## Introduction

## Methodology

### Data composition

The research done by Emma and I is based on the data collected in Bangaladesh as part of a field trail.   
They gathered blood samples from a total of 5352 patients and household contacts. The household contacts are the participants who live in the same house, same compound and share the same kitchen or are direct neighbours.   
From all of the patients they took the first blood sample between April 2013 and April 2018. From this cohort, the contacts who were clinically diagnosed with leprosy between 4-61 month were considered progressors, which were 85 contacts (t=1). After which, a second blood sample was taken from the progressors before they started multidrug, which ended up being 84 progressors (t=2).  
From these 84 progressors, the samples from 40 household contacts, 40 progressors of the same patients from the first sample and the second sample are used to perform RNA sequencing analysis on. These were processed by the opensource BIOWDL RNA-Seq pipeline v2.0 developed at the Leiden Medical Centre.  
The processed data by this pipeline will be used for the differential gene expression. In the processed data are 3 files with the sample information, counts information and feature information.

* The sample information contains information about the sample, the connected samples of the same progressor or household contact, the type of leprosy, their birthyear, their gender and the amount of months between the collection of the first and second sample.
* The count information contains all the different gene counts for a gene and their specific sample.
* The feature information contains information related to the gene and in specific their type of RNA and their gene name as well as their ensemble ID

### Differential Gene Expression

**Normalization**

**Library Comparison**

**Results Selection**After performing the DGE analysis, the data should be filtered in order to gain the genes that are confident enough to be expected to be differentially expressed. Therefore, a p adjusted value below 0.05 was used. This value indicates that less than 5% of the significant results are likely to be false positives. In addition to the p adjusted value, another filter is selected, namely the Log 2 Fold change value. This value represents the magnitude of change in gene expression between the two conditions researched; household contacts and early progressors. The threshold value for this condition is the log2 of 1.5, which indicates that the gene expression between the two conditions should differ at least be 50% cases. This will eliminate possible genes that are insignificant due to their small difference between the two conditions.

## Results

### Venn Diagrams

After the DGE analysis and applying the filters that I just discussed, I found the following results.   
In the Venn Diagram from left to right you can see that the Venn Diagram of all of the genes is the combined version of the Venn Diagram of the coding genes and the non-coding genes.   
Another observation that can be made is that the difference between the 3 methods. The DESeq2 and EdgeR method are rather close in terms of total gene count. In addition the intersect between these two methods are 276 genes of the 435 genes or 63%, found by these 2 methods. As discussed in the methodology, these 2 methods are using a rather similar algorithm to decide which gene is differentially expressed. The found difference can therefore be explained by the normalization method in addition to the small differences in the algorithm.   
These 2 methods, however, do not have a lot in common with the method LimmaVoom. Yes there is overlap, but the total number of genes found by LimmaVoom is considerably less than the total number of genes found by the other two methods. Although LimmaVoom has the same normalization method as EdgeR, the algorithm of LimmaVoom is vastly different than EdgeR and DESeq2. The impact of these differences in algorithm and which is best for this dataset can be determined by running a machine learning algorithm and seeing the difference in accuracy or AUC score.

### MDS plot

In order to find the outliers in the samples used for machine learning, I used a Multidimensional Scaling plot also known as an MDS plot in order to identify them. As shown in the plot on the screen, you can see three outliers of which 1 was identified as a progressor on t=1, 1 was identified as a household contact and 1 was identified as a progressors on t=2. Since, we are interested in the difference between the gene expression between household contacts and the progressors at t=1m two samples were removed from the samples taken from the data used for machine learning.   
(ADD SAMPLE IDS TO PICTURE)