Sumthin' cute here

Jean M. Macklaim[1,], Amy McMillan[2,], Name3 Surname[2,¤a], Name4 Surname[2,‡], Name5 Surname[2,‡], Gregor Reid[2], Gregory B. Gloor[3,*], with the Lorem Ipsum Consortium[¶]

1 Affiliation Dept/Program/Center, Institution Name, City, State, Country
2 Affiliation Dept/Program/Center, Institution Name, City, State, Country
3 Affiliation Dept/Program/Center, Institution Name, City, State, Country

‡Corresponding Author
¤a Insert current address of first author with an address update
¶Membership list can be found in the Acknowledgments section.
* CorrespondingAuthor@institute.edu

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

Introduction

The vaginal microbiome is important

The vaginal microbiome has been characterized using 16S rRNA gene sequencing with the result that ...

The vaginal meta-transcriptome was characterized using a very small sample size ...

The vaginal metabolome has been characterized with the result that (confirms and extends) ...

Here we report the integrated analysis of the vaginal microbiome, meta-transcriptome and metabolome on an overlapping set of samples from (N) women from London Ontario. We find that the ..., and .... We suggest that ...

Materials and Methods

Ethics.

Collection.

Metabolome determination.

RNA isolation and sequencing.

Data analysis.

Read Mapping. XX genomes comprising the set of species observed in the human vagina from 16S rRNA gene sequencing (REFS), culture studies (REFS) and additional genomes from organisms suspected to occur were downloaded from Genbank on ????. Open reading frames from these genomes were clustered using ?? with the following criteria (length, PID, etc) and a representative sequence was chosen to be the centroid using the following rule: ???. Centroid sequences were annotated by BLAST to SEED and KEGG databases and the best hit supplying the annotation. The taxonomy of the centroid sequences was taken to represent the taxonomy of the cluster. We will refer to the centroid sequence as the 'transcript'. Supplementary Table S1 contains the list of accession numbers. Supplementary Table S2 contains the set of centroid sequences and their inferred annotation.

Reads in fastq format were mapped to the library of centroid sequences using bowtie2 [1] and unmapped reads were assembled with ???. The assembled fragments were annotated as above, and added to the master table of counts. In the end there were an average of XXX million reads mapped to each sample, and Table S3 contains the pertinent information. Reads were further grouped by KEGG function or SEED level 4 subsystem to produce two tables of functions, these will be referred to as 'functions'.

Statistical Model for RNA-seq. The resulting table appears on the surface to be a table of counts that only requires normalization to a constant sequencing effort prior to analysis [2]. However, we and others have found this approach often fails in meta-RNA sequencing because of the interplay between the organism and transcript abundance [3–5]. Thus, it much more informative to model the observed count for a given gene in a sample as a distribution of probabilities that the count was observed given the total number of sequence reads obtained per sample [3]. Such a model falls into the count compositional data analysis paradigm [6] where only the relative differences between abundances in a sample are informative [7,8].

Compositional data has substantially different properties than do unconstrained data. The most important being that compositional data exhibits 'spurious correlation' between parts because the possible values for the parts are constrained by the constant sum [7,8]. Correlations are also unreliable because they change in unpredictable ways when a subset of the parts are examined. Thus, traditional correlation metrics are unsuitable when examining these data. Note that all RNA-seq datasets are subcompositions unless ribosomal RNA, tRNA, etc are not depleted and are included in the analysis [7,8].

Some of these problems can be ameliorated by log-ratio transformations such as the centered log-ratio transformation (clr). Given a vector of numbers that contains D parts, $x = [x_1, x_2, ...x_D]$, x can be converted to a vector of clr values, z, as follows:

$$\mathrm{clr}(x) = \left(\ln \frac{x_1}{g_x}, \ln \frac{x_2}{g_x}, \ldots, \ln \frac{x_D}{g_x} \right) = (z_1, z_2, \ldots, z_n) . \tag{1}$$

Where g_x is the geometric mean of vector x.

Exploratory data analysis was performed as point estimates with compositional biplots [9]. These show both the distances between samples and the variances of the transcripts. The 'count zero multiplicative' zero replacement method from the zCompositions R package [10] was used to adjust 0 values for the likelihood that the 0 represents a non-detect event in the sample.

When conducting quantitative analyses, the joint probability distributions for all genes in a sample were generated by Dirichlet multinomial sampling within the ALDEx2 Bioconductor R package using a uniform adjustment of Jeffrey's Prior [3]. These distributions were transformed using the centre log-ratio transformation prior to analysis citeAitchison:1986 which allows the value to be unconstrained. The expected value of ϕ [5] from the distribution was used to determine compositionally-linked transcripts or functions, since it has been shown that traditional correlation metrics give unpredictable results [7, 8]. The expected value of Kendall's Tau (b) was used when reporting the correlation between transcript abundance and metabolite abundance since it is not expected that the metabolome and meta-transcriptome tables share even passingly similar units or scaling.

Reproducibility. All R code needed to reproduce this analysis from the count tables can be accessed at: github...

Results and Discussion

High throughput sequencing experiments generate datasets where the total number of reads per sample are irrelevant, thus these data are compositional and contain only relative information about abundances [4]. Such data can be examined in a rigorous manner by examining the variation in ratios between all pairs of transcripts [7, 8]. However, this can be dramatically simplified using the centre log-ratio transformation, which reduces each count value to a ratio of that count to the geometric mean count within each sample. The logarithm of the ratio makes any change in relative abundance symmetrical. The clr transformation is functionally equivalent to a matrix of all vs all ratios, and compensates for differences in read abundance, thus eliminates the need for count normalization [5]. We and others have shown that clr-transforming the data constitutes the first step towards a compositional data analysis approach that is generally useful to characterize HTS [3–5, 11]. All data analysis is thus done in a compositional analysis framework with clr-transformed data.

Two caveats need to be expressed at the outset. First, the clr transformation cannot be performed in the presence of transcripts with 0 counts since it involves taking a logarithm. Initially all transcripts that have 0 reads in all samples, so-called 'essential zeros' are removed. We take two approaches for the remaining 0 value. When a point estimate of 0 is required, we use the count zero multiplicative method [10] that replaces each 0 value using a Bayesian estimate of the most likely value of 0 in each sample. This approach is used to adjust the data for the Singular Value Decomposition that is required for Principle Component Analysis (PCA). When a more robust estimate is required, we generate a Bayesian estimate of the range of values for each transcript (including 0 count transcripts) that are consistent with the original dataset [3]. All subsequent analyses are conducted on the resulting distribution of values, and the final reported value is the expected value of the p-value, ϕ [5], or effect-size metric.

0.1 Exploratory analysis

We collected NN samples for RNA-seq, and MM were sequenced on the Illumina HiSeq platform at The Centre for Applied Genomics in Toronto. We also included in the RNA-seq results four samples sequenced using the ABI SoLid platform and reported previously [?]. We examined these samples using taxonomic abundance profiles determined both from 16S rRNA gene sequencing and from the reference sequence abundance; using compositional biplots and unsupervised clustering on center-log-ratio transformed data; and by correlating the samples with the clinical phenotype. In the end, we excluded six samples (Supplemental Text) since they had either a very distinctive taxonomic abundance profile or that profile did not match the clinical phenotype. This left 22 samples for analysis composed of three samples sequenced by ABI-SoLid and 19 sequenced by Illumina HiSeq.

The twenty-two samples were filtered to remove all transcripts with an average of 2 or fewer reads across all 22 samples, this simplified the dataset from 48000 transcripts to ∼1000 transcripts. A compositional biplot of this dataset is shown in Fig. 1A. Here we can see that the samples partition nicely into two groups, healthy on the left and BV on the right.

Compositional biplots summarize in one plot both the relationships between the samples and the contributions of the variables to those relationships [9]. In this plot, the distance of a transcript from the origin is directly related to the standard deviation of the clr value of the transcript in the samples. The direction of the transcript relative to the samples shows the direction of greatest abundance. These two rules hold true up to the limit of the projection of the PCR. Moreover, we must interpret these plots with care, since there is *no information* regarding absolute abundance. Thus, two transcripts with widely varying absolute abundances could (and often do) have similar locations on the biplot.

Inspection of the location of the transcripts indicates that they are largely separable, with transcripts mapping to *Lactobacillus crispatus* furthest to the left and transcripts from several taxa associated with BV furthest to the right. As noted, their locations are proportional to their standard deviation in the entire dataset and so can be read out as the relative contribution of each transcript to the sample containing it. Interestingly, transcripts annotated as belonging to *Lactobacillus iners* are near the middle of PC1, indicating that knowledge of transcripts associated with this organism contribute little to the health-BV separation. We also note that the major lactobacillus groups separate much more strongly than do the taxa associated with BV. This could indicate that healthy microbiotas colonized by a near monoculture of one or the other lactobacillus species have disticnt ways of being healthy, or it could be an artefact of the non-overlapping gene content of these organisms.

Within BV, we observe that *Megasphaera* and *Prevotella* species form two or more distinct foci suggesting the presence of distinctive species or strains of these organisms in BV. Interestingly, for *Megasphaera* one of these foci is very close to the origin, suggesting that this strain or species is contributing little to the overall BV phenotype. Finally, several de-novo assembled transcripts appear to be major contributors to the BV phenotype.

Not surprisingly, the reference sequence-based biplot shows that taxonomic abundance is the major driver of the differences of transcript abundance between samples. We were interested to determine if the different states had different underlying functions regardless of the taxonomic composition. Thus the reads were grouped by SEED subsystem 4 function and Fig. **??**B shows the result. Here we can see that the functions partition more strongly along the major axis of variance with 56.2% of the variance explained on PC1, and only 8.5% on PC2. We observe that the samples group into the same groups as in Panel A, indicating that this split is robust

Figure 1. Exploratory analysis of the reference sequence transcripts. Panel A shows compositional biplots of the reference sequence transcripts filtered to include transcripts present at an average count of more than 2 reads per sample. Compositional biplots are Principle Component (PC) Plots generated from the singular value decomposition of the centre log-ratio transformed dataset (REF). The bottom and left axes show the unit scaled measures for PC1 and 2, and the top and right axes show unit scaled variances for the transcripts. These plots thus show the relationship between sample distance along with the contribution of the transcripts to that distance. In this plot component 1 and 2 explain over 58% of the variance in the dataset, which is exceptional for a dataset of this complexity. Sample names are shown in black text. All samples partitioning with negative values on PC1 are classified as clinically healthy, and all samples with positive PC1 values are classified as clinically BV. The location of each transcript is shown as a coloured point, with the colors corresponding to the taxonomic assignment of the centroid sequence. Panel B shows a biplot where the transcripts are summed by SEED subsystem 4 annotation. Panel C shows the taxonomic distribution of the samples inferred by sequencing the V6 variable region of the 16S rRNA gene , the inferred composition based on the taxonomic assignments of the RNA-seq reads, and a dendrogram and associated heat map based on the similarity between samples based on SEED subsystem 4 function abundance.

to aggregation of the data. Samples composing the healthy group on the left side of the biplot are associated with a relatively small set of functions that are proportionally increased in these samples, and the samples composing the BV group are associated with a large set of functions that are proportionally increased in the BV samples. We observe that the majority of functions are at the origin, indicating the presence of a core set of functions that are required regardless of condition. The health and BV samples also partition along PC2 similarly to the partitioning seen in Panel A. Correlating the organism abundance in each sample with the PC1 and PC2 location shows that the partitioning on component 2 is largely associated with the presence or absence of *L. iners*. Samples containing this organism have positive PC2 scores, and samples lacking or almost completely lacking this organism have negative PC2 scores.

What can we say about the dendrogram?

0.2 Differential abundance

Traditional statistical approaches, while widely used, are inappropriate when examining compositional data since these datasets contain information only on the ratios between the components (REFS). For this reason, we examined the correlation of functions using the ϕ-metric that assesses the stability of the ratios across samples of all possible pairs of transcripts (REFS). Samples that have stable ratios are said to be compositionally associated and have a value of *phi* near 0. In the original paper, the authors removed from analysis all genes that had a 0 count in any sample and then calculated the clr and ϕ as point estimates. We took a different approach and estimated the distribution of values that 0 could reasonably take and then clr-transformed the data using the aldex.clr function from the ALDEx2 Bioconductor package (REF). The ϕ measure was calculated for each independent element of the distribution and the expected value of ϕ was reported. This approach allowed us to determine compositional association values for functions with a dichotomous distribion, while still maintaining low false positive background.

Figure **??** shows a graphical depiction of ϕ groupings determined for the SEED subsystem 4 functional grouping (A) and the KEGG functional grouping (B). We found that the data partitioned very similarly with both of these functional

aggregation schemes underscoring the robustness of the approach. In both approaches, there were three major clusters that partitioned on PC1. The first group, composed of XX members was the small group of functions that were much more relatively abundant in the healthy than in the BV group. Inspection of these functions shows that they are belong to ???. These same functions were generally observed to be differentially abundant in our previous analysis (REF). The second group which is nearer the origin was also relatively more abundant in health than in BV, and were found to correspond largely to replication and gene expression housekeeping genes. This confirms our previous observation that these abundant housekeeping genes are slightly increased relative tot he remainder in the healthy microbiome. The third major group had a much greater relative abundance in BV than in health. These functions are largely functions that are absent in lactobacilli and correspond to functions involved in the greater biosynthetic and respiratory capacities of the majority of the BV-associated organisms (REF). In addition to the three major groups, there were a number of smaller clusters. These were generally found to correspond to functions in the same pathway or in many cases to co-expressed subunits. A full list of functions and their associated group is found in tabular format in the supplement.

Figure 2. Compositional association of microbial functions. Panel A shows a compositional biplot of the SEED subsystem 4 aggregated data with different groups of compositionally associated transcripts coloured by group membership. Panel B shows the same for data aggregated and analyzed by KEGG annotation. In both cases we determined the expected value of ϕ with a cutoff of 0.03.

Table 1. Table caption Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam.

Heading1				Heading2			
$cell1row1$	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
$cell1row2$	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
$cell1row3$	cell2 row 3	cell3 row 3	cell4 row 3	cell5 row 3	cell6 row 3	cell7 row 3	cell8 row 3

Table notes Phasellus venenatis, tortor nec vestibulum mattis, massa tortor interdum felis, nec pellentesque metus tortor nec nisl. Ut ornare mauris tellus, vel dapibus arcu suscipit sed.

Supporting Information

S1 Video

Bold the first sentence. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Text

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S2 Fig

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table

Lorem Ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357–359.

2. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013 Sep;8(9):1765–86.

3. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLoS ONE. 2013 July;8(7):e67019.

4. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2:15.

5. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol. 2015 Mar;11(3):e1004075.

6. Aitchison J. The Statistical Analysis of Compositional Data. Chapman & Hall; 1986.

7. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Modeling and Analysis of Compositional Data. John Wiley & Sons; 2015.

8. Pawlowsky-Glahn V, Buccianti A. Compositional data analysis: Theory and applications. John Wiley & Sons; 2011.

9. Aitchison J, Greenacre M. Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2002;51(4):375–392.

10. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems. 2015;143(0):85 – 96. Available from: `http://www.sciencedirect.com/science/article/pii/S0169743915000490`.

11. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012;8(9):e1002687.

References

1. Devaraju P, Gulati R, Antony PT, Mithun CB, Negi VS. Susceptibility to SLE in South Indian Tamils may be influenced by genetic selection pressure on TLR2 and TLR9 genes. Mol Immunol. 2014 Nov 22. pii: S0161-5890(14)00313-7. doi: 10.1016/j.molimm.2014.11.005

2. Huynen MMTE, Martens P, Hilderlink HBM. The health impacts of globalisation: a conceptual framework. Global Health. 2005;1: 14. Available: http://www.globalizationandhealth.com/content/1/1/14.