TAREFA 5 – análise de transcriptoma "sem genoma completo"

1.Quantos transcritos ou isoformas (Trinity junta a expressão quando de **isoform** em uma entrada "**gene**") foram montados e qual seu tamanho médio? E quantos genes totais foram encontrados (juntando isoformas numa contagem só)?

2. Entrando na pasta salmon_saida e na pasta da fase log de crescimento, olhando o arquivo quant.sf onde há a expressão das isoformas ("_i1" e "_i2"), quantos TPM aparecem para as duas isoformas do DN75? E do DN16? Agora compare com as contagens das isoformas de DN75 noa arquivo quant.sf da pasta da cultura em platô de crescimento.

3. Agora com os arquivos quant.sf.genes quais os 5 genes mais expresso, "DN??". Use:
cat quant.sf.genes | sort -k 4 -n (analise as duas fases, log versus platô)

4. Adicione -r à linha de sort para ver os menores valores, tem genes sem expressão, são os mesmos nas duas fases de crescimento ou diferentes?

5. Depois de fazer uma matriz, uma comparação das fases, olhe as contagens normalizadas por **TMM**, o jeito mais chique de medir expressão pois RNAseq não mede variáveis independentes, então TMM apaga o efeito de um gene muuuuito expresso que faria todos os outros diminuirem a expressão em TPM (transcritos por milhão). Capture quem tem expressão 0.00 usando grep!!
cat salmon.gene.TMM.EXPR.matrix | grep "0.00"

6. Com o edgeR a diferença de expressão (tabela DE_results) é avaliada com p-valor e com False Discovery Rate, e tem o log do fold change, quantas vezes cai pela metade na primeira amostra)valores negativos de log2) ou dobra na segunda amostra (valores positivos). Se fosse usado log na base 10 não seria dobrar, seria o quê?

7. Muitas pessoas usam FDR menor que 0,05 a sétima coluna ($7) da tabela DE_results acima. Tire o less e use wc:
cat salmon.isoform.counts.matrix.Log_Phase_vs_Plat_Phase.edgeR.DE_results | awk '$7 <= 0.05 {print $1,$4,$7}' | wc
Mas quantos genes seriam escolhidos se a gente usar valor de FDR de dez a menos três? Ou dez a menos 5? **Liste esses últimos trocando o wc por** less e salve. Agora diga, a maioria dos genes dessa lista cai várias vezes pela metade na fase log ou dobra várias vezes na fase platô? Quantas vezes dobra o que mais dobra? Quantos genes são **mais expressos** com esse corte de FDR no platô (dica, log do fold change positivo)?

8. No momento com uma reforma no uniprot.org idmapping service não se conecta ao Kegg, é preciso copiar o que vem em Entry (exemplo **SPBP35G2.12)** e abrir DBGET colar e dar [GO] e ai abrir o link de KEGG GENES pra ir à página do gene e nela pode ou não ter o link para Pathways (para essa tem). O EC number dela (Enzyme Catalog Number) vai estar em vermelho já: **3.6.1.13 ADP-ribose diphosphatase.** Analise também outras assim.

7. Use "pelo menos 10 ligações de primeiro e segundo nível da base" no String-DB como na aula, faça clusters e olhe o enriquecimento deles! O que mais teve de diferença?

YDD3_SCHPO
YI7H_SCHPO
PDC4_SCHPO
F16P_SCHPO
YDPH_SCHPO
YOF7_SCHPO
YF89_SCHPO
YF8B_SCHPO
YA7D_SCHPO
YQ61_SCHPO
YN8C_SCHPO
RL24A_SCHPO
RL24A_SCHPO
GLD1_SCHPO
HSP71_SCHPO
YC9F_SCHPO
HSP72_SCHPO
YG06_SCHPO
YOXA_SCHPO
YER5_SCHPO
PLB5_SCHPO
YGK3_SCHPO
YNS5_SCHPO
HSP31_SCHPO
CARA_SCHPO
INV1_SCHPO
NOL10_SCHPO
ZYM1_SCHPO
HSP90_SCHPO
YAAB_SCHPO
RRS1_SCHPO
FIO1_SCHPO
DAK2_SCHPO
YKN2_SCHPO
RS30A_SCHPO
LSD90_SCHPO
LAS1_SCHPO
YAKC_SCHPO
YAY8_SCHPO
SSDH2_SCHPO
URG1_SCHPO
HSP31_SCHPO
YBN4_SCHPO
YFE3_SCHPO
YDYE_SCHPO
YEPF_SCHPO
RL401_SCHPO

SAHH_SCHPO
PPID_SCHPO
ZHF1_SCHPO
HSP16_SCHPO
AGLU_SCHPO

8. Por fim, use o limite de FDR 0.00001 que é bem restrito, abra as entradas uniprot_ac no uniprot.org e escolha um com a qual faria uma pesquisa longa, leia tudo que puder sobre ela no uniprot.org, e justifique sua escolha, considerando o interesse do laboratório em estudar em detalhe um gene da S. pombe afetado pelas condições de ciclo celular.

Entregue um relatório sobre a análise com Trinity e suas respostas