Author’s response to reviews

Title: Characterization of APOBEC3 variation in a population of HIV-1 infected individuals in northern South Africa

Authors:

Nontokozo Matume (daphneymatume@yahoo.com)
Denis Tebit (demate10@gmail.com)
Laurie Gray (lrg4n@virginia.edu)
Stephen Turner (sdt5z@virginia.edu)
David Rekosh (dr4u@virginia.edu)
Pascal Bessong (Pascal.Bessong@univen.ac.za)
Marie-Louise Hammarskjold (mh7g@virginia.edu)

Version: 1 Date: 25 Nov 2018

Author’s response to reviews:

We are pleased to be able to submit a revised manuscript, in which we have addressed all of the concerns and comments by the Editor and the 3 reviewers. We want to thank the reviewers, since their remarks have enabled us to significantly improve the manuscript. Specifically, we have now included p-values in Tables 4 and 5, as requested by one of the reviewers, to point readers to the most important differences between our population and others. We have also added a comparison in Table 4 to the ExAC exome consortium database, since this contains data for over 60,000 unrelated individuals from all over the world. The editor pointed out, our statements about Linkage Disequilibrium (LD) values were in some cases incorrect. To remedy this and present a more complete picture for the readers, we now present the LD calculations for all the SNPs where this was possible in 4 new supplementary tables (S2-4, one for each A3 gene analyzed). Below, we highlight the changes that we made in response to each comment from the Editor and the reviewers.
Editor Comments:

Besides addressing reviewer questions, authors are suggested to:

1. Provide ethnic/race/population information of the study group; if unavailable for individuals, describe the possible composition in the local area (as you mentioned at P4, l57). This is important for the representation of the variant frequencies and meanings of their comparisons with other populations.

   Response: We extended table S1 and added gender, age, ethnicity, geography by district, viral loads and CD4+ cell count. The information in this table is now highlighted in the manuscript under the materials and methods section, study population and DNA extraction subsection.

2. Reviewer 2 suggested for comparison with sub-African populations, reviewer 1 suggested the African Genome Variant Project. When or if you do that, indicates which reference group is closest to your population, and whether it can serve as a valid control group for HIV infection. comparison of world population groups can be suppl.

   Response: We are unable to use the African Genome Variant Project, since the data are not yet publicly available but is controlled by by the African Partnership for Chronic Disease Research Data Access Committee. In addition in the published paper that reviewer 2 mentions, no allele frequencies are reported and since since the data is array data, they may not actually genotype the variants we would want to obtain allele frequency on

Minor:

3. Table 2 A3F, 108, 118,143 sites all HWE p<0.05. Was this local region sequencing quality low?

   Response: All data was generated with Illumina sequencing. Samples that did not pass QC or were of low quality as analyzed by FastQC and MultiQC for high throughput sequences were removed prior to analysis. We checked all the data from the A3F regions in question and the sequencing quality, as well as coverage was good.

4. Table 4. 5. A3H G105, 121, 140 have opposite allele freq comparing to other populations. check and discuss.
Response: We realized that the allele frequencies reported for our SA population in the original version of the manuscript were based on genotypes at each site rather than alleles. We have now recalculated all of the frequencies correctly. These corrected allele frequencies are now presented in the revised tables 4 and 5 and are no longer reversed.

5a Table 2. A3H N15 and R18L HWE $P < 0.00$, could these sequence site calls have problem?

Response: We checked the site calls in this region, they were called correctly.

5b Table 2, P13 L23-25, deviated from HWE ($P > 0.05$). should be: largely deviated from HWE ($P < 0.05$).

Response: This has been corrected.

6. $P12,155, LD R2 0.068$ are not strong

Response: We agree with this statement. We have also redone our discussion about the LD values for each gene and included tables S2-S5 to show all the LD values for each of the 4 genes in a table format.

7. p17 l21026, taqman and Rflp are not for variant discovery.

Response: The manuscript that was cited used Taqman and RFLP to determine the frequencies of the polymorphisms, not for variant discovery. The authors used PCR and Sanger sequencing, not NGS for variant discovery. We have now clarified this in the text as “The use of next generation sequencing also allowed the identification of SNP genotypes that were not previously identified, using methods such as TaqMan SNP genotyping assays, restriction fragment length polymorphism (RFLP) or Sanger sequencing [39]”

Reviewer reports:

Maria Chahrour (Reviewer 1): In their manuscript titled "Characterization of APOBEC3 variation in a population of HIV-1 infected individuals in northern South Africa", Matume and colleagues report on variants in the APOBEC3 genes in a cohort of HIV-1 infected individuals
from South Africa. Identifying variants in APOBEC3 genes and characterizing their population frequencies is important given the biological function of the encoded proteins and the significance of this in terms of HIV-1 pathogenesis. However, the major caveat in the study (that the authors also acknowledge), is the fact that they identify APOBEC3 variants and characterize their frequencies in a population of HIV-1-infected individuals rather than doing so in unaffected individuals. In my opinion this is a major flaw that needs to be addressed. They are comparing allele frequencies in their cohort of affected individuals to several other cohorts, all of which are not infected with HIV-1. The problem is compounded due to their small sample size. For this study to be meaningful, the authors need to identify variants and compare allele frequencies to a cohort of individuals who are not infected with HIV-1 and are from the same population in the Limpopo province. Furthermore, the authors rely on reference haplotype data from the 1000 Genome project which is not ethnically matched to their study cohort (despite some data from African ancestry in the 1000G project). It is more appropriate to use data from the African Genome Variant Project (see Gurdasani et al., 2015 Nature).

Response: As we described above, the reference paper on the African Genome project does not describe allele frequencies and the data from this projects is not not yet publicly available. Additionally, the data is array data. We do not know if they actually genotype the variants we would want to obtain allele frequency on.

As for the comment that we only used HIV positive individuals, we realize that it will be important in the future to compare this population to uninfected individuals in the same region, but these samples are currently not easy to access. In addition, there are currently no data in the literature to suggest that a certain ApoBec3 profile leads to a changed risk for HIV infection. This is in contrast to for example CCR5 polymorphisms, where the delta 32 allele has a strong protective effect. Thus there is no inherent reason to think that our SA population differs from any other population in this region. Thus even with the caveat that we are only reporting on HIV infected patients (which we clearly point out in the paper), we still think that the data that we present will be very valuable to other researchers, as they pursue further studies in Limpopo and other regions. This view seems to be shared by reviewers #2 and #3., both of whom comment that “report will be a valuable resource for designing and evaluating further studies of APOBEC3 genes” and “I am in favor of publication”.

I have additional comments that also need to be addressed, as detailed below:

1- The authors should provide more detail on the clinical picture of individuals in their cohort, and also the demographics of the cohort (age, sex, …).
Response: As we also responded above, we extended table S1 and added gender, age, ethnicity, geography by district, viral loads and CD4+ cell count. The info in this table is highlighted in the manuscript under the materials and methods section, study population and DNA extraction subsection.

2- In Table 2, in the "Genotypes" column, the authors provide the cDNA variant position but they do not mention that this is the cDNA position. Please include that in the Table description/title, and also include the NM transcript number for each gene since the cDNA positions are in reference to NM transcripts.

Response: We have remedied this by providing ENSEMBL transcript number and adding them to table 2 in the footnotes. In the case of APOBEC3F, we have also clarified that we detected amino acid changes in 2 different acknowledged isoforms and added this information in table 2.

3- The first column in Table 2 states "CDS codon position change & rs numbers". Please change this to "Amino acid change and variant ID".

Response: This has been corrected

4- In Table 3, in the "Variation" column, the authors indicate amino acid position. Please include that in the title for the column: "Variation (amino acid and its position)" so the readers can easily recognize.

Response: This has been changed

5- In Table 4, first column title needs to be "Amino acid change and variant ID".

Response: This has been changed

6- Same comment for Table 5, first column title.

Response: This has been changed
Jaroslav Bendl (Reviewer 2): In this study, the authors report on case-only study performed on 192 HIV-infected individuals of South Africa origin. Except for two variants, all other reported ones show a higher prevalence compared to known prevalence for African population. Noteworthy, some of those variants have been previously associated with HIV progression. Unfortunately, the absence of control cohort from Limpopo province has not allowed them to draw stronger conclusions about the importance of differences in frequencies of those variations in relation to HIV. On the other hand, their report will be a valuable resource for designing and evaluating further studies of APOBEC3 genes.

Response: See our comments in response to those of Reviewer 1 above.

Minor revision:

I would suggest using Fisher’s exact test to assess the differences in the allele frequencies between 192 HIV-infected South Africa’s individuals and the others world’s populations and Africa subpopulations (i.e., to extend Table 4 and 5). This would help to prioritize variations for further investigation.

Response: To address this, we applied Fisher’s exact test to the data in table 4 and 5 as recommended by the reviewer. We now show the p-values and have indicated those that are significant with (*). We also added information from the ExAc exome sequences base to Table 4, since this database examines the largest number of individuals. In this case, we used a Chi-squared test to assess the differences because the sample size is so large. Additionally, we now show the number of individuals in each population in each population. We agree with the reviewer that this additional information makes the manuscript stronger and more valuable to the readers. We thank the reviewer for this suggestion.

Typo

dpSNP -> dbSNP (Table 3 - legend)

Response: This has been corrected.

Nurten Akarsu, M.D, Ph.D (Reviewer 3): Matume et al. characterize polymorphisms in APOBEC3 genes, namely A3D, A3F, A3G and A3H, which are implicated in the restriction of HIV-1 replication. A total of 192 HIV-1 positive individuals from Limpopo province of South Africa are studied. In addition to known variants, novel variants and haplotypes, mainly in A3D and A3F, are now reported. APOBEC3 variants and its role in HIV infection are well discussed in the literature. Thus, the paper does not provide an extensive contribution to existing
knowledge. Furthermore, lack of functional characterization of these variants is the major weakness of the study as authors stated in the MS. On the other hand, population genetics and diversity are well studied and presented. The molecular data is strong. The data is traceable in Ensembl and other databases. Thus, I believe that these variant/haplotype informations will lead future functional studies in order to understand dual effect of APOBEC3 polymorphic changes on HIV infections. I am in favor of publication.