General tasks for report

Please create a folder under project folder (/proj/g2018024/nobackup/) with your user name in uppmax (e.g. /proj/g2018024/nobackup/czhang) and save all results there so you will not mess up your results with others.

A. Preprocessing (Kallisto) – Convert FASTQ files to Count and TPMs

Refer to the kallisto manual through the link https://pachterlab.github.io/kallisto/manual

1. Build an index file for mice.

To do this, you need to first download a mouse genome file from ensembl website (http://www.ensembl.org/index.html). Click the cute mice photo (GRCm38.p6) on the home page of ensembl so we will arrive http://www.ensembl.org/Mus_musculus/Info/Index, and click 'Download FASTA' under 'Genome annotation' and get to the ftp site ftp://ftp.ensembl.org/pub/release-94/fasta/mus_musculus/. In the ftp site, go to cdna folder, and click on the file 'Mus_musculus.GRCm38.cdna.all.fa.gz'. You can also download genome file from other organisms using similar approach. You can then upload this genome file onto uppmax and create an index file as you learned from the course.

2. Process the FASTQ files.

We have uploaded already 10 pairs of fastq file from mice in the project folder under (/proj/g2018024/nobackup/TranscriptomicWorkshop/FASTQ_Task). There are 5 sample pairs from control mice with file names starts with 'CMC_liver_6h_SHAM_', and 5 sample pairs from disease mice with file names start with 'CMC_liver_6h_MI_' and 5 sample. You will need to process them using kallisto with the index file you created. Note that you will need to create different folders each time you process a sample pair. After processing all samples, you need to merge them into a big table where each row represents a gene, and each column represents a sample so you can get to the next section.

Output: matrix with ensemble transcripts, counts and TPMs.

B. Analysis – Perform differential expression analysis and gene set enrichment

Refer to the file "Rcode.R" for help with the following steps.

1. Convert transcript ids to genes and pre-process transcriptomics data.

To help in your report, we've added a file that you can use to do the mapping between ensembl transcript (ENSMUST00000082423.1 or ENSMUST00000082423) to gene (ENSMUSG00000064372.1 or ENSMUSG00000064372) to gene name. This file is named "Mus_musculus_Ensembl2Gene.txt", and may also be retrieved from http://www.ensembl.org/biomart/. You must work with gene names from this point. For genes with multiple hits, sum its transcripts. Exclude transcripts with no corresponding gene name.

2. Merge files such that you have a single count matrix with all samples, and its associated metadata, ready to use by DESEq2.

[Optional: perform QC on all samples using PCA and add it to your report]

3. Perform DESeq and output it to another file.

How many genes are up- or downregulated in the disease state? What significance threshold do you choose and why? How many false positives to you estimate that you may have? What is the minimum absolute fold change that you find? And the maximum? What are the 10 genes showing the highest differential expression?

4. Perform GSEA for GO biological processes, and output it to another file.

Retrieve GO biological processes from GO2MSig (http://www.go2msig.org/cgi-bin/prebuilt.cgi?taxid=10090). Use this gene set list together with the fold change and p-values from DESeq2.

How many biological processes are enriched, and at which significance?

[Optional: identify the main up-/downregulated biological processes using DAVID]

Output: DESeq2 and PIANO output files.

C. Modeling – Integrate transcriptomics data with the reference model and perform standard modeling tasks

Refer to the file "Main.m" under 'MetabolicModelWorkshop' folder for help with the following steps.

1. Reconstruct a GEM for mice liver using the TPM you got from the sample you processed (Optional).

We have uploaded a reference model for mice named 'MMR_10.xlsx' under the project folder '/proj/g2018024/nobackup/MetabolicModelWorkshop/ReferenceMiceModel'. Together with that, you will need to create an expression file in the same format as 'HepG2_exp.txt' which you can find under the project folder '/proj/g2018024/nobackup/MetabolicModelWorkshop'. Since you have 10 samples, but you need only one expression value for each gene, what you need to do is to calculate the maximal expression value of each gene across all samples so that we could generate a mice liver model that covers all the active metabolic reactions. In addition, you will need the same task file 'common_tasks_growth_RPMI1640.xlsx' we used during the workshop again under the project folder. Finally, you will need to set some parameters for mosek as we did in the workshop.

Please note that this is optional task since that the reconstruction step will take around 9 hours, and you cannot finish that using interactive model in uppmax as you learned in the course. Therefore, in order to do that, you need to either learn how to submit a job which you could learn from http://uppmax.uu.se/support/user-guides/rackham-user-guide/, or you could download everything you need in your PC and do it locally.

After this section, please answer the following questions in your report: How many reactions, metabolites and genes have been removed from the reference model?

2. Gene essentiality analysis.

Using the model we provided ('MMR_10.xlsx') or you generated yourself, we could assess the gene essentiality by doing following steps. Firstly, we need to simplify the model to make it ready for flux balance analysis. This could be done using the code 'model = simplifyModel(model)'. In addition, you need to set the objective function as ATP production, which could be done by 'model = setParam(model,'obj','HMR_7799',1)'. Opening the model file in excel would let you see what is reaction "HMR_7799". After these two steps, you could perform the gene essentiality analysis by referring to the steps in 'Main.m' as we taught during the course.

After this section, please answer the following questions in your report: How many genes are essential? What are their functions (select 3 of them for example)?

3. Reporter metabolite analysis.

Again, using the model we provided ('MMR_10.xlsx') or you generated yourself, you should be able to identify the reporter metabolites which are the hubs in the metabolic model. For doing that, you should format the differential expression analysis result you obtained from section B.

After this section, please answer the following questions in your report: How many reporter metabolites are identified? What are the top 3 metabolites?

P.S. The gene names in the GEM are gene symbols, so you need to use gene symbols always in this section!

Output: The number of essential genes, and reporter metabolite lists as txt file.

D. Metagenomics – Analyze metagenomics data and report species abundances (Optional)

Please follow the steps on the metagenomic workshop slides on github and try to understand the commands. Examples and all intermediary results are provided in the uppmax folder.

Output: Heatmap or table to show the species and their relative abundance

Please send the report by the end of 10/19/2018 to muhammad.arif@scilifelab.se, rui.benfeitas@scilifelab.se and cheng.zhang@scilifelab.se.

PS: As mentioned before, if you have any question, you can contact or visit us at any time.
SciLifeLab, Alfa 6 (right when you enter the floor via the left doors, our office is on your left)