Multi-omic vaginal microbiome profiling reveals bacterial vaginosis subtypes

Jean M. Macklaim[1,2], Amy McMillan[1,3], Jo-Anne Hammond[4], Michael John[5], Mark Sumarah[6], Jonathan Swann[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
7 Division of Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, UK.
8 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

One major challenge of microbial investigation is to determine the role of the microbiome in the environment . We can divide microbiome analyses into categories of increasing resolution that correspond to the question answered: 'who is there' can be addressed by 16S rRNA gene sequencing (or other target gene sequencing method); 'what is the capacity of the microbiome' can be addressed by full metageomic sequencing or by an extrapolation method; 'what have they produced' can be addressed by small molecule profiling (metabolomics); and 'what are they doing now'

can be addressed by transcriptome profiling. The vast majority of microbiome analyses are designed to address the first question, and more recently methods have been developed to estimate the second question from target gene sequencing profiles. It is widely recognized that these two approaches are largely descriptive. However, the ability to address the latter two questions is less widespread because of their high complexity in both sample collection and in the analysis.

Amy - an introductory paragraph on metabolome

There are several barriers to collecting and analyzing whole microbial community transcriptomes (meta-transcriptomes). First, RNA is a much less stable molecule than DNA, with a significant fraction of microbial genes having a half-life under two minutes, and almost all having a half-life less than 10 minutes [1, 2]. Thus, careful sample collection and storage is extremely important to preserve the integrity of the sample. Second, unlike DNA abundance, RNA abundances are not necessarily tied to organism abundance because RNA production is induced or repressed in response to environmental cues. Thus, any method that discriminates based on the abundance of an RNA molecule will be at risk of confounding the effect of organism abundance with the abundance of the RNA species. The wide dynamic range of RNA abundances also makes assembly of the sequencing output problematic since many assembly algorithms assume a somewhat constant read coverage. Third, many RNA-seq experiments have a very high background of reads that do not map or assemble into sensible contigs: the origin of these phantom reads is unknown, but they greatly reduce the amount of information available on a meta-transcriptome at all sequencing depths. These constraints mean that many meta-transcriptome experiments yield less information than expected, and in some cases.

A final issue is that high-throughput sequencing data are compositional: that is, they are constrained to an arbitrary constant sum [3–5]. Compositional data have the property of linked correlations. This means that as one feature (gene or organism) becomes more abundant, one or more others *must* become relatively less abundant. This can be seen easily in the trivial example of a two part component $x = [1, 9]$ that originally has 10 parts and the two parts are partitioned such that x_1=10% and x_2=90%. . If, for example, the first part increases, $x_1 \rightarrow 9$, and the second part remains unchanged then the two parts are now partitioned such that x_1=50% and x_2=50%. Note that it appears that x_2 has become less abundant, when in fact it has remained unchanged. With a large number of such samples, we would conclude that the two abundances are negatively correlated, when in fact they are not correlated.

The problem of compositional data analysis is general no matter the number of features (genes or organisms). The linked correlation problem is the root cause of spurious correlation: where changing the membership of a composition by adding or removing one or more features, changes the correlation structure in unpredictable ways [6]. In many cases, the sign of correlation between features can change from strongly positive to strongly negative [3, 5]. This means that the observed correlations depend on group membership and not on any intrinsic correlation. There is not a current consensus on how to deal with this problem, and groups have attempted to minimize the problem [3, 7] or to approach the problem analytically [5]. Spurious correlation is very problematic because many multivariate tools such as ordination, principle component analysis and clustering are dependent upon it.

The spurious correlation problem was recognized very early by Pearson [8], yet a general solution was not elaborated until Aitchison suggested a log-ratio approach [6]. The approach advocated by Aitchison was to convert the data into ratios between the features and then to take the logarithm to make the data symmetrical such that the choice of numerator or denominator would change only the sign, and not the value of the result. Since then, much work has been done to place this approach onto a firm

theoretical and practical foundation (e.g. [9–12]), although these methods have only recently begun to be noticed and used by biologists [3–5, 7, 13, 14].

A final complication is that the measurement error of compositional data is different than that expected for non-compositional data being non-linearly related to the read depth [4]. The measurement error issue has led to a number of distinct approaches that aim to estimate the underlying distribution of these data. Zero-inflated negative binomial methods are currently popular since they appear to be able to provide good point estimates for the distributions of these highly sparse datasets. However, estimates derived from these approaches are highly parameterized and as such may be overfitting the data.

Rather than using point estimates, these datasets can be analyzed in a Bayesian manner, where the observations (count per feature) serve as the prior for the estimation of the posterior distribution of probabilities [4, 15, 16]. Using this approach a prior of zero can be converted to a posterior distribution of non-zero, probabilities that can be small or large depending on the sequencing depth and the number of features in the sample. We have shown that the posterior closely models the actual measurement error of compositional data and that that approach [4, 16]. One added benefit of this approach is that the need for sequencing-depth normalization is removed when the Bayesian estimates are used in concert with the log-ratio transformation [4, 5]. This is because the distribution of the posterior becomes broader when sequencing depth is low, reflecting the increased precision in this case. Thus, it is observed that low sequencing depth simply reduces power, as expected when less data is collected.

It has been argued that that microbial communities are best though of as ensembles of genes since the genetic makeup of individual bacterial genomes is rather fluid [17].

Materials and Methods

Ethics.

Collection.

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 [18] 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 [19]. However, we have found this approach fails in meta-RNA sequencing because of the interplay between the organism and transcript abundance [4, 5, 15]. 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 [4]. Such a model falls into the count compositional data analysis paradigm [6] where only the relative differences between abundances in a sample are informative [9, 20]. 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 [9, 20]. 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 [9, 20].

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:

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

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 [21] 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 [4] 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 [4, 16]. 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 [4]. 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 [22] 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 [9, 20]. 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 [13].

Metabolome determination.

Sample preparation GC-MS. Samples were collected from the mid-vaginal wall using the Cytobrush vaginal brush and sterile forceps and stored at -80°C until analysis. Vaginal brushes were pre-cut into 1.5 mL tubes and weighed prior to and after sample collection to determine the mass of vaginal fluid collected. After thawing, brushes were eluted in methanol-water (1:1) to a final concentration of 0.05 g/mL. This corresponded to a volume of 200-1500 μL, depending on the mass of vaginal fluid collected. Four control swabs were included which consisted of blank swabs eluted in 200, 400, 800, or 1500 μL of methanol-water. Samples and controls were vortexed for 10 sec to extract metabolites, centrifuged for 5 min at 10 000 rpm, vortexed again for 10 sec after which time the brushes were removed from tubes. Samples were centrifuged a final time to pellet cells and 150 μL of supernatant transferred to GC-MS vials. Remaining supernatant was transferred to a new 1.5 mL tube, frozen at -80°C and shipped to the University of Reading for NMR analysis. Next, 25 μL of 0.2 mg/ml ribitol standard was added to each GC-MS vial. Samples were then dried to completeness using a SpeedVac. After drying 100 μL of 2% methoxyamine-HCl in pyridine (MOX) was added to each sample for derivatization and samples were incubated at 500C for 90 min. 100 μL N-Methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) was then added to each vial and incubated at 500C for 30 min. After derivitization, an equal aliquot of each sample was combined to make the quality control (QC). Samples were then transferred to micro inserts before analysis by GC-MS (Agilent 7890A GC, 5975 inert MSD with triple axis detector). One μL of sample was injected using pulsed splitless mode into a 30 m DB5-MS column with 10 m duraguard, diameter 0.35mm, thickness 0.25 μm (JNW Scientific). Helium was used as the carrier gas at a constant flow rate of 1 mL/min. Oven temperature was held at 70°C for 5 min then increased at a rate of 5°C/min to 300°C and held for 10 min. Solvent delay was set to 13 min to avoid solvent and a large lactate peak, and total run time was 61 min. Masses between 25 m/z and 600 m/z were selected by the detector. All samples were run in random order and a the QC was run multiple times throughout the run to ensure machine consistency.

Data analysis GC-MS. Chromatogram files were de-convoluted and converted to ELU format using the AMDIS Mass Spectrometry software [23] with the sensitivity set to medium. Chromatograms were then aligned and integrated using Spectconnect [24] software with the Support Threshold set to low. All metabolites found in the blank swab, or believed to have originated from derivatization reagents were removed from analysis at this time. After removal of swab metabolites, the IS matrix from Spectconnect was transformed using the additive log ratio transformation (alr) and ribitol as a normalizing agent, $log2(x)/log2(\text{ribitol})$. Zeros were replaced with two thirds the minimum detected value on a per metabolite basis prior to transformation. All further metabolite analysis was performed using these alr transformed values.

A total of 90 metabolites were detected by GC-MS. Upon inspection it was determine that 50 of these metabolites were either redundant, background noise or present in controls and were removed from analysis. For redundant peaks, the sum of

each peak was combined resulting in a single value for each metabolite. Metabolites were initially identified by comparison to the NIST 11 standard reference database (http://www.nist.gov/srd/nist1a.cfm). Identities of metabolites of interest were then confirmed by authentic standards if available.

Independent Wilcox tests with a Benjamini-Hochberg correction to account for multiple testing were used to determine metabolites that differed significantly between healthy and BV ($p < 0.10.$). Groups were defined as healthy or BV using a percentage *Lactobacillus* cutoff of 75%

Sample preparation NMR. Samples were dried to completeness and resuspended in phosphate buffer saline (pH 7) containing sodium azide and the internal standard Trimethylsilyl propanoic acid (TSP). Samples were run on a Bruker 700 MHz cryoprobe spectrometer. A standard one-dimensional NMR spectrum was acquired for each sample with water peak suppression using a standard pulse sequence (recycle delay (RD)-90 -t1-90 -tm-90 -acquire free induction decay(FID)). The RD was set at 2 s and the mixing time (tm)100 ms. The pulse length was approximately 12 μs and t1 was set to 3 μs. For each sample, 8 dummy scans were followed by 64 scans and collected in 64 K data points using a spectral width of 20 ppm and an acquisition time per scan of 2.73 s.

Data analysis NMR. Spectra were processed according to the methods of Swann et al 2011 [25] with the following modifications. 1H NMR spectra were manually corrected for phase and baseline distortions and then referenced to the TSP resonanceat δ0.0. Spectra were digitized using an in-house MATLAB(version R2009b, The Mathworks, Inc.; Natwick, MA) script. To prevent baseline effects that arise from imperfect water saturation the region containing the water resonance was excised. An in-house peak alignment algorithm was then performed on each spectrum in MATLAB to adjust for shifts in peak position due to small pH differences between samples and then each spectrum was normalized using a sum normalization approach. Principal components analysis (PCA) using pareto scaling was applied in SIMCA (Umetrics, Umea). Orthogonal projection to latent structure discriminant analysis (OPLS-DA) models were constructed using unit variance scaling to aid the interpretation of the model and distinguish the metabolites that differed between the groups. Here, 1H NMR spectroscopic data were used as the descriptor matrix and class information (N or BV) as the response variable. The contribution of each variable (metabolite) to sample classification was visualized by back-scaling transformation, generating a correlation coefficient plot. These coefficient plots are colored according to the significance of correlation to "class" (e.g. N or BV), with red indicating high significance and blue indicating low significance. The direction and degree of the signals relate to covariation of the metabolites with the classes in the model. For all models, one orthogonal component was used to remove systematic variation unrelated to class. Predictive performance was assessed using the Q2^Y parameter.

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 [15]. Such data can be examined in a rigorous

manner by examining the variation in ratios between all pairs of transcripts [6, 9, 20]. The first step 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].

0.1 Exploratory analysis

We sequenced 27 vaginal mRNA samples enriched for microbial mRNA on the Illumina HiSeq platform at The Centre for Applied Genomics in Toronto, and for this analysis we also included in the RNA-seq results of four samples sequenced using the ABI SoLiD platform and reported previously [13]. One purpose of this work was to develop a robust approach for the exploration of RNA-seq datasets using compositional data, or CoDa, approach. We and others have shown that clr-transforming the data constitutes the first step towards a CoDa analysis approach that is generally useful to characterize HTS [3–5, 15]. All data analysis is thus done in a compositional analysis framework with clr-transformed data as outlined in the Materials and Methods.

The workhorse tool for CoDa is the compositional biplot, which summarizes in one plot the relationships between the samples and the contributions of the variables to those relationships [22]. A 'form' compositional biplot of the dataset at the level of individual refuses (corresponding to genes grouped by X percent identity, see Materials and Methods) is shown in Fig. 1refseq. This representation best preserves the relationships between the samples, and places the variables in relation to their contribution to sample location. Although a better representation of the variation in the refseqes is obtained with a 'variance' biplot, for the purposes of illustration, 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. Refseq location 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 for both the samples and the refseqs 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 40.6% of the variance on PC1 and 12.1% of the variance on PC2. We conclude that this representation is a good one since these two principle components explain over 52% of the variance in such a large dataset.

There are several observations. First, we can see that the samples partition along PC1 into several groups, with healthy on the left and BV on the right, and that several of the samples partition strongly on PC2. The major taxonomic groups to which the refseqs map is shown by color at the level of species for *Lactobacillus* and at the level of genus for the remainder. It is obvious that the distinction between samples on PC1 is driven by differential occurrence of *Lactobacillus* species on the left and non-lactobacillus anaerobes on the right. 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 [26–28] and we did not examine this further.

Interestingly, transcripts annotated as belonging to the ubiquitous*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.

Not surprisingly, the reference sequence-based biplot shows that taxonomic abundance is the major driver of the variation of transcript abundance between samples. Therefore, examining the difference between groups at the level of individual refseqs would provide little more information than could be obtained by knowing the genome of the organisms occurring in each sample. Thus, we were interested to determine if the different states had different underlying functions regardless of their taxonomic composition.

Refseqs were grouped by SEED subsystem 4 function [**?**] and Fig. 1SEED shows the result. Here we can see that the SEED functions partition more strongly along the major axis of variance with 52.2% of the variance explained on PC1, and 8.5% on PC2 when aggregated by SEED subsystem 4. Thus the first two components explain at least 60% of the variance. Interestingly, we observe that both the healthy and BV groups appear to partition into two subgroups

We examined the reason for this apparent partitioning using the bar plots and SEED subways4 functional expression profiles in Figure **??**. Here, we can see that many samples group strongly by similar functional expression profiles. The first group is the H1 group and is composed or largely *L. crispatus* by both the 16S rRNA gene sequencing and the mRNA fraction mapping. This group thus corresponds to the community state type 1 suggested by Ravel and coworkers [26]. The second group, H2, is composed largely of a mixed set of *Lactobacillus* species, often dominated by *L. iners* in total abundance, but with a substantial gene expression contribution from *L. jensenii*, *L. gasseri* or unknown *Lactobacillus* sp. This group could not be neatly put into a community state type. Interestingly we observed that the BV group partitioned into two groups. The BV1 group contained a substantial amount of the unclassified BVAB1 organism by 16S rRNA gene sequencing and had a large gene expression contribution from *Sneathia* sp, and also contained the largest contribution of expression from de-novo assembled contigs. These contigs were assigned to the BVAB1 group color in the figure. The final group, BV2, had only a small amount of BVAB1, and a generally larger amount of *Atopobium* sp. by 16S rRNA gene profiling, and a very small or absent contribution to gene expression by *Sneathia* and BVAB1.

We found that several samples did not fit neatly into these groups, and had long branches connecting them to all other samples. These samples were found to have very atypical community profiles. For example, samples 013B was composed largely by *Atopobium* sp. by 16S rRNA gene profiling, and by *Megasphaera* by gene expression contribution. Sample 015B was composed of *Bifidobacterium* sp. and *Streptococcus* by 16S rRNA gene profiling, and had a significant contribution from the latter to its gene expression profile. Thus, these two samples exhibited very atypical profiles. We also excluded four samples from analysis because they had long branches with their nearest neighbour or nearest clearly defined group in the clustered heat map, or because they contained substantial expression of both lactobacillus genes and of non-lactobacillus genes, or because they contained a substantial fraction of non-lactobacillus organisms by 16S rRNA gene sequencing yet did not cluster with either BV group. These samples include sample 31S which was embedded within the BV group with a long branch, and samples 001A, 003A, 008B. Finally, sample 019A was excluded even though it contained almost exclusively *L. iners* by both 16S rRNA gene profiling and by gene expression composition, yet it did not fit into either of the H groups. Interestingly, this sample was classed as clinically BV by Nugent scoring. We noted that the majority of these samples were found to be located very near the centre of

PC1 on either the refseq or SEED biplots (31S, 003A, 015B, 013B) the suggesting that they were placed inaccurately because of differences in gene expression, or were very far from all other samples (001A, 008B). All further analysis was done using the samples in the H1, H2, BV1 and BV2 groups.

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 [22]. 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.

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.

0.2 Exploring differential abundance

. 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.

The distinctions between the different *Megasphaera* and *Prevotella* types can be seen more clearly in Supplementary Figure 1.

or by KEGG function [**?**]

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 [26]. 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.

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 [4, 15]. 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.

Figure 3. Correlation between metabolite abundance and microbial function. Correlations between metabolite abundance and position of each sample on a compositional biplot built from SEED subsystem 4 aggregated data were calculated using the kendall method. These correlations were then scaled according to the first and second component of this biplot and plotted according to their scaled position. A. Metabolites detected by GC-MS. B. Metabolites detected by NMR. Only metabolites which differed significantly between BV and health are shown. GHB: gamma-hydroxybutyrate, 2HG: 2-hydroxyglutarate, 5AV: 5-aminovalerate, 2HIC: 2-hydroxyisocaproate, Tyr: tyramine, Cad: cadaverine, Pip: pipecolate, Gluc: glucose, Malt: maltose, Fruc: fructose, TMA: trimethylamine, Lact: lactate, Acet: acetate, Form: formate, 2HV: 2-hydroxyisovalerate, Chol: choline, Succ: succinate.

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. We thank Jennifer Reid, Shinthujah Arulanantham and Yohanna Emun for assistance with data compilation.

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. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. Proceedings of the National Academy of Sciences. 2002;99(15):9697–9702.

2. Andersson AF, Lundgren M, Eriksson S, Rosenlund M, Bernander R, Nilsson P. Global analysis of mRNA stability in the archaeon Sulfolobus. Genome biology. 2006;7(10):R99.

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

4. 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.

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. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015 May;11(5):e1004226.

8. Pearson K. Mathematical contributions to the theory of evolution. – on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London. 1897;60:489–498.

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

10. Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ, Pawlowsky-Glahn V, Buccianti A. The principle of working on coordinates. Compositional data analysis: Theory and applications. 2011;p. 31–42.

11. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcel O-Vidal C. Isometric logratio transformations for compositional data analysis. Math Geol. 2003;35(3):279–300.

12. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology. 2003;35(3):253–278.

13. 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.

14. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663.

15. 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.

16. 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.

17. Boon E, Meehan CJ, Whidden C, Wong DHJ, Langille MG, Beiko RG. Interactions in the microbiome: communities of organisms and communities of genes. FEMS microbiology reviews. 2014;38(1):90–118.

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

19. 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.

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

21. 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.

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

23. Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. Journal of the American Society for Mass Spectrometry. 1999;10(8):770–781.

24. Styczynski MP, Moxley JF, Tong LV, Walther JL, Jensen KL, Stephanopoulos GN. Systematic identification of conserved metabolites in GC/MS data for metabolomics and biomarker discovery. Analytical Chemistry. 2007;79(3):966–973.

25. Swann JR, Tuohy KM, Lindfors P, Brown DT, Gibson GR, Wilson ID, et al. Variation in antibiotic-induced microbial recolonization impacts on the host metabolic phenotypes of rats. J Proteome Res. 2011 Aug;10(8):3590–603.

26. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A. 2010;doi/10.1073/pnas.100611107.

27. Hummelen R, Fernandes AD, Macklaim JM, Dickson RJ, Changalucha J, Gloor GB, et al. Deep sequencing of the vaginal microbiota of women with HIV. PLoS One. 2010;5(8):e12078.

28. McMillan A, Rulisa S, Sumarah M, Macklaim JM, Renaud J, Bisanz JE, et al. A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. Scientific Reports. 2015 09;5:14174 EP –. Available from: http://dx.doi.org/10.1038/srep14174.

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.