Responses to reviewer comments: *Dos Santos et al*. 2024

Reviewer 1

In their manuscript titled “Vaginal metatranscriptome meta-analysis reveals functional BV subgroups and novel colonisation strategies”, the authors describe the re-analysis of three vaginal metatranscriptomic datasets using an updated version of the software “aldex2” which implements the scale reliant inference approach. They highlight several functional differences in the transcriptional activity of Lactobacillus dominated communities and BV-associated communities and even demonstrate the replication of these findings in the analysis of the third dataset. Other findings discussed include differences between BV-associated communities which originate from the expression of flagellar proteins by BVAB1. Overall, I found the study interesting and appreciate the push towards more mechanistic and functional understanding of the vaginal microbiome. However, I have some concerns, particularly about the removal of genes from the dataset via filtering and the use of language around the description of metatranscriptomic data.

Major issues

1. The inclusion of line numbers would have made the reviewing process significantly easier and faster.

We apologise to the reviewer for the omission of line numbers in our manuscript. These have now been added to the revised manuscript in bold blue text in the left margin.

2. Throughout the manuscript the authors sometimes refer to “abundances” or “relative abundances” of particular microbes, but the underlying data is from metatranscriptomics. These data do not measure the abundance, but rather the activity of the bacteria. This language needs to be corrected so as not to be misleading what is actually being measured.

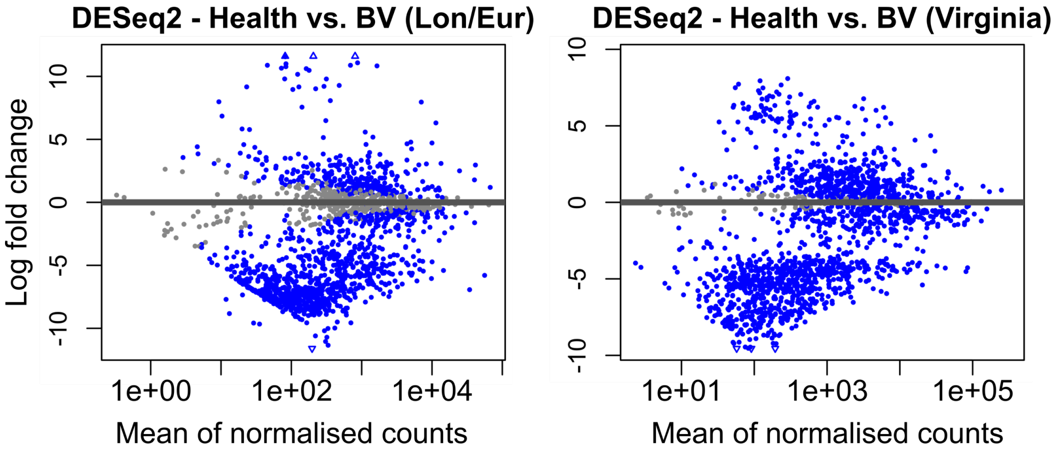
The reviewer is correct in that we are measuring the expression of genes across the microbial community- not the relative abundance of a given taxon in the microbiome. We have changed the wording across our manuscript accordingly. It is now explicitly clear that we are conducting differential expression analyses and that we are referring to the inferred taxonomic composition of the metatranscriptome based on proportional abundance of expressed transcripts in the dataset.

3. It’s curious that the original manuscript describing the Scale Reliant Inference method remains unpublished outside of preprint despite being on arxiv for more than 2 years. I understand the authors are not responsible for this work, but I wonder if they could comment on how much the underlying model/framework of SRI has changed over the 4 versions of the manuscript posted on arxiv. I would also point out that the implementation of SRI for Aldex2 also appears to be available only as a preprint.

We appreciate the concern regarding the delay in publication of the original SRI method paper; this is a source of constant frustration for us as well and can offer the following insight. Our collaborators, Drs. Silverman and Pistner-Nixon (Pennsylvania State University; authors of the paper in question and creators of the SRI approach) assure us that two years under review is unfortunately a common occurrence for statistical journals owing to lengthy review periods. We can confirm however that the framework underpinning SRI has not undergone any significant modifications that depart from how it is explained either in the current manuscript or in the pre-print explaining its implementation in ALDEx2. The latter paper is currently under review at the Annals of Applied Statistics.

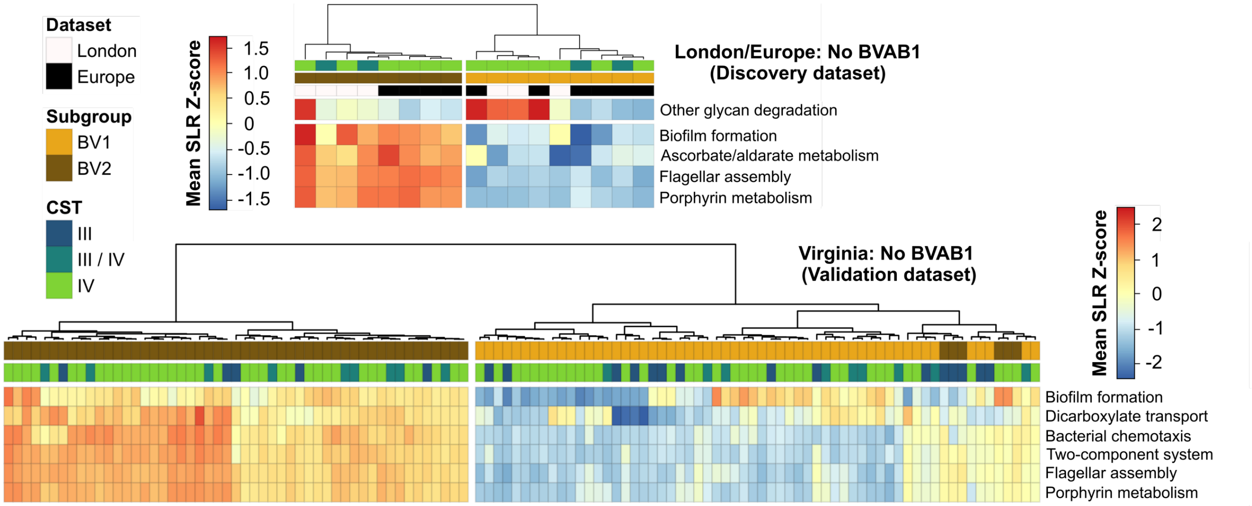
4. Many, if not all of the findings highlighted by the authors can be traced back to differences in the genomic content of the underlying microbes. E.g. Mobiluncus and BVAB1 have flagella, Gardnerella does not. These differences would have assumedly been apparent even using a method like deseq2. Could the authors point any particular differences that were enabled by the implementation of SRI to aldex2 or does this implementation only help to remove signal from overconfident estimates of differences?

We thank the reviewer for this comment and agree that many differences can be traced to gene content and are also detected by methods like DESeq2 (see Response Figure 1 below); however, **a major issue with DESeq2 is that it reports almost every single KO term as being significantly different between groups**, even after Benjamini-Hochberg correction. Unlike DESeq2, ALDEx2 does not suffer from this problem and instead returns a much smaller number of significantly different KOs (see the analogous plot from ALDEx2 in Supplemental Fig. S1). It is biologically counterintuitive that an overwhelming majority of genes will be different between groups. Therefore, using ALDEx2 with SRI allows users to be much more confident that the reported features are truly differential. Secondarily, the DESeq2 analyses of the Virginal dataset show a skew in the highly expressed genes such that many highly expressed genes in the BV group (bottom of the plot) are shown with a ~-2-fold difference (i.e. log2 fold change of ~-1) and are spurious false positives of the kind identified as being problematic in our scale ALDEx2 papers (see https://doi.org/10.1101/2023.10.21.563431 and https://doi.org/10.1371/journal.pcbi.1011659).



Response Figure 1: MA plot from DESeq2 comparing vaginal metatranscriptomes of healthy and BV groups for the discovery dataset (**left**) and the validation dataset (**right**). Input for DESeq2 was non-normalised read counts aggregated by KEGG orthology term (i.e. function). Blue data points represent significantly different KO terms.

Similarly, we present further evidence to show that the gene content of BVAB1 is not the primary driver of differences between the BV subgroups. We have now **repeated the differential abundance analyses between the BV subgroups excluding all genes assigned by VIRGO to BVAB1** (Response Figure 2). In both datasets, the main conclusions put forth in the original manuscript hold and functions such as 'Flagellar assembly', 'Biofilm formation' and 'Bacterial chemotaxis' remain significantly different between groups. This demonstrates that gene content differences between BVAB1 and – for example, *Gardnerella* spp.- are not the primary driver of the BV subgroup differences we reported here. We thank the reviewer for raising this important point and **we have now addressed this in the main text (lines 294-297), including the figure below in the supplementary material as Supplementary Figure S3** (all supplementary figure numbers have been updated accordingly).



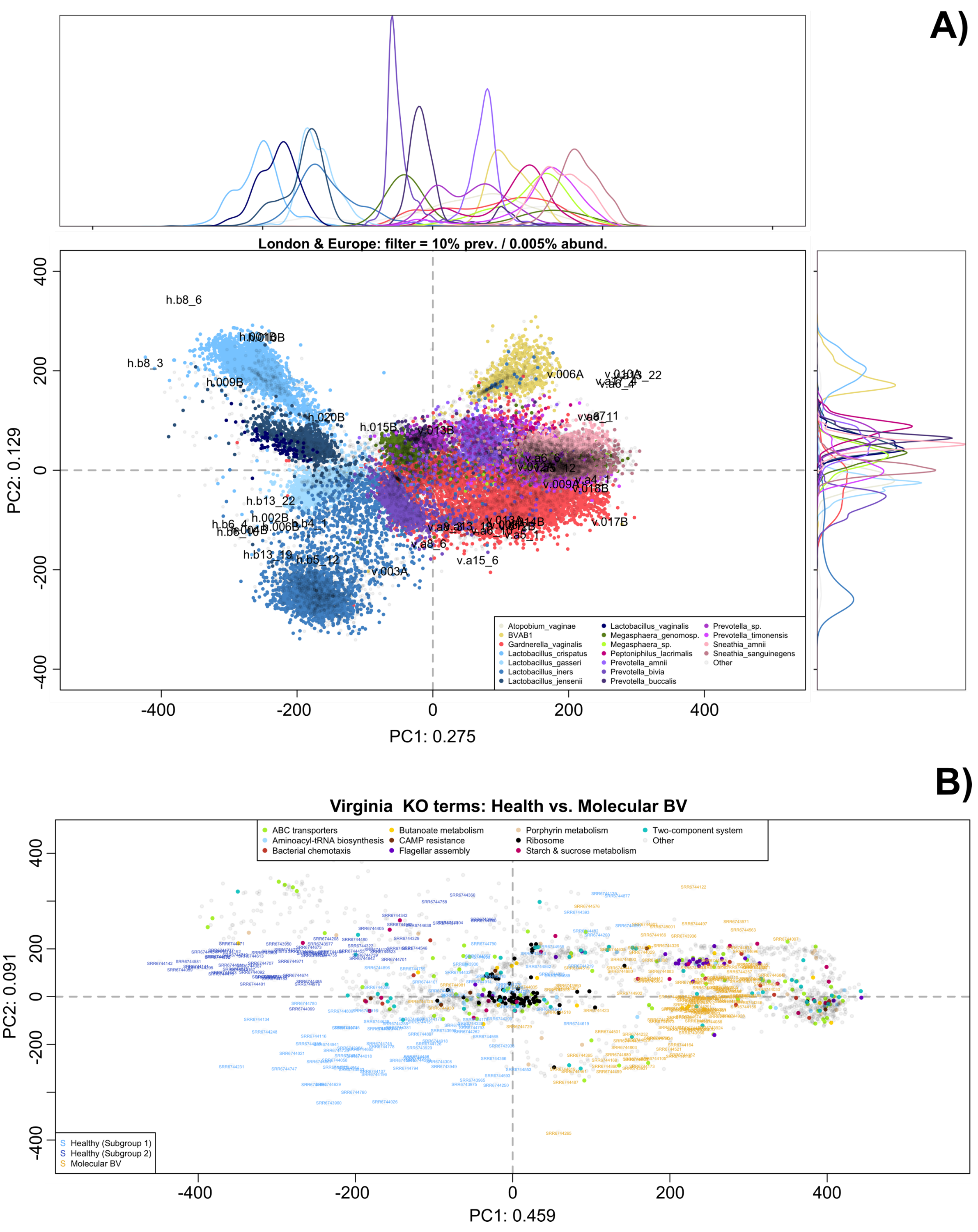
Response Figure 2: Differential expression analyses using ALDEx2 and scale-reliant inference were repeated for BV subgroups in both discovery (top) and validation (bottom) datasets at the functional level after removing all genes assigned to BVAB1, but prior to aggregating genes by KEGG orthology term. Data are expressed as a Z-score calculated from the mean scaled log-ratio (SLR) values of all KO terms assigned to a given pathway (individual SLR values represent a median across 128 Monte-Carlo instances). All post-SRI absolute effect sizes and false discovery rates calculated by ALDEx2 were >1 and <0.01, respectively).

5. The choice to remove genes present in less than 30% of the samples and with max observed relative expression of 0.005% seems too strict. Was there a particular reason to filter so deeply? This would mean a bacterial species present in ¼ of all women would have its entire transcriptome removed. You can see this effect in figure 3 where just the 16 named bacteria comprise approximately 98% of the trimmed metatranscriptome. What would this plot look like if the authors did not trim away so many of the genes? I imagine this could introduce tremendous bias in their estimation of the communities' transcriptional activity. The authors also do not report how many genes were removed filter, which would be useful in evaluating its effect.

We apologize for the oversight and recognise that we should have shown more of our initial exploratory information that led to our choosing these cutoff levels. We direct the reviewer to

Response Figure 3 below, and 'filtering_exploration.R' in the 'code' directory of study GitHub repository(https://github.com/scottdossantos/dossantos2024study; see the product of lines 97-115 and lines 159-177 of this R script). In Response Figure 3, we show that **even when using far less stringent filtering criteria** that encompass the reviewer's hypothetical situation of genes with a prevalence of 25%, **the taxonomic ordination for both datasets is almost identical**. Using the London/Europe dataset as an example, reducing the prevalence threshold from 30% to 10% added 10,847 genes (for a total of 30,811). **Of these >10,000 new genes, more than 95% of which have a sum across all samples <10,000 reads (i.e. ~200 reads/sample on average).** The situation is similar in the Virginia dataset: 20,682 genes were added by reducing the abundance filter threshold and >95% of these have a sum across all samples <40,000 (~130 reads/sample). Given the total number of genes across the entire, unfiltered datasets (330,123 London/Europe; 519,418 Virginia), this demonstrates that genes below both the filtering thresholds have relatively low read counts across all samples and including or excluding them has minimal effect on the ordination. **The methods section has been updated to reflect this (lines 647-649).**

Similarly, we would like to thank the reviewer for raising an issue regarding Figure 3 (and the related old Supplementary Figure S8)- specifically the fact that these 16 species comprise the majority of these vaginal metatranscriptomes. These stacked bar charts DID NOT include genes lacking a taxonomic assignment in VIRGO. **We have now corrected this and incorporated genes of unknown taxonomy into these plots (see Figure 3 and new Supplementary Figure S9)**. The species shown in these figures include the major species involved in states of both health and BV. Therefore, even when accounting for genes with an unknown taxonomic assignment, these species naturally comprise a majority of the metatranscriptome. This is consistent with the myriad culture-independent taxonomic studies of the vaginal niche. Furthermore, the 16 species in question are those represented by with more than 75 expressed genes across all metatranscriptomes. Below this threshold, there is a precipitous drop in occurrence, with many such taxa represented by only a few genes. **We can confirm that the genes below this threshold of 75 genes account for <1 % of reads in the less stringently filtered dataset** for both London/Europe and Virginia datasets (see the new R script, 'filtering_exploration.R' at lines 97-115 and lines 159-177). However, **any species that does not meet this threshold will still have its genes represented** in both the taxonomic PCA plots (under 'Other') and the differential expression analyses at a functional level (in which gene taxonomy is ignored and all genes are aggregated by functional KO terms). Therefore, this filtering introduces little to no bias into our analyses.



Response Figure 3: Both metatranscriptome datasets were re-filtered using a less stringent prevalence threshold of 10% and underwent log-ratio transformation with ALDEx2 as before. Species-level ordination is shown for London/Europe dataset (corresponds to Figure 1B) while functional-level ordination is shown for Virginia dataset (bottom; corresponds to old Supplementary Figure 5B). Ordination of genes/KO terms are the same regardless of filtering.

6. It was not clear to me how the "housekeeping genes" were selected. Were these chosen based on their functional annotations and our understanding of biology or did they get classified based on their expression patterns?

We thank the reviewer for raising this issue as we did not make it clear how these genes were selected in the manuscript. The reviewer is correct that these genes were chosen based on their functional annotations, and **we have now made it clear in the text that the housekeeping genes are "housekeeping functions"**. These housekeeping genes included those that are commonly targeted in qPCR assays for example- such as GAPDH and other glycolysis genes) or genes that play a role in essential cell functions AND are present in all species (i.e. aminoacyl-tRNA biosynthesis genes and ribosomal genes). These are, in general, not the major drivers of microbial behaviour of an ecosystem and- as per their use as loading/normalisation controls in qPCR assays- expected to be invariant between groups. Therefore, we can use these genes as a measure of how well 'centered' the data are (and by extension, how successful the data normalisation has been; see Supplementary Figure S1).

7. The authors spend a fair amount of time discussing two subgroups of BV that are distinguished by the presence of BVAB1 and the chemotactic and flagellar assembly proteins that it expressed. The splitting of BV into subtypes based on BVAB1 has been described in the literature and is included in the software used by many to assign samples to CSTs.
See: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00934-6.

We thank the reviewer for pointing out this oversight. We did not initially cite the above study- which employs clustering of vaginal microbiome profiles derived from sequencing of the 16S rRNA gene hypervariable regions- but now do so in both the introduction (lines 63-65) and results (lines 183-185) sections. We also highlight that BVAB1 is not solely responsible for these functional subgroup differences, as shown in Response Figure 2.

Minor

- In section 2.1, the authors should introduce the split between the initial and validation datasets.

  We now introduce the split between the discovery and validation datasets in section 2.1 as requested (lines 145-147).

- Figure 2. Color differences between III/IV and IV are miniscule. Difficult but not impossible to distinguish. Would suggest the authors choose different shades of green for these groups.

  We thank the reviewer for pointing this out and we agree that the colours are not readily distinguishable. We have altered the colour of CST IV for all heatmap plots. Similarly, we have also changed the colours representing 'Prevotella sp.' and 'Prevotella timonensis' on all plots as these were also not easily distinguished.

- Figure 3. Stacked bar plot labeled as "relative abundance" and described as "microbiome composition" but the data are from RNAseq. I would suggest at least clarifying that the what's being measures is the "relative abundance" of transcripts from the bacteria.

In accordance with major point 2, we now clarify this as requested.

- The names the authors use for the three datasets could be improved, three different geographic scales, city, state, and continent. I would suggest presenting them as "Canada", "USA", "Germany".

While we appreciate the request for standardisation, these terms are only labels and could equally be changed to an arbitrary 'A', 'B', and 'C'; therefore, we have opted to keep the labels as they are, reflecting how they are referred to internally by our study team.

- Page 2. You describe BV as being manifested as a depletion of Lactobacillus spp. Yet it is not clear if some women just never had lactobacilli to start with. I would suggest a more neutral phrase like a "paucity of lactobacilli"

We have made this requested change in the introduction (line 59).

- Page 2 and throughout: Atopobium has been reclassified as "Fannyhessea", see: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6113628/

We have changed the genus, '*Atopobium*', to '*Fannyhessea*' throughout the manuscript and relevant figures as requested and added the relevant citation; however, we note that this is the taxonomic assignment that exists in the VIRGO database.

- Page 6 The authors refer to "healthy" and "BV" patients but many two of the studies included in this analysis were derived from volunteer samples from participants.

We thank the reviewer for pointing this out and they are correct. We have now added a sentence in section 2.3 stating that use of this terminology is for convenience in distinguishing between groups of samples in differential expression and other analyses (lines 189-192).

- Page 6. Two additional Gardnerella species were given names, see: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.006140

We thank the reviewer for providing the reference, **which is now cited where we mention the new, named species of *Gardnerella* (lines 202-204 and 381-382)**. It is difficult to keep up to date with molecular bacterial taxonomics and we point out that 1) these figures were produced prior to the publication of the above study; and 2) the major databases of 16S rRNA genes- including VALENCIA- do not currently delineate any of the named species of *Gardnerella*, instead grouping everything under *G. vaginalis*.

For the revised manuscript, **we have repeated our analyses of *Gardnerella* species-specific genes, incorporating *G. greenwoodii* and *G. pickettii* as requested**. These are now found in Figure 4B and new Supplementary Figure S5. **We have also updated the full code used to perform the pangenome analysis and the R scripts for editing the VIRGO taxonomy** to reflect this change.

- Supplemental figure 1, the red points outlined in blue appear purple unless you zoom in quite a bit. Maybe just make them purple?

We thank the reviewer for their suggestion; however, it is intentional and necessary for interpreting the plot that the data points with a blue halo appear a distinct colour from those without a blue halo. Data points with a blue halo are those deemed to be differentially expressed based on effect size and BH-corrected P value (orange for housekeeping genes, red for non-housekeeping genes). Colouring all red points which are significant based on P value and effect size would make it very difficult to make out where the housekeeping genes are in the figure.

We agree that the plot colours are difficult to distinguish as is, not zoomed in. Therefore, we have made the blue halo around significant points larger and more transparent to emphasise the difference between non-housekeeping genes which do and do not meet the effect size threshold for significance. **We have replaced the old Supplementary Figure 1 with this image**. However, as we anticipate that most would-be readers of this article will be viewing the figure zoomed in on a screen, this figure has also been re-exported at a much higher resolution so that image quality won't be compromised while zoomed in.

Reviewer 2

Synopsis of the study:
The manuscript has multiple goals: 1) To establish the application of scale-reliant inference (SRI) to metatranscriptomic datasets. This reduces the rate of false positive detection through scaling of the dataset to remove the signal produced by house-keeping genes, which are not expected to be differentially abundant in distinct settings of the same microbial niche. 2) To identify genetic features of health and bacterial vaginosis (BV) by applying the SRI method to two distinct but aggregated datasets and validating these findings in an independent third dataset. The major reported findings are the successful application of SRI to the datasets evidenced by the same conclusions drawn from the initial and independent datasets, and the identification of two BV sub-groups driven by the expression of flagellar machinery of BVAB1, as well as the higher expression of cationic antimicrobial peptides (CAMPs) in healthy vaginal microbiomes compared to BV.

The manuscript demonstrates the application of scale-reliant inference (SRI) to metatranscriptomic datasets, Figure S1 is excellent. However, several points need clarification, such as the accurate description of the samples used from the Deng study, the study-design context, if any, of the Long and the Virginia samples (recent antibiotic use?), and the impact of recent metronidazole (MET) exposure on the observation of CAMP-related gene expression. Additionally, the novelty of designating BV subgroups requires further explanation, and the benefits of the SRI method over multi-project data aggregation alone should be more clearly demonstrated.

We thank the reviewer for these significant comments and they are addressed in the points outlined for reviewer 1. We have now more carefully described each cohort and clarified that only a subset of the Deng dataset was used: those that had been treated and been converted to a non-BV phenotype.

Specific Comments, (mostly) in order of appearance in the manuscript text:

- The reference regarding long-term recurrence should be altered to the original reference for this statement:
  - o Bradshaw CS Morton AN Hocking J, et al. High recurrence rates of bacterial vaginosis

over the course of 12 months after oral metronidazole therapy and factors associated with recurrence. J Infect Dis 2006;193:1478–86.

We have added in-text citations and the appropriate reference for this study as requested (line 57.

- Figure 1 caption: Were the reads in the feature tables corrected for gene length, an important step to ensure accurate gene expression values.

We thank the reviewer for their question regarding the data normalisation. No gene-length normalisation was performed as it is not needed if comparing between samples where gene length is simply a nuisance parameter. The ALDEx2 tool estimates the variability of measurement of the underlying data, and as such requires that the data be not preprocessed in any way. In the context of a metatranscriptome, the mRNA lengths for orthologous genes will be highly concordant because 1) most bacterial genes encode single-domain proteins, 2) orthology is commonly defined as having at least 90% overlap in protein sequence (shorter is often a sign of pseudogenes) along with some level of identity (usually >50%), 3) the same gene in different samples from the same organism will have identical length. Thus length is not likely to be a confounder. Read-length normalization, such as RPKM was first developed as a means of understanding the relative contribution of genes **in the same sample** to the transcriptome. **It has been shown to be not helpful in the general case and often introduces more bias than it solves** (see PMID: 22872506 and PMID: 20167110).

- If the study by Macklaim et al is from Canadian participants, why is it referred to as the London study? Are there any important aspects of this dataset (article not open-access or available on PUBMED) that are relevant to the interpretations of this manuscript (for example, the Deng study includes individuals post-MET).

We apologise to the reviewer for this confusion. We have now clarified that this was a simple convenience sample from a single clinic in London, Ontario (Canada). 'London' was the arbitrary label used internally by our team to refer to this dataset. This raw data for this dataset was used for two methodological papers as an example of difficult-to-analyze data and published as cited in the text, but formal analysis has not previously been performed or published for this dataset as we did not have the tools at the time to correctly normalise the data, as illustrated in Supplementary Figure 1. **We can confirm for the reviewer that there are no important aspects of this dataset that would change the interpretation of any findings in our manuscript.**

- The Deng study contains 37 individuals with BV that receive MET treatment, not 22, perhaps a typo (and in Table 1)? Later in the study they describe n=14 participants with V1 and V2 (before and after MET) samples that respond (n=8) and do not respond (n=6) to MET. This would be a total of n=28 samples. Please clarify the text and the table here to indicate which samples from each study were used for context and interpretation of the results as well as dependency between the samples (repeated measures).

- Figure 2C is also discordant with these numbers:

♣ n=15 healthy samples (7 from London, 8 from Europe).

♣ n =27 BV samples, (13 from London, 14 from Europe).

We apologise for the lack of clarity on the samples used and the group sizes. **We have now clarified this in the main text (lines 131-135)**. To answer the reviewer's question: we used only a subset of 22 samples out of the 40 samples listed in the Supplementary material of Deng *et al*. (2018). **The Deng samples classified as 'BV' in our study are pre-treatment samples for each individual patient, whereas the 'healthy' samples are post-treatment samples from patients classed as treatment responders** (column 'Metronidazole reaction' in the supplementary material of Deng *et al*. 2018). Specific ENA accession numbers for these samples can be found in the R script, 'merge_feature_tables.R'.

Regarding Figure 2C, there are 16 'healthy' samples in the figure: 15 cluster together, while the 16th (a sample from the London dataset) clusters with the BV samples. Health status is indicated by the middle colour bar beneath the dendrogram (sky-blue = healthy, orange = BV). There are a total of 26 BV samples and 16 healthy samples in Figure 2C (total n = 42). The **reviewer's question on sample numbers brought to light an additional typo**: we incorrectly stated that the total number of samples in the London/Europe (discovery) dataset was 44; we have now **corrected this to 42 throughout the manuscript and in Figure 1**.

- Did the analysis account for the dependency between BV and healthy samples of the Deng study? For example, if multiple samples from the same person are included in the BV group, this could incorrectly inflate the significance of gene up- or down-regulation, right? In the Deng study, those that did not respond to treatment have very similar before- and after treatment profiles (1-2 weeks). Should these be treated as independent samples/BV events?

We thank the reviewer for this question: the BV group did not include multiple samples from the same individual. All BV samples come from different participants as we only selected the first, pre-treatment sample. All samples were treated as unpaired during analyses; however, despite this, we were still able to replicate the results in a much larger, independent dataset whose demographics were drastically different from the studies by Deng et al. (2018) and Macklaim et al. (2018). Indeed, it was incredibly gratifying to observe that the three cohorts were concordant in so much of the analysis given their different populations, treatment statuses and sequencing dates. This leads us to believe that the results reported here are robust and we do not believe these conclusions are the result of over-inflated estimates of differential expression.

- The healthy samples in the Deng study were recently post-metronidazole. Couldn't this impact the interpretation of the CAMP finding as possibly not a phenotype of BV, but rather of recent MET exposure? Please clarify the text and the table here to indicate which samples from each study were used for context and interpretation of the results.

We thank the reviewer for their question and offer the following explanation. We report a number of CAMP resistance genes differentially expressed in health (e.g. DltABCD) and BV (various) in the discovery dataset (Deng dataset (met-exposed) combined with Macklaim dataset (no met exposure)). The same Dlt genes were also all found to be significantly over-expressed in healthy samples in the much larger validation dataset. Given the lack of exposure to metronidazole in both the London dataset and Virginia dataset (validation), and the fact that differentially expression of these Dlt genes was not specific to the Deng cohort, we do not think that metronidazole exposure is a convincing explanation for this finding.

o Figure 2A suggests that most of the healthy samples were from L. iners. Is this the source of the Dlt genes (VIRGO gives taxonomy and functional annotation, no?). In the Deng study, many of the post-MET responders ("healthy") samples are L. iners predominated, a commonly observed species after antibiotics. Similar to the CAMP findings, isn't it possible these genes are associated with a post-MET phenotype rather than a "healthy" phenotype?

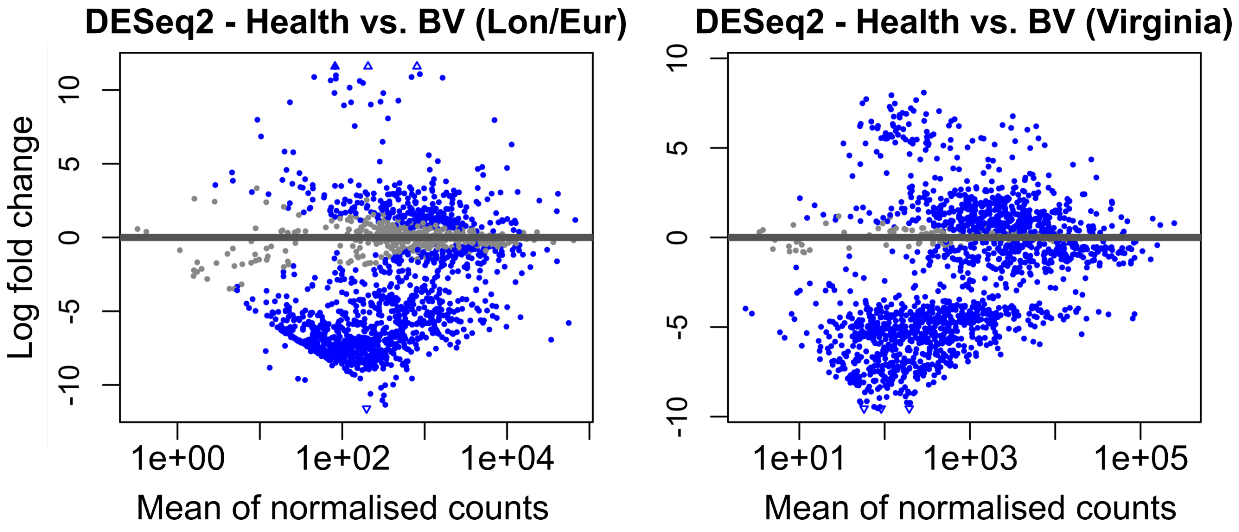
We thank the reviewer for their question as we had not previously included any information about the potential taxonomic origins of these genes. We expand upon this now in Section 2.4 for the London/Europe dataset (lines 225-227), and Section 2.8 for the Virginia dataset (lines 417-419). **We summarise the taxonomic assignment of these genes below in Response Table 1, but briefly, most of these genes and reads originated from non-*L. iners* lactobacilli. This table has now been incorporated into the supplementary material as Supplementary Table S1**. Regarding the metronidazole question, these four Dlt genes were identified as significantly overexpressed in the healthy groups for *both* the discovery and validation datasets and these genes represent the four KO terms comprising CAMP resistance (H) in the two corresponding heatmaps for these analyses (Fig 2C and Fig5A). Therefore, our response is the same as to the comment above: there was no metronidazole exposure in either the London or Virginia datasets, make metronidazole exposure an unconvincing explanation for the differential expression of these genes in our meta-analysis.

Response Table 1: Taxonomic assignments of genes assigned to the four KEGG orthology terms representing the DltABCD operon, in London/Europe (discovery) and Virginia (validation) datasets.

	London/Europe		**Virginia**	
Taxon	**No. genes**	**Read counts**	**No. genes**	**Read counts**
L. crispatus	5	115,858	8	7,191,584
L. iners	8	177,871	8	5,429,846
L. gasseri	2	14,043	2	144,203
L. jensenii	2	79,900	7	1,613,482
Unknown taxonomy	9	41,381	11	1,458,737

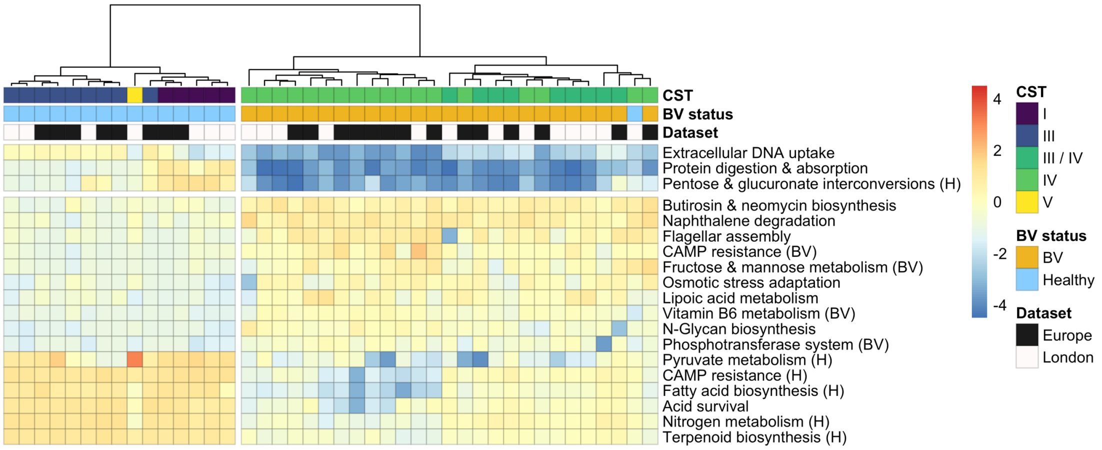
- Given that a major part of the paper is to show the efficacy of the SRI method, would the same results have been found in the absence of the method?

We thank the reviewer for their question and note that the other reviewer posed a similar one. Supplementary Figure S1 shows the consequence of NOT using SRI: a large number of false-positive and false-negative findings following the differential expression analysis. Many of the KO terms corresponding to housekeeping genes are incorrectly called as significantly differential. We show in response to Reviewer 1 (and below for convenience) that when using DESeq2 for differential expression analysis, almost all KO terms are called as significantly differential for discovery and validation datasets. When one compares this plot to the analogous plot for ALDEx2 (e.g. Supplementary Figure S1C for the discovery dataset), ALDEx2 using SRI does not suffer from this issue and one can be more confident in the results.



Response Figure 1: MA plot from DESeq2 comparing vaginal metatranscriptomes of healthy and BV groups for the discovery dataset (**left**) and the validation dataset (**right**). Input for DESeq2 was non-normalised read counts aggregated by KEGG orthology term (i.e. function). Blue data points represent significantly different KO terms.

We also show for the reviewer that not incorporating SRI into ALDEx2 greatly weakens the strength of the findings regarding CAMP resistance (compared to Figure 2C), using the discovery dataset as an example. For this analysis, no scale model was applied when running ALDEx2. The differences between healthy and BV groups become borderline (Response Figure 4). Instead, pathways such as ‘Pentose & glucoronate interconversions’ are the drivers of difference between these two groups. This figure and the corresponding code are both available at the study’s GitHub repository as ‘suppl_lon_eur_diffAbund_noSRI.png’ and ‘lon_eur_noSRI.R’, respectively. One major advantage of the scale model approach is that spurious false positive results are much reduced and true positive results are thus amplified. In the case of the DESeq2 analysis this is because of a misspecification of the location of all gene functions, that is observable most easily when examining the highly expressed housekeeping functions at the far right of the MA plots above.



Response Figure 4: Differential expression analysis in the London/Europe dataset without using SRI. The analysis was repeated without including a scale-model in ALDEx2 and the top 20 differentially expressed pathways are shown. Note the same colour scale for the heatmap as in Figure 2C.

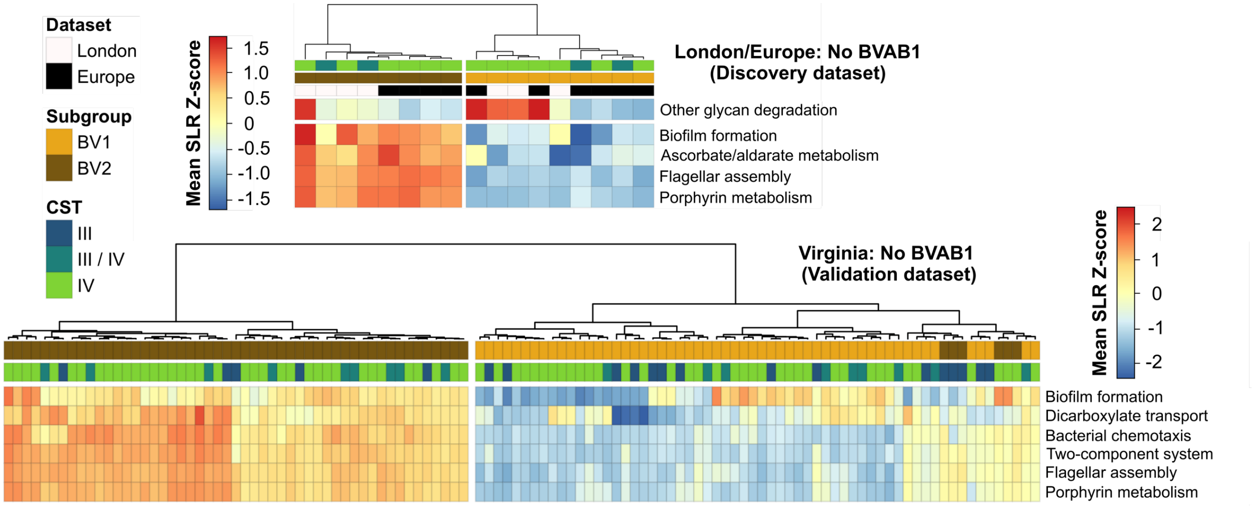
o Figure 2: The results in 2B greatly reflect those of the original study by Deng et al. (25) (Macklaim & Gloor (24) was not accessible to me, so I am unable to compare these results). Is it possible that it is not the application of the SRI method, but rather the increased power via data aggregation that produced the given results? It would be nice to see the effects of the SRI method on the aggregated (multi-project) analysis. For instance, in Figure S5, it doesn't seem that the addition of the SRI method altered the results and interpretation. It is a challenge to identify the benefit of the SRI method with the current presentation of analyses.

We thank the reviewer for their question and note that Figure 2B reflects almost every culture-independent study of the vaginal microbiome in the context of BV, not just the study by Deng *et al*. (2018). As in the response to the previous point raised by the reviewer, **one of the main advantages of SRI is proper normalisation of functional RNA-seq data**. Vaginal metatranscriptomics represents a worst-case scenario for data normalisation as there is a vast difference in gene content and total bacterial load between healthy vs. BV populations (i.e. a difference in scale), which introduces the issues outline in Supplementary Figure S1 (see also https://doi.org/10.1101/2023.10.21.563431 for more detail on this problem). In old Supplementary Figure S5 (new Supplementary Figure S6), one can see the effect of normalising the data properly, **whereby KO terms corresponding to housekeeping genes (ribosomal genes/ tRNA biosynthesis genes) only become properly centred when accounting for scale using SRI**. In our response to the previous comment, as well as in Supplementary Figure S1, we show the consequence of not using SRI- there are a large number of false-positive and -negative results and the strength of the findings from the differential expression analyses are considerably weaker to the point of becoming borderline.

- Regarding this statement: "A comparison of the two main BV subgroups at the functional level demonstrated that genes involved in motility (flagellar assembly), heme and cobalamin/vitamin B12 biosynthesis (porphyrin metabolism) and chemotaxis were the main drivers of this difference within the classical archetype of molecular BV (i.e. CST IV with/without BV symptoms; Figure 3, top). Based on the taxonomic composition of these samples, the presence or absence of BVAB1 (BV-associated bacterium-1; recently reclassified as "Candidatus Lachnocurva vaginae" [45]) appears to explain these differences almost entirely (Figure 3, bottom)."

o The designation of BV1 and BV2 as "the classical archetype of molecular BV (i.e. CST IV with/without BV symptoms)" is unfounded. What study has identified a molecular definition of asymptomatic vs. symptomatic BV? Instead, the difference appears to be simply CST IVA (BVAB1) vs CST IVB (Gardnerella), for example: doi 10.3389/fimmu.2021.730986 or 10.1186/s40168-020-00934-6, with the described functions unsurprising given those described of BVAB1 (doi: 10.3389/fcimb.2020.00117) and Mobiluncus (doi: 10.3389/fcimb.2021.759697), such as motility.

We thank the reviewer for raising this point and apologise for the poor wording of this sentence. We did not intend to imply that there is a molecular distinction between symptomatic and asymptomatic BV; however, we also acknowledge that the way this sentence is currently worded very much implies this. **We have now changed the wording to make it clear that BV1 and BV2 are two functional subgroups of what is commonly referred to as CST IV-A and CST IV-B (lines 290-293).** Furthermore, we refer this reviewer to a similar question posed by the other reviewer: we can be confident that **BVAB1 is not the sole driver of these subgroup differences as, when we remove all genes assigned to BVAB1, the same functions are still significantly overrepresented** (flagellar assembly, biofilm formation, chemotaxis etc.). Response Figure 2 is repeated below for convenience.



Response Figure 2: Differential expression analyses using ALDEx2 and scale-reliant inference were repeated for BV subgroups in both discovery (top) and validation (bottom) datasets at the functional level after removing all genes assigned to BVAB1, but prior to aggregating genes by KEGG orthology term. Data are expressed as a Z-score calculated from the mean scaled log-ratio (SLR) values of all KO terms assigned to a given pathway (individual SLR values represent a median across 128 Monte-Carlo instances). All post-SRI absolute effect sizes and false discovery rates calculated by ALDEx2 were >1 and <0.01, respectively).

- Figure 6: It would be ethical to refer to the other studies that contributed to the development of the hypothesized mechanisms within the figure.

The reviewer is correct to point this out and we have included references to the appropriate studies for each of the numbered mechanisms in Figure 6. **The following studies are now cited in the legend for Figure 6**:

1. CAMP resistance genes – Neuhaus et al. 2003; *Microbiol Mol Bio Rev*. **67**: 686-723.
2. Acid-activated urea channels - Strugatsky *et al*. 2013; Nature. **493** (7431): 225-8.
3. Iron acquisition – Chan et al. 2023; Biometals. **36** (3): 683-702.
4. Collagenase activity – Lithgow et al. 2022; *Am J Obstet Gynecol*. **226** (2): 302.
5. Biogenic amines - Nelson et al. 2015; *Front Physiol.* **6**: 253.
6. Biofilm formation – Choi *et al*. 2009; *J Bacteriol*. **191** (19): 5953-63.
6. Motility – Armitage *et al*. 2020; Ann Rev Microbiol. **74**: 181-200.
6. Chemotaxis – Miller *et al*. 2009; Adv Appl Microbiol. **66**: 53-75.
7. LL-37 transcription – Kiattiburut *et al*. 2021; *J Urol*. **206** (3): 491