Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea in Fisica

METODOLOGIE PER L'ANALISI STATISTICA DI DATI DI SEQUENCING RELATIVI A UN ESPERIMENTO DI RNA INTERFERENCE

Relatore:
Prof. Gastone Castellani

Correlatore:
Dott. Daniel Remondini

Presentata da:
Claudio Corte Coi

Sessione II
Anno Accademico 2012/2013

Sommario

La RNA interference è un processo attraverso il quale alcuni piccoli frammenti di RNA (19-25 nucleotidi) sono in grado di silenziare l'espressione genica. La sua scoperta, nel 1998, ha rivoluzionato le concezioni della biologia molecolare, minando le basi del cosiddetto Dogma Centrale. Si è visto che la RNAi riveste ruoli fondamentali in meccanismi di regolazione genica, nello spegnimento dell'espressione e funziona come meccanismo di difesa innata contro varie tipologie di virus. Proprio a causa di queste implicazioni richiama interesse non solo dal punto di vista scientifico, ma anche da quello medico, in quanto potrebbe essere impiegata per lo sviluppo di nuove cure. Nonostante la scoperta di tale azione desti la curiosità e l'interesse di molti, i vari processi coinvolti, soprattutto a livello molecolare, non sono ancora chiari.

In questo lavoro si propongono i metodi di analisi di dati di un esperimento prodotto dall'Istituto di Biologia molecolare e cellulare di Strasburgo. Nell'esperimento in questione vengono studiate le funzioni che l'enzima Dicer-2 ha nel *pathway* - cioè la catena di reazioni biomolecolari - della RNA interference durante un'infezione virale nel moscerino della frutta *Drosophila Melanogaster*. Per comprendere in che modo Dicer-2 intervenga nel silenziamento bisogna capire in quali casi e quali parti di RNA vengono silenziate, a seconda del diverso tipo di mutazione dell'enzima stesso. Dunque è necessario sequenziare l'RNA nelle diverse condizioni sperimentali, ottenendo così i dati da analizzare. Parte dei metodi statistici che verranno proposti risultano poco convenzionali, come conseguenza della peculiarità e della difficoltà dei quesiti che l'esperimento mette in luce.

Siccome le tematiche affrontate richiedono un approccio sempre più interdisciplinare, è aumentata considerevolmente la richiesta di esperti di altri settori scientifici come matematici, informatici, fisici, statistici e ingegneri. Questa collaborazione, grazie a una diversità di approccio ai problemi, può fornire nuovi strumenti di comprensione in ambiti che, fino a poco tempo fa, rientravano unicamente nella sfera di competenza dei biologi.

Indice

Introduzione

In questo lavoro di tesi viene discusso un esperimento relativo alla RNA interference nel moscerino della frutta *Drosophila Melanogaster*. Questo esperimento è stato condotto da un gruppo di ricerca facente capo all'Istituto di Biologia molecolare e cellulare, e al Dipartimento di Scienze della Vita, entrambi dell'Università di Strasburgo; si vuole sottolineare il fatto che il gruppo in questione è guidato da Jules Hoffmann, premio Nobel per la medicina nel 2011. Il gruppo di biofisica del Dipartimento di Fisica dell'Università di Bologna ha come obiettivo l'analisi dei dati di questo esperimento, quindi nelle pagine seguenti proponiamo dei metodi statistici di analisi, che saranno applicati a tali dati, per rispondere alle domande che il gruppo di Strasburgo si è posto.
È importante dare un ampia introduzione al problema, dal momento che questi argomenti sono normalmente estranei ai fisici.
Perciò nel capitolo 1 verrà discusso il contesto biologico in cui si andrà ad operare, ovvero gli acidi nucleici, sottolineando il ruolo chiave che i nuovi tipi di RNA stanno assumendo, in contrasto con le convinzioni del Dogma Centrale. Particolare risalto è ovviamente dato al processo della RNA interference, della quale verrà data una spiegazione prevalentemente "cronologica", ovvero seguendo passo passo i nuovi meccanismi appresi, a partire dalla prima scoperta e giungendo a considerare tutte le implicazioni che ne possono derivare.
Nel capitolo 2 si parlerà del sequenziamento, che è il metodo fondamentale tramite cui vengono effettuate le analisi relative a questo processo. In particolare verranno spiegati i metodi di Next Generation Sequencing (NGS), nati intorno al 2005; questi ultimi hanno portato a un rapido incremento degli studi riguardanti DNA e RNA, a causa dell'enorme aumento di dati forniti e dell'abbassamento dei costi. Si discuteranno i vantaggi di queste tecniche, ma anche i loro limiti, come ad esempio la minore lunghezza dei *reads*, i frammenti di codice che vengono sequenziati. Un altro problema relativo alle nuove tecniche

è quello dell'enorme mole di dati, che molto spesso richiede l'uso di reti di calcolatori per l'immagazzinamento e l'analisi. L'analisi effettuata è di tipo statistico, poiché, essendo il sequenziamento un processo casuale, i reads vengono considerati variabili aleatorie che seguono opportune distribuzioni: verrà fatto riferimento alla distribuzione di Poisson e alla binomiale negativa, e saranno forniti alcuni esempi di software. Inoltre verrà data un'ampia panoramica delle applicazioni del sequencing, molte delle quali stanno godendo di particolare fortuna grazie alle tecniche NGS.

Conclusa la spiegazione dei metodi di sequenziamento ci si concentrerà sull'esperimento vero e proprio: nel capitolo 3, infatti, verrà fornita una discussione più dettagliata sulla RNAi, sottolineando il ruolo che l'enzima Dicer-2 sembra avere, con particolare attenzione allo studio del dominio Elicasi, dal momento che quest'ultimo si è dimostrato avere comportamenti inaspettati. Sarà data una descrizione particolareggiata dell'esperimento, dei metodi utilizzati e dei risultati ottenuti, ovvero le domande su cui concentrarsi per l'analisi dati. Infine proponiamo i nostri modelli di analisi, che, a causa della peculiarità del problema, possono risultare differenti dai metodi classici che solitamente vengono utilizzati nell'analisi differenziale di espressione.

Capitolo 1

Il contesto biologico

Prima di sviluppare una trattazione approfondita del problema, è necessario descrivere l'ambito biologico della discussione, ovvero gli acidi nucleici e le loro funzioni (sezione 1.1), le teorie e i processi rilevanti ad essi legati (sezione 1.2), con particolare riguardo al meccanismo di RNA interference (sezione 1.3).

1.1 Gli acidi nucleici

Gli acidi nucleici sono delle macromolecole molto importanti a livello biologico poiché esse contengono e trasportano tutta l'informazione genetica di un organismo. Data la loro funzione si trovano principalmente all'interno del nucleo, la zona maggiormente protetta della cellula, ma in misura minore anche nel citoplasma. Si distinguono in acido ribonucleico (RNA) e desossiribonucleico (DNA). La forma in cui si trovano all'interno del nucleo è detta *cromatina*, che consiste sostanzialmente di filamenti di DNA ripiegati in molti modi diversi attorno a delle proteine chiamate *istoni*, formando delle strutture piuttosto complicate.

1.1.1 Il DNA

Il DNA è il biopolimero che contiene tutto il codice genetico di un organismo vivente. È strutturato in due filamenti antiparalleli (denominati *senso* e *antisenso*) che si avvolgono

su loro stessi a formare una doppia elica: questa catena è larga poco più di 2 nanometri, ma ha lunghezza che può arrivare a dimensioni dell'ordine del metro; per essere contenuta in una cellula, che ha dimensioni di qualche micron, deve, come già accennato, ripiegarsi attorno agli istoni, costituendo in tal modo la cromatina.
I suoi monomeri costituenti sono detti nucleotidi e sono formati ciascuno da un gruppo fosfato, uno zucchero a cinque atomi di carbonio, che nel caso del DNA è il desossiribosio, e una base azotata (cfr fig. 1.1). Le quattro basi sono adenina e guanina, dette purine, e timina e citosina, dette pirimidine. Una base può legarsi solo con la sua base complementare (A-T e C-G), tramite un legame idrogeno, che, quindi, può essere rotto con relativa semplicità. Gli atomi di carbonio sono numerati a partire da quello che si lega alla base. I legami tra gruppo fosfato e zucchero, e quindi tra i nucleotidi, si hanno ad opera degli atomi di Carbonio 3' e 5'.

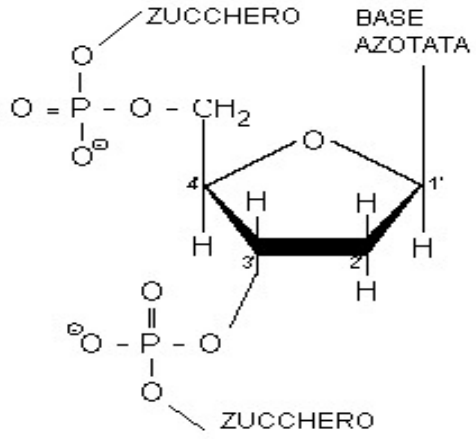

Figura 1.1: struttura di un nucleotide

Ciò che differenzia un nucleotide dall'altro è appunto la diversa base che lo costituisce e la sequenza in cui sono disposte le basi è quella che determina il codice genetico dell'individuo. A livello biologico il DNA è un componente essenziale della cellula, dal momento che in esso sono contenute tutte le informazioni per generare altre cellule. Dunque una sua funzione importante da considerare nella divisione cellulare è la capacità di replicarsi; infatti la cellula figlia deve avere esattamente lo stesso DNA della madre. Nella divisione cellulare il doppio filamento di DNA si rompe e ognuno dei due filamenti ne sintetizza uno complementare grazie all'enzima DNA polimerasi. Questa sintesi ha una direzione

preferenziale, che di solito è da 5' a 3'.
Per quanto riguarda la determinazione del fenotipo - cioè delle caratteristiche dell'organismo che effettivamente sono espresse e visibili - non tutto il materiale viene codificato, bensì si distingue in sequenze codificanti o *esoni*, che costituiscono solo il 5% del totale e non codificanti, gli *introni*. Questi ultimi sembravano non rivestire alcun ruolo rilevante fino a una ventina di anni fa, tanto che venivano definiti "DNA spazzatura", tuttavia si è scoperto che in realtà possiedono una certa importanza in meccanismi di regolazione genica, dal momento che vanno a formare i miRNA (cfr. sez. 1.1.2). Sequenze di introni ed esoni, mappate precisamente all'interno del codice, vengono dette *geni*; questi rappresentano le unità ereditarie fondamentali degli organismi e vengono ritrovati anche nell'RNA. Il DNA, mediante un processo di *trascrizione* viene trasformato in RNA, il quale svolge numerose funzioni tra cui quella di codificare le proteine.

1.1.2 L'RNA

L'RNA o acido ribonucleico è un altro biopolimero presente negli organismi viventi. A differenza del DNA, è presente molto spesso a singolo filamento, *ssRNA* (single stranded), piuttosto che nella forma *dsRNA* (double stranded); per quanto riguarda la struttura molecolare ha il ribosio come zucchero pentoso e l'uracile come base equivalente alla timina del DNA. La sua formazione avviene a partire dalla sintesi di molecole di DNA, attraverso un processo denominato trascrizione. Fino alla fine degli anni '90 (come si vedrà più in dettaglio nella prossima sezione) era consolidata l'idea che l'unica funzione dell'RNA fosse quella di tradurre l'informazione presente nel DNA e codificare le proteine: gli RNA conosciuti erano perciò mRNA, rRNA e tRNA. Tuttavia si scoprì che molti tipi di RNA, soprattutto quelli in filamenti brevi, avevano funzione di silenziamento genico e regolazione; questo diede una risposta anche ai quesiti sulla funzione degli introni.

L'mRNA

L'mRNA, o RNA messaggero è il principale responsabile della codifica del DNA in proteine, poiché contiene l'informazione genetica *trascritta* del DNA. Questo processo di trascrizione, portato avanti dall'enzima RNA polimerasi II, avviene per mezzo di triplette di basi, i *codoni* ($4^3 = 64$ in tutto), che codificano per gli amminoacidi[1]; siccome il codice

[1]Gli amminoacidi sono i monomeri delle proteine

è degenere, alcune triplette formano lo stesso amminoacido. I codoni vengono poi letti dai ribosomi, vere e proprie macchine catalitiche che eseguono il processo di sintesi proteica (cfr. fig. 1.2). In realtà negli organismi eucarioti si ha un passaggio intermedio, poiché il DNA viene prima trascritto in pre-mRNA, che è esattamente identico al DNA. Quindi per giungere all'RNA maturo sono necessari altri processi, di cui il principale è lo *splicing*; quest'ultimo serve tagliare e cucire il filamento, rimuovendo le sequenze introniche, in modo da mantenere solo la parte di RNA codificante.

Il tRNA

Questo tipo di RNA trasferisce un amminoacido specifico ad una catena polipeptidica in crescita nel sito ribosomiale in cui avviene la sintesi proteica durante il processo di traduzione, cioè assembla i vari amminoacidi per formare le proteine. E' piuttosto corto, dato che è formato da 74-93 nucleotidi e possiede un sito di attacco, l'*anticodone*, formato da tre nucleotidi, che consente il legame con il tripletto complementare dell'mRNA, permettendone quindi il trasporto. L'enzima che ne avvia la trascrizione, diversamente dall'mRNA, è la RNA polimerasi III.

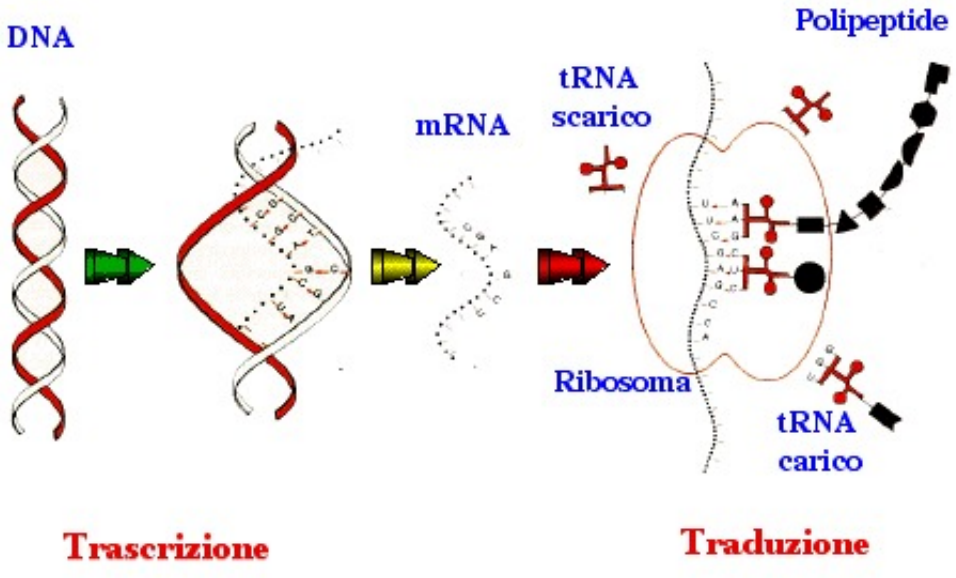

Figura 1.2: processo di sintesi proteica

L'rRNA

L'RNA ribosomiale è la tipologia più abbondante di RNA presente nella cellula. Essendo il componente essenziale (circa i due terzi) dei ribosomi, la sua funzione è quella di favorire l'assemblaggio delle proteine stesse. Perciò ha un ruolo di interazione con il tRNA durante la sintesi proteica. Una sua caratteristica peculiare è di essere il materiale genetico più conservato, cioè che ha subito meno modificazioni nel corso dell'evoluzione, e quindi viene spesso usato per identificare il gruppo tassonomico di appartenenza di un dato organismo.

I principali "nuovi" tipi di RNA sono i micro RNA (miRNA) e gli short interfering RNA (siRNA). Entrambi rivestono un ruolo fondamentale nella RNA interference (cfr. sez. 1.3); di seguito viene fornita una breve descrizione introduttiva.

Il miRNA

I microRNAs sono dei filamenti di circa 20-25 nucleotidi di RNA non codificante che svolgono una funzione di regolatori di espressione. Essi sono prodotti da RNA lungo, sia codificante, sia, in misura molto maggiore, da RNA non codificante; questa è una delle spiegazioni della funzionalità degli introni.

Il siRNA

Analogamente al miRNA, questo tipo di RNA non codificante ha strutture a filamenti molto brevi e ha anch'esso funzione di regolatore di espressione, intervenendo nel silenziare alcuni geni. A differenza del miRNA, però, il filamento è doppio e l'RNA è molto spesso esogeno, cioè non si trova naturalmente all'interno della cellula; infatti la sua presenza è causata da un'iniezione esterna, dovuta per esempio al contatto con virus. La lunghezza è variabile, ma molto spesso è di 21 nucleotidi, poiché in questi casi lo "spezzettamento" viene effettuato dall'enzima Dicer, che agisce come una sorta di righello molecolare.

1.2 Il dogma della biologia molecolare e il suo superamento

Quando, nel 1958, Francis Crick provò a spiegare il flusso dell'informazione genetica, propose un modello che può essere sostanzialmente sintetizzato con la frase "il DNA, fa l'RNA, che fa le proteine". Questa affermazione è la base di una concezione della funzionalità degli acidi nucleici, che viene conosciuta col nome *Dogma Centrale della biologia molecolare* (cfr. fig. 1.3), anche se Crick non lo riteneva affatto un dogma, quanto piuttosto una semplice ipotesi; inizialmente, però, i biologi fecero di questa ipotesi una verità fondamentale. In realtà quello che Crick non ammetteva non era il processo di tipo inverso tra RNA e DNA, ma solo la trasmissione di informazioni dalle proteine a uno dei due acidi nucleici. Come già accennato, quindi, si riteneva che l'unica funzione dell'RNA fosse quella

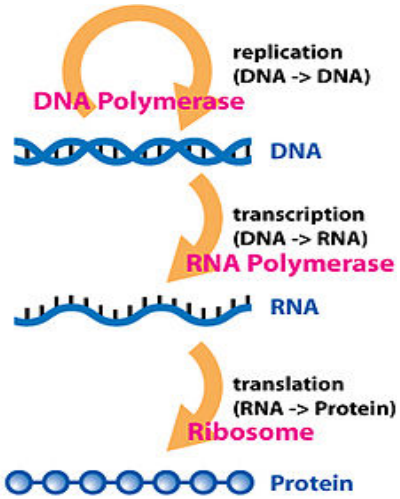

Figura 1.3: Flusso di informazioni secondo il Dogma Centrale

di tramite fra DNA e proteine, con differenze di compiti tra i tre tipi di RNA sopracitati. In realtà si scoprirono (e si stanno scoprendo tutt'ora) diverse varianti a questa teoria, come nel caso dei retrovirus, che conservano nell'RNA le proprie informazioni, e alcune contraddizioni. Già dai primi anni '90 furono rilevate negli eucarioti le prime discrepanze dal modello, grazie a studi condotti sul verme *Caenorhabditis elegans*, tuttavia le differenze erano considerate solo come casi particolari. Il punto di svolta si ebbe nel 2001, con la scoperta di un gran numero di microRNA, la quale ha permesso di rivoluzionare la concezione dell'RNA, con studi che si sono fatti sempre più frequenti. Un altro esempio

di discostamento dal dogma classico può essere la *metilazione* del DNA - l'aggiunta di un gruppo CH_3 - le cui variazioni producono cambiamenti significativi nell'espressione genica, senza tuttavia causare cambiamenti nella sequenza di codice. Queste problematiche rientrano all'interno di una classe più ampia, quella dell'epigenetica, che tratta appunto delle modificazioni dell'espressione genica prodotte dall'ambiente: in termini più tecnici si può dire che viene variato il *fenotipo* senza che avvengano modificazioni a livello del *genotipo*. Dunque il fenotipo della cellula, non è determinato dal solo codice genetico, quanto piuttosto da una sovrapposizione fra quest'ultimo e un fattore dovuto alle interazioni con l'esterno. Le modifiche epigenetiche durano per tutto l'arco della vita di una cellula e sono ereditabili, sempre senza che la sequenza di DNA venga mutata. Fra i vari meccanismi epigenetici possiamo trovare appunto la metilazione, ma anche modificazioni della cromatina e degli istoni, varianti ai fattori di trascrizione, nonché il silenziamento e il pathway dell'RNA interference, che meritano un discorso a parte, essendo il punto centrale di questo lavoro.

1.3 La RNA interference

La RNA interference è il processo attraverso il quale viene inibita, o silenziata, l'espressione genica. I principali autori di questo processo sono gli RNA brevi, siRNA e miRNA, i quali si legano a dei filamenti di mRNA e possono aumentarne o diminuirne l'attività. La RNA interference ha un ruolo fondamentale nella difesa delle cellule contro i virus, ma anche nel controllo dello sviluppo e nell'espressione genica in generale.

1.3.1 Scoperta della RNA interference

Il primo passo verso la scoperta di questo processo, avvenne quasi per caso con studi sulla pigmentazione della Petunia. In particolare, i ricercatori stavano lavorando sulla produzione di fiori di Petunia con colorazione più intensa: per raggiungere questo scopo, introdussero nelle piantine alcune copie aggiuntive di un gene noto per codificare un enzima che aumentasse colorazione dei petali. Sorprendentemente, il 42% delle piantine così trattate non aveva gli attesi colori vivaci, bensì presentava un colore bianco o sbia-

dito [1]. Nessuno dei transgeni usati come controllo aveva un tale fenotipo. Attraverso un'analisi più precisa, i ricercatori furono in grado di scoprire che sia il gene endogeno che il transgene erano stati soppressi. Per questo motivo, il fenomeno fu inizialmente definito come co-soppressione dell'espressione genica: il meccanismo molecolare, in ogni caso, rimaneva ignoto. Il silenziamento, in questo caso, poteva avvenire sia a livello di trascrizione, chiamato TGS (Transcriptional Gene Silencing), sia dopo, detto invece PT-GS (Post transcriptional Gene Silencing). Nonostante la scoperta di questi meccanismi di silenziamento, la chiave di volta per l'inizio della comprensione di questo processo si ebbe grazie a Fire e Mello nel 1998, che ricevettero, per il loro lavoro, il premio Nobel per la medicina nel 2006. Fu proprio in questo esperimento che venne coniato il termine RNA interference [2]. I due mostrarono i loro risultati in un articolo pubblicato su Nature [3]; il test si proponeva di studiare gli effetti fenotipici dell'inserimento di filamenti di RNA nel verme C. elegans. Le conclusioni a cui i ricercatori arrivarono furono le seguenti:

1. Il silenziamento viene attivato efficientemente quando viene inserito RNA a doppio filamento, mentre viene attivato debolmente o per nulla, dopo l'inserimento di RNA a singolo filamento, sia senso che antisenso (cfr. fig. 1.4).

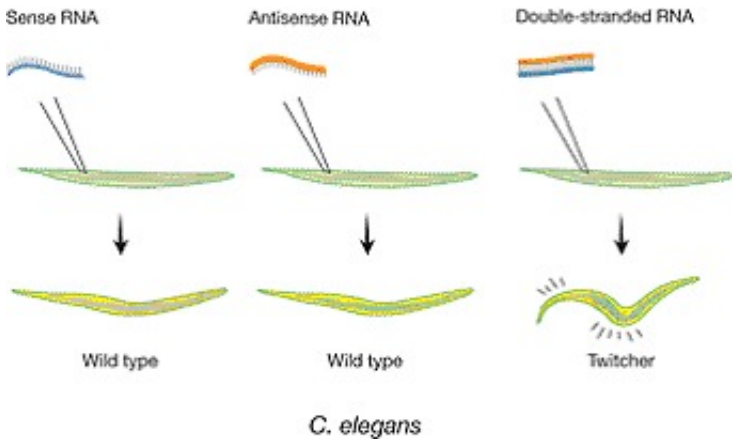

Figura 1.4: Solo il dsRNA produce mutazioni in C. elegans [2]

2. Il silenziamento è specico per mRNA omologo a quello inserito, cioè le basi azotate devono essere le stesse.

3. Il dsRNA deve corrispondere all'mRNA maturo, cioè né gli introni, né i promoter (regioni di DNA che avviano la trascrizione) attivano una risposta; questo indica

un meccanismo PTGS, cioè l'attivazione ha luogo dopo che è avvenuto lo splicing dell'mRNA, e quindi dopo che gli introni sono stati rimossi.

4. L'mRNA bersaglio scompare, segno che esso viene distrutto.

5. Bastano poche molecole di dsRNA per ottenere un silenziamento completo; questo è indice di un processo catalitico di amplificazione, piuttosto che di una reazione stechiometrica.

6. Gli effetti dell'inserimento del dsRNA possono espandersi anche in altri tessuti e alla progenie.

Inoltre tutto ciò costituiva un'analogia con i già visti nelle piante, come nel caso sopracitato della Petunia, e quindi ne forniva ulteriori strumenti di comprensione.

1.3.2 Prima spiegazione del meccanismo della RNAi

Nel giro di pochi anni la RNA interference venne osservata in altri animali, fra cui vermi e moscerini della frutta. Fu proprio con studi riguardanti uno di questi moscerini, il *Drosophila Melanogaster*, che si ricavarono le informazioni più importanti riguardo alla biochimica di questo meccanismo. Per prima cosa si vide che durante il processo di RNAi venivano trovati brevi filamenti di RNA (21-23 nucleotidi) [4]. Si ritiene che questi brevi filamenti, che sono i siRNAs (cfr. sez. 1.1.2), guidino la rottura dell'RNA lungo. In una coltura di cellule di Drosophila *in vitro* si è dimostrato che un vasto complesso, detto RISC (RNA induced silencing complex), viene attaccato all'mRNA attraverso un breve RNA antisenso; dopodiché l'RNA viene spezzettato e successivamente deteriorato (cfr. fig. 1.5). Questo perché nel RISC è sempre presente una proteina della famiglia argonauta, che agisce come una endonucleasi, tagliando l'mRNA. Inoltre anche *Dicer*, un enzima della classe ribonucleasi III, partecipa a funzioni di spezzettamento; si ritiene infatti che questo enzima sia la prima proteina che agisce all'interno del complesso RISC [5]. Nonostante la comprensione di questi fenomeni, c'è ancora molto da fare per comprendere a pieno il pathway della RNA interference. Infatti non è ancora noto di preciso quali enzimi e quali proteine siano essenziali per svolgere un certo tipo di funzione. Non è nemmeno chiaro come gli enzimi agiscano a livello molecolare. Inoltre, nel caso di infezioni virali, non si sa da dove venga esattamente il dsRNA che viene poi processato da Dicer-2 per avviare il

processo di RNAi, cioè non se ne conosce la *biogenesi*. Per tutti questi motivi, negli ultimi anni si sono intensificati sempre di più gli studi relativi a queste proteine e ai complessi che formano, e sono tutt'ora in corso di analisi.

1.3.3 Significatività e processi coinvolti nella RNAi

Fu chiaro da subito che il meccanismo della RNAi poteva avere importantissime implicazioni; di seguito vengono elencate le principali.

1. La RNAi agisce come un sistema immunitario genetico [6]. Da quando si è vista l'equivalenza fra il meccanismo di PTGS nelle piante e quello della RNA interference, sono state effettuate delle analisi più approfondite. Meccanismi di difesa si ritrovano in piante, vermi insetti, e anche alcuni mammiferi, nonostante alcuni sostengano che per questi ultimi le evidenze sperimentali a disposizione sono troppo deboli per poter parlare di sistema immunitario [7]. L'analogia principale tra il sistema immunitario tradizionale e quello "genico" è che entrambi attivano un meccanismo di difesa, che comprende riconoscimento e distruzione, in presenza di un organismo esterno.

2. La RNAi assicura la stabilità del genoma silenziando i *trasposoni*, cioè i geni che cambiano posizione all'interno della sequenza; in questo modo vengono inibite mutazioni genetiche indesiderate. Questo è stato verificato modificando alcuni componenti nel meccanismo di interferenza, sempre con esperimenti su C. elegans e Drosophila: in questo caso i trasposoni vengono riattivati, causando delle mutazioni. La biochimica coinvolta, tuttavia, non è ancora del tutto chiara.

3. Un meccanismo interferenziale opera anche relativamente alla regolazione genica e alla sintesi proteica. Infatti sono strati trovati dei frammenti di RNA brevi endogeni delle cellule, i miRNA (cfr. sez. 1.1.2). Il regolamento dell'espressione avviene tramite appaiamento di basi con l'mRNA, con il risultato di una distruzione di quest'ultimo oppure di una inibizione del processo di traduzione. Si stima che il 30% dei geni sia regolato dai micro RNA, anche se la portata di tale assunzione è ben lungi dall'essere chiara e definitiva.

4. L'interferenza è importante anche a livello trascrizionale e di controllo sulla cromatina. Quest'ultima, nella sua componente più condensata, detta *eterocromatina*,

impedisce la trascrizione, perciò si dice che tutti i geni contenuti in essa sono inattivi. Anche questo meccanismo, a livello molecolare, non è ancora ben compreso.

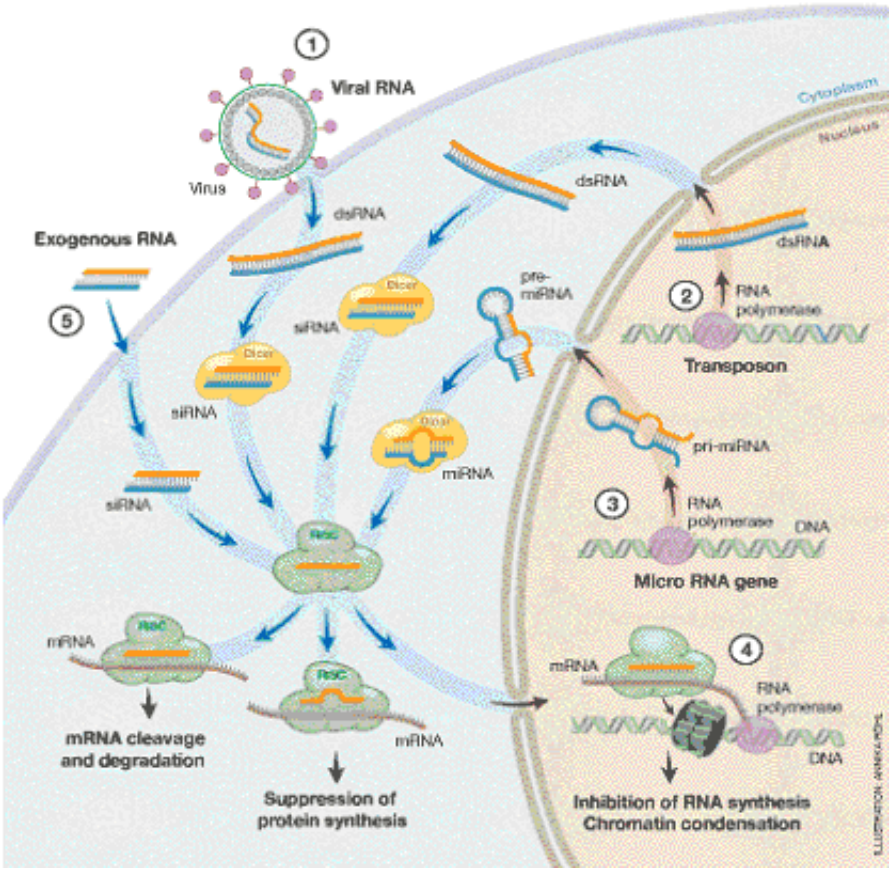

Figura 1.5: meccanismi principali nella RNA interference [2]

5. Da quanto detto finora dovrebbe essere ormai chiaro che la RNAi può essere usata come strumento di modificazione del fenotipo, attraverso la regolazione genetica. La generalità del processo consente di essere ottimisti sull'uso di questa tecnica nei mammiferi e anche negli esseri umani. A proposito di ciò si sono attivate ricerche per un futuro uso terapeutico, ma è ancora troppo presto per poterne prevedere sviluppi e reale applicabilità.

In conclusione questa scoperta è stata indubbiamente rivoluzionaria e porta con sé un potenziale di vastissime applicazioni; quindi non c'è da stupirsi del fatto che stia generan-

do un gran numero di nuove ricerche, e che queste richiedano l'aiuto di esperti di diversi settori. Le ricerche vertono soprattutto sulla comprensione dei processi biochimici coinvolti, che, come già affermato, sono tutt'altro che chiari.

Capitolo 2

Il sequenziamento

In questo capitolo verrà discusso il tema del sequenziamento: dopo aver introdotto la terminologia e i principali problemi (sez. 2.1), come la mole di dati con cui si ha a che fare, si analizzeranno i metodi di sequenziamento di nuova generazione (sez. 2.2). Nella terza sezione verrà data una panoramica delle applicazioni e degli ambiti di utilizzo, con alcuni approfondimenti. Infine, nella sezione 2.4 verrà affrontato il problema dell'analisi dei dati, discutendo le distribuzioni utilizzate, i loro vantaggi e i loro limiti, con un riferimento ai software che permettono l'allineamento e il confronto dei reads.

2.1 Cos'è il sequenziamento

Il sequenziamento è la lettura dell'esatta sequenza dei nucleotidi di un acido nucleico. Un filamento di DNA o RNA umano contiene qualche miliardo di nucleotidi, ma i dispositivi hanno una capacità limitata di dati per ogni lettura, consentendo l'analisi solo per frammenti. Il frammento di acido nucleico che viene letto è denominato *read*, e contiene, con le nuove tecnologie, al massimo un centinaio di nucleotidi; quindi i reads ottenuti vanno allineati e assemblati per formare la sequenza che si vuole conoscere. Il primo metodo utilizzato per il sequenziamento è il metodo Sanger, o metodo a terminazione di catena; tuttavia questo sta cedendo il passo ad altre tecnologie, a causa degli elevati costi e, soprattutto, del basso *throughput* (cioè la quantità di dati ottenuti) e dei tempi enormi che richiede. Con i metodi di sequenziamento di nuova generazione (NGS) il throughput si

è alzato considerevolmente e i costi sono crollati, rendendo accessibile una vasta gamma di studi che sfruttano queste tecnologie. Il principale svantaggio delle nuove tecniche è la ridotta lunghezza dei reads; ciò comporta, nella fase di allineamento e assemblaggio, la generazione di errori, dei quali, seppur presenti in numero piuttosto basso, bisogna tener conto.

Richieste computazionali

A causa delle elevate prestazioni dei metodi NGS anche le richieste computazionali sono aumentate in maniera non indifferente. Si deve considerare, infatti, che ad una base sequenziata possono corrispondere da 8 fino a 16 byte IC (ma anche oltre, a seconda del tipo di analisi) e che una corsa HiSeq 2000 (Illumina) può portare ad una richiesta di alcuni TB di spazio, senza considerare backup o eventuali ridondanze [8]. Per avere un'idea della quantità di dati prodotti da piattaforme NGS, ci si può riferire ai circa 9 petabyte (18x10^50 byte) generati nel 2010 solamente dal Sanger Institute, uno dei maggiori centri per il sequenziamento al mondo [8]. Naturalmente è importante che questi dati siano adeguatamente accessibili sia per la consultazione sia per il processamento, il che comporta l'utilizzo di sistemi di gestione dati e di rete ad alte prestazioni. Infatti un normale calcolatore possiede al massimo memorie dell'ordine dell'unità di Terabyte, quindi per poter gestire e analizzare correttamente i dati è necessario fare uso di *cluster* di computer.

Trattazione statistica

Essendo ormai i metodi di sequenziamento notevolmente avanzati, il problema concettuale e pratico che si riscontra nel sequencing è quello dell'analisi. Dal momento che la lunghezza dei reads, con i metodi NGS, è minore rispetto a quella che si poteva ottenere coi metodi tradizionali (anche se con il metodo Roche 454 la diminuzione è meno evidente [8]), ne risulta più difficile l'assemblaggio, a causa di una mappatura più incerta. Per ottenere un corretto assemblaggio si fa uso di distribuzioni statistiche, considerando i reads come delle variabili aleatorie. Infatti, poiché nel processo di sequenziamento si verificano degli errori (variabilità tecnica), dovuti alle tecniche utilizzate, è importante distinguere le fluttuazioni dovute alla misura da quelle intrinseche del sistema; questo è possibile proprio dal confronto dei dati ottenuti con le distribuzioni statistiche attese. In un processo di sequenziamento *de novo*, per esempio, l'adattamento alla distribuzione aspettata permette

di capire se ci sono dati da rigettare, oppure se è necessario un processo di risequenziamento. Nel caso di un'analisi differenziale, come quella relativa alla RNAi, invece, le distribuzioni rivelano se effettivamente si ha una modificazione nel livello di espressione, oppure se le piccole deviazioni ottenute rispetto ai valori attesi non sono significative.

2.2 Metodi di Next Generation Sequencing

I metodi di nuova generazione consentono di ottenere molti dati grazie a un sequenziamento in parallelo. Nello specifico questi metodi sono in grado di analizzare fino a 600 miliardi di basi (600 Gigabasi) con un ciclo di sequenziamento che dura una decina di giorni, contro 1 milione di basi per ciclo con il metodo tradizionale. Questo si traduce anche in una differenza di costi: 1 dollaro per Mb contro i 10000 dollari necessari nel 2000; in termini più intuitivi si pensi che ora bastano 2000 dollari per sequenziare un intero genoma. I problemi principali, come già accennato, riguardano la minore lunghezza di ciascun read (per motivi che saranno chiari a breve), con conseguenti difficoltà di mappatura. Una strategia per ovviare a questo inconveniente, largamente utilizzata, è il sequenziamento di librerie paired-end o mate-pair che, permettendo di determinare quali sequenze si trovano ad una specifica distanza fra loro, semplificano l'assemblaggio anche in presenza di lunghe porzioni altamente ripetitive. In generale, gli algoritmi di assemblaggio, partendo dalla sovrapponibilità (parziale) delle read sequenziate, operano inizialmente la ricostruzione dei primi assemblati o *contig*, e, successivamente, l'organizzazione di questi ultimi in strutture ordinate e orientate o *scaffold*. Sfruttando, quindi, l'informazione sulla distanza contenuta nelle librerie paired-end o mate-pair, si possono determinare in maniera univoca le posizioni reciproche dei contig all'interno degli scaffold. Diversi metodi NGS utilizzano diversi processi biochimici, esistono tuttavia delle linee guida comuni. Il primo passo consiste nella preparazione della libreria da andare poi a sequenziare; dopodiché si opera un meccanismo amplificazione, in modo da aumentare di parecchio il numero di copie, con la generazione di cluster (cfr. fig. 2.1). È proprio questa operazione a imporre delle limitazioni riguardo alla lunghezza dei reads: infatti ciascuna copia dello stesso frammento deve crescere alla stessa velocità delle altre, in modo tale che non ci siano disparità di lunghezza tra esse, fatto che ne comporterebbe

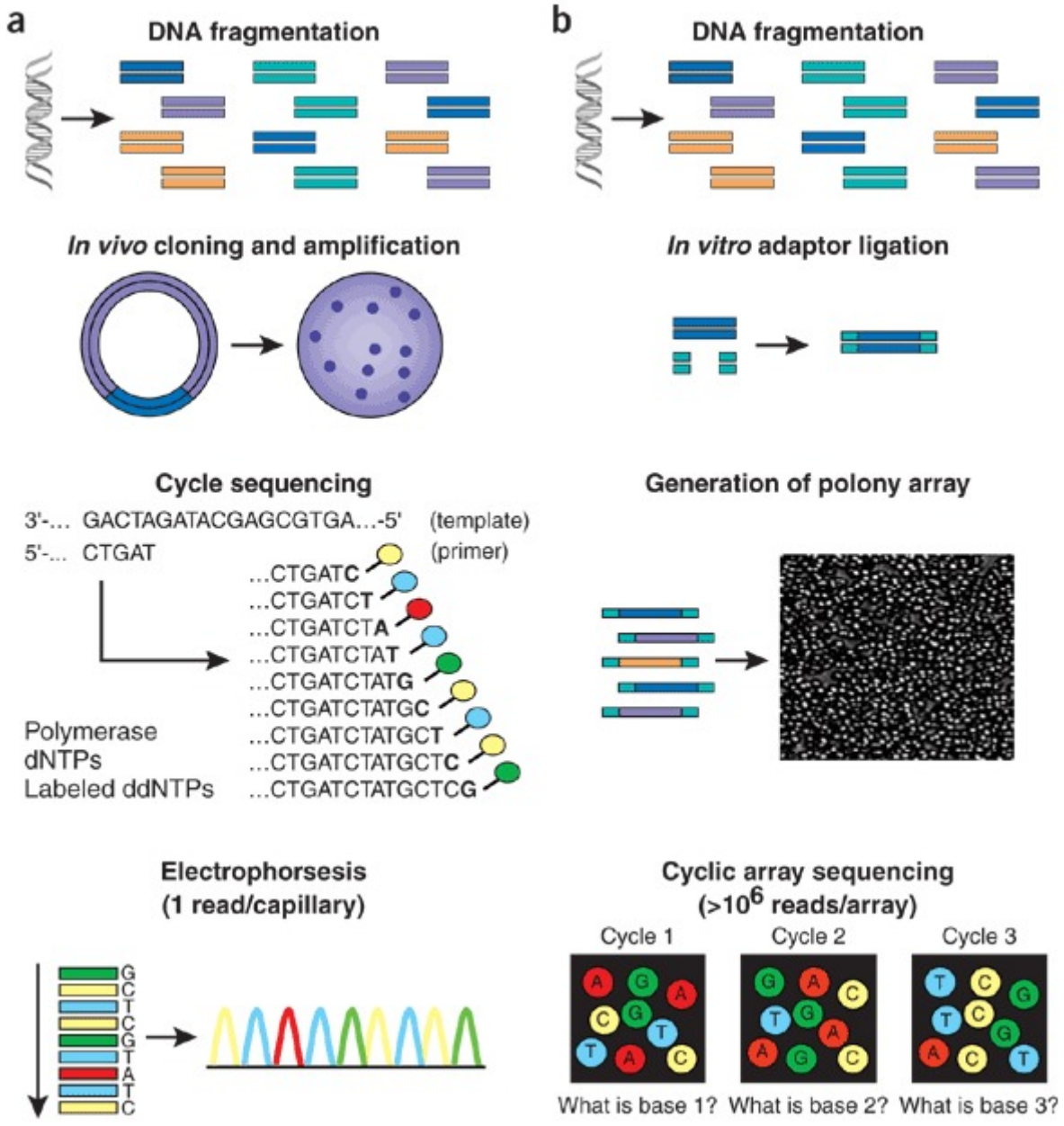

Figura 2.1: flusso di lavoro tipico dei metodi NGS (b) rapportato al metodo Sanger (a) [9]. È possibile notare il processo di creazione di cluster, o polony (polymerase colony)

una diversificazione, introducendo un errore nel sequencing. Infine c'è il processo di sequenziamento vero e proprio, il quale è fatto solitamente via sintesi, e il conseguente immagazzinamento dei dati.

Diamo ora una breve panoramica delle principali tecnologie di next generation sequencing, con una breve analisi finale di vantaggi e svantaggi.

2.2.1 *Genome Analyzer* di Illumina

Questo sequenziatore è stato per la prima volta messo sul mercato nel 2006 da Solexa, acquisita un anno dopo da Illumina. Il metodo su cui si basa è quello a terminazione di catena reversibile e il processo generale può essere suddiviso, come appena visto, in tre fasi: generazione della libreria, preparazione dei cluster e sequencing. Nella prima

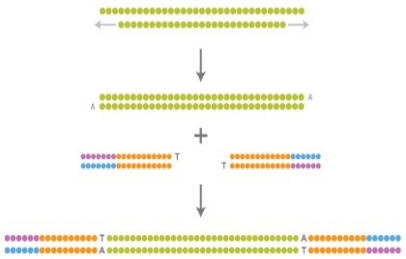

Figura 2.2: fase di preparazione della libreria [10]

fase il DNA viene frammentato e agli estremi di ogni frammento vengono aggiunti degli oligo-adattatori (cfr. fig. 2.2), per il successivo processo di amplificazione. Quest'ultimo è necessario dato che il sequenziamento richiede di avere molte copie dello stesso DNA, in modo da ottenere informazioni statistiche più robuste. La generazione di copie dei frammenti avviene su una *flow-cell,* una piastra su cui sono presenti oligonucleotidi di due diversi tipi. I frammenti di DNA si legano con entrambe le estremità a questi oligonucleotidi e ciascun estremo si lega con l'oligonucleotide ad esso complementare, formando delle strutture a ponte (cfr. fig. 2.3). A questo punto la DNA polimerasi sintetizza i filamenti complementari a quelli presenti, formando così strutture a doppio filamento, che vengono poi denaturate (chimicamente o termicamente). Con la rottura dei legami idrogeno che tengono uniti il doppio filamento, si ottengono nuovamente due filamenti separati, numericamente raddoppiati rispetto allo stato iniziale. La catena di processi appena descritta si ripete un gran numero di volte, fino ad ottenere un cluster di migliaia di frammenti; questi ultimi però contengono sia la copia identica all'originale, sia il filamento inverso: è necessario perciò ripulire il cluster dai frammenti antisenso prima di procedere con l'operazione di sequenziamento vera e propria. L'ultima fase prevede che ai frammenti di

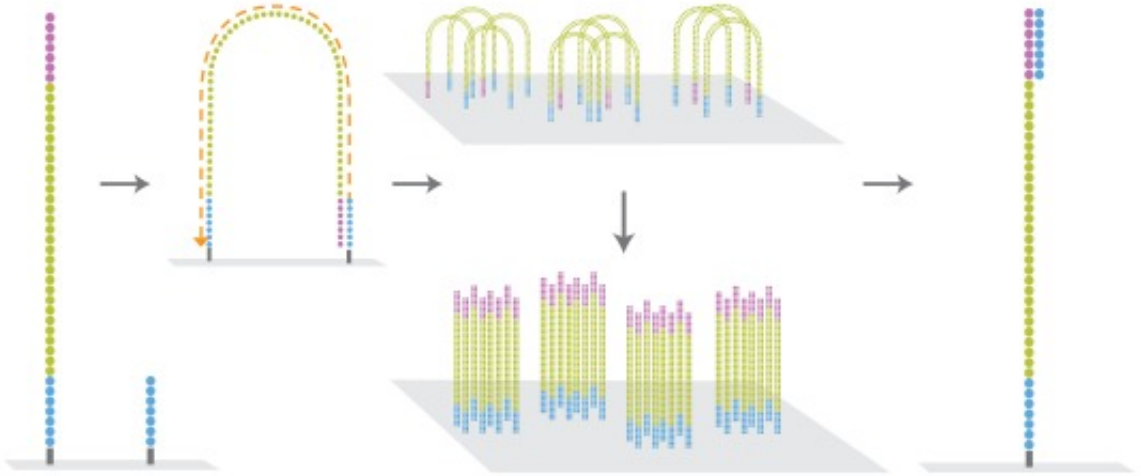

Figura 2.3: fase di amplificazione: è possibile notare le caratteristiche strutture a ponte [10]

ogni cluster venga eseguita la sintesi del *primer* (innesco), il frammento di DNA che dà il via alla reazione di sequenziamento; questa, grazie alla tecnologia può essere effettuata su centinaia di milioni di cluster in parallelo. Ogni ciclo coinvolge una DNA polimerasi e i quattro dNTP (deossi-nucleotidi, usati dalla polimerasi per formare la doppia catena) a cui sono state apportate due modificazioni:l'incorporazione di un marcatore fluorescente e l'aggiunta di un terminatore reversibile. Il marcatore reagisce in maniera diversa per ognuno dei quattro nucleotidi quando sottoposto a onde laser e permette dunque l'identificazione della base che è stata sequenziata. Il terminatore è una molecola che blocca il gruppo ossidrile impedendo l'ulteriore sintesi, in modo da garantire l'incorporazione di una sola base; è detto reversibile in quanto può essere dissociato chimicamente, riattivando la sintesi. Dopo ogni incorporazione, un laser eccita il fluorescente del dNTP generando un'emissione luminosa che ne permette l'identificazione. Dopodiché il terminatore e l'etichetta fluorescente vengono rimossi, in modo da abilitare il sequenziamento della base successiva. Con un solo ciclo i sequenziatori di Illumina sono in grado di leggere fino a 6 miliardi reads in pochi giorni, con un numero di basi per read variabile da 50 a 200 e, dunque, quasi mille Gigabasi in totale.

2.2.2 *454 Pyrosequencing* di Roche

Questo metodo fu il primo tra quelli NGS ad essere disponibile sul mercato (2005, ad opera di Life Sciences, ora di proprietà di Roche Diagnostic). Utilizza il pirosequenziamento, che è una tecnica alternativa a quella classica della terminazione di catena. Infatti, a differenza di quest'ultima, non si basa sull'incorporazione di un deossi-nucleotide che termina la reazione di sintesi, bensì sulla rivelazione di un segnale dovuto all'incorporazione di un nucleotide.
Lo schema del procedimento utilizzato è il seguente: i campioni costituiti da reads molto lunghi vengono divisi in filamenti più corti, di 300-800 basi. Gli adattatori essenziali per l'amplificazione, la purificazione e il sequenziamento vengono aggiunti a entrambe le estremità del filamento e, nel caso questo sia doppio, viene trasformato in filamento singolo. Il processo di amplificazione, in questo caso è una PCR (polymerase chain reaction) in

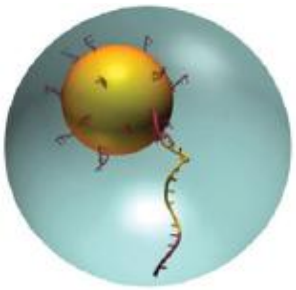

Figura 2.4: sfera pronta per il processo di emPCR [12]

emulsione. Alcuni template[1] di DNA sono posti su sferette, o *beads*, di $28\mu m$, e catturano i frammenti di DNA complementari. Il numero delle sferette deve essere molto grande, in modo tale che ciascuna contenga al massimo un frammento. Le sferette sono poste in un soluzione acquosa contenente reagenti per la polimerasi, e questo mix è a sua volta posto in una soluzione oleosa, che, essendo idrofoba, e quindi non miscibile con quella acquosa, permette la formazione di goccioline d'acqua dette *micelle*. Su ognuna di queste, dunque, si avrà la moltiplicazione dei filamenti (cfr. fig. 2.4). Il processo appena descritto fa in modo che in ogni goccia tutte le copie di DNA siano identiche, e quindi che copie diverse si trovino in gocce diverse. Ciò risulta molto vantaggioso, poiché l'amplificazione può essere fatta in vitro, ma mantenendo separate le reazioni dei vari frammenti. Suc-

[1] I template sono i filamenti di DNA stampo

cessivamente le sferette vengono deposte in una piastra *Pico Titer* costituita da circa 1,6 milioni di pozzetti, che hanno un diametro tale da poter accomodare un'unica gocciolina (cfr. fig. 2.5). A questo punto vengono aggiunte la DNA polimerasi ed altre sferette di dimensioni minori, a cui sono legati covalentemente gli enzimi responsabili della reazione di pirosequenziamento. Per la parte di sequencing la tecnica del pirosequenziamento richiede

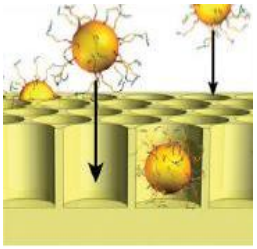

Figura 2.5: le *beads* si posizionano sulla piastra di sequenziamento [12]

l'uso, oltre che di DNA polimerasi, anche di ATP solforilasi, luciferasi e apirasi e di quattro deossinucleosidi trifosfati (dNTPs), cioè nucleotidi con tre gruppi fosfati. La DNA polimerasi viene fatta agire su un singolo filamento di DNA, introducendo solo un tipo di dNTPs: se la base azotata del nucleotide è quella complementare a quella che si ha sul filamento di DNA presente, allora si avrà un emissione luminosa (rilevata da un sensore CCD), altrimenti non si avrà emissione. Questo è spiegabile seguendo la reazione

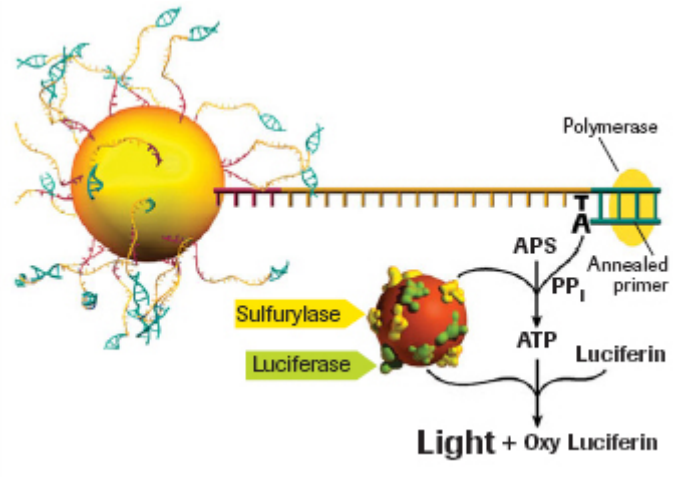

Figura 2.6: reazioni che portano all'emissione luminosa [12]

biochimica che avviene quando la base azotata è complementare a quella del filamento

già presente: si forma un legame fra le due basi e viene rilasciato un pirofosfato in quantità stechiometriche. Dopodiché l'ATP solforasi lo converte in ATP, il quale agisce da carburante per trasformare la luciferina in ossiluciferina. Quest'ultima genera a sua volta luce visibile (cfr. fig. 2.6), in quantità proporzionale all'ATP presente, e quindi anche al numero di basi che sono state aggiunte al filamento. I nucleotidi non incorporati e l'ATP vengono poi degradati dall'enzima apirasi, per poter procedere col sequenziamento delle basi successive. Lo svantaggio del pyrosequencing si ha soprattutto con omopolimeri (cioè sequenze formate dallo stesso tipo di base, p.e. AAAA); infatti, dato che in questo caso basi contigue vengono incorporate contemporaneamente, il loro numero può essere dedotto solamente dall'intensità della luce emessa, che in alcuni casi può essere fuorviante.
Nonostante lo sviluppo degli ultimi anni, come il minor tempo del sequenziamento (bastano 10 ore per ottenere un milione di sequenze) e la diminuzione dei costi (seppur rimangano i più alti fra le tecnologie NGS), tutto ciò non è stato accompagnato da un analogo miglioramento a livello qualitativo. Il margine di errore di questa tecnologia, infatti, ha un valore medio dell'1,07%, e, cosa più importante, il tasso non è equamente distribuito [11], ma si trova per più del 50% in determinate posizioni. I maggiori contributi agli errori sono gli omopolimeri, le dimensioni delle sequenze e la localizzazione delle goccioline nel piatto PT.

2.2.3 *SOLiD* di Applied Biosystems

In questo sistema, come per il 454 della Roche, ai frammenti da sequenziare sono legati degli adattatori necessari per il legame alle sferette; tali frammenti vengono successivamente amplificati mediante PCR in emulsione (cfr. fig. 2.7). Dopo la denaturazione le sferette vengono depositate su un supporto di vetro; la differenza tra quest'ultimo e il piatto TP del metodo 454, consiste nel fatto che in SOLiD non sono presenti delle cellette, quindi l'unica limitazione al numero di sfere per unità di superficie è dato dal loro diametro, che comunque è molto più piccolo (<1 micron) di quello della tecnologia 454. In questo caso il sequenziamento via sintesi è guidato dalla ligasi[2], non dalla DNA polimerasi, come invece avveniva per i due sistemi precedenti: da qui deriva l'acronimo SOLiD (Sequencing by Oligonucleotide Ligation and Detection). Ogni ciclo di sequenzia-

[2]La ligasi è l'enzima che permette l'ibridizzazione, ovvero una reazione biochimica in cui vengono legate due basi che, a processo avvenuto, sono adiacenti nello stesso filamento

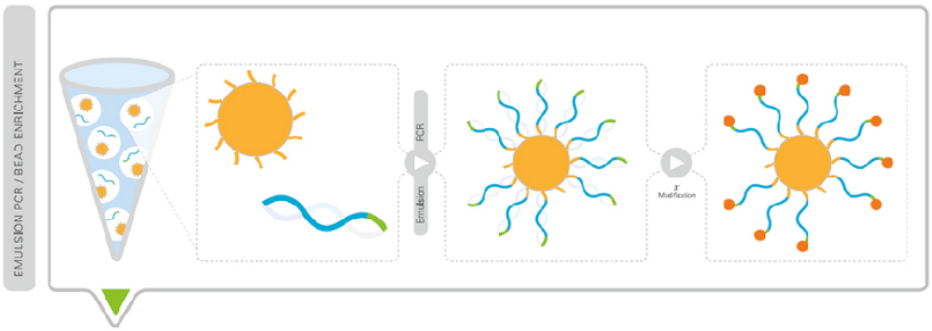

Figura 2.7: PCR in emulsione [13]

mento necessita di una sferetta, un primer degenere[3], una ligasi e quattro sonde di dNTP; queste sono composte da ottameri con due basi fissate e sei degeneri: le ultime tre sono rimovibili. Per prima cosa il primer si ibridizza con la sequenza adattatrice, dopodiché la

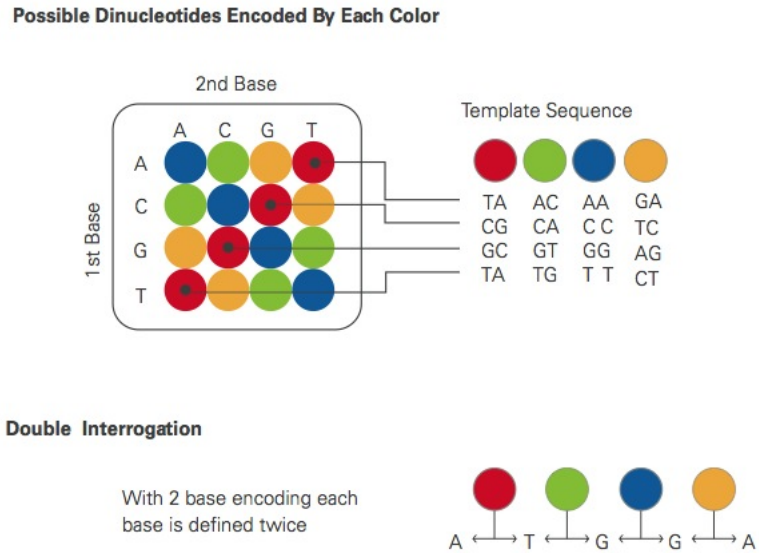

Figura 2.8: esempio di combinazione di colori per le basi [13]

ligasi permette l'ibridazione di una sonda, con conseguente emissione di fluorescenza da parte del marcatore; all'ottamero ligato vengono poi rimosse le ultime tre basi assieme alla marcatura, per le successive ripetizioni. A ciascuna coppia di basi è associato un colore, in modo tale da permetterne l'identificazione, tuttavia la marcatura non è univoca,

[3]Un primer degenere può legarsi a tutti e quattro i tipi di base

dato che i colori utilizzati sono 4 e le coppie di basi possibili 16 (cfr. fig. 2.8). Perché quindi associare un colore a una coppia di basi, invece di associarne uno a ognuna, rendendo così la corrispondenza biunivoca? Anche se la corrispondenza 1-1 sembra il metodo migliore per procedere in realtà non è così, infatti questa può portare più facilmente ad errori di sequenziamento (cfr. fig. 2.9), che in alcuni casi, come nella determinazione delle SNP (cfr. sez. 2.3.1), compromettono inevitabilmente i risultati dell'esperimento. Con l'utilizzo di due basi, invece, si riesce a distinguere facilmente tra errore di sequenziamento e differenza di espressione genetica intrinseca. In ogni caso, considerando che a

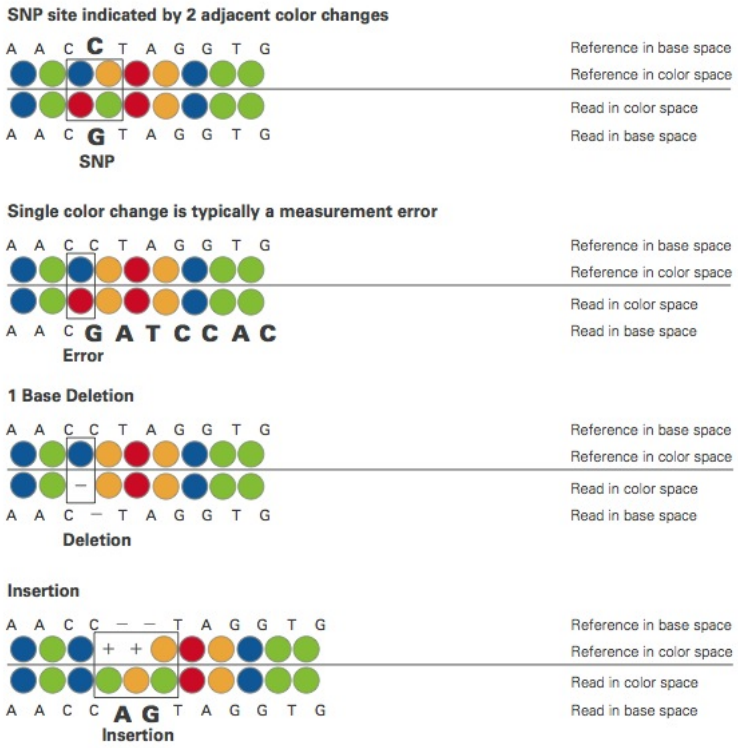

Figura 2.9: la combinazione è utile per rilevare errori di sequenziamento [13]

ogni iterazione riusciamo a sequenziare due nucleotidi ogni cinque, è necessario ripetere il ciclo di sequenziamento cinque volte (cfr. fig. 2.10), cosicché ogni base è univocamente determinata. Man mano che si procede con le reazioni di ligasi, gli ottameri non possono essere aggiunti in grande quantità; di solito si riesce a legarne sette, ma negli ultimi anni si è arrivati anche a dieci. Questo comporta ovviamente una lunghezza delle reads molto ridotta (35-50), ma, allo stesso tempo, una minimizzazione degli errori nella lettura di

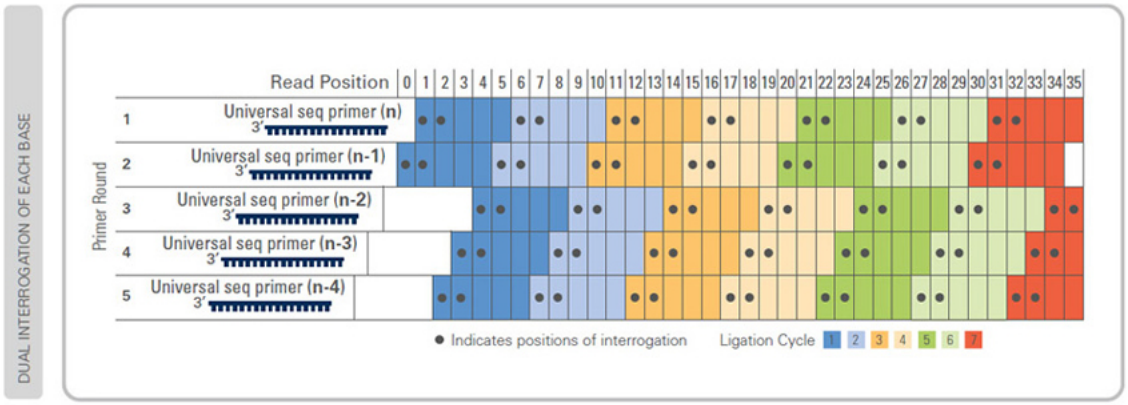

Figura 2.10: rappresentazione delle ripetizioni del ciclo di sequenziamento [13]

ciascun read.

2.2.4 Riassunto delle proprietà e confronto fra le tecnologie

Anche se le prestazioni delle varie tecnologie sono in continua evoluzione, e i dati a disposizione sono perciò molto spesso discordanti, è possibile fare comunque un confronto tra esse. Si notano infatti differenze di costo, accuratezza, lunghezza dei reads. Ciò indica che un metodo può essere più indicato per una determinata applicazione rispetto ad un altro. Tuttavia studi fatti con diverse tecnologie per lo stesso esperimento (per esempio nel caso di microbiomi) indicano che i risultati ottenuti sono gli stessi, entro il limite di errori statistici [14]. In figura 2.11 è riportato un riassunto delle prestazioni dei metodi analizzati precedentemente. Si può vedere come la tecnologia 454 sia quella con un numero

Technology	**Sanger Sequencing**	**Next Generation Sequencing**			
Manufacturer	**Applied Biosystems**	**Roche 454**	**Illumina**	**Life Technologies**	
Model	ABI 3730XL	GS FLX Titanium XL+	HiSeq 2000 dual flow cell	SOLiD 4 System	Ion PGM
Bases per RUN	~ 96 Kb	700 Mb	600 Gb	100 Gb	1 Gb
Time per RUN	2 h	~1 day	~11 days	~14 days	4.5 h
Reads per RUN	96	1 million	6 billions (paired-end)	1.4 billions	5 millions
Reads length	up to 1000 bp	up to 1000 bp (mode 700 bp)	2*100 bp	2*50 bp	up to 400 bp

Figura 2.11: tabella di confronto fra le varie tecnologie NGS [8]

di basi per reads più grande. Questo è particolarmente indicato per applicazioni come il sequenziamento de novo, mentre in altre, ad esempio il resequencing, è una caratteristica meno rilevante. Anche in esperimenti di interazione tra proteine e DNA la lunghezza non riveste grande importanza, mentre ciò che conta realmente è il numero di reads, per ottenere informazioni statistiche più consistenti. Infine bisogna tenere conto dei costi; ovviamente, progetti che necessitano di un gran numero di basi sequenziate, non possono prescindere dalla questione economica, quindi è necessario usare uno strumento che ha un costo per base contenuto. Tuttavia quest'ultimo parametro può risultare fuorviante, in quanto un minore costo per base rischia di tradursi in una minore accuratezza del sequenziamento della base stessa.

2.3 Applicazioni e ambiti di utilizzo del sequencing

Dato il loro rapido sviluppo e il crollo nei costi, i metodi di NGS stanno assumendo sempre maggior importanza non solo nel sequenziamento del DNA, bensì anche in molte altre applicazioni, per le quali finora è stata usata la tecnica dei *microarray*.

I microarray

I microarray di DNA sono dei dispositivi che permettono l'analisi di molti frammenti di genoma in parallelo. Sono costituiti da minuscole sonde di DNA, dette *spots*, poste su un supporto solido; queste sonde sono formate da template, ovvero delle sequenze note, in misure dell'ordine delle picomoli. Gli spots sono disposti a matrice, in modo tale che ogni elemento abbia una sua posizione ben determinata, a cui corrisponde un template noto. I frammenti vengono marcati con un elemento fluorescente, in modo tale da identificare la posizione in cui ciascun frammento si è legato, dunque anche la sua sequenza, dato che è nota quella dei template presenti su ogni cella della matrice. Il difetto principale di questa tecnica è che la matrice può contenere solo un certo numero di sequenze, perciò noi dobbiamo conoscere in anticipo le combinazioni che andiamo ad analizzare, cosa non applicabile in alcuni casi, per esempio nel sequenziamento *de novo*.

I metodi di sequenziamento di nuova generazione offrono una valida alternativa alle

procedure che utilizzano microarray per quanto riguarda molte applicazioni. Tuttavia il passaggio tra le due tecniche è tutt'altro che concluso, anzi, per alcuni ambiti è ancora ai primi passi [15]. Come mai? Nonostante gli obiettivi e i problemi per le due procedure siano molto simili, gli strumenti di analisi devono affrontare problemi tecnici molto diversi; quindi non è automatico che vengano cambiati immediatamente software progettati e collaudati già da diversi anni e dei quali è provata l'ottima funzionalità. Inoltre è parere comune che i microarray siano più facili da utilizzare e meno laboriosi nella preparazione dei campioni, rispetto ai metodi di NGS. A parte ciò, come già detto, l'enorme *troughput* e il costo sempre minore fanno intendere che i microarray verranno soppiantati anche nei campi in cui ancora oggi rimangono predominanti.

2.3.1 Descrizione degli ambiti di utilizzo

I principali ambiti di utilizzo del sequencing sono studi su analisi differenziale e di espressione genica, studi sulla metilazione del DNA, immunoprecipitazione della cromatina etc.; di seguito verranno brevemente discussi.

Sequenziamento *de novo*

Questo tipo di sequenziamento, già nominato nelle precedenti sezioni, si propone di determinare genomi di organismi non ancora sequenziati. Si presta particolarmente ai metodi NGS, dato l'abbattimento dei tempi di lavoro. Basti pensare che queste tecniche hanno bisogno di una decina di giorni per sequenziare un intero genoma, contro i dieci anni impiegati la prima volta col metodo Sanger.

Sequenziamento del trascrittoma (RNA-seq) e dei micro RNA

In questo caso si possono determinare le differenze di espressione nei vari tessuti, o in diverse fasi dello sviluppo cellulare e dell'organismo, per uno stesso tessuto. Si può studiare inoltre la variazione in differenti condizioni ambientali, legata ad attività di regolazione genica, come nella RNA interference o, più in generale, in molti studi di epigenetica. L'esperimento che sarà analizzato nel prossimo capitolo rientra proprio in questo ambito.

Resequencing

Può accadere che alcuni studi necessitino il risequenziamento di alcuni genomi, per esempio per identificare eventuali modificazioni, come le SNP (polimorfismo di singolo

nucleotide), o variazioni nel numero di copie espresse di un gene. Questo è utile per comprendere i meccanismi di sviluppo di alcune patologie e al fine di poterne individuare le direzioni per un futuro trattamento clinico.

Misure di biodiversità

Le moderne tecniche hanno permesso l'avvento su larga scala della metagenomica, cioè lo studio di comunità microbiche direttamente nel loro ambiente naturale, evitando così il problema del prelevamento e della coltivazione in laboratorio. Infatti, in molti casi, risulta difficile riprodurre le reali condizioni di un determinato ambiente e, molto spesso, alcuni organismi hanno particolari esigenze (temperature elevatissime, pressioni pari a quelle dei fondali oceanici, concentrazioni saline alte, etc). In questo caso ciò che viene sequenziato è l'RNA ribosomiale, poiché fra i vari tipi di RNA è quello che viene conservato più a lungo e quindi permette di risalire all'ordine tassonomico di appartenenza di un dato organismo, di ricostruire la diversificazione fra le specie, stimarne il tasso di divergenza e riconoscere gruppi correlati.

Durante il sequenziamento si ottengono campioni provenienti da diversi organismi, però questo non impedisce di riuscire a isolare i dati provenienti dal genoma di interesse. Ad esempio, cercando un particolare microrganismo produttore di petrolio, occorre un carotaggio in un terreno dove è noto trovarsi un giacimento. Si estrae il DNA che è presente in tutto il campione, e si avvia il sequenziamento. Saranno presenti genomi di svariati organismi (tra cui anche insetti e altre forme di vita), ma ciò che conta non è tanto andare a sequenziare un genoma in particolare, bensì soltanto identificare quel particolare gene produttore di petrolio. Questa tecnica ovviamente si può applicare anche per la ricerca di un nuovo antibiotico, di produttori di metano, etc. Un altro tipo di uso è la caratterizzazione delle comunità microbiche umane a fini clinici; per esempio si è studiata la variabilità biologica di virus patogeni come HIV e HCV.

Uno dei più colossali studi di metagenomica è stato fatto da Craig Venter, famoso biologo americano[4] che, insieme ai colleghi, ha analizzato un'intera popolazione marina [16], per un totale di un milione di miliardi di paia di basi sequenziate. I risultati di questo studio comprendono 1800 diversi genomi, tra cui la scoperta di 148 tipi di batteri, fino ad allora sconosciuti. Dunque ne è derivata un'importante svolta nelle conoscenze della

[4]Venter è noto per imprese a grande riscontro mediatico, come la sfida al Progetto Genoma Umano; questa è pero considerata da tanti come uno strumento molto più commerciale che scientifico

biodiversità marina, ma, soprattutto, una comprensione delle enormi potenzialità che i metodi di sequencing hanno in questo ambito.
Importante è sicuramente anche lo studio del microbioma intestinale dei mammiferi, inclusi gli esseri umani. Risultati di questi esperimenti [17],[18] hanno fornito indicazioni di influenza della dieta dei mammiferi sulle varie comunità batteriche, e viceversa, permettendo di ottenere importanti informazioni a favore di alcune teorie ecologiche ed evoluzionistiche.

Sequenziamento ad immunoprecipitazione della cromatina
Questa tecnica viene utilizzata per studiare l'interazione tra DNA e proteine ad attività regolativa. Infatti l'immunoprecipitazione permette di individuare i punti di DNA a cui si legano i fattori di trascrizione o gli istoni, o proteine che controllano la replicazione del DNA stesso. Un successivo sequenziamento di queste zone permette di determinare cosa succede al variare delle condizioni esterne, e cosa queste determinano nell'espressione del fenotipo. Ciò può essere di grande aiuto per l'identificazione e la comprensione di alcune malattie.

2.3.2 Il passaggio da microarray a NGS

Come già accennato sopra, i metodi di sequenziamento di nuova generazione stanno soppiantando i microarray in molti ambiti. La situazione fra le varie applicazioni, però, è piuttosto eterogenea, quindi è necessario analizzare caso per caso lo stato dell'arte.
Gli esperimenti ChIP sono stati i primi a effettuare la transizione da microarray (ChIP-chip) al sequenziamento (ChIP-Seq), dal momento che le tecnologie NGS forniscono una risoluzione di picco decisamente più elevata [15].
Per quanto riguarda il livello di espressione genica il passaggio da array a sequencing è molto vantaggioso. Infatti nel primo caso si ha un *design bias*, cioè un errore sistematico dovuto al metodo sperimentale: questo è causato dal fatto che le sonde sono state progettate solo per alcune regioni, e quindi riceveranno segnali solo da queste ultime, con una conseguente perdita di informazione. Perciò gli array mancano di generalità, dal momento che sono adatti solo a database per i quali sono stati progettati.
Nel caso di genotipi, siano essi sequenziati per la prima volta o sottoposti a resequencing, i microarray sono ancora molto utilizzati, poiché in media meno costosi e più facili da

utilizzare. Infatti per questo tipo di analisi, al contrario che per l'RNA-seq, sono meno soggetti ad errori, rispetto alle nuove tecnologie. Questo è vero per esempio nella rilevazione di una SNP, anche se i microarray, contenendo un numero limitato di stampi, sono efficienti nel riconoscere solo le mutazioni più comuni, ma non possono fare nulla per rilevare modificazioni non ancora note.
Per quanto riguarda le analisi di metilazione molto spesso i due metodi vengono usati in maniera combinata: le NGS vengono impiegate per nuove scoperte, mentre i microarray per avere un rapido profilo di espressione [15].
Le nuove tecnologie sono più lente a subentrare anche in ambito clinico, dal momento che i medici sono restii a sperimentare nuovi strumenti di cui non conoscono l'affidabilità.
Nonostante ciò queste tecniche stanno evolvendo di anno in anno e non è troppo azzardato pensare che presto subentreranno ai microarray in quasi tutti gli ambiti di utilizzo.

2.4 Analisi dati

Come già detto le procedure di acquisizione dati ed analisi risultano piuttosto complesse. Per prima cosa i vari reads vanno allineati, dopodiché viene eseguita l'analisi statistica per avere un'informazione sul profilo di espressione. Per fare ciò vengono utilizzati modelli statistici e software programmati appositamente. I dati ottenuti dal sequenziamento sono detti *count* e rappresentano il numero di volte che ciascun read (o gene, o trascritto) viene letto: va da sé che il valore dei count rappresenta una misura del livello di espressione del gene. I count vengono considerati come delle variabili aleatorie discrete, e perciò seguono opportune distribuzioni statistiche. Per esempio, in un'analisi differenziale del profilo di espressione (cfr. sez. 2.3.1), dobbiamo confrontare i count rilevati per un dato read con i dati di riferimento; ovviamente il numero di count non sarà mai perfettamente uguale nei due casi, a causa della variabilità. Quindi bisogna innanzitutto normalizzare i dati ottenuti rispetto al numero di count totali, e chiedersi se la variazione è da considerare significativa in riferimento alla propria distribuzione, oppure no. In pratica ciò che si deve fare è un test di ipotesi: ci si chiede se la probabilità che la differenza dai dati di riferimento sia quella effettivamente ottenuta è maggiore o minore di un dato valore, di solito fissato al 5%. La risposta a questo test consente di affermare se c'è effettivamente

differenza di espressione, o se le diversità osservate non sono significative dal punto di vista statistico. Metodi simili vengono utilizzati anche negli altri ambiti di applicazione del sequenziamento.

2.4.1 Modelli di analisi dati

Le distribuzioni principalmente utilizzate sono quella di Poisson e la binomiale negativa; la prima, seppur più facile da trattare matematicamente, risulta in molti casi troppo grezza, poiché la varianza coincide sempre con la media. Quando si ha a che fare con dati *sovradispersi* - in cui cioè la varianza è maggiore della media - si fa uso della binomiale negativa, che, avendo due parametri, consente una maggiore flessibilità. La sovradispersione dei dati è conseguenza della loro variabilità; questa può essere tecnica, cioè dipendente dagli strumenti e quindi ridotta migliorando le tecnologie, oppure biologica. Quest'ultima è dovuta al fatto che diversi individui, ma anche cellule differenti di uno stesso individuo, possono presentare un diverso livello di espressione. La variabilità biologica, ovviamente, non può essere ridotta in quanto è proprietà intrinseca del sistema. Ora verranno analizzate più nel dettaglio le due distribuzioni: ci si riferirà ai trascritti, in quanto caso di maggiore interesse in questo lavoro, ma la discussione è generalizzabile senza la necessità di modifiche sostanziali.

Poisson

Siano F l'insieme dei trascritti del campione sequenziato e $f \in F$ un particolare trascritto; definiamo poi l_f la lunghezza del trascritto f, cioè il numero di nucleotidi che lo compone e k_f il numero di copie di ogni read.

Quindi il totale dei nucleotidi sequenziati sarà $\sum_{f \in F} k_f l_f$.

La probabilità che un read provenga dallo specifico trascritto f è

$$p_f = \frac{k_f l_f}{\sum_{f \in F} k_f l_f}$$

Definendo poi l'abbondanza relativa del trascritto $\beta_f = \frac{k_f}{\sum_{f \in F} k_f l_f}$, possiamo riscrivere la probabilità come $p_f = \beta_f l_f$.

Chiamiamo inoltre la dimensione della libreria, cioè il numero di reads che compongono il campione con n. Sotto ipotesi di campionamento casuale, quindi, per uno specifico trascritto f, il numero dei reads mappati è descrivibile da una variabile aleatoria X_f, la quale segue una distribuzione binomiale di parametri n e p_f.

$$B(n, p_f) = \binom{n}{k} {p_f}^k (1 - p_f)^{n-k} \tag{2.1}$$

(In questo caso il successo è rappresentato dal risultato "il read appartiene al trascritto f", mentre il fallimento dal risultato "il read non appartiene al trascritto f") Tipicamente $p_f << 1$ e $n >> 1$ quindi, come noto, la binomiale può essere bene approssimata da una distribuzione di Poisson con parametro $\mu_f = p_f n$.

$$P(\mu_f) = e^{-\mu_f} \frac{{\mu_f}^n}{n!} \tag{2.2}$$

Il parametro μ_f identifica sia la media, sia la varianza della distribuzione. Da qui è possibile calcolare il valore di aspettazione[5] e la varianza.

$$E[X_f] = \sum_{s=1}^{n} s e^{-\mu_f} \frac{{\mu_f}^s}{s!} = e^{-\mu_f} \mu_f \sum_{s=1}^{n} \frac{{\mu_f}^{s-1}}{(s-1)!} = \mu_f \tag{2.3}$$

Omettiamo il calcolo della varianza, poiché, come già anticipato, questa coincide con la media.

In conclusione questa distribuzione è ampiamente utilizzata grazie alla sua semplicità, inoltre è dimostrato che rappresenta bene i dati dai quali si è riuscito a eliminare il rumore; tuttavia è poco adatta a trattare dati con una variabilità biologica elevata.

Binomiale negativa

La distribuzione binomiale negativa descrive la probabilità di ottenere y insuccessi prima del successo k-esimo (cfr. Appendice A), su un totale di $y + k$ prove. È una distribuzione di probabilità a due parametri, che indichiamo con p e y. Secondo questa notazione la distribuzione è

[5] Nel calcolo il limite per n molto grande ci consente di considerarlo infinito e quindi di fare lo sviluppo in serie dell'esponenziale

$$NB(y,p) = \binom{-y}{k} p^k (p-1)^y \tag{2.4}$$

dove rappresenta il numero di successi e p la probabilità di successo in una singola prova. A livello intuitivo possiamo interpretare la distribuzione come "probabilità che, nel momento in cui viene rilevato un numero k di reads del trascritto f, siano rilevati anche y reads di altri trascritti". Nel nostro caso, mantenendo la notazione usata per la distribuzione di Poisson, e definendo in aggiunta la dispersione[6] come $\Phi_f = \frac{{\sigma_f}^2 - \mu_f}{{\mu_f}^2}$, la variabile Y_f che descrive il trascritto f ha la forma

$$Y_f = B(\mu_f, \Phi_f) \tag{2.5}$$

dove la media è $\mu_f = p_f n$ e la varianza è $\sigma^2 = \mu_f(1 + \mu_f \Phi_f)$. Si può vedere che questo modello ricade in quello di Poisson per $\Phi_f = 0$. Il vantaggio nell'uso della binomiale negativa sta appunto nell'utilizzo della dispersione, che permette di tener conto della variabilità biologica dei dati.

2.4.2 Software di analisi dati

Nonostante gli obiettivi ed i metodi di analisi che sfruttano NGS siano abbastanza simili a quelli che usano microarray (limma, SMA, ema), i particolari sono piuttosto differenti e ciò richiede nuovi algoritmi di calcolo e nuovi software: in un'analisi di espressione genica la relativa abbondanza, che per un esperimento con microarray è determinata da un'intensità luminosa, con i nuovi metodi è data da un numero di count, quindi è una misura digitale. A causa di ciò, inizialmente, per procedure come quelle di RNA-seq e ChIP-seq, gli sperimentatori dovevano provvedere da soli alla costruzione di software che permettessero l'analisi dei dati. Con il crescente aumento di queste attività, però, sono stati sviluppati commercialmente alcuni software appositi: questi permettono un'adeguata analisi, tenendo conto delle distribuzioni a cui devono adattarsi i dati e facilitandone l'organizzazione. Non bisogna poi dimenticare il passo precedente, ovvero quello dell'allineamento; infatti l'operazione di assemblaggio per centinaia di milioni di reads

[6]Si fa notare che la dispersione è una misura dello scostamento della varianza rispetto al valore che avrebbe nella corrispondente distribuzione di Poisson

del tipico genoma di un mammifero risulta un compito molto oneroso. Molti pacchetti di analisi sono stati elaborati all'interno del software libero *R*, che è l'ambiente di sviluppo di analisi statistica dei dati maggiormente usato in biologia e bioinformatica, dal momento che contiene molte funzioni apposite, ed è più semplice da utilizzare rispetto a un linguaggio di programmazione di base. Di seguito verranno forniti due esempi di pacchetti progettati per l'analisi, che però hanno il limite di fornire il confronto solo fra due profili di espressione, e di un software che consente l'allineamento dei reads.

DEGseq

DEGseq è un pacchetto R, sviluppato da Wang e collaboratori [19], per l'identificazione dei geni differenzialmente espressi, in riferimento a dati di RNAseq. L'input di DEGseq sono i reads di un esperimento di RNA-seq mappati, con una annotazione del valore di espressione del gene corrispondente. L'output consiste in un file di testo con i valori di espressione e il p-value, e in una pagina XHTML contenente grafici di riassunto dell'analisi. L'analisi differenziale implementata da DEGseq utilizza il modello di Poisson, e come test di ipotesi fa riferimento al test esatto di Fisher e al test del rapporto di verosimiglianza. Si basa sullo strumento MAplot, utile nella diagnostica della qualità e per la necessità di normalizzare i dati; è già stato ampiamente impiegato anche nell'analisi di dati provenienti da microarray. Oltre ai metodi di test riferiti a campionamento casuale, MAplot permette anche di trattare statisticamente repliche tecniche; queste possono dare stime riguardo la variabilità dovuta a diversi macchinari o diverse piattaforme. Anche sulle repliche possono essere effettuati test di ipotesi e confronti nel profilo di espressione; in questo caso si implementa il metodo R samr. Poiché con questi metodi la variabilità tecnica è ridotta al minimo, la distribuzione di Poisson risulta adatta a descrivere l'andamento dei reads.

edgeR

edgeR (empirical analysis of DGE in R) è un pacchetto che fa parte del progetto Bioconductor[7] e che implementa il metodo di Robinson [20] per la quantificazione digitale di dati di espressione genica. Perciò è applicabile a qualsiasi studio in cui l'abbondanza dei trascritti è misurata in termini di count, cioè di dati digitali. Il pacchetto implementa

[7]Bioconductor è un software open-source per l'analisi di dati di sequenziamento high-throughput, che utilizza il linguaggio di programmazione statistica R

metodi statistici per il confronto fra due profili di espressione genica e utilizza il modello binomiale negativo per l'analisi dei dati. Può quindi risultare molto utile quando non si hanno a disposizione molte repliche, caso in cui c'è maggior rumore di fondo, con conseguente aumento della sovradispersione dei dati. La dispersione è stimata tramite il metodo di massima verosimiglianza condizionata e una funzione di Bayes empirica restringe la dispersione verso un valore fissato. La differenza di espressione viene trattata tramite una versione del test esatto di Fisher riarrangiata per dati sovradispersi.

Bowtie

Per quanto riguarda l'allineamento dei reads si menziona il software Bowtie, poiché esso è stato usato nell'esperimento di RNAi trattato nel prossimo capitolo. Questo short-read aligner è in grado di assemblare reads da 35 paia di basi a una velocità di 25 milioni di reads all'ora [21]. È in grado di preparare anche i reads per l'uso in software in step successivi, cioè, in pratica, per l'analisi dei dati. Data la sua velocità, Bowtie può analizzare solo frammenti piuttosto corti, quindi non si presta all'utilizzo per quanto concerne molti progetti, anche se rimane comunque molto potente in esperimenti che coinvolgono short reads.

Capitolo 3

Analisi di dati sull'azione di Dicer-2 in Drosophila

Una comprensione adeguata dei meccanismi con cui la RNA interference (cfr. sez. 1.3) opera durante un'infezione richiama particolare interesse non solo dal punto di vista scientifico e biologico, ma anche da quello medico, poiché potrebbe aprire la strada a un uso del processo di silenziamento per curare infezioni virali negli esseri umani. In questo capitolo verrà analizzato un esperimento di RNA interference condotto sul moscerino della frutta *Drosophila Melanogaster*. I dati, che dovranno essere analizzati al Dipartimento di Fisica dell'Università di Bologna, provengono da un gruppo che fa capo all'Istituto di Biologia molecolare e cellulare dell'Univeristà di Strasburgo, e alla facoltà di Scienze della Vita, sempre della stessa università. Sicuramente degno di nota è il fatto che il direttore di questo esperimento è Jules Hoffmann[1], premio Nobel per la Medicina nel 2011. Il nostro lavoro consiste nel proporre e discutere le metodologie per analizzare i dati prodotti dal suo gruppo di lavoro, in modo da capire se le distribuzioni ottenute possono condurre a qualche risultato statisticamente significativo. Nella prima sezione verrà descritto più nel dettaglio l'ambiente biologico in cui si opera, con particolare dettaglio alle azioni di Dicer-2 e ai complessi ad esso collegati. Poi (sez. 3.2) si analizzerà l'esperimento in sé:

[1] Jules Hoffmann (Echternach, 2 agosto 1941) è un immunologo francese, direttore di ricerca e membro del consiglio di amministrazione del CNRS. Le sue scoperte, rivelando alcuni fattori essenziali delle interazioni fra l'organismo e il suo ambiente, sono importanti per il futuro delle ricerche sul sistema immunitario, tanto che gli sono valse il Premio Balzan con studi riguardanti l'immunità innata nel 2007 e il premio Nobel (condiviso) per la medicina nel 2011.

verranno discussi i vari passi, le conclusioni e le domande che il gruppo di Strasburgo si è posto. Nella sezione 3.3 spiegheremo come rispondere a queste domande, cioè forniremo i metodi e i modelli utilizzabili, infine, discuteremo le possibili conclusioni che se ne possono trarre e le domande ancora irrisolte.

3.1 Meccanismo di RNAi in Drosophila

Vari studi condotti su piante e animali hanno dimostrato che la RNAi svolge un ruolo fondamentale nel processo di difesa immunitaria innata in risposta a infezioni virali. Molte delle conoscenze sul meccanismo sono state ricavate con il moscerino della frutta *Drosophila Melanogaster* (cfr. fig. 3.1); Drosophila è un insetto che si presta molto bene ad analisi genetiche, infatti è stato usato spesso in passato per la sua capacità di adattamento a diverse condizioni sperimentali e la relativa semplicità del suo genoma. Di conseguenza è comprensibile che i principali studi di RNAi vengano effettuati su questo insetto.

Figura 3.1: Drosophila Melanogaster

Il sequenziamento viene usato per studiare le risposte immunitarie negli esseri viventi ed è effettuato modificando geneticamente gli organismi in esame, per verificare quale enzima o proteina determina un certo tipo di funzione o risposta. Gli enzimi principali che catalizzano la reazione di silenziamento sono *Dicer-1* e *Dicer-2*, della classe ribonucleasi III. Nel caso di Drosophila, il primo si è dimostrato avere un ruolo predominante nella regolazione genica, quindi riguardante miRNA (endogeni), mentre il secondo agisce in risposta a fattori esogeni, come nel caso di infezioni da virus, formando i siRNA [23]; Dicer-2 riveste

quindi un ruolo centrale nel nostro lavoro. Per capire la funzione che questo ha nel processo in esame vengono operate delle modificazioni genetiche al moscerino, togliendo o variando alcuni geni che codificano per l'enzima Dicer-2. Il *pathway* biologico della RNAinterfer-

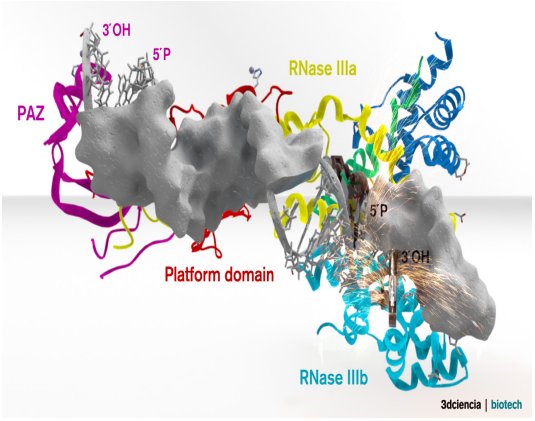

Figura 3.2: Struttura di Dicer

ence comincia quando Dicer-2, in associazione con un cofattore della proteina di legame dsRBP, *Loquacious* (Loqs-PD)[2], spezzetta i filamenti a doppia elica di RNA del virus in segmenti da 21 nucleotidi, gli short interfering RNA virali (vsiRNA); agisce perciò come un "righello molecolare". Il taglio, però, non avviene in maniera simmetrica per i due filamenti: ciò significa che si avrà la sporgenza di uno dei due, la quale, in condizioni normali, è di due nucleotidi all'estremità 3'. Questi siRNA vengono poi caricati, con un altro cofattore detto *R2D2*, su una proteina, *Argonauta-2* (Ago2), che ne permette il trasporto, per andare a silenziare l'espressione dei dsRNA complementari ai siRNA. Per fare ciò Ago2 rigetta uno dei due filamenti di RNA, quello detto *senso*, generando il RISC (cfr. sez. 1.3), che contiene quindi solo il siRNA guida (antisenso), o template. Quest'ultimo è quindi in grado di legarsi ad un RNA a singolo filamento, complementare ad esso, spezzarlo e poi distruggerlo. Si è visto [24] che Dicer è in grado di produrre siRNA anche in assenza di R2D2, ma, senza quest'ultimo, i siRNA non possono essere caricati su Ago2 e quindi perdono la loro funzionalità. Ciò può essere rilevato considerando che il caricamento su Ago2 provoca un arricchimento di citosina all'estremità 5', che, invece, non viene osservato quando si hanno mutazioni a carico di R2D2. Nonostante l'esistenza di una spiegazione relativamente chiara per quanto riguarda questa catena di processi,

[2]In realtà è stato dimostrato che Loqs-PD non è sempre necessario alla biogenesi dei vsiRNA [22]

mutazioni al dominio[3] Elicasi danno effetti non ancora compresi e, inoltre, rimane nebulosa l'origine dei dsRNA virali silenziati da Dicer, e cosa permetta a quest'ultimo di riconoscerli. Uno studio di Marques [22] afferma che possono essere generati durante la trascrizione, o come intermediari genoma-antigenoma durante la replicazione. Dice, inoltre, che la biogenesi dei siRNA virali è distinta da fonti esogene o endogene di siRNA e propone un modello in cui i trascritti virali sono i bersagli principali del silenziamento portato avanti da Ago2, che diventerebbe quindi l'interprete del riconoscimento dell'RNA del virus.

3.2 Esperimento

Per capire le varie funzioni di Dicer-2 si è studiata la risposta dei moscerini Drosophila Melanogaster a un'infezione con virus Sindbis (SINV), e, successivamente, sono stati ordinati i vsiRNA sequenziati. Sono state apportate delle modificazioni genetiche ad alcuni domini di Dicer e si è visto che le mutazioni genetiche più interessanti hanno luogo all'interno del dominio Elicasi. Infatti uno studio di Cenik e collaboratori [24], ha dimostrato che questo dominio lega l'idrolisi dell'ATP al trasporto di Dicer lungo il filamento; in altre parole, affinché Dicer riesca a spezzettare tutto il filamento di RNA, ci vuole ATP, che funge quindi da carburante, e viene prodotto grazie al dominio Elicasi. Modifiche a quest'ultimo permettono a Dicer di produrre dei siRNA, ma solo di un tipo: riescono cioè a spezzare solo un pezzo di RNA, il primo all'estremità 5', dopodiché cadono dal template e non sono più in grado di processare la restante parte di filamento. Questo fatto è stato testato in vitro, dove Dicer-2 ha un allele che porta una mutazione puntuale nel dominio: come già detto alcuni siRNA vengono prodotti, ma solo i primi all'estremità, dato che poi Dicer non è in grado di spostarsi lungo il filamento. Dunque molti processi di silenziamento vengono persi; in vivo si può notare che moscerini che normalmente dovrebbero avere occhi bianchi in realtà presentano un colore rosso: questo perché la modificazione all'enzima fa sì che un dsRNA lungo, detto $white^{IR}$, non venga più processato e quindi nemmeno i siRNA che dovrebbero silenziare l'espressione di tale gene, responsabile della pigmentazione rossa degli occhi. Un'altra funzione fondamentale

[3]Si può immaginare il dominio come un "pezzo" della proteina

dell'Elicasi è il riconoscimento di estremità anomale nei siRNA, come sporgenze 5' o assenza di sporgenze. Nell'esperimento si è cercato di chiarire il ruolo che questo dominio ha nel pathway di Dicer-2, e che cambiamenti determina alla processività dell'enzima.

3.2.1 Metodi sperimentali

Per comprendere il ruolo di questo dominio e il comportamento degli enzimi è necessario effettuare diverse modifiche genetiche a Dicer; occorre definire quindi i genotipi che sono stati utilizzati e come sono stati trattati.

Transgeni utilizzati

Sono stati impiegati i seguenti tipi di moscerini:
1) w^{IR}; +/+
2) w^{IR}; +/Df
3) w^{IR}; +/Df,Rescue
4) w^{IR};dcr-2L811fsx/Df (detto anche w^{IR};dcr-2null)
5) w^{IR};dcr-2L811fsx/Df,Rescue
6) w^{IR};dcr-2G31R/Df (detto w^{IR};dcr-2G31R)
7) w^{IR};dcr-2G31R/Df,Rescue (detto w^{IR};CantonS o wild-type)

- w^{IR} indica il gene $white^{IR}$, dove IR sta per Inverted Repeat.
- + indica un cromosoma *wild-type*, ovvero il genotipo naturale.
- Df sta per Deficiency e indica la mancanza di Dicer nel secondo cromosoma.
- *Rescue* significa che una copia dell'enzima Dicer è stata espressa in maniera esogena per sopperire alla mancanza del gene cancellato.

Dunque i primi tre e il settimo sono i genotipi usati come controlli, cioè per verificare se gli altri tipi presentano differenze di espressione rispetto a questi. Il quarto è il genotipo in cui Dicer è stato soppresso e il quinto la sua versione "rimediata". Infine il sesto è il mutante più interessante dal punto di vista biomolecolare, infatti è il genotipo con la mutazione puntuale (sulla glicina) nel dominio Elicasi.

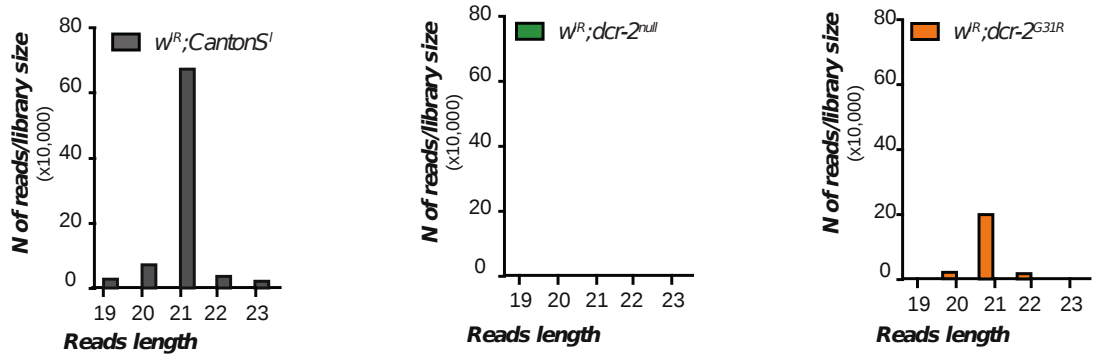

Figura 3.3: Lunghezza dei reads allineati sul transgene per i tre genotipi

Coltura e sequenziamento

Ciascun tipo sopra elencato è stato monitorato per 20 giorni a partire dall'infezione virale, in modo da effettuare un'analisi di sopravvivenza. Inoltre, dopo cinque giorni dall'infezione, gli RNA brevi sono stati selezionati (19-24 nt), prelevati e clonati tramite la T4 RNA ligasi. Per il sequenziamento degli RNA brevi si è usato il 2000 Sequence Analyzer di Illumina e le piattaforme BIOPUCES e SEQUENCAGE a IGBMC, Illkirch (Francia). I reads sono stati allineati confrontandoli con il genoma del virus e con il gene $white_{IR}$, usando il software Bowtie (cfr. sez. 2.4.2), e poi immagazzinati in file FastQ per le successive analisi.

3.2.2 Risultati e domande

Il primo risultato evidente è, come già accennato, il colore degli occhi nei vari genotipi. Mentre il wild-type possiede occhi bianchi, grazie all'avvenuta reazione di silenziamento del gene $white^{IR}$, i due tipi modificati (w^{IR};dcr-2null e w^{IR};dcr-2^{G31R}) presentano occhi rossi. Tuttavia tra w^{IR};dcr-2null e w^{IR};dcr-2^{G31R} si possono notare varie differenze: mentre nel primo non si rilevano reads sul gene $white^{IR}$, nel secondo alcuni vengono rilevati, e la conseguenza che se ne può trarre è che l'enzima possiede ancora qualche attività di taglio (cfr. fig. 3.4). In figura 3.3 si può notare la distribuzione in lunghezza di questi reads; come ci si aspetta il picco si ha per un valore di 21 nucleotidi.
Un altro risultato degno di nota è che la mutazione all'elicasi non influenza la sopravvivenza del moscerino (cfr. fig. 3.5), mentre quando Dicer non è presente si può notare come il moscerino muoia più in fretta.

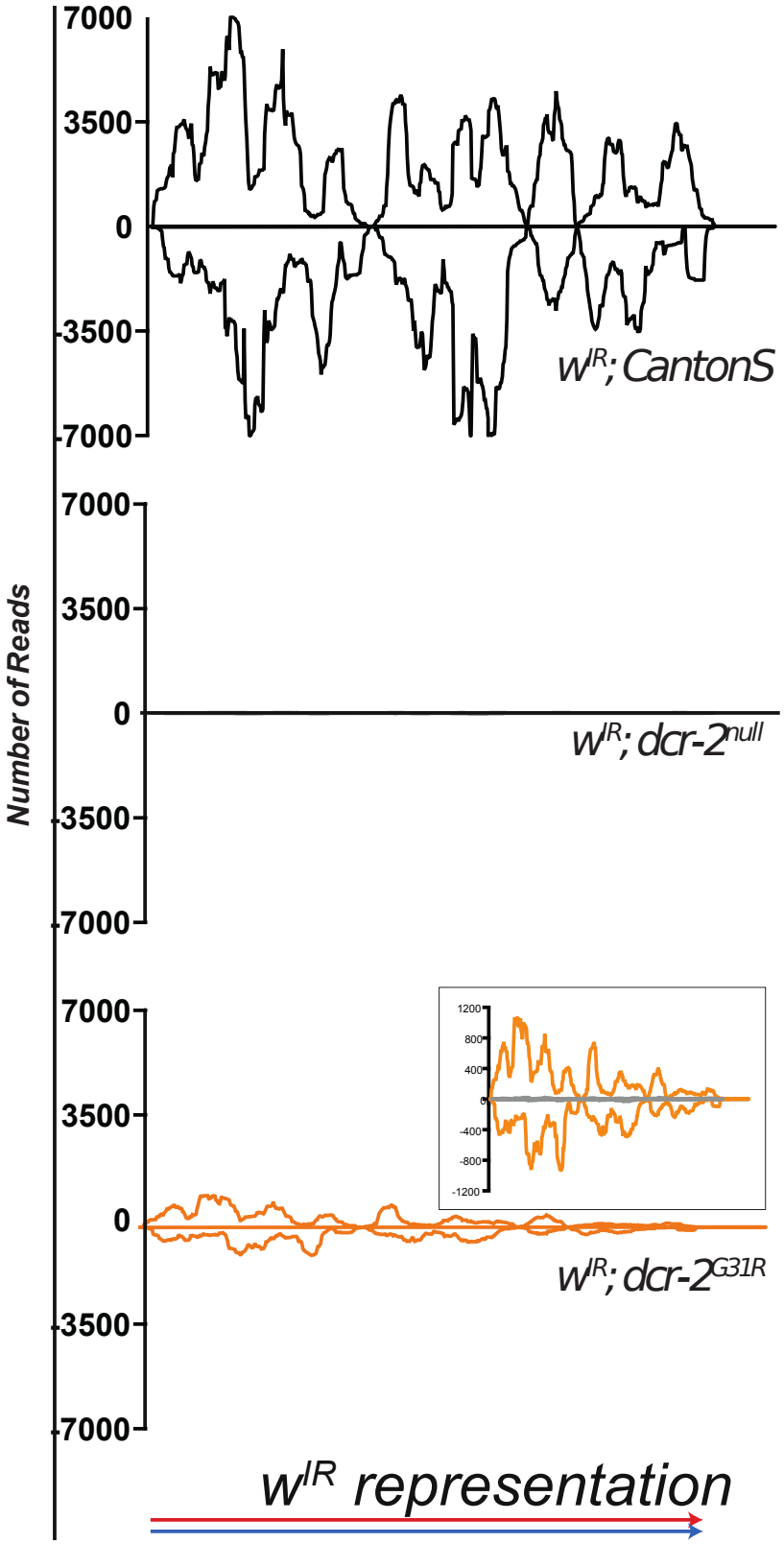

Figura 3.4: Allineamento dei reads sul transgene $white^{IR}$ per i tre genotipi

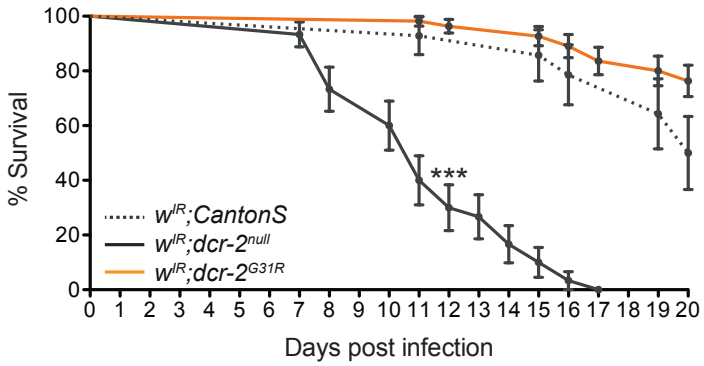

Figura 3.5: Analisi di sopravvivenza

Per quanto riguarda l'allineamento dei reads con il genoma di SINV, un ritrovamento di molti reads sta a significare che nel moscerino è attivo il meccanismo di difesa contro l'infezione. I risultati ottenuti dall'allineamento (cfr. fig. 3.6) sono in accordo con quanto derivato dall'analisi di sopravvivenza. Un'eccezione è rappresentata dal genotipo w^{IR};dcr-2null (in verde), in quanto è presente un numero minimo di reads: anche se afferma che potrebbero essere solamente residui del deterioramento dell'RNA, il gruppo di Strasburgo si è chiesto se in realtà questo abbia qualche rilevanza a livello biologico(cfr. domanda 3). Per quanto riguarda il w^{IR};dcr-2^{G31R}, vengono trovati molti reads , soprattutto nella parte iniziale del genoma. Il decremento nella parte finale può, forse, essere indice di una decrescita della processività di Dicer quando si arriva alla fine del filamento (cfr. domanda 2). Solo all'estremità 5' si può rilevare un match dei reads sia per quanto riguarda il genoma sia per l'antigenoma, mentre nelle restanti coordinate vengono rilevati quasi solo reads genomici; dunque questa è una differenza di espressione con il wild-type, il quale mostra un allineamento per entrambi i filamenti.
Questi risultati, anche se in linea generale confermano le previsioni attese, sono ricchi di sottigliezze e casi statistici dubbi. Si può provare a trarre qualche conclusione, che potrebbe portare a nuove conoscenze dei processi molecolari e a meccanismi finora non considerati.
In particolare ciò che il gruppo di lavoro di Strasburgo si chiede è:

1. Reads di $white^{IR}$ nel genotipo w^{IR};CantonS.
   Si vuole capire se il modello osservato nella distribuzione dei siRNAs (in sostanza,

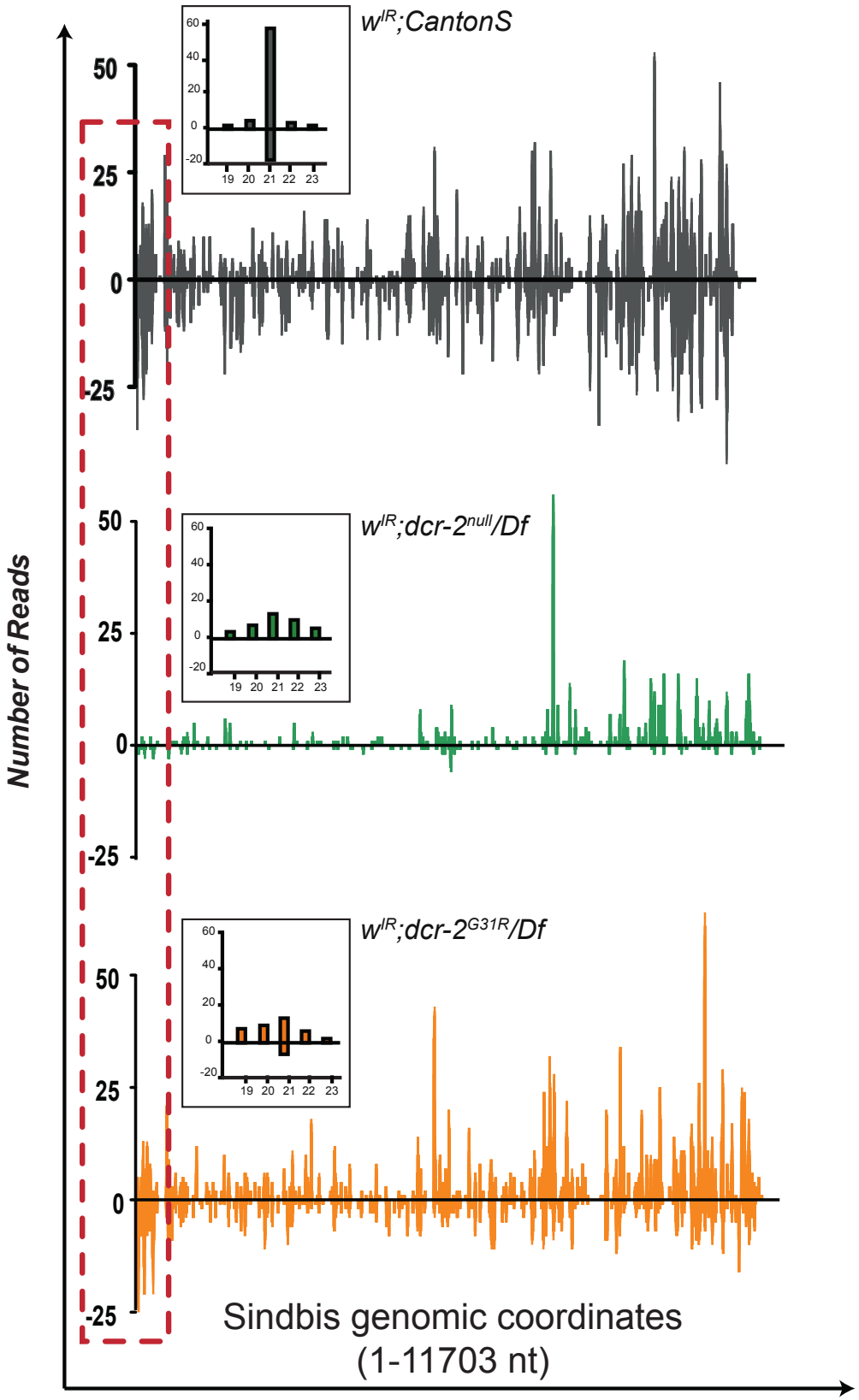

Figura 3.6: Allineamento dei reads sul genoma del virus

i reads sembrano raggrupparsi in tre regioni) ha qualche significato e se questo può essere riferito al pathway dell'enzima.

2. Reads di $white^{IR}$ nel genotipo w^{IR};dcr-2^{G31R}.
   La distribuzione dei reads sembra decrescere verso la fine del dsRNA processato da Dicer-2. Qual è il modo migliore di trattare la significatività statistica di questa osservazione, che, come anticipato prima, può indicare una decrescita nella processività del mutante dell'elicasi? Perciò, possiamo osservare una decrescita nel *phasing*, che è definito come la distanza tra le estremità 5' di reads provenienti dallo stesso filamento?

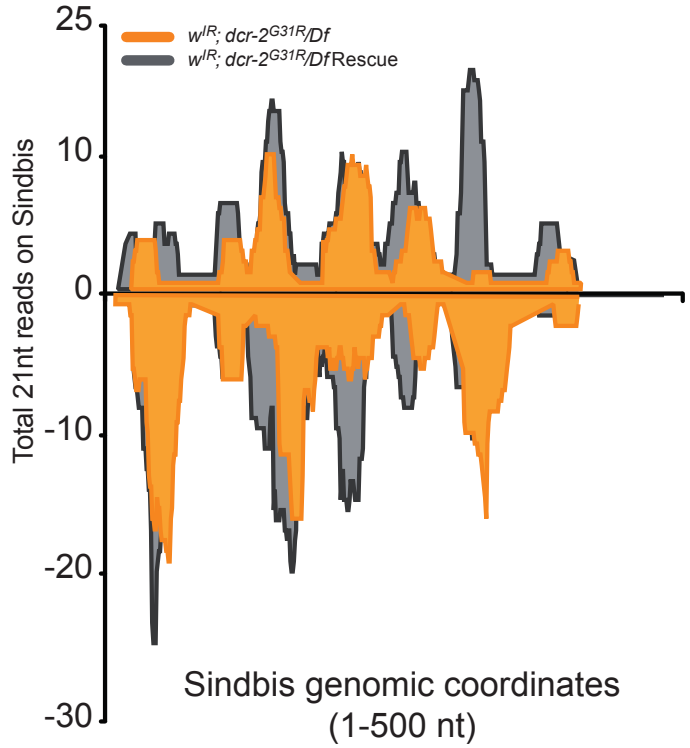

Figura 3.7: Confronto dell'espressione di CantonS e G31R all'estremità 5'

3. Reads di SINV.
   La somiglianza della distribuzione dei reads tra w^{IR};CantonS e w^{IR};dcr-2^{G31R} al 5' del genoma virale (cfr. fig. 3.7) è realmente significativa in termini statistici? Come può essere quantificata la differenza fra i profili degli RNA brevi per la coda 3' del

SINV in w^{IR};CantonS rispetto a w^{IR};dcr-2^{G31R} (prendendo in considerazione anche il fatto che ci sono reads in questo regione anche per il mutante nullo)?

4. È possibile osservare una *dicing signature* (firma di taglio) in uno dei dati di un wild-type di controllo? Definiamo come firma i reads che hanno una soglia minima di 18. Questo significa che dovremmo confrontare l'estremità 5' dei reads in un filamento con il 5' dei reads nel filamento opposto ottenendo una perfetta sovrapposizione fra essi. A causa della sporgenza di due nucleotidi in entrambi i filamenti, ci aspettiamo un offset al punto 18 (iniziando a numerare da zero). La firma di taglio è stata osservata in precedenza solo quando il complesso di Ago2 e dell'RNA breve ad esso associato è stato purificato. In questo caso tutto ciò che si associa ad Ago2 è da considerare siRNA e la firma di taglio è facile da vedere. La limitazione sta nel fatto che si sta cercando tutto il siRNA, dunque anche quello dei due filamenti che viene degradato in Ago2; quest'ultimo ovviamente non è rilevabile.

Nella prossima sezione proviamo a fornire le risposte a questi quesiti.

3.3 Modelli e metodi di analisi dati

Per rispondere alle questioni sollevate dal gruppo di ricerca dell'Università di Strasburgo proponiamo diversi metodi statistici. Siccome i problemi trattati e le relative domande sono piuttosto peculiari, cioè non abbiamo a che fare con un semplice problema classico di analisi differenziale, anche alcuni dei metodi di risoluzione che proponiamo risultano poco convenzionali. Per confrontare i profili di espressione facciamo uso del test di ipotesi di Kolmogorov-Smirnov (cfr. Appendice B), il quale consente di paragonare i due campioni senza fare ipotesi sulla loro distribuzione. Altri metodi che impieghiamo sono la somma di distribuzioni gaussiane, fit lineari e funzione di autocorrelazione.

Divisione in tre regioni dei reads di CantonS

Per capire se effettivamente i reads del genotipo di controllo si distribuiscono in tre regioni (cfr. fig. 3.3) e se questa distribuzione, quindi, è dovuta a una particolare caratteristica dell'enzima, sarebbe opportuno ripetere più volte l'esperimento in questione, per

avere un informazione statistica più robusta. Infatti svariati fattori possono influenzare i reads rilevati e la variabilità biologica è molto alta. Tuttavia, a causa dell'elevato costo dell'esperimento, comprendiamo le difficoltà che molte sue ripetizioni comporterebbero. Nonostante ciò, possiamo proporre un metodo con cui procedere, che può fornire una prima risposta. Ciò che intendiamo fare è l'adattamento dei dati a una *Mixture Distibution*, costruita, in questo caso, come combinazione lineare di tre Gaussiane. Per prima cosa è necessario normalizzare i dati in base al numero di reads presenti, in modo che quella che andiamo a utilizzare sia effettivamente una funzione di densità di probabilità. Bisogna poi dimostrare, tramite un test del χ^2, che questo adattamento risulta verosimile, mentre quello a una gaussiana singola no. Dobbiamo quindi vedere se la probabilità che il campione si adatti alla distribuzione singola è minore del 5% - in gergo si dice che dobbiamo verificare se il p-value è minore di 0,05 - mentre nel caso della distribuzione mista deve essere $p > 5\%$. Se ciò accade allora possiamo concludere che è altamente improbabile che il campione sia compatibile con la distribuzione singola, mentre c'è una probabilità non trascurabile che lo sia con quella mista, fatto che può indicare un pathway preferenziale dell'enzima.

Decrescita della processività in w^{IR};dcr-2^{G31R}

Con riferimento alla linea arancione della figura 3.4, ciò che dobbiamo fare è verificare se all'aumentare delle coordinate del transgene la distribuzione decresce, rispetto al caso usato come controllo. Per fare ciò, si può procedere facendo un fit lineare di tutti i picchi della distribuzione nei due diversi casi, mostrando che la retta del genotipo modificato ha pendenza più negativa di quella del controllo, se l'ipotesi è corretta. Tuttavia evidenziamo che, al di là del risultato dell'analisi del fit, a nostro parere la pendenza non è l'aspetto più significativo dei due grafici. Al contrario, ciò che si può notare immediatamente, è l'enorme decrescita nel numero di reads rilevati nel mutante, rispetto al controllo; infatti nel primo caso abbiamo picchi di reads tra 40 mln e 70 mln, nel secondo caso, invece, i picchi si trovano abbondentemente sotto i 10 milioni. Questo naturalmente indica una minore attività di Dicer nel genotipo w^{IR};dcr-2^{G31R}.

Somiglianza profilo di espressione 5'

Per testare se la somiglianza che c'è tra w^{IR};CantonS e il genotipo w^{IR};dcr-2^{G31R} all'estremità 5' del filamento è effettivamente significativa, facciamo uso del test di ipotesi

di Kolmogorov-Smirnov (cfr. Appendice B), che è particolarmente indicato non solo perché non è disponibile una stima di parametri di una distribuzione, ma perché non abbiamo alcuna informazione sulla distribuzione alla quale i dati dovrebbero adattarsi. Essendo i reads delle variabili discrete, si usa la versione discreta del test; quindi in questo caso le classi di frequenza sono definite da ciascun nucleotide del filamento. Ciò che si deve fare è confrontare il massimo scarto fra le due funzioni di distribuzione cumulative, definito come

$$D = \sup_{i=1,\ldots,N} \{|F(x_i) - F_0(x_i)|\}$$

Così facendo si vede se la differenza tra le due espressioni ha probabilità di verificarsi minore del 5%. Se ciò accade allora possiamo concludere che è altamente improbabile che la nostre due distribuzioni siano compatibili, ovvero i risultati indicano che ci sono effettive differenze nel livello di espressione, altrimenti l'ipotesi può essere considerata possibile. Anche in questo caso, ovviamente, la prima cosa da fare è la normalizzazione dei dati, in modo che la funzione che usiamo rappresenti realmente una distribuzione cumulativa.

Dicing signature e phasing

In questo caso è richiesta sostanzialmente un'analisi di periodicità, per vedere se le estremità dei reads si pongono su intervalli ben definiti, fatto che indica il corretto processamento del filamento da parte di Dicer-2, il quale taglia l'RNA in frammenti da 21nt. Per verificare la periodicità usiamo una funzione di autocorrelazione: quest'ultima, come dice il nome, dà la misura della correlazione tra diversi punti del dominio della funzione stessa. In particolare, nel nostro caso, indica dove vengono rilevate le estremità 5' dei reads, quindi, ciò che ci aspettiamo se il phasing avviene correttamente, sono dei picchi nella distribuzione delle estremità 5' a intervalli di 19 nucleotidi (considerando che c'è una sporgenza di due nucleotidi). Per fare questa analisi possiamo usare il programma *R*, che al suo interno ha già implementata una funzione di autocorrelazione, oppure possiamo usare qualsiasi altro linguaggio o applicazione a nostro piacimento, ma il lavoro risulta mediamente più lungo. Anche in casi in cui ci aspetteremmo una dicing signature piuttosto chiara, può accadere che non riusciamo a rilevarla, e cioè che la funzione di autocorrelazione o non abbia picchi (cfr. fig. 3.8) o li abbia in punti diversi da quelli attesi; questo può essere indice della presenza di qualche enzima che maschera il phasing, (come ad esempio R2D2) o della diversa biogenesi dei dsRNA tagliati da Dicer [22].

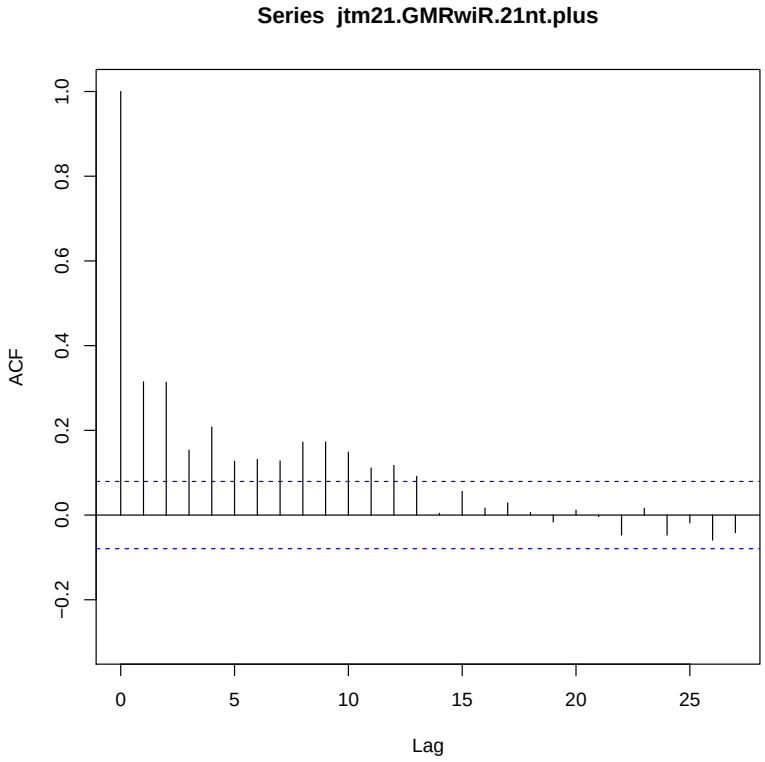

Figura 3.8: Esempio di analisi tramite funzione di autocorrelazione (ACF): in questo caso il phasing è mascherato [22]

Conclusioni

L'esperimento visto tratta dei ruoli che Dicer-2 ha nel pathway della RNA interference. Come si è potuto notare la mutazione puntuale nel dominio Elicasi di questo enzima è la più interessante, poiché porta a risultati non ancora ben compresi, relativamente a cui abbiamo proposto alcune metodologie di analisi statistica.
Si è mostrato che uno degli aspetti caratterizzanti di Dicer-2 è quello di tagliare i dsRNA in frammenti di esattamente 21 nucleotidi, con sporgenza di due nucleotidi all'estremità 3', quindi per verificarne la processività si implementa una funzione di autocorrelazione. Per quanto riguarda il confronto del livello di espressione del genotipo mutato nell'Elicasi con il genotipo di controllo, abbiamo proposto un test di Kolmogorov-Smirnov (nella versione discreta) a due campioni, che consente di non utilizzare una distribuzione teorica di paragone; la risposta a questo quesito può fornire informazioni sulla struttura dei dsRNA processati e su come, nello specifico, avvenga il loro taglio.
Inoltre, dal momento che nel genotipo di controllo i reads sembrano distribuirsi in tre regioni, abbiamo proposto un adattamento del campione a una distribuzione mista, formata da tre gaussiane. Un tale adattamento porrebbe molti nuovi interrogativi, dal momento che la distribuzione potrebbe essere conseguenza delle reazioni molecolari compiute dall'enzima, che sono tuttora sconosciute.
In conclusione, i metodi di analisi discussi e proposti in questa tesi possono fornire delle soluzioni ai quesiti posti, ma il lavoro da fare per giungere ad una completa comprensione del problema è ancora molto. Sottolineiamo nuovamente il contributo che fisici ed esperti di altri settori scientifici possono dare, poiché un'impostazione più rigorosa del problema e dell'analisi dei dati potrebbe evitare molti errori di valutazione. Inoltre, finora, le conoscenze che fanno riferimento esclusivamente al campo della biologia e della chimica non hanno fornito risposte soddisfacenti: quindi l'utilizzo di metodi statistici adeguati può fornire nuove e inaspettate chiavi di lettura riguardo molti meccanismi biomolecolari.

Appendice A:
la distribuzione binomiale negativa

La distribuzione binomiale negativa deriva dalla distribuzione di Pascal nel passaggio di variabile dai successi agli insuccessi. Quest'ultima è definita come distribuzione di probabilità discreta con due parametri, detti p ed x, che descrive il numero di fallimenti $x - k$ precedenti al successo k-esimo, in un processo di Bernoulli definito dal parametro p. Nello specifico, dato un processo bernoulliano, cioè una serie di variabili aleatorie indipendenti (tutte con la stessa distribuzione di Bernoulli), descrive la variabile aleatoria X che conta il numero di prove totali necessarie a ottenere k successi. Quindi se p è la probabilità di successo in una prova, allora $q = 1 - p$ è la probabilità di fallimento. La probabilità che si verifichino $x - k$ fallimenti prima di ottenere k successi totali è data dalla probabilità di ottenere un successo nella prova numero x, moltiplicata per la probabilità di ottenere $k - 1$ successi e $x - k$ fallimenti nelle precedenti prove. Cioè

$$P(x, p) = p\binom{x-1}{k-1}p^{x-1}q^k = \binom{x-1}{k-1}p^x q^k$$

Questa distribuzione può essere invertita, considerando al posto della variabile X, la variabile $Y = X - k$, che descrive l'andamento del numero dei fallimenti precedenti al successo k-esimo; in questo modo otteniamo la binomiale negativa. Sostituendo nell'equazione di Pascal si ha

$$NB(y, p) = \binom{y+k-1}{k-1}p^k q^y$$

$$NB(y,p) = (-1)^y \binom{y+k-1}{y} p^k (-q)^y = \frac{(-k)(-k-1)...(-k-y+1)}{y!} p^k (-q)^y$$

Otteniamo infine

$$NB(y,p) = \binom{-k}{y} p^n (-q)^k$$

Da quest'ultima risulta chiaro il nome della distribuzione: questa non è altro che una binomiale con coefficienti negativi. Si possono inoltre calcolare la media e la varianza, ottenendo

$$\mu = \frac{k}{p}$$

$$\sigma^2 = \frac{k(1-p)}{p^2}$$

.

Appendice B:
il test di Kolmogorov-Smirnov

Il test di ipotesi di Kolmogorov e Smirnov è un test non parametrico utilizzato per la sua semplicità di implementazione e per la sua validità in buona approssimazione già a partire da un numero di dati piccolo. Nella sua formulazione esatta prevede che la variabile considerata sia continua, tuttavia è presente anche una versione per variabili discrete. Se abbiamo un campione di N dati $\{x_1, x_2, ..., x_N\}$, questo test consente di confrontarlo con una distribuzione attesa, ma anche con un altro campione di dati. Il modo di procedere è il seguente: per prima cosa bisogna ordinare le variabili del campione, a seconda del loro valore in ascissa, dal minore al maggiore. Da qui si può costruire la funzione di distribuzione empirica $F_0(x)$, definita puntualmente come

$$F_0[x(i)] = \frac{i}{N}$$

Intuitivamente questa funzione è in grado di dare una misura della distribuzione cumulativa, tenendo conto dei punti rilevati dalla misura sperimentale. Cioè, se vengono rilevati molti punti in una determinata regione del dominio, allora lì la funzione di distribuzione sarà maggiore rispetto al caso di distribuzione uniforme. Ovviamente è definita in modo tale che nel punto N il suo valore sia uno, cioè è normalizzata. Per confrontare questa funzione di distribuzione con una distribuzione teorica nota o con la distribuzione empirica di un altro campione, detta $F(x)$ la funzione di distribuzione di confronto, si definisce la variabile

$$D = \sup_{i=1,...,N} \{|F(x_i) - F_0(x_i)|\}$$

Quest'ultima viene impiegata come statistica del test, e indica sostanzialmente la distanza massima tra due punti (con stessa ascissa) delle distribuzioni cumulative da confrontare. Quindi si procede in maniera analoga al test del χ^2, fissando un *p-value*, che rappresenta il livello di probabilità oltre al quale scartare l'ipotesi, cioè, per un dato n (o dati n ed m, nel caso del confronto fra due campioni):

$$P(D > d_p) = p$$

Di solito si fissa il valore p al 5%, ovvero $p = 0.05$. Data una distribuzione di confronto, i valori di d_p per un dato n si trovano tabulati. Quindi se otteniamo $D > d_p$ l'ipotesi di uguaglianza tra i campioni (o tra campione e distribuzione) è da rigettare, viceversa l'ipotesi è considerata possibile.
L'uso della funzione di distribuzione, che è cumulativa, permette di evitare la preventiva suddivisione in classi di frequenza, semplificando notevolmente il lavoro. Questo discorso vale nel caso di variabile continua; tuttavia quando andiamo a considerare la versione discreta del test, di fatto abbiamo già una divisione in classi di frequenza, e quindi il problema non si pone nemmeno. Ciò non significa che nel caso discreto il test K-S sia inutile, anzi, come già detto, permette di confrontare due campioni senza fare ipotesi sulla loro distribuzione o su altri parametri, dunque risulta molto potente.

Bibliografia

[1] C Napoli, C Lemieux, R Jorgensen, *Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans*, Plant Cell, vol. 2, pp. 279-289, (1990)

[2] (online) www.nobelprize.org, *The 2006 Nobel Prize in Physiology or Medicine - Advanced Information*, Nobel Media AB, (2013)

[3] Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, and Mello CC, *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans*, Nature, vol. 391, pp. 806-811, (1998)

[4] Tuschl T, Zamore PD, Lehmann R, Bartel DP, and Sharp PA, *Targeted mRNA degradation by double-stranded RNA in vitro*, Genes Development, vol. 13, pp. 3191-3197, (1999)

[5] Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R, *Human RISC couples microRNA biogenesis and posttranscriptional gene silencing*, Cell, vol. 123, pp. 631–640, (2005)

[6] Ronald H. A. Plasterk, *RNA Silencing: the Genome's Immune System*, Science, vol. 296, pp. 1263-1265, (2002)

[7] Cullen B, *Is RNA interference involved in intrinsic antiviral immunity in mammals?*, Nature Immunology, vol. 563 n.7, pp. 563-567, (2006)

[8] Giuseppe Aprea, Giulio Gianese, Marco Pietrella, Vittorio Rosato, Valentina Spedaletti, *Il ruolo dell'ICT nelle scenze omiche high-throughput*, EAI, Speciale, pp. 74-82, (2013)

[9] Jay Shendure and Hanlee Ji, *Next-generation DNA sequencing*, Nature Biotechnology, vol. 26, pp. 1135-1145, (2008)

[10] (online) www.illumina.com

[11] Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF., *Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing*, BMC Genomics, vol. 12, pp. 245-255, (2011)

[12] (online) www.roche.com

[13] (online) www.appliedbiosystems.com

[14] Chengwei Luo, Despina Tsementzi, Nikos Kyrpides, Timothy Read, Konstantinos T. Konstantinidis, *Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample*, PLoS ONE, vol. 7, (2012) (online: www.plosone.org)

[15] (online), http://www.genengnews.com, Shawn C. Baker, *Next-Generation Sequencing vs. Microarrays*

[16] J. Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, Hamilton O. Smith, *Environmental Genome ShotgunSequencing of the Sargasso Sea*, Science, vol. 304, pp. 66-74, (2004)

[17] Ruth E. Ley, Micah Hamady, Catherine Lozupone, Peter J. Turnbaugh, Rob Roy Ramey, J. Stephen Bircher, Michael L. Schlegel, Tammy A. Tucker, Mark D. Schrenzel, Rob Knight, Jeffrey I. Gordon, *Evolution of mammals and their gut microbes*, Science, vol. 320, pp. 1647-1651, (2008)

[18] Patricio Jeraldo, Maksim Sipos, Nicholas Chia, Jennifer M. Brulc, A. Singh Dhillon, Michael E. Konkel, Charles L. Larson, Karen E. Nelson, Ani Qu, Lawrence B. Schook, Fang Yang, Bryan A. White and Nigel Goldenfeld, *Quantification of the relative roles*

of niche and neutral processes in structuring gastrointestinal microbiomes PNAS, vol.109, pp.9692-9698, (2012)

[19] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang and Xuegong Zhang, *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data*, Bioinformatics vol. 26, pp. 136-138, (2010)

[20] Mark D. Robinson, Davis J. McCarthy and Gordon K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression*, Bioinformatics, vol. 26, pp. 139-140, (2010)

[21] (online) http://bowtie-bio.sourceforge.net

[22] Joao Trindade Marques, Ji-Ping Wang, Xiaohong Wang, Karla Pollyanna Vieira de Oliveira, Catherine Gao, Eric Roberto Guimaraes Rocha Aguiar, Nadereh Jafari and Richard W. Carthew, *Functional Specialization of the Small Interfering RNA Pathway in Response to Virus Infection*, PLOS Pathogens, vol. 9, (2013) (online: www.plospathogens.org)

[23] Young Sik Lee, Kenji Nakahara, John W. Pham, Kevin Kim, Zhengying He, Erik J. Sontheimer and Richard W. Carthew, *Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways*, Cell, vol. 117, pp. 69–81, (2004)

[24] Elif Sarinay Cenik, Ryuya Fukunaga, Gang Lu, Robert Dutcher, Yeming Wang, Traci M. Tanaka Hall, and Phillip D. Zamore, *Phosphate and R2D2 Restrict the Substrate Specificity of Dicer-2, an ATP-Driven Ribonuclease*, Molecular Cell, vol. 42, pp. 172-184, (2011)

Ringraziamenti

Questo lavoro di tesi è solo la fine di un percorso durato tre anni, che mi ha formato e cambiato sotto moltissimi aspetti.
Ringrazio innanzitutto i miei genitori, per avermi messo nelle condizioni di affrontare quest'esperienza e per il sostegno che mi hanno dato, nonostante la lontananza.
Ringrazio mio fratello e mia sorella, che sono sempre stati un punto fermo e che mi hanno permesso di avere davanti una strada già battuta in molte situazioni.
Ringrazio la professoressa Elda Gallo, determinante nella scelta di questo corso di laurea e sempre pronta ad ascoltarmi e consigliarmi.
Ringrazio il professor Giovanni Carlo Bonsignori, per l'incredibile entusiasmo e la vastità di conoscenze che è riuscito a trasmettere anche in poche ore di insegnamento.
Ringrazio i miei compagni di corso e gli amici dell'università, per avermi arricchito a livello didattico e personale, per la premura con cui mi hanno sempre dato consigli utili e sinceri quando ne ho avuto bisogno, per avermi sostenuto nei laboratori e per tutti i momenti condivisi durante le giorante in dipartimento.
Ringrazio coinquilini e coinquiline, presenti e passati, per l'affetto e la comprensione che ho ricevuto.
Infine un ringraziamento speciale va a Lorenzo, Gabriele e Tommaso, che mi hanno dato molto più di quanto io sia in grado di dimostrare loro.