Esercitazione RNA-Seq 2020/2021

Obiettivo

L'obiettivo dell'esercitazione è quello di analizzare i dati di 8 campioni di *Saccharomyces Cerevisiae* (lievito) in logica NGS, identificare possibili differenze nell'espressione genica tra le 2 condizioni (WT vs KO) e inferire il significato biologico dei geni differenzialmente espressi.

Svolgimento (punti 1-4 svolti in aula)

1. Carica gli 8 campioni su Galaxy (2 files **FASTQ** x campione), il genoma di riferimento (**genome.fasta**) e il file di annotazione (**genes.gtf**);
2. Crea una semplice pipeline per l'analisi di dati RNA-Seq per quantificare l'espressione dei geni noti (Allineamento con "STAR" + Quantificazione con "HTSeq-count").
3. Esporta per ogni campione i file "XXXX**.tabular**".
4. Crea un unico file testo in cui nella prima colonna ci siano i nomi dei geni e in quelle seguenti il valore di espressione estratto da tutti i file "XXXX**.tabular**":

gene_id	KO_1	KO_2	KO_3	KO_4	WT_1	WT_2	WT_3	WT_4
15S_rRNA	23	31	28	22	36	20	24	35
21S_rRNA	73	58	68	60	62	57	52	63
HRA1	20	18	22	16	17	21	19	12
ICR1	59	74	73	68	66	76	55	67
LSR1	19	25	23	21	24	27	14	22
…	…	…	…	…	…	…	…	…

5. Utilizzando MATLAB o qualunque ambiente di sviluppo, crea uno script che:
    i. importi importi il file;
    ii. filtri i geni non espressi (considera un gene espresso se #reads > 5 in tutti e 8 i soggetti);
    iii. trasformi i valori di espressione dei geni espressi in scala log2;
    iv. disegni boxplot per ogni campione prima e dopo la "log-trasformazione" per testare l'effetto della stessa;
    v. esegua l'analisi differenziale per ogni gene (WT vs KO – in MATLAB usa "`mattest`") ed effettua una correzione per test multipli (in MATLAB usa "`mafdr`")
    vi. disegni un heatmap con solo i geni con p_value <0.01.

6. Usando DAVID (https://david.ncifcrf.gov/tools.jsp), identifica i Processi Funzionali associati ai geni differenzialmente espressi (si consideri p_value <0.01):
    a. Cliccare su start Analysis;
    b. Incollare la lista di geni nel box "A: Paste a list";
    c. Selezionare "ENSEMBL_GENE_ID" come Identifier
    d. Cliccare su Gene List
    e. Cliccare su submit List
    f. Successivamente, cliccare su "Functional Annotation Tool", poi "Gene_Ontology" e infine sul tasto "Chart" a fianco della voce "GOTERM_BP_DIRECT"

Domande

Quanti sono i geni differenzialmente espressi (p_value <0.01)? E dopo aver corretto per test multipli (q_value <0.05)?

I geni differenzialmente espressi permettono di “clusterizzare” bene i soggetti delle 2 condizioni o qualche campione può essere considerato *outlier*?

Quali sono i 5 processi biologici più significativi, associati ai geni differenzialmente espressi?