**Participants and Study Design**

13 healthy donors and 19 HIV-infected donors. The age and gender are balanced.

**Clinical Parameter Measurement**

Clinical parameters such as blood CD4-T cell counts (cells/ul), plasma viral load, tissue HIV RNA (per CD4 T cell), Tissue CD4 T Cell Counts (number/g), IL-6 (pg/ml), CRP (ug/ml), iFABP (pg/ml), sCD27 (U/ml), CD14 (ng/ml), LPS (pg/ml), LTA (OD), IFNα, IFNβ, CD4 T cells (% viable CD45+ cells) were used in this study.

**Target Gene Lists**

Two gene-lists: Interferon Stimulated Genes (ISGs) which contains 230 genes and Interferon-beta related genes which contains 423 genes were pre-defined by preliminary *in-vitro* data.

**Statistical Analysis**

**Differential expression analysis**

The RNA-seq raw counts data had gene-level read counts for 13 healthy participants and 19 HIV-infected participants. Each gene was labeled by an Ensembl ID, together with a gene symbol and its gene length. 19890 out of 43297 genes which have at least five counts per sample on average were kept for differential expression (DE) analysis. The trimmed mean of M values (TMM) normalization method from edgeR (version 3.24.3) was chosen from several other methods such as: Transcripts Per Kilobase Million (TPM), DESeq2 (version 1.22.2) etc. Two-group comparison DE analysis 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-infected donors.

**Correlation with Clinical parameters**

Linear regression models were fit in a gene by gene fashion to test the correlation between RNA expression levels and clinical parameters, with one clinical parameter included in the model at a time. The model was also adjusted for age and gender. The normalized RNA-seq counts were further transformed by variance stabilization method “regularized log transformation” from DESeq2. The further transformed RNA-seq counts were used as the outcome in the model.

Two gene-lists: Interferon Stimulated Genes (ISGs) and Interferon-beta related genes were defined by preliminary *in-vitro* data. Based on the results of DE analysis, 117 out of 230 ISGs and 130 out of 423 Interferon-beta related genes with FDR less than 5% were selected to test the association with clinical parameters by the above model, respectively. Genes were considered as significantly associated with the corresponding clinical parameter with an FDR cutoff at 5%. The correlations with each clinical parameter of each gene list were compared by descriptive methods such as summary tables and volcano plots. In addition, the difference in proportions of significant genes in each gene list were tested with Chi-squared test in R, as well as the difference in proportions of positive correlations of each gene list.

**PCA analysis**

Principal component analysis (PCA) was performed with R (version 3.5.1) in Rstudio (Version 1.1.456). Different first principal components (PC1s) was extracted to represent gene sets at transcriptome level and targeted gene lists level, respectively. Thus, the correlations between clinical parameters and gene sets were evaluated using separate linear regression models with PC1 as the outcome and the corresponding clinical parameter as the primary predictor, while adjusting for age and gender. Principal component (PC) plots were also created to illustrate the clustering of participants.

**Others**

The Benjamini-Hochbergprocedure (BH step-up procedure) controlling the FDR was used as the multiple testing correction method in this study. All the plots were generated by ggplot2 (version 3.1.1) package in R (version 3.5.1).

**References:**

**For edgeR:**

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140.

McCarthy, J. D, Chen, Yunshun, Smyth, K. G (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research, 40(10), 4288-4297.

**Rstudio:**

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.

**R:**

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

**ggplot2:**

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

**DEseq2:**

Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” Genome Biology, 15, 550. doi: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).

**FDR:**  
Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Yoav Benjamini and Yosef Hochberg, Journal of the Royal Statistical Society. Series B