Exploratory Data Analysis of HIV infection dataset

Denis Torre

January 24th, 2017

Overview

Aim

The aim of this report is to perform Exploratory Data Analysis (EDA) on a HIV infection dataset provided by Dr. Weijia Zhang.

Data

The data was obtained by performing RNA-sequencing on 34 biological samples of podocytes under three experimental conditions: HIV infection, control GFP infection, and untreated. Raw readcounts were normalized using Variance Stabilizing Transform (VST) from the DESeq2 package.

The samples can be divided in two groups:

1. *13 samples* from **patient-derived cell-lines**, divided in three batches. 3 samples are untreated, the other 10 are infected with HIV at different timepoints (6h, 125, 24h, 48h).
2. *21 samples* from **patient-derived primary podocytes**, collected from 10 different patients. The samples are divided in HIV infection, control GFP infection and control.

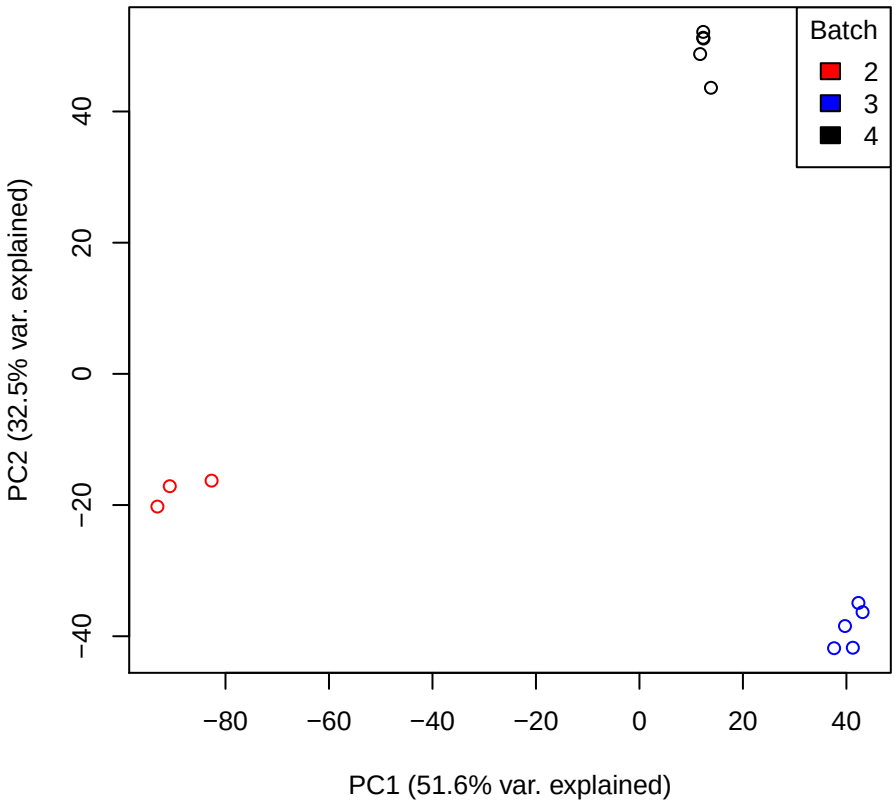
PCA Analysis

Rationale

A PCA analysis was performed to visualize the data, assess the separation of samples from different experimental conditions, and assess presence of batch effects.

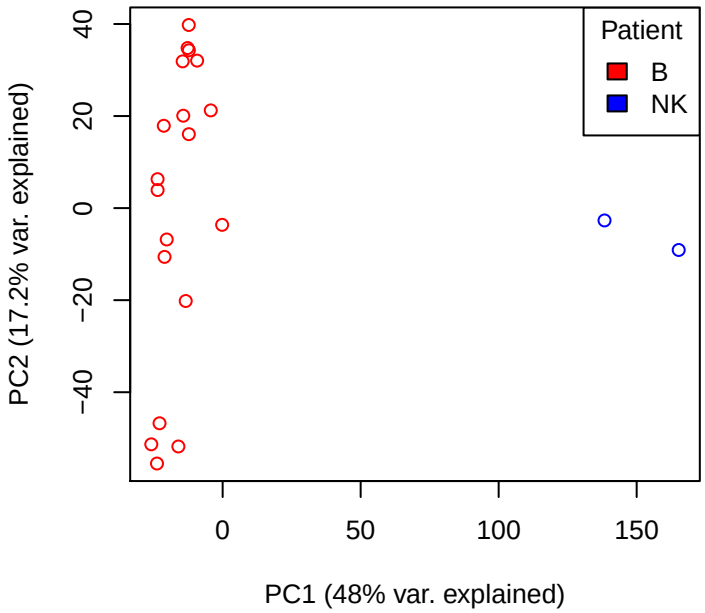
Cell-line Data

The following PCA displays the 13 samples from patient-derived cell lines, colored by batch. The plot reveals a very significant batch effect, which could be a potential confounding factor for further analyses.

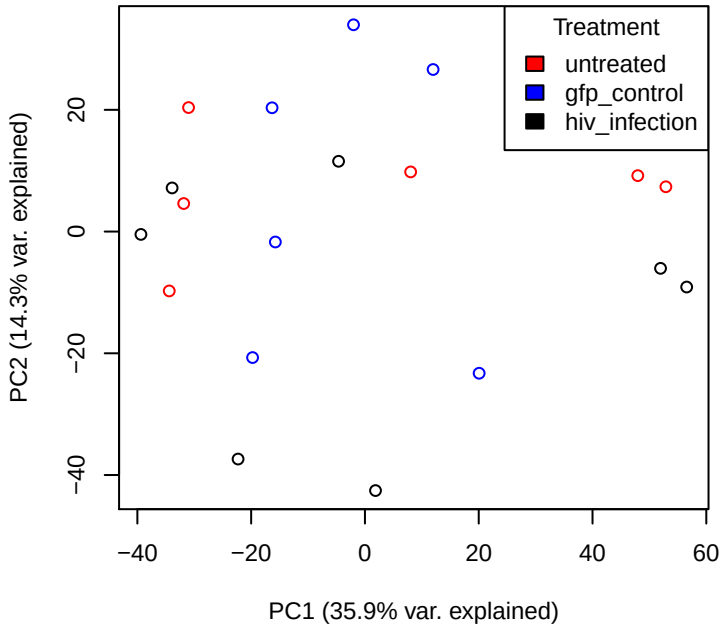


Primary Podocyte Data

The following PCA displays the 21 samples from patient-derived primary podocytes, colored by patient. The two samples from patient(s) NK are outliers.



To better visualize the separation between samples form B patient(s), the PCA was repeated after removing the two samples coming from NK patient(s). The following plot displays the results of the new analysis, colored by experimental condition.



The analysis shows that the samples don't clearly separate by experimental condition, potentially indicating the presence of additional unaccounted confounding factors.