Analisi Statistica dell'Espressione Genica Differenziale tra Pazienti Affetti da Autismo Macroencefalico e i loro Familiari non Affetti

Relazione di Biostatistica Computazionale e Bioinformatica

Camilla Cavaliere, Alessandro Clair, Giovanni Corradini

15 gennaio 2020

INTRODUZIONE

Il disturbo dello spettro autistico (ASD) è probabilmente causato da un'anomalia nello sviluppo cerebrale. Mutazioni rare (come quelle che causano l'alterazione delle connessioni a livello sinaptico), varianti genetiche comuni (che possono causare squilibri a livello neuronale inibitorio/eccitatorio) e certi fattori ambientali (mutazioni geniche indotte) contribuiscono al rischio di incorrere nella patologia, anche se in circa l'80% dei casi non sono note tuttora le cause che la originano.
Il nostro studio di riferimento, "Mariani J, Coppola G, Zhang P, Abyzov A et al. *FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders*", è incentrato sulla ricerca dei fattori genetici che contribuiscono maggiormente al rischio di ASD. Questo studio si è focalizzato su individui affetti da ASD che presentano macroncefalia, ovvero una dimensione della testa superiore al 90° percentile rispetto alla misura nella popolazione normale. Questa scelta è stata motivata dal fatto che la macroencefalia è uno dei fenotipi di autismo più coerentemente replicati e perchè è relazionata al gene FOXG1 (questo gene può causare un aumento della dimensione cranica se sovraespresso e viceversa se sottoespresso), ritenuto dagli sperimentatori un gene fondamentale per la spiegazione degli squilibri neuronali che differenziano sani e malati di ASD.
Al fine di analizzare l'espressione genica (e molto altro) gli sperimentatori hanno prelevato cellule staminali pluripotenti indotte (iPSCs), sia dai pazienti affetti da ASD scelti per lo studio che dai loro familiari sani (i controlli), successivamente coltivate in vitro fino ad ottenere cellule neuronali (organoidi) che hanno lo scopo di simulare il corredo genetico dei neuroni presenti durante la fase intermedia dello sviluppo embrionale nella corteccia cerebrale. Il nostro obiettivo è quello di trovare le principali differenze a livello di espressione genica tra sani e malati di ASD, analizzando statisticamente il trascrittoma ricavato dagli organoidi.

DISEGNO SPERIMENTALE

Originariamente i pazienti affetti da ASD candidati per questa analisi erano 9, ma sono stati ridotti a 4 poichè gli sperimentatori hanno voluto analizzare solo maschi, affetti da macroencefalia e con entrambi i genitori anch'essi affetti da macroencefalia: questa scelta è stata fatta sia per conferire più stabilità all'analisi, che per poter cogliere gli effetti del FOXG1 al "netto del suo effetto macroencefalia", ovvero per poter cogliere gli effetti del gene coinvolti nella deregolazione del GABA/glutammato e non quelli legati alla dimensione cranica. Infatti una delle analisi (analisi di rete genica) svolte nel nostro caso studio ha mostrato, come accennato nell'introduzione, come la differente espressione del FOXG1 influenzi la dimensione encefalica e come il FOXG1 sia coinvolto nello sviluppo del telencefalo (Hanashima et al., 2004; Martynoga et al., 2005; Xuan et al., 1995); inoltre, sempre secondo le analisi del nostro studio, è stato mostrato come il FOXG1 sia legato alla deregolazione dell'azione del GABA (principale neurotrasmettitore inibitorio, responsabile nell'attività di controllo dell'eccitabilità neuronale), una delle possibili cause dell'autismo ed "effetto" del gene che vorremmo cogliere.
Il gruppo di controllo comprende 4 padri, 3 madri e un fratello; in figura 1 vengono rappresentate le quattro famiglie: i quadrati rappresentano gli uomini, i cerchi rappresentano le donne mentre il quadrato pieno indica che l'individuo è malato.
Infine, mentre il trascrittoma delle iPSCs in stato di rosetta è stato misurato solo per alcuni soggetti, per i giorni 11 e 31 è stato analizzato il trascrittoma di tutti gli organoidi. Inoltre, probabilmente perchè gli

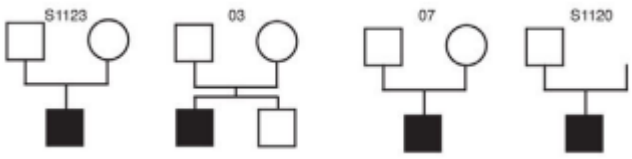


Figure 1: Famiglie Analizzate

sperimentatori volevano aumentare la precisione delle analisi, le misurazioni del trascrittoma sono state ripetute dalle 2 alle 4 volte per alcuni organoidi in alcuni giorni.

TECNOLOGIA E PIATTAFORMA

Per prima cosa sono stati presi dai pazienti dei campioni cellulari, utilizzati per produrre cellule staminali pluripotenti indotte (iPSCs), ottenute principalmente tramite metodi retrovirali. In seguito sono state create dalle 2 alle 3 linee di analisi per le singole iPSCs: alcune iPSCs non sono state prodotte al tempo 0 (in quanto probabilmente ritenute poco informative sull'espressione genica in quello stato) mentre sono state prodotte tutte, da 1 a 4 volte, per essere analizzate dopo 11 e dopo 31 giorni dalla loro "nascita".
Successivamente queste iPSCs sono state coltivate in vitro fino ad ottenere cellule neuronali (organoidi) che hanno lo scopo di simulare il corredo genetico dei neuroni presenti durante la fase intermedia dello sviluppo embrionale nella corteccia cerebrale. Inoltre sono state effettuate parecchie analisi all'interno del nostro articolo di riferimento che affermano che tutti gli organoidi preparati dagli sperimentatori simulano correttamente le caratteristiche genetiche dei neuroni del campione di riferimento.
In seguito, dopo aver amplificato il DNA tramite la PCR, è stata effettuata l'analisi di dati di RNA-seq contenuto nelle cellule dei vari organoidi (ai giorni 0, 11 e 31). In questa analisi è stato usato Tophat (Langmead et al., 2009) per mappare le reads dal genoma umano (hg19) all'annotazione transcrittomica GencodeV7 (Harrow et al., 2006); poi le reads (allineate) in formato BAM sono state convertite in formato SAM, grazie a SAMtools (Li et al., 2009).

PREPROCESSAMENTO DEI DATI

I dati, già appartenenti alla classe SummarizedExperiment, sono stati scaricati utilizzando la libreria recount di Bioconductor. La matrice dei conteggi delle reads contiene 58037 geni misurati su 48 campioni.
Per prima cosa abbiamo eliminato tutte le variabili riferite ai campioni contenute nei "colData" del summarized experiment, eccetto le "characteristics", in quanto ritenute superflue per le nostre analisi. Delle variabili ricavate dalle "characteristics" (mediante la funzione "geocharacteristics" di "recount") abbiamo tenuto solo quelle ritenute utili per le nostre analisi, ovvero: il genere, l'Id (fattore con 12 modalità), il giorno della rilevazione (0-11-31) e lo stato di salute (sano/malato). Successivamente abbiamo rimosso i geni (32511) che non hanno il nome in formato "SYMBOL": questo è stato fatto perché ritenuti geni non funzionali e perché, data l'assenza del nome, non sarebbero stati utili per l'interpretazione finale dei risultati.
In seguito abbiamo eliminato dal dataset le osservazioni del giorno 0 poiché sono solo 6 (non coinvolgono tutti i campioni) e poiché abbiamo dedotto dal nostro caso studio che un embrione a quello stadio non fosse sufficientemente maturo da poter evidenziare delle differenze interessanti tra sani e malati. Inoltre abbiamo deciso di non includere nelle analisi la persona con l'Id '03-04' (unico fratello), per poter confrontare direttamente figlio malato con genitori sani. Dato che nella matrice dei conteggi sono presenti, per lo stesso soggetto, più misurazioni relative allo stesso giorno, abbiamo deciso di accorparle utilizzando la media troncata all'intero più vicino. Come ultima operazione sono stati scartati i geni poco espressi (1867), ovvero i geni il cui numero medio di reads sequenziate per campione è minore di 10: questa operazione di filtraggio avrebbe in ogni caso filtrato quasi tutti i geni privi di nome in formato "SYMBOL", già filtrati in precedenza, in quanto non espressi (confermando l'ipotesi che l'assenza del nome in formato "SYMBOL" di un gene indicasse la sua non funzionalità). Il dataset dopo le operazioni di controllo qualità e filtraggio conta 23659 geni osservati ai giorni 11 e 31 per gli 11 campioni (4 figli malati, 4 padri e 3 madri sani).

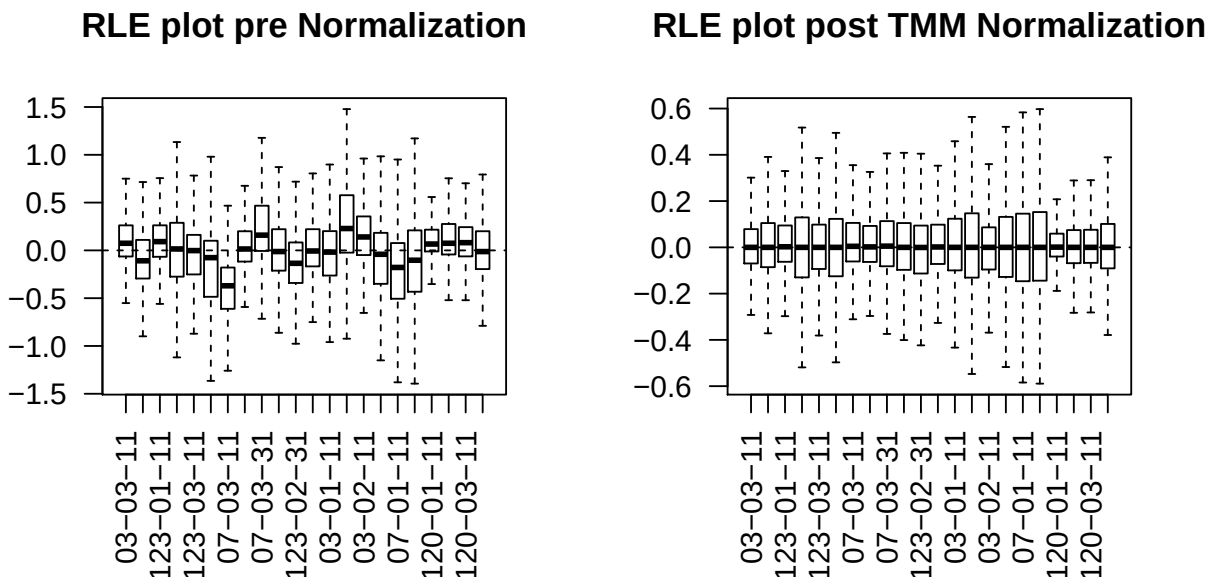


Figure 2: Normalizzazione TMM

ANALISI ESPLORATIVA E NORMALIZZAZIONE

il numero totale di reads mappate per ciascun campione sembra in generale non presentare differenze sostanziali tra campioni diversi. Un grafico che può essere d'aiuto nel verificare la presenza di asimmetria nel numero di reads tra i vari campioni è l'RLEplot della distribuzione dei conteggi (figura 2) :

Dal primo grafico notiamo che non sembrano esserci differenze così marcate tra i conteggi 'grezzi' dei vari campioni. In ogni caso procediamo alla normalizzazione per ridurre il più possibile gli errori sistematici, così che le differenze tra misurazioni rappresentino esclusivamente le differenze biologiche tra i campioni. Per confrontare le diverse normalizzazioni abbiamo osservato le distribuzioni degli RLE plots e i grafici delle componenti principali. In base ai grafici delle diverse normalizzazioni provate (UQ, TMM, FQ, RLE), è stata scelta la normalizzazione TMM, che è sembrata la più ragionevole e quella che ha alterato il meno possibile la struttura dei dati.

Tramite l'analisi delle componenti principali abbiamo visto che i pazienti sani da quelli malati non sono quasi per nulla discriminati, nemmeno dopo la normalizzazione (anche se la varianza spiegata dalla prima componente prinicpale è aumentata), sottolineando come l'analisi delle componenti principali non riesca a far emergere marcatamente la differenza tra i due gruppi. Ciò si intuisce anche dal fatto che le prime componenti principali (soprattutto pre normalizzazione) spiegano poco della varianza totale delle variabili originali. Quindi la PCA sui dati grezzi, soprattutto per la prima componente principale, dispone i campioni in base al numero di reads sequenziate senza considerare i fattori biologici a cui siamo interessati. Per la prima CP perciò stiamo valutando solo la variabilità tecnica e non quella biologica.

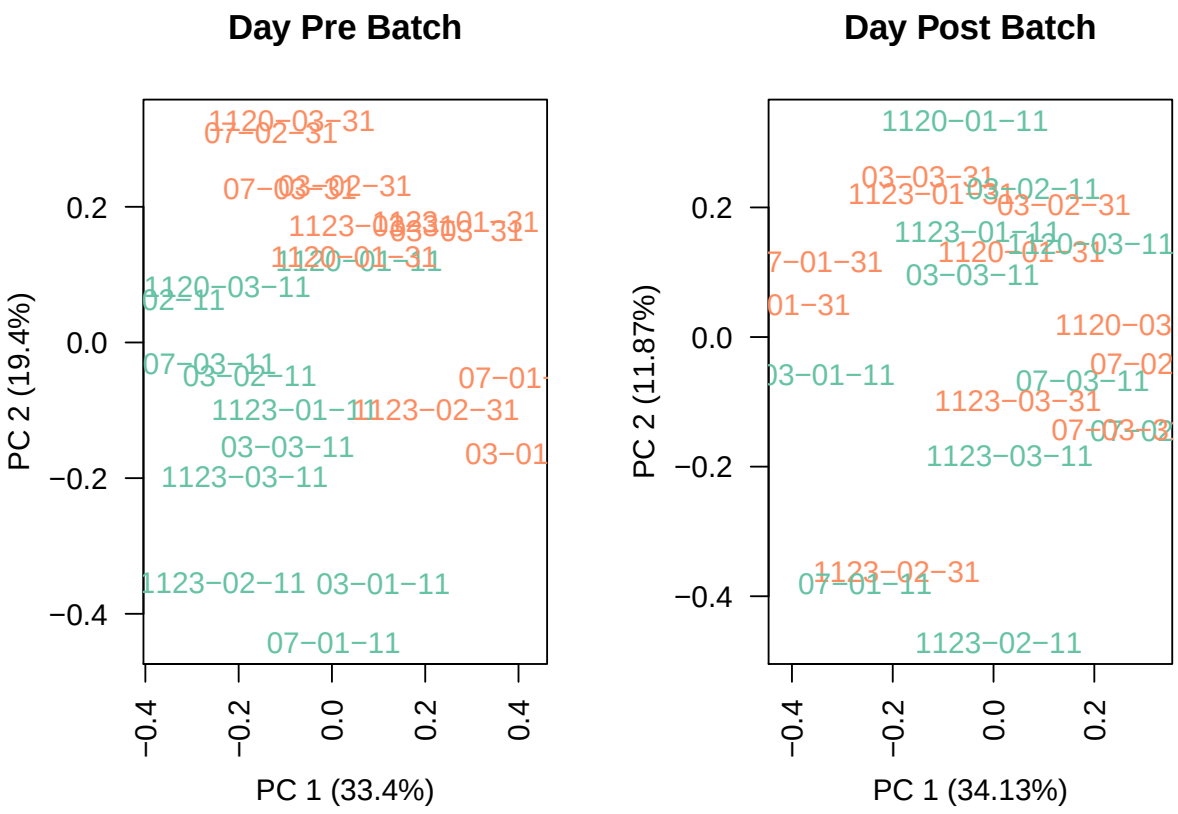


Figure 3: Batch Effect of TMM Normalized Counts

BATCH EFFECT

Molto più interessante sembra essere la PCA relativa alle variabili in cui viene sottolineata la dipendenza dal giorno della misurazione piuttosto che lo stato di salute. Infatti sembra esserci una distinzione marcata tra i due gruppi (giorno 11 in verde e giorno 31 in rosso). Questo si può attribuire al fatto che le cellule staminali risultino chiaramente più evolute nel giorno 31 rispetto al giorno 11. L' effetto del giorno sembra incidere molto di più dello stato di salute nella variabilità dei dati; per tenere conto di ciò si potrebbe includere il giorno come covariata del modello nella fase di inferenza, però abbiamo ritenuto che avrebbe potuto sovrastare l'effetto della 'salute'. Perciò abbiamo scelto di trattare il giorno come effetto di batch e di rimuovere dai dati la variabilità dovuta ad esso tramite il modello *Combat*.

I PCA plots precedenti (Figura 3) mostrano quanto l'effetto del giorno incidesse sulla variabilità dei dati e quanto sia stato ridotto (praticamente eliminato) dopo averlo trattato come effetto di batch.

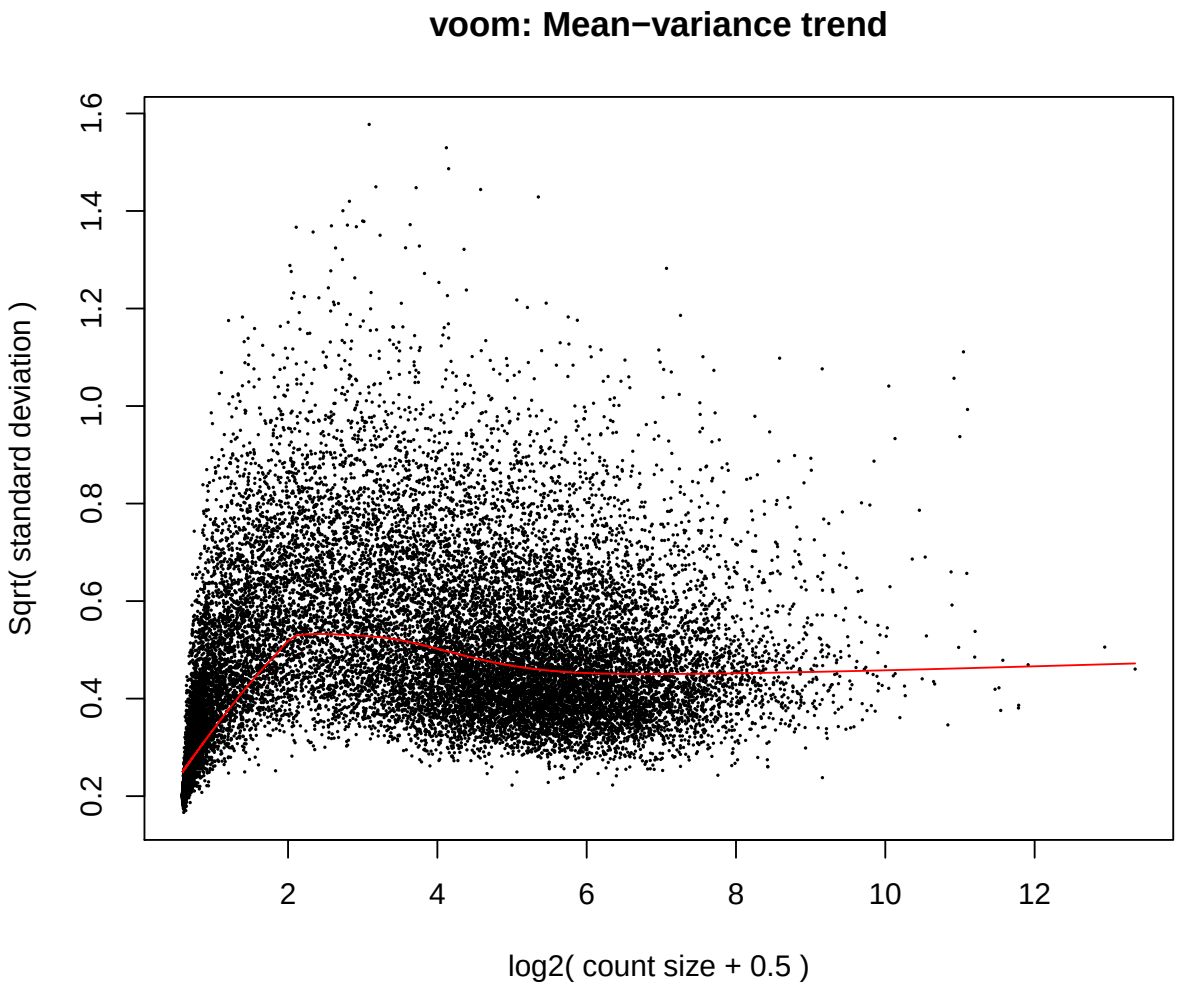


Figure 4: voom: Mean-variance trend

INFERENZA

Per prima cosa viene creato un oggetto DGEList a partire dai dati 'aggiustati' per l'effetto di batch. In seguito, nella costruzione della matrice del disegno, sono state scelte come esplicative: la famiglia di appartenenza del paziente (fattore con 4 modalità), il genere e lo stato di salute (sano/malato), quest'ultimo di importanza primaria nelle nostre analisi.
Per la fase di identificazione dei geni differenzialmente espressi abbiamo optato per il modello lineare coi pesi del pacchetto *limma* in quanto dovrebbe fornire risultati più stabili con una bassa numerosità campionaria (come nel nostro caso). Perciò abbiamo utilizzato la funzione *voom* per trasformare i dati in logCPM, stimare la relazione tra media e varianza (figura 4), e calcolare i pesi per ogni osservazione. I pesi stimati vengono poi incorporati nei modelli lineari, che vengono fittati per ogni gene. Infine si calcolano le stime Bayesiane empiriche dei parametri per "schiacciare" le stime verso un valore medio, e per calcolare la t moderata.

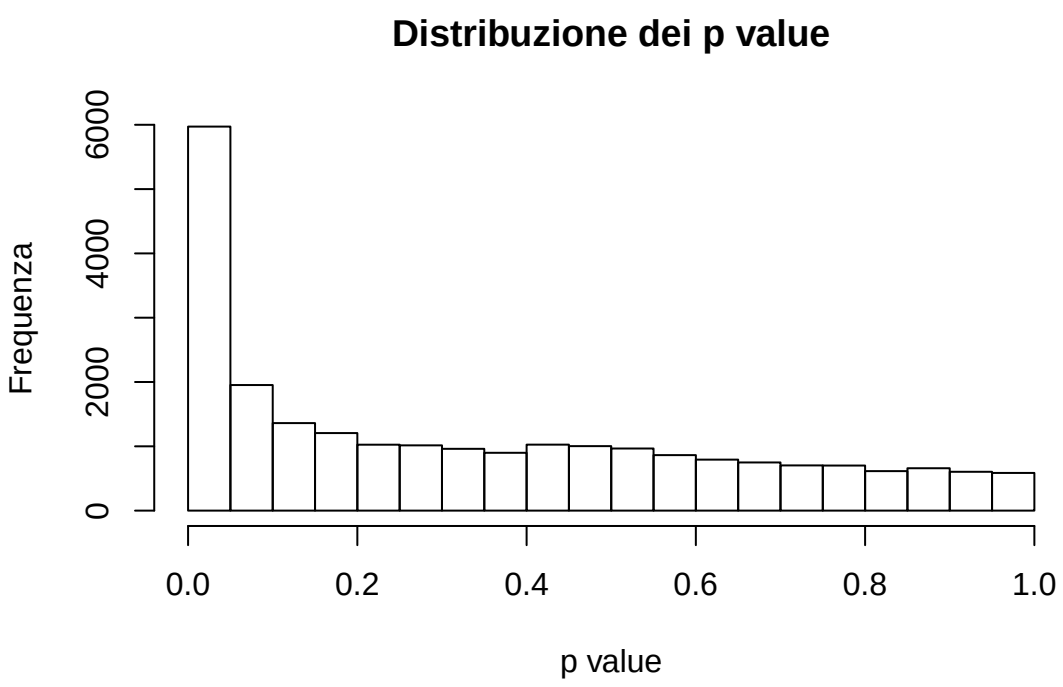


Figure 5: Distribuzione dei p value

L'istogramma in figura 5 mostra la distribuzione dei p value relativi ai coefficienti della variabile *salute* tra i vari geni; questa sembra essere soddisfacente in quanto vediamo una differenza molto marcata tra geni differenzialmente espressi e non differenzialmente espressi.

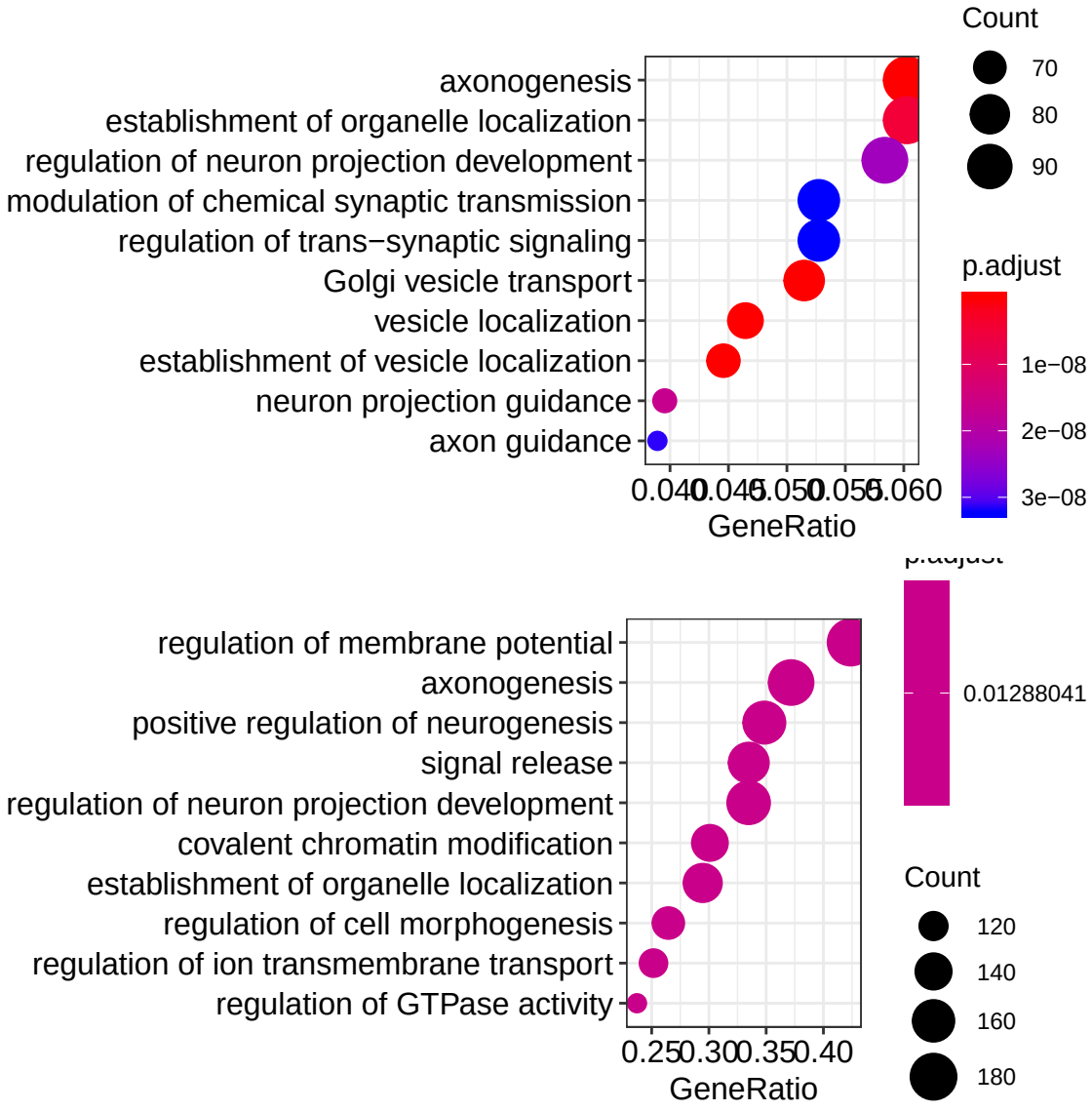
```
##
## FALSE  TRUE
## 21672  1987
```

```
##                      symbol     logFC        t    adj.P.Val
## ENSG00000176165.10   FOXG1 3.2072180 8.895904 0.0009077605
## ENSG00000162873.14 KLHDC8A 2.3857210 8.065976 0.0017071494
## ENSG00000130787.13   HIP1R 0.5906675 7.923954 0.0017071494
## ENSG00000198914.3   POU3F3 2.4666919 7.690920 0.0019168181
## ENSG00000012232.8    EXTL3 0.5138298 7.588020 0.0019168181
```

Il modello ha identificato 1987 geni differenzialmente espressi (ad un livello di significatività del 5%) su un totale di 23659 geni. La tabella mostra i 5 geni più significativamente differenzialmente espressi, con i relativi: nomi in formato 'SYMBOL', valori del log fold change, valori della t moderata e valori dei p value (aggiustati secondo la procedura di Benjamini Hochberg).

INTERPRETAZIONE DEI RISULTATI

Soffermandoci per ora sui primi 5 geni più significativi, abbiamo visto come il gene FOXG1 fosse il primo, anche se, come detto detto in precedenza, l'effetto FOXG1 collegato alla dimensione cerebrale è stato rimosso scegliendo tutti pazienti affetti da macroencefalia. L'elevata significatività di questo gene quindi è in linea con i risultati del nostro articolo di riferimento, nel quale FOXG1 è stato valutato come gene attivatore della deregolazione dei neuroni GABA/glutammato, ritenuti importanti nella differenziazione tra sani e malati di autismo. Inoltre i restanti 4 geni più significativi (KLHDC8A,HIP1R,POU3F3 e EXTL3) sono geni che codificano per proteine che sono coinvolte nello sviluppo cerebrale o che possono essere associate a disturbi neurali; anche i geni un po' più in basso 'in classifica' presentano, in generale, una funzione simile.



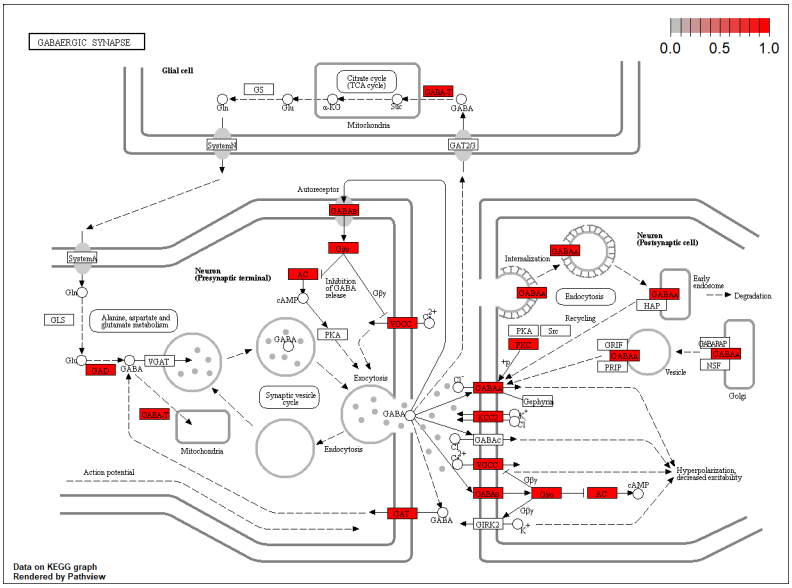


Figure 6: GABAergic Synapse Pathway

I dotplots nella pagina precedente, ottenuti tramite il test esatto di Fisher e tramite l'approccio GSEA, entrambi riferiti alla Gene Ontology, indicano le funzioni di attivazione genica risultate più significative nel nostro modello. Questi grafici indicano che le funzioni dei geni che differenziano principalmente sani e malati di ASD riguardano la costruzione, la regolazione e la modulazione degli assoni, delle sinapsi e degli organelli deputati alle principali funzioni neuronali: per comprendere questi risultati più nello specifico sarebbero stati necessari ulteriori approfondimenti di carattere prettamente biologico.
Infine abbiamo testato il pathway KEGG "GABAergic Synapse Pathway", in quanto abbiamo visto in letteratura che vi è un legame tra FOXG1 e l'azione del GABA nella regolazione sinaptica, che come detto in precedenza è una delle possibili cause dell'ASD macroncefalico; inoltre è stato scelto questo pathway in quanto è risultato essere uno dei più significativi (testato sia tramite il test esatto di Fisher che tramite l'approccio GSEA) tra quelli di KEGG. La maggior parte dei comunicatori cellulari presenti in questo pathway (figura 6) è rossa, indice del fatto che la maggior parte dei geni coinvolti in questi comunicatori sono significativamente sovraespressi ad un livello del 5%.
Concludendo possiamo affermare che le nostre analisi sembrano essere riuscite a differenziare (a livello genico) adeguatamente sani e malati: la evidente sovraespressione del gene FOXG1, la sovraespressione dei vari geni che possono codificare varie proteine coinvolte nello sviluppo cerebrale o che possono essere associate a disturbi neurali e la sovraespressione della pathway della sintesi GABA sembrano indicare una relazione tra la differenza nello sviluppo cerebrale a livello embrionale tra sani e malati di ASD e le deregolazioni di neurotrasmettitori legati all'eccitabilità neuronale come il GABA.