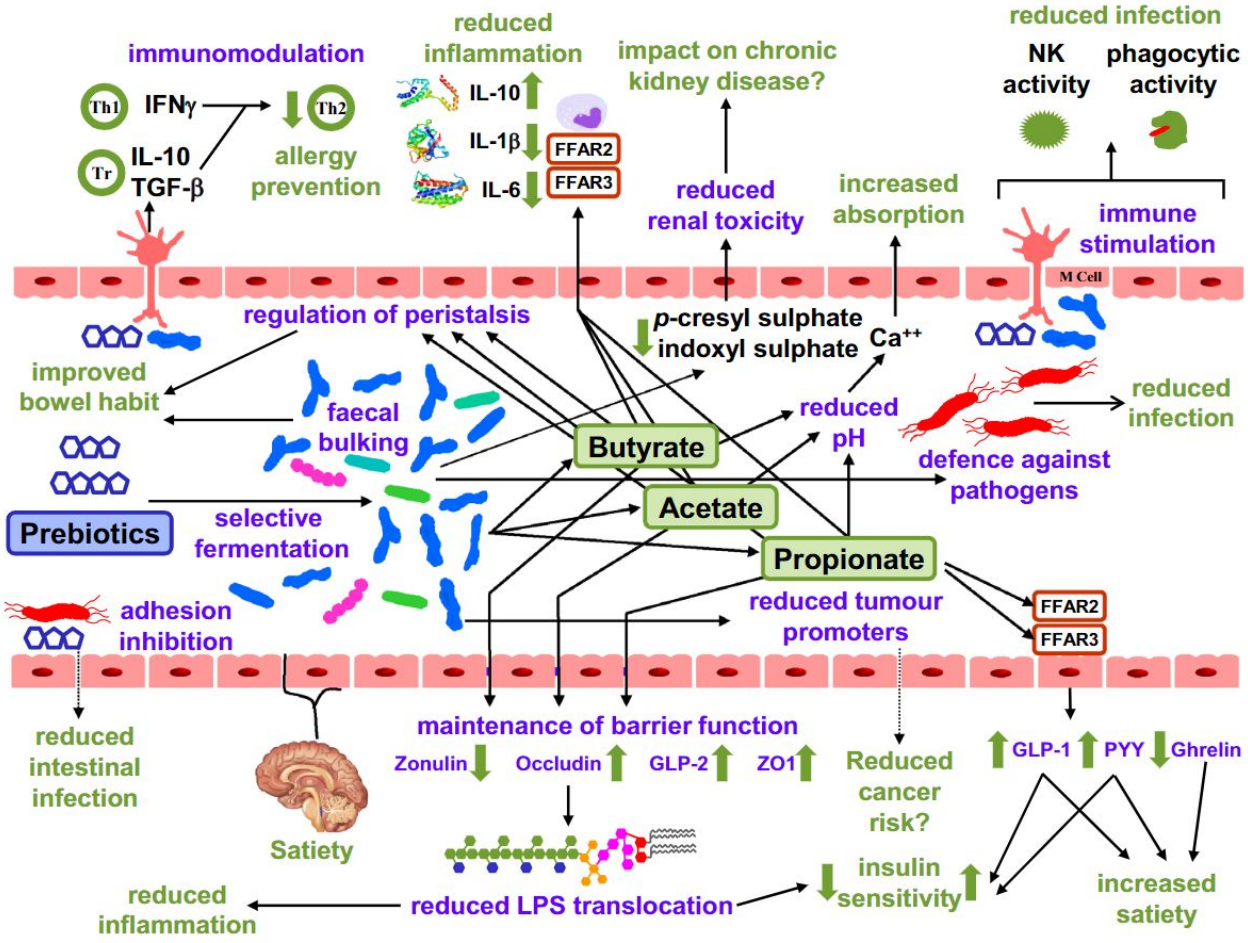
Characterizing sources of variation in bacterial sequence count data

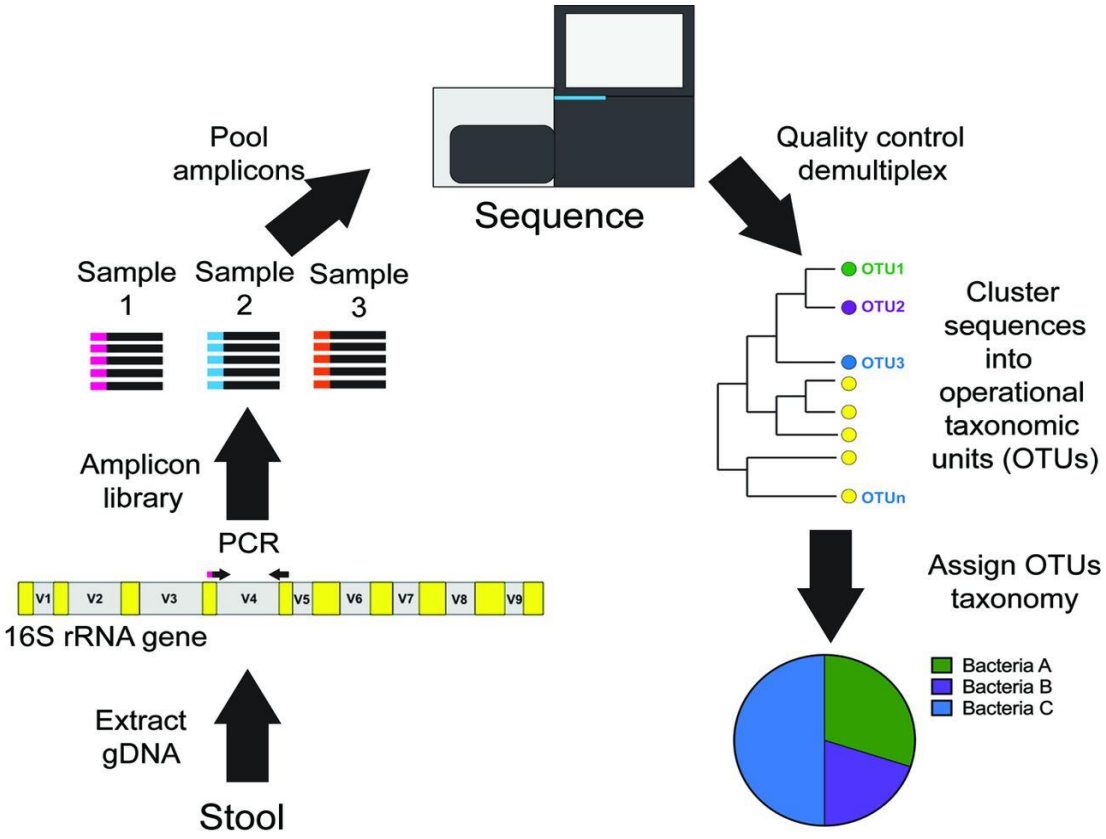
Nov. 13, 2019

The human gut microbiome plays roles in metabolism and host immune response.



Mohajeri et al. "The role of the microbiome for human health: from basic science to clinical applications." *European Journal of Nutrition* (2018)

We can profile the composition of the gut microbiome by 16S rRNA sequencing.

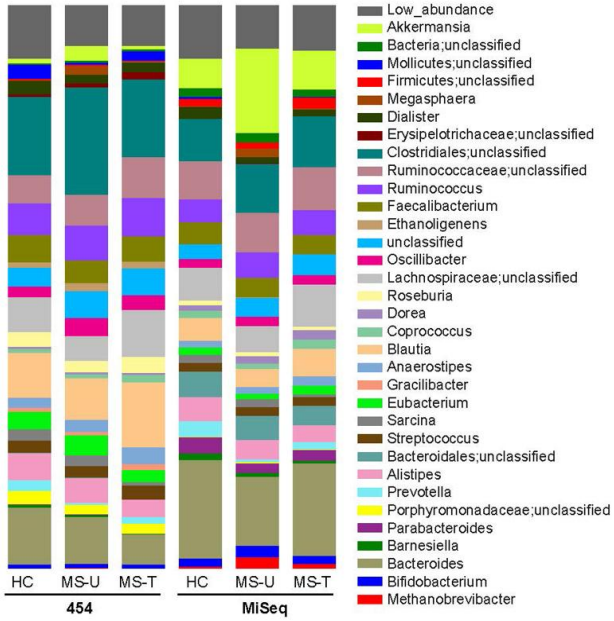


Koh, A. "Potential for Monitoring Gut Microbiota for Diagnosing Infections and Graft-versus-Host Disease in Cancer and Stem Cell Transplant Patients." *Clinical Chemistry* (2017).

The data are counts...

	Phylum 1	Phylum 2	Phylum 3	...	Phylum 30	Phylum 31	Phylum 32	...	Phylum 48	Phylum 49	Phylum 50
Sample 1	9351	18740	2548		14	47	0		29	15	0
Sample 2	28210	5029	449		0	9	106		70	0	0
Sample 3	31591	10762	559		40	191	13		0	0	0
Sample 4	6542	15030	4259		4	20	0		17	0	0
Sample 5	8333	4367	1157		7	435	208		9	0	0
Sample 6	12278	1426	379		0	16	108		77	0	0
Sample 7	7759	311	144		147	245	55		0	0	0
Sample 8	20609	56	4		101	89	13		0	0	0
Sample 9	28193	19442	321		0	5	50		0	0	0
Sample 10	16120	6318	257		59	28	90		5	0	0
Sample 11	552	98	951		147	0	9		52	0	55
Sample 12	1020	13553	1905		23	239	0		6	63	0
Sample 13	3163	610	1646		631	31	8		0	0	0
Sample 14	30054	909	55		39	1049	782		0	0	0
Sample 15	4122	94	263		38	0	102		14	0	0
Sample 16	12655	9327	385		3	105	0		0	0	0
Sample 17	26461	152	4		0	0	0		0	0	0
Sample 18	7707	7102	332		3	37	7		10	39	0
Sample 19	7181	7701	183		634	14	91		156	0	0
Sample 20	17341	3686	115		14	108	23		0	0	0

...that transmit proportional information.

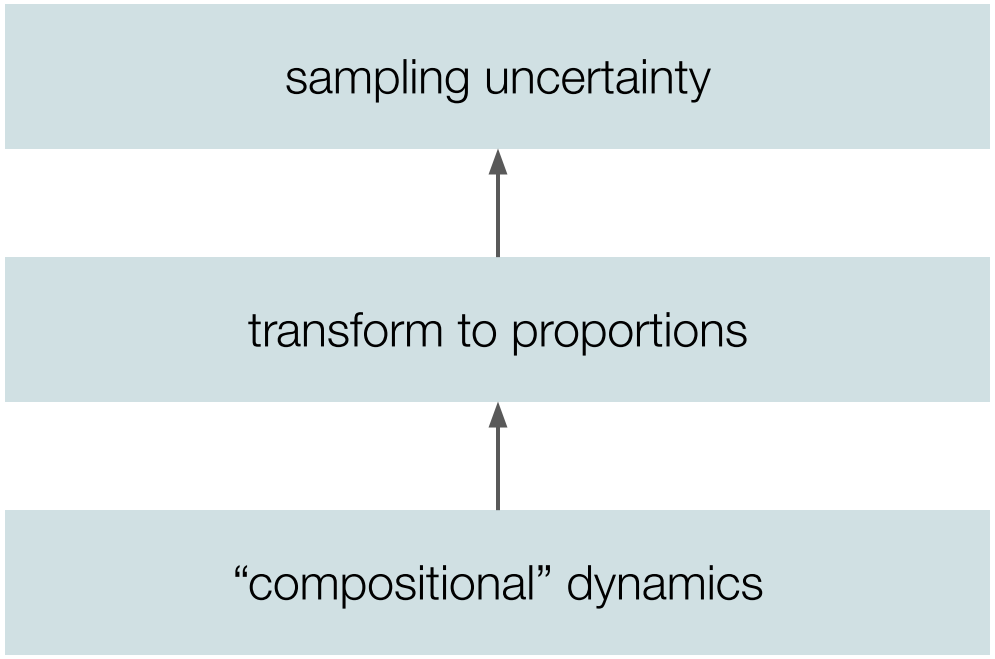


HC healthy control
MS-U multiple sclerosis (untreated)
MS-T multiple sclerosis (treated)

Jangi et al. "Alterations of the human gut microbiome in multiple sclerosis." Nature Communications (2016)

We're interested in biological and technical variation between samples.

Count and proportionality considerations motivate a multinomial logistic normal model.



Aims

A1 Performing batch effect correction in microbial sequence count data

A2 Characterizing population variation in gut microbial and functional “dynamics”

A3 Testing association of dynamics and host fitness

Aims

A1 Performing batch effect correction in microbial sequence count data

A2 Characterizing population variation in gut microbial and functional “dynamics”

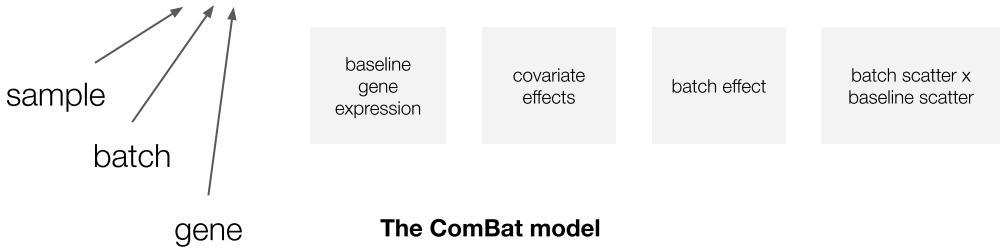
A3 Testing association of dynamics and host fitness

16S batch effect correction methods repurpose approaches from microarray and RNA-seq data.

	Compositional representation?	Model sampling uncertainty?	Model batch?
Linear models			
limma (Ritchie et al, 2015)	✓		✓
ComBat (Leek and Storey, 2007)	✓		✓
SVA (Leek et al., 2012)	✓		✓
Poisson/NB GLMs			
edgeR (Robinson et al., 2010)		✓	✓
DESeq2 (Love et al., 2014)		✓	✓
Differential expression in the microbiome			
ALDEx2 (Fernandes et al., 2014)	✓	✓	
BDMMA (Dai et al., 2018)	✓	✓	✓
McLaren et al., 2019	✓		✓
Non-parametric			
Percentile normalization (Gibbons et al., 2016)			✓

1 Enhance a classic approach by modeling sampling uncertainty & framing correction within compositional data POV

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$



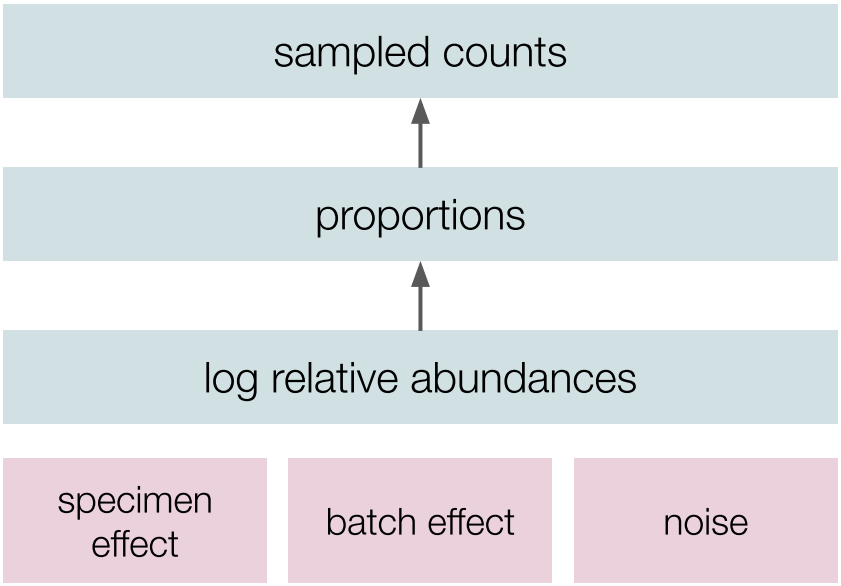
The ComBat model

Johnson, et al. "Adjusting batch effects in microarray expression using empirical Bayes methods." Biostatistics (2007)

2 Introduce a measure for quantifing goodness-of-correction

3 Evaluate performance on simulated and real data sets

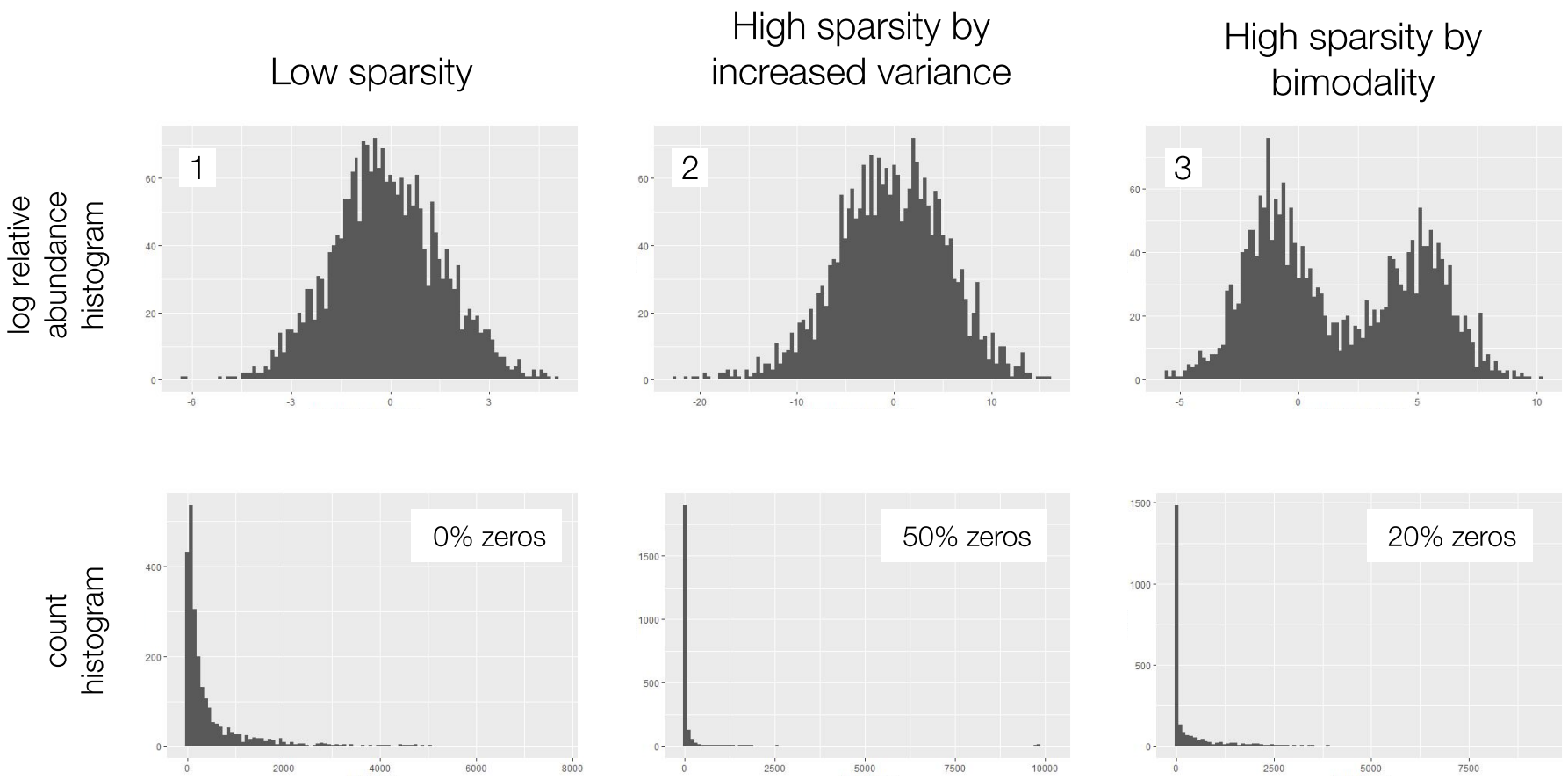
A simulation model in outline



A relative scoring scheme

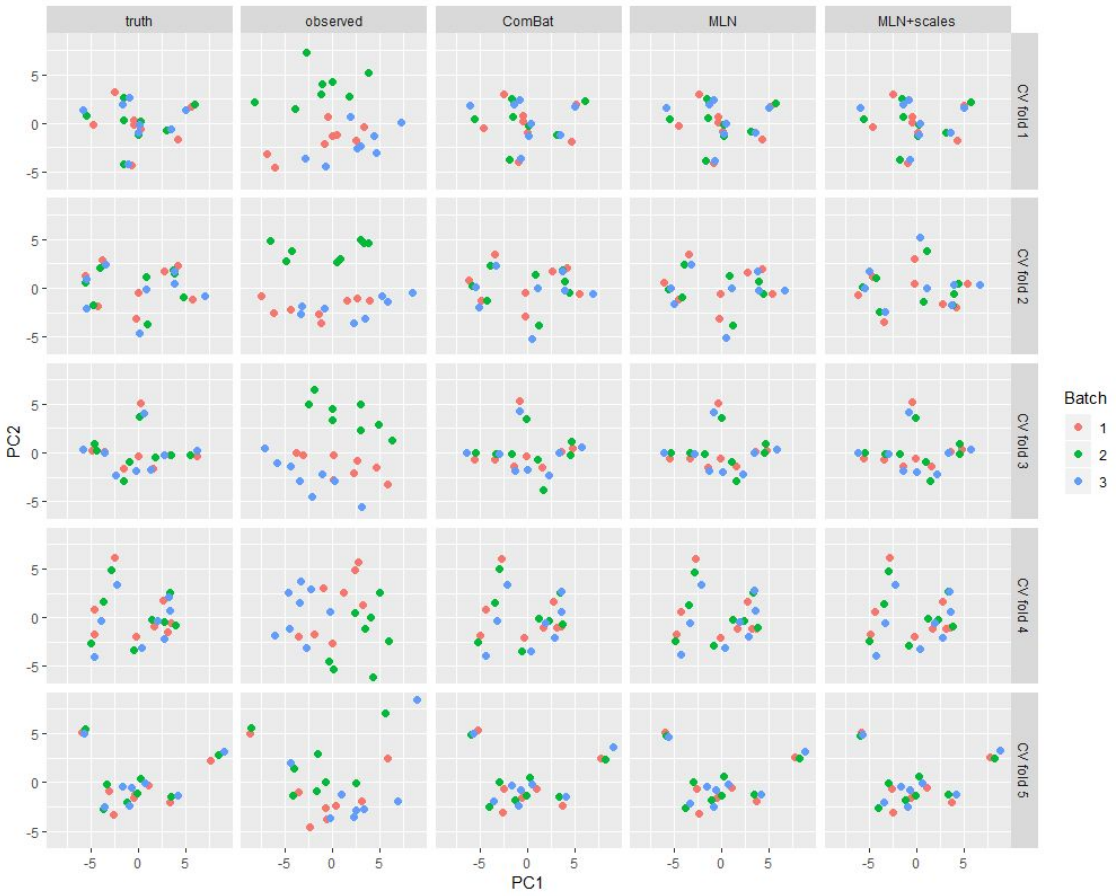
$$\text{score} = \frac{\text{average distance between batches}}{\text{average distance between specimens}}$$

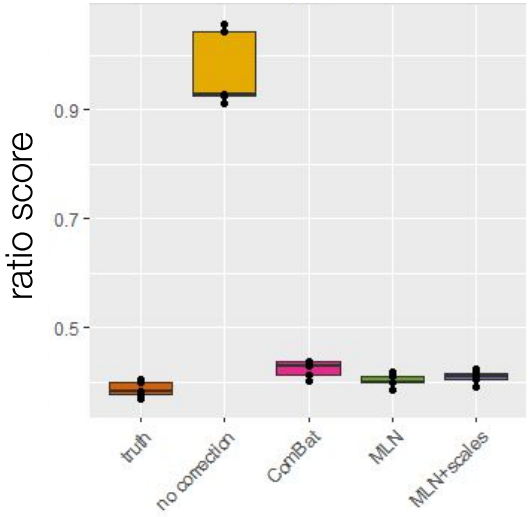
Simulating microbial sequence count data



Results on simulated data: low sparsity

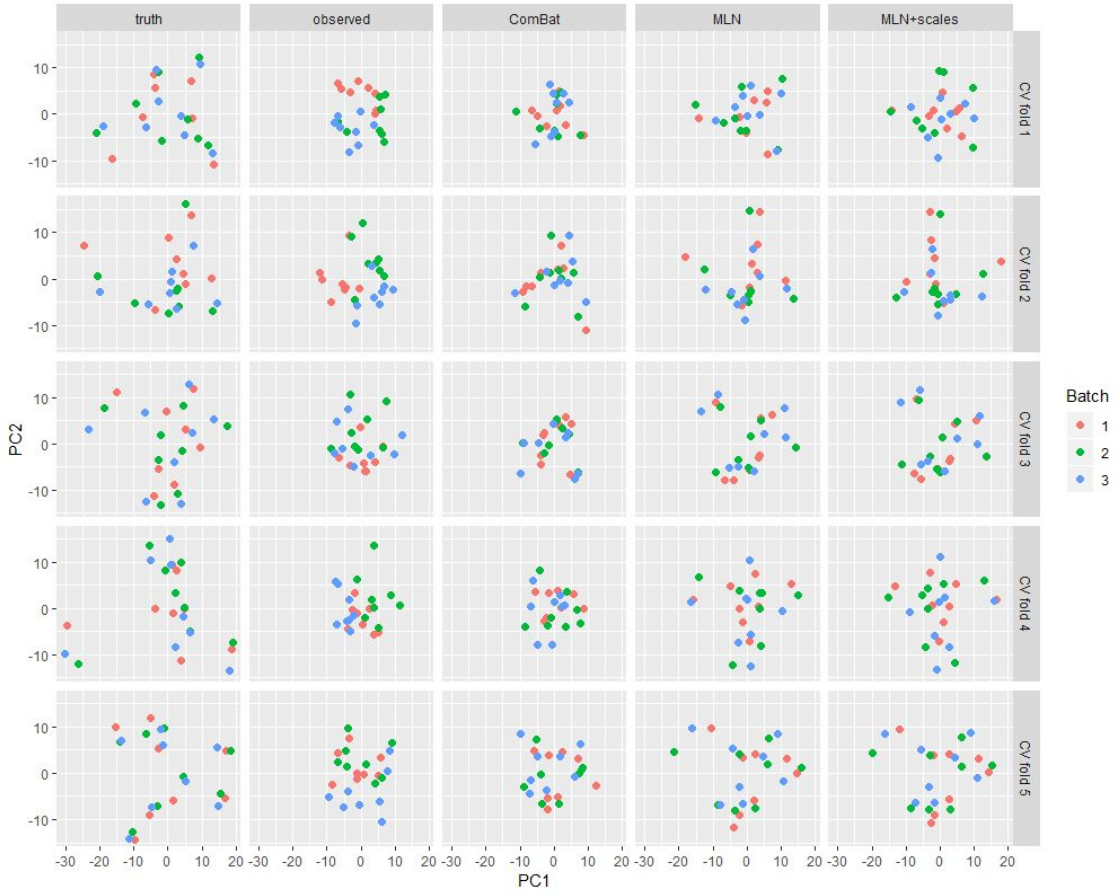
Batch correction over 5 folds

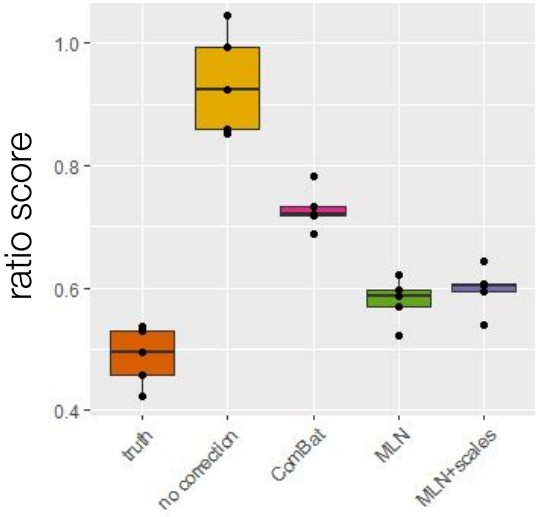




Results on simulated data: high sparsity

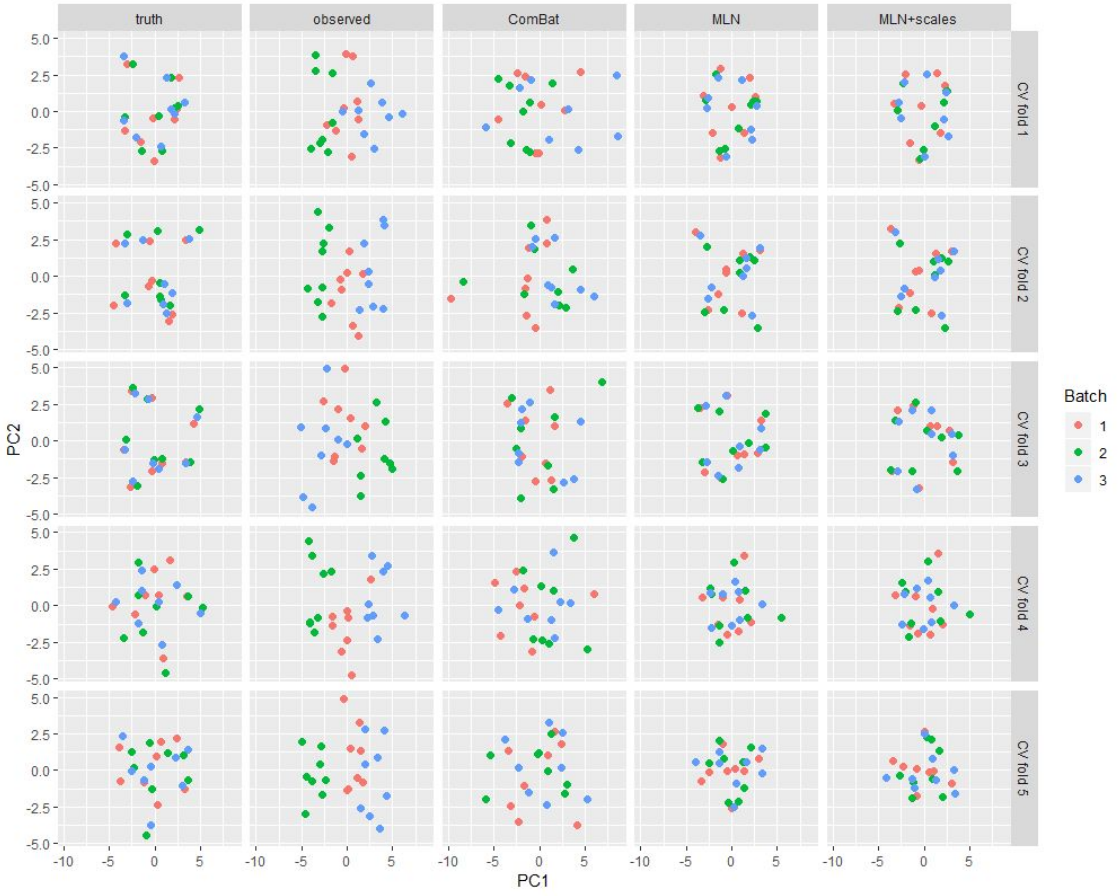
Batch correction over 5 folds

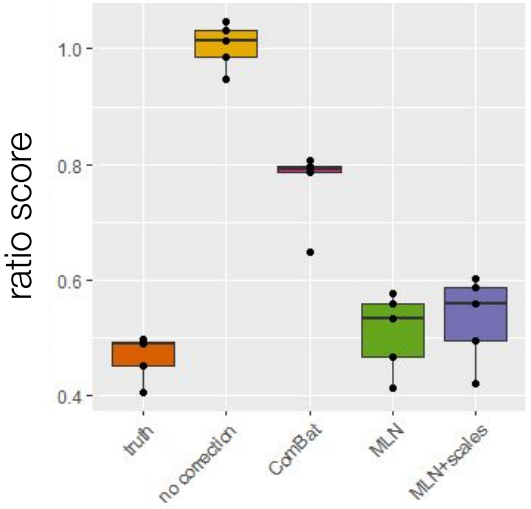




Results on simulated data: high sparsity (bimodal)

Batch correction over 5 folds





Real data set

Pediatric Obesity Metabolism & Microbiome Study (POMMS) baseline

20 patients

Replicates
One
Extraction
Qigen DNeasy PowerSoil Kit
Sequencing
MiSeq 250b PE

20 patients

Replicates
Two
Extraction
MagAttract 96-well plate
Sequencing
MiniSeq 150b PE

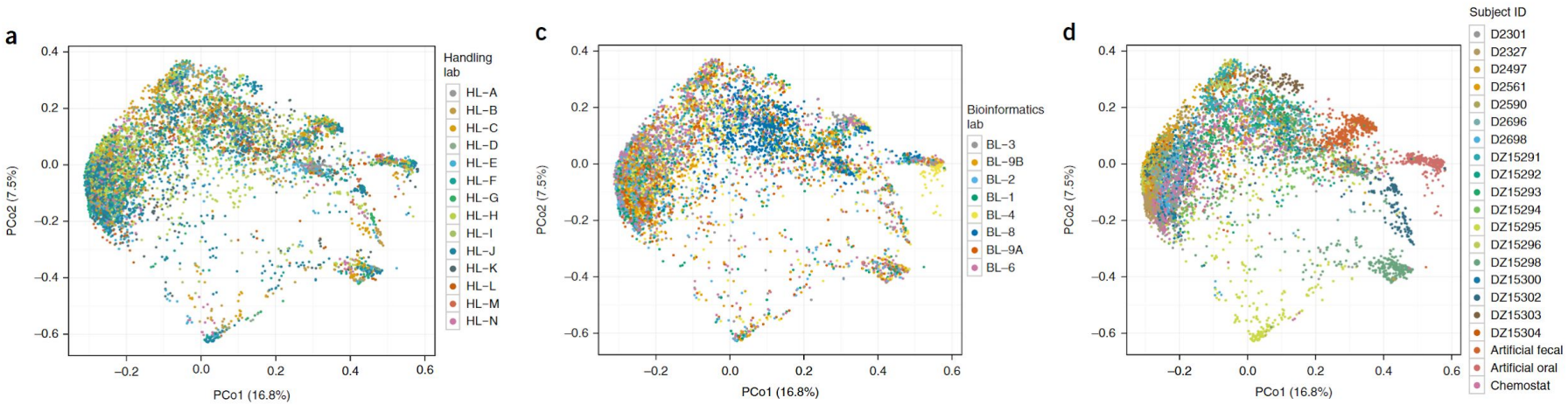
20 patients

Replicates
Two
Extraction
MagAttract 96-well plate
Sequencing
MiSeq 250b PE

Real data set

Microbiome Quality Control Project data

Sinha et al. Nature Biotechnology (2017)



Aims

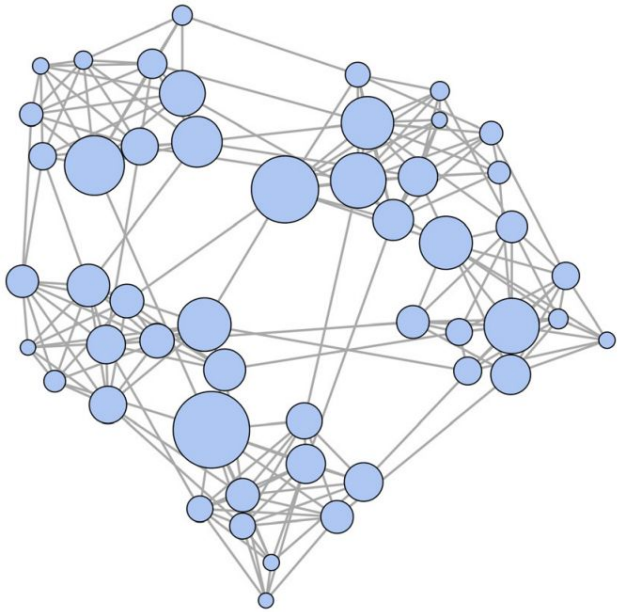
A1 Performing batch effect correction in microbial sequence count data

A2 Characterizing population variation in gut microbial and functional “dynamics”

A3 Testing association of dynamics and host fitness

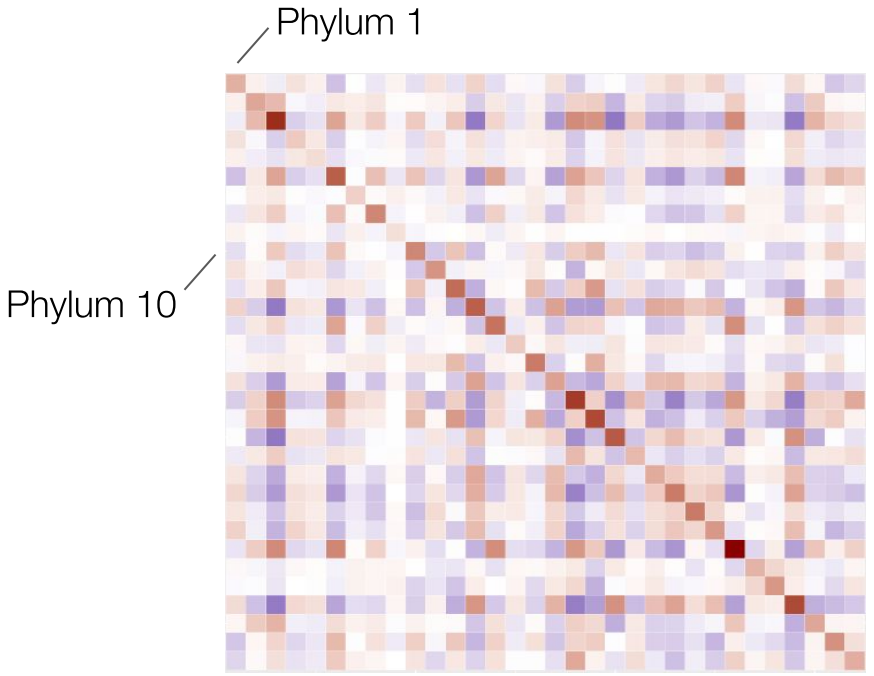
Do individuals' gut microbiomes evolve according to similar rules?

Interactions between microbes are likely to be context-dependent.



Layeghifard et al. "Disentangling Interactions in the Microbiome: A Network Perspective." Cell Press (2017).

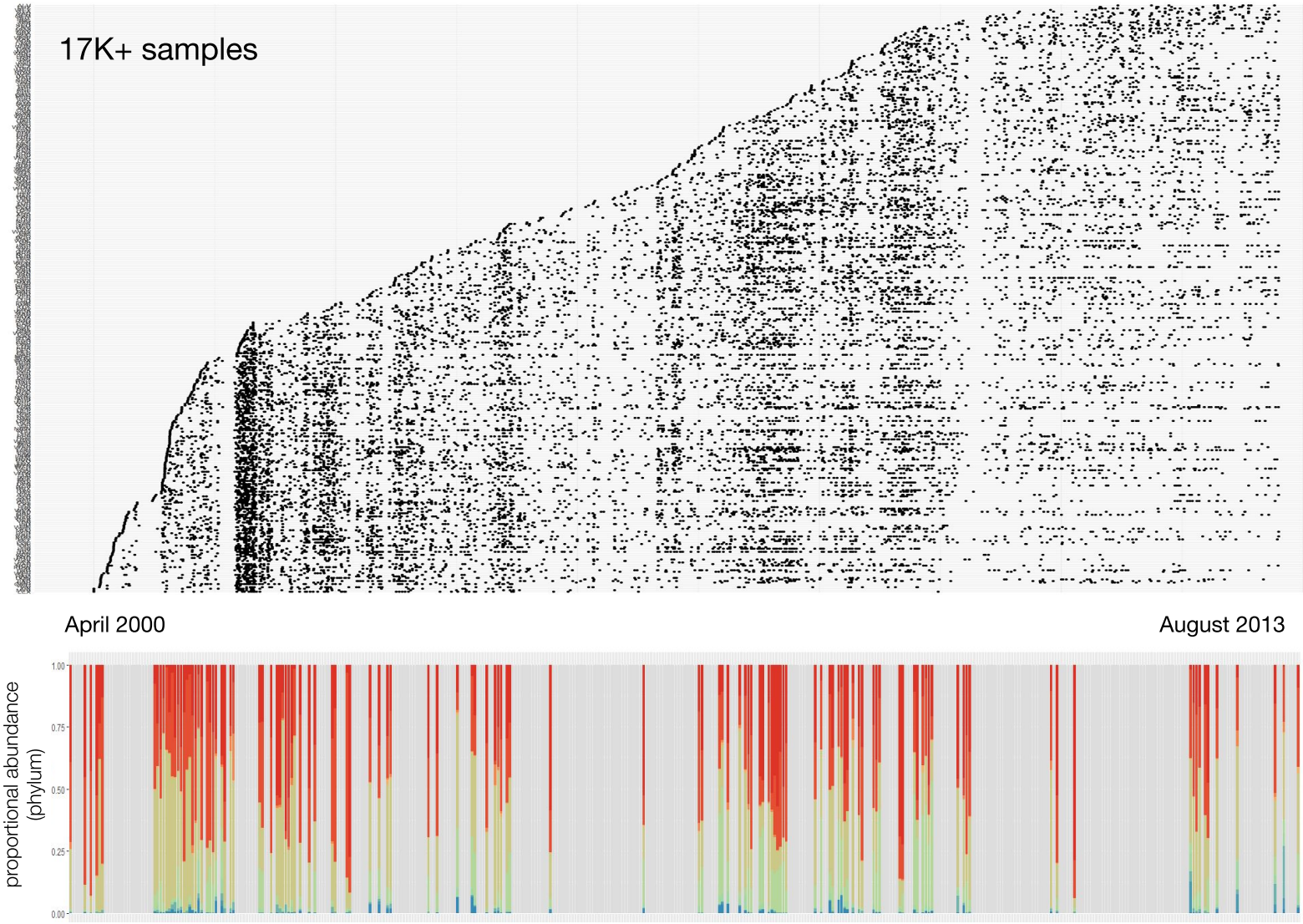
Associations are coarser but easier to characterize.



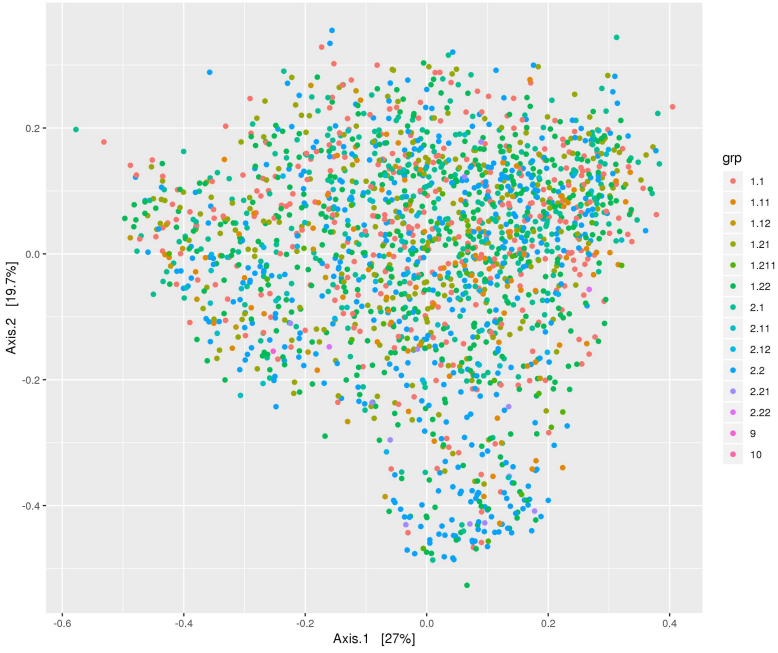
We'll call these "dynamics."

1. Amboseli Baboon Research Project 16S data set
2. Model & inference framework for estimating microbial covariance
3. Considerations: individual modeling, filtration, parameterization
4. Quantifying distances between individuals
5. Initial results

Amboseli Baboon Research Project 16S data set



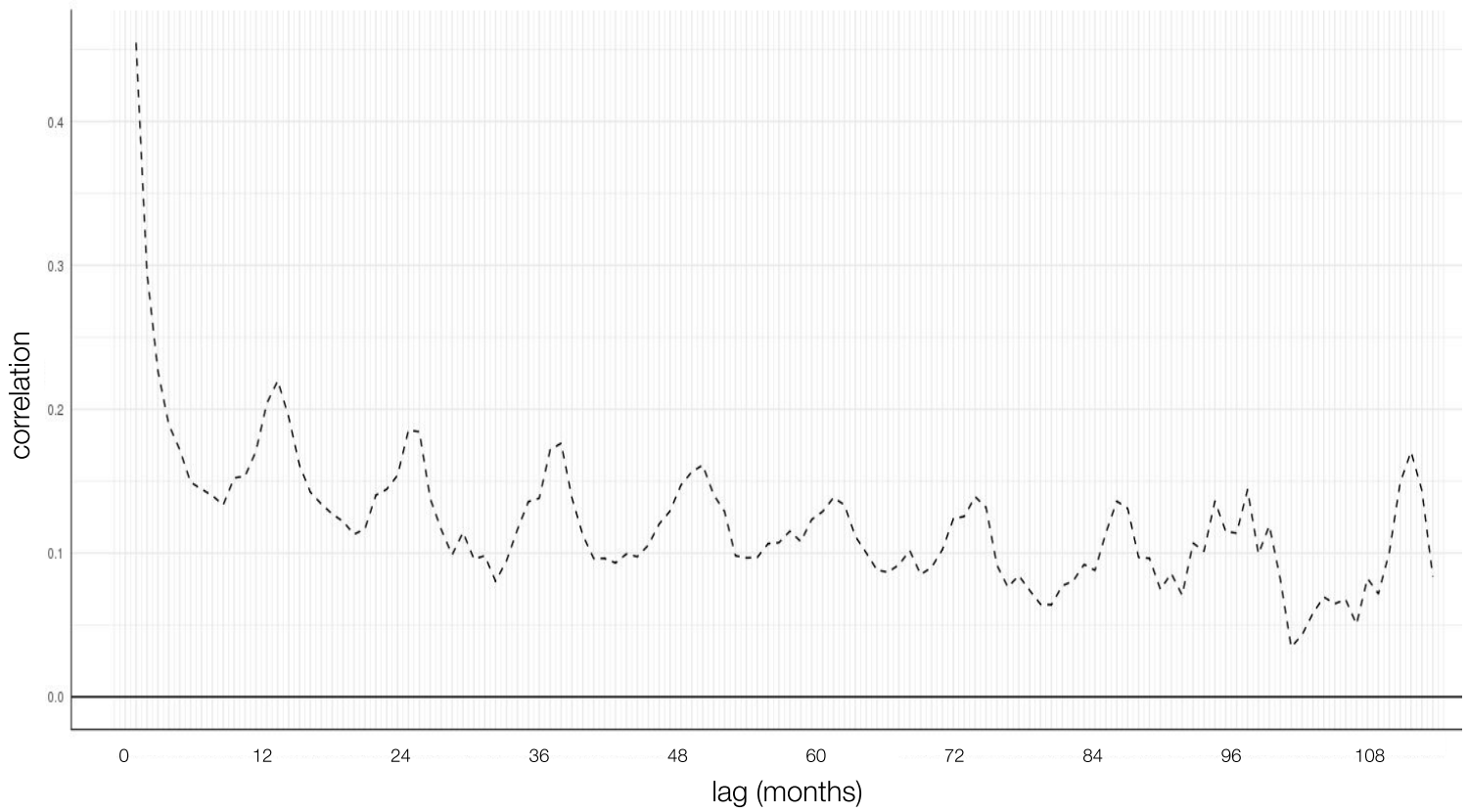
High-level associations of composition with social group and season are visible



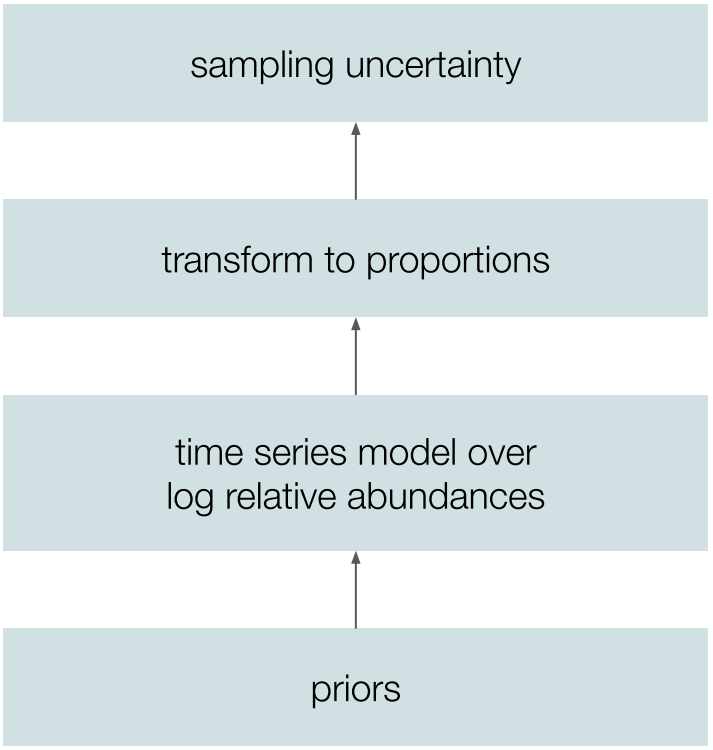
group at collection
dissimilarity by Bray-Curtis

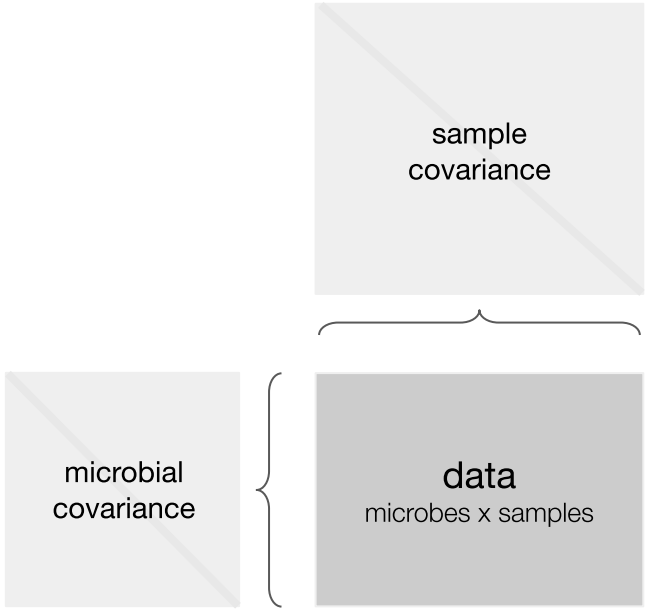
season
dissimilarity by Bray-Curtis

Autocorrelation plots show months-long memory and seasonal fluctuation



We will learn a covariance matrix for each individual via a MLN model housing a Gaussian process

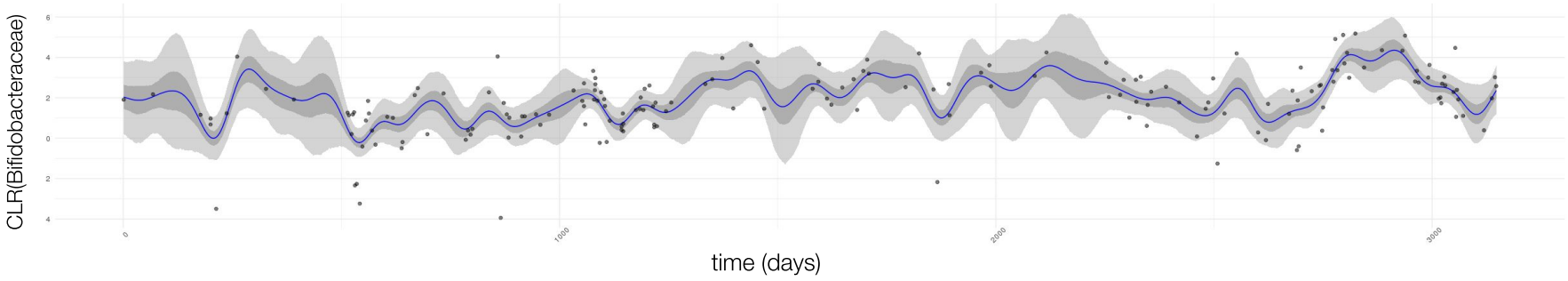




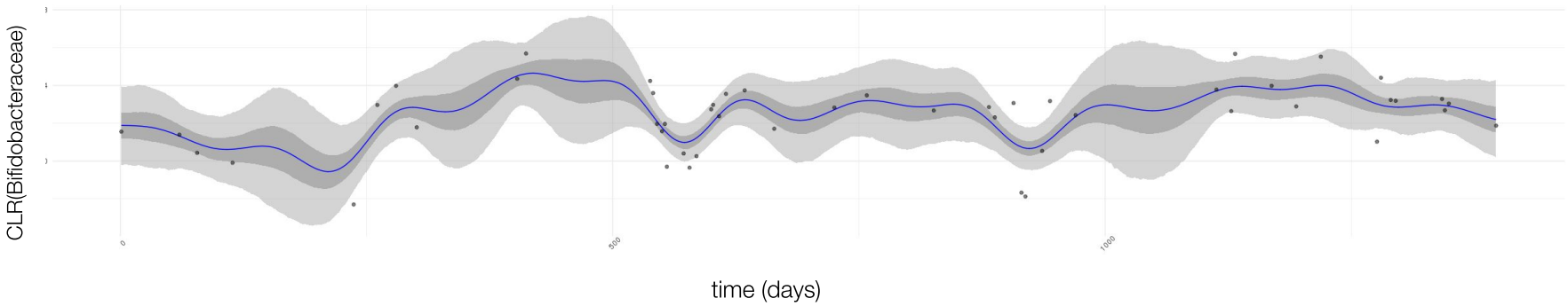
Sample covariance will be modeled by a sum of kernels over time and season.

Differences in sample number are a consideration

"DUIKER" (181 samples)

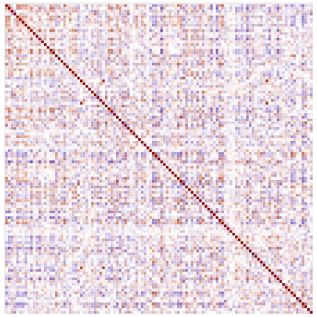


"POWELL" (50 samples)

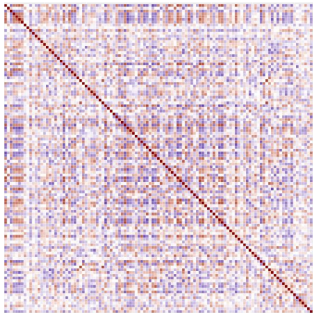


We can quantifying distance between individuals via the Riemannian metric

Individual A

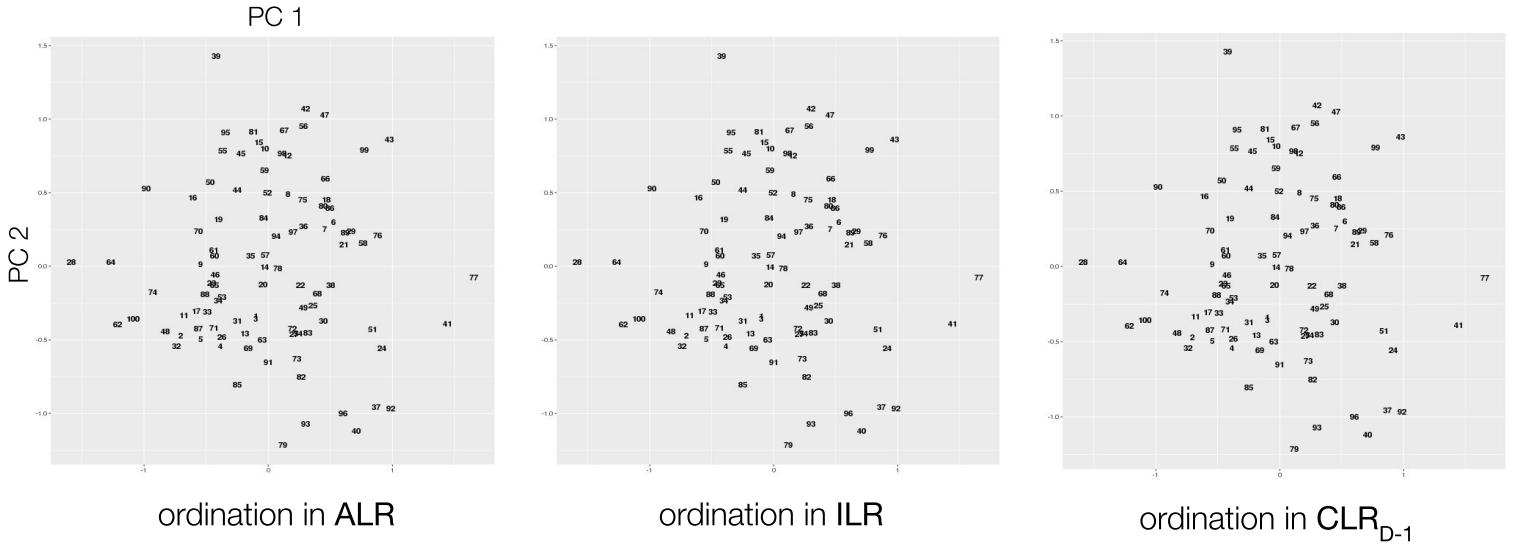


Individual B



$$\mathrm{d}(\boldsymbol{A},\boldsymbol{B}) = \sqrt{\mathrm{tr}\Big(\ln^2\left(\sqrt{\boldsymbol{A}^{-1}}\boldsymbol{B}\sqrt{\boldsymbol{A}^{-1}}\right)\Big)}$$

Distances are conserved across common log ratio representations



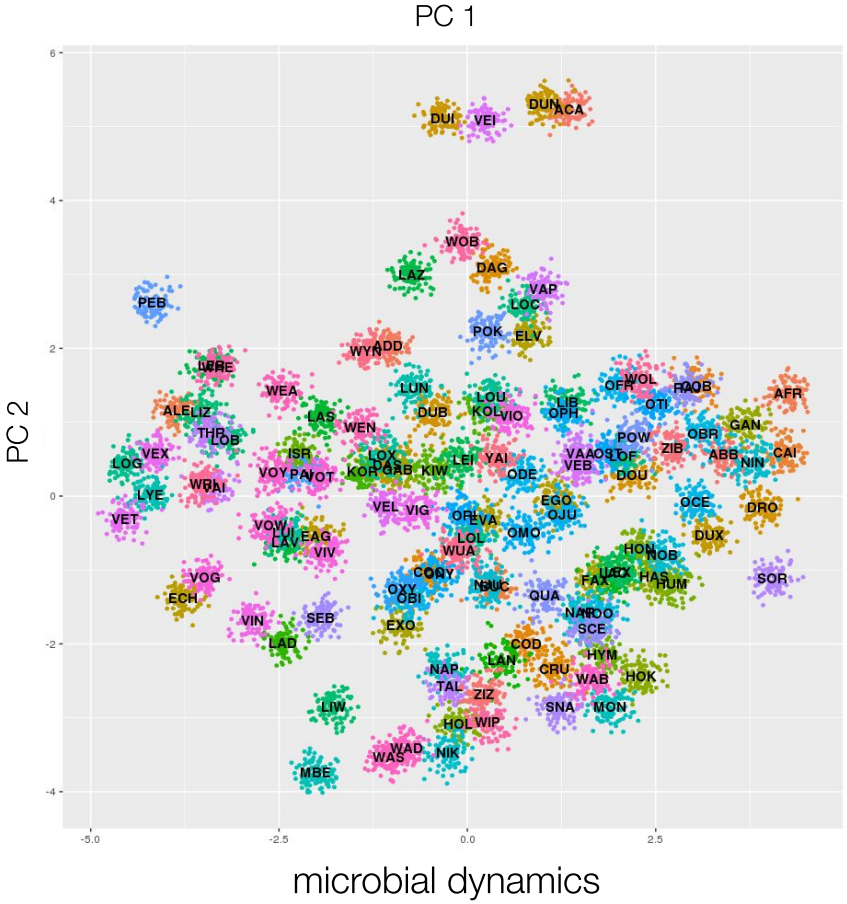
ordination in **ALR**

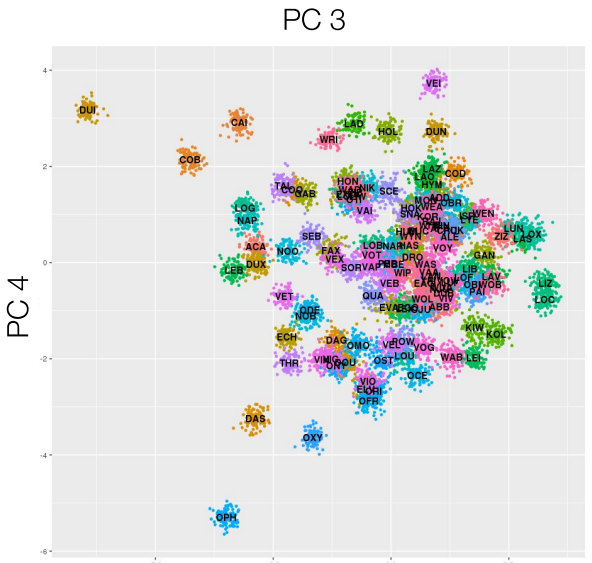
ordination in **ILR**

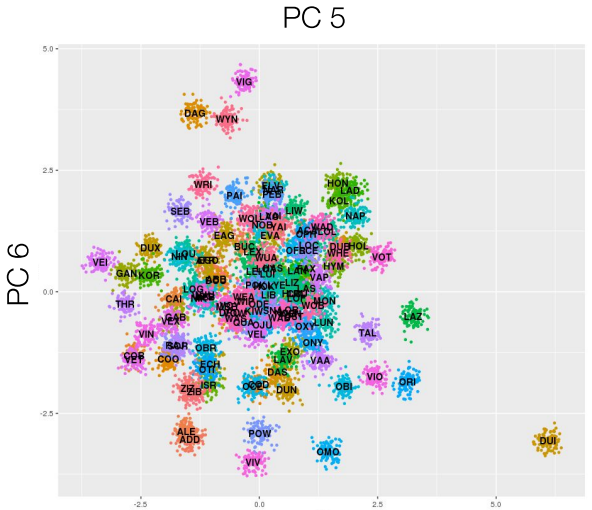
ordination in **CLR**$_{D\text{-}1}$

Visualizing distances in terms of microbial “dynamics”

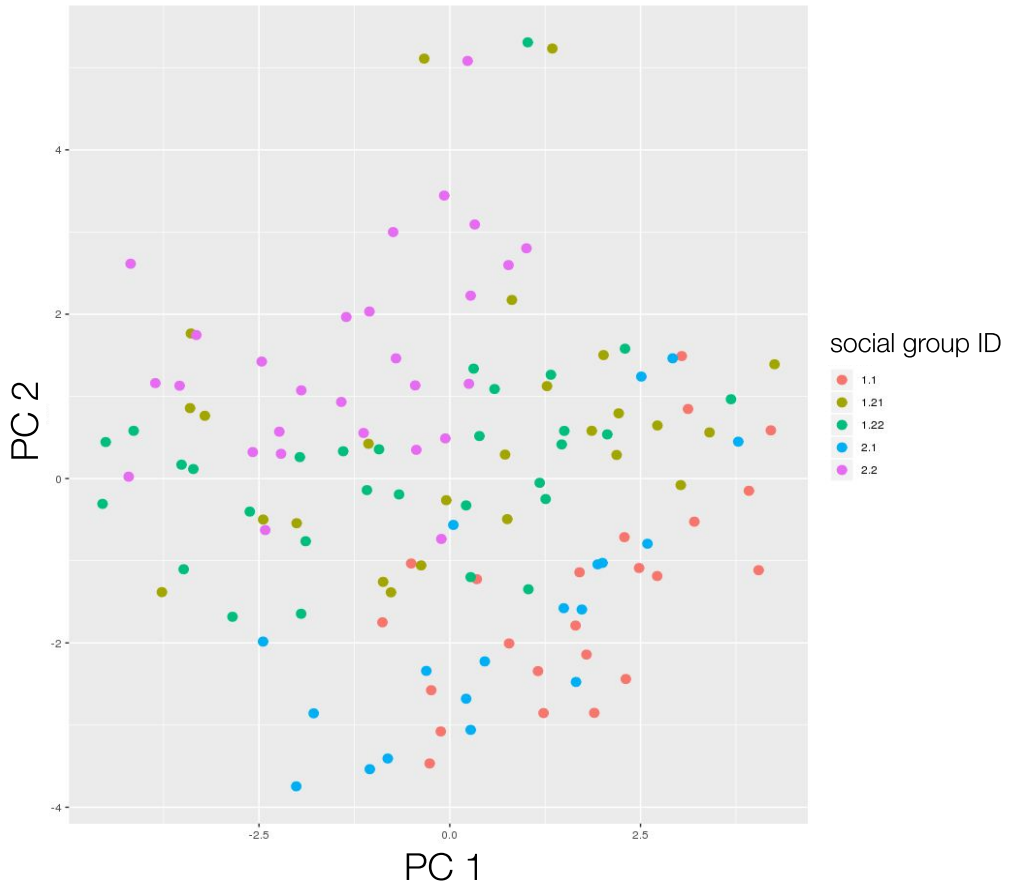
Genus level





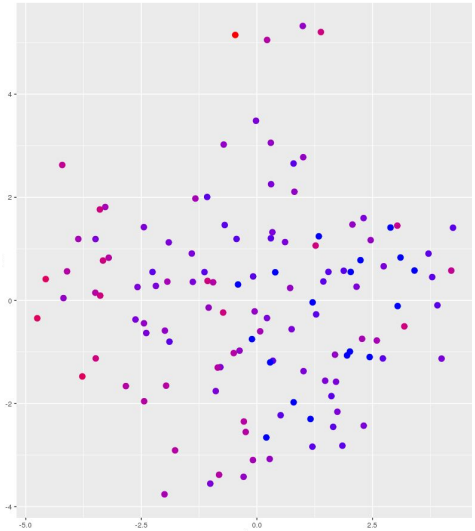


Social group associates with distances in terms of dynamics

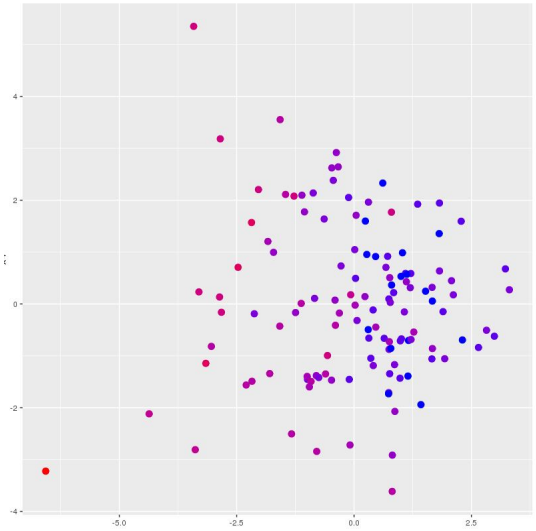


Note: Here individuals are represented by MAP estimate only.

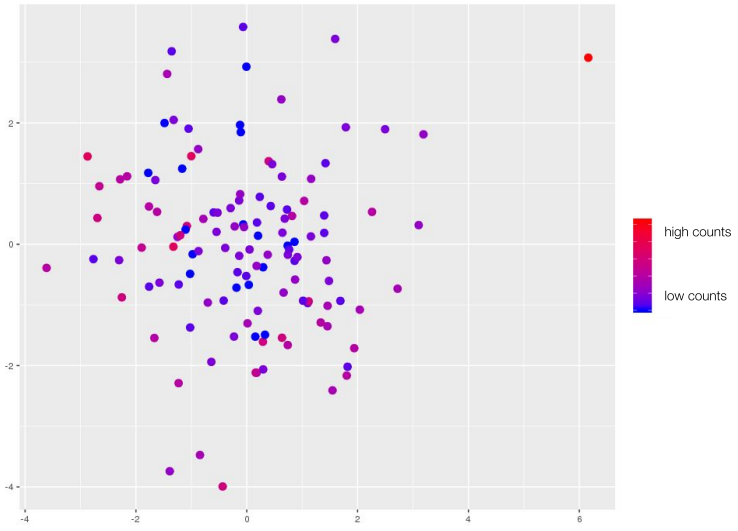
Sampling characteristics may influence estimates of dynamics



PC 1 x 2



PC 3 x 4



PC 5 x 6

Can we answer whether individuals have shared or unique dynamics?

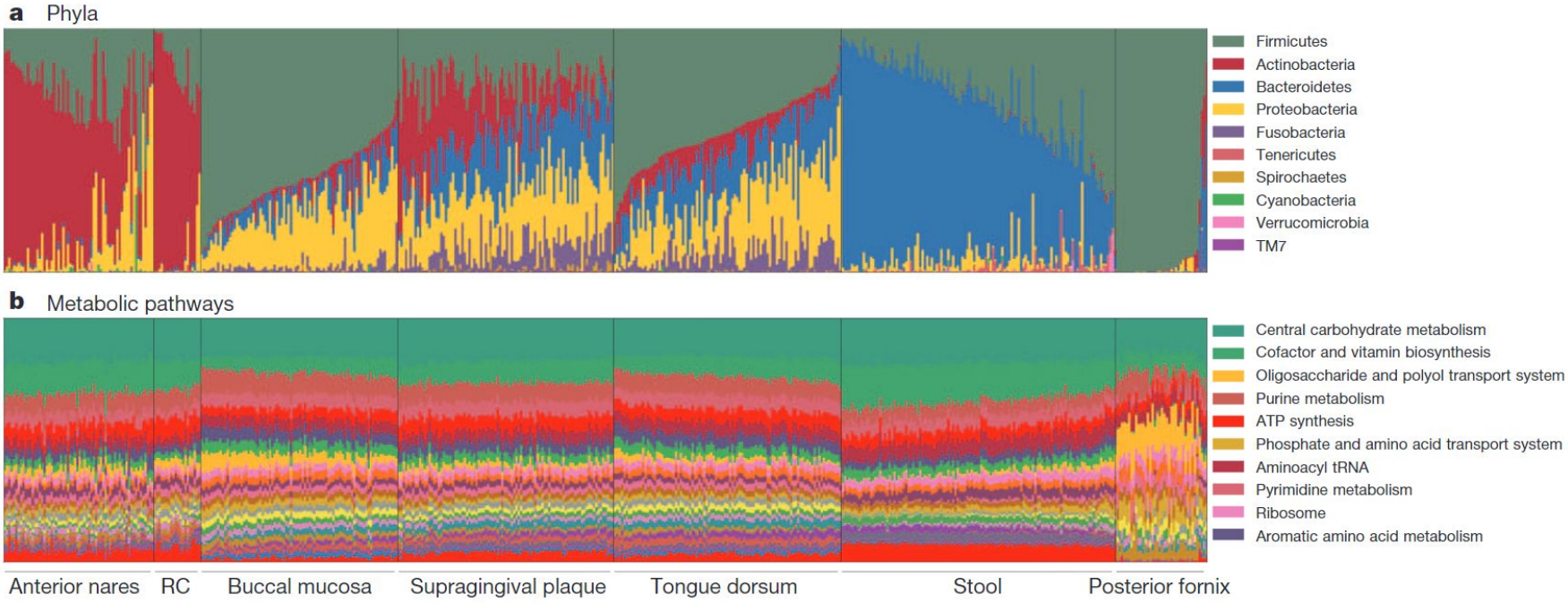
- Individuals can be resolved from each other in terms of their dynamics
- Variation appears continuous
- Need an interpretation for extreme difference
- Social group associates with covariance between microbes

These indicate that shared exposures could explain much of observed distances.

Further questions

- Are there subsets of microbial relationships in subsets of individuals that are shared?
- Are rare taxa the primary drivers of dissimilarity between individuals?
- Do functional metagenomics data reveal greater similarity between individuals than microbial abundances?

Comparing functional dynamics to compositional dynamics



Structure, function and diversity of the healthy human microbiome
The Human Microbiome Project Consortium in Nature (2012)

Aims

A1 Performing batch effect correction in microbial sequence count data

A2 Characterizing population variation in gut microbial and functional “dynamics”

A3 Testing association of dynamics and host fitness

Do individuals with atypical dynamics also have atypical fitness outcomes?

1. Define *fitness*: normalized number of offspring
2. Define a kernel ($\mathbf{K}$) between individuals based on similarity of dynamics
3. Model out other contributors to fitness (e.g. rank)
4. Interpret τ / (α + τ) as variance explained

$$g\left(\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}\right) \sim N(X\beta, \alpha I + \tau K)$$

$\mathbf{y_i}$ fitness outcome of individual i
$\mathbf{X\beta}$ baseline outcome explained by covariates
$\boldsymbol{\alpha}\mathbf{I}$ scaled noise
$\boldsymbol{\tau}\mathbf{K}$ variation due to similarity of dynamics

for σ in $\{\sigma_1, \ldots, \sigma_p\}$ **do**
  for each k of K folds **do**
    let $Y^{(k)}_{\text{train}} \sim N(X\beta, \alpha I + \tau K)$;
    optimize (α, τ) ;
    let $\hat{Y}^{(k)}_{\text{test}} = \mathbb{E}\left(Y^{(k)}_{\text{test}} | Y^{(k)}_{\text{train}}\right)$;
    let $\text{err}^{(k)}_{\text{pred}} = \|\hat{Y}^{(k)}_{\text{test}} - Y^{(k)}_{\text{train}}\|$;
  end
end
choose (α, τ, σ) that minimize err_{pred}

A1 Supplement

Common log relative abundance transformations

$$\mathrm{alr}(x_1, \ldots, x_D) = \boldsymbol{y} = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{X_{D-1}}{x_D} \right)$$
$$\mathrm{alr}^{-1}(\boldsymbol{y}) = \boldsymbol{x} = \mathcal{C}\big(\exp(\boldsymbol{y}; 0) \big)$$

$$\mathrm{clr}(x_1, \ldots, x_D) = \boldsymbol{y} = \left(\ln \frac{x_1}{g(\boldsymbol{x})}, \ldots, \ln \frac{X_D}{g(\boldsymbol{x})} \right) \qquad g(\boldsymbol{x}) = \sqrt[D]{x_1 \ldots x_D}$$
$$\mathrm{clr}^{-1}(\boldsymbol{y}) = \boldsymbol{x} = \mathcal{C}\big(\exp(\boldsymbol{y}) \big)$$

$$\mathrm{ilr}(x_1, \ldots, x_D) = \boldsymbol{y} = \ln(\boldsymbol{x}) \cdot \boldsymbol{V}$$
$$\mathrm{ilr}^{-1}(\boldsymbol{y}) = \boldsymbol{x} = \mathcal{C}\big(\exp(\boldsymbol{x} \cdot \boldsymbol{V}^T) \big)$$

Multinomial logistic normal model formulations used

linear regression

$$
\begin{aligned}
Y_j &\sim \text{Multinomial}(\pi_j) \\
\pi_j &= \text{ALR}_d^{-1}(\eta_j) \\
\eta_j &\sim N(\Lambda X_j, \Sigma) \\
\Lambda &\sim N(\Theta, \Sigma, \Gamma) \\
\Sigma &\sim W^{-1}(\Xi, \upsilon)
\end{aligned}
$$

linear regression+scales

$$
\begin{aligned}
Y_j &\sim \text{Multinomial}(\pi_j) \\
\pi_j &= \text{ALR}_d^{-1}(\eta_j) \\
\eta_j &\sim N(\Lambda X_j, \Sigma) \\
\Lambda &\sim N(\Theta, \Sigma, \Gamma) \\
\Sigma &\sim W^{-1}(\Xi, \upsilon) \\
\Gamma &= \sigma_0 \Gamma_0 + \sum_{i=1}^{q} \sigma_i \Gamma_i
\end{aligned}
$$

per-batch scatter

Batch effect simulation scheme

$$
\begin{aligned}
Y_j &\sim \text{Multinomial}(\pi_j) \\
\pi_j &= \text{ALR}_d^{-1}(\eta_j) \\
\eta &= MZ + BX + E \qquad (d-1 \times n) \\
M &\sim N(0, \Sigma_M, I_m) \qquad (d-1 \times m) \\
B &\sim N(0, \Sigma_B, \Lambda) \qquad (d-1 \times q) \\
E &\sim N(0, \Sigma_E, I_n) \qquad (d-1 \times n)
\end{aligned}
$$

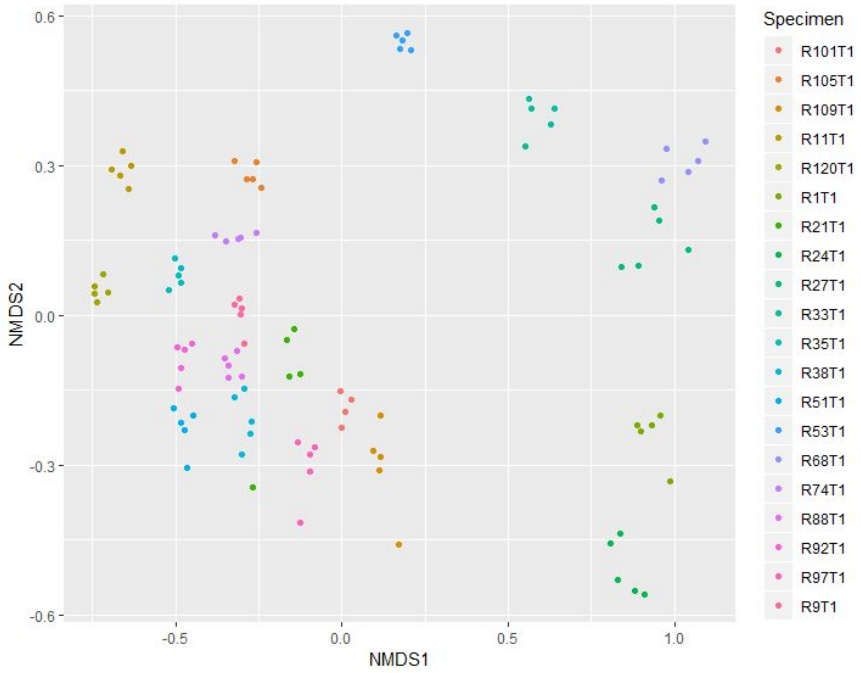
m = # individuals

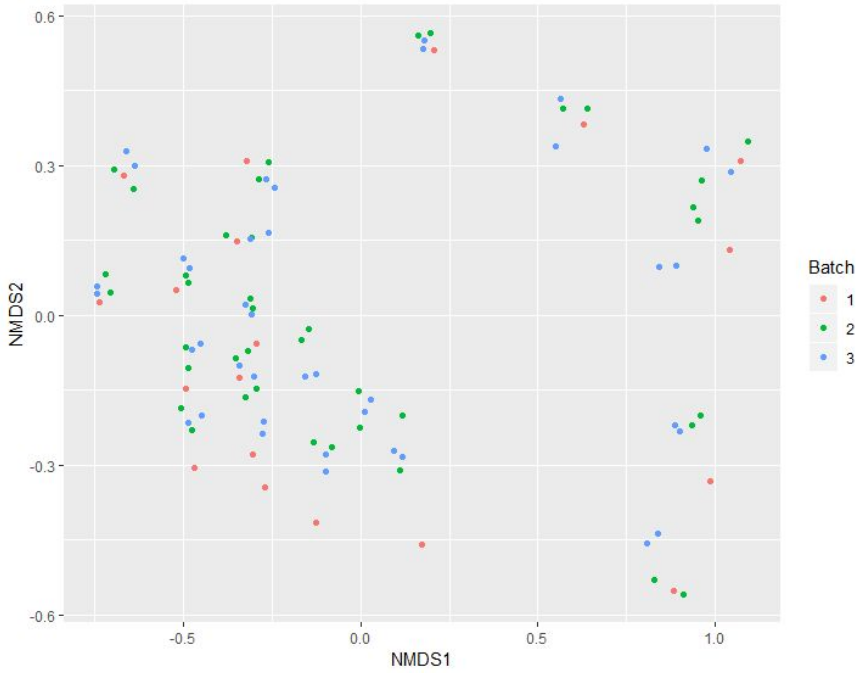
q = # batches

n = # total samples

Visualizing the POMMS baseline batch data set

Dissimilarity by Bray-Curtis

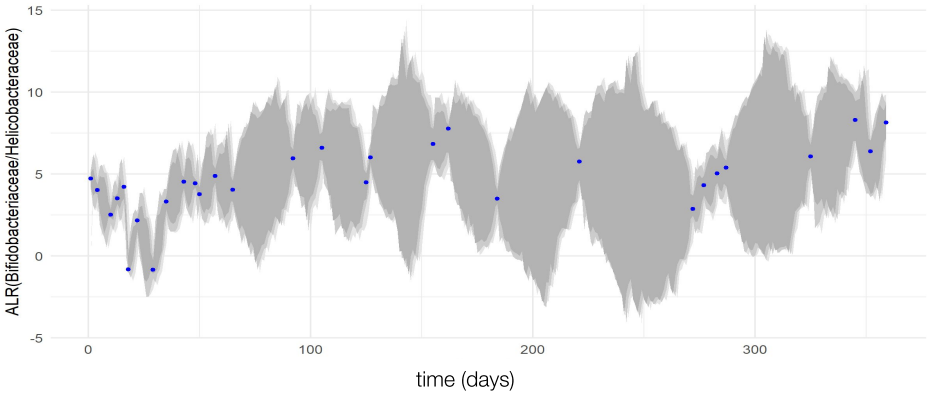




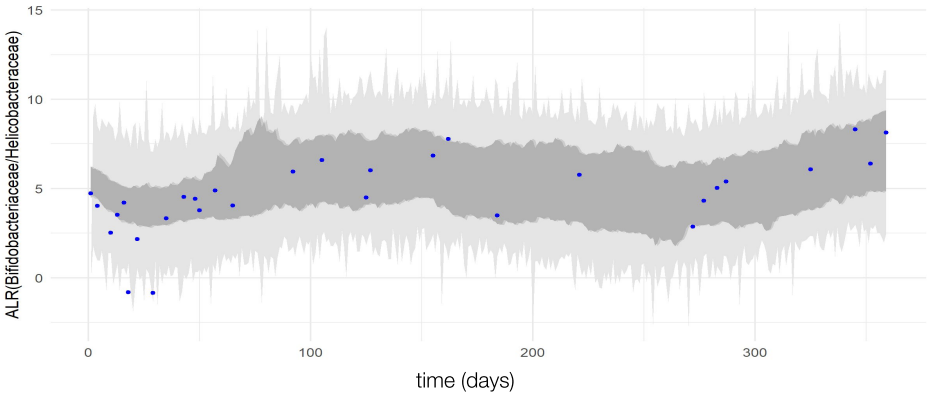
A2 Supplement

MLN + dynamic linear model

$$
\begin{aligned}
Y_t &\sim \text{Multinomial}(\pi_t) \\
\pi_t &= \text{ALR}_d^{-1}(\eta_t) \\
\eta_t^T &= F_t^T \Theta_t + v_t^T \qquad v_t \sim \text{N}(0, \gamma_t \Sigma) \\
\Theta_t &= G_t \Theta_{t-1} + w_t \qquad w_t \sim \text{N}(0, W_t, \Sigma) \\
\Theta_0 &\sim \text{N}(M_0, C_0, \Sigma) \\
\Sigma &\sim \text{Wishart}^{-1}(\Xi, \upsilon)
\end{aligned}
$$



Fixed low noise in v_t



Optimized noise in v_t and w_t

MLN + Gaussian process

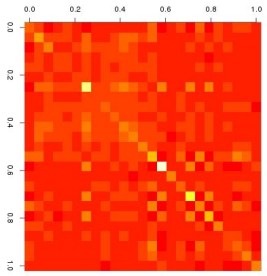
$$
\begin{aligned}
Y_t &\sim \mathrm{Multinomial}(\pi_t) \\
\pi_t &= \mathrm{ALR}_d^{-1}(\eta_t) \\
\eta &\sim \mathrm{N}(\Lambda[X], \Sigma, I_N), && \text{i.e. } \Sigma \otimes I_N \\
\Lambda[X] &\sim \mathrm{GP}(\Theta[X], \Sigma, \Gamma[X]), && \Sigma \otimes \Gamma \\
\Sigma &\sim \mathrm{Wishart}^{-1}(\Xi, \upsilon)
\end{aligned}
$$

$\mathbf{\Gamma}$: squared exp. kernel (time) +
periodic kernel (season) +
white noise (technical noise)

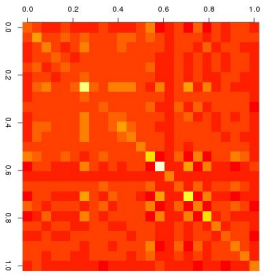
Downsampling can inform a lower informative sample number

EX: Start with all samples for a well-sampled individual, progressively downsample, and fit the model.

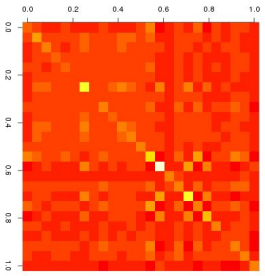
100% samples (181)



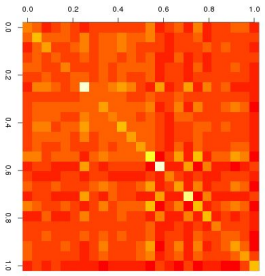
80% of samples



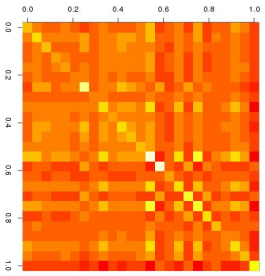
60% of samples



40% of samples



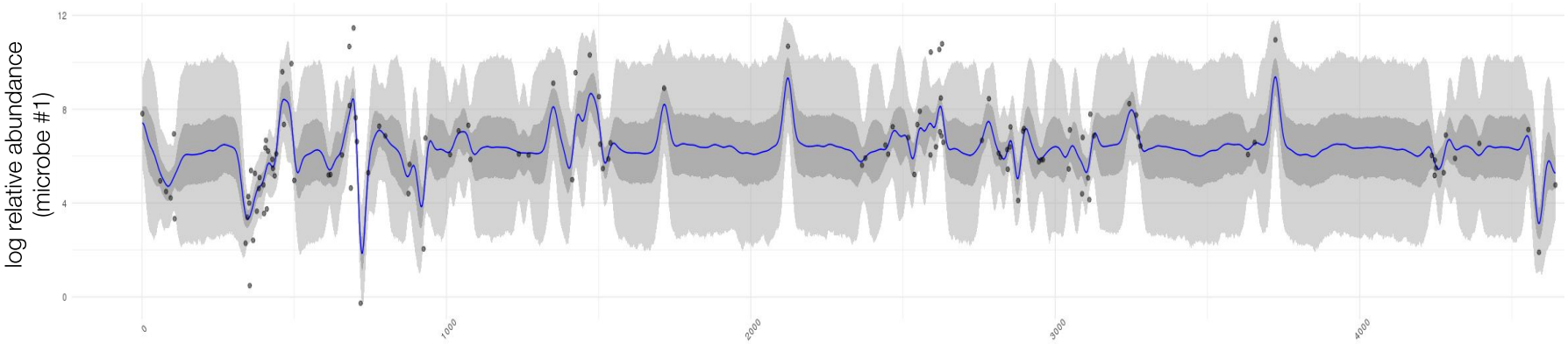
20% of samples



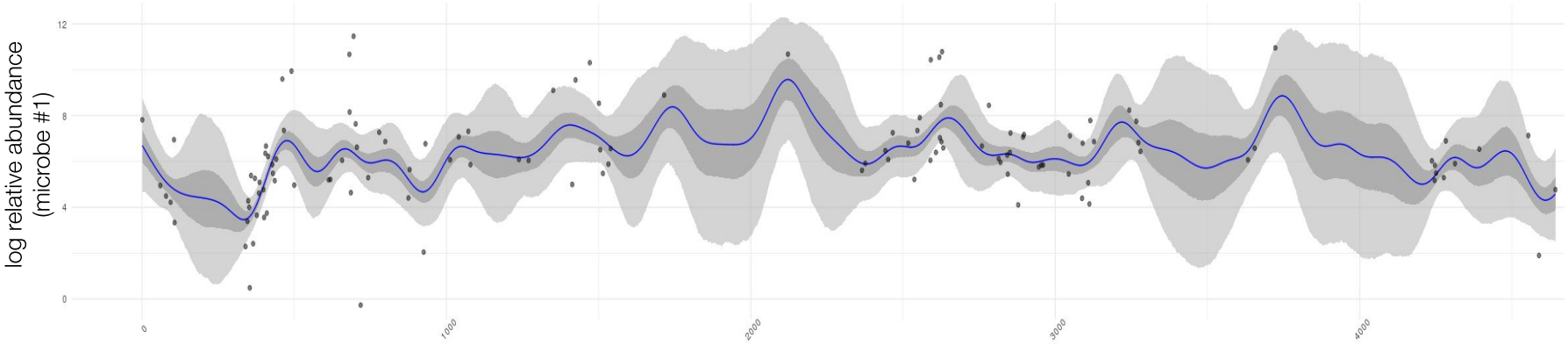
Determine minimum sample number where model fits exhibit undesirable behavior.

Posterior predictive fits, varying choice of GP bandwidth

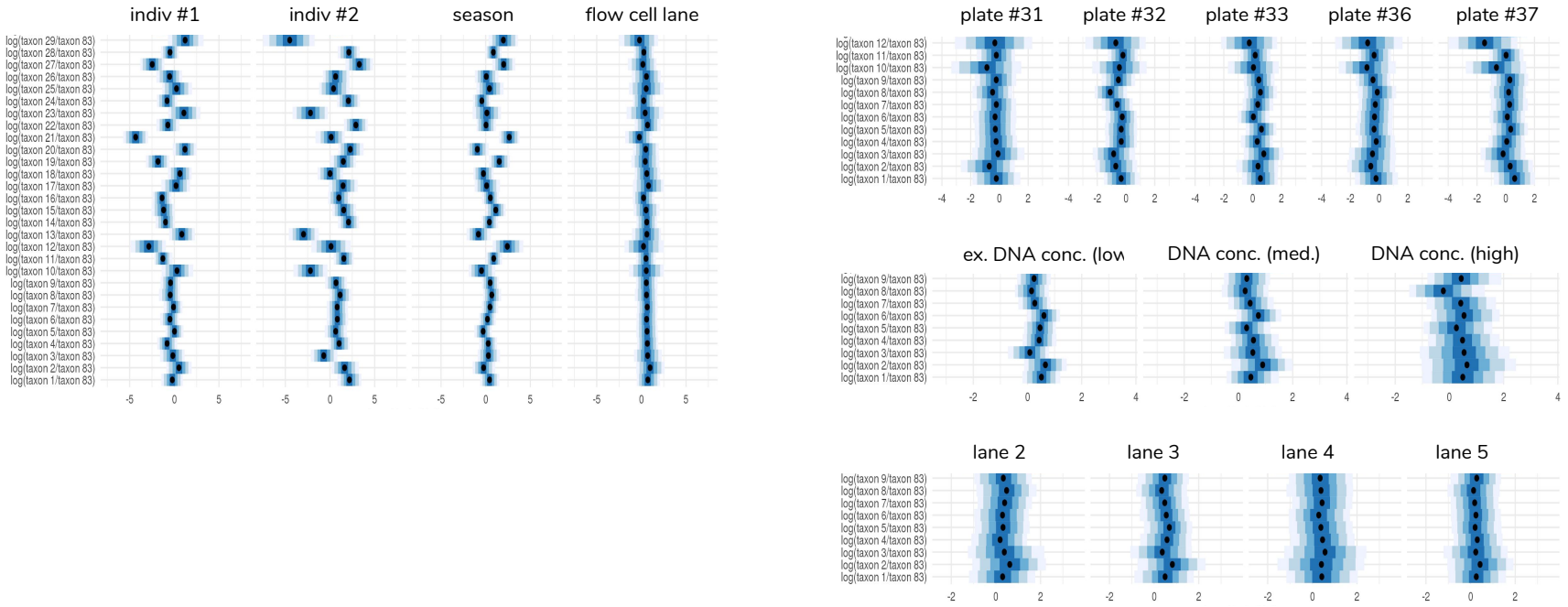
~30 day information window



~180 day information window



Regression shows reliable effects of individual ID and season on log relative abundance but not batch processing variables

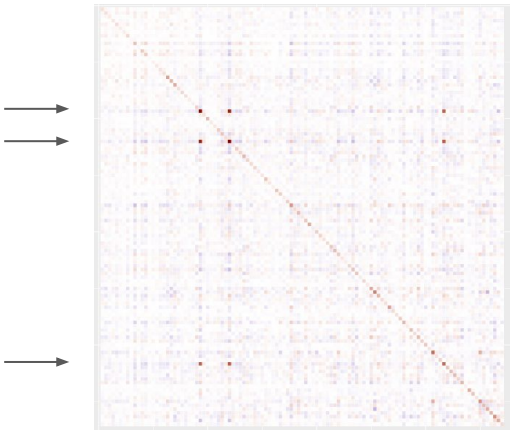


Excerpts of all variables and posterior intervals shown. Intervals that spanned zero were judged as *no effect*.

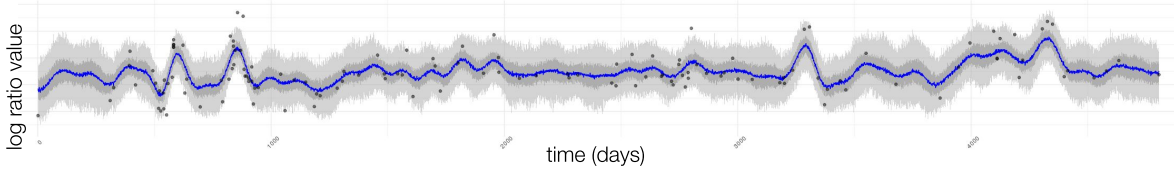
Correlated taxa seem reasonable

Shown: 3 highly variable, strongly (+) correlated Firmicutes

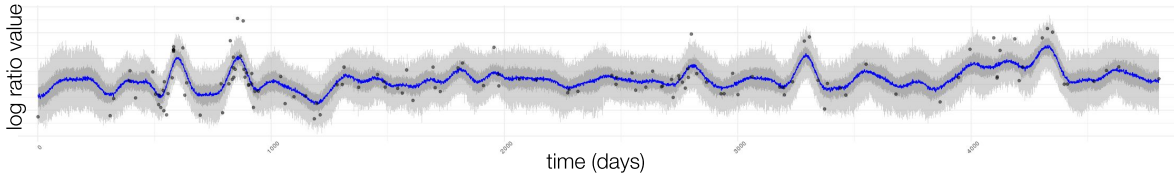
Individual “VET”



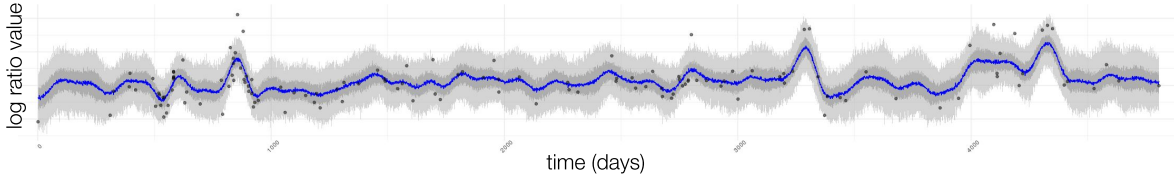
CLR(Planococcaceae/NA)



CLR(Planococcaceae/Solibacillus)

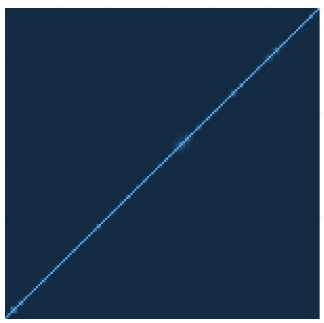


CLR(Bacillaceae/Bacillus)



A variance component model motivates relative contributions of GP kernels

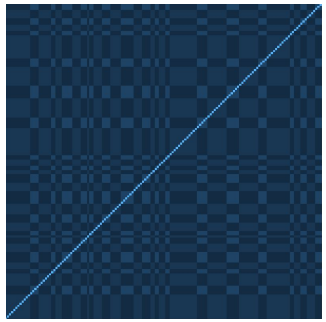
1 Variance components were constructed as in these toy examples


weekly sample correlation


age correlation


individual correlation

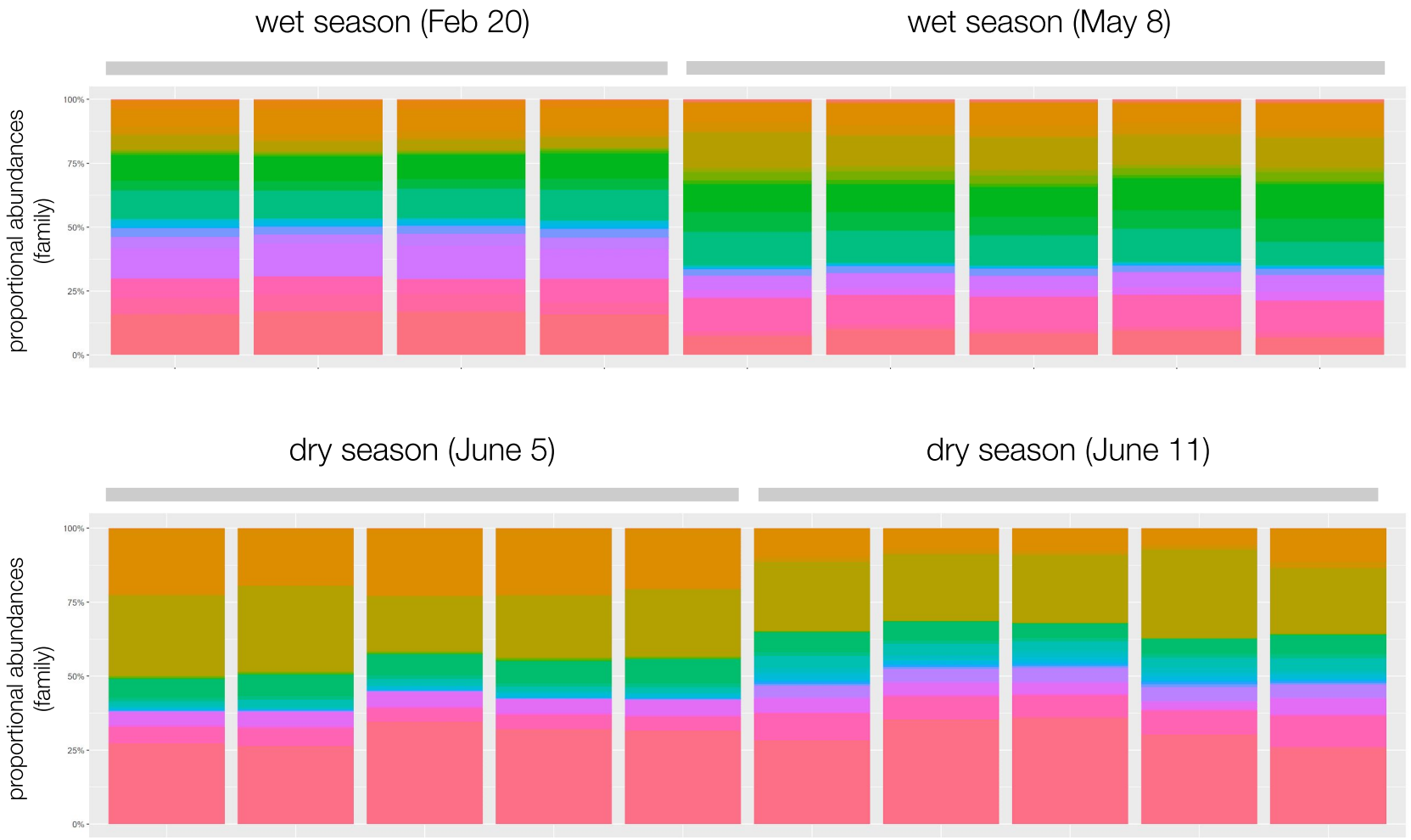

group correlation


seasonal correlation

2 Posterior density of this simple model was optimized to give component scales $\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_k$

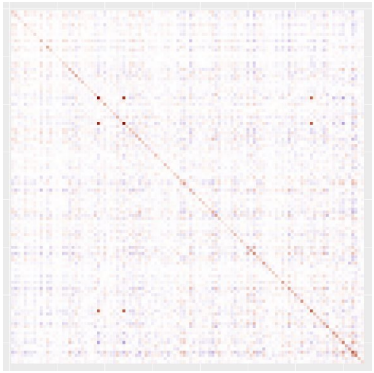
$$\eta \sim \mathrm{MN}(M, U, V)$$
$$U \sim \mathrm{W}^{-1}(\upsilon, \Psi)$$
$$V = \gamma_1 V_1 + \gamma_2 V_2 + \cdots + \gamma_k V_k$$

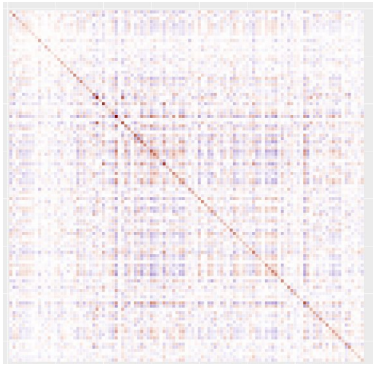
Variation in replicates can inform noise model



Microbial dynamics PC 1 may capture *net covariation*

individual "LOGAN" from extreme lower end of PC 1

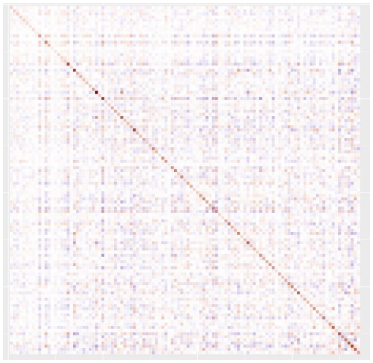


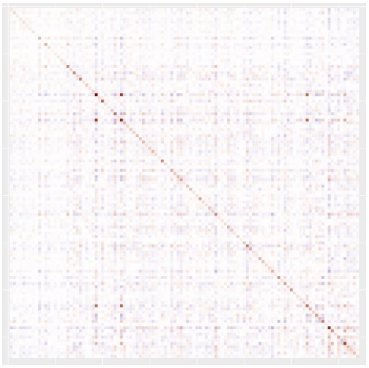


individual "AFRICA" from extreme upper end of PC 1

Microbial dynamics PC 2 may capture *net variation*

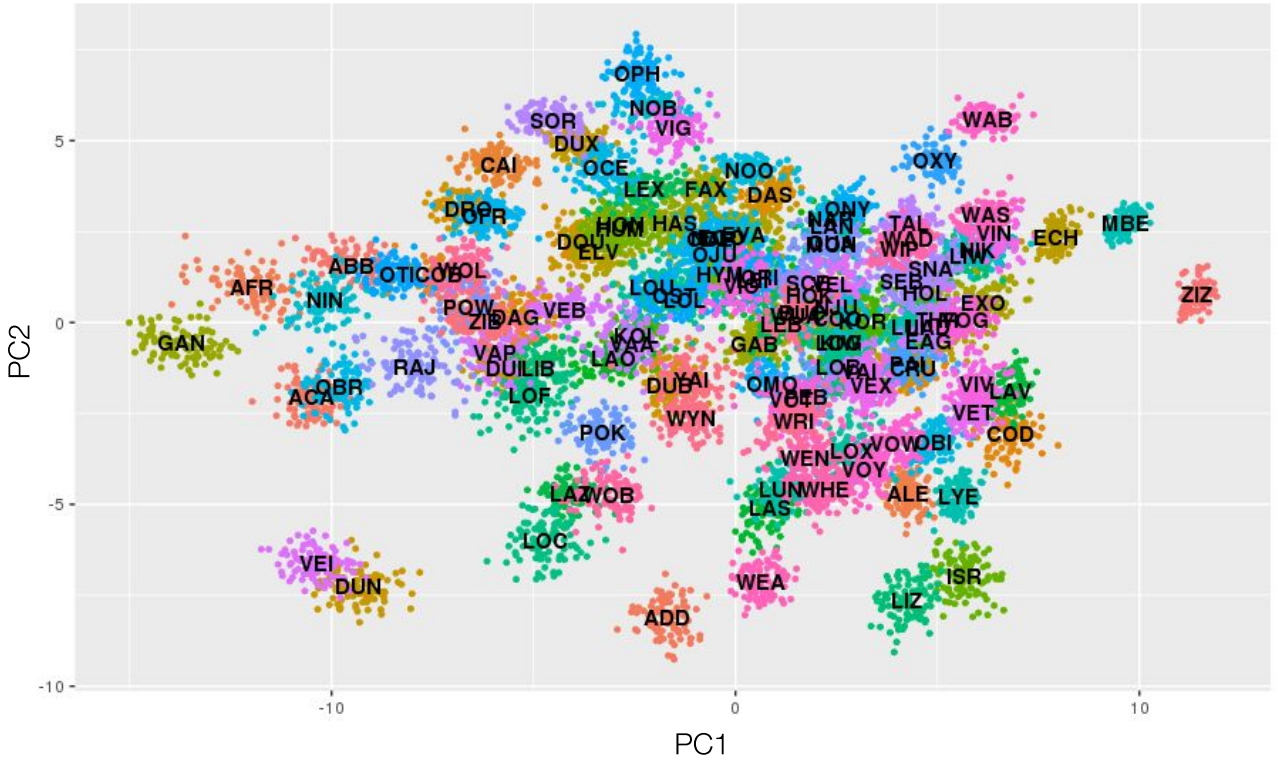
individual "WASP" from extreme lower end of PC 1





individual "DUIKER" from extreme upper end of PC 1

We can coarsely approximate a distance between individuals in terms of *compositional baseline* with this model

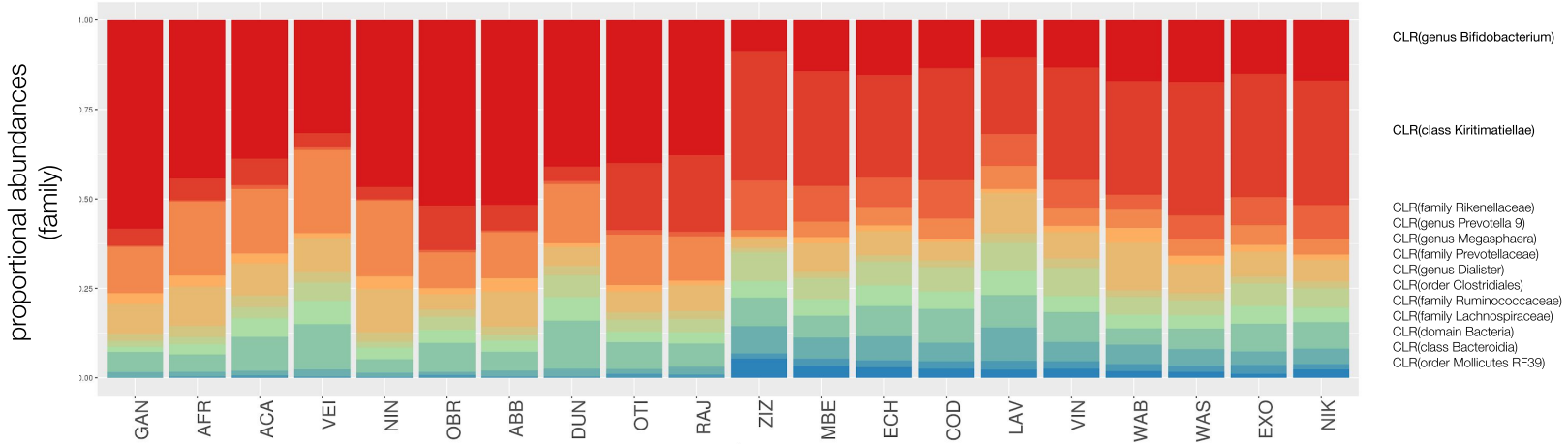


This whole quadrant is (seemingly) rich in Bifido + Kirimatiellae.; details follow

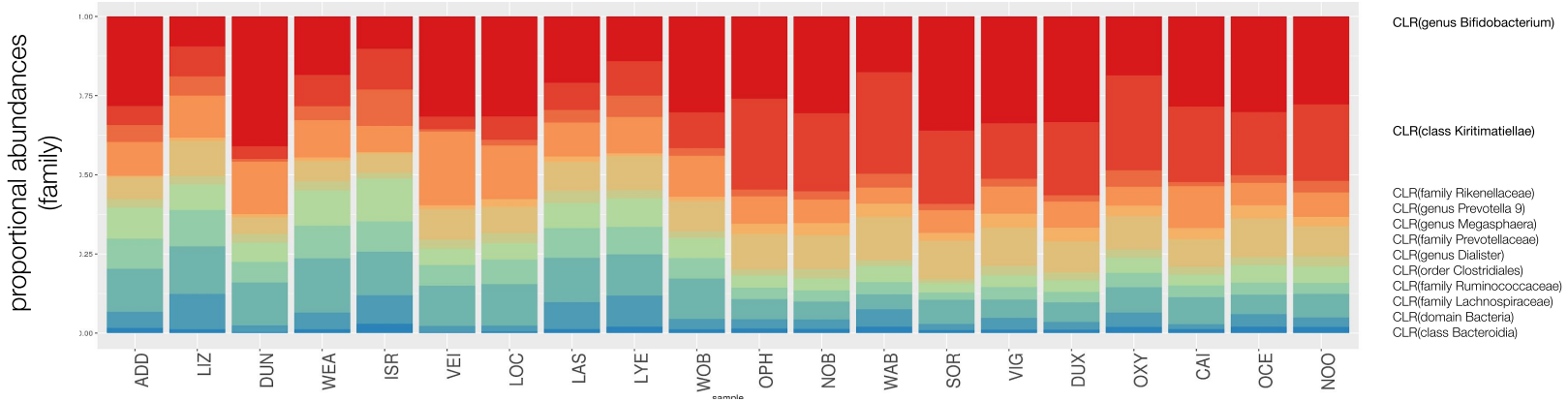
Each point is a posterior sample of $\boldsymbol{\Lambda}[\mathbf{X}]$ from model in slide 40, averaged across time.

Interpreting the extremes of difference in compositional baseline

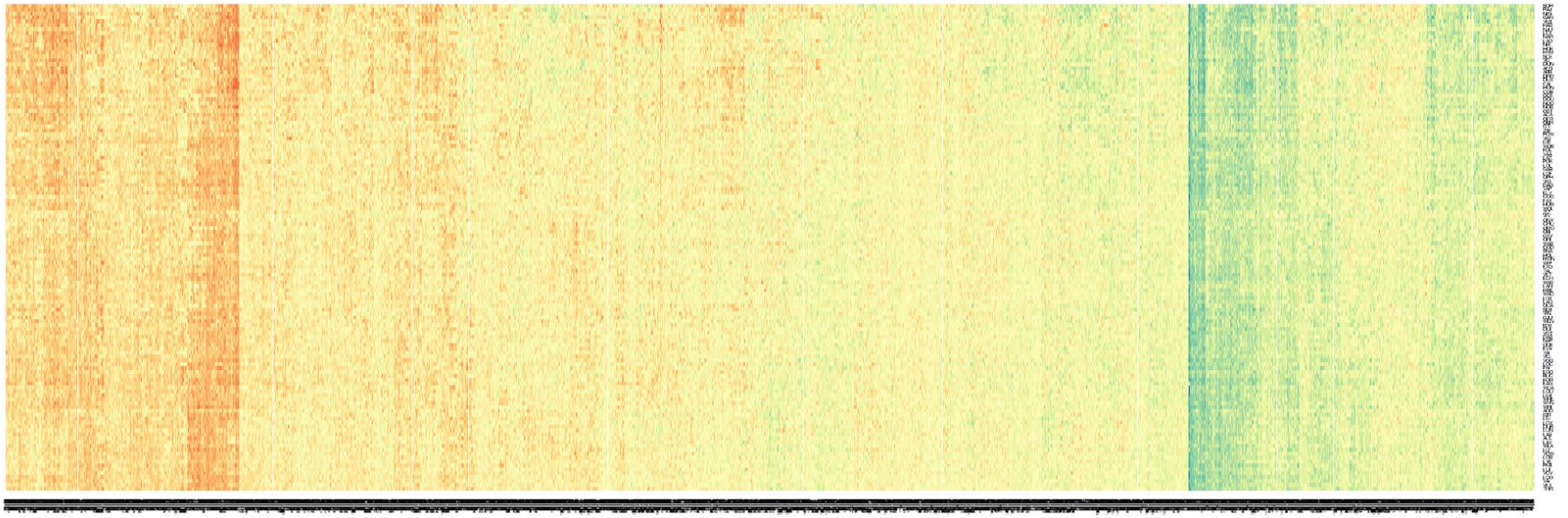
← individuals from opposite extremes of PC 1 →



← individuals from opposite extremes of PC 2 →



Subsets of microbe-microbe interactions appear to be shared between (subsets of) individuals



Functional metagenomics sample selection by partitioning around medoids (PAM)

Functional metagenomics analysis will proceed on 500-2000 selected samples

1. Collapse to *family* taxonomic level

2. Filter on reproductive annotation criteria (yields 86 individuals)

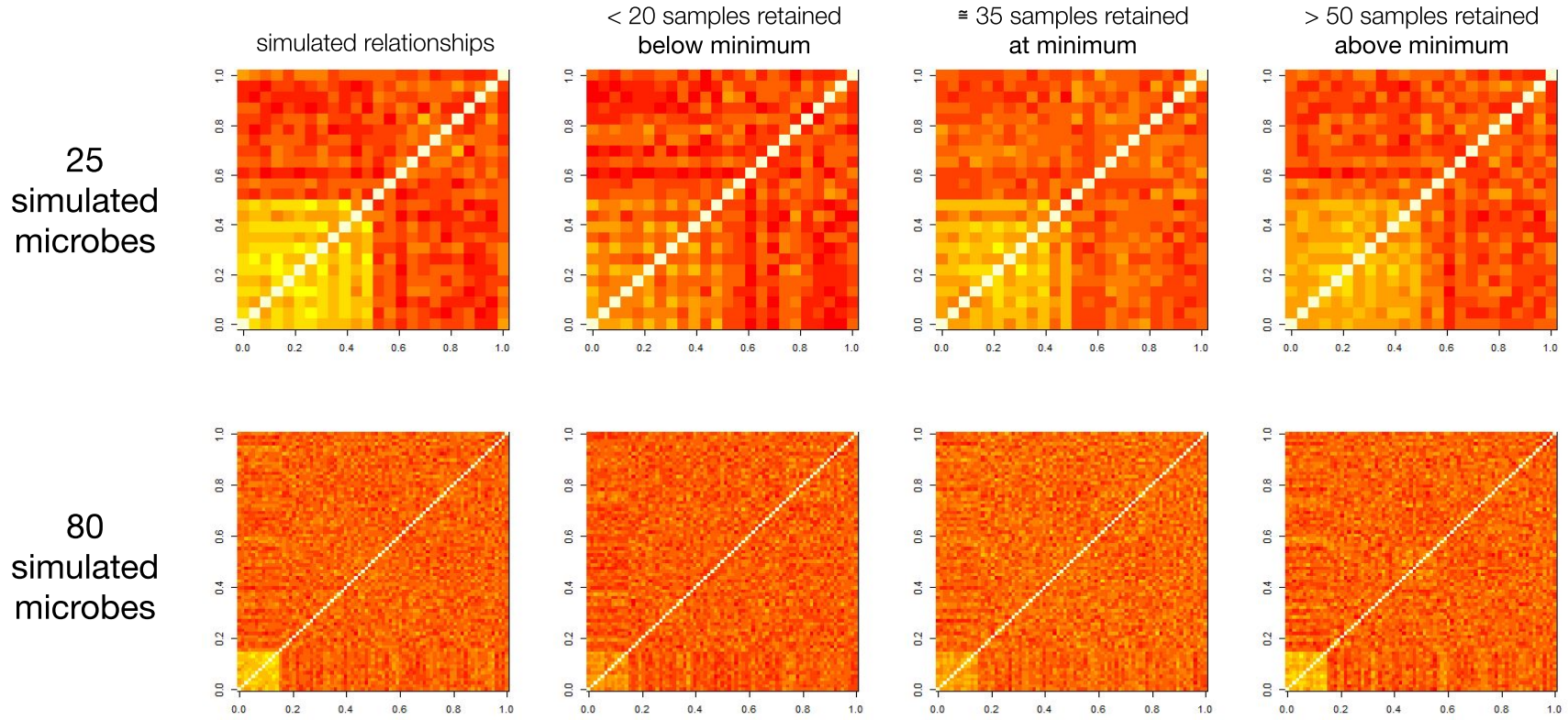
    a. number (or rate) of live births recorded
    b. number (or rate) of surviving births recorded
    c. lifespan recorded
    d. age at first live birth recorded

3. Filter on sampling inclusion criteria (yields 35 individuals); details follow

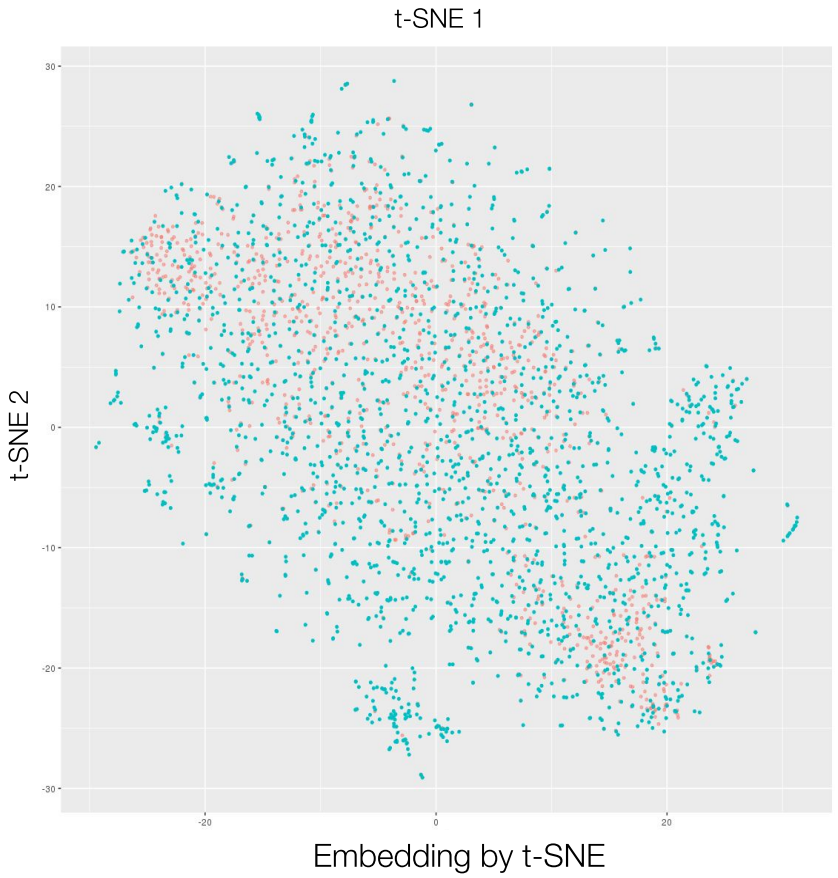
Minimum informative sample number threshold can be guided by simulation

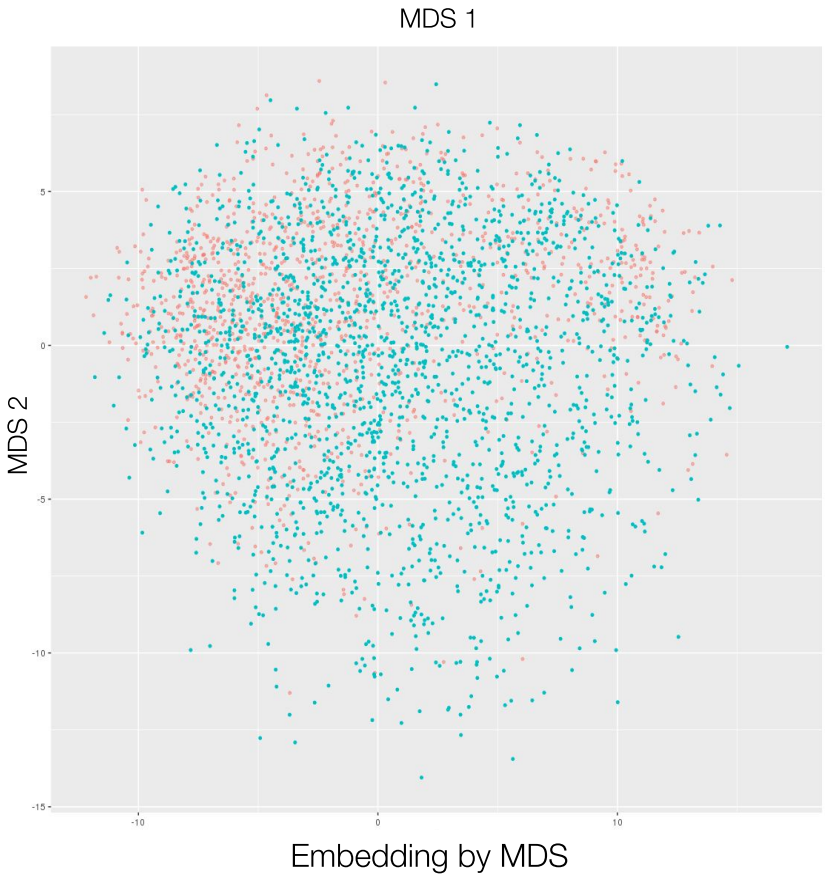
1. Simulate data with known microbial correlations, downsample data, and fit model.

2. A “minimum informative” sample number exists where microbial correlation can be recovered with some probability.



Functional metagenomics sample selection results: selected samples span the space of variation





Points represent samples from selected individuals. Turquoise points are those selected for functional metagenomics by PAM. These span the space of individual variation in gut microbial composition.

A3 Supplement

Distances in terms of microbial dynamics do not appear to associate with differences in selected “fitness” characteristics

	PERMANOVA R^2	PERMANOVA p-val
social group	0.047	0.000
maternal group	0.084	0.000
rank of mother	0.009	0.038
born in drought	0.008	0.580
born into large group	0.008	0.729
mother died	0.008	0.922
had competing sibling	0.009	0.019
early adversity	0.032	0.566
rate of live births (F)	0.174	0.271
rate of surviving births (F)	0.136	0.123