Scale Uncertainty in ALDEx2

Michelle Nixon

May 13, 2024

Overview

- These slides are by no means polished.
- Idea: Use a simulation, selex, and Vandputte to introduce SRI + SSRVs + modifications to ALDEx2
- My goals:
- Part 1: Introduce notation (W, Y, theta), apply this notation to the ALDEx2 model, show the source of unacknowledged bias in ALDEx2, connect to SRI/SSRVs
- Part 2: Discuss ALDEx2 as an SSRV and the modifications that we made to ALDEx2.
- Part 3: Real data examples

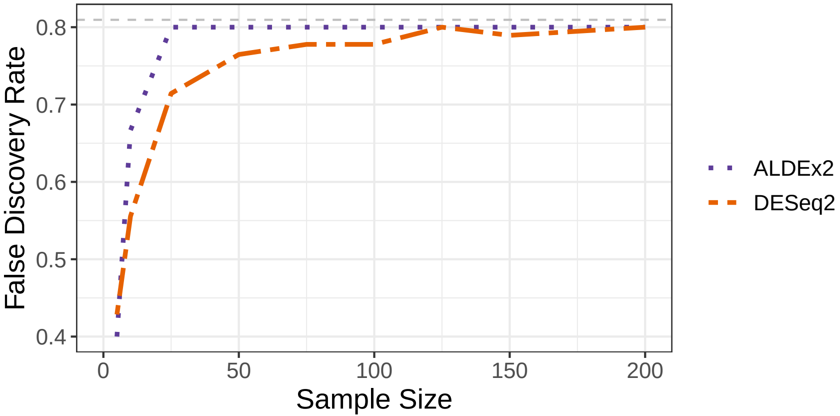
Recap: Sequencing depth can confound conclusions.

Observed data (Y)	Sample 1	Sample 2	Sample 3	Conclusion
Condition	Health	Health	Disease	
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	

Recap: This can mislead analyses.

System data (W)	Sample 1	Sample 2	Sample 3	Conclusion
Condition	Health	Health	Disease	
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	

Recap: ... and lead to unacknowledged bias.



Problem Set-Up

Observed Data as a Sample from the System

Differential Abundance/Expression Analysis

The Original ALDEx2 Model

Implied Assumptions about Scale

Unacknowledged bias in ALDEx2

Scale Reliant Inference (Informal)

Scale Reliant Inference: The Basics

- Y is a measurement of the underlying system W.
- Desired quantity depends on W (i.e., $\theta = f(W)$). However, W depends on both the composition ($W^{\parallel}_{dn}$) and system scale ($W^{\perp}_{n}$):

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

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

Scale Reliant Inference: The Basics

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

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

What if we have outside information on $W^{\perp}$?

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

Scale Reliant Inference: Theory Intro

Consider the case of 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!

The Updated ALDEx2 Software

Moving Past Normalizations to Scale

ALDEx2 as an SSRV

Coding Changes to ALDEx2

Including scale

Option 1: Default Scale Model

Option 2: More Complex Scale Models

Sensitivity Analyses

Real Data Examples

Real Example: SELEX

Real Example: Vandputte