Scale Uncertainty in ALDEx2

Michelle Nixon

May 13, 2024

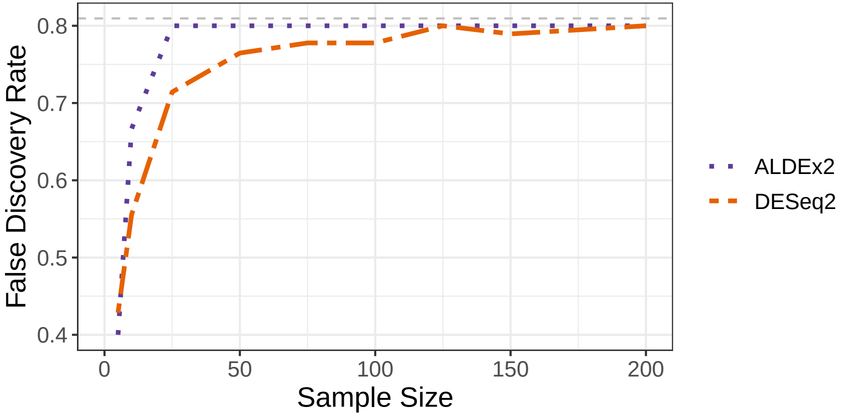
Recap: Sequencing depth can confound conclusions.

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
Sampling 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	

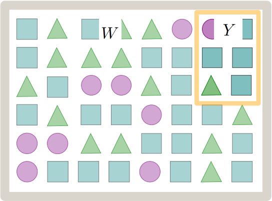
. . . and lead to unacknowledged bias.



Section 1

Problem Set-Up

Observed Data as a Sample from the System



■, ●, ▲ = 500 microbes

	Observed (Y)	Proportion ($Y^{\parallel}$)	Truth (W)	Proportion ($W^{\parallel}$)
●	500	0.17	4,500	0.19
■	2,000	0.66	12,000	0.50
▲	500	0.17	7,500	0.31
	3,000		24,000	

Sampling Depth ($Y^{\perp}$) Scale ($W^{\perp}$)

Notation

- Y is a measurement of the underlying system W.
- W depends on both the composition ($W^{\parallel}_{dn}$) and system scale ($W^{\perp}_{n}$):

$$W_{dn} = W^{\parallel}_{dn} W^{\perp}_{n}$$

$$W^{\perp}_{n} = \sum_{d=1}^{D} W_{dn}$$

- θ is what we want to estimate.

Differential Abundance/Expression Analysis

- Question: How do entities (e.g., taxa or genes) change between conditions?
- In this case, θ is the log-fold change (LFC):

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

The Original ALDEx2 Model

Step 1: Model Sampling Uncertainty

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

Step 2: Centered Log-Ratio Transformation

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

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

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

Implied Assumptions about Scale

Step 1: Model Sampling Uncertainty

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

Step 2: Centered Log-Ratio Transformation

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

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

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

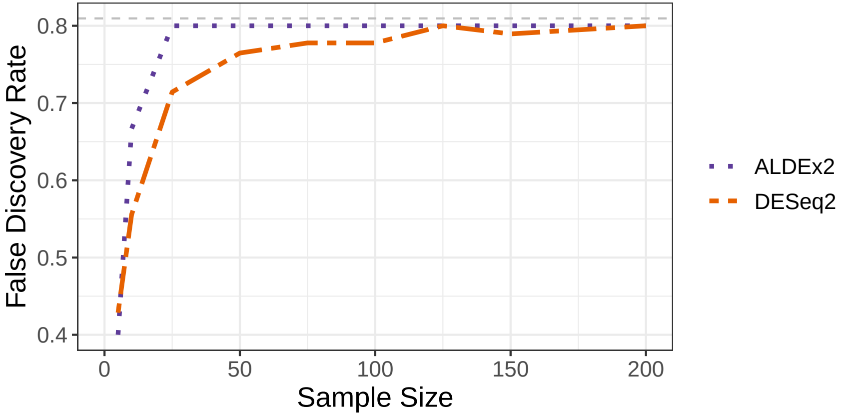
Implied Assumptions about Scale, cont.

Using the relationship $W_{dn} = W^{\parallel}_{dn} W^{\perp}_{n}$ and some math, the CLR normalization implies:

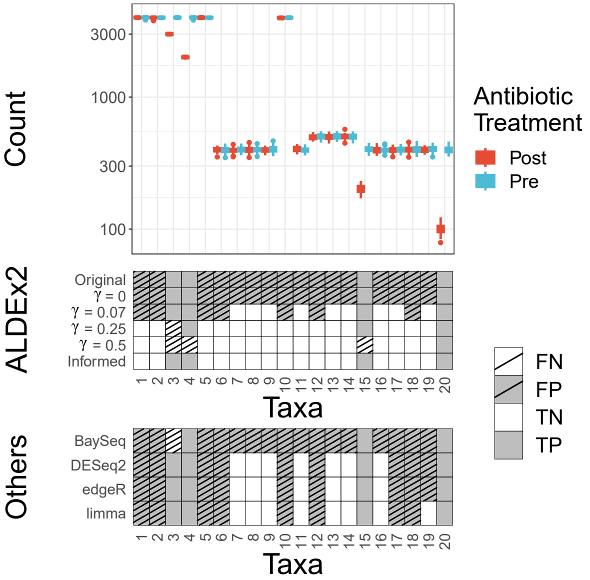
$$\log W^{\perp}_{n} = \text{mean}(\log W^{\parallel}_{\cdot n}).$$

What does this mean? What does this imply for analyses?

Unacknowledged bias!



Adding Uncertainty in Scale can Help.



Section 2

Scale Reliant Inference (Informal)

Scale Reliant Inference: The Basics

- **The CoDA perspective:** Research questions that depend on scale are not possible.
- **The Normalization perspective:** Research questions that depend on scale can be answered after normalization.
- Who is right?

Scale Reliant Inference: The Basics

- **The CoDA perspective:** Research questions that depend on scale are not possible.
- **The Normalization perspective:** Research questions that depend on scale can be answered after normalization.
- Who is right?
- **The CoDA perspective:** Technically yes, but limiting.
- **The Normalization perspective:** Technically no, 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?

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

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

Scale Reliant Inference: Theory Intro

Recall for LFCs:

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

- What can we say about θ from $\theta^{\parallel}$ alone?
- E.g. If $\theta^{\parallel} = 20$, what does that say about θ? If there are no restrictions, nothing!
- Statistical perspective: θ is not identifiable without $\theta^{\perp}$.
- Practical issues: unbiased estimators, calibrated confidence sets, and type-I error control *NOT* possible!

Scale Simulation Random Variables

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

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

Section 3

The Updated ALDEx2 Software

ALDEx2 as an SSRV

Step 1: Model Sampling Uncertainty

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

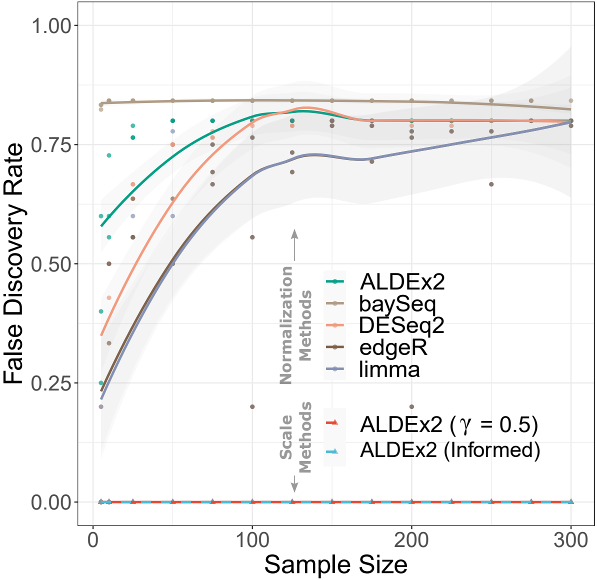
Step 2: Draw Samples from a Scale Model

$$\log W^{\perp}_{n} \sim Q$$
$$\log W_{\cdot n} = \log W^{\parallel}_{\cdot n} + \log W^{\perp}_{n}$$

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

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

Benefits of Moving Past Normalizations to Scale



Intro to Scale Models

Normalizations are replaced by a scale model:

$$\log W_n^{\perp} \sim Q$$

There are no restrictions on Q, although there are some helpful options:

1. Based on normalizations.
2. Based on biological knowledge.
3. Based on outside measurements.

Section 4

Coding Changes to ALDEx2

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.
- `gamma = NULL` returns the default 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 samples of $\Lambda^{\perp}$ fall within a factor of $\pm 2\gamma$.

Interpreting the Default Scale Model, cont.

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

}

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

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

Section 5

Real Data Examples

Simulation Study

Real Example: SELEX

Real Example: Vandputte