diffTop: Differential Topic Analysis Package Demo

BioC 2021

Pratheepa Jeganathan
McMaster University
(Joint work with Susan Holmes at Stanford University)

08/04/2021

Outline

- Data
- Differential abundance analysis
- Latent Dirichlet allocation - demo part I
- diffTop - demo part II

Molecular microbial data

- Container - A Bioconductor package phyloseq[1].

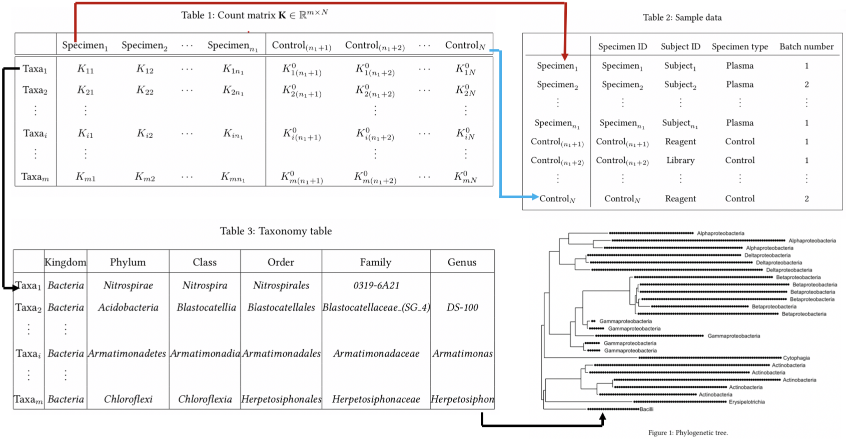


Table 1: Count matrix $\mathbf{K} \in \mathbb{R}^{m \times N}$

	Specimen$_1$	Specimen$_2$	$\cdots$	Specimen$_{n_1}$	Control$_{(n_1+1)}$	Control$_{(n_1+2)}$	$\cdots$	Control$_N$
Taxa$_1$	K_{11}	K_{12}	$\cdots$	K_{1n_1}	$K^0_{1(n_1+1)}$	$K^0_{1(n_1+2)}$	$\cdots$	K^0_{1N}
Taxa$_2$	K_{21}	K_{22}	$\cdots$	K_{2n_1}	$K^0_{2(n_1+1)}$	$K^0_{2(n_1+2)}$	$\cdots$	K^0_{2N}
$\vdots$	$\vdots$					$\vdots$		$\vdots$
Taxa$_i$	K_{i1}	K_{i2}	$\cdots$	K_{in_1}	$K^0_{i(n_1+1)}$	$K^0_{i(n_1+2)}$	$\cdots$	K^0_{iN}
$\vdots$	$\vdots$					$\vdots$		$\vdots$
Taxa$_m$	K_{m1}	K_{m2}	$\cdots$	K_{mn_1}	$K^0_{m(n_1+1)}$	$K^0_{m(n_1+2)}$	$\cdots$	K^0_{mN}

Table 2: Sample data

	Specimen ID	Subject ID	Specimen type	Batch number
Specimen$_1$	Specimen$_1$	Subject$_1$	Plasma	1
Specimen$_2$	Specimen$_2$	Subject$_2$	Plasma	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Specimen$_{n_1}$	Specimen$_{n_1}$	Subject$_{n_1}$	Plasma	1
Control$_{(n_1+1)}$	Control$_{(n_1+1)}$	Reagent	Control	1
Control$_{(n_1+2)}$	Control$_{(n_1+2)}$	Library	Control	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Control$_N$	Control$_N$	Reagent	Control	2

Table 3: Taxonomy table

	Kingdom	Phylum	Class	Order	Family	Genus
Taxa$_1$	*Bacteria*	*Nitrospirae*	*Nitrospira*	*Nitrospirales*	*0319-6A21*	
Taxa$_2$	*Bacteria*	*Acidobacteria*	*Blastocatellia*	*Blastocatellales*	*Blastocatellaceae_(SG_4)*	*DS-100*
$\vdots$	$\vdots$					
Taxa$_i$	*Bacteria*	*Armatimonadetes*	*Armatimonadia*	*Armatimonadales*	*Armatimonadaceae*	*Armatimonas*
$\vdots$	$\vdots$					
Taxa$_m$	*Bacteria*	*Chloroflexi*	*Chloroflexia*	*Herpetosiphonales*	*Herpetosiphonaceae*	*Herpetosiphon*

Figure 1: Phylogenetic tree.

[1] McMurdie and Holmes (2013)

Differential abundance analysis

Differential abundance analysis

- Find taxonomic differences across environments or groups or experimental conditions.

Existing methods

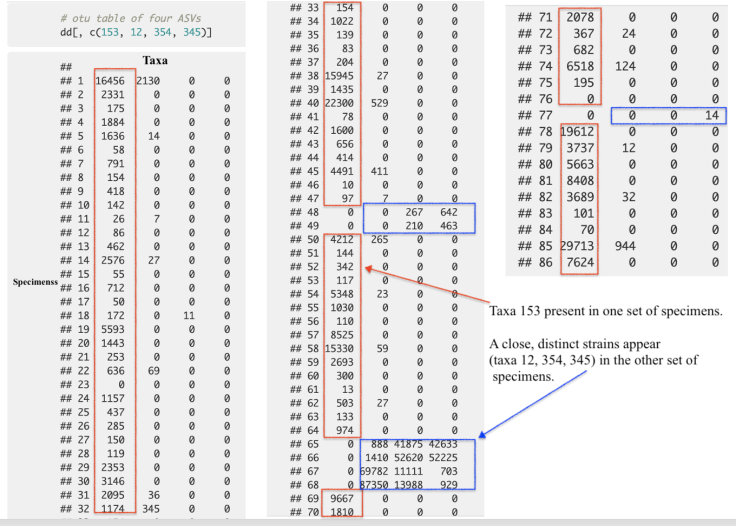
- Permutation tests using distance matrices - PERMANOVA.
- Differential abundance through generalized linear modeling and transformations.

Challenge

- High-resolution acquisition of data at the taxonomic strain level can result in different subjects or environments presenting slightly different taxa that are “functionally synonymous."

Challenge

- Strain switches



```
# otu table of four ASVs
dd[, c(153, 12, 354, 345)]
```

Taxa

Specimenss

```
##
## 1  16456  2130     0     0
## 2   2331     0     0     0
## 3    175     0     0     0
## 4   1884     0     0     0
## 5   1636    14     0     0
## 6     58     0     0     0
## 7    791     0     0     0
## 8    154     0     0     0
## 9    418     0     0     0
## 10   142     0     0     0
## 11    26     7     0     0
## 12    86     0     0     0
## 13   462     0     0     0
## 14  2576    27     0     0
## 15    55     0     0     0
## 16   712     0     0     0
## 17    50     0     0     0
## 18   172     0    11     0
## 19  5593     0     0     0
## 20  1443     0     0     0
## 21   253     0     0     0
## 22   636    69     0     0
## 23     0     0     0     0
## 24  1157     0     0     0
## 25   437     0     0     0
## 26   285     0     0     0
## 27   150     0     0     0
## 28   119     0     0     0
## 29  2353     0     0     0
## 30  3146     0     0     0
## 31  2095    36     0     0
## 32  1174   345     0     0
```

```
## 33   154     0     0     0
## 34  1022     0     0     0
## 35   139     0     0     0
## 36    83     0     0     0
## 37   204     0     0     0
## 38 15945    27     0     0
## 39  1435     0     0     0
## 40 22300   529     0     0
## 41    78     0     0     0
## 42  1600     0     0     0
## 43   656     0     0     0
## 44   414     0     0     0
## 45  4491   411     0     0
## 46    10     0     0     0
## 47    97     7     0     0
## 48     0     0   267   642
## 49     0     0   210   463
## 50  4212   265     0     0
## 51   144     0     0     0
## 52   342     0     0     0
## 53   117     0     0     0
## 54  5348    23     0     0
## 55  1030     0     0     0
## 56   110     0     0     0
## 57  8525     0     0     0
## 58 15330    59     0     0
## 59  2693     0     0     0
## 60   300     0     0     0
## 61    13     0     0     0
## 62   503    27     0     0
## 63   133     0     0     0
## 64   974     0     0     0
## 65     0   888 41875 42633
## 66     0  1410 52620 52225
## 67     0 69782 11111   703
## 68     0 87350 13988   929
## 69  9667     0     0     0
## 70  1810     0     0     0
```

```
## 71  2078     0     0     0
## 72   367    24     0     0
## 73   682     0     0     0
## 74  6518   124     0     0
## 75   195     0     0     0
## 76     0     0     0     0
## 77     0     0     0    14
## 78 19612     0     0     0
## 79  3737    12     0     0
## 80  5663     0     0     0
## 81  8408     0     0     0
## 82  3689    32     0     0
## 83   101     0     0     0
## 84    70     0     0     0
## 85 29713   944     0     0
## 86  7624     0     0     0
```

Taxa 153 present in one set of specimens.

A close, distinct strains appear (taxa 12, 354, 345) in the other set of specimens.

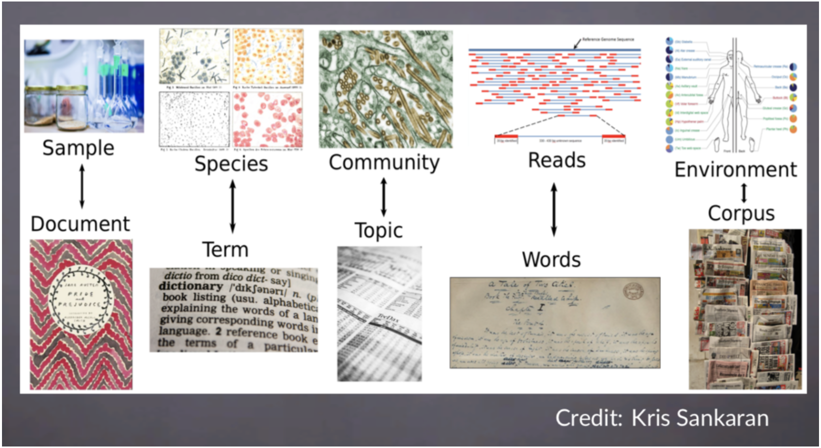
Challenge

- Strain switching decreases power to detect a difference using a Bray Curtis (for example) distance and `permanova` or generalized linear models.
  - A workflow for power analysis is available at Permutation tests using distance matrices.

This demo

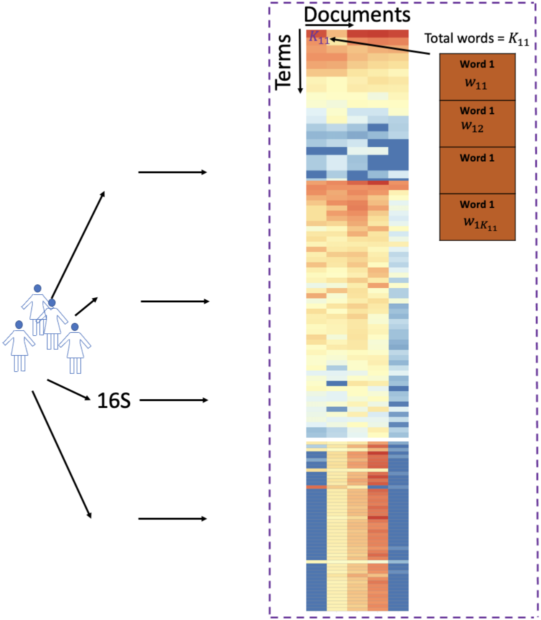
- Individual taxa may not be as important as their combination.
- Useful to consider co-occurrence of bacteria.
  - Similar to how synonyms occur in textual analyses.

Analogy between textual and microbiome data analysis[2]

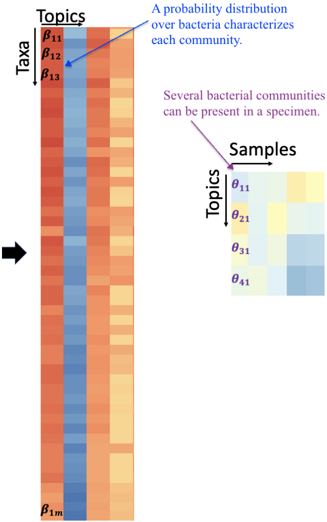


Credit: Kris Sankaran

[2] Latent variable modeling for the microbiome (Sankaran and Holmes 2019)


Documents
Terms
K_{11}
Total words = K_{11}
Word 1
w_{11}
Word 1
w_{12}
Word 1
Word 1
$w_{1K_{11}}$
16S

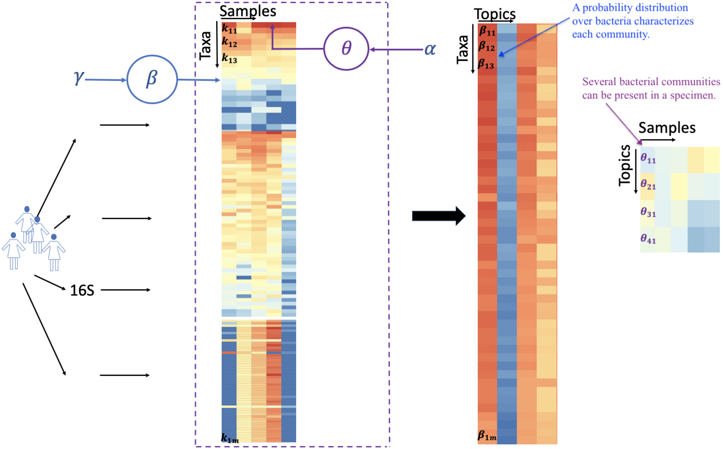
Generative model



A probability distribution over bacteria characterizes each community.

Several bacterial communities can be present in a specimen.

Latent Dirichlet allocation[3]



[3] Probabilistic topic models (Blei, Carin, and Dunson 2010)

Package demo - Part I

- Latent Dirichlet allocation in Stan

Differential topic analysis

diffTop

- Developed for differential topic analysis[4]
- Use the microbial community (topics) in each specimen to test whether topic memberships differ across conditions.

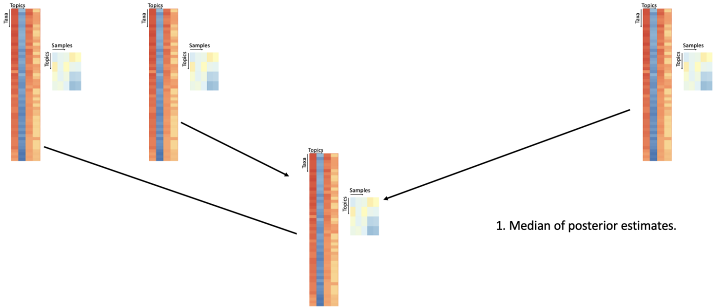
[4] *A Statistical Perspective on the Challenges in Molecular Microbial Biology* (Jeganathan and Holmes 2021)

Computing topic abundance in each specimen

Start with the posterior samples of parameters of the generative model

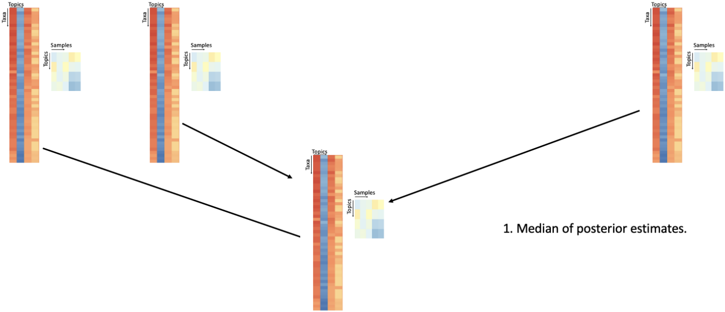
Topics
Taxa
Samples
Topics
Topics
Taxa
Samples
Topics
Topics
Taxa
Samples
Topics

Start with the posterior samples of parameters of the generative model



1. Median of posterior estimates.

Start with the posterior samples of parameters of the generative model

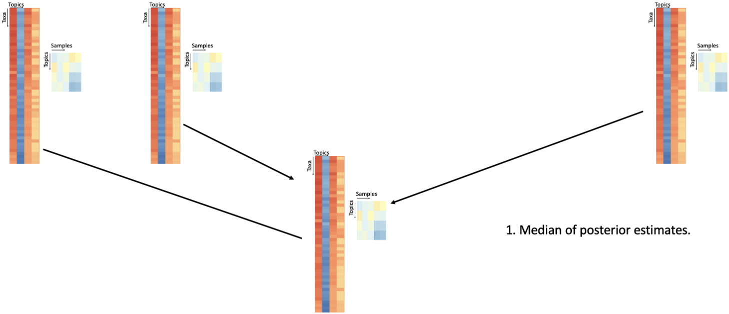


1. Median of posterior estimates.

	Sample 1	Sample 2	Sample 3	Sample 4
Topic 1	T_{11}	T_{12}	T_{13}	T_{14}
Topic 2	T_{21}	T_{22}	T_{23}	T_{24}
Topic 3	T_{31}	T_{32}	T_{33}	T_{34}
Topic 4	T_{41}	T_{42}	T_{43}	T_{45}

2. Multiply the proportions by the library size and round to an integer - gives abundance of the topic in each specimen.

Start with the posterior samples of parameters of the generative model



1. Median of posterior estimates.

	Sample 1	Sample 2	Sample 3	Sample 4
Topic 1	T_{11}	T_{12}	T_{13}	T_{14}
Topic 2	T_{21}	T_{22}	T_{23}	T_{24}
Topic 3	T_{31}	T_{32}	T_{33}	T_{34}
Topic 4	T_{41}	T_{42}	T_{43}	T_{45}

2. Multiply the proportions by the library size and round to an integer - gives abundance of the topic in each specimen.

3. Use DESeq2 to identify differentially abundant topics across groups or environmental conditions.

Package demo - Part II

- diffTop

diffTop results

- Conclusion for the exemplary data analysis
    - The authors of the dataset found little evidence of a difference between specimen β- diversity using PERMANOVA.
    - diffTop identifies some topics in the Asteraceae paired-specimens that are significantly different.

Conclusion

- Topic models are interpretable
    - Unmask rare and synonymous taxa.
    - Discover the assembly of microbial communities as topics which can be projected onto the phylogenetic tree.
        - Enhance our understanding of the differences in complex microbiome communities.

Future work

- diffTop will be submitted to Bioconductor.
  - We welcome your comments on improving the usability of diffTop and can be submitted to issue in Github.
- Website: https://pratheepaj.github.io/diffTop/
- Consider different topic models.

Thank You!

Email: jpratheepa31@gmail.com

References I

Blei, David, Lawrence Carin, and David Dunson. 2010. "Probabilistic Topic Models." *IEEE Signal Processing Magazine* 27 (6): 55–65.

Jeganathan, Pratheepa, and Susan P Holmes. 2021. "A Statistical Perspective on the Challenges in Molecular Microbial Biology." *Journal of Agricultural, Biological and Environmental Statistics* 26 (2): 131–60.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.

McMurdie, Paul J, and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PloS One* 8 (4): e61217.

References II

Morgan, Martin. 2021. *BiocManager: Access the Bioconductor Project Package Repository*. https://CRAN.R-project.org/package=BiocManager.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sankaran, Kris, and Susan P Holmes. 2019. "Latent Variable Modeling for the Microbiome." *Biostatistics* 20 (4): 599–614.

Stan Development Team. 2020. "RStan: The R Interface to Stan." http://mc-stan.org/.

References III

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Yu, Guangchuang. 2020. "Using Ggtree to Visualize Data on Tree-Like Structures." *Current Protocols in Bioinformatics* 69 (1): e96.