Reviewer's report

Title: Correlates of the molecular vaginal microbiota composition of African women

Version: 1 Date: 16 October 2014

Reviewer: Heidi Jones

Reviewer's report:

This is a very interesting paper looking at correlates of different vaginal microbiome characteristics in African women and will be an important addition to the literature. In general, the paper is very strong, but I have a few suggestions to further strengthen it:

Major Compulsory Revisions:

1. Please describe how many batches were run on the micro-array analyses, and include testing for/analysis of batch effects, especially for the analyses to identify/distinguish the VMB clusters.

2. Since timing in the menstrual cycle is likely to be highly correlated with characteristics of the VMB, I would also recommend running a sensitivity analysis for the correlates on women that were within 4 days of their LMP (by excluding the women from Tanzania and the pregnant women) to test whether the results are substantively similar.

3. Please clarify why only 216 samples were analyzed from the Biomarker study. Will there be a more complete data set in the future? If so, I would recommend waiting to analyze and publish data from the complete set of enrollment specimens. However, if the analysis for this subsample is due to funding constraints for the entire sample, please include this information in the text.

4. In the methods section, the authors should consider adjusting their p-value cut-off for a false discovery rate. Additionally, given that the smallest cluster had a sample size of 19, perhaps a less stringent critical value should be used prior to the false discover rate adjustment, e.g. p<0.10 or 0.20?

Minor Essential Revisions:

1. Line 249 needs to clarify whether evenness was based on the 5 most common genera of bacteria across the clusters – i.e. the calculation used the same 5 genera for the measure of evenness for each of the clusters?

2. Please describe the ordering of specimen collection in terms of when the swabs for VMB were taken in relationship to when the CVLs were taken for cytokine analyses in the methods section.

3. Table 1 – the authors should be sure to report all percentages to one decimal
point. Additionally, this table would be strengthened if it included a column with the n and column percentages as a new first column. Further, it would help if the first subcategory would start on a new line, as currently it is difficult to read both the category and the first group (e.g. Adults, KE should be on the next line in the text).

4. In this same table, I would not include variables that are missing a large proportion of cases – e.g. 124/208 cases missing reporting on sexual frequency.

5. In Table 1, the current contraceptive use categorization is confusing. Does ‘combined hormones’ refer to only COCs and then progestin-only pills, DMPA and Net-En are all included in the ‘progesterone’ category? Please clarify. The pregnant category should really be listed first as ‘none, pregnant’ followed by ‘none, not pregnant’. Additionally, I would be interested in the comparison of pregnant versus not pregnant with its own statistical test (rather than combining the data on pregnancy and contraceptive use into one measure).

6. Based on Table 1, it looks as though none of the women from the HIV-positive group from Rwanda were included in this analysis. However, in Figure 3, there were a few cases of HIV. Please clarify – are these women from follow up in Tanzania?

7. For the age category in this table, it would help to specify that the youngest age category was 16-17 (rather than <18) and the highest was 30-35? Please make it clear in this table that the numbers are ns and that the estimates in the parentheses are percentages.

8. Additionally, the readers need to understand how the composite scores for SES and sexual risk taking were calculated in order to interpret the low/medium/high categories.

9. For Table 2, how was ‘symptomatic BV’ defined – was this based on Nugent score or on Amsel criteria? In general, I would highly recommend including the RTI/UTI data in this table (rather than Figure 3, but keeping the supplemental figure S2A). I would be interested in seeing both the results from the Amsel criteria and the Nugent score (no/intermediate/BV), as well as the individual STIs - CT, GC, TV, HIV, HSV-2, TP and candidiasis. Were there any cases of Trichomoniasis based on culture (rather than wet mount)? If so, were these included with ‘bacterial’ STIs?

10. Given that processing of immunology was different in Tanzania, do the results in Table 3 change when the analysis is restricted to the other sites?

11. I find Figure 1A confusing – if co-values were calculated by looking at the 5 most common genera, why are there more than 5 points on some of the lines?

12. It seems worth noting in the limitations section in the discussion that for the majority of the specimens (those not in Tanzania), the day of STI and UTI diagnoses and reporting of sexual behaviors was not the same day as the specimen collection for the VMB (on average more than 3 weeks apart in time),
while for Tanzania it was on the same day.

13. What did the authors find in terms of the 18S probes for fungi and the 3 viral probes?

Discretionary Revisions

14. I would encourage the authors to add a dendrogram to the heat map in Figure 3B. Additionally, I find heat maps easier to interpret when they are done in black and white with shades of grey so that the intensity of presence of a type of bacteria is darker. This is just a suggestion but the color heat maps are harder to instantly interpret (especially for colors such as yellow and green) than heat maps with one color scheme (i.e. white to black) that increases in darkness with higher S/B ratios.

In general, the authors have done a great job presenting complex data and analyses in a clear manner and will add to the literature on our understanding of vaginal flora and its correlates. I hope these comments are helpful.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have not competing interests