Beyond Normalization: Incorporating Scale Uncertainty in ALDEx2

Michelle Nixon

May 13, 2024

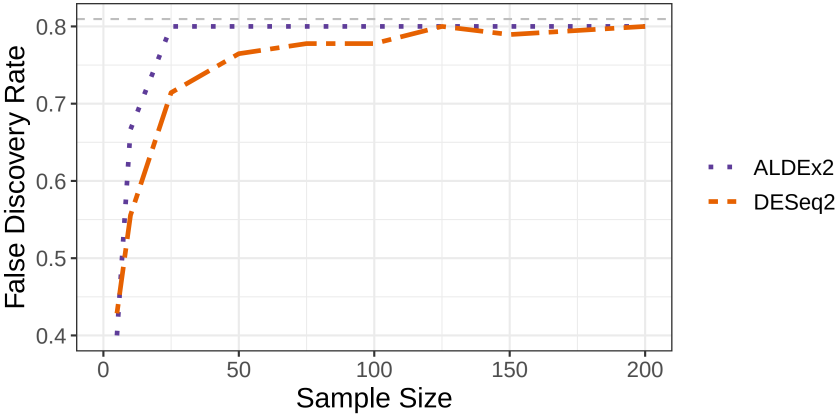
Recap: Sequencing depth can confound conclusions.

Observed data (Y)	Sample 1	Sample 2	Sample 3	
Condition	Health	Health	Disease	Conclusion
Entity 1	5	10	100	Increase
Entity 2	10	25	3	Decrease
Entity 3	0	1	8	Increase
Entity 4	0	0	19	Increase
Sequencing Depth	15	36	130	

This can mislead analyses.

System data (W)	Sample 1	Sample 2	Sample 3	
Condition	Health	Health	Disease	Conclusion
Entity 1	227	351	154	Decrease
Entity 2	684	891	3	Decrease
Entity 3	48	32	15	Decrease
Entity 4	43	39	27	Decrease
Scale ($W^{\perp}$)	1,002	1,313	200	

. . . and lead to unacknowledged bias.

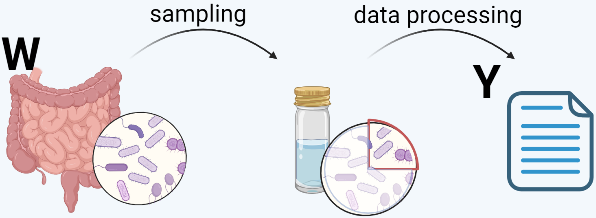


Section 1

Problem Set-Up

Observed Data as a Sample from the System

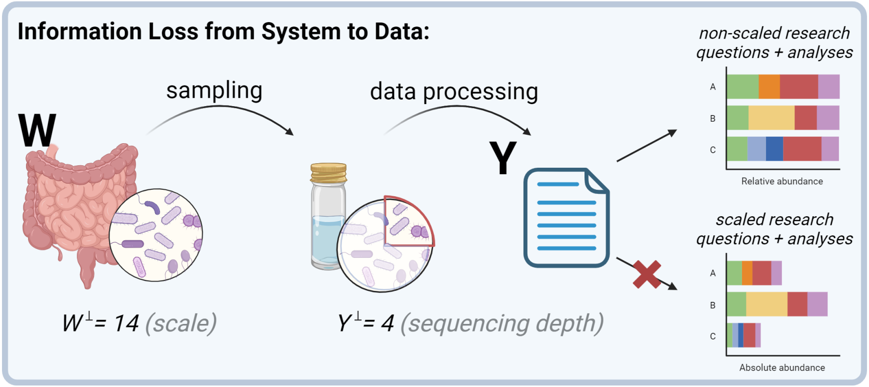
Information Loss from System to Data:



$W^{\perp}$= 14 (scale)

$Y^{\perp}$= 4 (sequencing depth)

Observed Data as a Sample from the System



Information Loss from System to Data:

W
sampling
data processing
Y
non-scaled research questions + analyses
A
B
C
Relative abundance
scaled research questions + analyses
A
B
C
Absolute abundance

$W^{\perp}= 14$ (scale) $Y^{\perp}= 4$ (sequencing depth)

Notation

- $\mathbf{Y}$: a measurement of the underlying system $\mathbf{W}$.

$$\mathbf{W}_{dn} = \underbrace{\mathbf{W}^{\parallel}_{dn}}_{\text{composition}} \times \underbrace{W^{\perp}_{n}}_{\text{scale}}$$

Notation

- **Y**: a measurement of the underlying system **W**.

$$\mathbf{W}_{dn} = \underbrace{\mathbf{W}^{\parallel}_{dn}}_{\text{composition}} \times \underbrace{W^{\perp}_{n}}_{\text{scale}}$$

- **Compostion:** $\mathbf{W}^{\parallel}_{dn} = \frac{\mathbf{W}_{dn}}{\sum_{d=1}^{D} \mathbf{W}_{dn}}$
- **Scale:** $W^{\perp}_{n} = \sum_{d=1}^{D} \mathbf{W}_{dn}$

Example: Notation

- Consider a simple study of the microbiome pre/post antibiotic administration.

System data ($W^{\parallel}$)	Sample 1	Sample 2	Sample 3
Condition	Pre	Pre	Post
Entity 1	0.27	0.27	0.77
Entity 2	0.68	0.68	0.02
Entity 3	0.05	0.02	0.08
Entity 4	0.04	0.03	0.13
Scale ($W^{\perp}$)	1,002	1,313	200

Differential Abundance/Expression Analysis

- **Question:** How do entities (e.g., taxa or genes) change between conditions?
- $\boldsymbol{\theta}$: what we want to estimate.

Differential Abundance/Expression Analysis

- **Question:** How do entities (e.g., taxa or genes) change between conditions?
- $\boldsymbol{\theta}$: what we want to estimate. (Log Fold Change)

$$\theta_d = \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn})$$

The Original ALDEx2 Model

Step 1: Model Sampling Uncertainty

$$\begin{aligned} \mathbf{Y}_{\cdot n} &\sim \text{Multinomial}(\mathbf{W}^{\parallel}_{\cdot n}) \\ \mathbf{W}^{\parallel}_{\cdot n} &\sim \text{Dirichlet}(\alpha) \end{aligned}$$

Step 2: Centered Log-Ratio Transformation

$$\log \mathbf{W}_{\cdot n} = \left[\log \mathbf{W}^{\parallel}_{1n} - \text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n}), \ldots, \log \mathbf{W}^{\parallel}_{Dn} - \text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n})\right]$$

Step 3: Calculate LFCs and Test if Different from Zero.

$$\theta_d = \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn})$$

Implied Assumptions about Scale

Step 1: Model Sampling Uncertainty

$$\mathbf{Y}_{\cdot n} \sim \text{Multinomial}(\mathbf{W}^{\parallel}_{\cdot n})$$
$$\mathbf{W}^{\parallel}_{\cdot n} \sim \text{Dirichlet}(\alpha)$$

Step 2: Centered Log-Ratio Transformation

$$\log \mathbf{W}_{\cdot n} = \left[\log \mathbf{W}^{\parallel}_{1n} - \text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n}), \ldots, \log \mathbf{W}^{\parallel}_{Dn} - \text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n})\right]$$

Step 3: Calculate LFCs and Test if Different from Zero.

$$\theta_d = \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn})$$

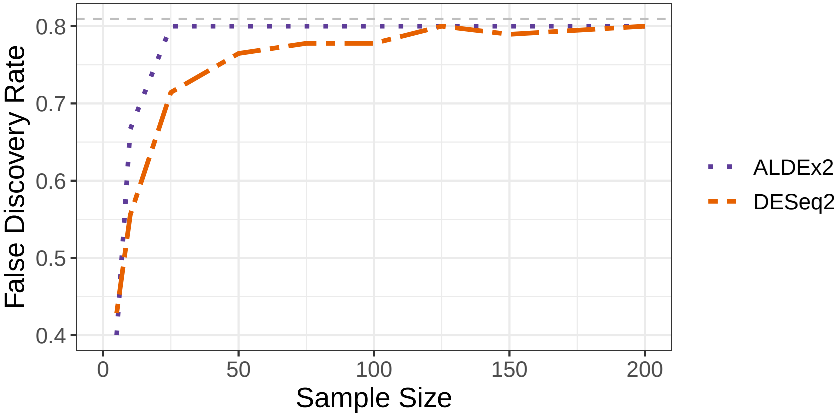
Implied Assumptions about Scale, cont.

Since $\log \mathbf{W}_{dn} = \log \mathbf{W}^{\parallel}_{dn} + \log W^{\perp}_{n}$, the CLR normalization implies:

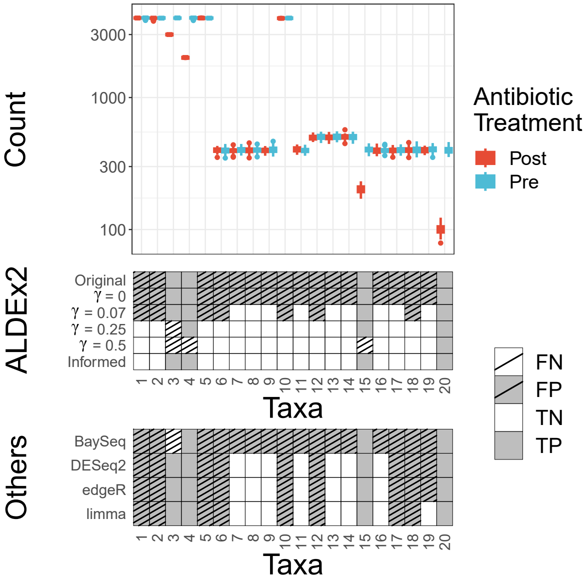
$$
\begin{aligned}
\log W_{dn} &= \log \mathbf{W}^{\parallel}_{dn} - \text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n}) \\
\log W^{\perp}_{n} &= -\text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n}).
\end{aligned}
$$

What does this mean?

Unacknowledged bias!



Adding Uncertainty in Scale can Help.



Section 2

Scale Reliant Inference

Scale Reliant Inference: The Basics

- **The CoDA perspective:** Research questions that depend on $W^{\perp}$ (scale) are not possible.
- **The Normalization perspective:** Research questions that depend on $W^{\perp}$ (scale) can be answered after normalization.
- Who is right?

Scale Reliant Inference: The Basics

- **The CoDA perspective:** Research questions that depend on $W^{\perp}$ (scale) are not possible.
- **The Normalization perspective:** Research questions that depend on $W^{\perp}$ (scale) can be answered after normalization.
- Who is right?
- **The CoDA perspective:** Yes, but this is limiting in practice.
- **The Normalization perspective:** Not correct, but attempting to answer relevant questions.

Scale Reliant Inference: The Basics

- What happens if θ depends on $W^{\perp}$?
- Consider LFCs: how are taxa changing between two conditions?

$$\begin{aligned}
\theta_d &= \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn}) \\
&= \ldots \\
&= \underbrace{\text{mean}_{\text{case}}(\log \mathbf{W}^{\parallel}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}^{\parallel}_{dn})}_{\theta^{\parallel}} \\
&\quad + \underbrace{\text{mean}_{\text{case}}(\log W^{\perp}_{n}) - \text{mean}_{\text{control}}(\log W^{\perp}_{n})}_{\theta^{\perp}}
\end{aligned}$$

Don't we need $\theta^{\perp}$?

Scale Reliant Inference: Theory Intro

Recall for LFCs:

$$\begin{aligned}\theta_d &= \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn}) \\ &= \theta^{\parallel} + \theta^{\perp}\end{aligned}$$

- What can we say about θ from $\theta^{\parallel}$ alone?

Scale Reliant Inference: Theory Intro

Recall for LFCs:

$$\begin{aligned}\theta_d &= \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn}) \\ &= \theta^{\parallel} + \theta^{\perp}\end{aligned}$$

- What can we say about θ from $\theta^{\parallel}$ alone?
- Statistical perspective: θ is not identifiable without $\theta^{\perp}$.
- Practical issues: unbiased estimators, calibrated confidence sets, and type-I error control **NOT** possible!
- See Nixon et al. (2023) for details.

Scale Simulation Random Variables

Goal: Estimate $\theta = f(\mathbf{W}^{\parallel}, W^{\perp})$.

1. Draw samples of $\mathbf{W}^{\parallel}$ from a measurement model (can depend on $\mathbf{Y}$).
2. Draw samples of $W^{\perp}$ from a scale model (can depend on $\mathbf{W}^{\parallel}$).
3. Estimate samples of $\theta = f(\mathbf{W}^{\parallel}, W^{\perp})$.

Section 3

Updated ALDEx2 Model

Rewind: $\theta^{\perp}$

$$\theta^{\perp} = \text{mean}_{\text{case}}(\log W_n^{\perp}) - \text{mean}_{\text{control}}(\log W_n^{\perp})$$

Rewind: $\theta^{\perp}$

$$\theta^{\perp} = \text{mean}_{\text{case}}(\log W_n^{\perp}) - \text{mean}_{\text{control}}(\log W_n^{\perp})$$

- The change in scales between conditions matters for estimating LFCs.
- The scale only needs to be known up to a constant (see Nixon et. al (2023)).

Rewind: $\theta^{\perp}$

$$\theta^{\perp} = \text{mean}_{\text{case}}(\log W_n^{\perp}) - \text{mean}_{\text{control}}(\log W_n^{\perp})$$

- The change in scales between conditions matters for estimating LFCs.
- The scale only needs to be known up to a constant (see Nixon et. al (2023)).
- Each normalization implies a value of $\theta^{\perp}$ (e.g., CLR):

$$\theta^{\perp}_{\text{CLR}} = \text{mean}_{\text{case}}(-\text{GM}(\mathbf{W}^{\parallel}_{\cdot n})) - \text{mean}_{\text{control}}(-\text{GM}(\mathbf{W}^{\parallel}_{\cdot n}))$$

ALDEx2 as an SSRV

Step 1: Model Sampling Uncertainty

$$\mathbf{Y}_{\cdot n} \sim \text{Multinomial}(\mathbf{W}^{\parallel}_{\cdot n})$$
$$\mathbf{W}^{\parallel}_{\cdot n} \sim \text{Dirichlet}(\alpha)$$

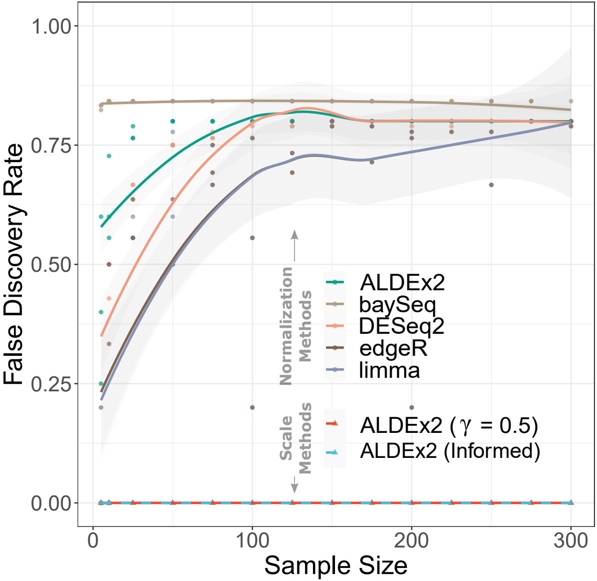
Step 2: Draw Samples from a Scale Model

$$\log W^{\perp}_{n} = -\text{mean}(\log \mathbf{W}^{\parallel}_{\cdot n}) + \epsilon,\ \epsilon \sim N(0, \gamma^2)$$
$$\log \mathbf{W}_{\cdot n} = \log \mathbf{W}^{\parallel}_{\cdot n} + \log W^{\perp}_{n}$$

Step 3: Calculate LFCs and Test if Different from Zero.

$$\theta_d = \text{mean}_{\text{case}}(\log \mathbf{W}_{dn}) - \text{mean}_{\text{control}}(\log \mathbf{W}_{dn})$$

Benefits of Moving Past Normalizations to Scale



Intro to Scale Models

There are no restrictions on what scale models can be, although there are some helpful options:

1. Based on normalizations. (Stochastic normalizations)
2. Based on biological knowledge.
3. Based on outside measurements.

Scale Models based on Biological Knowledge

What do past studies or biological mechanisms tell about the scale of the system?

Scale Models based on Biological Knowledge

What do past studies or biological mechanisms tell about the scale of the system?

1. You are confident that taking an antibiotic will kill at least some microbes in the gut.
2. A past study showed that a certain disease (e.g., Crohn's disease) leads to lower microbial load in the gut.
3. You believe the total microbial load in the mouth changes after brushing your teeth.

This type of information can be used in scale model building.

Scale Models based on Outside Measurements

How can outside measurements be used to quantify scale?

Scale Models based on Outside Measurements

How can outside measurements be used to quantify scale?

1. These measurements can be used *if* they relate to your scale of interest.
2. Examples include flow cytometry, qPCR, etc.
3. Scale models can incorporate measurement uncertainty.

Section 4

Coding Changes to ALDEx2

Including scale

The new ALDEx2 model removes normalizations in lieu of scale models.

Including scale

The new ALDEx2 model removes normalizations in lieu of scale models.

Major updates:

1. A new argument `gamma` which makes it easy to incorporate scale uncertainty.
2. A new function `aldex.senAnalysis` to see how analysis results change as a function of scale uncertainty.

The gamma argument

- Added as argument to the `aldex` and `aldex.clr` function.
- `gamma` can either be a single numeric or a matrix.
  1. Single numeric: controls the noise on the default scale model.
  2. Matrix: A $N \times S$ matrix of samples of $W^{\perp}$.
- `gamma = NULL` returns the original behavior of ALDEx2.

Option 1: Default Scale Model

The default scale model is based on errors in the CLR normalization.

$$\log \hat{W}_n^{\perp(s)} = -\mathrm{mean}\left(\log \hat{W}_{\cdot n}^{\parallel(s)}\right) + \Lambda^{\perp} x_n$$

$$\Lambda^{\perp} \sim \; N(0, \gamma^2).$$

Advantages of the Default Scale Model

1. It is built off the status quo for ALDEx2.
2. Any value of $\gamma > 0$ will reduce false positives compared to the CLR normalization.
3. It has a concrete interpretation to contextualize scale assumptions.

Interpreting the Default Scale Model

$$\log \hat{W}_n^{\perp(s)} = -\mathrm{mean}\left(\log \hat{W}_{\cdot n}^{\parallel(s)}\right) + \Lambda^{\perp} x_n$$

$$\Lambda^{\perp} \sim \; N(0, \gamma^2).$$

Empirical Rule: 95% of the samples of $\Lambda^{\perp}$ fall within $[-2\gamma, 2\gamma]$.

Interpreting the Default Scale Model, cont.

$$\log \hat{W}_n^{\perp(s)} = -\mathrm{mean}\left(\log \hat{W}_{\cdot n}^{\parallel(s)}\right) + \Lambda^{\perp} x_n$$

$$\Lambda^{\perp} \sim \; N(0, \gamma^2).$$

For case/control experiments:

1. If $x_n = 1$: add $\Lambda^{\perp}$ (95% of samples of $\log \hat{W}_n^{\perp(s)}$ fall within $[-\mathrm{mean}\left(\log \hat{W}_{\cdot n}^{\parallel(s)}\right) - 2\gamma, -\mathrm{mean}\left(\log \hat{W}_{\cdot n}^{\parallel(s)}\right) + 2\gamma]$.)
2. If $x_n = 0$: no noise added ($\log \hat{W}_n^{\perp(s)} = -\mathrm{mean}\left(\log \hat{W}_{\cdot n}^{\parallel(s)}\right)$).

Interpreting the Default Scale Model, cont.

$$\begin{aligned}\theta^{\perp}_{\text{Default Scale}} &= \text{mean}_{\text{case}}(-\text{GM}(\mathbf{W}^{\parallel}_{\cdot n})) - \text{mean}_{\text{control}}(-\text{GM}(\mathbf{W}^{\parallel}_{\cdot n})) + \Lambda^{\perp} \\ &= \theta^{\perp}_{\text{CLR}} + \Lambda^{\perp}\end{aligned}$$

Taken together, the default scale model implies that:

Interpreting the Default Scale Model, cont.

$$\begin{aligned}\theta^{\perp}_{\text{Default Scale}} &= \text{mean}_{\text{case}}(-\text{GM}(\mathbf{W}^{\parallel}_{\cdot n})) - \text{mean}_{\text{control}}(-\text{GM}(\mathbf{W}^{\parallel}_{\cdot n})) + \Lambda^{\perp} \\ &= \theta^{\perp}_{\text{CLR}} + \Lambda^{\perp}\end{aligned}$$

Taken together, the default scale model implies that:

1. The value of $\theta^{\perp}$ is within $\pm 2\gamma$ of the value of $\theta^{\perp}_{\text{CLR}}$ implied by the CLR normalization.
2. With 95% certainty, the true difference in scales falls within the the range $2^{\theta^{\perp}_{\text{CLR}} \pm 2\gamma}$.

Option 2: More Complex Scale Models

Alternatively, can pass a matrix of scale samples to `gamma` so long as:

1. The dimension is $N \times S$.
2. They are samples of $W^{\perp}$ not $\log W^{\perp}$.

Reasons to do this:

1. **Biological beliefs:** Scale is guided by the biological system or the researcher's prior beliefs.
2. **Outside Measurements:** These can be used in building a scale model *if* they are informative on the scale of interest (e.g., qPCR, flow cytometry).

Sensitivity Analyses

- Instead of picking γ, why not test over a range instead?

Sensitivity Analyses

Step 1: Model Sampling Uncertainty

$$\mathbf{Y}_{\cdot n} \sim \text{Multinomial}(\mathbf{W}^{\parallel}_{\cdot n})$$
$$\mathbf{W}^{\parallel}_{\cdot n} \sim \text{Dirichlet}(\alpha)$$

Step 2: Draw Samples from a Scale Model For a given γ:

$$\log W_n^{\perp,\gamma} = -\mathrm{mean}\left(\log \hat{W}^{\parallel(s)}_{\cdot n}\right) + \Lambda^{\perp} x_n$$
$$\Lambda^{\perp} \sim \ N(0, \gamma^2)$$
$$\log \mathbf{W}^{\gamma}{}_{\cdot n} = \log \mathbf{W}^{\parallel}_{\cdot n} + \log W_n^{\perp,\gamma}$$

Step 3: Calculate LFCs and Test if Different from Zero.

Step 4: Repeat for all desired values of γ.

Section 5

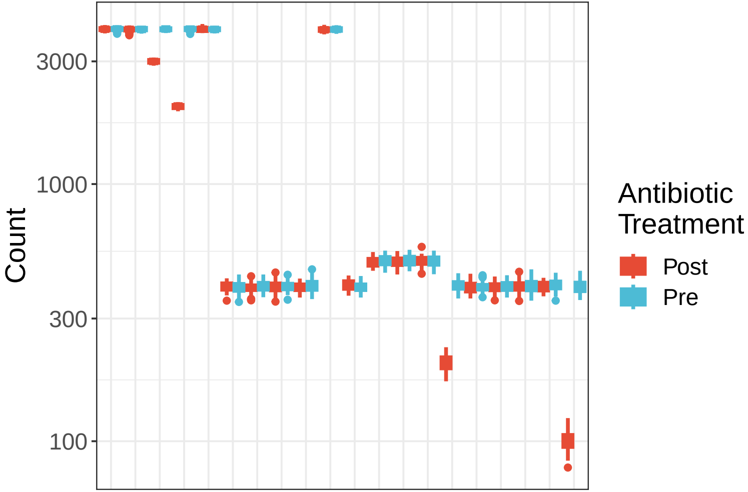
Data Examples

Simulation Study

Consider a simple study of the microbiome pre/post antibiotic administration.

- **Research question:** Which taxa change in absolute abundance after taking an antibiotic?
- 100 study participants, 50 in each condition (pre/post antibiotics).
- 20 taxa total with 4 taxa truly changing (decreasing)

Data



Adding Scale is Easy

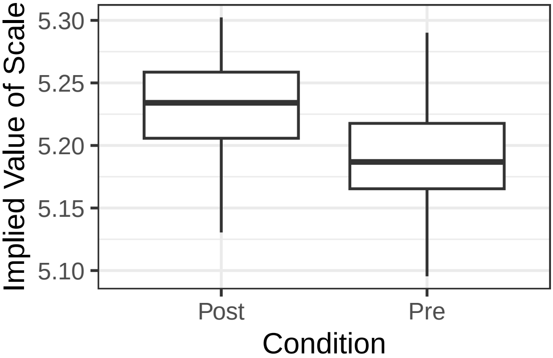
```r
## Adding noise via the default scale model
mod.ss.high <- aldex(Y, conds, gamma = 0.5)
```

Investigating Assumptions about Scale

```
## Looking at the implied scale
clr <- aldex.clr(Y, conds, gamma = 0.001)
clr@scaleSamps[1:6, 1:4]
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 5.174279 5.124890 5.199780 5.175163
## [2,] 5.175705 5.144470 5.184953 5.167715
## [3,] 5.178751 5.171188 5.130795 5.100749
## [4,] 5.158594 5.195139 5.164371 5.145696
## [5,] 5.120674 5.175533 5.189581 5.171154
## [6,] 5.208741 5.273464 5.207085 5.162631
```

Investigating Assumptions about Scale, cont.



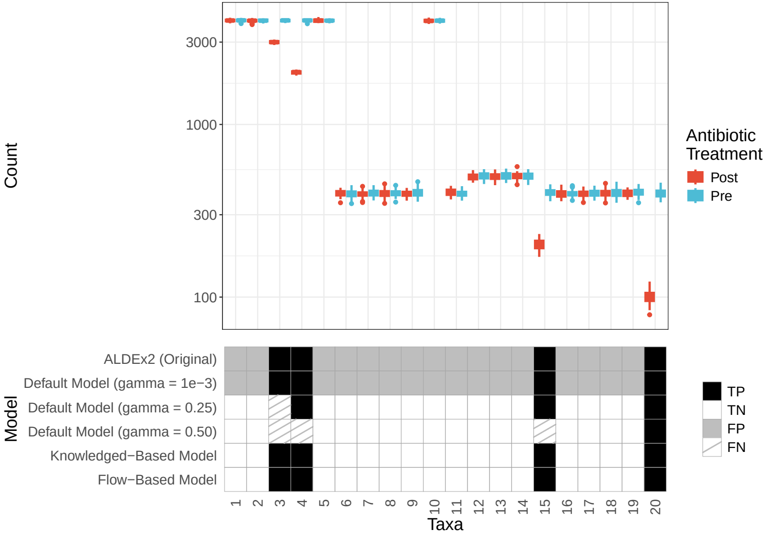
Scale Model based on Biology

```
## Creating an informed model using biological
↪  reasoning
scales <- c(rep(1, 50), rep(0.9, 50))
scale_samps <- aldex.makeScaleMatrix(gamma = 0.15, mu
↪  = scales,
    conditions = conds, log = FALSE)

mod.know <- aldex(Y, conds, gamma = scale_samps)
```

Scale Model based on Outside Measurements

```
flow_data_collapse <- flow_data %>%
    group_by(sample) %>%
    mutate(mean = mean(flow)) %>%
    mutate(stdev = sd(flow)) %>%
    dplyr::select(-flow) %>%
    ungroup() %>%
    unique()
scale_samps <- matrix(NA, nrow =
↪   nrow(flow_data_collapse), ncol = 128)
for (i in 1:nrow(scale_samps)) {
    scale_samps[i, ] <- rnorm(n = 128, mean =
↪   flow_data_collapse$mean[i],
        sd = flow_data_collapse$stdev[i])
}
mod.flow <- aldex(Y, conds, gamma = scale_samps)
```

Plotting Results

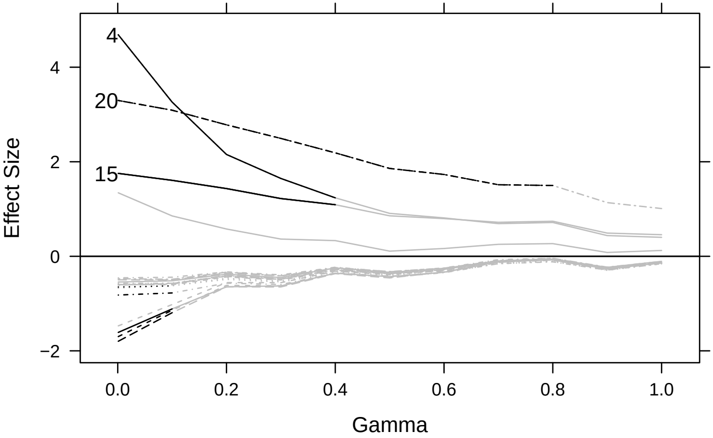


Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ALDEx2 (Original)	FP	FP	TP	TP	FP	FP	FP	FP	FP	FP	FP	FP	FP	FP	TP	FP	FP	FP	FP	TP
Default Model (gamma = 1e−3)	FP	FP	TP	TP	FP	FP	FP	FP	FP	FP	FP	FP	FP	FP	TP	FP	FP	FP	FP	TP
Default Model (gamma = 0.25)	TN	TN	FN	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TP
Default Model (gamma = 0.50)	TN	TN	FN	FN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	FN	TN	TN	TN	TN	TP
Knowledged−Based Model	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TP
Flow−Based Model	TN	TN	TP	TP	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TP	TN	TN	TN	TN	TP

Sensitivity Analyses

```
## First, specifying different values for the noise
↪  in the
## scale
gamma_to_test <- c(0.001, seq(0.1, 1, by = 0.1))

## Run the CLR function
clr <- aldex.clr(Y, conds)

## Run sensitivity analysis function
sen_res <- aldex.senAnalysis(clr, gamma =
↪  gamma_to_test)
plotGamma(sen_res, thresh = 0.1, blackWhite = TRUE,
↪  taxa_to_label = 3)
```

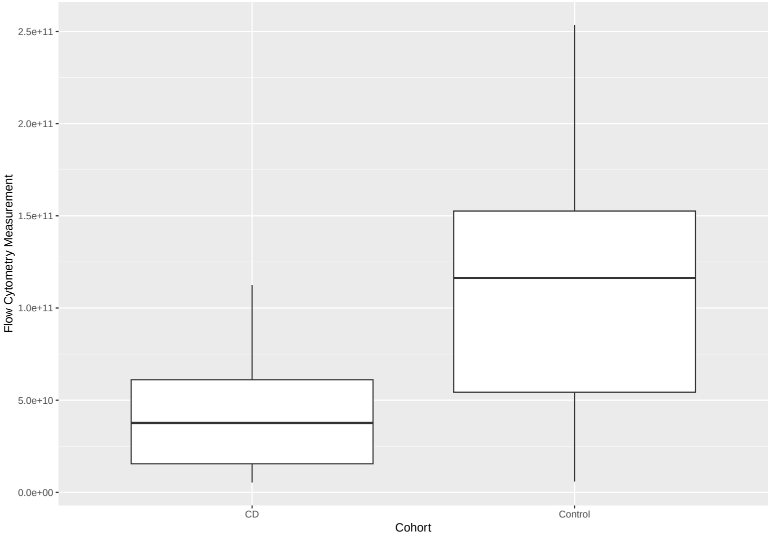
Sensitivity Analyses, cont.



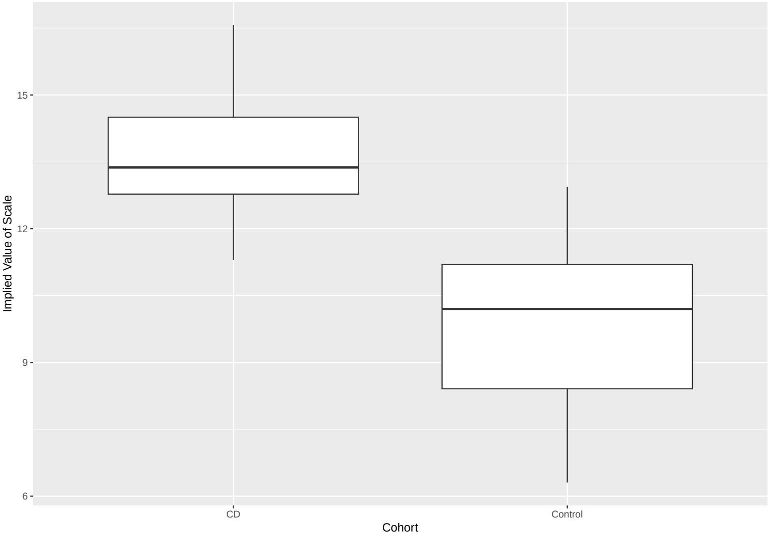
Real Example: Vandputte

1. Comparison study of 29 Crohn's disease patients and 66 healthy controls.
2. For each patient, they sequenced the fecal sample and obtained flow cytometry measurements.
3. Proposed an approach that supplemented sequence count data with flow cytometry measurements.

Difference in Scale Implied by Flow Cytometry



Difference in Scale Implied by CLR



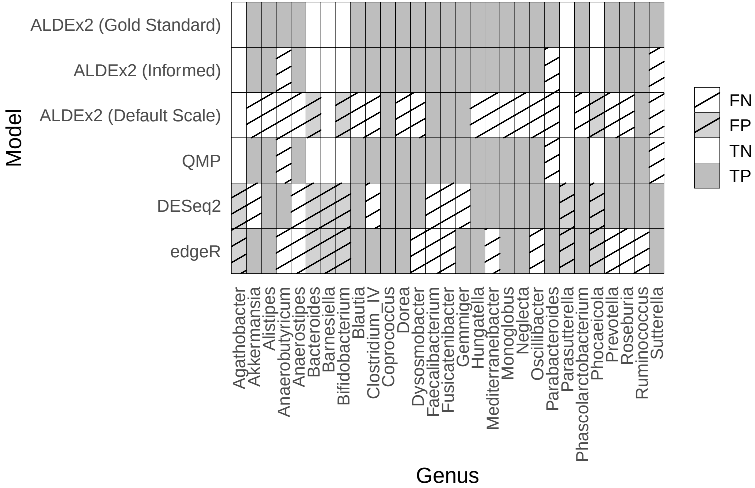
Creating a Gold Standard Model

```
scale_mean <- log2(sample_data(phylo)$CellCount)
scale_var <- rep(0.7, 95)

scale_samples <- matrix(NA, nrow = 95, ncol = 1000)
for (i in 1:95) {
    scale_samples[i, ] <- 2^rnorm(1000,
↪   scale_mean[i], scale_var[i])
}
```

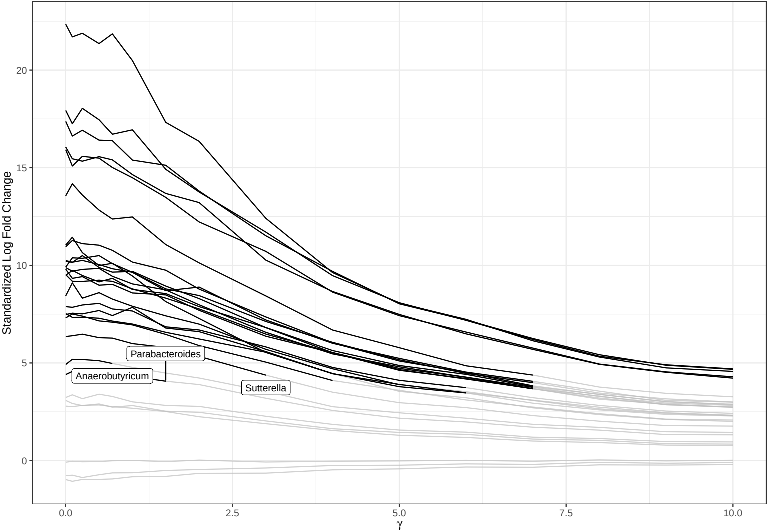
Creating an Informed Model

```
scale.cd <- 2^matrix(rnorm(1000 * 29, mean =
↪   log2(0.7), sd = 0.125),
    nrow = 29)
scale.control <- 2^matrix(rnorm(1000 * 66, mean =
↪   log2(1), sd = 0.125),
    nrow = 66)

scale.inf <- rbind(scale.cd, scale.control)
aldex_informed <- aldex(Y, X, mc.samples = 1000,
↪   gamma = scale.inf)
```

Comparing to Other Methods



Sensitivity Analyses



References

Scale Reliant Inference/Updates to ALDEx2:

- Nixon, et. al. (2023) "Scale Reliant Inference." *ArXiv Preprint 2201.03616*.
- Gloor, Nixon, and Silverman. (2023) "Scale is Not What You Think; Explicit Scale Simulation in ALDEx2." *BioRXiv Preprint 2023.10.21.563431*.
- Nixon, Gloor, and Silverman. (2024) "Beyond Normalizations: Incorporating Scale Uncertainty in ALDEx2." *BioRXiv Preprint 2024.04.01.587602*.
- Fernandes et. al. (2014). "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis." *Microbiome*.

References

Data Sources:

- McMurrough et. al. (2014)."Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues." *PNAS*.
- Vandputte et. al. (2017). "Quantitative microbiome profiling links gut community variation to microbial load." *Nature*.