RDS ICA

B242415

Importing Data and Libraries

Required files in workspace:

1. data_all.csv - .csv file containing sample names in row 1 and gene names in column 1.
2. gene_annotation.csv - .csv file containing gene names, their type and their long names.
3. genelist_72.csv - .csv files containing genes of interest to be included in analysis.
4. sample_annotation.csv - .csv file containing sample names and their treatment groups.

Required installed libraries:

1. pheatmap

```
## Warning: package 'pheatmap' was built under R version 4.3.2
```

```
## [1] "Imported files!"
```

Data Filtering

Extracting data from data_all.csv with only genes listed in genelist_72.csv into data_filtered dataframe.

```
## [1] "1 duplicate(s) found in genelist.csv"
```

```
## [1] "40 gene(s) seleced from data_all.csv for further processing."
```

Data Processing

log scale the data into logdata dataframe.

```
## [1] "0 NA values in data_filtered"
```

```
## [1] "Data log scaled."
```

Extracting annotations and long names

Extracting gene type and treatment type annotations from sample_annotation.csv and gene_annotation.csv.
Also renaming gene names into their long name for heatmap plots.

```
## [1] "All genes annotated with gene type"

## [1] "All samples annotated with treatment groups"

## [1] "All genes renamed to their long names"
```

Heatmaps

Figure 1: Heatmap of log mean gene counts of selected genes across samples clustered by genes and samples

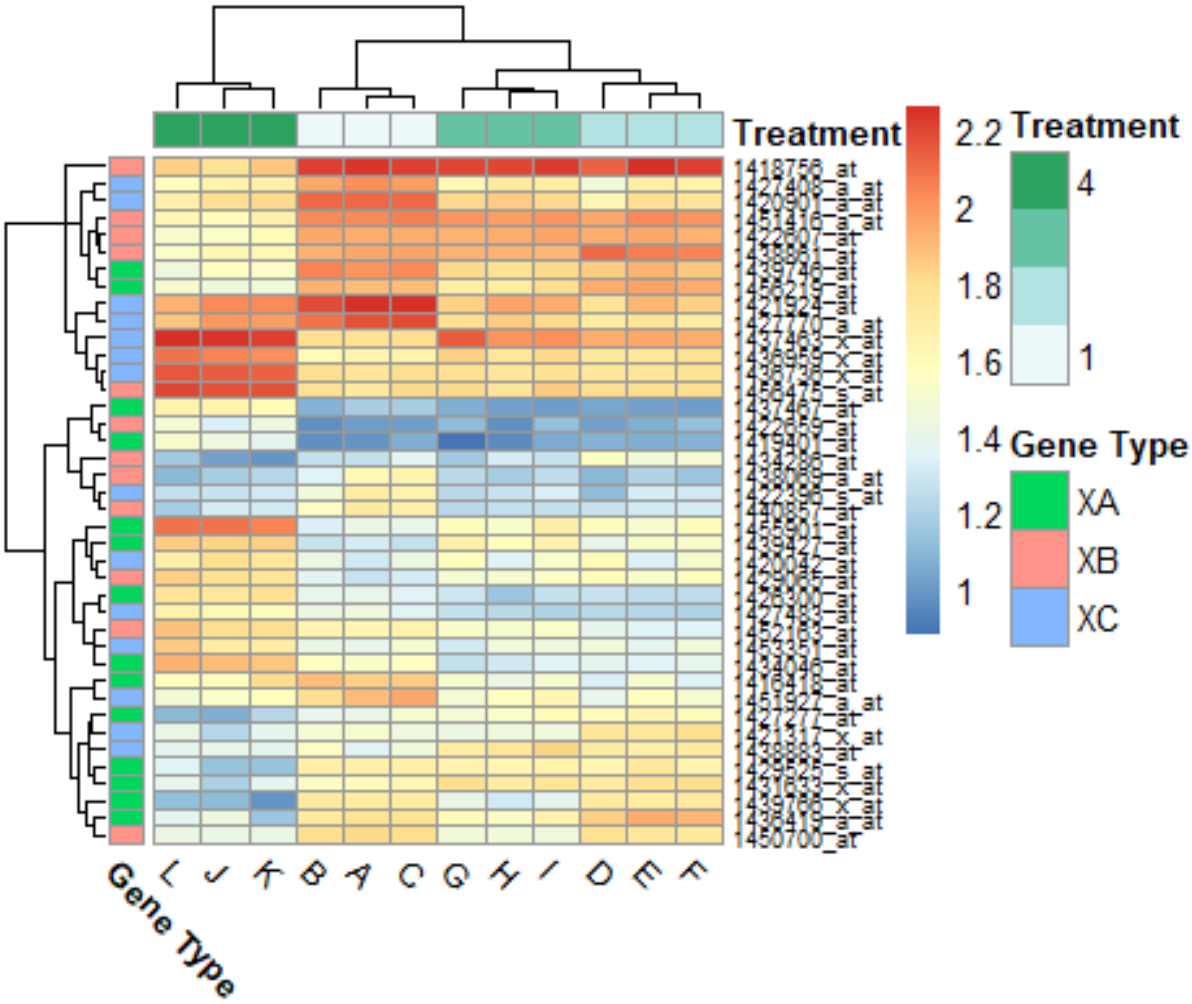
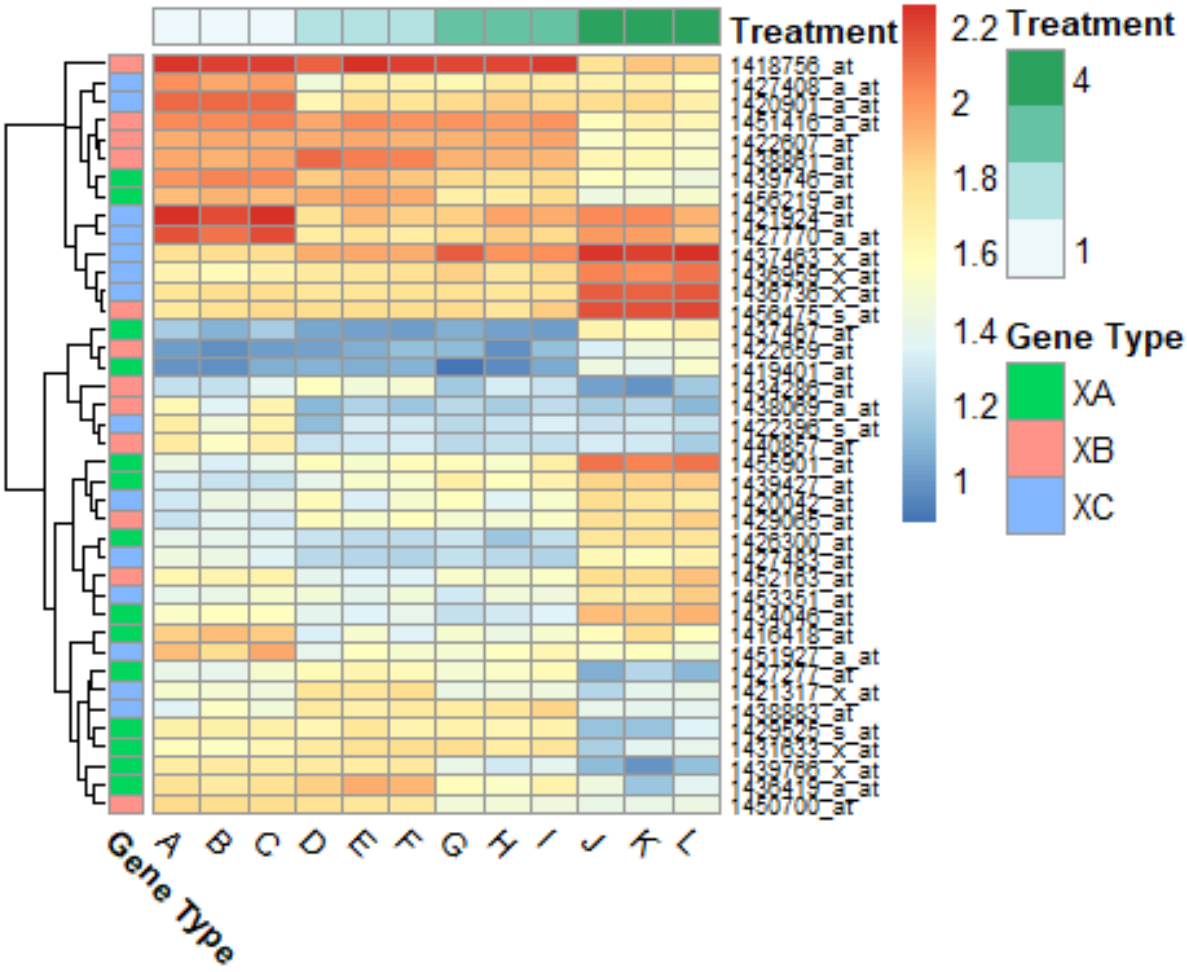


Figure 2: Heatmap of log mean gene counts of selected genes across samples clustered by genes



Additional Information on the Processing of Data

The numerical data has been assumed to be mean gene counts as they are not integers, but floats.

For further processing of the data, normalization of gene counts is suggested (Not done here). If housekeeping genes have been included in the experimental design, relative gene abundances for valid comparisons can be gained via methods such as the median of ratios used by DESeq2 (Love, Huber & Anders, 2014). If not, a relative proportion of transcripts in the pool of gathered RNA can be used for valid comparisons with methods such as TPM (transcripts per million) (Zhao, Ye & Stanton, 2020). However, due to the lack of context in the given data, it may already have been performed.

Defensive R techniques have been utilized where checkpoints have been implemented to fail fast (check data before next steps eg. Checking for duplicates, checking that all required data is included in all the files etc.), fail conspicuously (eg. returns descriptive error

messages during checkpoint failures), and fail appropriately (eg. Continuing even if datapoint returns NA, but still reporting NAs to the user).

One improvement for defensive R techniques not implemented is to involve creative failures. For example, continuing the analysis even if there is missing required data by replacing the missing data with a placeholder, then flagging the datapoint at the end.

Interpretation of Heatmaps

The heatmaps show the upregulation of certain genes after a treatment has been administered to the samples. The sample clustering in figure 1 shows that the samples with the same treatment groups are most related, which is expected. This can also be shown in the similarity of log mean gene counts between samples of the same treatment group.

Some genes immediately stand out as interesting due to the large differences in gene counts between treatment groups. For example, 1418756_at, 1451416_a_at, 1422607_at, 1427277_at, 1429525_s_at genes seems to be largely downregulated after treatment 4 compared to the others as shown by their relatively low gene counts compared to other treatment groups while 1436959_x_at, 1436736_x_at, 1456475_s_at genes all seem to be upregulated in treatment 4 only.

There does not appear to be a correlation between gene type and the gene count changes between treatments. This can be shown by the clustering of the genes, where gene types are not clustered together but appear to be dispersed between clusters. This can also be seen in how genes of the same gene type do not all follow a similar trend; not all genes of the same type all increase or decrease, but instead genes of the same gene type can either increase or decrease between the same treatment groups.

References

Love, M. I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology.* 15 (12), 550. 10.1186/s13059-014-0550-8.

Zhao, S., Ye, Z. & Stanton, R. (2020) Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.).* 26 (8), 903-909. 10.1261/rna.074922.120.