**Computational analysis of gene expression signatures induced by HIV infection in patient-derived podocyte cells**

**Overview**

*Data analyzed*

The following report contains analyses of two datasets generated by gene expression profiling of human podocyte cells, extracted from HIVAN patients (??), following infection by HIV. Data was generated in Dr. Paul Klotman’s lab.

Datasets were generated by infection of two different cell types:

* Cell lines derived from podocytes. The data was generated by RNA-seq and consists of 13 samples, analyzed before infection and at multiple timepoints after HIV infection (6h, 12h, 24h, 48h). Samples were processed in three separate batches.
* Primary podocytes extracted from HIVAN patients (???). The data was generated by RNA-seq and consists of 19 samples from 10 different patients, analyzed before infection and at (timepoint???) after infection by HIV, or by a control GFP virus.

*Report aims*

The overarching goal of the report is to analyze and compare the differential gene expression signatures induced by infection of HIV in podocytes.

The report is divided in three sections: sections 1 and 2 concern individual analyses of each dataset, section 3 compares the results across datasets . 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.

**Results**

**1. Cell Lines**

**Overview of the dataset**

The cell line dataset consists of 13 RNA-seq samples generated by infection of podocyte-derived cell lines in three batches. Samples were analyzed before infection and at multiple timepoints (6h, 12h, 24h, 48h). Each group has the following number of samples: 3 for pre-infection controls, 3 for 6h timepoint, 2 for 12h timepoint, 2 for 24h timepoint, 3 for 48h timepoint.

**1.1 PCA Analysis**

*Original Data*

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

*Figure 1* | PCA analysis of the podocyte cell line 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 batches. The plot reveals a very strong batch effect, indicating that the variance between samples in the dataset is largely due to the batches, rather than the timepoint of HIV infection.

*Batch effect removal*

In order to remove the batch effects in the dataset, the batch effect removal algorithm ComBat (Leek et al., 2012) was used. The PCA analysis was repeated on the corrected dataset, to assess the effect of the algorithm on sample separation

*Figure 2* | PCA analysis of the podocyte cell line RNA-seq dataset processed by ComBat. Point colors correspond to batches. The plot shows the algorithm efficiently removed sample batch effects, while recovering the separation due to the timepoint of infection, to an extent.

**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). Values represent gene expression levels, transformed by Z-scores 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 podocytes, 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 12h and 24h timepoints have 2 samples each, they were merged in a single ’12-24h’ 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.

**1.4 Enrichment analysis**

To investigate the biological pathways associated to the differentially expressed genes, an enrichment analysis was performed using Enrichr. A summary of the most significantly enriched terms across the signatures is shown in the heatmaps below.

*Figure 3* | Interactive visualization of the most significantly enriched pathways in the upregulated and downregulated genes following HIV infection at multiple timepoints. Columns represent differential gene expression signatures calculated by comparing samples of the corresponding timepoint to the pre-infection timepoint, rows represent biological terms, and values represent enrichment scores as calculated by Enrichr (log10P).

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 HIV 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 mitoxantrone (Carlson et al., 2000), QL-XII-47 (de Wispelaere et al., 2017), and emetine (Chaves Valadão et al., 2015).

Full results are available below:

**2. Primary Podocytes**

**Overview of the dataset**

The primary podocyte dataset consists of 19 RNA-seq samples generated by infection of patient-derived primary podocytes by HIV and GFP-control viruses. Samples were analyzed before infection and at (timepoint???) after infection. Each group has the following number of samples: 6 pre-infection controls, 6 HIV-infected samples, 6 GFP-infected samples.

**1.1 PCA Analysis**

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

*Figure 9* | PCA analysis of the primary podocyte 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 treatment. The plot reveals a poor separation between samples generated by the same experimental conditions.

*Batch effect removal*

In order to remove the batch effects in the dataset, the batch effect removal algorithm ComBat (Leek et al., 2012) was used. The PCA analysis was repeated on the corrected dataset, to assess the effect of the algorithm on sample separation

*Figure 10* | PCA analysis of the primary podocyte RNA-seq dataset processed by ComBat. Point colors correspond to batches. The plot shows the algorithm efficiently rescues sample separation based on experimental conditions.

**1.2 Interactive heatmap**

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

*Figure 11* | 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). Values represent gene expression levels, transformed by Z-scores 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 samples from different experimental conditions. To achieve this, the Characteristic Direction (CD) method was used on normalized expression data.

*Figure 12* | 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.

**1.4 Enrichment analysis**

To investigate the biological pathways associated to the most upregulated and downregulated genes induced by HIV infection, an enrichment analysis was performed using Enrichr. The HIV infection signature used in all following analyses is generated by comparing HIV-infected samples to GFP-infected control samples.

*Figure 12* | Selected enrichment results for the top 500 most overexpressed and underexpressed genes induced by HIV infection, when compared to GFP control 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 13* | Summary of the most relevant small molecules which mimic or reverse HIV 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 mitoxantrone (Carlson et al., 2000), QL-XII-47 (de Wispelaere et al., 2017), and emetine (Chaves Valadão et al., 2015).

Full results are available below:

**3. Identification of Common Pathways and Drugs**

The final part of the analysis aims to uncover pathways and small molecules which are associated to HIV infection signatures across the two datasets.

**1.1 Common Pathways**

In order to identify pathways which are common to both signatures, the enrichment results for corresponding genesets of each signature were compared. The analysis revealed that the Cell Line 48h post-infection timepoint shared the largest amount of enriched terms to the Primary Podocyte HIV-infection vs GFP-infection signature.

*Figure 14* | Comparison of enrichment results of the HIV infection signatures computed in the cell line dataset and the podocyte dataset. Points represent biological terms or pathways, axes represent enrichment scores (calculated by –log10p), the red lines represent significance threshold of FDR = 0.05. Panel A indicates that the upregulated genes across the two signatures share enrichment for terms involved in TNF signaling and response to virus; whereas Panel B indicated that the downregulated genes across the two signatures share enrichment for targets of SUZ12, EED, KDM2B, and other key members of the Polycomb repressive complex.

**1.2 Common Molecules**

In order to identify small molecules which reproducibly mimic or reverse both signatures, the results of the L1000CDS2 analyses for each signature were compared.

*Figure 15* | Summary of the small molecules which most reproducibly mimic (red) or reverse (blue) the different HIV infection signatures across datasets. Bars correspond to different HIV infection signatures, Y axis indicates how often each small molecule appears within the top 50 most similar (or opposing) signatures, subplots indicate different small molecules or drugs. Many of these have been reported to be involved in the inhibition of viral replication, such as mitoxantrone (Carlson et al., 2000), QL-XII-47 (de Wispelaere et al., 2017), and emetine (Chaves Valadão et al., 2015).

**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 Preprocessing**

Raw readcount expression for the cell line and primary podocyte datasets was provided by Dr. Weijia Zhang. Raw data was normalized using the Variance Stabilizing Transformation method, available in the DESeq R package (Love et al., 2014).

**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).

**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.

Carlson, H.A., Masukawa, K.M., Rubins, K., Bushman, F.D., Jorgensen, W.L., Lins, R.D., Briggs, J.M., and McCammon, J.A. (2000). Developing a dynamic pharmacophore model for HIV-1 integrase. J. Med. Chem. *43*, 2100–2114.

Chaves Valadão, A.L., Abreu, C.M., Dias, J.Z., Arantes, P., Verli, H., Tanuri, A., and de Aguiar, R.S. (2015). Natural Plant Alkaloid (Emetine) Inhibits HIV-1 Replication by Interfering with Reverse Transcriptase Activity. Mol. Basel Switz. *20*, 11474–11489.

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.

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

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinforma. Oxf. Engl. *28*, 882–883.

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.

de Wispelaere, M., Carocci, M., Liang, Y., Liu, Q., Sun, E., Vetter, M.L., Wang, J., Gray, N.S., and Yang, P.L. (2017). Discovery of host-targeted covalent inhibitors of dengue virus. Antiviral Res. *139*, 171–179.