# Elaborazione dati e analisi statistiche - HDLPFC

## Elaborazione dei dati grezzi Visium

I file FASTQ grezzi e le immagini istologiche sono stati elaborati con il software Space Ranger 1.0.0 rispetto al genoma di riferimento Cell Ranger hg38.

## Elaborazione e segmentazione delle immagini istologiche

Le immagini istologiche sono state elaborate e i nuclei sono stati segmentati utilizzando la "Color-Based Segmentation K-Means Clustering" in MATLAB vR2019a.

La funzione MATLAB rgb2lab è stata utilizzata per convertire l'immagine dallo spazio di colore RGB allo spazio di colore CIELAB, chiamato anche spazio di colore L\*a\*b:

* L - livello di luminosità misura la luminosità dal nero al bianco,
* a - livello di cromaticità misura il colore lungo l'asse rosso-verde,
* b - livello di cromaticità misura il colore lungo l'asse blu-giallo).

Lo spazio di colore CIELAB quantifica le differenze visive causate dai diversi colori dell'immagine. Lo spazio di colore a\*b viene estratto dall'immagine convertita in L\*a\*b e viene fornito alla funzione di clustering Kmeans imsegkmeans insieme al numero di colori che l'utente identifica visivamente nell'immagine. La funzione imsegkmeans produce una maschera binaria per ogni colore identificato. Poiché i nuclei nelle immagini istologiche hanno un colore brillante che può essere facilmente differenziato dallo sfondo, la maschera binaria generata per il colore dei nuclei viene utilizzata per la segmentazione dei nuclei.

La maschera binaria segmentata è stata utilizzata per stimare il numero di nuclei in ciascun punto. Per ogni immagine istologica è disponibile un file JSON che descrive alcune proprietà dell'immagine, tra cui il diametro della macchia in pixel a piena risoluzione. Inoltre, per ogni immagine è disponibile un file di testo in formato tabellare che include una riga per ogni spot con un codice a barre identificativo, una riga, una colonna e le coordinate dei pixel per il centro dello spot sull'immagine a piena risoluzione. Utilizzando queste informazioni, è stato implementato il seguente protocollo:

* Per ogni punto, tutti i pixel della maschera binaria sono stati impostati a zero, tranne quelli che si trovavano nel raggio del centro del punto. La maschera binaria risultante è stata quindi etichettata con un numero intero unico per ogni cluster contiguo di pixel. Il massimo di questa maschera etichettata è stato memorizzato come stima del numero di nuclei all'interno di quello spot.

## Elaborazione dei dati a livello di spot

1. I file Visium grezzi per ciascun campione sono stati letti in R in una struttura personalizzata utilizzando il pacchetto SummarizedExperiment R per mantenerli abbinati alle immagini istologiche a bassa risoluzione ai fini della visualizzazione.
2. Sono stati poi combinati in un singolo oggetto SingleCellExperiment v1.8.0 (sce) per eseguire analisi utilizzando i dati di espressione genica di tutti i campioni.
3. È stato aggiunto all'oggetto sce informazioni quali il numero di cellule stimate, la somma di UMI per spot, il numero di geni espressi per spot e i risultati del clustering basato su grafici (calcolati per campione) forniti dal software 10x Genomics Space Ranger.
4. Sono state valutate le metriche di qualità per spot utilizzando la funzione perCellQCMetrics del pacchetto scran v1.14.3 R Bioconductor e non è stato eliminato nessuno spot dato il modello spaziale che presentavano. Sono state utilizzate le funzioni *scran* quickCluster, bloccando le sei coppie di repliche spazialmente adiacenti, computeSumFactors e logNormCounts di scater per calcolare i conteggi log-normalizzati dell'espressione genica a livello di spot.
5. Modellando l'espressione media e la varianza dei geni con la funzione modelGeneVar di scran e bloccando nuovamente le sei coppie di repliche spazialmente adiacenti, seguite da getTopHVGs, è stato identificato il 10% di geni altamente variabili (HVGs): 1.942 geni.
6. Utilizzando questo sottoinsieme di HVG, sono state calcolate le componenti principali (PC) con runPCA di scater v1.14.3 per produrre 50 componenti. Utilizzando queste 50 componenti principali, sono stati calcolati i metodi di riduzione delle dimensioni tSNE e UMAP utilizzando runTSNE (perplessità 5, 20, 50, 80) e runUMAP (15 vicini) da scater.
7. Con i primi 50 PC è stata eseguita una clusterizzazione basata su un grafo su tutti i campioni utilizzando 50 vicini più prossimi con buildSNGraph da scran e il metodo Walktrap implementato da igraph, ottenendo 28 cluster (da snn\_k50\_k4 a snn\_k50\_k28).
8. è stato tagliato ulteriormente il grafico per produrre cluster da 4 a 28 con incrementi di 1. È stato usato spatialLIBD v0.99.0 per assegnare i cluster basati sul grafico da 10x Genomic agli strati anatomici più vicini per ciascun campione.

Tutte queste informazioni sono state combinate e visualizzate attraverso un'applicazione web all'indirizzo http://spatial.libd.org/spatialLIBD in modo tale che è possibile visualizzare l'espressione di un determinato gene o un determinato insieme di risultati di clustering, su tutti i campioni o su ogni singolo campione. Per ogni campione scelto, spatialLIBD consente agli utenti di visualizzare l'espressione genica e i risultati dei cluster selezionati sia nel contesto dell'istologia spaziale sia in base ai risultati della riduzione delle dimensioni (PCA, tSNE, UMAP) utilizzando plotly.

Utilizzando questa applicazione web per visualizzare la citoarchitettura in combinazione con un metodo di riduzione della dimensionalità (t-Distributed Stochastic Neighbor Embedding (t-SNE)) e i pattern di espressione di MBP e PCP4 (noti geni marker di WM e L5) un singolo sperimentatore ha assegnato manualmente ogni spot a uno strato corticale per ciascun campione per tutti i 47.681 spot, tranne 352, su tutti i campioni. Questi 352 spot erano localizzati su piccoli frammenti di tessuto danneggiato scollegati dalla sezione principale del tessuto. Queste annotazioni di strato supervisionate sono state aggiunte all’oggetto sce e la versione finale è disponibile per il download attraverso la funzione fetch\_data in spatialLIBD.

## Elaborazione dei dati a livello di strato

1. Per il soggetto con ID cerebrale Br5595, privo di L1 e di una chiara citoarchitettura per L2 e L3, sono stati rietichettati tutti gli spot ambigui "L2/L3" come L3 ed eliminati i 352 spot non assegnati.
2. Sono stati pseudo-bulkati gli spot in dati a livello di strato, sommando i conteggi grezzi di espressione genica in tutti gli spot di un dato campione e di un dato strato e ripetendo questa procedura per ogni combinazione di gene, campione e strato. Il risultato è stato 47.329 geni quantificati in 76 combinazioni strato-campione (7 \* 12 = 84, perché non tutti gli strati sono stati osservati chiaramente in ogni campione, dato che Br5595 non presentava L1 o L2 distinti in tutte e quattro le repliche spaziali). Ne è risultato un altro oggetto SingleCellExperiment chiamato sce\_layer.
3. Sono state utilizzate le funzioni librarySizeFactors e logNormCounts da scater per calcolare i valori di espressione genica normalizzati a livello di strato.
4. Sono stati eliminati tutti i geni mitocondriali e mantenuti i geni espressi per almeno il 5% (4 / 76 combinazioni strato-campione) e che avevano un conteggio medio superiore a 0,5 come calcolato da calculateAverage di scater, ottenendo un set finale di 22.331 geni.
5. Sono stati identificati 1.280 top HVG a livello di strato e calcolato 20 PC (componenti principali) che sono stati poi utilizzati nei calcoli tSNE (perplessità = 5, 15 e 20) e UMAP (15 vicini) come nell’elaborazione dei dati a livello di spot.
6. Sono stati poi raggruppati i dati a livello di strato utilizzando diversi approcci basati su grafi e k-means. Questi sono i dati di sce\_layer, disponibili per il download attraverso la funzione fetch\_data in spatialLIBD

## Analisi dell'arricchimento del neuropilo

1. È stata eseguita l'analisi dell'espressione differenziale a livello di spot nei dati Visium, confrontando i 4.855 spot con 0 corpi cellulari con gli altri 42.474 spot contenenti almeno un corpo cellulare, aggiustando per gli effetti fissi di strato e replica spaziale.
2. Sono state scaricate e utilizzate le statistiche di espressione differenziale da RNA-seq di sinaptosomi arricchiti di vGLUT1+ nel cervello di topo da Hafner et al, e allineati questi dati a livello di geni utilizzando gli ID di entrez omologhi tra topo e uomo (tramite http://www.informatics.jax.org/downloads/ reports/HMD\_HumanPhenotype.rpt).
3. Sono stati confrontati gli effetti degli spot contenenti 0 cellule nei dati con quelli delle cellule arricchite di vGLUT1+ di Hafner et al, sia sull'intero trascritto omologo sia all'interno dei geni significativi nel set di dati di Hafner con FDR < 0,05.

## Modellazione genica a livello di strato

I dati a livello di strato sono stati adattati a tre tipi di modelli:

1. **ANOVA**: per questo modello è stato testato per ogni gene se i conteggi di espressione genica normalizzati in log sono variabili tra gli strati calcolando la statistica F. Sono state utilizzate le funzioni lmFit ed eBayes da *limma* dopo aver bloccato le sei coppie di repliche spazialmente adiacenti e tenendo conto della correlazione calcolata da duplicateCorrelation.

2. **Enrichment**: Utilizzando le stesse funzioni lmFit ed eBayes e tenendo conto della stessa struttura di correlazione, sono state calcolate le statistiche t confrontando uno strato con gli altri sei utilizzando i dati a livello di strato. Ne sono risultate sette serie di statistiche t (una per strato) con valori P a due facce. Sono stati considerati i geni con statistiche t positive (espressi in misura maggiore in uno strato rispetto agli altri), poiché si tratta di geni arricchiti anziché di geni impoveriti.

3. **Pairwise**: Utilizzando le stesse funzioni lmFit ed eBayes e tenendo conto della stessa struttura di correlazione, oltre a usare contrasts.fit di limma, sono state calcolate le statistiche t per ogni coppia di strati, ottenendo statistiche t con valori P a due facce.

## Registrazione spaziale snRNA-seq

1. Per ogni set di dati snRNA-seq, sono stati utilizzati i dati pubblici ed elaborati di conteggio degli indici molecolari univoci (UMI) per ogni gene e nucleo e fornite annotazioni di cluster/sottotipi di cellule.
2. All'interno di ciascun set di dati, è stato eseguito uno 'pseudo-bulking' degli UMI a livello di nucleo in conteggi normalizzati logtrasformati specifici per tipo di cellula per ciascun soggetto unico.
3. Sono state calcolate le statistiche di "enrichment" del tipo di cellula per ciascun gene e tipo di cellula fornito dal set di dati all'interno dei loro profili pseudo-bulk eseguendo una modellazione lineare a effetti misti che confronta ciascun tipo di cellula con tutti gli altri tipi di cellule, trattando il donatore come un'intercetta casuale e aggiustando per le covariate specifiche dello studio. (strategia analoga alle statistiche di "enrichment" degli strati descritte per i dati Visium nella modellazione genica a livello di strato).
4. Sono stati calcolati i coefficienti di correlazione di Pearson tra le statistiche di "enrichment" a livello di strato (ottenute precedentemente per i dati Visium) e le statistiche di "enrichment" specifiche del tipo di cellula snRNAseq tra i 700 geni più arricchiti a livello di strato (combinando i 100 geni più significativi per ciascuno dei sei strati e WM nei dati Visium) espressi in ciascun set di dati snRNA-seq.

## Analisi del clustering arricchito di strati

È stata eseguita la selezione delle caratteristiche in due modi per identificare i modelli laminari e non laminari nei dati.

1. Il primo metodo di selezione delle caratteristiche utilizzato è stato SpatialDE per identificare i geni che presentano modelli di espressione spazialmente variabili (SVG).

SpatialDE (in Python versione 3.8.0) è stato eseguito individualmente su ciascuno dei 12 campioni ed ha restituito un insieme di SVG statisticamente significativi per ogni campione (false discovery rate < 0,05). È stata inclusa un'ulteriore fase di filtraggio per rimuovere i geni poco espressi (meno di 1.000 UMI totali sommati in tutti gli spot per campione), oltre a rimuovere i geni mitocondriali. In questo modo sono rimasti tra i 521 e i 2.217 geni per campione, per un totale di 2.775 geni unici tra i campioni; per confronto, abbiamo anche eseguito metodi di clustering utilizzando questo elenco raggruppato.

1. Il secondo metodo di selezione delle caratteristiche ha utilizzato il pacchetto scran di R Bioconductor per identificare i geni altamente variabili (non spaziali) in tutti i campioni combinati, ed ha identificato 1.942 HVG. {A causa della lentezza dei tempi di esecuzione, non è stato possibile eseguire SpatialDE sugli spot in pool di tutti i campioni combinati}.

Nell'approccio "non supervisionato" per definire sottogruppi di spot con profili di espressione simili e in modo completamente guidato dai dati, sono state considerate le possibili combinazioni di

1. due tipi di metodi per la riduzione della dimensionalità (le prime 50 componenti principali (PC) con il pacchetto *BiocSingular* v1.2.2 Bioconductor e le prime 10 componenti UMAP con il pacchetto *uwot* v0.1.5 calcolate sulle prime 50 PC),
2. gli insiemi di geni definiti dopo aver applicato la selezione delle (geni SpatialDE per ciascun campione, liste di geni SpatialDE in pool per tutti i 12 campioni e HVG)
3. l'inclusione (o meno) delle due coordinate spaziali (coordinate x e y) di ciascun punto come caratteristiche aggiuntive per il clustering.
4. Per l'algoritmo di clustering, è stato costruito un grafo dei vicini più prossimi con il pacchetto Bioconductor scran e poi applicato il metodo Walktrap del pacchetto igraph di R per ottenere le etichette di cluster previste. Sono state impostate tutte le implementazioni del clustering in modo che restituissero otto cluster finali (cioè uno in più rispetto ai sei strati della DLPFC più la materia bianca

È stato anche implementato un approccio "semi-supervisionato", in cui sono stati utilizzati gli insiemi di geni arricchiti di strati identificati utilizzando i modelli di "enrichment" DE descritti in precedenza e un approccio "a marcatori" utilizzando geni marcatori noti di Zeng et al. Per valutare le prestazioni degli approcci di clustering è stato utilizzato l'indice Rand aggiustato (ARI), che misura la somiglianza tra le etichette dei cluster previsti e le etichette dei cluster "gold standard" (i livelli annotati manualmente). Valori di ARI più elevati corrispondono a migliori prestazioni di clustering, con un valore massimo di 1 che indica un perfetto accordo di clustering. Per valutare il miglioramento dell'ARI quando si includono le coordinate spaziali nei metodi di clustering, è stato applicato un modello lineare ai punteggi ARI, confrontando questi metodi con quelli privi di coordinate spaziali per tutti i metodi e i campioni e registrando il valore di p.