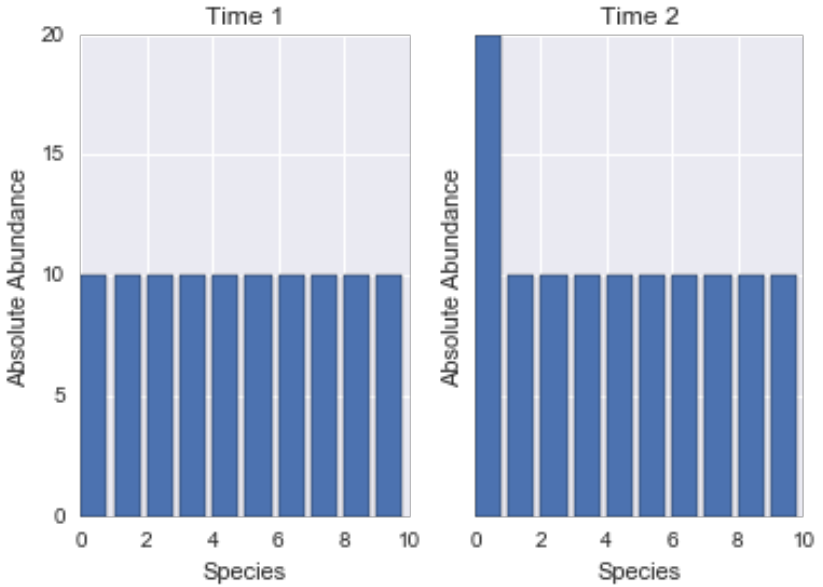
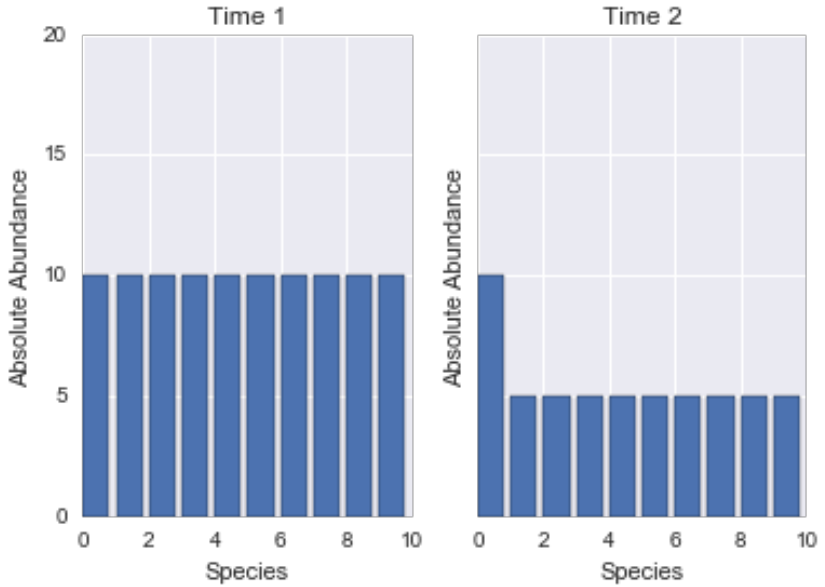
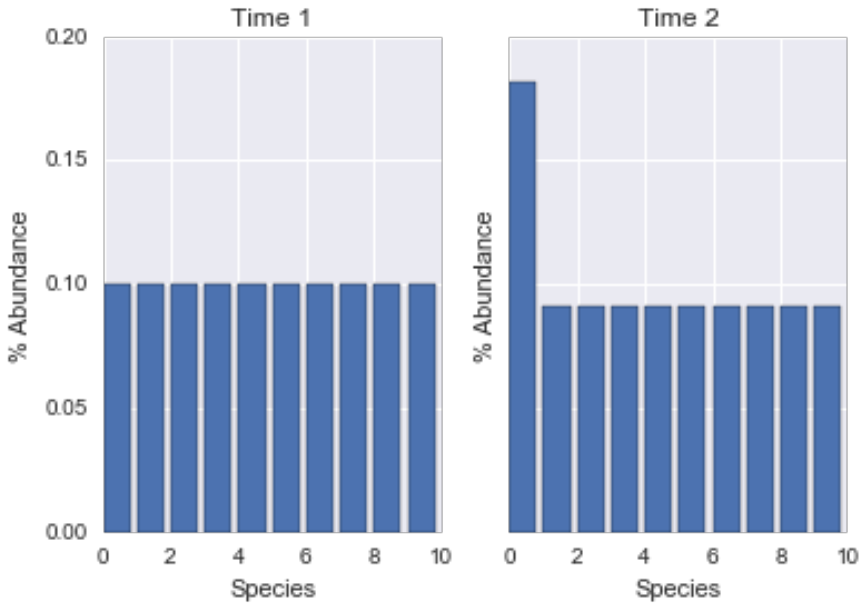
# Compositionality

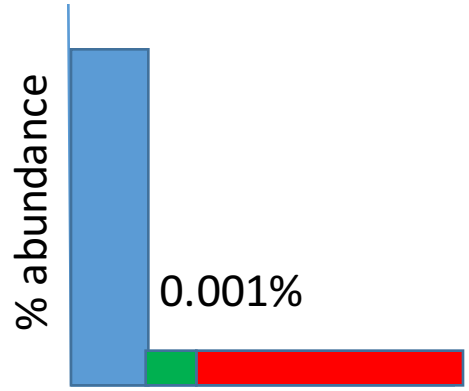
- The problem
  - We make inferences on microbial proportions
  - Proportions are unstable due dependence
    - Proportions need to sum to 1
  - Scaling issues

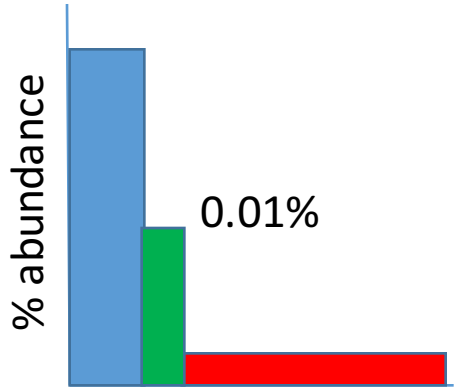

Time 1
Time 2
Absolute Abundance
Species
0
5
10
15
20
0
2
4
6
8
10


Time 1
Time 2
Absolute Abundance
Species
0
5
10
15
20
0
2
4
6
8
10


Time 1
Time 2
% Abundance
Species
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10

# Scaling




0.01% change in proportions -- Insignificant

**10x FOLD CHANGE!!!**

# Centre log ratio transform

$$clr(x) = \left[\ln\frac{x_1}{g(x)}, \ldots, \ln\frac{x_D}{g(x)}\right] = \ln x - \overline{\ln x}$$

$$g(x) = \sqrt[n]{\prod_{i=0}^{n} x_i}$$

- Log scaling
- Centers samples around mean log

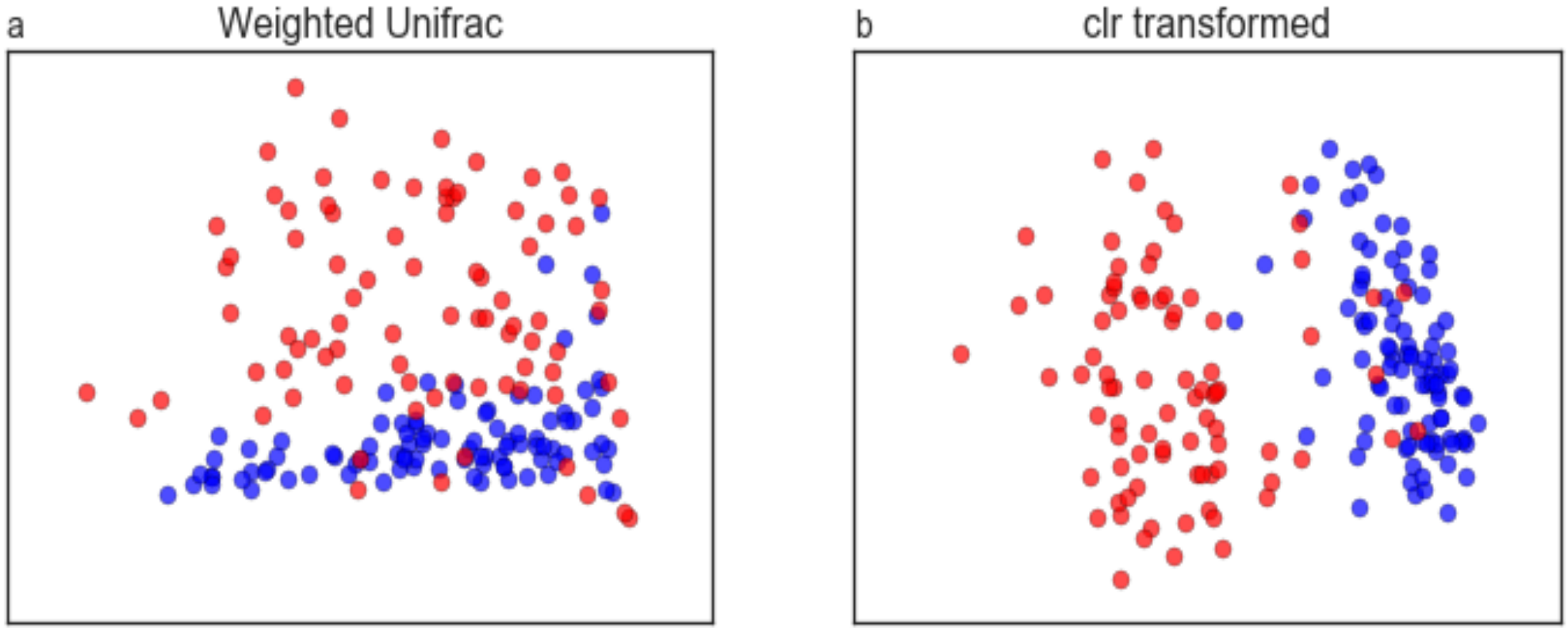
# Aitchison distance

$$d_a(x,y) = \sqrt{\frac{1}{2D}\sum_{i,j=0}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2} = d_e\big(clr(x), clr(y)\big)$$

- Isometry
  - We can apply PCA directly on clr transformed data

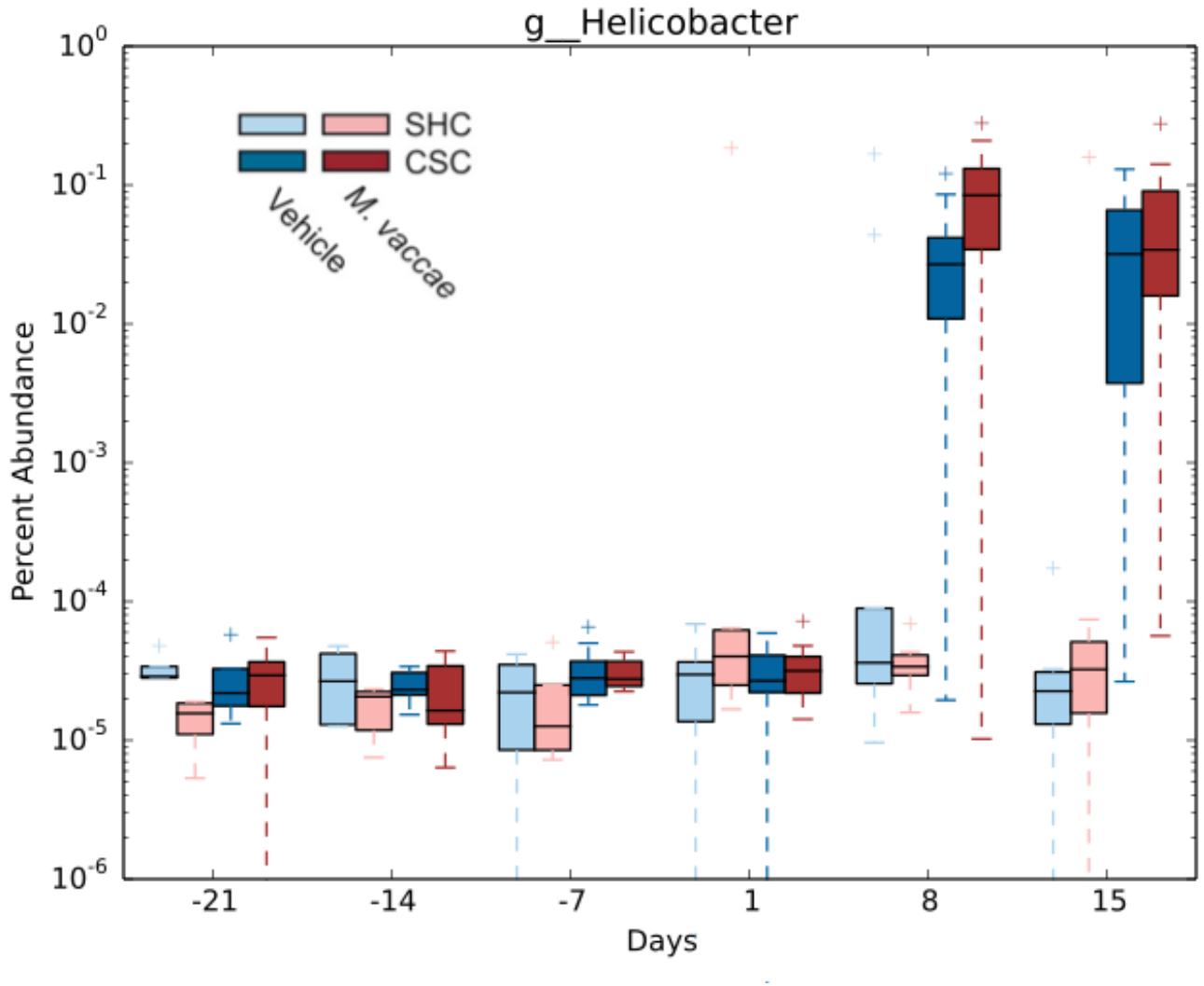
# PTSD study

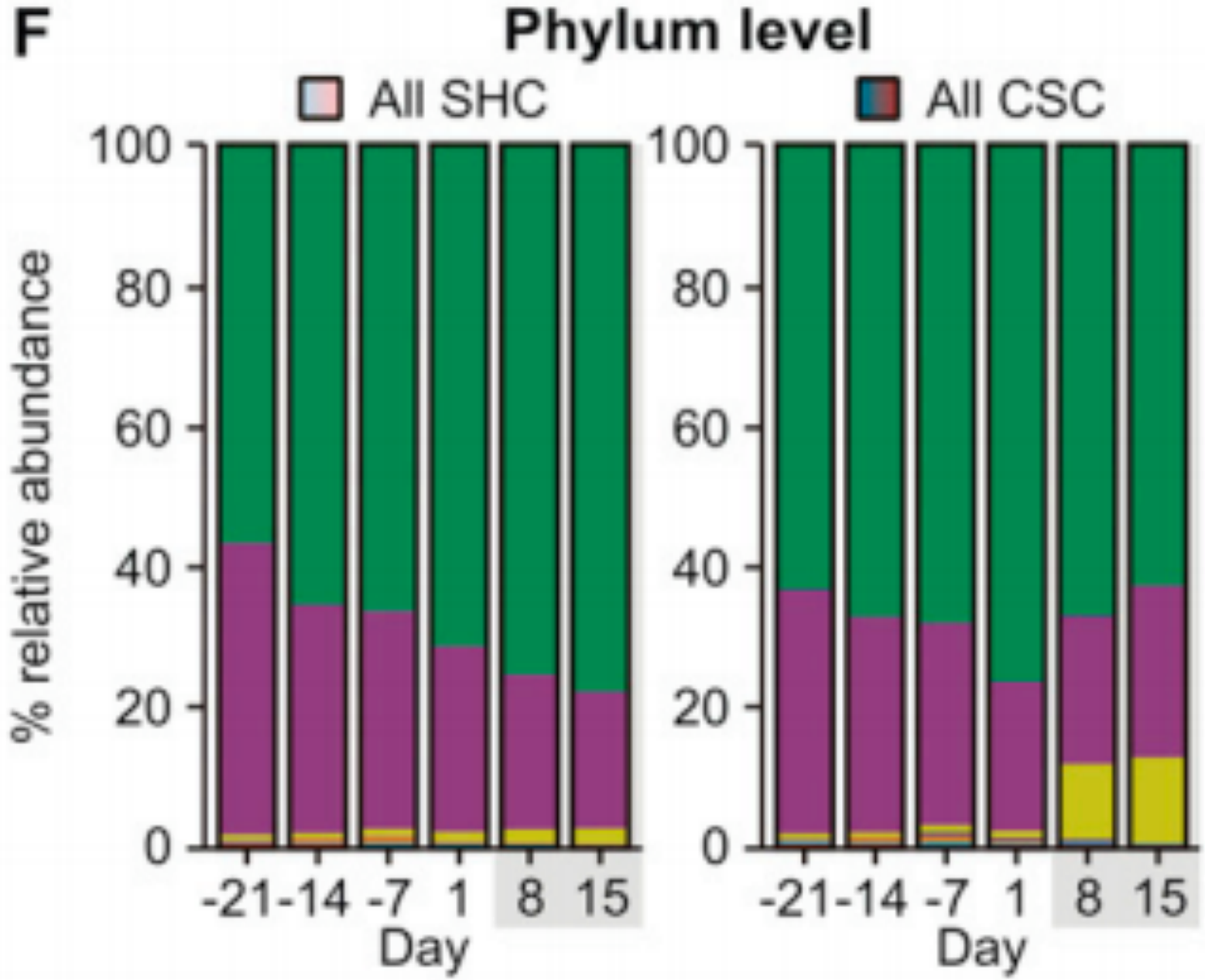

A
Experiment 1 (N = 65)
5 min EPM
+ 90 min single housing,
or 95 min single housing
(1800 - 2100 h)
4% para-
formaldehyde
perfusion (N = 30)
collect brain
s.c. M. vaccae or Vehicle
single
housing
CSC or SHC housing
Arrival
Time [days] –23 –21 –14 –7 1 8 15 19 20
rapid decapitation (N = 35)
collect brain, plasma,
pituitary, thymus,
adrenals and colon
Stool pellet collection


a
Weighted Unifrac
b
clr transformed

# What happened?

# Helicobacter grew

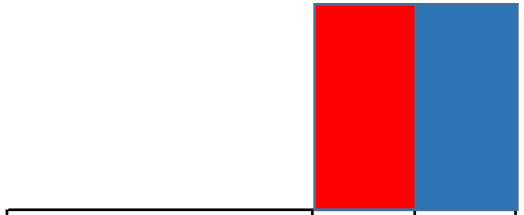



F
Phylum level
All SHC
All CSC
% relative abundance
100
80
60
40
20
0
-21
-14
-7
1
8
15
Day

# Balances

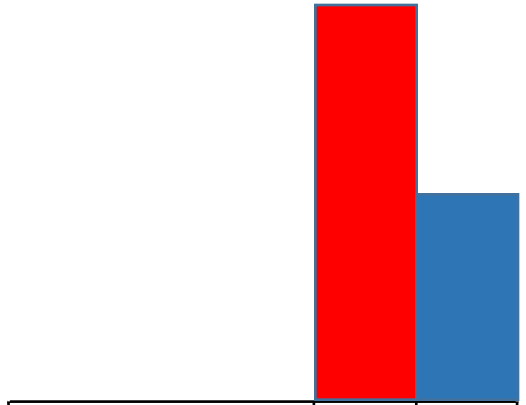
- More interpretable axes

- More correct interpretation



| Proportions | 1/2 | 1/2 |
|---|---|---|
| Counts | 100 | 100 |

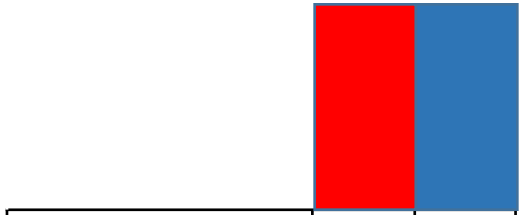
Red doubled



| Proportions | 2/3 | 1/3 |
|---|---|---|
| Counts | 200 | 100 |

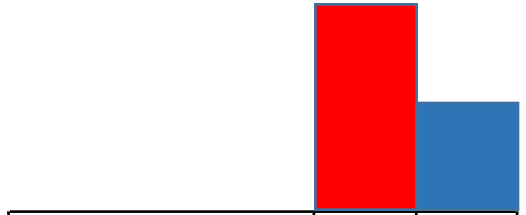
Time point 1

Time point 2



| Proportions | 1/2 | 1/2 |
|---|---|---|
| Counts | 100 | 100 |

Blue halved

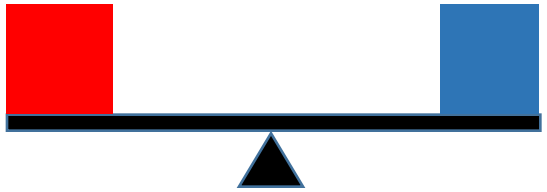


| Proportions | 2/3 | 1/3 |
|---|---|---|
| Counts | 100 | 50 |

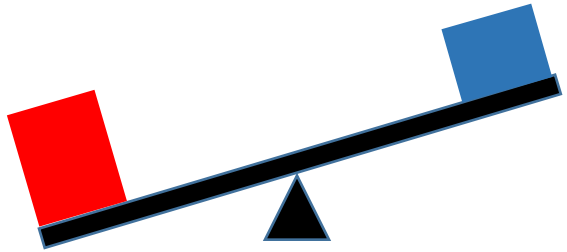
Time point 1

Time point 2

# Possible Solution? Balances



Red doubled
or
Blue halved



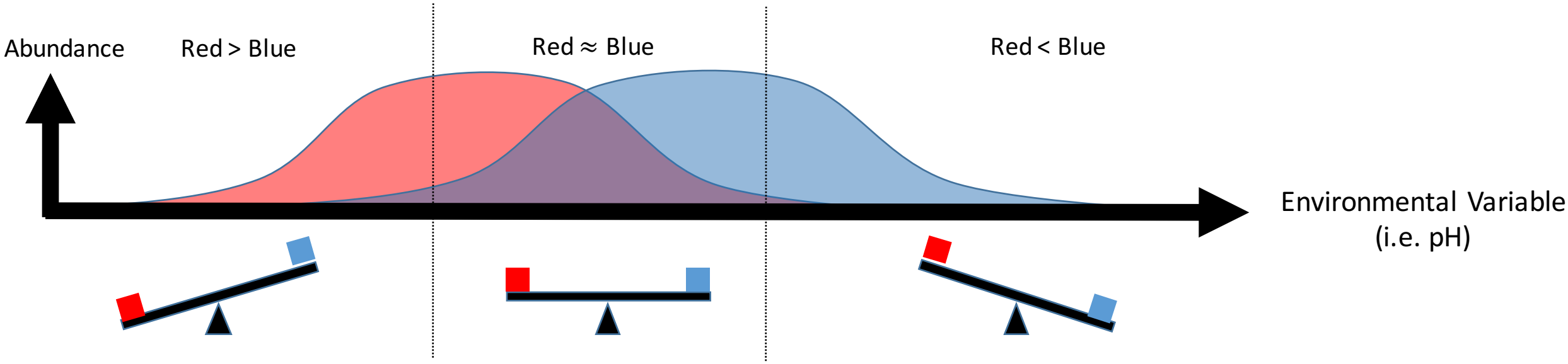
$$balance = \log\left(\frac{100}{100}\right) = 0$$

$$balance = \log\left(\frac{100}{50}\right) = \log\left(\frac{200}{100}\right) = \log\left(\frac{2/3}{1/3}\right) = \log 2$$

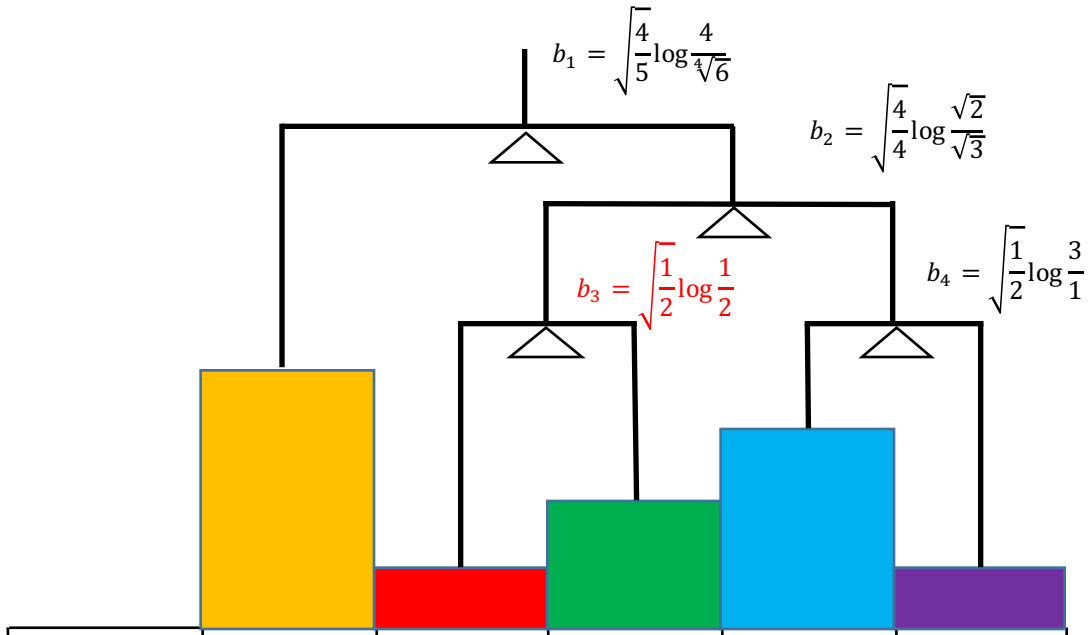
Time point 1

Time point 2

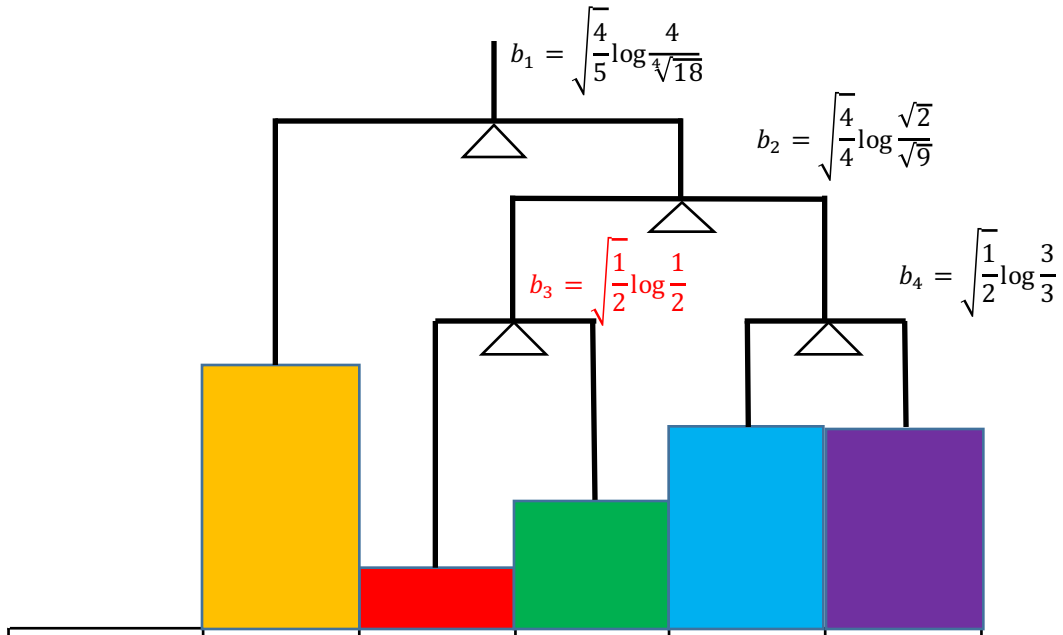
# Balances vs Environmental Variables



# Balances for multiple proportions




| Proportions | 4/11 | 1/11 | 2/11 | 3/11 | 1/11 |
|---|---|---|---|---|---|
| Percentages | 36% | 9% | 18% | 27% | 9% |
| Counts | 4 | 1 | 2 | 3 | 1 |




| Proportions | 4/13 | 1/13 | 2/13 | 3/13 | 3/13 |
|---|---|---|---|---|---|
| Percentages | 30% | 7% | 14% | 21% | 21% |
| Counts | 4 | 1 | 2 | 3 | 3 |

$$b_i = \sqrt{\frac{|i_L||i_R|}{|i_L| + |i_R|}}\log\left(\frac{g(i_L)}{g(i_R)}\right)$$

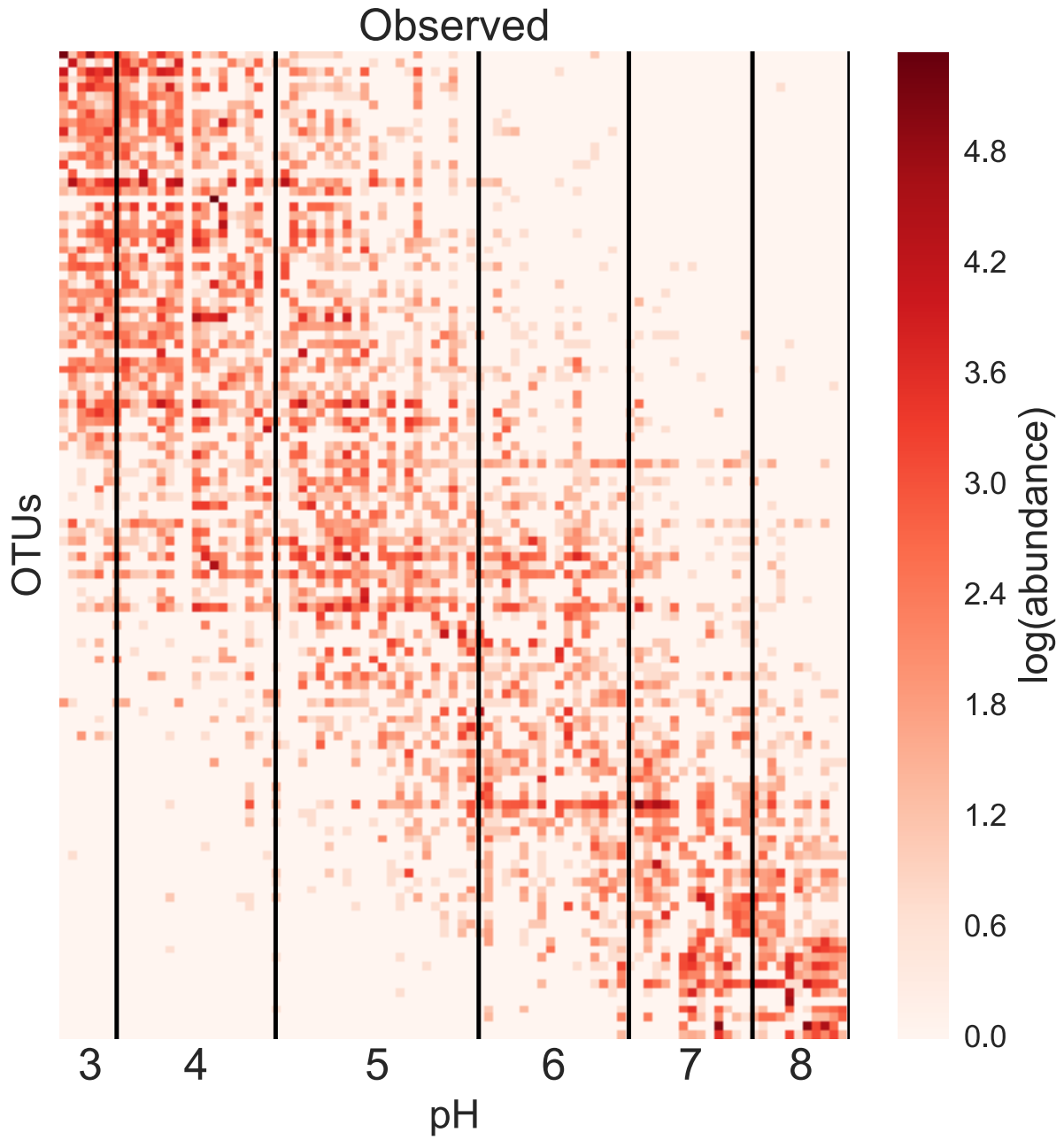
# Properties

- Isometry
  - $d_a(x, y) = d_e(clr(x), clr(y)) = d_e(ilr(x), ilr(y))$
- Scale invariance
- Independence

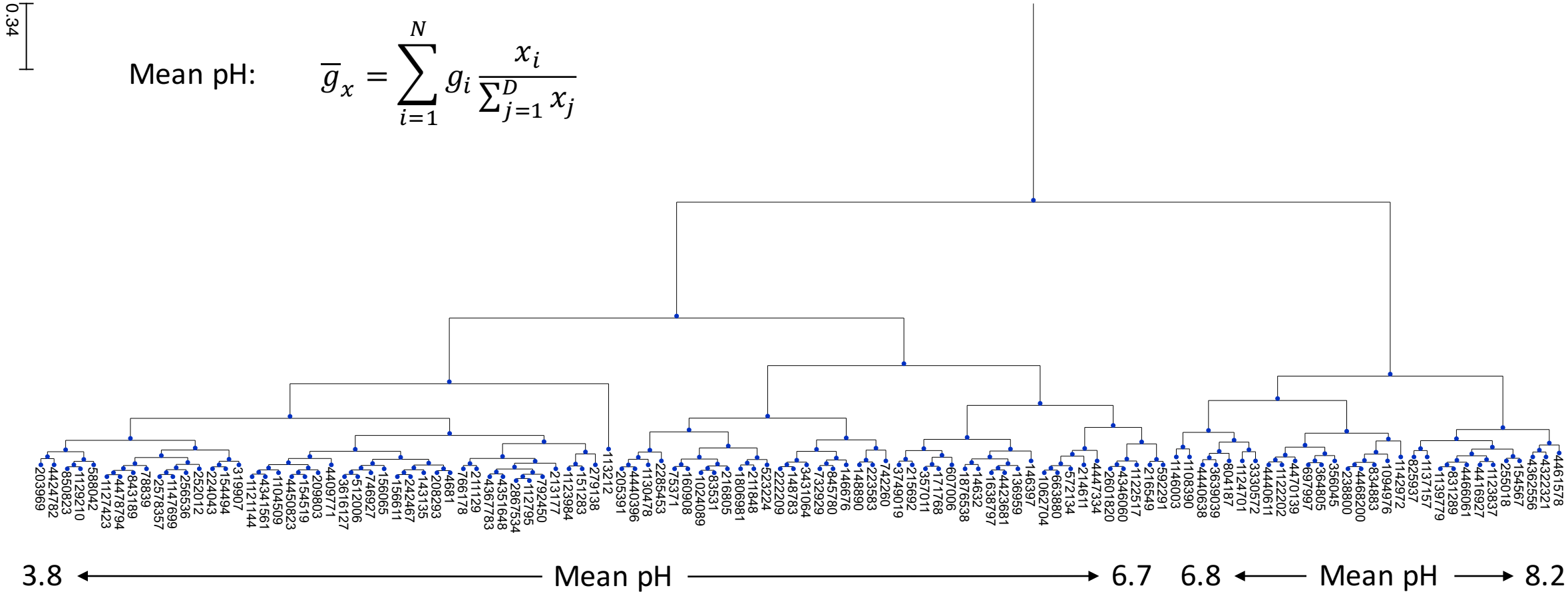
# Case Study: 88 soils

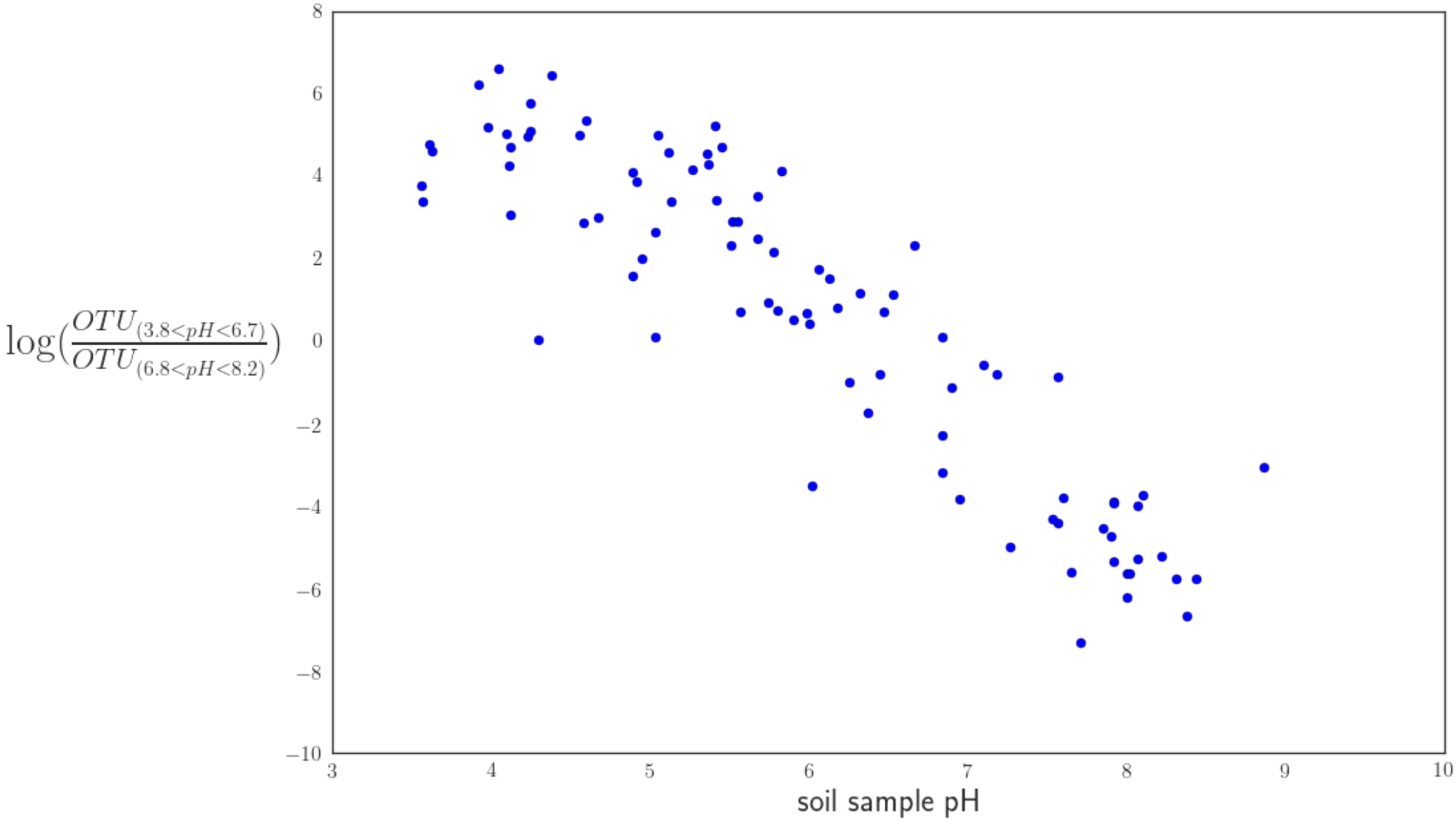
- 88 soil samples from North/South America
- Previous study found - pH changes microbial community

Mean pH: $$\overline{g}_x = \sum_{i=1}^{N} g_i \frac{x_i}{\sum_{j=1}^{D} x_j}$$

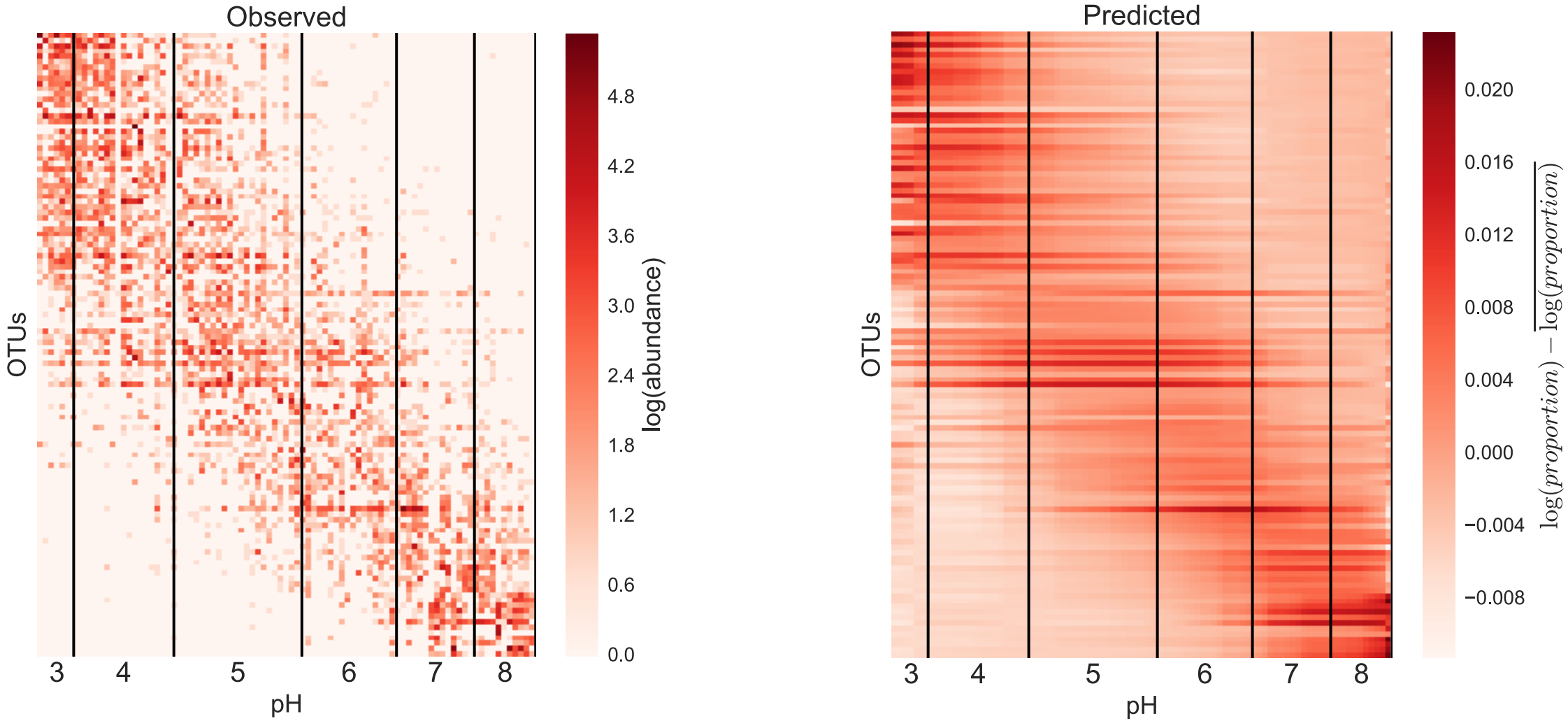


# Hierarchical clustering of OTUs




$\log(\frac{OTU_{(3.8<pH<6.7)}}{OTU_{(6.8<pH<8.2)}})$
8
6
4
2
0
−2
−4
−6
−8
−10
3
4
5
6
7
8
9
10
soil sample pH

# Linear Regression on Balances



# Conclusion

- Hypothesis : pH is a driving factor for the microbial communities
- The microbes live in a very specific pH range
- Microbial communities can be predicted using pH