**Detailed description of our analysis pipeline**

To reproduce all results in this paper, run all scripts in the order in which they are listed in this file.

**Part A**

run: generating\_RNA\_RawCount.R

generates

- a\_results / rnaRawData.Rda

- a\_results / rnaMatrix\_meta.csv

- a\_results / rnaMatrix\_mRNA.csv

- a\_results / rnaMatrix\_RNA.csv

- a\_results / rnaMatrix\_rRNA.csv

- a\_results / rnaMatrix\_tRNA.csv

- a\_results / rnaMatrix.csv

- a\_figures / log2AmountDensity.pdf

run: generating\_Protein\_RawCount.R

generates

- a\_results / proteinMatrix\_w\_NA.csv

- a\_results / proteinMatrix\_wo\_NA.csv

- a\_results / proteinMatrix.csv

run: generating\_Meta\_data.R

generates

- a\_results / sampleSizeDf.csv

- a\_results / metaData.csv

- a\_results / metaProtein.csv

- a\_results / metaRNA.csv

- a\_results / metaRawData.Rda

run: data.normalization.R with parameters.R (RUNS 1 to 15) with given parameters

|  |  |  |  |
| --- | --- | --- | --- |
|  | All Data | Exp | Sta |
| mRNA | 1 DONE | 2 DONE | 3 DONE |
| Protein | 4 DONE | 5 DONE | 6 DONE |
| Int\_mRNA | 7 DONE | 8 DONE | 9 DONE |
| Int\_protein | 10 DONE | 11 DONE | 12 DONE |
| Int\_mrna\_protein | 13 DONE | 14 DONE | 15 DONE |

With parameters

* Data type: TABLE
* Bad data set: set00 (No bad sample is excluded)
* Reference parameter order: growthPhase, Mg\_mM\_Levels, Na\_mM\_Levels, carbonSource, experiment
* Referece levels: exponential(stationary), baseMg, baseNa, glucose, glucose\_time\_course
* experimentVector: allEx
* carbonSourceVector: SYAN
* MgLevelVector: allMg
* NaLevelVector: allNa
* growthPhaseVector: TABLE
* filterGenes: noMatchFilter
* threshold: NA
* roundData: TRUE
* sumTechnicalReplicates: TRUE
* deSeqSfChoice: P1Sf
* normalizationMethodChoice: vst

those runs generate

output of run No 1:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 2:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 3:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 4:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 5:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 6:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 7:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 8:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 9:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 10:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 11:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 12:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 13:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 14:

- a\_results / metaData\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_protein\_trT\_set00\_StcYtcNasAgrNgrMgh\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

output of run No 15:

- a\_results / metaData\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

- a\_results / resDf\_mrna\_trT\_set00\_StcAllEx\_SYAN\_baseMgAllMg\_baseNaAllNa\_ExpAllPhase\_noMatchFilter\_p1Sf\_noNorm.csv

**PART B**

**PART B**

B\_code\_MLRN

This folder does the machine learning part.

It composes of four subfolders.

starterFilesDiscreate: run the machine learning codes the file names are explanatory for sub categories.

For example: code\_int\_mrna\_protein\_Mg.R runs the machine learning pipeline for intersection data and for only Mg samples and for both mrna and protein data. The results of the individual runs are saved under   
b\_results folder. The callAll.R file call all the files one by one.

The pipeline: When the starter file runs in finds the data that will be used and send it to the pipeline the pipeline composes of multiple steps. Main steps are checks if two distinct data sets need to be combined like mRNA and protein

|  |  |  |
| --- | --- | --- |
| Individual | combined | What the code does |
| machineLearning\_subCode\_initDfprep.R | same | It prepares the data  Changes the information in DF into human readable form  Define abbreviations Put the column labels to correct order  Define F1 function |
| machineLearning\_mainCode.R | machineLearning\_mainCodeComb.R | Runs the main pipeline in parallel  # Divide the data as train and test  # do batch correction  # insert data preparation function  # do pca  # Generate Class Weight Vector  # Generate train tune data frames in trainTuneDFs # Generate Gamma and Cost vectors for tuning  # tunes the models  # make predictions on test data  # calculate the scores and save them |
| dataPreperation\_func.R | dataPreperationComb\_func.R | Prepears the data for the run   # Preperation of Data for PCA and PCoA  # 1.) Correct batch effects  # 2.) Remove rows (genes) with 0 sd from train (PCA fails without this)  # 3.) remove same genes from test set data  # 4.) transpose the data frame |

As said previously the code in the individual runs calls the files inside pipeline and the results are saved into b\_results. (the runs take time)

figureRegenerateFilesDiscreate: The folder generates figures associated with individual runs. The folder is there because the individual machine learning runs too long time (a day). Instead running all the machine learning pipeline, the codes there takes the output results and use them to generate desired figures.

File names are self explanatory: i.e. int\_mRNA\_analyze.R finds the results for intersecting mRNA data and sends to mainAnalyzeCode.R the file generate all the associated figures and generate some summary tables as well. The code that calls all figure generation codes one by one is called callAll.R

The associated figures goes to \b\_figures and associated tables goes to \b\_results.

additionalFigures: Most of the time we need to generate figures that are produced by multiple subsets of data. For example increase in preformace after combining mRNA and protein data figure (Figure 5) requires combining analyze of 3 sub data sets (intersection mRNA, intersection protein, intersection mRNA and protein). For those figures we need to call summary tables generated by the files inside \figureRegenerateFilesDiscreate folder. The files call the summary files from \figureRegenerateFilesDiscreate and generate the publication figures are located at \additionalFigures folder. The files produce the figures under \b\_figures folder.

One can see which file generates which figure from the file

GitHub/ecoli\_learning\_bacterial\_response\_optimization/text/figures/figures\_and\_tables\_preparation\_information.txt