**Computational analysis of gene expression signatures of peripheral blood mononucleated cells (PBMCs) following infection of different viruses in *Macaca mulatta***

**Overview**

*Data analyzed*

The following report contains analyses of two datasets generated by gene expression profiling of peripheral blood mononucleated cells (PBMCs) in the Rhesus macaque (*Macaca mulatta*) following viral infection in the Messaoudi lab.

Datasets were generated by infecting macaques with two viruses:

* Ebola virus (*Zaire ebolavirus*), Makona strain. The data was generated by RNA-seq and consists of 30 samples generated by infection of 10 different animals, extracted at multiple timepoints between 0 and 6 days post infection (DPI).
* Yellow fever virus (gen. *Flavivirus*). The data was generated by microarray and consists of 12 samples generated by infection of 3 animals by two virus strains, by wild-type Yellow Fever Virus (wtYFV) and 17D-mutated virus, extracted at DPI 0 and 3. Data is publicly available on GEO under the accession GSE51972.

*Report aims*

The overarching goal of the report is to analyze the differential gene expression signatures induced by infection of Ebola and YFV in Macaque PBMCs.

The report is divided in two sections, one for each virus type. Each section addresses the following specific aims:

1. Investigate sample separation in an unbiased manner by visualizing the dataset using Principal Components Analysis (PCA) and interactive heatmaps.
2. Calculate differential gene expression signatures generated by virus infection, comparing gene expression levels at different timepoints to the pre-infection levels.
3. Characterize the gene expression signatures by calculating the enrichment of relevant biological pathways and terms.
4. Identify small molecules which reproducibly mimic or reverse the observed gene expression signatures, by querying a small molecule perturbation database.
5. Investigate the impact of viral infection on clinical blood parameters, by investigating their variation over time and their correlation to gene expression levels.

**Results**

**1. Ebola Makona**

**Overview of the dataset**

The Ebola Makona infection dataset consists of 30 RNA-seq samples generated by infection of 10 animals followed by PBMC sampling at multiple timepoints between 0 and 6 DPI. The samples are divided in the following timepoints: 10 samples before infection (day 0); 5 from day 1; 4 from day 2; 3 from day 3; 2 for day 4; 2 for day 5; 4 from day 6.

**1.1 PCA Analysis**

To visualize the dataset and assess separation of samples, a PCA analysis was performed on the dataset.

*Figure 1* | PCA analysis of the Ebola RNA-seq dataset. The plot displays a projection of the gene expression dataset in a 3-dimensional plane, where each point represents a different sample. Axes represent a certain fraction of the variance observed in the dataset, and account for independent sources of variation in the data. Point colors correspond to different timepoints of infection, measured in DPI. Principal Component 1 (PC1), which accounts for the largest fraction of variance in the dataset, shows a strong association to the sample’s timepoint in DPI.

**1.2 Interactive heatmap**

To interactively visualize the dataset, an interactive heatmap was generated using Clustergrammer (http://amp.pharm.mssm.edu/clustergrammer/).

*Figure 2* | Interactive visualization of the expression levels of the most variable genes.  The tool allows to interactively explore the expression of the top 1000 most variably expressed genes (displayed on rows) across the 30 RNA-seq samples (displayed on columns). Color bars above the columns show the sample’s timepoint in DPI. The tool allows to identify clusters of coexpressed genes, and to calculate enriched biological terms using Enrichr (Chen et al., 2013).

**1.3 Differential gene expression analysis**

To investigate the effect of viral infection on the transcriptome of PBMCs, differential gene expression signatures were calculated by comparing gene expression levels at different timepoints to the one at the zero timepoint using the Characteristic Direction method (Clark et al., 2014). The method requires a minimum of 3 samples per timepoint; as the day 4 and 5 timepoints have 2 samples each, they were merged in a single ‘day4-5’ timepoint.

*Figure 3* | Interactive visualization of the differential expression levels of the most variably differentially expressed genes. Columns represent differential gene expression signatures calculated by comparing samples of the corresponding timepoint to the 0 timepoint. The heatmap reveals the presence of two main clusters of genes, distinguished by patterns of overexpression in early timepoints (days 1 and 2) or in later timepoints (days 3 to 6).

**1.4 Enrichment analysis**

To investigate the biological pathways associated to the differentially expressed genes, an enrichment analysis was performed using Enrichr.

*1.4.1* *Early and late genesets*

The first section of the analysis aims to broadly characterize the differential gene expression patterns across timepoints by calculating the enrichment in the two clusters of genes identified in Figure 3, labeled *early* and *late* genesets.

*Figure 4* | Selected enrichment results for the early and late genesets induced by Ebola infection. The figure displays some of the terms which are most significantly associated to the genesets, as ranked by Enrichr.

*1.4.2 Timepoint analysis*

The second section of the analysis aims to characterize the enrichment of gene expression signatures for each individual timepoint. For each signature, the top 500 most overexpressed and underexpressed genes were analyzed.

Full enrichment results are available below:

*Figure 5* | Enrichment across the upregulated genesets of separate timepoints for selected terms. The plot displays the combined enrichment score, as calculated by Enrichr, of two of the most significantly associated pathways.

Full enrichment results are available below:

**1.5 Small molecule profiling**

To identify small molecules which can mimic or reverse the observed differential gene expression signatures when used for treatment on cell lines, the L1000 small molecule perturbation database was queried using the L1000CDS2 tool (Duan et al., 2016).

For each timepoint, the top 50 most similar and 50 most opposite signatures were extracted, and the names of the small molecules which generated them were extracted. For each timepoint, the number of times each small molecule appears within the top 50 signatures was calculated and plotted below.

*Figure 6* | Summary of the most relevant small molecules which mimic or reverse Ebola infection signatures. The barchart represents the number of times each molecule appears within the top 50 most similar or opposite signatures. Colors represent different timepoints. Some of the identified small molecules have been reported to be involved in immune modulation and inhibition of viral replication, such as withaferin-A and emetine.

Full results are available below:

**1.6 Analysis of clinical data**

To explore the impact of Ebola infection on clinical blood parameters, the variation of their levels across timepoints were investigated. In order to provide further insight into this relationship, the association between blood parameter levels and gene expression levels was also investigated.

*1.6.1 Timepoint analysis*

The first section of the analysis aims to investigate the variation of clinical blood parameters across timepoints of viral infection. To facilitate visualization, samples were divided in three groups according to time: early timepoints (days 0 and 1, n=15), intermediate timepoints (days 2 to 4, n=7), and late timepoints (days 5 and 6, n=6).

*Figure 6 |* Variation of blood parameters across timepoints of viral infection. The boxplots display the levels of clinical blood parameters across three groups of samples.

*1.6.2* *Correlation to gene expression*

The second section of the analysis aims to investigate the association between clinical blood parameters and gene expression. To achieve this, the correlation between gene expression levels and blood parameter levels across the 30 RNA-seq samples were measured using Spearman’s index.

*Figure 7* | Network displaying correlation between gene expression and clinical blood parameter levels. Red nodes represent blood parameters, blue nodes represent genes, edges represent correlations and are colored by sign (red indicates positive correlation, blue indicates negative correlation). Correlations above R=0.6 are displayed. Note that the geneset shared between GR (granulocytes) and LY (lymphocytes) show opposite correlation signs.

**2. Yellow Fever**

**Overview of the dataset**

The Yellow Fever infection dataset consists of 12 microarray samples generated by infection of 3 animals by two virus strains, wild-type Yellow Fever Virus (wtYFV) and 17D-mutated strain, followed by PBMC sampling before infection and at 3 DPI.

While displaying few genetic differences, the two viral strains produce distinctly different disease outcomes in human and monkey hosts. To investigate the biological effects of the wild-type strain, only samples infected by wtYFV were used in the analysis.

**1.1 PCA Analysis**

To visualize the dataset and assess separation of samples, a PCA analysis was performed on the dataset.

*Figure 8* | PCA analysis of the Ebola RNA-seq dataset. The plot displays a projection of the gene expression dataset in a 3-dimensional plane, where each point represents a different sample. Axes represent a certain fraction of the variance observed in the dataset, and account for independent sources of variation in the data. Point colors correspond to different timepoints of infection, measured in DPI. The combination of Principal Components show a discrete separation between samples pre-infection and at day 3 post-infection.

**1.2 Interactive heatmap**

To interactively visualize the dataset, an interactive heatmap was generated using the Clustergrammer tool.

*Figure 9* | Interactive visualization of the expression of the most variable genes.  The tool allows to interactively explore the expression of the top 1000 most variably expressed genes (displayed on rows) across the 6 microarray samples (displayed on columns). Color bars above the columns show the sample’s treatment type, timepoint of sampling, and replicate number. The tool allows to identify clusters of coexpressed genes, and to calculate enriched biological terms using Enrichr.

**1.3 Differential gene expression analysis**

Differential gene expression signatures were calculated by comparing gene expression levels at day 3 post infection to the pre-infection levels. To achieve this, the Characteristic Direction (CD) method was used on normalized expression data.

*Figure 10* | Summary of the differential expression results for infection of wtYFV. The X axis displays mean gene expression, the Y axis displays the differential gene expression signature, points represent genes. Positive CD values indicate overexpression in infected samples, negative values indicate underexpression.

**1.4 Enrichment analysis**

To investigate the biological pathways associated to the most upregulated and downregulated genes, an enrichment analysis was performed using Enrichr.

*Figure 11* | Selected enrichment results for the top 500 most overexpressed and underexpressed genes induced by wtYFV infection. The figure displays some of the terms which are most significantly associated to the genesets, as ranked by Enrichr.

**1.5 Small molecule profiling**

To identify small molecules which can mimic or reverse the observed differential gene expression signatures when used for treatment on cell lines, the L1000 small molecule perturbation database was queried using the L1000CDS2 tool.

The top 50 most similar and 50 most opposite signatures were extracted, and the names of the small molecules which generated them were extracted. Following this, the number of times each small molecule appears within the top 50 signatures was calculated and plotted below.

*Figure 12* | Summary of the most relevant small molecules which mimic or reverse wtYFV infection signature. The barchart represents the number of times each molecule appears within the top 50 most similar or opposite signatures. Some of the small molecules which reverse the signature have been reported to be involved in inhibition of viral replication, such as narciclasine and emetine.

Full results are available below:

**1.6 Analysis of clinical data**

To explore the impact of wtYFV infection on clinical blood parameters, the variation of their levels across timepoints were investigated. In order to provide further insight into this relationship, the association between blood parameter levels and gene expression levels was also investigated.

*1.6.1 Timepoint analysis*

The first section of the analysis aims to investigate the variation of clinical blood parameters across timepoints of viral infection. Samples were divided in three groups according to time: pre-infection (Day0, n=3) and day 3 post-infection (Day3, n=3).

*Figure 6 |* Variation of blood parameters across timepoints of viral infection. The boxplots display the levels of clinical blood parameters across three groups of samples.

*1.6.2* *Correlation to gene expression*

The second section of the analysis aims to investigate the association between clinical blood parameters and gene expression. To achieve this, the correlation between gene expression levels and blood parameter levels across the microarray samples were measured using Spearman’s index.

*Figure 7* | Network displaying correlation between gene expression and clinical blood parameter levels. Red nodes represent blood parameters, blue nodes represent genes, edges represent correlations and are colored by sign (red indicates positive correlation, blue indicates negative correlation). Correlations above R=0.7 are displayed.

**Methods**

**Data Preprocessing**

**Ebola Data**

FASTQ files containing single-end, 100 base pair reads were generated and kindly provided by Ilhem Messaoudi’s lab. Kallisto (Bray et al., 2016) was used to align the reads to the latest Ensembl *Macaca mulatta* transcriptome build (Mmul\_8.0.1.cdna.all) and to quantify transcript expression. Total gene expression was calculated by summing the readcounts of corresponding transcripts. Raw readcount data was normalized using the Variance Stabilizing Transformation method, available in the DESeq R package (Love et al., 2014).

**Yellow Fever Dataset**

Raw expression data was downloaded from the GEO series GSE51972. Probes were matched to gene symbols using the platform annotation file for GPL3535 downloaded from GEO. Expression levels of genes with multiple matching probes were calculated by averaging the expression of corresponding probes. Expression data was then normalized using the Quantile Normalization method (Bolstad et al., 2003).

**Data Analysis**

**Principal Components Analysis**

An R implementation of the Principal Components Analysis (PCA) algorithm was used to generate the three-dimensional visualizations of samples. The analysis was performed by taking a subset of the top 5000 most variables genes in each dataset.

**Differential expression analysis**

Differential gene expression signatures were calculated by applying the Characteristic Direction method (Clark et al., 2014) on normalized gene expression data. The signatures were generated by comparing samples from each timepoint to the pre-infection timepoint.

**Enrichment and small molecule identification**

Enrichment analysis was performed on the top 500 most overexpressed and underexpressed genesets for each signature using Enrichr (Chen et al., 2013). Small molecule identification was performed on the complete gene expression signatures using L1000CDS2 (Duan et al., 2016).

**Clinical data analysis**

Associations between clinical parameters and normalized gene expression levels were calculated across each samples by Spearman’s index. Network visualization was generated with Cytoscape (Shannon et al., 2003). Displayed networks show the most significant correlations between genes and clinical parameters; a cutoff of 0.6 and 0.7 for Spearman R was used respectively for the Ebola and Yellow Fever datasets.

**References**

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma. Oxf. Engl. *19*, 185–193.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma’ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

Clark, N.R., Hu, K.S., Feldmann, A.S., Kou, Y., Chen, E.Y., Duan, Q., and Ma’ayan, A. (2014). The characteristic direction: a geometrical approach to identify differentially expressed genes. BMC Bioinformatics *15*, 79.

Duan, Q., Reid, S.P., Clark, N.R., Wang, Z., Fernandez, N.F., Rouillard, A.D., Readhead, B., Tritsch, S.R., Hodos, R., Hafner, M., et al. (2016). L1000CDS2: LINCS L1000 characteristic direction signatures search engine. Npj Syst. Biol. Appl. *2*, 16015.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

http://amp.pharm.mssm.edu/clustergrammer/ Clustergrammer.