempre un possibile arresto del ciclo cellulare e induzione della morte programmata.

Conclusioni

Le osservazioni riportate nella prima parte del presente lavoro di tesi, permettono di concludere che l'identificazione dei profili di *Alu* RNA autonomi, tipo cellulare-specifici, consentirà di analizzare a fondo il loro controllo trascrizionale, e ciò risulta rilevante in quanto stanno emergendo evidenze, seppur sporadiche, del fatto che la loro espressione sia connessa con diverse patologie[340,376].

A questo proposito, lo studio degli effetti della sovraespressione di elementi *Alu* sul trascrittoma cellulare, condotto nella seconda parte del presente elaborato, risulta innovativo, dal momento che finora sono poche, se non assenti, le evidenze di ciò che può succedere incrementando notevolmente l'espressione dei retrotrasposoni nelle cellule umane. Vi sono infatti, evidenze di un aumento di espressione degli *Alu* RNA in alcune condizioni patologiche, come riportato sopra, e in seguito a particolari condizioni di stress[200], ma non si sono ancora indagati con precisione gli effetti di tale aumento e, soprattutto, i meccanismi alla base di tali effetti, sul trascrittoma cellulare.

Molto interessante è la scoperta che a seguito della sovraespressione di *Alu* RNA si ha una deregolazione dell'espressione di molti geni sia codificanti che non codificanti per proteine. In particolare la diminuzione dei livelli di trascritti codificanti per proteine con più elementi *Alu* antisenso (rispetto alla loro direzione di trascrizione) nella loro 3'UTR, evidenzia la probabile esistenza di un meccanismo senso-antisenso alla base di tale diminuzione, e, in ultima analisi, un possibile nuovo livello di regolazione genica post-trascrizionale dipendente dall'appaiamento di *Alu* RNA alla 3'UTR di mRNA bersaglio.

Al contrario, l'evidenza che il livello dei lunghi trascritti non codificanti (lncRNA) diminuisca solo se presenti al loro interno sequenze *Alu* senso (nello stesso orientamento della loro trascrizione), lascia intravedere un possibile meccanismo di regolazione post-trascrizionale di tali elementi (opposto a quello degli mRNA), che potrebbe coinvolgere il sequestro del complesso SRP9/14 da parte degli *Alu* RNA e consentire la destabilizzazione miRNA-dipendente dei lncRNA.

Le alterazioni del trascrittoma *Alu*-dipendenti messe in luce saranno utilizzate per cercare alterazioni comparabili nei set di trascritti delle linee cellulari tumorali esaminate nella prima parte del programma, con la speranza di identificare alterazioni simili che potrebbero derivare dall'espressione di un sottogruppo di *Alu* RNA cancro-specifici.

L'aver identificato, infatti, mediante l'analisi dei dati di RNA-seq di ENCODE relativi a 7 linee cellulari (sia tumorali che non tumorali), profili di *Alu* RNA tipo cellulare-specifici, potrebbe essere utile dal momento che la loro espressione potrebbe essere correlata con un'alterazione

dell'epigenoma cellulare. L'attività trascrizionale dei loci *Alu* nelle cellule tumorali e non tumorali non è mai stata esaminata a fondo, nonostante la notevole influenza degli elementi sull'evoluzione del genoma e alcune sporadiche evidenze del loro coinvolgimento nei tumori. Pertanto l'identificazione di profili di *Alu* RNA, potrebbe fornire, in ultima analisi, un nuovo tipo di biomarcatori del cancro e aprire la strada verso studi futuri circa il coinvolgimento dell'espressione di elementi *Alu* nei tumori. Ciò è in accordo con il sempre maggior interesse rivolto verso i profili di espressione sia dei miRNA[399] che dei lncRNA[400] come biomarcatori del cancro, pertanto, dal momento che gli *Alu* RNA hanno lo stesso potenziale di tali RNA non codificanti di alterare gli stati epigenomici, la loro espressione potrebbe essere correlata con le alterazioni cellulari che accompagnano i tumori. Tali considerazioni potrebbero essere alla base di studi futuri volti a verificare se differenti *Alu* RNA (visti come marcatori cancro-specifici) si esprimono regolarmente non solo in linee cellulari tumorali (come quelle utilizzate nel nostro studio) ma anche in campioni provenienti da tumori umani, con la speranza di poter definire un quadro più completo circa gli *Alu* RNA espressi in differenti condizioni e le alterazioni che essi possono provocare sul trascrittoma cellulare.

Inoltre, avendo evidenziato anche un possibile meccanismo di azione degli *Alu* RNA espressi nelle cellule, si è fatta luce circa ciò che può avvenire in un contesto cellulare a seguito della loro derepressione e ciò risulta importante in quanto uno dei primi esempi in tale ambito. Il fatto che gli *Alu* RNA possano esercitare una regolazione negativa sull'espressione del trascrittoma sia codificante che non codificante, aggiunge un tassello verso la conoscenza di questi piccoli trascritti e dei loro meccanismi di azione.

Materiali e metodi

Realizzazione dei costrutti *Alu*Sq2_pSUPER e *Alu*Y_pSUPER

Si sono trasfettati in cellule HeLa l'elemento intergenico *Alu*Sq2_ chr1: 61523296-61523586, il quale è risultato espresso in 5 diverse linee cellulari delle 7 esaminate della prima parte del programma (H1-hESC, HeLa-S3, Hep G2, K562, NHEK) e che manca del motivo interno A_5TACA_6 (sostituito da A_3G), che normalmente hanno tutti gli elementi *Alu*, e *Alu*Y_chr22:41932115-41932411, il quale non è stato trovato espresso in nessuna delle linee cellulari analizzate ma che è risultato interessante in quanto all'interno dell'introne 2 del gene per la subunità RPC8 dell'enzima RNA polimerasi III, in orientamento antisenso. I due elementi *Alu* sono stati amplificati mediante PCR condotta su DNA genomico estratto da saliva umana, subclonati nel vettore pGEM-T Easy (come riportato nei Materiali e metodi della sezione precedente) e successivamente clonati nel vettore di espressione pSUPER.basic (Oligoengine). Per il clonaggio in tale vettore, si è reso necessario progettare dei primers che consentissero l'amplificazione della sola unità trascrizionale dei due elementi *Alu* (dall'ipotetico TSS al primo terminatore della trascrizione nella regione genomica fiancheggiante; in pGEM-T Easy era stata clonata l'unità trascrizionale con circa 200 bp a monte del TSS e circa 100 bp a valle), nonché introducessero siti di taglio per il successivo clonaggio direzionato (le unità trascrizionali, infatti, sono state inserite in orientamento senso a valle del promotore forte H1, presente nel vettore pSUPER.basic). Si sono inseriti i siti di restrizione per gli enzimi BglII e XhoI e, per facilitare la restrizione degli ampliconi ottenuti, aventi alle estremità i siti di taglio prescelti, è stata aggiunta una piccola coda ricca in A e T, in quanto tale tratto di DNA, piuttosto rigido, aumenta l'efficienza della reazione di digestione. La PCR è stata eseguita sui costrutti *Alu*Sq2_pGEM-T Easy e *Alu*Y_ pGEM-T Easy e i primers utilizzati sono riportati in Appendice II. In seguito si sono eseguite digestioni preparative degli ampliconi ottenuti e del vettore pSUPER.basic con gli enzimi di restrizione BglII e XhoI (Thermo Scientific), e si è realizzata la ligazione degli inserti nel vettore. Si è verificata la realizzazione dei costrutti ricombinanti mediante sequenziamento.

Trasfezione transiente dei costrutti in cellule HeLa

I plasmidi ricombinanti *Alu*Sq2_pSUPER e *Alu*Y_pSUPER e il plasmide vuoto pSUPER.basic sono stati purificati con il kit Qiagen Plasmid Midi Kit (Qiagen) prima dell'esecuzione delle reazioni di trasfezione transiente in cellule HeLa. Tali reazioni sono state condotte con il Lipofectamine® LTX Reagent (Life Technologies) e Opti-MEM® I Reduced Serum Medium (Gibco®, Life Technologies)

secondo le indicazioni della casa produttrice. Sono stati trasfettati 12,5 µg di ogni plasmide in piastre contenenti 10 ml di coltura. Le trasfezioni transienti sono state eseguite con cellule HeLa alla confluenza di circa il 50%, a 37°C (5% CO_2, 95% di umidità) per 48 ore. Al termine dell'incubazione, si è controllata la vitalità cellulare utilizzando il colorante Trypan Blue, le cellule sono state contate mediante l'impiego della camera di Burker (per calcolare la percentuale di cellule morte e di cellule vive, riportate in **Tabella 3**, Risultati e discussione), in seguito sono state raccolte, lavate con PBS (Life Technologies), sedimentate e conservate a -80°C in attesa dell'estrazione dell'RNA.

Gli esperimenti sono stati eseguiti indipendentemente tre volte e ogni replicato biologico è stato condotto trasfettando le cellule rispettivamente con il plasmide vuoto pSUPER.basic, *Alu*Sq2_pSUPER e *Alu*Y_pSUPER. Inoltre è stato eseguito anche un controllo non trasfettato, corrispondente alle cellule HeLa non trattate, alla stessa confluenza delle trasfettate prima dell'incubazione per 48 ore. Le cellule HeLa prima della trasfezione transiente, sono state coltivate in terreno DMEM (Gibco®, Life Technologies) in presenza di FBS alla concentrazione finale del 10% (Gibco®, Life Technologies) e Penicilin/Streptomicin dell'1% a 37°C (5% CO_2, 95% di umidità).

Le trasfezioni sono state condotte in collaborazione con la Prof. Annamaria Buschini del Dipartimento di Bioscienze dell'Univeristà degli Studi di Parma.

Estrazione dell'RNA totale dalle cellule trasfettate

Si è eseguita l'estrazione dell'RNA totale dalle cellule conservate a -80°C, mediante l'utilizzo di TRIzol® Reagent (Life Technologies), e si è trattato con TURBO DNA-*free*™ Kit (Ambion), secondo le indicazioni della casa produttrice. Gli RNA totali estratti sono stati quantificati utilizzando il NanoDrop ND-1000 Spectrophotometer V3.7 (Thermo Scientific) e ne è stata verificata l'integrità mediante l'utilizzo di Agilent 2200 TapeStation (Agilent Technologies). Gli esperimenti successivi sono stati condotti con RNA aventi rapporti A260/A280 > 2, A260/A230 > 2 e con RIN > 9,6.

Retrotrascrizione dell'RNA totale

Si è eseguita la retrotrascrizione degli RNA totali estratti, dai quali sono state rimosse eventuali contaminazioni di DNA genomico, utilizzando il kit iScript cDNA Synthesis Kit (Bio-Rad), che contiene una trascrittasi inversa Rnasi H+ e una miscela di primers oligo dT e *random examers*, secondo quanto indicato dalla casa produttrice. Sono stati retrotrascritti 500 ng di templato ad RNA e si sono eseguiti, per ogni campione, controlli no RT, ovvero retrotrascrizioni dei medesimi

templati in assenza dell'enzima retrotrascrittasi, necessari per poter evidenziare, nelle reazioni di PCR successive, eventuali contaminazioni da DNA genomico che potrebbero falsare l'attendibilità dei dati.

RT-PCR semiquantitativa

I cDNA ottenuti dalla retrotrascrizione dei campioni in esame sono stati amplificati mediante l'utilizzo di coppie di primers progettati:

- sulla sequenza 3'trailer, caratteristica per ognuno dei due elementi *Alu* trasfettati (*Alu*Sq2 e *Alu* Y);
- sul corpo centrale di ognuno dei due elementi *Alu* tasfettati;
- sul gene housekeeping per la β2 microbulina (B2M), espresso costitutivamente a livelli costanti nel contesto sperimentale utilizzato.

Le coppie di primers sono riportate in Appendice II. I prodotti di PCR sono stati visualizzati su gel di poliacrilammide non denaturante al 6% a seguito di uno staining in Etidio Bromuro.

Arraystar Human LncRNA Microarray V3.0

L'Arraystar Human LncRNA Microarray V3.0 utilizzato nelle analisi di espressione è disegnato per determinare profili globali di lncRNA e trascritti codificanti per proteine (mRNA). Possono essere rilevati 30586 lncRNA e 26109 mRNA. I lncRNA inseriti nell'array derivano da database di sequenze nucleotidiche (per esempio Refseq, UCSC Genome Browser, Gencode) o da pubblicazioni precedenti[189,401]. Ogni trascritto è rappresentato da una probe, corrispondente a un esone o a una giunzione di splicing, che permette di identificare accuratamente un solo trascritto e distinguerlo, pertanto dagli altri. Per il controllo qualità dell'ibridazione sono disegnate sull'array anche probe positive per i geni housekeeping e probe negative.

Marcatura dei campioni e ibridazione allo Human LncRNA Expression Microarray

Una volta verificata l'avvenuta sovraespressione di entrambi i costrutti, gli RNA totali sono stati ibridati allo Human LncRNA Expression Microarray (Arraystar, Inc.), il quale permette di tracciare un profilo di espressione di 26109 trascritti codificanti per proteine e di 30586 lncRNA, come riportato sopra. Tutta la procedura è stata svolta presso i laboratori di Arraystar Inc. Per la preparazione dei campioni e l'ibridazione all'array è stato seguito il protocollo dell'Agilent One-Color Microarray-Based Gene Expression Analysis (Agilent Technology) apportando solo alcune

piccole modifiche. In breve, sono stati rimossi gli rRNA dall'RNA totale mediante l'utilizzo del kit mRNA-ONLY™ Eukaryotic mRNA Isolation Kit (Epicentre) e, in seguito, ogni campione è stato amplificato e trascritto in un cRNA fluorescente per tutta la sua lunghezza, utilizzando un metodo *random priming* (Arraystar Flash RNA Labeling Kit, Arraystar). I cRNA marcati sono stati purificati con il kit RNeasy Mini Kit (Qiagen) e sia la loro concentrazione che la loro attività specifica (calcolata secondo la formula: pmol dye (Cy3)/µg RNA) sono state misurate utilizzando il NanoDrop ND-1000 (Thermo Scientific). 1 µg di ogni cRNA marcato è stato successivamente frammentato per aggiunta di 10X Blocking Agent e di 25X Fragmentation Buffer, si sono incubate le reazioni a 60°C per 30 minuti e si è addizionato il 2X GE Hybridization buffer per diluire i cRNA marcati. Un volume corrispondente a 50 µl di soluzione di ibridazione è stato aggiunto alla *gasket slide* e assemblato al vetrino del LncRNA expression microarray. I vetrini sono stati incubati 17 ore a 65°C nell' Agilent Hybridization Oven. Infine, gli array ibridati sono stati lavati, fissati e scansiti utilizzando l'Agilent DNA Microarray Scanner.

Acquisizione dei dati e normalizzazione quantilica

Per analizzare le immagini acquisite degli array (9 array in totale) è stato utilizzato l'Agilent Feature Extraction software (versione 11.0.1.1) ed è stata fatta una normalizzazione quantilica dei dati ottenuti in R utilizzando la suite di Bioconductor[402]. Sono stati considerati per le analisi successive solo trascritti rilevati in almeno 2 delle 3 repliche biologiche dell'esperimento. Pertanto, dei 56695 trascritti nell'Arraystar Human LncRNA Microarray V3.0, le analisi successive si sono concentrate solo su 39978.

Analisi sugli mRNA

I trascritti differenzialmente espressi sono stati identificati con il test LIMMA, per distinguere il sottogruppo di geni up-regolati e il sottogruppo di geni down-regolati nelle trasfezioni con l'elemento *Alu*Sq2 o *Alu*Y, rispetto al controllo (trasfezione con plasmide vuoto). Le analisi di GSEA (Gene Set Enrichment Analysis) sono state condotte mediante l'impiego della versione Java del software[403,404], utilizzando i parametri consigliati. In particolare "gene set" è stato scelto per il parametro "permutation type" e "149" è stato scelto per il parametro "seed for permutation". Gli parametri sono stati lasciati invariati. I valori di P-value a 0.01 e FDR a 5% sono stati scelti come soglie di significatività. Gene sets già disponibili al MSigDB (CP, per i pathway), o costruiti come descritto nel capitolo Risultati e Discussione, sono stati utilizzati per le analisi. Per facilitarne la

lettura, i risultati di GSEA sono stati visualizzati mediante l'utilizzo dell'interfaccia grafica di Cytoscape, utilizzando il programma Enrichment map.

Analisi sugli mRNA con elementi *Alu* nelle regioni non tradotte

In seguito si è scelto di considerare solo gli mRNA che presentavano almeno un elemento *Alu* senso o antisenso nelle loro regioni non tradotte. Per ottenere i trascritti che rispettassero i requisiti richiesti, si sono incrociate le coordinate delle 5'UTR e 3'UTR degli mRNA (ottenute mediante l'utilizzo di UCSC Table Browser) con il dataset di elementi *Alu* che include, oltre agli elementi completi (full-length), anche gli elementi *Alu* tronchi (in 5' e/o in 3'), ma non gli elementi FLAM, FRAM o FAM (monomeri da cui gli elementi *Alu* si sono originati[200]). Si è quindi ottenuta la lista di trascritti codificanti per proteine con almeno un elemento *Alu* nella 5'UTR, formata da 43 mRNA e, analogamente, si sono ottenuti i 2867 mRNA con almeno un elemento *Alu* nella 3'UTR. Inoltre, si sono determinati il numero e gli orientamenti degli elementi *Alu* presenti nelle regioni non tradotte e, dal momento che il numero di trascritti con elementi nella 5'UTR era molto basso, si è continuata l'analisi solo sui 2867 mRNA con almeno un elemento *Alu* nella 3'UTR.

Si sono realizzati, poi, dei gene sets di tali trascritti sulla base, non solo dell'orientamento, ma anche del numero di elementi *Alu* presenti, e si è scelto di considerare mRNA con un gruppo di almeno 3 o almeno 4 elementi *Alu* nella loro 3'UTR per valutare gli effetti che la presenza di più elementi poteva avere sulla modulazione della loro espressione a seguito delle trasfezioni. Innazitutto si è risaliti ai geni corrispondenti a tali trascritti e si è eseguita un'analisi GSEA (http://www.broadinstitute.org/gsea) sui gene sets in **Tabella 4** (Risultati e discussione), confrontando ogni esperimento di sovraespressione con il controllo rappresentato dalla trasfezione con il pasmide vuoto. Sono stati considerati significativamente deregolati solo i geni dei gene sets con P-value < 0.01 e FDR < 5% , scelti come soglia (**Tabella 5**, Risultati e discussione).

Inoltre, per valutare se fossero maggiormente deregolati gli mRNA con elementi *Alu* nella 3' UTR appartenenti alla stessa famiglia degli elementi trasfettati, si è risaliti alle famiglie di appartenza di vari elementi *Alu* nelle regioni 3' non tradotte (nel database di elementi *Alu*, utilizzato inizialmente, per ogni elemento era anche riportata la famiglia di appartenenza) si sono creati dei gene sets formati dai geni corrispondenti ai trascritti con elementi AluJ (senso, antisenso o entrambi), AluS (senso, antisenso o entrambi) o AluY (senso, antisenso o entrambi) (oltre ai gene sets sulla base dell'orientamento degli elementi *Alu*, sono stati realizzati anche gene sets formati da geni con gruppi di almeno 2 elementi Alu per famiglia), e si è eseguita un'analisi GSEA

(http://www.broadinstitute.org/gsea) su tali gene sets e sui gene sets creati in precedenza (i gene sets sui quali è stata condotta la GSEA sono riassunti in **Tabella 6A**, Risultati e discussione). I valori di P-value a 0.01 e FDR a 5% sono stati scelti come soglie di significatività (**Tabella 6B**, Risultati e discussione).

Analisi sui lncRNA

Innanzitutto si è osservato che dei 39978 trascritti rilevati, 19441 corrispondono a trascritti non codificanti per proteine. In seguito si è scelto di considerare solo i lncRNA che presentavano almeno un elemento *Alu* senso o antisenso nella sequenza della loro forma matura (post-splicing). Per ottenere i trascritti che rispettassero i requisiti richiesti, in primo luogo si sono determinate le posizioni degli esoni dei lncRNA rilevati (mediante l'utilizzo di UCSC Table Browser). Così facendo, sono stati scartati dall'analisi 2028 lncRNAs per i quali non si avevano informazioni certe sulla posizione degli esoni. In seguito si sono incrociate le coordinate dei 17413 lncRNA ottenuti con il dataset di elementi *Alu* che include, oltre agli elementi completi (full-length), anche gli elementi *Alu* tronchi (in 5' e/o in 3'), ma non gli elementi FLAM, FRAM o FAM (monomeri da cui gli elementi *Alu* si sono originati[200]). Si è quindi ottenuta la lista di trascritti non codificanti con almeno un elemento *Alu* nella regione interna. Inoltre, si sono determinati il numero e gli orientamenti degli elementi *Alu* presenti in tali regioni interne ai lncRNA (**Figura 30**, Risultati e discussione), e si sono creati dei gene sets di tali trascritti sulla base, non solo dell'orientamento, ma anche del numero di elementi *Alu* presenti, scegliendo di considerare lncRNA con un gruppo di almeno 2 elementi *Alu* nella loro sequenza. Anche in questo caso si volevano valutare gli effetti che la presenza di più elementi poteva avere sulla modulazione della loro espressione a seguito delle trasfezioni. Innazitutto si è risaliti ai geni corrispondenti a tali trascritti e si è eseguita un'analisi GSEA (http://www.broadinstitute.org/gsea) sui gene sets descritti in **Tabella 7** (Risultati e discussione), confrontando ogni esperimento di sovraespressione con il controllo rappresentato dalla trasfezione con il pasmide vuoto. I valori di P-value a 0.01 e FDR a 5% sono stati scelti come soglie di significatività (**Tabella 8**, Risultati e discussione).

Analisi dei processi cellulari in cui sono coinvolti gli mRNA e i lncRNA down-regolati

L'analisi GSEA è stata utilizzata anche per individuare il significato regolativo della risposta alle sovraespressioni (v. Risultati e discussione).

Nelle nostre analisi sono stati utilizzati i gene sets di geni codificanti e non codificanti per proteine contenenti elementi *Alu* (in giallo in **Figura 32** e in **Figura 33**, Risultati e discussione), descritti in precedenza, significativi all'analisi GSEA (**Tabella 5** per gli mRNA e **Tabella 8** per i lncRNA, Risultati e discussione). Nel caso dei lncRNA i gene sets per la visualizzazione delle connessioni con i pathway sono stati creati dai gene sets originali, sostituendo i nomi dei trascritti lncRNA con i nomi degli mRNA ad essi "sovrapposti" (v. Risultati e discussione), ovvero i lncRNA all'interno dei trascritti codificanti per proteine. Le connessioni con i nodi dei pathway rappresentano la significatività al Fisher's exact test, con soglia al 5%. Sono riportati soltanto i pathway con FDR < 1% in GSEA.

Le analisi GSEA condotte in questa parte della presente tesi di dottorato sono state eseguite in collaborazione con la Dr.ssa Barbara Montanini del Dipartimento di Bioscienze dell'Univeristà degli Studi di Parma.

Appendice II

• Sono riportati i primers utilizzati nello studio:

Primers utilizzati per le RT-PCR	Sequenze dei primers (5' → 3')
***Alu*Sq2_3'trailer _fw**	AAGTGTCACCTCCCCATCTG
***Alu*Sq2_3'trailer _rev**	ACCACCGTTTCCTGAGCTT
***Alu*Sq2_body_fw**	CACATGGTGAAACCCCGTCTCT
***Alu*Sq2_body_rev**	GGTACAAGCGATTCTCCTGCCT
***Alu*Y_3'trailer_fw**	AGTTCTGGAGGTGGATGGC
***Alu*Y _3'trailer _rev**	AAGTCAGTGGCATTAGGCACA
***Alu*Y _body_fw**	CATGATGAAACCCCGTCTCAAC
***Alu*Y _body_rev**	TTCTCCTGCCTCAGCCTCCCGA
B2M _fw	GGCTATCCAGCGTACTCCAAAG
B2M _rev	TGGATGAAACCCAGACACATAGC

• Sono riportati risultati della GSEA sui pathway (P-value <0.01 e FDR <1%), visualizzati nel capitolo Risultati e discussione mediante l'interfaccia grafica di Cytoscape:

Esperimento di trasfezione con *Alu*Y (confrontato con il controllo trasfettato con plamide vuoto):

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY_CCT_TF	19	-0,86655	-2,32914	0	0
REACTOME_PREFOLDIN_MEDIATED_TRANSFER_OF_SUBSTRATE_TO_CCT_TRIC	25	-0,76768	-2,14295	0	0
REACTOME_OLFACTORY_SIGNALING_PATHWAY	126	0,561855	2,830932	0	0
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	207	0,49518	2,715249	0	0
KEGG_OLFACTORY_TRANSDUCTION	163	0,511614	2,574627	0	0
KEGG_SPLICEOSOME	120	-0,59318	-2,09764	0	2,59E-04
REACTOME_MRNA_SPLICING	103	-0,59446	-2,06307	0	4,12E-04
REACTOME_PROTEIN_FOLDING	47	-0,64544	-2,03346	0	0,001202
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	131	-0,57147	-2,02897	0	0,001328
PID_HIF1_TFPATHWAY	53	-0,62603	-2,01757	0	0,001416
REACTOME_METABOLISM_OF_NUCLEOTIDES	65	-0,59511	-1,99899	0	0,001601
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	232	0,39362	2,222819	0	0,001695
REACTOME_MITOTIC_PROMETAPHASE	81	-0,58244	-1,9854	0	0,001967
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	47	-0,63493	-1,98592	0	0,002164
REACTOME_MRNA_PROCESSING	150	-0,54753	-1,95708	0	0,002396
REACTOME_ENDOSOMAL_SORTING_COMPLEX_REQUIRED_FOR_TRANSPORT_	24	-0,70916	-1,97214	0	0,0024
KEGG_PYRIMIDINE_METABOLISM	88	-0,57199	-1,96227	0	0,002422
REACTOME_CELL_CYCLE_MITOTIC	292	-0,52124	-1,96283	0	0,002531
PID_ILK_PATHWAY	41	-0,61708	-1,94962	0	0,002756
REACTOME_DNA_REPLICATION	179	-0,53542	-1,93374	0	0,003498
REACTOME_KINESINS	22	-0,70593	-1,92932	0	0,003762
REACTOME_MRNA_SPLICING_MINOR_PATHWAY	40	-0,62486	-1,9208	0	0,004209
REACTOME_MITOTIC_M_M_G1_PHASES	159	-0,53342	-1,91775	0	0,00446
REACTOME_CLEAVAGE_OF_GROWING_TRANSCRIPT_IN_THE_TERMINATION_F	42	-0,61314	-1,91074	0	0,004663
PID_MTOR_4PATHWAY	64	-0,57504	-1,91281	0	0,00469
REACTOME_GLUCOSE_METABOLISM	56	-0,58094	-1,89645	0	0,005975
REACTOME_PURINE_METABOLISM	29	-0,64937	-1,89118	0	0,00615
REACTOME_POST_CHAPERONIN_TUBULIN_FOLDING_PATHWAY	15	-0,76222	-1,88229	0	0,006643
REACTOME_GLUCONEOGENESIS	28	-0,64834	-1,86927	0	0,008165

Esperimento di trasfezione con *Alu*Sq2 (confrontato con il controllo trasfettato con plamide vuoto):

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
REACTOME_FORMATION_OF_TUBULIN_FOLDING_INTERMEDIATES_BY	19	-0,83993	-2,21836	0	0
REACTOME_MRNA_SPLICING	103	-0,65315	-2,18216	0	0
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_M	131	-0,64072	-2,17562	0	0
REACTOME_MRNA_PROCESSING	150	-0,62597	-2,14475	0	0
REACTOME_PREFOLDIN_MEDIATED_TRANSFER_OF_SUBSTRATE_TO_C	25	-0,77222	-2,10819	0	0
REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION	22	-0,78905	-2,10247	0	0
REACTOME_CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COM	52	-0,67127	-2,09016	0	0
REACTOME_OLFACTORY_SIGNALING_PATHWAY	126	0,547272	2,894607	0	0
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	207	0,44995	2,773845	0	0
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	232	0,424521	2,516112	0	0
REACTOME_DESTABILIZATION_OF_MRNA_BY_AUF1_HNRNP_D0	49	-0,65404	-2,03821	0	1,46E-04
REACTOME_CELL_CYCLE_CHECKPOINTS	106	-0,60514	-2,04585	0	1,57E-04
REACTOME_SYNTHESIS_OF_DNA	86	-0,61931	-2,04677	0	1,70E-04
KEGG_SPLICEOSOME	120	-0,60582	-2,05326	0	1,86E-04
REACTOME_GLUCONEOGENESIS	28	-0,7242	-2,05927	0	2,04E-04
REACTOME_MITOTIC_M_M_G1_PHASES	159	-0,59926	-2,06034	0	2,27E-04
REACTOME_M_G1_TRANSITION	74	-0,6281	-2,02351	0	2,73E-04
REACTOME_P53_DEPENDENT_G1_DNA_DAMAGE_RESPONSE	52	-0,63697	-2,00609	0	2,79E-04
REACTOME_P53_INDEPENDENT_G1_S_DNA_DAMAGE_CHECKPOINT	47	-0,65083	-2,00824	0	2,93E-04
REACTOME_SIGNALING_BY_WNT	61	-0,61963	-1,99534	0	2,98E-04
REACTOME_CDK_MEDIATED_PHOSPHORYLATION_AND_REMOVAL_OI	45	-0,66032	-2,01055	0	3,07E-04
REACTOME_ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX	59	-0,63228	-1,9998	0	3,11E-04
KEGG_CYSTEINE_AND_METHIONINE_METABOLISM	27	-0,72032	-2,02002	0	3,19E-04
REACTOME_DNA_REPLICATION	179	-0,58324	-2,01235	0	3,24E-04
REACTOME_CELL_CYCLE_MITOTIC	292	-0,55769	-1,99174	0	3,27E-04
REACTOME_GLUCOSE_METABOLISM	56	-0,63319	-2,01548	0	3,42E-04
REACTOME_GPCR_LIGAND_BINDING	309	0,386913	2,325258	0	3,46E-04
REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE_ODC	47	-0,64879	-1,98764	0	3,54E-04
REACTOME_S_PHASE	102	-0,60146	-2,0177	0	3,62E-04
REACTOME_HIV_INFECTION	181	-0,56857	-1,98258	0	4,17E-04
KEGG_OLFACTORY_TRANSDUCTION	163	0,470695	2,375318	0	4,33E-04
REACTOME_AUTODEGRADATION_OF_THE_E3_UBIQUITIN_LIGASE_CO	46	-0,64778	-1,97994	0	4,38E-04
REACTOME_MRNA_SPLICING_MINOR_PATHWAY	40	-0,6585	-1,97073	0	4,93E-04
REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G	47	-0,64585	-1,97271	0	5,10E-04
KEGG_AMINO_SUGAR_AND_NUCLEOTIDE_SUGAR_METABOLISM	43	-0,65034	-1,97321	0	5,27E-04
REACTOME_HOST_INTERACTIONS_OF_HIV_FACTORS	114	-0,57827	-1,96897	0	5,42E-04
REACTOME_TRNA_AMINOACYLATION	39	-0,65544	-1,96402	0	5,88E-04
REACTOME_PROTEIN_FOLDING	47	-0,63706	-1,95674	0	6,43E-04
REACTOME_ORC1_REMOVAL_FROM_CHROMATIN	61	-0,61757	-1,95918	0	6,62E-04
REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES	170	-0,56198	-1,95137	0	6,72E-04
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	47	-0,6407	-1,95328	0	6,81E-04
REACTOME_POST_CHAPERONIN_TUBULIN_FOLDING_PATHWAY	15	-0,79275	-1,95273	0	6,90E-04
KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	37	-0,65624	-1,94536	0	7,07E-04
REACTOME_AUTODEGRADATION_OF_CDH1_BY_CDH1_APC_C	55	-0,62247	-1,94367	0	7,15E-04
REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTI	45	-0,62998	-1,93725	0	7,97E-04
REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1	48	-0,61757	-1,92247	0	0,001019
KEGG_PROTEASOME	42	-0,63119	-1,9253	0	0,00102
REACTOME_G1_S_TRANSITION	102	-0,56992	-1,92126	0	0,001042
REACTOME_LOSS_OF_NLP_FROM_MITOTIC_CENTROSOMES	52	-0,61101	-1,91729	0	0,001065

REACTOME_AMINE_LIGAND_BINDING_RECEPTORS	25	0,607139	2,238333	0	0,001123
REACTOME_ER_PHAGOSOME_PATHWAY	57	-0,60261	-1,91292	0	0,001198
REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	51	-0,61264	-1,91194	0	0,001216
REACTOME_APC_C_CDH1_MEDIATED_DEGRADATION_OF_CDC20_AN	63	-0,59545	-1,90156	0	0,001593
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_S	32	-0,66273	-1,89829	0	0,001644
REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_	64	-0,59321	-1,88839	0	0,001998
REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCLE	94	-0,55818	-1,87188	0	0,002638
REACTOME_RECRUITMENT_OF_MITOTIC_CENTROSOME_PROTEINS_A	59	-0,58492	-1,86817	0	0,002763
REACTOME_G2_M_CHECKPOINTS	38	-0,62494	-1,85981	0	0,002986
REACTOME_MITOTIC_G1_G1_S_PHASES	123	-0,54663	-1,86166	0	0,003002
REACTOME_RNA_POL_III_TRANSCRIPTION	30	-0,64374	-1,86072	0,001184834	0,003022
REACTOME_ACTIVATION_OF_NF_KAPPAB_IN_B_CELLS	59	-0,58574	-1,86219	0	0,00304
REACTOME_DNA_REPAIR	95	-0,56187	-1,85398	0	0,003381
REACTOME_DOWNREGULATION_OF_SMAD2_3_SMAD4_TRANSCRIPTI	16	-0,72765	-1,85272	0	0,003445
PID_HIF1APATHWAY	17	-0,71507	-1,85154	0,005235602	0,003491
REACTOME_MITOTIC_PROMETAPHASE	81	-0,56269	-1,848	0	0,003577
REACTOME_INTERACTIONS_OF_VPR_WITH_HOST_CELLULAR_PROTEI	31	-0,63864	-1,84814	0	0,003636
REACTOME_REGULATION_OF_APOPTOSIS	53	-0,58756	-1,84376	0	0,003682
REACTOME_CELL_CYCLE	370	-0,51828	-1,84184	0	0,003694
REACTOME_DNA_STRAND_ELONGATION	27	-0,65672	-1,83689	0	0,003711
BIOCARTA_MTA3_PATHWAY	15	-0,72642	-1,83826	0	0,003722
REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX_	39	-0,61506	-1,83873	0	0,003747
REACTOME_TRANSCRIPTION_COUPLED_NER_TC_NER	42	-0,60592	-1,84189	0,001180638	0,003753
REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION	68	-0,56627	-1,83303	0,001082251	0,003771
REACTOME_SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21	52	-0,5865	-1,8337	0	0,003796
REACTOME_MITOTIC_G2_G2_M_PHASES	74	-0,56171	-1,83376	0	0,003852
REACTOME_METABOLISM_OF_NON_CODING_RNA	46	-0,59424	-1,83016	0	0,003877
REACTOME_RNA_POL_II_TRANSCRIPTION	96	-0,55681	-1,82922	0	0,00388
REACTOME_HIV_LIFE_CYCLE	107	-0,54928	-1,83047	0	0,003932
REACTOME_REGULATION_OF_MRNA_STABILITY_BY_PROTEINS_THAT_	79	-0,56056	-1,82289	0	0,004337
REACTOME_FORMATION_OF_THE_HIV1_EARLY_ELONGATION_COMPL	30	-0,62685	-1,81905	0,001190476	0,004579
REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PE	19	-0,69166	-1,81366	0,004920049	0,004935
REACTOME_CLEAVAGE_OF_GROWING_TRANSCRIPT_IN_THE_TERMIN	42	-0,59447	-1,80374	0	0,005626
REACTOME_CYCLIN_E_ASSOCIATED_EVENTS_DURING_G1_S_TRANSIT	61	-0,56741	-1,80267	0	0,005685
KEGG_CELL_CYCLE	113	-0,53421	-1,79739	0	0,005698
REACTOME_METABOLISM_OF_RNA	246	-0,51333	-1,79931	0	0,0057
KEGG_PYRIMIDINE_METABOLISM	88	-0,5422	-1,79793	0	0,005718
REACTOME_MITOCHONDRIAL_PROTEIN_IMPORT	47	-0,58576	-1,79995	0	0,00572
REACTOME_GLOBAL_GENOMIC_NER_GG_NER	32	-0,61317	-1,80069	0	0,005754
REACTOME_ANTIVIRAL_MECHANISM_BY_IFN_STIMULATED_GENES	59	-0,56863	-1,79246	0	0,006079
REACTOME_MRNA_3_END_PROCESSING	33	-0,62562	-1,7915	0	0,006103
REACTOME_NUCLEOTIDE_EXCISION_REPAIR	47	-0,58055	-1,78883	0	0,006232
REACTOME_UNFOLDED_PROTEIN_RESPONSE	71	-0,54672	-1,7895	0	0,006245
BIOCARTA_PROTEASOME_PATHWAY	28	-0,63093	-1,78355	0	0,006544
REACTOME_EXTENSION_OF_TELOMERES	26	-0,64563	-1,7841	0,002469136	0,006561
KEGG_BASAL_TRANSCRIPTION_FACTORS	28	-0,63518	-1,78289	0	0,006584
KEGG_BASE_EXCISION_REPAIR	31	-0,61959	-1,78446	0	0,006624
REACTOME_ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX	27	-0,63959	-1,77886	0,00120048	0,006933
REACTOME_GLYCOLYSIS	23	-0,64984	-1,77252	0,001231527	0,007538
REACTOME_ELONGATION_ARREST_AND_RECOVERY	28	-0,62846	-1,76988	0	0,007784
REACTOME_ASSOCIATION_OF_TRIC_CCT_WITH_TARGET_PROTEINS_D	25	-0,63681	-1,7675	0	0,007937
KEGG_NUCLEOTIDE_EXCISION_REPAIR	44	-0,57581	-1,76472	0	0,008069
REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX	53	-0,55945	-1,76494	0	0,008121
BIOCARTA_P53HYPOXIA_PATHWAY	19	-0,67084	-1,76528	0,001297017	0,008163
PID_ILK_PATHWAY	41	-0,56934	-1,75315	0	0,009368
PID_TGFBRPATHWAY	46	-0,57211	-1,75377	0	0,00937
REACTOME_TRANSPORT_OF_RIBONUCLEOPROTEINS_INTO_THE_HOS	25	-0,63528	-1,75083	0,002436054	0,00963
REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_	20	-0,66053	-1,74987	0,002478315	0,009694
KEGG_DNA_REPLICATION	33	-0,60582	-1,74865	0,001175088	0,009829

Parte III

Analisi dell'espressione di elementi *Alu* e LINE-1 in risposta a trattamenti con idrochinone

Numerose evidenze riportate in letteratura mettono in luce che l'esposizione a differenti fattori ambientali, noti come causa di differenti patologie, provoca una ipometilazione degli elementi ripetuti (non solo *Alu* ma anche LINE-1)[257,322,328,339,405,406] (si rimanda all'Introduzione). Avendo messo in luce, mediante l'applicazione della nostra strategia sperimentale (v. prima parte della presente tesi di dottorato), che esistono sottogruppi di elementi *Alu* espressi in linee cellulari tumorali e che questi sono specifici per ogni linea, è possibile supporre, dal momento che gli elementi ripetuti sono normalmente silenziati (il che correla in genere con uno stato di ipermetilazione), che la loro derepressione sia connessa con una variazione nel loro stato di metilazione. Tale ipotesi può valere anche per gli elementi LINE-1 (L1) in quanto anch'essi elementi ripetuti silenziati in cellule normali, e in quanto vi sono casi riportati di ipometilazione di tali elementi associata all'insorgenza e alla progressione tumorale[230,353,407]. Pertanto si potrebbe ipotizzare che, a seguito dell'esposizione a determinati fattori ambientali, gli elementi ripetuti siano derepressi con conseguenti aumenti della loro espressione, in correlazione con una loro ipometilazione. La nostra analisi si è concentrata sul benzene, ed in particolare sul suo metabolita idrochinone, come paradigma dei fattori ambientali che potrebbero alterare l'espressione degli elementi ripetuti *Alu* e L1. È ben noto, infatti, come il benzene sia associato all'aumento del rischio di insorgenza della leucemia mieloide acuta (AML)[335], ma non sono stati spiegati in modo esauriente i meccanismi attraverso cui ciò avviene[333]. Vi sono, però, evidenze che l'esposizione di soggetti umani a bassi livelli di benzene porta ad una ipometilazione di elementi *Alu* e L1[339], suggerendo che tali retrotrasposoni potrebbero essere coinvolti nella sua ematotossicità. Essendo l'idrochinone in grado di creare danno ossidativo al DNA, far arrestare il ciclo cellulare e indurre l'apoptosi[332,408], si pensa che la tossicità del benzene sia mediata proprio da questo derivato metabolico particolarmente reattivo. Si è deciso quindi, innanzitutto, di analizzare gli effetti che l'idrochinone ha sull'espressione degli elementi *Alu* e L1, in quanto tali effetti non sono ad oggi ancora stati esplorati, e di indagare, qualora tale metabolita avesse indotto una variazione nella loro espressione fisiologica, i meccanismi attraverso cui ciò avviene. In ultima analisi, si è cercato di capire in che misura gli effetti sull'espressione dei retroelementi possano rappresentare una delle cause, finora ignorate, che contribuiscono alle patologie associate alla esposizione ambientale.

Risultati e discussione

Trattamento delle cellule THP-1 con il metabolita del benzene idrochinone

Dal momento che si hanno evidenze dell'associazione tra il benzene e l'aumentato rischio di insorgenza della leucemia[335] (in particolare tra esposizione a bassi livelli di benzene e leucemia mieloide acuta[409], come riportato nell'Introduzione), poiché si ritiene che il benzene eserciti la sua tossicità attraverso il metabolita idrochinone, si è inizialmente scelto di condurre gli esperimenti sulla linea cellulare immortalizzata THP-1 (rappresentativa della leucemia mieloide acuta) e si è deciso di eseguire dei trattamenti con idrochinone (HQ) alle concentrazioni 1 µM, 5 µM e 15 µM.
La scelta tali concentrazioni è stata motivata dal fatto che le prime due possono essere considerate come rappresentative del valore limite di esposizione al benzene (1ppm) attualmente accettabile, mentre per l'ultima concentrazione si è intorno ad una esposizione di benzene pari a 5ppm, comunque al di sotto dei livelli ritenuti di alta esposizione[409,410] (v. Appendice III per la scelta delle concentrazioni di HQ impiegate).
Le cellule sono state sottoposte a trattamento con le tre concentrazioni di HQ per 24 e per 48 ore e si è anche eseguito un controllo non trattato ad entrambi i tempi per valutare gli effetti dei trattamenti sull'espressione degli elementi ripetuti. L'esperimento è stato eseguito indipendentemente tre volte. La scelta dei due tempi è motivata dal fatto che si voleva osservare se vi fosse una variazione dell'espressione di *Alu* e L1 dose- e tempo-dipendente. Al termine dei trattamenti si sono eseguite le analisi di vitalità cellulare per essere certi di analizzare variazioni di espressione imputabili a un trattamento che non avesse come effetto preponderante la morte cellulare. Si è verificato, pertanto, che la maggior parte delle cellule fosse vitale individuando così le concentrazioni non letali di HQ per le cellule THP-1. Non vi sono, infatti, esempi riportati delle concentrazioni di tale metabolita che inducono la morte in queste cellule e, nonostante le concentrazioni siano differenti a seconda della differente suscettibilità cellulare, le nostre analisi rivelano che anche le THP-1 potrebbero andare incontro alla morte programmata a concentrazioni superiori a quelle utilizzate, come riportato per altri tipi cellulari[408,411]. Una volta verificato che la maggior parte delle cellule fosse vitale, si è proceduto all'estrazione dell'RNA totale del quale è stata controllata l'integrità e la purezza.

Analisi quantitativa dei livelli di espressione di *Alu* RNA e LINE-1 RNA

In seguito all'estrazione dell'RNA totale dalle cellule, si è eseguita la sua retrotrascrizione mediante l'utilizzo del kit *iScript cDNA Synthesis Kit* (Bio-Rad). Per ogni campione, alle reazioni di trascrizione inversa, è stata affiancata una reazione in assenza di retrotrascrittasi (controllo no RT) in modo da poter evidenziare, nelle successive reazioni di PCR, eventuali contaminazioni di DNA genomico, che avrebbero potuto falsare l'attendibilità del dato. Si è in seguito eseguita una PCR quantitativa (qPCR) con primers specifici per gli elementi *Alu* e L1. In particolare i primers sulle sequenze di elementi *Alu* sono stati progettati in modo da avere sia un appaiamento prevalente su *Alu* della famiglia *Alu*J (chiamati coppia *Alu*J) che *Alu* delle famiglie evolutivamente più recenti *Alu*S e *Alu*Yb8 (chiamati coppia *Alu*Y), mentre i primers sulle sequenze di elementi L1 sono stati progettati sia sulla 5'UTR sia su ORF1 (Le sequenze dei primers sono riportate in Appendice III). L'utilizzo di due coppie di primers su sostanzialmente gli stessi elementi, voleva essere una sorta di ulteriore controllo interno per aumentare la significatività dei risultati. I dati ottenuti sono stati rielaborati secondo il metodo di Livak[412], ovvero si è eseguita la quantificazione relativa dei livelli di espressione dei retroelementi, normalizzati per il gene endogeno costante B2M (gene codificante per la β2-microtubulina), confrontando ogni trattamento con il corrispettivo controllo non trattato (mantenendo separati i due tempi di esposizione). L'analisi dei dati di espressione ha permesso di dimostrare che il trattamento delle cellule THP-1 con HQ genera un aumento di espressione sia degli elementi *Alu* che L1, in particolare a 24 ore, portando a ipotizzare che il trattamento oltre le 24 ore porti in questo caso alla attivazione di alcuni sistemi cellulari che si oppongono all'aumento di espressione degli elementi ripetuti. Inoltre, si osserva un incremento di espressione dose-dipendente particolarmente evidente per gli elementi L1, i quali mostrano un aumento di espressione statisticamente significativo a seguito del trattamento con HQ alla concentrazione 15 μM per 48 ore (**Figura 35**). Tale aumento di espressione dei retroelementi, per esposizione a idrochinone, è il primo esempio della loro derepressione indotta da un "fattore ambientale" e risulta di particolare rilievo, dal momento che un aumento dei livelli di trascritti di *Alu* e L1 può generare instabilità genomica e interferire con i normali processi cellulari[200].

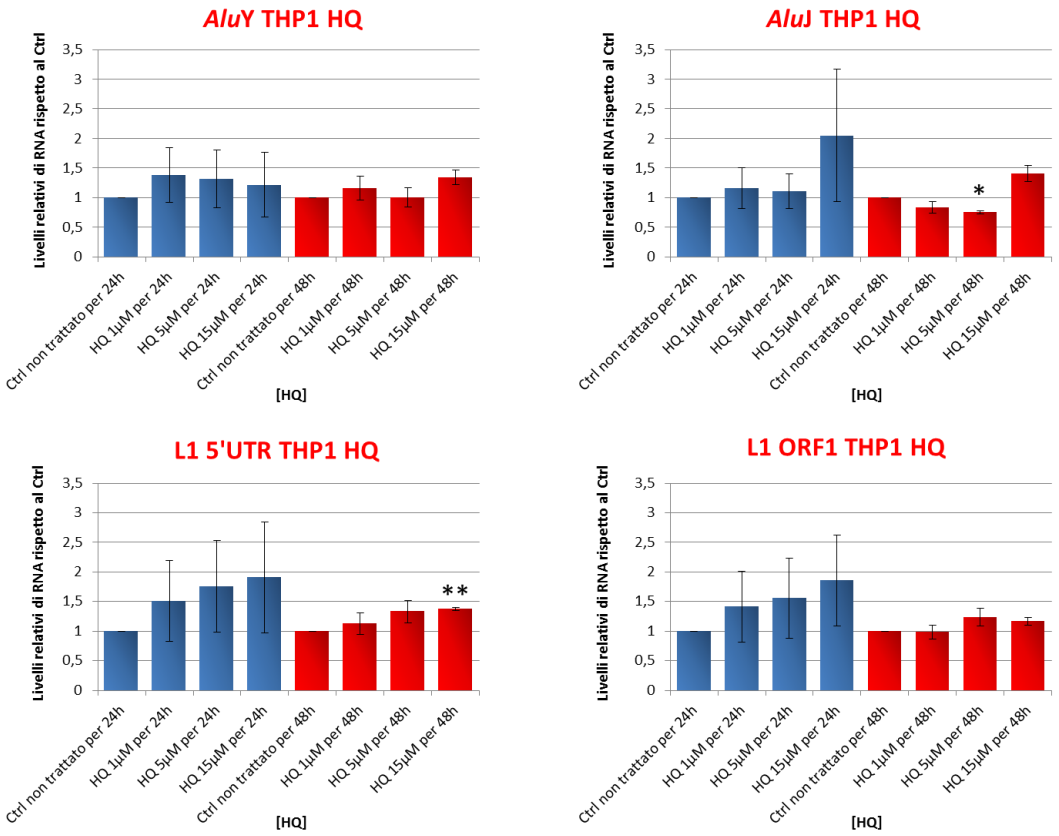


Figura 35. Rappresentazione dei livelli relativi di espressione degli elementi *Alu* e L1 per esposizione delle cellule THP-1 a HQ 1 µM, 5µ M e 15 µM per 24 h (blu) e 48 h (rosso). I livelli relativi di espressione sono stati calcolati usando il metodo di Livak ($2^{-\Delta\Delta Ct}$) e i dati rappresentano la media ± l'errore standard di tre replicati indipendenti. Gli asterischi corrispondono a P-value < 0.05 (*) e P-value < 0.01 (**), applicando il test t di Student per campioni dipendenti accoppiato.

Analisi dei livelli di metilazione degli elementi *Alu* e LINE-1 in cellule THP-1 esposte a HQ

Un recente studio ha messo in correlazione per la prima volta l'esposizione umana a bassi livelli di benzene con l'ipometilazione globale del DNA[339], portando a concludere che tale variazione epigenetica, riscontrata anche nelle cellule tumorali, potrebbe essere alla base dell'ematotossicità del benzene stesso. Dal momento che è noto che il benzene esercita i suoi effetti attraverso i suoi metaboliti e che uno di questi, particolarmente reattivo, è l'idrochinone, sono sempre maggiori i quesiti in merito al ruolo dell'idrochinone nella tossicità del benzene, per capire se sia proprio tale metabolita ad indurre l'ipometilazione globale del genoma. Alcune evidenze del fatto che esso produce una generale ipometilazione del genoma sono state recentemente riportate[347,413], e si è assunto che, essendo gli elementi ripetuti la maggior fonte di siti CpG nel genoma umano, ed essendo essi normalmente silenti (ovvero altamente metilati, v. Introduzione), essi possano essere i maggiormente colpiti da tale fenomeno, nonostante non vi siano dati certi in merito, se non un caso di aumento di 5-idrossimetilcitosine negli elementi L1[338]. A tal proposito, per verificare se l'osservato aumento di espressione indotto da idrochinone nelle cellule THP-1, fosse imputabile ad una derepressione degli elementi *Alu* e LINE-1 mediante loro ipometilazione, dalle stesse cellule è stato estratto il DNA genomico, è stato trattato con bisolfito e amplificato, ed è stato eseguito il pirosequenziamento dei prodotti di PCR[414], in collaborazione con la Prof. Valentina Bollati del Dipartimento di Scienze Cliniche e di Comunità dell'Università degli Studi di Milano. L'analisi è stata condotta su tre famiglie di elementi L1 (L1Pa5, L1Pa2 e L1Hs, riportate dalla evolutivamente più antica alla più recente, v. Introduzione) e due famiglie di elementi *Alu* (*Alu*Yb8 e *Alu*Sx). Le sequenze dei primers sono riportate in Appendice III. Il trattamento con idrochinone ha rivelato che nella maggior parte degli elementi considerati non si hanno apprezzabili variazioni di metilazione. In alcuni casi si ha una leggera ipermetilazione e solo per gli elementi L1Pa5 si osserva una ipometilazione per 48 ore di trattamento (**Figura 36**).

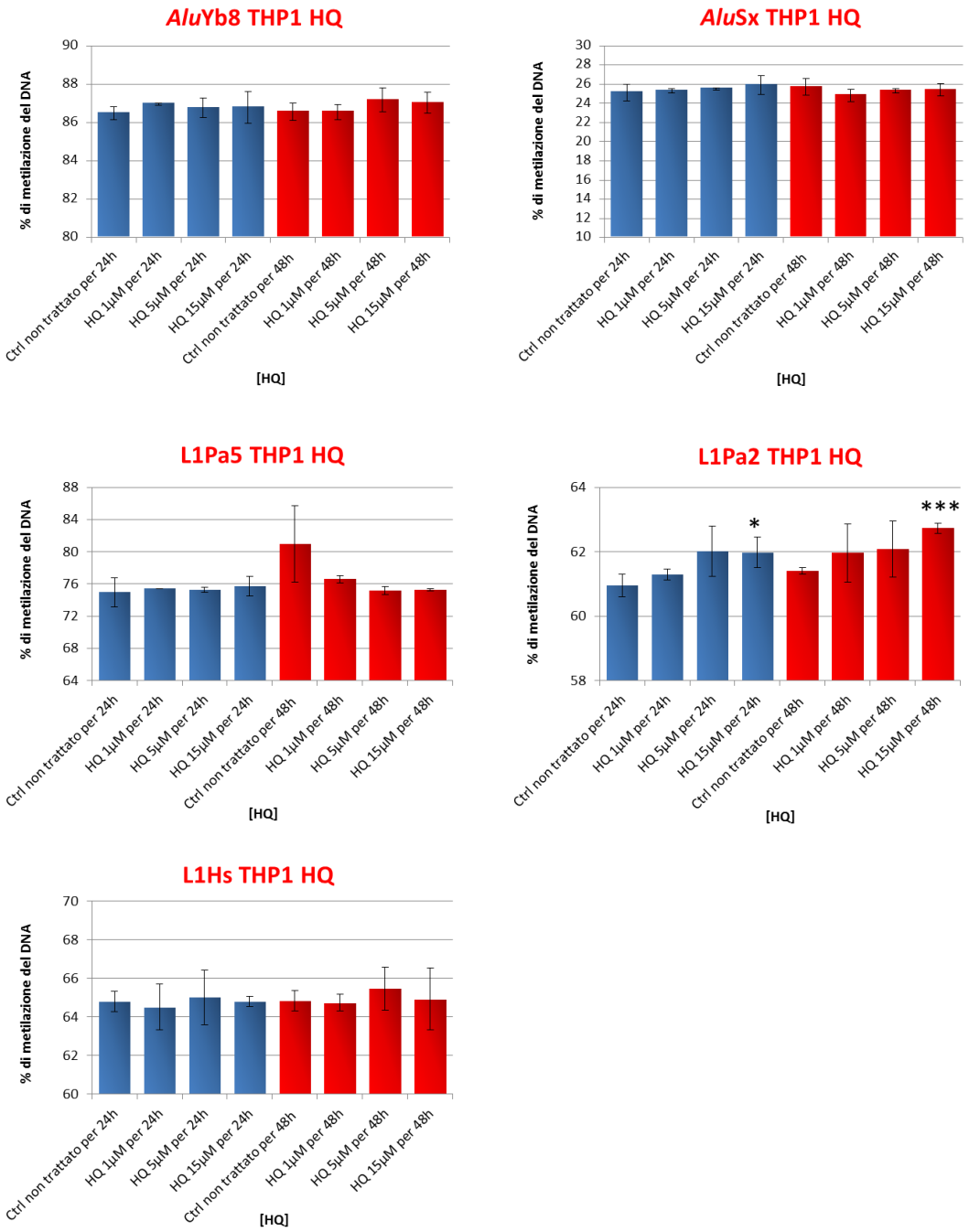


Figura 36. Rappresentazione della percentuale di metilazione degli elementi *Alu* e L1 per esposizione delle cellule THP-1 a HQ 1 µM, 5 µM e 15 µM per 24 h (blu) e 48 h (rosso). I dati rappresentano la media ± la deviazione standard di tre replicati indipendenti. Gli asterischi corrispondono a P-value < 0.05 (*) e P-value < 0.001 (***), applicando il test t di Student per campioni indipendenti a varianza non dissimile.

La nulla o leggera ipermetilazione potrebbe essere spiegata in primo luogo ammettendo che l'HQ, alle concentrazioni utilizzate, non sia in grado di indurre un cambiamento rilevante nello stato di metilazione genomica delle cellule in esame, e ciò è molto interessante, dal momento che si sono utilizzate concentrazioni verosimili di metabolita. Tali concentrazioni, infatti, si avvicinano alle stime delle concentrazioni di idrochinone che dovrebbero essere presenti nel sangue per esposizione a bassi livelli di benzene[347,410] (v. razionale per la somministrazione delle concentrazioni di HQ utilizzate nello studio in Appendice III, realizzato grazie alla collaborazione con la Prof. Silvia Fustinoni dell'Università degli Studi di Milano), pertanto si potrebbe ipotizzare che, alle concentrazioni utilizzate, gli effetti sulla metilazione globale del genoma non siano significativi. Gli sporadici casi riportati della ipometilazione indotta da tale metabolita, infatti, si riferiscono solo ad alcuni geni (effetto locale sulla metilazione e non globale)[409] e inoltre per esposizioni *in vitro* a concentrazioni generalmente maggiori di quelle utilizzate nel nostro contesto sperimentale[413] (pertanto non paragonabili alle esposizioni umane). Al contrario, l'unica evidenza di leggera ipometilazione degli elementi ripetuti, in particolare di L1, è stata riportata per esposizioni a benzene in soggetti umani[339] e non è detto che l'effetto osservato sulle cellule del sangue periferico, sia imputabile al solo idrochinone e non anche ad altri metaboliti.

Analizzando le diverse famiglie di elementi ripetuti, inoltre, è emerso che non tutte presentano uno stesso profilo di metilazione. La nulla o lieve ipermetilazione della maggior parte delle famiglie di elementi *Alu* e L1, e la sola ipometilazione della famiglia L1Pa5, potrebbe essere spiegata dal fatto che ogni famiglia di elementi ripetuti è metilata da una differente isoforma di DNA metiltrasferasi (DNMTs) e che le diverse isoforme hanno una differente sensibilità ai fattori ambientali, come riportano Byun e collaboratori[342]. In questo caso, l'attività delle isoforme che metilano le diverse famiglie di elementi *Alu* e L1 potrebbe rimanere inalterata o essere leggermente incrementata dalla esposizione all'idrochinone. Al contrario, la ipometilazione osservata per gli elementi L1Pa5 per 48 ore di trattamento, può essere spiegata ammettendo che sia ridotta l'espressione o l'attività dell'isoforma di DNMT deputata alla loro metilazione, per esposizione a idrochinone (la risposta di tale isoforma potrebbe essere differente da quella della altre isoforme, come riportano Byun e collaboratori[342]). A tal proposito, gli stessi autori mettono in luce una diversa risposta delle famiglie L1 all'esposizione ambientale (esposizione al PM_{10} (polveri con diametro <10 μm), al benzene e al BC (*Black Carbon*)), ed in particolare che le famiglie evolutivamente più antiche sono più suscettibili. Ciò è in accordo con quanto osservato nel nostro caso, in quanto gli elementi della famiglia L1Pa5 sono ipometilati, tale famiglia è la più antica, e

risponde al trattamento con effetti più marcati rispetto alle famiglie evolutivamente più giovani (L1Pa2 e L1Hs). Infine, un recente studio riporta una ipometilazione indotta da HQ e la correla con l'aumento di 5-idrossimetilcitosine e un incremento della attività delle proteine TET1 (Ten-Eleven protein 1), ovvero diossigenasi in grado di ossidare le 5-metilcitosine e attivare la via di demetilazione del DNA[338]: anche nel nostro contesto sperimentale l'HQ induce ipometilazione degli elementi L1Pa5, e tale ipometilazione potrebbe essere dovuta non solo alla ridotta espressione o attività delle DNMTs, come notato sopra, ma anche all'attivazione della via di demetilazione TET1-dipendente, così come la lieve ipermetilazione, osservata per gli altri elementi ripetuti, potrebbe essere spiegata da una inibizione della stessa via.

Detto ciò, i dati di metilazione fanno supporre che esista un altro meccanismo in grado di portare all'aumento di espressione degli elementi ripetuti, che non coinvolga una loro attivazione mediata dal controllo epigenetico, e che quindi l'idrochinone eserciti un effetto sulla espressione dei retrotrasposoni non correlato con la ipometilazione del DNA. In accordo con tale ipotesi, vi è lo studio di Niesen e collaboratori[218] in cui viene riportato l'aumento di espressione, e quindi l'attivazione trascrizionale, dei geni di classe II del complesso maggiore di istocompatibilità MHC (MHC II, v. Introduzione) nonostante la metilazione del suo promotore. In tale studio, infatti, viene descritto come i geni MHC II siano trascritti dalla RNA polimerasi II e, l'analisi del possibile meccanismo alla base di tale fenomeno, ha svelato il coinvolgimento della proteina RFX, la quale lega sequenze specifiche di DNA metilato. Tale proteina, infatti, media l'attivazione trascrizionale dei geni MHC II facilitando la transizione dallo stato di eterocromatina allo stato di eucromatina a livello del loro promotore metilato, ed è stato ipotizzato che ciò avvenga in quanto RFX sembra in grado di impedire il legame o rimuovere le proteine MBD (le proteine di legame a siti CpG metilati) legate al DNA metilato. La rimozione di tali proteine fa sì che non possano più reclutare il complesso delle istone deacetilasi e di rimodellamento della cromatina, con conseguente perdita dello stato altamente condensato della cromatina stessa, e successiva trascrizione dei geni, indipendentemente dalla metilazione del loro promotore. Il meccanismo riportato per tali geni, potrebbe essere esteso anche ai retrotrasposoni *Alu* e L1, e permetterebbe di spiegare la loro derepressione, indotta da idrochinone, che non si accompagna alla ipometilazione delle loro sequenze (eccezion fatta per gli elementi della famiglia L1Pa5, ipometilati per 48 ore di trattamento e derepressi). L'idrochinone potrebbe aumentare l'affinità della proteina RFX, o di suoi analoghi funzionali, per i siti metilati negli elementi ripetuti e far sì che essi vengano trascritti, spiegando l'aumento di espressione osservato.

Un'altra possibile spiegazione, probabilmente più plausibile della precedente, potrebbe essere data dal fatto che l'analisi di metilazione è estesa a tutti gli elementi ripetuti del genoma che, come detto nella parte introduttiva, sono molto più numerosi dei trascritti (generalmente solo lo 0.1% degli elementi ripetuti è trascritto)[54,200]. Pertanto il sottoinsieme di elementi ripetuti analizzato potrebbe risultare espresso, in quanto derepresso a seguito di un calo di metilazione (conseguente alla ridotta espressione delle DNMTs o alla attivazione delle proteine TET1 da parte dell'HQ), ma la inalterata o leggermente aumentata metilazione della maggior parte degli elementi, potrebbe "mascherare" tale ipometilazione riguardante un sottoinsieme minuscolo di elementi.

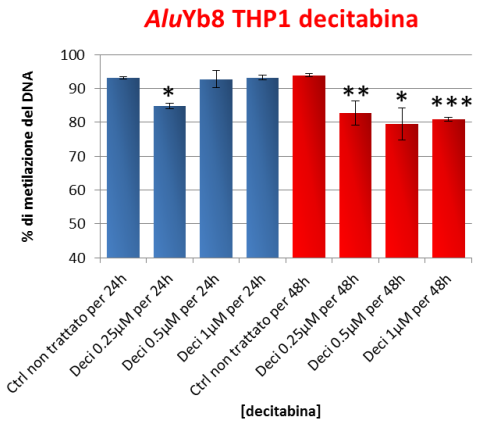
Analisi degli effetti del trattamento delle cellule THP-1 con l'agente ipometilante 5-aza-2'deossicitidina

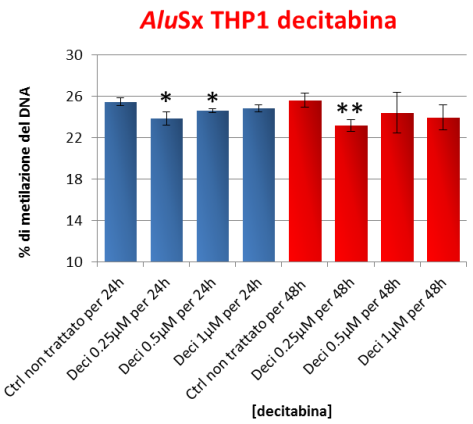
Per verificare le ipotesi avanzate nel paragrafo precedente, si è scelto di condurre trattamenti che dessero una ipometilazione certa: in caso di aumento di espressione dei retrotrasposoni sarebbe stato possibile imputarla al calo di metilazione, in caso contrario si sarebbero potuti ritenere non correlabili i fenomeni di aumento di espressione e di ipometilazione del DNA. Si sono trattate le stesse cellule (THP-1) con tre diverse concentrazioni di 5-aza-2'-deossicitidina (comunemente definita decitabina), un potente inibitore delle DNMTs, pertanto un agente ipometilante noto[415]. La decitabina esplica la sua funzione dopo essere penetrata nella cellula, aver subito una modificazione chimica per aggiunta di tre gruppi fosfato ed essere stata incorporata nel DNA in quanto analogo della base azotata citosina[416]. Questa molecola è anche il principio attivo del farmaco DACOGEN, impiegato nel trattamento dei pazienti con sindromi mielodisplastiche, ed approvato anche per il trattamento della leucemia mieloide acuta[417]. L'utilizzo della decitabina ci è pertanto servito sia come controllo per correlare il calo di metilazione dei retroelementi *Alu* e L1 con la loro eventuale derepressione trascrizionale, sia per indagare gli effetti del principio attivo di un farmaco antileucemico che, qualora incrementasse la trascrizione degli elementi ripetuti, potrebbe avere un effetto secondario, non trascurabile, sui pazienti affetti da sindromi mielodisplastiche e AML. Si è deciso di eseguire dei trattamenti con le concentrazioni di decitabina 0.25 µM, 0.5 µM e 1 µM per 24 e 48 ore, le quali sono state scelte sulla base di esperimenti di ottimizzazione, condotti al fine di individuare quelle che permettessero di ottenere il maggior effetto ipometilante sul genoma (osservando in particolare gli elementi ripetuti *Alu* e LINE-1) in assenza di alterazioni significative della vitalità cellulare.

Come nel caso precedente, dalle stesse cellule THP1 sono stati estratti il DNA genomico e l'RNA totale: il primo è stato trattato con bisolfito, amplificato e sequenziato, mentre il secondo è stato

retrotrascritto ed è stata condotta una qPCR volta a definire le variazioni di espressione degli retroelementi *Alu* e L1.

L'analisi di metilazione (condotta in collaborazione con la Prof. Valentina Bollati del Dipartimento di Scienze Cliniche e di Comunità dell'Università degli Studi di Milano) ha evidenziato, come atteso, che tutti gli elementi ripetuti, indipendentemente dalla famiglia (e quindi dalla DNMT in grado di metilarli), risultano ipometilati (la decitabina inibisce tutte le isoforme di DNMTs indistintamente seppur con diversa efficienza). In particolare si osserva (**Figura 37**) che la ipometilazione è tempo-dipendente sia per gli elementi L1 che per gli elementi *Alu* (maggiore ipometilazione per 48 ore di trattamento) e ciò è in accordo con il meccanismo di azione della decitabina, la quale per integrarsi nel DNA, come analogo della base citosina, necessita di almeno una replicazione cellulare[416] (le cellule THP-1 si replicano in circa 24 ore).





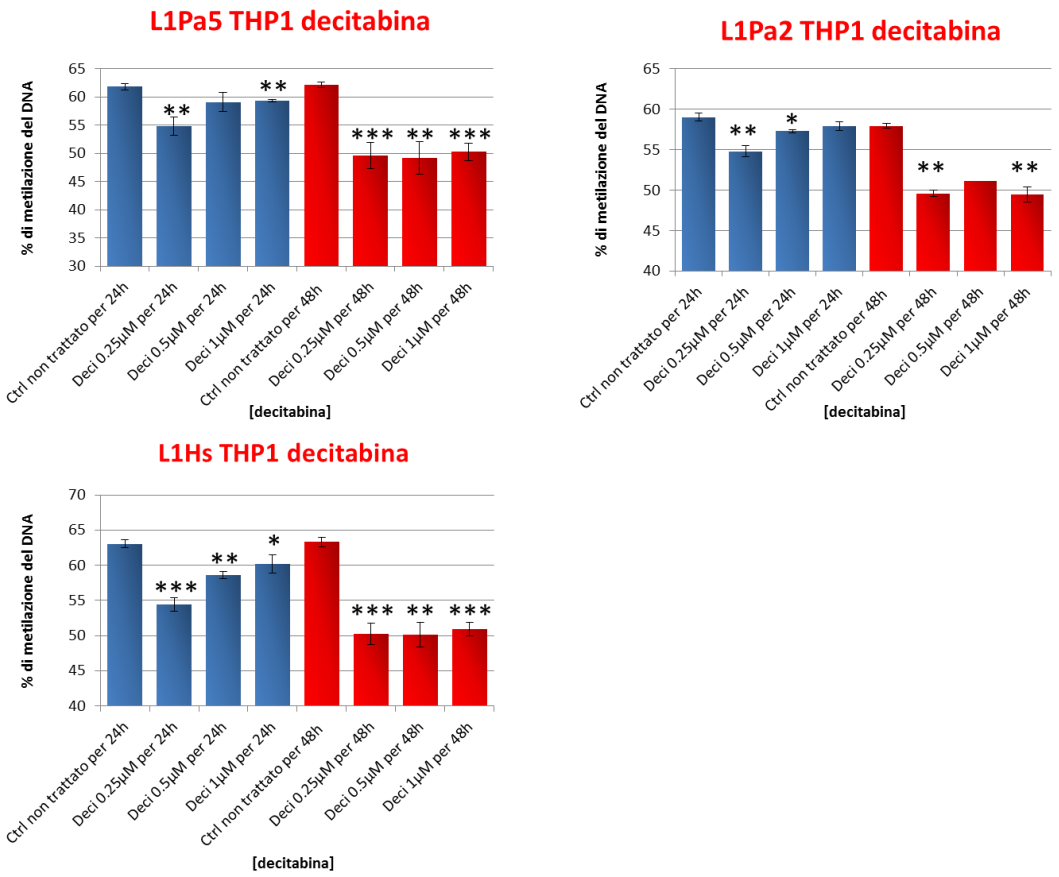


Figura 37. Rappresentazione della percentuale di metilazione degli elementi *Alu* e L1 per esposizione delle cellule THP-1 a decitabina 0.25 µM, 0.5 µM e 1 µM per 24 h (blu) e 48 h (rosso). I dati rappresentano la media ± la deviazione standard di tre replicati indipendenti. Gli asterischi corrispondono a P-value < 0.05 (*), P-value < 0.01 (**) e P-value < 0.001 (***), applicando il test t di Student per campioni indipendenti a varianza non dissimile.

Inoltre, come mostrato in **Figura 38**, viene messo in luce che il trattamento con decitabina porta ad un aumento dell'espressione dei retrotrasposoni, pertanto la ipometilazione del DNA può essere correlata con un aumento dei livelli di espressione di tali elementi. Per gli elementi L1 l'aumento di espressione è tempo-dipendente in accordo con la maggiore ipometilazione per 48 ore di trattamento, mentre per gli elementi *Alu* non sempre si ha una correlazione perfetta tra ipometilazione e aumento di espressione, a causa del limite tecnico rappresentato dal fatto che si considera l'espressione di tutti gli elementi *Alu*, non solo quelli autonomi trascritti dalla Pol III, ma anche quelli facenti parti di trascritti Pol II di maggiori dimensioni.

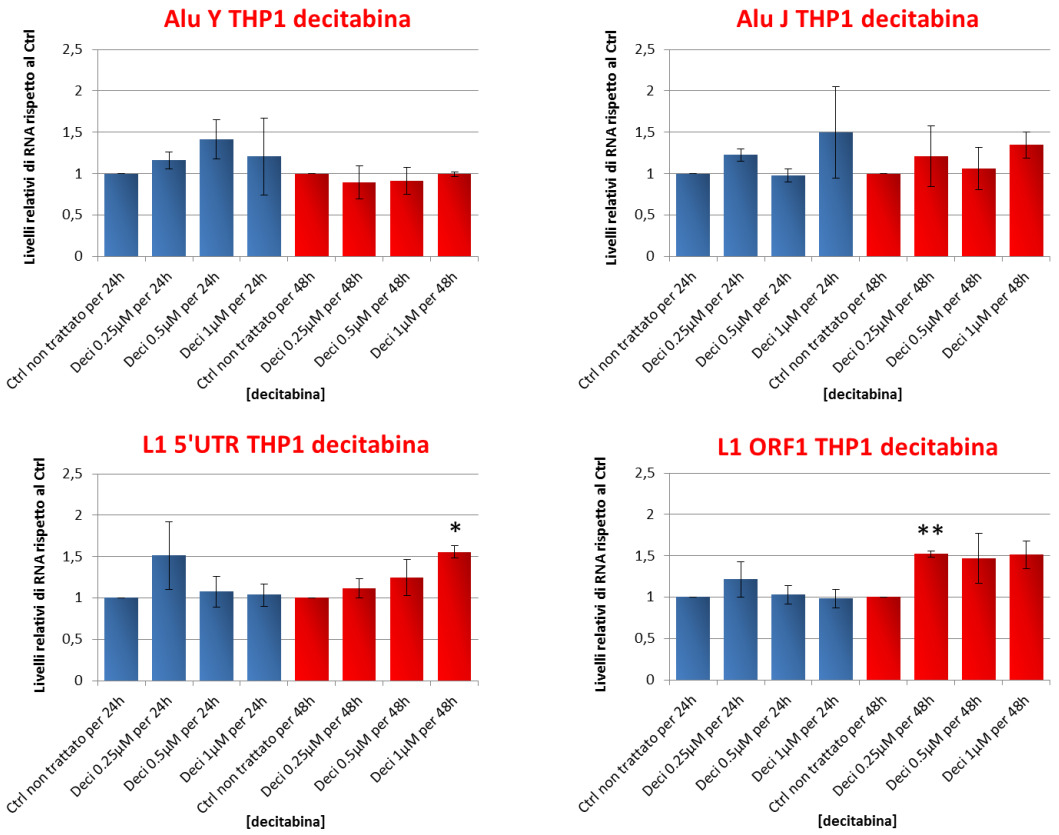


Figura 38. Rappresentazione dei livelli relativi di espressione degli elementi *Alu* e L1 per esposizione delle cellule THP-1 a decitabina 0.25 µM, 0.5 µM e 1 µM per 24 h (blu) e 48 h (rosso). I livelli relativi di espressione sono stati calcolati usando il metodo di Livak ($2^{-\Delta\Delta Ct}$) e i dati rappresentano la media ± l'errore standard di tre replicati indipendenti. Gli asterischi corrispondono a P-value < 0.05 (*), P-value < 0.01 (**), applicando il test t di Student per campioni dipendenti accoppiato.

L'aumento di espressione dei retroelementi *Alu* e L1 messo in evidenza a seguito del trattamento con decitabina, è rilevante in quanto uno dei primi esempi in tale ambito. Vi sono, infatti, solo sporadiche evidenze *in vitro* sia per gli elementi *Alu*[69] che per gli elementi L1[405,418]. Generalmente, infatti, vengono studiati gli effetti di tale agente sulla metilazione dei geni codificanti per proteine[419,420]; per questo, non solo non sono stati finora studiati a fondo gli effetti sulla metilazione degli elementi ripetuti, ma non sono neanche stati analizzati in dettaglio gli effetti sulla loro espressione. L'alterazione dell'espressione degli elementi ripetuti messa in luce utilizzando la decitabina, è rilevante in quanto, trattandosi del principio attivo di un farmaco antileucemico, come accennato all'inizio del paragrafo, il suo effetto non può essere trascurato. Se infatti tale agente provoca un aumento di espressione dei retrotrasposoni come effetto secondario, si potrebbe riscontrare un accumulo dei loro trascritti nelle cellule con conseguente

possibilità di inserzione degli elementi in nuovi siti genomici (v. Introduzione), nonché interferenza con i normali processi cellulari. Le nostre analisi hanno quindi messo in luce come la decitabina sia in grado di dereprimere gli elementi ripetuti e si ritiene necessario indagare a fondo l'aspetto mezionato, in quanto rilevante dal momento che il farmaco viene impiegato normalmente nella cosiddetta "terapia epigenetica" per le sindromi mielodisplastiche e per la leucemia mieloide acuta.

Analisi quantitativa dei livelli di espressione di *Alu* RNA e LINE-1 RNA in cellule staminali ematopoietiche trattate con il derivato del benzene idrochinone

Dal momento che i trattamenti con decitabina hanno permesso di correlare l'aumento di espressione con una riduzione della metilazione del DNA, per fare maggior chiarezza circa gli effetti dell'idrochinone e i suoi meccanismi di azione, si è scelto di condurre gli stessi trattamenti, eseguiti sulle cellule tumorali THP-1, anche su un tipo cellulare non tumorale, in quanto le cellule tumorali potrebbero presentare già di per sé un ridotto livello di metilazione del DNA.

Si sono utilizzate a questo scopo le cellule staminali ematopoietiche da cordone ombelicale (Hematopoietic Stem Cells, HSC), non solo importanti in quanto con fenotipo non tumorale, ma anche considerate uno dei migliori modelli disponibili in mancanza di cellule staminali ematopoietiche di midollo osseo (ardue da reperire e di difficile utilizzo nelle condizioni sperimentali comunemente impiegate), la cui alterazione conduce a leucemia. Il cordone ombelicale, infatti, è noto per essere ricco in cellule progenitrici ematopoietiche e cellule staminali, e può, pertanto, essere un'alternativa conveniente al midollo osseo, in quanto tali cellule hanno la stessa frequenza (ovvero la stessa abbondanza relativa rispetto ad altri tipi cellulari nello stesso distretto), se non maggiore, di quella che avrebbero nel midollo osseo stesso[421]. Inoltre, un recente studio ha messo in luce che l'idrochinone può causare tossicità nelle HCS[422] e che questa potrebbe essere correlata con l'aumentato rischio di leucemia. In questo sistema sperimentale, quindi, l'analisi di espressione dei retroelementi potrebbe far luce su un possibile, non ancora indagato, effetto alla base della sua tossicità. Tali cellule sono state trattate alle stesse tre concentrazioni di HQ (1 µM, 5 µM e 15 µM, sempre per 24 ore e 48 ore) impiegate per i trattamenti delle cellule THP-1, al fine di valutare se gli effetti osservati in precedenza sono o meno tipo cellulare-specifici. L'esperimento è stato condotto indipendentemente tre volte e si è verificata la vitalità cellulare, per essere sicuri di analizzare gli effetti del trattamento con idrochinone, non dovuti, invece, al fatto che le cellule fossero in apoptosi. Ci si è quindi accertati che la maggior parte delle cellule fosse vitale (in collaborazione con la Dr. ssa Lazzari, direttore di "Cell factory R&D", Fondazione

IRCCS Ca' Granda, Ospedale Maggiore Policlinico, Milano), e si è proceduto con l'analisi di espressione (condotta mediante qPCR, come riportato sopra) la quale ha rivelato che, analogamente a quanto osservato nelle cellule THP-1, si ha un generale aumento di espressione dei retroelementi *Alu* e LINE-1 (statisticamente significativo per questi ultimi) che, però, in tale contesto cellulare risulta più marcato a 48 ore che a 24 ore (**Figura 39**).

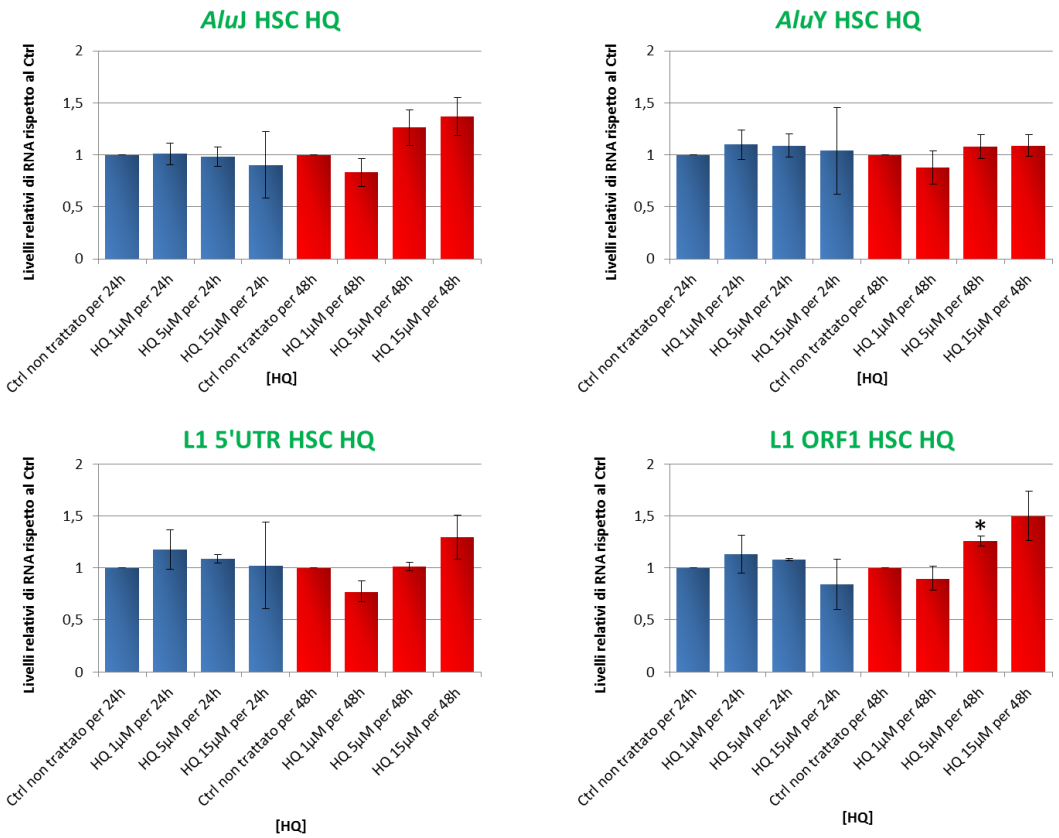


Figura 39. Rappresentazione dei livelli relativi di espressione degli elementi *Alu* e L1 per esposizione delle cellule HSC a HQ 1µM, 5µM e 15µM per 24h (blu) e 48h (rosso). I livelli relativi di espressione sono stati calcolati usando il metodo di Livak ($2^{-\Delta\Delta Ct}$) e i dati rappresentano la media ± l'errore standard di tre replicati indipendenti. Gli asterischi corrispondo al P-value < 0.05 (*), applicando il test t di Student per campioni dipendenti accoppiato.

Si è quindi potuta fare una prima considerazione circa gli effetti dell'HQ sui due diversi tipi cellulari: in entrambi determina un aumento di espressione dei retroelementi *Alu* e L1. Ciò è rilevante, in quanto, anche in questo caso, si sono utilizzate concentrazioni paragonabili a quelle di idrochinone nel sangue, corrispondenti all'esposizione di soggetti umani al benzene entro i limiti

attualmente fissati. I nostri dati confermano, pertanto, che, nonostante si abbia una risposta individuale, tali limiti potrebbero essere ritenuti accettabili per ridurre il rischio di insorgenza della leucemia (qualora un possibile causa di tale patologia sia proprio la derepressione dei retrotrasposoni). Inoltre, si mette in luce che, nonostante in entrambi i tipi cellulari il metabolita del benzene provochi un aumento di espressione dei retroelementi indotto da HQ, è necessario un maggior lasso di tempo affinché questa risposta si produca nelle cellule staminali ematopoietiche. Sulla base delle osservazioni precedenti, è possibile quindi concludere che l'idrochinone, indipendentemente dal tipo cellulare, induce un aumento di espressione dei retroelementi analizzati. In particolare, l'induzione dell'espressione dei retroelementi nelle cellule staminali ematopoietiche, con le sue possibili conseguenze anche in termini di destabilizzazione del genoma, potrebbe essere rilevante come meccanismo attraverso il quale l'idrochinone (e quindi l'esposizione ambientale al benzene a monte) può contribuire alla trasformazione tumorale alla base della leucemia mieloide acuta.

Analisi dei livelli di metilazione degli elementi *Alu* e LINE-1 nelle cellule HSC

Analogamente a quanto riportato per le cellule THP-1, per valutare se l'aumento di espressione degli retroelementi *Alu* e L1 osservato sia imputabile alla loro ipometilazione, si è eseguita l'analisi di metilazione sul DNA estratto dalle stesse cellule e si è osservata una generale tendenza alla ipermetilazione specifica per le diverse famiglie e sottofamiglie di elementi ripetuti (**Figura 40**). Nonostante ciò, anche in questo contesto cellulare si è evidenziata la ipometilazione di una sottofamiglia di elementi L1 (L1Pa2) ad entrambi i tempi di trattamento con HQ. Tale ipometilazione riguarda una famiglia evolutivamente antica e permette di fare considerazioni analoghe a quelle fatte nel caso di cellule THP-1 trattate con HQ[342]. Anche in questa circostanza l'ipotesi più plausibile potrebbe essere che l'idrochinone determini un calo della metilazione dei retrotrasposoni *Alu* e L1 portando a un loro aumento di espressione, e che tale ipometilazione sia "mascherata" dalla ipermetilazione dei più numerosi elementi ripetuti genomici. Tuttavia, potrebbe anche aumentare l'affinità di alcune proteine per i siti CpG metilati degli elementi ripetuti, con conseguente impossibilità di legame per le proteine MBD e successivo loro richiamo delle istone deacetilasi e dei complessi di rimodellamento della cromatina. Come osservato da Niesen e collaboratori[218] per i geni MHC II, si potrebbe ipotizzare una transizione da eterocromatina ad eucromatina e successiva derepressione trascrizionale dei retroelementi *Alu* e L1 che risultano, a livello di sequenza genomica, comunque metilati.

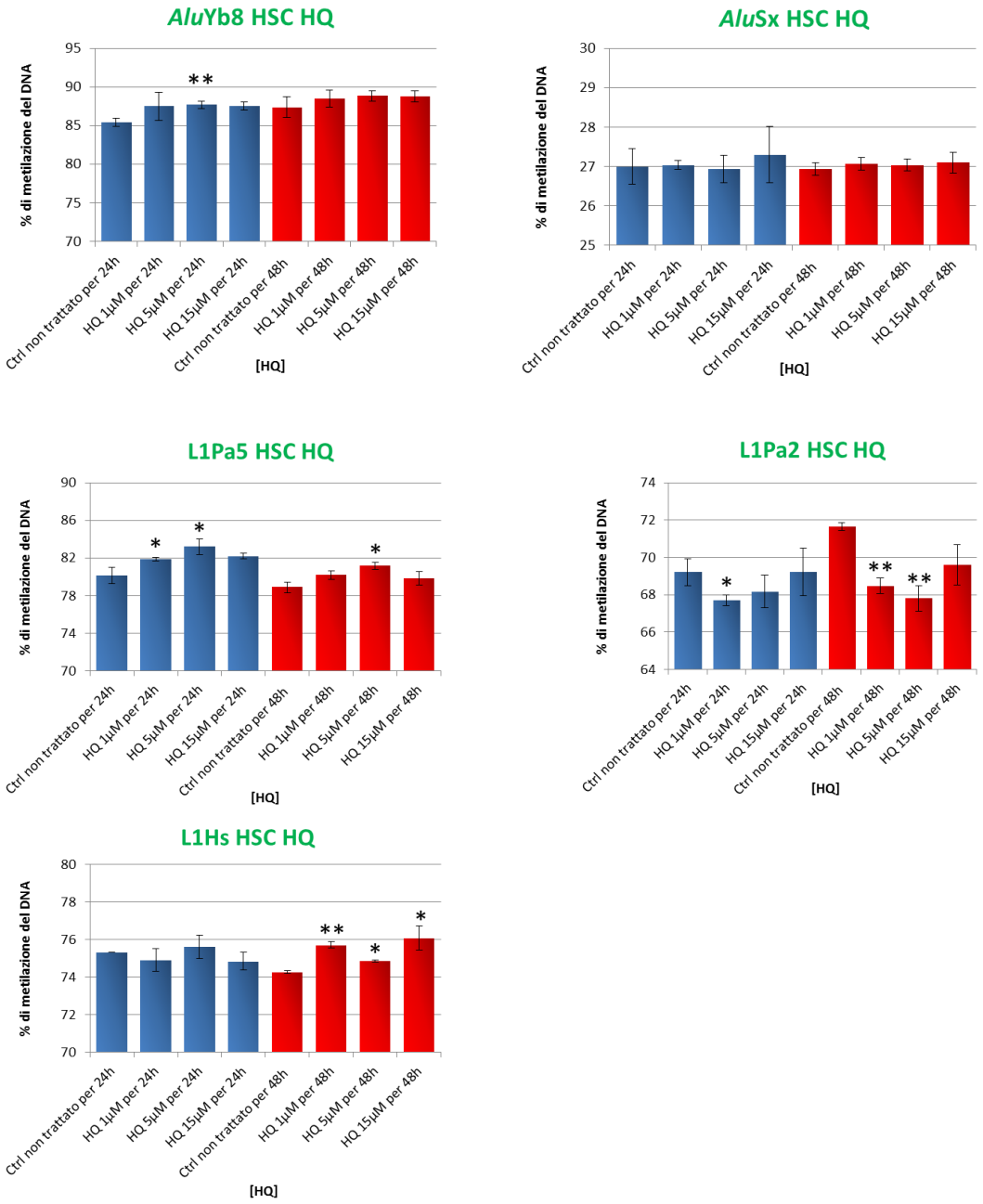


Figura 40. Rappresentazione della percentuale di metilazione degli elementi *Alu* e L1 per esposizione delle cellule HSCs a HQ 1µM, 5µM e 15µM per 24h (blu) e 48h (rosso). I dati rappresentano la media ± la deviazione standard di tre replicati indipendenti. Gli asterischi corrispondono a P-value < 0.05 (*) e P-value < 0.01 (**), applicando il test t di Student per campioni indipendenti a varianza non dissimile.

Conclusioni

Ad oggi sono sempre maggiori i casi riportati di un calo della metilazione dei retrotrasposoni in contesti patologici[407,423,424], e il fatto che possano generare instabilità genomica o alterare i processi cellulari fisiologici ha spinto a studiare più approfonditamente tali elementi, a lungo considerati DNA "spazzatura". Le nostre analisi condotte su 7 linee cellulari tumorali, hanno messo in luce che esistono sottogruppi di elementi *Alu* espressi specifici a seconda della linea cellulare tumorale, portando a concludere che lo studio dei profili di espressione di *Alu* RNA (normalmente silenti perché metilati) può essere utile per monitorare le alterazioni epigenomiche cellulari. È infatti stato ipotizzato, e confermato con le analisi riportate[373], che gli elementi derepressi sono associati ad un particolare intorno cromatinico e presentano uno stato di metilazione del DNA favorevole alla loro espressione, e dal momento che tali caratteristiche sono diverse a seconda del tipo cellulare, è ragionevole supporre che lo studio dei profili di espressione degli *Alu* RNA possa aiutare a comprendere il loro controllo trascrizionale. Ciò è molto importante, anche a fronte delle nostre evidenze che la sovraespressione di questi retroelementi sia in grado di down-regolare sia i geni codificanti che non codificanti per proteine, interferendo in primo luogo con i processi biologici fondamentali per la vita di una cellula (ciclo cellulare, trascrizione, traduzione). A fronte di tali evidenze, e a fronte delle recenti analisi che hanno messo in luce una ipometilazione degli elementi ripetuti per esposizione a diversi fattori ambientali (v. Introduzione), si è deciso di analizzare gli effetti che l'idrochinone può avere sull'espressione dei retroelementi. Tale metabolita del benzene è stato scelto come paradigma dei fattori ambientali a cui l'uomo può essere esposto quotidianamente e che potrebbero alterare l'espressione degli elementi ripetuti *Alu* e L1. Sono infatti oggigiorno numerosi i casi riportati di associazione tra ipometilazione del DNA (per esposizione a fattori ambientali) e patologie, ma non si sono ancora considerati gli effetti esercitati dai fattori ambientali sull'espressione dei retroelementi, in tali contesti patologici. Pertanto, è sembrato interessante rivolgere l'attenzione sull'espressione di elementi *Alu* e L1 e, soprattutto, rilevante, in quanto un aumento dei loro trascritti potrebbe rappresentare un evento potenzialmente mutagenico alla base dell'insorgenza di numerose malattie. Le analisi condotte con idrochinone, a concentrazioni paragonabili a quelle nel sangue per esposizione a livelli di benzene intorno ai valori limite sanciti, ci hanno portato alla scoperta che tale metabolita induce un aumento di espressione dei retroelementi, fenomeno finora inesplorato ma degno della massima attenzione.

Innanzitutto, la derepressione dei retrotratrasposoni indotta da idrochinone è risultata tipo cellulare-indipendente, dal momento che effetti simili si sono osservati sia nelle cellule tumorali THP-1 sia nelle cellule staminali ematopoietiche (non tumorali). Ciò fa supporre, pertanto, che il derivato del benzene eserciti effetti paragonabili in differenti contesti cellulari e che quindi, in ultima analisi, la derepressione degli elementi *Alu* e L1 possa essere un suo effetto peculiare, proprio come verificato per la sua tossicità[422,425,426]. Nonostante il comune effetto, tuttavia, è stata messa in luce anche una diversa tempo-dipendenza nella risposta delle cellule esaminate (si ha una maggiore espressione dei retroelementi per 24 ore di trattamento nelle cellule THP-1 e per 48 ore nelle HSC) e ciò è in accordo con la differente sensibilità cellulare a diversi stimoli (osservata in numerosi studi ed anche messa in luce, nella prima sezione dell'elaborato, dalla nostra analisi delle 7 linee cellulari, le quali presentano l'espressione di sottoinsiemi specifici di *Alu* RNA).

Inoltre, il fatto che il trattamento con idrochinone, a concentrazioni non letali, porti a una derepressione trascrizionale degli elementi *Alu* e L1, è rilevante dato che esso sembra esercitare *in primis* gli effetti tossici del benzene, ma non si è ancora definito con precisione attraverso quale meccanismo. Le nostre analisi permettono, quindi, di delineare un effetto che potrebbe avere un ruolo nell'azione di tale metabolita e che potrebbe, in ultima analisi, rappresentare un possibile meccanismo alla base della connessione tra l'esposizione al benzene e l'aumentato rischio di insorgenza della leucemia mieloide acuta, messa in luce da tempo. A tal proposito, si potrebbe ipotizzare che la generazione di trascritti in grado di retrotrasporsi (provocata dall'idrochinone), possa essere correlata con l'ematotossicità del benzene, dal momento che un aumento della loro espressione potrebbe interferire con i normali processi cellulari, come riportato da Berger e Strub[200], e come evidenziano anche gli esperimenti di sovraespressione di *Alu* RNA, descritti nella seconda sezione della presente tesi di dottorato. È ragionevole supporre che si possa verificare un aumento dell'instabilità genomica per retrotrasposizione degli elementi, che l'abbondanza di trascritti possa influenzare negativamente la trascrizione, la traduzione ed anche numerosi processi post-trascrizionali (come la maturazione degli RNA) ed inoltre, avendo definito che determinati profili di *Alu* RNA possono essere il risultato di alterazioni epigenetiche, si potrebbe avere un cambiamento nell'assetto epigenetico del genoma, con alterazioni dell'espressione anche dei geni non necessariamente ripetuti.

Le analisi volte a definire una possibile correlazione tra la derepressione trascrizionale indotta da HQ e il calo di metilazione nelle sequenze degli elementi ripetuti hanno messo in luce, in primo luogo, che, nel complesso, il metabolita non genera variazioni apprezzabili di metilazione, nonostante si sia osservata la ipometilazione degli elementi della famiglia L1Pa5 (per trattamento

delle cellule THP-1) e L1Pa2 (per trattamento delle cellule HSC), e che quindi possa generare un aumento di espressione di *Alu* e L1 secondo un meccanismo indipendente dalla metilazione del DNA. Tuttavia, come riportato nei paragrafi precedenti, essendo il numero di trascritti di elementi ripetuti nelle cellule circa lo 0.1% dei loci totali presenti nel genoma (sui quali si è eseguita l'analisi di metilazione), è più plausibile supporre che l'idrochinone eserciti una ipometilazione sugli elementi che risultano derepressi, ma che tale effetto venga "mascherato" dalla loro ridotta abbondanza. Infine, dato che numerosi agenti ambientali sono in grado di ipometilare tali elementi, è possibile supporre che inducano anch'essi un aumento dei livelli di espressione dei retrotrasposoni (anche in questo caso non è ancora stato esplorato il meccanismo di azione di tali agenti, ma potrebbe essere simile a quello identificato per l'idrochinone), e che questo possa giustificare, in ultima analisi, la loro abbondanza in contesti patologici[340,427]. In ogni caso, le analisi condotte con idrochinone hanno permesso di approfondire un aspetto molto peculiare dell'espressione del genoma (quello epigenetico nell'accezione più ampia del termine) e gli effetti a questo livello potrebbero rappresentare uno dei meccanismi, finora ignorati, che contribuiscono alle patologie associate alla sua esposizione. Inoltre, dato che l'idrochinone è stato scelto come paradigma di fattori ambientali che possono alterare l'espressione di elementi *Alu* e L1, le stesse analisi (secondo il metodo di identificazione di retroelementi messo a punto) potrebbero in linea di principio essere condotte per tutte le sostanze a cui l'uomo risulta essere esposto, per indagare in modo approfondito gli effetti di tali esposizioni. Ad oggi è noto, infatti, che solo alcuni dei numerosissimi fattori ambientali, inducono una ipometilazione degli elementi ripetuti (osservata soprattutto negli elementi L1) ed è anche riportato che un calo della metilazione potrebbe riattivare tali elementi, come riportato per i geni in singola copia. Detto ciò, analizzare se effettivamente tali retroelementi siano riattivati e aumenti l'espressione dei corrispondenti RNA, potrebbe aggiungere un tassello verso la conoscenza di un possibile meccanismo alla base dell'insorgenza delle patologie indotte dai fattori ambientali in esame, in quanto un incremento di *Alu* RNA e L1 RNA può aumentare la probabilità di inserzioni di tali retrotrasposoni in nuovi loci genomici (si rimanda all'Introduzione) con conseguente generazione dell'instabilità genomica che potrebbe promuovere, in ultima analisi, l'inizio e la progressione dello stato tumorale[50,427]. Inoltre un accumulo di trascritti dei retroelementi può anche interferire con i normali processi cellulari, come riportato soprattutto per gli *Alu* RNA[200].

In conclusione, anche gli esperimenti di controllo condotti con decitabina, agente ipometilante noto, hanno permesso di mettere in luce un aspetto peculiare della sua azione connesso con la derepressione dei retroelementi. I trattamenti, infatti, hanno permesso di constatare che tale

molecola non è solo in grado di indurre la ipometilazione di geni codificanti per proteine normalmente silenziati nelle cellule tumorali (come riportato in letteratura[428]), ma anche un calo di metilazione degli elementi ripetuti, che provoca un loro aumento di espressione. Tale aumento, come ampiamente descritto in precedenza, è plausibile supporre che abbia potenzialità mutageniche e possa essere, insieme ad altri fattori, una delle possibili cause di insorgenza delle patologie, e ciò è da tenere presente essendo la decitabina il principio attivo di un farmaco antileucemico (DACOGEN). Si è, infatti, dimostrata la sua efficacia nel trattamento di sindromi mielodispastiche e leucemia mieloide acuta, ma l'effetto secondario rappresentato da una derepressione degli elementi ripetuti, che le nostre analisi hanno svelato, non può essere trascurato.

A lungo termine, tale effetto potrebbe portare ad un accumulo incontrollato di trascritti non codificanti nelle cellule e provocare l'insorgenza di altre forme tumorali, probabilmente di natura differente rispetto alle leucemie. Le nostre analisi si sono, pertanto, rilevate interessanti anche a fronte di questo fenomeno, e saranno necessarie indagini più approfondite per chiarire un effetto finora ignorato, ma degno della massima attenzione.

Materiali e metodi

Colture cellulari

Le cellule THP-1 (ATCC® TIB-202™) sono state coltivate in terreno RPMI-1640 (Gibco®, Life Technologies) con l'aggiunta di 10% FBS (Gibco®, Life Technologies), 1% Penicillin/Streptomicin (Gibco®, Life Technologies) e 0,05 mM 2-mercaptoetanolo.

Le cellule staminali ematopietiche sono state estratte da sangue di cordone ombelicale e sono state coltivate in terreno CellGro con l'aggiunta di citochine (SCF, Flt3-ligand, TPO, IL6). Si è resa necessaria l'aggiunta di citochine ogni 3-4 giorni per due settimane al fine di stabilizzare la linea cellulare.

Entrambi i tipi cellulari sono stati coltivati a 37°C (5% CO_2, 95% di umidità) sia prima che dopo i trattamenti (v. Risultati e discussione), in flask da 25 cm^2 (Corning®) in un volume finale di terreno pari a 5ml.

Trattamenti con idrochinone e decitabina

Per eseguire i trattamenti delle cellule THP-1 e delle cellule staminali ematopoietiche da cordone ombelicale (HSC) l'idrochinone (Sigma-Alrich) e la decitabina (Abcam) sono stati risospesi nei rispettivi terreni di coltura. Per una completa risospensione, le soluzioni sono state anche sonicate in bagnetto per 10' prima di eseguire i trattamenti. I trattamenti delle cellule THP-1 e HSC sono stati eseguiti con idrochinone alle concentrazioni di 1µM, 5µM e 15µM (v. razionale per la somministrazione di tali concentrazioni in Appendice III), mentre i trattamenti delle cellule THP-1 con decitabina sono stati eseguiti alle concentrazioni di 0.25 µM, 0.5 µM e 1 µM. Tutti i trattamenti sono stati condotti per 24 e 48 ore incubando le cellule a 37°C (5% CO_2, 95% di umidità). Inoltre in parallelo ad ogni set di trattamenti, è stato realizzato anche un controllo non trattato (formato dalle stesse cellule nel solo terreno di coltura). Tutte le cellule sono state piastrate (prima dei trattamenti) alla concentrazione di 300 mila cellule/ml in 5ml di terreno, e sono state trattate dopo un'incubazione di circa 16 ore (overnight), pertanto alla confluenza del 60% - 70%.

Vitalità cellulare

Al termine dei trattamenti, la vitalità delle cellule THP-1 è stata misurata mediante citometria a flusso dopo aver trattato le cellule con propidio ioduro (PI), mentre la vitalità delle cellule staminali ematopoietiche da cordone ombelicale, nonché l'analisi delle cellule apoptotiche, mediante l'utilizzo del kit FITC Annexin V/Dead Cell Apoptosis Kit with FITC annexin V and PI, for Flow Cytometry (Invitrogen), secondo quanto consigliato dalla casa produttrice.

Estrazione e quantificazione del DNA genomico

Il DNA è stato estratto da entrambi i tipi cellulari mediante l'utilizzo del kit Maxwell® 16 LEV Blood DNA Kit (Promega), secondo quanto suggerito dalla casa produttrice e apportando solo alcune modifiche. In breve, per ogni campione le cellule sono state contate nella camera di Burker e sono state sedimentate a 1200 rpm per 5'. In seguito sono state risospese in 300 µl di PBS e sono stati aggiunti 30 µl di Proteinase K Solution e 300 µl si Lysis Buffer. Si è vortexato per 10'' e i campioni sono stati incubati a 56°C per 20'. In seguito si è eseguita l'estrazione automatica del DNA mediante l'utilizzo del AS2000 Maxwell® 16 Instrument (Promega), secondo quanto riportato dalla casa produttrice, e si sono rispesi in 50 µl di Elution Buffer (Promega). Si è eseguita la quantificazione del DNA estratto mediante l'utlizzo del NanoDrop ND-1000 Spectrophotometer V3.7 (Thermo Scientific) e si sono utilizzati per gli esperimenti successivi solo DNA con rapporti A260/A280 >1,8 e A230/A280 >2.

Trattamento con bisolfito di sodio

Per poter mantenere i profili di metilazione durante la PCR successiva, il DNA è stato trattato con il bisolfito di sodio, il quale è in grado di deamminare selettivamente le citosine (ma non le 5-metilcitosine) portando alla formazione di uracile e creando, quindi, una modificazione stabile nel DNA. In tal modo, le citosine che erano originariamente metilate saranno amplificate come citosine, mentre le citosine non metilate verranno trasformate in uracili e in seguito amplificate come timine. La modificazione indotta dal bisolfito sul DNA genomico è stata descritta in precedenza[429]. La procedura prevede, in breve, che 1 µg di DNA venga denaturato in 50 µl di NaOH 2M a 37°C per 20'. In seguito, si aggiungono 30 µl di idrochinone 10 mM (preparato fresco; Sigma-Aldrich) e 520 µl di bisolfito di sodio 3M (Sigma-Aldrich) pH 5 e si incubano i campioni 16h a 50°C. Il DNA trattato con bisolfito è in seguito purificato mediante l'utilizzo del kit Wizard DNA Clean-Up System (Promega), è eluito in 50 µl di acqua e si aggungono 5,5 µl di NaOH per 5'. In seguito il DNA

è lavato con etanolo, precipitato con glicogeno e risospeso in 20 µl di acqua ed è stato conservato a -20°C fino all'utilizzo. Tale procedura è stata condotta come riportato in precedenza[339] e in collaborazione con la Prof. Valentina Bollati del Dipartimento di Scienze Cliniche e di Comunità dell'Università degli Studi di Milano.

PCR e pirosequenziamento (analisi della metilazione del DNA)

Il pirosequenziamento è una tecnica che permette di quantificare accuratamente ogni citosina nell'amplicone sequenziato, nonché di ridurre al minimo la variabilità intra-sequenziamento, e ciò è molto importante per la riproducibilità dei dati ottenuti. Tale tecnica è sostanzialmente un sequenziamento di DNA che è basato sull'identificazione di pirofosfato (PPi) rilasciato durante la sintesi del DNA stesso. A seguito di una cascata di reazioni enzimatiche, viene, infatti, generato un fascio di luce proporzionale al numero di nucleotidi incorporati secondo la sequenza del templato[430]. I saggi sono stati eseguiti in collaborazione con la Prof. Valentina Bollati del Dipartimento di Scienze Cliniche e di Comunità dell'Università degli Studi di Milano.

Per ogni saggio si sono eseguiti dei controlli per verificare che fosse avvenuta efficientemente la conversione con il bisolfito (sia il duplicato della conversione con bisolfito sul DNA genomico di ogni campione che il pirosequenziamanto di ampliconi universali, amplificati da altre linee cellulari), i campioni sono stati sequenziati dopo amplificazione e si sono condotti due replicati tecnici per ogni assay, per avere una maggior affidabilità dei dati.

Per poter avere una stima delle variazioni di metilazione degli elementi ripetuti *Alu* e LINE-1, si sono progettati dei primers specifici per la loro amplificazione e per il successivo pirosequenziamento. I primers sono stati progettati in regioni conservate di differenti famiglie di elementi ripetuti e in regioni prive di polimorfismi a singolo nucleotide (SNP). Si sono progettati primers sulle famiglie L1Pa5, L1Pa2, L1Hs (per gli elementi LINE-1) e *Alu*Yb8, *Alu*Sx (per gli elementi *Alu*), e in ogni saggio sono state valutati siti CpG da un minimo di 1 a un massimo di 5. Le PCR sono state condotte in condizioni standard utilizzando GoTaq Hot Start Polymerase (Promega) secondo quanto indicato dalla casa produttrice. In breve, le reazioni sono condotte con 50 ng di DNA trattato con bisolfito e 1 pmol di entrambi i primers. Uno dei due primers (forward o reverse a seconda di come si disegna il saggio) è legato alla biotina ed è utilizzato per purificare l'amplicone finale usando delle biglie di sefarosio. Il prodotto di PCR, infatti, è legato alla Streptavidin Sepharose High Performance (Amersham Biosciences) e le biglie di sefarosio, contenenti il prodotto di PCR, sono purificate, lavate e denaturate utilizzando una soluzione di NaOH 0,2M.

Inoltre, si è eseguito un secondo lavaggio mediante l'impiego del Pyrosequencing Vacuum Prep Tool (Pyrosequencing, Inc.), secondo quanto indicato dalla casa produttrice, ed infine è stato aggiunto il primer di sequenziamento (0,3 µM), il quale si appaia al prodotto di PCR a singolo filamento. Il pirosequenziamento è stato condotto mediante l'utilizzo del PSQ Q96 MD pyrosequencing System (QIAGEN) come descritto in precedenza[339].
Il grado di metilazione è espresso per ogni locus di DNA come la percentuale di citosine metilate rispetto alla somma delle citosine metilate e non metilate.
I primers utilizzati per le PCR e il pirosequenziamento sono riportati in Appendice III.

Estrazione e quantificazione dell'RNA totale

L'RNA totale è stato estratto mediante l'utilizzo del kit *mir*Vana™ miRNA Isolation Kit (Ambion), secondo quanto suggerito dalla casa produttrice. Al termine della procedura di estrazione dell'RNA totale e dei soli RNA < 200 nt (eseguite separatamente), si sono riunite le due aliquote per ottenere un campione di RNA totale arricchito dei piccoli RNA che, estraendo con Trizol, sarebbero scarsamente rappresentati. L'RNA totale è stato eluito in 52 µl di Elution Solution del kit (Ambion), e si è trattato con TURBO DNA-*free*™ Kit (Ambion), secondo le indicazioni della casa produttrice. In seguito si è eseguita la quantificazione degli RNA al NanoDrop ND-1000 Spectrophotometer V3.7 (Thermo Scientific) e ne è stata verificata l'integrità mediante l'utilizzo del 2100 Agilet Bioanalyser (Assay: Eukaryote total RNA Nano assay; Kit: RNA 6000 Nano kit, Agilet Technologies). Si sono utilizzati per gli esperimenti successivi solo RNA totali con rapporti A260/A280 > 1,9, A230/A280 > 2 e con RIN > 8.

Retrotrascrizione dell'RNA totale

Si è eseguita la retrotrascrizione degli RNA totali estratti, ai quali sono state rimosse eventuali contaminazioni di DNA genomico, utilizzando il kit iScript cDNA Synthesis Kit (Bio-Rad), che contiene una trascrittasi inversa Rnasi H+ e una miscela di primers oligo dT e *random examers*, secondo quanto indicato dalla casa produttrice. Sono stati retrotrascritti 500 ng di templato ad RNA e si sono eseguiti, per ogni campione, controlli no RT, ovvero retrotrascrizioni dei medesimi templati in assenza dell'enzima retrotrascrittasi, necessari per poter evidenziare, nelle reazioni di qPCR successive, eventuali contaminazioni da DNA genomico che potrebbero falsare l'attendibilità dei dati.

PCR quantitativa (qPCR)

Sono state condotte reazioni di PCR quantitativa con lo strumento ABI 7300 (Applied Biosystem 7300 Real Time PCR System) sui cDNA ottenuti dalla retrotrascrizione di 500ng di RNA totale, per ogni campione (v. paragrafo precedente). Le reazioni sono state condotte con coppie di primers specifici per le sequenze di elementi *Alu* e di elementi LINE-1. In particolare, i primers per gli elementi *Alu* sono stati progettati in modo da avere sia un appaiamento prevalente su *Alu* della sottofamiglia *Alu*J (chiamati coppia *Alu*J) che *Alu* delle sottofamiglie evolutivamente più recenti *Alu*S e *Alu*Yb8 (chiamati coppia *Alu*Y), mentre i primers sulle sequenze di elementi L1 sono stati progettati sia sulla 5'UTR sia su ORF1. Inoltre sono stati progettati primers sul gene endogeno della β2-microtubulina (B2M), utilizzato come standard interno. Tutte le coppie di primers per i geni di interesse sono stati disegnati con il Software PrimerExpress® (version 3.0) (Life Technologies) in modo che avessero Tm comprese tra 58°C e 60°C e in grado di produrre degli ampliconi di taglie comprese tra 100 bp e 102 bp. Le concentrazioni finali dei primers utilizzate nelle reazioni di qPCR sono state decise a seguito di prove di efficienza condotte realizzando 5 diluizioni seriali di cDNA (dalle cellule non trattate di controllo) di 10 in 10, a partire dalla diluizione 1:5 di templato. Ogni reazione di qPCR, necessaria per valutare i livelli di espressione, è stata eseguita in un volume finale di 20 µl con 1X Power SYBR® Green PCR Master Mix (Life technologies), 2 µl di cDNA templato (diluito 1:50 nei saggi condotti per gli elementi *Alu* e 1:10 per gli elementi LINE-1) e l'opportuna concentrazione di primers forward e reverse (ottenute mediante le prove empiriche volte a determinare la concentrazione di primers che desse una efficienza vicina al 100%, come riportato sopra). Il programma di PCR che è stato utilizzato è: 50°C per 2', 95°C per 10', 40 cicli a 95°C per 15'' e 60°C per 1', e in seguito una fase di analisi degli ampliconi mediante ricostruzione della curva di dissociazione (95°C per 15'', 60°C per 30'', 95°C per 15''). Tutte le reazioni sono state allestite in triplicato tecnico e si sono calcolate la media e la deviazione standard dei singoli valori di Ct ottenuti. I dati sono stati analizzati in Microsoft Excel, secondo il metodo di Livak[412], ovvero mediante una quantificazione relativa dei livelli di espressione. Si è potuto applicare il metodo di Livak dal momento che sia i geni target che il gene endogeno sono amplificati con efficienze vicino al 100%, e comunque entro il 5% di differenza una dall'altra (requisiti necessari per applicare tale metodo). Per determinare la differenza relativa nel livello di espressione di ciascun gene target nei diversi campioni (a seguito dei diversi trattamenti) dapprima si è calcolata la media dei Ct (Ct) del triplicato tecnico del gene target e del gene endogeno in ciascun campione (controllo non trattato e trattamenti). In seguito la media del triplicato tecnico di ciascun gene target è stata normalizzata

rispetto alla media del triplicato tecnico dello standard interno (B2M), sia nel controllo non trattato che nei trattamenti, ottendo un valore di ΔCt:

$$\Delta Ct \text{ (non trattato)} = Ct \text{ gene target} - Ct \text{ standard interno}$$
$$\Delta Ct \text{ (trattato)} = Ct \text{ gene target} - Ct \text{ standard interno}$$

In seguito il ΔCt del campione trattato è stato normalizzato rispetto al ΔCt del controllo non trattato, ottendo un valore di ΔΔCt:

$$\Delta\Delta Ct = \Delta Ct \text{ (trattato)} - \Delta Ct \text{ (non trattato)}$$

Infine si è calcolato il livello di espressione convertendo in funzione lineare il valore di ΔΔCt:

$$\text{Fold change} = 2^{-\Delta\Delta Ct}$$

Si ottiene, pertanto un valore di fold change (di aumento o di diminuzione) dei geni target nei trattamenti rispetto al controllo non trattato, e tale valore è normalizzato per l'espressione di uno standard interno, ovvero un gene housekeeping che non varia a seguito dei trattamenti. La normalizzazione dell'espressione dei geni target rispetto a quella dello standard interno permette di correggere alcune differenze nel quantitativo di cDNA iniziale nei vari campioni (si compensano gli errori di pipettatura, le differenti efficienze di retrotrascrizione e di amplificazione). Le coppie di primers utilizzate sono riportate in Appendice III.

Analisi statistiche

Per verificare la significatività dei dati di metilazione degli elementi *Alu* e LINE-1 (suddivisi in base alla famiglia di appartenenza) si è applicato il test t di Student per campioni indipendenti a varianza non dissimili (a due code), mentre per la significatività dei dati di espressione dei medesimi elementi, si è applicato il test t di Student per campioni dipendenti accoppiato (a due code). Sono stati ritenuti significativi i dati con valori di P-value < 0.05 ($*$), P-value < 0.01 ($**$)e P-value < 0.001 ($***$).

Appendice III

• Sono riportati i primers utilizzati nello studio.

Primers per le analisi di metilazione di elementi *Alu* e LINE-1:

Primers per PCR e pyrosequenziamento	Primer Forward 5'→3'	Primer Reverse 5'→3'	Primer di sequenziamento 5'→3'
L1Pa5	TTAGTTAAGGGAAGAGGGGATAA	Biotina)ATAAACATAAAACCCTCTAAACCAAACA	TTAGTTAAGGGAAGA
L1Pa2	TTAGATAGTGGGYGTAGGTTAGTGGGT	Biotina)CCTCCRAACCAAATATAAAATATAATCT	GAGTTAAAGAAAGGG
L1Hs	TTTTGAGTTAGGTGTGGGATATA	Biotina)AAAATCAAAAAATTCCCTTTC	AGTTAGGTGTGGGATATAGT
*Alu*Yb8	Biotina)AGATTATTTTGGTTAATAAG	AACTACYAACTACAATAAC	AATAACTAAAACTACAAAC
*Alu*Sx	Biotina)TTTTTATTAAAAATATAAAAATT	CCCAAACTAAAATACAATAA	AATAACTAAAATTACAAAC

Primers per le analisi di espressione di elementi *Alu* e LINE-1:

Primers utilizzati per le qPCR	Sequenze dei primers (5' → 3')
***Alu*J_qRT_fw**	CATAGCGAGACCCCGTCTCTAC
***Alu*J_qRT_rev**	CGATCCTCCTGCCTCAGC
***Alu*Y_qRT_fw**	CAAGGTGAAACCCCGTCTCTAC
***Alu*Y_qRT_rev**	CCATTCTCCTGCCTCAGCC
L1_5'UTR_qRT_fw	CTCAAGTGGGTCCCTGACTCC
L1_5'UTR_qRT_rev	TCTGTTGGAATACCCTGCCG
L1_ORF1_qRT_fw	GGACATTCAAACCAAAGGCAA
L1_ORF1_qRT_rev	ATCAGCTCCTTTAAGCACTTCTCTG
B2M_qRT_fw	GGCTATCCAGCGTACTCCAAAG
B2M_qRT_rev	TGGATGAAACCCAGACACATAGC

• È riportato il razionale alla base della somministrazione delle concentrazioni di idrochinone.

1	concentrazione di benzene inalato (ppm)	1	10	30
1a	concentrazione di benzene inalato (mg/m3)	3,2	32,0	96,0
2	quantità di benzene assorbito (mg) se: assorbimento 70%, ventilazione 1 m3/h, ore lavoro 8/giorno	17,9	179,2	537,6
2a	quantità di benzene assorbito (mmoli)	0,2	2,3	6,9
3	quantità di idrochinone (se il 6.75% del benzene viene convertito a idrochinone, mmoli)	0,016	0,2	0,5
4	concentrazione di idrochinone nel sangue (μM)	3,1	31,0	93,0

Pertanto, facendo il calcolo a ritroso:

1	**concentrazione di idrochinone nel sangue (µM)**	**1**	**5**	**15**
2	quantità di idrochinone (in 5 L sangue, mmoli)	0,005	0,025	0,075
3	quantità di benzene assorbito (mmoli) (se l'idrochinone corrisponde al 6.75% del benzene)	0,07	0,37	1,11
3a	quantità di benzene assorbito (mg) (PM benzene 78)	5,78	28,89	86,67
4	concentrazione di benzene inalato (mg/m3) (l'assorbito è il 70% dell'inalato, la ventilazione polmonare 1 m3/h e la durata 8 h)	1,0	5,2	15,5
4a	concentrazione di benzene inalato (ppm)	0,3	1,6	4,9

Sono riportati i calcoli che giustificano le concentrazioni di idrochinone utilizzate negli esperimenti: le prime due concentrazioni sono una (1 µM) al di sotto e l'altra (5 µM) lievemente al di sopra del valore limite di esposizione al benzene (1ppm) attualmente accettabile, mentre l'ultima concentrazione (15 µM) è intorno ad una esposizione di benzene pari a 5ppm, comunque al di sotto dei livelli ritenuti di alta esposizione[347,431]. Le concentrazioni da somministrare sono state definite in accordo con la Prof. Silvia Fustinoni dell'Università degli Studi di Milano.

Progetti collaterali

Identificazione dei geni trascritti dalla RNA polimerasi III negli eucarioti

Il trascrittoma dell'enzima RNA polimerasi III è molto vasto e comprende numerosi trascritti differenti dagli *Alu* RNA analizzati nelle precedenti sezioni. Pertanto, nell'ambito di un progetto collaterale, a margine del progetto della presente tesi di dottorato, si è voluto approfondire anche questo aspetto, ed in particolare si è rivolta l'attenzione su RNA non codificanti con un promotore di tipo III (le unità trascrizionali *Alu* hanno un promotore di tipo II, sebbene trascritte dallo stesso macchinario trascrizionale) coinvolti in pathway fisiopatologici e nella regolazione dello splicing alternativo, in collaborazione con il professor Aldo Pagano, del Dipartimento di Medicina Sperimentale dell'Università di Genova e dell' IRCCS-AOU San Martino-IST, Genova.

Di rilevante importanza sono le scoperte di nuovi ncRNAs prodotti dalla Pol III che sembrano svolgere ruoli regolativi: il ncRNA 17A[141], il ncRNA 38A[143] ed lo snRNA NMD29[142], di cui si è parlato nell'introduzione e che saranno ripresi brevemente a questo punto della trattazione.

Il ncRNA 17A è sintetizzato da parte della Pol III a seguito di stimoli infiammatori, e determina la produzione della variante B del recettore GABAB R2 con conseguente alterazione delle vie dipendenti dal neurotrasmettitore GABA nelle cellule cerebrali, ed un aumento della secrezione del peptide β amiloide, tipicamente presente nelle cellule cerebrali di individui affetti da Alzheimer (AD, Alzheimer's Disease).

Al contrario, lo snRNA NMD29 riduce lo sviluppo del neuroblastoma aumentando la sensibilità delle cellule tumorali ai trattamenti farmacologici.

Da ultimo, ma non per importanza, è da menzionare il ncRNA 38A, il quale è sintetizzato dalla Pol III (presenta un promotore di tipo 3) sempre a seguito di uno stimolo infiammatorio, e regola lo splicing alternativo del trascritto primario del gene codificante per la proteina che interagisce con i canali potassio di tipo A (KCNIP4), con conseguente alterazione della conduttanza ionica attraverso tali canali e l'aumento della secrezione del peptide β amiloide. Il ncRNA 38A risulta sovraespresso nelle cellule della corteccia cerebrale di pazienti affetti da AD, e ciò conferma il suo ruolo, come nel caso del ncRNA 17A, nei processi neurodegenerativi.

La scoperta e l'analisi del ruolo del ncRNA 38A umano ci hanno spinto a mettere in luce una possibile correlazione tra unità trascrizionali Pol III umane e murine, dal momento che la maggior parte degli studi, volti a chiarire i meccanismi molecolari potenzialmente in grado di contribuire alle patologie umane, si avvalgono del topo come organismo modello.

Risultati e discussione

Identificazione di potenziali unità trascrizionali Pol III murine con promotore di tipo 3

Al fine di determinare se i trascritti umani e murini presentano la medesima collocazione genomica, nonché condividono un'omologia funzionale, si è deciso di analizzare il genoma murino per identificare le unità trascrizionali Pol III con promotore di tipo 3. Mediante l'utilizzo di un recente algoritmo che permette di ottenere tutte le sequenze del genoma murino contenenti la sequenza *consensus* del tipico promotore 3 riconosciuto dalla Pol III, si sono identificate 702 unità trascrizionali Pol III putative in topo, la cui analisi *in silico* supporta un potenziale coinvolgimento nella regolazione dello splicing alternativo e nella fisio-patologia del sistema nervoso. Inoltre, sono state anche individuate 121 unità trascrizionali Pol III con promotore di tipo 3 in comune tra uomo e topo che, peraltro, presentano la medesima collocazione genomica all'interno di specifici geni ortologhi, evidenza che suggerisce una loro possibile omologia funzionale. Tra le unità trascrizionali comuni ai due organismi identificate, si è scelto di concentrare l'attenzione sulla "coppia" ncRNA 38A umano-ncRNA 38B murino, essendo il gene umano già stato ampiamente caratterizzato.

Analisi dell'unità trascrizionale 38B murina e sua trascrizione *in vitro*

L'analisi delle sequenze genomiche corrispondenti al ncRNA 38A umano e al ncRNA 38B murino ha permesso di constatare che essi non presentano la medesima sequenza, ma nonostante ciò, essi mostrano la medesima collocazione genomica: entrambi in antisenso nell'introne 1 del gene codificante per la proteina KCNIP4 (Potassium Channel Interacting Protein 4). In seguito, per verificare che l'unità trascrizionale murina identificata fosse effettivamente trascritta dalla Pol III, la sua sequenza è stata amplificata dal genoma murino, clonata nel vettore pEGFP-N1 e si è eseguita la sua trascrizione *in vitro* in estratto nucleare da cellule HeLa, in presenza di α-amanitina in concentrazione tale da inibire l'attività della RNA polimerasi II. L'analisi dei profili di trascrizione ottenuti ha permesso di concludere che il gene codificante per il ncRNA 38B murino è a tutti gli effetti un'unità trascrizionale riconosciuta e trascritta dalla Pol III (come la sua controparte 38A umana) e che, inoltre, si ha la sintesi di trascritti di taglie in accordo con la presenza di terminatori non canonici all'intero dell'unità trascrizionale stessa[53](**Figura 41**).

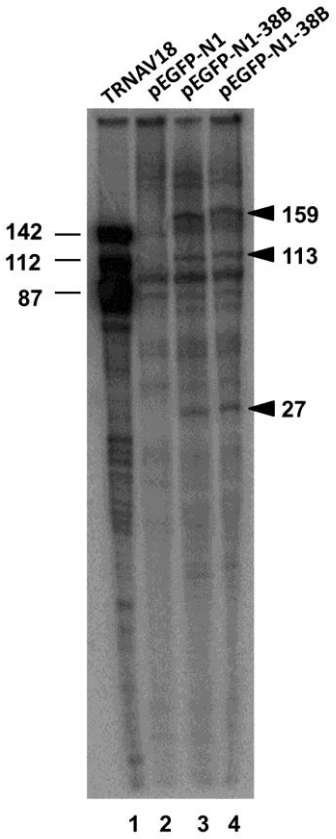


Figura 41. Saggi di trascrizione *in vitro* condotti in estratto nucleare da cellule HeLa. Le reazioni sono state eseguite con 1.5 µg di DNA plasmidico contenente l'unità trascrizionale 38B murina (corsie 3 e 4) e sono state eseguite anche reazioni di controllo con il gene per il $tRNA^{Val}$ (AAC) wt (TRNAV18) che genera tre trascritti di taglia nota (87, 112 e 142 nt, corsia 1) e con il vettore vuoto (corsia 2) per monitorare i livelli di trascrizione aspecifici. I saggi sono stati condotti indipendentemente due volte.

Analisi degli effetti della sovraespressione del ncRNA 38B in cellule neuronali murine

Una volta caratterizzata l'unità trascrizionale 38B murina, per valutare se il corrispondente RNA ragolasse lo splicing alternativo della proteina KCNIP4 come il ncRNA 38A umano, si è analizzato il potenziale aumento di espressione della variante di splicing C della proteina KCNIP4 nei neuroni dell'ippocampo, a seguito della sovraespressione del ncRNA 38B. A tal proposito, il plasmide ricombinante pEGFP-N1-38B (contenente l'unità trascrizionale 38B e il gene per la GFP) è stato trasfettato in colture primarie di neuroni dell'ippocampo ed è stata eseguita anche una trasfezione col plasmide vuoto come controllo (Mock). Al termine delle 48 ore di trasfezione si è confermata non solo la sovraespressione del ncRNA mediante PCR quantitativa, ma si è eseguita anche l'analisi dei livelli di espressione delle varianti di splicing A e C della proteina KCNIP4. I nostri dati

confermano l'avvenuta sovraespressione del ncRNA 38B e inoltre aggiungono che, a seguito di un suo accumulo nelle cellule, si ha una forte inibizione della variante A della KCNIP4, e, al contrario un notevole aumento della variante di splicing C (**Figura 42**).

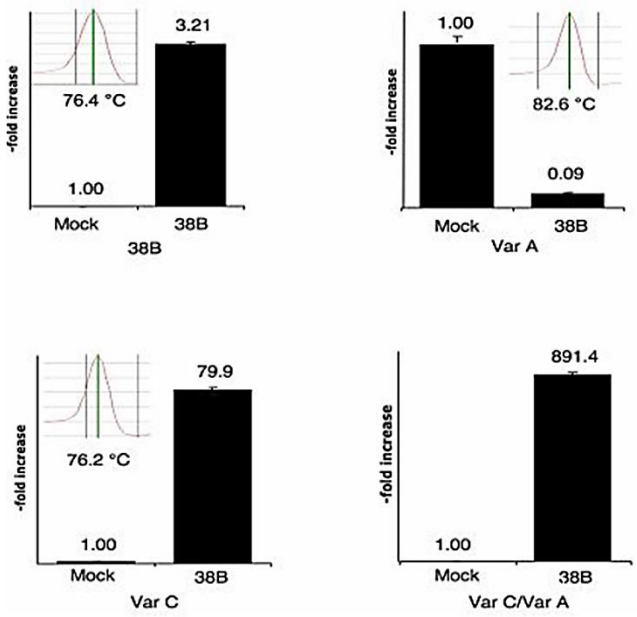


Figura 42. Rappresentazione dei livelli relativi di espressione degli elementi 38B, KCNIP4 varianti A e B a seguito della sovraespressione del ncRNA 38B nelle cellule neuronali di ippocampo murino per 48 h. I livelli relativi di espressione sono stati calcolati usando il metodo di Livak ($2^{-\Delta\Delta Ct}$) e i dati rappresentano la media ± l'errore standard di tre replicati indipendenti.

Quanto riportato permette di concludere che nelle cellule primarie dei neuroni ippocampali murini la sovraespressione del ncRNA 38B determina un accumulo del trascritto il quale a sua volta regola lo splicing alternativo del trascritto primario KCNIP4 con conseguente formazione della variante di splicing C. Ciò è in accordo con quanto riporato per il ncRNA 38A umano, e permette di ipotizzare che la sovraespressione del ncRNA 38B possa avere le stesse conseguente biologiche osservate per la sua controparte umana. Inoltre, la misura della cinetica di inattivazione dei canali potassio di tipo-A, nelle colture primarie delle cellule ippocampali murine trasfetatte con il costrutto pEGFP-N1-38B, ha permesso di osservare un'inibizione della inattivazione veloce di tali canali, proprio come precedentemente scoperto in uomo[143], ed ha portato a concludere che i due RNA possono esercitare lo stesso effetto e quindi che potrebbero presentare un'omologia funzionale. Tali evidenze sostengono fortemente l'ipotesi di un'omologia funzionale tra i ncRNA con promotore di tipo 3 tra topo e uomo.

Conclusioni

L'analisi delle potenziali unità trascrizionali Pol III con promotore di tipo 3 murine ha rivelato l'esistenza di 702 unità che possono essere trascritte dalla Pol III nel genoma murino e che circa un centinaio di queste presentano la stessa localizzazione genomica tra uomo e topo, avvalorando l'ipotesi di una loro possibile omologia funzionale. Di rilievo, inoltre, sono le evidenze di una consistente frazione di unità trascrizionali all'interno degli introni dei geni codificanti per proteine la cui funzione biologica è associata con il sitema nervoso e, in particolare, con la fisio-patologia del cervello, soprattutto a fronte di recenti scoperte del coinvolgimento di ncRNA nella regolazione dello splicing alternativo[143], nonché del cervello come organo in cui si verifica, appunto, la maggior parte degli eventi di splicing alternativo, associati all'attività di unità trascrizionali Pol III con promotore di tipo 3[432]. All'interno di questo contesto, l'isolamento e la caratterizzazione dell'unità trascrizionale 38B murina nei neuroni dell'ipoccampo ha permesso di dimostrare che, nonostante il gene codificante per tale trascritto presenti una sequenza differente rispetto alla controparte 38A umana, i due RNA sono entrambi trascritti dalla Pol III, e condividono non solo la stessa collocazione genomica ma anche la medesima funzione: entrambi sono posti nel primo introne del gene codificante per la proteina KCNIP4 e regolano lo splicing alternativo dell'mRNA ospite. Il fatto che i due trascritti non abbiano la stessa sequenza ma svolgano la stessa funzione nel cervello di uomo e topo, è in accordo con quanto recentemente riporto da Tsirigos e collaboratori, i quali sostengono che gli introni di uomo e topo sono implicati in funzioni comuni indipendentemente dalla loro sequenza[433].

Le nostre evidenze hanno permesso di aprire la strada all'identificazione di numerosi ncRNA murini omologhi a quelli umani. A tal proposito si potrebbero estendere le analisi condotte per la "coppia" ncRNA 38B murino-ncRNA 38A umano su molte unità trascrizionali Pol III identificate in topo, al fine di poter utilizzare topi transgenici per lo studio del ruolo di molti ncRNA, e di poter trasferire le stesse analisi sui rispettivi omologhi funzionali umani. Il fatto che tali ncRNA siano coinvolti nei processi di splicing alternativo aventi luogo nel cervello, con conseguenti effetti patologici rilevanti[144], potrebbe porre le basi per l'utilizzo del silenziamento genico come forma di terapia da affiancare alle terapie farmacologiche classiche.

Inoltre indagando i meccanismi alla base della regolazione epigenetica di tali ncRNA si potrebbe valutare se, a seguito di particolari stimoli (per esempio per esposizione a fattori ambientali connesi con alcune patologie), si abbia una loro derepressione trascrizionale, come osservato per

gli *Alu* RNA e LINE-1 RNA, che porti al loro accumulo a livello cellulare con conseguenti effetti tossici indagabili in topi transgenici e riportabili, in un secondo momento, all'uomo.

Materiali e metodi

Analisi del genoma murino ed identificazione delle potenziali unità trascrizionali Pol III

Per individuare le potenziali unità trascrizionali murine che avessero un promotore riconosciuto dalla Pol III di tipo 3, si è impiegato l'algoritmo COMPASS (COMplex Pattern of Sequence Search Software)[434] utilizzando come input le sequenze dei contig dei 21 cromosomi murini (ottenuti mediante la consultazione della National Centre for Biotechnology Information, NCBI). Per ottenere, in seguito, i promotori di tipo 3 che fossero al'interno di geni Pol II si è impiegato l'algoritmo BLAST (Basic Local Alignment Search Tool) impostando i seguenti parametri: "Nucleotide", database "Mouse genomic + transcript", program "Megablast"[435].

Realizzazione del plasmide ricombinante pEGFP-N1_38B

L'unità trascrizionale corrispondente al ncRNA 38B murino è stata amplificata da DNA murino mediante l'utilizzo dei primers:

38B_ fw: GTCATTAATCCTAGATATCAAAAGGATGG

38B_ rev: AGTCATTAATTACAGATTTAGCATTATACTAA

L'amplicone ottenuto è stato in seguito digerito con l'enzima di restrizione AseI (New Labs) e clonato nel vettore pEGFP-N1 (Clontech), secondo quanto indicato dalla casa produttrice.

Saggi di trascrizioni *in vitro*

Il plasmide pEGFP-N1_38B e il corrispondente plasmide vuoto sono stati purificati con il kit Qiagen Plasmid Mini Kit (Qiagen) prima dell'esecuzione delle reazioni di trascrizione *in vitro*. Queste sono state eseguite in un volume finale di 25 µl con 1500 ng di templato plasmidico, 70 mM KCl, 5mM $MgCl_2$, 1mM DTT, 2.5% glicerolo, 20 mM Tris-HCl pH 7.9, 5mM fosfocreatina, 2 µg/ml α-amanitina, 0.4 U/ml SUPERase-In RNase Inhibitor (Ambion), 20ng TBP di *A. thaliana*, 40 µg di estratto nucleare da cellule HeLa[375], 0.5 mM ATP, CTP e GTP, 0.025 mM UTP e 10 µCi di [α^{32} P] UTP (PerkinElmer). Le miscele di reazione sono state incubate per 60 minuti a 30°C e in seguito interrotte con 75 µl di nuclease-free water e 100 µl di fenolo:cloroformio pH 4.2 (1:1). Gli RNA marcati radioattivamente così purificati, sono stati caricati su gel di poliacrilammide denaturante al 6% (urea 7M) e sono stati visualizzati e quantificati mediante l'utilizzo del Cyclone Phosphor Imager (PerkinElmer) e del software Quantity One (Bio-Rad).

Coltivazione e trasfezione delle cellule neuronali di ippocampo murino

I neuroni ippocampali murini sono stati isolati da 4 topi C57/BL6 di 7 giorni e coltivati in terreno di coltura MEM, con l'aggiunta di glucosio (0.6%) e 10% FBS (Life Technologies). Dopo un giorno, le cellule sono state trasferite in una nuova piastra e coltivate con il terreno Neurobasal/B27 (Life Technologies), preparato secondo quanto indicato dalla casa produttrice. Dopo 8 giorni di coltura i neuroni sono stati trasfettati utilizzando il kit Lipofectamine 2000 (Life Technologies) con i plasmidi pEGFP-N1 vuoto e pEGFP-N1_38B, secondo quanto riportato dalla casa produttrice. Il terreno è stato sostituito con terreno fresco dopo 7 ore e dopo 48 ore di trasfezione si sono sedimentate le cellule per la successiva estrazione dell'RNA mediante l'utilizzo di TRIzol® Reagent (Life Technologies).

Retrotrascrizione e qPCR

Gli RNA totali estratti sono stati retrotrascritti mediante l'impiego del kit Transcriptor First Strand cDNA Synthesis Kit (Roche) secondo quanto indicato dalla casa produttrice e i cDNA ottenuti sono stati amplificati mediante qPCR utilizzando lo strumento PE ABI PRISM@ 7700 Sequence Detection System (Perkin Elmer). I primers utilizzati sono riportati nella tabella alla pagina seguente.

Nome del primer	**Sequenza del primer (5'→3')**
38B_qRT_fw	TGAGTATGTCGATGGAATG TTTTCTAA
38B_qRT_rev	GGGTCTCTAGTGGCAGGTTTGT
KCNIP4_A_qRT_fw	GTGAGAAGGGTGGAAAGCATCT
KCNIP4_A_qRT_rev	AATGCT GCGCTTGGTGTTG
KCNIP4_C_qRT_fw	GGAACAGTTTGGGCTGATTGA
KCNIP4_C_qRT_fw	GTAGCCATCTCCAGCTCATCTTC
mG3PDH_qRT_fw	TGTGTCCGTCGTGGATCTG
mG3PDH_qRT_rev	GATGCCTGCTTCACCACCTT

Dapprima si sono determinate le efficienze dei primers utilizzando diluizioni seriali di cDNA, e in seguito si sono determinati i livelli relativi di espressione dei trascritti target, normalizzati per l'RNA del gene endogeno G3PDH e confrontati rispetto al controllo non trasfettato, secondo quanto riportato precedentemente[436].

Il lavoro descritto in questa parte della tesi di dottorato, svolto nell'ambito di un progetto collaterale, ha portato alla pubblicazione dell'articolo: Bruzzone, M.J. *et al.* The Murine PSE/TATA-dependent transcriptome: evidence of functional homologies with its human counterpart. *Int J Mol Sci* **13**, 14813-27 (2012)[144].

Come riportato all'inizio della sezione, le scoperte sul trascrittoma Pol III umano degli ultimi anni, mettono in luce quanto sia vasto l'insieme di trascritti sintetizzato da tale enzima (che inizialmente si pensava fosse ristretto solo ad alcuni geni costitutivamente espressi e facilmente rilevabili nel genoma) e quanto siano rilevanti i ruoli che tali RNA possono potenzialmente svolgere nelle cellule. A tal proposito, nell'ambito di un progetto collaterale, si è eseguita una revisione della letteratura volta a riassumere tutti i geni trascritti dalla RNA polimerasi III non solo nel genoma umano, ma anche negli altri genomi eucariotici, la quale ha comportato la consultazione di numerosi studi genome-wide condotti negli ultimi tre anni e basati su esperimenti di ChIP-seq[356-358]. Tali studi basati sul crosslinking dei frammenti di DNA *in vivo* con la Pol III e i suoi fattori di trascrizione, immunoprecipitazione cromatinica e sequenziamento del DNA immunoprecipitato, hanno permesso di identificare numerosi nuovi potenziali geni trascritti dalla Pol III soprattutto nelle cellule umane, aprendo la strada verso l'identificazione di ncRNA con un possibile ruolo fisio-patologico.

L'analisi sperimentale da noi condotta su alcuni di tali loci genomici, identificati come associati al macchinario trascrizionale della Pol III dagli studi genome-wide, che sembravano i migliori candidati per una trascrizione *in vitro*, ha tuttavia messo in luce che essi per la maggior parte non vengono trascritti in estratti nucleari da cellule HeLa, facendo concludere, che l'associazione *in vivo* del macchinario trascrizionale a monte di un locus genomico non è sempre condizione sufficiente affinchè avvenga la sua trascrizione (A. Conti e G. Dieci, dati non pubblicati). Si è quindi messo in evidenza, nell'ambito della revisione letteraria, il limite maggiore di tali tecniche genome-wide, ovvero quello di creare falsi positivi e falsi negativi. Con falsi positivi si intendono i loci genomici associati *in vivo* al macchinario trascrizionale Pol III negli studi genome-wide ma che non sono trascritti *in vitro*, mentre con falsi negativi si intendono i loci genomici notoriamente trascritti dalla Pol III non rivelati dalle analisi di ChIP-seq associate al macchinario trascrizionale. A questo proposito si ricorda che la metà dei geni per tRNA trascritti dalla Pol III non è stata trovata associata al macchinario trascrizionale, come invece atteso. Nonostante il limite messo in evidenza, l'analisi della letteratura corrente in merito a tali studi di sequenziamento massivo, ci ha permesso di realizzare un elenco di nuovi loci Pol III individuati nei genomi eucariotici per i quali

sono documentate anche evidenze di trascrizione *in vitro* da parte della Pol III, nonché di un ruolo a livello cellulare, con i quali si potrebbe pensare di ripercorrere il lavoro condotto con altri ncRNA (v. sezioni precenti).
Infine si sono definite anche importanti linee guida per l'identificazione di unità trascrizionali Pol III, come il riconoscimento dei motivi conservati degli elementi del promotore, la posizione del sito di inizio e del terminatore della trascrizione, nonché per l'identificazione di trascritti maturi, al fine di facilitare la loro individuazione, sempre ostica ed ambigua.

Il lavoro descritto in questa parte della tesi di dottorato, inerente alla revisione della letteratura in merito all'identificazione dei geni trascritti della RNA polimerasi III nei genomi eucariotici, ha portato alla pubblicazione dell'articolo: Dieci, G., Conti, A., Pagano, A. & Carnevali, D. Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. *Biochim Biophys Acta* **1829**, 296-305 (2013)[52].

Conclusioni

Nell'ambito della presente tesi di dottorato si sono identificati elementi *Alu* trascritti autonomamente dalla RNA polimerasi III mediante analisi dei dati di RNA-sequencing ottenuti all'interno del progetto ENCODE e si è verificata l'esistenza di un sottogruppo di elementi espressi in un contesto cellulare rispetto agli altri, che potrebbe essere utile in quanto possibile indicatore di un determinato stato cellulare e, in ultima analisi, delle alterazioni dell'epigenoma cellulare. L'idea di trattare modelli cellulari con agenti epigenotossici e in seguito analizzare i trascrittomi di elementi *Alu* e LINE-1 con risoluzione a singolo locus potrebbe essere rilevante in quanto permetterebbe di analizzare su larga scala gli effetti delle perturbazioni indotte da agenti a cui normalmente l'uomo risulta esposto, e ovviare al problema del carattere altamente ripetuto di tali retrotrasposoni. Inoltre, le medesime analisi potrebbero essere estese a soggetti umani direttamente esposti ai fattori ambientali che aumentano il rischio di insorgenza di patologie, con la speranza di identificare profili di espressione dei retrotrasposoni umani associati ad una particolare esposizione. Ciò permetterebbe di poter analizzare in dettaglio il sottogruppo di retroelementi derepresso per esposizione a uno specifico fattore ambientale e determinare, pertanto, uno dei possibili meccanismi alla base dell'azione dell'agente in esame. Si potrebbe, altresì, misurare l'espressione differenziale di singoli *Alu* RNA e L1 RNA in un gruppo di individui esposti rispetto a un gruppo di soggetti non esposti di controllo, con la possibilità di impiegare tale espressione (e quindi tali elementi) come possibile biomarcatore dell'esposizione ambientale. L'impiego della tecnica di RNA-seq, altamente sofisticata e in grado di analizzare i livelli di espressione di singoli *Alu* RNA e L1 RNA potrebbe permettere, inoltre, di superare i limiti dell'approccio seguito nella terza parte della presente tesi di dottorato, per monitorare l'espressione dei retroelementi: è nota infatti la difficoltà di interpretazione dei dati di RT-qPCR condotta sui retroelementi, a causa del loro carattere altamente ripetuto e della loro notevole somiglianza in sequenza. Infine, dal momento che il sequenziamento massivo del trascrittoma cellulare permette anche di analizzare l'espressione differenziale dei geni codificanti e non codificanti per proteine, si potrebbe ripercorrere l'analisi eseguita nella seconda parte del presente elaborato al fine di indagare, nei soggetti esposti, gli effetti che la derepressione dei retrotrasposoni può avere sul trascrittoma cellulare, dal momento che stanno emergendo evidenze di come l'espressione, e il conseguente accumulo, di *Alu* RNA e L1 RNA possono interferire con i processi cellulari fisiologici.

Bibliografia

1. Mc, C.B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**, 344-55 (1950).
2. McClintock, B. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21**, 197-216 (1956).
3. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
4. Smit, A.F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**, 657-63 (1999).
5. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765-8 (1996).
6. Pace, J.K., 2nd & Feschotte, C. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res* **17**, 422-32 (2007).
7. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-82 (2007).
8. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements are active in the human genome? *Trends Genet* **23**, 183-91 (2007).
9. Belancio, V.P., Hedges, D.J. & Deininger, P. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res* **18**, 343-58 (2008).
10. Kazazian, H.H., Jr. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164-6 (1988).
11. Abecasis, G.R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
12. Huang, C.R. *et al.* Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**, 1171-82 (2010).
13. Iskow, R.C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253-61 (2010).
14. Swergold, G.D. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10**, 6718-29 (1990).
15. Babushok, D.V. & Kazazian, H.H., Jr. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28**, 527-39 (2007).
16. Scott, A.F. *et al.* Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**, 113-25 (1987).
17. Martin, S.L. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. *RNA Biol* **7**, 706-11 (2010).
18. Feng, Q., Moran, J.V., Kazazian, H.H., Jr. & Boeke, J.D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905-16 (1996).
19. Moran, J.V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917-27 (1996).

20. Kulpa, D.A. & Moran, J.V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* **13**, 655-60 (2006).
21. Hohjoh, H. & Singer, M.F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *Embo j* **15**, 630-9 (1996).
22. Kulpa, D.A. & Moran, J.V. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet* **14**, 3237-48 (2005).
23. Wei, W. *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**, 1429-39 (2001).
24. Grimaldi, G., Skowronski, J. & Singer, M.F. Defining the beginning and end of KpnI family segments. *Embo j* **3**, 1753-9 (1984).
25. Szak, S.T. *et al.* Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3**, research0052 (2002).
26. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**, 5280-5 (2003).
27. Beck, C.R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-70 (2010).
28. Athanikar, J.N., Badge, R.M. & Moran, J.V. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* **32**, 3846-55 (2004).
29. Becker, K.G., Swergold, G.D., Ozato, K. & Thayer, R.E. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Hum Mol Genet* **2**, 1697-702 (1993).
30. Kuwabara, T. *et al.* Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nat Neurosci* **12**, 1097-105 (2009).
31. Minakami, R. *et al.* Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res* **20**, 3139-45 (1992).
32. Tchenio, T., Casella, J.F. & Heidmann, T. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res* **28**, 411-5 (2000).
33. Yang, N., Zhang, L., Zhang, Y. & Kazazian, H.H., Jr. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* **31**, 4929-40 (2003).
34. Nigumann, P., Redik, K., Matlik, K. & Speek, M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628-34 (2002).
35. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol* **21**, 1973-85 (2001).
36. Matlik, K., Redik, K. & Speek, M. L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **2006**, 71753 (2006).
37. Yang, N. & Kazazian, H.H., Jr. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol* **13**, 763-71 (2006).
38. Faulkner, G.J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**, 563-71 (2009).

39. Macia, A. *et al.* Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol Cell Biol* **31**, 300-16 (2011).
40. Martin, S.L. & Bushman, F.D. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* **21**, 467-75 (2001).
41. Martin, S.L. *et al.* LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. *J Mol Biol* **348**, 549-61 (2005).
42. Basame, S. *et al.* Spatial assembly and RNA binding stoichiometry of a LINE-1 protein essential for retrotransposition. *J Mol Biol* **357**, 351-7 (2006).
43. Goodier, J.L., Ostertag, E.M., Engleka, K.A., Seleme, M.C. & Kazazian, H.H., Jr. A potential role for the nucleolus in L1 retrotransposition. *Hum Mol Genet* **13**, 1041-8 (2004).
44. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
45. Cost, G.J., Feng, Q., Jacquier, A. & Boeke, J.D. Human L1 element target-primed reverse transcription in vitro. *Embo j* **21**, 5899-910 (2002).
46. Luan, D.D., Korman, M.H., Jakubczak, J.L. & Eickbush, T.H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595-605 (1993).
47. Dewannieux, M. & Heidmann, T. Role of poly(A) tail length in Alu retrotransposition. *Genomics* **86**, 378-81 (2005).
48. Boeke, J.D. LINEs and Alus--the polyA connection. *Nat Genet* **16**, 6-7 (1997).
49. Repanas, K. *et al.* Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res* **35**, 4914-26 (2007).
50. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
51. Batzer, M.A. & Deininger, P.L. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**, 370-9 (2002).
52. Dieci, G., Conti, A., Pagano, A. & Carnevali, D. Identification of RNA polymerase III-transcribed genes in eukaryotic genomes. *Biochim Biophys Acta* **1829**, 296-305 (2013).
53. Orioli, A. *et al.* Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res* **39**, 5499-512 (2011).
54. Deininger, P. Alu elements: know the SINEs. *Genome Biol* **12**, 236 (2011).
55. Paolella, G., Lucero, M.A., Murphy, M.H. & Baralle, F.E. The Alu family repeat promoter has a tRNA-like bipartite structure. *Embo j* **2**, 691-6 (1983).
56. Perez-Stable, C., Ayres, T.M. & Shen, C.K. Distinctive sequence organization and functional programming of an Alu repeat promoter. *Proc Natl Acad Sci U S A* **81**, 5291-5 (1984).
57. Kramerov, D.A. & Vassetzky, N.S. SINEs. *Wiley Interdiscip Rev RNA* **2**, 772-86 (2011).
58. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T. & Geiduschek, E.P. Analysis of transcription of the human Alu family ubiquitous repeating element by eukaryotic RNA polymerase III. *Nucleic Acids Res* **9**, 6439-56 (1981).

59. Perez-Stable, C. & Shen, C.K. Competitive and cooperative functioning of the anterior and posterior promoter elements of an Alu family repeat. *Mol Cell Biol* **6**, 2041-52 (1986).
60. Elder, J.T., Pan, J., Duncan, C.H. & Weissman, S.M. Transcriptional analysis of interspersed repetitive polymerase III transcription units in human DNA. *Nucleic Acids Res* **9**, 1171-89 (1981).
61. Chesnokov, I. & Schmid, C.W. Flanking sequences of an Alu source stimulate transcription in vitro by interacting with sequence-specific transcription factors. *J Mol Evol* **42**, 30-6 (1996).
62. Roy, A.M. *et al.* Upstream flanking sequences and transcription of SINEs. *J Mol Biol* **302**, 17-25 (2000).
63. Shaikh, T.H., Roy, A.M., Kim, J., Batzer, M.A. & Deininger, P.L. cDNAs derived from primary and small cytoplasmic Alu (scAlu) transcripts. *J Mol Biol* **271**, 222-34 (1997).
64. Orioli, A., Pascali, C., Pagano, A., Teichmann, M. & Dieci, G. RNA polymerase III transcription control elements: themes and variations. *Gene* **493**, 185-94 (2012).
65. Dieci, G. & Sentenac, A. Detours and shortcuts to transcription reinitiation. *Trends Biochem Sci* **28**, 202-9 (2003).
66. Lunyak, V.V. & Atallah, M. Genomic relationship between SINE retrotransposons, Pol III-Pol II transcription, and chromatin organization: the journey from junk to jewel. *Biochem Cell Biol* **89**, 495-504 (2011).
67. Chang, D.Y. & Maraia, R.J. A cellular protein binds B1 and Alu small cytoplasmic RNAs in vitro. *J Biol Chem* **268**, 6423-8 (1993).
68. Liu, W.M., Chu, W.M., Choudary, P.V. & Schmid, C.W. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23**, 1758-65 (1995).
69. Liu, W.M., Maraia, R.J., Rubin, C.M. & Schmid, C.W. Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. *Nucleic Acids Res* **22**, 1087-95 (1994).
70. Maraia, R.J., Driscoll, C.T., Bilyeu, T., Hsu, K. & Darlington, G.J. Multiple dispersed loci produce small cytoplasmic Alu RNA. *Mol Cell Biol* **13**, 4233-41 (1993).
71. Matera, A.G., Hellmann, U. & Schmid, C.W. A transpositionally and transcriptionally competent Alu subfamily. *Mol Cell Biol* **10**, 5424-32 (1990).
72. Roy-Engel, A.M. LINEs, SINEs and other retroelements: do birds of a feather flock together? *Front Biosci (Landmark Ed)* **17**, 1345-61 (2012).
73. Russanova, V.R., Driscoll, C.T. & Howard, B.H. Adenovirus type 2 preferentially stimulates polymerase III transcription of Alu elements by relieving repression: a potential role for chromatin. *Mol Cell Biol* **15**, 4282-90 (1995).
74. Jagadeeswaran, P., Forget, B.G. & Weissman, S.M. Short interspersed repetitive DNA elements in eucaryotes: transposable DNA elements generated by reverse transcription of RNA pol III transcripts? *Cell* **26**, 141-2 (1981).
75. Ohshima, K., Hamada, M., Terai, Y. & Okada, N. The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Mol Cell Biol* **16**, 3756-64 (1996).

76. Okada, N., Hamada, M., Ogiwara, I. & Ohshima, K. SINEs and LINEs share common 3' sequences: a review. *Gene* **205**, 229-43 (1997).
77. Wallace, N., Wagstaff, B.J., Deininger, P.L. & Roy-Engel, A.M. LINE-1 ORF1 protein enhances Alu SINE retrotransposition. *Gene* **419**, 1-6 (2008).
78. Hsu, K., Chang, D.Y. & Maraia, R.J. Human signal recognition particle (SRP) Alu-associated protein also binds Alu interspersed repeat sequence RNAs. Characterization of human SRP9. *J Biol Chem* **270**, 10179-86 (1995).
79. West, N., Roy-Engel, A.M., Imataka, H., Sonenberg, N. & Deininger, P. Shared protein components of SINE RNPs. *J Mol Biol* **321**, 423-32 (2002).
80. Muddashetty, R. *et al.* Poly(A)-binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. *J Mol Biol* **321**, 433-45 (2002).
81. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**, 41-8 (2003).
82. Roy-Engel, A.M. *et al.* Active Alu element "A-tails": size does matter. *Genome Res* **12**, 1333-44 (2002).
83. Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. *J Mol Biol* **354**, 994-1007 (2005).
84. Ostertag, E.M., Goodier, J.L., Zhang, Y. & Kazazian, H.H., Jr. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**, 1444-51 (2003).
85. Orgel, L.E. & Crick, F.H. Selfish DNA: the ultimate parasite. *Nature* **284**, 604-7 (1980).
86. Kidwell, M.G. & Lisch, D.R. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**, 1-24 (2001).
87. Kazazian, H.H., Jr. Mobile elements: drivers of genome evolution. *Science* **303**, 1626-32 (2004).
88. Britten, R.J. Transposable element insertions have strongly affected human evolution. *Proc Natl Acad Sci U S A* **107**, 19945-8 (2010).
89. Gibbs, R.A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-34 (2007).
90. Deininger, P.L. & Batzer, M.A. Alu repeats and human disease. *Mol Genet Metab* **67**, 183-93 (1999).
91. Cordaux, R. The human genome in the LINE of fire. *Proc Natl Acad Sci U S A* **105**, 19033-4 (2008).
92. Sen, S.K. *et al.* Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* **79**, 41-53 (2006).
93. Carroll, M.L. *et al.* Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311**, 17-40 (2001).
94. Roy-Engel, A.M. *et al.* Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* **159**, 279-90 (2001).
95. Roy-Engel, A.M. *et al.* Non-traditional Alu evolution and primate genomic diversity. *J Mol Biol* **316**, 1033-40 (2002).
96. Roy, A.M. *et al.* Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res* **10**, 1485-95 (2000).

97. Roy, A.M. *et al.* Recently integrated human Alu repeats: finding needles in the haystack. *Genetica* **107**, 149-61 (1999).
98. Smit, A.F., Toth, G., Riggs, A.D. & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246**, 401-417 (1995).
99. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* **16**, 78-87 (2006).
100. Sheen, F.M. *et al.* Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res* **10**, 1496-508 (2000).
101. Myers, J.S. *et al.* A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71**, 312-26 (2002).
102. Badge, R.M., Alisch, R.S. & Moran, J.V. ATLAS: a system to selectively identify human-specific L1 insertions. *Am J Hum Genet* **72**, 823-38 (2003).
103. Boissinot, S., Entezam, A., Young, L., Munson, P.J. & Furano, A.V. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res* **14**, 1221-31 (2004).
104. Wang, J. *et al.* dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat* **27**, 323-9 (2006).
105. Chen, J.M., Stenson, P.D., Cooper, D.N. & Ferec, C. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117**, 411-27 (2005).
106. Mathews, L.M., Chi, S.Y., Greenberg, N., Ovchinnikov, I. & Swergold, G.D. Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am J Hum Genet* **72**, 739-48 (2003).
107. Mills, R.E. *et al.* Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet* **78**, 671-9 (2006).
108. Xing, J. *et al.* Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J Mol Biol* **344**, 675-82 (2004).
109. Hedges, D.J. *et al.* Modeling the amplification dynamics of human Alu retrotransposons. *PLoS Comput Biol* **1**, e44 (2005).
110. Boissinot, S., Chevret, P. & Furano, A.V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17**, 915-28 (2000).
111. Ovchinnikov, I., Rubin, A. & Swergold, G.D. Tracing the LINEs of human evolution. *Proc Natl Acad Sci U S A* **99**, 10522-7 (2002).
112. Salem, A.H. *et al.* LINE-1 preTa elements in the human genome. *J Mol Biol* **326**, 1127-46 (2003).
113. Goodman, M. *et al.* Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9**, 585-98 (1998).
114. Lutz, S.M., Vincent, B.J., Kazazian, H.H., Jr., Batzer, M.A. & Moran, J.V. Allelic heterogeneity in LINE-1 retrotransposition activity. *Am J Hum Genet* **73**, 1431-7 (2003).

115. Seleme, M.C. *et al.* Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proc Natl Acad Sci U S A* **103**, 6611-6 (2006).
116. Deininger, P.L., Batzer, M.A., Hutchison, C.A., 3rd & Edgell, M.H. Master genes in mammalian repetitive DNA amplification. *Trends Genet* **8**, 307-11 (1992).
117. Jurka, J. & Milosavljevic, A. Reconstruction and analysis of human Alu genes. *J Mol Evol* **32**, 105-21 (1991).
118. Jurka, J. & Smith, T. A fundamental division in the Alu family of repeated sequences. *Proc Natl Acad Sci U S A* **85**, 4775-8 (1988).
119. Batzer, M.A. *et al.* Standardized nomenclature for Alu repeats. *J Mol Evol* **42**, 3-6 (1996).
120. Kidwell, M.G. & Lisch, D. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A* **94**, 7704-11 (1997).
121. Lukens, L.N. & Zhan, S. The plant genome's methylation status and response to stress: implications for plant improvement. *Curr Opin Plant Biol* **10**, 317-22 (2007).
122. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
123. Mundade, R., Ozer, H.G., Wei, H., Prabhu, L. & Lu, T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle* **13**, 2847-52 (2014).
124. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
125. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-8 (2012).
126. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
127. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242-6 (2004).
128. Mercer, T.R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30**, 99-104 (2012).
129. Mattick, J.S. The genetic signatures of noncoding RNAs. *PLoS Genet* **5**, e1000459 (2009).
130. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**, 5320-5 (2006).
131. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-7 (2009).
132. Banfai, B. *et al.* Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* **22**, 1646-57 (2012).
133. Tisseur, M., Kwapisz, M. & Morillon, A. Pervasive transcription - Lessons from yeast. *Biochimie* **93**, 1889-96 (2011).
134. Liu, Q. & Paroo, Z. Biochemical principles of small RNA pathways. *Annu Rev Biochem* **79**, 295-319 (2010).

135. Moazed, D. Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457**, 413-20 (2009).
136. Taft, R.J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nat Genet* **41**, 572-8 (2009).
137. Seila, A.C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-51 (2008).
138. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851-4 (2008).
139. Taft, R.J. *et al.* Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol* **17**, 1030-4 (2010).
140. Muse, G.W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39**, 1507-11 (2007).
141. Massone, S. *et al.* 17A, a novel non-coding RNA, regulates GABA B alternative splicing and signaling in response to inflammatory stimuli and in Alzheimer disease. *Neurobiol Dis* **41**, 308-17 (2011).
142. Massone, S. *et al.* NDM29, a RNA polymerase III-dependent non coding RNA, promotes amyloidogenic processing of APP and amyloid beta secretion. *Biochim Biophys Acta* **1823**, 1170-7 (2012).
143. Massone, S. *et al.* RNA polymerase III drives alternative splicing of the potassium channel-interacting protein contributing to brain complexity and neurodegeneration. *J Cell Biol* **193**, 851-66 (2011).
144. Bruzzone, M.J. *et al.* The Murine PSE/TATA-dependent transcriptome: evidence of functional homologies with its human counterpart. *Int J Mol Sci* **13**, 14813-27 (2012).
145. Mercer, T.R., Dinger, M.E. & Mattick, J.S. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155-9 (2009).
146. Wilusz, J.E., Sunwoo, H. & Spector, D.L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23**, 1494-504 (2009).
147. Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M. & Mattick, J.S. Non-coding RNAs: regulators of disease. *J Pathol* **220**, 126-39 (2010).
148. Brown, C.J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527-42 (1992).
149. Clemson, C.M., McNeil, J.A., Willard, H.F. & Lawrence, J.B. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol* **132**, 259-75 (1996).
150. Swiezewski, S., Liu, F., Magusin, A. & Dean, C. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* **462**, 799-802 (2009).
151. Houseley, J., Rubbi, L., Grunstein, M., Tollervey, D. & Vogelauer, M. A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. *Mol Cell* **32**, 685-95 (2008).

152. Bernstein, H.D., Zopf, D., Freymann, D.M. & Walter, P. Functional substitution of the signal recognition particle 54-kDa subunit by its Escherichia coli homolog. *Proc Natl Acad Sci U S A* **90**, 5229-33 (1993).
153. Reeves, M.B., Davies, A.A., McSharry, B.P., Wilkinson, G.W. & Sinclair, J.H. Complex I binding by a virally encoded RNA regulates mitochondria-induced cell death. *Science* **316**, 1345-8 (2007).
154. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).
155. Pang, K.C., Frith, M.C. & Mattick, J.S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**, 1-5 (2006).
156. Wang, J. *et al.* Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature* **431**, 1 p following 757; discussion following 757 (2004).
157. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**, e1000371 (2010).
158. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-27 (2011).
159. Clark, M.B. & Mattick, J.S. Long noncoding RNAs in cell biology. *Semin Cell Dev Biol* **22**, 366-76 (2011).
160. Ponting, C.P., Oliver, P.L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629-41 (2009).
161. Martianov, I., Ramadass, A., Serra Barros, A., Chow, N. & Akoulitchev, A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445**, 666-70 (2007).
162. Rinn, J.L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-23 (2007).
163. Rintala-Maki, N.D. & Sutherland, L.C. Identification and characterisation of a novel antisense non-coding RNA from the RBM5 gene locus. *Gene* **445**, 7-16 (2009).
164. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**, 925-38 (2010).
165. Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**, 756-69 (2008).
166. Campalans, A., Kondorosi, A. & Crespi, M. Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. *Plant Cell* **16**, 1047-59 (2004).
167. Kloc, M. *et al.* Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. *Development* **132**, 3445-57 (2005).

168. Mourtada-Maarabouni, M., Hedge, V.L., Kirkham, L., Farzaneh, F. & Williams, G.T. Growth arrest in human T-cells is controlled by the non-coding RNA growth-arrest-specific transcript 5 (GAS5). *J Cell Sci* **121**, 939-46 (2008).
169. Sheik Mohamed, J., Gaughwin, P.M., Lim, B., Robson, P. & Lipovich, L. Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *Rna* **16**, 324-37 (2010).
170. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113-7 (2010).
171. Espinoza, C.A., Allen, T.A., Hieb, A.R., Kugel, J.F. & Goodrich, J.A. B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. *Nat Struct Mol Biol* **11**, 822-9 (2004).
172. Shamovsky, I., Ivannikov, M., Kandel, E.S., Gershon, D. & Nudler, E. RNA-mediated response to heat shock in mammalian cells. *Nature* **440**, 556-60 (2006).
173. Tano, K. & Akimitsu, N. Long non-coding RNAs in cancer progression. *Front Genet* **3**, 219 (2012).
174. Wapinski, O. & Chang, H.Y. Long noncoding RNAs and human disease. *Trends Cell Biol* **21**, 354-61 (2011).
175. Hangauer, M.J., Vaughn, I.W. & McManus, M.T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9**, e1003569 (2013).
176. Khalil, A.M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-72 (2009).
177. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-10 (2010).
178. Prensner, J.R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* **29**, 742-9 (2011).
179. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-19 (2010).
180. Lee, J.T. Epigenetic regulation by long noncoding RNAs. *Science* **338**, 1435-9 (2012).
181. Hirota, K. *et al.* Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**, 130-4 (2008).
182. Wang, K.C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-4 (2011).
183. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-93 (2010).
184. Chu, C., Qu, K., Zhong, F.L., Artandi, S.E. & Chang, H.Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell* **44**, 667-78 (2011).
185. Nguyen, V.T., Kiss, T., Michels, A.A. & Bensaude, O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414**, 322-5 (2001).

186. Yakovchuk, P., Goodrich, J.A. & Kugel, J.F. B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes. *Proc Natl Acad Sci U S A* **106**, 5569-74 (2009).
187. Mariner, P.D. *et al.* Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol Cell* **29**, 499-509 (2008).
188. Goodrich, J.A. & Kugel, J.F. Dampening DNA binding: a common mechanism of transcriptional repression for both ncRNAs and protein domains. *RNA Biol* **7**, 305-9 (2010).
189. Orom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).
190. Tsuiji, H. *et al.* Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes Cells* **16**, 479-90 (2011).
191. Parrott, A.M. *et al.* The evolution and expression of the snaR family of small non-coding RNAs. *Nucleic Acids Res* **39**, 1485-500 (2011).
192. Pagano, A. *et al.* New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. *PLoS Genet* **3**, e1 (2007).
193. Gong, C. & Maquat, L.E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**, 284-8 (2011).
194. Sumazin, P. *et al.* An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* **147**, 370-81 (2011).
195. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358-69 (2011).
196. Feng, J. *et al.* The RNA component of human telomerase. *Science* **269**, 1236-41 (1995).
197. Hellwig, S. & Bass, B.L. A starvation-induced noncoding RNA modulates expression of Dicer-regulated genes. *Proc Natl Acad Sci U S A* **105**, 12897-902 (2008).
198. Hung, T. & Chang, H.Y. Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol* **7**, 582-5 (2010).
199. Hu, Q. *et al.* DICER- and AGO3-dependent generation of retinoic acid-induced DR2 Alu RNAs regulates human stem cell proliferation. *Nat Struct Mol Biol* **19**, 1168-75 (2012).
200. Berger, A. & Strub, K. Multiple Roles of Alu-Related Noncoding RNAs. *Prog Mol Subcell Biol* **51**, 119-46 (2011).
201. Piriyapongsa, J., Marino-Ramirez, L. & Jordan, I.K. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**, 1323-37 (2007).
202. Hadjiargyrou, M. & Delihas, N. The Intertwining of Transposable Elements and Non-Coding RNAs. *Int J Mol Sci* **14**, 13307-28 (2013).
203. Karginov, F.V. & Hannon, G.J. The CRISPR system: small RNA-guided defense in bacteria and archaea. *Mol Cell* **37**, 7-19 (2010).

204. Sigurdsson, M.I., Smith, A.V., Bjornsson, H.T. & Jonsson, J.J. The distribution of a germline methylation marker suggests a regional mechanism of LINE-1 silencing by the piRNA-PIWI system. *BMC Genet* **13**, 31 (2012).
205. Ohms, S. & Rangasamy, D. Silencing of LINE-1 retrotransposons contributes to variation in small noncoding RNA expression in human cancer cells. *Oncotarget* **5**, 4103-17 (2014).
206. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-43 (2008).
207. Garcia-Bellido, A. Genetic control of wing disc development in Drosophila. *Ciba Found Symp* **0**, 161-82 (1975).
208. Guibert, S. & Weber, M. Functions of DNA methylation and hydroxymethylation in mammalian development. *Curr Top Dev Biol* **104**, 47-83 (2013).
209. Pradhan, S. & Esteve, P.O. Mammalian DNA (cytosine-5) methyltransferases and their expression. *Clin Immunol* **109**, 6-16 (2003).
210. Jin, B., Li, Y. & Robertson, K.D. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* **2**, 607-17 (2011).
211. Probst, A.V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol* **10**, 192-206 (2009).
212. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3**, 662-73 (2002).
213. Egger, G. *et al.* Identification of DNMT1 (DNA methyltransferase 1) hypomorphs in somatic knockouts suggests an essential role for DNMT1 in cell survival. *Proc Natl Acad Sci U S A* **103**, 14080-5 (2006).
214. Riggs, A.D. & Xiong, Z. Methylation and epigenetic fidelity. *Proc Natl Acad Sci U S A* **101**, 4-5 (2004).
215. Schaefer, M. *et al.* RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev* **24**, 1590-5 (2010).
216. Varriale, A. DNA methylation, epigenetics, and evolution in vertebrates: facts and challenges. *Int J Evol Biol* **2014**, 475981 (2014).
217. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-22 (2011).
218. Niesen, M.I. *et al.* Activation of a methylated promoter mediated by a sequence-specific DNA-binding protein, RFX. *J Biol Chem* **280**, 38914-22 (2005).
219. Rubin, C.M., VandeVoort, C.A., Teplitz, R.L. & Schmid, C.W. Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res* **22**, 5121-7 (1994).
220. Ichiyanagi, K. Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, SINEs. *Genes Genet Syst* **88**, 19-29 (2013).
221. Hata, K. & Sakaki, Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227-34 (1997).
222. Hedges, D.J. & Deininger, P.L. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* **616**, 46-59 (2007).

223. Baillie, J.K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534-7 (2011).
224. Paulson, K.E. & Schmid, C.W. Transcriptional inactivity of Alu repeats in HeLa cells. *Nucleic Acids Res* **14**, 6145-58 (1986).
225. Zhang, Y.W. *et al.* Methylation Contributes to the Imbalance of PRDM1alpha/PRDM1beta Expression in Diffuse Large B-cell Lymphoma. *Leuk Lymphoma*, 1-26 (2014).
226. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-5 (2011).
227. Carreira, P.E., Richardson, S.R. & Faulkner, G.J. L1 retrotransposons, cancer stem cells and oncogenesis. *Febs j* **281**, 63-73 (2014).
228. Esteller, M. & Herman, J.G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J Pathol* **196**, 1-7 (2002).
229. Belkhiri, A. & El-Rifai, W. 5-Methylcytosine hydroxylation-mediated LINE-1 hypomethylation: a novel mechanism of proto-oncogenes activation in colorectal cancer? *Gut* **63**, 538-9 (2014).
230. Sunami, E., de Maat, M., Vu, A., Turner, R.R. & Hoon, D.S. LINE-1 hypomethylation during primary colon cancer progression. *PLoS One* **6**, e18884 (2011).
231. Schulz, W.A. L1 retrotransposons in human cancers. *J Biomed Biotechnol* **2006**, 83672 (2006).
232. Slotkin, R.K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**, 272-85 (2007).
233. Munzel, M. *et al.* Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angew Chem Int Ed Engl* **49**, 5375-7 (2010).
234. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-5 (2009).
235. Ko, M. *et al.* TET proteins and 5-methylcytosine oxidation in hematological cancers. *Immunol Rev* **263**, 6-21 (2015).
236. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-3 (2011).
237. Shen, L., Song, C.X., He, C. & Zhang, Y. Mechanism and function of oxidative reversal of DNA and RNA methylation. *Annu Rev Biochem* **83**, 585-614 (2014).
238. Nabel, C.S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat Chem Biol* **8**, 751-8 (2012).
239. Heyn, H. & Esteller, M. DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* **13**, 679-92 (2012).
240. Esteller, M. *et al.* Inactivation of glutathione S-transferase P1 gene by promoter hypermethylation in human neoplasia. *Cancer Res* **58**, 4515-8 (1998).
241. Esteller, M. *et al.* Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med* **343**, 1350-4 (2000).
242. Pepenella, S., Murphy, K.J. & Hayes, J.J. Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma* **123**, 3-13 (2014).

243. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-95 (2011).
244. Allfrey, V.G., Faulkner, R. & Mirsky, A.E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci U S A* **51**, 786-94 (1964).
245. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419**, 407-11 (2002).
246. Lachner, M. & Jenuwein, T. The many faces of histone lysine methylation. *Curr Opin Cell Biol* **14**, 286-98 (2002).
247. White, R.J. Transcription by RNA polymerase III: more complex than we thought. *Nat Rev Genet* **12**, 459-63 (2011).
248. Yan, C. & Boyd, D.D. Histone H3 acetylation and H3 K4 methylation define distinct chromatin regions permissive for transgene expression. *Mol Cell Biol* **26**, 6357-71 (2006).
249. Fahrner, J.A., Eguchi, S., Herman, J.G. & Baylin, S.B. Dependence of histone modifications and gene expression on DNA hypermethylation in cancer. *Cancer Res* **62**, 7213-8 (2002).
250. Kondo, Y. & Issa, J.P. Enrichment for histone H3 lysine 9 methylation at Alu repeats in human cells. *J Biol Chem* **278**, 27658-62 (2003).
251. Karimi, M.M. *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* **8**, 676-87 (2011).
252. Huda, A., Marino-Ramirez, L. & Jordan, I.K. Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. *Mob DNA* **1**, 2 (2010).
253. Keller, C. & Buhler, M. Chromatin-associated ncRNA activities. *Chromosome Res* **21**, 627-41 (2013).
254. Krishnan, J. & Mishra, R.K. Emerging trends of long non-coding RNAs in gene activation. *Febs j* **281**, 34-45 (2014).
255. Wogan, G.N. Molecular epidemiology in cancer risk assessment and prevention: recent progress and avenues for future research. *Environ Health Perspect* **98**, 167-78 (1992).
256. Reamon-Buettner, S.M., Mutschler, V. & Borlak, J. The next innovation cycle in toxicogenomics: environmental epigenetics. *Mutat Res* **659**, 158-65 (2008).
257. Bollati, V. & Baccarelli, A. Environmental epigenetics. *Heredity (Edinb)* **105**, 105-12 (2010).
258. Dolinoy, D.C. & Jirtle, R.L. Environmental epigenomics in human health and disease. *Environ Mol Mutagen* **49**, 4-8 (2008).
259. Baccarelli, A. & Bollati, V. Epigenetics and environmental chemicals. *Curr Opin Pediatr* **21**, 243-51 (2009).
260. Hou, L., Zhang, X., Wang, D. & Baccarelli, A. Environmental chemical exposures and human epigenetics. *Int J Epidemiol* **41**, 79-105 (2012).
261. Ho, S.M. *et al.* Environmental epigenetics and its implication on disease risk and health outcomes. *Ilar j* **53**, 289-305 (2012).

262. Valko, M., Morris, H. & Cronin, M.T. Metals, toxicity and oxidative stress. *Curr Med Chem* **12**, 1161-208 (2005).
263. Valinluck, V. *et al.* Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res* **32**, 4100-8 (2004).
264. Turk, P.W., Laayoun, A., Smith, S.S. & Weitzman, S.A. DNA adduct 8-hydroxyl-2'-deoxyguanosine (8-hydroxyguanine) affects function of human DNA methyltransferase. *Carcinogenesis* **16**, 1253-5 (1995).
265. Nordstrom, D.K. Public health. Worldwide occurrences of arsenic in ground water. *Science* **296**, 2143-5 (2002).
266. Lokuge, K.M., Smith, W., Caldwell, B., Dear, K. & Milton, A.H. The effect of arsenic mitigation interventions on disease burden in Bangladesh. *Environ Health Perspect* **112**, 1172-7 (2004).
267. Das, N.K. & Sengupta, S.R. Arsenicosis: diagnosis and treatment. *Indian J Dermatol Venereol Leprol* **74**, 571-81 (2008).
268. Pruss-Ustun, A., Vickers, C., Haefliger, P. & Bertollini, R. Knowns and unknowns on burden of disease due to chemicals: a systematic review. *Environ Health* **10**, 9 (2011).
269. Ratnaike, R.N. Acute and chronic arsenic toxicity. *Postgrad Med J* **79**, 391-6 (2003).
270. Straif, K. *et al.* A review of human carcinogens--part C: metals, arsenic, dusts, and fibres. *Lancet Oncol* **10**, 453-4 (2009).
271. Kitchin, K.T. & Conolly, R. Arsenic-induced carcinogenesis--oxidative stress as a possible mode of action and future research needs for more biologically based risk assessment. *Chem Res Toxicol* **23**, 327-35 (2010).
272. Kligerman, A.D. & Tennant, A.H. Insights into the carcinogenic mode of action of arsenic. *Toxicol Appl Pharmacol* **222**, 281-8 (2007).
273. Bailey, K.A. & Fry, R.C. Arsenic-Associated Changes to the Epigenome: What Are the Functional Consequences? *Curr Environ Health Rep* **1**, 22-34 (2014).
274. Drobna, Z., Styblo, M. & Thomas, D.J. An Overview of Arsenic Metabolism and Toxicity. *Curr Protoc Toxicol* **42**, 4.31.1-4.31.6 (2009).
275. Zhao, C.Q., Young, M.R., Diwan, B.A., Coogan, T.P. & Waalkes, M.P. Association of arsenic-induced malignant transformation with DNA hypomethylation and aberrant gene expression. *Proc Natl Acad Sci U S A* **94**, 10907-12 (1997).
276. Fu, H.Y. *et al.* Arsenic trioxide inhibits DNA methyltransferase and restores expression of methylation-silenced CDKN2B/CDKN2A genes in human hematologic malignant cells. *Oncol Rep* **24**, 335-43 (2010).
277. Wilson, A.S., Power, B.E. & Molloy, P.L. DNA hypomethylation and human diseases. *Biochim Biophys Acta* **1775**, 138-62 (2007).
278. Mass, M.J. & Wang, L. Arsenic alters cytosine methylation patterns of the promoter of the tumor suppressor gene p53 in human lung cells: a model for a mechanism of carcinogenesis. *Mutat Res* **386**, 263-77 (1997).

279. Davis, C.D., Uthus, E.O. & Finley, J.W. Dietary selenium and arsenic affect DNA methylation in vitro in Caco-2 cells and in vivo in rat liver and colon. *J Nutr* **130**, 2903-9 (2000).
280. Pilsner, J.R. *et al.* Influence of prenatal lead exposure on genomic methylation of cord blood DNA. *Environ Health Perspect* **117**, 1466-71 (2009).
281. Majumdar, S. *et al.* Arsenic exposure induces genomic hypermethylation. *Environ Toxicol* **25**, 315-8 (2010).
282. Chanda, S. *et al.* DNA hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy. *Toxicol Sci* **89**, 431-7 (2006).
283. Zhang, A.H., Bin, H.H., Pan, X.L. & Xi, X.G. Analysis of p16 gene mutation, deletion and methylation in patients with arseniasis produced by indoor unventilated-stove coal usage in Guizhou, China. *J Toxicol Environ Health A* **70**, 970-5 (2007).
284. Marsit, C.J. *et al.* Carcinogen exposure and gene promoter hypermethylation in bladder cancer. *Carcinogenesis* **27**, 112-6 (2006).
285. Chen, W.T., Hung, W.C., Kang, W.Y., Huang, Y.C. & Chai, C.Y. Urothelial carcinomas arising in arsenic-contaminated areas are associated with hypermethylation of the gene promoter of the death-associated protein kinase. *Histopathology* **51**, 785-92 (2007).
286. Zhong, C.X. & Mass, M.J. Both hypomethylation and hypermethylation of DNA associated with arsenite exposure in cultures of human cells identified by methylation-sensitive arbitrarily-primed PCR. *Toxicol Lett* **122**, 223-34 (2001).
287. Cui, X., Wakai, T., Shirai, Y., Hatakeyama, K. & Hirano, S. Chronic oral exposure to inorganic arsenate interferes with methylation status of p16INK4a and RASSF1A and induces lung cancer in A/J mice. *Toxicol Sci* **91**, 372-81 (2006).
288. Huang, Y.C., Hung, W.C., Chen, W.T., Yu, H.S. & Chai, C.Y. Effects of DNMT and MEK inhibitors on the expression of RECK, MMP-9, -2, uPA and VEGF in response to arsenite stimulation in human uroepithelial cells. *Toxicol Lett* **201**, 62-71 (2011).
289. Chen, H. *et al.* Chronic inorganic arsenic exposure induces hepatic global and individual gene hypomethylation: implications for arsenic hepatocarcinogenesis. *Carcinogenesis* **25**, 1779-86 (2004).
290. Benbrahim-Tallaa, L. *et al.* Molecular events associated with arsenic-induced malignant transformation of human prostatic epithelial cells: aberrant genomic DNA methylation and K-ras oncogene activation. *Toxicol Appl Pharmacol* **206**, 288-98 (2005).
291. Cui, X. *et al.* Arsenic trioxide inhibits DNA methyltransferase and restores methylation-silenced genes in human liver cancer cells. *Hum Pathol* **37**, 298-311 (2006).
292. Chervona, Y. & Costa, M. The control of histone methylation and gene expression by oxidative stress, hypoxia, and metals. *Free Radic Biol Med* **53**, 1041-7 (2012).
293. Huang, B.W., Ray, P.D., Iwasaki, K. & Tsuji, Y. Transcriptional regulation of the human ferritin gene by coordinated regulation of Nrf2 and protein arginine methyltransferases PRMT1 and PRMT4. *Faseb j* **27**, 3763-74 (2013).

294. Li, J., Gorospe, M., Barnes, J. & Liu, Y. Tumor promoter arsenite stimulates histone H3 phosphoacetylation of proto-oncogenes c-fos and c-jun chromatin in human diploid fibroblasts. *J Biol Chem* **278**, 13183-91 (2003).
295. Mishra, K.P. Lead exposure and its impact on immune system: a review. *Toxicol In Vitro* **23**, 969-72 (2009).
296. Navas-Acien, A., Guallar, E., Silbergeld, E.K. & Rothenberg, S.J. Lead exposure and cardiovascular disease--a systematic review. *Environ Health Perspect* **115**, 472-82 (2007).
297. Kim, R. *et al.* A longitudinal study of low-level lead exposure and impairment of renal function. The Normative Aging Study. *Jama* **275**, 1177-81 (1996).
298. Sallmen, M. Exposure to lead and male fertility. *Int J Occup Med Environ Health* **14**, 219-22 (2001).
299. Bellinger, D., Leviton, A., Allred, E. & Rabinowitz, M. Pre- and postnatal lead exposure and behavior problems in school-aged children. *Environ Res* **66**, 12-30 (1994).
300. Olympio, K.P., Goncalves, C., Gunther, W.M. & Bechara, E.J. Neurotoxicity and aggressiveness triggered by low-level lead in children: a review. *Rev Panam Salud Publica* **26**, 266-75 (2009).
301. Flora, S.J. Lead exposure: health effects, prevention and treatment. *J Environ Biol* **23**, 25-41 (2002).
302. Sanders, T., Liu, Y., Buchner, V. & Tchounwou, P.B. Neurotoxic effects and biomarkers of lead exposure: a review. *Rev Environ Health* **24**, 15-45 (2009).
303. Kovatsi, L. *et al.* p16 promoter methylation in Pb2+ -exposed individuals. *Clin Toxicol (Phila)* **48**, 124-8 (2010).
304. Wright, R.O. *et al.* Biomarkers of lead exposure and DNA methylation within retrotransposons. *Environ Health Perspect* **118**, 790-5 (2010).
305. Spiegel, S.J. Occupational health, mercury exposure, and environmental justice: learning from experiences in Tanzania. *Am J Public Health* **99 Suppl 3**, S550-8 (2009).
306. Park, J.D. & Zheng, W. Human exposure and health effects of inorganic and elemental mercury. *J Prev Med Public Health* **45**, 344-52 (2012).
307. Holmes, P., James, K.A. & Levy, L.S. Is low-level environmental mercury exposure of concern to human health? *Sci Total Environ* **408**, 171-82 (2009).
308. Hanna, C.W. *et al.* DNA methylation changes in whole blood is associated with exposure to the environmental contaminants, mercury, lead, cadmium and bisphenol A, in women undergoing ovarian stimulation for IVF. *Hum Reprod* **27**, 1401-10 (2012).
309. Goodrich, J.M., Basu, N., Franzblau, A. & Dolinoy, D.C. Mercury biomarkers and DNA methylation among Michigan dental professionals. *Environ Mol Mutagen* **54**, 195-203 (2013).
310. Onishchenko, N., Karpova, N., Sabri, F., Castren, E. & Ceccatelli, S. Long-lasting depression-like behavior and epigenetic changes of BDNF gene expression induced by perinatal exposure to methylmercury. *J Neurochem* **106**, 1378-87 (2008).

311. Gadhia, S.R., Calabro, A.R. & Barile, F.A. Trace metals alter DNA repair and histone modification pathways concurrently in mouse embryonic stem cells. *Toxicol Lett* **212**, 169-79 (2012).
312. Jarup, L. & Akesson, A. Current status of cadmium as an environmental health problem. *Toxicol Appl Pharmacol* **238**, 201-8 (2009).
313. Godt, J. *et al.* The toxicity of cadmium and resulting hazards for human health. *J Occup Med Toxicol* **1**, 22 (2006).
314. Kippler, M. *et al.* Maternal cadmium exposure during pregnancy and size at birth: a prospective cohort study. *Environ Health Perspect* **120**, 284-9 (2012).
315. Benton, M.A., Rager, J.E., Smeester, L. & Fry, R.C. Comparative genomic analyses identify common molecular pathways modulated upon exposure to low doses of arsenic and cadmium. *BMC Genomics* **12**, 173 (2011).
316. Waisberg, M., Joseph, P., Hale, B. & Beyersmann, D. Molecular and cellular mechanisms of cadmium carcinogenesis. *Toxicology* **192**, 95-117 (2003).
317. Sanders, A.P. *et al.* Cadmium exposure and the epigenome: Exposure-associated patterns of DNA methylation in leukocytes from mother-baby pairs. *Epigenetics* **9**, 212-21 (2014).
318. Severson, P.L., Tokar, E.J., Vrba, L., Waalkes, M.P. & Futscher, B.W. Agglomerates of aberrant DNA methylation are associated with toxicant-induced malignant transformation. *Epigenetics* **7**, 1238-48 (2012).
319. Huang, D., Zhang, Y., Qi, Y., Chen, C. & Ji, W. Global DNA hypomethylation, rather than reactive oxygen species (ROS), a potential facilitator of cadmium-stimulated K562 cell proliferation. *Toxicol Lett* **179**, 43-7 (2008).
320. Benbrahim-Tallaa, L., Waterland, R.A., Dill, A.L., Webber, M.M. & Waalkes, M.P. Tumor suppressor gene inactivation during cadmium-induced malignant transformation of human prostate cells correlates with overexpression of de novo DNA methyltransferase. *Environ Health Perspect* **115**, 1454-9 (2007).
321. Zhou, Z.H., Lei, Y.X. & Wang, C.X. Analysis of aberrant methylation in DNA repair genes during malignant transformation of human bronchial epithelial cells induced by cadmium. *Toxicol Sci* **125**, 412-7 (2012).
322. Hossain, M.B., Vahter, M., Concha, G. & Broberg, K. Low-level environmental cadmium exposure is associated with DNA hypomethylation in Argentinean women. *Environ Health Perspect* **120**, 879-84 (2012).
323. Ronco, A.M., Llaguno, E., Epunan, M.J. & Llanos, M.N. Effect of cadmium on cortisol production and 11beta-hydroxysteroid dehydrogenase 2 expression by cultured human choriocarcinoma cells (JEG-3). *Toxicol In Vitro* **24**, 1532-7 (2010).
324. Somji, S. *et al.* Differences in the epigenetic regulation of MT-3 gene expression between parental and Cd+2 or As+3 transformed human urothelial cells. *Cancer Cell Int* **11**, 2 (2011).
325. Brook, R.D. *et al.* Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation* **109**, 2655-71 (2004).
326. Baccarelli, A. *et al.* Exposure to particulate air pollution and risk of deep vein thrombosis. *Arch Intern Med* **168**, 920-7 (2008).

327. Tarantini, L. *et al.* Effects of particulate matter on genomic DNA methylation content and iNOS promoter methylation. *Environ Health Perspect* **117**, 217-22 (2009).
328. Bellavia, A. *et al.* DNA hypomethylation, ambient particulate matter, and increased blood pressure: findings from controlled human exposure experiments. *J Am Heart Assoc* **2**, e000212 (2013).
329. Baccarelli, A. *et al.* Rapid DNA methylation changes after exposure to traffic particles. *Am J Respir Crit Care Med* **179**, 572-8 (2009).
330. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400-13 (2002).
331. Castro, R. *et al.* Increased homocysteine and S-adenosylhomocysteine concentrations and DNA hypomethylation in vascular disease. *Clin Chem* **49**, 1292-6 (2003).
332. Peng, D., Jiaxing, W., Chunhui, H., Weiyi, P. & Xiaomin, W. Study on the cytogenetic changes induced by benzene and hydroquinone in human lymphocytes. *Hum Exp Toxicol* **31**, 322-35 (2012).
333. Smith, M.T. Advances in understanding benzene health effects and susceptibility. *Annu Rev Public Health* **31**, 133-48 2 p following 148 (2010).
334. Monks, T.J., Butterworth, M. & Lau, S.S. The fate of benzene-oxide. *Chem Biol Interact* **184**, 201-6 (2010).
335. Snyder, R. Leukemia and benzene. *Int J Environ Res Public Health* **9**, 2875-93 (2012).
336. Traver, R.D. *et al.* NAD(P)H:quinone oxidoreductase gene expression in human colon carcinoma cells: characterization of a mutation which modulates DT-diaphorase activity and mitomycin sensitivity. *Cancer Res* **52**, 797-802 (1992).
337. Zhang, L., Yang, W., Hubbard, A.E. & Smith, M.T. Nonrandom aneuploidy of chromosomes 1, 5, 6, 7, 8, 9, 11, 12, and 21 induced by the benzene metabolites hydroquinone and benzenetriol. *Environ Mol Mutagen* **45**, 388-96 (2005).
338. Coulter, J.B., O'Driscoll, C.M. & Bressler, J.P. Hydroquinone increases 5-hydroxymethylcytosine formation through ten eleven translocation 1 (TET1) 5-methylcytosine dioxygenase. *J Biol Chem* **288**, 28792-800 (2013).
339. Bollati, V. *et al.* Changes in DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer Res* **67**, 876-80 (2007).
340. Belancio, V.P., Roy-Engel, A.M. & Deininger, P.L. All y'all need to know 'bout retroelements in cancer. *Semin Cancer Biol* **20**, 200-10 (2010).
341. Schulz, W.A., Steinhoff, C. & Florl, A.R. Methylation of endogenous human retroelements in health and disease. *Curr Top Microbiol Immunol* **310**, 211-50 (2006).
342. Byun, H.M. *et al.* Evolutionary age of repetitive element subfamilies and sensitivity of DNA methylation to airborne pollutants. *Part Fibre Toxicol* **10**, 28 (2013).
343. Yang, J. *et al.* PTEN methylation involved in benzene-induced hematotoxicity. *Exp Mol Pathol* **96**, 300-6 (2014).

344. Zhong, Y. *et al.* Signal transducer and activator of transcription 3 (STAT3) gene polymorphisms are associated with treatment outcomes in acute myeloid leukemia. *Int J Lab Hematol* **34**, 383-9 (2012).
345. Li, P. *et al.* Stat3 activates the receptor tyrosine kinase like orphan receptor-1 gene in chronic lymphocytic leukemia cells. *PLoS One* **5**, e11859 (2010).
346. Yang, J., Bai, W., Niu, P., Tian, L. & Gao, A. Aberrant hypomethylated STAT3 was identified as a biomarker of chronic benzene poisoning through integrating DNA methylation and mRNA expression data. *Exp Mol Pathol* **96**, 346-53 (2014).
347. Ji, Z. *et al.* A comparison of the cytogenetic alterations and global DNA hypomethylation induced by the benzene metabolite, hydroquinone, with those induced by melphalan and etoposide. *Leukemia* **24**, 986-91 (2010).
348. Figueroa, M.E. *et al.* DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17**, 13-27 (2010).
349. Forrest, M.S. *et al.* Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. *Environ Health Perspect* **113**, 801-7 (2005).
350. Zhang, L., Eastmond, D.A. & Smith, M.T. The nature of chromosomal aberrations detected in humans exposed to benzene. *Crit Rev Toxicol* **32**, 1-42 (2002).
351. Tenen, D.G. Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer* **3**, 89-101 (2003).
352. Li, T.H., Kim, C., Rubin, C.M. & Schmid, C.W. K562 cells implicate increased chromatin accessibility in Alu transcriptional activation. *Nucleic Acids Res* **28**, 3031-9 (2000).
353. Saito, K. *et al.* Long interspersed nuclear element 1 hypomethylation is a marker of poor prognosis in stage IA non-small cell lung cancer. *Clin Cancer Res* **16**, 2418-26 (2010).
354. Kroutter, E.N., Belancio, V.P., Wagstaff, B.J. & Roy-Engel, A.M. The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. *PLoS Genet* **5**, e1000458 (2009).
355. Cheetham, S.W., Gruhl, F., Mattick, J.S. & Dinger, M.E. Long noncoding RNAs and the genetics of cancer. *Br J Cancer* **108**, 2419-25 (2013).
356. Canella, D., Praz, V., Reina, J.H., Cousin, P. & Hernandez, N. Defining the RNA polymerase III transcriptome: Genome-wide localization of the RNA polymerase III transcription machinery in human cells. *Genome Res* **20**, 710-21 (2010).
357. Oler, A.J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**, 620-8 (2010).
358. Moqtaderi, Z. *et al.* Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat Struct Mol Biol* **17**, 635-40 (2010).
359. Oler, A.J. *et al.* Alu expression in human cell lines and their retrotranspositional potential. *Mob DNA* **3**, 11 (2012).
360. Kojima, K.K. Alu monomer revisited: recent generation of Alu monomers. *Mol Biol Evol* **28**, 13-5 (2011).

361. Sassa, T. The role of human-specific gene duplications during brain development and evolution. *J Neurogenet* **27**, 86-96 (2013).
362. Martignetti, J.A. & Brosius, J. BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. *Proc Natl Acad Sci U S A* **90**, 11563-7 (1993).
363. Parrott, A.M. & Mathews, M.B. snaR genes: recent descendants of Alu involved in the evolution of chorionic gonadotropins. *Cold Spring Harb Symp Quant Biol* **74**, 363-73 (2009).
364. Giuliodori, S. *et al.* A composite upstream sequence motif potentiates tRNA gene transcription in yeast. *J Mol Biol* **333**, 1-20 (2003).
365. Su, M., Han, D., Boyd-Kirkup, J., Yu, X. & Han, J.D. Evolution of Alu elements toward enhancers. *Cell Rep* **7**, 376-85 (2014).
366. Pascali, C. & Teichmann, M. RNA polymerase III transcription - regulated by chromatin structure and regulator of nuclear chromatin organization. *Subcell Biochem* **61**, 261-87 (2013).
367. Dieci, G., Preti, M. & Montanini, B. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics* **94**, 83-8 (2009).
368. Hesselberth, J.R. Lives that introns lead after splicing. *Wiley Interdiscip Rev RNA* **4**, 677-91 (2013).
369. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
370. Treangen, T.J. & Salzberg, S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36-46 (2012).
371. Anders, S., Pyl, P.T. & Huber, W. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* (2014).
372. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
373. Conti, A. *et al.* Identification of RNA polymerase III-transcribed Alu loci by computational screening of RNA-Seq data. *Nucleic Acids Res* (2014).
374. Preti, M. *et al.* The telomere-binding protein Tbf1 demarcates snoRNA gene promoters in Saccharomyces cerevisiae. *Mol Cell* **38**, 614-20 (2010).
375. Dignam, J.D., Lebovitz, R.M. & Roeder, R.G. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res* **11**, 1475-89 (1983).
376. Tang, R.B. *et al.* Increased level of polymerase III transcribed Alu RNA in hepatocellular carcinoma tissue. *Mol Carcinog* **42**, 93-6 (2005).
377. Kaer, K. & Speek, M. Intronic retroelements: Not just "speed bumps" for RNA polymerase II. *Mob Genet Elements* **2**, 154-157 (2012).
378. Chen, X. & Macara, I.G. Par-3 controls tight junction assembly through the Rac exchange factor Tiam1. *Nat Cell Biol* **7**, 262-9 (2005).
379. Brummelkamp, T.R., Nijman, S.M., Dirac, A.M. & Bernards, R. Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-kappaB. *Nature* **424**, 797-801 (2003).

380. McCabe, N. *et al.* Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Res* **66**, 8109-15 (2006).
381. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).
382. Fitzpatrick, T. & Huang, S. 3'-UTR-located inverted Alu repeats facilitate mRNA translational repression and stress granule accumulation. *Nucleus* **3**, 359-69 (2012).
383. Mandal, A.K., Pandey, R., Jha, V. & Mukerji, M. Transcriptome-wide expansion of non-coding regulatory switches: evidence from co-occurrence of Alu exonization, antisense and editing. *Nucleic Acids Res* **41**, 2121-37 (2013).
384. Chen, L.L. & Carmichael, G.G. Nuclear Editing of mRNA 3'-UTRs. *Curr Top Microbiol Immunol* **353**, 111-21 (2012).
385. Kaneko, H. *et al.* DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. *Nature* **471**, 325-30 (2011).
386. Gong, C., Tang, Y. & Maquat, L.E. mRNA-mRNA duplexes that autoelicit Staufen1-mediated mRNA decay. *Nat Struct Mol Biol* **20**, 1214-20 (2013).
387. Chen, L.L., DeCerbo, J.N. & Carmichael, G.G. Alu element-mediated gene silencing. *Embo j* **27**, 1694-705 (2008).
388. Capshew, C.R., Dusenbury, K.L. & Hundley, H.A. Inverted Alu dsRNA structures do not affect localization but can alter translation efficiency of human mRNAs independent of RNA editing. *Nucleic Acids Res* **40**, 8637-45 (2012).
389. Baryakin, D.N. *et al.* Alu- and 7SL RNA Analogues Suppress MCF-7 Cell Viability through Modulating the Transcription of Endoplasmic Reticulum Stress Response Genes. *Acta Naturae* **5**, 83-93 (2013).
390. Smalheiser, N.R. & Torvik, V.I. Alu elements within human mRNAs are probable microRNA targets. *Trends Genet* **22**, 532-6 (2006).
391. Lehnert, S. *et al.* Evidence for co-evolution between human microRNAs and Alu-repeats. *PLoS One* **4**, e4456 (2009).
392. Daskalova, E., Baev, V., Rusinov, V. & Minkov, I. 3'UTR-located ALU elements: donors of potential miRNA target sites and mediators of network miRNA-based regulatory interactions. *Evol Bioinform Online* **2**, 103-20 (2006).
393. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G.D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
394. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat Methods* **9**, 1069-76 (2012).
395. Hua, F. *et al.* Effect of downregulated histone deacetylase 2 expression on cell proliferation and cell cycle in cervical cancer. *J buon* **19**, 497-501 (2014).
396. Brooks, C.L. & Gu, W. p53 ubiquitination: Mdm2 and beyond. *Mol Cell* **21**, 307-15 (2006).
397. Kleiblova, P. *et al.* Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. *J Cell Biol* **201**, 511-21 (2013).

398. Pandey, R., Mandal, A.K., Jha, V. & Mukerji, M. Heat shock factor binding in Alu repeats expands its involvement in stress through an antisense mechanism. *Genome Biol* **12**, R117 (2011).
399. Di Leva, G. & Croce, C.M. miRNA profiling of cancer. *Curr Opin Genet Dev* **23**, 3-11 (2013).
400. Li, J. *et al.* lncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* **63**, 1700-10 (2014).
401. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4.1-9 (2006).
402. Tippmann, S. Programming tools: Adventures with R. *Nature* **517**, 109-10 (2015).
403. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J.P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251-3 (2007).
404. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
405. Aporntewan, C. *et al.* Hypomethylation of intragenic LINE-1 represses transcription in cancer cells through AGO2. *PLoS One* **6**, e17934 (2011).
406. Filkowski, J.N. *et al.* Hypomethylation and genome instability in the germline of exposed parents and their progeny is associated with altered miRNA expression. *Carcinogenesis* **31**, 1110-5 (2010).
407. Szpakowski, S. *et al.* Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene* **448**, 151-67 (2009).
408. Moran, J.L., Siegel, D., Sun, X.M. & Ross, D. Induction of apoptosis by benzene metabolites in HL60 and CD34+ human bone marrow progenitor cells. *Mol Pharmacol* **50**, 610-5 (1996).
409. Zhang, L. *et al.* Systems biology of human benzene exposure. *Chem Biol Interact* **184**, 86-93 (2010).
410. Kerzic, P.J. *et al.* Analysis of hydroquinone and catechol in peripheral blood of benzene-exposed workers. *Chem Biol Interact* **184**, 182-8 (2010).
411. Lee, J.S., Yang, E.J. & Kim, I.S. Hydroquinone-induced apoptosis of human lymphocytes through caspase 9/3 pathway. *Mol Biol Rep* **39**, 6737-43 (2012).
412. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-8 (2001).
413. Hu, J. *et al.* Effects of benzene and its metabolites on global DNA methylation in human normal hepatic L02 cells. *Environ Toxicol* **29**, 108-16 (2014).
414. Yang, A.S. *et al.* A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Res* **32**, e38 (2004).
415. Jones, P.A. & Taylor, S.M. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20**, 85-93 (1980).
416. Stresemann, C. & Lyko, F. Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine. *Int J Cancer* **123**, 8-13 (2008).

417. Zhang, R. & Li, H.M. [Research Progress of Epigenetic Drug Decitabine in AML]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi* **22**, 1467-71 (2014).
418. Woodcock, D.M., Lawler, C.B., Linsenmeyer, M.E., Doherty, J.P. & Warren, W.D. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J Biol Chem* **272**, 7810-6 (1997).
419. Lund, K. *et al.* DNMT inhibitors reverse a specific signature of aberrant promoter DNA methylation and associated gene silencing in AML. *Genome Biol* **15**, 406 (2014).
420. Liang, G., Gonzales, F.A., Jones, P.A., Orntoft, T.F. & Thykjaer, T. Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. *Cancer Res* **62**, 961-6 (2002).
421. Broxmeyer, H.E. *et al.* Human umbilical cord blood as a potential source of transplantable hematopoietic stem/progenitor cells. *Proc Natl Acad Sci U S A* **86**, 3828-32 (1989).
422. Zhu, J. *et al.* Comparison of toxicity of benzene metabolite hydroquinone in hematopoietic stem cells derived from murine embryonic yolk sac and adult bone marrow. *PLoS One* **8**, e71153 (2013).
423. Rodic, N. & Burns, K.H. Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? *PLoS Genet* **9**, e1003402 (2013).
424. Hur, K. *et al.* Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut* **63**, 635-46 (2014).
425. Pyatt, D.W., Stillman, W.S. & Irons, R.D. Hydroquinone, a reactive metabolite of benzene, inhibits NF-kappa B in primary human CD4+ T lymphocytes. *Toxicol Appl Pharmacol* **149**, 178-84 (1998).
426. Kolachana, P., Subrahmanyam, V.V., Meyer, K.B., Zhang, L. & Smith, M.T. Benzene and its phenolic metabolites produce oxidative DNA damage in HL60 cells in vitro and in the bone marrow in vivo. *Cancer Res* **53**, 1023-6 (1993).
427. Aoki, Y. *et al.* Genomic vulnerability to LINE-1 hypomethylation is a potential determinant of the clinicogenetic features of multiple myeloma. *Genome Med* **4**, 101 (2012).
428. Jones, P.A. & Baylin, S.B. The fundamental role of epigenetic events in cancer. *Nat Rev Genet* **3**, 415-28 (2002).
429. Clark, S.J., Harrison, J., Paul, C.L. & Frommer, M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res* **22**, 2990-7 (1994).
430. Ronaghi, M. Pyrosequencing sheds light on DNA sequencing. *Genome Res* **11**, 3-11 (2001).
431. Waidyanatha, S. *et al.* Rapid determination of six urinary benzene metabolites in occupationally exposed and unexposed subjects. *Anal Biochem* **327**, 184-99 (2004).
432. Mills, J.D. & Janitz, M. Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. *Neurobiol Aging* **33**, 1012.e11-24 (2012).

433. Tsirigos, A. & Rigoutsos, I. Human and mouse introns are linked to the same processes and functions through each genome's most frequent non-conserved motifs. *Nucleic Acids Res* **36**, 3484-93 (2008).
434. Maccari, G., Gemignani, F. & Landi, S. COMPASSS (COMplex PAttern of Sequence Search Software), a simple and effective tool for mining complex motifs in whole genomes. *Bioinformatics* **26**, 1777-8 (2010).
435. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
436. Pagano, A. *et al.* CALbeta, a novel lipocalin associated with chondrogenesis and inflammation. *Eur J Cell Biol* **81**, 264-72 (2002).