REPORTE

En este reporte encontraremos el pipeline, ya curado y mejorado, para el procesamiento de datos realizado por ADM-Valencia, usando los datos de FTA y Midgut de Albert. Pero antes, quisiera dedicar una página al **método de normalización** que aconsejo usemos de ahora en adelante:

Seleccioné como método de normalización el que se hace con DeSeq2. La selección de este método de normalización está respaldada por los mismos desarrolladores de phyloseq package en R, quienes NO recomiendan normalizar por rarefacción usando phyloseq (no recomiendan su propio paquete para normalizar por rarefacción) sino que recomiendan la normalización de DeSeq2 para trabajar con los datos (*https://github.com/joey711/phyloseq/issues/229 : "1)* ***Rarefying should never be used. Ever.*** *A statistician would never recommend this approach. It is not justifiable. I described rarefying as sometimes tolerable for whole-sample clustering, but I meant this in the ultra-cautious fair-to-the-point-of-insanity sense of the word tolerable -- that rarefying is not giving you the completely wrong answer, but it is very often (approaching always) giving you a worse, less-reliable answer, because you are throwing away data. I can describe rarefying this bluntly because we are here on my Issues Tracker, but we used a much more diplomatic description in the manuscript. I'm sorry that it appeared that rarefying should ever be used. Ever again. Please feel free to point out sections of the manuscript where this was less than clear, in case we have not caught this during the revisions. ; 2) Between an analysis framework for sample-wise comparisons, or testing differential abundance, the effect of* ***rarefying is much worse for differential abundance****. However, rarefying should not be used for either analysis framework."),* y tras pruebas con otros dos métodos de normalización (TMM y TSS).

Cabe recalcar que **corrí todo con 4 métodos de normalización diferentes: rarefacción con phyloseq, DeSeq2, TMM y TSS.** Y, en mis pruebas, ni el número de ASVs asignados ni las significancias de separación por grupos para las pruebas de Alpha y Beta diversidad cambian usando los métodos DeSeq2, TMM y TSS. Sólo cambian con el método de rarefacción. Esto tiene sentido si entendemos que el último método genera un submuestreo (O sea que elimina ASVs de algunas muestras para "emparejar" los datos para después compararlos, lo cual no es correcto… aunque mucha gente lo usa, no es nada recomendable/correcto). Por esta razón, recalco la importancia de no usar este método, especialmente para nuestras preguntas y tipo de estudio (con tamaños de n pequeños, menos profundidad de secuenciación / tamaños de librerías no tan grandes, 16S con dos grupos donde nuestra pregunta biológica es ¿son diferentes entre sí?, etc).

En conclusión: DeSeq2 es el más robusto y a ese le sigue TMM. Yo sólo he usado TMM en el pasado. TMM sería una buena opción sólo si los datos son tan grandes que el poder de cómputo es insuficiente para procesar con DeSeq2 o si no usamos R para procesar nuestros datos. Pero este no es nuestro caso, especialmente lo del poder de cómputo porque los datos de microbioma y de amplicones de 16S no son tan diversos y grandes como para presentarnos esta limitante (menos con la computadora nueva que está en camino).

Esto es una mejora importante del método de Valencia. Pero es mucho más confiable y mejor hacerlo como lo recomiendo.

WORKFLOW STEPS

From Illumina paired end amplicon sequence, raw reads:

a) Adapter trimming and min length of 100pb filtering, using Rfastp

b) DADA2 Quality trimming of min Q 20 and quality analysis report

c) DADA2 pre-processing:

- Learn error rates of Forward & Reverse reads
- Derreplication
- Sample error inference
- Merge paired end reads
- Sequence table construction
- Chimera Removal (checkpoint: extract reads without chimeras)
- DADA 2 REPORT

d) DADA 2 taxonomic assignation, using SILVA train set and species database.

e) Generating phyloseq object (ps)

f) Data normalization, **using DeSeq2 *** AND ADDITIONAL filter of minimum ASV observed count of 5, to eliminate possible false positives in assignation. ***** **mejoras del método de Valencia**

g) Generating new normalized phyloseq object (ps_norm_deseq)

h) Statistical analysis and generation of publication plots, using vegan:

- Rarefaction curves
- Alpha Diversity **(Observed, Shannon, Simpson, Pielou)** and boxplots **(Wilcoxon with Benjamini Hochberg p-value correction) * mejoras del método de Valencia**
- Beta Diversity (Bray-Curtis dissimilarity distance) and NMDS (stress score) and PCoA **(Beta dispersion and PERMANOVA)*** plots ***** **mejoras del método de Valencia**
- Bar Plots that include de LCBD value (beta: grouping influence of each sample)

i) Differential Abundance Plot, with DeSeq2

RESULTS

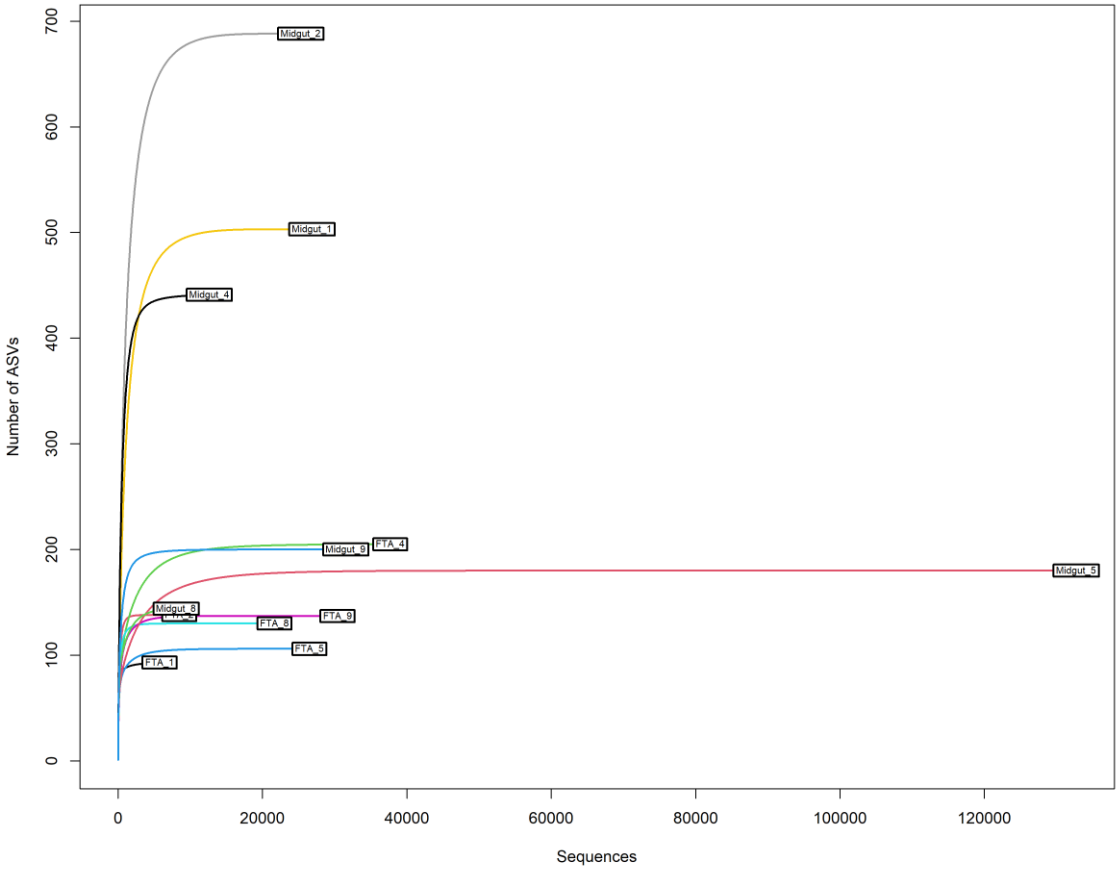


FIGURE. 1. Rarefaction curves from ASV counts.

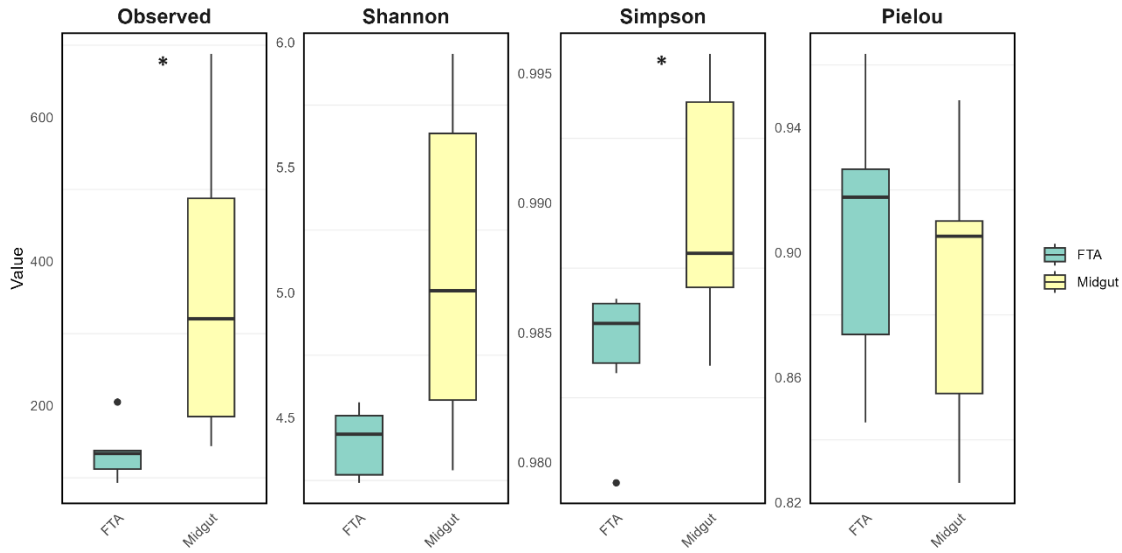
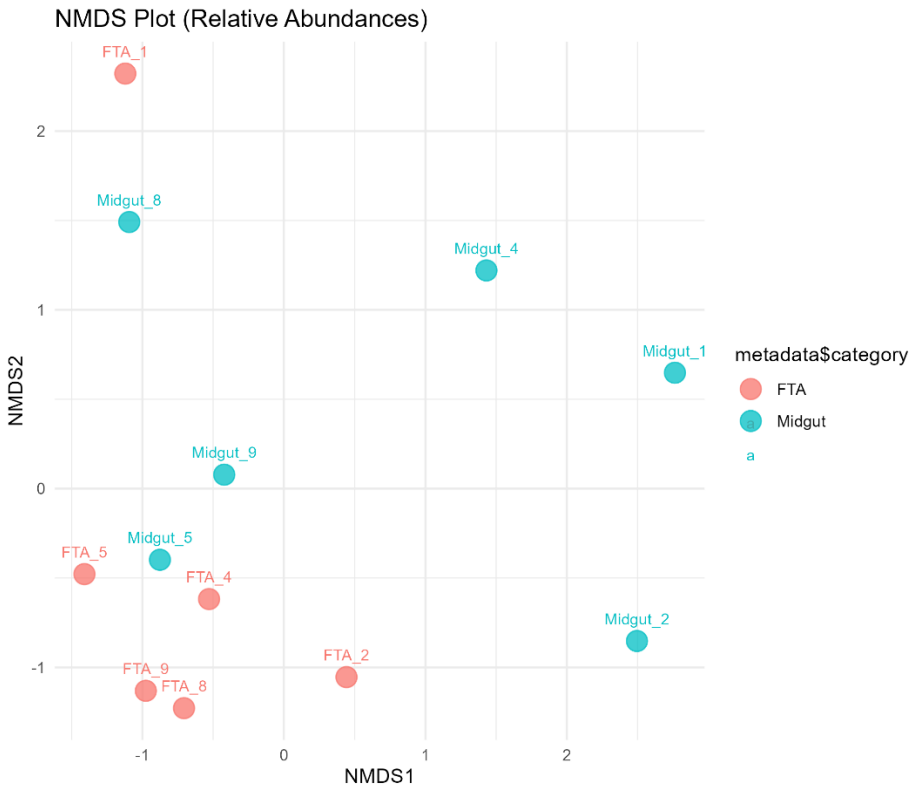


FIGURE. 2. Alpha diversity boxplots. **Significant differences between Midgut and FTA groups for Observed and Simpson (*observados podría ser un problema de profundidad de secuenciación en algunos vs otros. Y Simpson indica dominancias diferenciales. Discuto más debajo, en los Bar Plots).**

A.



B.

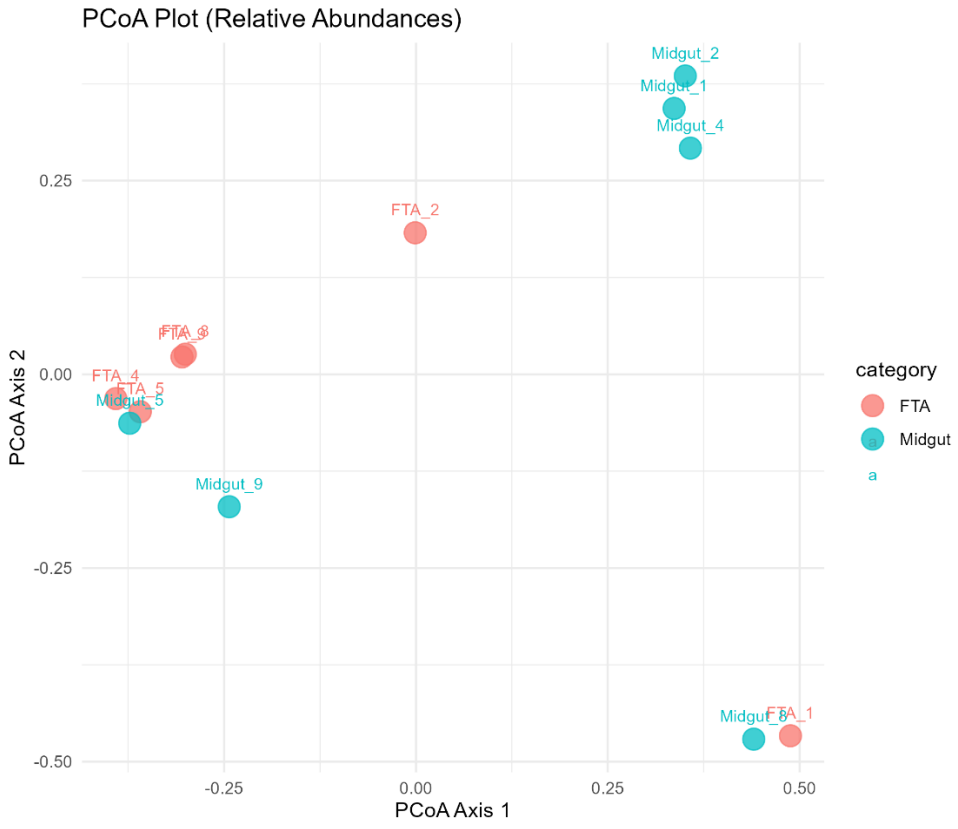


FIGURE. 3. Beta diversity plots: **(A)** NMDS; **(B)** PCoA. No significant differences found.

Tablas con las salidas de estadísticos **(MENCIONANDO que nuestra n es en realidad muy pequeña como para que los estadísticos se consideren robustos)**:

ALPHA DIVERSITY

Var1	Var2	value	metric
Midgut	FTA	0.01515152	Observed
Midgut	FTA	0.09307359	Shannon
Midgut	FTA	0.02597403	Simpson
Midgut	FTA	0.39393939	Pielou

BETA DIVERSITY

Method	stat	P.value
PERMANOVA	0.10462127	0.3298
Betadisper	1.31381354	0.27839277

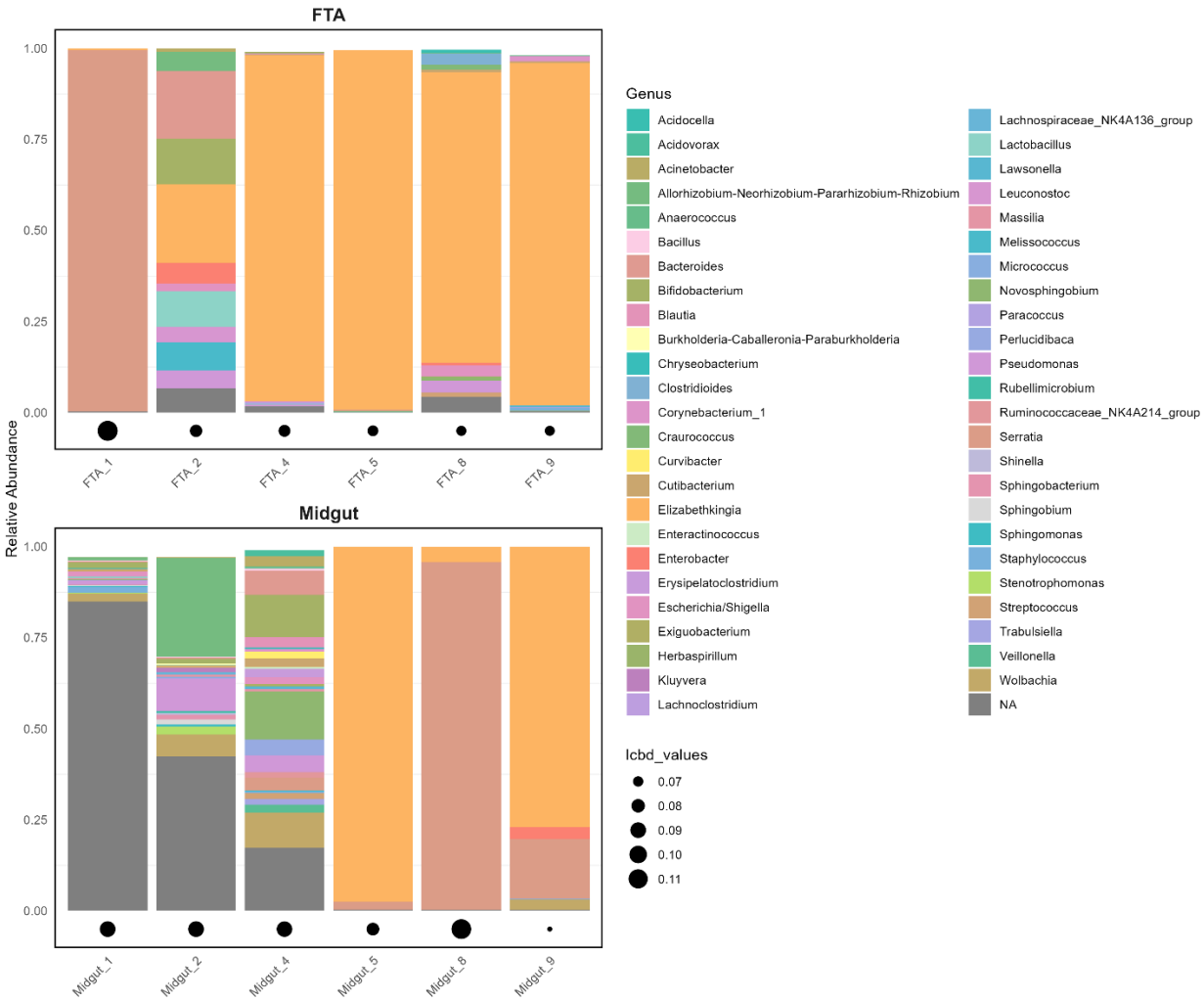


FIGURE. 4. Top 50 most abundant genera Bar Plot, with LCBD values added.

DISCUSION: Ok, aquí hay mucho que me gustaría discutir.

1. **Vemos mucha diversidad en Midgut (tres muestras con un patrón completamente diferente a las otras tres muestras). Las que tienen dominancia en Elizabethkingia y en Serratia (el naranja y durazno-beige, respectivamente) se parecen más a lo que vemos en las muestras de FTA. Las otras tres muestras de Midgut tienen mucha más diversidad y además tienen una gran proporción de NA. Esto es muy importante de discutir entre nosotros porque yo pienso que podría ser contaminación. Podría ser que se amplificó algún gen/genes inespecíficos del mosquito mismo y/o haya algo humano por ahí. También querría decir que no estamos sólo viendo midgut microbiome sino microbioma de lo que sea que haya contaminado. Saliva? Sudor? Otra parte del mosquito. Lo dejo completamente a su consideración pero no me parecería algo imposible, ya que disectar a un insecto tan chiquito debe ser muy difícil.**
2. **Si lo anterior es así, entonces creo que nuestras significancias entre Alpha diversidad de Observados y Simpson (dominancias) estarían explicadas por estas tres muestras de Midgut. Lo podemos ver en el valor del LCBD. Las tres muestas con NAs en Midgut, tienen los puntos más gordos. Dejando eso de lado, tanto en Midgut como en FTA, la dominancia en Serratia (que creo que sí es un observable real del microbioma) son lo que explican diferencias entre muestras.. aunque las observamos en ambos grupos.**
3. **Sobre el mismo hilo de pensamiento y añadiendo que tenemos una n pequeñita de sólo 6 muestras de mosquito por grupo (6 Midgut, 6 FTA), yo esperaría entonces que a mayor n, NO esperaríamos diferencias entre grupos. Sólo diría que muy posiblemente Midgut tenga un poco más de diversidad representada (menos dominancias.. aunque no lo suficiente como para ser significativo), comparado con FTA.**
4. **Por último diré que jugué con otro filtro más que se puede hacer con DeSeq2 que es un filtro de PREVALENCIA de los ASVs entre muestras. Este filtro de prevalencia lo hice de la siguiente manera: "elimina los ASVs con menos de 5 counts (conteo) & que no estén en mínimo el 10% de las muestras (prevalencia)". No me quedé con este filtro porque pasa algo parecido que con la normalización por rarefacción (elimina ASVs de algunas muestras y de otras no y es quitar información para comparar... no es aceptable). Pero, a pesar de que no lo terminé usando, me pareció interesante que, emparejando los grupos (con este submuestreo), los géneros más cambiantes (en términos de abundancia relativa) eran: Elizabethkingia, Serratia, Wolbachia, Novosphingobium, Bifidobacteria y Allo/Neo/Para/Rhizobium. Lo menciono nada más para que tengamos esos géneros en mente al analizar Bar Plots y el análisis de Abundancia Diferencial (debajo).**

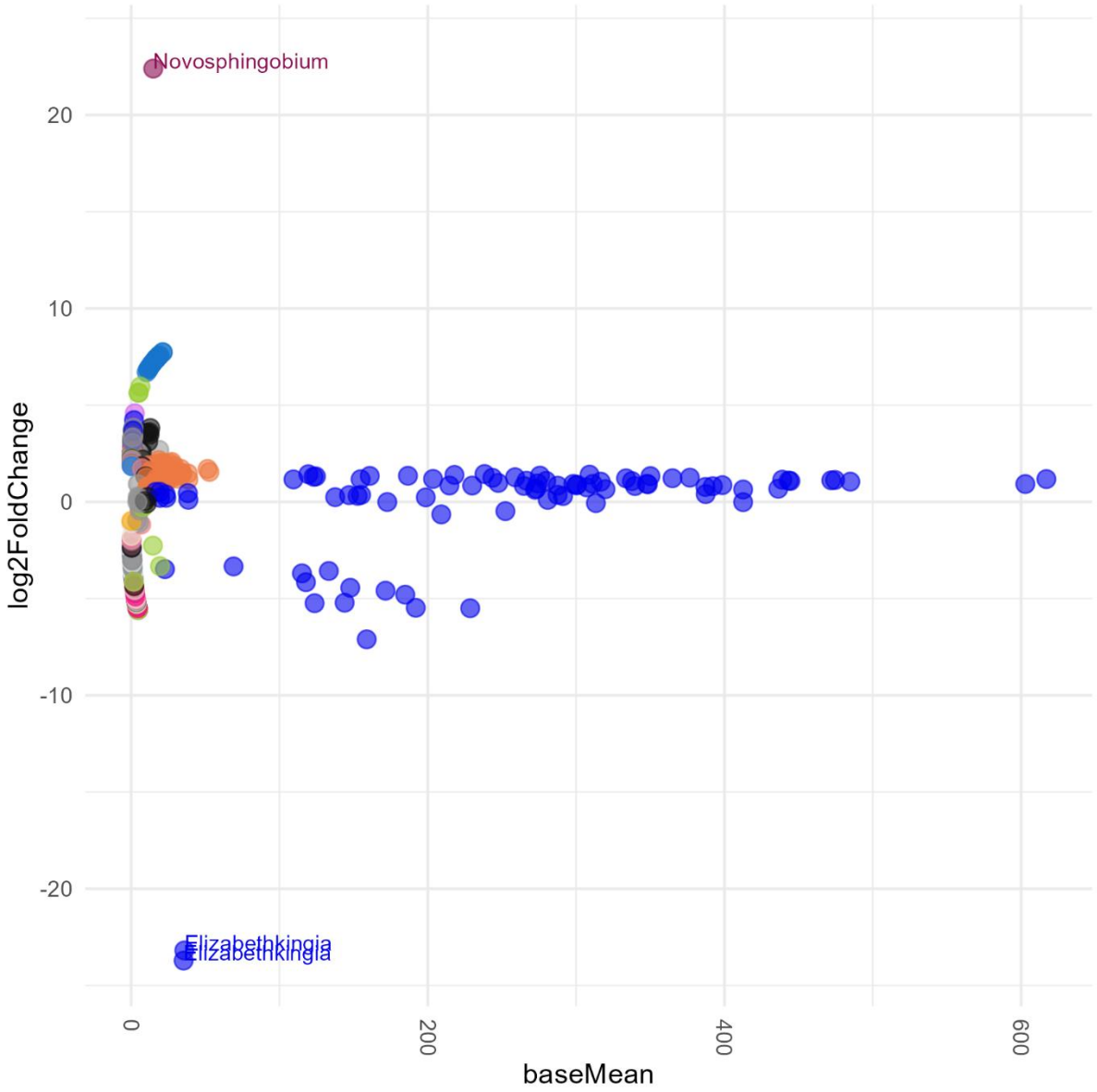


FIGURE. 5. Differential abundance analysis. Labels on differentially abundant genera with significant support.

DISCUSION: Pues aquí solo tengo que decir que yo esperaría que, si tengo razón con las muestras de Midgut y con un mayor tamaño de muestreo (yo sugeriría una n mínima de 30… y si es muy muy difícil tener tanta muestra, pues lo más acercado posible), veríamos a Serratia y Wolbachia también anotadas en este análisis.