**Differential Gene Expression using Kallisto and Degust**

Based on tutorial found in:

<http://sepsis-omics.github.io/tutorials/modules/kallisto/>

1. **Obtain Genbank file for both organism:**

E coli: <https://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521>

P aeruginosa: <https://www.ncbi.nlm.nih.gov/genome/187?genome_assembly_id=299953>

Graphical user interface, text, application

Description automatically generated

1. **Download and run the following Python script (from now on the example commands will be for the e\_coli analysis, but it’s the same for the other bacteria):**

<https://github.com/AnnaSyme/genbank_to_kallisto.py>

$ python3 genbank\_to\_kallisto.py <Genbank\_file> e\_coli.transcripts e\_coli.table

Note: the P. aeruginosa Genbank file fails to run with this script. I manually deleted the gene that was crashing it. The deleted gene di dn’t specify a sequence and I think that was the problem.

1. **Install Kallisto:**

$ conda install kallisto

1. **Index the reference file:**

$ kallisto index -i e\_coli\_genbank.idx e\_coli.transcripts

1. **Run Kallisto quant on each sample:**

$ kallisto quant -i e\_coli\_genbank.idx -b 100 -o e\_coli\_ctrl\_1 ./e\_coli/SRR12820309/SRR12820309\_1.fastq ./e\_coli/SRR12820309/SRR12820309\_2.fastq

-o is the output folder and should be changed depending on sample to analyze

This will produce a folder per sample. Each folder will have 3 files with the results.

1. **Extract the columns of interest for each run:**

cut -f4 -d$'\t' abundance.tsv | tail -n +2 > e\_coli\_ctrl\_1\_headless.tsv

echo -e "CTRL1" | cat – e\_coli\_ctrl\_1\_headless.tsv > e\_coli\_ctrl1.tsv

These commands will produce two files. The final file will consist of a single column with the header being the name of the run (in this case “ctrl\_1”) and the est\_counts of each gene.

After producing all these single column files for each run, copy them to the main folder and run the following command:

$ paste <result\_1> <result\_2> <result\_3> e\_coli.table > counts.tsv

Where each <result> is the single column file for each sample and the e\_coli.table file is the one produced by the script in step 2. The .table file contains the information of each sequence in the transcript file.

NOTE: the tutorial does not mention this, so I don’t know if I did something wrong, but if you paste the .table file as is, the information for each transcript shifts one row up because the .table file does not contain headers for each column. So I manually added headers to the .table file (ID, type, name, ???, description).

1. Go to <https://degust.erc.monash.edu/>:

* Click on Upload

Text

Description automatically generated

* Click “choose file” and select the counts.tsv file.
* Click “Upload”
* Specify a name for the run.
* Select “RNA-seq counts” on Input type
* On info columns select the column with the gene name
* Specify “10” on “min gene read count”
* Click on “Add condition”
* Write “Control” and select the columns that belong to controls
* Click on “Add condition” again
* Write “Treatment” and select the columns that belong to treatment
* Click on “Save changes”
* Click “View”

Chart, scatter chart

Description automatically generated

Note: in the files section you will find:

* The Genbank file I used for the genbank\_to\_kallisto python script
* The files generated by the python script
* The index file generated by Kallisto
* The outputs of each run
* The final counts.tsv file for both datasets
* List of differentially expressed genes for both datasets with a FDR of 0.05 and 0.01

That way you can start the “pipeline” at any point to check for my results.

E coli results:

<https://degust.erc.monash.edu/degust/compare.html?code=5719b6002251d31150e2bc998a96f228#/>

P aeruginosa results:

<https://degust.erc.monash.edu/degust/compare.html?code=7c964c8ed5b0c5f58f3a4bd590cbaa87#/>