**Table of Contents**

[Abstract 2](#_Toc19020032)

[Background 2](#_Toc19020033)

[Overview of Multi-Omics data integration 2](#_Toc19020034)

[Human Microbiome 3](#_Toc19020035)

[Host-Transcriptome 5](#_Toc19020036)

[Clinical Phenotypes 5](#_Toc19020037)

[Host-Transcriptome-Microbiome Integration 5](#_Toc19020038)

[Network analysis 6](#_Toc19020039)

[Primary Aims 7](#_Toc19020040)

[The Network Analysis Method Applied 9](#_Toc19020041)

[Canonical Correlation analysis 9](#_Toc19020042)

[SmCCNet 10](#_Toc19020043)

[Enrichment analysis 12](#_Toc19020044)

[Methods 13](#_Toc19020045)

[Overview 13](#_Toc19020046)

[Data Preprocessing: Normalization 14](#_Toc19020047)

[Data Preprocessing: Transformation 14](#_Toc19020048)

[Data Preprocessing: Down-sampling 15](#_Toc19020049)

[Grubbs’ outlier test 15](#_Toc19020050)

[Z-score standardization 16](#_Toc19020051)

[SmCCNet Model Hyperparameter Tuning 16](#_Toc19020052)

[Summary of Hubs from Networks Generated 17](#_Toc19020053)

[Motivating Example 17](#_Toc19020054)

[Dataset Generation and Preprocessing 17](#_Toc19020055)

[SmCCNet: Networks generation 19](#_Toc19020056)

[Preliminary Results 19](#_Toc19020057)

[Conclusions 24](#_Toc19020058)

[References 24](#_Toc19020059)

Thesis Proposal:

Integrating Host Gut Transcriptome and Intestinal Mucosal Microbiome with Soluble CD14 Through Multi-Omics Network Analysis in Chronic HIV-1 Infection

# Abstract

The human gastrointestinal (GI) tract plays a unique role in immunological protection against exposure of pathogens from the outside environment[1, 2]. Evidence shows that Human Immunodeficiency Virus-1 (HIV) infection targets gut-associated lymphoid tissues (GALT) as major sites of viral transmission, replication and seeding, resulting in gut mucosal damage and CD4+ T cell depletion[1]. [Introduce microbial translocation here]. It is increasingly recognized that the interaction between microbes and host responses influences disease progression[3]. Meanwhile, a wide spectrum of relevant clinical phenotypes is involved in the biological mechanisms as well. Thus, there is a need to explore the associations among the alterations in microbial communities, host physiology and disease stages, with respect to different phenotypic traits[4]. Particularly, it has also been shown that plasma soluble CD14 (sCD14), a marker of microbial translocation, is a vital clinical parameter during HIV-1 infection[5, 6]. However, integrating multi-Omics datasets with a specific phenotype is a daunting task[4]. Here, we use the sparse multiple canonical correlation network analysis (SmCCNet) to integrate intestinal microbiome and the host gut transcriptome along with a quantitative phenotype of interest, sCD14[7, 8]. With this approach, host-transcriptome-microbiome networks which are specific to sCD14 are constructed. This study will help provide insights of host-microbe interactions during chronic HIV-1 infection. To our knowledge, there is no published research on integrating host-transcriptome and gut microbiome in chronic HIV-1 infection using correlation-based network analysis approach.

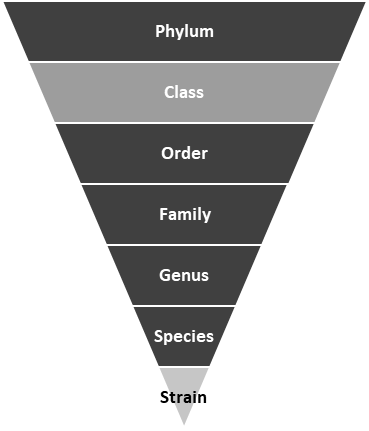
# Background

## Overview of Multi-Omics data integration

Since the publication of the first human genome, the Omics era has accelerated all aspects of biological research. However, one-dimensional Omics-data can only provide limited information regarding disease etiology, such as oncogenesis[9]. With the advent of high-throughput sequencing (HTS) and other technologies, nowadays, it is possible to obtain multi-Omics data, which consists of two or more datasets which were sampled from the same cohort of subjects but containing different sets of biological features such as genes, metabolites, or operational taxonomic units (OTUs)[10]. Therefore, it is challenging but necessary to develop novel data integration algorithms to incorporate different biological layers of information with phenotypic outcomes. Data integration algorithms can be classified into two types: unsupervised methods and supervised methods[3, 9]. The main difference between the two types is that a supervised algorithm is performed using labeled response variables for training, such as the phenotype labels of samples, whereas unsupervised methods are performed without labeled response variables and rely on the data structure[9]. Network analysis or network-based method are used to understand and represent biological entities at the systems level using tools and concepts derived from graph theory, and can either be unsupervised or supervised[3, 9, 11, 12]. Integrative network analysis using multi-Omics data has been shown as a promising tool in describing the underlying ecosystem, generating testable hypothesis and gaining mechanistic insights for designing subsequent experiments[13].

## Human Microbiome

In recent years, the studies of human microbiome have been an important source of novel findings and open questions[14]. The human microbiome carries an estimated 150-fold more genes than the human genome[3]. The collective genome of the microbiome, termed as metagenome, is also referred to as the second genome of the human body[3]. The microbial community can either be monitored by low throughput methods such as quantitative PCR (qPCR) and flow cytometry, or high throughput sequencing methods such as 16S ribosomal RNA (rRNA) (amplicon) sequencing and shotgun metagenomic sequencing[11]. The 16S rRNA sequencing method using the 1.5 Kbp ribosomal RNA subunit gene as the marker gene, is less expensive, and can work with low biomass and host-contaminated samples (such as colon biopsies), at the sacrifice of resolution, compared with shotgun metagenomic sequencing[15-18]. The 16S rRNA sequencing consists of two steps: 1) Amplification and sequencing of the16S rRNA gene and 2) Binning, which is the assignment of sequences to OTUs[15, 18]. The taxonomic hierarchy of the microbiome data usually contains Phylum, Class, Order, Family, Genus, Species and Strain as shown in **Figure 1**. However, the resolution of 16S pipelines is usually limited to the Genus level[15, 16, 18].



**Figure 1.** The taxonomic hierarchy can be built with the microbiome sequencing data. Through 16S rRNA pyrosequencing, the resolution is typically limited to the Genus level. Levels included in our dataset are in dark. gray.

The GI tract is the largest mucosal surface, and is composed of epithelial cells, immune cells and the microbiota[2]. The gut microbiome (**Figure 1.**) is a complex ecosystem that performs several vital functions contributing to host fitness and well-being, such as the production of essential nutrients, digestion and degradation of macromolecules and training of the host immune system[10, 13]. A loss of gut microbial diversity and complexity can lead to various microbiota-associated diseases, such as obesity, type 2 diabetes, and inflammatory bowel disease (IBD)[2]. The NIH Human Microbiome Project (HMP) which has been carried out over ten years and two phases, can serve as the icon of cutting edge microbiome research[14]. The microbiome datasets collected by HTS have been widely used and researchers has become increasingly aware of several challenges introduced by high-throughput microbiome data. Due to an arbitrary total read (library size) imposed by the instrument, the HTS methods can only provide us with the relative counts of taxa and this type of microbiome data should always be normalized by the corresponding library size[19]. As a result, the normalized count which is known as relative abundance (RA) is compositional, and taxa are not [19, 20]. Moreover, this type of data contains many zeros, resulting in high sparsity. Additionally, the distribution of the RA of taxa is almost always highly right/positively skewed, because the presence of several dominant taxa in the ecosystem. Fortunately, the above issues can be addressed with proper statistical methods as applied in this study. However, there are still several other limitations that need be considered when modelling the human gut microbiota. One limitation is that the simplified networks are difficult for explaining the complexity and diversity of the gut microbiota, and network reconstruction methods do not scale well[13]. Finally, it is also time consuming to validate the networks by subsequent experiments, and existing knowledge may not be sufficient to validate newly generated networks either.

## Host-Transcriptome

The host transcriptome can be quantified via two high-throughput techniques: microarrays, which quantify a set of predetermined sequences, and RNA sequencing (RNA-Seq), which uses high-throughput sequencing to capture all sequences[21]. Since first introduced in 2006, RNA-Seq has continually improved and has overtaken microarrays as the dominant transcriptOmics technique in 2015[21]. Read counts obtained by RNA-Seq can be used to accurately model the relative gene expression level. The integrative Omics analysis can be conducted at both candidate genes and genome-wide level. In this study, interferon (IFN) stimulated genes (ISGs *aka interferome*) were selected, which were identified in the preliminary RNA-Seq results and correlated with clinical outcomes during chronic HIV-1 infection (*Kejun Guo et al. 2019 manuscripts*). Particularly, type I Interferon (IFN-Is) responses serve as the first line of defense against viral infections, however, IFN subtypes trigger different sets of genes (*Kejun Guo et al. 2019 manuscripts*). Our study reveals that genes stimulated by IFNα subtypes (1, 2, 5, 8 and 14) and IFNβ have qualitative differences in their expression level and during HIV-1 infection (*Kejun Guo et al. 2019 manuscripts*). Hence, the integrative analysis in our initial work was limited to these candidate genes.

## Clinical Phenotypes

Structural impairment of the epithelial barrier and a disruption of intestinal homeostasis occurs early in HIV-1 infection[5, 6]. Consequently, the accompanying translocation of microbial products and potentially microbes themselves from GI tract into the systemic circulation has been linked to immune dysfunction, inflammation and HIV-1 disease progression[5, 6]. Soluble CD14 is a co-receptor for lipopolysaccharide (LPS)[8]. Moreover, sCD14 is a marker of monocyte activation, not restricted to activation by LPS[8]. The elevated plasma levels of sCD14 are associated with severe disease progression and higher risk of death in HIV-infection[8, 22, 23]. Soluble CD14 is one of microbial translocation markers as well[5, 6]. In one study, bacterial diversity in the stool was negatively associated with markers of systemic microbial translocation, including sCD14[6]. Thus, in this study, we propose to utilize sCD14 as an “anchor” to determine associations between the microbiome and host transcriptome.

## Host-Transcriptome-Microbiome Integration

Indeed, the human gut microbiome has proven difficult to study because its close relations with the physiology of the host[18]. The collective research results highlight the need to develop integrative Omics analysis tools in order to study the highly diverse host–microbiome responses. Network analysis has been widely used in exploring microbiome and integrating microbiome with host Omics data, such as co-occurrence networks, microbe-microbe metabolic networks, host-metabolic networks, host-genome microbiome networks, host-epigenome microbiome networks, host-transcriptome microbiome networks and host-metabolome microbiome networks[3]. The host-metabolome and microbiome integration are most widely studied among those host-microbiome interactions Despite the fact that the integrative analysis of host transcriptome and microbiome can help identify novel mechanisms, very few studies have examined the host-microbe interactions in the Omics integration fashion, mainly due to the lack of proper designed cohort study and computational tools[2, 4]. Previously, methods such as multivariate linear model, additive general linear model, rules-based model (RBM), logistic regression, unsupervised PCA, Bayesian factor analysis (FA), generalized linear mixed model (GLMM) and sparse, weighted correlation network analysis (WGCNA), transkingdom network (TransNet) and compositionally robust partial least squares regression (CompPLS) have been applied to integrate host-transcriptome and microbiome[3, 4, 24-29].

## Network analysis

With the rapid development of high-throughput techniques, the large-scale identification of components (genes, RNAs, proteins, metabolites and microbes) are more common and achievable[7, 9, 30]. Typically, the goal of network analysis is to assemble large-scale functional interactions or expression similarities into a network format, whose topological structure contains significant biological properties represented as interactions in the form of edges that connect biological components in the form of nodes[13, 30]. To date, several types of biological networks have been characterized extensively, such as: transcription factor-binding network, protein–protein interactions network, metabolic network, genetic network, MicroRNA-transcriptome network and microbiota co-occurrence network[7, 29-31]. Networks can be generated from single Omics such as weighted gene co-expression network analysis (WGCNA) method, or from multiple Omics data using joint non-negative matrix factorization (NMF)[9, 29]. Networks can either be created top-down as data-driven networks as in this work, or bottom-up as knowledge-driven networks [7, 13].

Within a network, the biological components such as genes are termed vertices or nodes, and connections between them are defined as edges (in undirected networks) or arcs (in directed networks). Due to the properties of host-microbe interactions, only undirected networks were generated in this study. The edges are usually based on strength of correlations or similarities[12]. The strength of edges can vary considerably, by orders of magnitude, and there is no objective threshold of edges’ strengths to quantitatively define whether the edges exist or not[30]. The degree feature in a network is defined as the number of links connected to a node. A node with high degree is likely to play a more important role in maintaining the network structure, which can be named as “hub”[12, 30].

It is generally accepted that biological networks typically have a scale-free format where several hubs have many connections and most nodes only have one or a few connections. It has been shown that the microbial co-occurrence networks, like protein interaction and human social networks, are scale-free[9]. The scale-free network is more robust, since the dysfunction of non-hub nodes is less disruptive in a scale-free network compared with a random network. Thus, hub components may play essential roles in biological systems. Biological networks also have dynamic behaviors as adaptions to various conditions[12, 30]. Unfortunately, our data only contains static information.

In this work, we will apply SmCCNet integrative network analysis method with host-gut-transcriptome, mucosal microbiome and clinical meta datasets from the chronic HIV-1 infection cohort. Then, we will develop guidelines for data preprocessing and algorithm tuning for SmCCNet method, when incorporating microbiome data. Our specific aims are:

Primary Aims

**Aim 1:** Construct integrative networks of host-gut-transcriptome and gut-microbiome specific to sCD14 for the chronic mucosal HIV-1 infection cohort through the SmCCNet method

Hypothesis: Test if the SmCCNet algorithm can be applied to discover associations between the microbiome, transcriptome, and host phenotype, with datasets from the above cohort.

Aim 1a: Apply SmCCNet integrative network analysis approach to the chronic mucosal HIV-1 infection cohort.

Aim 1b: Compare networks generated with/without the sCD14 through hubs in the networks

Aim 1c: Evaluate and interpret networks through enrichment analysis and literature search

It has recently been shown that HIV-1 infection dramatically alters the intestinal microbial communities and triggers microbial translocation affecting systemic circulation. Hence, there has been an increasingly interest in exploring the complex host-microbiome interplays during HIV-1 disease progression. However, no studies have investigated these mechanisms at the systems level. The SmCCNet network analysis method will be applied in this study, which provides the unique exploratory capability through an unsupervised approach integrating gut microbiome and host-gut-transcriptome with respect to sCD14, the biomarker of microbial translocation.

**Aim 2:** Develop guidelines for data preprocessing and algorithm tuning for SmCCNet method, when incorporating microbiome data.

Hypothesis: Demonstrate that a properly tuned SmCCNet pipeline will be widely applicable in incorporating microbiome with host-transcriptome and phenotype data.

Aim 2a: Sensitivity analysis on microbiome preprocessing parameters such as the relative abundance threshold and different microbiome taxa levels.

Aim 2b: Sensitivity analysis on SmCCNet algorithm hyperparameters.

As the profiling of human microbial communities is becoming more accessible, there is an urgent need to understand the interactions between the alterations in microbiota and the human genetic response. However, the SmCCNet analysis pipeline was not developed for microbiome data initially. In this study, we will test the stability of this pipeline with microbiome data and provide specific guidelines to integrate microbiome with host-transcriptome.

# The Network Analysis Method Applied

## Canonical Correlation analysis

Canonical correlation analysis (CCA) was first introduced by Hotelling, which is a traditional set correlation-based modelling method and has been modified and applied for data integration[3, 32-34]. This method is developed to investigate the relationship between two set of variables by finding the optimum linear combination of features which maximize the correlations[7, 32, 33].

**Equation 1**

subject to

Where the canonical correlation is expressed as , , are two data matrices of subjects, each with dimension and . The vectors and are the canonical weights (loading factors)[7, 9]. In ***Equation 1***, if we assume that both columns of and have been standardized to have mean 0 and standard deviation 1 and the covariance matrix of the features is diagonal, then the model constraint can be simplified to . Traditional CCA doesn’t apply dimension reduction techniques. Hence, penalization and regularization terms are added to generate sparse models with high-dimensional data. Particularly, the objective function of L1-penalized sparse CCA can be expressed as[7, 32, 33]:

**Equation 2**

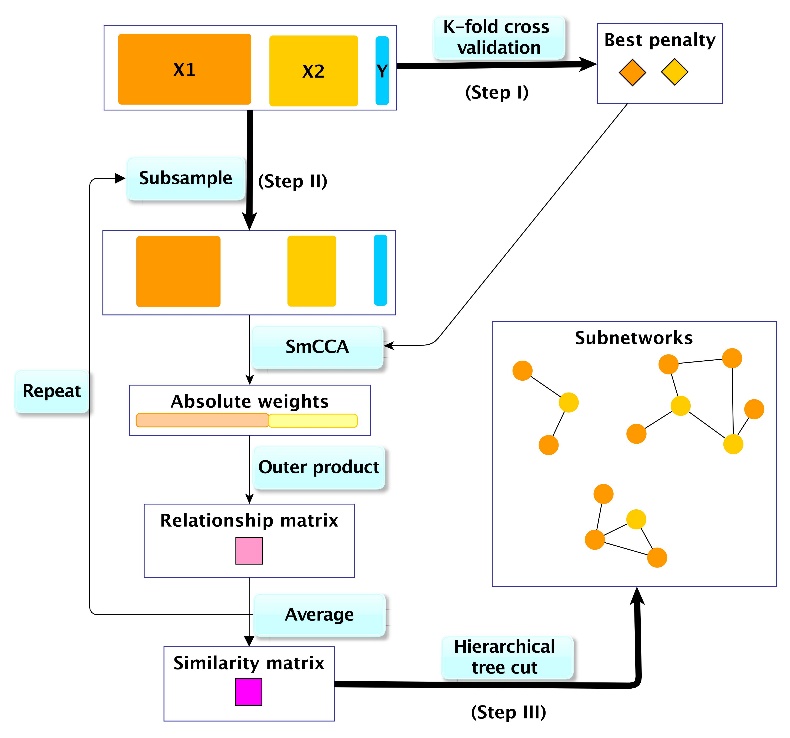
subject to

where the is the pre-selected sparse penalty constant. The least absolute shrinkage and selection operation (LASSO) regularization has been shown to be an effective variable selection method. Under the L1 regularization from LASSO, the penalty term is the of weights . In this case, the sparse penalty can range from to , where the is the number of features in the two Omics datasets, . In practical, is re-parametrized to with a range , such that . Therefore, large corresponds to less stringent criteria which has more features left in the subnetworks and vice versa.

Other newly modified CCA methods includes structure constrained CCA (ssCCA) and group sparse CCA, which consider grouped effects of features as structure embedded within the datasets[35, 36].

## SmCCNet

Unlike previous CCA-based methods, SmCCNet is a semi-supervised machine learning technique to integrate more than two datasets on the same subjects. A quantitative phenotype of interest can also be incorporated into the framework which guarantees constructed multi-Omics networks are specific to the phenotype[7]. n this framework, besides , , we also have a quantitative vector (Here a clinical parameter) whose dimension is . In this way, the new objective function becomes the sum of three pairwise set correlations:



***Figure 2.*** *Overview of SmCCNet workflow. In this study, , and represent host transciptome, microbiome and LPS levels. The whole procedure contains three steps: 1) Utilize K-fold CV to identify the best penalty pair; 2) Repeatedly randomly subsampling of omics data without replacement, apply SmCCA with identified penalties, obtain the weights and then the relationship matrix. The similarity matrix is the average of all feature relationship matrices; 3) Apply a hierarchical tree cut to the similarity matrix to find subnetworks.*

**Equation 3**

subject to

where the is the pre-selected sparse penalty constant as above. In our study, this is to maximize the sum of pairwise Pearson correlations by finding the optimal linear combinations of each Omics data matrix[7].

The unweighted version , is a special case of sparse multiple CCA (SmCCA) introduced by Witten and Tibshirani[7, 32]. The weighted SmCCNet (a, b and c are not all equal) is useful to prioritize weaker correlations among all three pairwise correlations.

As shown in ***Figure 2***, the very first step of SmCCNet is to identify the best penalty parameters through K-fold cross validation (CV), which is to decide the scale of L1 regularization via internal validation. In particular, the data is broken into K folds of test and training sets; then for each training data and penalty option, pseudo weights are aggregated and estimated through feature subsampling and SmCCA; the total prediction error for each penalty option is calculated subsequently by summing across K folds; finally, the best penalty pair is the one with the minimal prediction error[7]. In general, the best penalties are given by:

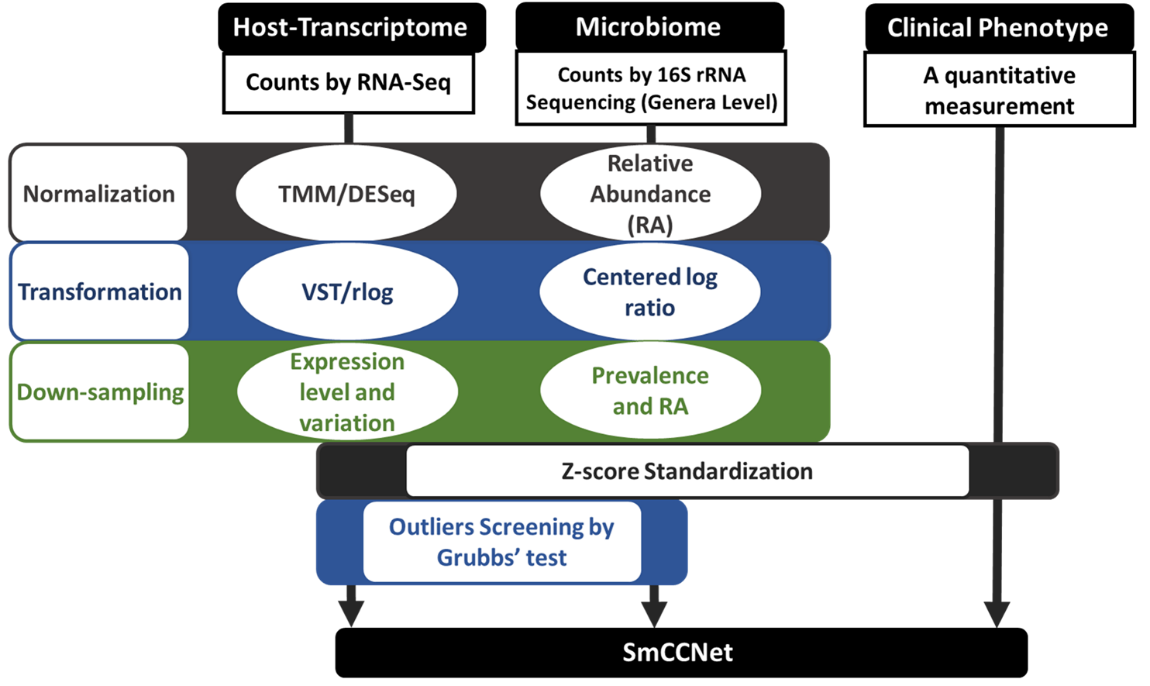
**Equation 4**

where the two Omics data sets are split to , and , for the th fold, the pseudo weights are calculated based on the training set given the penalty .

Given a penalty pair, the SmCCNet uses a repeated random feature subsampling scheme to generate robust canonical weights through SmCCA (***Figure 2***). The relationship matrix A is created for each subsampling by the outer product of canonical weights , where, a vector of length . The th entry, , measures the relatedness of the th and th features. The similarity matrix is calculated by taking the average of all sub-samplings and then rescaled to a maximum entry of 1. Utilizing the complete linkage method, where the distance between two classes is defined as the greatest distance that could be obtained, a hierarchical tree can be constructed based on . The hierarchical tree is further cut by a loose height threshold () and singletons are trimmed off. To obtain trimmed modules, all edges with weights less than d and nodes without any edge are removed from the similarity matrix. In terms of weighted SmCCNet, the weights in the weighted SmCCNet can be defined empirically by the density plot of the distribution of pairwise Pearson correlations. We applied SmCCNet, which was implemented as an R software package and followed the package vignettes. For each given combination of datasets, several hyperparameters of SmCCNet need be tuned to generate robust subnetworks, including subsampling proportion and , the penalty pair and , and weights of weighted SmCCNet. The subsampling procedure is typically repeated 1000 times with respect to certain penalty pair and datasets.

# Methods

## Overview



***Figure 3.*** *The data preprocessing workflow for SmCCNet given with RNA-Seq, 16S rRNA amplicon sequencing and quantitative phenotypes data. For the RNA-Seq data and 16S rRNA sequencing data, the read counts are normalized, transformed and down-sampled. Eventually, through the outlier screening and standardization, the whole dataset is SmCCNet ready.*

The SmCCNet network analysis approach developed by Shi et al 2019 was applied in this study to integrate host-transcriptome, gut microbiome with respect to a clinical trait[7]. As shown in ***Figure 3***, three data sets on the same cohort were prepared for SmCCNet. However, for all three datasets, there are special considerations, especially for the microbiome data: 1) The dataset should also be screened for missing data, because the algorithm only works for complete case analysis, where none of features and subjects has missing values. 2) Although SmCCNet does not require normality, it calculates the Pearson correlation between linear combinations of Omics features and the phenotype, which assumes finite variances and finite covariance, which holds for typical data. 3) The results of SmCCNet model can be impacted disproportionally by highly skewed data and/or zero-inflated data. Thus, it is necessary to include a transformation if the data are skewed. Omics4) CCA is designed for linear relationships among variables, which means one should be careful of curvilinear patterns. 5) Gene expression data obtained with HTS methods may be heteroscedastic, where the variability of gene expression is unequal across genes and highly expressed genes tend to have higher variability. 6) For microbiome data, the data is compositional where the RA of taxa sum to 1. 7) As a CCA based method, outliers can cause severe problems to the SmCCNet algorithm. In summary, the above challenges will be addressed through our data preprocessing framework.

Below we describe the different data preprocessing steps for the two data types transcriptome or microbiome respectively.

## Data Preprocessing: Normalization

We applied the trimmed mean of M values (TMM) normalization method from edgeR (version 3.24.3) or the DESeq2 normalization method (version 1.22.2) and compared the methods using relative log expression (RLE) plots[37, 38]. The TMM is a simple yet robust way to estimate the ratio of RNA production[39]. In general, the TMM method calculates the scaling normalization factor for each sample by accounting for library size and observed counts, while under the assumption that the majority of genes are not differentially expressed[39].

To account for the issue of different library sizes, the counts of OTUs are divided by corresponding library size to generate RA.

## Data Preprocessing: Transformation

The normalized gene counts still have skewness and heteroskedasticity issues. Hence, the regularized logarithm (rlog) transformation was applied to stabilize the variance[38]. The rlog-transformed values are the fitted values defined as:

**Equation 5**

Where the is the shrunken logarithmic fold change (LFC) on the base 2 scale for the th sample. All these procedures should be done at the global level of the transcriptome, while only subset of genes might then be selected for SmCCNet. The variance-stabilizing transformation (VST) method can be applied instead of rlog transformation if the estimated size factors are similar across libraries.

The microbiome dataset used in this study was generated by 16S rRNA amplicon sequencing. Hence, this dataset is compositional, regardless of being normalized to relative abundance (RA), and therefore the RA of taxa sum to 1. As a result, the assumption of true independence across features (taxa) cannot hold for relative counts or RA of OTUs[19, 20]. The centered log-ratio (clr) transformation was then applied on the counts to account for the compositional nature of the data. Given a sample consists of counts of OTUs (taxa), which is sample , the clr transformation for this sample is calculated as:

**Equation 6**

where , is the geometric mean and the output is a vector of transformed zero-centered values. The clr transformation is necessary to remove spurious correlations in the compositional data [40].

## Data Preprocessing: Down-sampling

For the transcriptome data, genes with low average counts and variation were filtered out before normalization. The transcriptome data may have much more features than the microbiome data, by orders of magnitude. Hence, stringent thresholds can be applied to address this unbalanced dataset issue. In practice, elbow plots can be used to control the number of genes after filtering.

For the microbiome data generated by 16S rRNA sequencing, the genus level should be selected for most detailed and accurate taxonomy according to the resolution of this method. Besides the compositional nature of the microbiome data, several other characteristics of microbiome data may cause problems to correlation-based network analysis methods as well, including the high sparsity (zero inflated) and the large dynamic range (heavy-tailed distribution). Thus, the data should be filtered by RA values (at least one library has RA of the given taxon more than the cutoff) and proportion of non-zero counts (prevalence of the given taxon is higher than the cutoff) to remove low abundance (including zero abundance) taxa. In practice, elbow plots can be used to control the number of taxa after filtering. Through this double thresholding procedure, the data is much less zero-inflated, contains less noise and more informative.

## Grubbs’ outlier test

As a Pearson correlation-based method, outliers can cause severe problems to SmCCNet. The human microbiome data generated by 16S rRNA amplicon sequencing contains many zeros and has a wide dynamic range, as several dominant taxa have large relative abundances[19, 20]. Therefore, the microbiome dataset may introduce the problem of outliers into the SmCCNet framework. Moreover, outlying observations can also be found in the host-transcriptome data, where extreme values may be observed for certain genes. Hence, outlier test will be applied to remove genes and taxa which contain extreme values.

Grubbs’s test tests whether the most extreme value from the mean is an outlier[41]. This test is based on the normal distribution assumption and can test an outlier at a time. The Grubbs test statistic is the largest standardized absolute deviation from the sample mean, which is defined as where is the sample mean, is the sample standard deviation and is the sample size. This is a two-sided test, and its significance level can be determined using where denoting the upper critical value of the t-distribution with degrees of freedom and a significance level of .

## Z-score standardization

Eventually, all three datasets should be standardized to mean 0 and variance 1 by feature. For a given feature, the standardization is defined as:

**Equation 7**

where is the mean of the given feature and is the corresponding standard deviation.



## SmCCNet Model Hyperparameter Tuning

The tuning parameters used in this study were summarized in **Table 1**. These hyperparameters were used by the model through the workflow described in **Figure 2**. The optimized values are only applicable to the motivating example described below. The sub-sampling proportions for two Omics datasets are decided based on the number of features in each Omics data, where larger Omics dataset tend to have more noise thus a relatively smaller proportion should be used in each subsampling. The total number of repeated subsampling can be 500, 1000 and 2000, and 1000 was chosen according to the manual. The results of SmCCNet are very sensitive to the penalties, and . Therefore, the best penalty pair should always be identified through cross validation with respect to specific datasets and values of all the other tuning parameters. The K in the K-fold cross validation (CV) is decided based on the sample size, that is a small K should be selected for a small sample size to have a stable average metric. Here, the best penalty pair should have the smallest average prediction error from the CV, which can be easily identified through the contour plot. The SmCCNet can be applied either as unweighted () or as weighted framework. In this study, both unweighted and weighted versions were used to test the impact of weights on the generated subnetworks. Currently, there is no mathematically rigorous method to identify the weights. However, the weights can be implied empirically with the summary Pearson correlations density plot. Eventually, the networks generated by SmCCNet will be compared to the networks generated by sparse CCA without a clinical phenotype.

## Summary of Hubs from Networks Generated

It is generally accepted that biological networks typically have a scale-free format where several hubs have many connections and most nodes only have one or a few connections. Our preliminary results revealed hubs in the integrative host-transcriptome-microbiome networks. Hubs from each network will be summarized and compared across different settings to evaluate the robustness of networks.

## Enrichment analysis

Enrichment analysis will be performed on the list of genes to provide valuable insight on biological function by systematically mapping genes to their associated biological annotations (such as gene ontology [GO] terms or Kyoto Encyclopedia of Genes and Genomes [KEGG] pathway)[42, 43]. We wil use clusterProfiler, which is implemented in R using the Bioconductor annotation data in GO.db and KEGG.db[6, 42].

# Motivating Example

## Dataset Generation and Preprocessing

The intestinal mucosal transcriptome was obtained through RNA-Seq from colon biopsies of a clinical study comparing HIV-1-uninfected () and age- and gender-matched, antiretroviral-therapy naïve persons living with HIV-1 (PLWH; ). The RNA-Seq library quality was verified using 2100 BioAnalyzer (Agilent, Santa Clara, CA) before loaded into a HiSeq 2500 by the Genomic Sequencing Core facility at the University of Colorado Anschutz Medical Campus. The quality screened FASTQ files were mapped to the current human genome assembly GRCh38 using *Hisat2* (Johns Hopkins University). The targeted gene lists core IFN-stimulated genes (core-ISGs) and IFNβ specific ISGs (beta-ISGs) were pre-defined by a preliminary *ex vivo* study, which were used as the candidate genes for integrative network analysis. Particularly, primary gut CD4 T cells ( donors) were treated for 6 hours *ex vivo* with individual IFN-Is normalized for IFNAR signaling activity. 1,980 IFN-regulated genes were detected by downstream quantitative PCR analysis pipeline. The core-ISGs were induced by all IFN-Is tested, while many other ISGs were not shared between the IFNα subtypes. Since IFNβ, but not IFNα, was upregulated during chronic HIV-1 infection, a broader spectrum of ISGs was induced by IFNβ by our observation. The IFN-stimulated genes were defined by 1.5-fold change cutoff, where the genes are only stimulated by IFNβ not by IFNα belong to the beta-ISGs. Two-group comparison differential expression (DE) analysis at the global level was called using normalized counts with edgeR according to the package vignettes and with an FDR (false discovery rate) of 5%. This analysis is to test whether a gene is significantly altered between healthy controls and HIV-1-infected donors.

The intestinal microbiome was obtained from colon biopsies of the same cohort but with a slightly different panel of donors. In our study, we used the 16S rRNA sequencing to generate taxonomic profiling. The microbiome contains data from HIV-1-uninfected () and age- and gender-matched, antiretroviral-therapy naïve persons living with HIV-1 (PLWH; ). DNA was extracted from colon biopsies and then 16S rRNA was generated then sequenced by Illumina paired-end sequencing on the Miseq platform with version 2.0 of the Miseq Control Software, using a 500-cycle version 2 reagent kit[5]. OTUs were produced by clustering sequences with identical taxonomic assignments at similarity threshold 99%. Finally, OTUs with relative abundance (RA) were generated at phylum, order, family, genus and species levels[5].

Clinical traits such as plasma biomarkers and gut immune cell counts were measured to characterize the chronic HIV-1 infection in the gut, including virological, microbial and immunological parameters. In particular, the panel of clinical phenotypes includes clinical readouts (plasma viral loads, blood and gut CD4+ T cell counts), markers of microbial translocation ( soluble CD27 (sCD27), lipoteichoic acid (LTA), sCD14), inflammatory biomarkers (C-reactive protein (CRP), IL-6) and intestinal epithelial barrier dysfunction (intestinal fatty acid binding protein (iFABP))[5]. Soluble CD14 was selected as the quantitative phenotype of interest in this study since it is the crucial indicator of gut leakage and microbial translocation in chronic HIV-1 infection[6].

There are 27 subjects that are shared across all the three datasets. This shared cohort contains HIV-1-uninfected () and PLWH (). To meet all these assumptions and requirements, all three datasets were further preprocessed as shown in ***Figure 4***.

The microbiome data was filtered by RA and prevalence at the same time, according to **Figure 4**. Proper cutoffs were detected with elbow plots and shown in results.

Then the Omics data were screened for outliers by Grubbs’ test at the feature level. Multiple testing correction was considered to recognize outliers by p value cutoff. Eventually, all three datasets are standardized.

## SmCCNet: Networks generation

With the SmCCNet-ready dataset. The model was fine-tuned according to the methods described previously. 4-fold cross validation was used in this study since the sample size here is 27.

**Generalizability**

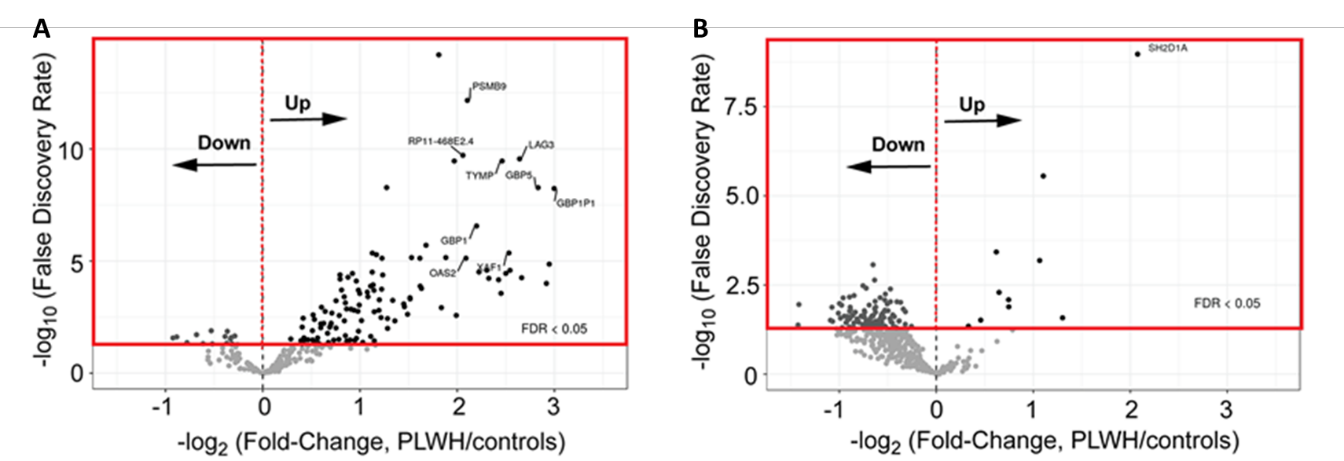
All the analyses were done by R (version 3.5.1) in RStudio (Version 1.1.456). All the R code can be found at <https://github.com/Guannan-Shen/Omics_Integration>.

# Preliminary Results



In this study, the two Omics datasets and the clinical phenotype data share a cohort of 27 subjects, which consists of 13 uninfected and 14 untreated, chronic HIV-1-infected subjects (**Table 2**). The two groups are age, sex and race matched.

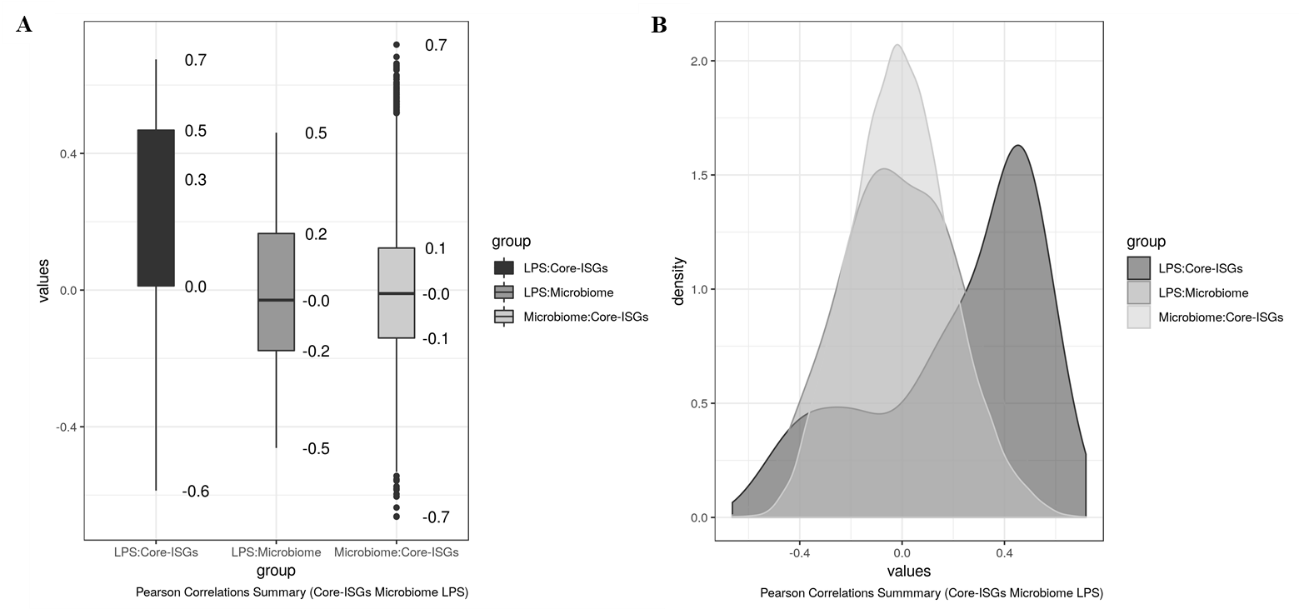
Two subsets of genes were selected from the host gut transcriptome, core-ISGs ( after filtering) and beta-ISGs ( after filtering). Both subsets have shown obvious alterations during HIV-1 infection and there is a qualitative difference in expression patterns between these two gene lists. Of these 246 core ISGs, 126 (51%) were significantly altered in HIV-1 infected versus uninfected individuals at 5% FDR (**Fig. 4A**). Majority (89%) of these altered core ISGs were upregulated in chronic HIV-1 infection. For beta-ISGs, 112 out of 406 genes were significantly altered in HIV-1 infection in the gut (**Fig. 4B**). In contrast, majority () of altered beta-ISGs were downregulated. Therefore, the ISGs may have an important role during chronic HIV-1 immunopathogenesis in the gut, which may be further addressed using the integrative network analysis approach.



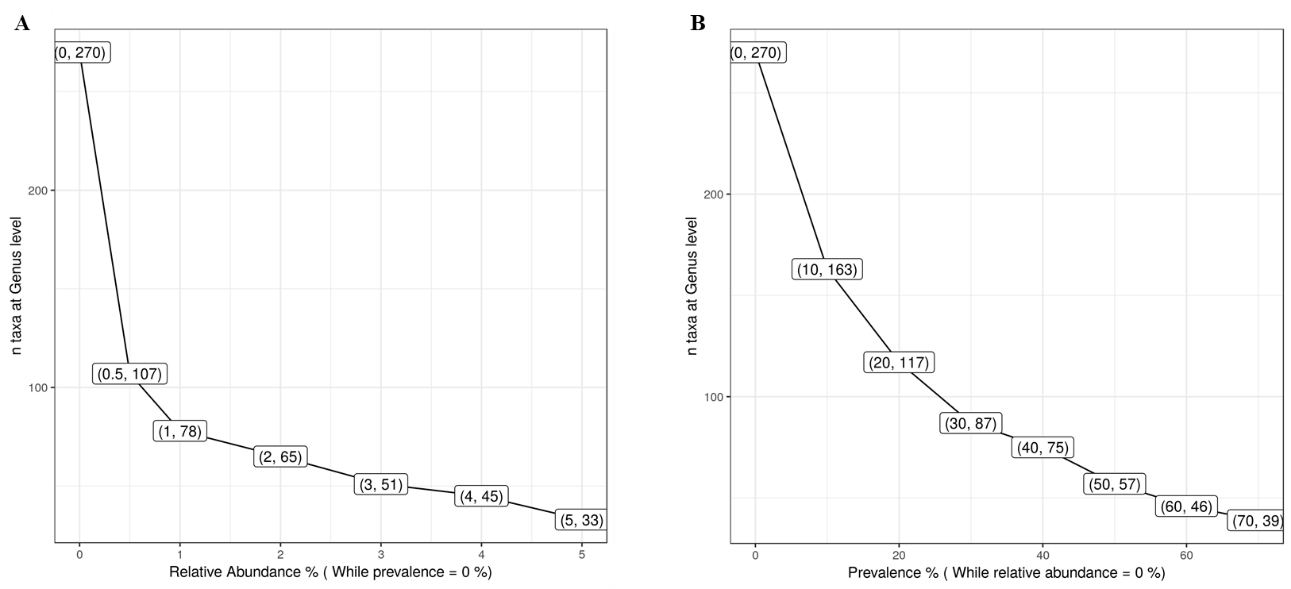
***Figure 4.*** *(A) Volcano Plot showing the FDR and fold-change criteria for altered core-ISGs. Majority of the core-ISGs were upregulated in PLWH relative to HIV-uninfected controls. (B) Volcano Plot for beta-ISGs. Majority of the beta-ISGs were downregulated in PLWH relative to HIV-uninfected controls.*

For the transcriptome data, compared with other popular normalization methods, the TMM normalization is the most effective method on removing the unwanted variations in this study, according to the RLE plots (figures not shown here). Through the rlog transformation, the overall distribution of the transformed read counts is roughly symmetric, compared with the heavily right skewed distribution before transformation (figure not shown here).

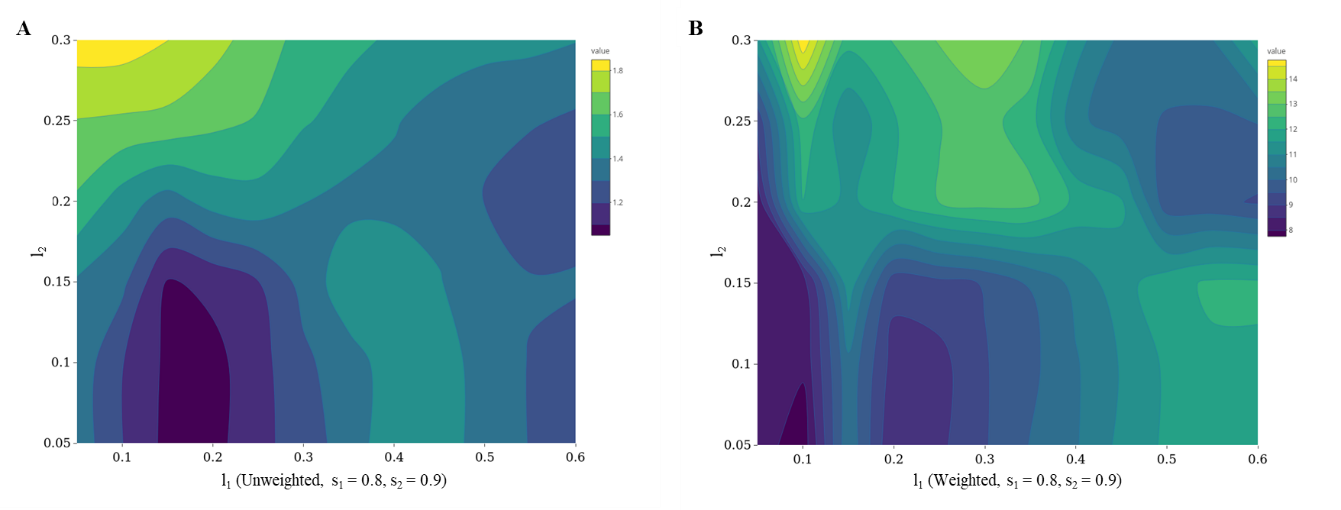
In terms of the microbiome data, we focused on the Genus level in this study, since the resolution of the data is limited to this level. The data was filtered by RA and prevalence at the same time. The RA thresholding requires that, for a given taxon, RA of at least one library (sample) is greater than the cutoff. The prevalence cutoff requires a proportion of samples have non-zero counts for a given taxon. The idea of elbow plot comes from the elbow method for the K-means clustering, which is a heuristic method to identify the optimal cutoff. The two cutoffs were tested separated by the elbow method, where the threshold of the RA was identified as 2% and the prevalence as 40% (**Figure 5, Table 3**). As a result, 54 out of 270 taxa at the Genus level were kept for SmCCNet.

The weights were decided empirically by the strengths of pairwise Pearson correlations among the three datasets (**Figure 6**). The results of core-ISGs subset of host transcriptome were shown here as the example. For each pair of datasets, the Pearson’s correlation coefficients were calculated for all possible pairs of features from two different datasets, and then the results were summarized by a boxplot and density plot. The summary plots suggest that the overall correlation between core-ISGs expression level and the plasma LPS level is much stronger than the others and the gut microbiome is not closely related to the other datasets. Therefore, we can upweight the microbiome data in the SmCCNet framework, which is one of the advantages offered by this algorithm. A very similar pattern is also observed for the beta-ISGs subset while the other two datasets were kept as the same. Depending on these results, weights , and were tested for weighted SmCCNet.

***Figure 5.*** *(A) Elbow plot showing the number of taxa left as the RA cutoff becoming more stringent, while the prevalence cutoff remains at 0%. (B) Elbow plot showing the number of taxa left as the prevalence cutoff becoming more stringent, while the RA cutoff remains at 0%.*

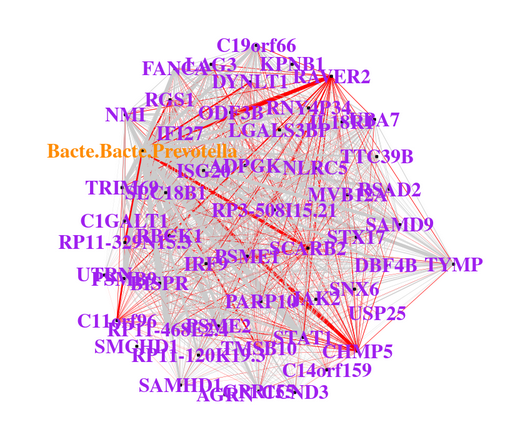


***Figure 7.*** *Contour plot to identify the best penalty pair. The average prediction errors from 4-fold cross validation were shown for each combination of penalties, while the subsampling proportions were always 0.8 for transcriptome and 0.9 for microbiome. (A) Contour plot of unweighted SmCCNet, using beta-ISGs transcriptome subset. (B) Contour plot of weighted SmCCNet, using beta-ISGs transcriptome subset. The dark blue represents smaller prediction errors, whose corresponding penalties are recommended for the algorithm.*



***Figure 6.*** *(A, B) Boxplot showing the quantiles and the median of pairwise Pearson’s correlation coefficients among three datasets. For each pair of datasets, all correlation coefficients at the feature level were summarized. The core-ISGs subset of host transcriptome was used here.*

In this study, the host transcriptome subsets contain features, while the gut microbiome only contains 54 taxa. This indicates the transcriptome may contain more noise, thus requires a greater down-sampling in the repeated subsampling procedure to guarantee the robustness of the algorithm. Eventually, through the grid search method (data not shown here), the subsampling proportion for (transcriptome) was confirmed as 0.8 and 0.9 for of (microbiome). The total number of repeated subsampling can be 500, 1000 and 2000, and 1000 was chosen according to the manual. The results of SmCCNet are very sensitive to the penalties. Therefore, the best penalty pair should always be identified through cross validation with respect to specific datasets and values of all the other tuning parameters. 4-fold CV was always used in this study since the sample size of the cohort is 27. The best penalty pair has the smallest average prediction error from the CV, which can be easily identified through the contour plot (**Figure 7**). In this case, penalty terms of for the microbiome and for the beta-ISGs transcriptome should be applied in the unweighted SmCCNet. Meanwhile, penalty pair should be applied in the weighed SmCCNet with weights . The smaller penalty term (ranges from 0 to 1) means the corresponding Omics data will be sparser in the generated subnetworks.



***Figure 8.*** *The network generated with preprocessed data and given hyperparameters of the SmCCNet model. The network is the integrated network of outliers removed core-ISGs and Genus level* microbiome data, with respect to LPS.

Given all the hyperparameters, the networks were generated through the SmCCNet. One of the networks was shown in **Figure 8**, where the red line represents the positive correlation and the gray line represents the negative correlation, and the strength of the correlations were measured with the width of edges. The Prevotella taxon at the genus level together with transcripts such as CHMP5, SCARB2 and RAVER2 might play an important role in the microbial translocation occurred during the chronic HIV-1 infection.

# Conclusions

Through the preliminary study, the pipeline of data preprocessing and model tuning has been established. The networks generated shows the taxon Prevotella and several transcripts such as CHMP5, SCARB2 and RAVER2 might be vital in HIV pathogenesis.

For the next step, the SmCCNet algorithm will be applied on the global level of host-gut-transcriptome, genus level of microbiome and clinical phenotypes such as sCD14, LPS and Lipoteichoic acid (LTA) respectively. Pathway analysis of gene sets and literature search on hubs will be used to explore the biological mechanisms of generated sub-networks.

Eventually, I will compare networks generated by different thresholds in microbiome down-sampling process and networks generated by different hyperparameters of SmCCNet algorithm such as the sub-sampling proportions, which hopefully will show the internal robustness of the SmCCNet pipeline when incorporating the microbiome data.

# References

1. Crakes KR, Jiang G: **Gut Microbiome Alterations During HIV/SIV Infection: Implications for HIV Cure.** *Front Microbiol* 2019, **10:**1104.

2. Barr T, Sureshchandra S, Ruegger P, Zhang J, Ma W, Borneman J, Grant K, Messaoudi I: **Concurrent gut transcriptome and microbiota profiling following chronic ethanol consumption in nonhuman primates.** *Gut Microbes* 2018, **9:**338-356.

3. Wang Q, Wang K, Wu W, Giannoulatou E, Ho JWK, Li L: **Host and microbiome multi-omics integration: applications and methodologies.** *Biophys Rev* 2019, **11:**55-65.

4. Byrd AL, Segre JA: **Integrating host gene expression and the microbiome to explore disease pathogenesis.** *Genome Biol* 2015, **16:**70.

5. Dillon SM, Lee EJ, Kotter CV, Austin GL, Dong Z, Hecht DK, Gianella S, Siewe B, Smith DM, Landay AL, et al: **An altered intestinal mucosal microbiome in HIV-1 infection is associated with mucosal and systemic immune activation and endotoxemia.** *Mucosal Immunol* 2014, **7:**983-994.

6. Dillon SM, Frank DN, Wilson CC: **The gut microbiome and HIV-1 pathogenesis: a two-way street.** *AIDS* 2016, **30:**2737-2751.

7. Shi WJ, Zhuang Y, Russell PH, Hobbs BD, Parker MM, Castaldi PJ, Rudra P, Vestal B, Hersh CP, Saba LM, Kechris K: **Unsupervised discovery of phenotype specific multi-omics networks.** *Bioinformatics* 2019.

8. Shive CL, Jiang W, Anthony DD, Lederman MM: **Soluble CD14 is a nonspecific marker of monocyte activation.** *AIDS* 2015, **29:**1263-1265.

9. Huang S, Chaudhary K, Garmire LX: **More Is Better: Recent Progress in Multi-Omics Data Integration Methods.** *Front Genet* 2017, **8:**84.

10. Chong J, Xia J: **Computational Approaches for Integrative Analysis of the Metabolome and Microbiome.** *Metabolites* 2017, **7**.

11. Faust K, Raes J: **Microbial interactions: from networks to models.** *Nat Rev Microbiol* 2012, **10:**538-550.

12. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG: **Using graph theory to analyze biological networks.** *BioData Min* 2011, **4:**10.

13. Bauer E, Thiele I: **From Network Analysis to Functional Metabolic Modeling of the Human Gut Microbiota.** *mSystems* 2018, **3**.

14. Integrative HMPRNC: **The Integrative Human Microbiome Project.** *Nature* 2019, **569:**641-648.

15. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, Wong GK: **Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics.** *Front Microbiol* 2016, **7:**459.

16. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE: **Conducting a microbiome study.** *Cell* 2014, **158:**250-262.

17. Ames NJ, Ranucci A, Moriyama B, Wallen GR: **The Human Microbiome and Understanding the 16S rRNA Gene in Translational Nursing Science.** *Nurs Res* 2017, **66:**184-197.

18. Morgan XC, Huttenhower C: **Chapter 12: Human microbiome analysis.** *PLoS Comput Biol* 2012, **8:**e1002808.

19. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ: **Microbiome Datasets Are Compositional: And This Is Not Optional.** *Front Microbiol* 2017, **8:**2224.

20. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vazquez-Baeza Y, Birmingham A, et al: **Normalization and microbial differential abundance strategies depend upon data characteristics.** *Microbiome* 2017, **5:**27.

21. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T: **Transcriptomics technologies.** *PLoS Comput Biol* 2017, **13:**e1005457.

22. Sandler NG, Wand H, Roque A, Law M, Nason MC, Nixon DE, Pedersen C, Ruxrungtham K, Lewin SR, Emery S, et al: **Plasma levels of soluble CD14 independently predict mortality in HIV infection.** *J Infect Dis* 2011, **203:**780-790.

23. Lien E, Aukrust P, Sundan A, Muller F, Froland SS, Espevik T: **Elevated levels of serum-soluble CD14 in human immunodeficiency virus type 1 (HIV-1) infection: correlation to disease progression and clinical events.** *Blood* 1998, **92:**2084-2092.

24. Ruan JW, Statt S, Huang CT, Tsai YT, Kuo CC, Chan HL, Liao YC, Tan TH, Kao CY: **Dual-specificity phosphatase 6 deficiency regulates gut microbiome and transcriptome response against diet-induced obesity in mice.** *Nat Microbiol* 2016, **2:**16220.

25. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R, Stempak JM, Gevers D, Xavier RJ, Silverberg MS, Huttenhower C: **Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease.** *Genome Biol* 2015, **16:**67.

26. Castro-Nallar E, Shen Y, Freishtat RJ, Perez-Losada M, Manimaran S, Liu G, Johnson WE, Crandall KA: **Integrating microbial and host transcriptomics to characterize asthma-associated microbial communities.** *BMC Med Genomics* 2015, **8:**50.

27. Perez-Losada M, Castro-Nallar E, Bendall ML, Freishtat RJ, Crandall KA: **Dual Transcriptomic Profiling of Host and Microbiota during Health and Disease in Pediatric Asthma.** *PLoS One* 2015, **10:**e0131819.

28. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, Belzer A, Bolourchi S, Caldera S, Fung M, et al: **Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults.** *Proc Natl Acad Sci U S A* 2018, **115:**E12353-E12362.

29. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9:**559.

30. Zhu X, Gerstein M, Snyder M: **Getting connected: analysis and principles of biological networks.** *Genes Dev* 2007, **21:**1010-1024.

31. Jackson MA, Bonder MJ, Kuncheva Z, Zierer J, Fu J, Kurilshikov A, Wijmenga C, Zhernakova A, Bell JT, Spector TD, Steves CJ: **Detection of stable community structures within gut microbiota co-occurrence networks from different human populations.** *PeerJ* 2018, **6:**e4303.

32. Witten DM, Tibshirani RJ: **Extension of Sparse Canonical Correlation Analysis with Applications to Genomic Data.** *Statistical Applications in Genetics and Molecular Biology* 2009, **8**.

33. Parkhomenko E, Tritchler D, Beyene J: **Sparse canonical correlation analysis with application to genomic data integration.** *Stat Appl Genet Mol Biol* 2009, **8:**Article 1.

34. Hotelling H: **Relations between two sets of variants.** *Biometrika* 1936, **28:**321-377.

35. Lin D, Calhoun VD, Wang YP: **Correspondence between fMRI and SNP data by group sparse canonical correlation analysis.** *Med Image Anal* 2014, **18:**891-902.

36. Chen J, Bushman FD, Lewis JD, Wu GD, Li H: **Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis.** *Biostatistics* 2013, **14:**244-258.

37. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

38. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.

39. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.

40. Cao Y, Lin W, Li H: **Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding.** *arXiv Statistics Methodology* 2016.

41. Grubbs FE: **Sample Criteria for Testing Outlying Observations.** *The Annals of Mathematical Statistics* 1950, **21:**27-58.

42. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *OMICS* 2012, **16:**284-287.

43. Tipney H, Hunter L: **An introduction to effective use of enrichment analysis software.** *HUMAN GENOMICS* 2010, **4:**202-206.