Università degli Studi di Padova
Corso di Laurea in Biologia Molecolare

Elaborato di Laurea

Analisi informatica di dati di RNA-seq per l'identificazione dei siti di poliadenilazione in *Saccharomyces cerevisiae*

Tutor: Dott. Stefano Campanaro
Dipartimento di Biologia

Laureanda: Benedetta Scarpari

Anno Accademico 2011/2012

INDICE

ABSTRACT

Negli ultimi anni l'analisi del trascrittoma è stata rivoluzionata dalle nuove tecniche di sequenziamento del RNA (RNA-seq) su piattaforme di sequenziamento *Next-Generation*. Queste tecnologie innovative permettono, insieme alla determinazione del profilo di espressione, una precisa caratterizzazione del 3' e 5' terminale dei trascritti. Il nuovo problema che ora si pone è la trattazione della grande quantità di dati prodotta dai sequenziatori di nuova generazione. In questo contesto diventa molto utile saper utilizzare gli strumenti bioinformatici e saper sviluppare nuovi programmi di analisi. Per l'identificazione dei siti di poliadenilazione di *Saccharomyces cerevisiae* si è scelto quindi di analizzare dati di sequenziamento Illumina con alcuni *script* redatti nel linguaggio di programmazione Perl. Lo scopo è determinare le posizioni delle terminazioni dei trascritti, per concentrarsi poi su quelli che presentano possibili terminazioni alternative. Questi ultimi sembrano essere correlati funzionalmente e presentano un coinvolgimento nel processo di traduzione. Sono state inoltre ricercate sequenze segnale per la poliadenilazione nelle regioni genomiche adiacenti alle terminazioni per comprendere il loro effetto nella determinazione della posizione finale del trascritto. L'analisi è stata effettuata su dati ottenuti con tre diversi metodi di costruzione di librerie per RNA-seq, tra cui quello che utilizza i primer oligo(dT) si è dimostrato il più sensibile.

STATO DELL'ARTE

La mia analisi si concentra sulla struttura del 3' terminale dei trascritti, in particolare sulla determinazione dei siti di terminazione della trascrizione. La posizione terminale di un trascritto può variare a seconda della localizzazione subcellulare o delle condizioni fisiologiche della cellula. Può essere quindi molto interessante studiare i siti di poliadenilazione di un trascritto in diverse condizioni sperimentali. Per questa analisi è stato scelto l'organismo modello *S. cerevisiae* ceppo S288c, che possiede un genoma interamente sequenziato e accuratamente annotato. Si tratta di un organismo eucariote che ha però il vantaggio di possedere un genoma compatto di circa 6000 geni. I dati utilizzati sono dati di RNA-seq ottenuti con piattaforme di sequenziamento di nuova generazione (NGS: *Next Generation Sequencing*). L'RNA-seq è oggi all'avanguardia nello studio del trascrittoma, poiché ha un'alta resa in termini di sequenze prodotte e permette di studiare anche geni con bassi livelli di espressione. Inoltre è un metodo molto sensibile per lo studio della struttura del trascritto, in quanto le *read* generate dalle piattaforme NGS, corte e molto numerose, permettono una precisa identificazione del 3' e 5' terminale.
Durante la mia analisi sono stati messi a confronto tre diversi metodi per la costruzione delle librerie di RNA-seq: "Illumina RNA ligation", cioè una libreria *shotgun* ottenuta per frammentazione del mRNA e altre due librerie che hanno utilizzato primer oligo(dT) e primer random esameri (il cui utilizzo genera una libreria *shotgun*) per la retrotrascrizione del mRNA in cDNA.

Il lievito: *Saccharomyces cerevisiae*

S. cerevisiae è un fungo unicellulare utilizzato sia nel campo della ricerca sia in processi industriali come la panificazione, la vinificazione, la birrificazione e la produzione di bioetanolo (fig. 1).

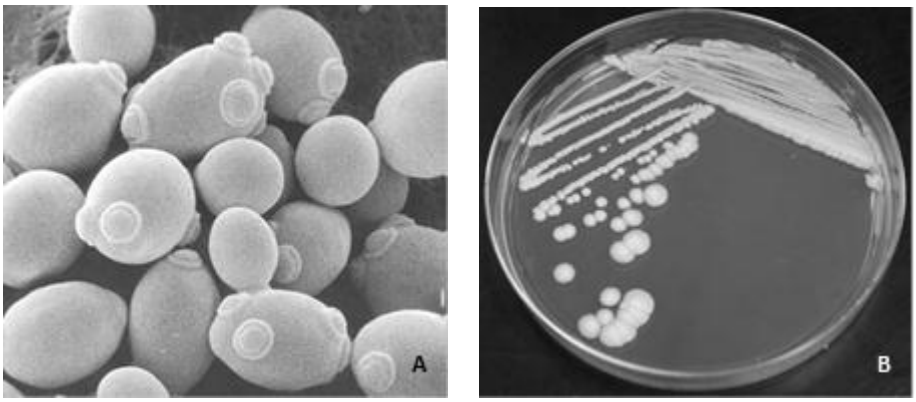

Figura 1: A. Immagine di *S. cerevisiae* ottenuta con microscopia elettronica. B. Coltura di lievito in mezzo solido; sono visibili alcune colonie isolate nella parte inferiore della piastra.

S. cerevisiae è un ottimo organismo modello in biologia: presenta dimensioni ridotte (5-30 μm), è facile da coltivare in terreno liquido o solido, ha un ciclo vitale breve della durata di 90 minuti e si divide ogni 2 ore a 30°C per gemmazione (per questo è chiamato "budding yeast", lievito gemmante). Presenta uno stadio aploide e uno diploide e in condizioni favorevoli dà vita a popolazioni clonali dividendosi per mitosi. In condizioni di stress può sporulare, producendo spore aploidi con due diversi *mating type*: a e α, che possono combinarsi tra loro permettendo lo scambio di materiale genetico.

Uno dei più grandi vantaggi di *S. cerevisiae* è che il suo genoma, nonostante sia contenuto in una cellula eucariote complessa, è piuttosto compatto. Esso (lungo 12 Mbp e contenente circa 6000 geni) è strutturato in 16 cromosomi ed è più semplice rispetto a quello degli eucarioti superiori, infatti solo 263 geni su 6000 (4% circa) presentano introni. È possibile inoltre sfruttare la capacità di lievito di fare ricombinazione omologa per inserire sequenze di DNA in specifici loci e per generare ceppi *knockout*.
Il ceppo *S. cerevisiae* S288c è stato il primo organismo eucariote sequenziato tramite una collaborazione internazionale (Goffeau *et al.* 1996) e oggi il suo genoma è uno dei meglio caratterizzati. Esso è quindi il più utilizzato in laboratorio, e si lavora spesso su ceppi da esso derivati, come ad esempio BY4741.

La poliadenilazione negli eucarioti: il lievito

La terminazione al 3' di un mRNA eucariote si genera mediante taglio endonucleolitico della sequenza e successiva poliadenilazione, che consiste nell'aggiunta al 3' terminale di una coda di poli(A) da parte dell'enzima poli(A) polimerasi. In questo processo sono coinvolte numerose proteine che collaborano per formare un complesso. Alcune di esse sono omologhe ad alcune subunità degli enzimi CstF (*Cleavage Stimulatory Factor*) e CPSF (*Cleavage and Polyadenylation Specific Factor*) che si legano alle principali sequenze segnale dei mammiferi, le quali presentano una certa similarità rispetto ad alcuni dei siti trovati in lievito.
In *S. cerevisiae* i segnali di poliadenilazione sono quattro. Il più a monte, cioè tra -80 e -25 basi dal sito di inizio, è l'*Efficiency Element* (EE), la sequenza chiave di tutto il sito di poliadenilazione. La sua funzione consiste nell'aumentare l'efficienza dei siti più a valle. Tra -35 e -15 basi invece si trova il *Positioning Element* (PE), necessario per un corretto posizionamento del sito di taglio; tra -20 e +5 basi è presente il *Pre-cleavage Element* (PrCl) e tra +5 e +30 basi il *Downstream Element* (DE). Se l'EE e il PE sono "ottimali", cioè possiedono le sequenze maggiormente rappresentate nelle regioni genomiche più a monte delle terminazioni dei trascritti, non sono richiesti segnali "forti" nel PrCl o nel DE (Graber *et al.* 1999).

Efficiency (I)		**Positioning (II)**		**Pre-cleavage (III)**		**Downstream (IV)**	
Sequenza	Score	Sequenza	Score	Sequenza	Score	Sequenza	Score
TATATA	1.55	**AAAATA**	0.97	**TTTTAT**	0.72	**TTTTCT**	0.46
ATATAT	1.15	**AATAAA**	0.92	**TTTTTT**	0.72	**CTTTTT**	0.44
TATGTA	0.63	**ATAATA**	0.87	**TATTCT**	0.67	**TTTTTC**	0.44
TGTATA	0.62	TAATAA	0.77	**TTTCTT**	0.6	**TTTCAT**	0.37
TACATA	0.54	AATATA	0.77	**TTCTTT**	0.6	TATTCT	0.3
GTATAT	0.47	AAATAA	0.67	ATTTTT	0.55	TTCATT	0.3
CATATA	0.46	AAAAAA	0.62	TTTTTA	0.46	TTTATT	0.26
ACATAT	0.38	AAGAAA	0.59	TATTAT	0.46	TATTTC	0.25
ATGTAT	0.37	AAAAAT	0.57	TTCTTC	0.44	TCTTTT	0.24
ATATAA	0.37	ATAAAA	0.51	TTTTTC	0.42	TCATTT	0.24

Tabella 1: I quattro siti che compongono il segnale di poliadenilazione. Le sequenze indicate in grassetto rappresentano i segnali "forti" di poliadenilazione ("parole ottimali").

I siti sopra elencati, a differenza dei segnali di poliadenilazione dei mammiferi, presentano una forte degenerazione (tabella 1). Le sequenze in grassetto rappresentano i segnali "forti" di poliadenilazione, che indicherò come "parole

ottimali". Le sequenze dei quattro segnali di poliadenilazione elencate in tabella 1 sono state utilizzate da uno degli *script* da me sviluppati per la ricerca delle sequenze segnale presenti nelle regioni trascritte del genoma.
La poliadenilazione ha un'influenza fondamentale nel metabolismo del mRNA: conferisce stabilità al trascritto impedendone la degradazione da parte di esonucleasi agenti in direzione 3'-5', assicura l'efficienza di traduzione e ha un ruolo nel trasporto del mRNA processato dal nucleo al citoplasma. L'emivita del trascritto ed il suo decadimento sono regolati dal tasso di deadenilazione, per questo la lunghezza iniziale della coda è controllata (in lievito arriva a circa 80 basi, mentre nei mammiferi raggiunge una lunghezza di 200-250 basi).

RNA-seq e *Next Generation Sequencing*: nuove tecnologie per l'analisi del trascrittoma

Il trascrittoma è costituito dalla totalità dei trascritti presenti in una cellula, ciascuno con la propria abbondanza, e varia a seconda dello stadio cellulare e delle condizioni fisiologiche. Lo studio del trascrittoma è fondamentale per interpretare gli elementi funzionali del genoma catalogando i diversi tipi di trascritti (mRNA, RNA non codificanti, small RNA), per determinare il loro livello di espressione e la loro struttura (terminazioni al 5' e al 3', struttura esoni-introni, splicing alternativi).
L'introduzione delle tecnologie di sequenziamento *Next-Generation* (NGS) ha rivoluzionato la caratterizzazione e la quantificazione dei trascrittomi superando le limitazioni dei metodi precedenti (per esempio i *microarray*), come la necessità di una conoscenza pregressa della sequenza genomica, il rumore di fondo (dovuto alla cross-ibridazione) e il *range* dinamico limitato.
L'RNA-seq utilizza le nuove tecnologie di sequenziamento. A partire dal mRNA si costruisce una libreria di cDNA e si procede alla ligazione di particolari adattatori ad entrambe le estremità di ciascuna sequenza. Ogni molecola viene poi sequenziata dopo un'eventuale amplificazione tramite PCR con una delle piattaforme NGS disponibili: "SOLiD Sequencer" (*Life Technologies Applied Biosystems*), "GS FLX Pyrosequencer" (*Roche 454 Life Sciences*) e "Illumina Genome Analyzer" (*Illumina/Solexa*). Il sequenziamento produce una serie di corte sequenze, chiamate *read*, che possono essere allineate sul genoma o sul trascrittoma di riferimento. Queste sono utilizzate per costruire un profilo di espressione per ogni gene, con una risoluzione che può arrivare fino alla singola base. Se le *read* sono corte e molto numerose (fino a diversi milioni per corsa) è possibile identificare trascritti con un basso livello di espressione, dato che esso è calcolato contando il numero di sequenze che mappano in quella regione, e si può inoltre determinare in modo accurato la loro struttura. L'RNA-seq è altamente riproducibile e richiede meno RNA per la sintesi della libreria, poiché manca lo *step* di clonaggio (Wang *et al.* 2009).
La tecnologia di sequenziamento di nuova generazione utilizzata per ottenere le *read* nei tre metodi analizzati è quella SBS: *Sequencing by synthesis* (Illumina/Solexa). Innanzitutto si ligano alle sequenze di cDNA della libreria gli adattatori specifici Illumina e le si trasferisce su un supporto solido a cui esse si legano tramite oligonucleotidi complementari agli adattatori. Qui vengono amplificate con un metodo particolare detto "bridge amplification", che restituisce gruppi di molecole di DNA identiche tra loro, ognuno derivato dall'amplificazione di una singola molecola. Il sequenziamento si basa sul metodo della terminazione ciclica reversibile, con un approccio "by-synthesis", che comprende tre passaggi: l'incorporazione del nucleotide, il rilevamento dell'immagine a fluorescenza e il taglio (Mardis *et al.* 2008). Nella prima fase del ciclo una DNA polimerasi legata allo stampo allunga uno

specifico primer aggiungendo un nucleotide legato covalentemente ad un fluoroforo. Questo presenta un blocco sul 3'-OH del ribosio che non gli permette la polimerizzazione con altri nucleotidi. Ogni base azotata è legata ad un fluoroforo di un colore specifico. L'incorporazione del nucleotide marcato avviene solo se sulla sequenza stampo è presente la base complementare. Segue lo *step* di rilevamento dell'immagine che riconosce la specifica lunghezza d'onda di emissione del fluoroforo. Si procede quindi con il taglio, che rimuove sia il fluoroforo sia il gruppo inibitore presente al 3'-OH e si ricomincia il ciclo. Ad ogni base è inoltre associato un punteggio che ne denota la qualità tramite un processo chiamato "base-calling", che consiste nella conversione dei dati del rilevamento dell'immagine in sequenze e punteggi di qualità. I dati di bassa qualità sono infine rimossi da ogni corsa di sequenziamento.

APPROCCIO SPERIMENTALE

Per l'analisi dei siti di poliadenilazione sono stati analizzati e confrontati dati di sequenziamento in piattaforma Illumina ottenuti con tre diverse procedure di costruzione delle librerie. Esse sono state costruite retrotrascrivendo l'mRNA poliadenilato di *S. cerevisiae* ceppo BY4741 in fase semi-logaritmica. La prima è "Illumina RNA ligation" (Levin *et al.* 2010), una libreria di tipo *shotgun*, cioè una libreria di sequenze ottenute per frammentazione casuale del mRNA. Le *read* generate dal sequenziamento con "Illumina Genome Analyzer II" sono lunghe 76 basi.
Le altre due librerie sono state costruite da Nagalakshmi e colleghi (2008) in questo modo: l'mRNA poliadenilato è stato isolato da cellule di lievito e retrotrascritto in cDNA a partire da primer randomici di 6 basi (random esameri) o da primer oligo(dT). Questi primer sono complementari alla coda di poli(A), perciò, dato che la sintesi della prima elica di cDNA inizia a valle della terminazione del trascritto, il 3' terminale dovrebbe essere definito in modo preciso. Ci si aspetta quindi che il metodo che utilizza primer oligo(dT) generi, dopo il sequenziamento, un numero maggiore di *read* poliadenilate.
Il cDNA a doppia elica viene poi frammentato e sequenziato. Le *read* ottenute dal sequenziamento con "Illumina 1G high throughput sequencing" sono lunghe 36 basi. I dati del sequenziamento Illumina dei tre esperimenti sono stati scaricati dai database GEO e SRA[1].
Il database SRA (*Sequence Read Archive*) del NCBI (*National Center for Biotechnology Information*) è stato sviluppato proprio per depositare le numerose *read* derivanti dal sequenziamento con le nuove piattaforme NGS. GEO (*Gene Expression Omnibus*) è un database pubblico del NCBI che archivia i dati ottenuti dalle moderne piattaforme NGS e da esperimenti di *microarray*. Indicando l'*accession number* si accede al *record* relativo alla pubblicazione, che presenta le principali informazioni sull'organismo utilizzato, il metodo di estrazione del materiale genetico, la costruzione della libreria e la piattaforma di sequenziamento. I file Solexa scaricati da questi database sono compressi in formato "SRA"; devono essere quindi decompressi con un *toolkit* ("Illumina-dump") che li converte dal formato "SRA" in "FASTQ", formato adatto

[1] Illumina RNA ligation: GEO accession: GSE21739, Sample GSM542248, SRA: SRX022776, Oligo(dT) e Random esameri (RH): GEO accession: GSE11209, Sample GSM282598, SRA: SRR002051 e SRR002059, rispettivamente.

per un file di input di PASS, il programma di allineamento che ho utilizzato durante la mia analisi.

Per trattare i dati di sequenziamento, ora in formato "FASTQ", e per analizzare l'output di PASS, ho elaborato alcuni *script* nel linguaggio di programmazione Perl.

Il linguaggio di programmazione Perl

Perl è un linguaggio di programmazione che si è progressivamente diffuso come linguaggio di *scripting*, cioè come linguaggio interpretato. Se dal punto di vista della velocità può rappresentare un problema, poiché l'interprete deve tradurre il linguaggio di programmazione in linguaggio macchina, questo è stato invece uno dei suoi punti di forza in quanto si possono facilmente sviluppare i programmi sul proprio computer e correggere errori di compilazione modificando direttamente lo *script*. Inoltre si tratta di un linguaggio flessibile (lo stesso problema può essere risolto in svariati modi) che presenta mezzi potenti come il "pattern matching", che riconosce un *pattern* specifico all'interno di una stringa e il "pattern substitution", che può sostituire uno specifico *pattern* con un altro. L'interprete Perl è disponibile con licenza "open source" ed è comodamente scaricabile dal web. Per facilitare la scrittura degli *script* è stato utilizzato un programma di supporto, Komodo Edit, che include un *debugger*.

PASS: un programma per l'allineamento di sequenze

Per allineare le *read* derivate dai dati di sequenziamento Illumina sul genoma di riferimento (in questo caso il genoma di *S. cerevisiae* ceppo S288c in formato "FASTA") è stato utilizzato il programma PASS (Campagna *et al.* 2009). Questo programma è stato sviluppato specificamente per l'allineamento di milioni di *read* molto corte sul genoma di riferimento. L'algoritmo procede attraverso la ricerca nella sequenza *query* di "parole" lunghe 11-12 basi che hanno una posizione specifica nel genoma. Quando si identifica una corrispondenza ("match") si procede ad un allineamento dinamico della regione circostante. L'esecuzione è veloce e permette l'inserimento di *gap* e *mismatch*. PASS può essere utilizzato per l'analisi della struttura esoni-introni dei geni, poiché è in grado di allineare le *read* sulle giunzioni di splicing. Il file di input è generalmente in formato "FASTQ", un formato "FASTA" che contiene anche la qualità relativa ad ogni sequenza. Essa consiste in un punteggio in codice ASCII assegnato a ciascuna base a seconda della bontà del sequenziamento. Il file di output è in formato "SAM" (*Sequence Alignement Map format*), un formato che presenta una riga per ogni sequenza allineata, elencando in 11 campi una serie di informazioni (come il nome della *read*, lo *strand*, il cromosoma, la posizione, la sequenza, la qualità della sequenza, la qualità dell'allineamento) separate da un carattere di tabulazione. Per poter sfruttare i dati contenuti in questo file si procede alla "parserizzazione", cioè all'estrazione delle informazioni di interesse, ottenuta separando ("splittando") ogni riga del file in corrispondenza del carattere di tabulazione.

Artemis

Durante la mia analisi si è rivelato molto utile visualizzare la posizione delle terminazioni predette dal mio *script* direttamente sul genoma di riferimento. Ciò è stato possibile grazie ad Artemis, un browser di visualizzazione e di annotazione della sequenza di DNA implementato in Java che permette di visualizzare i risultati delle analisi nel contesto della sequenza e dei suoi sei *frame* di lettura. Artemis è usato

soprattutto per analizzare i genomi compatti di batteri, archaea e eucarioti unicellulari (Rutherford *et al.* 2000). Si utilizza un file in formato "FASTA" o "GenBank" come riferimento e i dati da analizzare sono forniti come input nei formati "EMBL", "GenBank" e "GFF". Con questo strumento è possibile ad esempio visualizzare l'allineamento delle *read* sul genoma di riferimento, evidenziare certe posizioni particolari e affiancare alla sequenza un grafico costruito dall'utente.

GoMiner e REViGO

Al termine della mia analisi ho cercato di comprendere meglio la funzione dei geni codificanti trascritti con possibile terminazione alternativa. Ho scelto dunque di utilizzare GoMiner (Zeeberg *et al.* 2003), un software che permette di classificare dal punto di vista funzionale i geni di interesse secondo il vocabolario definito dalla Gene Ontology. L'insieme di termini che costituisce il vocabolario viene utilizzato dal Gene Ontology Consortium per organizzare i geni in categorie gerarchiche secondo il processo biologico, la funzione molecolare e la localizzazione subcellulare. GoMiner utilizza come input due liste di geni: la lista totale e una sottolista che contiene i geni con caratteristiche interessanti che si vogliono analizzare e classifica automaticamente i geni in input secondo le categorie di GO. Le categorie più rappresentate nella lista di geni in analisi sono identificate grazie ai test statistici eseguiti dal programma, che assegna loro un p-value.
REViGO (*Reduce + Visualize Gene Ontology*) invece è un web server che riassume lunghe liste di termini di GO utilizzando un semplice algoritmo di *clustering* simile al metodo di "hierarchical clustering", basato sulla similarità di campo semantico. Allo scopo di formare dei gruppi di termini di GO con alta similarità, REViGO elimina i termini meno significativi (secondo il p-value), quelli molto generali e quelli "figli", cioè sottocategorie di altri. I termini che rimangono in lista non superano una certa soglia di affinità, settabile dall'utente (Supek *et al.* 2011).
REViGO utilizza come input la lista dei termini di GO separati dal p-value da un carattere di tabulazione e dopo la procedura di raggruppamento ("clusterizzazione") che elimina i ridondanti, mette a disposizione quattro metodi di visualizzazione che permettono di interpretare velocemente i risultati. Lo "scatterplot" indica le relazioni semantiche tra i vari gruppi ("cluster") di geni disponendoli in uno spazio bidimensionale, la visualizzazione a grafico mostra per ogni categoria di GO un nodo collegato alle categorie ad essa correlate, le mappe ad albero presentano i livelli gerarchici tra i *cluster* e le "tag clouds" identificano le parole chiave più rappresentate tra le descrizioni delle varie categorie funzionali.

RISULTATI E DISCUSSIONE

Allo scopo di analizzare la struttura dei trascritti al 3' terminale ho elaborato alcuni *script* nel linguaggio di programmazione Perl. Di seguito presento uno schema riassuntivo che spiega in sintesi la funzione di ogni *script* ed elenca i vari passaggi, che saranno poi trattati punto per punto (fig. 2).

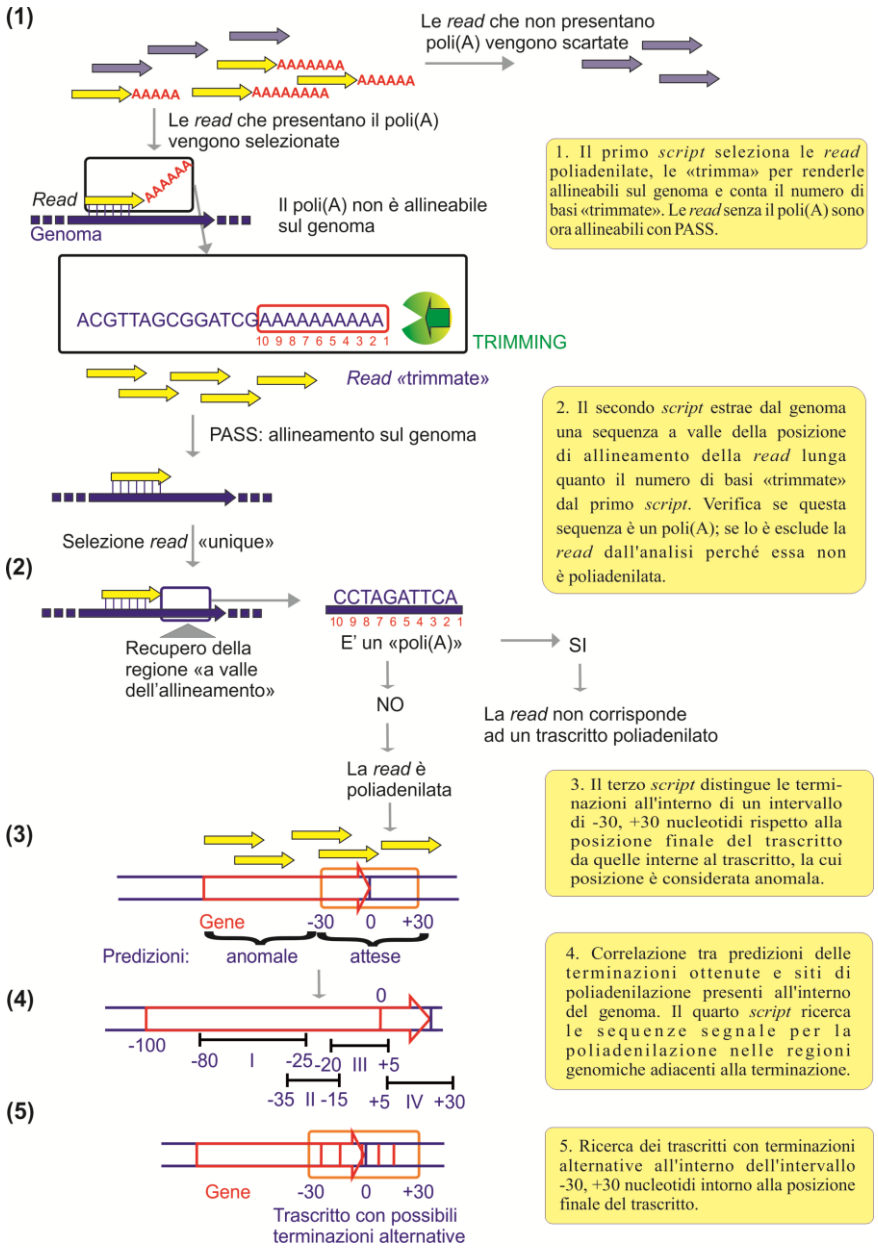

Figura 2: Schema riassuntivo delle principali funzioni degli *script* in Perl.

1. Identificazione delle *read* poliadenilate, *trimming* e allineamento sul genoma

I dati di RNA-seq scaricati dai database GEO e SRA consistono in un gran numero di *read*, cioè corte sequenze derivate dal sequenziamento del cDNA. Le *read* utili per lo studio del 3' terminale dei trascritti sono quelle che presentano una coda di poli(A). Per questo è necessario eliminare dai dati quelle non poliadenilate. Ho elaborato quindi uno *script* che identifica e seleziona le *read* che possiedono un poli(A)

lungo almeno 5 basi al 3' terminale (ritengo infatti che la coda di poli(A) per essere tale debba presentare questa lunghezza minima). Come input sono stati utilizzati i file in formato "FASTQ" ottenuti dalla decompressione dei file "SRA" scaricati per la libreria "Illumina RNA ligation" (Illumina *shotgun*) e per le librerie costruite con primer oligo(dT) e random esameri. Eseguendo lo *script* per tutti e tre gli esperimenti, si ottengono le seguenti percentuali di *read* che posseggono il poli(A) rispetto al totale (tabella 2):

Esperimento:	***Read* totali:**	***Read* con poli(A) (% rispetto al totale):**	***Read* "trimmate" (% sulle *read* con poli(A)):**
Illumina *shotgun*	23'400'111	213'225 (0,91%)	121'831 (57,13%)
Oligo(dT)	3'854'691	70'330 (1,82%)	21'880 (31,11%)
Random esameri	3'681'701	80'393 (2,18%)	18'659 (23,2%)

Tabella 2: Percentuale di *read* poliadenilate rispetto al totale e percentuale di *read* il cui poli(A) è stato "trimmato" dallo *script* in quanto più lungo di 5 basi.

Si nota che la percentuale di *read* poliadenilate è più abbondante per i metodi oligo(dT) e random esameri, in particolare l'oligo(dT) presenta un numero maggiore di *read* con un poli(A) più lungo di 5 basi rispetto al metodo con i random esameri. Questo è dovuto al fatto che l'utilizzo dei primer oligo(dT) per la retrotrascrizione rileva in modo più preciso il 3' terminale dei trascritti.
Una volta selezionate le *read* di interesse la coda di poli(A) va eliminata per rendere la sequenza allineabile sul genoma, in quanto il poli(A) non è allineabile essendo aggiunto alla fine della trascrizione. Si procede quindi con il "trimming" della coda di poli(A), che consiste nell'eliminare una ad una le basi che la compongono, a partire dall'ultima (fig. 2, pt. 1). Questo processo è ripetuto sia sulla sequenza sia sulla qualità, contando il numero di basi "trimmate", valore che verrà salvato in un file di output necessario allo *script* successivo. Tuttavia in alcuni casi il poli(A) presente al 3' terminale delle *read* potrebbe essere determinato da un omopolimero di A presente all'interno del genoma; questo caso, che sarà analizzato in seguito, va quindi distinto dalle poliadenilazioni dovute alla terminazione della trascrizione. La soglia di *trimming* utilizzata va da un minimo di 5 basi di lunghezza del poli(A) fino ad un massimo di 40 per le *read* llumina *shotgun*, lunghe 76 basi, e di 15 per le *read* oligo(dT) e random esameri, lunghe 36 basi. Il secondo file di output è un file in formato "FASTQ" contenente le sequenze e la rispettiva qualità "trimmate". Ora si procede all'allineamento delle *read* "trimmate" sul genoma di *S. cerevisiae* ceppo S288c con PASS (Campagna *et al.* 2009). Prima dell'allineamento il programma filtra le *read* mantenendo solamente quelle di alta qualità. Non tutte le *read* sono allineate con successo, e, tra quelle allineate, solo una frazione di esse è allineata in un'unica posizione ("unique"). Solo le *read* "unique" sono utili per la mia analisi, perché sono quelle che possono essere usate per calcolare il profilo di espressione di ogni sequenza. Le altre si allineano in più posizioni, cioè nelle regioni ripetute del genoma.

2. Selezione delle *read* poliadenilate e visualizzazione delle terminazioni con Artemis

Il secondo programma che ho elaborato analizza l'output di PASS in formato "SAM" generato dall'allineamento sul genoma di riferimento delle sequenze il cui poli(A) è stato "trimmato". PASS nel suo output fornisce dei parametri che permettono di definire la qualità e l'unicità dell'allineamento per ogni *read*. Questi sono sfruttati dal mio *script* che con un "pattern matching" per ogni riga del file "SAM" controlla se la

read possiede un'unica corrispondenza nel genoma ("match"). In questo caso la *read* è "unique" ed è adatta per la mia analisi. Lo scopo del secondo *script* è verificare se l'omopolimero di A presente alla fine della *read* è effettivamente il prodotto di un evento di poliadenilazione o se si tratta invece di un omopolimero presente all'interno del genoma.
Lo *script* recupera quindi dal genoma di riferimento una sequenza a valle della posizione in cui la *read* si allinea lunga tante basi quante sono state "trimmate" dal primo programma (fig. 2, pt. 2). Infatti, se sul genoma vi è un omopolimero di A (o di T, se consideriamo le *read reverse*) subito a valle dell'allineamento con la *read* (o a monte), la *read* rappresenta un "falso positivo" in quanto pur possedendo una coda di poli(A) non identifica la terminazione del trascritto. La sequenza estratta dal genoma è definita non omopolimerica quando supera una soglia di basi diverse da A (o da T) pari al 30% della sua lunghezza in basi. La percentuale di *read* "veri positivi" ottenuta è indicata nella tabella 3.

Esperimento:	***Read* "unique":**	***Read* poliadenilate (% sulle *read* "unique"):**
Illumina *shotgun*	81,9%	85,35%
Oligo(dT)	57,9%	46,2%
Random esameri	48%	77%

Tabella 3: Percentuale di *read* "unique" selezionate dallo *script* rispetto al totale e percentuale di *read* poliadenilate identificate dallo *script* rispetto alle *read* analizzate (cioè le "unique").

Le *read* poliadenilate ("veri positivi") sono quindi selezionate e la posizione sul genoma corrispondente alla loro ultima base (prima del poli(A), dato che le *read* sono state "trimmate") dovrebbe identificare la posizione di terminazione del trascritto.
La struttura dei trascritti è abbastanza studiata in lievito e per la mia analisi mi sono avvalsa sia dell'annotazione dei geni e relative CDS (*Coding DNA Sequence*) sul genoma di *S. cerevisiae* ceppo S288c, sia su di una predizione della struttura del trascritto effettuata nel laboratorio dove ho svolto lo stage. Questa predizione è stata fatta sulla base di dati di RNA-seq per cui è stata utilizzata la tecnica di sequenziamento SBS (*Sequencing by Synthesis*, Illumina). Le *read* ottenute sono state allineate sul genoma di riferimento e dalla continuità della loro sovrapposizione, cioè considerando come un unico trascritto solo le sequenze genomiche coperte senza interruzioni da almeno una *read* è stata dedotta la struttura dei trascritti. Nel caso in cui fosse presente un'interruzione ma la CDS del *reference* in quel punto fosse invece continua la sequenza è stata considerata come un unico trascritto. Grazie a queste predizioni sono stata in grado di confrontare le posizioni delle terminazioni che ho ottenuto con le terminazioni "effettive" dei trascritti (cioè quelle predette dalla predizione appena descritta) e di visualizzarle con il *tool* Artemis. Lo *script* infatti genera dei file in formato "GFF" che sono leggibili da questo programma. In figura 3 le mie predizioni sono visualizzate con un rettangolino rosso, mentre in alto è mostrata la struttura del trascritto sul genoma di riferimento.

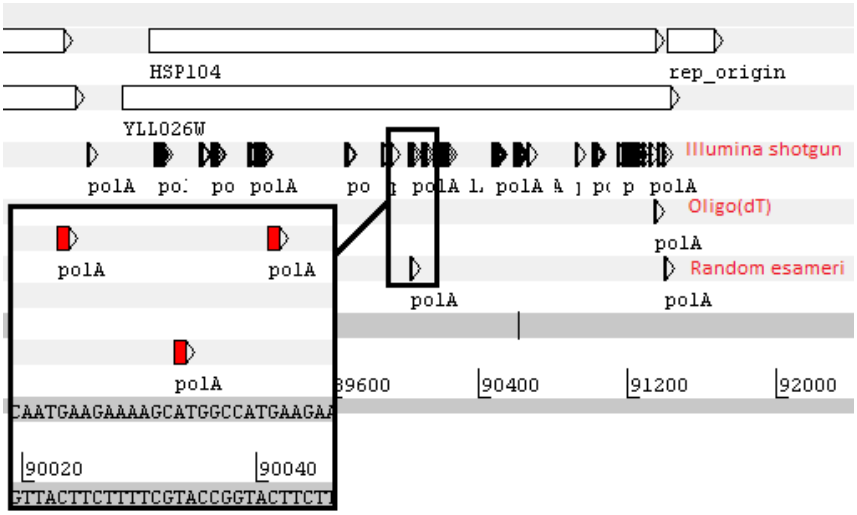

Figura 3: *HSP104* (YLL026W), localizzato in posizione 88486-91428 sul cromosoma 12. Esempio di gene con al suo interno numerose predizioni della posizione finale del trascritto visualizzato con Artemis. Le predizioni sono indicate con "polA". Sono mostrate le predizioni ottenute con i tre esperimenti, una per linea. L'ingrandimento indica tre terminazioni visualizzate in rosso.

A differenza di quanto ci si aspettava, cioè una distribuzione delle predizioni intorno alla posizione finale del trascritto, si nota che per alcuni trascritti le terminazioni non si concentrano sul 3' terminale ma sono distribuite lungo tutto il gene (fig. 3). Sono infatti presenti (soprattutto per quanto riguarda i dati Illumina *shotgun* che possiedono un numero di *read* molto più elevato rispetto agli altri due metodi) numerose terminazioni interne ai trascritti, lontano dal 3' terminale. Per comprendere la causa di questo andamento, forse dovuto ad anomalie nel processo di poliadenilazione, ho effettuato ulteriori analisi con alcuni *script* in Perl, che descriverò in seguito.

3. Identificazione delle terminazioni predette intorno alla posizione finale del trascritto

Le *read* "trimmate" che si allineano in corrispondenza di regioni genomiche trascritte ma lontano dalle regioni 3' terminali potrebbero indicare la posizione di terminazioni alternative, tuttavia ritengo più probabile che una terminazione alternativa cada entro un certo intervallo dalla posizione finale del trascritto. Per questo ho definito le terminazioni identificate da queste *read* in posizione "anomala" poiché troppo distanti, a mio avviso, dalla terminazione "effettiva" del trascritto. Esse potrebbero essere il risultato di un errore di sequenziamento, di anomalie nel processo di poliadenilazione o essere il frutto di segnali interni di poliadenilazione, probabilmente segnali "spuri", quindi incompleti. Trascritti con terminazione così precoce rispetto alla sequenza del gene corrispondente sarebbero probabilmente privi di segnali di STOP, quindi molto instabili e non funzionali. Perciò ho ritenuto opportuno elaborare un ulteriore *script* per distinguere le terminazioni predette in posizione attesa, cioè correttamente posizionate sul 3' terminale, da queste ultime, in quanto esse non sono utili per l'analisi del 3' terminale dei trascritti. Per fare questa distinzione il programma confronta la posizione del sito di poliadenilazione da me riscontrato con la predizione della struttura dei trascritti di cui sopra. La trascrizione spesso non termina in una posizione precisa, ma tale posizione può variare di qualche base in un certo intorno. Ho quindi considerato come possibile regione di terminazione un intervallo compreso tra 30 basi a monte e 30 basi a valle della terminazione "effettiva" del trascritto (-30 e +30 nucleotidi se si considera tale

terminazione in posizione 0). Inoltre anche la posizione di questa terminazione "effettiva" può variare di qualche base, poiché è stata predetta con la sovrapposizione di *read* derivate da dati di RNA-seq e quindi ha un certo margine di imprecisione. Secondo questo modello, se la predizione cade all'interno dell'intervallo considerato la sua posizione corrisponde all'atteso (fig. 2, pt. 3).
Confrontando i dati ottenuti per i tre esperimenti (fig. 4A) si nota che la libreria costruita con primer oligo(dT) presenta la percentuale più alta di predizioni vicine alla posizione finale del trascritto. Ciò è dovuto al fatto che il primer oligo(dT), appaiandosi alla coda di poli(A), è più preciso nella determinazione della posizione finale del trascritto, cioè dell'ultima base prima del poli(A). Per questo motivo retrotrascrivendo l'mRNA con questo metodo la maggior parte delle *read* che si ottengono si allinea in corrispondenza della parte terminale del trascritto. Il grafico B della figura 4 mostra in blu la percentuale di trascritti di *S. cerevisiae* ceppo S288c per cui è stato riscontrato almeno un sito di poliadenilazione e in rosso la percentuale di tali trascritti per cui è stato riscontrato un sito di poliadenilazione nella posizione attesa, cioè nell'intervallo di terminazione considerato. Si nota che i dati Illumina *shotgun* hanno rilevato siti di terminazione per un'alta percentuale di trascritti ma che pochi di questi si trovano nella regione 3' terminale.

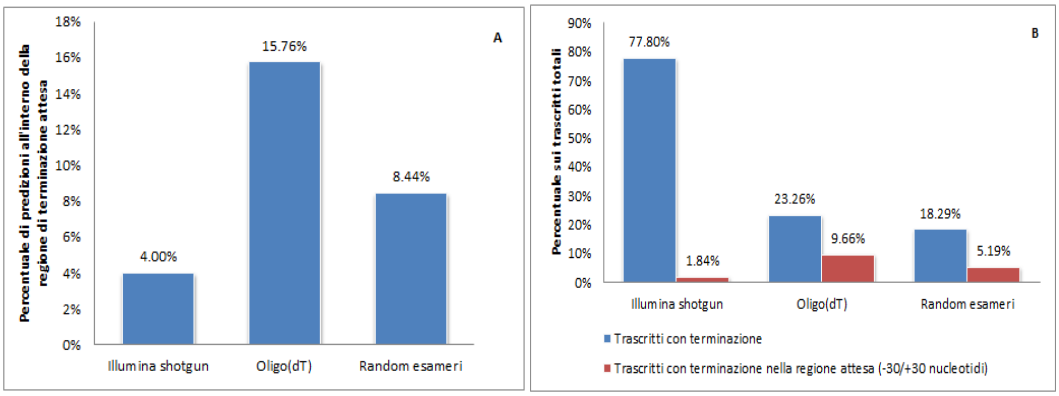

Figura 4: Grafico A: percentuale di siti di terminazione rilevati all'interno della regione attesa (-30, +30 nucleotidi) per i tre esperimenti. Grafico B: in blu è indicata la percentuale di trascritti sul totale dei trascritti di *S. cerevisiae* ceppo S288c per cui è stato riscontrato almeno un sito di poliadenilazione e in rosso la percentuale di tali trascritti per cui è stato riscontrato un sito di poliadenilazione nella regione attesa.

I diagrammi di Venn in figura 5 costruiti con il programma Venn Master rappresentano i trascritti (divisi per ognuno dei tre metodi di costruzione delle librerie) per cui è stata predetta almeno una terminazione nella regione attesa (diagramma 5A) e per cui è stata predetta almeno una terminazione in posizione anomala (diagramma 5B). Le intersezioni dei tre insiemi (uno per esperimento) indicano che per alcuni trascritti sono stati riscontrati siti di poliadenilazione da più di una libreria di RNA-seq. Alcuni hanno terminazioni predette da tutti e tre gli insiemi di dati (in fig. 5: intersezione verde). Per quanto riguarda il grafico 5A si nota che per 105 trascritti sono state riscontrate terminazioni nella regione attesa (-30, +30 nucleotidi dalla posizione finale) da tutti e tre i metodi di costruzione delle librerie. Il grafico 5B invece mostra che le predizioni anomale, cioè quelle che cadono all'interno del trascritto, sono molto numerose per la libreria "Illumina RNA ligation" (4317 trascritti) e meno numerose per gli altri due metodi. Ciò è dovuto sicuramente alla superiorità numerica delle *read* di questa libreria, tuttavia in proporzione si riscontrano molte più predizioni interne al trascritto con questo

metodo rispetto agli altri due (fig. 4). Questo mi porta ad ipotizzare che per questi dati ci possa essere stato un errore di sequenziamento o un bias dovuto al metodo di costruzione della libreria. In ogni caso il motivo per cui si presenta questa anomalia resta di difficile comprensione.
Si può affermare inoltre che i trascritti con predizioni anomale rilevate dalle librerie oligo(dT) e random esameri presentano predizioni anomale anche nei dati della libreria "Illumina RNA ligation", poiché gli insiemi nel diagramma sono quasi del tutto sovrapposti. È probabile che questi geni siano molto espressi e che siano rappresentati quindi da un alto numero di *read*, cosa che aumenta la possibilità di trovare *read* poliadenilate che mappano all'interno di quel gene.

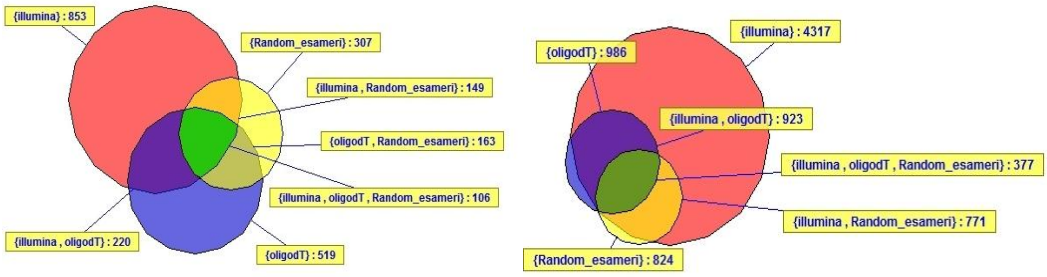

Figura 5: Diagramma A, a sinistra: ogni insieme rappresenta i trascritti che hanno riscontrato una terminazione nella regione attesa (-30, +30 nucleotidi dalla posizione finale del trascritto) per ciascuno dei tre esperimenti. In rosso è indicata la libreria "Illumina RNA ligation", in blu la libreria oligo(dT) e in giallo la libreria random esameri. L'intersezione in verde rappresenta quei trascritti la cui terminazione è stata riscontrata da tutti e tre gli esperimenti. Diagramma B, a destra: i tre insiemi rappresentano quei trascritti che presentano almeno una terminazione in posizione anomala, cioè interna al trascritto e lontana dal 3' terminale. Si nota che per la libreria "Illumina RNA ligation" il numero di trascritti con predizioni anomale è molto maggiore rispetto alle altre due librerie.

4. Correlazione tra terminazioni e segnali di poliadenilazione

Una possibile ipotesi che spieghi la presenza di un elevato numero di predizioni in posizione anomala è che esse possano essere determinate dalla presenza di segnali di poliadenilazione, probabilmente parziali, all'interno delle regioni genomiche trascritte. Invece per le terminazioni identificate da quelle *read* che si allineano in corrispondenza del 3' terminale ci si aspetta che il sito di poliadenilazione sia determinato da un segnale di poliadenilazione pressoché completo. Ho quindi elaborato uno *script* che per ogni terminazione predetta estrae dal genoma di riferimento le sequenze corrispondenti alle quattro regioni in cui dovrebbero essere presenti i quattro segnali di poliadenilazione identificati in *S. cerevisiae* (Graber *et al.* 1999). La posizione delle quattro regioni (rispetto alla terminazione predetta) è: -80/-25 per *l'Efficiency Element* (EE), -35/-15 per il *Positioning Element* (PE), -20/+5 per il *Pre-cleavage Element* (PrCl) e +5/+30 per il *Downstream Element* (DE) (fig. 2 pt. 4). Lo *script* ricerca all'interno di ciascuna regione il corrispettivo segnale di poliadenilazione con un "pattern matching", cioè scorrendo la sequenza genomica della regione corrispondente alla ricerca di precise "parole", in questo caso le corte sequenze componenti i quattro segnali di poliadenilazione (tabella 1). L'analisi è stata ripetuta due volte: la prima ricercando nelle regioni genomiche sopra elencate solo i segnali "forti" di poliadenilazione (indicati in grassetto in tabella 1), che d'ora in avanti indicherò come "parole ottimali"; la seconda ricercando l'insieme totale dei segnali di poliadenilazione, che indicherò come "parole totali". I dati dei tre esperimenti sono stati unificati in un unico file di input. Per ogni predizione il programma fornisce in

output una matrice che riporta "0" se la "parola" non è stata trovata all'interno della regione considerata e "1" in caso contrario. Questi valori sono stati quindi sommati per ottenere una stima della completezza del sito di poliadenilazione e del numero di predizioni che presentano un segnale di poliadenilazione parziale o completo. Innanzitutto vengono messi a confronto i risultati ottenuti per le terminazioni in posizione anomala, cioè interne al trascritto, e quelle in posizione attesa, cioè nell'intervallo intorno all'effettiva terminazione del trascritto (fig. 6A). Si nota una percentuale più alta di predizioni aventi una o più porzioni del segnale "totale" di poliadenilazione per le terminazioni in posizione attesa e una percentuale più alta con zero segnali per le predizioni anomale (interne al trascritto). Questo esito era atteso, in quanto le predizioni intorno alla posizione finale del trascritto hanno una maggior probabilità di essere il risultato di un effettivo evento di poliadenilazione che avviene in seguito al rilevamento di un segnale di poliadenilazione.
Il grafico 6B invece compara la sensibilità delle due analisi effettuate prima con la ricerca delle "parole totali" e poi con la ricerca delle "parole ottimali". La prima modalità, come ci si poteva aspettare, rileva più segnali di poliadenilazione della seconda.

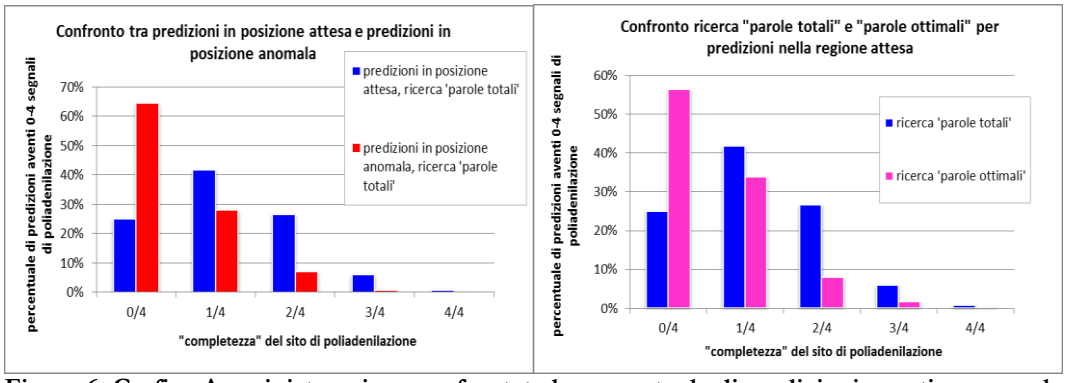

Figura 6: Grafico A, a sinistra: viene confrontata la percentuale di predizioni aventi un segnale di poliadenilazione parziale o completo tra le predizioni riscontrate in posizione attesa (-30, +30 nucleotidi dalla posizione finale del trascritto), in blu, e quelle riscontrate in posizione anomala, cioè interne al trascritto e lontane dal 3' terminale, in rosso. Le terminazioni entro il 3' terminale del trascritto presentano segnali di poliadenilazione più completi nelle regioni genomiche circostanti. Grafico B, a destra: viene confrontata la sensibilità delle analisi eseguite ricercando le "parole ottimali" (in rosa) e le "parole totali" (in blu) per le predizioni nella regione attesa. Si nota che la ricerca delle "parole totali" ha una maggior sensibilità nel rilevare segnali di poliadenilazione.

Come accennato in precedenza, le terminazioni interne al trascritto rilevate dai dati Illumina *shotgun* della libreria "Illumina RNA ligation" (fig. 3) potrebbero essere il risultato di anomalie nel processo di poliadenilazione o essere il frutto di segnali di poliadenilazione parziali ("spuri") presenti nelle regioni trascritte del genoma. Per rispondere a questa questione ho deciso di confrontare queste terminazioni in posizione anomala con una serie di posizioni prese randomicamente all'interno delle regioni trascritte. Se tali terminazioni fossero il frutto di segnali di poliadenilazione interni alle regioni trascritte la loro posizione dipenderebbe dalla distribuzione di questi ultimi. L'analisi di posizioni casuali invece rileva solo segnali di poliadenilazione distribuiti nel genoma in posizione del tutto casuale. Uno *script* in Perl ha quindi generato una serie di posizioni casuali all'interno delle regioni trascritte del genoma, per le quali sono stati ricercati i segnali di poliadenilazione nelle regioni circostanti con il quarto *script*, come era stato fatto per le predizioni ottenute in precedenza (per semplicità è stata considerata solo la ricerca delle "parole totali", che

è la più sensibile). L'esito è mostrato in fig. 7A. L'abbondanza dei segnali EE, PE, PrCl e DE nelle regioni genomiche considerate è analoga a quella rilevata per le predizioni interne al trascritto. Questo indica che i siti di poliadenilazione interni individuati dalle *read* Solexa sono il risultato di eventi casuali determinati forse da anomalie nel processo di poliadenilazione. Infatti, se fossero dovuti a segnali di poliadenilazione interni al trascritto, avremmo dovuto ottenere risultati diversi rispetto all'analisi casuale.
Successivamente è stata analizzata l'abbondanza dei singoli siti EE, PE, PrCl e DE per tutte le analisi considerate. Osservando gli andamenti in fig. 7B si vede che i quattro siti sono nettamente più abbondanti quando si tratta di predizioni intorno alla posizione finale del trascritto. Si nota una certa prevalenza dell'EE per quanto riguarda l'analisi delle predizioni in posizione attesa. La presenza di un segnale "forte" nell'EE è infatti segno di un vero sito di poliadenilazione.

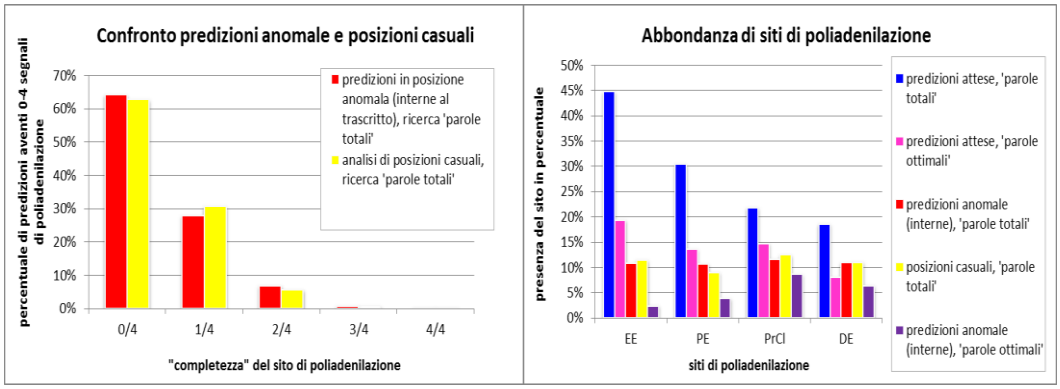

Figura 7: Grafico A, a sinistra: confronto tra i segnali di poliadenilazione rilevati per le terminazioni interne al trascritto, quindi in posizione anomala rispetto all'atteso (in rosso), e i segnali rilevati per terminazioni con posizioni casuali e interne al trascritto (in giallo). Come si nota i risultati sono quasi del tutto sovrapponibili. Grafico B, a destra: Abbondanza di ciascun segnale di poliadenilazione (EE, PE, PrCl, DE) in ciascuno dei casi considerati: predizioni in posizione attesa o in posizione anomala, analisi con posizioni casuali, ricerca di "parole totali" o "ottimali". La ricerca di "parole totali" per le terminazioni riscontrate in posizione attesa (-30, +30 nucleotidi dalla posizione finale del trascritto) presenta una percentuale di presenza più alta per ogni segnale di poliadenilazione, in particolare per l'EE.

5. Analisi delle terminazioni alternative

Come analisi conclusiva ho ritenuto interessante ricercare trascritti con possibili terminazioni alternative. Esse non sono molto studiate in lievito e si conosce ancora poco riguardo al motivo per cui certi geni presentano una posizione di terminazione della trascrizione che può variare entro un numero considerevole di basi. Inoltre non è chiara la correlazione con la funzione che tali regioni trascritte ricoprono all'interno della cellula.
Ho scelto di analizzare solo i trascritti con più di una predizione del sito di poliadenilazione entro l'intervallo -30, +30 nucleotidi dalla posizione finale del trascritto (fig. 2, pt. 5), trascurando le predizioni lontane dal 3' terminale. Ho elaborato quindi uno *script* che scorre questi trascritti ricercando quelli che presentano più di una terminazione nell'intervallo considerato, mantenendo solo quelli in cui la prima e l'ultima sono distanti almeno trenta basi (ho scelto di impostare questo valore soglia, ma essa può essere modificata dato che questo parametro è settabile al momento dell'esecuzione del programma). Due terminazioni, per essere considerate come "alternative", devono distare almeno un certo numero di basi poiché corrispondono a siti di poliadenilazione differenti. Esse non vanno confuse con le

variazioni di poche basi che si riscontrano per la posizione di un singolo sito di poliadenilazione. Lo *script* stampa in un file di output il nome del gene, lo *strand* e il cromosoma con le rispettive terminazioni ordinate per posizione e genera un file visualizzabile in Artemis con le terminazioni alternative trovate (fig. 8).

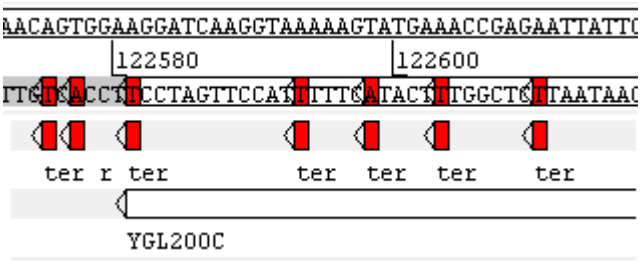

Figura 8: *EMP24* (YGL200C), localizzato in posizione 122581-123379 sul cromosoma 7. Esempio di trascritto con possibile terminazione alternativa visualizzato con Artemis. È mostrata solo la parte terminale del trascritto e le predizioni, in rosso, sono indicate con "ter". Questo gene presenta numerose predizioni nella regione circostante la posizione finale del trascritto, con la prima e l'ultima predizione distanti almeno 30 basi.

Sono stati individuati 101 trascritti con terminazione alternativa con una soglia di distanza di trenta basi tra la prima e l'ultima terminazione su un totale di 1833 trascritti con terminazioni nell'intervallo che va da 30 basi a monte a 30 basi a valle dalla posizione finale del trascritto (i dati dei tre esperimenti sono stati unificati in un unico file di input). A questo punto è stata ricercata la funzione di questi particolari geni all'interno della cellula e valutato se appartengono a delle categorie funzionali specifiche utilizzando il software GoMiner (Zeeberg *et al.* 2003). Questo programma utilizza come input due file: una lista totale di geni e una sottolista. Nel nostro caso consideriamo come lista totale la lista di geni che presentano una terminazione all'interno della regione attesa (-30, +30 nucleotidi) e come sottolista la lista di geni con putative terminazioni alternative. Le categorie di Gene Ontology più significative rilevate nell'analisi sono indicate nella tabella sottostante (tabella 4).

HYPERLINKED GO CATEGORY	TOTAL GENES	CHANGED GENES	ENRICHMENT	LOG10(p)	p value
GO:0006412_translation	267	52	2.426508	-11.8867	**1.29814E-12**
GO:0009059_macromolecule_biosynthetic_process	438	63	1.792074	-8.6378	**2.30252E-09**
GO:0034645_cellular_macromolecule_biosynthetic_process	434	62	1.779883	-8.27555	**5.3021E-09**
GO:0006417_regulation_of_translation	93	24	3.215273	-7.36913	**4.27434E-08**
GO:0032268_regulation_of_cellular_protein_metabolic_process	97	24	3.082685	-6.978	**1.05197E-07**
GO:0010608_posttranscriptional_regulation_of_gene_expression	98	24	3.051229	-6.88399	**1.3062E-07**
GO:0051246_regulation_of_protein_metabolic_process	108	25	2.88407	-6.64765	**2.25086E-07**
GO:0010467_gene_expression	467	61	1.627431	-6.33591	**4.61412E-07**
GO:0010556_regulation_of_macromolecule_biosynthetic_process	216	37	2.134212	-6.19863	**6.32946E-07**
GO:0031326_regulation_of_cellular_biosynthetic_process	216	37	2.134212	-6.19863	**6.32946E-07**
GO:0009889_regulation_of_biosynthetic_process	217	37	2.124377	-6.14344	**7.18719E-07**
GO:0044267_cellular_protein_metabolic_process	475	61	1.600021	-6.02369	**9.46913E-07**
GO:0044249_cellular_biosynthetic_process	562	68	1.507517	-6.00653	**9.85068E-07**
GO:0009058_biosynthetic_process	568	68	1.491592	-5.78998	**1.62189E-06**

Tabella 4: La tabella mostra le 15 categorie di Gene Ontology con p-value al di sotto di 1.65E-06 ottenute dall'analisi di GoMiner. La colonna "Total genes" indica il numero di geni della lista totale (che comprende i geni con terminazioni nella regione attesa) appartenenti a quella categoria di GO. La colonna "Changed genes" invece rappresenta il numero dei geni della sottolista, cioè la lista dei geni con possibile terminazione alternativa, appartenenti a quella categoria di GO. La categoria di GO con il p-value più basso (1.298E-12) è GO:0006412, *translation*.

Nel caso di geni molto espressi, avendo essi un numero di *read* molto elevato, è maggiore la probabilità di ottenere delle terminazioni alternative rispetto ai geni poco espressi sui quali mappano un numero molto ridotto di sequenze. Per questo motivo è stato scelto di usare come lista totale di geni quelli con le terminazioni individuate

nella regione attesa (-30, +30 nucleotidi) e non la lista totale dei geni di lievito. Se avessimo usato la lista totale dei geni di lievito probabilmente avremmo identificato nell'analisi di GO categorie funzionali di geni molto espressi.
REViGO (Supek *et al.* 2011) utilizza come input la lista dei termini di GO separati con un carattere di tabulazione dal p-value, valore assegnato ad ogni categoria di GO dal test statistico di GoMiner, e li "clusterizza" (cioè li raggruppa) mettendo a disposizione dei grafici che presentano in modo chiaro la correlazione funzionale tra i vari "cluster" (gruppi) di geni. Inoltre è possibile aumentare la stringenza della procedura di "clustering" settando il parametro *Allowed similarity* sull'interfaccia web. In figura 9 è rappresentato il grafico "Tree map" di REViGO, che presenta le varie categorie di GO in modo gerarchico. Ogni rettangolo rappresenta un "cluster", cioè un gruppo di geni appartenenti alla stessa categoria di GO. Queste categorie sono a loro volta raggruppate in "supercluster", ciascuno indicato da un diverso colore. La dimensione di ogni rettangolo dipende dal logaritmo negativo in base 10 del p-value associato a quel "cluster": i rettangoli più grandi presentano un p-value minore.

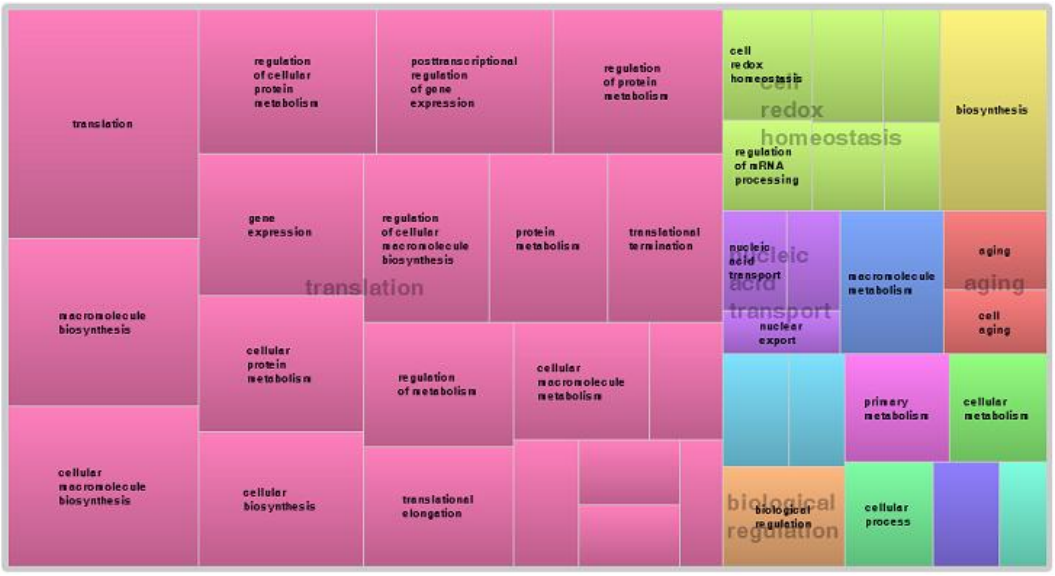

Figura 9: REViGO Tree map. Ogni rettangolo rappresenta un "cluster". I più rappresentativi sono unificati in "supercluster" di termini correlati, visualizzati in differenti colori. L'abbondanza delle categorie di GO, cioè la dimensione dei rettangoli, si basa sul logaritmo negativo in base 10 del p-value. Il "supercluster" più abbondante è *translation*. Ciò indica che i geni con possibile terminazione alternativa sono prevalentemente coinvolti nel processo di traduzione, poiché codificano per proteine che partecipano a questo processo spesso come elementi regolatori.

L'analisi di REViGO mostra che i geni con terminazioni alternative sono prevalentemente coinvolti nella traduzione. Inoltre se si ricerca nell'SGD (*Saccharomyces Genome Database*) il codice identificativo di alcuni di questi geni si trova che essi codificano per proteine componenti le subunità ribosomali oppure fattori di allungamento o d'inizio della traduzione. Spesso si tratta quindi di proteine regolatrici implicate nel processo di traduzione.

BIBLIOGRAFIA

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S.G. (1996) **Life with 6000 genes**. *Science*, 274: 563-567.

Graber, J.H., Cantor, C.R., Mohr, S.C. & Smith, T.F. (1999) **Genomic detection of new yeast pre-mRNA 3'-end-processing signals**. *Nucleic Acids Research*, 27: 888–894.

Wang, Z., Gerstein, M. & Snyder, M. (2009) **RNA-seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics*, 10: 57–63.

Mardis, E.R. (2008) **Next-Generation DNA Sequencing Methods**. *Annual Review of Genomics and Human Genetics*, 9: 387-402.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A. & Regev, A. (2010) **Comprehensive Comparative analysis of strand-specific RNA sequencing methods**. *Nature Methods*, 7: 709–715.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. (2008) **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science*, 320: 1344–1349.

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., Vitulo, N. & Valle, G. (2009) **PASS: a program to align short sequences**. *Bioinformatics*, 25: 967-968.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. & Barrell, B. (2000) **Artemis: sequence visualization and annotation**. *Bioinformatics*, 16: 944-945.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C. & Weinstein, J.N. (2003) **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome Biology*, 4: R28.

Supek, F., Bošniak, M., Škunca, N., Šmuc, T. (2011) **REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms**. *PLoS ONE*, 6: e21800.