If you have any questions or need help, just ask!!!

RNASEQ – Huntington's disease

1. https://www.nature.com/articles/s41598-022-06970-6#Sec21

*Recent publication collecting RNASEQ data from HD patients and controls

Here you need to see where the data comes from: female/male, where the tissue was taken from, what ethnic minority, origin, etc.

The data will have to be downloaded in the format available in NCBI GEO. The ID of the data in this publication is given in the supplement of table 1. You can download the data to your computer and check that they are all in the same or at least similar format so that they can be analyzed right away. Once you've prepared your content description data, you can proceed to the next step. If the data are in different formats, they must be converted to one format (e.g. xlsx, tsv, csv, etc.).

Data for RNASEQ must be normalized (bring them to a state where the state of gene expression will be reduced to real values). This is due to the fact that different genes have different transcript lengths, so during sequencing we have different numbers of reads per gene related to different lengths, and not to the expression itself. This is because during sequencing, transcripts are cut into smaller fragments and longer transcripts will produce more fragments. Example:

Gen1 - length 1500, after fragmentation of 11 fragments, gene copy number per cell 3 (3*11 = 33 reads for this gene).

Gen2 – length 600, after fragmentation of 5 fragments, copy number of genes per cell 3 (3*6=18 reads for this gene)

18 > 33 -> apparent expression for gene 1 is higher, but it is not true, because the real expression equals 3, i.e. the number of copies. Therefore, different types of normalization are used!!!

2. Familiarize yourself with the types of data normalization:

   https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
   https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

   Do not learn by heart, rather choose a method for the analysis, you can always look at the cheat sheet ^^

3. Download RStudio and R

   https://www.rstudio.com/

   https://cran.r-project.org/bin/windows/base/old/

4. For the analysis of RNASEQ data in R, the most widely used library is DESEQ2, so I recommend downloading it and familiarizing yourself with its functions.

   How download: https://bioconductor.org/packages/release/bioc/html/DESeq2.html

   How to work with data in DESEQ2:
   http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Have fun!