Gene Expression Varies Pre- and Post- HIV Infection

Rachel Edidin
edidi003

Abstract

In the almost 45 years since HIV became an epidemic, 40 million people have died of HIV and another 40 million people currently live with HIV. HIV is a lifelong illness that develops into AIDS if left untreated, which causes death via immune system breakdown and increased susceptibility to infections. HIV can be treated with antiretroviral drugs, which lower viral load and increase life expectancy back to non-infected levels, but there is no cure. Understanding how HIV affects gene expression is critical to treatment development and further understanding of the disease overall. Here, we analyzed gene expression data from 14 individuals before and after HIV infection. We found evidence of significantly differentially expressed genes following infection, although changes in expression were not homogenous over time. Additionally, we found that differentially expressed genes were significantly enriched for functions relating to immune response and mitochondria.

Introduction

HIV, or the Human Immunodeficiency Virus, has caused more than 25 million deaths since its first recognized appearance in humans in the early 1980s (Sharp and Hahn, 2011). HIV is a lentivirus, or a retrovirus responsible for lifelong illnesses with long incubation periods (Milone and O'Doherty, 2018). Left untreated, HIV develops into Acquired Immunodeficiency Syndrome (AIDS), which is characterized by depletion of T cells, frequent infections, high risk of Kaposi's Sarcoma, pneumonia, and death (Seale, 1985). With medication, HIV progression can be controlled, and quality of life for people living with HIV remains high. Although medication has lowered HIV incidence in the developed world, access to treatment is not available in all places and thus AIDS has persisted as a pandemic for the last four decades.

There is no cure or preventative vaccine for HIV. HIV is a retrovirus, which is characterized by a single strand of RNA containing viral genetic material, which uses reverse transcriptase to form DNA from RNA and integrate into the host genome (Coffin et al., 1997). The reverse transcriptase of HIV is very error prone, and thus the genetic structure of HIV can vary rapidly, not just from person to person but in one person throughout the course of an infection (Barouch, 2008). Frequent genetic changes make HIV a challenging candidate for a vaccine, and no vaccine has led to a successful immune response against HIV.

Despite the challenges of producing a vaccine for HIV, the virus can be well managed with antiretroviral (ART) therapy. ART drugs, first introduced in the 1990's, target and inhibit enzymes necessary for HIV replication such as integrase, reverse transcriptase, and protease. Generally, several different inhibitors are used in combination to minimize the chances of resistance. With a daily ART regimen, HIV viral loads can drop to undetectable levels, preventing transmission and progression of the disease (Arts and Hazuda, 2012).

Although HIV can be effectively managed, infection is lifelong. HIV can affect host gene expression in different ways depending on the stage of infection. Analyses of HIV gene expression literature have shown that CD4+ cells, the primary target of HIV, experience changes in cell cycle and apoptosis regulation following HIV infection (Judge et al., 2020). However, HIV has a high mutation rate and infections can vary genomically in different individuals, making it hard to quantify HIV-mediated gene expression changes on a broad level. Additionally, small sample sizes and lack of pre-infection measurements can pose challenges to further understanding of gene expression and HIV.

This paper aims to investigate changes in gene expression shortly before (<100 days) and shortly after (<1 year) HIV infection. Using data from Mackelprang et al (2023) that contained

samples from patients in Sub-Saharan Africa at pre- and several intervals post-infection, gene expression differences across individuals were compared (data accessible at NCBI GEO database (Edgar, 2002), accession GSE195434 (Mackelprang et al., 2023)). These participants were not treated with antiretroviral therapy. We hypothesize that with increased time since infection, more genes will be significantly differentially expressed, although the small sample size for the study may limit the statistical power with which to conclude whether gene expression has significantly changed.

Methods

Data was read into R using the GEOquery package. All gene expression values were log-transformed, and quantile normalized. A large data matrix was made such that each column was a sample (representing a certain patient at a certain point in time either pre or post infection) and each row was an Illumina ID. Ninety total samples were present, representing measurements taken between 314 days before infection and up to 514 days post infection. Each cell in the matrix contained the expression value for each patient at that Illumina ID. 26 participants were in the original dataset; the data was filtered so that only patients with both a pre-infection and post-infection measurement remained (n=14). For patients with more than one pre-infection measurement (n=3), the measurement closer to infection date (i.e., less negative) was kept. Illumina IDs were mapped to genes using the g:Convert tool on the g:Profiler site (Raudvere et al., 2019). Only rows that corresponded with IDs associated with known, named genes were kept (n=20237). All further analysis was also done in R.

Four different time intervals were assessed. Pre-infection values were always involved in analysis, and they were compared to expression values between 0-100 days post infection (DPI)

(n=10 patients with post-infection measurement in this time interval), 100-200 DPI (n= 9 patients), 200+ DPI (n=11), and anytime (n=14; all patients were included in this analysis. For patients with more than one post-infection measurement, the latest measurement was kept). For each time interval, data was filtered so that only the samples corresponding to the desired time intervals were included. Pre-infection values were subtracted from the post-infection value in the time interval (for each patient included in the interval, for each gene) and a moderated t-test was performed on the differences using the mod.t.test function in the MKmisc package (based on the limma package). The null hypothesis of no difference in expression following HIV infection was tested against the alternative hypothesis of a change in gene expression following infection. The moderated t test assumes equal variance and, for each gene, calculated both a p-value and an adjusted p value using the Benjamini-Hochberg adjustment method.

Further analysis was done on the time interval that compared pre-infection expression to all-time post infection expression. Genes with an adjusted p-value <=0.2 (n=358) were tested for enhanced functional enrichment using DAVID (Dennis et al., 2003).

Results

The distribution of adjusted p-values was compared for each time interval. Results are in Figure 1. The distribution of adjusted p-values for the time interval of pre-infection to 100-200 days post-infection varied the most from the other time intervals, with a heavy right skew compared to the left skewed distributions of the other three time intervals.

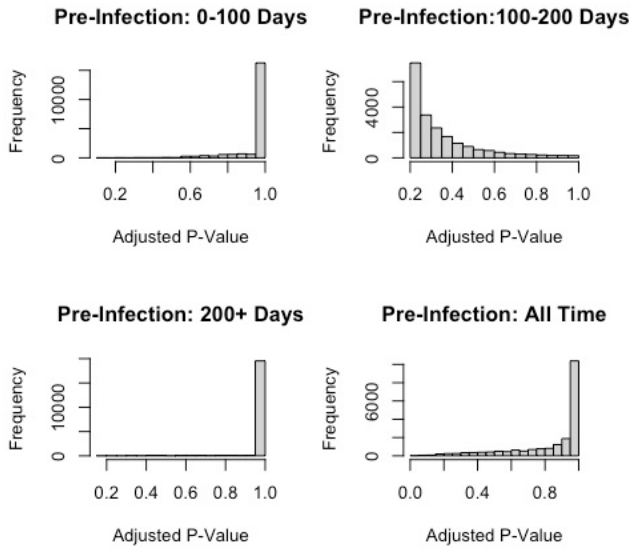

Figure 1: Distribution of adjusted p-values for each time interval. An adjusted p-value was calculated for each gene in each time interval using a moderated t-test and the Benjamini-Hochberg adjustment.

The number of significant adjusted p-values for each time interval at alpha levels of 0.2, 0.10, and 0.05 is displayed in Figure 2. Unsurprisingly, more genes were significant with larger alpha levels. Additionally, the pre-infection to any time post-infection time interval had the most significant adjusted p-values and was the only group that had significant p-values at a cutoff of 0.05.

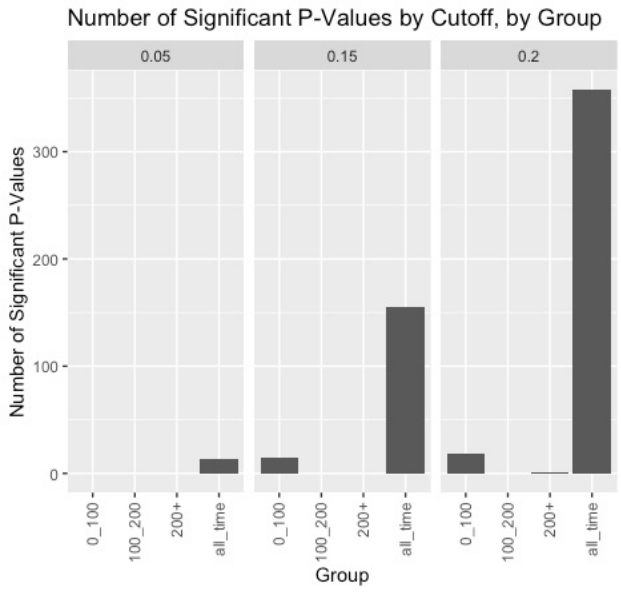

Figure 2: Number of significant p-values by cutoff, by group. The number of significant adjusted p-values at alpha levels of 0.05, 0.15, and 0.2 was compared for each time interval.

Functional annotation results from DAVID are shown in Tables 1 and 2. These results represent the annotation clusters for the genes in the pre-infection to all-time post infection group with an adjusted p-value less than or equal to 0.2. Table 1 shows the cluster with the highest enrichment score (17.86). Genes relating to response to virus and immunity were significantly enriched (compared to what would be expected with a random set of genes), with Benjamini-adjusted p-values less than 1.4e-8. Table 2 shows the cluster with the second highest enrichment score (4.57). Genes relating to mitochondria were significantly enriched, with Benjamini-adjusted p-values less than 1.1e-2.

Table 1: Annotation Cluster 1. Top functional annotation cluster for genes with adjusted p-values less than 0.2 when comparing pre-infection expression to post-infection expression.

Annotation Cluster 1	Enrichment Score: 17.86	G		Count	P_Value	Benjamini
GOTERM_BP_DIRECT	defense response to virus	RT		39	2.8E-26	4.1E-23
UP_KW_BIOLOGICAL_PROCESS	Antiviral defense	RT		33	9.8E-26	8.8E-24
GOTERM_BP_DIRECT	negative regulation of viral genome replication	RT		19	1.9E-20	1.4E-17
UP_KW_BIOLOGICAL_PROCESS	Innate immunity	RT		40	1.1E-16	4.9E-15
GOTERM_BP_DIRECT	response to virus	RT		21	3.0E-15	1.5E-12
UP_KW_BIOLOGICAL_PROCESS	Immunity	RT		57	1.5E-14	4.6E-13
GOTERM_BP_DIRECT	innate immune response	RT		38	3.9E-11	1.4E-8

Table 2: Annotation Cluster 2. Second functional annotation cluster for genes with adjusted p-values less than 0.2 when comparing pre-infection expression to post-infection expression.

Annotation Cluster 2	Enrichment Score: 4.57	G		Count	P_Value	Benjamini
GOTERM_CC_DIRECT	mitochondrial inner membrane	RT		27	2.1E-7	6.4E-5
UP_KW_CELLULAR_COMPONENT	Mitochondrion	RT		47	2.0E-5	2.3E-4
UP_KW_DOMAIN	Transit peptide	RT		24	4.1E-5	7.8E-4
UP_SEQ_FEATURE	TRANSIT:Mitochondrion	RT		24	8.5E-5	1.3E-1
GOTERM_CC_DIRECT	mitochondrial matrix	RT		19	1.3E-4	8.0E-3
GOTERM_CC_DIRECT	mitochondrion	RT		44	2.0E-4	1.1E-2

Discussion

The distributions of adjusted p-values for each time interval were very similarly skewed left (with most p-values quite high), except for the time interval between 100-200 days post infection (when compared to pre-infection expression values). It is possible that flaws in the experimental approach could have caused this result, as described later in this section. However, HIV antibody production and immune response is not uniform at all time intervals following infection. Rather, for a period following infection, the body has not yet started making antibodies, and a test for HIV infection would return a false negative result. The duration of this pre-antibody period has been estimated to be 2-3 months (Longini et al., 1989a, 1989b). Thus, gene expression may not appear to be significantly different during the initial period post infection because viral response in the body has not yet started, which may explain lack of significantly differentially expressed genes in the first 100 days following infection. However, because the patients in this study were not treated with antiretroviral therapy, antibody levels would be expected to persist over time, which would not explain why the 100-200 time interval p-value distribution was so different from the 200+ and all-time distributions. Lastly, significantly differentially expressed genes are not necessarily a proxy for antibody production or viral response.

The most genes with the most significant changes in expression pre- and post-infection (in the 'all time' group) had several significant functional clusters. The top cluster (Table 1) contained annotation terms relating to immunity and viral response, with very low Benjamini-adjusted p-values, indicating that the number of terms relating to immunity in the group of genes is much more than would be expected with a random group of genes. This result corroborates understanding about not just HIV, but many viruses. Following infection, immune response occurs. Although this response takes more time to become observable in HIV infection compared to other viral infections, partly because HIV RNA copies peak around 25 DPI, immune response is still observable with gene expression data as in this paper (McMichael et al., 2010).

The second highest functional annotation cluster (Table 2) contained annotation terms relating to mitochondria. HIV is not particularly thought of as a disease that affects mitochondria. However, previous research has shown that HIV infection can affect mitochondria, especially when antiretroviral therapy is not administered (Hulgan and Gerschenson, 2012). The *env* gene which encodes for the env protein in HIV has been found to cause apoptosis (cell death) via mitochondrial facilitation (Schank et al., 2021). HIV DNA has also been found integrated into mitochondria of infected cells (in addition to being integrated into the DNA of the cells themselves) (White, 2001). Further research could investigate the specific genes associated with these GO terms and how their function changes following HIV infection.

Although this paper found significant results for gene expression following HIV infection, including significant functional enrichment for genes associated with immune response and mitochondria, this study was likely overall underpowered due to small sample size and possible loss of power due to p-value adjustments with large gene lists. The largest group

investigated in this study, which was the group that compared pre-infection to any time post-infection, only had 14 participants. The smallest group in this study had nine. These sample sizes are small, and thus it may not be possible to get sufficient power with groups of that size.

One benefit of this study was that each participant was able to act as their own control because their own pre-infection expression values were compared to post-infection expression values. It can be challenging to have a large sample size when collecting data in this manner, because generally it's hard to identify people who can be reasonably predicted to contract HIV within less than a year. The participants from which this data came all had a sexual relationship with a person who already had HIV (who was not currently receiving HIV treatment), thus they were likely to contract it themselves. However, it's likely hard to find couples that meet the parameters for this study or a similar one (and it would be extremely unethical to intentionally cause HIV infection to study gene expression), so although this study has the benefit of each person being their own control, it is bound by a small sample size.

When testing many genes at a time for significant gene expression, it is important to adjust for multiple hypothesis testing because, with 20,000 gene expression levels to test, some will appear statistically significant just by chance. Adjusting p-values takes this into account while attempting to maintain high power. However, because there were so many genes tested, even with a t-test that assumed equal variance (increasing statistical power), p-values were inflated quite a bit due to the 20,000 hypotheses tested and thus (even if genes were truly significantly differentially expressed), adjusted p-values appeared much higher than if a smaller subset of genes had been tested.

Despite challenges involved with obtaining significant statistical power, analysis of the gene expression data presented showed signals for differential gene expression following HIV

infection, including two significant functional clusters (immune response and mitochondria) for differentially expressed genes. Further research could focus on a more specific subset of genes, which would not only produce more specific results about which genes are involved in immune response but would increase statistical power. Research on how HIV affects gene expression is an important element to understanding how HIV behaves and how it can be targeted, which are essential aspects to developing even better treatments and prevention.

References

Arts, E.J., and Hazuda, D.J. (2012). HIV-1 Antiretroviral Drug Therapy. Cold Spring Harb. Perspect. Med. *2*, a007161–a007161.

Barouch, D.H. (2008). Challenges in the development of an HIV-1 vaccine. Nature *455*, 613–619.

Coffin, J., Hughes, S., and Varmus, H. (1997). Retroviruses (Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press).

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, R60.

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Hulgan, T., and Gerschenson, M. (2012). HIV and Mitochondria: More Than Just Drug Toxicity. J. Infect. Dis. *205*, 1769–1771.

Judge, M., Parker, E., Naniche, D., and Le Souëf, P. (2020). Gene Expression: the Key to Understanding HIV-1 Infection? Microbiol. Mol. Biol. Rev. *84*.

Longini, I., Clark, W.S., Haber, M., and Horsburgh, R. (1989a). The Stages of HIV Infection: Waiting Times and Infection Transmission Probabilities. In Mathematical and Statistical Approaches to AIDS Emidemiology, pp. 111–137.

Longini, I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F., and Hethcote, H.W. (1989b). Statistical analysis of the stages of HIV infection using a Markov model. Stat. Med. *8*, 831–843.

Mackelprang, R.D., Filali-Mouhim, A., Richardson, B., Lefebvre, F., Katabira, E., Ronald, A., Gray, G., Cohen, K.W., Klatt, N.R., Pecor, T., et al. (2023). Upregulation of IFN-stimulated genes persists beyond the transitory broad immunologic changes of acute HIV-1 infection. IScience *26*, 106454.

McMichael, A.J., Borrow, P., Tomaras, G.D., Goonetilleke, N., and Haynes, B.F. (2010). The immune response during acute HIV-1 infection: clues for vaccine development. Nat. Rev. Immunol. *10*, 11–23.

Milone, M.C., and O'Doherty, U. (2018). Clinical use of lentiviral vectors. Leukemia *32*, 1529–1541.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198.

Schank, M., Zhao, J., Moorman, J.P., and Yao, Z.Q. (2021). The Impact of HIV- and ART-Induced Mitochondrial Dysfunction in Cellular Senescence and Aging. Cells *10*, 174.

Seale, J. (1985). AIDS Virus Infection: Prognosis and Transmission. J. R. Soc. Med. *78*, 613–615.

Sharp, P.M., and Hahn, B.H. (2011). Origins of HIV and the AIDS Pandemic. Cold Spring Harb.

Perspect. Med. *1*, a006841–a006841.

White, A.J. (2001). Mitochondrial toxicity and HIV therapy. Sex. Transm. Infect. *77*, 158 LP – 173.