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 annotation.

Reads in fastq format were mapped to the library of centroid sequences using bowtie2 (REF) 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 (REF) that only requires normalization to a constant sequencing effort prior to analysis. However, we and others have found this approach often fails in meta-RNA sequencing because of the interplay between the organism and transcript abundance. Thus, we find 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 (REF). Such a model falls into the count compositional data analysis paradigm (REF) where only the relative differences between abundances in a sample are informative (REF).

Exploratory data analysis was performed as point estimates with compositional biplots (REF). 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 (REF) was used to adjust 0 values for the likelihood that the 0 represents a non-detect event in the sample. We have shown that for the purposes the 0 replacement method does not contribute significantly to the outcome of these plots (REF).

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 (REF). These distributions were transformed using the centre log-ratio transformation prior to analysis (REF) which allows the value to be unconstrained. The expected value of phi (REF) from the distribution was used to determine compositionally-linked transcripts or functions, since it has been shown that traditional correlation metrics give

unpredictable results. The expected value of Kendall's Tau (b) was used when reporting the correlation between transcript abundance and metabolome 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 (REFS). Such data can be examined in a rigorous manner by examining the variation in ratios between all pairs of transcripts. However, this can be dramatically simplified using the centre log-ratio transformation, or clr, which has the following formula:

FORMULA HERE

The clr has a one to one mapping ... Functionally equivalent to all vs all ratios (REF). Compensates for differences in read abundance, thus eliminates the need for count normalization (REF). and has been shown to be a generally useful method to characterize HTS (REFS). All data analysis is thus done in a compositional analysis framework with clr-transformed data.

0.1 Exploratory analysis

We collected NN samples for RNA-seq, and MM were sequenced on the Illumina HiSeq platform at TCAG. We also included in the RNA-seq results four samples sequenced using the ABI SoLid platform. Examination of these samples using taxonomic abundance determined both from 16S rRNA gene sequencing and from the reference sequence abundance, with compositional biplots and unsupervised clustering on center-log-ratio transformed data, and correlating the samples with the clinical phenotype, six samples were excluded (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.

These 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 $\sim$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.

Inspection of the location of the transcripts indicates that they are largely separable, with *Lactobacillus crispatus* transcripts furthest to the left and several taxa associated with BV furthest to the right. Their locations are proportional to their standard deviation in the entire dataset in these two dimensions 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 distint 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.

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.

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. **??** 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. Samples composing the healthy group on the left side of the biplot are associated with a relatively small set of functions that are strongly increased in these samples, and the samples composing the BV group are associated with a large set of functions that are strongly increased in the Bv samples. 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

Compositional data contains information only on the ratios between the components (REFS). (error profile, correlation and covariance). We next examined the differential abundance of functions using the ϕ-metric that measures the strength of association between

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.

LOREM and IPSUM Nunc blandit a tortor.

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. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit. Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat.

Sed ac quam id nisi malesuada congue.

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Subsection 1

Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Subsection 2

3rd Level Heading. Nulla mi mi, venenatis sed ipsum varius, volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

Discussion

Nulla mi mi, venenatis sed ipsum varius, Table 1 volutpat euismod diam. Proin rutrum vel massa non gravida. Quisque tempor sem et dignissim rutrum. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi at justo vitae nulla elementum commodo eu id massa. In vitae diam ac augue semper tincidunt eu ut eros. Fusce fringilla erat porttitor lectus cursus, vel sagittis arcu lobortis. Aliquam in enim semper, aliquam massa id, cursus neque. Praesent faucibus semper libero.

LOREM and IPSUM Nunc blandit a tortor.

CO_2 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. Quisque augue sem, tincidunt sit amet feugiat eget, ullamcorper sed velit.

Sed non aliquet felis. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris commodo justo ac dui pretium imperdiet. Sed suscipit iaculis mi at feugiat. Ut neque ipsum, luctus id lacus ut, laoreet scelerisque urna. 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. Nam condimentum sem eget mollis euismod. Nullam dui urna, gravida venenatis dui et, tincidunt sodales ex. Nunc est dui, sodales sed mauris nec, auctor sagittis leo. Aliquam tincidunt, ex in facilisis elementum, libero lectus luctus est, non vulputate nisl augue at dolor. For more information, see S1 Text.

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. 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.