Molecular Microbial Data Analysis

(RLadies Dallas Meetup)

Pratheepa Jeganathan
Department of Mathematics and Statistics
McMaster University

07/12/2021

About me

- Faculty at McMaster University.
  - Assistant Professor - Department of Mathematics and Statistics.
  - Associate Member - School of Computational Sciences and Engineering.
- Research - solving challenges arise in applications.
- Teaching - statistics and data science courses.

Jeganathan research group

- Started - May 3rd, 2021.
- Uses, extends, develops statistical methods and computational tools to integrate multi-domain data.
  - Mixture models (unsupervised, supervised, and semi-supervised).
  - Bayesian modeling.
  - Topic models.
  - Methods for spatial point pattern analysis.
  - Resampling methods for dependent data.
  - Approximation theory in statistics (saddlepoint apporoximations)
- Applications: microbiome multi-omic, spatial multi-omic, non-life insurance research.

- Undergraduate summer research students.
- Ph.D. students, starting Fall 2021.
- Watch out at https://twitter.com/PratheepaJ for openings in my lab.

Roadmap

- Motivation
- Molecular microbial data
- Components of microbiome data
- Example data set
- Models
- Transformations
- Visualization
- Multivariate and network analysis
- Differential abundance analysis
- Topic model using LDA
- Differential topic analysis

Disclaimer

- I'm not an expert in biology.
- I didn't develop all the tools that I used in the analysis - Thank you to all the developers and maintainers of the software packages that I use today.

Motivation

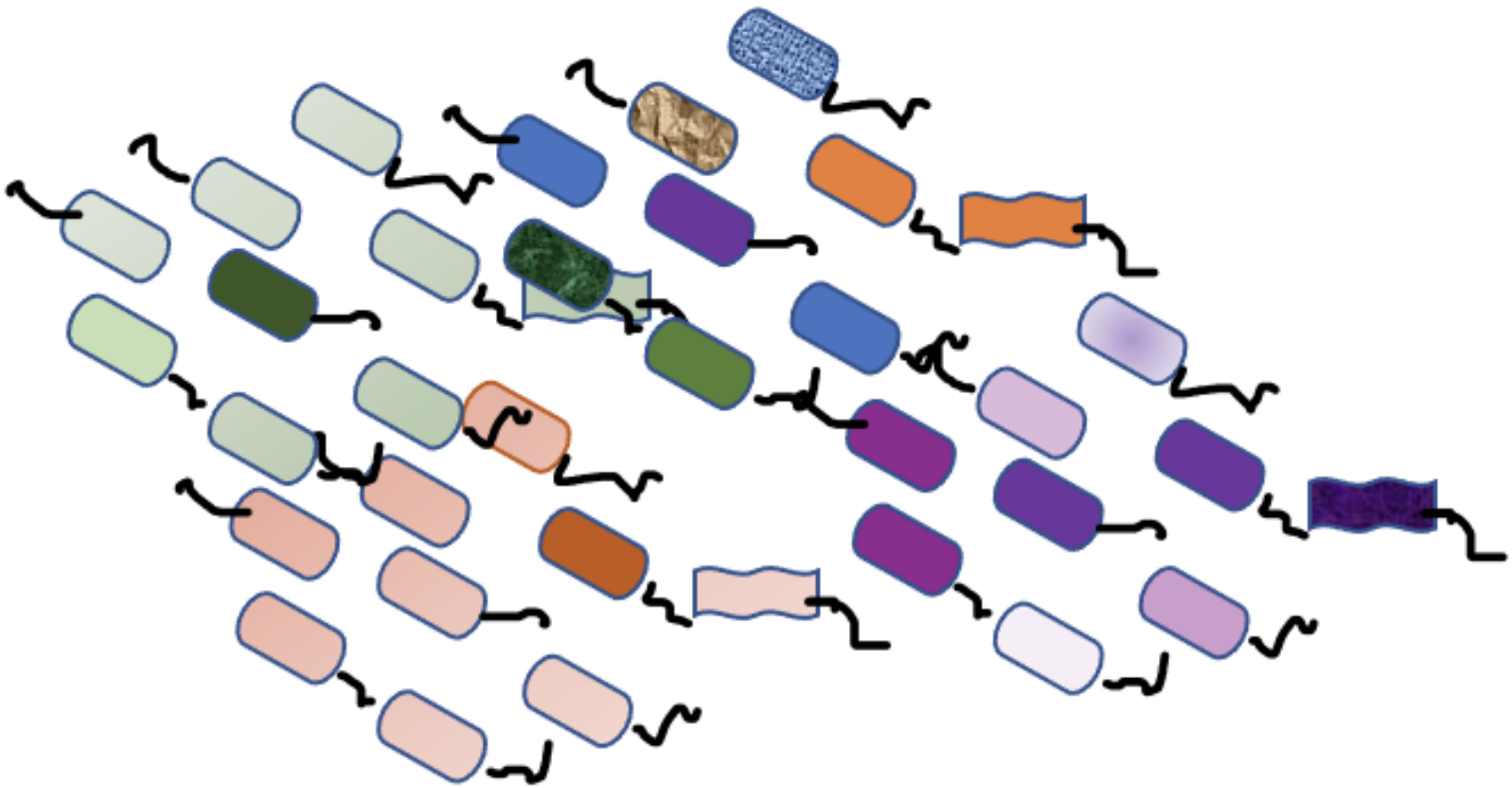
Projects in microbiome data analysis

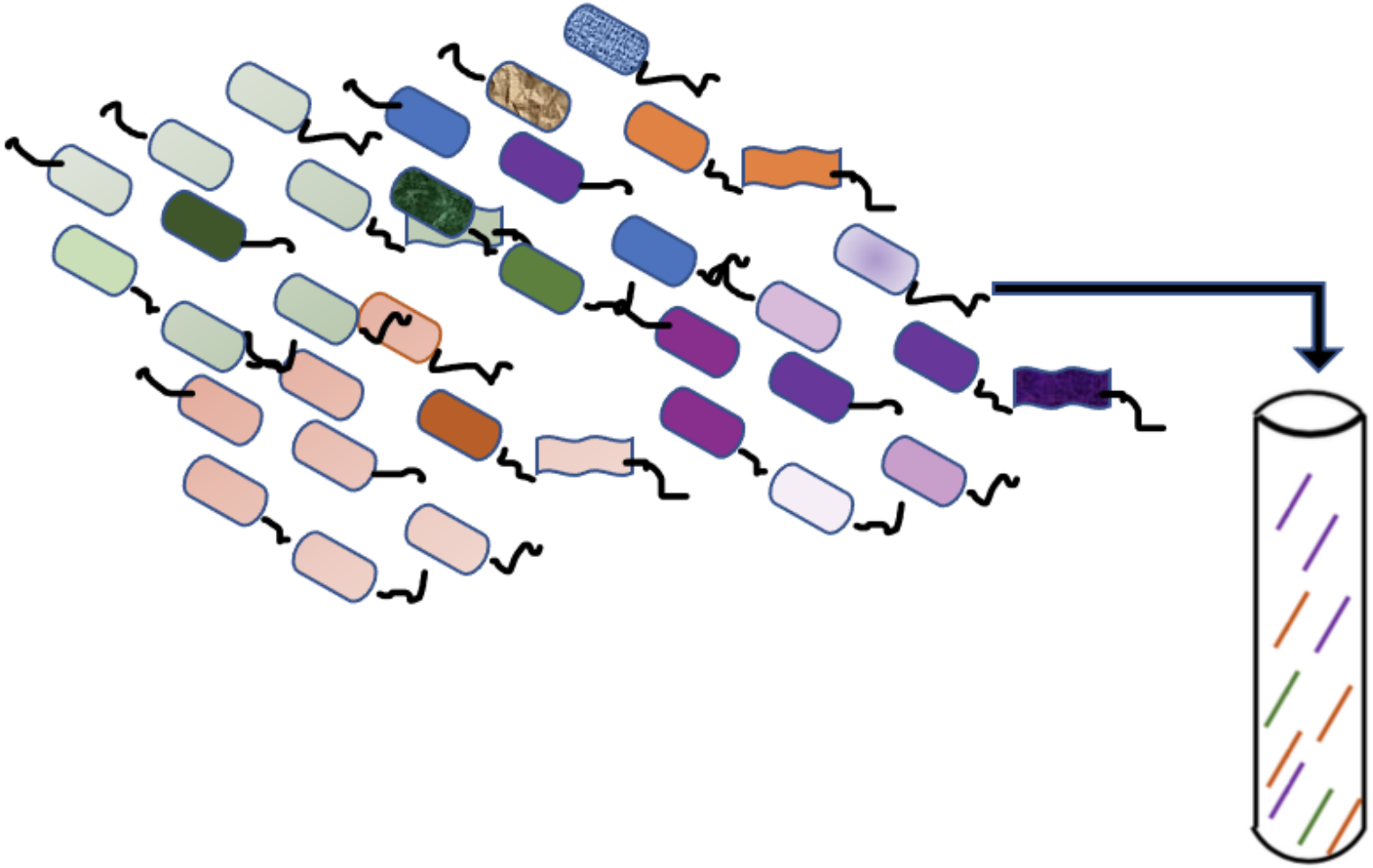
- Resilience study (gut microbiome).
- Preterm birth study (vaginal microbiome).
- Pregnant and reporductive-aged women study (vaginal microbiome multi-omic).
- Inflammatory bowel disease study in children and adult (gut microbiome).
- Marine mammal study (stool microbiome).
- Sepsis study (blood).
- Healthy flossing study (oral and blood).
- Plant microbiome study (endosphere microbiome).

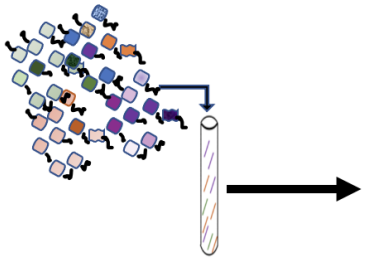
Molecular microbial data

Sample collection

- Use small regions (V4-V6) of a particular gene (the 16S rRNA gene, occasionally others) that serves as a "fingerprint" or signature for each species or strain of bacteria.







```
@M00779:331:000000000-C43T8:1:1101:11340:1096 1:N:0:359
GCCTATGGGGCGCTGCAGTGGGGAATTTTGGCCAATGGGCGAAAGCCTGATCCAGCAACGCCGCGTGGGTGATGAAGGCCTTCGGGTTGTAAAGCACTTTTGTCCGGAAAGAAATGGCACTGGTGTATCCCTGGGGTCAATGACGGTACCGGCAGACAAAAGCACGGCTAACTACGGGCCAGCAGCCGCGGTAATACGTAGGTT
CCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGCAGTCGGTTGTGAACGACCGAAGTGAAATCCCCGAGCTTAACCAGGGCGTTGCATTGGTGCACGCACAGCTAGAGTGTG
+
CCCCCFGFC7;@:@C7FCC,CCFGGG@EGGGBCE,CFCF<CFE+@FGFFG9,6CAFDD,C+@CGG:4+8BF4,,5AA,BFFFG@FC<F:FG9F9FE<E<FF<,FC=C+FCF@@,,5,3,:3>,,,,:,87,=8BD+3+,,,,,>CFFF+,5**35*4,,2,6,,>57**=F8,2?8*==*=/CFCCC;*:/
B5*1<@EF<C8:+02+<+0/;;C=*22;5:C5<EDC+<FCC71/99*775>F57D))1:)/99:*2:CGDD<)))0*****32CF3*7)8**07**..7((()))(.**)6.>*,*(]-).,8:693/./))
@M00779:331:000000000-C43T8:1:1101:21155:1272 1:N:0:359
GCCTACGGGGCGCTGCAGTGGGGAATCTTGGACAATGGGGGCAACCCTGATCCAGCAATGCCGCGTGGGGGAAGAAGGCCTTCGGGTTGTAAAGCCCTTTCGGCAGGGAAGAAATCGGTGGAGTGAATAACTCCACTGGATGACGGTACCTGCAGAAGAAGAACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAAGGT
GGGAGCGTTAATCGGATTTCCTGGGCGTAAAGCGTGCGCAGGAGGCGCGGTAAGCACGGTGTGAAATCACTGGGCTAAACCTAGGCATTGCGATTGTGACTGGACAGCGCGAGTCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGCGGDFCFGGGGGGGGGDGGCFGGGEEFGGGGEEFGGGGGGGDG7F+BCGGFGGFGGGGGGGGGGGGGGA=FGGC<FF7@7@:FGGGGDA>>+3@@5FC7,@D,=DA,8FG,@:33DE9:35@*?@>FGGC6,::C,22,<*:7CA,,BFE+
++3>ECFEEFC3:C8*8+C75C6CCECC*8@CDE/5:::95**19C+0+3+;CGCCE*1:7;<5B)993)*/*)/))7>***+2)9><>7*;**0+***27<:*)9955295(.)7@26).(06C*6@45()),(09(73<)6
@M00779:331:000000000-C43T8:1:1101:15530:1326 1:N:0:359
ACCTACGGGGGGCTGCAGTGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCGACGCCGCGTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAGCTCCGAAGAAGCGAAAGTGACGGTAGGAGCAGAAGAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTTGTCCGGAATTA
TTGGGCGTAAAGAGCTCGTAGGTGGTTCGTCGCGTCTGCTGTGAAAGTCCGAGGCTTAACCTCGGATCTGCAGTGGGTACGGGCGGACTAGCGTACGGTAAAGGAGATCGGAATTCC
+
CCCCCGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGCGGGGGGGGGGEAFGGGGGGGGGGGGGGGGGGGGFGFGGGGGGGGGGGGGGGGGGGFFGGFGFFFGB*FGGG@GGGGGGGGGGGGEEBFGGFGGGGCCCGGGGGFGGDGGGGGFCGGGGGFGGEGFF>CGGCEGG7ECFFGGGGGEGGFGDGGGDG79EEC=FFF
GEEFG4E=8C=DCFFEC*/;CC7@GDGGGGGFGGBGEFFGGFF:CCCGGFFDDCEFFGFGF/095C0*<<9*+29F4776BD>37>65B@:(57(0,)2,).97>5(4)<<65:FB2
@M00779:331:000000000-C43T8:1:1101:14716:1481 1:N:0:359
TCCTATGGGGCGCAGCAGTGGGGAATTTTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGCGGGAAGAAGGCCTTCGGGTTGTAAACCGCTTTTGTCAGGGAAGAAAAGGTTCTGGCTAATACCTGGGACTCATGACGGTACCTGAAGAATAAGCACCGGCTAACTACGTGCCAGCAGCCACGGTAATACGTCGGGT
GCAAGCGTTAAGCGGAATTACAGGGAGTAAAGCGTGCGCAGGCGGTTATGCAAGACAGAGGTGAAATCCCCGGGCTCAACCTGGGAACGGCCTTGTGACGACATGGCTAGAGTACG
+
CCCCCGFFGC;CEGCGFGCFGGGGGGFDCGGGGGDFGGGGGGGGCFCFGFGEGEGGCFAFFFFGGGGGGCGCEGGGGGGG,FF7FCFF7CFEGGBFE7F@FGGFFGGFG,CDFFGF:,7DFGGFCF,9DABB=,=D3=@FF9CFGGEG@CC:F@3=:DDGC;,DC1*44*74;E8C4:F57?=,,,
2,=CGCGC79@F9*:*=CC3++95;***;*3*2:**<EC+2:**17+<>9C387))0)9@BGDCD3**/>**:00*29*?C*9:67FCF3:70)19/75),))325)4007)9550))(0(]/..47>)*-94?
@M00779:331:000000000-C43T8:1:1101:19287:1506 1:N:0:359
GCCTACGGGAGGCTGCAGTAAGGAATATTGGTCAATGGAGGCAACTCTGAACCAGCCATGCCGCGTGCAGGAAGACAGCCCTCTGGGTCGTAAACTGCTTTTATTCGGGAATAAACCTATCTACGTGTAGATAGCTGAATGTACCGAAGGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCAAG
CGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGCTTATTAAGTCAGGGGTGAAAGACGGTGGCTCAACCATCGCAGTGCCTTTGATAATGATGAGCTAGAATATACTAGA
+
CCCCCGGGG7FGGCGGAEFFGGFGFGGGFGFEGGGGGFGGGGEFGGFGGGGGGGGGGGGGGC@C:F>CFGGGGGGGFGGGCFF,EFFGFGGGFDEGEGFGFFGGGGGGEE@F<FGGDGGGGAGGFFGGBBEFGE9FFEFFFGGFEG7F+CDFCFF9FFFGED=EEGGGGGGGGG+>DCCFFGDGGGGGECFFFGGDCG?,=C<A
FGGEGFFEGG8C*A7E7DCCDE*<>@@DEGGFG*.C>@@GG3>FFG7://204C37CF9:@FFDG:FG4)7+6@4>=6)00,8924*9*-)**]/.:C:=)*5**544@*)*-/4)
@M00779:331:000000000-C43T8:1:1101:22508:1545 1:N:0:359
TCCTACGGGAGGCTGCAGTGGGGAATCTTGCGCAATGGGCGGAAGCCTGCCGCGCAGCAACGCCGCGTGGGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCGCGATCGACGAAGGCCCACTTTCTGTGGGTTGACGGTAGGTCGAGAAGAAGCACCGGCCAACTACGTGCCAGCAGCCGCGGGAATACGTAGGGTACGAGCGTT
GTCCGGAATTATTGGTCGTAAAGAGCTCGTAGGCGGTGTGTCGCGTCGGCCGTGAAAACCTGGGGATCAACTCTGGGCGGGCGGTCGATACGGGCATCAACTGAGTTCGGCAGGGGG
+
CCCCCGE@BFGDEGCEGFFE@F@7FGGGGGGGGEG?EFC:FGCEGGGFEGGGGGGGGGGGGEGGCFGGGCFCFFGGDGGEFFFEFGEGECGFFFGFFGGGGGGGGGF@<FF+FFG7FGGGGGA9,,,*5:F:CCGEGDFGC***6<7CFGBF::ECGGGGCGGB29CFGCGFFGFGGG5*:CCGGG<+/
A**+31;*/:*8CECFGD****+2<+<2+1:C**+<C4>4EEG3**)17).*>B>FC)>3CD>3C<*77*>DGFC3C)*225*A*19*05]8>9>)((0))((0](8,>*))6).)926()/486)3-
@M00779:331:000000000-C43T8:1:1101:17545:1568 1:N:0:359
CCCTACGGGAGGCTGCAGTGGGGAATTTTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTGTCCGGAAAGAAATGGCTCTGGTTAATACCTGGGGTCGATGACGGTACCGGAAGAATAAGGACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTCGGGT
CCAAGCGTTAATCGGAATTACTGGGCGTACAGCGTGCGCAGGCGGTTGGTCAAGACCGATGTGAACTCCCGAGCTTAACTTGGGAATTGCATTGGTGACTGCACGGCTCGAGTGGG
+
CCCCCFGGGGGGGGGGF@FCDGGGGGFGGGDGGGFFFGCFGFGGGGBFFGGGGGGGFEFFDFGGGEGGGGGGGGGGGGGBE9FFEFCFGGFG9DAF<E<EAFGGGGGGGGGGGFGGGGDFGBC<5:AFFE95=E7FFEGDE;CF=FF:FGE++3>:8>,EF38+@C>8@FGCBC>2>>>F6*,44:CE*=B*;+<>;
0**:B*:*5>C38/*2AC7C75/5++*2+<59<5=*19>))1.7)/7:DD5)195.*29*C7FDD3*)2*020+293)*7>:**9@/*4<).4)9*)18)/617)96.-)5(030925).((
@M00779:331:000000000-C43T8:1:1101:18410:1654 1:N:0:359
TCCTATGGGCGCAGCAGTGGGGAATTTTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTGTCCGGAAAGAAATGGCTCTGGTTAATACCTTGGGTCGATGACGGTACCGGGAGAATAAGGACCGGCTAACTACGTGCCAGCGGCCGCGGCAATACGTAGGGG
CCAAGCGTTAATCGGAATGACTGGGCGGGAAGCGTGCGCAGGCGGGTGTGCAAGACCGATATGAACTTCCCTAACTGACATGTGAAGTGCGTTAGTGGCTACACGGCCAGAGTGTG
+
<CCCCEGGCEECGGGGGGFGFGGGGGBECFEGGBFGFGGBF:FGGGGGGGFGGGGGGGGCGGGGCGG:FGG,AFDGGEGGDC9FGGGGGEGGGGDGGGGGFGGDGGD+>CEEF??9:,,:?F,5C9?BDDAE,86D+AB7+C==+883BCEG++5>EEF,743@@F7>*17DFG6:,:*4>,/
***1=C5>**28A@=2CFEE*9*/:8CE855;>21***969:C08A*/)**10*7)/.)7<<5BC*/1/9.07>C=5*<5]))0:***/+3*)*0*0=C*9)*1(*1-6799>()),04**-.)).)<(((]/(8./)6
@M00779:331:000000000-C43T8:1:1101:12482:1689 1:N:0:359
GCCTACGGGAGGCTGCAGTTAGGAATCTTCGTCAATGGGCGAAAGCCTGAACGAGCGACGCCGCTTGAGGGACGAAGCCCTTCGGGGTGTAAACCTCTGAACAGGGGACGATGATGACGGTACCCTGCTAATAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCGAACGTTACCCGGATTTACTGGGCG
TACAGGGCGGGTAGGCGGTTTTCTAAGTCCGGGGCTAAAGACCAGGGCTCAACCCTGGGTTTGCCGTTGATACTGGAAAGCCCGACGGCGGGAGGGGGCAGCGGAATTTCCGGGGTG
+
BCCCCFGGGGGEE<@FEGGFGGGGGGGGGGGGGGGGGFEGGGGGGGGGFGFGEEEECEEEGEGGGGGGG7=FGGGGGGGFFE<CCFGGEE@EEFFFGGCFFFBFF>FBFE7+E99BFGGGFEDCGGGFE=,DFGGFGEGGGGGGCECFGCCGGFCFCFEEEEEEG7CGGGGGCED8CGGCEEEGGCCGCEG58;EFGG?
F5@89C*;7<8BE**//1:9/5:E*C9FGCGF7EGFDGEGFFFGDF;CFG<C5>*9CF55>5>55@GGG]8*2**237=FC6/.))-5717BE)(4.)-<BG395()-0)4).7C(7(.90
@M00779:331:000000000-C43T8:1:1101:11954:1825 1:N:0:359
TCCTATGGGGCGCTGCAGTAGGGGAATATTGCGCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGGAGGATGACACTTTTCGGAGCGTAAACTCCTTTTGTTAGGGAAGAATAATGACGGTACCTAACGAATAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTCATACGGAGGGTGCCAGCGTTACTCGGAATCACTGG
GCGTAAAGGACGCGTAGGCGGATTATCAAGTCTCTTGTGAAATCTAACGGCTTAACCGTTAAACTGCTTGGGAAACTGATAATCTAGAGTAAGGGAGAGGCAGATGGAATTCTTGGG
+
CCCCCGEFBCFGEFGEGGEFGGGGGGGGFFGGGGCEGGGGGGGGGFGGCFDGDGGGGGGGGGGGGGGCFGFFGGGGGGGGGGGGGGGGGGFFGCFGGGGGGGGGGGGGGGGGGG
```

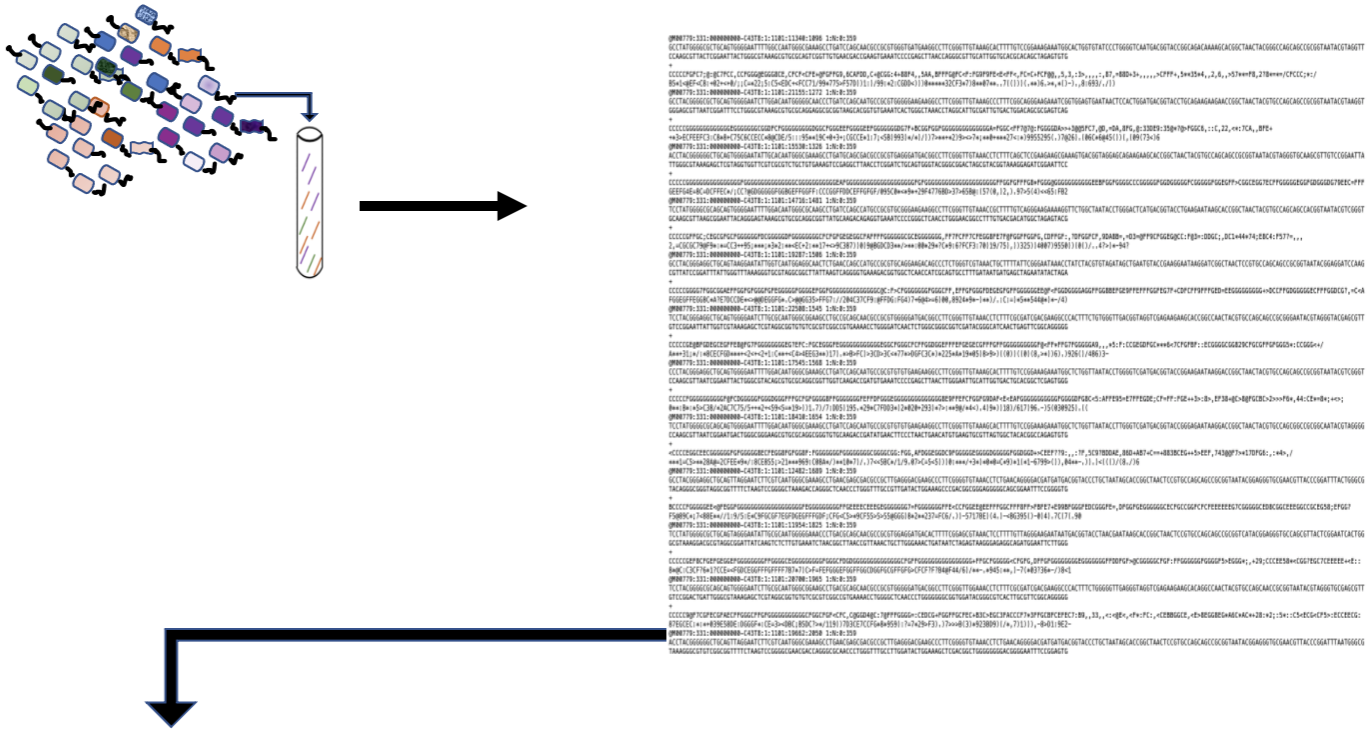


Table 1: Count matrix $\mathbf{K} \in \mathbb{R}^{m \times N}$

	Specimen_1	Specimen_2	$\cdots$	Specimen_{n_1}	$\text{Control}_{(n_1+1)}$	$\text{Control}_{(n_1+2)}$	$\cdots$	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}

Prepare the data for the analysis

Raw sequences to sequencing table

- Use DADA2 package in Bioconductor[1]

[1] DADA2: high-resolution sample inference from Illumina amplicon data (Callahan et al. 2016)

- Install required packages.

```
pkgs <- c(
  "dada2",
  "tidyverse",
  "phyloseq"
  )
BiocManager::install(
  setdiff(
    pkgs,
    installed.packages()
    ),
  update = TRUE
  )
```

- Load required packages.
- Print the raw sequence files for one sample.

```
library(dada2); library(tidyverse)
library(phyloseq)
path <- "data"
list.files(path)
```

```
## [1] "in1299_1_R1.fastq" "in1299_1_R2.fastq" "psE.rda"
```

- Forward and reverse fastq filenames have format: SAMPLENAME_R1.fastq and SAMPLENAME_R2.fastq.

```
fnFs <- sort(
  list.files(
    path,
    pattern="_R1.fastq",
    full.names = TRUE
    )
  )
fnRs <- sort(
  list.files(
    path,
    pattern="_R2.fastq",
    full.names = TRUE
    )
  )
```

- Extract sample names, assuming filenames have format: SAMPLENAME_R1.fastq.

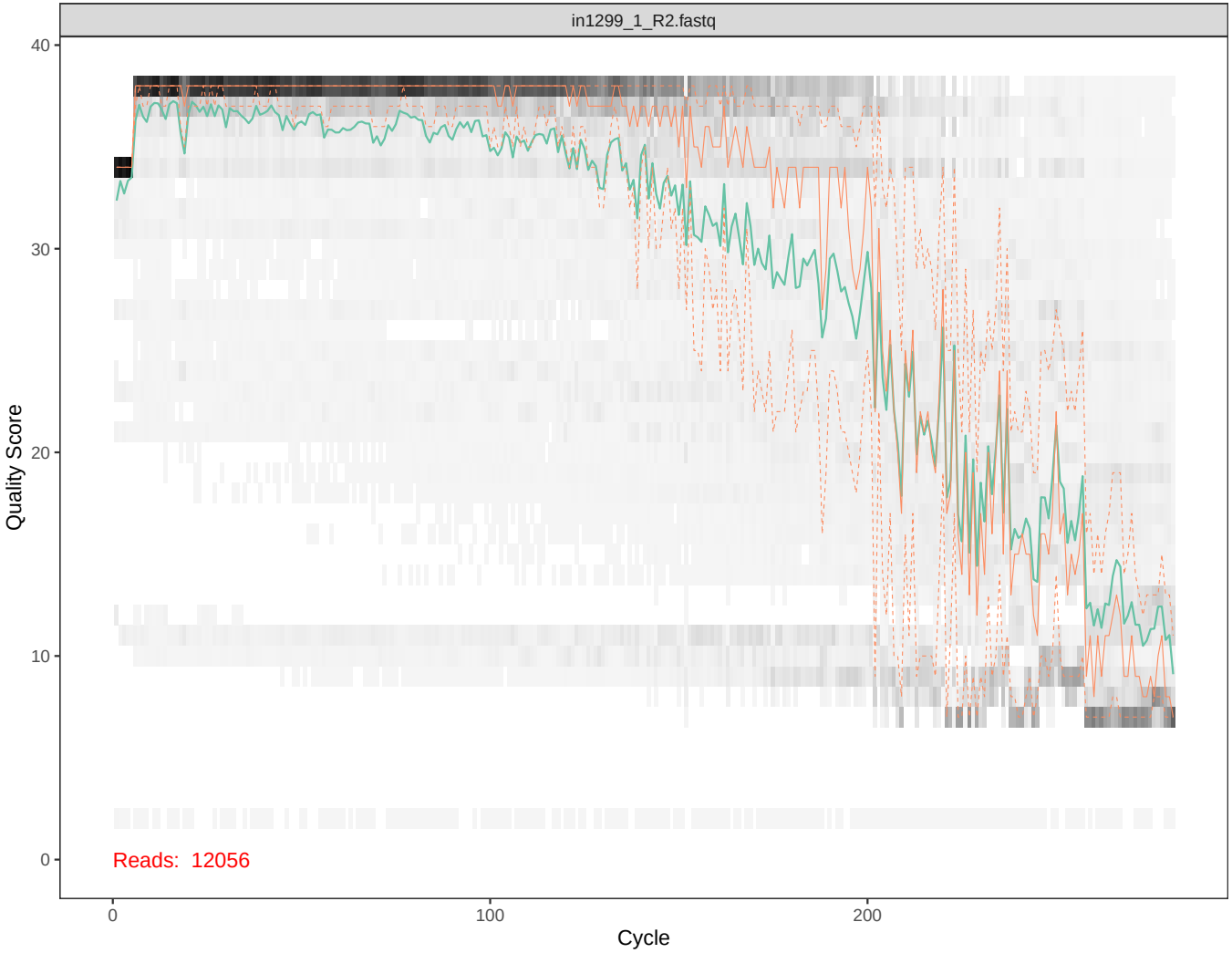
```
sample.names <- sapply(
  strsplit(
    sub('(^[^_]+_[^_]+)_(.*)$',
        '\\1 \\2',
        basename(fnFs)),
    " "
    ),
  `[`,
  1
  )

sample.names <- gsub(
  "\\_",
  ".",
  sample.names
  )
```

- Quality profile of forward reads.

```
plotQualityProfile(fnFs)
```

- Quality profile of reverse reads.

```
plotQualityProfile(fnRs)
```



- Define directory for filtered sequences.

```
filtFs <- file.path(
  path,
  "filtered",
  paste0(
    sample.names, "_F_filt.fastq.gz"
          )
  )
filtRs <- file.path(
  path,
  "filtered",
  paste0(
    sample.names, "_R_filt.fastq.gz"
    )
  )
```

- Filtering raw sequences
  - Primers are at the start of the reads and are a constant length (16 nt, 24 nt).
  - Use trimLeft = c(FWD_PRIMER_LEN = 16, REV_PRIMER_LEN = 24) to remove the primers.
  - Based on the quality scores plots (and Zymo suggestion), we remove FWD reads length less than 320 nt and REV reads length less than 180 nt.

```r
out <- filterAndTrim(
  fnFs,
  filtFs,
  fnRs,
  filtRs,
  truncLen=c(320, 180),
  trimLeft = c(16, 24),
  maxEE = c(10, 10),
  maxN = 0,
  truncQ = 6,
  rm.phix=TRUE,
  compress=TRUE,
  multithread=TRUE)

out
```

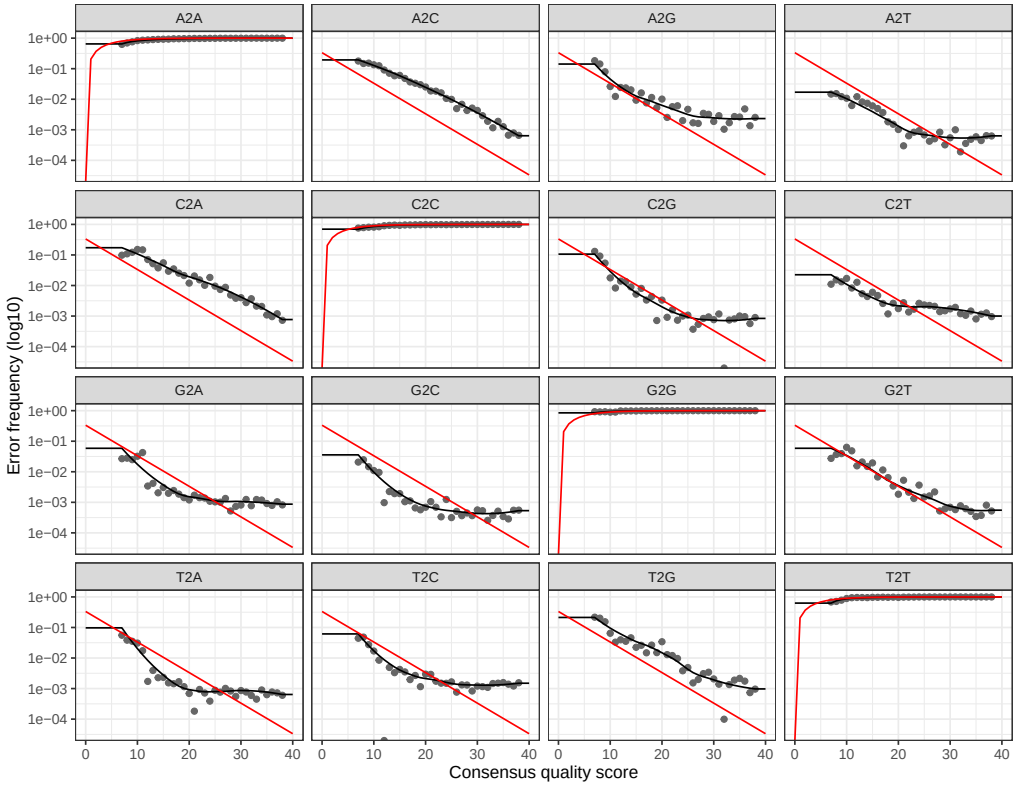
```
##                     reads.in reads.out
## in1299_1_R1.fastq    12056     10502
```

Learn the error rates

```
errF <- learnErrors(filtFs, multithread=TRUE)
```

```
## 3192608 total bases in 10502 reads from 1 samples will be used for learning the error rates.
```

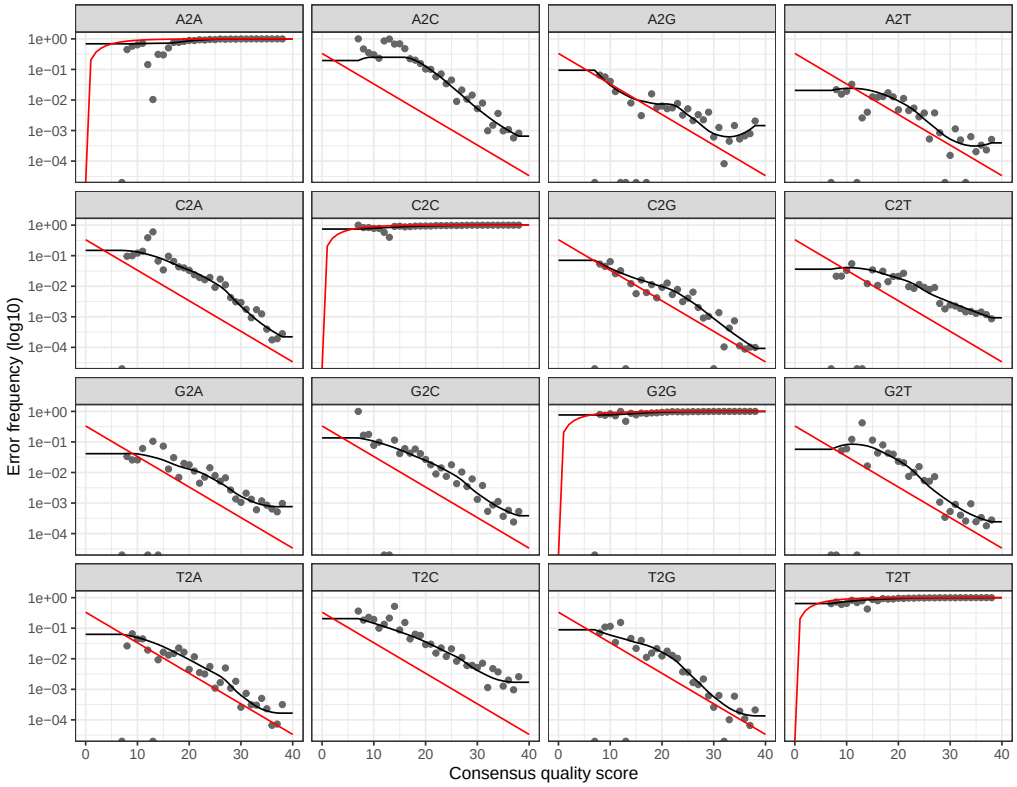
```
plotErrors(errF, nominalQ=TRUE)
```



```r
errR <- learnErrors(filtRs, multithread=TRUE)
```

```
## 1638312 total bases in 10502 reads from 1 samples will be used for learning the error rates.
```

```r
plotErrors(errR, nominalQ=TRUE)
```



- Dereplication - combines all identical sequencing reads into "unique sequences" with a corresponding "abundance" equal to the number of reads with that unique sequence.

```
derepFs <- derepFastq(
  filtFs,
  verbose=TRUE
  )
derepRs <- derepFastq(
  filtRs,
  verbose=TRUE
  )
```

- Sample inference

```
dadaFs <- dada(
  derepFs,
  err=errF,
  pool = FALSE,
  multithread=TRUE
  )
```

```
## Sample 1 - 10502 reads in 9678 unique sequences.
```

```
dadaRs <- dada(
  derepRs,
  err=errR,
  pool = FALSE,
  multithread=TRUE
  )
```

```
## Sample 1 - 10502 reads in 4022 unique sequences.
```

- Inspecting the returned dada-class object.

```
dadaFs
```

```
## dada-class: object describing DADA2 denoising results
## 18 sequence variants were inferred from 9678 input unique sequences.
## Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
```

```
dadaRs
```

```
## dada-class: object describing DADA2 denoising results
## 24 sequence variants were inferred from 4022 input unique sequences.
## Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
```

- Merge paired reads.

```
mergers <- mergePairs(
  dadaFs,
  derepFs,
  dadaRs,
  derepRs,
  verbose=TRUE,
  minOverlap = 12,
  maxMismatch = 0
  )
head(mergers[[1]])
```

```
## [1] "AGCCAGGAATCTTGCGCAATGGGCGAAAGCCTGACGCAGCAACGCCGCGTGGGCGATGAAGGCCTTCGGGTCGTAAAGCCCTGTTGTCGGGGACGAAC
## [2] "AGTTAGGAATCTTCGTCAATGGGCGAAAGCCTGAACGAGCGACGCCGCTTGAGGGACGAAGCCCTTCGGGGTGTAAACCTCTGAACAGGGGACGATGA
## [3] "AGTGGGGAATTTTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTGTCCGGAAAGAAA
## [4] "AGTGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCGACGCCGCGTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAGCTCCGAAGAAC
## [5] "AGTGGGGAATTTTCCGCAATGGGCGAAAGCCTGACGGAGCAAGACCGCGTGAGGGAGGAAGGCTCTTGGGTTGTAAACCTCTTTTCTTGGGGAAGAAC
## [6] "AGTAAGGAATATTGGTCAATGGAGGCAACTCTGAACCAGCCATGCCGCGTGCAGGAAGACAGCCCTCTGGGTCGTAAACTGCTTTTATTCGGGAATAA
```

- Construct sequence table.

```
seqtab <- makeSequenceTable(mergers)
```

- Examine distribution of sequence length.

```
table(nchar(getSequences(seqtab)))
```

```
##
## 400 403 406 407 408 419 422 426 427
##   2   1   1   4   1   1   1   1   6
```

- Remove non-target-length sequences.

```
seqtab <- seqtab[,nchar(colnames(seqtab)) %in%
                    seq(310,430)]
```

- Remove chimeras (two sequences incorrectly join together).

```
seqtab.nochim <- removeBimeraDenovo(
  seqtab,
  method = "consensus",
  multithread=TRUE,
  verbose=TRUE
  )
```

- Assign taxonomy.
  - We choose to hit multiple species (multithread = TRUE).

```
taxa <- assignTaxonomy(
  seqtab.nochim,
  "./silva_nr_v132_train_set.fa.gz",
  multithread = TRUE
  )

taxa <- addSpecies(
  taxa,
  "./silva_species_assignment_v132.fa.gz",
  allowMultiple = TRUE
  )
```

Construct phyloseq[2]

```
samples.out <- rownames(
  seqtab.nochim
  )
sampleType <- ifelse(
  sampleID %in% as.character(seq(1,10)),
  "Negative", "Standard"
  )
samdf <- data.frame(
  sampleID = samples.out,
  sampleType = sampleType
  )
rownames(samdf) <- samples.out

ps <- phyloseq(
  otu_table(
    seqtab.nochim,
    taxa_are_rows=FALSE
    ),
  sample_data(samdf),
  tax_table(taxa))
```

[2] phyloseq (McMurdie and Holmes 2013)

```r
ps
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 6929 taxa and 111 samples ]
## sample_data() Sample Data:       [ 111 samples by 13 sample variables ]
## tax_table()   Taxonomy Table:    [ 6929 taxa by 6 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 6929 tips and 6928 internal nodes ]
```

Components of microbiome data

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*

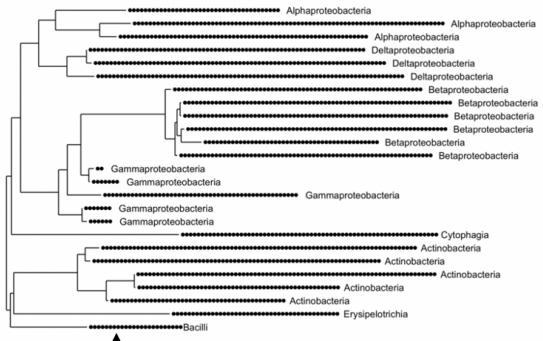


Figure 1: Phylogenetic tree.

Unknown parameter of interest

$\boldsymbol{p}_1 = (p_{i1}, i = 1, 2, \ldots, m)$ $\boldsymbol{p}_{n_1} = (p_{in_1}, i = 1, 2, \ldots, m)$

Prevalence matrix

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

Challenges

Unequal library sizes

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}
	S_1	S_2		S_{n_1}	S^0_1	S^0_2		S^0_N

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*

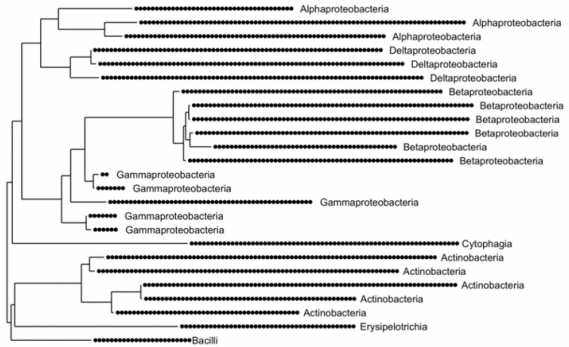


Figure 1: Phylogenetic tree.

Undetected taxa

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*

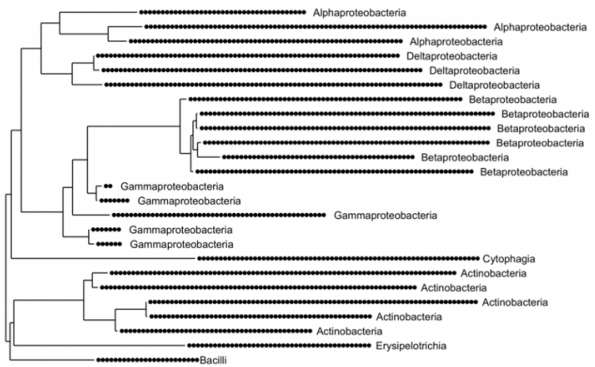


Figure 1: Phylogenetic tree.

Sparsity

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}		$\cdots$		$K^0_{2(n_1+1)}$	$K^0_{2(n_1+2)}$	$\cdots$	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}		$\cdots$	K_{mn_1}	$K^0_{m(n_1+1)}$	$K^0_{m(n_1+2)}$	$\cdots$	K^0_{mN}
	S_1	S_2		S_{n_1}	S^0_1	S^0_2		S^0_N

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*

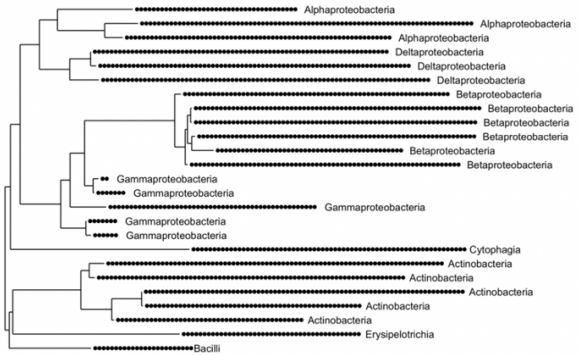


Figure 1: Phylogenetic tree.

Count table is sparse with a large proportion of zeros (both technological and biological.)

Heterogeneity

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}		$\cdots$		$K^0_{2(n_1+1)}$	$K^0_{2(n_1+2)}$	$\cdots$	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}		$\cdots$	K_{mn_1}	$K^0_{m(n_1+1)}$	$K^0_{m(n_1+2)}$	$\cdots$	K^0_{mN}
	S_1	S_2		S_{n_1}	S^0_1	S^0_2		S^0_N

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*

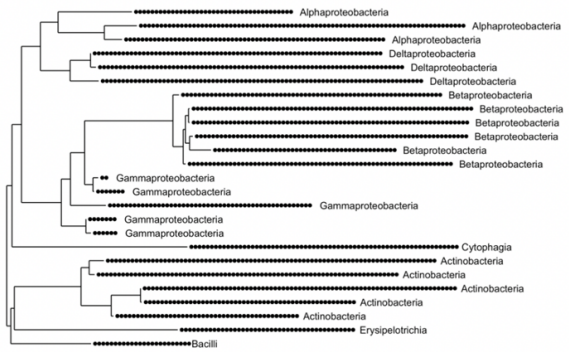
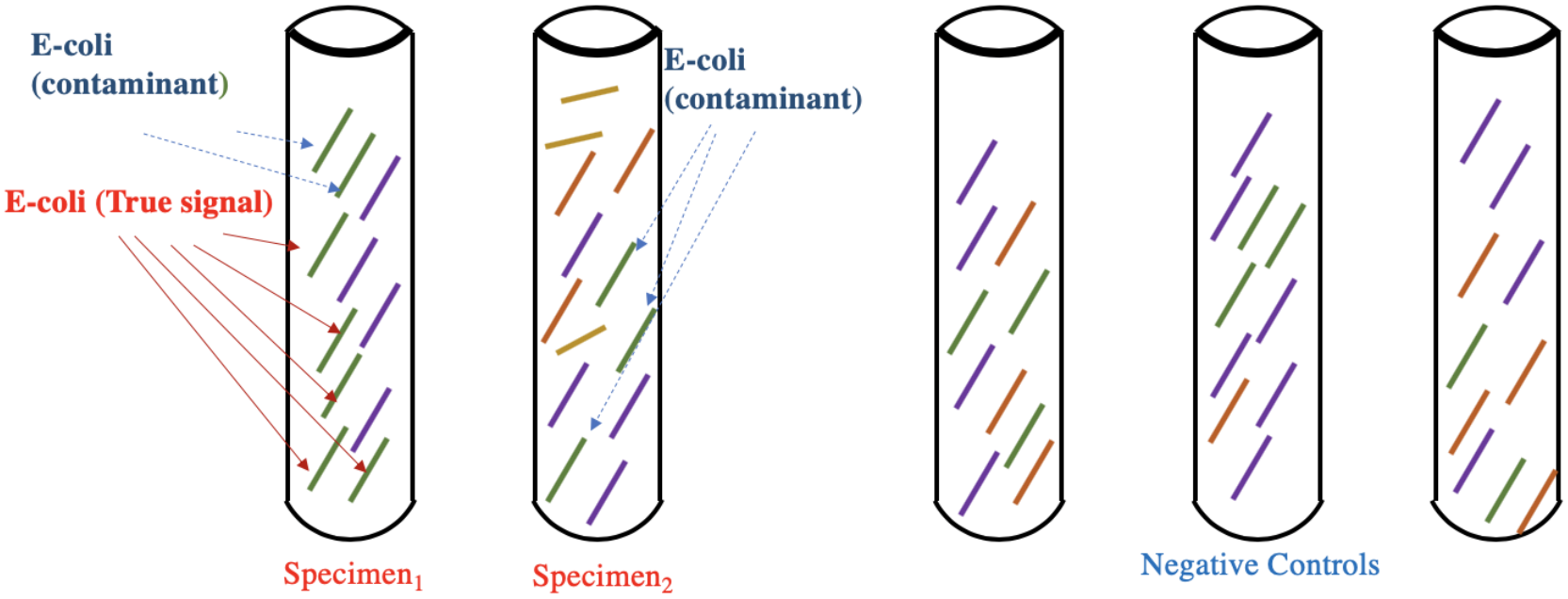


Figure 1: Phylogenetic tree.

DNA contamination

- Contaminating DNA from external sources.
  - extraction reagents, DNA from library preparation, and amplification kits; lab environments, and skin.



Example data set

Plant microbiome study

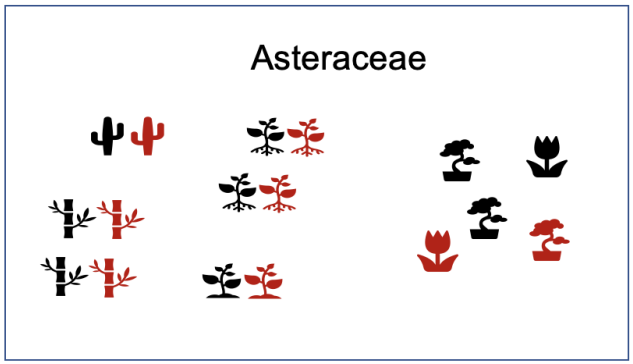
- Published dataset of 16S rRNA gene sequencing of the root endosphere specimens[3] .
  - Retrieved from Google Dataset Search.
- Started from the sequencing table available in the above database.

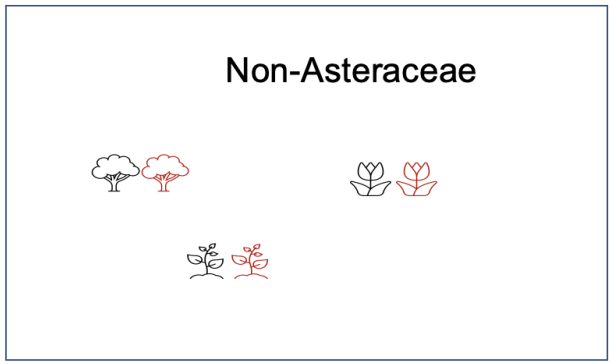
[3] Fitzpatrick et al. (2018)

Background of the study

- In plants, host organellar sequences (mitochondrial and plastid) share similar bacterial lineages.
  - Reduces the efficiency of quantifying microbial sequences.
  - Universal peptide nucleic acids (O-PNA) have been used to limit the amplification of host-derived mitochondrial and plastid sequences.
- Study goal.
  - Evaluate the efficacy of O-pPNA and Asteraceae-modified pPNA (M-pPNA) in limiting plastid host contaminants.

- Experimental design
  - Two different plant families - Asteraceae or non-Asteraceae.
  - Six Asteraceae and three non-Asteraceae plant types.
  - Four out of six Asteraceae and all three non-Asteraceae plant types were sequenced with both pPNA (**O** and **M**) and had 18 paired specimens.





Processing[4]

- Removed low coverage host plant type samples, removed Archaea, Eukaryota or unknown, remove unkown Phylum.
- Removed mitochondira and Chloroplast.
  - M-pPNA reduced plastid host contaminants in Asteraceae plants, but increased contaminants in non-Asteraceae plants.
- Goal: Compare microbial distribution in two different pPNA types (original and Asteraceae-modified).

[4] Section 3: EXAMPLE OF 16S rRNA GENE SEQUENCING DATA

Models[7]

- Removed DNA contamination using BARBI[5], in the presence of negative controls.
- Goodness of fit tests.
  - A negative binomial distribution (or equivalently gamma-Poisson) fits our example data set for the ASV counts well.
- Library size scaling factor.
  - Account for the library sizes (the column sums of the ASV contingency table) using median-of-ratios algorithm implemented in DESeq2[6].

[5] https://pratheepaj.github.io/BARBI/
[6] DESeq2 (Love, Huber, and Anders 2014)
[7] Section 4: MODELS

Transformations[8]

- Use transformed data to explore pattern in specimens and ASVs.
  - Anscombe's transformation of abundance data.
- Use rank-based methods for exploring patterns.

[8] Section 5: TRANSFORMATIONS

Visualization[9]

- Use barplots and heatmaps to identify taxonomic patterns or differences in various phenotypes.

[9] Section 6: VISUALIZATIONS FOR HETEROGENEOUS DATA

Multivariate methods and network analysis[10]

- Use ordination, ordination with phylogentic information, correspondence analysis, and network analysis.
  - Identify outliers, clusters, and relative position or gradients of specimens in low dimensions.

[10] Section 7: MULTIVARIATE AND NETWORK ANALYSES

Differential abundance analysis[11]

- Find taxonomic differences across environments or groups.
- Permutation tests using distance matrices.
  - In 16S rRNA data, the distance based tests are very sensitive to the choice of distance and the presence of **strain switching** can substantially decrease the power of the test.
- Strain switching - present as one ASV in one set of specimens and a close, distinct strain appears in the other set of specimens.
- In the presence of strain switching, power to detect a difference using a Bray Curtis distance and `permanova` is considerably diminished.

[11] Section 8: DIFFERENTIAL ABUNDANCE ANALYSIS

Differential abundance analysis contd.

- Differential abundance through generalized linear modeling and transformations suffer from strain switching.
- We proposed that topic models provide useful aggregates that can be used for differential abundance analysis based on topics rather than individual strains.

Differential topic analysis

Latent Dirichlet allocation[14]

- `diffTop` package is developed for differential topic analysis[12].
- Why differential topic analysis?
  - Metagenomic assembly of samples from diverse population shows that many species split into different SGBs (species-level genome bins), some of which where polyphyletic and had genetic diversity on the order of genus and family[13].
  - Topic analysis is useful to infer granular changes in microbiome communities when we combine data from different demographic or geographic regions or omics over time.

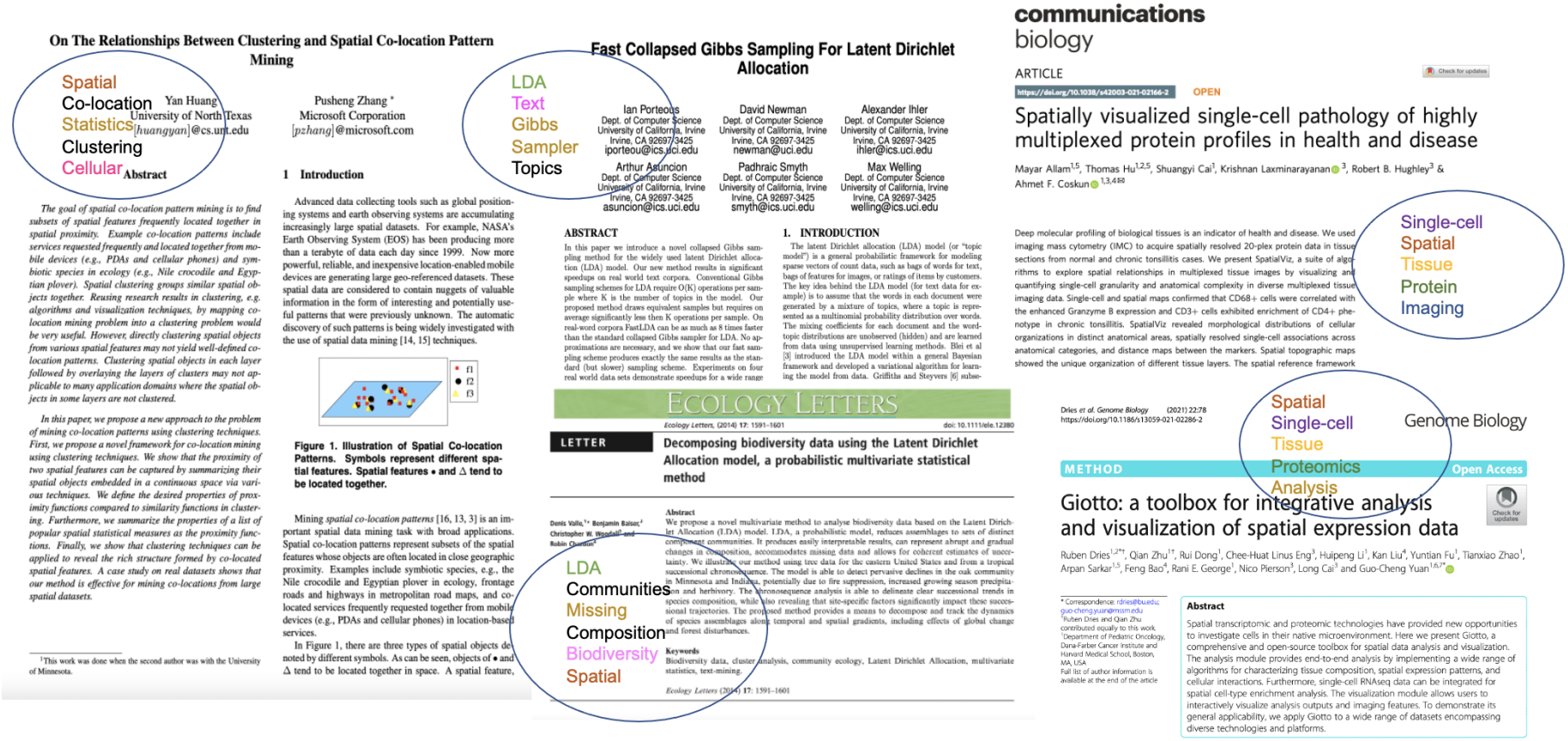
[12] *A Statistical Perspective on the Challenges in Molecular Microbial Biology* (Jeganathan and Holmes 2021)

[13] Species-level genome bins (SGBs) from the human microbiome

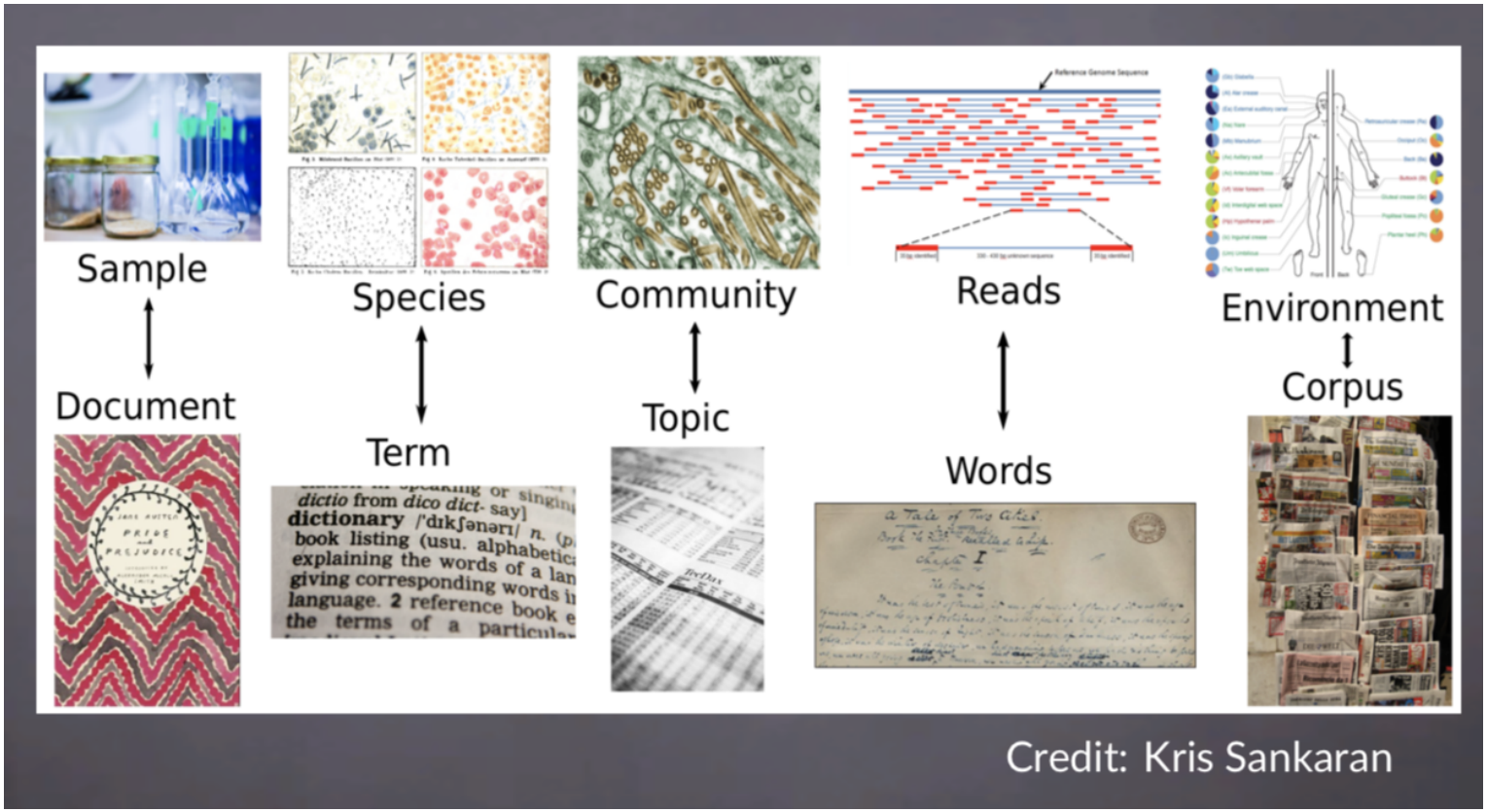
[14] Section 9: LATENT DIRICHLET ALLOCATION

Topic analysis

- Consider abstracts of some journal articles and create a word cloud for each article using the top five frequent words.



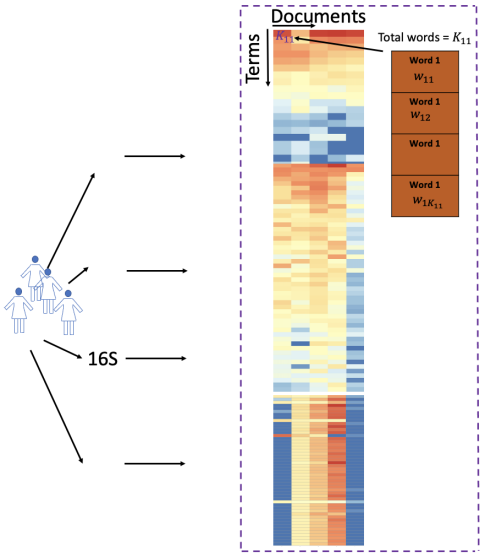
Topic analysis for molecular microbiome data[15]



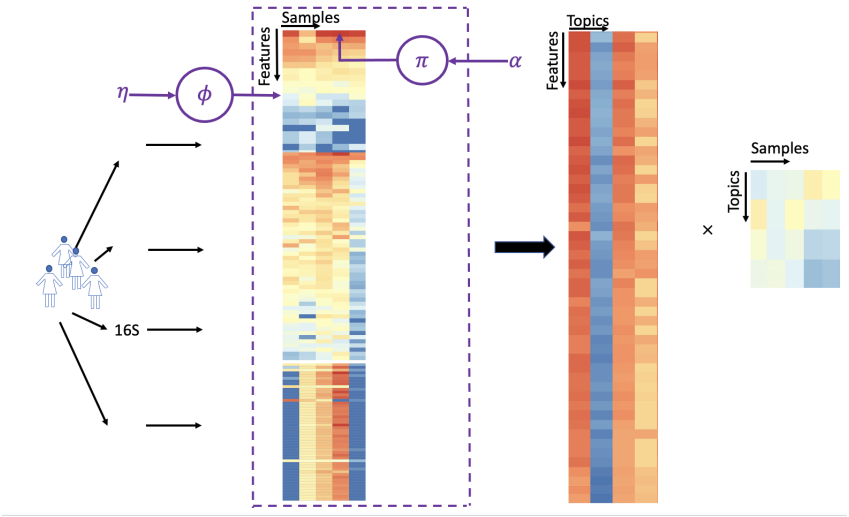
[15] Latent variable modeling for the microbiome (Sankaran and Holmes 2019)

Notations

- D documents, V terms, K underlying topics, N_d total number of words in the d-th document, w_{dn} term in the n-th word of d-th document, and $z_{dn} \in \{1, 2, \ldots, K\}$ the associated topic, k-th topic places weight π_{vk} on the v-th term.



Latent Dirichlet allocation[16]



$w_{dn} | (\boldsymbol{\pi}_k)_{k=1}^K, z_{dn} \sim \text{Multi}(1, \boldsymbol{\pi}_{z_{dn}})$ for $d = 1, \ldots, D$ and $n = 1, \ldots, N_d$.

$z_{dn} | \boldsymbol{\psi}_d \sim \text{Multi}(1, \boldsymbol{\psi}_d)$ for $d = 1, \ldots, D$ and $n = 1, \ldots, N_d$.

$\boldsymbol{\pi}_k \sim \text{Dir}(\alpha)$ for $k = 1, \ldots, K$.

$\boldsymbol{\psi}_d \sim \text{Dir}(\eta)$ for $d = 1, \ldots, D$.

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

Stan (probabilistic programming language) model

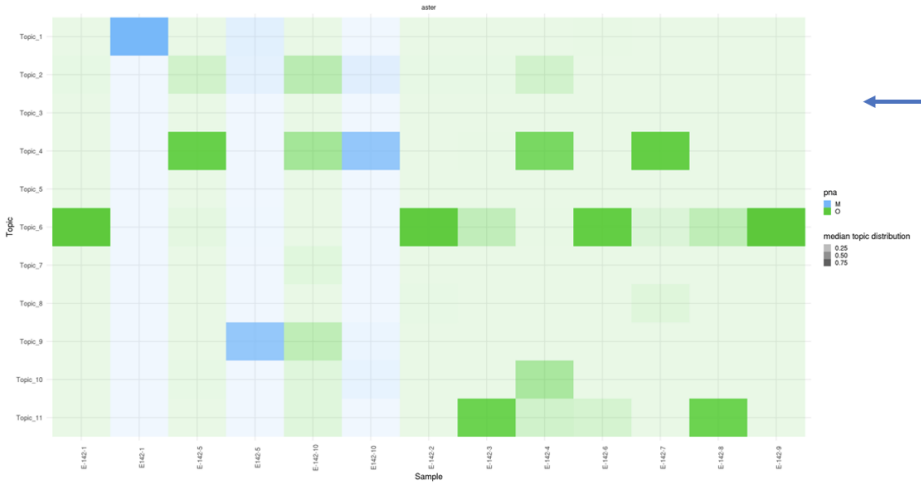
lda.stan

Check on Save Check

```
// copyright to Kris Sankaran https://github.com/krisrs1128/boot_expers/blob/master/
data {
  int<lower=1> K; // num topics
  int<lower=1> V; // num words
  int<lower=1> D; // num docs
  int<lower=0> n[D, V]; // word counts for each doc

  // hyperparameters
  vector<lower=0>[K] alpha;
  vector<lower=0>[V] gamma;
}

parameters {
  simplex[K] theta[D]; // topic mixtures
  simplex[V] beta[K]; // word dist for t^th topic
}

model {
  for (d in 1:D) {
    theta[d] ~ dirichlet(alpha);
  }

  for (k in 1:K) {
    beta[k] ~ dirichlet(gamma);
  }

  for (d in 1:D) {
    vector[V] eta;
    eta = beta[1] * theta[d, 1];

    for (k in 2:K) {
      eta = eta + beta[k] * theta[d, k];
    }

    n[d] ~ multinomial(eta);
```

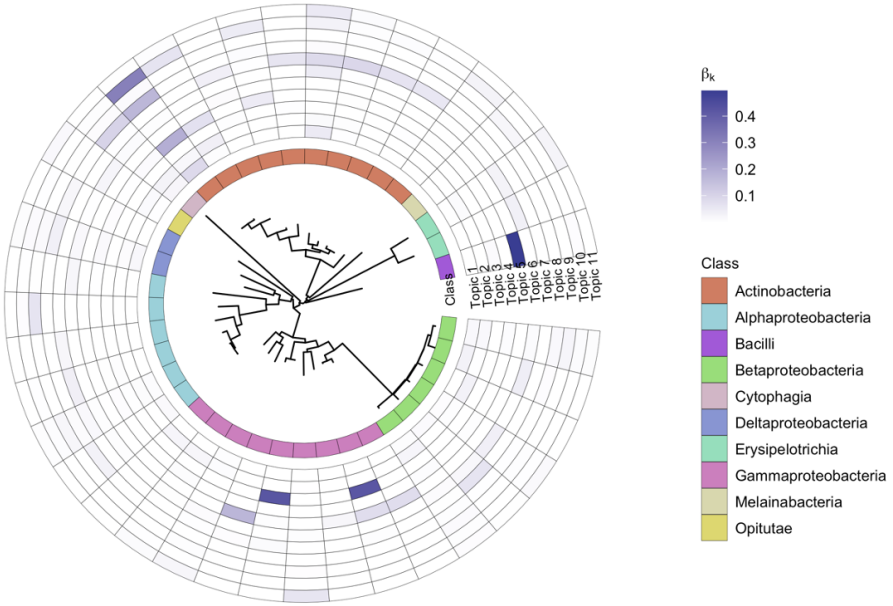
15:41 Parameters Stan

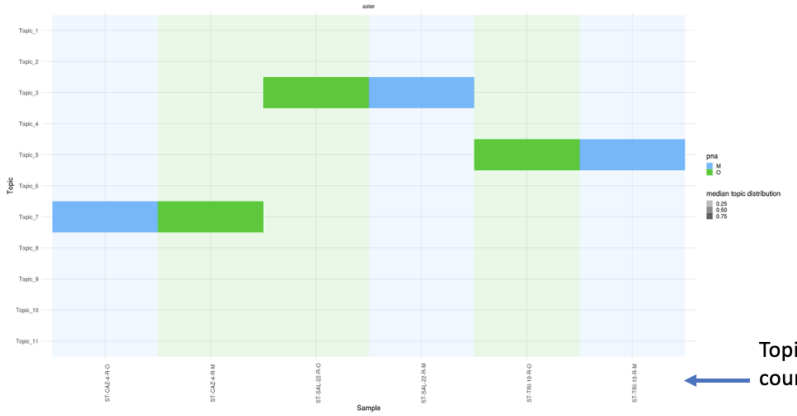
LDA results[17]



Topic distribution in specimens from *Sonchus Arvensis* plant. The color gradient represents the median topic distribution.

ASV distribution over topics in all specimens.





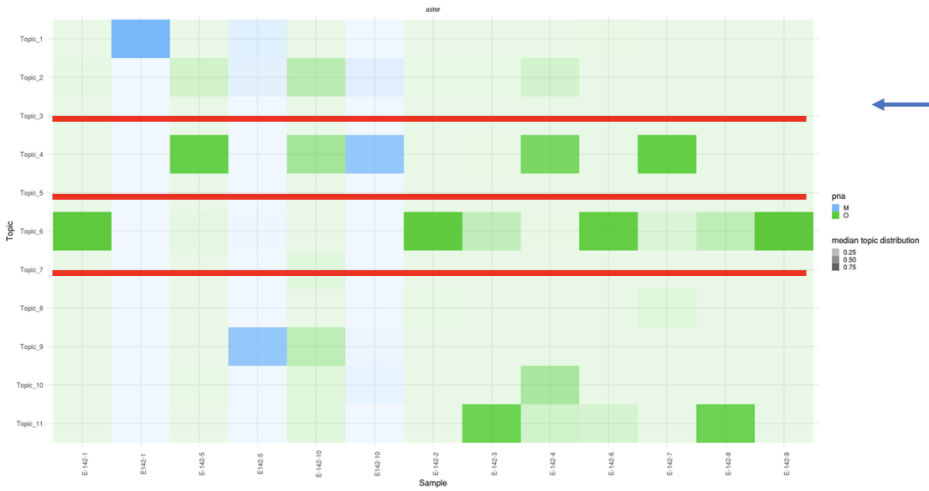
Topic distribution in specimens from *Centaurea solstitialis*, which is from three different countries. This plant has three paired specimens sequenced with O and M pPNA types.

[17] LDA workflow

LDA results

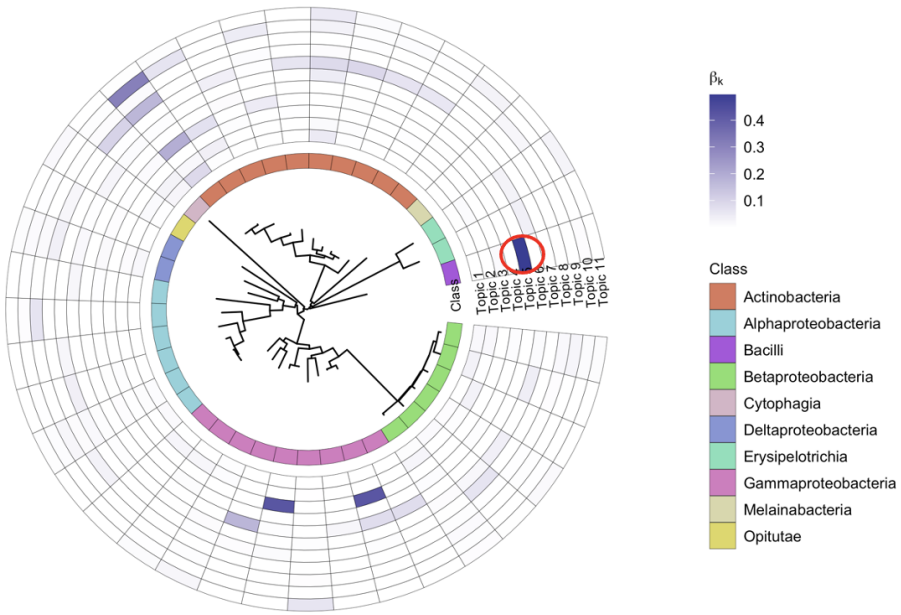
- Among the eleven topics for all plants, seven topics were predominately present in specimens from *Sonchus Arvensis*, an Asteraceae plant with 13 replicates specimens.
- Ten specimens were amplified with O and three with M-pPNA types, respectively.
  - Among these specimens, E-142-1, E-142-5, and E-142-10 are paired because both pPNAs were used for the same DNA specimens.
  - Paired-specimens (E-142-1, E142-1) and (E-142-5, E142-5) have a different mixture of topics.

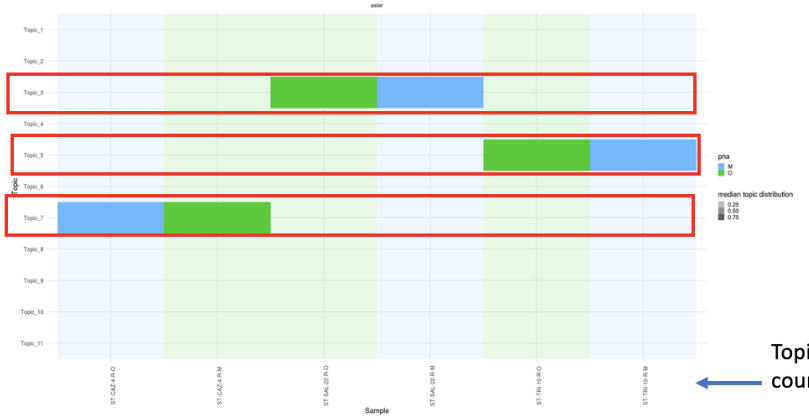
LDA results



Topic distribution in specimens from *Sonchus Arvensis* plant. The color gradient represents the median topic distribution.

ASV distribution over topics in all specimens.





Topic distribution in specimens from *Centaurea solstitialis*, which is from three different countries. This plant has three paired specimens sequenced with O and M pPNA types.

LDA results

- The replicates of *Centaurea solstitiali* that sampled across three countries show variation in microbial distribution, but the paired-specimens have similar distributions.
- We choose eleven topics based on the exploratory analysis.
  - Stability analysis is preferred.

diffTop[19]

- Use the microbial community-topics in each specimen to do a differential topic analysis to test whether topic memberships differ across conditions.
  - 1. compute the median Bayesian posterior of topic proportions in each specimen.
  - 2. multiply the proportions by the library size and round to an integer - gives abundance of the topic in each specimen.
  - 3. Apply `DESeq2`[18] to identify differentially abundant topics across O and M pPNA types.

[18] DESeq2 (Love, Huber, and Anders 2014)
[19] diffTop vignette

diffTop results

- Topics 4, 5 and 10 are differentially abundant in O and M pPNA types.
- There is a large proportion of Topics 1, 9, and 11 in Asteraceae paired-specimens.
  - These microbial communities are different in pPNA types.
- Topics 1, 2, 8, 9, 10 dominates in non-Asteraceae paired-specimens and we do not reject the hypothesis that these topics are differentially abundant.
- Conclusion
  - microbial communities are significantly different in some Asteraceae plants across pPNA types.

Comments

- By applying a linear mixed-effects model for each taxon, original paper concluded that the pPNA types do not affect the identification of individual bacterial taxa and within specimen α- diversity for different metrics.
- The authors found little evidence of a difference between specimen β- diversity using PERMANOVA
- Differential topic analysis identifies some topics in the Asteraceae paired-specimens that are significantly different.

Conclusion

- Strain switching precludes simple differential abundance analysis at the taxonomic level.
- Topic models
  - interpretable as they unmask rare and synonymous taxa and can enhance our understanding of the assembly of microbial communities as topics which can be projected onto the phylogenetic tree.
- The ASV distributions over interpretable topics demonstrate that differential topic analysis can enhance our understanding of the differences in complex microbiome communities (here across different plants and pPNA types).

Future research direction

- Sources of biological variability can be incorporated into the hierarchical model
  - For example, subject or location variation can be modeled as random effects.
- Microbiome multi omic integrative analysis can be benefited from differential topic analysis.
- Many other directions suggested in our paper[20]

[20] A statistical perspective on the challenges in molecular microbial biology (Jeganathan and Holmes 2021)

Thank You!

Email: jeganp1@mcmaster.ca

References I

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

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "Dada2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581.

Fitzpatrick, Connor R, Patricia Lu-Irving, Julia Copeland, David S Guttman, Pauline W Wang, David A Baltrus, Katrina M Dlugosch, and Marc TJ Johnson. 2018. "Chloroplast Sequence Variation and the Efficacy of Peptide Nucleic Acids for Blocking Host Amplification in Plant Microbiome Studies." *Microbiome* 6 (1): 144.

References II

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.

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

References III

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.

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.