Microbiome data analysis

Kim Dill-McFarland, PhD (kadm@mail.ubc.ca)
March 05, 2019

Experiential
Data science for
Undergraduate
Cross-disciplinary
Education

Learning objectives

- Define microbiome and microbiota
- Describe computational and statistical challenges in microbiome research
- Assess microbiome data using beta-diversity

What is a microbiome?

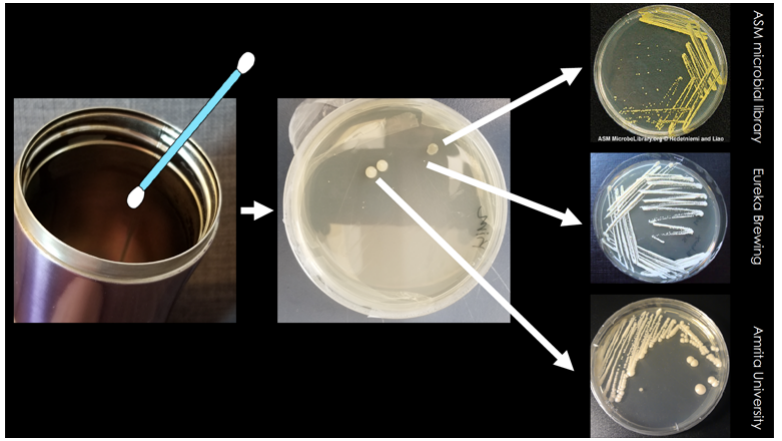
Term clarification

Microbial community present in a particular environment

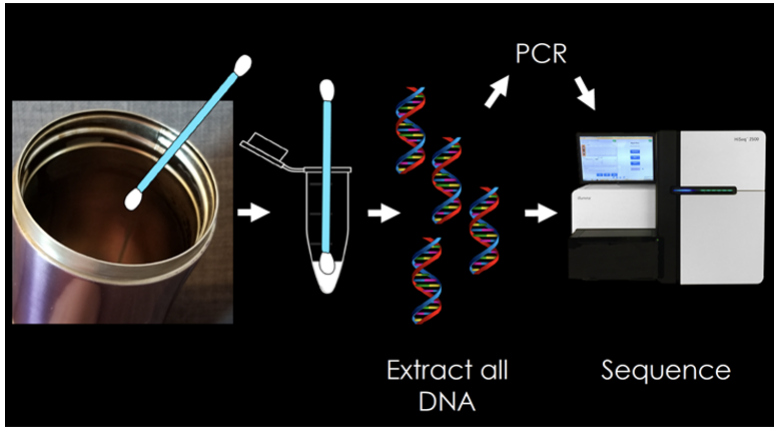
- **Microbiome**: all combined genetic material of the microorganisms in a particular environment

- **Microbiota**: the microorganisms in a particular environment

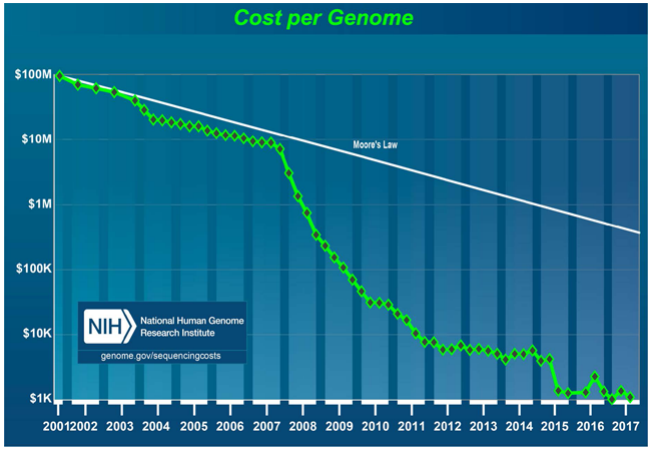
Studying microbial communities - Past



Studying microbial communities - Now



Moore's Law



Waypoints in history

2000: Lynx Therapeutics massively parallel signature sequencing (MPSS)

2004: Roche 454 parallel pyrosequencing

2007: Illumina buys Solexa which bought Lynx

2015: Roche shuts down 454 sequencing

The "microbiome" is born

Let's talk numbers

Sanger

- 150,000 bp / day
- Up to 1,000 bp long

Roche GX-FLX 454

- Up to 4.5 million bp / day
- Up to 700 bp long

Illumina MiSeq or HiSeq

- Up to 1.2 trillion bp / day
- Up to 300 bp long

- **++ Number**
- **— Length**

Challenges with microbiome data

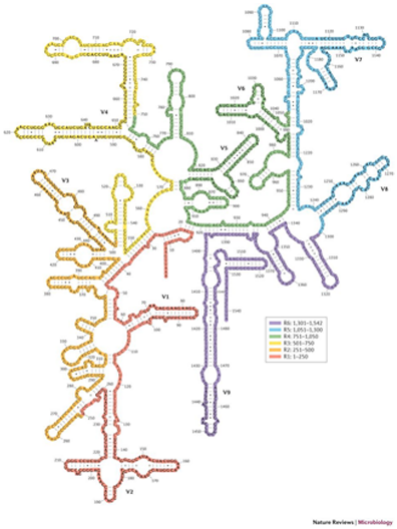
1. Big data, high complexity

- Solution: obtain high coverage and assemble into longer sequences
- **Solution: targeted sequencing of specific gene(s)**

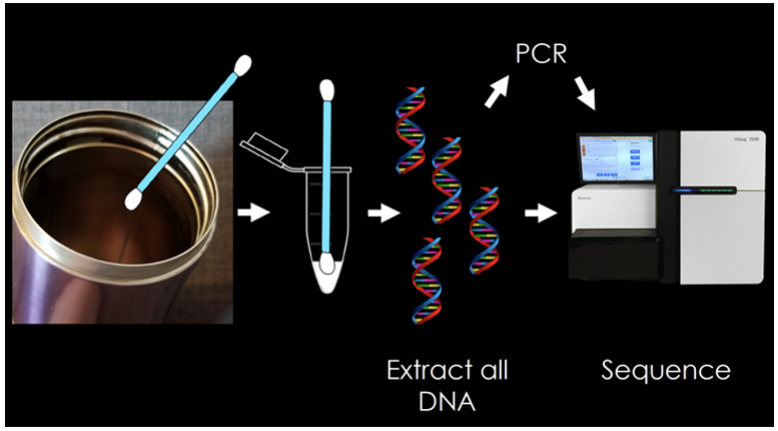
Identity genes

- Requirements:
- All species (within a group) must have the gene
- The gene must contain enough differences to tell related species apart
- The gene must have both conserved and variable regions

Bacterial 16S rRNA gene



16S pipeline



Raw 16S data

DNA sequences separated by observation

```
## >SRR5809668.19649
## CTGGGGGGTGCCAGCCGCCGCGGTAATACGGAGGGGGTTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGTAC
## >SRR5809668.21945
## CTGTGCCGTGCCAGCAGCCGCGGTAATACGGAGGGGGTTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGTAC
## >SRR5809668.41549
## CTGGTGGGTGCCAGCCGCCGCGGTAATACGGAGGGGGTTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCAC
## >SRR5809668.33746
## TTATGAGTGCCAGCAGCCGCGGTAATACGGAGGGGGTTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGTACG
## >SRR5809668.53410
## CTGGGGGGTGCCAGCCGCCGCGGTAATACGGAAGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGC
```

16S data cleaning

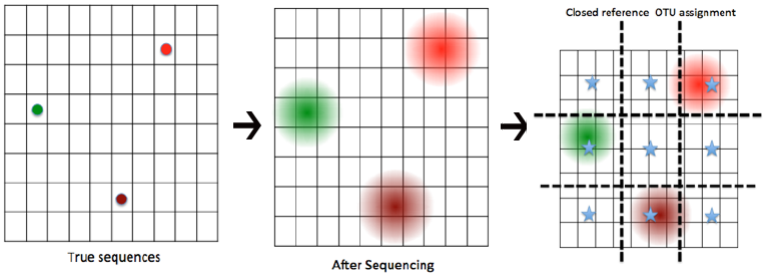
- Remove sequencing errors
- Align to a database
- Cluster very similar sequences (assumed error)
- Remove chimeras
- Classify to a database**

Challenges with microbiome data

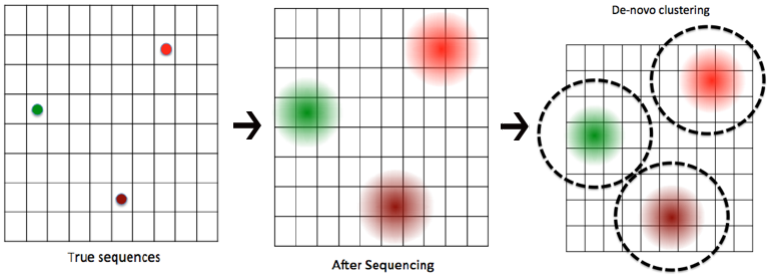
1. Big data, high complexity
2. Microbial species definition

- Solution: Operational taxonomic units (OTUs)
- Solution: Amplicon sequence varients (ASVs)

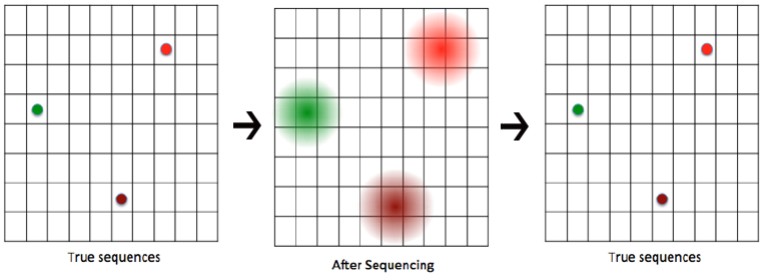
Operational taxonomic units (OTUs)



Operational taxonomic units (OTUs)



Amplicon sequence variants (ASVs)



OTU vs. ASV

OTU

- More established in the literature (2000-)
- Keeps more data
- Uses representative sequence of each OTU to determine taxonomy
- Various levels (*e.g.* 97% = species)

ASV

- Less established and newer (2017-)
- Discards more data
- Treats each ASV as a “species”

16S OTU data

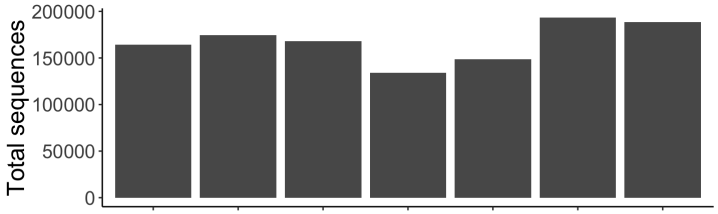
```
## # A tibble: 7 x 7
##   Sample      Otu0001 Otu0002 Otu0003 Otu0004 Otu0005 Otu0006
##   <chr>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Saanich_010     263       0    3210      26     108       1
## 2 Saanich_100    6489       0   18405     779    2713       6
## 3 Saanich_120   24380       0    8221    3404    5110      56
## 4 Saanich_135   39519       3    2793    5368    8216     198
## 5 Saanich_150   55812      12     596    7032    6478     119
## 6 Saanich_165   49362       6     178   10689    3583     113
## 7 Saanich_200    8140   41438      60     273      68   17913
```


Challenges with microbiome data

1. Big data, high complexity
2. Microbial species definition
3. Unequal sequencing coverage

Challenges with microbiome data

1. Big data, high complexity
2. Microbial species definition
3. Unequal sequencing coverage

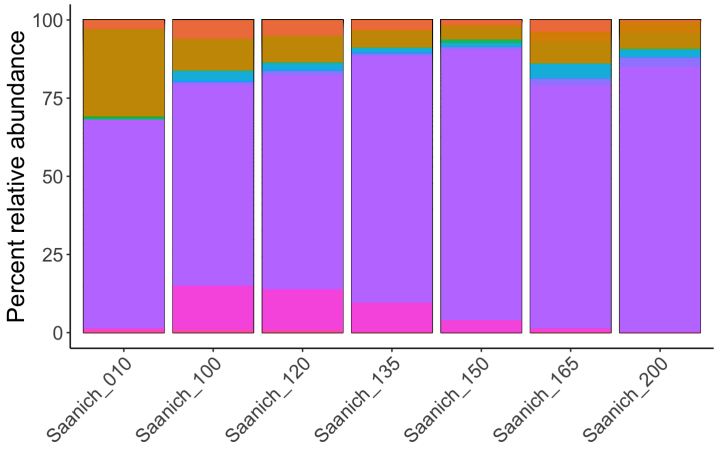


Challenges with microbiome data

1. Big data, high complexity
2. Microbial species definition
3. Unequal sequencing coverage

- **Solution: Percent relative abundance**
- Solution: Subsampling / rarefying
- Solution: Variant stabilization
- Solution: Inference of missing data

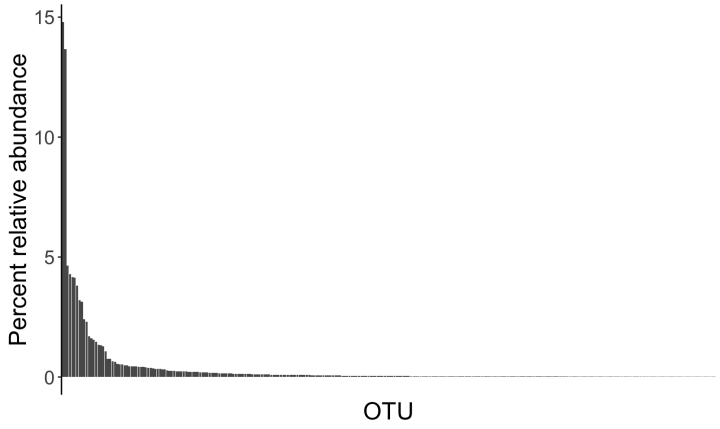
16S relative abundance OTU data



Challenges with microbiome data

1. Big data, high complexity
2. Microbial species definition
3. Unequal sequencing coverage
4. Skewed, sparse, dependent data

Skewed data



Sparse data

```
## # A tibble: 7 x 7
##   Sample      Otu0001 Otu0002 Otu0003 Otu0004 Otu0005 Otu0006
##   <chr>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Saanich_010     263       0    3210      26     108       1
## 2 Saanich_100    6489       0   18405     779    2713       6
## 3 Saanich_120   24380       0    8221    3404    5110      56
## 4 Saanich_135   39519       3    2793    5368    8216     198
## 5 Saanich_150   55812      12     596    7032    6478     119
## 6 Saanich_165   49362       6     178   10689    3583     113
## 7 Saanich_200    8140   41438      60     273      68   17913
```

- 7 samples, 3,754 OTUs
- 26,278 OTU counts
- **75% zeros**

Sparse data

```
## # A tibble: 7 x 8
##   Otu0999 Otu1000 Otu1001 Otu1002 Otu1003 Otu1004 Otu1005 Otu1006
##     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1      11       0      10       8       0       7      10      10
## 2       0       0       0       0       6       0       0       2
## 3       0       0       0       0       2       2       1       0
## 4       0       0       0       0       0       0       1       0
## 5       0       0       0       0       0       0       0       0
## 6       0       0       0       0       0       0       0       0
## 7       0      12       0       0       0       0       0       0
```

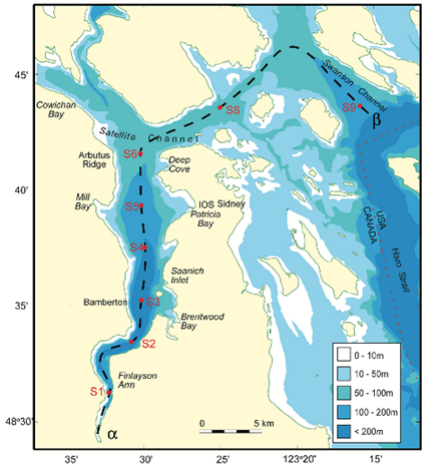
- Rely on statistical analyses
- Still searching for solutions

Pipeline recap

1. Isolate DNA
2. Amplify and sequence a piece of the bacterial 16S
3. Remove poor quality and erroneous sequences
4. Cluster sequences into OTUs (*i.e.* species)
5. Normalize to relative abundance

What next?

More on our example dataset



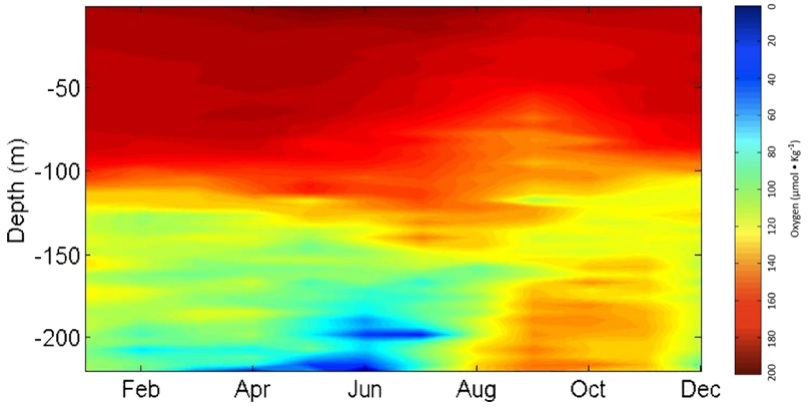
Saanich Inlet sampling

Saanich Inlet sampling

https://www.youtube.com/embed/XYQSm2Me86I

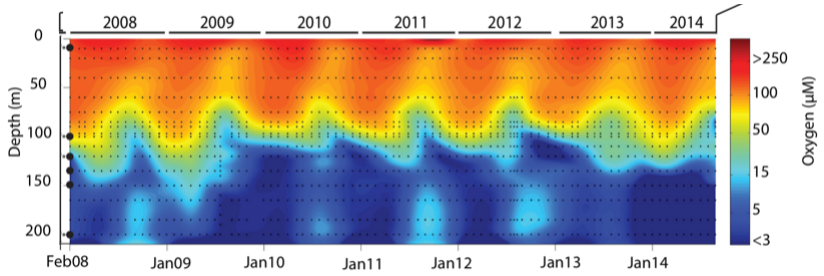
The Saanich Inlet model system

Oxygen stratification

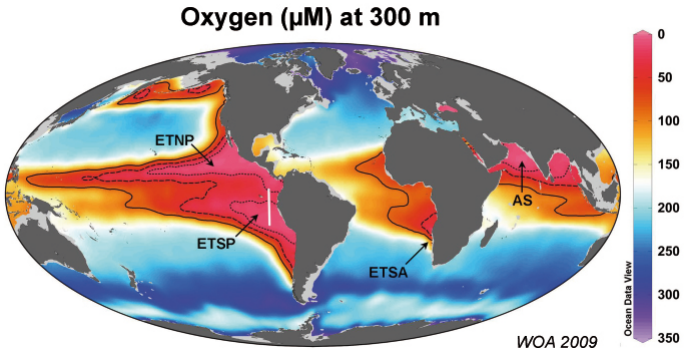


The Saanich Inlet model system

Annual cycles of stratification and renewal

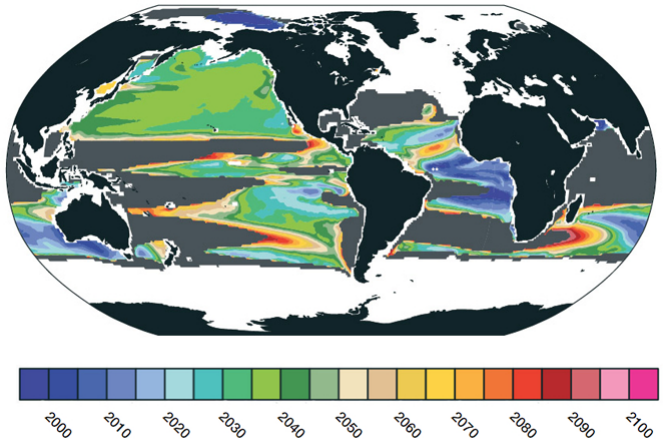


Oxygen in the ocean



Global ocean deoxygenation

Year (expected) deoxygenation will be detectable



Global ocean deoxygenation

- 20%+ reduction in O_2 over past 50 years at several monitoring stations worldwide
- Widespread deoxygenation by 2030-2040

Consequences of deoxygenation?

- Unknown how deoxygenation will impact the ocean ecosystem including:
- Global nutrient cycles
- Primary production like photosynthesis
- Fish populations
- Marine mammal survival
- …

How is global ocean deoxygenation impacting marine microbial communities?

Beta-diversity

- Between sample diversity
- Measure of overall microbiome differences

- Can take into account:
- Presence/absence of OTUs
- Relative abundance of OTUs
- Taxonomic relatedness of OTUs

Calculate beta-diversity

Bray-Curtis: Presence/absence and abundance but no taxonomy

```
library(vegan)
```

```
vegdist(dat[,-1], method="bray")
```

```
##         1       2       3       4       5       6
## 2 0.78426
## 3 0.84537 0.33392
## 4 0.85566 0.54948 0.26826
## 5 0.93583 0.73835 0.45606 0.24560
## 6 0.97132 0.82054 0.56106 0.38327 0.25888
## 7 0.99041 0.91716 0.89464 0.88614 0.88652 0.88483
```


Visualize beta-diversity

Non-metric multidimensional scaling (nMDS)

- Reduce all pairwise comparisons to distances within 2-dimensional plane

```
nMDS <- metaMDS(dat[,-1], k=2, trymax=100)
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.002299741
## Run 1 stress 0.007185395
## Run 2 stress 0.119503
## Run 3 stress 0
## ... New best solution
## ... Procrustes: rmse 0.2731915  max resid 0.4673939
## Run 4 stress 0.002299736
## Run 5 stress 0.02796809
## Run 6 stress 0.119503
## Run 7 stress 0.001523812
```


Visualize beta-diversity

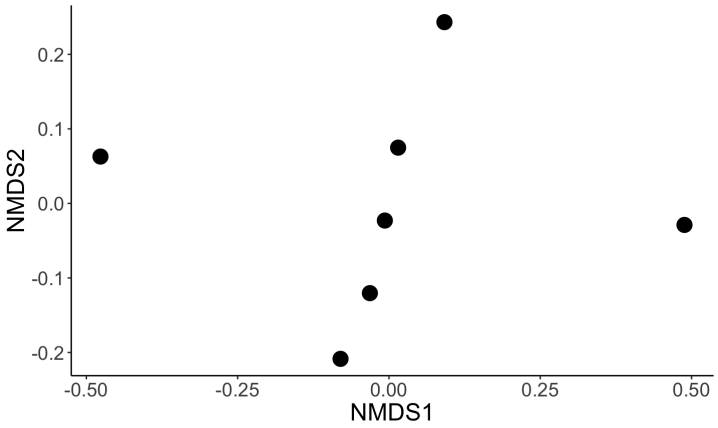
```
##
## Call:
## metaMDS(comm = dat[, -1], k = 2, trymax = 100)
##
## global Multidimensional Scaling using monoMDS
##
## Data:     wisconsin(sqrt(dat[, -1]))
## Distance: bray
##
## Dimensions: 2
## Stress:     0
## Stress type 1, weak ties
## No convergent solutions - best solution after 100 tries
## Scaling: centring, PC rotation
## Species: expanded scores based on 'wisconsin(sqrt(dat[, -1]))'
```


Visualize beta-diversity

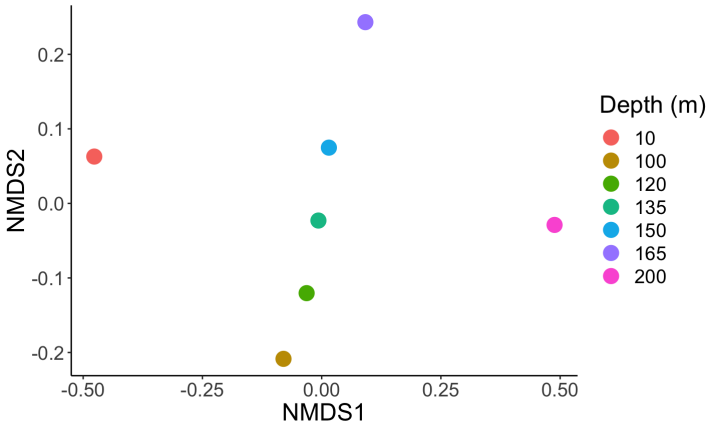
```
scores(nMDS)
```

```
##          NMDS1       NMDS2
## 1 -0.476730913  0.06285042
## 2 -0.080032659 -0.20855781
## 3 -0.031528047 -0.12039132
## 4 -0.006807545 -0.02300302
## 5  0.015144826  0.07480475
## 6  0.091644619  0.24317037
## 7  0.488309719 -0.02887339
```

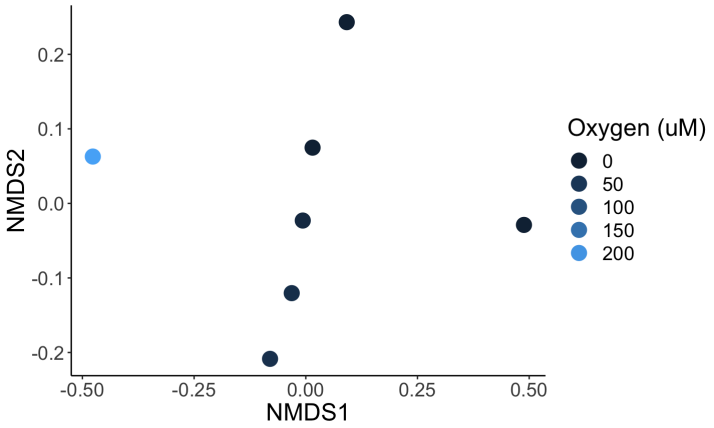
Visualize beta-diversity



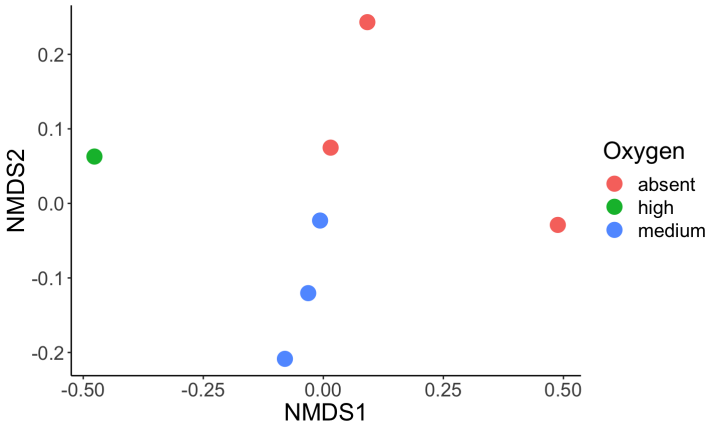
Add metadata



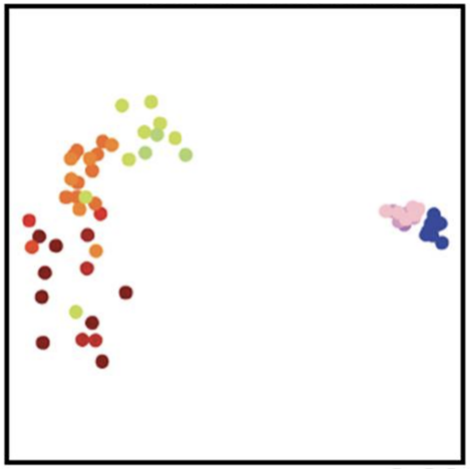
Add metadata



Add metadata



A more exciting example



Dill-McFarland *et al* 2017 Sci Rep

Drawing conclusions from nMDS

- Clear separation =/= significant
- Lack of clear separation =/= not significant

- **A sometimes useful way to visualize beta-diversity**

Statistically assess beta-diversity

Permutational ANOVA (PERMANOVA)

- ANOVA on all pairwise beta-diversity values (*not* nMDS data points)
- Same assumptions as ANOVA
- Allows complex models

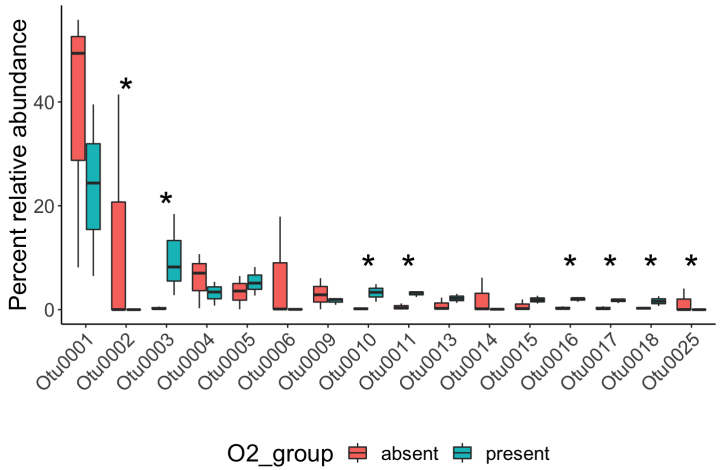
Statistically assess beta-diversity

```
adonis(dat[,-1] ~ meta$O2_uM, method="bray")
```

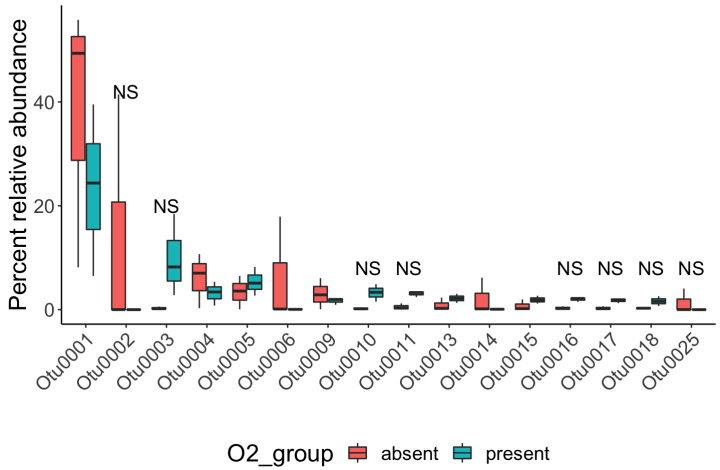
```
##
## Call:
## adonis(formula = dat[, -1] ~ meta$O2_uM, method = "bray")
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## meta$O2_uM  1    0.5727 0.57270  2.7273 0.35294  0.038 *
## Residuals   5    1.0499 0.20999         0.64706
## Total       6    1.6226                1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Which microbes are causing this difference?

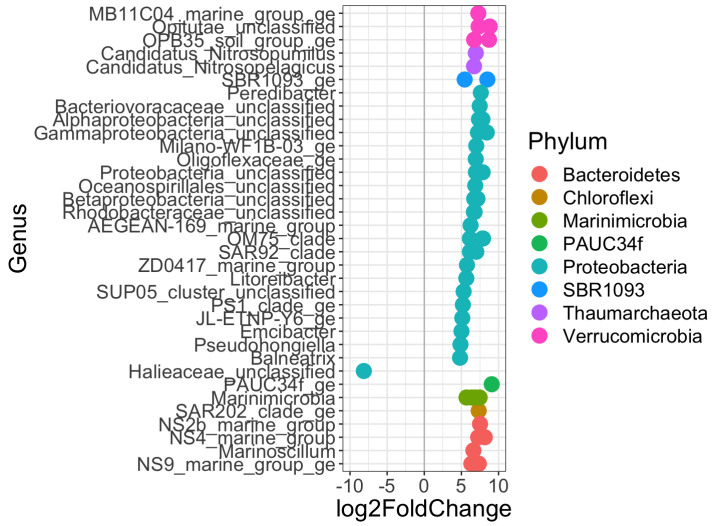
Kruskal-Wallis of abundant OTUs



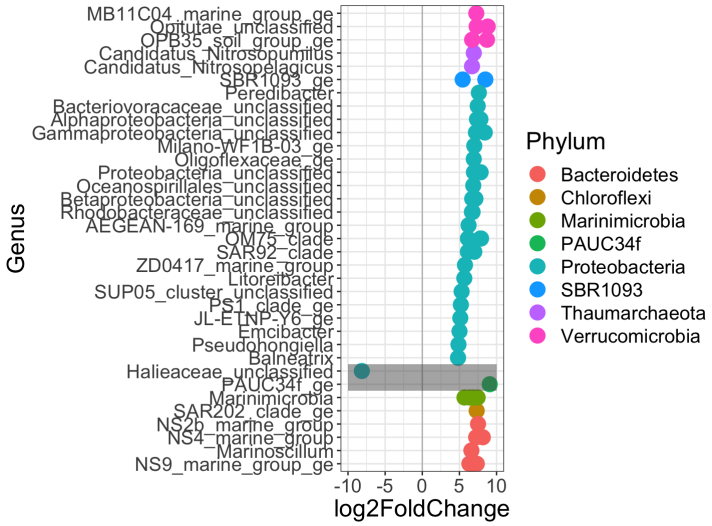
Kruskal-Wallis with FDR correction



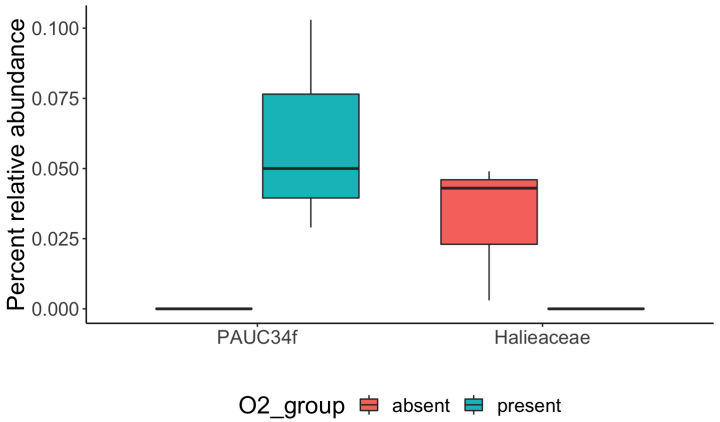
Differentially expressed OTUs (DESeq2)



Differentially expressed OTUs (DESeq2)



Further investigation



Deeper into beta-diversity

- Indicator species analysis
- Co-occurrence networks
- Machine learning models like random forest
- Etc…

On-going challenges

- More sequencing isn't always better
- What is a species?
- Normalization methods
- Biological relevance