Sumthin' cute here

Jean M. Macklaim[1,2], Amy McMillan[1,3], Jo-Anne Hammond[4], Michael John[5], Mark Sumarah[6], Johnathan Swan[7], Gregor Reid[1,3,8], Gregory B. Gloor[1,2,*], with the VOGUE Research Group[¶]

1 Canadian Centre for Human Microbiome and Probiotic Research, Lawson Health Research Institute, The University of Western Ontario, London, Ontario, Canada
2 Department of Biochemistry, The University of Western Ontario, London, Ontario, Canada
3 Department of Microbiology and Immunology, The University of Western Ontario, London, Ontario, Canada
4 Department of Family Medicine, The University of Western Ontario, London, N6A 5C1, Canada
5 Department of Pathology and Laboratory Medicine, The University of Western Ontario, London, N6A 5C1, Canada
6 Agriculture and Agri-food Canada, London, Ontario, Canada
7Department of Food and Nutritional Sciences, School of Chemistry, Food and Pharmacy, University of Reading, Reading RG6 6AP, United Kingdom
7 Department of Surgery, The University of Western Ontario, London, Ontario, Canada

¶Membership list can be found in the Acknowledgments section.
*** ggloor@uwo.ca**

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 used high-throughput sequencing of 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 per sequence feature in each sample. 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 table of counts mapped to reference sequences, and aggregated to either the SEED subsystem or KEGG KO level, appears on the surface to be a table of counts. Common practice is to normalize the reads per feature such that the 'sequencing effort' is a constant or each sample [2]. However, we have found this approach 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 feature 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]. Upon reflection, this approach makes intrinsic sense because the number of reads obtained in any high throughput sequencing experiment is constrained by the capacity of the instrument. We know intuitively that the same library will give a more accurate depiction of reality with the ~ 200 million reads obtained from a HiSeq lane than with the ~ 20 million reads obtained from a MiSeq.

Compositional data has substantially different properties than do unconstrained data. The most important being that compositional data exhibits 'spurious correlation' between features because the possible values for the features 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 [5]. 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 features, $x = [x_1, x_2, ...x_D]$, x can be converted to a vector of clr values, z, as follows:

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

Where g_x is the geometric mean of vector x.

One complication is that the geometric mean cannot be computed when the x contains a value of 0. Thus two approaches were taken. First, when a point estimate was required the 'count zero multiplicative' zero replacement method from the zCompositions R package [9] was used to adjust 0 values for the likelihood that the 0 represents a non-detect event in the sample. Second, when conducting quantitative analyses, the joint probability distributions for all features in a sample were generated by Dirichlet multinomial sampling within the ALDEx2 Bioconductor R package [3] using a uniform adjustment of Jeffrey's Prior. This returns a distribution of possible values for the probability of the feature given the feature count and the total count for the sample [3, 10]. We have shown previously that the relative error in these probability distributions is largest for features with 0 counts and smallest for those with the largest number of counts in both real and simulated data [3]. The distributions were subsequently transformed using the centre log-ratio transformation prior to further analysis.

Exploratory data analysis was performed as point estimates with compositional biplots [11] following clr transformation of the data. These show both the distances between samples and the variances of the transcripts.

The expected value of ϕ [5] from the clr 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. We used an effect size cutoff of 2 when evaluating differential abundance between conditions, a cutoff that was used in our previous investigations [12].

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 [6–8]. 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, 13]. All data analysis is thus done in a compositional analysis framework with clr-transformed data as outlined in the Materials and Methods.

0.1 Exploratory analysis

We collected NN samples for RNA-seq and metabolomic analysis, 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 [12]. 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 clr-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 had a microbiome profile did not match the clinical phenotype. This left 22 samples for RNA-seq analysis composed of three samples sequenced by ABI-SoLid and 19 sequenced by Illumina HiSeq. The table of counts for 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 10052 transcripts.

Figure **??** shows ...

A compositional biplot of this dataset is shown in Fig. 1A. Compositional biplots summarize in one plot both the relationships between the samples and the contributions of the variables to those relationships [11]. In this plot, the distance of a transcript from the origin is related to the standard deviation of the clr value of the transcript in the samples: it is *not* directly related to the absolute abundance of the transcript. The direction of the transcript relative to the samples shows the sample group with the greatest abundance of the transcript. Our ability to interpret the distance and direction information is only as good as the projection of the PCR, which is determined by the proportion of variance explained on the plot. The biplot in Figure 1 explains 47.7% of the variance on PC1 and 11.2% of the variance on PC2. We conclude that this representation is a good one since these two principle components explain nearly 60% of the variance in such a large dataset.

There are several features of the data that can be gleaned. First, we can see that the samples partition nicely into two groups, healthy on the left and BV on the right. This separation is driven by several taxa known to be associated with health (*Lactobacillus* species) on the left and of BV furthest to the right (non-lactobacillus anaerobes). Interestingly, transcripts annotated as belonging to *Lactobacillus iners* are near the middle of PC1, indicating that transcripts associated with this organism contribute little to the health-BV separation. Within BV, we observe that *Megasphaera* and *Prevotella* species form two 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 on PC1, 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.

We next observe that there are potential groups of healthy and of BV samples that separate along PC2. The first healthy group, H1, is composed of samples 30S, 4S, 1B, 9B, and 10B. The second healthy group, H2, is composed of samples 2B, 4B, 0B, 1A and 6B. The first BV group, BV1, is composed of samples 16B, 27S, 14B, 13A, 8A, 17B, 18B, 12B, and the second, BV2, is composed of samples 9A, 6A, 10A and 12A. The reason for these groupings is obvious when we examine the taxonomic origin of

each transcript. Here we must recall that there is *no information* regarding absolute abundance for a transcript and that two transcripts with widely varying absolute abundances but similar ratio abundances will have similar locations on the biplot. Inspection of the location of the transcripts in the biplot indicates that they are largely separable by taxonomy. Transcripts mapping to *Lactobacillus crispatus* furthest to the left and near the bottom correspond to the H1 group. Transcripts derived to the remaining lactobacilli are more closely associated with the H2 group. Such a split has been observed before, and group H1 corresponds to vaginal community state type 1 [**?**]. The samples in H2 are much less homogenous than those in H1, and correspond to a mixture of community state types two through 4.

The BV1 group is associated with an increased ratio abundance of *Leptotrichia* sp., one *Megasphaera* and one *Prevotella* species and the unmapped assembled transcripts. The BV2 group is more associated with *L. iners*, *Gardnerella vaginalis* and the other strain or species of *Prevotella* sp. and *Megasphaera* sp.

We note that the major lactobacillus groups separate much more strongly on PC2 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 distinct ways of being healthy, or it could be an artefact of the non-overlapping gene content of these organisms. The partitioning of different lactobacillus dominated healthy microbiota types is well documented in the literature [**?**, **?**, **?**] and we did not examine this further. However, it is rarely observed that the microbiota of the BV condition partitions in a meaningful way, and we examine the causes of this in a later section.

Figure 1. Exploratory analysis of the reference sequence transcripts. 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 [11]. 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. The location of each transcript is shown as a coloured point, with the colors corresponding to the taxonomic assignment of the centroid sequence. The second panel shows a blow-up of the area of the biplot on the right side. Taxa are labeled with the color corresponding to the points and the following legend: Lc-*L. crispatus*, Li-*L. iners*, Lje-*L. jensenii*, Ljg-*L. johnsonii* or *L. gasseri*, Gv-*G. vaginalis*, Me-*Megasphaera* sp., Pr-*Prevotella* sp., AS- assembled transcripts.

Not surprisingly, the reference sequence-based biplot shows that taxonomic abundance is the major driver of the variation 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 or by KEGG function and Fig. **??** 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 8.5% on PC2 when aggregated by SEED subsystem and 57.3% and 8.6% when aggregated by KEGG function. Thus the first two components explain at least 64% of the variance regardless of the functional classification used. We observe that in both cases the samples group into the same four groups as observed for the non-aggregated data. This indicates that the observed splits are robust to aggregation of the data, and are not simply driven by changes in taxonomic abundance. 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. The large number of functions with small variance indicates 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.

We note that that the majority of functions are at or near the origin with the greatest density being near -5 on PC1. A limitation of the a centered log-ratio approach, or indeed of any ratio-based approach is that the geometric mean of a sample depends on the density of 0 values in the sample. Samples will have a high density of 0 values in the group that has fewer expressed genes. The H1 and H2 group samples, composed of *Lactobacillus* species will have a much smaller set of functions than will the BV1 and BV2 groups that comprise a mixed bag of organisms, and which generally include at least some functions found only in *Lactobacillus*. Thus the geometric mean for the H groups will be closer to 0 than will the geometric mean for the BV group samples, and the resulting ratio values will be higher.

0.2 Exploring differential abundance

Traditional statistical approaches, while widely used, are inappropriate when examining compositional abundances using the ϕ metric that assesses the stability of the abundance ratios across samples for all possible pairs of transcripts [5]. Samples that have stable ratios are said to be compositionally associated and have a value of *phi* near 0.

As with any compositional data analysis approach, values of 0 are not compatible with the transformation required. 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 assume and then clr-transformed the data using the aldex.clr function from the ALDEx2 Bioconductor package [3, 4]. The ϕ measure was then calculated for each independent element of the distribution and the expected value of ϕ was reported. This approach allowed us to determine reasonable estimates for compositional association values even for functions with a dichotomous distribution, while still maintaining a low false positive background.

Figure 2 shows a graphical depiction of ϕ groupings determined for the SEED subsystem 4 functional grouping and the KEGG functional grouping. In both approaches, there were four major clusters that partitioned on PC1, and Supplementary Tables 3 and 4 contain the clusters of SEED subsystem 4 and KO annotations. 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**			
*cell*1*row*1	cell2 row 1	cell3 row 1	cell4 row 1	cell5 row 1	cell6 row 1	cell7 row 1	cell8 row 1
*cell*1*row*2	cell2 row 2	cell3 row 2	cell4 row 2	cell5 row 2	cell6 row 2	cell7 row 2	cell8 row 2
*cell*1*row*3	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.

Acknowledgments

The VOGUE Research Group is Deborah Money, Alan Bocking, Sean Hemmingsen, Janet Hill, Gregor Reid, Tim Dumonceaux, Gregory Gloor, Matthew Links, Kieran O'Doherty, Patrick Tang, Julianne Van Schalkwyk and Mark Yudin.

Funding Statement Financial support for this study was provided by a joint Canadian Institutes of Health Research (CIHR) Emerging Team Grant and a Genome British Columbia (GBC) grant awarded to DM, SMH, GR and JEH (grant reference #108030). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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. 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.

10. Gloor GB, Macklaim JM, Vu M, Fernandes AD. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. Austrian Journal of Statistics. 2015;under review.

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

12. Macklaim MJ, Fernandes DA, Di Bella MJ, Hammond JA, Reid G, Gloor GB. Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. Microbiome. 2013;1:15.

13. 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.