Università Campus Bio-Medico di Roma

Corso di Dottorato di ricerca in

Bioingegneria e Bioscienze

XXXI ciclo anno 2019

Approccio bioinformatico per l'integrazione di dati array CGH e Genome Wide Expression per la diagnosi molecolare e la terapia personalizzata nel Disturbo dello Spettro Autistico

Pasquale Tomaiuolo

Controrelatore
Prof. Giulio Iannello

Relatore
Prof. Antonio M. Persico

Alla mia tenacia e testardaggine,
A tutte le persone che mi hanno
sostenuto sempre e comunque
A tutti i miei familiari…
A chi non c'è più, affinché, dall'alto,
possa vigilare e illuminare il mio
cammino, sempre.

Indice

1. INTRODUZIONE

Con il termine *"Bioinformatica"* s'intende l'applicazione delle tecniche computazionali per comprendere e organizzare le informazioni associate alle macromolecole biologiche (Hagen JB, 2000). La Bioinformatica è caratterizzata dall'applicazione di metodi matematici, statistici e computazionali all'analisi di dati biologici, biochimici e biofisici. I suoi obiettivi sono molteplici: il primo è quello di organizzare e aggiornare i dati in modo da permettere ai ricercatori un facile accesso alle informazioni esistenti. Il secondo obiettivo è quello di sviluppare strumenti e risorse che aiutino nell'analisi e nel confronto dei dati; lo sviluppo di tali risorse richiede competenze in teoria computazionale, nonché una comprensione approfondita dei processi biologici che sottendono l'oggetto della ricerca. Infine, obiettivo ultimo del processo è quello di utilizzare questi strumenti per analizzare i dati e interpretare i risultati in maniera biologicamente significativa (Hagen JB, 2000, Luscombe NM, 2000).

Le sue principali attività riguardano la costruzione ed il mantenimento di una varietà di banche dati, lo sviluppo di algoritmi per l'allineamento di sequenze di DNA, RNA e proteine, l'identificazione dei geni e l'assemblaggio dei genomi, la predizione di strutture molecolari e delle interazioni di acidi nucleici e proteine, la ricostruzione e l'analisi di pathway biologiche, lo sviluppo e l'applicazione di metodi adeguati per l'immagazzinamento, l'interrogazione, l'integrazione e l'analisi dei dati biologici.

Storicamente lo sviluppo della Bioinformatica è andato di pari passo con i progressi scientifici acquisiti sia nel campo della biologia molecolare che nel campo dell'informatica degli ultimi quaranta anni. E' pertanto importante sottolineare che i livelli di sviluppo che la bioinformatica ha raggiunto sono dovuti anche al concomitante aumento da un lato delle tecnologie *-omiche*, che hanno portato alla realizzazione di importanti progetti con risultati di grande rilevanza, dall'altra delle tecnologie informatiche che hanno enormemente facilitato l'archiviazione di grandi quantità di dati, consentendo la diffusione delle informazioni attraverso le reti telematiche.

1.1 Storia della bioinformatica

La bioinformatica nasce circa dieci anni prima del sequenziamento del DNA (Diniz WJ, 2017). La pubblicazione, nel 1953, della struttura del Dna da parte di Watson e Crick (Watson JD, 1953) e l'accumulo di dati e conoscenze biochimiche sulla struttura delle proteine originati dagli studi di Pauling (Pauling L, 1951), Coren e Ramachandran (Ramachandran GN, 1963) negli anni Sessanta, hanno rappresentato due delle tappe storiche fondamentali (Diniz WJ, 2017). Tuttavia è sicuramente la pubblicazione, nel 1967, dell'*"Atlas of protein sequences"* da parte di Margareth Dayhoff a porre le basi per la nascita di questa disciplina scientifica (Hunt LT, 1984). Infatti, il lavoro della Dayhoff

rappresentava la prima collezione completa, seppur cartacea, delle sequenze proteiche conosciute all'epoca. Pochi anni dopo, nel 1972, il suo Atlante fu convertito nella versione elettronica tuttora conservata nella banca dati NBRF (National Biomedical Research Foundation). Nello stesso anno Paul Berg descrisse la prima molecola di DNA ricombinante, mentre Maxam e Gilbert (Maxam AM, 1977) e Sanger (Sanger F, 1977) definirono, indipendentemente, due metodi per ottenere la sequenza del filamento del DNA. Parallelamente, Peter Metcalfe descrisse uno standard, chiamato Ethernet (Metcalfe RM, 1976), specializzato nella trasmissione dei dati tra computer connessi via cavo; mentre Cerf e Kahn, l'anno successivo, svilupparono un protocollo di comunicazione tra computer chiamato Transmission Control Protocol (TCP). Queste scoperte diedero avvio in particolare alla creazione dell'attuale rete web e nel complesso diedero il via ad una imponente trasformazione delle tecniche informatiche.

Verso la fine degli anni Settanta furono pubblicate le prime sequenze nucleotidiche; nacque allora l'esigenza di avere a disposizione sistemi informatici per l'archiviazione e l'analisi dei dati di sequenza, prodotti in maniera sempre più massiva. Ai biologi molecolari apparve sempre più evidente la necessità di utilizzare tecnologie informatiche a supporto della biologia e divenne di fatto impensabile condurre grandi progetti di ricerca senza il supporto di archivi informatici in cui poter immagazzinare tali dati e di codici per poterli analizzare.

Agli inizi degli anni Ottanta, l'European Molecular Biology Laboratory (EMBL) di Heidelberg promosse la costituzione dell'***EMBL data library***, una banca dati di sequenze di DNA e RNA e più tardi diede inizio al progetto di biocomputing, contribuendo, nel tempo, allo sviluppo della bioinformatica europea con la messa a punto di strumenti e metodologie. Nel 1992, come conseguenza degli investimenti economici fatti da alcuni Stati Europei nascerà l'Istituto di Bioinformatica Europeo (**EBI**) (Cook CE, 2017). Contemporaneamente, negli Stati Uniti fu creato un archivio simile: si tratta della banca dati da cui ha avuto origine *GenBank* (Benson DA, 2018), una delle più grandi banche date di sequenze nucleotidiche. Si dovrà aspettare il 1986 per vedere realizzata la banca dati Giapponese DDBJ (DNA Data Bank of Japan) (Kodama Y, 2018), cui seguirà un accordo tra queste 3 grandi società per un continuo scambio di dati. Ed ancora, sempre nella seconda metà degli Anni '80 si formarono le prime banche dati specializzate come PROSITE (Database of Protein Domains, Families and Functional Sites) (Sigrist CJ, 2012) ed EPD (The Eukaryotic Promoter Database) (Périer RC, 2000). A oggi sono centinaia i Database specializzati accessibili dalla rete (Baxevanis AD, 2015) che permettono ai ricercatori di condurre numerosi studi e di condividere i dati.

Da questi importanti presupposti nacque la bioinformatica, un branca della scienza che ad oggi si pone obiettivi di ricerca anche molto ambiziosi come ad esempio il progetto Genoma Umano. Avviato nel corso degli anni '90 dal Dipartimento dell'Energia e dal National Institute of Health degli

USA, il progetto **Genoma Umano** si poneva l'obiettivo di determinare la sequenza dei nucleotidi che formano il DNA e di identificare e mappare i geni del Genoma Umano dal punto di vista sia fisico che funzionale. Completato nel 2000 (Venter JC, 2001), il 26 giugno dello stesso anno l'allora presidente degli Stati Uniti Bill Clinton, insieme ai due principali investitori, i ricercatori Craig Venter e Francis Collins, annunciarono, congiuntamente, il completamento della sequenza preliminare del genoma umano. Questo annuncio fu salutato come il primo grande trionfo del ventunesimo secolo. Anche se il progetto Genoma Umano è stato un'impresa di successo, numerosi aspetti e quesiti restano ancora irrisolti, interrogativi ai quali numerose altre iniziative hanno iniziato a lavorare per trovare risposta. Tra questi ricordiamo, ad esempio, il progetto ENCODE (Consortium EP, 2004), il cui scopo è quello di identificare e descrivere in maniera completa tutte le regioni funzionali della sequenza del genoma umano, ed il 1000 Genomes Project (Consortium, 2015), il cui scopo è quello di studiare la variabilità genetica nelle popolazioni umane.

Più recentemente le nuove tecnologie genomiche high-throughput (Slatko BE, 2018) hanno ulteriormente rivoluzionato la ricerca medica generando un volume straordinario di informazioni mai raggiunto prima nella storia della Biologia. La necessità di analizzare questa enorme quantità di dati ha assegnato alla bioinformatica un ruolo sempre più centrale nel settore biomedico.

1.1.1 La Bioinformatica e l'open source

La capacità di programmare, di impartire cioè a un computer una serie di comandi e istruzioni al fine di ottenere il risultato desiderato, costituisce certamente una parte importante del bagaglio di nozioni e abilità necessarie per operare in ambito bioinformatico. Infatti, nonostante la quantità sempre crescente di software e strumenti disponibili, molto spesso i dati di cui si dispone devono essere analizzati in maniera peculiare, attraverso l'utilizzo di tools e script sviluppati e implementati ad hoc. Così, buona parte del lavoro di un bioinformatico si svolge attraverso un'interfaccia "*command line*" che al contrario dell'interfaccia grafica, ha bisogno di istruzioni di testo esatte.

Diversi sono i linguaggi di programmazione usati a questo scopo, fatti propri dalla bioinformatica e nel corso del tempo diverse comunità online sono nate con l'obiettivo di raccogliere script e pipeline, e offrono soluzioni alle svariate domande biologiche. Non vi è, però, nessuna indicazione specifica nella scelta di un linguaggio ma molto dipende dalla tipologia di problema da affrontare, dalle conoscenze e competenze personali e dalle risorse informatiche disponibili.

Uno dei primi linguaggi di programmazione a essere utilizzato in bioinformatica è stato **Perl** (https://www.theperlreview.com/). Si è diffuso quando alcuni gruppi dell'EBI (European Bioinformatics Institute) hanno cominciato a usarlo ed a rilasciare i codici sotto Gnu/GPL (GNU

General Public License). Perl è un tipo di linguaggio di programmazione interpretato che consente di processare file e manipolare testi con estrema facilità, pertanto è particolarmente adatto per la realizzazione di programmi in cui è necessaria un'intensa interazione con il sistema operativo. Perl, può interagire con altri programmi ed è comunemente usato per pubblicare dati sui siti Web. Ad oggi, sono presenti numerose librerie di funzioni e pipeline disponibili su portali online, quali ad esempio CPAN (https://www.cpan.org/) e BioPerl (Stajich JE, 2007).

BioPerl (https://bioperl.org/) è un insieme di moduli che semplificano lo sviluppo di script Perl per le applicazioni bioinformatiche, come ad esempio la conversione di un file in un formato diverso, la ricerca di un articolo, l'avvio di tools di allineamento su database pubblici, il calcolo delle proprietà di una sequenza, etc. BioPerl è stato ed è, un importante progetto per lo sviluppo di script, però nel campo della trascrittomica non ha una grande rilevanza, per cui si preferisce l'uso di software come Bioconductor, BioPython, BioJava, BioC.

R è una suite integrata di software per la manipolazione, l'analisi statistica e la rappresentazione grafica dei dati (Venables LWN, 2018). Le caratteristiche che rendono R particolarmente versatile e attraente sono la sua libera disponibilità, l'essere un linguaggio di programmazione e quindi non essere limitato alle sole funzioni rese disponibili dagli sviluppatori, la capacità di operare anche a livello *"batch"* cioè in forma non interattiva, il sottosistema grafico molto potente e adatto per chiarire i fenomeni sotto indagine e i risultati ottenuti mediante immagini esplicative. Inoltre, la comunità dei ricercatori che ne cura lo sviluppo e l'aggiornamento è molto attiva e dinamica e questo permette di avere una notevole quantità di documenti e di aggiornamenti disponibili che permettono di aumentare costantemente la potenzialità del sistema (https://cran.r-project.org/).

Bioconductor (Gentleman RC, 2004; Huber W, 2015), è un progetto *open source* che, sfruttando il ricco ambiente statistico e di programmazione offerto da R, ha lo scopo di sviluppare tools e metodologie per l'analisi e la comprensione di dati genomici high-throughput. Supporta varie tipologie di dati di sequenziamento high-throughput (compreso DNA, RNA, immunoprecipitazione della cromatina, Hi-C, e metiloma) e risorse per l'annotazione; contiene script per l'analisi di dati provenienti da studi di microarray, proteomica, metabolomica, citometria a flusso. Bioconductor consente una rapida creazione di workflow che supportano il ricercatore in tutte le varie fasi di analisi, interpretazione e pubblicazione dei dati prodotti da esperimenti high-throughput.

Infine il linguaggio di programmazione **Python**, grazie alla sua semplicità e chiarezza, oltre che alla sua potenza, è tra i linguaggi che più si stanno diffondendo negli ultimi anni nel campo della Bioinformatica. **Biopython** (Cock, 2018) è una libreria di moduli focalizzata sul trattamento e l'analisi bioinformatica di dati biologici, scaricabile gratuitamente dal sito http://biopython.org. Le funzioni,

implementate in opportune classi, sono molteplici, e consentono l'interpretazione di numerosi formati di file comuni; la capacità di compiere query per l'accesso ai servizi in rete (NCBI, SRS) e numerose altre analisi.

1.2 DNA e variabilità Genetica

Lo strumento di lavoro del bioinformatico è il computer, che usa per raccogliere, consultare e analizzare i dati biologici al fine di comprendere i meccanismi alla base, mentre i principali oggetti di studio sono il DNA, l'RNA e le proteine.

Il DNA è il veicolo deputato all'immagazzinamento e alla trasmissione dell'informazione genetica, codificata nella sequenza lineare dei nucleotidi, utile al funzionamento ed alla sopravvivenza della cellula e dell'organismo. Il gene è l'unità fondamentale, fisica e funzionale dell'informazione genetica e contiene le istruzioni per l'assemblaggio di proteine ed RNA funzionali. I geni rappresentano la porzione codificante del Genoma e corrispondono solo a circa il 2% dell'intera sequenza. Il resto è costituito da sequenze non codificanti (ripetizioni, sequenze introniche, regioni intrageniche) e per gran parte di esse non è ancora chiara la funzione. L'espressione dell'informazione codificata in un gene avviene in due stadi: la trascrizione e la traduzione. Nella trascrizione il filamento di DNA fa da stampo per produrre RNA messaggero (mRNA), l'elemento deputato al trasferimento dell'informazione dal nucleo cellulare al citoplasma dove, a sua volta, funge da stampo per l'assemblaggio di proteine (traduzione), i veri effettori di tutti i processi fisiologici. Molti RNA, tuttavia, non vengono tradotti e quindi non codificano per proteine, ma fungono da regolatori della trascrizione, dello splicing o della replicazione. Tra questi abbiamo long non-coding RNA (lncRNA) come 3' overlapping ncRNA, antisense, macro-lncRNA, sense intronic, sense overlapping e ncRNA come miRNA, piRNA, rRNA, siRNA per citare alcuni dei più importanti.

La salvaguardia del materiale genetico richiede meccanismi estremamente precisi sia di sintesi che di riparazione (Wolters S, 2013). Nonostante ciò, nel DNA di una cellula possono insorgere delle variazioni casuali rispetto alla normale sequenza nucleotidica ("mutazioni"). La più semplice tra le variazioni che si possono osservare nel genoma umano è il risultato della sostituzione di un singolo nucleotide. In funzione della frequenza con cui una determinata variazione è presente nella popolazione, la stessa sarà definita polimorfismo (frequenza allelica minore o MAF > 1%) o mutazione (MAF < 1%). I polimorfismi di singolo nucleotide (Single Nucleotide Polymorphism, SNP) (Karki R, 2015), sono definiti come differenze di una singola base in sedi specifiche. Gli SNP costituiscono fino al 90% delle differenze di sequenza tra individui e nel genoma umano si osservano

mediamente ogni 200 basi nucleotidiche. Sebbene la maggior parte di essi non abbia effetti diretti sulla funzione cellulare molti sono strettamente associati con alleli o geni causa di malattie.

Fino a pochi anni fa si pensava che la maggior parte della variabilità genetica fosse rappresentata unicamente dagli SNP. Tuttavia, una tra le maggiori scoperte del progetto Genoma Umano è stata l'identificazione di un gran numero di variazioni strutturali submicroscopiche chiamate Copy Number Variation (CNV). Il termine CNV è stato introdotto nel 2006 da Redon per definire *un segmento di DNA maggiore o uguale a 1 Kb presente in un numero variabile di copie rispetto ad un genoma di riferimento* (Redon R, 2006). Gli studi effettuati hanno stimato che più del 12% del Genoma Umano è interessato da CNV di dimensioni intermedie, comprese tra 1 Kb e 3 Mb, e ciò rappresenta la maggiore causa di variabilità genetica tra gli individui (Feuk L, 2006). Poiché tali variazioni strutturali hanno dimensioni anche di diverse megabasi e quindi coprono interi geni e regioni regolatorie, possono determinare variabilità biochimica, fisiologica e morfologica tra gli individui e persino essere causa di patologia.

L'avvento dell'array-CGH (Array Comparative Genomic Hybridization) (Kallioniemi A, 1992; Pinkel D, 1998), consentendo di esaminare l'intero genoma umano in un singolo esperimento con una risoluzione molto più elevata rispetto alle tecniche di bandeggio classiche (Shaw-Smith C, 2004), ha incrementato notevolmente la possibilità di individuare CNV, in forma di delezioni o duplicazioni, sia patogene, sia polimorfiche. La maggior parte delle CNV sono variazioni neutre, ovvero senza apparenti implicazioni nel fenotipo, e perciò possono fissarsi nel genoma, risultando quindi ricorrenti nella popolazione. Alcune variazioni strutturali, però, possono influire sull'espressione genica e sulla variabilità fenotipica, e diversi studi hanno messo in luce il loro possibile coinvolgimento come fattori causativi o di suscettibilità in diverse malattie complesse, fino a quel momento ad eziologia ignota (Grayton HM, 2012; Chong WW, 2014). Le CNV possono, infatti, alterare geni dosaggio-sensibili, interrompere sequenze codificanti, generare geni di fusione, modificare la regolazione genica e, in presenza di una delezione, smascherare un allele recessivo (Feuk L, 2006; Colnaghi R, 2011).

1.3 Array CGH

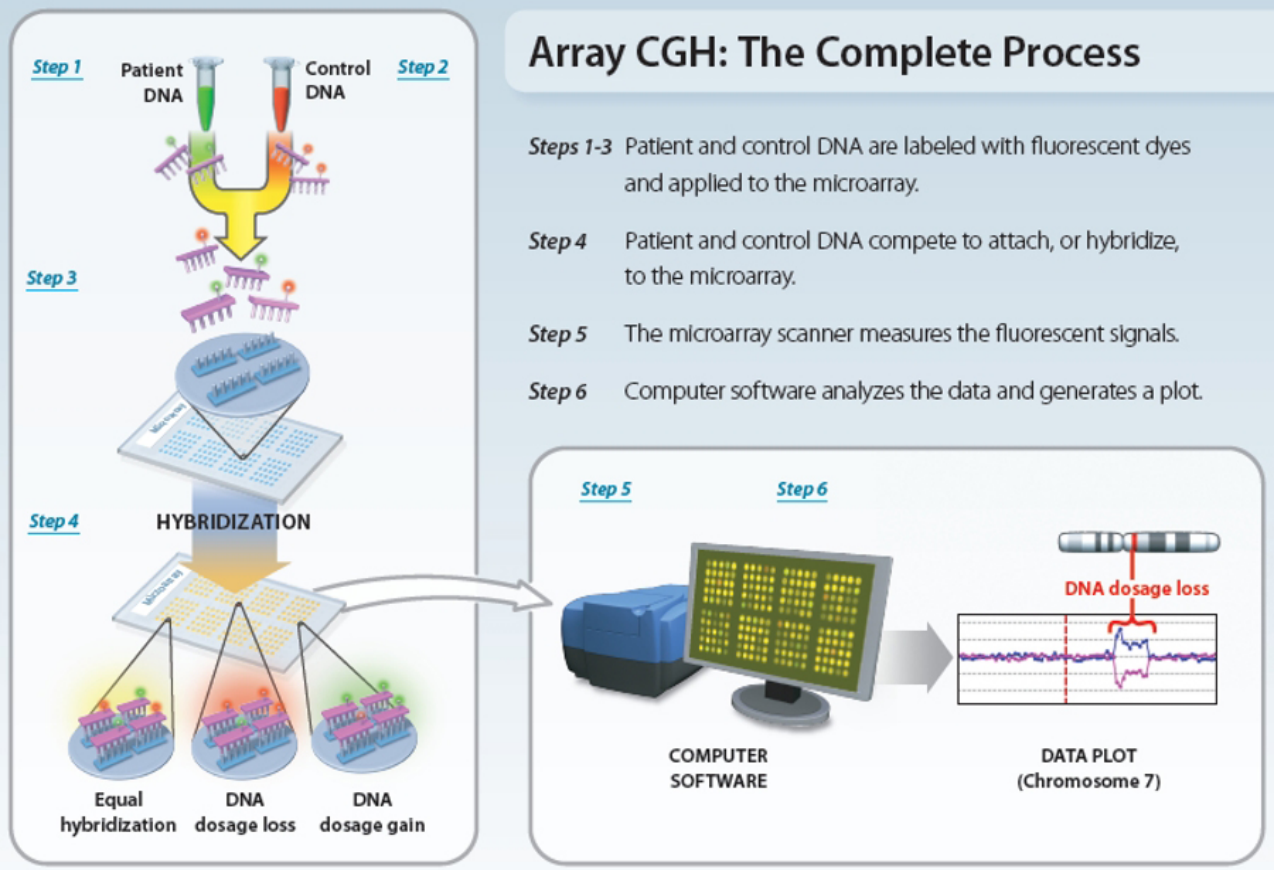
L'ibridazione genomica comparativa (Comparative Genomic Hybridization o CGH) è un approccio di citogenetica molecolare che ha il potenziale di individuare, a seconda della piattaforma utilizzata, sbilanciamenti nell'ordine delle 5-100 Kb (Kallioniemi A, 1992). E' basata sugli stessi principi della CGH classica, ma anziché utilizzare matrici su cui sono ibridati preparati metafasici normali, utilizza piattaforme in cui sono spottati cloni BAC di 150-160 Kb (*BAC arrays*) o sonde oligonucleotidiche di 10-100 bp (*oligo arrays*) corrispondenti a sequenze specifiche presenti solo in un determinato punto

del genoma, distribuiti su ogni singolo cromosoma, che nell'insieme coprono l'intero genoma umano. Alcune piattaforme utilizzano sonde equidistribuite sull'intero genoma, altre invece concentrano le sonde maggiormente nelle regioni codificanti, dove si vuole ottenere una maggiore sensibilità, mentre le sonde sono in numero molto inferiore nelle regioni non codificanti

Il metodo, raffigurato nella Figura 1, consiste nella co-ibridazione fluorescente di un DNA test ("campione") con un DNA di controllo ("reference"). Il rapporto tra le intensità di fluorescenza dei due fluorocromi utilizzati per marcare in maniera specifica campione e reference, riflette il rapporto quantitativo tra DNA test e DNA di controllo per ogni locus cromosomico considerato ibridizzato da una sonda. L'incremento e il decremento di questo rapporto espresso in scala logaritmica rappresenteranno, rispettivamente, la duplicazione (gain) e la delezione (loss) di una determinata regione cromosomica (Theisen, 2008). E' importante considerare che con questa tecnica in ogni punto del genoma corrispondente ad ogni sonda si sta quindi quantificando il numero di alleli (normale N=2, delezione <2, duplicato o amplificato >2) e non la loro sequenza. Il limite della tecnica è rappresentato dal fatto che non consente l'identificazione di riarrangiamenti cromosomici bilanciati, mosaicismi inferiori al 30%, quantificazione di regioni ipervariabili e mutazioni di singolo nucleotide.

La risoluzione genomica operativa dipende dalla lunghezza delle sonde utilizzate e dalla distanza tra una sonda e l'altra.

Figura 1. Diagramma del processo di CGH array (Theisen, 2008)



1.4 Analisi Bioinformatica dei CGH array

A dispetto delle differenti piattaforme array utilizzate nei diversi Laboratori, il workflow generale per determinare il ruolo di ciascuna CNV è simile. Il processo inizia con la scansione ad alta risoluzione dell'array, seguita dallo "storage" dei raw data e da una fase di pre-processamento. Questo step include la valutazione della qualità dell'immagine, la normalizzazione dei valori grezzi per rimuovere le distorsioni sistematiche e l'annotazione. La procedura di elaborazione converte le immagini grezze in valori relativi di intensità per ogni sonda presente sull'array. In seguito i dati sono normalizzati per rimuovere gli errori sistematici, così da lasciare solo i segnali rilevanti dal punto di vista biologico (Yang YH, 2001).

Identificati gli spot (gridding), è necessario separare il contributo del segnale specifico ("foreground") da quello del segnale di fondo ("background") tramite un processo noto come "segmentazione".

Lo step finale di questo processo consiste nell'identificazione delle CNV o "chiamata delle varianti", tramite algoritmi "proprietari" specifici di ogni ditta produttrice di piattaforme array-CGH

(ad esempio, l'algoritmo ADM-2 sviluppato dall'Agilent Technologies). Questo processo permette di categorizzare le differenti regioni del Genoma in uno dei quattro stati discreti possibili ossia "normali" (2 copie), "deleto" (<2 copie), "duplicato" (3–4 copie), "amplificato" (>4 copie). In questo modo i risultati di un esperimento di CGH sono trasformati in una forma più facilmente interpretabile. Questo processo ha anche il vantaggio di ridurre il numero di CNV, eliminando falsi positivi e diminuendo il numero di falsi negativi, rendendo l'analisi a valle più semplice e precisa.

Gli esperimenti di array-CGH richiedono quindi una serie complessa di operazioni computazionali per trasformare un'immagine ad altissima risoluzione composta da centinaia di migliaia di spot fluorescenti in dati discreti, comprensibili ed interpretabili. Gli sforzi compiuti dai bioinformatici nel corso degli anni hanno permesso di sviluppare tools "open source" come CGHPRO (Chen W, 2005), CAPweb (Liva S, 2006), CGH-Explorer (Lingjaerde OC, 2005), ma soprattutto hanno portato alla commercializzazione di software commerciali sempre più *"user friendly"*. Grazie a questi mezzi, oggi la maggior parte del lavoro non consiste più nella definizione delle CNV, quanto piuttosto nell'interpretazione dei dati ottenuti, ovvero nell'analisi della correlazione "genotipo-fenotipo". Anche in questo caso i grandi consorzi bioinformatici hanno messo a disposizione della comunità scientifica una serie di risorse come tools e database che permettono di portare a termine analisi dettagliate per rispondere ai più svariati quesiti biologici.

1.4.1 Interpretazione dei dati

I principali step per l'interpretazione dei dati di CGH sono comuni alla maggior parte dei laboratori: comparazione dei dati su database di controllo per lo studio della frequenza delle singole CNV; comparazione con dataset di individui affetti e analisi del contenuto genico e confronto con la Letteratura per cercare una correlazione genotipo-fenotipo. Queste risorse possono essere consultate seguendo una determinata pipeline, utilizzando sia dataset pubblici che locali. Significativo è il fatto che negli ultimi anni sono stati sviluppati pacchetti software che hanno reso possibile, almeno in parte, l'automatizzazione di questo processo. Però, se da un lato la possibilità di accedere ad un numero sempre più crescente di fonti ha permesso di diminuire significativamente il tempo di analisi, rimane altresì vero che in molti casi è necessaria una interpretazione "manuale" delle CNV.

I database disponibili per l'interpretazione delle CNV possono essere suddivisi in tre categorie:

- Database interni creati dal laboratorio stesso; questi variano in complessità da un foglio di calcolo excel a un database relazionale.

- Database specializzati ossia repository che raccolgono singoli casi o informazioni di controllo riguardanti i dettagli genetici e fenotipici, come ad esempio, una raccolta di CNV da una particolare popolazione di controllo o database specifici per una data malattia.
- Database internazionali di grandi dimensioni, in cui sono aggregati dati provenienti da diverse fonti. Le queries interrogano l'intero set di dati consentendo all'utente di visualizzare i risultati richiesti tra i numerosi a disposizione.

1.4.2 Il Database of Genomic Variants

I database di controllo sono risorse utili ed importanti per interpretare i dati array-CGH, in quanto permettono di effettuare un primo filtraggio di quelle varianti che molto probabilmente non svolgono un ruolo determinante nella definizione del fenotipo indagato, proprio perché presenti in forma polimorfica (cioè con alta frequenza) nella popolazione generale.

Il *Database of Genomic Variants* (DGV) (MacDonald JR, 2013) è una risorsa bioinformatica facilmente consultabile on-line (http://dgv.tcag.ca/dgv/app/home), molto utile sia per i laboratori clinici che di ricerca, che raccoglie le varianti comuni presenti nel DNA di soggetti sani o con fenotipi clinici non tali da richiedere una diagnosi ed un intervento medico. Questo database permette di capire se una determinata CNV rappresenta un evento raro e quindi potenzialmente patogeno, oppure se è "comune", poiché frequente nella popolazione generale (Iafrate AJ, 2004; Zhang J, 2006). Il DGV è stato creato subito dopo la pubblicazione dei primi articoli sulle CNV che descrivevano la loro presenza in individui clinicamente sani (Sebat J, 2004). Nella sua prima versione, il DGV comprendeva i dati di poche centinaia di individui per un totale di circa 1000 CNV e alcune inversioni (Zhang J, 2006). Questo Database è costantemente aggiornato con nuovi dati provenienti da studi di ricerca peer reviewed e molti degli studi iniziali sono stati rimossi, proprio a causa del continuo processo di aggiornamento. Dopo l'avvento degli array ad alta risoluzione e dei dati prodotti dal sequenziamento di nuova generazione, le "entries" sono notevolmente aumentate e migliorate in qualità. Oggi il DGV comprende dati provenienti da 72 studi che corrispondono a circa 552.586 regioni polimorfiche e 3164 regioni interessate da inversioni.

I dati sono disponibili sia in formato grafico, sia in forma di tabelle che di file di testo scaricabili. L'analisi sul browser si presenta con un'interfaccia grafica, che utilizza la piattaforma gmod/Gbrowse (Stein LD, 2002). I dati vengono visualizzati come traccia di diverso colore a seconda della tipologia della variante. E' possibile visualizzare altre annotazioni per consentire l'interpretazione dei dati nel contesto genomico. Queste includono tracce di annotazione standard come RefSeq e geni OMIM, duplicazioni segmentali (Bailey JA, 2002), sonde appartenenti a diverse piattaforme array e

varianti clinicamente rilevanti. Le opzioni di filtraggio sono state sviluppate per permettere una visione personalizzata del browser. E' anche possibile scaricare l'intero database del DGV in modo da poterlo "interrogare", usando i vari linguaggi di programmazione, mediante script ad hoc.

Come si diceva, l'obiettivo principale del database è fornire informazioni circa la frequenza con cui una determinata CNV è presente nella popolazione. Questo dato permette all'utilizzatore di classificare come "benigna" o "rara" ogni variante riscontrata in clinica e presente nel database. Come vedremo, la classificazione delle varianti in rare e benigne è solo uno dei vari steps che nel loro insieme costituiscono la pipeline interpretativa delle varianti. Per tale motivo, per comprendere il significato biologico di una CNV è necessario estendere ulteriormente l'analisi e comparare il fenotipo su altre risorse disponibili, come ad esempio DECIPHER (Firth HV, 2009), Human Phenotype Ontology (HPO) oltre che su database malattia-specifici come ad esempio SFARI (Abrahams BS, 2003) e AutismKb (Yang C, 2018) nel caso dell'autismo, ADHDgene (Zhang L, 2012) nel caso dell'ADHD e tanti altri ancora.

1.4.3 Database specializzati e Internazionali

Decipher (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources) è un database interattivo web-based (https://decipher.sanger.ac.uk/) che raccoglie dati genetici di pazienti affetti da vari tipi di patologie ed è quindi molto utile per valutare una CNV identificata in altri soggetti affetti. Decipher (Firth HV, 2009) contiene un potente motore di ricerca che permette di aggregare i dati in modo che possano essere rapidamente ottenute informazioni in base alle coordinate genomiche, banda citogenetica, o nome del gene da indagare. Incorpora risorse bioinformatiche innovative ed una traccia che mostra i collegamenti con le CNV descritte e le relative fonti bibliografiche. E' stato creato nel 2004 con lo scopo di fornire una risorsa bioinformatica disponibile per clinici e ricercatori impegnati nell'interpretazione di dati genomici provenienti da array CGH e facilitarne la collaborazione e lo scambio di dati, con lo scopo di accelerare i progressi per la definizione di nuove sindromi e funzioni per i diversi geni.

Oltre a DECIPHER ci sono altri database utili:

- **ECARUCA** (http://www.ecaruca.net) (Feenstra I, 2006) è un database online interattivo, dinamico che raccoglie e fornisce informazioni cliniche e molecolari dettagliate sulle aberrazioni cromosomiche sbilanciate rare. Memorizza solo gli squilibri genomici che sono considerati come causa del fenotipo clinico. Questa repository contiene dati provenienti da oltre 4800 pazienti per un totale di oltre 6600 aberrazioni, di cui 2500 sono alterazioni cromosomiche uniche.

- Il Consorzio **Isca** (International Standards for Cytogenomic Arrays Consortium, http://dbsearch.clinicalgenome.org/search/) ha creato una banca dati pubblica per classificare i pazienti con ritardi dello sviluppo, anomalie congenite e altri fenotipi con lo scopo di accelerare la comprensione della CNV nella popolazione clinica (Kaminsky EB, 2011). I dati all'interno del database Isca sono curati e continuamente rivisti. Periodicamente, l'intero database viene aggiornato da un gruppo di esperti che rivaluta le regioni con chiamate discordanti usando le informazioni raccolte dalla letteratura, da peer-review e dati di studio caso-controllo su grande scala.
- Infine, il database **OMIM** (Online Mendelian Inheritance in Man, https://omim.org/) è molto utile sia per studiare il ruolo funzionale dei geni contenuti nella CNV d'interesse, sia per vedere se un determinato gene è già stato associato a patologie note. A seconda della patologia studiata è poi molto utile la consultazione di database specifici oppure per le malattie rare di **ORPHANET** (https://www.orpha.net) (Pavan S, 2017)**.** La consultazione di questi database permette di stabilire se uno o più geni contenuti all'interno della CNV d'interesse sono presenti nelle liste dei geni candidati per un disturbo specifico. Molti database contengono inoltre annotazioni sugli studi che sono stati fatti su un determinato gene e sull'esistenza di eventuali modelli animali e/o cellulari. Ad esempio, nel caso dell'autismo, i database più aggiornati e ricchi d'informazioni sono quelli della Fondazione Simon "SFARIgene" (Abrahams BS, 2003) e il database "AutismKB" (Yang C, 2018) del Dipartimento di Bioinformatica dell'Università di Pechino (http://autismkb.cbi.pku.edu.cn/).

1.4.4 Classificazione delle Copy Number Variant

Dopo aver consultato questi ed eventualmente altri database, le CNV vengono classificate in 3 categorie principali (Kearney HM, 2011; Vermeesch JR, 2012):

- PATOLOGICHE quando è ben documentata in Letteratura l'associazione tra la stessa e sindromi da microdelezione/microduplicazione;
- BENIGNE quando sono state osservate in più dell'1% della popolazione generale e sono descritte, con lo stesso orientamento (duplicazione / delezione), in almeno tre soggetti riportati nel Database delle Varianti Genomiche (DGV);
- DI SIGNIFICATO INCERTO (VOUS: Variants of uncertain clinical significance) quando non rientrano nelle due classi precedenti. Vengono a loro volta ripartite in 3 sottoclassi:
    - VOUS propriamente dette, quando non sono mai state descritte e comprendono geni la cui funzione è sconosciuta, o sono state osservate sia in individui sani che affetti e non vi sono evidenze sufficienti per una classificazione più certa.

- VOUS VEROSIMILMENTE PATOLOGICHE quando sono riportate in uno o pochi casi con fenotipo simile, o si sovrappongono parzialmente con quelle riportate in individui affetti in cui il gene causativo non è ancora stato identificato, o non sono mai state riportate ma comprendono geni la cui funzione potrebbe essere causativa del fenotipo clinico.
- VOUS VEROSIMILMENTE BENIGNE quando non sono mai state descritte, ma vengono ereditate da un genitore sano, o non comprendono geni, oppure si sovrappongono parzialmente con quelle riportate in individui sani o sono state osservate solo in pochi di essi, o comprendono geni la cui funzione non è verosimilmente causativa del fenotipo clinico.

1.4.5 Il Genome browser

Dopo aver consultato varie risorse per classificare le CNV, può essere utile aumentare il dettaglio delle informazioni, in modo da ottenere chiarimenti sulla funzione, espressione e struttura di uno specifico gene.

Il **Genome Browser della UCSC** (Haeussler M, 2019) è una risorsa grafica online interattiva (https://genome.ucsc.edu/) che assolve pienamente questo scopo. E' attualmente il più usato dai laboratori perché facilmente accessibile, stabile ed ha collegamenti con tutti i principali database. Contiene al suo interno diversi database annotati con dati di espressione genica, associazioni tra geni e malattia, mappatura degli elementi ripetitivi del DNA, informazioni strutturali. L'utilizzatore può semplicemente inserire all'interno di una finestra iniziale il numero del cromosoma, le coordinate (inizio e fine) della regione che vuole visualizzare e la sequenza di riferimento del genoma umano al quale si rifà il referto array-CGH (ad esempio "build hg19"). Una volta visualizzata la regione d'interesse, sarà possibile accedere a tutte le informazioni riguardanti quella regione semplicemente cliccando sui vari link degli altri database. Al suo interno sono inoltre presenti "tools" e tabelle informative che permettono di applicare specifici algoritmi per mostrarne l'espressione, l'omologia di sequenza ed altre informazioni.

Nel complesso, questi database, quando vengono consultati possono risultare spesso molto lenti rendendo così l'analisi lunga e tediosa. Vi è quindi la necessità che i bioinformatici sviluppino delle risorse completamente integrate in grado di ricercare tra più database e fonti contemporaneamente e fornire output più esaustivi. Alcune delle banche dati pubbliche (ad esempio, Decipher) e dei prodotti commerciali (ad esempio, Cartagenia bench e Nexus DB ™) stanno iniziando a creare risorse che vanno verso questa direzione.

1.5 Copy Number Variant e trascrittomica

Le numerose risorse bioinformatiche sviluppate e descritte finora hanno permesso di standardizzare il work-flow di un'analisi dei dati di array CGH. Tuttavia alcune delezioni/duplicazioni devono essere validate con una seconda metodica indipendente come la Real-Time PCR. L'interpretazione dei dati non è univoca, una delezione o una duplicazione non necessariamente si traducono in una diminuzione o in un aumento dell'espressione, rispettivamente, e quindi non sempre hanno conseguenze funzionali. E' noto che le variazioni strutturali possono influenzare il fenotipo, compresa l'espressione genica, attraverso diversi meccanismi che dipendono dalla tipologia (delezioni/duplicazioni) e dal grado di sovrapposizione con la porzione codificante del gene e il suo promotore. Infatti, alcune CNV coinvolgono il gene interamente, altre si sovrappongono parzialmente ed altre infine sono presenti all'interno di regioni introniche (Schlattl A, 2011).

Sono stati descritti diversi meccanismi in grado di spiegare l'influenza che le CNV esercitano sull'espressione genica (Feuk L, 2006). Uno studio del 2007 ha indagato gli effetti delle CNV sull'espressione genica in linfoblasti umani rilevando che le CNV spiegano circa il 20% delle variazioni trovate. Questo è il risultato di un alterato dosaggio nei geni presenti all'interno della CNV, e probabilmente sui geni vicini (Merla G, 2006; Stranger BE, 2007). In circa il 2-15% dei geni, invece, si osserva una correlazione inversa, tra CNV ed espressione, per cui, ad esempio, una duplicazione, non si traduce in un aumento dell'espressione, ma in livelli di espressione normali o ridotti. Quest'ultima osservazione può essere spiegata: (a) da un loop a feedback negativo che riduce l'espressione della CNV; (b) dall'ingombro sterico di una copia extra di un gene che impedisce l'accesso a specifici fattori trascrizionali (Stranger BE, 2007; Henrichsen CN, 2009)

Una CNV può influenzare il trascrittoma non solo regolando l'espressione dei geni rigorosamente colocalizzati al suo interno (effetto in cis), ma anche, con un meccanismo di regolazione a distanza (effetto in trans), alterando l'espressione di geni distanti parecchie centinaia di Kb dai punti di rottura (Kleinjan DJ, 1998). Infatti, in uno studio è emerso che più della metà delle sonde di espressione che mostravano associazione con un clone CGH, mappavano numerose kb al di fuori dell'intervallo definito da una CNV (Stranger BE, 2007).

Una CNV, spesso, non ha alcun effetto sull'espressione: (a) a causa di un meccanismo di compensazione operato da altri geni coinvolti nel pathway; (b) per inclusione di elementi regolatori durante gli eventi di formazione della CNV. Meccanismi come l'imprinting, inoltre, possono anche inibire l'espressione del locus colpito da CNV (Sexton T, 2007). Infine le CNV possono alterare il fenotipo smascherando mutazioni recessive oppure influenzare l'espressione alterando la struttura dei trascritti (Reymond A, 2007).

La presenza, piuttosto che il cambiamento del numero di copie alleliche di una CNV, può avere effetti sorprendenti sull'espressione genica. Jacquemont et al. (Jacquemont S, 2011) hanno studiato l'impatto fenotipico dell'intervallo CNV 16p11.2; in particolare, i geni nell'intervallo della regione centromerica del riarrangiamento non hanno mostrato differenze significative tra i casi e i controlli, in netto contrasto con i geni telomerici dell'intervallo, che hanno mostrato variazioni significative. Questi ultimi geni sono stati, tuttavia, ugualmente sovra-regolati sia nelle delezioni che nelle duplicazione, suggerendo che la presenza dell'intervallo, piuttosto che il cambiamento del numero di copie, abbiano causato gli effetti sui livelli trascrizionali.

Anche se solitamente si tende a dare meno importanza alle alterazioni introniche, queste rappresentano la più frequente causa di variazione nei geni umani con circa 12.986 delezioni descritte, che colpiscono circa 4.147 geni (tra cui 1.157 geni essenziali e 1.638 geni associati a malattie). Le delezioni introniche possono risultare sia in un aumento che in una riduzione dell'espressione genica, mentre delezioni in regioni esoniche sono più comunemente associate ad una down-regulation (Rigau M, 2019).

Da un punto di vista meccanicistico, mutazioni in regioni non codificanti possono influire non solo sulla trascrizione in maniera allele-specifica, ma anche a livello post-traduzionale andando ad interferire con fenomeni rilevanti per i disturbi del neurosviluppo quali il funzionamento sinaptico (Zhou et al., 2019). In tal senso, esercitano una influenza convergente con quella delle varianti collocate nelle regioni codificanti, influenzando nei pazienti autistici caratteri fenotipici importanti quali il quoziente intellettivo (Zhou et al., 2019).

Poiché la lunghezza intronica può influenzare la formazione dei trascritti (Roy M, 2008) è possibile che una delezione possa influenzare la relativa porzione di trascritti alternativi in molti geni. Infine, dal momento che il riconoscimento degli introni e degli esoni da parte del macchinario di splicing si basa sul loro contenuto intronico di GC (Amit M, 2012; Gelfman S, 2013) è possibile che cambiamenti, a seguito di delezione di porzioni introniche, possano favorire la formazione di trascritti alternativi o di trascritti alterati.

Al fine di valutare se una CNV ha conseguenze funzionali e può pertanto svolgere un ruolo patogeno, può essere utile effettuare, nello stesso soggetto, oltre ad uno studio genomico, un profilo di espressione genica. Lo studio dell'espressione genica genome-wide, infatti, consente di misurare in un determinato tessuto o tipo cellulare, più spesso i globuli bianchi, la quantità di mRNA trascritto per ciascun gene presente nell'intero genoma e permette così di trovare gruppi di geni iper- o ipo-espressi. In alternativa, sarà possibile correlare l'alterazione o mediante un approccio genome-wide, oppure si andrà a vedere se i geni contenuti nel CNV mostrano una espressione alterata.

Questi due approcci possono permettere di chiarire se e fino a che punto una duplicazione/delezione a livello genomico modifica i livelli di espressione soprattutto per quelle malattie in cui è difficile spiegare, con il solo dato genomico, il fenotipo clinico.

La necessità di studiare gli effetti delle CNV sull'espressione genica e la loro manifestazione a livello del fenotipo clinico stimola lo sviluppo e l'implementazione di risorse bioinformatiche e metodi statistici in grado di correlare il dato genomico con quello trascrittomico. Ad oggi, nonostante si tratti di una tematica di grande attualità e con importanti ricadute potenziali, non vi è univocità rispetto all'approccio bioinformatico da seguire.

1.6 Metodiche per lo studio del trascrittoma

L'analisi del trascrittoma può essere condotta secondo due tecnologie: (a) l'ibridazione su microarray e (b) il sequenziamento dell'RNA (tabella 1).

(a) L'ibridazione si fonda sulla proprietà dei nucleotidi di appaiarsi con sonde complementari fissate su un supporto solido. A questa categoria appartengono i microarray, da anni largamente utilizzati per ottenere informazioni sull'espressione genica (Gonzalo R, 2018). Quest'ultimi sono costituiti da un supporto solido a cui sono ancorate delle sonde di DNA, in numero molto elevato per ogni gene e disposte in posizioni note, come già descritto per gli array-CGH. L'RNA estratto dalla cellula viene retrotrascritto in cDNA, marcato con molecole fluorescenti e ibridizzato con il microarray. L'intensità della fluorescenza è una misura di quante molecole di cDNA si sono legate alle sonde, ovvero quanto il gene associato ad ogni singola sonda è espresso nella cellula.

(b) Una tecnica più recente per la quantificazione del trascrittoma, basata sulle tecnologie di sequenziamento *NGS (Next Generation Sequencing)* è il sequenziamento dell'*RNA (RNA-seq*) (Wang Z, 2009) che rispetto agli array presenta diversi vantaggi. Infatti, consente la caratterizzazione dei trascritti senza una conoscenza a priori dei siti di inizio della trascrizione, rendendo possibile la scoperta di trascritti non noti o poco espressi. Permette di ottenere una risoluzione a livello di singoli nucleotidi e mostra maggior capacità di distinguere tra le varie isoforme (biotipi) di RNA, di determinare l'espressione allelica e di rilevare le variazioni di sequenza. I livelli di espressione sono dedotti dal numero totale di sequenze lette (*"reads"*) che mappano su un determinato gene, normalizzate per la lunghezza degli esoni che possono essere mappati in modo univoco. Inoltre, l'intervallo dinamico dell'RNA-seq per determinare i livelli di espressione è di 3-4 ordini di grandezza maggiore rispetto ai 2 ordini di grandezza degli array di espressione (Voelkerding KV, 2009).

Negli ultimi anni, grazie alla diffusione di software sempre più user-friendly e di tools bioinformatici gratuiti, l'analisi dei dati di microarray di espressione è diventata molto più semplice,

lineare e standardizzata. Per l'analisi dei dati di RNA-seq, invece, nonostante siano disponibili numerosi metodi di analisi, non c'è ancora un protocollo univoco. Per tali motivi questa metodica richiede ancora grande esperienza, risorse informatiche e competenze bioinformatiche specifiche che derivano anche dalle molteplici applicazioni possibili.

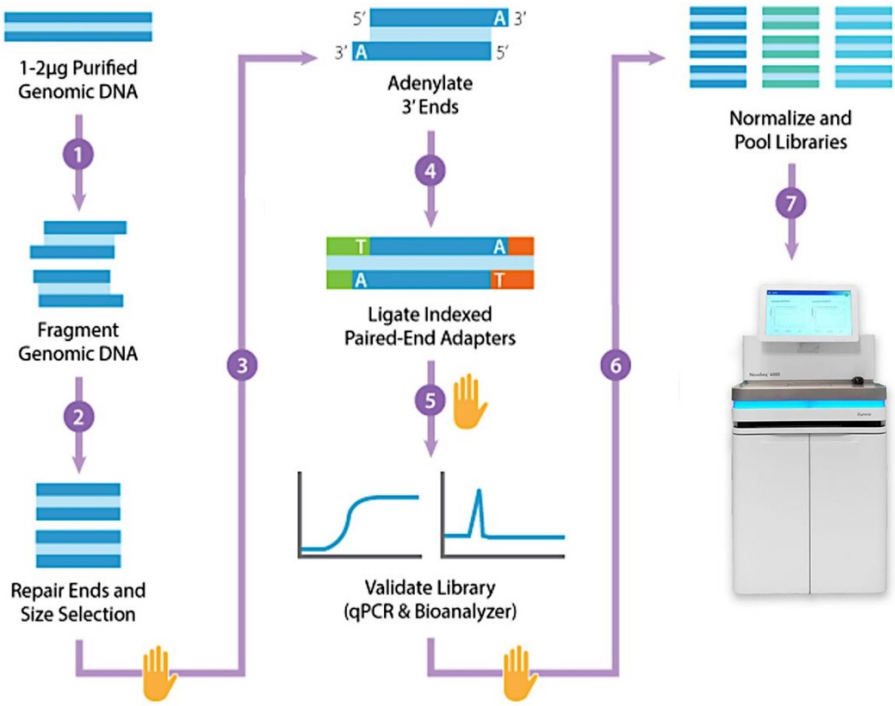
Tabella 1. Confronto tra microarray di espressione ed RNA-seq

Microarray	Rna-seq
Vantaggi - Protocollo ben definito - Pipeline di analisi consolidata - Approccio standardizzato per la sottomissione dei dati - Costi relativamente bassi	Vantaggi - Range altamente dinamico (non va in saturazione) - Non necessita di una conoscenza a priori della sequenza - Allineamento diretto della sequenza, non necessita di ibridizzazione; - Permette di trovare eventi di splicing; - I geni paraloghi possono essere evidenziati; - Può essere usata per l'identificazione di SNP - Approccio standardizzato - Costi di sequenziamento in calo
Svantaggi - Analisi solo per sequenze predefinite - Range dinamico limitato dalla risoluzione dello scanner - Basata sull'ibridazione - Ibridazione potenzialmente non specifica - Potrebbe non dare informazioni sui geni paraloghi - Alta varianza per geni poco espressi - Non identifica gli eventi di splicing	Svantaggi - Richiede specifiche risorse computazionali - Costi per l'allestimento della strumentazione elevati - Analisi delle varianti di splicing molto complessa - L'analisi può essere complessa se sono presenti geni paraloghi

1.7 L'Rna-Seq

L'Rna-seq rappresenta un insieme di esperimenti e metodiche computazionali per determinare l'identità e l'abbondanza delle sequenze di RNA in un campione biologico. Il protocollo prevede una parte svolta in laboratorio ed una parte di analisi ed interpretazione bioinformatica. Varia in base alla tecnologia utilizzata ma è comunque possibile descriverne in linea generale i passaggi principali (Wang Z, 2009).

Figura 2. Diagramma del processo di Rna-seq eseguito in laboratorio (www.illumina.com, modificato)



I campioni di RNA possono essere estratti da cellule, tessuti o sangue mediante kit commerciali, come ad esempio RNeasy® (Qiagen) e TRIzol™ (Thermo Fisher Scientific) che sono facili da usare e consentono rese elevate. Anche se questi kit sono specifici per eliminare contaminazioni da DNA, tracce residue potrebbero interferire con la quantificazione degli acidi nucleici compromettendone gli step successivi e i risultati. E' pertanto necessario trattare l'RNA estratto con una DNAsi e testarne la qualità, per valutare degradazione, purezza e concentrazione. L'RNA totale estratto dovrebbe mostrare chiaramente le bande 28S e 18S in un rapporto di 2:1. L'Agilent Technologies ha sviluppato vari anni fa uno strumento, il Bioanalyzer, che permette di attribuire un parametro qualitativo all'RNA, definito come R.I.N. (RNA Integrity Number). Tale parametro presenta una scala di valori che va da 0, indice di completa degradazione, a 10 RNA di ottima qualità (Dundar F, 2018). Non si consiglia di utilizzare per esperimenti di RNA-seq campioni di RNA con R.I.N. inferiore a 7.

Dopo averne testata la qualità, l'RNA è convertito in una libreria di cDna tramite retrotrascrizione che rappresenta tutte le molecole di RNA presenti nel campione. Questo step è necessario perché l'RNA non è direttamente sequenziabile, mentre il retrotrascritto lo è. La preparazione della libreria prevede vari passaggi in cui gli RNA vengono frammentati, ibridati e legati tramite una RNA ligasi, ad una miscela di adattatori. Successivamente le molecole di RNA, legate agli adattatori, vengono retrotrascritte in cDna a singolo filamento e purificate utilizzando un sistema di

cattura con biglie magnetiche. La libreria di cDna viene amplificata mediante PCR per ottenere la quantità di campione richiesta per i successivi passaggi di lavorazione e, al contempo, per permettere l'aggiunta delle sequenze terminali necessarie per il sequenziamento.

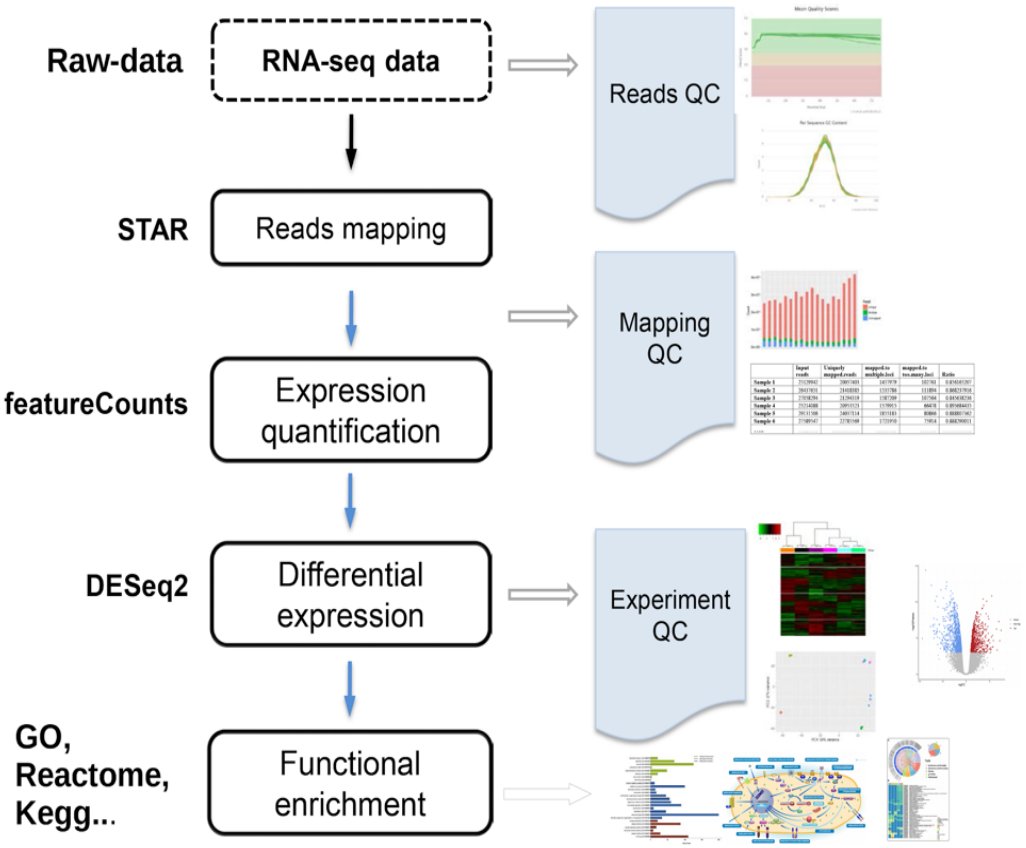
In seguito, i frammenti sono ibridizzati ad una flow cell composta di "lane indipendenti", sulla cui superficie sono immobilizzati due diversi oligonucleotidi con sequenza complementare a quella degli adattatori della libreria. I frammenti di cDNA della libreria sono immessi sulla piastra di sequenziamento, permettendo l'ibridizzazione tra i loro adattatori e gli oligonucleotidi della piastra. Il legame si forma ad entrambe le estremità dei frammenti, che sono così immobilizzati sulla superficie della flow cell, assumendo una forma "a ponte". Dopo l'immobilizzazione ha inizio il processo di amplificazione tramite PCR "a ponte" ("bridge PCR") da un primer all'altro. Al termine di vari cicli di PCR si vengono a formare sul vetrino circa 120 milioni di cluster, ciascuno contenente a sua volta milioni di frammenti di DNA identici. Nello step successivo viene eseguito l'annealing del primer di sequenziamento ai frammenti di ogni cluster che permette l'avvio della reazione di sequenziamento vera e propria. In questo modo è possibile sequenziare contemporaneamente fino a 120 milioni di frammenti di DNA per ogni lane. Il sequenziamento avviene attraverso l'incorporazione di nucleotidi (dNTPs) a cui è aggiunto un terminatore reversibile. Il terminatore è una molecola che blocca il gruppo ossidrile impedendo l'ulteriore allungamento della catena nucleotidica di nuova sintesi dopo l'incorporazione di un nucleotide. Esso è detto "reversibile" poiché può essere dissociato chimicamente, riattivando la sintesi al ciclo successivo. Dopo ogni incorporazione, un laser eccita il fluoroforo del dNTP incorporato generando un'emissione luminosa che ne permette l'identificazione. Quindi il terminatore viene rimosso, e il sequenziamento prosegue con l'aggiunta della base successiva.

In un esperimento di RNA-Seq più un gene è espresso, più numerose sono le copie del trascritto genico, più numerosi sono i frammenti di cDNA generati. Quindi il numero di reads prodotte è direttamente proporzionale all'espressione del gene corrispondente (Korpelainen E, 2014).

1.8 Analisi Bioinformatica dei dati di RNA-seq

L'analisi bioinformatica dei dati provenienti da un esperimento di Rna-seq comprende diverse fasi che possono essere eseguite con tool diversi e prevedono i seguenti passaggi generali: controllo di qualità, allineamento delle reads ad un genoma o trascrittoma di riferimento, stima dei livelli di espressione genica individuale, normalizzazione e identificazione dei geni differenzialmente espressi (Dundar F, 2018).

Figura 3. Analisi bioinformatica dei dati di Rna-seq



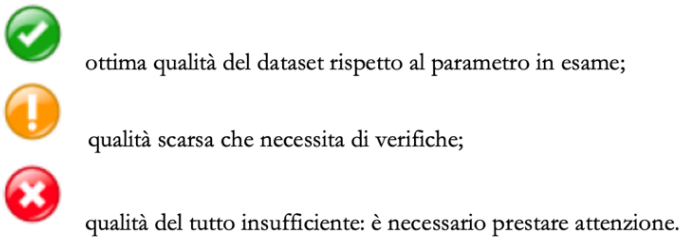
1.8.1 Controllo di qualità

I files ***BCL*** (Binary Base Call) rappresentano i dati di partenza dell'analisi bioinformatica. Sono generati direttamente dal sequenziatore dell'Illumina. Il software Real Time Analysis (RTA), durante ogni ciclo di sequenziamento, scrive le singole basi e l'affidabilità nella chiamata come un quality score in un file in formato *.bcl*. Successivamente, mediante l'utilizzo di software specifici come CASAVA o di bcl2fastq viene generato il file FASTQ che racchiude sia la sequenza grezza che il suo quality score. I file FASTQ saranno utilizzati per le successive fasi dell'analisi.

Questo primo step, prevede, attraverso dei software ad hoc, un controllo generale della qualità delle reads generate. Infatti, il *"base calling",* è un processo ad opera dello strumento di sequenziamento NGS, che associa ad ogni nucleotide letto un valore di probabilità per ogni base azotata. Poiché la stessa sequenza viene letta più volte, per ovviare alla mancanza di accuratezza nelle letture e a valori di probabilità non soddisfacenti, è necessario compiere un controllo di qualità per confermare che i frammenti siano stati letti correttamente.

Esistono numerosi programmi per compiere questo tipo di analisi, e uno dei più completi è **FastQC** (Andrews, 2010). Questo tool permette di eseguire, in modo semplice, alcuni controlli di qualità su dati di sequenze provenienti da pipeline di sequenziamento high-throughput. Fornisce una serie di analisi che è possibile, utilizzare per capire se i dati presentano problemi di cui è necessario essere a conoscenza prima di effettuare gli step successivi. L'output è molto dettagliato e permette di spiegare anche le possibili cause della scarsa affidabilità dei dati. Il programma riceve come input un file in formato *.fastq* contenente sia la sequenza nucleotidica sia il corrispondente *"quality score"* e valuta, tramite opportuni algoritmi, differenti aspetti per il controllo della qualità complessiva dei frammenti. In particolare considera: la qualità delle reads delle singole basi; la verosimiglianza nell'assegnazione di ciascuna base della sequenza al nucleotide scelto; le duplicazioni delle sequenze. Per ogni voce, poi, il programma crea dei grafici per ogni singolo controllo di qualità eseguito e secondo opportuni standard, assegna automaticamente un simbolo (fig.4) che indica il livello di qualità associato: "ottima qualità del dataset rispetto al parametro in esame"; "qualità scarsa che necessita di verifiche"; "qualità del tutto insufficiente: è necessario prestare attenzione".

Figura 4. Simboli usati per indicare la qualità dei dati con FastQC



Bisogna prestare molta attenzione nell'interpretazione dei vari parametri presenti in un report di qualità. Infatti, alcuni scostamenti potrebbero essere causati dalla natura stessa del dato (es. aumento di CG o di sequenze duplicate) piuttosto che da problemi sperimentali (Dundar F, 2018).

MultiQC (Ewels P, 2016) è un tool scritto in linguaggio Python che può essere lanciato da riga di comando. Cerca in modo ricorsivo dentro le directory file *.log*, analizza le informazioni già presenti prodotte da altri tool e le aggrega generando un singolo output, per più campioni, in formato HTML. Rispetto alla maggior parte degli altri tool che producono report dettagliati su un singolo campione, MultiQC presenta il grande vantaggio di unire i risultati di analisi proventi da diversi tool, permettendo così all'utente di scegliere e visualizzare le statistiche di interesse. I grafici condivisi consentono un confronto accurato tra i campioni, consentendo il rilevamento di differenze sottili non evidenti quando

si passa da un file all'altro. La visualizzazione dei dati aiuta il rilevamento di effetti batch e minimizza il rischio di fattori confondenti che influenzano i risultati dello studio.

RSeQC (Wang L, 2012) è un pacchetto scritto in Python che contiene un numero di moduli utili per poter valutare in modo completo i dati di RNA-seq. I "moduli di base" analizzano rapidamente problemi nella qualità generale, come la composizione nucleotidica della sequenza, artifizi di PCR e contenuto in GC, mentre i "moduli specifici dell'RNA-seq" studiano il profilo e la distribuzione delle reads mappate, l'uniformità della copertura di un gene, la riproducibilità, la specificità del filamento e l'annotazione delle giunzioni di splicing. In base ai risultati ottenuti, è possibile decidere se "ripulire" il dataset dalle reads di scarsa qualità, per facilitare e rendere più accurate le analisi successive. Le possibili operazioni sono:

- Scartare le estremità di bassa qualità;
- Eliminare le reads con bassa qualità media;
- Eliminare le reads che dopo le operazioni di trimming rimangono troppo corte;
- Rimuovere gli adattatori.

Al fine di eseguire una o più delle procedure elencate sopra, è necessario applicare un processo che prende il nome di *"**trimming delle reads**"*, che consiste nell'accorciare i frammenti. Si possono utilizzare due strategie di trimming:

1) Trimming statico: si tagliano tutte le reads nello stesso punto;

2) Trimming dinamico (flessibile): le reads sono tagliate sia all'estremità 5' sia all'estremità 3', finché la qualità resta sotto un valore di soglia definito. Le reads, quindi, non avranno più tutte la stessa lunghezza. Il vantaggio di questa seconda tecnica è quello di ottenere reads più corte, ma di maggiore qualità media. **Trimmomatic** (Bolger AM, 2014), è uno strumento capace di leggere e tagliare i dati prodotti in formato *.fastq* e di rimuovere gli adattatori aggiunti nella generazione della libreria. Se questi adattatori non sono rimossi, possono causare un assemblaggio non corretto.

1.8.2 L'allineamento delle reads

In un esperimento di RNA-Seq, le reads rappresentano i dati grezzi dai quali ricavare l'informazione sul livello di espressione dei geni nel campione. Più numerose sono le copie di un trascritto in un campione, più probabilità avrà quel trascritto di essere sequenziato e quindi di generare reads. Per quantificare il numero di reads riferite a ciascun gene, le reads di ogni campione devono essere mappate a un genoma o trascrittoma di riferimento. La fase di allineamento presenta un aspetto critico: idealmente si vorrebbe trovare l'univoca posizione nel genoma in cui il riferimento sia identico alla reads. In realtà, il reference non sarà mai una rappresentazione perfettamente identica del campione

biologico, a causa di errori di sequenziamento e/o dell'imperfetta similarità tra la sequenza del campione d'interesse e quella del riferimento. La mappatura delle reads provenienti da un esperimento di RNA-seq, comporta, inoltre, una sfida informatica aggiuntiva che deriva dalla struttura non contigua dei trascritti lungo le coordinate genomiche come conseguenza dell'eliminazione degli introni dovuti al processo di splicing. Lo scopo dell'allineamento diventa così l'identificazione della posizione del genoma in cui ogni read ottiene il miglior appaiamento con il reference (Ye H, 2015).

I diversi programmi di allineamento usano algoritmi euristici e differiscono per il metodo di indicizzazione che usano per rendere rapido il processo. Gli allineatori che usano le tabelle di hash presentano il vantaggio di essere facilmente estendibili per rilevare le differenze tra reads e reference, ma impongono una notevole richiesta computazionale. I metodi basati sulla trasformata di Burrows Wheeler (BWA), possono mappare reads che corrispondono bene col reference, ma sono molto lenti quando sono presenti allineamenti complessi. Gli allineatori differiscono anche nel modo di maneggiare le reads che mappano in maniera identica a diverse localizzazioni (multimaps). Pertanto, alcuni allineatori eliminano queste reads, (Langmead B, 2009), altri le collocano in maniera casuale (Li H, 2008), altri sulla base di una stima del coverage locale (Mortazavi A, 2008). Tra i software più utilizzati e standardizzati abbiamo Bowtie, BWA, TopHap, STAR. Tuttavia, sta aumentando l'utilizzo di altri software come Kallisto e Salmon, che si basano sui nuovi concetti di pseudo-allineamento e quasi-mapping. Questi software risultano essere più veloci e hanno meno necessità computazionale

Bowtie è un software gratuito, open source, che si può facilmente scaricare al seguente indirizzo web http://bowtie.cbcb.umd.edu (Langmead B, 2009). E' uno dei software più utilizzati per l'allineamento in quanto molto veloce e richiede per il suo utilizzo poca memoria. Bowtie allinea brevi sequenze di DNA al genoma umano ad una velocità di 25 milioni di reads di circa 35 bp all'ora. Indicizza il genoma usando la trasformata di Burrows-Wheeler per mantenerlo in memoria. Bowtie estende la trasformata di Burrows-Wheeler con un algoritmo di ricerca quality aware che consente di utilizzare i mismatch. È possibile inoltre utilizzare contemporaneamente più processori per ottenere una maggiore velocità di allineamento.

TopHat è un programma per l'allineamento delle reads di RNA-Seq ad un genoma di riferimento utile per identificare i siti di splicing (Trapnell C, 2009). È basato sul programma di mappatura Bowtie e funziona su piattaforme Linux e OS X.

STAR è un software open source che può essere lanciato su piattaforme Linux, Unix e McOS X (Dobin A, 2013). Permette un allineamento accurato e rapido delle reads. Il suo workflow prevede un primo step in cui prende in input il genoma di riferimento in formato *.FASTA* e un file *.GTF* di annotazione per generare un genoma indicizzato che sarà poi utilizzato nel processo di mappatura successivo. Nel secondo step, STAR mappa le reads al genoma indicizzato e fornisce come output

diversi file come il file di allineamento sotto forma di *BAM/SAM*, un sommario delle reads mappate e non mappate e dei file di log. Questa fase è controllata da una serie di parametri opzionali scelti dall'utente. I risultati ottenuti dalle operazioni di allineamento e assemblaggio sono archiviati in file di output in formato *SAM (Sequence Alignment Map) e BAM (Binary Alignment Map),* che saranno usati negli step successivi.

Kallisto (Bray, 2016) è un programma per quantificare l'abbondanza dei trascritti dei dati di RNA-Seq, e più in generale dei dati provenienti da esperimenti high-throughput. Si basa sulla nuova idea di *pseudo-allineamento* per determinare rapidamente l'accuratezza delle reads con il target, senza necessità di allineamento. Kallisto può quantificare 30 milioni di reads in meno di 3 minuti su un computer desktop Mac utilizzando solo le reads e un trascrittoma di riferimento che viene costruito in meno di 10 minuti. Lo pseudo-allineamento conserva le informazioni principali, per cui kallisto non è solo veloce, ma anche molto accurato.

Salmon (Patro, 2017) è uno strumento per quantificare l'abbondanza dell'espressione dei trascritti. Utilizza nuovi algoritmi basati sul concetto di quasi-mapping per fornire stime di espressione accurate in maniera rapida, utilizzando poca memoria. Salmon opera in due fasi: indicizzazione e quantificazione. Il passaggio dell'indicizzazione è indipendente dalle letture e deve essere eseguito solo una volta per un determinato set di trascrizioni. Il passaggio di quantificazione, invece, è specifico dell'insieme delle reads di RNA-seq.

1.8.3 Quantificazione dell'espressione genica

Dopo aver determinato le posizioni delle reads su un DNA di riferimento, è possibile contare il numero di reads allineate ad un determinato gene, trascritto o esone. La stima del livello di espressione genica/trascritto è l'applicazione più comune degli studi di RNA-seq. L'idea alla base della quantificazione a livello di espressione è che il numero di reads che mappa su ogni sequenza trascritta è una valida stima del livello di espressione. Diversi tools come **HTSeq, BEDTools, Qualimap, featureCounts** sono stati sviluppati. In generale, questi programmi prendono come input uno o più file SAM/BAM e un file di annotazione in formato *GTF, GFF o BED* contenente le coordinate cromosomiche e forniscono in output il numero di reads assegnate a ciascuna features o meta features. Differiscono tra loro nel modo di maneggiare le reads che mappano a loci multipli. Ad esempio HTSeq le ignora, Qualimap divide equamente le conte tra le differenti regioni di appaiamento, mentre Cufflink contiene un'opzione per dividere ciascuna reads sulla base dell'abbondanza del gene su cui mappa.

I counts prodotti in questo step rappresentano i dati finali in un esperimento di RNA-seq e sono memorizzati in matrici in cui le righe rappresentano i geni e le colonne i campioni sequenziati. Per loro

stessa definizione, dal punto di vista statistico, *i counts rappresentano una somma di eventi aleatori indipendenti* (la mappatura di ogni read sui geni). Possono quindi essere descritti da una variabile aleatoria che segue una determinata distribuzione statistica. I due modelli di distribuzione più utilizzati per descrivere i counts sono il modello di Poisson e il modello Binomiale Negativo.

1.8.4 Normalizzazione

Alcuni lavori (Bullard JH, 2010; Robinson MD, 2010) hanno dimostrano che la normalizzazione è uno step essenziale per l'analisi dell'espressione differenziale. La normalizzazione è il processo di ridimensionamento dei valori dei conteggi non elaborato per tenere conto dei fattori "non interessanti". Il suo scopo è quello di eliminare gli effetti sistemici che non sono associati alle differenze biologiche di interesse. Oltre ad essere essenziale per l'analisi delle espressione differenziale, la normalizzazione è anche utile per l'analisi esplorativa e la visualizzazione dei dati, ogni volta che si esplorano o confrontando i conteggi tra o all'interno dei campioni (Evans C, 2017). I principali fattori da considerare durante la normalizzazione sono: la lunghezza dei geni, la profondità del sequenziamento e la composizione dell'RNA. Normalizzare per **la lunghezza del gene** è necessario per confrontare l'espressione tra geni diversi all'interno dello stesso campione. **La profondità di sequenziamento**, invece, serve per il confronto dell'espressione genica tra i campioni. Corrisponde al numero di round di sequenza che vengono effettuati. Infatti, più sono i rounds (maggiore è la profondità) più è possibile, in RNA-seq, identificare e quantificare RNA a bassa e bassissima espressione. Quando, invece, si sequenzia il Dna, una maggiore profondità di sequenziamento permette di identificare percentuali di mosaicismo molto basse. Infine, geni espressi in modo altamente differenziato tra i campioni, differenze nel numero di geni espressi tra i campioni o la presenza di contaminazioni possono distorcere i risultati di alcuni metodi di normalizzazione. E' quindi utile tenere conto della **composizione dell'RNA** per avere un confronto accurato dell'espressione tra i campioni ed è particolarmente importante quando si eseguono analisi di espressione differenziale.

Le unità utilizzate più frequentemente per misurare l'espressione genica dai dati di RNA-seq sono: (a) RPKM, ossia reads-per-kilobase per milione di reads mappate; (b) FPKM, ossia frammenti-per-kilobase per milione di reads mappate; (c) CPM, conta per milione (CPM). Questi metodi, implementati nel pacchetto R edgeR permettono di rimuovere gli effetti dovuti alla profondità totale del sequenziamento e alla lunghezza del gene. In tabella 2 un riassunto dei metodi descritti e del loro utilizzo.

Tabella 2. Metodi di normalizzazione dei counts

Metodo di normalizzazione	Descrizione	Fattori considerati	Utilizzo
Conte per milione (CPM)	Conteggi normalizzati in base al numero totale di letture	Profondità di sequenziamento	- Confrontare le conte **tra** replicati dello stesso campione; - NON per confronti tra campioni o per analisi di Espressione Differenziale
Trascritti per milione di kilobasi(TPM)	Conteggi normalizzati per lunghezza della trascrizione (kb) e per milione di letture mappate	Profondità di sequenziamento e lunghezza del gene	- Confrontare le conte geniche **all'interno** di un campione o tra campioni dello stesso gruppo; - NON per analisi di Espressione Differenziale
Letture/frammenti per kilobase di esone per milioni di reads/ frammenti mappati (RPKM/FPKM))	Simile al TPM	Profondità di sequenziamento e lunghezza del gene	- Confrontare le conte dei geni tra geni **all'interno** di un campione; - NON per confronto tra campioni o analisi di Espressione Differenziale
La mediana dei rapporti di DESeq2	Conteggi diviso fattori dimensionali campione-specifici, ossia rapporto tra conteggi (N) e la media geometrica del campione per ogni singolo gene	Profondità di sequenziamento e composizione dell'RNA	- Confrontare le conte dei geni tra campioni e analisi dell'espressione differenziale; - NON per confronti tra campioni
EdgeR's trimmed mean of M values (TMM)	Utilizza una media "trimmed" ponderata dei log expression ratios tra i campioni	Profondità di sequenziamento, lunghezza del gene e composizione dell' RNA	- Confrontare le conte dei geni tra e all'interno di campioni; - Analisi di Espressione Differenziale

1.8.5 Analisi dell'espressione genica differenziale

Lo scopo dell'analisi di espressione differenziale è quello di identificare geni il cui livello di espressione cambia tra le condizioni oggetto di studio (ad esempio, casi vs controlli, pre- e post-trattamento, ecc). Si deve quindi usare un test statistico per decidere se, per un dato gene, una differenza osservata nelle conte delle reads tra due condizioni è statisticamente significativa, cioè se è maggiore o minore di quello atteso, se fosse dovuta unicamente ad una variazione casuale. Pertanto, un aspetto cruciale è rappresentato dalla procedura statistica usata per individuare i geni differenzialmente espressi. Diversi metodi statistici sono disponibili per testare l'espressione differenziale tra le diverse condizioni oggetto di studio (Costa-Silva J, 2017)

Tutti i principali metodi, tranne il pacchetto R Limma (Ritchie ME, 2015), lavorano direttamente sulle conte DESeq2 (Love MI, 2014), edgeR (Robinson DM, 2010) , baySeq (Hardcastle TJ, 2010), EBSeq (Leng N, 2013), ShrinkSeq (Van de Wiel MA, 2012), NOISeq (Tarazona S, 2015), e SAMseq (Li J, 2013). Questi metodi possono essere divisi, a loro volta in:

• parametrici, come baySeq , edgeR, DESeq2, EBSeq che usano un modello Binomiale Negativo (NB) per tenere conto della sovra-dispersione. DESeq2 ed edgeR prevedono un classico test d'ipotesi mentre EBSeq e baySeq utilizzano un approccio Bayesiano (Soneson C, 2013).

• non parametrici come SAMseq, NOISeq che non assumono alcuna distribuzione a priori dei dati, ma classificano i geni in base all'espressione e usano test e permutazioni randomizzate di queste liste per identificare i geni differenzialmente espressi (Soneson C, 2013).

Non c'è un consenso univoco riguardo la miglior procedura da seguire ma i risultati di alcuni studi comparativi, hanno permesso di definire una serie di raccomandazioni da seguire nella scelta della metodica appropriata. In primo luogo, bisogna prestare cautela quando il numero di replicati o il livello dei geni espressi è piccolo, (Robles JA, 2012). Tra i vari tools a disposizione, Limma ha dimostrato di funzionare bene in diversi disegni sperimentali ed è molto veloce da eseguire. DESeq e edgeR lavorano in maniera simile, ma in riferimento al controllo dell'errore statistico di tipo I mediante False Discovery Rate (FDR, definito come la proporzione attesa del numero di risultati falsi positivi sul totale di tutti i risultati positivi o, in altri termini, del numero di ipotesi nulle erroneamente rifiutate sul totale di quelle rifiutate), sono rispettivamente, troppo conservatori o troppo liberali (Nookaew I, 2012; Robles JA, 2012). SAMseq funziona bene in termini di FDR e presenta una buona sensibilità quando il numero di replicati è relativamente alto, almeno 10 o più (Nookaew I, 2012). Altre accortezze da tener presente nella scelta della metodologia sono rappresentate dalla facilità di installazione e utilizzo dei relativi software, dai requisiti computazionali e dalla qualità della

documentazione e delle istruzioni. Infine, un aspetto cruciale quando si sceglie un metodo di analisi è il disegno sperimentale. Infatti, mentre alcuni degli strumenti di espressione differenziale possono eseguire solo un confronto tra coppie, altri come edgeR (Robinson DM, 2010), Limma-Voom (Law CW, 2014), DESeq (Anders S, 2010), DESeq2 (Love MI, 2014) possono eseguire più confronti, includere diverse covariate o analizzare dati provenienti da studi con time-point multipli.

1.8.6 Interpretazione biologica dei dati

Una volta identificati i geni differenzialmente espressi in maniera significativa nei gruppi sperimentali e di controllo, si passa all'ultima fase dello studio, che tiene conto del ruolo biologico dei geni individuati. Esistono numerosi strumenti bioinformatici che permettono di indagare la funzione dei geni e i processi biologici nei quali sono implicati. Il più diffuso e utilizzato è sicuramente il database della "**Gene Ontology**" (GO) (Gene Ontology Consortium, 2004) che raccoglie tutti i dati e le informazioni inerenti i geni e i loro trascritti. La Gene Ontology è un sistema di classificazione che raccoglie migliaia di termini che descrivono i geni e i loro prodotti. Tali termini sono suddivisi in tre grandi categorie: funzione molecolare (MF), processi biologici (BP), e componenti cellulari (CC). La prima categoria raggruppa tutti i geni in base alla funzione molecolare dei loro prodotti e ha lo scopo di descriverne semplicemente l'attività cellulare. I processi biologici, invece, sono più complessi poiché comprendono una o più funzioni molecolari. Infine, con i termini delle componenti cellulari, i geni vengono raggruppati secondo la localizzazione dei loro prodotti nelle varie strutture cellulari. In ciascuna delle categorie di GO, i termini sono ordinati e collegati in un grafico diretto ed aciclico, in cima al quale si trovano i termini più generici, e procedendo verso livelli più bassi, si trovano termini via via più specifici. Grazie a questa struttura è possibile caratterizzare un gene, o un gruppo di geni, secondo diversi livelli di profondità. Inoltre, un gene può essere rappresentato da più di un termine per ciascuna classe in quanto localizzato in diversi compartimenti cellulari, avere una o più funzioni molecolari oppure partecipare a uno o più processi biologici (Ashburner M, 2000; The Gene Ontology Consortium., 2019).

Strettamente collegata alla Gene Ontology vi è l'applicazione web **DAVID** (Database for Annotation, Visualization and Integrated Discovery) (https://david.ncifcrf.gov/), che permette di operare numerose analisi su set di dati di espressione genica. La grande utilità di questi strumenti è di consentire l'analisi sia di liste, sia di singoli geni. La classificazione funzionale, infatti, permette di suddividere i geni in cluster omogenei in base alle annotazioni funzionali, ovvero i geni vengono raggruppati se hanno funzioni cellulari simili, partecipano agli stessi processi biologici o pathway.

Per l'interpretazione dei dati di espressione genica un'altra importante metodologia analitica è la Gene Set Enrichment Analysis (GSEA) (Subramanian A 2005). La GSEA è un metodo

computazionale che determina se un set di geni definito a priori mostra differenze statisticamente significative e concordanti tra due stati biologici. Dato un insieme definito a priori di geni S, ad esempio, geni che codificano i prodotti in una pathway metabolica, o situati nella stessa banda citogenetica o che condividono la stessa categoria GO, l'obiettivo di GSEA è di determinare se i membri di S sono distribuiti casualmente in L o se si trovano esclusivamente nella parte superiore o inferiore. Le strategie tradizionali per l'analisi dell'espressione genica si sono concentrate sull'identificazione di singoli geni che presentano differenze tra due stati di interesse, che sebbene utili, non riescono a rilevare processi biologici, quali pathway metabolici, trascrizionali e risposte allo stress. La GSEA, invece, considera tutti i geni di un esperimento, non solo quelli al di sopra di un limite arbitrario in termini FoldChange. Inoltre, preserva le correlazioni gene-gene e, quindi, fornisce un modello nullo molto accurato.

Altre due utili risorse sono REACTOME (https://reactome.org/) e KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://genome.jp/kegg/kegg1.html). Il primo è un database di pathways open source, open access, curato manualmente e peer-reviewed con lo scopo di fornire strumenti intuitivi per la visualizzazione, l'interpretazione e l'analisi delle conoscenze sulle pathway utili per supportare la ricerca di base e clinica. KEGG, invece, è una risorsa utile per comprendere le funzionalità di alto livello dei sistemi biologici, come cellule, organismi e ecosistemi partendo da informazioni genomiche e molecolari. Contiene anche informazioni su malattie e farmaci.

Queste analisi possono essere effettuate direttamente online, come ad esempio GSEA (http://software.broadinstitute.org/gsea/index.jsp) del Broad Institute, Enrichr (http://amp.pharm.mssm.edu/Enrichr/) o gProfiler (https://biit.cs.ut.ee/gprofiler/gost), oppure implementando script ad hoc mediante librerie dedicate di Bioconductor come, ad esempio, clusterProfiler (Yu G, 2012), TopGO (Alexa A. & Rahnenfuhrer J 2019).

1.9 Importanza dell'approccio integrato "CGH + RNA-seq" nelle malattie complesse

I recenti sviluppi tecnologici hanno permesso di produrre dati a livello genomico per molteplici tipi di varianti geniche. Ogni singolo dataset fornisce generalmente informazioni su un solo tipo di variante, e le analisi di questi singoli dataset hanno portato alla comprensione di una lunga lista di patologie genetiche. Tuttavia, è sempre più evidente come le analisi isolate delle singole varianti geniche patogene possono fornire solo una visione lineare e indipendente di un panorama in realtà complesso e multidimensionale. Pertanto, l'integrazione dei vari dati prodotti è spesso necessaria per raggiungere una comprensione approfondita sia delle malattie comuni, sia delle malattie rare, così da

misurare le possibili interazioni tra i fattori di rischio identificati (Thingholm LB, 2016). L'integrazione dei dati può essere suddivisa in due principali categorie:

- integrazione dello stesso tipo di dato tra i diversi studi (es. solo dati di espressione o solo dati di Copy Number Variation)
- integrazione di differenti tipi di dati *"-omici"* per la stessa coorte di campioni (es. analisi integrata di dati di espressione con dati di metilazione, o dati di espressione con dati genomici). (Thingholm LB, 2016)

Un esempio concreto è rappresentato dalle Copy Number Variation e il loro impatto funzionale. Infatti, un'importante questione nella ricerca biomedica attuale è di stabilire una relazione tra le varianti genomiche, che comprendono sia le variazioni polimorfiche apparentemente neutre, sia le variazioni patologiche che causano o predispongono alle malattie e il loro impatto funzionale, allo scopo di rendere possibile la correlazione genotipo-fenotipo. Come già discusso, nonostante i numerosi studi condotti per creare una mappatura completa delle CNV umane, solo ad una frazione di queste è possibile attribuire un chiaro significato biologico. Tuttavia, poiché le CNV alterano il numero di copie e potrebbero agire alterando l'espressione genica, alcuni studi hanno esplorato gli effetti delle Copy Number Variant sul trascrittoma dimostrando che le CNV potrebbero spiegare molte delle alterazioni trascrittomiche rilevanti per la patogenesi di varie patologie umane (Schlattl A, 2011).

Il Disturbo dello Spettro Autistico (DSM-5, 2014) rappresenta un buon esempio per spiegare la complessa relazione tra CNV ed espressione genica. Dal punto di vista clinico, si può distinguere un "**autismo sindromico**", che comprende le sindromi monogeniche associate all'autismo, come la Sclerosi Tuberosa, la sindrome da X-Fragile, la Neurofibromatosi ed alcune malattie citogenetiche come la sindrome di Phelan McDermid, e l' "**autismo idiopatico**" in cui sono presenti i classici segni clinici della sindrome autistica, ma non è noto il gene responsabile.

1.9.1 Studio della componente ereditaria nelle malattie complesse: il paradigma dell'autismo

Una delle sindromi dell'età evolutiva più difficile da spiegare è rappresentata dal Disturbo di Spettro Autistico (DSA) (DSM-5, 2014). Si tratta di un eterogeneo disturbo del neurosviluppo caratterizzato da deficit sociale e di comunicazione, movimenti e pattern comportamentali stereotipati, interessi ristretti ed anomalie sensoriali. I primi sintomi sono potenzialmente osservabili già nei primi 12-18 mesi di vita e nella maggior parte dei casi vengono identificati entro i 3 anni. I segni comportamentali precoci vanno colti con attenzione, sono spesso poco specifici, dal momento che i "tratti autistici" sono distribuiti in maniera continua nella popolazione generale (Willfors C, 2017).

L'autismo colpisce circa l'1% della popolazione ed è 4 volte più comune nei maschi rispetto alle femmine. Tra i disturbi neuropsichiatrici è considerato uno dei disordini con più alto background genetico. Questa affermazione nasce dall'elevatissima concordanza osservata negli studi condotti sui gemelli omozigoti (Muhle R, 2004; Berg JM, 2012; Persico AM & Napolioni V, 2013). L'ereditabilità, ovvero la quota di variabilità di un certo carattere riconducibile alla genetica, varia infatti dal 50 al 95% (Pinto D, 2010; Colvert E, 2015). Inoltre le stime sul rischio di ricorrenza tra i fratelli di bambini autistici oscillano tra il 3 ed il 18%, valori ben superiori alla prevalenza dell'1% circa, osservata nella popolazione generale (Ozonoff S, 2011; Sandin S, 2014). Come la maggior parte delle malattie neuropsichiatriche, presenta un'eziologia complessa, che prevede il coinvolgimento di fattori genetici ed ambientali, la cui interazione determina la predisposizione individuale alla malattia.

Un importante tassello nella comprensione del DSA è rappresentato dalle CNV. L'attenzione verso questa particolare classe di variante genica deriva dall' eccesso pari a circa 3-5 volte di CNV, sia *de-novo*, che ereditate, riscontrata in soggetti autistici rispetto ai controlli ed ai fratelli sani (Pinto D, 2010). Analisi di grosse coorti hanno dimostrato il coinvolgimento di almeno otto loci cromosomici nell'aumento della suscettibilità all'autismo (1q21.1, 2p16, 3q29, 7q11.23, 16p11.2, 15q11.2q13, 22q13.33) (Sanders SJ, 2015), ma il loro impatto funzionale rimane ancora largamente inesplorato. Al di fuori di questi loci ricorrenti, senza opportune validazioni e repliche, è però difficile stabilire se una CNV *de-novo* possa effettivamente contribuire all'eziogenesi dell'autismo (Amir RE, 1999). Stime statistiche suggeriscono però che i loci potenzialmente capaci di contribuire alla suscettibilità potrebbero essere centinaia (Sanders SJ., 2011). Ad oggi sono noti infatti più di 800 geni coinvolti nell'autismo, ma le evidenze che supportano l'associazione di ciascun gene con la malattia sono estremamente variabili (Basu SN, 2009). A dispetto della grande eterogeneità di distribuzione delle CNV, i geni coinvolti e i rispettivi prodotti proteici sono evidentemente coinvolti nei diversi processi biologici che sostengono l'attività della cellula nervosa e sembrano convergere verso specifici aspetti funzionali dei processi coinvolti nel neurosviluppo, in particolare proliferazione cellulare, migrazione, elongazione del neurite, sinaptogenesi e funzionamento sinaptico, rimodellamento della cromatina, metabolismo energetico (Mahfouz A 2015; Lintas C, 2017; DeRosa BA 2018). Oltre all'analisi delle CNV, la trascrittomica è stata usata per testare l'esistenza di pathway alterate nei pazienti con DSA. Gli studi trascrittomici rappresentano, infatti, un collegamento essenziale tra la quantificazione dei livelli di espressione delle proteine e l'analisi delle informazioni genetiche. Uno studio trascrittomico su 116 individui affetti da autismo idiopatico, suddivisi in 3 gruppi a seconda della gravità dei sintomi, ha identificato 123 geni differenzialmente espressi. Dalla pathway analysis è emerso che questi geni convergevano verso comuni target funzionali come trasmissione sinaptica, neurogenesi, neurulazione, apprendimento, ubiquitinazione delle proteine e funzioni cerebrali (Hu VW, 2009).

Uno studio trascrittomico condotto sulla corteccia temporale di 6 cervelli prelevati post-mortem da soggetti autistici e dai corrispondenti controlli, ha rivelato un notevole aumento dell'espressione dei geni correlati al sistema immunitario oltre ad altre pathway come quelle dei geni coinvolti nella comunicazione cellulare, differenziazione, regolazione del ciclo cellulare e morte cellulare (Garbett KA, 2008).

Una dettagliata descrizione delle pathways comunemente alterate nell'autismo è presente in una revisione dei diversi studi di trascrittomica su vari tessuti del 2017 (Ansel A, 2017), da cui emerge che le pathways più comunemente alterate coinvolgono ciclo cellulare, malattia gastrointestinale, geni immunitari, neurogenesi, geni coinvolti nelle funzioni sinaptiche e nel trasporto vescicolare

Nonostante questi studi abbiano evidenziato delle pathways in comune nei soggetti affetti da DSA, per una comprensione esaustiva dei meccanismi coinvolti è necessario considerare anche le alterazioni a livello genomico. Infatti, identificare 189 geni differenzialmente espressi in 20 coppie di autistici e di fratelli/sorelle sani non ci informa sull'origine di questa differenza, ossia non ci dice se l'alterazione trascrittomica ha una origine genetica (ad esempio, da CNV) oppure epigenetica (da iper-/ipo-metilazione delle citosine site nei promotori e/o in altre regioni regolatrici della trascrizione prodotta da fattori ambientali) (Kong SW, 2013).

La correlazione dei dati provenienti da analisi genomiche di 389 soggetti autistici con i loro rispettivi dati trascrittomici ha permesso, per esempio, di trovare un arricchimento in pathway neuronali come la sinaptogenesi, la segnalazione neuropeptidica e l'adesione cellulare. L'intersezione dei dati di espressione genica con dati provenienti da CNV o SNV (variazione del singolo nucleotide) ha rappresentato un approccio efficiente per identificare nuove mutazioni e priorizzare i geni che predispongono all'autismo con le variazioni cromosomiche strutturali (Luo R, 2012). L'importanza di questo approccio è stata confermata successivamente; infatti, l'integrazione dei dati trascrittomici di 36 pazienti autistici con i relativi dati provenienti da studi genomici ha permesso di chiarire il difetto molecolare nel 19% dei pazienti coinvolti nello studio (Codina-Solà M, 2015).

1.10 Approcci statistici e bioinformatici per l'integrazione e la correlazione dei dati (CGH e RNA-seq)

Differenti approcci statistici sono stati usati per l'integrazione dei dati di espressione genica con quelli genomici. Nella situazione più semplice, questi metodi si basano sull'identificazione di CNV ricorrenti su più pazienti e valutano il loro impatto sull'espressione genica. In studi caso-controllo è spesso sufficiente eseguire una comparazione dei due dataset sperimentali attraverso un diagramma di Venn (Ali Hassan NZ, 2014), usando la correlazione di Pearson o di Spearman o

un'analisi di Covarianza (ANCOVA) per identificare aberrazioni che impattano sull'espressione e sono associate con il fenotipo in esame (Ding L, 2008).

Nel caso in cui non sia presente un gruppo di controllo, possono essere usati metodi come l'analisi delle componenti indipendenti (ICA) o la Decomposizione Generale in Valori Singoli (GSVD). Il primo è un metodo di elaborazione computazionale che permette di separare un segnale multivariante nelle diverse sotto-componenti additive, assumendo l'esistenza di una reciproca indipendenza statistica nella sorgente dei segnali non Gaussiani, mentre il secondo metodo è una particolare fattorizzazione (decomposizione) di una matrice basata sull'uso di autovalori e autovettori. Questo modello riduce la dimensione dei dati genomici e considera complessivamente i dati biologici osservati su scala genomica come il risultato di una semplice rete lineare.

Un altro approccio usato in vari studi (Pinto D, 2014; Blumenthal 2016,) è di descrivere i dati secondo un modello lineare mediante la formula generale $Y= X\beta+U$, in cui la variabile dipendente è rappresentata dai dati di espressione, mentre la variabile indipendente è rappresentata dai dati di Copy Number. Il modello lineare è molto utile poiché permette di aggiungere una o più covariate come sesso, età e diversi parametri clinici.

Altre due metodiche usate sono la Partial Least Square (PLS) che è una tecnica di regressione lineare usata per identificare le relazioni (covarianza) tra due matrici (Geladi, P., 1986; Bjørn-Helge 2019) e la Canonical Correlation Analysis (CCA) che analizza la relazione tra due set di dati misurati sullo stesso individuo. Diversamente dall'approccio PLS, per la CCA non è necessario definire quale dei due gruppi contiene le variabili di risposta. Questo rende il metodo più adatto per un'integrazione totale senza particolari assunti.

Questi metodi descritti sono fruibili come funzioni statistiche di base o devono essere caricate da "librerie" specifiche come "*geepack*" o "*gee*" per i metodi basati sul modello lineare o "*CCA*" per la Canonical Correlation Analysis nel software statistico, R. Infine bisogna citare anche la libreria di R *"SegCorr"* che utilizza una procedura statistica per l'identificazione di geni co-espressi adiacenti.

2. SCOPO DELLA TESI

Gli array-CGH rappresentano lo standard per la caratterizzazione dei soggetti affetti da disturbo di spettro autistico. Tuttavia non è ancora ben chiaro quale sia l'impatto delle singole Copy Number Variation a livello trascrittomico. Mancano, o non sono ancora ben standardizzati, strumenti bioinformatici per l'analisi combinata dei dati genomici relativi alle CNV e trascrittomici relativi alla espressione genica nelle patologie neuropsichiatriche.

Lo scopo del mio lavoro di tesi è quello di:

- definire una pipeline che permetta di effettuare uno studio trascrittomico completo;
- produrre uno script che sia in grado di correlare i dati provenienti dai studi genomici di CNV con dati di espressione provenienti da Rna-seq;
- testare gli script prodotti su un gruppo ben definito di soggetti affetti da autismo idiopatico e rispettivi fratelli/sorelle appaiati per sesso e per età, per ampliare le conoscenze sulle impatto che le CNV hanno sull'espressione genica.

3. MATERIALI E METODI

3.1 Campione selezionato

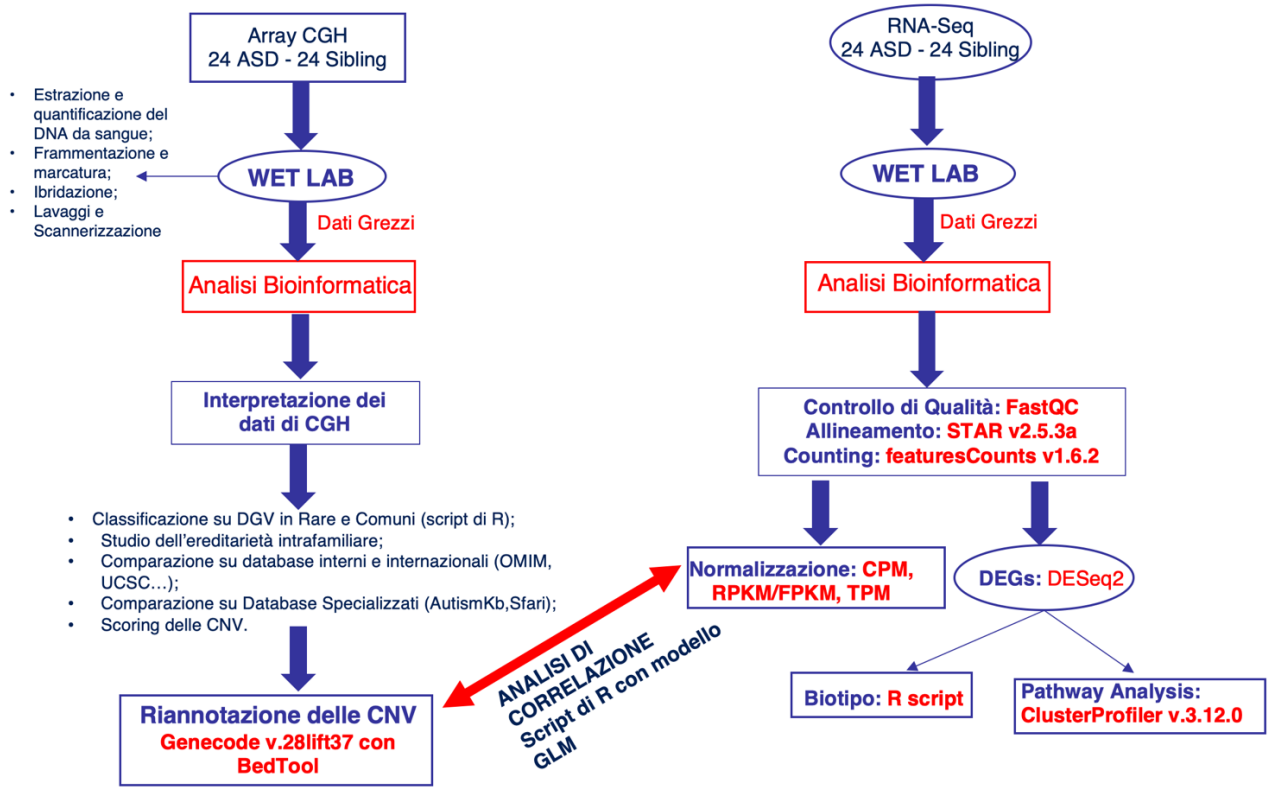
Il campione sperimentale consiste di 24 coppie costituite da un paziente affetto da autismo idiopatico e il rispettivo fratello/sorella non affetto ("sibling"). Il campione selezionato è etnicamente italiano. Le coppie di fratelli sono state selezionate escludendo eventuali comorbidità o sindromi genetiche note e sono appaiate per sesso e per età. L'appaiamento per età considera i prepuberi fino agli 8 anni mentre i post-puberi dagli 8 anni in poi.

Il rapporto M:F è di 4:1 (20 M e 4 F), il numero di coppie pre-puberi è di 7 mentre per le post-puberi è 17. L'uso dei farmaci non è un criterio di esclusione ed è stata tenuta in considerazione l'assunzione fino al giorno del prelievo di sangue. Dei 24 pazienti con DSA, 9 sono in trattamento farmacologico, 11 pazienti non assumono farmaci, mentre per 4 pazienti non è stato possibile recuperare questa informazione in modo affidabile. Nessun fratello/sorella assume farmaci.

Comportamento autistico, funzioni adattative e Q.I. sono stati valutati solo nei soggetti con DSA, utilizzando i questionari ADOS [Autism Diagnostic Observation Schedule, (Gotham K, 2002)], ADI-R [Autism Diagnostic Interview- Revised, (Lord C, 2004)], VABS [Vineland Adaptive Behavior Scales, (Sparrow SS, 1984)], e i test WISC [Wechsler Intelligence Scales for Children], GMDS [Griffith Mental Developmental Scales], Colored Raven Matrices, o Leiter International Performance Scale in funzione dell'età e del livello cognitivo del paziente. Storia familiare e clinica sono state raccolte, come già pubblicato in precedenza (Sacco et al., 2012). Tutte le famiglie hanno fornito il consenso informato scritto per loro stessi e per i propri bambini/ragazzi. Il consenso informato è stato approvato dal Comitato Etico dell'Università "Campus Bio-Medico" di Roma e dal comitato etico dell'Università degli Studi di Messina.

3.2 Workflow sperimentale

Figura 5. Workflow sperimentale



3.3 Array CGH

A ciascun paziente sono stati prelevati circa 3 ml di sangue periferico, raccolti in provette sterili contenenti EDTA da cui è stato estratto il DNA mediante il kit commerciale della Qiagen Gentra Puregene Blood Kit. (Appendice 1, Estrazione del Dna).
L'analisi array Comparative Genomic Hybridization (CGH) è stata eseguita su tutti i soggetti autistici, i rispettivi fratelli/sorella sani e ove possibile, nei genitori utilizzando il vetrino 4 x 180K (SurePrint G3 Human Genome CGH Microarray Kit, Agilent Technologies) consistente di circa 180.000 sonde oligonucleotidiche (60-mer) con una risoluzione media di circa 50 Kb (appendice 1, Array CGH).

3.4 Classificazione e interpretazione delle Copy Number Variant

Le CNV identificate mediante array CGH sono state innanzitutto classificate in "Rare" e "Comuni" sulla base dei dati di MAF [Minor Allele Frequency] disponibili nell'ultima release di DGV [Database of Genomic Variants, (MacDonald, 2013) (R Core Team, 2014). La soglia per definire una duplicazione o delezione come "Rara" è stata fissata a ≤ 3 duplicazioni o delezioni descritte in DGV, con grado di sovrapposizione > 80% con la CNV sperimentale. Questa classificazione è stata eseguita implementando uno script di R (R core team, 2012), descritto nel dettaglio in Appendice 3.
Lo script è stato creato usando le funzioni del pacchetto base di R, ed avvalendosi del pacchetto R xlsx per la lettura e scrittura degli output.

- Nella prima parte viene caricato il set di dati GRCh37hg19variants2016-05-15.txt, precedentemente scaricato dal sito web del Database of Genomic Variants (DGV) (http://dgv.tcag.ca/dgv/app/downloads?ref=GRCh37/hg19). In input viene fornito un file di testo (.txt) che riporta, per ogni regione cromosomica duplicata o deleta, i valori di start – end e Dup/Del.
- Nella parte centrale lo script confronta ogni singola regione fornita in input con il database totale. Ogni CNV deve mostrare una sovrapposizione per almeno l' 80% con la rispettiva CNV presente nel database DGV.
- Successivamente viene posta una condizione per cui ogni delezione o duplicazione, viene classificata come *"Rara"*, se descritta in non più di 3 soggetti sani o *"Comune"* se descritta in 4 soggetti o più.

- L'output è un file excel in cui per ogni delezione o duplicazione è riportato il numero di volte in cui è presente nel database e la classificazione in Rara e Comune.

L'origine familiare o *de-novo* delle CNV è stata stabilita, dove possibile, estendendo l'analisi (CGH e/o Real Time-PCR) ai membri della famiglia.

Per determinare il potenziale ruolo eziopatogenetico delle varianti identificate, è stata eseguita una comparazione sistematica di tutti i geni coinvolti da CNV identificate nelle coppie di paziente autistico-fratello/sorella sano con i principali database online (Decipher, OMIM, USCS, etc) e sul database interno al nostro laboratorio contenente le principali liste di geni candidati (elencati nella tabella 3) per l'autismo presenti in letteratura.

Tabella 3. Principali liste di geni candidati per l'autismo utilizzate nello studio.

Liste di geni candidati per autismo	Citazione bibliografica
Simons Foundation ASD list (1089 genes)	Basu et al. 2009 (https://id.sfari.org/)
List of de novo (Suppl. Table S7a) and inherited (Suppl. Table S7b) CNV	Pinto et al. 2010
List of ASD genes (Table 1)	Betancur, 2011
ASD gene list (Suppl. Table 5 suppl.) and ID gene list (Suppl. Table 6)	Neale et al. 2012
Rare and de novo CNV identified in 122 trios (Suppl. Table 6)	O'Roak et al. 2012
List of de novo variants (Suppl. Table S2)	Sanders et al. 2012
Autism KB core dataset (171 genes)	Xu et al. 2012 (http://autismkb.cbi.pku.edu.cn/)
ASD gene list (Suppl. Table S6A) and ID gene list (Suppl. Table S6C)	Pinto et al. 2014

Per determinare la rilevanza funzionale dei CNV, abbiamo inoltre usato un sistema di classificazione delle CNV basato su cinque diverse categorie:

- CNV certamente causale (categoria 5),
- CNV probabilmente causale o in grado di influire significativamente sul quadro clinico (categoria 4),
- CNV di interpretazione clinica incerta (categoria 3),
- CNV fattore di rischio o modulatore, ma sicuramente non patogeno (categoria 2),
- Negativo per CNV patogeni (categoria 1).

Ogni categoria si basa su diversi criteri obbligatori e facoltativi che devono essere soddisfatti al fine di classificare il risultato del test genetico in una determinata categoria. I criteri principali includono il ruolo funzionale del gene o dei geni all'interno del CNV, la presenza di una precedente descrizione del CNV nella letteratura, la presenza/assenza del CNV nella popolazione generale, ereditato rispetto al CNV de novo, le porzioni introniche rispetto a quelle esoniche del gene incluso nella CNV ecc. Questa procedura è difficile da automatizzare, in quanto legato ad un rigido schema. Pertanto, la classificazione in una categoria o in un'altra è stata eseguita indipendentemente da tre diversi ricercatori. Ogni volta che c'era un disaccordo tra la categoria da assegnare ad una CNV, i casi erano ulteriormente messi in revisione al fine di raggiungere una categoria finale condivisa.

Tabella 4. Parametri considerati per categorizzare le Copy Number Variants

Categoria	Score	Criteri
5. CNV certamente causale (a-d obbligatori; 1-2 opzionale) (Cut-off score 4.5)	1 1 1.5 1 1 1/0,5	Obbligatori: a. Evidenze definitive che almeno un gene della CNV giochi un ruolo importante nel neurosviluppo. b. (A) Almeno una pubblicazione descrive questa CNV mutata/inattivata in un paziente e/o la stessa CNV è fortemente associata con lo specifico disordine o con un disturbo simile, e/o (B) l'intera CNV o il singolo gene della CNV è incluso in uno specifico database o lista di geni relativa al disturbo. c. Almeno una pubblicazione descrive modelli animali e/o cellulari compatibili con un ruolo eziologico nel disordine per lo stesso gene della CNV. d. ≤ 3 duplicazioni/delezioni simili in DGV. Opzionali: 1. Uno o più pazienti con CNV in gran parte o parzialmente sovrapposte e fenotipo presente nel database Decipher. 2. Può essere o De novo (1) o ereditata (0,5).
4. CNV probabilmente causale (a-c obbligatori; 1-3 opzionale) (Cut-off score 4.5)	1 1 1 1 1 0,5/0,5	Obbligatori: a. Evidenze induttive che almeno un gene della CNV giochi un importante ruolo nel neurosviluppo o nella funzione del cervello. b. La CNV coinvolge preferenzialmente l'intero gene candidato, esone o regione del promotore o (B) se intronica, il segmento contiene almeno uno dei seguenti: (a) un sito di legame istonico, (b) un antisenso funzionale o miRNA, (c) un cluster di elementi funzionali. c. ≤ 3 duplicazioni/delezioni simili in DGV. Opzionali : 1. Almeno un gene della CNV è presente in specifici database, pubblicazioni o liste di geni connessi al disordine. 2. Uno o più pazienti con CNV in gran parte o parzialmente in sovrapposizione e fenotipo presente nel database Decipher. 3. Può essere o *de novo* (0,5) o ereditata (0,5).
3. CNV possibilmente causale ma di interpretazione incerta (a. obbligatoria) (Cut-off score 4.5)	1 1 1 1 1 1/0,5	Obbligatori: a. Almeno un gene della CNV è incluso in specifici database connessi al disturbo o è descritto in Letteratura. Opzionali: b. Delezioni/duplicazioni di segmenti intronici che non soddisfano il criterio 4b o sono vicini a un importante gene (~ 100Kb). c. Evidenze inferenziali o ambigue del coinvolgimento nel neurosviluppo o in neuropatologie. d. ≤ 5 duplicazioni/delezioni simili in DGV. e. Nessun paziente con CNV simile nel database Decipher (tipicamente un CNV molto più grande). f. Può essere o *de novo* (1) o ereditata (0,5).
2. CNV Rare o Comuni fattore di rischio o modulatore, ma sicuramente non patogeno		a. Per varianti rare (per es. ≤ 5 duplicazioni/delezioni simili in DGV), evidenze inferenziali di un ruolo funzionale nel neurosviluppo e in processi biologici rilevanti; per le varianti comuni, evidenze conclusive di un ruolo funzionale per uno specifico gene dentro la CNV. b. Per varianti rare, nessun gene della CNV è incluso in specifici database collegati al disturbo o è descritto in Letteratura come associato al disordine (se si, categoria 3). Per le varianti comuni, il gene della CNV deve essere incluso in specifici database collegati al disturbo o essere descritto in letteratura come associato con un disturbo del neurosviluppo (se no, categoria 1). c. Può essere *de novo* o ereditato.
1. Risultato negativo o CNV comuni con nessun ruolo causale, possibili fattori di rischio.		1. Nessuno o almeno un gene della CNV gioca un ruolo noto nel neurosviluppo e/o funzioni neuronali. 2. La CNV è localizzata in una regione altamente variabile (≥ 6 duplicazioni/delezioni presenti in DGV). 3. Evidenze inferenziali di un ruolo funzionale per le varianti comuni. 4. Può essere de novo o ereditato.

3.5 Estrazione dell'RNA

A ogni paziente sono stati prelevati circa 3 ml di sangue periferico mediante provetta Tempus™ specifica per la stabilizzazione e l'isolamento dell'RNA totale da sangue intero. Subito dopo il prelievo, i campioni sono stati miscelati vigorosamente e conservati a -80 °C gradi. L'RNA è stato estratto l'Rna mediante Tempus™ Spin RNA Isolation Kit seguendo il protocollo descritto in Appendice 2.

3.6 Controllo di qualità dell'Rna estratto

La qualità dell'RNA estratto è stata testata con lo strumento Agilent 2100 Bioanalyzer System (Agilent Technologies). Questa verifica è effettuata con lo scopo di evidenziare le varie frazioni di RNA. Infatti, l'RNA totale intatto dovrebbe mostrare chiaramente le bande 28S e 18S in un rapporto di 2:1. Il Bioanalyzer restituisce un parametro il R.I.N (RNA Integrity Number) che va da 0, indice di completa degradazione a 10, RNA di ottima qualità. Tutti i campioni mostravano dei valori ottimali di R.I.N. compresi tra 7.8 e 9.4 e sono stati pertanto sottoposti a sequenziamento.

3.7 RNA-seq

Il sequenziamento dell'RNA è stato eseguito presso il *"Centro di Genomica Traslazionale e Bioinformatica dell'Ospedale San Raffaele di Milano"* usando il protocollo standard *TruSeq Stranded mRNA* dell'Illumina descritto dettagliatamente nell'Appendice 2. Il protocollo prevede i seguenti step:

1- Purificazione e frammentazione dell'mRNA;

2- Sintesi del primo filamento di cDNA;

3- Sintesi del secondo filamento di cDna;

4- Adenilazione dell'estremità al 3';

5- Ligazione degli adattatori;

6- Arricchimento dei frammenti di cDna;

7- Normalizzazione e pooling della libreria.

3.8 FastQC e MultiQC

Il controllo di qualità dei dati è un passaggio fondamentale nelle pipeline di Next Generation Sequencing. Tutti i dati prodotti dal sequenziamento sono stati analizzati mediante **FastQC** che è progettato per eseguire una serie di verifiche sui file di sequenza. Ulteriori controlli di qualità sono stati effettuati mediante **MultiQC** che è uno strumento di uso generale, perfetto per sintetizzare l'output di numerosi strumenti bioinformatici.

3.9 STAR

Per l'allineamento dei dati abbiamo usato STAR2.5.3a (Dobin A, 2013), implementando uno script in linguaggio Bash, costituito dai blocchi di codici descritti nell'Appendice 3:
Lo script è costituito da:

- Una prima parte in cui si ha il manuale che fornisce un utile riferimento e promemoria delle funzioni e opzioni usate;
- Il costrutto **"case…esac",** una particolare struttura che permette di dirigere il flusso del programma a uno dei diversi blocchi di codice, in base alle condizioni di verifica. Nella parte centrale dello script, mediante questo particolare costrutto, definiamo tutte le opzioni e i percorsi alle varie directory che saranno necessarie per l'allineamento; inoltre verifichiamo, in maniera automatica e ricorsiva, l'esistenza delle directory del Genoma indicizzato, e dei file FASTQ che saranno usati come input per l'allineamento.

Infine passiamo a STAR i seguenti parametri specifici per l'allineamento:

- --runThreadN per definire il numero di thread da usare per la generazione del genoma indicizzato, nel nostro caso 1 thread;
- --genomeDir per specificare il percorso della cartella dove è presente il genoma indicizzato;
- --readFilesIn per indicare il percorso della cartella dove sono presenti i file di input in formato BAM;
- --outSAMstrandField intronMotif per eliminare le reads con introni non canonici o inconsistenti;
- --outFileNamePrefix per indicare la directory dove devono essere scritti i file di output e il prefisso con cui tali file devono essere rinominati;
- --outputSAMtype BAM, Sorted ByCoordinate queste due opzioni permettono di produrre come output un file con il risultato dell'annotazione in formato .BAM ordinato in base alle coordinate genomiche;
- --outSAMunmapped Within raccoglie in un file .SAM tutte le reads che non sono state mappate;

- --outFilterMismatchNmax 10 rappresenta il numero massimo di mismatches permessi per l'allineamento;
- --readFilesCommand zcat è il comando che permette di leggere in input un file compresso e di restituire in output il file non compresso.

3.10 FeatureCounts

FeatureCounts è un programma che ha lo scopo generale di contare le reads mappate per le regioni genomiche come ad esempio geni, esoni, promotori, porzioni genomiche e posizioni cromosomiche. E' disponibile sia come pacchetto di Bioconductor Rsubread che come codice sorgente da riga di comando.

La conta è stata eseguita come codice sorgente usando le seguenti opzioni:

Il formato generale da riga di comando è il seguente:

featureCounts [opzioni] –a inputfile.gtf –o outputfile.txt file.bam

Le opzioni che abbiamo usato sono le seguenti:

- -T numero di thread = 4;
- -g per specificare il tipo di attributo da usare per raggruppare le features quando viene fornito un file di annotazione GTF. Nel nostro caso gene_name;
- -Q indica il punteggio minimo di qualità che ogni reads deve soddisfare per essere conteggiata. Abbiamo usato il valore di default 0;
- -a per indicare il percorso ai file di annotazione in formato *.gtf;*
- -s 2 per indicare che la conta deve essere filamento (strand) specifica e deve essere fatta su entrambi i filamenti (reversely stranded);
- -t specifica il tipo di features da usare, nel nostro caso exon;
- -F specifica il formato del file di annotazione, nel nostro caso *GTF*;
- -o indica il nome del file di output;
- -$inputfile specifica il percorso dove leggere i file di input in formato BAM.

3.11 DESeq2

L'analisi dell'espressione differenziale è stata eseguita usando il pacchetto R di Bioconductor DESeq2 (Love MI, 2014) versione 1.24.0 includendo come covariate la presenza di un disturbo dello Spettro autistico, il nucleo familiare, il sesso, la pubertà e l' eventuale uso di farmaci.

Questo pacchetto fornisce metodi per testare l'espressione differenziale mediante l'uso di un modello lineare generalizzato (GLM), modellando le counts secondo una distribuzione binomiale negativa. Il Log2 Fold Change è stimato con una procedura bayesiana empirica, la significatività è valutata con un test di Wald, mentre il p-value viene corretto per test multipli usando il metodo descritto da Benjamini-Hochberg.

Il pacchetto DESeq2 necessita di tre elementi ossia: un metadata, un file coi counts e una formula di design.

Questi parametri sono stati passati secondo la seguente formula generale:

dds = DESeqDataSetFromMatrix(CountData = "counts" ,

colData = "metadata",

design = ~ cond1 + cond2...)

L'analisi di espressione differenziale è fatta mediante l'unica funzione base

DESeq(dds)

con l'aggiunta dei seguenti parametri opzionali:

betaPrior = FALSE : che consente di inserire (TRUE) o meno (FALSE) una media normale pari a zero prima dei coefficienti di non- intercetta;

test = "Wald" per usare il test di significatività di Wald;

fitType = "parametric", per il tipo di adattamento delle dispersioni all'intensità media;

minReplicatesForReplace = Inf il numero minimo di replicati richiesti per utilizzare una sostituzione dei valori anomali (ReplaceOutlier) sui campioni.

I risultati sono stati generati usando la funzione

results(dds, contrast = c("cond1", "var1", "var2")

L'output finale è un file in formato .csv con le seguenti 7 colonne:

GeneName, in cui c'è l'elenco dei geni e dei trascritti;

BaseMean che rappresenta la media dei conteggi normalizzati di tutti i campioni, normalizzando per profondità di sequenziamento;

Log2FoldChange che indica la stima della dimensione dell'effetto. Ci dice quanto l'espressione del gene sembra essere cambiata tra le diverse condizioni usate;

lfcSE che rappresenta l'errore standard stimato per la stima del log2FoldChange;

stat calcolato tramite test di Wald test.

3.12 ClusterProfiler

La pathway analysis è stata eseguita mediante specifiche funzioni presenti nel pacchetto di Bioconductor clusterProfiler v.3.12.0. Inoltre è stato necessario usare funzioni provenienti dalle seguenti librerie:

- org.Hs.eg.db v3.8. 2 ;
- optparse v1.6.2;
- DOSE v3.10.2;
- ReactomePA v1.28.0;
- enrichplot v1.4.0;
- WriteXLS v5.0.0

Lo script prodotto è costituito dai seguenti blocchi di codici:

- Nella prima parte, mediante il pacchetto R optparse, si ha il *"pairsing"* di tutti gli argomenti e vengono impostati tutti i parametri che poi verranno richiamati nella parte centrale dello script, mediante il simbolo $;
- In seguito si ha la conversione delle liste dei geni dal formato "SYMBOL" al formato "ENTREZID", necessario per l'analisi e il subset dei geni in "up" e "down" regolati, in modo da ottenere 2 liste distinte di geni;
- nella parte centrale dello script, le due liste vengono analizzate mediante le seguenti funzioni specifiche:

- *groupGO( )* usata per la classificazione genica basata sulla distribuzione GO a diversi livelli specifici, nel nostro caso abbiamo considerato i 3 livelli di profondità ossia 3, 4 e 5;
- *enrichGO( )* testa l'intero corpus della GO e il risultato è una "enrichment analysis" con il controllo per False Discover Rate (FDR); poichè si ottengono risultati spesso ridondanti, per rendere il risultato dell'arricchimento più chiaro e semplificato è stata usata la funzione *simplify( ) ;*
- *enrichKEGG( )* testa l'intero corpus di KEGG e il risultato è una "Kegg enrichment analysis", dopo opportuno controllo per test multipli tramite FDR;
- *enrichPathway( )* questa funzione restituirà i pathway arricchiti con controllo FDR provenienti dal confronto con il database Reactome.

- Nell'ultima parte del codice, tutti i risultati prodotti vengono rinominati e salvati, in maniera ricorsiva, in formato *.xlxs.* Inoltre, mediante il costrutto *if* viene posta una condizione che permette di plottare solo risultati con p-value < 0.05. Successivamente, i risultati plottati vengono salvati in formato *.pdf.*

3.13 Analisi del biotipo

Il biotipo è la classificazione di un gene o di un trascritto. Lo script utilizzato ha permesso, attraverso l'uso di alcuni pacchetti di Bioconductor specifici per l'annotazione, di creare una tabella contenente, oltre alla colonna con i Biotipi, le colonne con i geni in formato Ensembl, Entrez, Gene Symbol ed HGN, in modo da essere fruibile per le diverse annotazioni.
La tabella prodotta, contiene più di 65 mila query, è stata interrogata attraverso la seguente funzione

bioTypes[bioTypes$HGNC%in%list,]

dove la funzione %in% permette di confrontare la lista dei nostri geni (list) con l'intera tabella e di trovare, in ciascuna riga (",") il gene di interesse, restituendo in output un file .xlxs con la lista dei biotipi dei nostri geni input classificati secondo Ensembl (https://www.ensembl.org/info/genome/genebuild/biotypes.html).

3.14 Analisi di correlazione mediante regressione lineare

L'analisi di correlazione tra i valori di logRatio dei geni e i rispettivi valori di espressione è stata eseguita applicando un modello di regressione lineare, dove la variabile dipendente è rappresentata dai valori di espressione, mentre la variabile indipendente dai valori di Copy Number Variant espressi come valori normalizzati (-1 Delezioni, 1 Duplicazioni e 0 quando il numero di alleli è normale).

Sono stati usati i seguenti pacchetti Bioconductor:

- multtest v 2.40.0
- edgeR v 3.26.7
- ggplot2 v 3.2.1

Lo script prodotto è costituito dai seguenti blocchi di codici (in Appendice 3 il codice in dettaglio):

- Nella prima parte dello script vengono caricate le librerie necessarie per l'analisi e le matrici con i valori di logRatio e counts rispettivamente per le CNV e l'espressione genica. Le matrici vengono subsettate sulla base dei campioni di interesse (solo ASD o solo SIB, per esempio);
- In seguito, i counts sono normalizzati, prima in CPM (Counts Per Milion), quindi viene calcolato il logaritmo in base 2 (log(CPM). Dopo aver verificato che i geni delle 2 matrici hanno

lo stesso ordine, viene applicato un modello di regressione lineare mediante la funzione glm( ) del pacchetto R base. E' inoltre possibile, caricando il metadata aggiungere le diverse covariate.

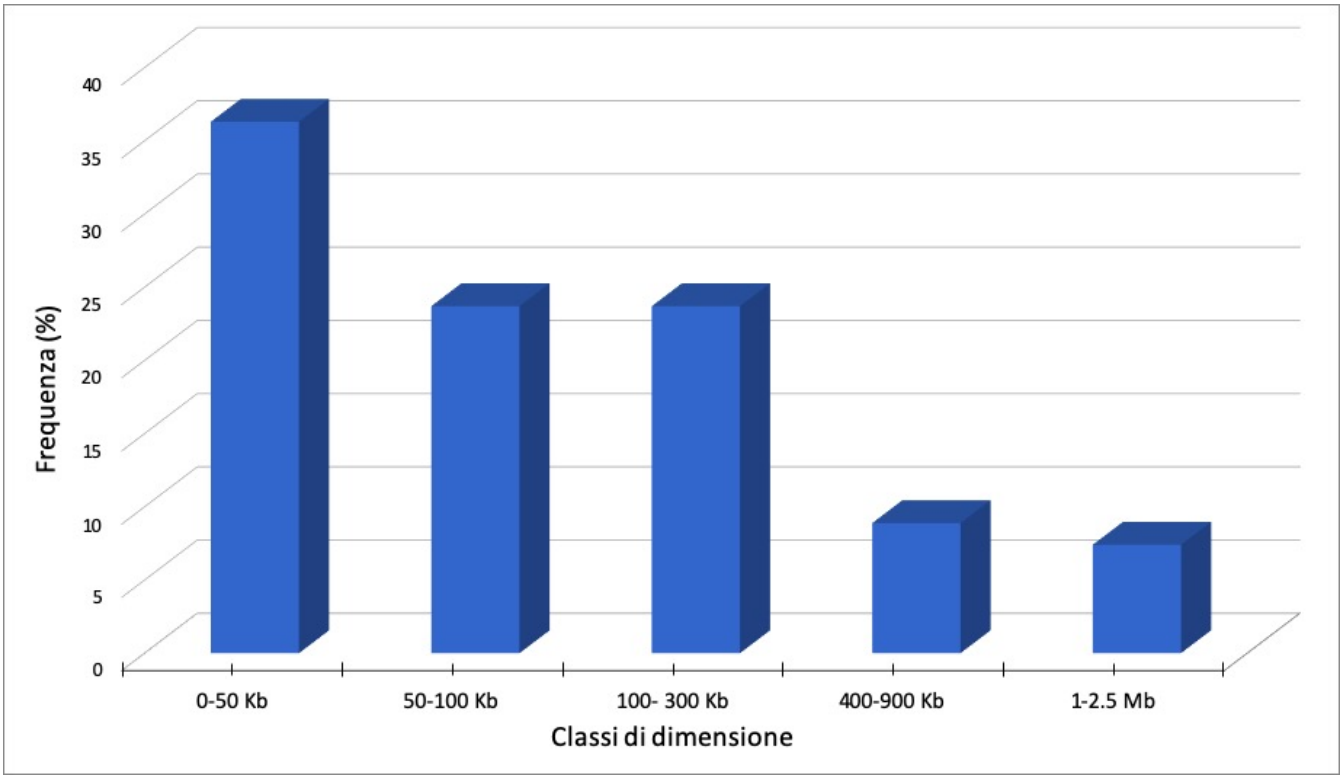
- I risultati vengono salvati in un dataframe, per poi essere ordinati per p-value, corretti per test multipli (BH) e salvati in un file in formato .csv;
- In output, si ottiene un foglio di calcolo excel con l'elenco dei geni, il corrispondente Beta (indice di correlazione), un p-value e un pvalue corretto (BH) da cui sono selezionati i risultati con BH <0.10 e creati dei grafici ordinati per p-value in un unico pdf multipage.

4. RISULTATI

4.1 Caratterizzazione genetica

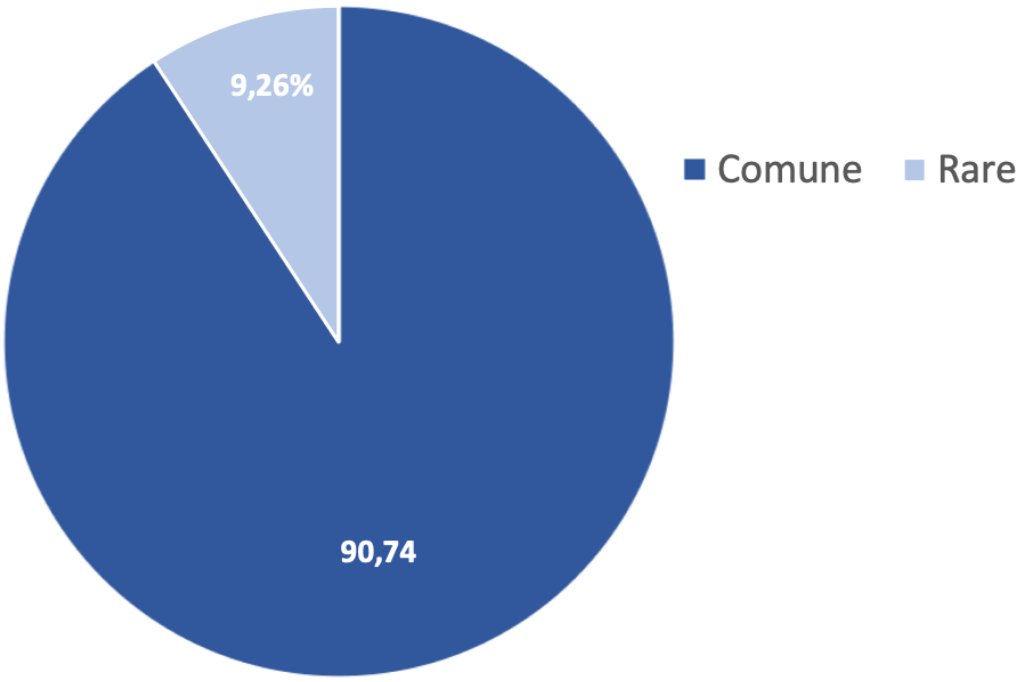
L'array-CGH è stato eseguito su tutte le 24 coppie DSA-sibling. Focalizzando la nostra attenzione solo sui 24 pazienti affetti da autismo, questa analisi ha permesso di identificare un totale di 270 regioni cromosomiche con CNV, di cui 96 duplicazioni e 174 delezioni. Di queste 270 CNV, 196 (73%) sono regioni con all'interno uno o più geni, mentre le restanti 74 (27%) sono localizzate in regioni intergeniche. L'estensione media delle CNV si attesta intorno alle 228 kb, con una distribuzione in base alla dimensione rappresentata nella Figura 6.

Figura 6. Estensione delle CNV



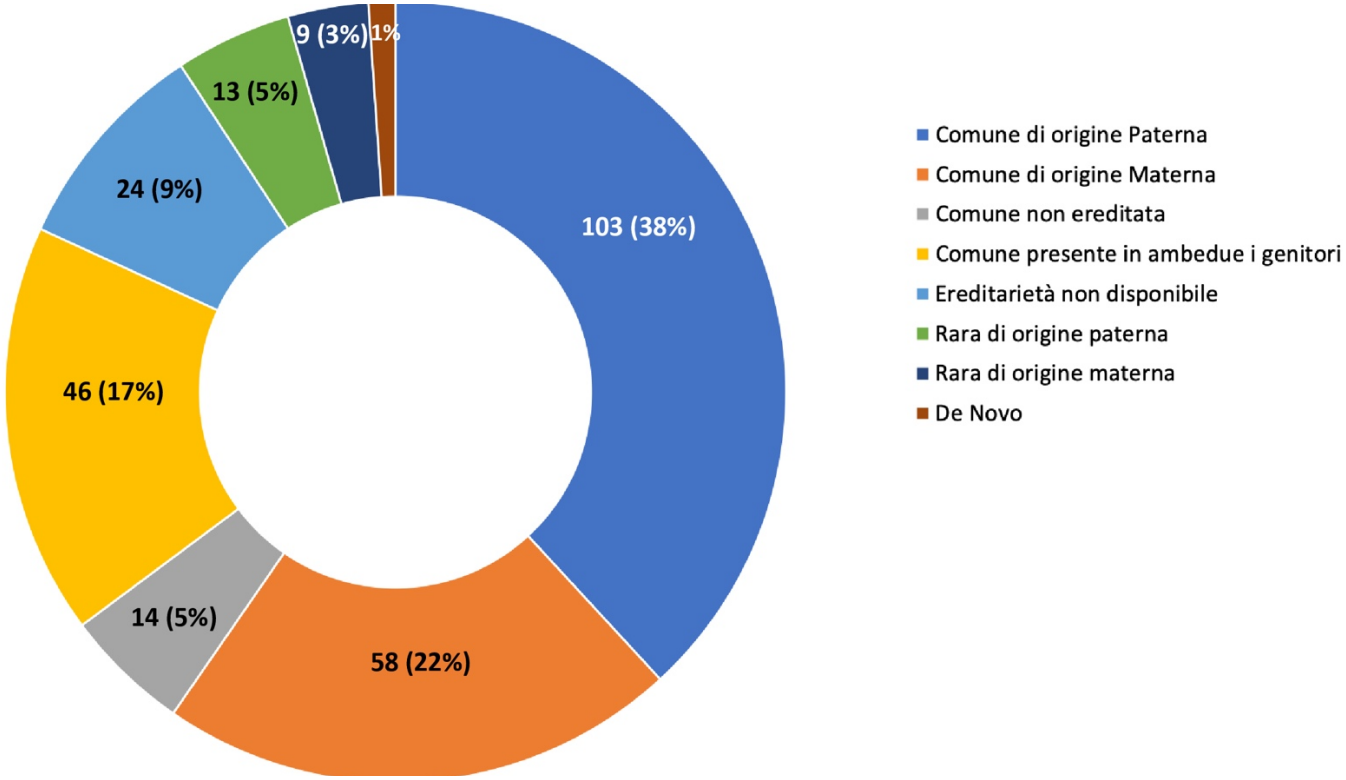
Tutte le CNV individuate sono state classificate come "Comuni" o "Rare" sulla base dei dati presenti in DGV utilizzando uno script di R; con questo strumento è stato possibile effettuare una classificazione automatica delle CNV molto rapida, accurata, e obiettiva. Inoltre, ripetendo l'analisi direttamente sul database online, (http://dgv.tcag.ca/dgv/app/home?ref=), non si è rilevata nessuna discrepanza nei risultati. Da questa analisi il 90,74 % delle varianti sono state classificate come "Comuni", mentre il restante 9,26% come "Rare" (Fig. 7).

Figura 7. Grafico a torta con le percentuali di CNV Rare o Comuni



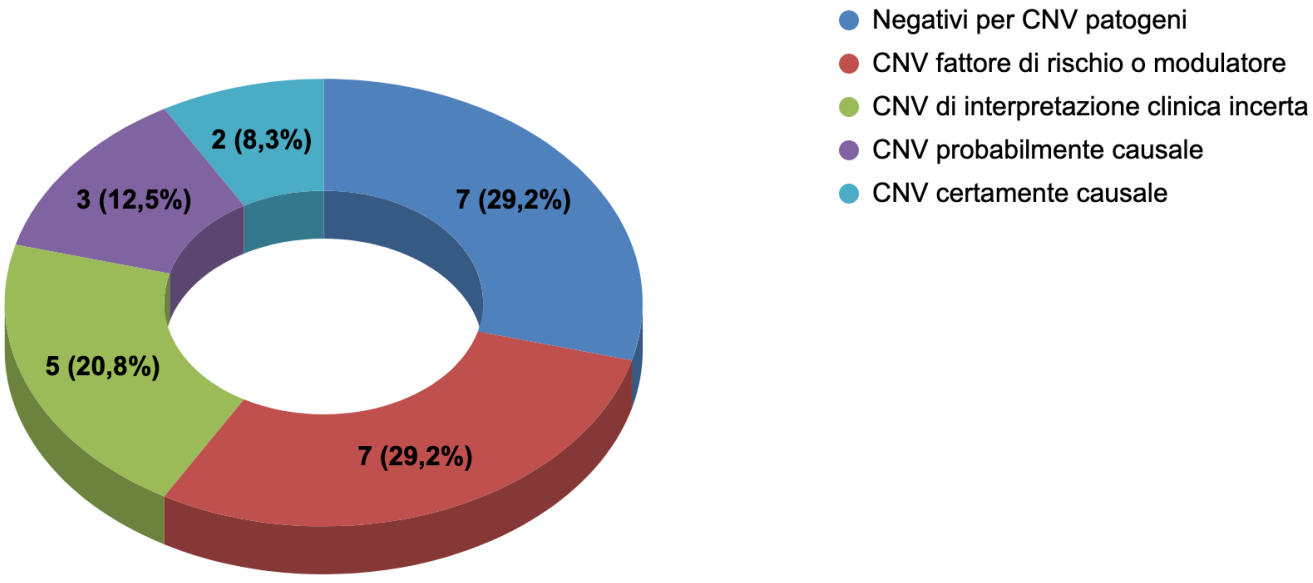
Per una piccola proporzione 24/270 (8,8%) di CNV non è stato possibile stabilire la trasmissione intra familiare per difficoltà di raccolta del campione o per condizioni familiari (es. paziente adottato, rifiuto prelievo). Tra le CNV identificate solo 25/270 (circa 9,26%) sono risultate essere "*rare*", di cui 3 (1%) De Novo, ossia *"descritte in non più di 3 soggetti sani e non ereditata dai genitori"*, mentre 14/270 (5,19%) sono state classificate come *"Comuni non ereditate, ossia, presenti con una frequenza elevata nella popolazione, ma non ereditate da nessuno dei due genitori"*. Le restanti 207/270 (76,67%) mostravano, una origine familiare con una quota maggiore di CNV trasmesse, come atteso, per via patrilineare rispetto alla via matrilineare. La figura 8 mostra in dettaglio la trasmissione delle CNV considerate nello studio.

Figura 8. Classificazione delle CNV sulla base della frequenza e del pattern di trasmissione



Infine, le varianti del numero di copie alleliche (CNV) identificate sono state valutate, per ogni paziente, in merito alla loro rilevanza funzionale per il DSA, in base ai criteri precedentemente elucidati.

Figura 9. Classificazione dei CNV sulla base della patogenicità (vedi criteri in Tab. 4)



Nel complesso, nel 20,8 % dei soggetti studiati è stato possibile riscontrare ristrutturazioni genomiche certamente o probabilmente correlate con la patologia autistica, mentre in un altro 20,8% la variante identificata è stata classificata come di interpretazione clinica incerta. Infine, nel restante 58,4% le varianti trovate sono state classificate come non patogene (29,2%) o come fattore di rischio o modulatore (29,2%).

4.2 Controllo di qualità, allineamento e counting delle reads

Dopo aver effettuato il sequenziamento dell'RNA mediante Novaseq, il primo step consiste nel controllo di qualità dei dati grezzi in formato *.fastq*. Questa analisi è stata eseguita usando il software FastQC (Andrews, 2010) su tutte le 24 coppie sequenziate. I risultati ottenuti possedevano, complessivamente, una buona qualità. Anche la successiva analisi con MultiQC ha confermato l'ottima qualità delle reads generate. Così, sulla base di questi risultati, si è proseguito alla successiva fase di allineamento con il genoma di riferimento.

L'allineamento delle reads è stato effettuato mediante STAR v2.5.3a (Dobin A, 2013), implementando uno script in Bash. Nello script prodotto è stato inserito una piccola descrizione di tutte le funzioni usate e degli step di controllo. Questo ha permesso un facile "debugging" durante lo sviluppo dello script; inoltre favorirà future implementazioni rendendolo facilmente accessibile e favorendo la comprensione delle varie funzioni utilizzate per le analisi. Nel complesso questo script ha permesso di allineare velocemente le reads al genoma di riferimento (GRCh37/hg19).

Successivamente si è proseguito alla fase di conteggio delle reads. Questa fase consiste nel contare, rispetto ad un file di annotazione *"homosapiens.grch37.75.gtf",* il numero di reads allineate ad un determinato gene, trascritto o esone. A tale scopo è stato usato il tool featureCounts (Liao Y, 2013) che implementa tecniche di hashing cromosomico e blocco delle funzionalità altamente efficienti. Funziona con letture single o paired-end e offre una vasta gamma di opzioni appropriate per diverse applicazioni di sequenziamento. Nel nostro caso, attraverso uno script adeguato alle nostre esigenze, le reads sono state contate per gene (*geneid*), tralasciando le diverse isoforme, e considerando la lettura paired-end. Lo script è risultato molto veloce ed ha richiesto l'impiego di poca memoria del computer (nThred =1). I counts prodotti in questo step rappresentano i dati finali di un esperimento di RNA-seq e sono memorizzati in una matrice, in cui le righe rappresentano i geni e le colonne i campioni sequenziati.

4.3 Analisi esplorativa dei dati di RNA-seq

Prima di poter proseguire con le analisi di espressione differenziale è utile compiere un'analisi esplorativa per capire come il nostro dataset è strutturato, quali sono e di che tipo sono le variabili, che tipo di relazione può esistere tra due variabili, e cominciare a creare alcuni grafici esplorativi per una prima analisi. Questo passaggio è stato eseguito applicando due dei metodi descritti in letteratura, tra quelli considerati più stabili, ossia lo scaling multidimensionale (MDS, MultiDimensional Scaling) (Heng T.S., 2009) e la Principal Component Analysis (PCA) (Jolliffe I, 2002).

Lo scaling multidimensionale è una generalizzazione del concetto di ordinamento. Partendo da una matrice quadrata, contenente la "somiglianza" di ogni elemento di riga (gene) con ogni elemento di colonna (campione sequenziato), l'algoritmo assegna a ogni "somiglianza" una posizione in uno spazio N-dimensionale, con N stabilito a priori. Se N è sufficientemente piccolo, questo spazio può essere rappresentato graficamente con uno scatterplot, un grafico nel quale le distanze, rappresentate sul plot, approssimano le differenze registrate tra i campioni, espresse in log2 Fold Change.

L'idea centrale dell'analisi dei componenti principali (PCA), invece, è quella di ridurre la dimensionalità di un set di dati costituito da un gran numero di variabili correlate, pur mantenendo il più possibile la variazione presente nel set di dati. Ciò si ottiene trasformando in una nuova serie di variabili, i componenti principali (PC) variabili che non sono correlate e sono ordinate in base al grado di variazione mantenuta nella trasformazione.

Come mostrato nella Figura 10.A, l'analisi MSD, usando sia l'intero dataset dei geni *(i counts)*, sia un gruppo di 5000 geni con valori di espressione pari a 1 in valore assoluto, non ha mostrato una netta separazione tra i pazienti affetti da autismo (in nero), rispetto ai fratelli/sorelle sani (in rosso). Questo risultato è stato confermato anche, figura 10.B, dalla analisi PCA.

Figura 10.A MSD plot dei dati di RNA-seq. Sono mostrate le caratteristiche del campione considerando l'espressione genica (CPM) e lo status (ASD o sibling). I pazienti affetti da autismo sono colorati in nero, mentre i fratelli/sorelle sani in rosso.

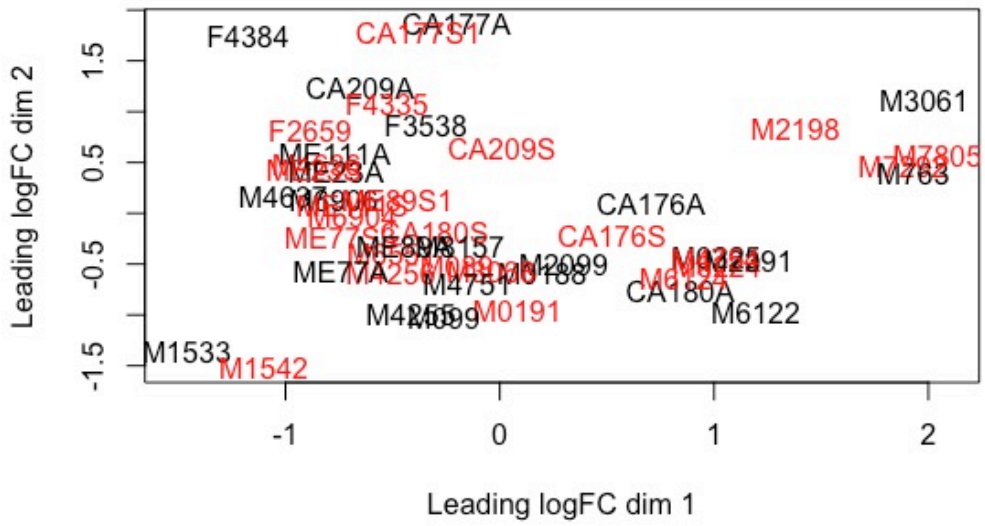
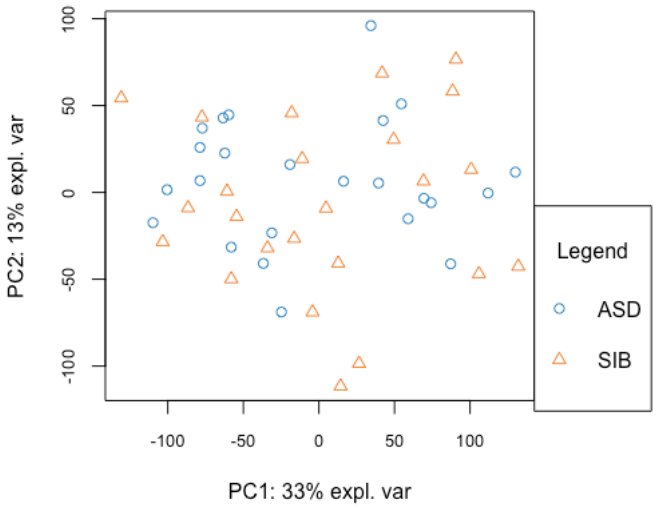


Figura 10.B PCA plot dei dati di RNA-seq. Sono mostrate le caratteristiche del campione considerando l'espressione genica (RPKM) e lo status (ASD o sibling). I cerchietti in blu rappresentano i pazienti affetti da autismo, mentre i triangolini in rosso i rispettivi fratelli/sorella.



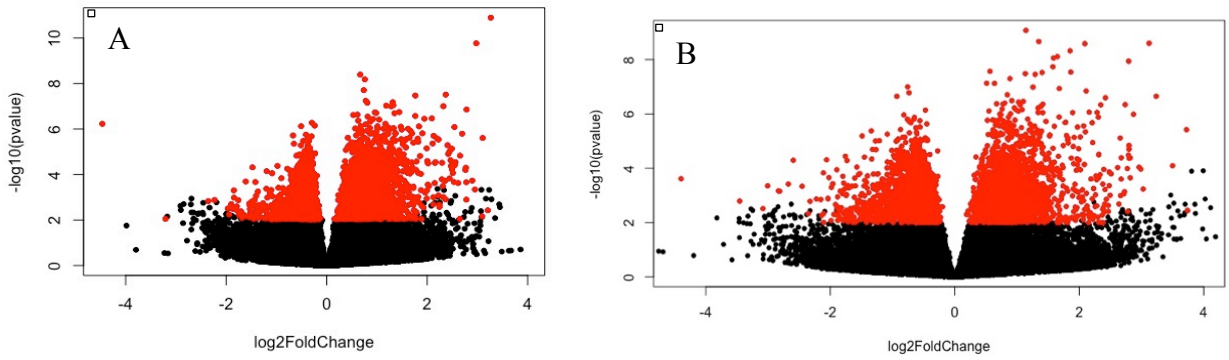
4.4 Analisi di espressione differenziale

L'analisi di espressione differenziale è stata eseguita mediante script di R usando la libreria Bioconductor DESeq2 (Love MI, 2014). Sono stati testati diversi design e covariate ed effettuati vari contrasti tra diversi sottogruppi.

4.4.1 Gli effetti del trattamento farmacologico sul trascrittoma

Il primo test effettuato si poneva l'obiettivo di valutare se il trattamento farmacologico avesse un'influenza sull'espressione genica del campione in esame ed in qual misura. È stato effettuato confrontando i pazienti autistici con e senza trattamento farmacologico: dapprima è stato valutato tutto il campione in esame (9 vs 11) ed in un secondo momento limitando l'analisi ai solo DSA post puberi (7 vs 6). I risultati sono stati rappresentati attraverso Volcano Plot, un grafico che assume sull'asse delle X la misura dell'effetto (Fold Change) e sull'asse Y la significatività statistica (-log10 del raw p-value). Da questa analisi è emersa una forte influenza del trattamento farmacologico sulla trascrittomica. Infatti sono risultati differenzialmente espressi in maniera statisticamente significativa (Padj < 0.05), 3806 geni di cui 2147 up-regolati e 1659 down-regolati (9 vs 11) e 4120 geni di cui 2033 down regolati e 2087 up-regolati, quando corretti per la covariata pubertà (7 vs 6).

Figura 11: *Volcano plo*ts dei geni differenzialmente espressi confrontando 9 pazienti autistici in trattamento e 11 senza trattamento farmacologico (A) e 7 pazienti autistici post-puberi in trattamento e 6 pazienti autistici post-puberi non in trattamento farmacologico (B). *I geni la cui trascrizione viene significativamente modificata dall'azione farmacologica sono indicati in rosso.*



La successiva analisi di pathway, eseguita tramite script R , mediante specifiche funzioni presenti nella library clusterProfiler, ha permesso di confrontare le due liste di geni con i database Kegg, Reactome e della Gene Ontology. L'analisi è stata eseguita suddividendo i geni in Up e Down regolati nei soggetti autistici a seguito di una terapia farmacologica sulla base dei $Log_2FoldChange$ significativi (Padj <0.05). Nelle tabelle seguenti sono mostrati i risultati significativi. In particolare, nelle tabelle 4a (geni down regolati) e 4b (geni up-regolati) sono elencate le pathway significativamente arricchite

confrontando ASD con e senza trattamento farmacologico nel campione totale (9 vs 11) usando la pubertà come covariata:

Tabella 5. Analisi di arricchimento dei geni (A) down-regolati e (B) up-regolati dal trattamento farmacologico.

A	ID Number Pathway	Description	GeneRatio	p-value	p.adjust
	GO:0140098	Catalytic activity, acting on RNA	51/1341	3,12E-09	2,46E-06
	GO:0008135	Translation factor activity, RNA binding	21/1341	2,32E-08	8,6E-06
	GO:0003743	Translation initiation factor activity	16/1341	3,27E-08	8,6E-06
	GO:0140101	Catalytic activity, acting on a tRNA	25/1341	1,34E-07	2,65E-05
	GO:0008094	DNA-dependent ATPase activity	20/1341	4,55E-07	7,18E-05
	GO:0043021	Ribonucleoprotein complex binding	24/1341	2,4E-06	0,00027
	GO:0003697	Single-stranded DNA binding	21/1341	5,21E-06	0,000514
	GO:0001882	Nucleoside binding	49/1341	1,22E-05	0,000873
	GO:0005525	GTP binding	45/1341	6,51E-05	0,003425

B	ID Number Pathway	Description	GeneRatio	p-value	p.adjust
	GO:0051020	GTPase binding	85/1677	1,86E-10	1,67E-07
	GO:0030695	GTPase regulator activity	52/1677	1,92E-08	5,77E-06
	GO:0004674	protein serine/threonine kinase activity	68/1677	3,72E-07	5,54E-05
	GO:0008047	enzyme activator activity	68/1677	2,78E-05	0,002787
	GO:0004407	histone deacetylase activity	9/1677	8,58E-05	0,00773

Nella tabella 5a e b (geni down- e up-regolati) vengono mostrate invece le pathway ottenute confrontando solo DSA post-puberi in trattamento contro DSA post puberi non in trattamento farmacologico (7 vs 6).

Tabella 6a. Analisi di arricchimento dei geni down regolati. In grigio sono evidenziate le categorie GO condivise con quelle ottenute tramite analisi sull'intero campione (tabella 5)

A	ID Number Pathway	Description	GeneRatio	p-value	p.adjust
	GO:0001882	nucleoside binding	62/1518	1,24E-08	5,44E-06
	GO:0019003	GDP binding	21/1518	1,57E-08	5,44E-06
	GO:0005525	GTP binding	59/1518	2,76E-08	5,44E-06
	GO:0001883	purine nucleoside binding	60/1518	2,9E-08	5,44E-06
	GO:0032549	ribonucleoside binding	60/1518	3,26E-08	5,44E-06

GO:0003743	translation initiation factor activity	15/1518	1,18E-06	0,000104
GO:0003924	GTPase activity	43/1518	1,25E-06	0,000104
GO:0043021	ribonucleoprotein complex binding	26/1518	1,87E-06	0,000142
GO:0051082	unfolded protein binding	22/1518	2,45E-06	0,00017
GO:0003697	single-stranded DNA binding	23/1518	2,88E-06	0,000185
GO:0140098	catalytic activity, acting on RNA	47/1518	5,6E-06	0,000334
GO:0016874	ligase activity	28/1518	1,72E-05	0,000896
GO:0140101	catalytic activity, acting on a tRNA	22/1518	5,02E-05	0,002463
GO:0016853	isomerase activity	25/1518	6,68E-05	0,003096
GO:0008143	poly(A) binding	7/1518	0,000153	0,006702
GO:0019787	ubiquitin-like protein transferase activity	49/1518	0,000269	0,010853
GO:0004722	protein serine/threonine phosphatase activity	13/1518	0,000273	0,010853

Tabella 6b. Analisi di arricchimento dei geni up regolati.

B	ID Number Pathway	Description	GeneRatio	p-value	p.adjust
	GO:0051020	GTPase binding	79/1522	3,12E-10	2,7E-07
	GO:0030695	GTPase regulator activity	47/1522	1,24E-07	2,68E-05
	GO:0004674	protein serine/threonine kinase activity	63/1522	5,39E-07	7,78E-05
	GO:0005085	guanyl-nucleotide exchange factor activity	35/1522	6,08E-05	0,005851
	GO:0003779	actin binding	54/1522	0,000128	0,010083

Dal confronto tra le tabelle 4 e 5 (vedi categorie GO evidenziate in grigio in tabella 6) emerge la sostanziale sovrapposizione tra gli effetti trascrizionali del trattamento farmacologico prima e dopo la pubertà.

4.4.2 Gli effetti della pubertà sul trascrittoma

Nonostante sembri meno rilevante rispetto al trattamento farmacologico, la pubertà rappresenta una ulteriore influenza prevedibile sul trascrittoma che abbiamo inteso saggiare conducendo una analisi di espressione differenziale solo su 11 coppie DSA-sibling che non assumevano farmaci (5 vs 6). Dall'analisi sono risultati 61 geni differenzialmente espressi.

4.4.3 Gli effetti dell'autismo sul trascrittoma

Sono stati testati differenti design (dati appaiati e non appaiati), e aggiunte le covariate della pubertà ma da queste analisi non è emerso nulla di statisticamente significativo, probabilmente a causa dell'esiguo numero di individui testati (11 coppie ASD-Sibling prive di trattamento farmacologico).

4.5 Normalizzazione dei dati di Copy Number Variant e di RNAseq

Prima di effettuare l'analisi di correlazione tra dati di trascrittomica e i dati genomici di Copy Number Variant si è reso necessario effettuare alcuni step preliminari. Infatti, tutti i geni che sono risultati essere deleti o duplicati dall'analisi di array-CGH sono stati riannotati mediante script Bash usando il tool BedIntersect (https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html) e in maniera automatica, tutte le regioni delete o duplicate riscontrate sono state confrontate con il file di annotazione *"gencode.v28lift37.annotation.gtf.gz",* precedentemente scaricato da Genecode (https://www.gencodegenes.org). Dopo l'analisi con BedIntersect è risultato un totale di 951 geni per il gruppo degli ASD e 1020 geni, per il gruppo dei sibling.

Bisogna sottolineare che questo step di riannotazione ha permesso di uniformare l'identificativo (Geneid) tra la matrice dei counts e la matrice dei logRatio delle CNV; ha permesso inoltre di arricchire l'analisi di quei geni non presenti nella piattaforma Agilent Cytogenomics, poiché fa riferimento ad un reference diverso e più semplificato, utilizzato di routine in molti laboratori di genetica.

Infine, tutti i valori corrsipondenti ad una delezione, duplicazione e alla normalità sono stati convertiti e uniformati in -1, 1 e 0, rispettivamente. La matrice di conte, invece, è stata normalizzata usando le conte per milioni (CPM), implementata nel pacchetto edgeR. Questo metodo normalizza le conte in base alla profondità di sequenziamento ed è usata per confrontare le conte tra replicati dello stesso gruppo di campioni. I cpm ottenuti sono stati convertiti in log CPM. Il logaritmo ha permesso di uniformare i dati e di renderli confrontabili e paragonabili con i logRatio delle CNV. Abbiamo così ottenuto una matrice costituita da valori di espressione negativi (geni down-regolati), valori di espressione positivi (geni up-regolati) e valori di espressione invariati. Questa procedura di normalizzazione è stata inserita nello stesso script usato per la correlazione, con lo scopo di ottimizzare gli step.

Sono stati testati anche altri metodi di normalizzazione (RPKM/FPKM), tuttavia si è deciso di usare la trasformazione dei counts in log(cpm) in quanto fornisce risultati più affidabili.

4.6 Correlazione tra dati di trascrittomica e di Copy Number Variant

La correlazione tra i valori di logRatio dei CNV e dei valori di logCpm è stato eseguita su tutte le 24 coppie (ASD e SIB) prese in esame nello studio. L'analisi è stata eseguita sui due gruppi divisi con lo scopo di studiare separatamente l'impatto delle CNV sulla trascrittomica e poterli così confrontare successivamente studiando meglio le somiglianze e differenze tra i due gruppi appaiati. E' stato testato un modello di regressione lineare, in cui la variabile dipendente è rappresentata dai dati di espressione, mentre la variabile indipendente è rappresentata dai dati di Copy Number.
Il modello lineare ha permesso di aggiungere le diverse covariate considerate nello studio.
Sono stati effettuati 3 test diversi. Nel primo test i dati non sono stati corretti per covariate, ma è stata effettuata una analisi di regressione mediante script di R sui due dataset separatamente. Solo 3 geni, ADGRG7, LMLN, ZNF630 per gli ASD sono risultati positivamente correlati in maniera significativa (padj <0.05) con la trascrittomica (CNV duplicata, elevata espressione). Per i sibling invece dei 9 geni trovati ADGRG7, TMEM255B, LL22NC03-33B6.4, IGKV6-21, HSPD1P3, UGT2B28, AC135893.2, CH17-264L24.1, ZNF37B, 2 geni (IGKV6-21 e UGT2B28) presentavano una correlazione inversa, CNV deleta, espressione aumentata. Il gene ADGRG7 è regolato in maniera simile tra i due gruppi.
Successivamente i dati sono stati corretti aggiungendo la pubertà come covariata. Sono risultati correlati in maniera significativa (padj <0.05) 247 geni per i soggetti affetti da autismo e 229 per i relativi fratelli/sorelle. Tra questi geni un esempio importante è rappresentato dalla NRXN3 (che codifica un membro di una famiglia di proteine con funzioni nel sistema nervoso sia come recettori che come molecole di adesione). La delezione in questo gene è stata trovata solo in due pazienti ASD e l'analisi di correlazione ha evidenziato una diminuzione dell'espressione genica significativa.
Infine usando come fattori correttivi tutte le covariate dello studio (pubertà, sesso, appartenenza ad un nucleo famigliare, uso di farmaci), è emerso che il numero dei geni influenzati sia dal numero di copie che da altri fattori esterni è di 165 per i soggetti ASD e 229 per i sibling.

5. DISCUSSIONE

Questa tesi riporta i risultati ottenuti da uno studio condotto mediante due tecnologie differenti, CGH-array e RNA-seq, e successiva correlazione dei risultati ottenuti su un campione etnicamente omogeneo costituito da 24 coppie di soggetti italiani affetti da disturbo dello spettro autistico (DSA) e rispettivo fratello/sorella non affetto ("sibling") appaiato per sesso e per età.

Le analisi bioinformatiche sono state condotte in due fasi ben distinte. Infatti, prima è stata condotta

l'analisi e lo studio dei dati genomici di array CGH, successivamente è stato effettuato lo studio trascrittomico mediante Rna-seq, per poi effettuare l'analisi finale di correlazione.

La valutazione genetica costituisce, nel contesto della prassi diagnostica, uno step fondamentale dell'iter necessario per un corretto inquadramento di molti disturbi neuropsichiatrici infantili. La citogenetica tradizionale, effettuata tramite cariotipo su cellule in metafase con il metodo del bandeggio G, pur utilissima nell'identificare un gran numero di anomalie cromosomiche numeriche e strutturali, ha un potere di risoluzione limitato, che può raggiungere al massimo 3-5 Mb. Per tale motivo, allo scopo di aumentare la sensibilità di "detection" delle anomalie cromosomiche ed effettuare un'analisi degli sbilanciamenti genomici presenti in tutto il genoma, è stata messa appunto una tecnica comparativa e quantitativa, gli array CGH, appunto, che permette di identificare la presenza di eventuali anomalie (delezioni e duplicazioni) a livello dell'intero genoma con una risoluzione fino ad appena ~15Kb. L'array CGH, come già ampiamente descritto in letteratura, si è confermato pertanto come test di primo livello *("gold standard")* nella diagnosi genetica del Disturbo di Spettro Autistico e dei disturbi nel neurosviluppo, in generale. Questo dato è confermato anche in questo campione "selezionato". Infatti nel 20,8% dei soggetti studiati è stato possibile riscontrare alterazioni genomiche certamente o probabilmente correlate con la patologia.

L'analisi trascrittomica, condotta partendo da RNA estratto da sangue prelevato mediante provette Tempus, ha richiesto una attenta fase di studio preliminare, sia per individuare tutte le variabili biologiche che possono influenzare la trascrittomica (es. sesso, l'appartenenza a nuclei familiari diversi, lo stato di ASD o di Sibling, l'assunzione di farmaci) e quindi da usare per modellare e correggere i dati. Nel caso presente, i nostri dati mostrano l'influenza profonda del trattamento farmacologico e, in seconda battuta, della pubertà sul trascrittoma del paziente autistico. Stiamo sviluppando approcci aggiuntivi per far uso di tutto il campione anziché limitare le analisi ai pazienti non trattati farmacologicamente. Stiamo, ad esempio, cercando di "condizionare" i dati di espressione rispetto alla presenza/assenza di una terapia farmacologica al momento del prelievo, sebbene la sommatoria degli effetti dei singoli farmaci rappresenti una ipersemplificazione a fronte degli effetti reali che sono presumibilmente farmaco-specifici.

La scelta dei tools e delle library , nonchè delle funzioni e delle opzioni da usare ha richiesto una intensa fase di studio preliminare. Come già sottolineato precedentemente, mentre per le analisi di array CGH il protocollo è molto standardizzato tra i diversi laboratori, per le analisi bioinformatiche dei dati di RNA-seq, pur essendoci un work-flow definito, sono disponibili numerose risorse per i vari step analitici. Queste spaziano da piattaforme online a pagamento come per es. BaseSpace (https://basespace.illumina.com) sviluppato da Illumina o geneInvestigator (https://genevestigator.com/gv/) a Galaxy (https://galaxyproject.org). Queste piattaforme online

supportano risorse e work-flow per analisi NGS come RNA-Seq, CHIP-seq, DNA-seq e molti altri. Galaxy è stata inizialmente utilizzata per alcune prove sui nostri dati, ma non si è rivelata utile per i nostri scopi. Per cui alla fine si è deciso *"di scrivere e sviluppare"* degli script utilizzando la potenza e la libertà dei linguaggi di programmazione disponibili, in particolare di R (Venables LWN, 2018) e delle library specifiche presenti in Bioconductor (Gentleman RC, 2004; Huber W, 2015) e di Bash. La scelta del linguaggio di programmazione è stata meno ardua, in quanto dipendente esclusivamente dalle competenze e conoscenze personali. Così si è optato per due tra i linguaggi di programmazione più usati in bioinformatica. Il linguaggio R, è un linguaggio *"orientato ad oggetti"*, inizialmente ostico da imparare, ma che grazie al grande contributo della comunità scientifica e delle risorse messe a disposizione, ha permesso di sviluppare una pipeline complessa, ma sicuramente ottimizzata, chiara, completa e utile ai nostri scopi, ma anche facilmente adattabile per altri set di dati.

La modalità *"command line"* (Bash) ha inoltre permesso di compiere facilmente operazioni ripetute su diversi dataset (allineamento, merging di dataset, counting e per rinominare file in modo ricorsivo) e di effettuare analisi su un cluster di computer localizzato su una macchina remota. Nel mio caso, il linguaggio bash è stato utile per condurre le analisi di allineamento e di counting, sfruttando il cluster (48 nodi di calcolo con 12 processori e oltre 76GB RAM) del Centro di Genomica Traslazionale e Bioinformatica dell'Ospedale San Raffaele di Milano.

I codici degli script prodotti sono tutti commentati in maniera chiara o mediante manuale ad inizio dello script, per gli script in linguaggio Bash, o mediante # per gli script in R, permettendo così di essere "leggibili" e quindi, fruibili ai vari utenti ma anche facilmente implementabili con ulteriori funzioni e opzioni.

L'analisi di espressione differenziale ha dimostrato l'influenza significativa del trattamento farmacologico e della pubertà , mentre non ha permesso di evidenziare differenze significative tra i soggetti affetti da autismo e i rispettivi sibling. *Questo risultato è essenzialmente dovuto alla drastica riduzione di dimensione del campione nel momento in cui si escludono le coppie DSA-sibling in cui al paziente è stato prescritto un farmaco e le coppie vengono divise in relazione alla pubertà.* Questo limite verrà superato aumentando il campione fino a 50 coppie appaiate e mantenendo i casi con trattamento farmacologico in atto tramite opportuno condizionamento dei dati sulla base dei risultati ottenuti nei soggetti trattati. Almeno tre aspetti possono ulteriormente contribuire a questo risultato: (1) L'epigenetica è tessuto-specifica. Lo studio è stato condotto usando sangue periferico che se da un lato offre il grande vantaggio di essere facilmente reperibile, dall'altro potrebbe averci fatto perdere differenze in espressione per i geni espressi esclusivamente a livello del sistema nervoso centrale. (2) Considerando la patogenicità dei CNV, il 20,8% dei soggetti è stato classificato come portatore di CNV certamente o probabilmente patogena, il 20,8% con CNV di dubbio significato, e il 58,4% con

CNV non patogene o modulatrici del fenotipo. Questa relativa uniformità dei sottogruppi è stata voluta, ma insieme al numero di coppie (N=24), può aver influenzato i risultati dell'analisi differenziale, non arricchendo il campione di casi genetici a causa certa e unitaria. Questo limite verrà superato analizzando un campione di pazienti con sindrome di Phelan-McDermid derivante da una causa unitaria (delezione del cromosoma 22q13.33) (Phelan K, 2005). Infine, (3) dal momento che i "tratti autistici" sono distribuiti in maniera continua nella popolazione generale (Willfors, 2007), questa analisi potrebbe aver mostrato, in un campione così selezionato, un valore predittivo positivo basso tra autistici e fratelli/sorelle non affetti. Riteniamo tuttavia che questa spiegazione non sia soddisfacente, in quanto i siblings dovrebbero essere portatori di fattori di resilienza differenzialmente espressi tra fratelli affetti e non affetti che riteniamo possano emergere con un campione di dimensioni appropriate.

Infine, il modello di correlazione descritto in questo lavoro vuole rappresentare un modo di descrivere e interpretare la complessa interazione tra genotipo e fenotipo. I primi risultati confermano che le CNV hanno sicuramente un impatto sull'espressione ma probabilmente numerosi altri fattori sia genetici, come SNP (Polimorfismi a singolo nucleotide) ed SNV (Polimorfismo a singolo nucleotide), sia epigenetici di natura ambientale, la influenzano.

6. CONCLUSIONI E PROSPETTIVE FUTURE

I risultati di questo lavoro di tesi si collocano in un progetto più ampio. Infatti il reclutamento di nuove coppie ASD-sibling è ancora in corso. Le analisi su questo primo gruppo hanno sicuramente fornito importanti indicazioni.

Gli scopi che ci eravamo dati sono stati raggiunti, in quanto abbiamo:

- Definito e descritto una pipeline per effettuare uno studio trascrittomico genome-wide;
- Prodotto uno script in grado di correlare i dati di CNV provenienti dagli studi genomici con i dati di espressione provenienti dall'Rna-seq;
- Testato gli script prodotti sulle 24 coppie DSA-sibling appena descritte. Tutti i risultati presentati in questa tesi, derivanti dai suddetti script, sono stati replicati in maniera indipendente presso il "Centro di Genomica Traslazionale e Bioinformatica" dell'Ospedale San Raffaele di Milano (Dott.ssa Simona Baghai) e/o il "Translational Genomics Research Institute" (TGen) di Phoenix, AZ (Dott. Ignazio Piras), dimostrando la correttezza degli script stessi.

La pipeline sviluppata è molto articolata ma ha permesso di effettuare tutto il lavoro di analisi ed interpretazione dei dati di RNA-seq in maniera rapida, chiara e automatica richiedendo l'intervento

dell'utente solo in poche fasi. Tuttavia il campo della Bioinformatica è in continua e costante evoluzione per cui questa pipeline può essere ulteriormente migliorata con lo scopo di renderla ulteriormente efficiente e veloce.

E' già in fase di studio la modalità di utilizzare il tool BedIntersect per ottimizzare alcune fasi dell'analisi di CNV nella routine diagnostica, e di altri tools di Bioconductor, come BioMart (Durinck, 2019) per array-CGH o WCGNA (Weighted Correlation Network Analysis) (Langfelder, 2008) per dati di RNAseq con lo scopo di ampliare, completare e arricchire ulteriormente questa pipeline bioinformatica. Nonostante la scelta del software featureCount non abbia mostrato particolari problemi, si sta valutando la possibilità di sostituire questo step con l'opzione di STAR --*quantMode GeneCounts* in modo da rendere lo step di counting più veloce e leggero in termini di richiesta computazionale. Inoltre, per la GO si sta valutando anche l'utilizzo di altri tool come gProfiler (https://biit.cs.ut.ee/gprofiler/gost), ClueGO/Cytoscape (http://apps.cytoscape.org/apps/cluego) o TopGo in R (https://bioconductor.org/packages/release/bioc/html/topGO.html).

I risultati delle analisi di correlazione, infine, dovranno essere validati sperimentalmente mediante RT-PCR, sia per eliminare falsi positivi che per fornire continue indicazioni per l'implementazione degli script. Infatti, solo un continuo scambio di informazioni tra biologi molecolari e bioinformatici può rendere gli script realmente efficienti e utili per la routine clinica, aprendo in futuro nuove prospettive di terapia personalizzata.

Bibliografia

- Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, Packer A(2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Molecular Autism 4(36).
- Alexa A, Rahnenfuhrer J (2019). topGO: Enrichment Analysis for Gene Ontology. R package version 2.37.0.
- Ali Hassan NZ, Mokhtar NM, Kok Sin T, Mohamed Rose I, Sagap I, Harun R, Jamal R. (2014) Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. PLoS One 9(4):e92553.
- Amir RE, Van den Veyver IB,Wan M,Tran CQ,Francke U, Zoghbi HY (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nature Genetics 23(2):185–188.
- Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, Pupko T, Ast G(2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Reports 1(5):543-556.
- American Psychiatric Association. DSM-5 Manuale diagnostico e statistico dei disturbi mentali. Raffaello Cortina Editore. 2014.
- Anders S, Huber W (2010). Differential expression analysis for sequence count data. Genome Biology, 11:R106.
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
- Ansel A, Rosenzweig JP, Zisman PD, Melamed M, Gesundheit B (2017). Variation in Gene Expression in Autism Spectrum Disorders: An Extensive Review of Transcriptomic Studies. Frontiers in Neuroscience, 5 (10), 601.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 25(1):25-9.
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002). Recent segmental duplications in the human genome. Science 297(5583):1003-1007.
- Basu SN, Kollu R, Banerjee-Basu S (2009). AutDB: a gene reference resource for autism research. Nucleic Acids Research 37(Database issue):D832-6.

- Baxevanis AD (2015). The Importance of Biological Databases in Biological Discovery. Current Protocols in Bioinformatics 27(1):1.1.1–1.1.6.
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2018). Genbank. Nucleic Acids Research 46(D1):D41-D47
- Berg JM, Geschwind DH, Autism genetics: searching for specificity and convergence. Genome biology 13(7):247.
- Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B 57(1):289-300.
- Betancur C, (2011). Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Research 1380:42-77.
- Blumenthal I, Ragavendran A, Erdin S, Klei L, Sugathan A, Guide JR, Manavalan P, Zhou JQ, Wheeler VC, Levin JZ, Ernst C, Roeder K, Devlin B, Gusella JF, Talkowski ME (2014).Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. American Journal of Human Genetics 94(6):870-83.
- Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114-20.
- Bray Nl, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Brief Communication 34(5):525-7.
- Bjørn-Helge M, Wehrens R (2019). The pls Package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software 18:22.
- Bullard JH, Purdom E, Hansen KD, Dudoit S (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11:94.
- Chang J, Chapman B, Friedberg I, Hamelryck T, de Hoon M, Cock P, Antao T, Talevich E, Wilczyński B (2019). Biopython Tutorial and Cookbook. Last Update – 16 July 2019 (Biopython 1.74)
- Chen W, Erdogan F, Ropers HH, Lenzner S, Ullmann R (2005). CGHPRO -- a comprehensive data analysis tool for array CGH. BMC Bioinformatics 6:85.
- Chong WW, Lo IF, Lam ST, Wang CC, Luk HM, Leung TY, Choy KW (2014). Performance of chromosomal microarray for patients with intellectual disabilities/developmental delay, autism, and multiple congenital anomalies in a Chinese cohort. Molecular Cytogenetics 7:34.
- Codina-Solà M, Rodríguez-Santiago B, Homs A, Santoyo J, Rigau M, Aznar-Laín G, Del Campo M, Gener B, Gabau E, Botella MP, Gutiérrez-Arumí A, Antiñolo G, Pérez-Jurado LA,

Cuscó I (2015). Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. Molecular Autism 6:21.

- Colnaghi R, Carpenter G, Volker M, O'Driscoll M (2010). The consequences of structural genomic alterations in humans: genomic disorders, genomic instability and cancer. Seminars in Cell and Developmental Biology 22(8):875-85.
- Colvert E, Tick B, McEwen F, Stewart C, Curran SR, Woodhouse E, Gillan N, Hallett V, Lietz S, Garnett T, Ronald A, Plomin R, Rijsdijk F, Happé F, Bolton P (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample.JAMA Psychiatry 72(5):415-23.
- Cook CE, Bergman MT, Cochrane G, Apweiler R, Birney E (2018). The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Research 46(D1):D21-D29.
- Costa-Silva J, Domingues D, Lopes FM (2017). RNA-Seq differential expression analysis: An extended review and a software tool.PLoS One.12(12):e0190152.
- DeRosa BA, El Hokayem J, Artimovich E, Garcia-Serje C, Phillips AW, Van Booven D, Nestor JE, Wang L, Cuccaro ML, Vance JM, Pericak-Vance MA, Cukier HN, Nestor MW, Dykxhoorn DM (2018). Convergent Pathways in Idiopathic Autism Revealed by Time Course Transcriptomic Analysis of Patient-Derived Neurons. Scientific Reports 8:8423.
- Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455(7216):1069-75.
- Diniz WJ, Canduri F (2017). Bioinformatics: an overview and its applications. Genetics and Molecular Research 16(1).
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. 9(1):15-21.

- Dündar F, Skrabanek L, Zumbo P (2015). Introduction to differential gene expression analysis using RNA-seq. Applied Bioinformatics 1–67.
- ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306(5696):636-40.
- Evans C, Hardin J, Stoebel DM (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics 19(5):776–792.
- Ewels P, Magnusson M, Lundin S, Käller M (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32(19):3047–3048.
- Feenstra I, Fang J, Koolen DA, Siezen A, Evans C, Winter RM, Lees MM, Riegel M, de Vries BB, Van Ravenswaaij CM, Schinzel A (2006). European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations (ECARUCA); an online database for rare chromosome abnormalities. European Journal of Medical Genetics 49(4):279-91.
- Feuk L, Carson AR, Scherer SW (2006). Structural variation in the human genome. Nature Reviews Genetics 7(2):85-97.
- Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. American Journal of Human Genetics 84(4):524-533.
- Garbett KA, Ebert PJ, Mitchell A, Lintas C, Manzi B, Mirnics K, Persico AM (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. Neurobiology of Disease, 30:303-311.
- Gelfman S, Ast G (2013). When epigenetics meets alternative splicing: the roles of DNA methylation and GC architecture. Epigenomics 5(4):351-353.
- Geladi P, Kowalski BR (1986). Partial least-squares regression: a tutorial. Analytica Chimica Acta 185:1-17.
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5(10):R80.
- Gonzalo R, Sanchez A. Chapter Three - Introduction to Microarrays Technology and Data Analysis. In "Data Analysis for Omic Sciences: Methods and Applications" (Jaumot J,Bedia C, Tauler R, Eds), Elsevier Ed, pp. 37-69, 2018.

- Gotham K, Risi S, Pickles A, Lord C (2006). The Autism Diagnostic Observation Schedule: Revised Algorithms for Improved Diagnostic Validity. Journal of Autism and Developmental Disorders 37:613.
- Grayton HM, Fernandes C, Rujescu D, Collier DA (2012). Copy number variations in neurodevelopmental disorders. Progress in Neurobiology 99(1):81-91.
- Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, Gibson D, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Research 47(D1):D853-D858.
- Hagen JB. (2000). The origins of bioinformatics. Nature Reviews Genetics.1(3):231-236.
- Hardcastle TJ, Kelly KA (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 10;11:422.
- Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium (2004).The Gene Ontology (GO) database and informatics resource.Nucleic Acids Research.32(Database issue):D258-61.
- Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, Ruedi M, Kaessmann H, Reymond A (2009). Segmental copy number variation shapes tissue transcriptomes. Nature Genetics 41(4):424-429.
- Heng TS. Multidimensional Scaling. In "Encyclopedia of Database Systems" (Ling L & Ozsu M.T, Eds), Springer US, Boston (MA), pp 1784-1784, 2009.
- Hu VW, Sarachana T, Kim KS, Nguyen A, Kulkarni S, Steinberg ME, Luu T, Lai Y, Lee NH. (2009). Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. Autism Research 2(2):78-97.
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum

D, Waldron L, Morgan M (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods 12(2):115-21.

- Hunt LT (1984) Margaret Oakley Dayhoff, 1925–1983. Bulletin of Mathematical Biology 46: 467–472.
- Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004). Detection of large-scale variation in the human genome. Nature Genetics 36(9):949-51.
- Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, Martinet D, Shen Y, Valsesia A, Beckmann ND, Thorleifsson G, Belfiore M, Bouquillon S, Campion D, de Leeuw N, de Vries BB, Esko T, Fernandez BA, Fernández-Aranda F, Fernández-Real JM, Gratacòs M, Guilmatre A, Hoyer J, Jarvelin MR, Kooy RF, Kurg A, Le Caignec C, Männik K, Platt OS, Sanlaville D, Van Haelst MM, Villatoro Gomez S, Walha F, Wu BL, Yu Y, Aboura A, Addor MC, Alembik Y, Antonarakis SE, Arveiler B, Barth M, Bednarek N, Béna F, Bergmann S, Beri M, Bernardini L, Blaumeiser B, Bonneau D, Bottani A, Boute O, Brunner HG, Cailley D, Callier P, Chiesa J, Chrast J, Coin L, Coutton C, Cuisset JM, Cuvellier JC, David A, de Freminville B, Delobel B, Delrue MA, Demeer B, Descamps D, Didelot G, Dieterich K, Disciglio V, Doco-Fenzy M, Drunat S, Duban-Bedu B, Dubourg C, El-Sayed Moustafa JS, Elliott P, Faas BH, Faivre L, Faudet A, Fellmann F, Ferrarini A, Fisher R, Flori E, Forer L, Gaillard D, Gerard M, Gieger C, Gimelli S, Gimelli G, Grabe HJ, Guichet A, Guillin O, Hartikainen AL, Heron D, Hippolyte L, Holder M, Homuth G, Isidor B, Jaillard S, Jaros Z, Jiménez-Murcia S, Helas GJ, Jonveaux P, Kaksonen S, Keren B, Kloss-Brandstätter A, Knoers NV, Koolen DA, Kroisel PM, Kronenberg F, Labalme A, Landais E, Lapi E, Layet V, Legallic S, Leheup B, Leube B, Lewis S, Lucas J, MacDermot KD, Magnusson P, Marshall C, Mathieu-Dramard M, McCarthy MI, Meitinger T, Mencarelli MA, Merla G, Moerman A, Mooser V, Morice-Picard F, Mucciolo M, Nauck M, Ndiaye NC, Nordgren A, Pasquier L, Petit F, Pfundt R, Plessis G, Rajcan-Separovic E, Ramelli GP, Rauch A, Ravazzolo R, Reis A, Renieri A, Richart C, Ried JS, Rieubland C, Roberts W, Roetzer KM, Rooryck C, Rossi M, Saemundsen E, Satre V, Schurmann C, Sigurdsson E, Stavropoulos DJ, Stefansson H, Tengström C, Thorsteinsdóttir U, Tinahones FJ, Touraine R, Vallée L, van Binsbergen E, Van der Aa N, Vincent-Delorme C, Visvikis-Siest S, Vollenweider P, Völzke H, Vulto-van Silfhout AT, Waeber G, Wallgren-Pettersson C, Witwicki RM, Zwolinksi S, Andrieux J, Estivill X, Gusella JF, Gustafsson O, Metspalu A, Scherer SW, Stefansson K, Blakemore AI, Beckmann JS, Froguel P (2011). Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478(7367):97-102.
- Jolliffe, IT. Principal Component Analysis. Springer-Verlag, Berlin (Ger), 1986.

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992).Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258(5083):818-821.
- Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, Moreno-De-Luca D, Moreno-De-Luca A, Mulle JG, Warren ST, Richard G, Compton JG, Fuller AE, Gliem TJ, Huang S, Collinson MN, Beal SJ, Ackley T, Pickering DL, Golden DM, Aston E, Whitby H, Shetty S, Rossi MR, Rudd MK, South ST, Brothman AR, Sanger WG, Iyer RK, Crolla JA, Thorland EC, Aradhya S, Ledbetter DH, Martin CL (2011). An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genetics in Medicine 13(9):777-784.
- Karki R, Pandya D, Elston RC, Ferlini C (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. BMC Medical Genomics 8:37.
- Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST; Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genetics Medicine 13(7):680-685.
- Kleinjan DJ, van Heyningen V (1998). Position effect in human genetic disease. Human Molecular Genetics 7(10):1611-8.
- Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, Okubo K, Nakamura Y, Takagi T (2018). DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Research 46(D1):D30-D35.
- Korpelainen E, Tuimala J,Somervuo P, Huss M, Wong G. RNA-seq Data Analysis. A pratical Approach. Chapman & Hall/CRC Mathematical & Computational Biology. 2014.
- Kong SW, Shimizu-Motohashi Y, Campbell MG, Lee IH, Collins CD, Brewster SJ, Holm IA, Rappaport L, Kohane IS, Kunkel LM (2013). Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. Neurogenetics. 14(2):143-52.
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3):R25.
- Law CW, Chen Y, Shi W, Smyth GK (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15(2):R29.
- Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, Haag JD, Gould MN, Stewart RM, Kendziorski C (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29(8),1035-43.

- Li H, Ruan J, Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Resource 18(11):1851-1858.
- Li J, Tibshirani R (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Statistical Methods in Medical Research 22(5):519–536.
- Lingjaerde OC, Baumbusch LO, Liestøl K, Glad IK, Børresen-Dale AL (2008). CGH-Explorer: a program for analysis of array-CGH data. Bioinformatics 21(6):821-822.
- Liao Y, Smyth GK, Shi W (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30(7):923-30.
- Lintas C, Picinelli C, Piras IS, Sacco R, Brogna C, Persico AM (2017). Copy number variation in 19 Italian multiplex families with autism spectrum disorder: Importance of synaptic and neurite elongation genes. American Journal of Medical Genetics 174(5):547-556.
- Liva S, Hupé P, Neuvial P, Brito I, Viara E, La Rosa P, Barillot E (2006). CAPweb: a bioinformatics CGH array Analysis Platform. Nucleic Acids Research 34(Web Server issue):W477-81.
- Lord C, Rutter M, Le Couteur A (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of Autism and Developmental Disorders. (24) 5:659–685.
- Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12):550.
- Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, Klei L, Cai C, Ou J, Lowe JK, Hurles ME, Devlin B, State MW, Geschwind DH (2012). Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. American Journal of Human Genetics 91(1):38-55.
- Luscombe NM, Greenbaum D, Gerstein M (2001). What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine 40(4):346-358.
- MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Research 42(Database issue):D986-92.
- Maxam AM, Gilbert W (1977). A new method for sequencing DNA. Proceedings of the National Academy of Sciences 74(2):560-4.
- Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, Antonarakis SE, Reymond A (2006). Submicroscopic deletion in patients with Williams-Beuren syndrome influences

expression levels of the nonhemizygous flanking genes. American Journal of Human Genetics 79(2):332-41.

- Metcalfe, RM, Boggs DR (1976). Ethernet: Distributed Packet Switching for Local Computer Networks, Communications of the ACM 19(7), 395–404.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5(7):621-628.
- Muhle R, Trentacoste SV, Rapin I (2004). The genetics of autism. Pediatrics113(5):e472-86.
- Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Research, 40(20)10084–10097.
- Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH Jr, Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485(7397):242-245.
- O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485(7397):246-50.
- Ozonoff S, Young GS, Carter A, Messinger D, Yirmiya N, Zwaigenbaum L, Bryson S, Carver LJ, Constantino JN, Dobkins K, Hutman T, Iverson JM, Landa R, Rogers SJ, Sigman M, Stone WL. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. Pediatrics 128(3):e488-95.
- Patro R, Duggal G, Love MI, Irizarry R A, Kingsford C (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14(4):417-419.
- Pauling L, Corey RB, Branson HR (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. PNAS 37(4):205-11.

- Pavan S, Rommel K, Marquina MEM, Höhn S, Lanneau V Rath A (2017).Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS One12(1): e0170365.
- Périer RC, Praz V, Junier T, Bonnard C, Bucher P (2000). The eukaryotic promoter database (EPD). Nucleic Acids Research 28(1):302-3.
- Persico AM, Napolioni V (2013). Autism genetics. Behavioural Brain Research 251:95-112.
- Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics 20(2):207-11.
- Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bölte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BH, Cochrane L, Corsello C, Crawford EL, Crossett A, Cytrynbaum C, Dawson G, de Jonge M, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green A, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le Couteur A, Leventhal BL, Lionel AC, Liu XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Pilorge M, Piven J, Ponting CP, Posey DJ, Poustka A, Poustka F, Prasad A, Ragoussis J, Renshaw K, Rickaby J, Roberts W, Roeder K, Roge B, Rutter ML, Bierut LJ, Rice JP, Salt J, Sansom K, Sato D, Segurado R, Sequeira AF, Senman L, Shah N, Sheffield VC, Soorya L, Sousa I, Stein O, Sykes N, Stoppioni V, Strawbridge C, Tancredi R, Tansey K, Thiruvahindrapduram B, Thompson AP, Thomson S, Tryfon A, Tsiantis J, Van Engeland H, Vincent JB, Volkmar F, Wallace S, Wang K, Wang Z, Wassink TH, Webber C, Weksberg R, Wing K, Wittemeyer K, Wood S, Wu J, Yaspan BL, Zurawiecki D, Zwaigenbaum L, Buxbaum JD, Cantor RM, Cook EH, Coon H, Cuccaro ML, Devlin B, Ennis S, Gallagher L, Geschwind DH, Gill M, Haines JL, Hallmayer J, Miller J, Monaco AP, Nurnberger JI Jr, Paterson AD, Pericak-Vance MA, Schellenberg GD, Szatmari P, Vicente AM, Vieland VJ, Wijsman EM, Scherer SW, Sutcliffe JS, Betancur C (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466(7304):368-72.

- Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z, Vorstman JA, Thompson A, Regan R, Pilorge M, Pellecchia G, Pagnamenta AT, Oliveira B, Marshall CR, Magalhaes TR, Lowe JK, Howe JL, Griswold AJ, Gilbert J, Duketis E, Dombroski BA, De Jonge MV, Cuccaro M, Crawford EL, Correia CT, Conroy J, Conceição IC, Chiocchetti AG, Casey JP, Cai G, Cabrol C, Bolshakova N, Bacchelli E, Anney R, Gallinger S, Cotterchio M, Casey G, Zwaigenbaum L, Wittemeyer K, Wing K, Wallace S, van Engeland H, Tryfon A, Thomson S, Soorya L, Rogé B, Roberts W, Poustka F, Mouga S, Minshew N, McInnes LA, McGrew SG, Lord C, Leboyer M, Le Couteur AS, Kolevzon A, Jiménez González P, Jacob S, Holt R, Guter S, Green J, Green A, Gillberg C, Fernandez BA, Duque F, Delorme R, Dawson G, Chaste P, Café C, Brennan S, Bourgeron T, Bolton PF, Bölte S, Bernier R, Baird G, Bailey AJ, Anagnostou E, Almeida J, Wijsman EM, Vieland VJ, Vicente AM, Schellenberg GD, Pericak-Vance M, Paterson AD, Parr JR, Oliveira G, Nurnberger JI, Monaco AP, Maestrini E, Klauck SM, Hakonarson H, Haines JL, Geschwind DH, Freitag CM, Folstein SE, Ennis S, Coon H, Battaglia A, Szatmari P, Sutcliffe JS, Hallmayer J, Gill M, Cook EH, Buxbaum JD, Devlin B, Gallagher L, Betancur C, Scherer SW (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. American Journal of Human Genetics 94(5):677-94.
- Phelan K, Rogers RC, Boccuto L (2005) Phelan-McDermid Syndrome. [Updated 2018 Jun 7]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews Seattle (WA): University of Washington, Seattle; 1993-2019.
- R Development Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing, Vienna.
- Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963). Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology 7:95-9.
- Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13(484).
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H,

Lee C, Jones KW, Scherer SW, Hurles ME (2006). Global variation in copy number in the human genome. Nature 444(7118):444-54.

- Reymond A, Henrichsen CN, Harewood L, Merla G (2007). Side effects of genome structural changes. Current Opinion in Genetics & Development 17(5):381-6.
- Rigau M, Juan D, Valencia A, Rico D (2019). Intronic CNVs and gene expression variation in human populations. PLoS Genetics. 2019 Jan 24;15(1):e1007902.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2005). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7):e47.
- Robinson DM, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11(3):R25.
- Robinson MD, Smyth GK (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23(21):2881-7.
- Roy M, Kim N, Xing Y, Lee C (2008). The effect of intron length on exon creation ratios during the evolution of mammalian genomes. RNA 14(11):2261-73.
- Sacco R, Lenti C, Saccani M, Curatolo P, Manzi B, Bravaccio C, Persico AM (2012). Cluster Analysis of Autistic Patients Based on Principal Pathogenetic Components. Autism Research, 5(2) 137-47.
- Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH, Moreau MP, Gupta AR, Thomson SA, Mason CE, Bilguvar K, Celestino-Soper PB, Choi M, Crawford EL, Davis L, Wright NR, Dhodapkar RM, DiCola M, DiLullo NM, Fernandez TV, Fielding-Singh V, Fishman DO, Frahm S, Garagaloyan R, Goh GS, Kammela S, Klei L, Lowe JK, Lund SC, McGrew AD, Meyer KA, Moffat WJ, Murdoch JD, O'Roak BJ, Ober GT, Pottenger RS, Raubeson MJ, Song Y, Wang Q, Yaspan BL, Yu TW, Yurkiewicz IR, Beaudet AL, Cantor RM, Curland M, Grice DE, Günel M, Lifton RP, Mane SM, Martin DM, Shaw CA, Sheldon M, Tischfield JA, Walsh CA, Morrow EM, Ledbetter DH, Fombonne E, Lord C, Martin CL, Brooks AI, Sutcliffe JS, Cook EH Jr, Geschwind D, Roeder K, Devlin B, State MW (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron , 70 (5):863-85.
- Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, Keaney JF 3rd, Klei L, Mandell JD, Moreno-De-Luca D, Poultney CS, Robinson EB, Smith L, Solli-Nowlan T, Su MY, Teran NA, Walker MF, Werling DM, Beaudet AL, Cantor RM, Fombonne E, Geschwind DH, Grice DE, Lord C, Lowe JK, Mane SM, Martin DM, Morrow EM, Talkowski ME, Sutcliffe JS, Walsh CA, Yu

TW; Autism Sequencing Consortium, Ledbetter DH, Martin CL, Cook EH, Buxbaum JD, Daly MJ, Devlin B, Roeder K, State MW (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87(6):1215-1233.

- Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. Journal of American Medical Association 311(17):1770-7.
- Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74(12):5463-7.
- Schlattl A, Anders S, Waszak SM, Huber W, Korbel J (2011). Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Research 21(12):2004-13.
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004). Large-scale copy number polymorphism in the human genome. Science 305(5683):525-528.
- Sexton T, Umlauf D, Kurukuti S, Fraser P (2007). The role of transcription factories in large-scale structure and dynamics of interphase chromatin. Seminars in Cell and Developmental Biology. 18(5):691-7.
- Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L, Bobrow M, Carter NP (2004). Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. Journal of Medical Genetics 41(4):241-8.
- Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2013). New and continuing developments at PROSITE. Nucleic Acids Research 41(Database issue):D344-7.
- Slatko BE, Gardner AF, Ausubel FM (2018). Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology 122(1):e59.
- Soneson C, Delorenzi M (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics,14:91.
- Sparrow SS, Balla DA, Cicchetti DV (1984). Vineland adaptive behavior scales: Interview edition, survey form manual. Circle Pines, MN: American Guidance Service.
- Stajich JE (2007). An Introduction to BioPerl. Methods Molecular Biology 406:535-48.

- Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002). The generic genome browser: a building block for a model organism system database. Genome Research. 12(10):1599-1610.
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848-53.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS USA, 102(43):15545-50
- Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, Conesa A (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Research 43(21):e140.
- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526:68–74.
- The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research 47(D1):D330-D338.
- Theisen, A. (2008) Microarray-based Comparative Genomic Hybridization (aCGH). Nature Education 1(1):45.
- Thingholm LB, Andersen L, Makalic E, Southey MC, Thomassen M, Hansen LL (2016). Strategies for Integrated Analysis of Genetic, Epigenetic, and Gene Expression Variation in Cancer: Addressing the Challenges. Frontiers in Genetics 7:2.
- Trapnell C, Pachter L, Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105-11.
- Van de Wiel MA, Leday GGR, Pardo L, Rue H, Van der Vaart AW, Van Wieringen WN (2012). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics 14, 113-128.
- Venables WN, Smith DM and the R Core Team. An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics. R Development Core Team 2006.
- Vermeesch JR, Brady PD, Sanlaville D, Kok K, Hastings RJ (2012). Genome-wide arrays: quality criteria and platforms to be used in routine diagnostics. Human Mutation 33(6):906-15.
- Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clinical Chemistry55(4):641-58.

- Wang L, Wang S, Li W (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics 28(16):2184-5.
- Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 10(1):57-63.
- Watson JD, Crick FH (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 171(4356):737-8.
- Willfors C, Carlsson T, Anderlid B-M, Nordgren A, Kostrzewa E, Berggren S, Ronald A, Kuja-Halkola R, Tammimies K, Bölte S (2017). Medical history of discordant twins and environmental etiologies of autism. Translational Psychiatry. 7(1): e1014.
- Wolters S, Schumacher B (2013). Genome maintenance and transcription integrity in aging and disease. Frontiers in Genetics 4: 19.
- Yang C, Li J, Wu Q, Yang X, Huang AY, Zhang J, Ye AY, Dou Y, Yan L, Zhou WZ, Kong L, Wang M, Ai C, Yang D, Wei L (2018). AutismKB 2.0: a knowledgebase for the genetic evidence of autism spectrum disorder. Database (Oxford) 2018.
- Yang YH, Buckley MJ, Speed TP (2001). Analysis of cDNA microarray images.Brief Bioinformatics 2(4):341-9.
- Ye H, Meehan J, Tong W, Hong H (2015). Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. Pharmaceutics 7(4): 523–541.
- Yu G, Wang L, Han Y, He Q (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." OMICS: A Journal of Integrative Biology, 16(5), 284-287.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg

S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001). The sequence of the human genome. Science 291(5507):1304-51.

- Xu LM, Li JR, Huang Y, Zhao M, Tang X, Wei L (2012). AutismKB: an evidence-based knowledgebase of autism genetics. Nucleic Acids Research 40(Database issue):D1016-22.

- Zhang L, Chang S, Li Z, Zhang K, Du Y, Ott J, Wang J (2012). ADHDgene: a genetic database for attention deficit hyperactivity disorder. Nucleic Acids Research 40(Database issue):D1003-9.

- Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenetics and Genome Research 115(3-4):205-14.

- Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y, Packer A, Darnell RB, Troyanskaya OG (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nature Genetics 51(6):973-980.

Appendice 1

A1.1 Estrazione del DNA

1) In una provetta da 1,5 ml aliquotare 300 µl di sangue intero e 900 µl di RBC Lysis Solution e lasciare a temperatura ambiente per 12 minuti invertendo ogni 3 minuti circa.
2) Centrifugare per 2 minuti a temperatura ambiente a 16000xg. Rimuovere il surnatante prestando attenzione a lasciarne circa 10 µl che serviranno per risospendere il pellet di globuli bianchi.
3) Dopo aver vortexato vigorosamente, aggiungere 300 µl di Cell Lysis Solution e risospendere fino a completa lisi delle cellule.
4) Aggiungere 1,5 µl di RNase A solution, mescolare per inversione, centrifugare per 10 secondi ed infine incubare a 37° C per 15 minuti.
5) Trasferire la provetta in ghiaccio per 1 minuto, aggiungere 100 µl di Protein Precipitation, vortexare vigorosamente per 20 secondi e centrifugare per 2 minuti a temperatura ambiente a 16000xg.
6) Trasferire il surnatante in una nuova provetta da 1,5 ml in cui sono stati precedentemente dispensati 300 µl di isopropanolo freddo. In questa fase, invertendo più volte, compare il pellet di DNA.
7) Dopo aver centrifugato per 2 minuti a 16000xg a temperatura ambiente, rimuove il surnatante ed asciugare accuratamente.
8) Lavare il pellet con 300 µl di etanolo al 70% e centrifugare. Rimuovere il surnatante e lasciare essiccare il pellet per 45-60 minuti a temperatura ambiente.
9) Aggiungere 50 µl di TE e favorire l'idratazione del pellet mantenendolo a 65°C per 5 minuti. Vortexare delicatamente e centrifugare per 1 minuto a 6000xg. Mantenere a temperatura ambiente per una notte.
10) Procedere con la quantificazione del campione in spettrofotometria.

A1.2 Array CGH

La metodica si compone di 4 step principali:

1. **MARCATURA**
2. **PURIFICAZIONE**
3. **IBRIDAZIONE**

4. LAVAGGI

STEP 1: MARCATURA

Il primo passaggio ha lo scopo di marcare il DNA reference ed il DNA test con fluorocromi specifici (Cy5 e Cy3):

1) Preparare N provette da 0,2 ml (n per test ed n per reference, dove N=n+n) contenenti 1000 ng di DNA genomico in 26 µl di eluente.

2) Aggiungere 5 µl di Random Primers.

3) Trasferire in termociclatore per 10 minuti a 98°C e poi immediatamente in ghiaccio.

Component	Per reaction (µL)	× 8 rxns (µL) (including excess)	× 24 rxns (µL) (including excess)	× 48 rxns (µL) (including excess)
Nuclease-Free Water	2.0*	17*	50*	100*
5× Reaction Buffer	10.0	85	250	500
10× dNTPs	5.0	42.5	125	250
Cyanine 3-dUTP *or* Cyanine 5-dUTP	3.0	25.5	75	150
Exo (-) Klenow	1.0	8.5	25	50
Final volume of Labeling Master Mix	**19.0 or 21.0***	**161.5 or 178.5***	**475 or 525***	**950 or 1050***

Tabella 6. Componenti della labeling Master Mix

4) Procedere alla preparazione della Cy3/Cy5-Labeling Master Mix come indicato in Tabella 4.

5) Dispensare 19 µl di Labeling Master Mix in ciascun tubo di reazione che contiene il DNA genomico per ottenere un volume finale di 50 µl.

6) Transferire i campioni in termociclatore per 2 ore a 37° C, 10 minuti a 65°C e riporre in ghiaccio.

STEP 2: PURIFICAZIONE

1) Aggiungere 430 µl di 1X TE (pH 8) in ciascuna provetta e caricare i campioni marcati nel filtro di purificazione (Amicon 30kDa) precedentemente inserito in una provetta di raccolta da 2 ml.

2) Centrifugare per 10 minuti a 14000xg a temperatura ambiente, svuotare la provetta e re-inserire il filtro.

3) Caricare altri 480 µl di TE in ciascun filtro e ripetere la centrifugazione come al punto 2.

4) Recuperare la colonnina ed inserirla, rovesciata, in una nuova provetta da 2 ml, centrifugare per 1 minuto a 1000xg a temperatura ambiente per raccogliere il campione purificato.

5) Concentrare il campione in Speed Vac Concentrator fino a portarlo all'essicazione e reidratarlo con 21 µl di H2O nuclease-free.

6) Misurare in spettrofotometria concentrazione di DNA (ng/µl), grado di purezza (260/280) e pmol per µl di marcatura. Calcolare attività specifica e resa secondo le formule:

yield (µg) = concentrazione DNA (ng/µl) X volume del campione (µl) / 1000 (ng/µg)

Attività specifica = pmol per µl di DNA marcato/ µg per µl DNA genomico.

7) Sulla base di questi valori combinare test e rispettivo reference.

STEP 3: IBRIDAZIONE

1) Preparare l'Hybridization Master Mix come indicato in Tabella 5.

Component	Volume (µL) per hybridization	× 8 rxns (µL) (including excess)	× 24 rxns (µL) (including excess)	× 48 rxns (µL) (including excess)
Cot-1 DNA (1.0 mg/mL)*	5	42.5	125	250
10× aCGH Blocking Agent†	11	93.5	275	550
2× HI-RPM Hybridization Buffer†	55	467.5	1,375	2,750
Final Volume of Hybridization Master Mix	71	603.5	1,775	3,550

* Use Cot-1 DNA (1.0 mg/mL) from the appropriate species.

Tabella 7. Componenti dell'Hybridization Master Mix

1) Aggiungere 71 µl di mix di ibridazione ad ogni campione composto da DNA test e DNA reference.

2) Trasferire i campioni in termociclatore per 3 minuti a 98°C e 30 minuti a 37°C.

3) Centrifugare per 1 minuto a 37°C a 6000xg.

4) Caricare ciascun campione sul rispettivo pozzetto ed ibridare.

5) Trasferire la cameretta di ibridazione in stufa rotante per 24 ore a 67°C e 20 r.p.m.

STEP 4: LAVAGGI

1) Riempire una glass dish identificata con il numero #1 con Agilent Oligo aCGH / ChIP-on-Chip Wash Buffer 1 a temperatura ambiente.

2) Inserire un rack in una glass dish (#2), aggiungere una barra di agitazione magnetica e riempiere con Agilent Oligo aCGH / ChIP-on- Chip Wash Buffer 1 a temperatura ambiente. Posizionare su agitatore magnetico.

3) Riempire una glass dish (#3) con soluzione Agilent Oligo aCGH / ChIP-on-Chip Wash Buffer 2 a 37°C ed aggiunge una barra di agitazione.

4) Preparare le glass dish #4 e #5 riempendole rispettivamente con acetonitrile e Stabilization and Drying Solution.

5) Rimuovere la camera di ibridazione dall'incubatore e procedere con il disassemblaggio del vetrino nella glass dish #1.

6) Immergere l'array nella glass dish #2 e lavare per 5 minuti in agitazione.

7) Transferire il rack nel glass dish #3 e lasciare in agitazione per 1 minuto.

8) Transferire il rack nel glass dish #4 lasciare in agitazione per 10 secondi.

9) Ripetere l'operazione nel glass dish #5 e lasciare in agitazione per 30 secondi.

10) Rimuovere lentamente il rack, trasferire l'array nella cameretta di scansione e coprire con un ozone-barrier slide. Inserire nello scanner.

La scansione degli array è stata eseguita con DNA Microarray Scanner with Sure Scan High-Resolution Technology (Agilent), in grado di rilevare i livelli di fluorescenza (Cy5 e Cy3) emessi da ogni singolo spot, e di fornire l'immagine raw su cui sono stati eseguiti controllo di qualità tramite Agilent Feature Extraction v10.7 ed analisi dei dati.

L'elaborazione dei dati è stata eseguita con il software Agilent CytoGenomics v.4.0.3.12, la chiamata delle CNV con l'algoritmo ADM-2 settato per riconoscere alterazioni coinvolgenti tre sonde consecutive con una MAALR (absolute average Log Ratio) di ± 0,25.

Il software elabora per ogni spot il valore log2 del rapporto tra le intensità di fluorescenza emesse da Cy5 e Cy3. I dati così ottenuti sono poi sottoposti ad un processo di normalizzazione (sottrazione del background e correzione mediana) in grado di rimuovere le distorsioni sistematiche eventualmente presenti nell'esperimento.

Il rapporto logaritmico risultante tra le intensità delle fluorescenze è proporzionale al rapporto tra il numero di copie delle sequenze di DNA test e reference.

Valori attorno al valore "zero" del log2 campione / reference indicano un uguale dosaggio della regione per test e controllo. Sonde che si spostano verso valori di "+0.58" (log2 3/2) sono indicativi di una duplicazione e regioni che si spostano verso il valore "-1" (log2 ½) indicano una delezione. Tutte

le CNV de-novo e le varianti ereditate che mostravano un profilo array ambiguo sono state validate in Real-Time PCR mediante TaqMan o SybrGreen assay.

Appendice 2

A2.1 Estrazione dell'Rna

1) Versare il contenuto della provetta in una falcon da 50 ml e aggiungere PBS 1X fino a raggiungere il volume totale di 12 ml.
2) Vortexare la falcon vigorosamente per 30 secondi e successivamente centrifugare per 30 minuti a 4 °C a 4000xg.
3) Eliminare il surnatante.
4) Lasciare la falcon capovolta per 1 o 2 minuti su carta assorbente ed assorbire il liquido rimanente.
5) Aggiungere 440 µl di RNA Purification Resuspension Solution e risospendere il pellet di RNA.
6) Mantenere i campioni in ghiaccio mentre si preparano gli step di purificazione.
7) Aggiungere 100 µl di RNA Purification Wash solution 1 nel filtro di purificazione.
8) Risospendere l'RNA nel filtro e centrifugare per 30 secondi a 16000 x g.
9) Eliminare il surnatante, aggiungere 500 µl di Purification Wash Solution 1 e centrifugare per 30 secondi a 16000 x g.
10) Eliminare il surnatante e aggiungere 500 µl di Wash Solution 2 e centrifugare per 30 secondi a 16000 x g.
11) Eliminare il surnatante, aggiungere 100 µl di Absolute RNA Wash Solution e incubare per 15 minuti a temperatura ambiente.
12) Aggiungere 500 µl di RNA Purification Wash Solution2 e incubare per 5 minuti a temperatura ambiente e centrifugare 30 s a 16000 x g.
13) Svuotare il surnatante e centrifugare per 30 secondi a 16000 x g.
14) Trasferire il filtro di purificazione in una nuova provetta, aggiungere 90µl di Nucleic Acid Purification Elution Solution e incubare per 2 minuti a 70 °C gradi.
15) Centrifugare per 30 secondi a 16000 x g.
16) Riconcentrare l'eluito nel filtro e centrifugare per 30 secondi a 16000 x g.
17) Eliminare il filtro, trasferire l'eluito in una nuova provetta e conservare a -80 °C.

A2.2 Rna-Seq: TruSeq® Stranded mRNA

1.1 Purificazione e frammentazione dell'mRNA

1) Impostare nel termociclatore i seguenti programmi:
   - 65° C per 5 minuti e mantenere a 4°C e salvare come mRNA Denaturation
   - 80° C per 2 minutes, mantenere a 25°C e salvare mRNA Elution 1
   - 94° C per 8 minuti, mantenere a 4°C e salvare come Elution 2 - Frag - Prime

2) Diluire l'Rna totale in acqua ultrapura nuclease-free fino ad un volume finale di 50 µl in ogni pozzetto di della piastra RBP.
3) Aggiungere 50 µl di RNA Purification Beads in ciascun pozzetto della piastra RBP. Pipettare delicatamente l'intero volume e miscelare accuratamente.
4) Posizionare la piastra sigillata RBP sul termociclatore e selezionare il programma mRNA denaturation.
5) Incubare a temperatura ambiente per 5 minuti.
6) Rimuovere la pellicola adesiva dalla piastra RBP.
7) Posizionare la piastra RBP sul supporto magnetico per 5 minuti a temperatura ambiente.
8) Eliminare il surnatante e togliere la piastra RBP dal supporto magnetico.
9) Lavare le beads aggiungendo 200 µl of Bead Washing Buffer in ciascun pozzetto della piastra RBP.
10) Collocare la piastra RBP sul supporto magnetico per 5 minuti a temperatura ambiente.
11) Centrifugare l'Elution Buffer a 600 xg per 5 secondi.
12) Eliminare il supernatante da ciascun pozzetto della piastra RBP.
13) Rimuovere la piastra RBP dal supporto magnetico e aggiungere 50 µl di Elution Buffer in ciascun pozzetto. Pipettare e miscelare l'intero volume.
14) Chiudere la piastra con una guarnizione adesiva, posizionare la piastra sigillata RBP sul termociclatore e selezionare il programma mRNA Elution 1.
15) Posizionare la piastra a temperatura ambiente e rimuovere la membrana adesiva.
16) Centrifugare il Bead Binding Buffer a 600 × g per 5 secondi.
17) Aggiungere 50 µl di Bead Binding Buffer in ciascun pozzetto della piastra RBP pipettando e miscelando l'intera soluzione.
18) Incubare la piastra RBP per 5 minuti a temperatura ambiente.

19) Posizionare la piastra RBP sul supporto magnetico per 5 minuti a temperatura ambiente.
20) Eliminare il surnatante.
21) Rimuovere la piastra RBP dal supporto magnetico e lavare le beads aggiungendo 200 µl of Bead Washing Buffer in ciascun pozzetto della piastra.
22) Mettere la piastra RBP sul supporto magnetico per 5 minuti a temperatura ambiente.
23) Eliminare il surnatante e togliere la piastra dal supporto magnetico.
24) Aggiungere 19.5 µl di Fragment, Prime, Finish Mix in ciascun pozzetto della piastra RBP.
25) Chiudere la piastra con una guarnizione adesiva, collocare la piastra sul termociclatore e selezionare il programma Elution 2 - Frag – Prime.
26) Centrifugare brevemente la piastra RBP e procedere immediatamente agli step successivi.

2. Sintetisi del primo filamento di cDNA.

1) Impostare nel termociclatore il seguente programma:
   - 25°C per 10 minuti, 42°C per 15 minuti, 70°C per 15 minuti • lasciare a 4°C e salvare come Synthesize 1st Strand.
2) Rimuovere la membrana adesiva dalla piastra RBP e posizionare la piastra sul supporto magnetico per 5 minuti a temperatura ambiente.
3) Trasferire 17 µl di surnatante dalla piastra RBP alla nuova piastra CDP.
4) Aggiungere 50 µl SuperScript II nella mix First Strand Synthesis Act D.
5) Aggiungere 8 µl of First Strand Synthesis Act D Mix e SuperScript II mix in ciascun pozzetto della piastra CDP e pipettare e miscelare bene.
6) Chiudere la piastra CDP con una guarnizione adesiva, centrifugare brevemente.
7) Mettere la piastra sul termociclatore e correre il programma Synthesize 1st Strand.

3. 1 Sintesi del secondo filamento di cDNA

1) Rimuovere la membrana adesiva dalla piastra CDP e aggiungere 5 µl di End Repair Control (diluito 1/50 in Resuspension Buffer).
2) Aggiungere 20 µl Second Strand Marking Master Mix a ciascun pozzetto della piastra CDP.
3) Chiudere la piastra CDP con la membrana adesiva e incubare a 16°C per 1 ora.
4) Lasciare la piastra a temperatura ambiente.
5) Aggiungere 90 µl AMPure XP in ciascun pozzetto della piastra CDP.

6) Incubare la piastra per 15 minuti a temperatura ambiente.
7) Collocare la piastra sul supporto magnetico per 5 minuti a temperatura ambiente.
8) Eliminare 135 µl di surnatante da ciascun pozzetto della piastra CDP.
9) Mettere la piastra sul supporto magnetico e aggiungere 200 µl di EtOH 80% a ciascun pozzetto.
10) Incubare a temperatura ambiente per 30 secondi e ripetere questi passaggi per 2 volte.
11) Lasciare la piastra per 15 minuti a temperatura ambiente e rimoverla dal supporto magnetico.
12) Aggiungere 17.5 µl Resuspension Buffer a ciascun pozzetto della piastra CDP pipettando e miscelando la soluzione.
13) Incubare per 2 minuti a temperatura ambiente.
14) Trasferire la piastra sul supporto magnetico e lasciarla a temperatura ambiente per 5 minuti.
15) Trasferire 15 µl (ds cDNA) di surnatante dalla piastra CDP ad una piastra marcata ALP.

4. Adenylate 3′ Ends

1) Impostare nel termociclatore il seguente programma:
   37 °C per 30 minuti, 70°C per 5 minuti lasciare a 4°C e salvarlo come ATAIL70.
2) Aggiungere 2.5 µl di Resuspension Buffer in ogni pozzetto della piastra ALP.
3) Aggiungere 12.5 µl A-Tailing Mix in ogni pozzetto, pipettare e miscelare bene.
4) Chiudere la piastra con la membrana adesiva, disporre la piastra nel termociclatore e correre il programma ATAIL70.

5. 1 Ligate Adapters

1) Lasciare la piastra ALP per 10 minuti a temperatura ambiente e centrifugarla per 1 minuto a 280 x g.
2) Rimuovere la membrana adesiva e aggiungere 2.5 µl di Resuspension Buffer a ciascun pozzetto.
3) Aggiungere 2.5 µl di Ligation Mix in ciascun pozzetto della piastra ALP.
4) Chiudere la piastra con la membrana adesiva e centrifugare la piastra a 280 x g per 1 minuto.
5) Collocare la piastra nel termociclatore e incubare a 30°C per 10 minuti.
6) Rimuovere la membrana adesiva e aggiungere 5 µl di Stop Ligation Buffer a ciascun pozzetto della piastra pipettando e miscelando bene.

7) Aggiungere 42 µl di AMPure XP Beads a ciascun pozzetto della piastra ALP e incubare per 15 minuti a temperatura ambiente.
8) Collocare la piastra sul supporto magentico per 5 minuti. Rimuovere ed eliminare 79.5 µl di surnatante da ciascun pozzetto.
9) Lasciare la piastra sul supporto magnetico e aggiungere 200 µl di EtOH 80%, incubare per 30 secondi a temperatura ambiente. Ripetere questo step due volte.
10) Lasciare la piastra sul supporto magnetico per 15 minuti a temperatura ambiente.
11) Rimuovere la piastra ALP dal supporto magnetico e aggiungere 52.5 µl di Resuspension Buffer. Incubare per 2 minuti a temperatura ambiente.
12) Mettere la piastra sul supporto magnetico per 5 minuti a temperatura ambiente.
13) Trasferire 50 µl di surnatante in ciascun pozzetto della nuova piastra marcata come CAP.
14) Aggiungere 50 µl di AMPure XP Beads in ciascun pozzetto della piastra CAP, pipettando e miscelando la soluzione.
15) Incubare la piastra CAP a temperatura ambiente per 15 minuti.
16) Lasciare la piastra per 5 minuti sul supporto magnetico ed eliminare 95 µl di surnatante da ciascun pozzetto della piastra CAP.
17) Aggiungere 200 µl a ciascun pozzetto della piastra posizionata sul supporto magnetico.
18) Incubare la piastra per 30 secondi a temperatura ambiente ed eliminare il surnatante. Ripetere questi step per due volte.
19) Lasciare la piastra sul supporto magnetico per 15 minuti.
20) Aggiungere 22.5 µl di Resuspension Buffer, pipettare e miscelare bene l'intero volume.
21) Incubare la piastra per 2 minuti a temperatura ambiente.
22) Trasferire la piastra sul supporto magnetico per 5 minuti.
23) Trasferire 20 µl di surnatante nei rispettivi pozzetti della nuova piastra marcata come PCR.

6. Arricchimento dei frammenti di cDna

1) Impostare nel termociclatore il seguente programma:

   - 98 °C per 30 secondi, 15cicli: 98 °C per10 secondi, 60 °C per 30 secondi, 72 °C per 30 secondi, 72 °C per 5 minuti, mantenere a 4°C e salvalo come PCR.
2) Aggiungere 5 µl di PCR Primer Cocktail a ciascun pozzetto della piastra PCR.
3) Aggiungere 25 µl di PCR Master Mix.

4) Chiudere la piastra con la membrana adesiva, posizionare la piastra nel termociclatore e correre il programma PCR.
5) Rimuovere la membrana adesiva, aggiungere 47.5 µl AMPure XP Beads in ciascun pozzetto della piastra e incubarla a temperatura ambiente per 15 minuti.
6) Posizionare la piastra sul supporto magnetico per 5 minuti.
7) Eliminare 95 µl di surnatante da ogni pozzetto della piastra.
8) Lasciare la piastra sul supporto magnetico e aggiungere 200 µl di EtOH 80%.
9) Incubare la piastra a temperatura ambiente per 30 secondi, rimuovere il surnatante. Ripetere questi passaggi per 2 volte.
10) Lasciare la piastra sul supporto magnetico per 15 minuti.
11) Rimuovere la piastra e aggiungere 32.5 µl Resuspension Buffer in ogni pozzetto della piastra pipettando l'intera soluzione.
12) Incubare la piastra PCR per 2 minuti a temperatura ambiente.
13) Posizionare la piastra sul supporto magnetico per 5 minuti.
14) Trasferire 30 µl di surnatante da ciascun pozzetto nella piastra PCR nel rispettivo pozzetto della piastra marcata TSP1.

Quantifica le librerie

Per ottenere dati di altissima qualità sulle piattaforme di sequenziamento Illumina, è importante creare densità di cluster su ogni corsia della flow-cell. L'ottimizzazione delle densità di cluster richiede un'accurata quantificazione delle librerie di DNA.

La qualità della libreria è stata testata usando Bioanalyzer 2100 correndo 1 µl di libreria non diluita.

7. Normalizzazione e unione delle librerie

1) Trasferire 10 µl di library da ciascun pozzetto della piastra TSP1 ai corrispondenti pozzetti della piastra marcata MIDI.
2) Normalizzare la concentrazione della libreria in ciascun pozzetto della piastra DCT a 10 nM usando Tris-HCl 10 mM, pH 8.5 with 0.1% Tween 20.
3) Pipettare e miscelare bene l'intera libreria normalizzata.
4) Procedere alla generazione dei cluster sulla piattaforma **NovaSeq** dell'Illumina.

Appendice 3

A3.1 DGV

Le CNV identificate mediante array CGH sono state innanzitutto classificate in "Rare" o "Comuni". Questa classificazione è stata eseguita mediante il seguente script di R (R core team, 2012)

```
library("xlsx")
patient <- readline(prompt="Type patient ID: ")
print("Loading GRCh37hg19variants2016-05-15",quote=F)
print("15 seconds needing",quote=F)

# Change work directory
Setwd("path/to/folder ")
load("GRCh37hg19variants2016-05-15.Rdata")

print("Databased loaded",quote=F)
print("Cutoff 80% overlap (row 38 to change)",quote=F)
cutoff <- 80

#adding the first column in the region file: 1:nrow
# Change work directory
setwd("path/to/folder")
print("Reading input file",quote=F)
regions <- read.table("regions.txt",header=F)
righe <- nrow(regions)
col1 <- seq(1,righe,1)
regions <- data.frame(col1,regions)
diag <- data.frame()

for (i in 1:nrow(regions)){
```

```
region <- regions[i,1]
chr <- as.vector(regions[i,2])
start <- regions[i,3]
end <- regions[i,4]
cns <- as.character(regions[i,5])

dgvch <- dgv[dgv$chr == chr,]
dgvch <- dgvch[,c(1:7,14:17)]

#computing lenght input CNV
len <- (end - start)

#selecting CNV type A
selectedA <- dgvch[dgvch$start >= start & dgvch$end <= end,]
#computing lenght of CNV type A
lenA <- (selectedA$end - selectedA$start)
#computing overlap between patient CNV and DGV cnv
overlap <- (lenA/len)*100
#preparing dataframe and selecting CNV with cutoff
selectedA <- cbind(selectedA,overlap)
similarA <- selectedA[selectedA$overlap > cutoff,]

#selecting CNV type B: overlap is always 100%
selectedB <- dgvch[dgvch$start < start & dgvch$end > end,]
lenB <- (selectedB$end - selectedB$start)
overlap <- rep(100,nrow(selectedB))
selectedB <- cbind(selectedB,overlap)

#selecting CNV type C
selectedC <- dgvch[dgvch$start < start,]
selectedC1 <- selectedC[selectedC$end >= start & selectedC$end <= end,]
```

```
#computing overlap between patient CNV and DGV cnv
overlap <- ((selectedC1$end - start)/len)*100
#preparing dataframe and selecting CNV with cutoff
selectedC <- cbind(selectedC1,overlap)
similarC <- selectedC[selectedC$overlap > cutoff,]

#selecting CNV type D
selectedD <- dgvch[dgvch$start >= start & dgvch$start <= end,]
selectedD1 <- selectedD[selectedD$end > end,]
overlap <- ((end - selectedD1$start)/len)*100
selectedD <- cbind(selectedD1,overlap)
similarD <- selectedD[selectedD$overlap > cutoff,]

#report of total CNV
totalCNV <- rbind(selectedA,selectedB,selectedC1,selectedD1)
reportCNV <- summary(totalCNV$variantsubtype,maxsum=1000)

#sum of Del,Losses,Gain,Dup,Inser,Gain+Loss,
#Tandem Dup in that region
#1 complex
#2 deletion
#3 duplication
#4 gain
#5 gain+loss
#6 insertion
#7 inversion
#8 loss
#9 mobile element insertion
#10 novel sequence insertion
#11 sequence alteration
#12 tandem duplication
```

```
tot <- sum(reportCNV[c(2,3,4,5,6,8,12)])
gain <- sum(reportCNV[c(3,4,5,6,12)])
loss <- sum(reportCNV[c(2,5,8)])
insertion <- sum(reportCNV[c(6)])
otherElm <- sum(reportCNV[c(1,7,9,10,11)])

#report of similar CNV
similarCNV <- rbind(similarA,selectedB,similarC,similarD)
reportSimilar <- summary(similarCNV$variantsubtype,maxsum=1000)
totS <- sum(reportSimilar[c(2,3,4,5,6,8,12)])
#sum of Del,Losses,Gain,Dup,Inser,Gain+Loss in that region
gainS <- sum(reportSimilar[c(3,4,5,6,12)])
lossS <- sum(reportSimilar[c(2,5,8)])
insertionS <- sum(reportSimilar[c(6)])
otherElmS <- sum(reportSimilar[c(1,7,9,10,11)])

#report for diagnosis
diag <-  data.frame(Chr = chr,Start = start, End = end,
                    cns = cns,TotalGains = gain,SimilarGains = gainS,
                    TotalLosses = loss, Similarlosses = lossS,TotIns=insertion,
                    SimilarInsertions=insertionS)
write.table(diag,paste(region,"dia.txt",sep=""),quote=F,
          col.names=T,row.names=F,sep="\t")
print(paste("Region",i,"analyzed",sep=" "),quote=F)
}
filenames <- list.files(pattern="*dia.txt")
results <- data.frame()

for (j in 1:length(filenames)){
file <- filenames[j]
```

```
dgvinfo <- read.table(paste(file),header=T,sep="\t")
results <- rbind(results,dgvinfo)
}
print("**Removing tmp files**",quote=F)
file.remove(filenames)

#adding column rare/common
rareCom <- data.frame()
for (i in 1:nrow(results)){
cns <- results[i,4]
simgains <- results[i,6]
simlosses <- results[i,8]

if (cns == "Dup" & simgains <= 3) rareCom[i,1] <-c("Rara")
if (cns == "Del" & simlosses <= 3) rareCom[i,1] <-c("Rara")
if (cns == "Dup" & simgains > 3) rareCom[i,1] <-c("Comune")
if (cns == "Del" & simlosses > 3) rareCom[i,1] <-c("Comune")
}
colnames(rareCom)<- ("Status")
results <- data.frame(results,rareCom)

print("Writing result file",quote=F)
write.xlsx(results,paste(patient,"results.xlsx",sep=""),col.names=T, row.names=F)
print("**Deleting workspace**",quote=F)
print(paste("Results in ***",patient,"results.xlsx***",sep=""),quote=F)
rm(list=ls())
```

A3.2 STAR 2.5.3a

Per l'allineamento dei dati abbiamo usato STAR2.5.3a (Dobin A D. C., 2013), implementando il seguente script nel linguaggio Bash:

```
#!/bin/bash
STARversion="STAR2.5.3a"
function man() {
cat << EOFMAN
usage: ./staraligner.sh [-n|--nThread] THREADNUMBER [-g|--GenDir] GENOMEDIRECTORY [-p|--paired] [-f|--FastQ] PATHTOFASTQ (R1 and R2 if paired) [-b|--outSAMsf] FILTEROUTINTRON [-o|--OutDir] OUTPUTDIRECTORY [-c|--outSAMt] OUTPUTTYPE [-d|--outSAMu] OUTPUTUNMAPPEDREADS [-m|--MisMatch] MISMATCHNUMBER [-o|--OutDir] [-e|--readFileCommand] READCOMMAND
OPTION
 -h| --help                                  Print this usage
-n| --nThread (1-8)                          Number of threads
-g| --genomeDir path/to/genome               Directory where genome file are stored
-p| --paired                                 Data is paired-end, -f is followed by R1 and R2
-f| --readFilesIn path/to/FASTQfile          Paths to files that contain input read1 (and read2 for paired-end data)
-b| --outSAMstrandField (None/intronMotif)   Reads with inconsitent or non canonical introns are filtred out
-o| --outFileNamePrefix path/to/dir          Output files name prefix
 -c| --outSAMtype
              (SAM                           Default
              BAM                            Output BAM without sorting
               None                          No BAM/SAM output
               Unsorted                      Standard unsorted
               SortedByCoordinate)           Sorted by coordinate
-d| --outSAMunmapped                        Output of unmapped reads
               (None                         Default
             Within)                         Output unmapped reads within the main SAM file
   -m| --outFilterMismatchNmax (-n)          Alignment will be output only if it has no more mismacthes than this value.
(n=14 QuantSeq; n=10 default)


EOFMAN
exit
}

if [ $# -lt 1 ]
then
echo "No arguments were given." && man && exit;
fi
```

```
while [ $# -ge 1 ]
do
key=$1
nThread=1
MisMatch=10
outSAMt=BAM
outSAMsf=intronMotif
outSAMu=Within

        case $key in

        -h | --help )
        man && exit
         ;;
        -g | --GenDir )
        GENDIR=$2
        shift
        if [ ! -d "$GENDIR" ] ;
        then
        echo "The Genome directory $GENDIR does not exist." && exit;
        fi
        ;;
        -p | --paired )
        $paired=Y
        shift/
        ;;
        -f | --FastQ )
        if [ -z "$paired" ] ;
        then    # single-end mode
                FastQ=$2
                outputsubdir=$(basename $FastQ | sed 's/.fastq.gz/star/')
                shift
        else
```

```
            # paired-end mode
            if [ ! -e "$2" ] || [ ! -e "$3" ] ; then echo "Two fastQ are expected in paired-end mode!"
&& exit; fi
            FastQ="$2 $3"
            outputsubdir=$(basename $2 | sed 's/.fastq.gz/star/')
            shift; shift
        fi
        ;;
        -o | --OutDir )
        OutDir=$2
        shift
        ;;
        -n | --nThread )
        nThread=$2
        shift
        ;;
        -m | --MisMatch )
        MisMatch=$2
        shift
        ;;
        -b | --outSAMsf )
        outSAMsf=$2
        shift
        ;;
        -c | --outSAMt )
        outSAMt=$2
        shift
        ;;
        -d | --outSAMu )
        outSAMu=$2
        shift
        ;;
        -e | --readFileCommand )
        readFileCommand=$2
```

```
        shift
        ;;
        * )
        echo "Command not found, read usage" && man && exit
        ;;
esac
shift
done

if [ ! -d $OutDir/$outputsubdir ] ;
then
mkdir -p $OutDir/$outputsubdir
fi
#exit;
condactivate; \
if [ $STARversion != `STAR --version` ]; \
then \
echo "Your STAR version is not correct. You need $STARversion"; \
exit ; \
fi  ; \
STAR \
--runThreadN $nThread \
--genomeDir $GENDIR \
--readFilesIn $FastQ \
--outSAMstrandField $outSAMsf \
--outFileNamePrefix $OutDir/$outputsubdir/ \
--outSAMtype $outSAMt SortedByCoordinate \
--outSAMunmapped $outSAMu \
--outFilterMismatchNmax $MisMatch \
--readFilesCommand zcat
```

A3.3 Features Count

```
#       -inputfiles             The name of input files that include the read mapping results (BAM or SAM)
#       -a (file.GTF o .GFF o .SAF)     Give the name of an annotation file
#       -o <string>             Give the name of the output file.
#
#       [options]
#
#-g (GTF .attrType)             Specify the attribute type used to group features when GTF annotaion is provided.
                                geneid default
#-Q <int>                       The minimum mapping quality score a read must satisfy in order to be counted. 0 by default.
#-T <int>                       Number of the Threads. The value should be between 1 and 32. 1 by default.


inputfile=$inputbam
outputfile=$outputcounts
nthread=4
attrgtf=genename
qs=0
strand=2
featureType=exon

cd $mydir
condactivate subread
#if ! [[ `featureCounts -v` =~ "v1.6.2" ]]; then
#echo "Your featureCounts version is not correct. You need version 1.6.2"; \
#exit
#fi
FEATURESCOUNTS
featureCounts --tmpDir $TMPDIR \
        -T $nthread \
        -g $attrgtf \
        -Q $qs \
        -a $pathgtf \
        -s $strand \
        -t $featureType \
        -F GTF \
```

```
        -o $outputfile \
         $inputfile \
        2> ${outputfile}.log
```

A3.4 Analisi Esplorativa dei Dati

A3.4.1 Multidimensional Scaling (MSD)

```
library(limma)

library(edgeR)

setwd("path/to/file")

counts <- read.table("counts", header = T, row.names = "GeneID")


#delete sample 1 sample

counts <- counts[,colnames(counts) != "    "]


logR <- read.table("asdsib.logr", header = F)

metadata <- read.table("metadata.txt", header = T)


#Checking metadata and counts samples order

all(colnames(counts) == metadata$patient)


#Normalizing counts

d0 <- DGEList(counts)

d0 <- calcNormFactors(d0)


#Filtering low-expressed genes
```

```
cutoff <- 1

drop <- which(apply(cpm(d0), 1 , max) < cutoff)

d <- d0[-drop,]

dim(d)

#MDS plot coloring by status

plotMDS(d, col = as.numeric(as.factor(metadata$status)))


mm <- model.matrix(~ 0 + as.factor(metadata$status) + metadata$age +

                   metadata$sex + metadata$puberty + metadata$drugs)


colnames(mm) <- c("ASD", "SIB","Age", "Sex", "Puberty", "Drugs")


#Normalization

v <- voom(d, design = mm, plot = T)
```

A3.4.2 Principal Component Analysis

```
library(mixOmics)

library(edgeR)

count = read.table("counts.gz", header = T, check.names=F, row.names = 1, stringsAsFactors=F)
meta = read.table("metadata.txt", header = T, row.names = 1)
#keep = row.names(meta[which(meta$gender == "M"),])
#countmale = count[,keep]
count <- count[,colnames(count) != "  "]

count2  <-DGEList(counts=count, genes=row.names(count))
count2 = calcNormFactors(count2, method="RLE")
gene.len=read.table("genelengths.v28.txt",stringsAsFactors=F,row.names=1, header=T)
row.names(gene.len)=gene.len$GeneID
```

```
sorted.gene.len=gene.len[match(row.names(count),row.names(gene.len)),]
rpkmM = rpkm(count2, log=T, gene.length=sorted.gene.len$Length)
isexpr=rowSums(cpm(count2)>1) >= 4
rpkmM = rpkmM[isexpr,]

write.table(rpkmM, sep="\t", row.names=TRUE, quote=FALSE, file="rpkmMfit4.txt")

data = read.table("rpkmMfit4.txt", check.names=F,stringsAsFactors=F,h=T,row.names=1)
pca=mixOmics::pca(t(data) ,ncomp=24 ,center=T,scale=T)

var = round(matrix(((pca$sdev^2)/(sum(pca$sdev^2))), ncol=1)*100,1)

metaM= read.table("metadata.txt", header = T, row.names = 1)
sorted.meta=metaM[match(row.names(pca$x),row.names(metaM)),]
meta = sorted.meta

mixOmics::plotIndiv(pca, group=meta$status, pch=as.integer(meta$drugs), pch.levels=meta$drugs,
legend=T, comp=c(1,2), style="graphics", legend.position="top")
```

A3.5 DESeq2

```
# set work directory and load counts matrix, metadata and library
 setwd("/path/to/folder")
#suppressMessages(library(DESeq2))


 count = read.table("countsmatrix.gz", header = T, stringsAsFactors = F, row.names = 1, check.names
= F)
 meta = read.table("metadata.txt", header = T)
 #delete sample F3537
counts <- counts[,colnames(counts) != "F3537"] dim(count)
```

```
dds24 = DESeqDataSetFromMatrix(countData = counts, colData = meta, design = ~ covariate +
covariate2 + )

deseq = DESeq(dds, betaPrior = FALSE, test = "Wald", fitType = "parametric",
minReplicatesForReplace = Inf)

res = results(deseq, contrast = c("covariate", "status1", "status2"))


# results number with padj < 0.05

sum(resF$padj < 0.05, na.rm = T)

# order the results based on padj and write it on .csv file

resorder = resF[order(resF$padj),]

 write.csv(resFo, "ris.csv")


# Make a  Volcano plot

res <- read.csv("resFo ", header=TRUE)

head(res)

# Make a basic volcano plot

with(res, plot(log2FoldChange, -log10(pvalue), pch=20, main="Volcano plot", xlim=c(-4.5,4)))


# Add colored points: red if padj<0.05, orange of log2FC>1, green if both)

with(subset(res, padj<.05 ), points(log2FoldChange, -log10(pvalue), pch=20, col="red"))

with(subset(res, abs(log2FoldChange)>1), points(log2FoldChange, -log10(pvalue), pch=20,
col="orange"))

with(subset(res, padj<.05 & abs(log2FoldChange)>1), points(log2FoldChange, -log10(pvalue),
pch=20, col="green"))


# Optional: Label points with the textxy function from the calibrate plot
```

```
library(calibrate)

with(subset(res, padj<.05 & abs(log2FoldChange)>1), textxy(log2FoldChange, -log10(pvalue),
labs=Gene, cex=.8))
```

A3.6 Analisi del Biotipo

L'analisi del Biotipo dei geni differenzialmente espressi è stata automatizzata mediante il seguente script di R:

```
Load required library
suppressMessages(library(AnnotationDbi)
suppressMessages(library(org.Hs.eg.db))
suppressMessages(library(dplyr))
suppressMessages(library(EnsDb.Hsapiens.v86))
suppressMessages(library(TxDb.Hsapiens.UCSC.hg19.knownGene))
ENTREZIDorg <- keys(org.Hs.eg.db, keytype = "ENTREZID")

Rename the column
colnames(genedataframeorgDb) <- c("Entrez", "Ensembl", "HGNC")
colnames(genedataframeorgDb) <- paste(colnames(genedataframeorgDb), "orgDb", sep = "")

 Additional column with HGNC added for merging
genedataframeorgDb$HGNC <- genedataframeorgDb$HGNCorgDb
ENSEMBLEnsDb <- keys(EnsDb.Hsapiens.v86, keytype = "GENEID")
genedataframeEnsDb <- ensembldb::select(EnsDb.Hsapiens.v86, keys=ENSEMBLEnsDb,
columns=c("ENTREZID", "SYMBOL", "GENEBIOTYPE"), keytype="GENEID")
colnames(genedataframeEnsDb) <- c("Ensembl", "Entrez", "HGNC", "GENEBIOTYPE")
colnames(genedataframeEnsDb) <- paste(colnames(genedataframeEnsDb), "EnsDb", sep = "")
genedataframeEnsDb$HGNC <- genedataframeEnsDb$HGNCEnsDb
ENTREZIDTxDb <- keys(TxDb.Hsapiens.UCSC.hg19.knownGene, keytype = "GENEID")
ENTREZIDTxDbdf <- data.frame(ENTREZID = as.character(ENTREZIDTxDb), EntrezTxDb =
ENTREZIDTxDb)
write.csv(genedataframeEnsDb, "biotypes.csv")
```

```
BioT = read.csv("biotypes.csv", header = T)
list = read.table("geni.txt")
list = as.character(list$V1)
bioT2 <- bioT[bioT$HGNC%in%list,]
WriteXLS(bioT2, "names.xls")
```

A3.7 clusterProfiler

Load required library

```
#!/usr/bin/Rscript

suppressMessages(library(clusterProfiler))

suppressMessages(library(optparse))

suppressMessages(library(org.Hs.eg.db))

suppressMessages(library(DOSE))

suppressMessages(library(ReactomePA))

suppressMessages(library(enrichplot))

suppressMessages(library(WriteXLS))



cat RISULTATI.csv | tr “,” “\t” | tr -d “”" | awk ‘$7<0.05’ | cut -f1,3 | sed “1d” > Gene

Create a list with options

option = list (

makeoption(c("-v", “–verbose”), action=“storetrue”, default=TRUE, help=“Should the program print extra stuff out? [default %default]”),

makeoption(c("-i", “–input”), action = “store”, help = “A list of genes SYMBOL and FoldChange”),

makeoption(c("-c","–cat"), action=“store”, default = 10, help = “Show number of Category to plot”),

makeoption(c("-p","–pvalue"), action = “store”, default = 0.05, help = “Set p-value cutoff”),
```

makeoption(c("-q", "–qvalue"), action = "store", default = 0.01, help = "Set q-value cutoff"),

makeoption(c("-a", "–padj"), action = "store", default = "BH", help = "Method to adjust p-value: one of holm, hochberg, hommel, bonferroni, BH, BY, fdr, none"),

makeoption(c("-f", "–fun"), action = "store", default = "MF", help = "Subontologies: it can be CC, MF, BP"),

makeoption(c("-k", "–ktype"), action = "store", default = "ENTREZID", help = "Format of gene annotation"),

makeoption(c("-o", "–output"), action ="store", default = " outputenrich", help = "Output files in .pdf and .xls format")

)

opt = parseargs(OptionParser(optionlist = option))

I load the list, convert to EntrezID and divide the genes into up and down regulated

input.df= read.table("geniconditiononly", stringsAsFactors = F, header=F)

outputname = as.character("RISconditionOnly")

input.df.entrez = bitr(input.df[,1], fromType = "SYMBOL", toType = "ENTREZID", OrgDb = "org.Hs.eg.db")

row.names(input.df)=input.df[,1]

input.df.entrez$logFC<-input.df[input.df.entrez$SYMBOL,][,2]

geneList=input.df.entrez$logFC

names(geneList)=input.df.entrez$ENTREZID

genidw = names(geneList[geneList<0])

geniup = names(geneList[geneList>0])

geneListorder = sort(geneList, decreasing = TRUE)

Gene Ontology, Enrichment Gene Ontology, Kegg and Reactome analysis for down-regulated genes

GOdw = groupGO(gene = genidw, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, level = 3, readable = TRUE)

GOl4dw = groupGO(gene = genidw, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, level = 4, readable = TRUE)

GOl5dw = groupGO(gene = genidw, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, level = 5, readable = TRUE)

EGOdw = enrichGO(gene = genidw, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, pAdjustMethod = optpadj,pvalueCutoff=optpadj, pvalueCutoff = optpadj,pvalueCutoff=optpvalue, qvalueCutoff = optqvalue,readable=TRUE)

EGOdw=simplify(EGOdw,cutoff=0.7,by="p.adjust",selectfun=min)keggdw=enrichKEGG(gene=genidw,organism="hsa",pvalueCutoff=optqvalue, readable = TRUE) EGOdw = simplify(EGOdw, cutoff=0.7, by="p.adjust", selectfun=min)

keggdw = enrichKEGG(gene = genidw, organism = "hsa", pvalueCutoff = optqvalue,readable=TRUE)EGOdw=simplify(EGOdw,cutoff=0.7,by="p.adjust",selectfun=min)

reapadw = enrichPathway(genidw, pAdjustMethod = optpadj,pvalueCutoff=optpadj, pvalueCutoff = optpadj,pvalueCutoff=optpvalue, qvalueCutoff = opt$qvalue, readable = TRUE)

Gene Ontology, Enrichment Gene Ontology, Kegg and Reactome analysis for up-regulated genes

GOup = groupGO(gene = geniup, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, level = 3, readable = TRUE)

GOl4up = groupGO(gene = geniup, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, level = 4, readable = TRUE)

GOl5up = groupGO(gene = geniup, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, level = 5, readable = TRUE)

EGOup = enrichGO(gene = geniup, OrgDb = org.Hs.eg.db, keyType = optktype,ont=optktype, ont = optktype,ont=optfun, pAdjustMethod = optpadj,pvalueCutoff=optpadj, pvalueCutoff = optpadj,pvalueCutoff=optpvalue, qvalueCutoff = optqvalue,readable=TRUE)

EGOup=simplify(EGOup,cutoff=0.7,by="p.adjust",selectfun=min)keggup=enrichKEGG(gene=geniup
,organism="hsa",pvalueCutoff=optqvalue, readable = TRUE)

keggup = enrichKEGG(gene = geniup, organism = "hsa", pvalueCutoff = opt$pvalue)

reapaup = enrichPathway(geniup, pAdjustMethod = optpadj,pvalueCutoff=optpadj, pvalueCutoff =
optpadj,pvalueCutoff=optpvalue, qvalueCutoff = opt$qvalue, readable = TRUE)

Save myself files produced in excel format

WriteXLS(GOdw@result, ExcelFileName = paste0(opt$output,"GOdw.xls"))

WriteXLS(GOl4dw@result, ExcelFileName = paste0(opt$output,"GOl4dw.xls"))

WriteXLS(GOl5dw@result, ExcelFileName = paste0(opt$output,"GOl5dw.xls"))

WriteXLS(EGOdw@result, ExcelFileName = paste0(opt$output,"EGOdw.xls"))

WriteXLS(keggdw@result, ExcelFileName = paste0(opt$output,"Keggdw.xlsx"))

WriteXLS(reapadw@result, ExcelFileName = paste0(opt$output,"Pathdw.xls"))

WriteXLS(GOup@result, ExcelFileName = paste0(opt$output,"GOup.xls"))

WriteXLS(GOl4up@result, ExcelFileName = paste0(opt$output,"GOl4up.xls"))

WriteXLS(GOl5up@result, ExcelFileName = paste0(opt$output,"GOl5up.xls"))

WriteXLS(EGOup@result, ExcelFileName = paste0(opt$output,"EGOup.xls"))

WriteXLS(keggup@result, ExcelFileName = paste0(opt$output,"Keggup.xls"))

WriteXLS(reapaup@result, ExcelFileName = paste0(opt$output,"Pathup.xls"))

Put condition on whether or not to plot the results

if (!(EGOdw@resultp.adjust<0.05||EGOup@resultp.adjust<0.05 ||
EGOup@resultp.adjust<0.05||EGOup@resultp.adjust <0.05 ||
keggdw@resultp.adjust<0.05||keggup@resultp.adjust<0.05 || keggup@resultp.adjust<0.05||keggu
p@resultp.adjust <0.05 || reapadw@resultp.adjust<0.05||reapaup@resultp.adjust<0.05 ||

```
reapaup@resultp.adjust<0.05||reapaup@resultp.adjust<0.05 ||
GSEAGO@resultpadjust<0.05||GOdw@resultpadjust<0.05 || GOdw@resultpadjust<0.05||GOd
w@resultGeneRatio !=0 ||

GOup@result$GeneRatio !=0 )){

stop("There is no significant enrichment (pvalue adjusted <0.05).")

}

Plot and save it in pdf format
pdf(paste0(opt$output,".pdf"))
if(GOdw@result$GeneRatio !=0){barplot(GOdw, x = "Count", color = "p.adjust", drop = TRUE,
showCategory = opt$cat, title = "Gene Ontology geni down-regulati")}
if(EGOdw@result$p.adjust<0.05){barplot(EGOdw, x = "Count", color = "p.adjust", drop = TRUE,
showCategory = opt$cat, title = "Geni down-regulati")}
if(keggdw@result$p.adjust<0.05){barplot(keggdw,showCategory = opt$cat)}
if(EGOdw@result$p.adjust<0.05){heatplot(GOdw, foldChange = geneList)}
if(reapadw@result$p.adjust<0.05){cnetplot(reapadw, foldChange = geneList)}

if(GOup@result$GeneRatio != 0){barplot(GOup, x = "Count", color = "p.adjust", drop = TRUE,
showCategory = opt$cat, title = "Gene Ontology geni up-regulati")}
if(keggup@result$p.adjust<0.05){barplot(keggup,showCategory = opt$cat)}
if(EGOup@result$p.adjust>0.05){barplot(EGOup, drop = TRUE, showCategory=opt$cat)}
if(reapaup@result$p.adjust<0.05){cnetplot(reapaup, foldChange = geneList, showCategory =
opt$cat)}
if(EGOup@result$p.adjust<0.05){heatplot(GOup, foldChange=geneList)}
dev.off()
```

A3.8 BedIntersect

```
sed -i 's/^/chr//g' file.bed # aggiunge chr all'inizio del file Bed;
sed -i '$dì file.bed    # elimino l'ultima riga contenente solo chr
sed -i  's/[[:punct:]]//g' file.bed   # elimino tutti i punti
 # per il download
```

```
wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencodehuman/release28/GRCh37mapping/gencode.v28lift
37.annotation.gtf.gz
bedtools intersect -loj -a file.bed -b gencode.v28lift37.annotation.gtf.gz > results.txt
awk  {print$13 } file.bed > results.txt  #mi seleziona solo la colonna di interesse

sed -i 's/"//g' results.txt
cat results.txt | cut -d . -f1 | sort -n | uniq > file.txt
#mi elimina tutto ciò che c'è dopo il punto, me li ordina numericamente, e mi toglie i duplicati.
```

A3.9 Correlazione mediante regressione lineare

```
suppressMessages(library(multtest))
suppressMessages(library(edgeR))
suppressMessages(library(ggplot2))

#importo la matrice di count
#setwd(dataDir)

counts = read.table("counts.gz", header = T, row.names = 1, check.names = F, stringsAsFactors = F)

#delete sample F3537
counts <- counts[,colnames(counts) != "F3537"]

#********ordine alfabetico per le colonne
counts <- counts[,order(colnames(counts))]

#importo la matrice dei coi valori di log-ratio CNV
cnvasd = read.table("LogRSIB.txt", header = T, row.names = 1, check.names = F, stringsAsFactors =
F)
```

```
dim(cnvasd)

#********ordine alfabetico per i pazienti
cnvasd <- cnvasd[ ,order(colnames(cnvasd))]

CNV = as.matrix(cnvasd)
keep = as.vector(colnames(CNV))

#********importo il file di metadata
metadata <- read.table("metadata.txt", header = T)

#*********seleziono campioni di interesse nel file di metadata
metadata2 <- metadata[metadata$patient%in%keep,]
metadata2 <- metadata2[order(metadata2$patient),]

#cnvsib = read.table("logratiosib.txt", header = T, row.names = 1, check.names = F, stringsAsFactors
= F)
#dim(cnvsib)

#CNV = as.matrix(cnvsib)
#keep = as.vector(colnames(CNV))

#subsetto i campioni per ottenere solo i campioni di interesse (o ASD o SIB)
counts = counts[,keep]

# normalizzo i counts in CPM e successivamente in log(CPM)
dgl = DGEList(counts=counts, genes=row.names(counts))
dgl = calcNormFactors(dgl, method="RLE")

gene.len = read.table("genelengths.v28.txt",stringsAsFactors=F,row.names=1, header=T)
row.names(gene.len) = gene.len$GeneID
```

```
sorted.gene.len =gene.len[match(row.names(counts),row.names(gene.len)),]

logCPM = cpm(counts, normalized.lib.sizes=sorted.gene.len, log=T, prior.count=0.25)

#******Compara colnames delle counts e del file logRatio, e poi dei metadata
all(colnames(logCPM) == colnames(CNV))
all(colnames(logCPM) == metadata2$patient)

#******Seleziona i geni nella count table che sono presenti nel file CNV
keepGenes <- rownames(CNV)
logCPM2 <- logCPM[which(rownames(logCPM)%in%keepGenes), ]
dim(logCPM2)
dim(CNV)

#*******in CNV ci sono piu' geni che nella count table, quindi vanno selezionati sono quelli della count table
CNV2 <- CNV[rownames(logCPM2),]

#*******Verifica se i geni sono ordinati allo stesso modo
all(rownames(CNV2) == rownames(logCPM2))

#Applico un modello di regressione lineare usando o la funzione glm o lm (che sono equivalenti.
results <- data.frame() # Crea un dataframe vuoto per i risultati

for (i in 1:nrow(logCPM2)) {
  print(paste("Computing Regression ", i))

  exp <- logCPM2[i,] #valore di expression per un gene, e ordina ID
  logR <- CNV2[i,] #log Ratio per il gene corrispondente, e ordina ID
  sex <- metadata2$sex #sex
  age <- metadata2$puberty #age
```

```
  drugs <- metadata$drugs
  family <- metadata$family

  Gene <- rownames(logCPM2)[i] #simbolo gene per RNA counts
  GeneB <- rownames(CNV2)[i]# simbolo gene per CNV
  stopifnot(Gene == GeneB)# questo e' un test per verificare se l'ordinamento e' corretto

  data.for.lm <- data.frame(exp, logR, age, sex)#crea il dataframe per la regressione

  res.lm =  lm(exp ~ logR + age + sex, data = data.for.lm) # qui si potrebbero aggiungere le covariate
come sex
  summary.res.lm <- summary(res.lm) # queste righe sono per estrarre coefficiente e p value
  beta <- summary.res.lm$coefficients[2]
  p <- summary.res.lm$coefficients[8]

  res.tmp <- data.frame(Gene, beta, p)
  results <- rbind(results, res.tmp) #aggiunge il risultato per il gene a data frame totale
}
results2 <- results[order(results$p), ] # ordina per pvalue

#*** correzione per test multipli
resAdj.tmp <- mt.rawp2adjp(results2$p, proc = "BH")

#***unisce il dataframe dei risultati con i p value corretti
results3 <- data.frame(results2, resAdj.tmp$adjp)

#****seleziona le colonne utili
results4 <- results3[colnames(results3) != "rawp", ]

#****salva il file
write.csv(results4, "resTest.csv", row.names = F)
```

```
# creo dei grafici ordinati per pvalue in un unico pdf multipage


# seleziona i risultati con BH < 0.10
results5 <- na.omit(results4[results4$BH < 0.05,])
nrow(results5)


pdf("ResultsPlot.pdf", width = 8, heigh = 8)
for (i in 1:nrow(results5)) {
 print(paste("Plot ", i))


 GeneForPlot <- results5$Gene[i]
 exp <- logCPM2[GeneForPlot,] #valore di expression per un gene
 logR <- CNV2[GeneForPlot,] #valore di log ratio
 adjp <- results5$BH
 dataPlot <- data.frame(exp, logR)


 p <- ggplot(dataPlot, aes(logR, exp)) +
  geompoint() +
  xlab("Log Ratio") +
  ylab("Log CPM counts") +
  ggtitle(paste(GeneForPlot, "; adj p = ", formatC(adjp[i], digits = 2, format = "e"))) +
  themebw() +
  theme(panel.grid=elementblank(),
       plot.title = elementtext(hjust=0.5, size = 25),
       axis.title.x =elementtext(size=15),
       axis.title.y=elementtext(size=15),
       axis.text.x = elementtext(size=15),
       axis.text.y = elementtext(size = 15),
       axis.ticks.x = elementblank(),#remove tics
       legend.position = "none",
```

```
    plot.margin = unit(c(1,1,1,1), "cm"))
print(p)
}
dev.off()
```