Differential Gene Expression Analysis Between Actively and Latently HIV-Infected Hematopoietic Stem and Progenitor Cells

Authors: Callie Swanepoel, Maria Virgilio, Yidi Qin, Yuwei Bao

Abstract

Human Immunodeficiency Virus (HIV) is a life-long infection eventually resulting in death. Current methods of HIV therapy effectively eliminate all actively infected cells, however, latently infected cells will remain and drive the persistence of HIV infection. Therefore, identification and elimination of latently infected cells is critical to the treatment of HIV. Here, we sought to identify differentially expressed genes (DEGs) between HIV active and latent cells that may be used as biomarkers and/or targets of latently active cells. Using single cell RNAseq data (scRNAseq) generated from HIV infected hematopoietic stem and progenitor cells (HSPCs), we classified cells HIV active or latent based on an mCherry transgene marker, and performed differential gene expression between these two groups. We performed QC and normalization of scRNAseq data and surveyed three different DE tools: edgeR, limma and seurat. In addition we evaluated the type I error rate of edgeR and limma using permuted labels under our null hypothesis. Our analysis revealed variation in DEGs identified and type I error rate across the different DE models. EdgeR was found to have the lowest type I error rate, while limma predicted DEGs with greater power. Overall we identified 21 DEGs that by edgeR and limma that may explain the regulation of different cellular pathways between active and latent cells. Gene ontology revealed common DE pathways, including leukocyte activation tumor necrosis factor production, both of which are involved in HIV active states. All together our data suggests that latent cells are more likely to remain in an undifferentiated state and variation between DE tools should be carefully considered when conducting analysis of scRNAseq data.

1.Introduction

Human Immunodeficiency Virus (HIV) causes a persistent infection that is life-long and ultimately results in death. Routine HIV treatment with combination antiretroviral therapy (cART) eliminates nearly all actively infected cells. Nevertheless, the small reservoir of residual cells, some of which can remain dormant for long periods of time before becoming active and producing new virus particles, represents a crucial barrier to completely curing the disease. Discovering markers that identify latently infected cells or the biochemical factors that control latency activation could enable the effective use of approaches to target or activate latently infected cells and eliminate the viral reservoir.

Hematopoietic stem and progenitor cells (HSPCs) are long-lived cells that can differentiate into all of the cells in the hematopoietic lineage and populate the hematopoietic compartment throughout the life of an individual[1]. Importantly, they can also become infected by HIV and contribute to a persistent presence of virus in the blood (viremia) in patients by forming a latent population. Recent work from the Collins laboratory suggests that the global transcriptomic and epigenomic changes during hematopoietic differentiation affect viral latency and activation[2]. Therefore, we set out to identify transcriptional differences between actively and latently infected HSPCs.

To identify actively and latently infected HSPCs in an unperturbed state, we used a latency probe VT1 (Figure 1A). VT1 is a modified version of the HIV clone 89.6, derived from peripheral blood. The genome is triple-deleted for structural proteins necessary for infection with truncations in gag, pol, and env, making the virus exclusively capable of single-round infection. This virus expresses mCherry as a Gag-mCherry fusion protein, and eGFP driven by the spleen focus forming virus promoter (pSFFV) inserted in place of the HIV accessory protein nef open reading frame (ORF). In validation studies, VT1 expressed GFP in nearly all infected cells (both active ($mCherry^{high}$) and latent ($mCherry^{low}$). Thus, VT1 allows positive selection of both latently and actively infected cells without perturbing the natural state of the infected cells.

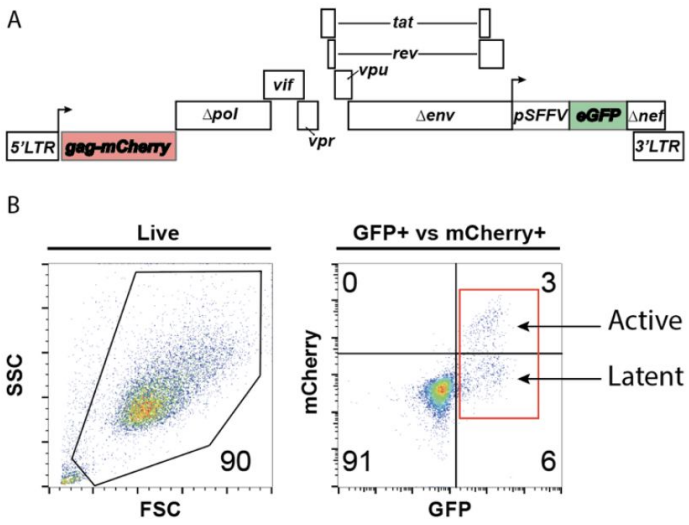


Figure 1
A: Map of HIV Dual Reporter Construct, VT1. HSPCs. Cells analyzed by scRNA-seq outlined in the red box;
B: Fluorescence activated cell sorted VT1 infected.

To investigate gene expression differences between latently and actively infected HSPCs, we performed Single-cell RNA sequencing (scRNA-seq) from a mixture of latently and actively infected HSPCs (Figure 1B). We then used three different methods to investigate differential gene expression between latently and actively infected HSPCs: Seurat, EdgeR, and limma. Top differentially expressed genes from each analysis were to determine gene ontology. Using EdgeR, the biological processes related to an increase in regulation of tumor necrosis factor genes and cellular activation in active HIV expressing HSPCs compared to latent cells. Similarly, the top biological processes in actively infected cells revealed positive regulation of viral entry in host cells, and regulation of leukocyte chemotaxis. Top DEGs that overlapped from all three analyses revealed a set of genes related to B and T cell development, cell cycle progression and apoptosis, among others.

2. Methods

2.1 Cell culture and scRNA-seq data generation

Briefly, CD133+ HSPCs were isolated from a de-identified donor of cord-blood. Cells were cultured for 4 days at 37ºC, spin-infected with replication-incompetent HIV, VT1, and cultured for

3 days post infection. After 7 days in culture, some of the cells have begun to differentiate and have lost expression of CD133 (also known as *PROM1).* Cells expressing GFP (infection marker) were sorted and sent for scRNA-seq processing using the 10X Chromium v3 kit and sequenced on the Illumina Novaseq 6000. Transcripts were aligned to hg38nc and VT1 (mCherry, GFP, HIV) genomes. Transcript counts were used for downstream analysis.

2.2 Quality control, data filtering and normalization

All quality control, data filtering and normalization was conducted using *Seurat* (version 2.4)[3]. In the first round of filtration, all genes expressed in less than 3 cells and cells with less than 200 detected genes were removed. We then calculated the number of genes, unique molecular identifiers, and the the percentage of mitochondrial genes present for each cell sample. Cells expressing more than 8500 genes, and cells with mitochondrial gene percentage higher than 12.5% were further filtered out. After removing unwanted cells from the dataset, we normalized the gene expression measurements for each cell by the total expression, multiplied this by 10,000 as the scaling factor, and log-transforms the result. Next, we calculated the average expression and dispersion for each gene, placed these genes into bins, and then calculated a z-score for dispersion within each bin. This step helped us find the top 3000 genes that are highly variable in gene expression.

To remove unwanted source of variation, such as technical noise, batch effects, or even biological sources of variation (cell cycle stage), Seurat also constructs linear models to predict gene expression based on user-defined variables. We then regressed these signals out of the analysis to improve downstream dimensionality reduction and clustering. Since the nature of Single-Cell RNA-Seq data matrix is very sparse, we performed Principal Component Analysis on the data matrix and reduced the dimensions to 20. Finally, a TSNE, a non-linear dimensional reduction algorithm was run to detect cell clusters.

2.3 Differential Gene Expression analysis

Single cells were divided into one of two groups based on expression of mCherry, a transgene marker for HIV expression in our construct. Cells expressing high levels of HIV will also express high levels of mCherry expression and are considered “HIV active” cells. Cells expressing low or no HIV will express low or no mCherry and are considered “HIV latent” cells. Differential

expression (DE) analysis was conducted between HIV active and HIV latent cell types using three different DE tools, two of which were designed for bulk RNAseq: *edgeR* (version 3.24.3)[4,5] and *limma* (version 3.38.3)[6], and one tool specific to scRNAseq : *Seurat* (version 2.4). Additionally, DE was assessed both with and without cluster as a covariate in the edgeR and limma analyses. For edgeR and limma DE analysis, normalized and filtered scRNAseq metadata was read into R, and read count data was extracted as a data frame to generate a DGEList data object used for downstream DE analysis. The DGEList contains read counts, gene annotations and sample information. Finally, each cell in our dataset represents an individual sample.

2.3.1 Seurat

After the Seurat object was pre-processed, it was divided into two groups, Clusters 0, 1, 4, 6, and 10 as the low HIV expression group, and the rest as high HIV expression group. Differentially expressed genes were found between between these two cluster groups, and the false discovery rate (FDR) adjusted p-values were calculated for top genes. Genes with an adjusted p-value <0.05 were used for further downstream analysis.

2.3.2 EdgeR

Since read counts were normalized and filtered using Seurat, no additional filtering or normalization was necessary in edgeR. Instead, our DGEList was used to estimate dispersion. EdgeR uses the Cox-Reid profile-adjusted likelihood method to estimate dispersion for experiments with multiple factors (such as HIV active/latent and cluster), and accounts for these factors by fitting a generalized linear model (GLM) with a design matrix. Our design matrix included an intercept, HIV active/latent status, and cluster. To calculate the common dispersion, trended dispersion and tagwise dispersion tagwise dispersion the estimateDisp() function with the DGEList and design matrix as arguments. Finally, a negative binomial GLM was fitted using the glmFit() function with the DGEList and design matrix as arguments.

Once dispersion estimates were calculated and a negative binomial GLM was fitted, significant DEGs were determined using likelihood ratio tests . The null hypothesis of testing is no DE in genes. Using the function glmLRT() with our fitted model as the argument, DE was determined between HIV active and latent cell types. The output of our DE analysis was a DGELRT, but

relevant data (i.e. logFC, PValues, FDR and gene annotations) were extracted as a dataframe. Genes with a FDR less than 0.05 were selected for further downstream analysis.

2.3.3 Limma

Similar to edgeR, no additional filtering or normalization was necessary using limma. Additionally, our design matrix included an intercept, HIV active/latent status, and cluster as in edgeR. Unlike edgeR, we followed the limma-trend approach which accounts for variance in the data by converting read counts to log2 counts per million (logCPM) and the mean-variance relationship is modeled using an empirical Bayes prior trend. Read counts were converted to logCPM values using the function logCPM() with the DGEList as an argument and the log argument set to FALSE (since our input data is already log-transformed). To determine differential expression, the logCPM values were used in the standard limma pipeline that gives more weight to fold-changes in gene ranking. The functions used to determine differential expression by fold-change are as follows: lmFit() with logCPM values and the design matrix as arguments, followed by tret() with the fitted model and the default log fold-change cutoff used as arguments. Relevant data (i.e. logFC, p-values, FDR and gene annotations) were extracted from the output as a dataframe. Genes with adjusted p-values (we used FDR correction methods to keep consistency with edgeR method) less than 0.05 were used for further downstream analysis.

2.3.4 Result visualization

To visualize the different methods used to determine DE, volcano plots for each method were generated as well as venn diagrams demonstrating the overlap of significantly DE genes between all methods used. Volcano plots were generated using the plot() function with logFC and and -log10(p-values) as the x- and y-axes respectively. The points() function was used to label data points with a p-value <0.05 (red) and FDR <0.05 (green) and the texty() function was used to label points with a p-value <0.05 and FDR <0.05 with their gene symbol. Volcano plots were generated for edgeR and limma DE methods using data analyzed either with or without cluster as a covariate.

To determine the overlap of significantly DE genes (p-value <0.05 and FDR <0.05) from all methods used, Venn diagrams were generated using Venny (version 2.1)[7], an online bioinformatics tool. Since we were not able to use cluster as a covariate when calculating DE

using Seurat, a comparison of DEGs across all three methods was done using DE data without cluster as a covariate. In addition, it is also important to note that DE genes determined by Seurat were considered significant only if p-value < 0.05 since we did not have FDR data available for consideration. Finally, a Venn diagram was generated to compare significantly DE genes between edgeR and limma using DE data with cluster as a covariate.

2.3.5 Method Evaluation

To evaluate the results using differential gene expression analysis approaches, we permuted group labels under null hypothesis H0, which assumed there is no expression difference of genes between two groups. To be more specific, we used random binomial distribution with the same group proportion as real data to generate group labels. Then for each of the methods, we fitted models and carried out tests using the permuted data. Distributions of p-value and QQ plots were conducted to validate if assumption of uniform distribution of p-value under null hypothesis was met. In addition, by 1000 times data label permutations, we calculated the rate of false-positive predictions (type I error rate) identified by each method at an adjusted P value cutoff (or FDR) of 0.05.

2.4 Gene Ontology Analysis

Based on the results of differential gene expression analysis, we took the top highly differentially expressed gene (DEG) subsets using Serutat, limma, and EdgeR and performed Gene Ontology (GO) analysis. To do so, we used the GO analysis tool, Gene Ontology enRIchment anaLysis and visuAlization tool (GORILA; http://cbl-gorilla.cs.technion.ac.il/)[8,9]. Using the running mode of two unranked list of genes, we took the top DEGs with an FDR <0.05 from each analysis and called them the “target set” and the entire list of genes as the “background set”. We then chose the biological process as the ontology. Taking the GO analysis terms and their adjusted p-values from GORILA, we used *reduce and visualise gene ontology* (REVIGO; http://revigo.irb.hr/) to visualize the GO term results. The GO terms and the associated FDR q-value calculated from GORILA were used to visualise the significant GO pathways as a scatterplot. Results were summarized in Table1.

3. Results

3.1 Quality Control

The raw data obtained includes 19433 genes and 4418 cell samples. After the first round of rough filtration, the violin plot of number of genes, unique molecular identifiers, and the the percentage of mitochondrial genes present for each cell sample was shown in Figure 2A. Most of the cells expressed about 3000 to 8500 genes, and 100,000 unique RNA copies, but there is also a group of cells that only expressed less than 1000 genes and 10,000 unique RNA copies. These cells are likely to be the undifferentiated stem cells, which have lower gene expression activity than the differentiated ones. Therefore, we kept all the cells with less than or equal to 8500 expressed genes. Also in the same plot, most of the cells have less than 12.5% mitochondrial genes expressed, which indicates the good quality of our raw data. We removed the cell samples with mitochondrial gene percentage higher than 12.5%. At the end of filtration, we obtained 3000 genes and 4294 cell samples.

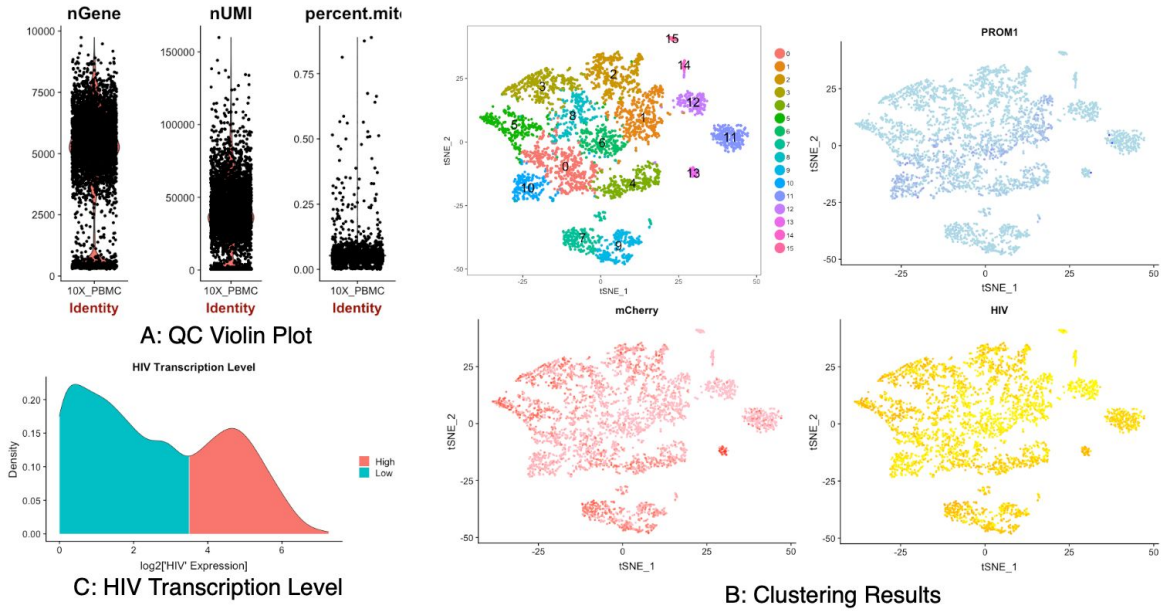


Figure 2

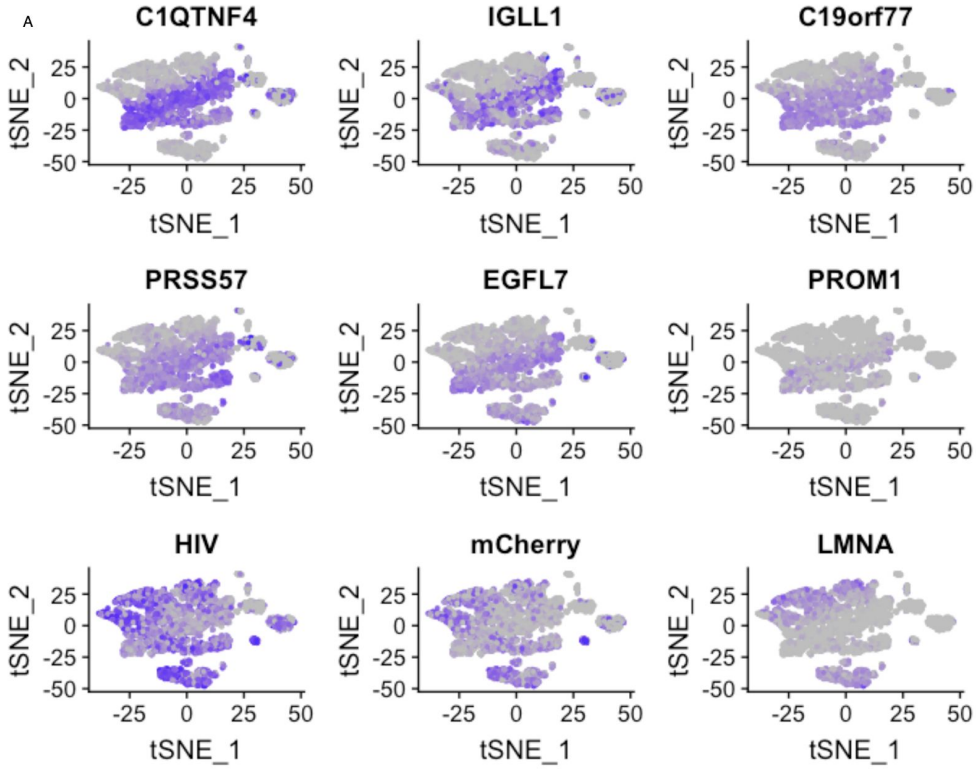
A: QC violin plot. Number of genes, number of UMI, and mitochondrial gene percentage for each cell; B: Clustering results. All cells were clustered into 16 groups. They were also colored based on gene expression level of PROM1, mCherry, and HIV; C: HIV transcription level and the cut off between high VS. low.

The clustering result shows Figure 2B there are 16 distinctive groups of cells. We also colored the same TSNE plot according to the expression level of gene HIV, PROM1, and mCherry, where darker color indicates higher expression. It is clear that the upper and lower clusters have high expression of HIV and mCherry, whereas the middle clusters have high expression of PROM1. The whole dataset was then split into two according to the log HIV expression level with a cutoff at 3.5 Figure 2C, with 2812 cell samples in the low HIV expression group, and 1482 cells in the high HIV expression group.

3.2 Results of DE analysis

3.2.1 Results of Seurat

The DE analysis between the two cluster groups output 274 significant genes. After adjusting for False Discovery Rate (FDR), 233 genes are shown as significant. As mentioned in the quality control session, HIV and mCherry have higher expression in the upper and lower clusters in the Figure 2B, whereas PROM1 has higher expression in the middle clusters. In Figure 3A, we also plotted top 6 significant genes, together with HIV, mCherry and PROM1. It shows that these top genes, C1QTNF4, IGLL1, C19orf77, PRSS57, EGFL7, and LMNA also showed to be highly expressed in the same area as either HIV, mCherry or PROM1.



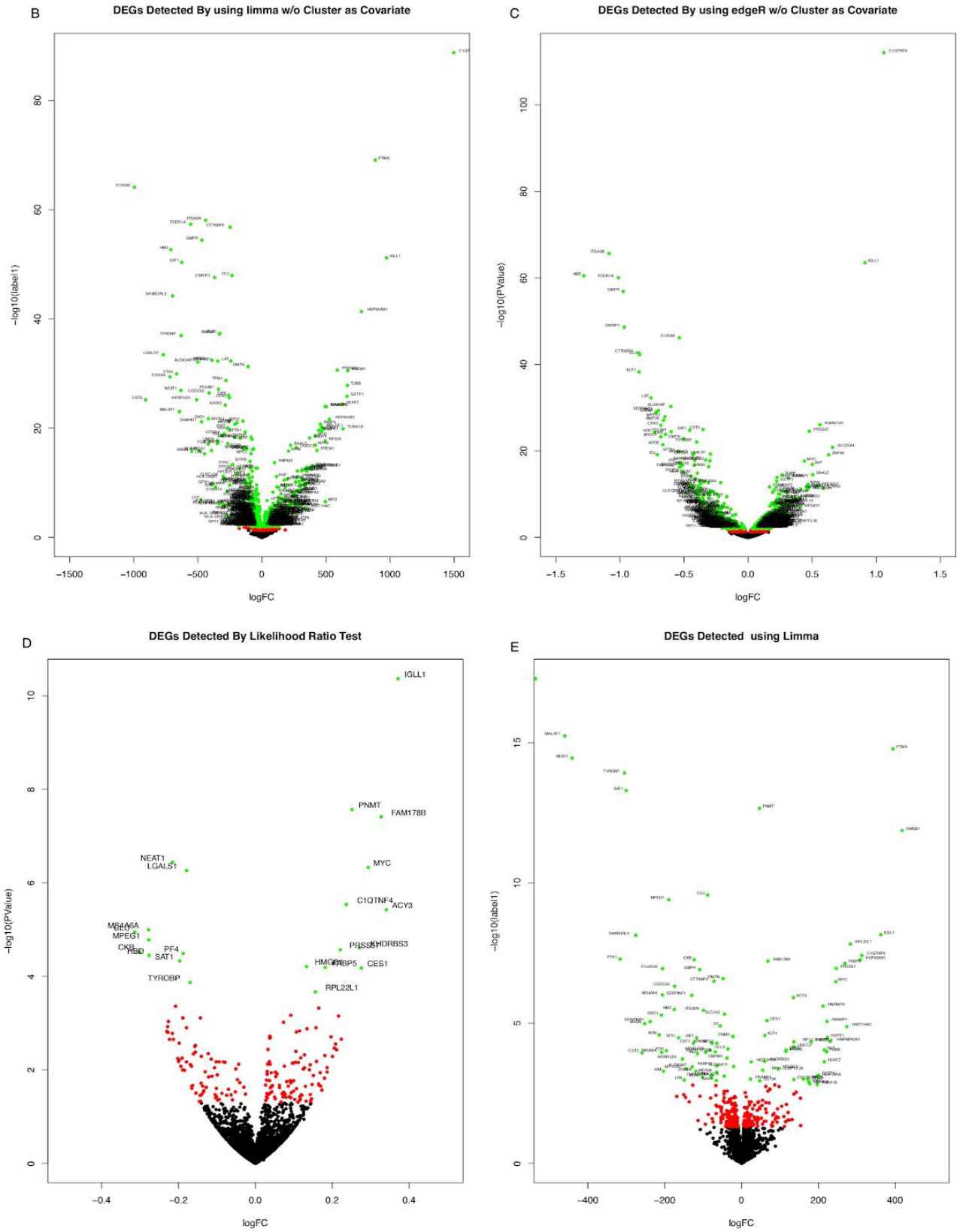


Figure 3

A:Results of Seurat. The expression level of op 6 significant genes together with HIV, mCherry, and PROM1 in the clustering graph; B: DEGs detected by using edgeR without cluster as covariate; C: DEGs detected by using limma without cluster as covariate; D: DEGs detected by likelihood ratio test using edgeR; E: DEGs detected using limma.

3.2.2 Result of edgeR and limma

By fitting the model without adjusting for cluster, the log fold change of each genes and their significant levels using edgeR and limma were presented in Figure 3B and Figure 3C, respectively. After cluster adjustment, the DE analysis results of edgeR and limma were shown in Figure 3D and Figure 3E, respectively. In all of the four volcano plots, significantly DEGs (FDR < 0.05) were colored by green.

3.2.3 Result comparison

Using the datasets generated without adjusting for cluster, we identified 537, 635, and 233 DEGs using edgeR, limma and Seurat respectively. Among all three methods, 61 DEGs were shared (Figure 4). When including cluster in our model, edgeR identified 22 DEGs and limma identified 95 DEGs between HIV active and HIV latent cell types. Of those genes 21 were overlapping between the two methods.

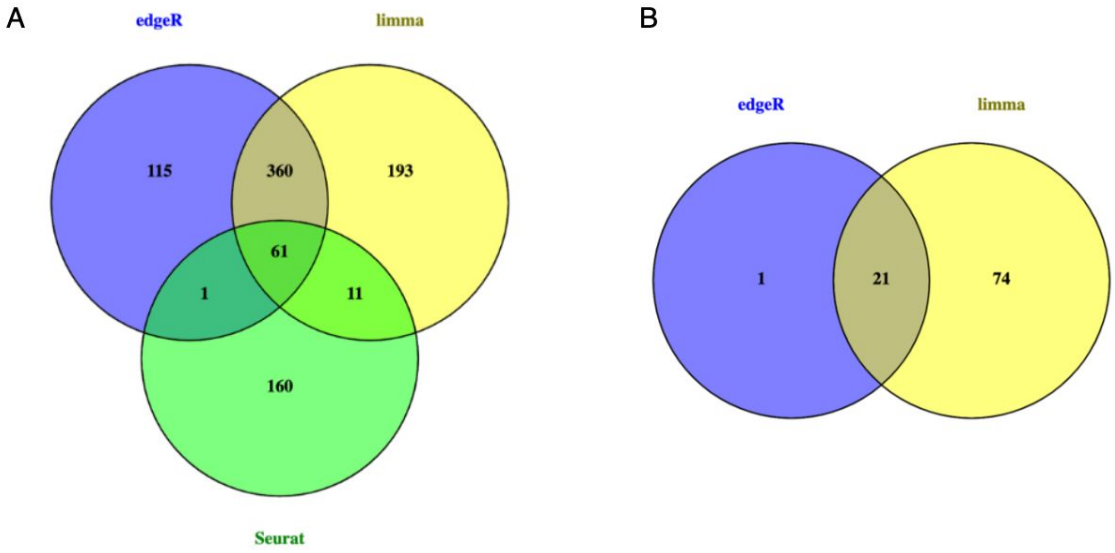


Figure 4

A: DEG comparison among all three methods without clusters as covariate; B: DEG comparison among two methods with clusters as covariate

We also observed that 421 genes were detected by using unadjusted models of edgeR and limma methods, which indicates that about 400 genes did not significantly differentially express after adjusting for cluster. Among the 21 significantly differentially expressed genes identified by

the adjusted models, 18 of them were also included in the 421 overlapped genes using unadjusted models. The adjusted p-value for all of these 18 genes are much larger after adding the cluster into fitted models. Therefore, both of the large difference in counts of significant DE genes and the increase of p-values of 18 final selected genes illustrated the high confounding effect of cluster.

Interestingly, immunoglobulin lambda like polypeptide 1, *IGLL1* was one of the top up-regulated genes in HIV latent cells compared to HIV active cells using both edgeR and limma. Literature suggest that this gene is involved in transduction of signals in cellular proliferation and differentiation from proB to preB cell stage[12], thus suggesting the HIV active cells have already completed differentiation, while HIV latent cells lag behind and are still undergoing differentiation.

3.2.4 Methods Evaluation

Since Seurat method defined its own group labels according to the relationship of clusters and HIV expression level label, we did not compared this method with edgeR and limma methods together. In regards to the evaluation and comparison of edgeR and limma methods, we found that the p-values for both methods were largely uniform according to the histogram of Figure 5A. Moreover, QQ plots in Figure 5B also provided us with more powerful evidence of distribution of p-value. For each of genes, the observed -log p-values (Y-axis) were plotted against the -log expected P-values under the null hypothesis (X-axis). The black diagonal line denotes the pattern under null hypothesis (uniform distribution of p-value). In the QQ plot of the two methods, most of the points are observed to be located in 95% CI interval (grey shaded area) of diagonal line and there is a slight deviation below the line at the tail. This illustrates that by using limma and edgeR method, it is more like to have a type II error (false negative) rather than type I error and thus the power to find detect differentially expressed genes may not as large as our expectation. Type I error rates using different number of simulation were calculated (Figure 5C). As simulation numbers increases from 100 to 1000 by 100, the type I error becomes stable around 0.05 for limma and around 0.047 for edgeR.

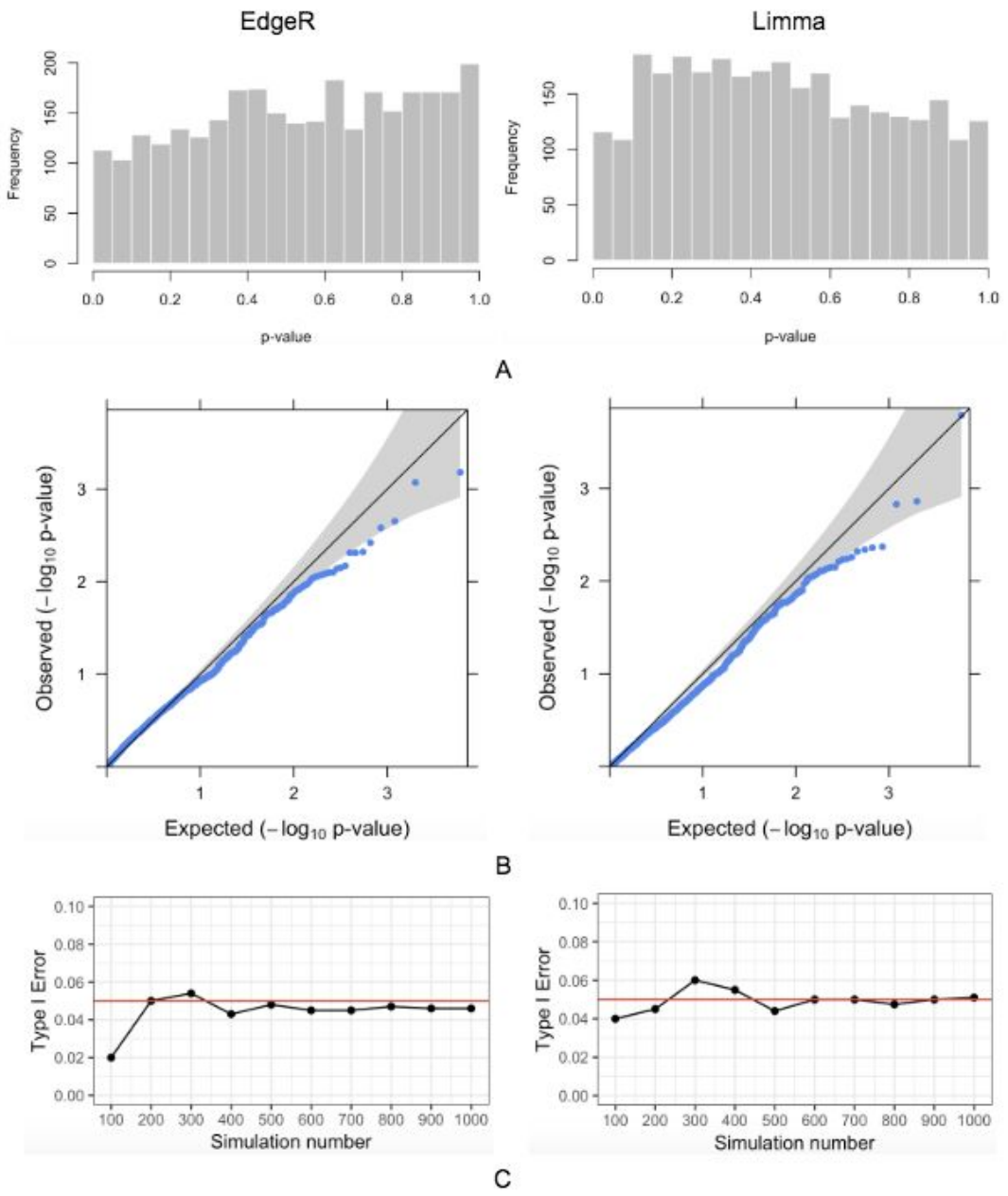


Figure 5

*Evaluation results comparing performances of edgeR and limma methods were based on permutation data (generated under null hypothesis) A: Distribution of p-value B: QQ plot comparing observed p-value distribution with expected p-value distribution (uniform distribution) C: Type I error rate of two methods by different simulation numbers

3.3 Gene ontology

Differential gene analysis between HIV high- and low-expressing HSPCs by Seurat, limma, and edgeR produced a set of DEGs that passed an FDR cutoff of 0.05. Performing gene ontology using GORILA with these top DEG genes yielded similar biological pathways that were important for HIV control between the HIV high and low cells. Top Seurat GO terms primarily involved immune system activation (Table1-Seurat), similarly with limma (Table1-limma). EdgeR top GO terms were more specific to increases in tumor necrosis factor (TNF) cytokine production and cell cycle regulation (Table1-edgeR). TNF cytokines, including TNFa, are powerful stimulants for hematopoietic and immune related cells [10]. As HIV relies on the host cellular machinery in order to replicate itself, like many viruses, HIV “hijacks” the cellular environment for its own purposes. This involves takeover of cellular trafficking networks and regulation of cell cycle progression in infected cells, particularly stopping progression at the G2/M stage, which promotes early steps in HIV infection [11]. Interestingly, GO analysis using the combined 22 genes that overlapped between all three analysis types appeared to closely resemble the top GO terms identified using edgeR (Table1-Combined). Overall, each of the analysis methods used revealed similar genetic pathways related to immune activation and modifications to the cellular environment.

Analysis Type	GO term	Description
Seurat	GO:0002376	immune system process
	GO:0045055	regulated exocytosis
	GO:0002252	immune effector process
	GO:0002366	leukocyte activation involved in immune response
	GO:0002263	cell activation involved in immune response
limma	GO:0045055	regulated exocytosis
	GO:0002252	immune effector process
	GO:0006887	exocytosis
	GO:0030162	regulation of proteolysis
	GO:0002376	immune system process
edgeR	GO:0032760	positive regulation of tumor necrosis factor production
	GO:1903557	positive regulation of tumor necrosis factor superfamily cytokine production
	GO:0032680	regulation of tumor necrosis factor production
	GO:1903555	regulation of tumor necrosis factor superfamily cytokine production
	GO:0043277	apoptotic cell clearance
Combined	GO:0032760	positive regulation of tumor necrosis factor production
	GO:1903557	positive regulation of tumor necrosis factor superfamily cytokine production
	GO:0032680	regulation of tumor necrosis factor production
	GO:1903555	regulation of tumor necrosis factor superfamily cytokine production
	GO:0050867	positive regulation of cell activation

Table 1. Top five GO terms from each analysis type and the overlapping genes from combined top-DEGs. GO terms produced using GORILA.

4. Discussion

To conclude our findings, a total number of 21 significantly differentially expressed genes based on high and low HIV expression level were detected in this study. The results of the DE analyses and methods evaluation indicate that the method chosen will result in the identification of different DE genes with a subset that are shared among all methods. In addition, the power and type I error varies depending on method used.

For differential expression analysis, cluster of cell was a major confounder of HIV high vs low expressing groups and had a large effect on the testing results if not included as a covariate. This clustering of cell may be explained by different cell type or other similar features of cells. Therefore, one of the main focuses of our study design was to eliminate the confounding effect of cluster. In this way, Seurat method may not perform as well as edgeR and limma methods, given that it failed to conduct DE adjusted for cell cluster so that led to a result with higher false positive rate. Comparing the performance of limma and edgeR methods, we found that by edgeR has a relative low type I error, while limma detected more significant genes than edgeR, which could be explained by a higher power. These results corroborate findings from Soneson et al. (2017).

In hindsight, it may have been more appropriate to use the voom approach from limma to conduct our DE analysis. The voom approach is used in situations where there is greater than 3-fold variation in library size for each sample, which is likely the case for scRNAseq when each cell represents a sample and there is significant drop-out. Furthermore the voom approach is designed for DGELists in which the read counts have already been normalized and filtered. Together this may help to explain the abnormally large logFC results obtained from the limma-trend approach. In future experiments the limma voom approach should be considered.

One major limitation of our study was the limited number of differentially expressed genes finally found, which was not large enough to support informative pathway analysis using GO. Since we selected the top 3000 highly expressed genes and did not considering the low expression genes for differential expression analysis, this may account for limited number of DEGs included in our final result. In addition, dispersion calculations in edgeR were computationally time consuming, thus limiting the number of simulations that could be used to calculate type I error. This may fail

to give a stable value and impact our results interpretation. Finally, due to the framework limitations, differential expression analysis for Seurat was performed on two groups separated by visual cluster selection. The difference in active VS. latent separation methods between Seurat and the other two methods might account for some difference in significant genes found.

In future experiments, improvements in Seurat analysis could be made. In addition to using cluster label as a covariate, we could also perform differential expression analysis for active vs. latent cells within each cluster to understand cell type specific gene expression differences introduced by HIV infection. From a biological perspective, it may be interesting to use the differential biological pathways we identified through scRNAseq to manipulate latent cells in order to force them to become active and differentiate. Once active and differentiate these cells could be more readily targeted by current HIV therapies and the pool of latent cells could be depleted.

Acknowledgements

We want to thank Kathleen Collins for permission to use the data for this project; Josh Welch for his expertise and advice working with scRNA-seq data; Maureen Sartor for her statistical aid and advice; and all of the teaching staff for Bioinformatics 545 winter 2019 term.

Project contributions from each person

Maria: generated raw data, performed differential gene expression analysis between HIV high and HIV low expressing cells using limma, and Gene Ontology analysis.
Yuwei: performed quality control on raw data, differential gene expression analysis between HIV high and HIV low expressing cells using Seurat, and Gene Ontology analysis.
Yidi: Evaluated DE methods by permutation data and calculated type I error for edgeR and limma methods.
Callie: performed differential gene expression analysis between HIV high and HIV low expressing cells using edgeR, created volcano plots and Venn diagrams.

References

[1] Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. Cell. 2008;132(4):631–644. doi:10.1016/j.cell.2008.01.025

[2] Sebastian, N. T. et al. CD4 is expressed on a heterogeneous subset of hematopoietic progenitors, which persistently harbor CXCR4 and CCR5-tropic HIV proviral genomes in vivo. PLoS Pathog 13, e1006509, doi:10.1371/journal.ppat.1006509 (2017)

[3] Stuart and Butler et al. Comprehensive integration of single cell data. bioRxiv (2018)

[4] 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.

[5] 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

[6] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research, 43(7), e47.

[7] Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html

[8] Eran Eden*, Roy Navon*, Israel Steinfeld, Doron Lipson and Zohar Yakhini. "GOrilla: A Tool For Discovery And Visualization of Enriched GO Terms in Ranked Gene Lists", BMC Bioinformatics 2009, 10:48.

[9] Eran Eden, Doron Lipson, Sivan Yogev, Zohar Yakhini. "Discovering Motifs in Ranked Lists of DNA sequences", PLoS Computational Biology, 3(3):e39, 2007.

[10] Dembic, Z. & Dembic, Z. Cytokines Important for Growth and/or Development of Cells of the Immune System. Cytokines Immune Syst. 263–281 (2015). doi:10.1016/B978-0-12-419998-9.00008-0

[11] Groschel, B. & Bushman, F. Cell cycle arrest in G2/M promotes early steps of infection by human immunodeficiency virus. J. Virol. 79, 5695–704 (2005).

[12] Nomura, K. et al. Genetic defect in human X-linked agammaglobulinemia impedes a maturational evolution of pro-B cells into a later stage of pre-B cells in the B-cell differentiation pathway.

[13] Soneson, C., & Robinson, M. D. (2017). Bias, Robustness And Scalability In Differential Expression Analysis Of Single-Cell RNA-Seq Data. doi:10.1101/143289

[14] Supek F, Bošnjak M, Škunca N, Šmuc T. "REVIGO summarizes and visualizes long lists of Gene Ontology terms" PLoS ONE 2011. doi:10.1371/journal.pone.0021800