Differential gene analysis on SARS-COV-2 and SARS-COV infected cells

Introduction

Learning about the gene expression in SARS-COV-2 infected cells as opposed to mock cells can help provide insight as to which genes may be most impacted by the disease. Comparing the genes from SARS-COV-2 with SARS-COV allows us to get a better understanding of how the two differ in which genes they may affect, and some idea of how much. Many of the affected genes are shared however, both finding genes which are expressed differently between the two, and the degree of how greatly the gene counts vary can potentially give some new insight into SARS-COV-2.

Data

In this paper we use a Covid-19 Dataset containing gene counts of various samples under 3 main conditions. Samples labeled S2 are from SARS-COV-2 infected cells, samples labeled S1 are from SARS-COV infected cells, and mock samples are from mock infected cells. Measurements for both S1 and S2 conditions were taken at 4 hours, 12 hours and 24 hours post-infection. The mock infected cells only have measurements at 4 hours and 24 hours.

The Covid-19 Dataset used was provided in the following directory:
/mnt/biosoft/praktikum/genprakt/DiffAnalysis/Corona

Methods

After loading in the dataset the first step is to do some data wrangling and get the data into an understood format. In this case it required adjusting some mismatched column names and utilizing a DESeq container. This specialized container does checks for uniformity of the data up front and ensures that there should be no discrepancies later down the pipeline.

To analyze the differential gene expression between samples infected with SARS-COV-2 and SARS-COV several steps were taken.

Results

A first plot to consider is a histogram of p-values of the various samples calculated by DESeq. The background level at around 300-350 is indicative of the non-differentially expressed genes. The peak we see at the left hand side likely relates to differentially expressed genes. The peaks on the right hand side are likely due to genes with small counts. Both the hisat and star datasets follow the same pattern with only minute differences.

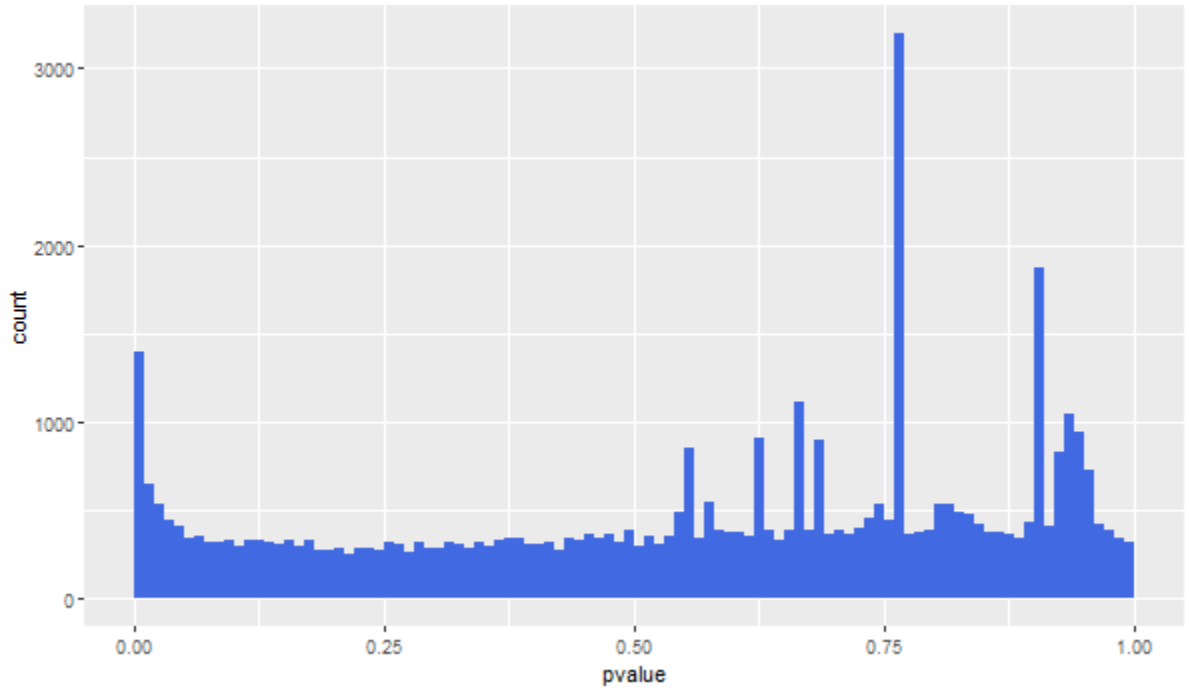

Figure 1: histogram (hisat)

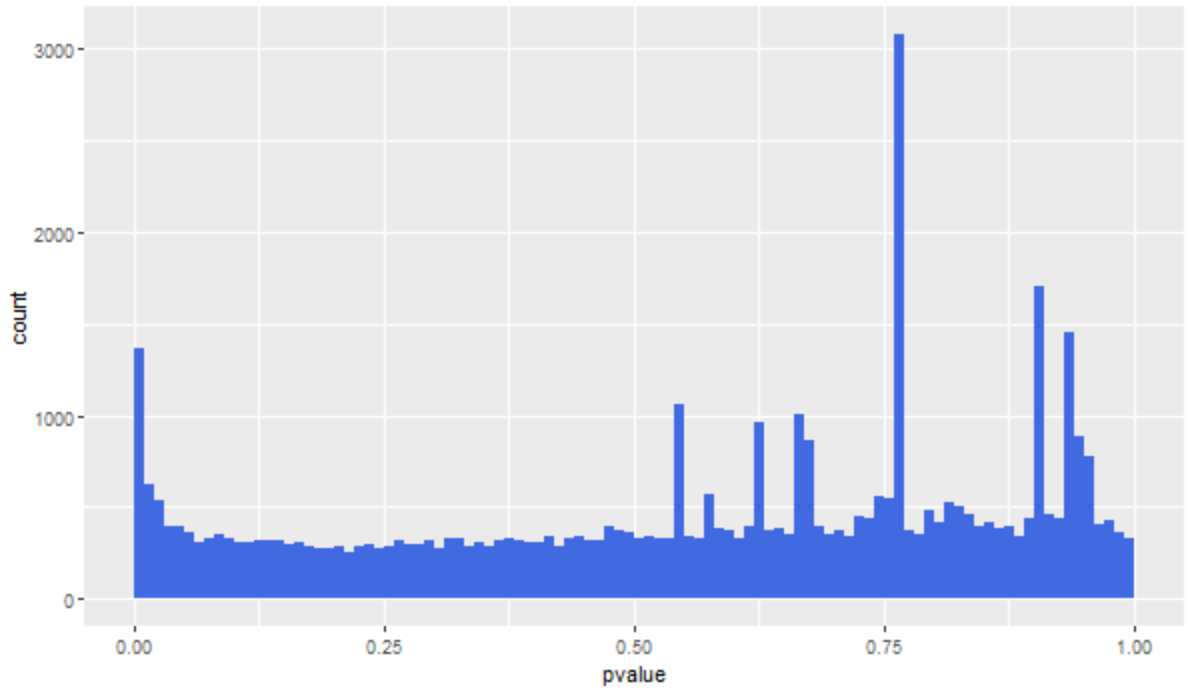

Figure 2: histogram (star)

Depicted in the MA plots are the mean of the normalized counts at various log fold changes. Marked in blue are those points which have a p value less than 0.1. The small arrows depicted towards the edges indicate points which fall off the y scale. Again both hisat and star look remarkably similar with few differences.

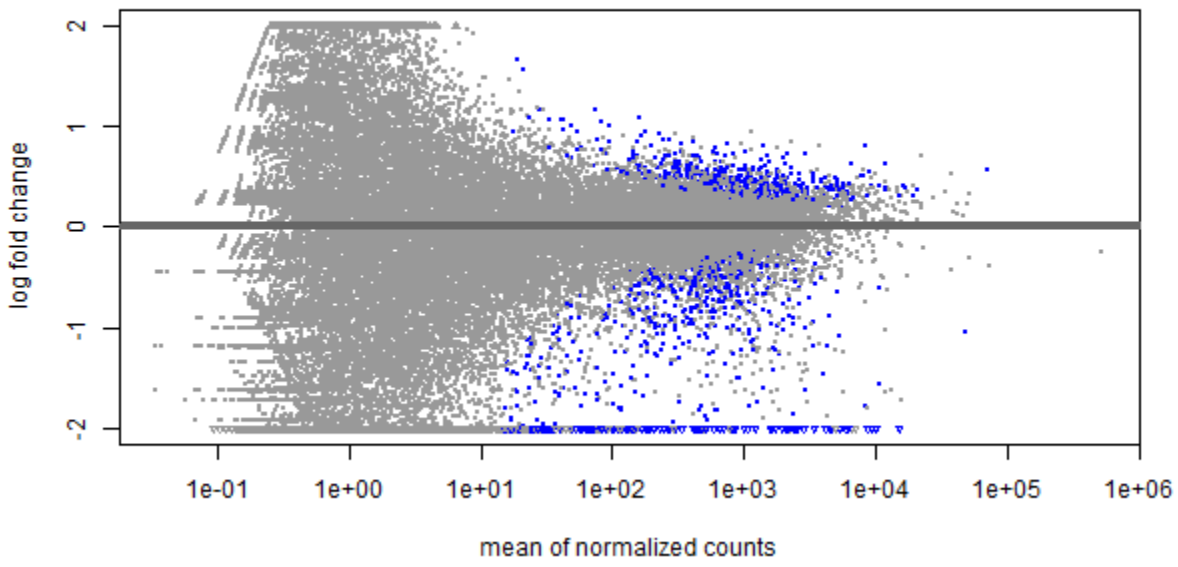

Figure 3: MA-plot (hisat)

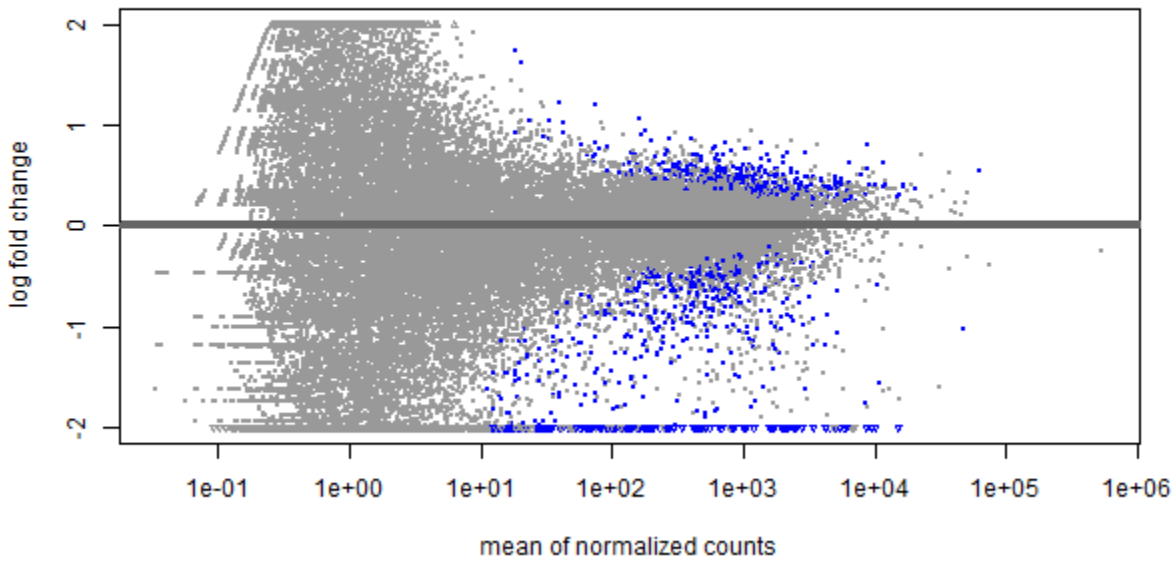

Figure 4: MA-plot (star)

The PCA analysis we conducted is shown below once in **Figure 5** split by each individual sample, and in **Figures 6** and **Figure 7** grouped into S1, S2 and mock. The S2 samples were more spread than the S1 samples.
The biggest apparent difference between hisat and star is visible in the rightmost S1 group.

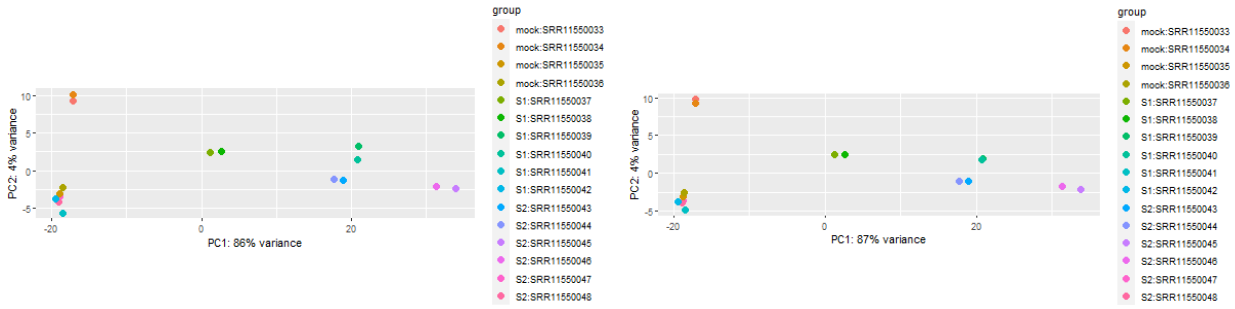

Figure 5: PCA (hisat), PCA (star) by sample

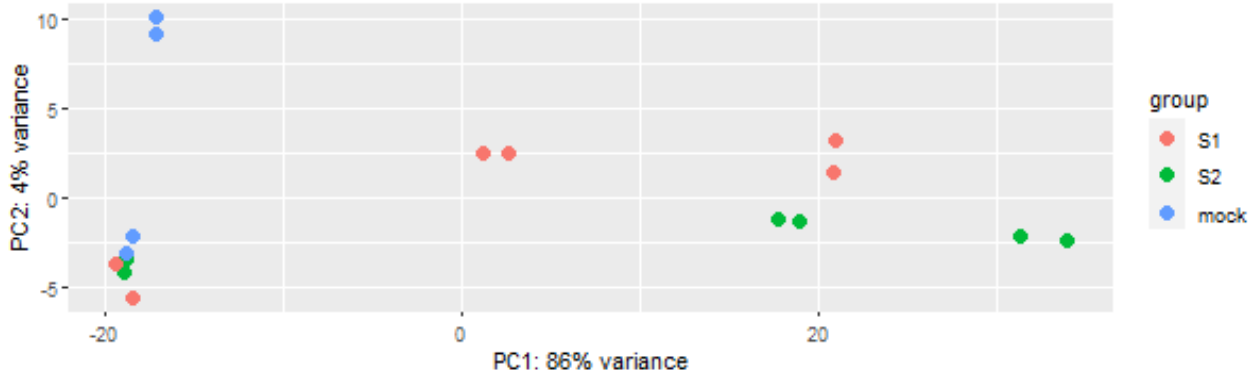

Figure 6: PCA(hisat) by condition

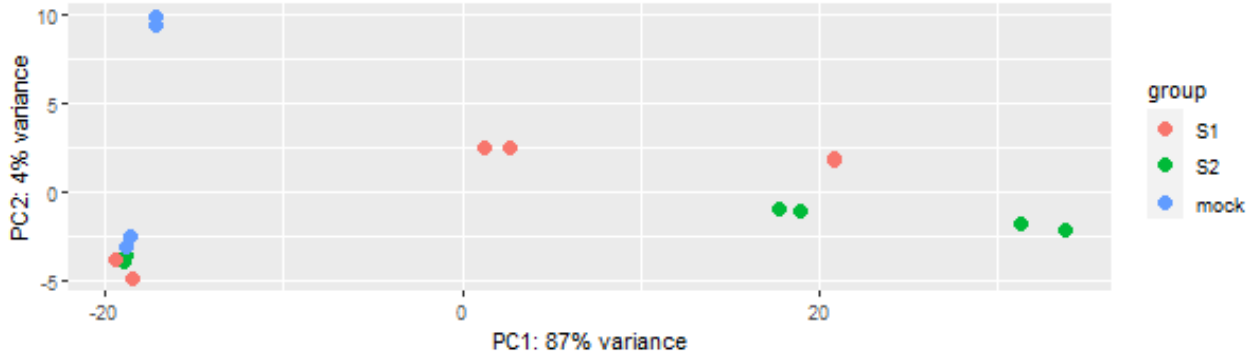

Figure 7: PCA(star) by condition

The heatmaps in **Figures 8** and **9** show regularized log transformed data with the various samples as columns and the top 30 genes as rows. The Samples are color coded to show what condition they belong to (S1,S2,mock). One can see the mock samples have values >2 which mark them as outliers.

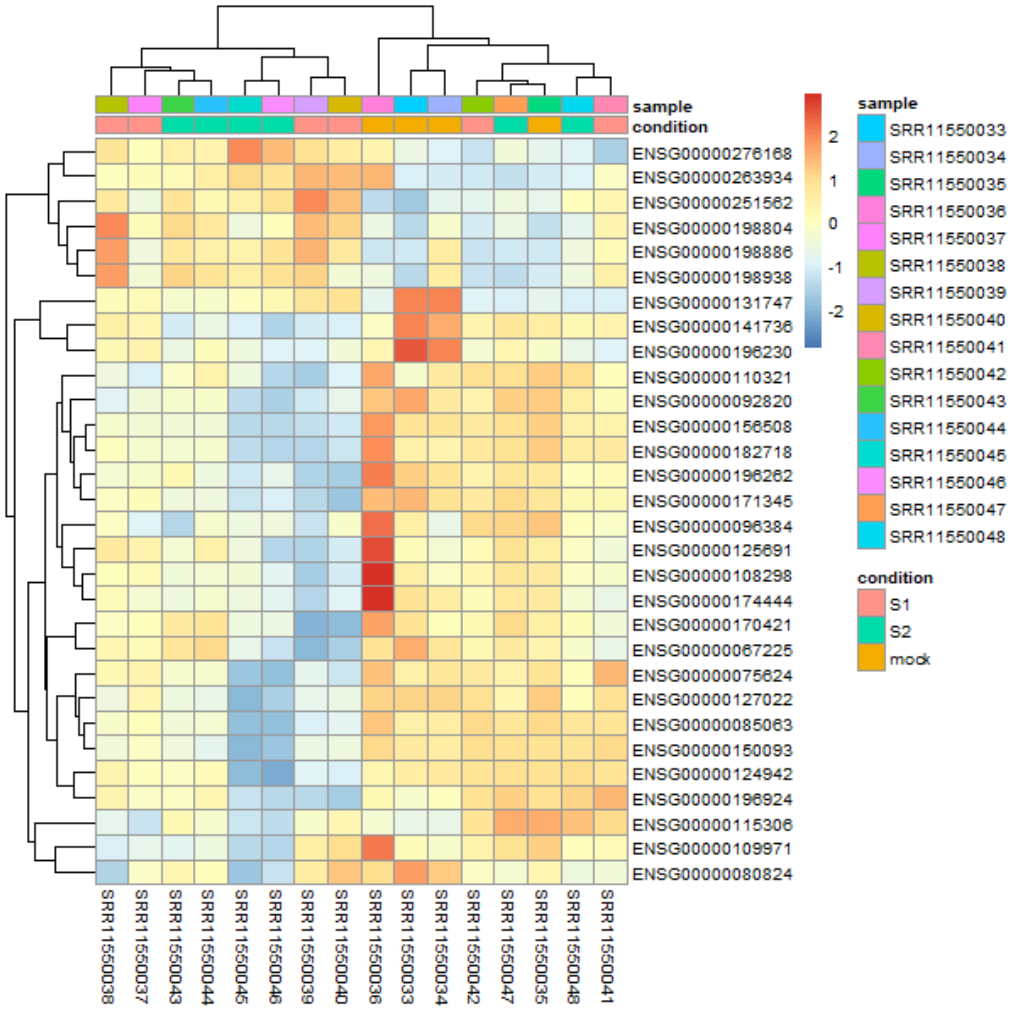

Figure 8: Heatmap (hisat)

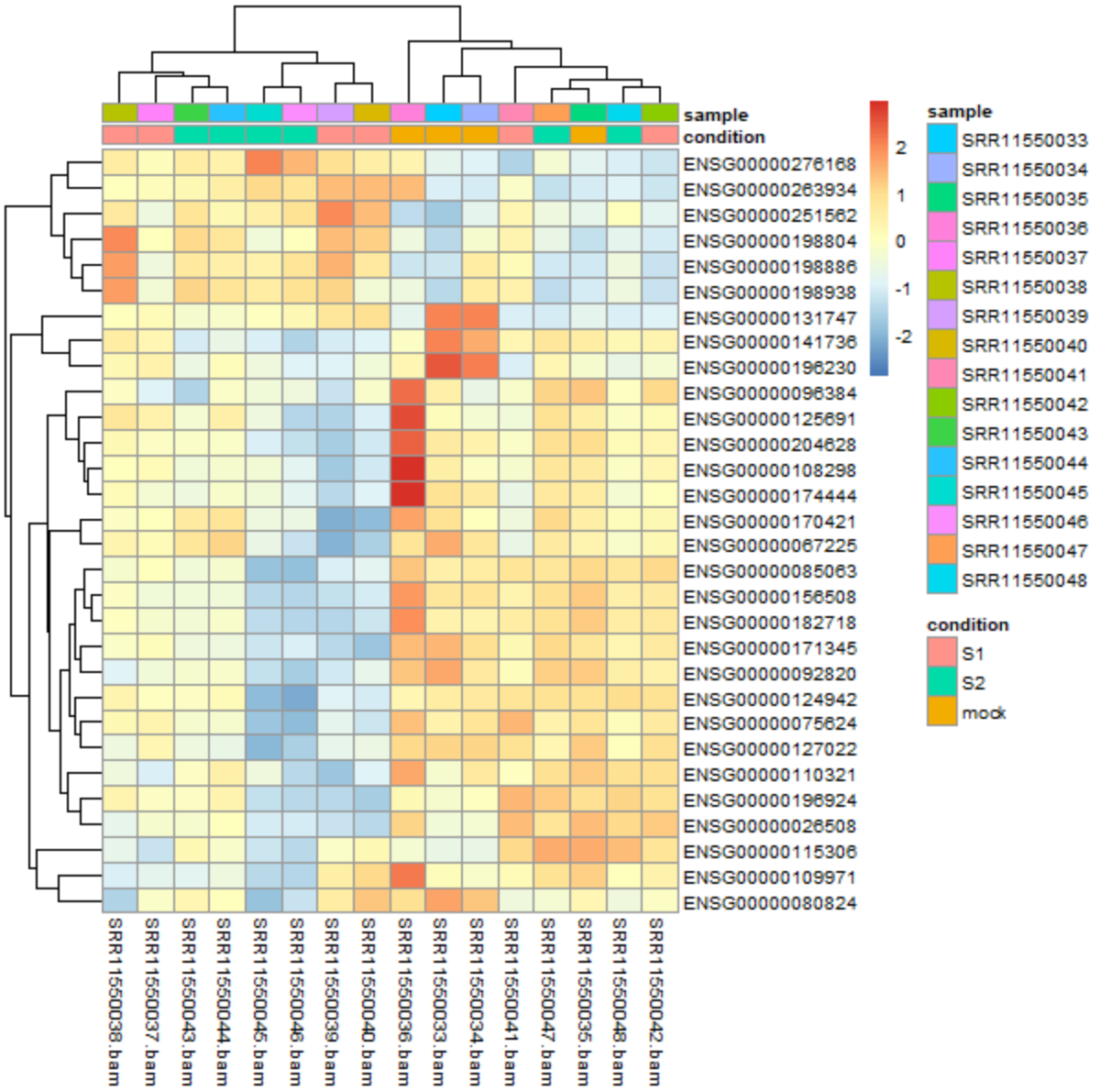

Figure 9: Heatmap (hisat)

References

Data: /mnt/biosoft/praktikum/genprakt/DiffAnalysis/Corona

R Packages:
dslabs,data.table,magrittr,tidyr.dplyr,ggplot2,BiocManager,ReportingTools,DESeq2,pheatmap

Additional references:
https://www.huber.embl.de/msmb/Chap-CountData.html
https://www.rdocumentation.org/