# BioKit: An application for gene expression data analysis

Christopher LoBianco '19 | Advisor: Dr. Chris Armen

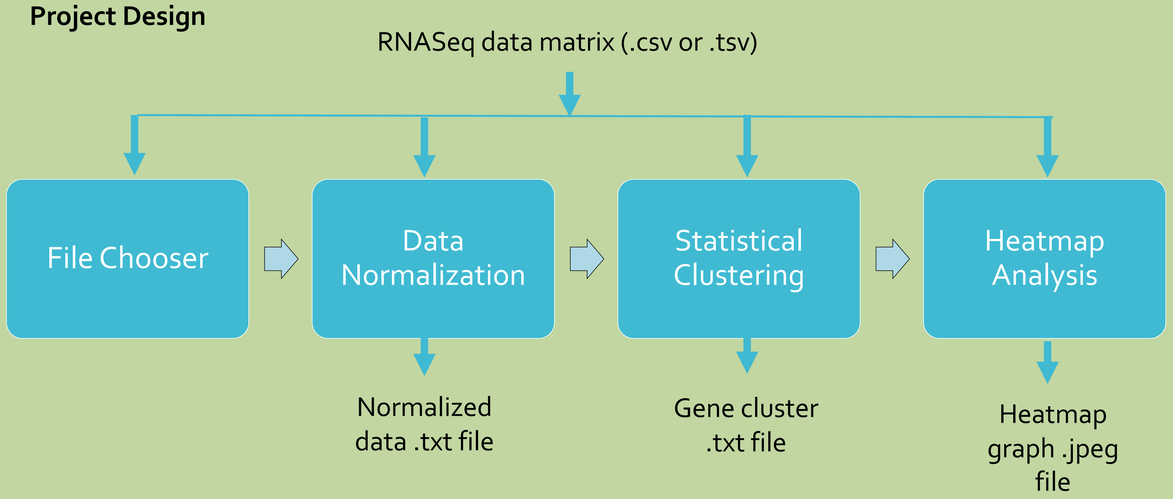
## Motivations

Biological research today is characterized by large, complex datasets. These datasets require the algorithms, data structures, and processing power of bioinformatics to efficiently analyze.

Many bioinformatics tools and software modules are freely available to implement these algorithms. However, they are often inaccessible to researchers due to unfamiliarity with these resources or insufficient coding background necessary to implement them.

This project, BioKit, provides an interface for biology researchers to access these modules. It encapsulates three modules for gene expression data analysis and provides a GUI for a user to navigate through the steps of data processing.

## Project Design




## Results

Figure 2: Raw RNAseq reads matrix

| | SRR1039508 | SRR1039509 | SRR1039512 | SRR1039513 | SRR1039516 |
|---|---|---|---|---|---|
| ENSG00000000003 | 679 | 448 | 873 | 408 | 1138 |
| ENSG00000000005 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000000419 | 467 | 515 | 621 | 365 | 587 |
| ENSG00000000457 | 260 | 211 | 263 | 164 | 245 |
| ENSG00000000460 | 60 | 55 | 40 | 35 | 78 |
| ENSG00000000938 | 0 | 0 | 2 | 0 | 1 |
| | SRR1039517 | SRR1039520 | SRR1039521 | | |
| ENSG00000000003 | 1047 | 770 | 572 | | |
| ENSG00000000005 | 0 | 0 | 0 | | |
| ENSG00000000419 | 799 | 417 | 508 | | |
| ENSG00000000457 | 331 | 233 | 229 | | |
| ENSG00000000460 | 63 | 76 | 60 | | |
| ENSG00000000938 | 0 | 0 | 0 | | |

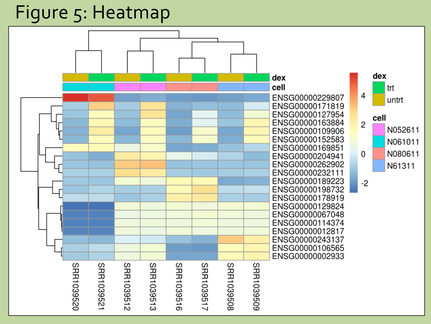
Figure 3: Normalized data matrix

| | SRR1039508 | SRR1039509 | SRR1039512 | SRR1039513 | SRR1039516 |
|---|---|---|---|---|---|
| ENSG00000000003 | 9.399151 | 9.142478 | 9.501695 | 9.320796 | 9.757212 |
| ENSG00000000005 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| ENSG00000000419 | 8.901283 | 9.113976 | 9.032567 | 9.063925 | 8.981930 |
| ENSG00000000457 | 7.949897 | 7.882371 | 7.834273 | 7.916459 | 7.773819 |
| ENSG00000000460 | 5.849521 | 5.882363 | 5.486937 | 5.770334 | 5.940407 |
| ENSG00000000938 | -1.638084 | -1.637483 | -1.558248 | -1.636072 | -1.597606 |
| | SRR1039517 | SRR1039520 | SRR1039521 | | |
| ENSG00000000003 | 9.512183 | 9.617378 | 9.315309 | | |
| ENSG00000000005 | 0.000000 | 0.000000 | 0.000000 | | |
| ENSG00000000419 | 9.108531 | 8.894830 | 9.052303 | | |
| ENSG00000000457 | 7.886645 | 7.946411 | 7.908338 | | |
| ENSG00000000460 | 5.663847 | 6.107733 | 5.907824 | | |
| ENSG00000000938 | -1.639362 | -1.637608 | -1.637724 | | |

Figure 4: Differential expression analysis

```
## log2 fold change (MAP): dex trt vs untrt
## Wald test p-value: dex trt vs untrt
## DataFrame with 6 rows and 8 columns
##                   baseMean log2FoldChange     lfcSE      stat
##                  <numeric>      <numeric> <numeric> <numeric>
## ENSG00000152583   997.4398       4.316100 0.1724127  25.03354
## ENSG00000165995   495.0929       3.188698 0.1277441  24.96160
## ENSG00000101347 12703.3871       3.618232 0.1499441  24.13054
## ENSG00000120129  3409.0294       2.871326 0.1190334  24.12201
## ENSG00000189221  2341.7673       3.230629 0.1373644  23.51868
## ENSG00000211445 12285.6151       3.552999 0.1589971  22.34631
##                          pvalue          padj      symbol      entrez
##                       <numeric>     <numeric> <character> <character>
## ENSG00000152583 2.637881e-138 4.755573e-134      SPARCL1        8404
## ENSG00000165995 1.597973e-137 1.440413e-133       CACNB2         783
## ENSG00000101347 1.195378e-128 6.620010e-125       SAMHD1       25939
## ENSG00000120129 1.468829e-128 6.620010e-125        DUSP1        1843
## ENSG00000189221 2.627083e-122 9.472210e-119         MAOA        4128
## ENSG00000211445 1.311440e-110 3.940441e-107         GPX3        2878
```

Figure 5: Heatmap




## Features

- User can upload RNAseq data file
- Process data using a regularized logarithmic transformation
- Cluster similarly expressed genes using hierarchical clustering
- Visualize results using a graphical heatmap
- Output all results as .txt or .jpeg files to system directory

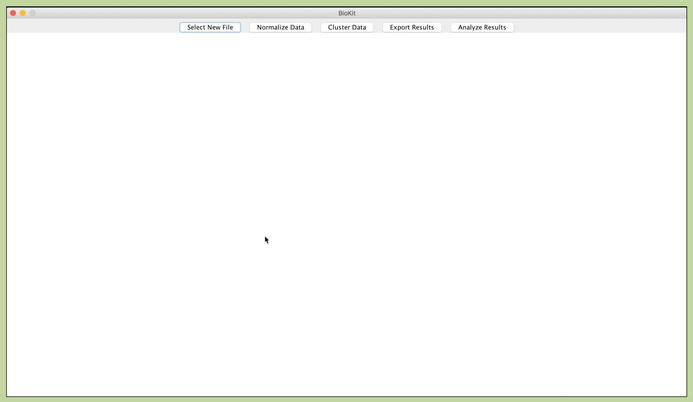
## Project Implementation

- Desktop application
- GUI components
- File chooser

- RUVnormalize
- multiClust
- DESeq2

*R Packages

Figure 1: User Interface & Main Menu




## Modules

- RUVnormalize: implements a log-transformation to remove experimental variation from sequence reads
- multiClust: implements hierarchical clustering to identify genes expressed under similar experimental conditions
- DESeq2: provides probability testing and heatmap output

## References


- Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA- seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.
- Jacob, L., Gagnon-Bartsch, J., Speed, P. T (2016). "Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed." *Biostatistics*.
- Lawlor N, Guan P, Fabbri A, Karuturi K, George J (2018). *multiClust: multiClust: An R-package for Identifying Biologically Relevant Clusters in Cancer Transcriptome Profiles*. R package version 1.12.0.


## Acknowledgements


- Thank you to my advisors in the CS and biology department
- Thank you to the authors of the BioConductor modules