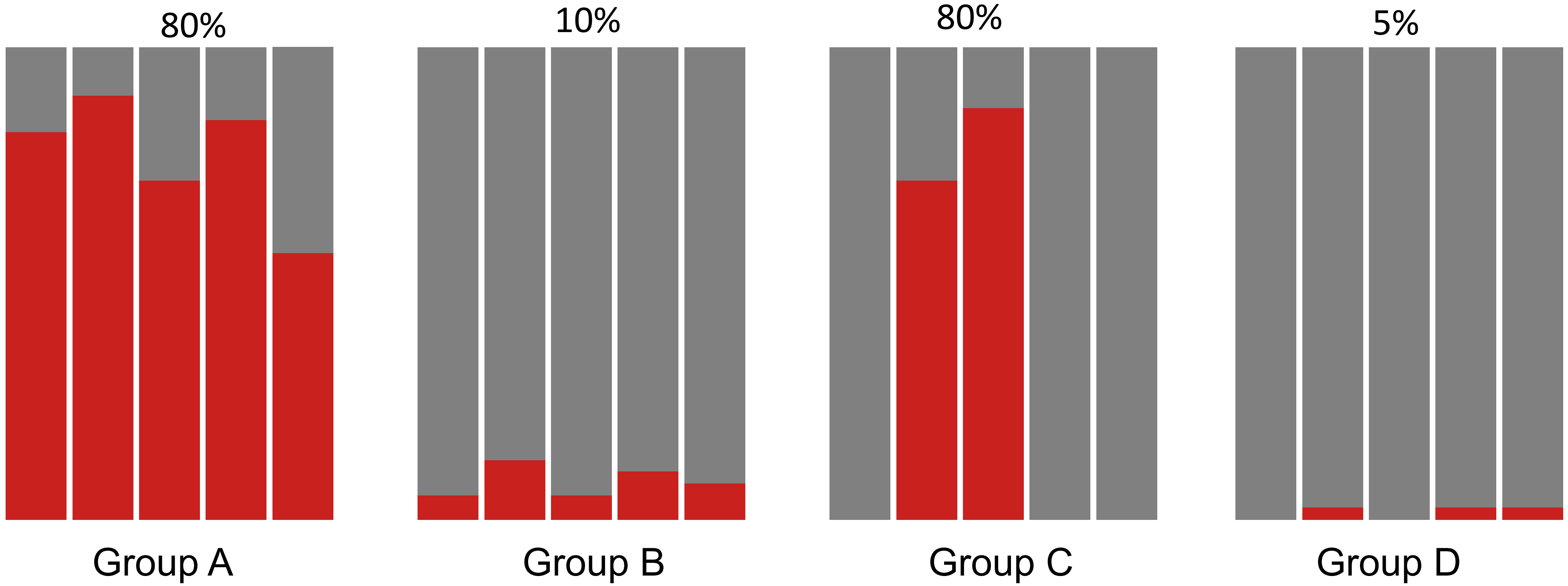
Module 16:
Identifying interesting ASVs

Module 16: Learning Outcomes

- Explain the difference between abundance and prevalence
- Identify "core microbiome" members (by specifying and justifying abundance and prevalence thresholds)
    - Create Venn Diagrams to visualize member overlap
- Explain the theory behind Indicator Species Analysis and conduct Indicator Species Analysis in R
    - Explain why regular t-tests are inadequate for identifying "indicator species/ASVs" in different treatment groups
- Explain the theory behind DESeq and conduct DESeq analysis in R
    - Interpret Volcano plots and expression bar graphs

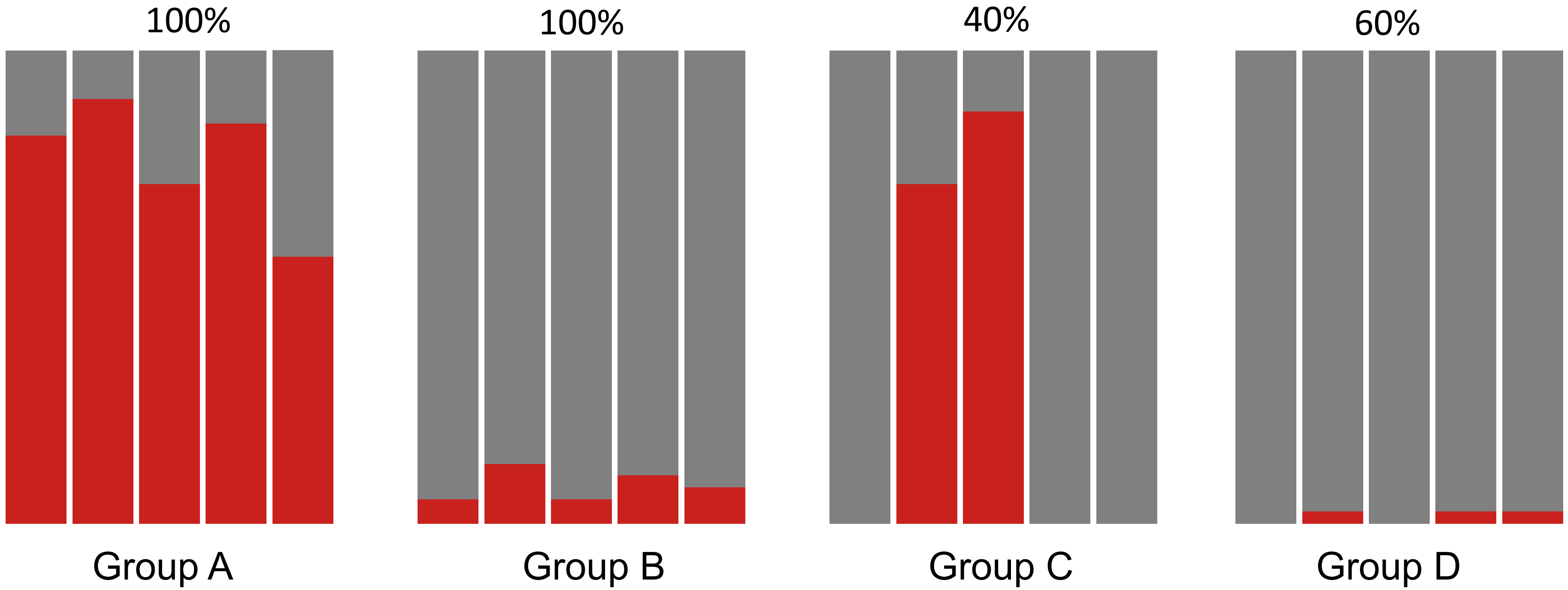
Most methods in microbial ecology use a combination of prevalence and abundance

Abundance is how much there is in a single group



Most methods in microbial ecology use a combination of <u>prevalence</u> and <u>abundance</u>

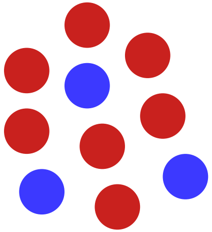
Prevalence is the frequency of presence in the samples



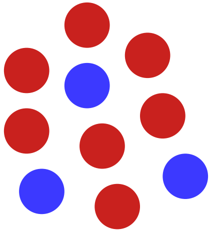
Statistical challenges with microbial data

- Changes in relative abundance of one thing will affect relative abundance of another

70% red



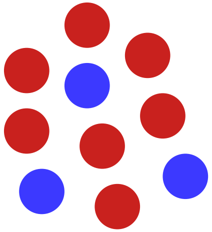
70% red



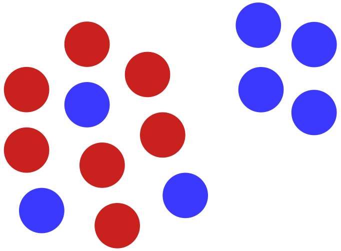
Statistical challenges with microbial data

- Changes in relative abundance of one thing will affect relative abundance of another

70% red



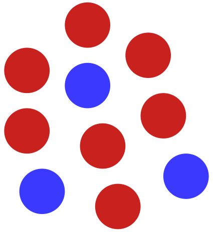
50% red



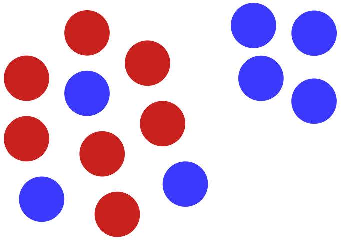
Statistical challenges with microbial data

- Changes in relative abundance of one thing will affect relative abundance of another

70% red



50% red



- Abundances are zero-inflated
  - How do you know if something is truly zero, or if we just didn’t detect it?

Three methods we will cover:

I. Set thresholds for abundance and prevalence

“Core microbiome

II. Calculate a score that incorporates both abundance and prevalence

“Indicator species analysis”

III. Fit a distribution that accounts for strange sample distributions in relative abundance, ignore zeros.

“DESeq”

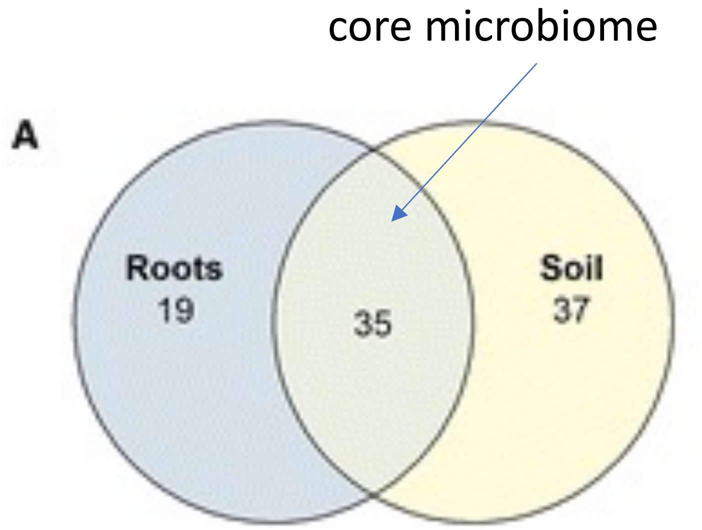
(Other methods include Lefse, ANCOM, etc)

Core Microbiome

Determine shared and unique ASVs (or taxa)

The “core microbiome”

- A set of microbial taxa (or microbial functions) that are associated with a treatment, host, or environmental condition
- Classic way of calculating core microbiome is by setting thresholds for abundance and prevalence



https://apsjournals.apsnet.org/doi/10.1094/PBIOMES-04-22-0024-R

What thresholds should I use?

- Typical abundance thresholds include:
  - 0 (presence/absence)
  - 0.001 (0.1% relative abundance filters out rare things)
  - 0.01 (1% relative abundance is considered “abundant”)
- Typical prevalence thresholds include:
  - 0 (present in at least one sample)
  - 0.5 (present in at least half of samples)
  - 0.8-0.9 (present in almost all samples)

USE YOUR BEST JUDGEMENT GIVEN YOUR OWN DATA

Core microbiome function with library(microbiome)

Usage:

```
library(microbiome)
vector_of_ASVs_treat1 <- core_members(phyloseq_treat1,
detection=0, prevalence = 0.8)
vector_of_ASVs_treat2 <- core_members(phyloseq_treat2,
detection=0, prevalence = 0.8)
```

Venn Diagrams with library(ggVennDiagram)

After creating a **vector** of microbes “associated” with each environment, you can create Venn diagrams with the VennDiagram package in R

```
ggVennDiagram( list(vec1, vec2) )
```

Indicator Taxa/Species Analysis

Determine taxa that could be predictors of a certain independent variable

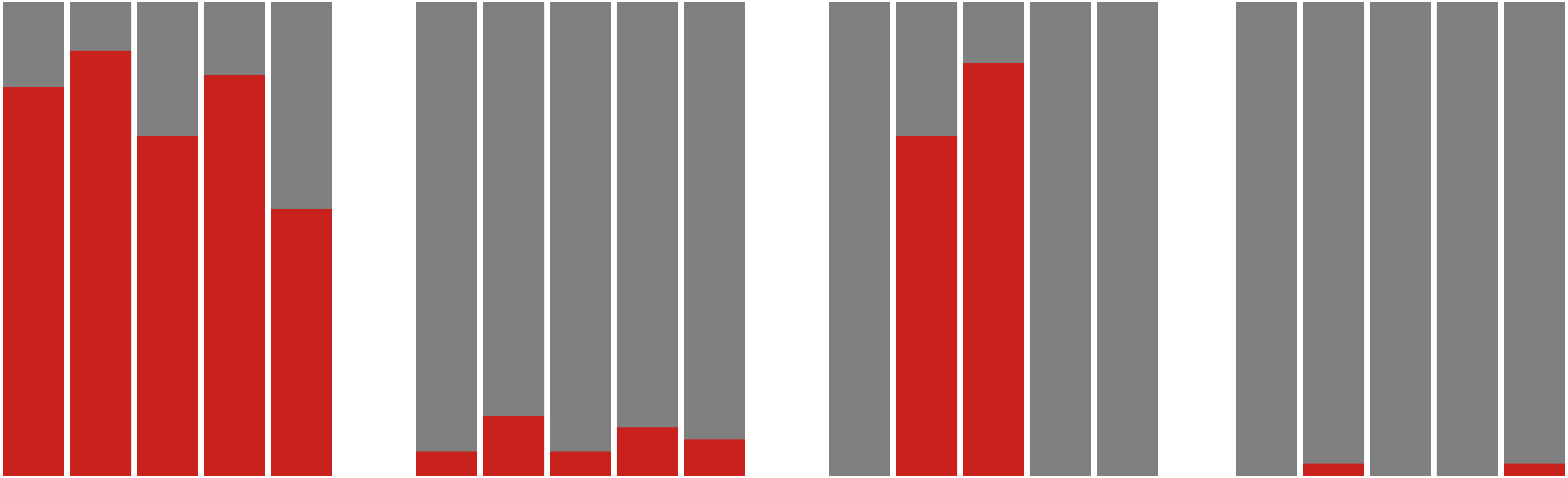
Indicator Species Analysis

- Statistical tool that uses abundance and prevalence to “score” each ASV in how associated it is with a group

- **Indicator Value: 100*RA*RK**
  - RA = relative abundance (how many individuals are in group)
  - RK = relative frequency (proportion of sites in group that have the individual)

Indicator Species Analysis

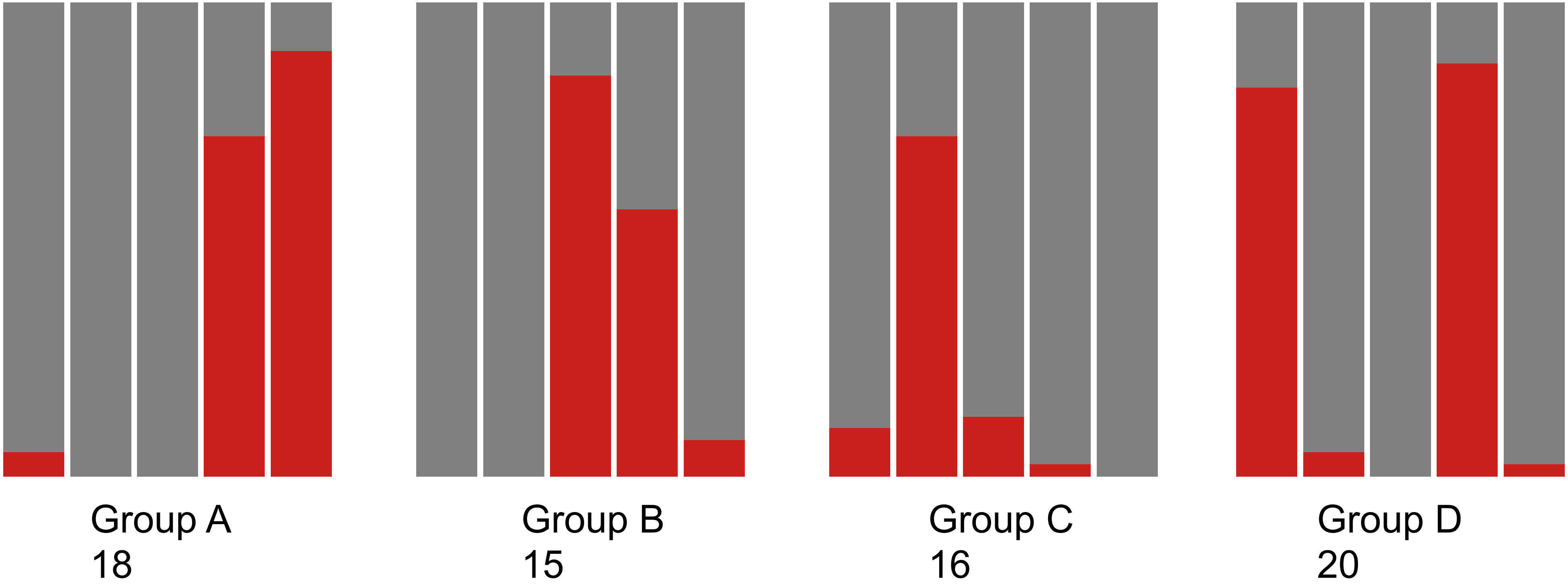
Indicator Value: 100 * RA * RK



	Group A	Group B	Group C	Group D
IVmax	100*0.8*1 = 80	100*0.1*1 = 10	100*0.8*0.4 = 32	100*0.05*0.4 = 2

Indicator Species Analysis

Compare to randomized group:



Indicator Species Analysis

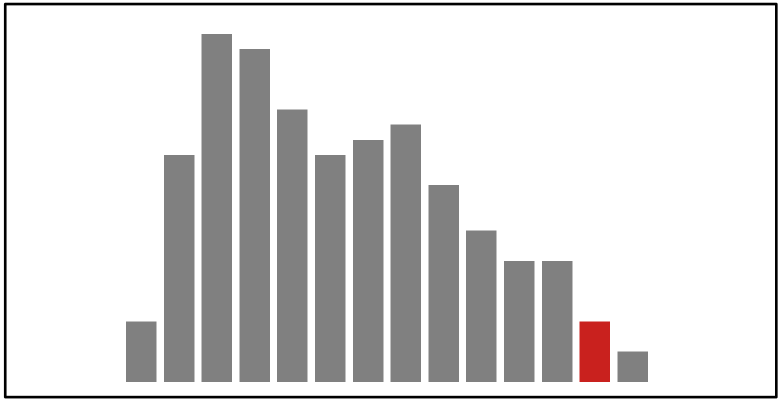
Indicator Value: 100 * RA * RK

Shuffles samples among groups

Re-calculated indicator values for each new group

Compares your “real” IV against distribution of IV values to see if it is “unusually large” given all other possibilities

ISA is therefore NON-PARAMETRIC

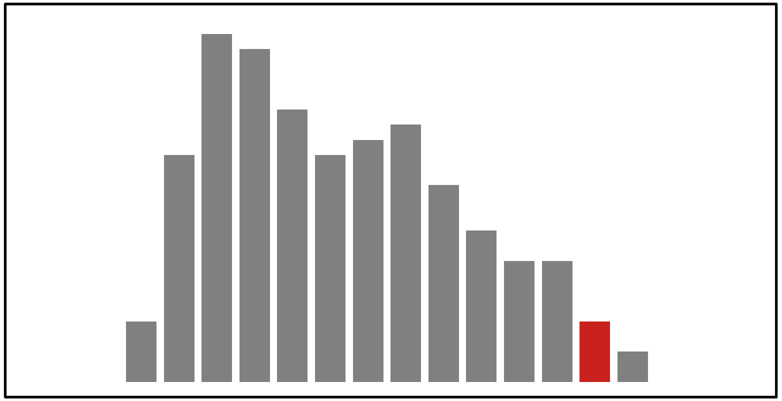


Indicator Species Analysis

Indicator Value: 100 * RA * RK

Downside:

Considered abundance and prevalence “equally important”-- is it though? Difficult to say.
library(indicspecies)

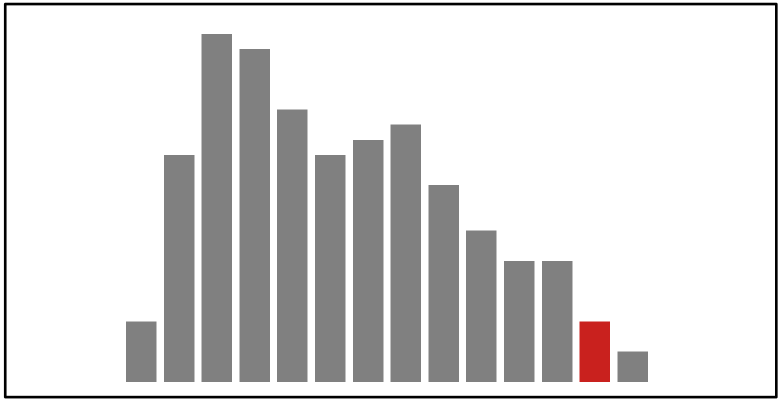


Indicator Species Analysis

Indicator Value: 100 * RA * RK

Usage:

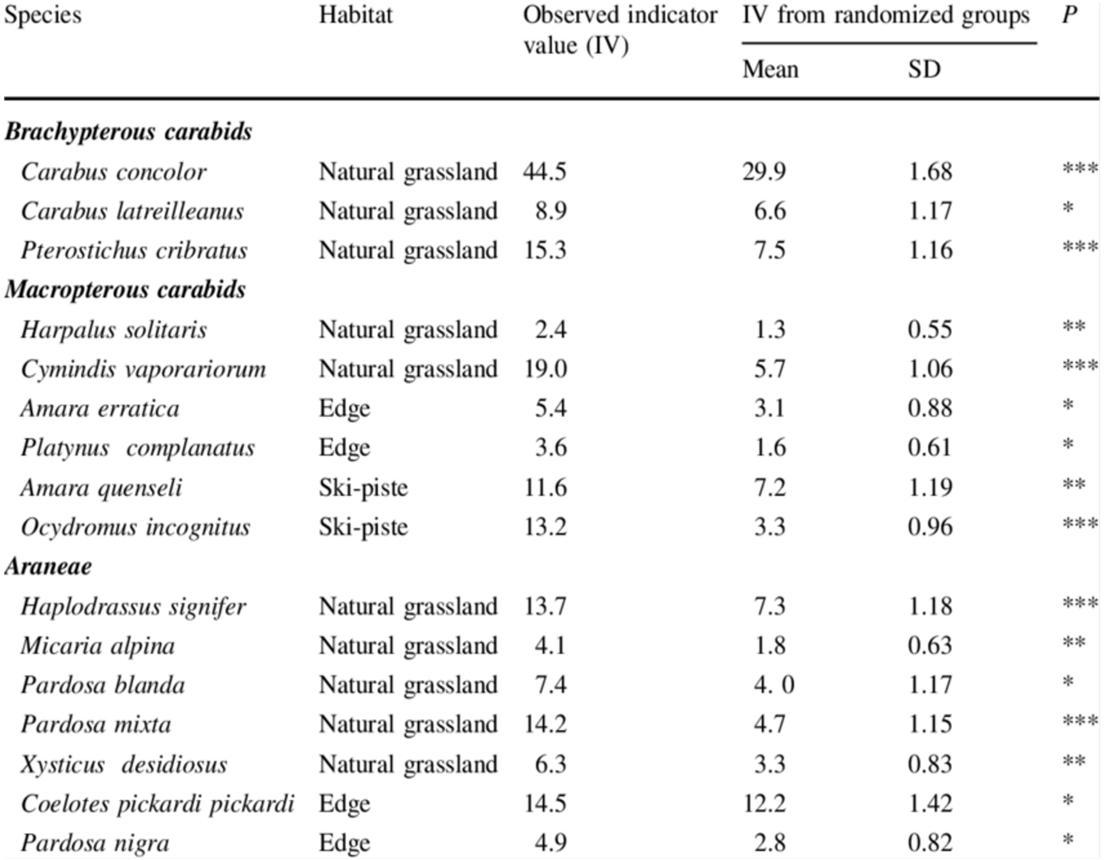
```
library(indicspecies)

multipatt( t(otu_table), cluster= vec_of_groups )
```



Indicator Species Analysis

Best way to visualize:

Table of indicator species in each group



Species	Habitat	Observed indicator value (IV)	IV from randomized groups		P
			Mean	SD	
Brachypterous carabids					
Carabus concolor	Natural grassland	44.5	29.9	1.68	***
Carabus latreilleanus	Natural grassland	8.9	6.6	1.17	*
Pterostichus cribratus	Natural grassland	15.3	7.5	1.16	***
Macropterous carabids					
Harpalus solitaris	Natural grassland	2.4	1.3	0.55	**
Cymindis vaporariorum	Natural grassland	19.0	5.7	1.06	***
Amara erratica	Edge	5.4	3.1	0.88	*
Platynus complanatus	Edge	3.6	1.6	0.61	*
Amara quenseli	Ski-piste	11.6	7.2	1.19	**
Ocydromus incognitus	Ski-piste	13.2	3.3	0.96	***
Araneae					
Haplodrassus signifer	Natural grassland	13.7	7.3	1.18	***
Micaria alpina	Natural grassland	4.1	1.8	0.63	**
Pardosa blanda	Natural grassland	7.4	4. 0	1.17	*
Pardosa mixta	Natural grassland	14.2	4.7	1.15	***
Xysticus desidiosus	Natural grassland	6.3	3.3	0.83	**
Coelotes pickardi pickardi	Edge	14.5	12.2	1.42	*
Pardosa nigra	Edge	4.9	2.8	0.82	*

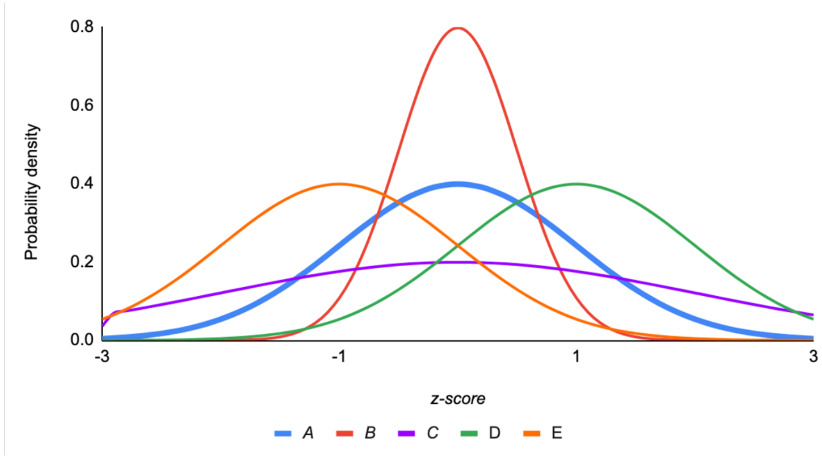
Differential abundance

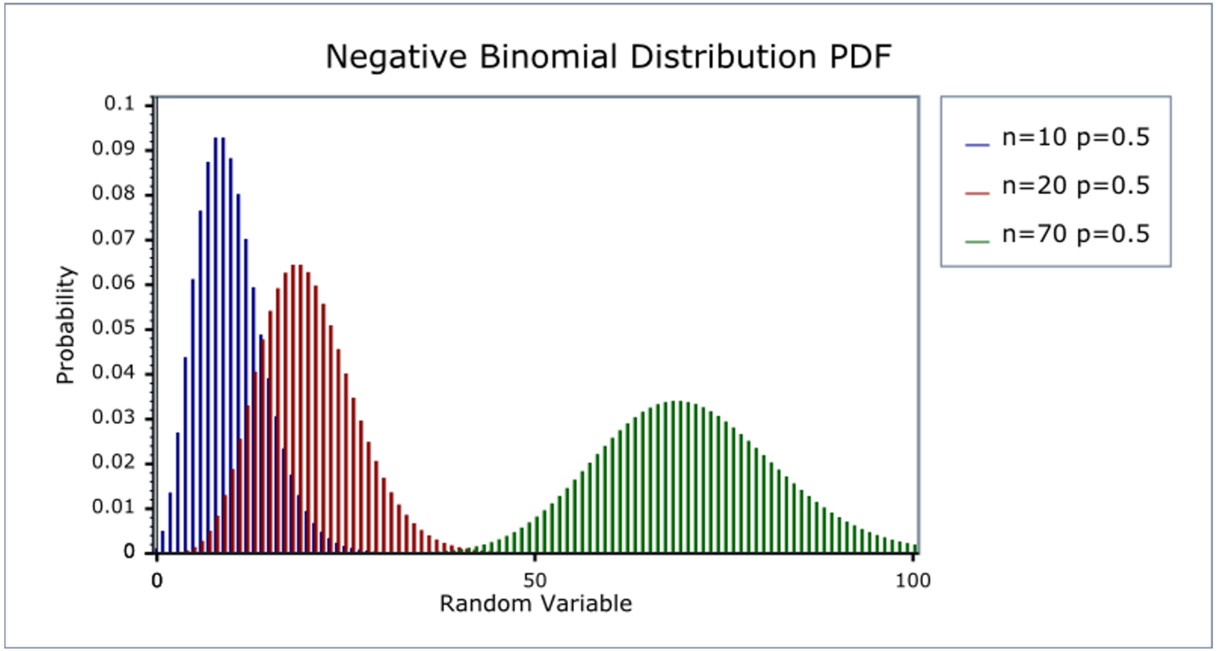
Determine ASVs (or Taxa) that have decreased or increased in presence relative to a reference

DESeq (Differential expression sequence analysis)

- Parametric test that models read counts with a negative binomial distribution

Normal distributions are not "bound" by zero like read counts are





DESeq (Differential expression sequence analysis)

- Parametric test that models read counts with a negative binomial distribution
- Powerful because it models “real” data well
- Does not handle zeros (you can add +1 to help errors)

DESeq (Differential expression sequence analysis)

Usage:

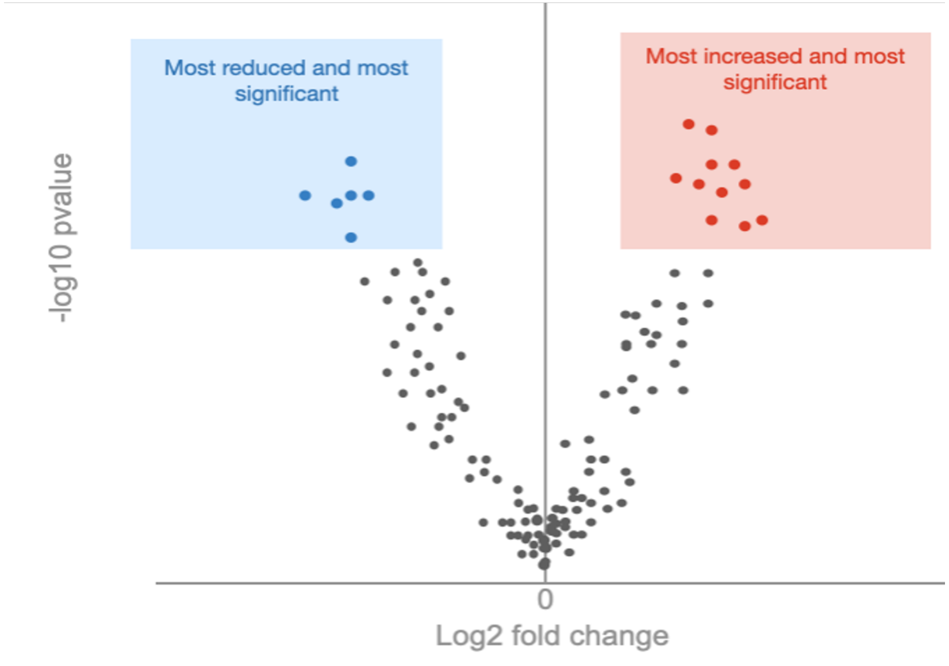
- library(DESeq2)
- Can convert phyloseq object to DESeq object using command: phyloseq_to_deseq2()
- Then run DESeq() command

Visualing DESeq results

- Two ways to visualize:

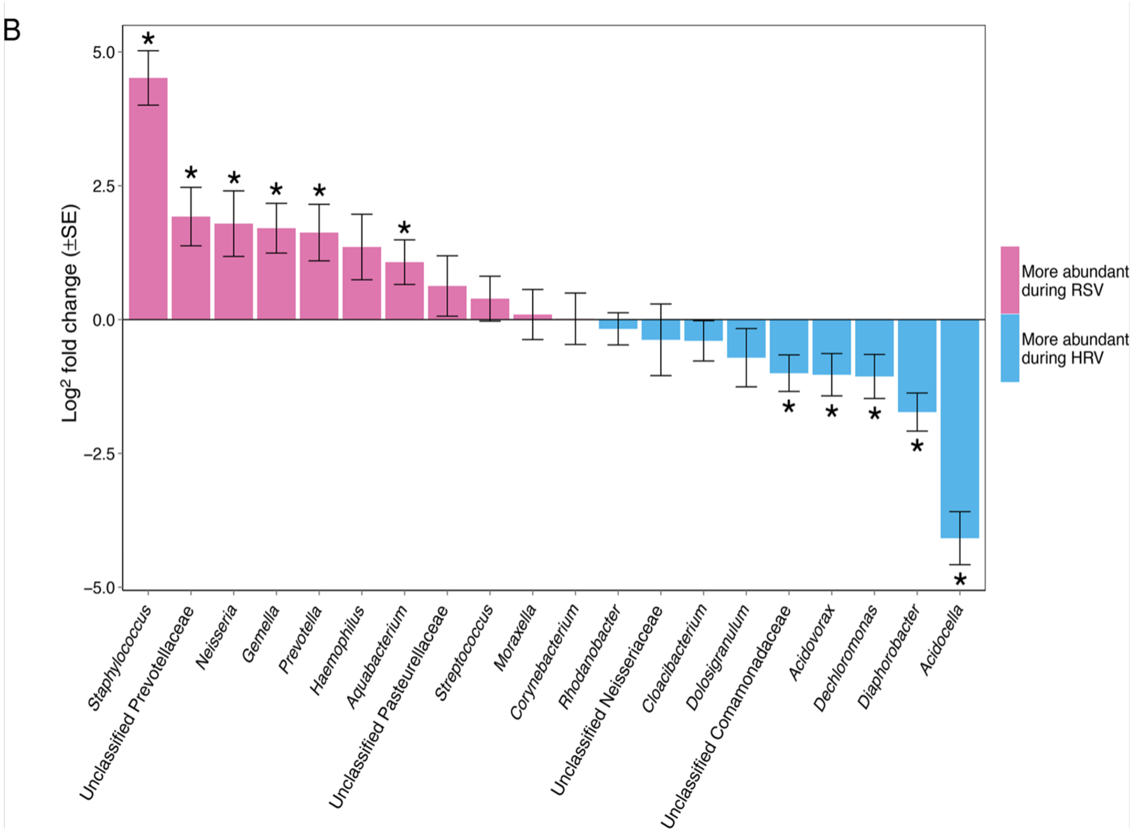
Visualing DESeq results

- Volcano plot
  - X axis is **<u>EFFECT SIZE</u>** (is there a big difference?)
  - Y axis is **<u>SIGNIFICANCE</u>** (is there a significant difference?)



Visualing DESeq results

- Bar plot
  - Show which ASVs are increased/decreased (log2 fold change) between two groups



SUMMARY

- Identifying important ASVs can be accomplished using abundance or prevalence
- Three (of many) options are:
  - Core microbiome comparisons (abundance/prevalence thresholds), visualized with Venn diagram
  - Indicator Species Analysis (combined abundance/prevalence score), visualized in table
  - DESeq2 (abundance modelling), visualized with volcano plots or bar plots