**GC Microbiome Manuscript Analysis: README**

**Author: Nelly Amenyogbe**

**29-July-2020**

Table of Contents

[Introduction 2](#_Toc55913238)

[Data file descriptors 2](#_Toc55913239)

[Human data 2](#_Toc55913240)

[**Microbiome data** 2](#_Toc55913241)

[**Luminex data** 3](#_Toc55913242)

[**Metadata** 3](#_Toc55913243)

[Mouse Data 3](#_Toc55913244)

[**Luminex** 3](#_Toc55913245)

[**Microbiome** 3](#_Toc55913246)

[Analytical script descriptors 4](#_Toc55913247)

[Human Data 4](#_Toc55913248)

[**Demographics** 4](#_Toc55913249)

[**Microbiome analysis** 4](#_Toc55913250)

[**scripts/microbiome/ms\_gc\_ordination.R** 4](#_Toc55913251)

[**scripts/microbiome/ms\_gc\_alpha\_diversity.R** 4](#_Toc55913252)

[**scripts/microbiome/ms\_gc\_ordiR2step.R** 5](#_Toc55913253)

[**scripts/microbiome/ms\_gc\_taxa\_barplot.R** 5](#_Toc55913254)

[**scripts/microbiome/ms\_gc\_deseq\_heatmap.R** 5](#_Toc55913255)

[**Luminex Analysis** 5](#_Toc55913256)

[**scripts/luminex/ms\_gc\_filter\_lmx\_sPLS.R** 5](#_Toc55913257)

[**scripts/luminex/ms\_gc\_kruskalwallis.R** 6](#_Toc55913258)

[**scripts/luminex/ms\_gc\_lmx\_splsda.R** 6](#_Toc55913259)

[**sPLS Microbiome-Immune Integration** 6](#_Toc55913260)

[**scripts/sPLS\_mb\_immun/ms\_gc\_prepare\_otu\_spls.R** 6](#_Toc55913261)

[**scripts/sPLS\_mb\_immun/ms\_gc\_prepare\_otu\_lmx\_spls.R** 7](#_Toc55913262)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_all\_cohorts.R** 7](#_Toc55913263)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_blg\_cad\_ecd.R** 7](#_Toc55913264)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_blg.R** 7](#_Toc55913265)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_cad.R** 8](#_Toc55913266)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_ecd.R** 8](#_Toc55913267)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_saf.R** 8](#_Toc55913268)

[**scripts/sPLS\_mb\_immun/ms\_gc\_saf\_model\_comparison.R** 9](#_Toc55913269)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_stimulus\_hypergeometric\_test.R** 9](#_Toc55913270)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_cytokine\_hypergeometric\_test.R** 9](#_Toc55913271)

[**scripts/sPLS\_mb\_immun/ms\_gc\_spls\_otu\_hypergeometric\_test.R** 9](#_Toc55913272)

[**sPLS microbiome-immune-host demographics** 10](#_Toc55913273)

[**scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_spls\_demo\_cad.R** 10](#_Toc55913274)

[**scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_spls\_demo\_ecd.R** 10](#_Toc55913275)

[**scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_spls\_demo\_cad\_ecd.R** 10](#_Toc55913276)

[**scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_bfeed\_duration.R** 11](#_Toc55913277)

[Germ-Free Mouse Data 11](#_Toc55913278)

[**Luminex analysis** 11](#_Toc55913279)

[**scripts/gf\_mouse/luminex/ms\_gc\_gfmouse\_lmx\_pca\_wilcox.R** 11](#_Toc55913280)

[**Microbiome analysis** 12](#_Toc55913281)

[**scripts/gf\_mouse/microbiome/ms\_gc\_gfmouse\_mb\_ordination.R** 12](#_Toc55913282)

[**scripts/gf\_mouse/microbiome/ms\_gc\_gfmouse\_deseq2.R** 12](#_Toc55913283)

[**scripts/gf\_mouse/microbiome/ms\_gc\_gfmouse\_lmratio.R** 12](#_Toc55913284)

# **Introduction**

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

**Biogeography of the relationship between the child gut microbiome and innate immune system**

The **ms\_global\_cohort\_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. Original data files are saved in the **raw\_data** directory. 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

Package versions used for analysis are given in ms\_gc\_session\_info.docx. The R markdown file used to produce these is given in scripts/ ms\_gc\_session\_info.Rmd.

# **Data file descriptors**

## **Human data**

All human is located in the **Rdata/human\_raw\_data** directory. A data dictionary is provided here as **gc\_human\_data\_dictionary.xlsx.**

### **Microbiome data**

* **Rdata/raw\_data/gc\_physeq.rds:** Microbiome data is supplied as a phyloseq object. Load this package using the readRDS() command in R statistical software. 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/raw\_data/gc\_otu\_table.csv:** OTU count table, identical to data embedded in the phyloseq object
* **Rdata/raw\_data/gc\_taxonomy\_table.csv**: Taxonomic identifiers of OTUs listed in the OTU table, identical to data embedded in the phyloseq object

### **Luminex data**

* **Rdata/raw\_data/gc\_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.

### **Metadata**

* **Rdata/human\_raw\_data/gc\_metadata.csv:** all study participant metadata utilized throughout the manuscript
* **Rdata/human\_raw\_data/gc\_bfeed\_duration.csv:** metadata for all Canadian and Ecuadorean study subjects, indicating duration of breastfeeding in months. For South Africans, indicates whether children were ever breastfed only

## **Mouse Data**

All mouse is located in the **Rdata/mouse\_raw\_data** directory. A data dictionary is provided here as **gc\_mouse\_data\_dictionary.xlsx.**

### **Luminex**

* **Rdata/mouse\_raw\_data/gc\_gfmouse\_lmx.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.

### **Microbiome**

* **Rdata/mouse\_raw\_data/gc\_gfmouse\_phyloseq.rds:** Microbiome data is supplied as a phyloseq object. For useage, see human data gc\_physeq.rds.
* **Rdata/mouse\_raw\_data/gc\_gfmouse\_otu\_table.csv:** OTU count table, identical to data embedded in the phyloseq object
* **Rdata/mouse\_raw\_data/gc\_gfmouse\_taxonomy\_table.csv:** Taxonomic identifiers of OTUs listed in the OTU table, identical to data embedded in the phyloseq object
* **Rdata/mouse\_raw\_data/gc\_gfmouse\_metadata.csv:** experiment metadata, as embedded in the phyloseq object
* **Rdata/mouse\_raw\_data/** **gc\_gfmouse\_lactulose\_mannitol.csv:** Experiment data for lactulose\_mannitol ratio data, presented in Fig. 6G

# **Analytical script descriptors**

## **Human Data**

Human data scripts include all analyses performed in the manuscript, excluding analyses specific to germ-free mouse experiments.

### **Demographics**

#### **scripts/ms\_gc\_days\_between\_blood\_stool.R**

In this script, we will plotted the time difference between blood draw and stool sample collection, in days.

* **Output figures:**
  + **Supl. Figure 1:** days\_between\_blood\_stool.pdf

### **Microbiome analysis**

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

In this script, we performed an ordination of all samples via Bray-Curtis Distance, and visualized using non-metric multidimensional scaling.

* **Output figures:** 
  + **Fig. 1C:** microbiome/all\_samples\_nmds\_ordination.pdf
  + **Fig. 1D:** microbiome/gc\_birthcountry\_nmds\_ordination.pdf
* **Output files:** none

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

In this script, we determined the observed richness and Shannon Diversity of all samples, and applid the Kruskal-Wallis test to determine if either differ across cohorts. Further, we performed linear regression across cohorts to determine if any host factors affect richness and diversity measurements.

* **Reported statistics (linear regressions):**
  + **Cohort:** Shannon diversity ~ Cohort: Adj. R2 = 0.01352, p = 0.2352
  + **Cohort:** Observed richness ~ Cohort: Adj. R2 = 0.1544, p = 0.0002979, ECD p = 0.0428, NS SAF/ECD
  + **CAD: Shannon Diversity:** DELIVERY R2 = 0.1537992; p = 0.02910, MOM\_AGE R2 = 0.2575; p = 0.00356
  + **CAD: Shannon Diversity ~ maternal age + delivery:** MomAge p = 0.0141, DeliveryMode p = 0.1192, Model R2 = 0.271
  + **SAF: Shannon Diversity:** no significant interactions
  + **ECD: Shannon Diversity:** MOM\_AGE R2 = 0.2219; p = 0.001639, GEST\_AGE R2 = 0.1161, p = 0.04852
  + **ECD: Shannon Diversity ~ maternal age + gestational age:** MOM\_AGE p = 0.00346, GEST\_AGE p = 0.08317, Model R2 = 0.2894
  + **ECD + CAD: Shannon Diversity ~ maternal age splines:** Adjusted R2 = 0.09. p = 0.01374
* **Output Figures:**
  + **Fig. 1A:** microbiome/boxplot\_shannon\_diversity.pdf
  + **Fig. 1B:** microbiome/boxplot\_observed\_richness.pdf
  + **Supl. Fig. 2A:** microbiome/boxplot\_momage\_diversity.pdf
  + **Supl. Fig. 2B:** microbiome/cad\_ecd\_momage\_diversity.pdf
  + **Supl. Fig. 2C:** microbiome/ms\_matage\_uplot.tiff
* **Output files:** none

#### **scripts/microbiome/ms\_gc\_ordiR2step.R**

In this script, we determined contribution of host factors to microbiome community composition among all children, and within each cohort separately. Statistical results are included within the script, and were presented as text in the manuscript results.

* **Reported statistics:** 
  + **Cohort ordiR2step**: Cohort only has significant finding with R2 = 0.12655 (p = 0.002). R2 total is 0.12655. Remainder of variables contribute an R2 of 0.011 (0.11 variance) for a total R2 of 0.13777, or 13% total variance explained.
  + **CAD ordiR2step**: WLZ R2 = 0.030; p = 0.012
  + **SAF ordiR2step:** WLZ R2 = 0.174, p = 0.016
  + No significant findings for ECD/BLG
* **Output figures:** none
* **Output files:** none

#### **scripts/microbiome/ms\_gc\_taxa\_barplot.R**

In this script, we prepared figures to visualize taxonomic composition of all samples in barplots, colored by the taxonomic representation of samples.

* **Reported statistics:** none
* **Output Figures: Fig. 2A:** microbiome/genus\_barplot.pdf
* **Output files:** none

#### **scripts/microbiome/ms\_gc\_deseq\_heatmap.R**

In this script, we determines differentially abundant OTUs across cohorts using DESeq2. The most discriminatory taxa were selected using sPLS-DA, and visualized in a heatmap.

* **Reported statistics: DESeq2**: 462 OTUs differ across cohorts; p < 0.01
* **Output Figures: Fig. 2B:** microbiome/ms\_gc\_heatmap.png
* **Output files:** none

### **Luminex Analysis**

#### **scripts/luminex/ms\_gc\_filter\_lmx\_sPLS.R**

In this script, we prepared the luminex data for use with sPLS and sPLS-DA by selecting only stimulus-cytokine combinations that were significantly produced above baseline. These were determined using the Fliger-Kileen test. Unstimulated values were selected if over 70% of values were above the assay threshold.

* **Reported statistics:** none.
* **Output Figures:** none.
* **Output files:** Filtered luminex datasets for sPLS-DA and sPLS: R\_export/ms\_gc\_lmx\_filtered\_sPLS.rds

#### **scripts/luminex/ms\_gc\_kruskalwallis.R**

In this script, we performed the kruskal-wallis test for all stimulus-cytokines that passed the fligner-kileen test. Cytokines that significantly differ across cohort were then passed to sPLS-DA analysis, to determine the cohort-specific signatures in a multivariate space.

* **Reported statistics:** none.
* **Output Figures:** none.
* **Output files:** results table for Kruskal\_Wallis test: R\_export/ms\_gc\_lmx\_cohort\_kruskalwallis\_res.csv

#### **scripts/luminex/ms\_gc\_lmx\_splsda.R**

In this script, we performed sPLS-DA analysis for the luminex cytokine dataset including all children, excluding South African children, to identify discriminatory cytokine signatures for Belgian, Canadian, and Ecuadorean infants. Our previous findings illustrated that South African infants under-responded to PRR stimulation compared to all other site. This dramatic effect failed to highlight more subtle differences among the other three sites. For this reason, we focus ed on the other three sites here.

* **Reported statistics:**
* **Output figures:** 
  + **Fig. 3A:** luminex/cad\_features\_down.pdf
  + **Fig. 3B:** luminex/cad\_features\_up.pdf
  + **Fig. 3C:** luminex/blg\_features\_up.pdf
  + **Fig. 3D:** luminex/blg\_features\_down.pdf
  + **Fig 3E:** luminex/gc\_lmx\_splsda12\_ord.pdf
  + **Fig 3F:** luminex/gc\_lmx\_splsda\_error.pdf
* **Output files:** none

### **sPLS Microbiome-Immune Integration**

For sPLS analysis, OTU data are first filtered to remove near zero variance features, and normalized using the centered log ratio transformation. luminex and microbiome data are then prepared into data matrices with matching row names, and these data are saved. sPLS is then performed in six groups: i. all samples together, ii. Belgian Canadian, and Ecuadorean children, and iii-vi. each cohort individually.

#### **scripts/sPLS\_mb\_immun/ms\_gc\_prepare\_otu\_spls.R**

In this script, we prepared microbiome data for use with sPLS. This includes filtering the OTU tables to only select OTUs repesented across at least 30% of subjects within each cohort, and normalizing using the CLR transformation.

* **Reported statistics:** none
* **Output figures:** none
* **Output files:** OTU tables for sPLS: R\_export/otu\_clr.rds

#### **scripts/sPLS\_mb\_immun/ms\_gc\_prepare\_otu\_lmx\_spls.R**

In this script, we prepared microbiome and OTU data for sPLS. This involves creating data matrices from luminex data, and ensuring both data types have matching row names. These matrices are saved for ease of use for sPLS models.

* **Reported statistics:** none
* **Output figures:** none
* **Output files:** R\_export/spls\_data.rds

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_all\_cohorts.R**

In this script, we performed sPLS for all children together. We then refined the model to only include features that vary across at least 30% of features in the respective data frame, and plot selected examples alongside their univariate correlation strength.

* **Reported statistics:** none.
* **Output figures:** 
  + **Supl. Fig. 3C:** sPLS\_mb\_immun/all\_subjects/ms\_gc\_spls\_hm\_all\_subjects.png
  + **Supl. Fig. 3D:** sPLS\_mb\_immun/all\_subjects/ms\_gc\_spls\_rn\_all\_subjects.png
* **Output files:** R\_export/spls\_res/all\_subjects\_spls\_res.rds

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_blg\_cad\_ecd.R**

In this script, we perform sPLS for BLG, CAD, and ECD children. We then refine the model to only include features that vary across at least 30% of features in the respective data frame, and plot selected examples alongside their univariate correlation strength.

* **Reported statistics:** none
* **Output figures:**
  + **Fig. 4A:** sPLS\_mb\_immun/blg\_cad\_ecd/ms\_gc\_spls\_hm\_blg\_cad\_ecd.png
  + **Fig. 4B:** sPLS-mb\_immun/blg\_cad\_ecd/blg\_cad\_ecd\_cor\_circle\_sig.pdf
  + **Fig 4C:** sPLS-mb\_immun/blg\_cad\_ecd/blg\_cad\_ecd\_rn1.graphml [edited in cytoscape]
  + **Fig. 4D**: sPLS-mb\_immun/blg\_cad\_ecd/blg\_cad\_ecd\_example\_plots.pdf
  + **Fig. 4D:** sPLS-mb\_immun/blg\_cad\_ecd/all\_ecd\_example\_plots.pdf
* **Output files:** 
  + **sPLS model:** spls\_res/blg\_cad\_ecd\_spls\_res.rds
  + **Statistics to annotate plots:** figures/sPLS- figures/sPLS-mb\_immun/blg\_cad\_ecd/example\_plot\_annotations.csv
  + **Statistics to annotate plots**: figures/sPLS-mb\_immun/blg\_cad\_ecd/ecd\_example\_plot\_annotations.csv

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_blg.R**

In this script, we perform sPLS for BELGIAN children. We then refine the model to only include features that vary across at least 30% of features in the respective data frame, and plot selected examples alongside their univariate correlation strength.

* **Reported statistics:** none
* **Output figures:** 
  + **Fig. 5A:** spls\_mb\_immun/blg/gc\_ms\_spls\_hm\_blg.pdf
  + **Supl. Fig. 4A.A:** sPLS-mb\_immun/blg/blg\_cor\_circle\_sig.pdf
  + **Supl. Fig. 4A.B:** sPLS-mb\_immun/blg/ms\_gc\_spls\_blg\_rn.png/ blg\_rn12.graphml
  + **Supl. Fig. 4A.C:** sPLS-mb\_immun/blg/blg\_example\_plots.pdf
* **Output files: sPLS model:** spls\_res/blg\_spls\_res.rds
  + **Statistics to annotate plots**: figures/sPLS-mb\_immun/blg /blg\_example\_plot\_annotations.csv

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_cad.R**

In this script, we perform sPLS for CANADIAN children. We then refine the model to only include features that vary across at least 30% of features in the respective data frame, and plot selected examples alongside their univariate correlation strength.

* **Reported statistics:**
* **Output figures:** 
  + **Fig. 5B:** spls\_mb\_immun/cad/gc\_ms\_spls\_hm\_cad.pdf
  + **Supl. Fig. 4B.A**: sPLS-mb\_immun/cad/cad\_cor\_circle\_sig.pdf
  + **Supl. Fig. 4B.B:** sPLS-mb\_immun/cad/cad\_rn12.graphml/ ms\_gc\_spls\_rn\_cad.png
  + **Supl. Fig. 4B.C**: sPLS-mb\_immun/cad/cad\_example\_plots.pdf
* **Output files: sPLS model:** pls\_res/cad\_spls\_res.rds
  + **Statistics to annotate plots**: figures/sPLS-mb\_immun/cad /cad\_example\_plot\_annotations.csv

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_ecd.R**

In this script, we perform sPLS for ECUADOREAN children. We then refine the model to only include features that vary across at least 30% of features in the respective data frame, and plot selected examples alongside their univariate correlation strength.

* **Reported statistics:**
* **Output figures:** 
  + **Fig. 5C:** spls\_mb\_immun/ecd/gc\_ms\_spls\_hm\_ecd.pdf
  + **Supl. Fig. 4C.A**: sPLS-mb\_immun/ecd/ecd\_cor\_circle\_sig.pdf/ sPLS-mb\_immun/ecd/ecd\_cor\_legend.pdf
  + **Supl. Fig. 4C.B:** sPLS-mb\_immun/ecd/ecd\_rn13.graphml/ ms\_gc\_spls\_rn\_ecd.png
  + **Supl. Fig. 4C.C**: sPLS-mb\_immun/ecd/ecd\_example\_plots.pdf
* **Output files: sPLS model:** pls\_res/ecd\_spls\_res.rds
  + **Statistics to annotate plots**: figures/sPLS-mb\_immun/ecd/ecd\_example\_plot\_annotations.csv

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_saf.R**

In this script, we perform sPLS for SOUTH AFRICAN children. We then refine the model to only include features that vary across at least 30% of features in the respective data frame, and plot selected examples alongside their univariate correlation strength.

* **Reported statistics:** none
* **Output figures:** 
  + **Fig. 5D:** spls\_mb\_immun/saf/gc\_ms\_spls\_hm\_saf.pdf
  + **Supl. Fig. 4D.A-B**: sPLS-mb\_immun/saf/saf\_cor\_circle\_sig.12pdf/ saf\_cor\_circle\_sig.13pdf
  + **Supl. Fig. 4D.C:** sPLS-mb\_immun/saf/saf\_rn13.graphml/ ms\_gc\_spls\_rn\_saf.png
  + **Supl. Fig. 4D.D**: sPLS-mb\_immun/saf/saf\_example\_plots.pdf
* **Output files: sPLS model:** pls\_res/saf\_spls\_res.rds
  + **Statistics to annotate plots**: figures/sPLS-mb\_immun/saf/saf\_example\_plot\_annotations.csv

#### **scripts/sPLS\_mb\_immun/ms\_gc\_saf\_model\_comparison.R**

In this script, we compared sPLS results with and without SAF included, to illustrate overlap of features selected at the expense of significant findings

* **Reported statistics:** N
  + OTU overlap: 61. Number of OTUs selected with SAF: 3. Without: 21
  + LMX overlap: 67. Number of cytokines selected with SAF: 12. Without: 23
* **Output figures:**
  + **Supl. Fig 3A:** OTU selection with/without SAF: sPLS-mb\_immun/model\_comparison/percent\_sig\_comparison\_OTU.pdf
  + **Supl. Fig 3B:** LMX selection with/without SAF: sPLS-mb\_immun/ model\_comparison/percent\_sig\_comparison\_LMX.pdf
  + **Supl. Fig 3E-F:** OTU-cytokine plots with and without South African children: sPLS-mb\_immun/ model\_comparison/all\_subjects\_comparison\_plots.pdf/ sPLS-mb\_immun/blg\_cad\_ecd\_comparison\_plots.pdf
* **Output files:**
  + **Annotations for Supl. Fig 3E:** sPLS-mb\_immun/ model\_comparison/ms\_gc\_all\_subjects\_comparison\_annotation.csv
  + **Annotations for Supl. Fig 3F:** sPLS-mb\_immun/ model\_comparison/ms\_gc\_blg\_cad\_ecd\_comparison\_annotation.csv

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_stimulus\_hypergeometric\_test.R**

In this script, we performed the hypergeometric test to determine whether any stimuli were enriched for in the sPLS results for each cohort

* **Reported statistics:** none
* **Output figures: Fig. 5E.** sPLS-mb\_immun/ms\_gc\_stim\_phypher\_sig.pdf
* **Output files:** none

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_cytokine\_hypergeometric\_test.R**

In this script, we performed the hypergeometric test to determine whether any stimuli were enriched for in the sPLS results for each cohort

* **Reported statistics:** none
* **Output figures: Fig. 5F:** sPLS-mb\_immun/ms\_gc\_cytokine\_phypher\_sig.pdf
* **Output files:** none

#### **scripts/sPLS\_mb\_immun/ms\_gc\_spls\_otu\_hypergeometric\_test.R**

In this script, we performed the hypergeometric test to determine whether any bacterial families were enriched for in the sPLS results for each cohort

* **Reported statistics:** Significant findings are Prevotellaceae for BLG\_CAD\_ECD and ECD alone.
* **Output figures:** none
* **Output files:** none

### **sPLS microbiome-immune-host demographics**

For this section, we utilize the microbiome and immune data prepared above, but also include host demographic data as a third layer, and perform block-sPLS for Canadian and Ecuadorean children both together, and for each cohort separately.

#### **scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_spls\_demo\_cad.R**

In this script, we determined whether host demographic factors associated with microbiome-immune correlations among Canadian children.

* **Reported statistics:** none
* **Output figures:**
  + Heatmaps for microbiome immune demographic associations:
    - **Supl. Fig. 7A:** sPLS\_mb\_im\_demographics/cad/cad\_dm\_hm\_comp1.pdf
    - **Supl. Fig. 5B:** sPLS\_mb\_im\_demographics/cad/cad\_dm\_hm\_comp2.pdf
    - **Supl. Fig. 7B:** sPLS\_mb\_im\_demographics/cad/cad\_dm\_hm\_comp3.pdf
  + Plots of example correlations:
    - **Supl. Fig. 7E:** sPLS\_mb\_im\_demographics/cad/cad\_delivery\_plots.pdf
      * annotations in graph\_annotations/ cad\_dm\_plot\_annotations.csv
    - **Supl. Fig. 7F:** sPLS\_mb\_im\_demographics/cad/cad\_sex\_plots.pdf
      * annotations in graph\_annotations/ cad\_sex\_plot\_annotations.csv
    - **Supl. Fig. 7G:** sPLS\_mb\_im\_demographics/cad/cad\_waz\_plots.pdf
      * annotations in graph\_annotations/ cad\_waz\_plot\_annotations.csv
* **Output files:** none.

#### **scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_spls\_demo\_ecd.R**

In this script, we determined whether host demographic factors associated with microbiome-immune correlations among Ecuadorean children.

* **Reported statistics:** none
* **Output figures:**
  + Heatmaps for microbiome immune demographic associations:
    - **Supl. Fig 7C:** sPLS\_mb\_im\_demographics/ecd/ecd\_dm\_hm\_comp1.pdf
    - **Supl. Fig. 7D:** sPLS\_mb\_im\_demographics/ecd/ecd\_dm\_hm\_comp2.pdf
  + Plots of example correlations:
    - **Supl. Fig. 7H:**sPLS\_mb\_im\_demographics/ecd/ecd\_comp2\_dm\_plots.pdf
      * annotations in graph\_annotations/ecd\_dm\_plot\_annotations.csv
    - **Supl. Fig. 7I:** sPLS\_mb\_im\_demographics/ecd/ecd\_comp1\_momage\_plots.pdf
      * Annotations in graph\_annotations/ecd\_ma\_plot\_annotations.csv
* **Output files:** none

#### **scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_spls\_demo\_cad\_ecd.R**

In this script, we determined whether host demographic factors associated with microbiome-immune correlations among both Canadian and Ecuadorean children.

* **Reported statistics:** none
* **Output figures:**
  + Heatmaps for microbiome immune demographic associations:
    - **Supl. Fig. 6A:** sPLS\_mb\_im\_demographics/cad\_ecd/ cad\_ecd\_comp1\_hm.pdf
    - **Supl. Fig. 6B:** sPLS\_mb\_im\_demographics/cad\_ecd/cad\_only\_comp1.pdf
    - **Supl. Fig. 6C:** sPLS\_mb\_im\_demographics/cad\_ecd/ecd\_only\_comp1.pdf
  + Plots of example correlations:
    - **Supl. Fig. 6D:** sPLS\_mb\_im\_demographics/cad\_ecd/cad\_ecd\_comp1\_plot.pdf
      * Annotations in graph\_annotations/cad\_ecd\_otu\_plot\_annotations.csv
    - **Supl. Fig. 6E:** sPLS\_mb\_im\_demographics/cad\_ecd/cad\_ecd\_comp1\_plot\_separate.pdf
      * Annotations in graph/annotations/cad\_otuplot\_annotations.csv, ecd\_otuplot\_annotations.csv
* **Output files:** none

#### **scripts/sPLS\_mb\_immun\_demographics/ms\_gc\_bfeed\_duration.R**

* **Reported statistics:** Significance levels for OTU-time since bfeed correlations for Canadian and Ecuadorean children:
  + **Supl. Fig. 7C: OTU\_61 G\_Roseburia**
    - CAD: p = 0.014, Adj. R2 = 0.264. ECD: p = 0.76, Adj. R2 = 0.76
  + **Supl. Fig. 7D: OTU\_48 F\_Lachnospiraceae**
    - CAD: p = 0.017, Adj. R2 = 0.249. ECD: p = 0.3, Adj. R2 = 0.31
* **Output figures:** 
  + **Supl. Fig. 5A:** breast feeding duration: sPLS\_mb\_im\_demographics/cad\_ecd/bfeed\_duration\_spls\_subjects.pdf
  + Plots of select OTU correlations to breast feeding duration:
    - **Supl. Fig. 5C:** sPLS\_mb\_im\_demographics/cad\_ecd/otu61roseburia\_bf.pdf
    - **Supl. Fig. 5D:** sPLS\_mb\_im\_demographics/cad\_ecd/otu48lachno\_bf.pdf
* **Output files:** none

## **Germ-Free Mouse Data**

Germ-Free data scripts include analyses specific to germ-free mouse experiments.

### **Luminex analysis**

#### **scripts/gf\_mouse/luminex/ms\_gc\_gfmouse\_lmx\_pca\_wilcox.R**

In this script, we performed PCA for luminex cytokine data, to visualize the variance due to stimulus, cohort, and stool donor. We also performed the wilcoxon test to determine which cytokine responses differed between SAF and CAD mice.

* **Reported statistics:** none.
* **Output figures:** 
  + **Fig. 6D**: gf\_mouse/luminex/gf\_sig\_cytokines.pdf
  + **Supl. Fig. 8A:** gf\_mouse/luminex/gf\_all\_cytokines.pdf
  + **Fig. 6A:** gf\_mouse/luminex/gfmouse\_lmx\_pca\_12.pdf
  + **Fig. 6B:** gf\_mouse/luminex/gfmouse\_lmx\_pca\_13.pdf
* **Output files:** none.

### **Microbiome analysis**

#### **scripts/gf\_mouse/microbiome/ms\_gc\_gfmouse\_mb\_ordination.R**

In this script, we determined the Bray-Curtis distance between all microbiome samples, and visualized via Non-metric multidimensional scaling (NMDS). We then performed this analysis for the Ielum, Jejunum, and Feces separately, and performed the Adonis test to determine the variance explained by fecal donor country on microbial community composition.

* **Reported statistics: Adonis test for Feces, Jejunum and Ileum:**
  + **Jejunum:** R2 = 0.17, p = 0.019
  + **Ileum:** R2 = 0.21, p = 0.002
  + **Feces**: R2 = 0.20, p = 0.001
* **Output figures:**
  + **Fig. 6E**: gf\_mouse/microbiome/gfmouse\_ord\_alltissues.pdf
  + **Fig. 6F:** gf\_mouse/microbiome/gfmouse\_ord\_cohort.pdf
* **Output files:** none

#### **scripts/gf\_mouse/microbiome/ms\_gc\_gfmouse\_deseq2.R**

In this script, we pefored the DESeq2 test to identify differentially-abundant OTUs between mice gavaged with either CAD or SAF feces. We then visualized these specific OTUs using boxplots.

* **Reported statistics:** none
* **Output figures:** 
  + **Supl. Fig 8B:** gf\_mouse/microbiome/mouse\_feces\_deseq.pdf
  + **Supl. Fig 8C:** gf\_mouse/microbiome/mouse\_ileum\_deseq.pdf
  + **Supl. Fig 8D:** gf\_mouse/microbiome/mouse\_jejunum\_deseq.pdf
  + **Supl. Fig 8E:** gf\_mouse/microbiome/gf\_saf\_otus\_boxplot.pdf
  + **Supl. Fig 8F:** gf\_mouse/microbiome/gf\_cad\_otus\_boxplot.pdf
* **Output files:** none

#### **scripts/gf\_mouse/microbiome/ms\_gc\_gfmouse\_lmratio.R**

In this script, we compared the lactulose:mannitol ratio in urine collected from mice gavaged with Canadian or South African stool samples using a t-test, and graphed the results.

* **Reported statistics:** t-test LM ratio CAD vs SAF: p = 0.0033
* **Output figures: Fig. 6G:** gf\_mouse/microbiome/ms\_gc\_gfmouse\_lmratio.pdf
* **Output files:** none