Kellen Tyrrell

Dr Brandon Briggs

BioA455

26 April 2022

Identifying similarities in the Env HIV gene within the African Continent

Introduction

As of 2020, over half a million people worldwide died of AIDS. That year, over 50% of people living with HIV lived in Africa. As of 2017, the top five countries in terms of HIV related deaths were all located in Africa ("The Global Impact", 2020). Much of the burden of the AIDS disease in low-income countries, such as those in Africa can be attributed to high rates of unprotected sex, low access to treatment and a lack of health education. However, due to the high rates of mortality in these countries, it seems that factors other than those stated above may affect the rate of infection and of death.

Although HIV is more diverse in Africa then in any other continent (Bbosa, 2019), it is possible a common genetic trait that leads to increase in rate of infection is shared between subtypes in Africa. To investigate this, I compared the HIV envelope gene between sequences from within Africa, and outside of Africa. If envelope gene sequences from the African continent share more similarity based on location than subtype, there is likely a shared trait in the envelope gene between HIV strains from Africa that may lead to higher mortality rates.

Methods

Sequences that were used in this analysis were gathered from the HIV database (Leitner, 2006). For comparison, I used samples originating between 2005 and 2006. To identify the major HIV subtype of each country and their phylogeny I created a phylogenetic tree. To do this, I gathered 22 sequences representing every continent, and used Muscle (Edgar, 2004) to create a multi sequence alignment, using a maximum of 100 iterations. Using the multi sequence alignment, a phylogenetic tree based on the generalized time-reversible evolutionary model was created using FastTree (Price, 2009). This tree was visualized using Dendroscope (Huson, 2007) and was rooted using a sequence from Estonia.

In order to represent the similarities between samples (countries), based on continent and based on subtype, I created two Non-Metric Multidimensional Scaling plots. I first downloaded one envelope gene sequence per patient documented in 2006. Then, I aligned the sequences using muscle using a maximum of 2 iterations, as suggested by the author of Muscle (Edgar, 2004). I then used vsearch (Rognes, 2016) to dereplicate the sequences and create OTUs from the data at 94% identity. In order to visualize the OTU table generated, I used the QIIME2R and vegan packages within R (R Core team, 2021).

Results

The major subtypes of each country in Africa described shared more similarity between each other than among countries of other continents, with the notable exceptions of sequences from Thailand and Russia. Within the African continent, the subtypes C, D, and A were prominent. Outside of the African continent, the B subtype was prominent (Fig. 1).

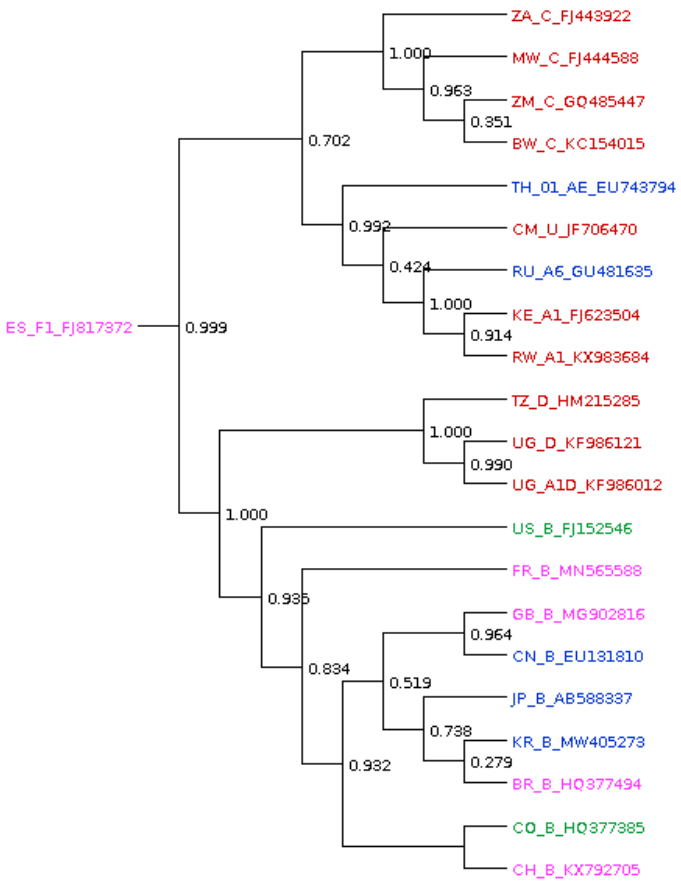

Figure 1. Dendroscope visualization of muscle alignment of major subtypes within each country analyzed. The labels are labeled in the format of twolettercountrycodes_subtype_accession number. Samples are labeled in red, blue, green, and purple for the African continent, Asian continent, the Americas and the European continent respectively

Based on the NMDS plots, envelope sequences clustered significantly more by subtype than by location, although clustering of both types were observed (Fig. 2, Fig. 3). Subtypes that were significantly clustered were type B, type C and type A, although type A showed multiple smaller clusters, whereas type B and type C showed one larger cluster (Fig. 2). Within every type, there were large numbers of outliers. The African continent was the only location to have one significant cluster, although this cluster was shared between only two subtypes. No other continent maintained a significant cluster (Fig. 3).

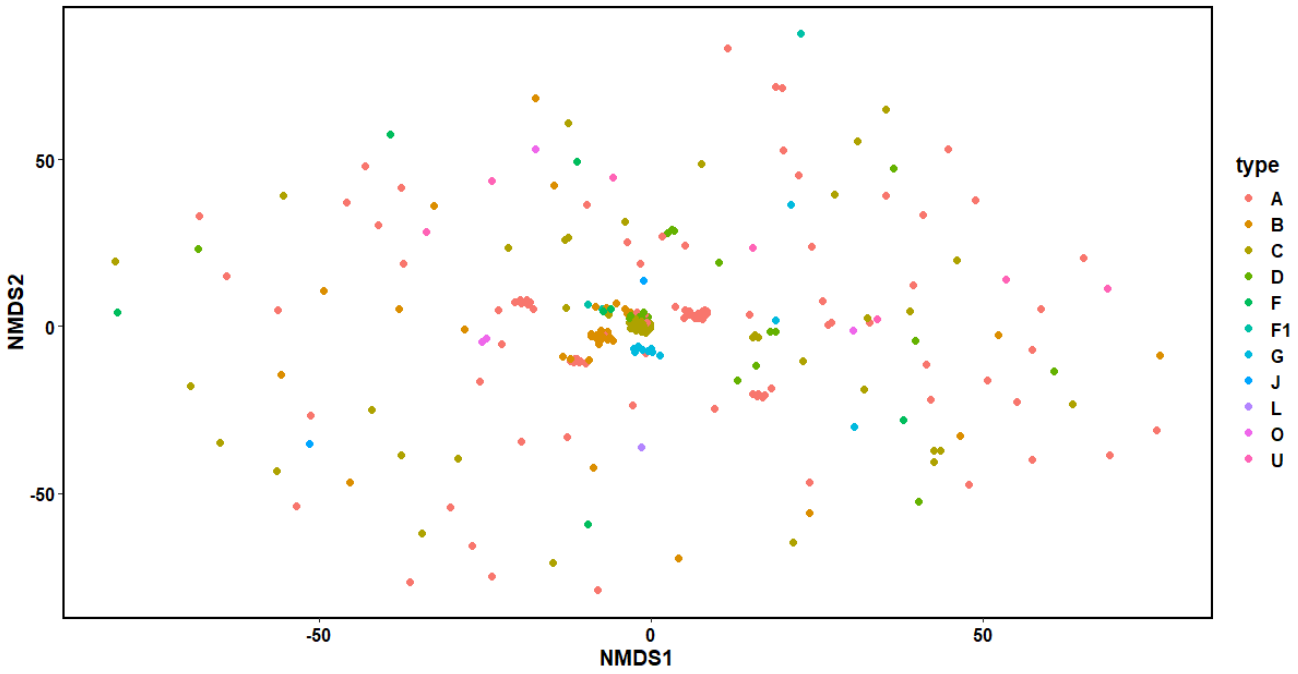

Figure 2. NMDS (non-metric multidimensional scaling) plot, with coloration based on HIV subtype.

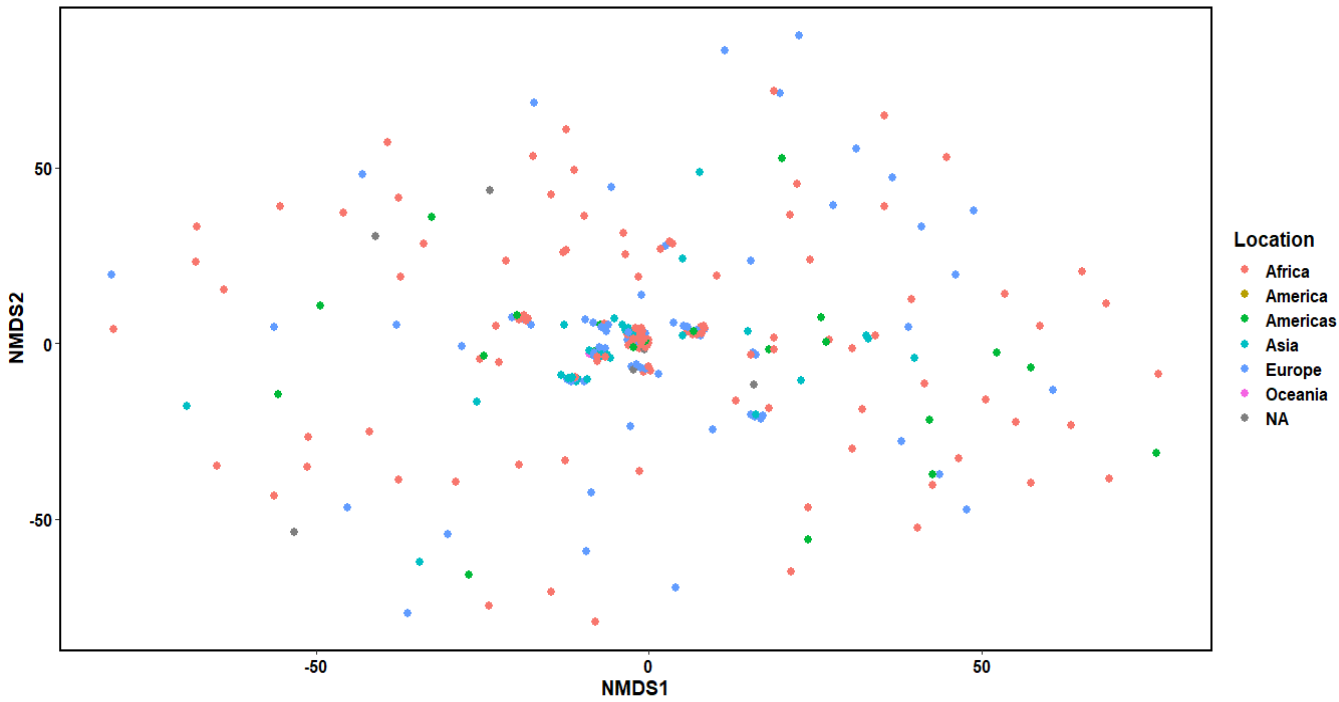

Figure 3. NMDS (non-metric multidimensional scaling) plot, with coloration based on continent.

References

Bbosa, Nicholasa; Kaleebu, Pontianoa,b; Ssemwanga, Deogratiusa,b (2019) HIV subtype diversity worldwide, Current Opinion in HIV and AIDS: May 2019 - Volume 14 - Issue 3 - p 153-160 doi: 10.1097/COH.0000000000000534

Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113 (2004).

Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.

Huson D., Richter D., Rausch C., Dezulian T., Franz M., and Rupp R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics.* 8, 460 (2007).

Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, and Korber B, Eds (2006). HIV Sequence Compendium 2006/2007 Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 07-4826. http://www.hiv.lanl.gov/

Price M., Dehal P., Arkin A. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. Molecular *Biology and Evolution*. 26(7), 1641-1650.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. https://doi.org/10.7717/peerj.2584

The Global Impact of Hiv & Aids (2020). Retrieved from https://www.hiv.gov/hiv-basics/overview/data-and-trends/global-statistics