Esercitazione - Interpretare i dati di espressione genica a partire da un articolo scientifico

L'articolo di nostro interesse si intitola "Comprehensive transcriptional landscape of aging mouse liver" ed è stato pubblicato da White e colleghi nel 2015 sulla rivista BMC Genomics.

Task 1

Aprite l'articolo (seguendo il link fornito su Moodle) e date uno sguardo all'abstract per comprendere quale è l'argomento trattato. In particolare poi scendete alla sezione relativa ai materiali e metodi per verificare quale è il disegno sperimentale, in particolare per quanto riguarda il numero di repliche biologiche per ciascuna condizione (cosa che dovrebbe risultare evidente anche dalla Figura 1D).

Task 2

Recuperate dal Moodle il file **DEG_table_mouse_aging_liver.xlsx**, che contiene tutte le informazioni relative ai livelli di espressione dei geni differenzialmente espressi (DEG) identificati nell'articolo. In particolare sono riportati i livelli di espressione nei 6 campioni biologici analizzati (3 repliche per i fegati di topi giovani + 3 repliche per i fegati di topi anziani).

Chiedetevi:

-A che cosa corrispondono i livelli di espressione mostrati nel file?

-Questi valori permettono una comparazione dei livelli tra i diversi campioni? Se sì, con quali limitazioni?

-Quali sono stati i criteri utilizzati dagli autori per selezionare i DEG?

Suggerimento: potete ordinare le righe nel file excel sulla base delle diverse colonne, selezionando quelle che a vostro avviso sono quelle che avrebbe senso considerare per effettuare la scelta dei DEG.

Task 3

Che tipo di rappresentazione grafica potrebbe essere utile per verificare se le 3 repliche biologiche dei due campioni sono coerenti tra loro?

Scaricate da Moodle i seguenti files:

-aging_mouse_PCA_data.txt

-aging_mouse_PCA_data_logscale.txt

Collegatevi dunque al link "PCA online" e da qui cliccate su "data import". Da qui, tramite "Upload data", oppure tramite "Paste data" potrete procedere con l'analisi degli stessi dati che avete già visto nel file excel, che in questo caso sono stati formattati in modo da renderli compatibili con lo strumento che utilizzeremo. Partite con l'analisi dei dati "grezzi" contenuti nel primo file e verificate i risultati, andando a cliccare prima su "data pre-processing" e poi su "PCA". E' importante cliccare su "change plot labels" e poi su "show sample IDs" per visualizzare le corrispondenze tra punti e campioni.

Task 4

Chiedetevi a questo punto, ripensando a quanto detto a lezione, se possa essere opportuno adoperare una trasformazione dei dati. Aprite dunque il secondo file, che riporta i livelli di espressione dei DEG trasformati in scala Log10. Il logaritmo è stato calcolato sui valori di espressione a cui è stato aggiunto un +1, di modo da evitare il calcolo di logaritmi per valori inferiori ad 1.

Da "Data import", recuperate i dati dal file

-aging_mouse_PCA_data_logscale.txt

Utilizzate lo strumento esattamente come fatto in precedenza ed osservate I risultati. Notate dei cambiamenti significativi? Secondo voi, come è interpretabile questo risultato? Riflettete nuovamente sulla necessità (o meno) di trasformare i valori di espressione in scala logaritmica per questo particolare tipo di grafico.

Task 5

Ritorniamo dunque sull'articolo e visualizziamo la figura 1D. Vi sembra un'immagine informativa, così come è proposta? C'è qualcosa che manca per renderla comprensibile? Su quali aspetti si potrebbe, secondo voi, fare un'azione migliorativa, per quanto riguarda la visualizzazione grafica?

Proveremo a generare ora una heat map simile a quella presente in questa figura, magari tentando di migliorarla.

Scaricate i seguenti files:

- aging_mouse_original_values.txt

- aging_mouse_transformed_values.txt

- aging_mouse_top100_transformed_values.txt

Collegatevi a questo punto al link su Moodle "Heat map online".

Task 6

Caricate dunque il primo file che contiene i valori normalizzati di espressione nei 6 campioni (Upload file -> Browse -> selezionate il file con i valori originali). A che cosa corrispondono le righe e le colonne in questo grafico? Notate che la colorazione non rispecchia i livelli di espressione assoluti, ma è legata ad uno Z-score, come indicato dall'opzione "scale type -> row". Questo sta a significare che viene assegnato di default un colore giallo al campione (tra i 6) dove il valore di espressione è il più alto e blu a quello dove il valore di espressione è più basso, indipendentemente dal livello di espressione assoluto.

Cambiate questa impostazione selezionando "none" dal menu a tendina. Come è cambiato il grafico? Vi sembra informativo? Cosa potremmo fare per migliorare la visualizzazione?

Task 7

Prima di cambiare il file analizzato, modifichiamo alcuni altri parametri di visualizzazione, che ci serviranno tra poco.

-Notate che il clustering è stato effettuato solo per i geni, ma non per i campioni. Questa opzione può essere modificata andando sulle voci “apply clustering to” e “show dendrogram” ed aggiungendo alla voce “rows” anche “columns”.

-E’ possibile anche cambiare i colori da assegnare ai geni espressi a livelli bassi, intermedi ed elevati. Potete scegliere diverse combinazioni da “colour scheme” (la scelta resta comunque arbitraria). E’ possibile impostare una scala di colori a piacimento tramite “custom”.

Task 8

Caricate a questo punto, dopo aver cliccato su “clear”, il file **aging_mouse_transformed_values.txt**

Si tratta dello stesso dataset appena analizzato, che però in questo caso riporta i valori di espressione trasformati in scala log10, analogamente a quanto visto sopra per la PCA.

Come è cambiato il grafico? Vi sembra che sia più informativo?

Ricordate che, per default, la colorazione non è basata si livelli di espressione assoluti dei singoli campioni, ma su degli z-score e questa opzione può essere modificata.

Provate ad utilizzare diversi schemi di colorazione per ottimizzare la visualizzazione.

Task 9

I geni mostrati sono comunque molti, troppi per permetterci di visualizzare i nomi dei singoli geni a fianco di ciascuna riga della heat map.

Provate dunque a caricare il file **aging_mouse_top100_transformed_values.txt**

Notate che in ogni caso non è ancora possibile assegnare ad ogni riga il nome del gene corrispondente.

E’ tuttavia possibile modificare le dimensioni del grafico, cliccando su “advanced options” ed aumentando le opzioni “Plot Width (pixels)” e soprattutto “Plot Height (pixels)”, fino a quando non trovate l’opzione migliore per visualizzare i nomi dei singoli DEG.

Task 10

Notate in particolare un gruppo di geni che mostrano una sovraespressione nel fegato dei topi anziani? Si tratta di geni che appartengono evidentemente ad una stessa famiglia genica, quelle delle MUP (Major Urinary proteins).

Se volessimo capire qualcosa di più riguardo a questa famiglia genica potremmo collegarci ad un genome browser, ed in particolare ad Ensembl, seguendo questo link:
https://uswest.ensembl.org/Mus_musculus/Info/Index

Proviamo ad effettuare una ricerca per quanto riguarda il gene Mup10 ad esempio. Dalla scheda relativa al gene cliccate su “Go to region in detail” e provate ad interpretare ciò che state vedendo.

Può essere anche molto interessante cliccare, dalla scheda relativa a ciascun gene appartenente a questa famiglia, su “gene expression”, per verificare se effettivamente l’espressione di questi geni era attesa nel fegato o meno.

Per ulteriori informazioni sulle MUP naturalmente potremmo fare affidamento a Pubmed, che potremmo utilizzare ad esempio per ricavare articoli di interesse che contengono le parole chiave “Major Urinary Proteins” nel titolo. In ogni caso, per gli scopi di questa esercitazione possiamo limitarci alla pagina wikipedia: https://en.wikipedia.org/wiki/Major_urinary_proteins

Sulla base di quanto leggete, vi sembra che le informazioni che gli autori hanno ricavato per quanto riguarda il topo possano essere anche applicate all’uomo?

Task 11

A questo punto non ci resta che procedere ad uno studio di arricchimento funzionale dei DEG, visto che i geni differenzialmente espressi sono molti, cercando di replicare quanto fatto dagli autori (che a differenza nostra avevano a disposizione uno strumento molto avanzato come IPA). Noi ci affideremo a g:Profiler, uno strumento accessibile online per effettuare dei test ipergeometrici sulle annotazioni.

Partiamo però dal file excel con i DEG e tentiamo di separare i geni sovraespressi nei fegati di topi anziani dai fegati dei topi giovani sulla base della colonna relativa al fold change.

Task 12

Seguendo il link “functional enrichment online” collegatevi a g:Profiler.

Sotto a “query”, incollate la lista di geni differenzialmente espressi, iniziando da quelli sovraespressi in uno dei due gruppi.

Da “ Vi verrà data a questo punto la possibilità di selezionare il nome scientifico della vostra specie di riferimento, quindi selezionate “Mus musculus”.

Da “data sources” sarà possibile selezionare quali sono i database funzionali che intendiamo utilizzare per effettuare le analisi di arricchimento. Questi verranno interrogati automaticamente, collegando la lista dei codici identificativi dei geni differenzialmente espressi alle corrispondenti annotazioni.

Notate come, di default, siano utilizzate le tre categorie di Gene Ontology, KEGG, Reactome e WikiPathways. Oltre a questi, è possibile anche prevedere quali potrebbero essere i principali fattori di trascrizione coinvolti nella regolazione (con TRANSFAC) ed i miRNA coinvolti nel network, oltre ad altri database selezionabili.

Cliccate dunque su “Run query” e, nel caso vengano evidenziate alcune ambiguità nel recupero delle associazioni (cosa possibile nel caso in cui siano presenti geni paraloghi), risolvetele, cliccando poi su “re-run query).

Task 13

In basso comparirà un grafico nella sezione “results”, che ci riporta schematicamente i p-value (in scala log) delle annotazioni sovra-rappresentate. I risultati però possono essere analizzati in maggior dettaglio cliccando su “detailed results”. Qui le annotazioni sono ordinate prima per database e poi per significatività.

Task 14

Ripetete le stesse analisi con il secondo set di DEGs ed analizzate i risultati.

Task 15

Verificate se i vostri risultati vi sembrano coerenti con quelli riportati nella Figura 3 e nella tabella 1 dell'articolo.