**HEU Manuscript Analysis: README**

**Author: Nelly Amenyogbe**

**21-July-2020**

Table of Contents

[Introduction 2](#_Toc46919060)

[Data file descriptors 2](#_Toc46919061)

[Luminex Data 2](#_Toc46919062)

[Flow Cytometry Data 2](#_Toc46919063)

[Microbiome data 3](#_Toc46919064)

[Metadata 3](#_Toc46919065)

[R\_export 3](#_Toc46919066)

[Analytical script descriptors: 4](#_Toc46919067)

[Luminex analysis 4](#_Toc46919068)

[**scripts/luminex\_cytokines/ms\_lmx\_pca.R** 4](#_Toc46919069)

[**scripts/luminex\_cytokines/ms\_luminex\_lm.R** 4](#_Toc46919070)

[**scripts/luminex\_cytokines/ms\_luminex\_splsda.R** 4](#_Toc46919071)

[Flow Cytometry Analysis 4](#_Toc46919072)

[**scripts/flow\_cytometry/ms\_flow\_population\_lm.R** 4](#_Toc46919073)

[**scripts/flow\_cytometry/ms\_flow\_icc\_lm.R** 4](#_Toc46919074)

[Microbiome Analysis 5](#_Toc46919075)

[**scripts/microbiome/ms\_microbiome\_alpha\_diversity.R** 5](#_Toc46919076)

[**scripts/microbiome/ms\_microbiome\_ordination.R** 5](#_Toc46919077)

[**scripts/microbiome/ms\_microbiome\_differential\_abundance.R** 5](#_Toc46919078)

[DIABLO 5](#_Toc46919079)

[**scripts/DIABLO/data\_preparation/ms\_luminex\_diablo\_preparation.R** 6](#_Toc46919080)

[**scripts/DIABLO/data\_preparation/ms\_otu\_diablo\_preparation.R** 6](#_Toc46919081)

[**scripts/DIABLO/data\_preparation/ms\_mlx\_otu\_data\_matrix\_preparation.R** 6](#_Toc46919082)

[**scripts/DIABLO/cohort\_models/ms\_DIABLO\_blg.R** 6](#_Toc46919083)

[**scripts/DIABLO/cohort\_models/ms\_DIABLO\_cad.R** 6](#_Toc46919084)

[**scripts/DIABLO/cohort\_models/ms\_DIABLO\_saf.R** 6](#_Toc46919085)

[**scripts/DIABLO/cohort\_models/splsda\_error\_comparison.R** 7](#_Toc46919086)

# **Introduction**

This folder contains all raw numerical data presented in the manuscript:

**Innate immune responses and gut microbiomes distinguish HIV exposed from unexposed children in a population-specific manner**

The HEU Manuscript Analysis directory contains all input data, scripts, and figures associated with the study manuscript. The data is organized into three main folders:

* **Rdata:** Contains all data utilized for manuscript analysis, in assay-type specific folders. Intermediate files generated by the raw input data, that are then passed into later analyses, are saved in the **R\_export**directory.
* **scripts:** This directory contains all R scripts associated with analyses performed for this manuscript. Each script can be run independently of others. Where intermediate files are required that do not include the raw input data, these have been saved in Rdata/R\_export and are appropriately refereced in each script
* **figures:** This directory contains all figures used to generate figures for the manuscript, in data-type specific folders

# **Data file descriptors**

This section pertains to all the folders located in the Rdata directory. The data files are further explained below:

## **Luminex Data**

* Rdata/luminex/gc\_heu\_luminex.csv: Luminex cytokine data. This data is in long format, with each cytokine in response to each stimulus for each individual in a separate row. The final concentration values were fitted to the standard curve for the corresponding analyte and raw, non-normalized final concentration values are included in this file.
* Rdata/luminex/ heu\_pca\_colours.csv: Colour codes used for HEU and HUU children within each respective site.

## **Flow Cytometry Data**

* Rdata/flow/gc\_heu\_fct\_icc.csv: Flow cytometry, intracellular cytokine data. This data is in long format, with each cytokine for each cell type for each individual in a separate row. Values correspond to the percent of each indicated cell type gated as positive for the indicated cytokine.
* Rdata/flow/gc\_heu\_fct\_count.csv: Flow cytometry, count data. This data is in wide format, with the count of each cell type of interest in a separate column. This data was used to calculate the percent of total cells (cells in the size gate) presented in Supplementary Figure 4.

## **Microbiome data**

* Rdata/microbiome/gc\_heu\_phyloseq.rds: Microbiome data is supplied as a phyloseq object. To access the data, R statistical software must be installed. The latest version can be found here:

<https://cran.r-project.org/bin/windows/base/>

R studio is a user-friendly tool to facilitate the use of R, and can be downloaded here:

<https://rstudio.com/products/rstudio/download/>

The Phyloseq R package must be installed to load the phyloseq object. Instructions to download phyloseq can be found here:

<https://bioconductor.org/packages/release/bioc/html/phyloseq.html>

* Rdata/microbiome/ gc\_heu\_otu\_table.csv: OTU count table, identical to data embedded in the phyloseq object
* Rdata/microbiome/gc\_heu\_otu\_taxonomy.csv: Taxonomic identifiers of OTUs listed in the OTU table, identical to data embedded in the phyloseq object
* Rdata/microbiome/ gc\_phylum\_cols.csv: Color codes used to plot phylum-level identifiers

## **Metadata**

* Rdata/metadata/gc\_heu\_metadata.csv: all study participant metadata utilized throughout the manuscript
* Rdata/metadata/gc\_heu\_data\_dictionary.xlsx: Excel file with explanations for each of the variables in all the data files listed above
* ms\_HEU\_session\_info.docx: R markdown export of all package versions used for analysis. The Rmd file is located under scripts/ms\_HEU\_session\_info.Rmd

## **R\_export**

This directory contains all intermediate data files generated using analytical scripts, which are applied to a later point in the analysis. The generation of these files is indicated in in the “Output files” section for each respective script.

# **Analytical script descriptors:**

## **Luminex analysis**

### **scripts/luminex\_cytokines/ms\_lmx\_pca.R**

Principal components analysis of luminex cytokine data and Adonis test for sample clustering

* **MS figures generated**: Figure 1, Supplementary Figure 1, Supplementary Figure 2
* **Output files**: none
* **Output figures (figures/luminex\_cytokines)**:
  + PCAs by site: blg\_lmx\_pca, cad\_lmx\_pca, saf\_lmx\_pca
  + PCA by stimulus: heu\_pca\_lps, heu\_pca\_pam, heu\_pca\_pgn, heu\_pca\_pic, heu\_pca\_r848
  + PCA legends: heu\_pca\_stim\_legend, heu\_pca\_unstim\_legend
  + PCA of all stimuli:heu\_pca\_stims
  + Adonis results: pca\_ad\_res, pca\_ad\_res\_stim
  + PCAs for LPS only: lps\_pcas,

### **scripts/luminex\_cytokines/ms\_luminex\_lm.R**

Linear regression to determine effect of HIV exposure on cytokine responses

* **MS figures generated:** Figure 2
* **Output files:** none
* **Output figures (figures/luminex\_cytokines):**
  + Boxplot Canadian Cohort: cad\_lm\_c, cad\_lm\_h
  + Boxplot Belgian Cohort: blg\_lm\_h

### **scripts/luminex\_cytokines/ms\_luminex\_splsda.R**

sPLS-DA analysis for luminex cytokine data

* **MS figures generated: None, precursor to Figure 6**
* **Output files:** HEU\_manuscript\_analysis/Rdata/R\_export/heu\_lmx\_perf.csv
* **Output figures:** none

## **Flow Cytometry Analysis**

### **scripts/flow\_cytometry/ms\_flow\_population\_lm.R**

Linear regression for flow cytometry cellular populations

* **MS figures generated:** Supplementary Figure 3
* **Output files:** none
* **Output figures (figures/flow\_cytometry):**
  + Boxplot for cellular populations: gcHEU\_flow\_populations

### **scripts/flow\_cytometry/ms\_flow\_icc\_lm.R**

Linear regression for flow cytometry intracellular cytokine staining

* **MS figures generated:** Figure 3
* **Output files:** none
* **Output figures (figures/flow\_cytometry):**
  + Boxplot for IL-12 responses by cDC and monocytes: cad\_heu\_flow\_il12.pdf
  + Boxplot for other significant ICC findings: cad\_heu\_flow\_others.pdf

## **Microbiome Analysis**

### **scripts/microbiome/ms\_microbiome\_alpha\_diversity.R**

Microbiome alpha diversity computed via the Shannon index, and associated Wilcoxon test to determine if alpha diversity differs between HEU and HUU children

* **MS Figures Generated:** Figure 4
* **Output files:** none
* **Output figures (figures/microbiome):**
  + Boxplot for Shannon Diversity: heu\_shannon\_boxplots.pdf

### **scripts/microbiome/ms\_microbiome\_ordination.R**

Beta-diversity via Bray-Curtis Index, and Adonis test to assess effect of HIV exposure on community composition

* **MS Figures Generated:** Figure 4
* **Output files:** none
* **Output figures (figures/microbiome):**
  + Ordination of all samples: heu\_nmds.pdf
  + Cohort-specific ordinations: cad\_heu\_nmds.pdf, blg\_heu\_nmds.pdf, saf\_heu\_nmds.pdf

### **scripts/microbiome/ms\_microbiome\_differential\_abundance.R**

Microbiome Differential Abundance with DESeq2. Also included is sPLS-DA for all significant OTUs to determine the classification accuracy of these features in a multivariate space.

* **MS Figures Generated:** Figure 5
* **Output files:**
  + sPLS-DA classification error rate: Rdata/R\_export/heu\_microbiome\_splsda\_error.csv
* **Output figures (figures/microbiome):**
  + DESeq2 Differentially Abundant OTUs by cohort: cad\_heu\_deseq.pdf, blg\_heu\_deseq.pdf, saf\_heu\_deseq.pdf
  + Boxplots of discriminatory OTUs selected by sPLS-DA: blg\_plot\_taxa\_controls.pdf, cad\_plot\_taxa\_controls.pdf, saf\_plot\_taxa\_controls.pdf
  + Annotations:
    - CAD R2 = 0.03527, p = 0.043
    - BLG R2 = 0.03824, p = 0.579
    - SAF R2 = 0.0724, p = 0.311

## **DIABLO**

Scripts for DIABLO analysis are located in two directories; data\_preparation, where microbiome and luminex data are independently prepared for DIABLO analysis, and a third script where data matrices from both data types are prepared together. The cohort\_models directory contains the performed DIABLO analyses.

### **scripts/DIABLO/data\_preparation/ms\_luminex\_diablo\_preparation.R**

In this script, we prepared luminex data for use with DIABLO and sPLS-DA analyses. This includes excuding any stim-cytokine combinations where most of the data is below the detection threshold

* **MS Figures Generated:** none
* **Output files:** Rdata/R\_export/heu\_lmx\_filtered\_sPLS.RDS
* **Output figures:** none

### **scripts/DIABLO/data\_preparation/ms\_otu\_diablo\_preparation.R**

In this script, we prepared microbiome data for use with DIABLO and sPLS-DA analyses. This includes filtering data to remove rare OTUs and normalizing the OTU data using total sum of squares followed by the centered log ratio.

* **MS Figures Generated:** none
* **Output files:** Rdata/R\_export/heu\_otu\_clr.rds
* **Output figures:** none

### **scripts/DIABLO/data\_preparation/ms\_mlx\_otu\_data\_matrix\_preparation.R**

In this script, we prepared microbiome and luminex data for use with DIABLO and sPLS-DA analyses. This includes ensuring that the row names for both data sets match, and that the metadata also matches the output files.

* **MS Figures Generated:** none
* **Output files:** Rdata/R\_export/spls\_data.rds
* **Output figures:** none

### **scripts/DIABLO/cohort\_models/ms\_DIABLO\_blg.R**

In this script, we performed DIABLO intgration for Luminex-cytokine data and OTU data for the BELGIAN cohort. The DIABLO model is saved for analysis of error rates

* **MS Figures Generated:** none
* **Output files:** Rdata/R\_export/blg\_sgccda\_res.rds
* **Output figures:** none

### **scripts/DIABLO/cohort\_models/ms\_DIABLO\_cad.R**

In this script, we performed DIABLO intgration for Luminex-cytokine data and OTU data for the CANADIAN cohort. The DIABLO model is saved for analysis of error rates

* **MS Figures Generated:** none
* **Output files:** Rdata/R\_export/cad\_sgccda\_res.rds
* **Output figures:** none

### **scripts/DIABLO/cohort\_models/ms\_DIABLO\_saf.R**

In this script, we performed DIABLO intgration for Luminex-cytokine data and OTU data for the SOUTH AFRICAN cohort. The DIABLO model is saved for analysis of error rates

* **MS Figures Generated:** none
* **Output files:** Rdata/R\_export/saf\_sgccda\_res.rds
* **Output figures:** none

### **scripts/DIABLO/cohort\_models/splsda\_error\_comparison.R**

In this script, we loaded the error rate results generated for sPLS-DA and DIABLO integration results for OTU and luminex cytokines data, for each cohort separately. We then constructed a figure comparing these directly.

* **MS Figures Generated:**Figure 6
* **Output files:** none
* **Output figures (figures/DIABLO):**
  + Cohort-specific error rates: blg\_err\_clr.pdf, cad\_err\_clr.pdf, saf\_err\_clr.pdf