A Transcriptomics Analysis
CMPE 492 Project - Boğaziçi University

Project Coordinators: Adil Mardinoğlu, Özlem Altay
Project Advisor: Taflan Gündem
Ömer Faruk Özdemir farukozderim@gmail.com

INTRODUCTION

- Transcriptomics analysis focuses on analysis of RNA data.
- Analyzing transcript data from different diets may show us possible effects on some diseases and thus possible biomarkers for drugs.
- Functional analysis and GEM analysis are popular ways for this task.
- We aim to analyze transcript data from muscle and liver based on fish, lard and soy diets on mus musculus.

DATA

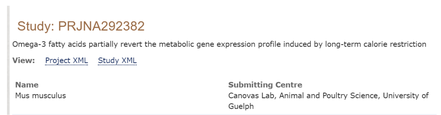

Figure 1:Dataset web screenshot [4]

- The dataset contains 29 instances.
- Examples are taken from muscle and liver tissues.
- There 4 class:
  1-Soy-oil diet
  2-Fish-oil diet
  3-Lard-oild diet
  4-Control diet

- **Data Processing Pipeline**

1. RNA sequencing from muscle and liver tissues of mus musculus
2. Quality control and filtering
3. Genome mapping
4. Quantification
5. Sample exploration
6. Differential Expression
7. Functional Analysis / GEM Analysis / Network Analysis

Data Structure And Analysis

- First raw transcript data is obtained from the dataset's website, which are in fastq format.

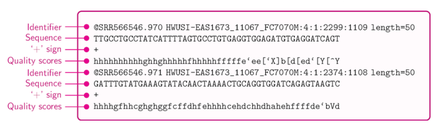

Figure 2:Fastq format[1]

- Quantification is carried out with kallisto which generates number of gene counts in the raw transcript data. Raw data contains only nucleotide sequences, and it is fractions in random way, we extract number of genes using psudoalingment method with kallisto using a reference index dataset which contains gene IDs and gene sequences. This process creates abundance data. The abundance data is in normalized form TPM(transcripts per kilobase million).
- Abundance datas are consolidated into one table.
- Data is clustered using PCA for visualizing different diet classes.
- Differential Expression Analysis is applied taking control group as reference using deseq2 library. Which creates gene affect ratios and probability of null hypotesis using negative binomial distribution.

"geneID"	"baseMean"	"log2FoldChange"	"lfcSE"	"stat"	"pvalue"	"padj"
"0610009B22Rik"	212.62584072533	-0.328166129603727	0.161330894542038	-2.03411832888722	0.0419396663828995	0.48046777824344
"0610010F05Rik"	239.055010810135	0.161868475201437	0.212812194667943	0.760616540109488	0.446886137575549	0.981743048536826
"0610010K14Rik"	111.304184560309	-0.206145241197334	0.318943732930199	-0.646337331363864	0.518069916280909	0.986343170791891

Figure 3:Deseq2 entry examples[1]

- Functional analysis is applied via piano library using deseq2 outputs. It uses statistical methods for analyzing gene level statistics. It combines the deseq2 results with the biological pathway info it has. It shows effect of gene activation info on human diseases and pathways.

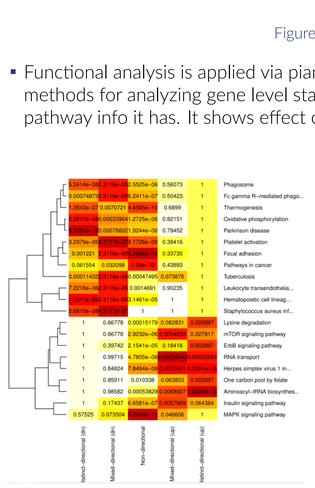

Figure 4:Heatmap [3]

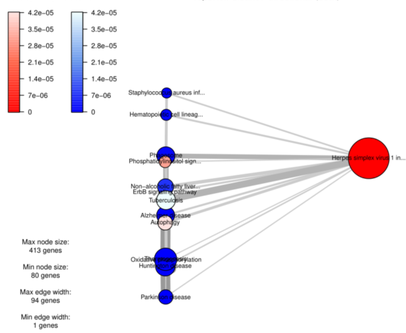

Figure 5:Network Plot [3]

Our initial plan was to apply GEM(Genome Scale Metabolic Model) based analysis using a cancer model. GEM model contains metabolic equations and cell level flux constraints. Using deseq2 outputs it applies a constaint based flux analysis using Linear Problem techniques. It uses an objective function representing ATP production or growth. But this failed due to the difference in naming styles of genes, which created inconsistency with deseq2 files and the model. Thus this is left as a future work.

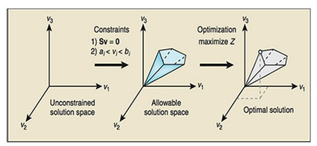

Figure 6:Constraint based modeling[2]

RESULTS

- Applying Functional Analysis we showed effects on pathways and diseases using gene activation info obtained from the dataset. Network Plots were generated for each diet for each tissue. It shows the significant gene-sets and the overlap of genes between sets.

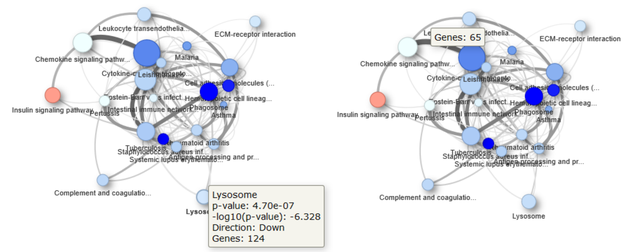

Figure 7:Network Plot Graph [3]

CONCLUSION

We applied Differential gene expression analysis which shows effects of diets on gene expression. Generated results for each diet for both liver and muscle tissues.

Using these results we applied Functional analysis, calculated effects of gene expression on biological pathways and diseases using gene-set analysis(GSA) via piano. For highly effected pathways or diseases drugs can be tested with the elements diets contain.

ACKNOWLEDGEMENTS

I want to express my gratitude to Adil Mardinoğlu and Özlem Altay for their assistance and advisory in the project. I also would like to thank Taflan Gündem for his allowance to make this project as my Bachelors Degree Senior Project.

REFERENCES

[1] Hosseini, M.; Pratas, D.; Pinho, A.J. A Survey on Data Compression Methods for Biological Sequences. Information 2016, 7, 56.

[2] Orth, J., Thiele, I. Palsson, B. What is flux balance analysis?. Nat Biotechnol 28, 245–248 (2010). https://doi.org/10.1038/nbt.1614

[3] https://github.com/FarukOzderim/cmpe492-Transcriptomics_Analysis/tree/master/results

[4] https://www.ebi.ac.uk/ena/data/view/PRJNA412001