Relative Abundance Estimation with Microbiome Data

David Clausen
5/17/2021

High-Throughput Sequencing of Microbial Communities

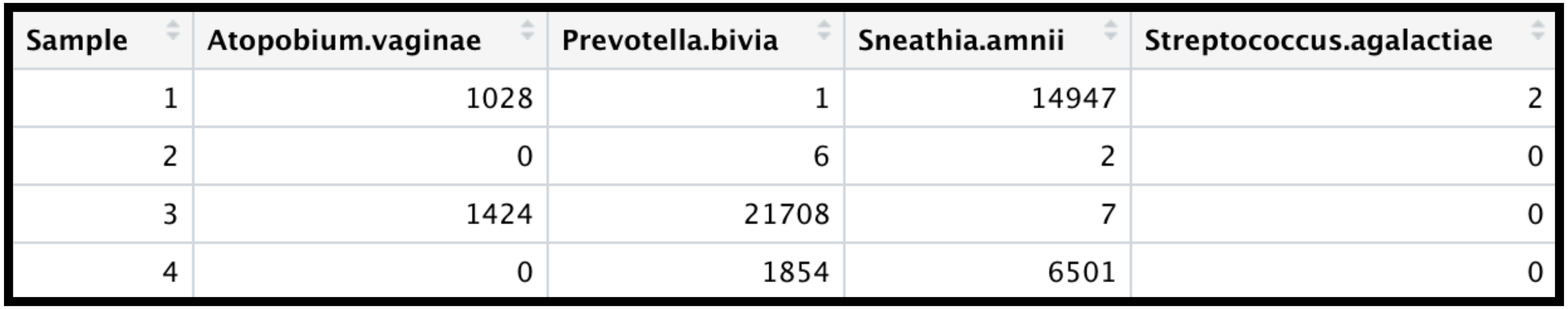
- Motivation: how can we determine what microbes live in a given environment?
  - **Isolation/culturing**
    - Useful but limited: most microbes aren’t culturable!
  - **Sequencing**
    - Attempt to detect and identify microbes by sequencing genetic material in samples
    - Feasible with high-throughput sequencing techniques developed and refined over ~ the past two decades

High-Throughput Sequencing of Microbial Communities

- High-throughput sequencing (whole-genome or marker gene)
  - Complex measurement process with many steps
    - Sample collection and storage
    - DNA extraction
    - DNA amplification
    - Sequencing
    - Taxonomic assignment

High-Throughput Sequencing of Microbial Communities

- Measurement output: table $W_{n \times J}$ of taxon counts
  - W_{ij}: count of reads assigned to taxon j in sample i

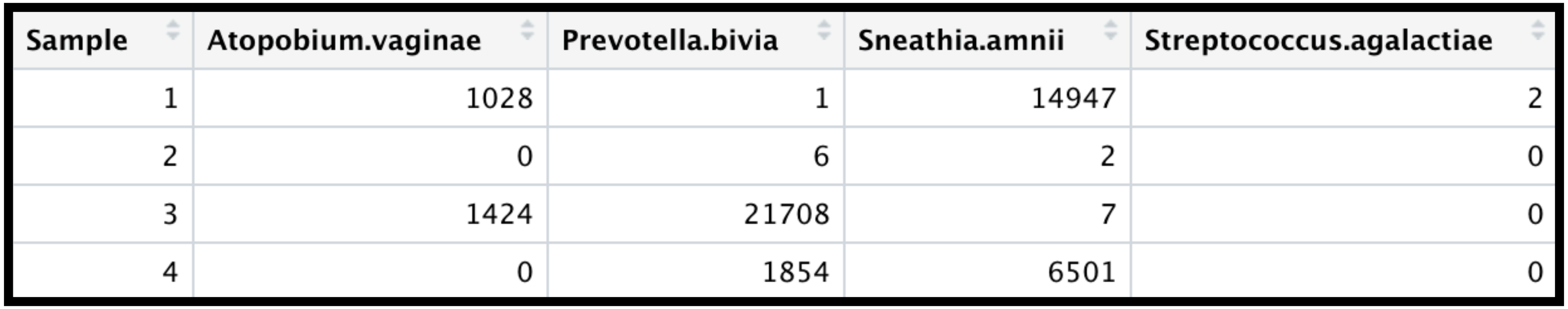


Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

What Do Read Counts (Allegedly) Tell Us?

According to microbiome folk wisdom, W_{ij} (# of reads assigned to taxon j and sample i)

- Does not to reflect “absolute abundance”
    - i.e., $W_{i'j} > W_{ij}$ does not imply that taxon j is present in higher concentration in sample i' than in sample i

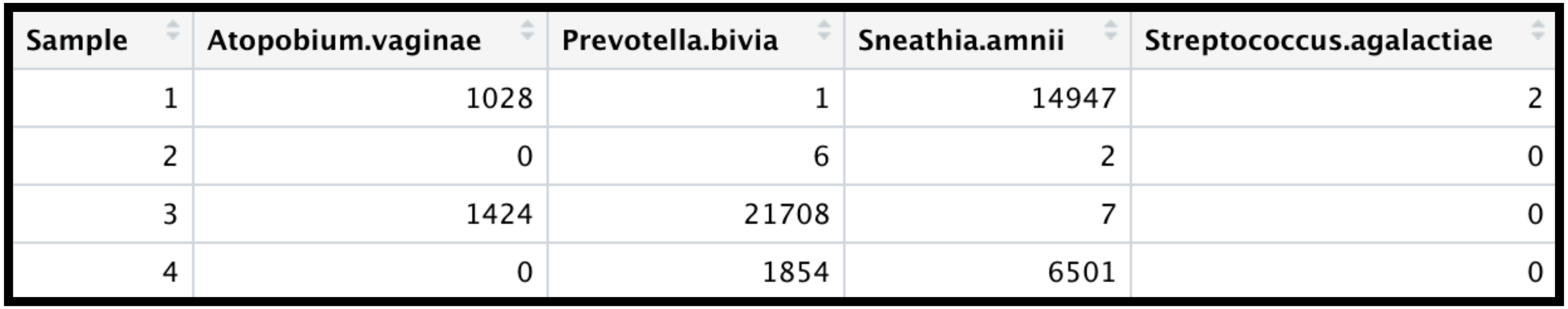


Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

What Do Read Counts (Allegedly) Tell Us?

Also according to microbiome folk wisdom, W_{ij} (# of reads assigned to taxon j and sample i)

- reflects “relative abundance” in sense that $W_{ij} \propto p_{ij}$, where p_{ij} is the true proportion of detectable microbes in sample i belonging to taxon j

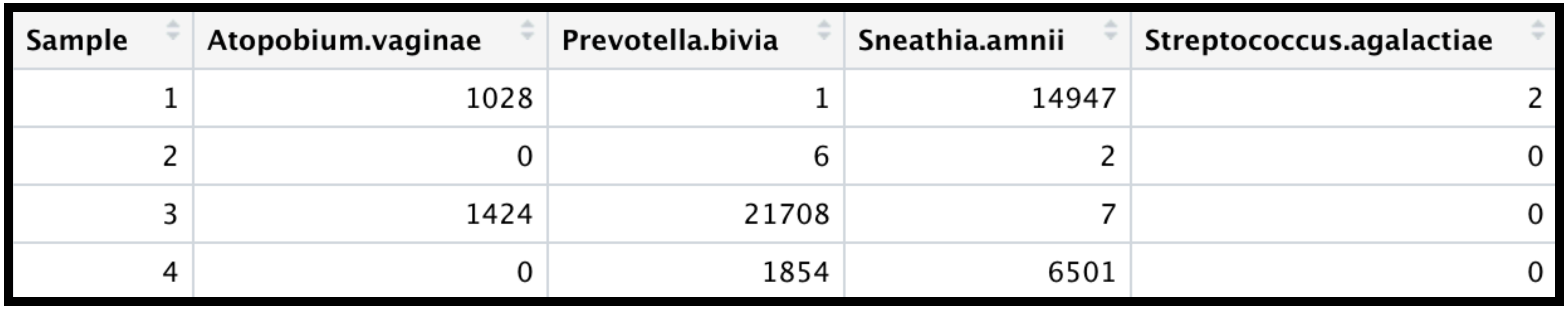


Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

What Do Read Counts (Allegedly) Tell Us?

- “Relative abundance” interpretation motivates estimator for p_{ij} (true prop. of microbes in sample i belonging to taxon j)

$$\hat{p}_{ij} = \frac{W_{ij}}{\sum_{j=1}^{J} W_{ij}}$$



Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

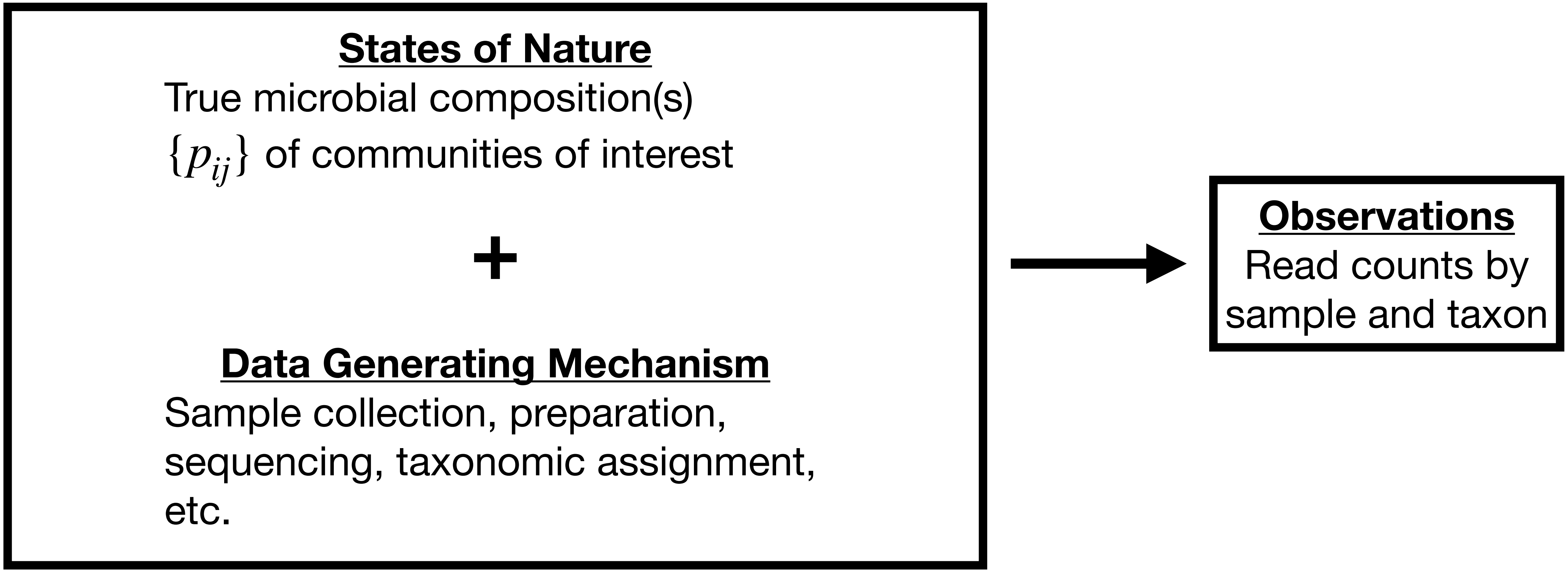
What Do Read Counts (Allegedly) Tell Us?

- "Relative abundance" interpretation motivates estimator for p_{ij} (true prop. of microbes in sample i belonging to taxon j)

$$\hat{p}_{ij} = \frac{W_{ij}}{\sum_{j=1}^{J} W_{ij}}$$

- **Focus of this talk**: $\hat{p}_{ij}$ a reasonable estimator of p_{ij}?
  - How can we evaluate performance?
  - Can we do better?

Some Statistical Framing



States of Nature

True microbial composition(s) $\{p_{ij}\}$ of communities of interest

+

Data Generating Mechanism

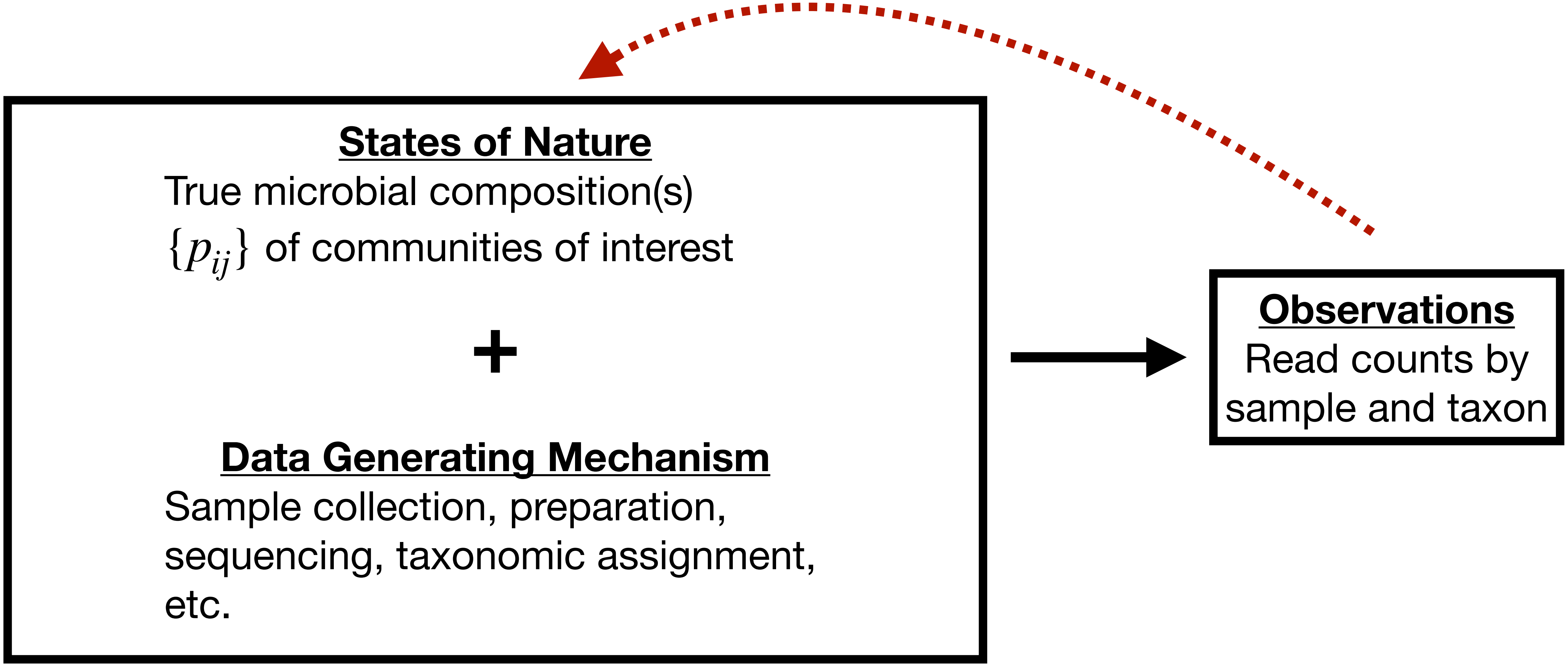
Sample collection, preparation, sequencing, taxonomic assignment, etc.

Observations

Read counts by sample and taxon

Some Statistical Framing

Goal: reason about states of nature / data-generating mechanism using observations + what we know about how they were generated



States of Nature

True microbial composition(s) $\{p_{ij}\}$ of communities of interest

+

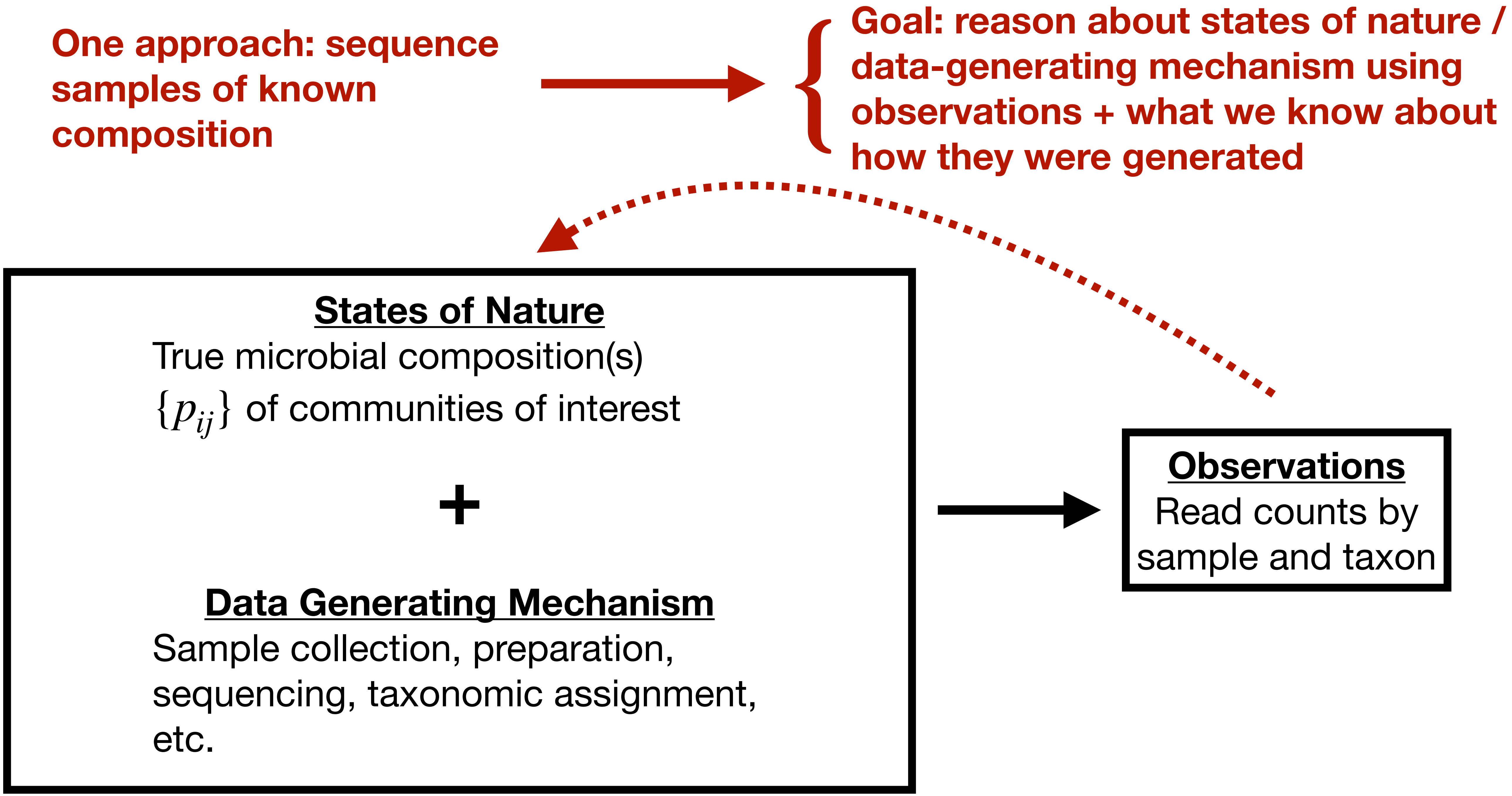
Data Generating Mechanism

Sample collection, preparation, sequencing, taxonomic assignment, etc.

Observations

Read counts by sample and taxon

Some Statistical Framing



One approach: sequence samples of known composition

Goal: reason about states of nature / data-generating mechanism using observations + what we know about how they were generated

States of Nature

True microbial composition(s) $\{p_{ij}\}$ of communities of interest

+

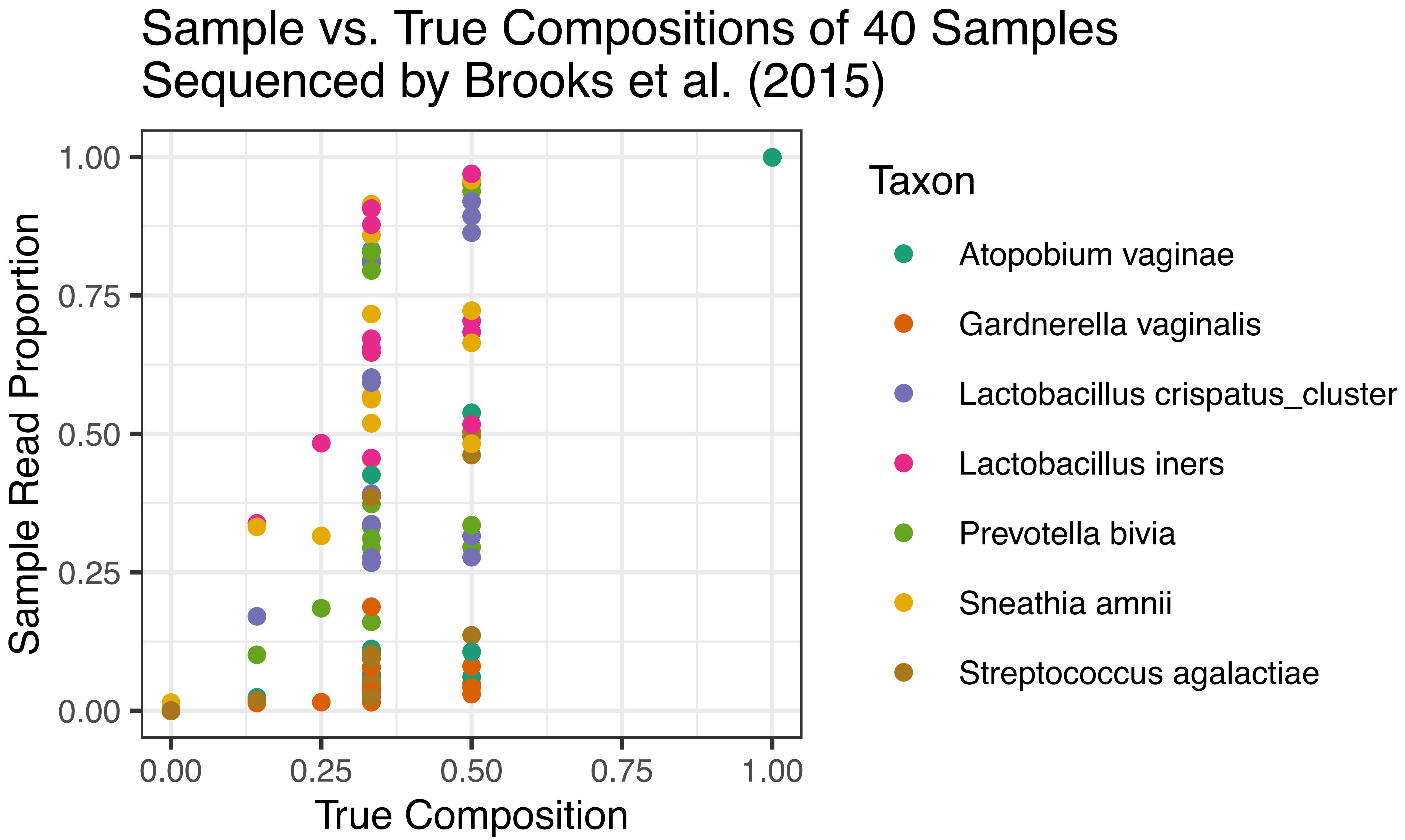
Data Generating Mechanism

Sample collection, preparation, sequencing, taxonomic assignment, etc.

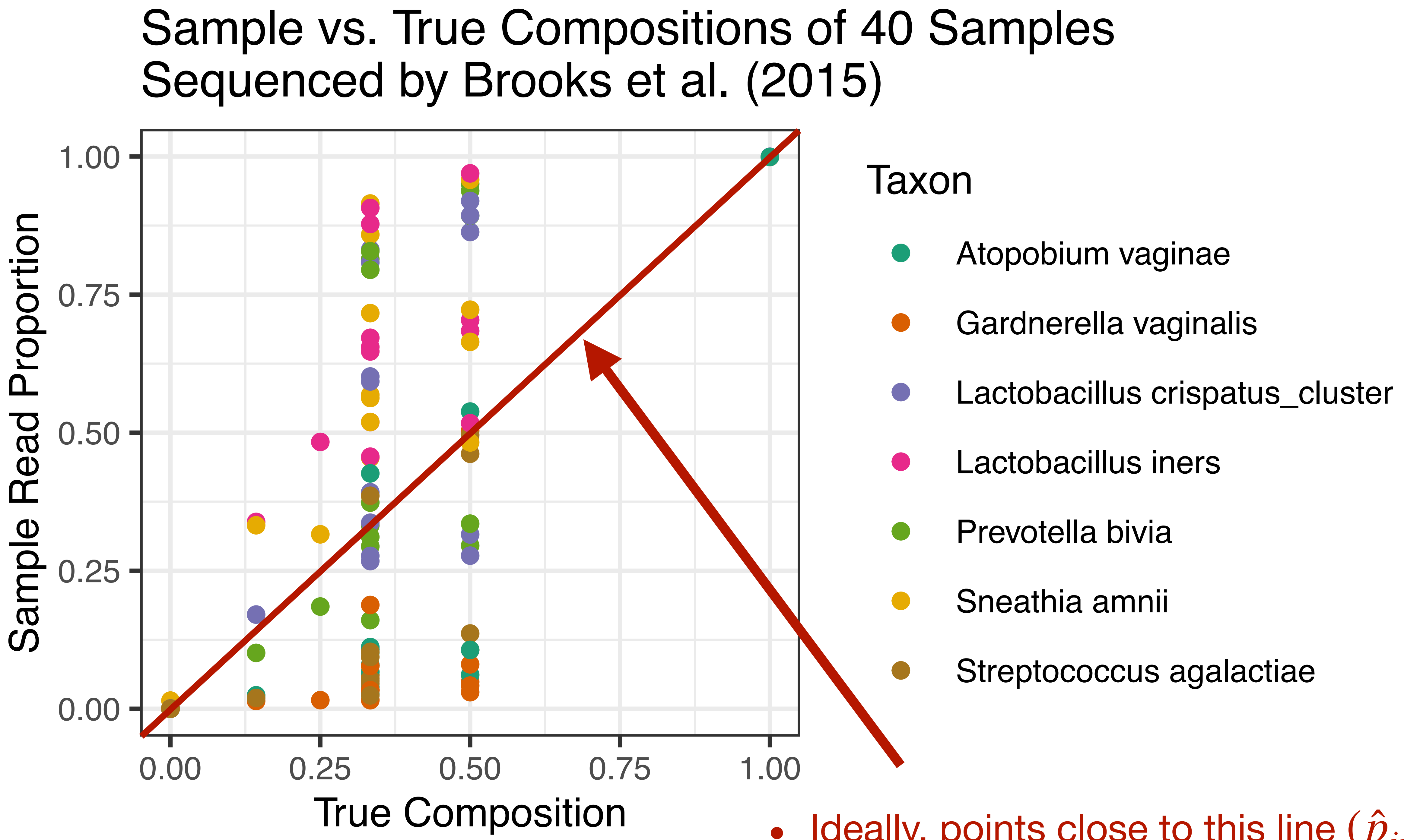
Observations

Read counts by sample and taxon

How Well Does the Naive Estimator Perform?

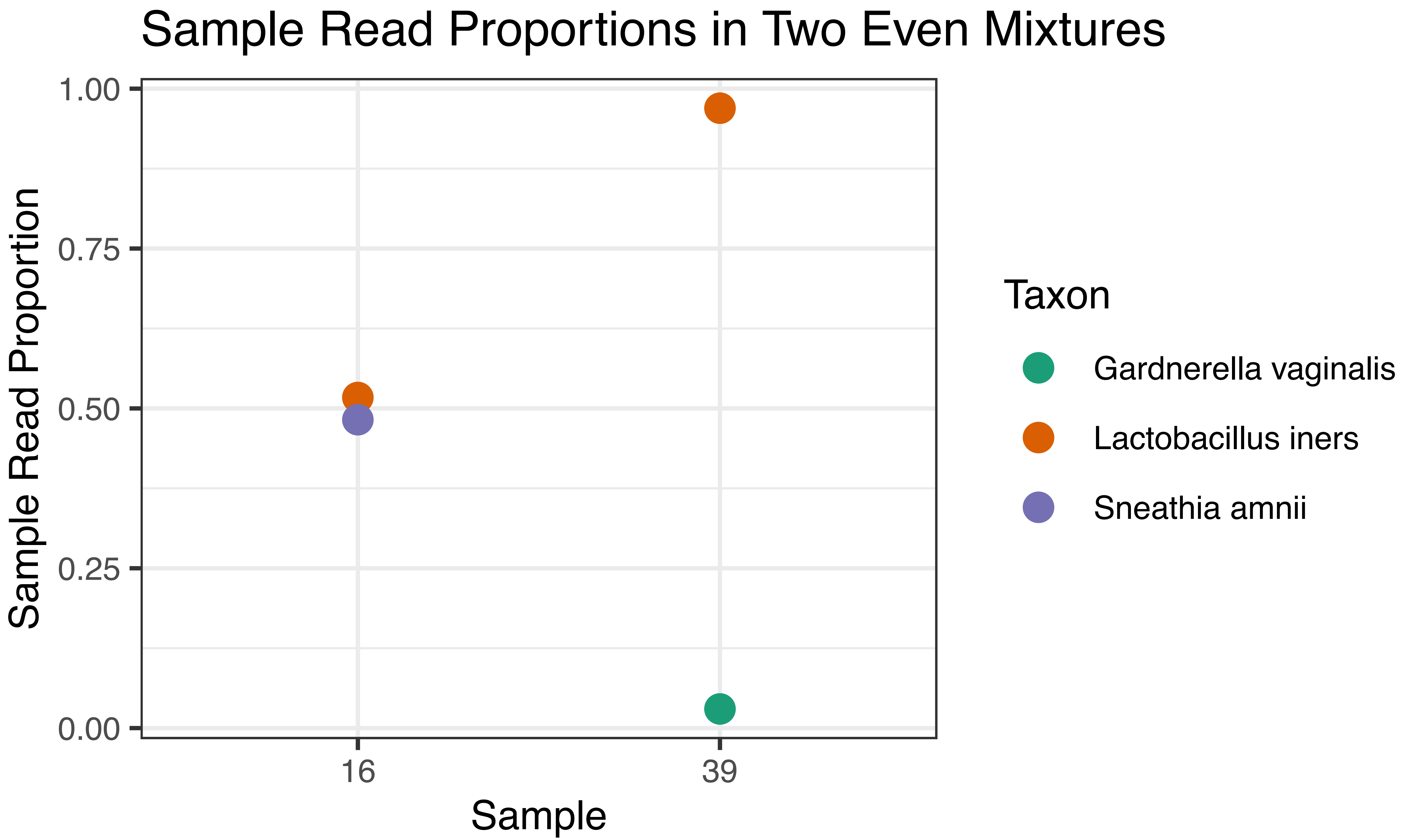


How Well Does the Naive Estimator Perform?

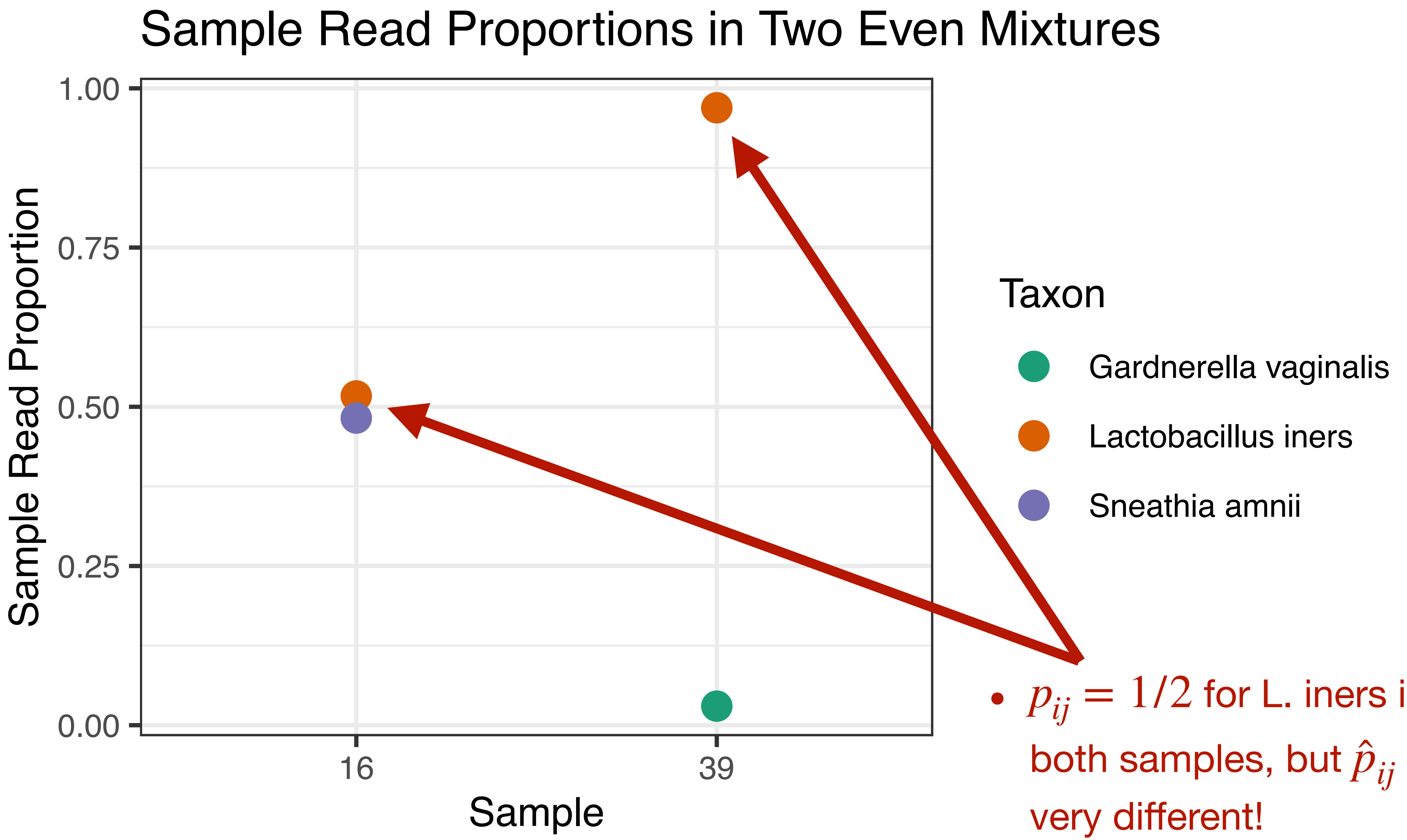


- Ideally, points close to this line $(\hat{p}_{ij} \approx p_{ij})$

How Well Does the Naive Estimator Perform?



How Well Does the Naive Estimator Perform?



- $p_{ij} = 1/2$ for L. iners in both samples, but $\hat{p}_{ij}$ very different!

What’s Going On?

- McLaren et al. (2019)
    - Observe $\hat{p}_{ij}$ does not perform well as an estimator of p_{ij}
        - Suggesting that $W_{ij} \propto p_{ij}$ does not hold in general
            - i.e., read counts across taxa in a sample are **not** approximately proportional to true relative abundances
    - Hypothesis: any given sequencing protocol will be better at detecting some microbial taxa than others

Multiplicative Distortion: “Efficiencies”

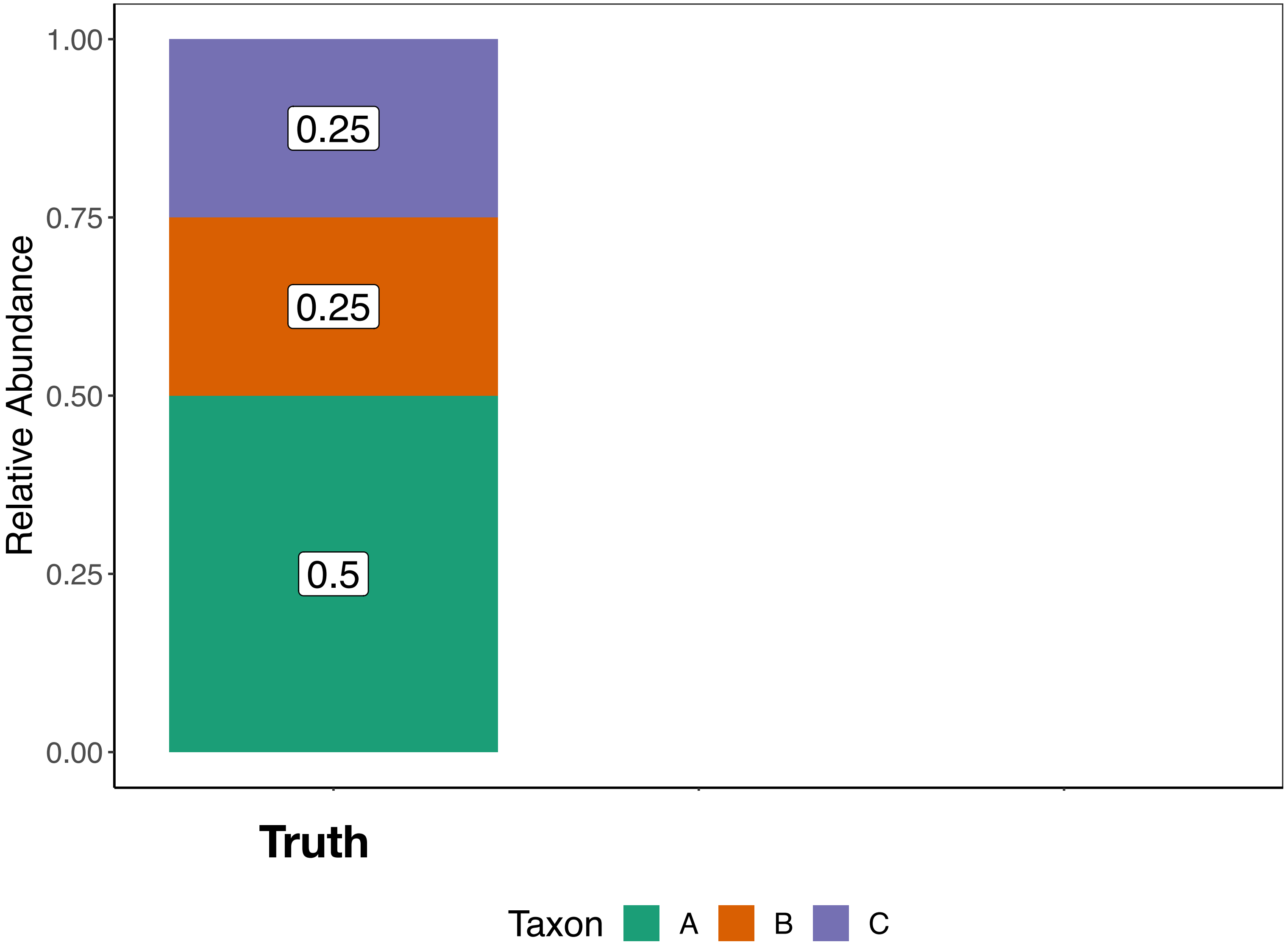
- McLaren et al. (2019)
  - A given sample handling/sequencing/postprocessing protocol will preferentially detect some microbes over others
    - Formalize this idea in terms of a detection “**efficiency**” e_j of taxon j
    - Instead of $W_{ij} \propto p_{ij}$, posit $W_{ij} \propto e_j p_{ij}$ (at least approximately)

$$\Rightarrow \hat{p}_{ij} = \frac{W_{ij}}{\sum_{j=1}^{J} W_{ij}} \approx \frac{e_j p_{ij}}{\sum_{j=1}^{J} e_j p_{ij}} \neq p_{ij} \text{ (in general)}$$

Efficiencies: An Example

$$\underbrace{\mathbb{E}[W_{i\cdot}]}_{\substack{\textbf{Expected}\\ \textbf{counts in}\\ \textbf{sample } i}} = \Big(\underbrace{\rho_{i\cdot}}_{\substack{\textbf{True relative}\\ \textbf{abundance}\\ \textbf{profile of}\\ \textbf{sample } i}} \circ \mathbf{exp}\underbrace{(\boldsymbol{\beta})}_{\substack{\textbf{Taxon-}\\ \textbf{specific}\\ \textbf{``efficiencies''}}}\Big) \cdot \underbrace{k_i}_{\substack{\textbf{Proportionality}\\ \textbf{term}}}$$

Simulated Sequencing Data on a Simple Community

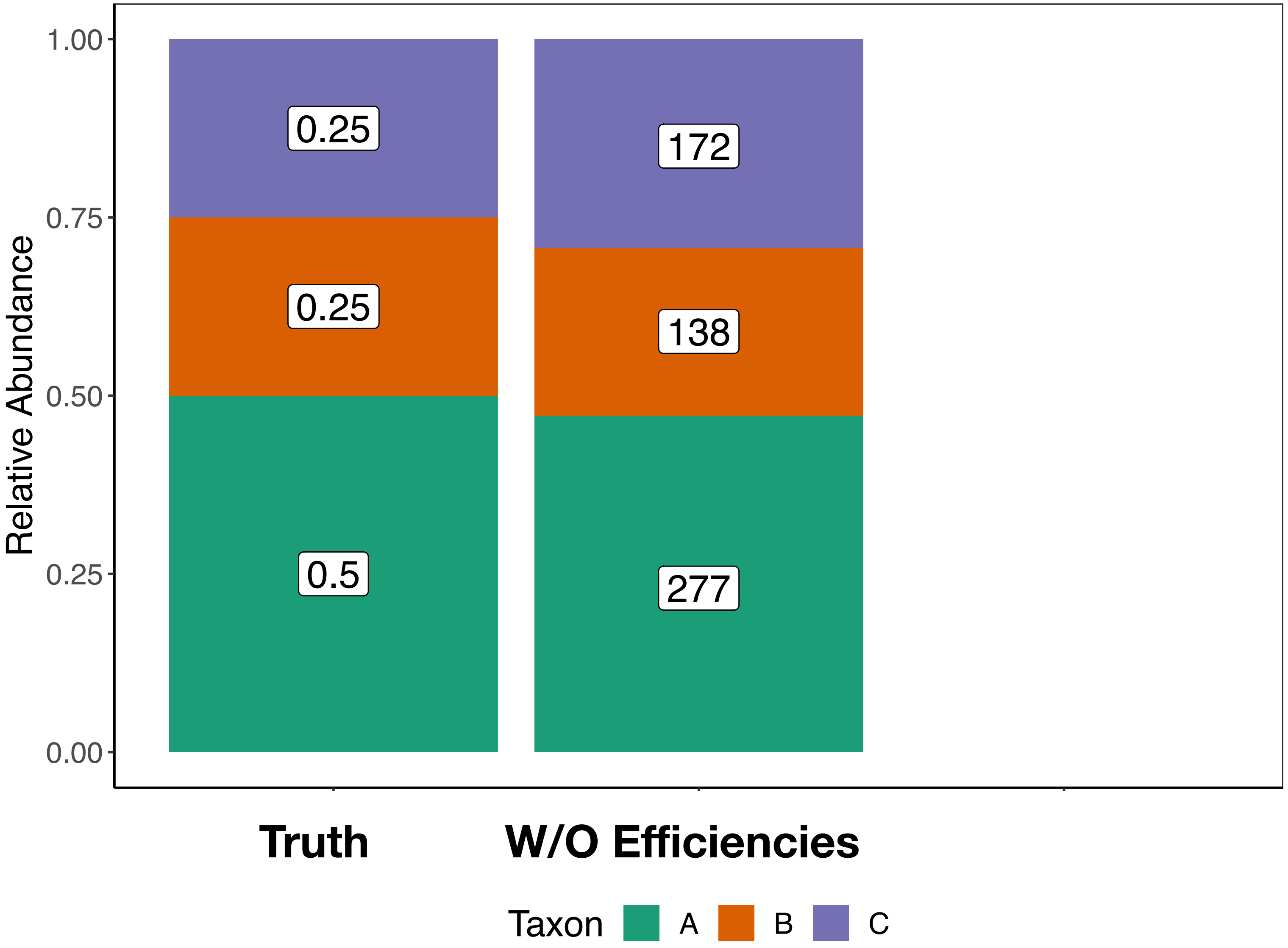


Consider a specimen

- containing taxa A, B, and C
  - in relative abundances 0.5, 0.25, and 0.25, respectively

Efficiencies: An Example

$$\underbrace{\mathbb{E}[W_{i\cdot}]}_{\text{Expected counts in sample } i} = \Big(\underbrace{\rho_{i\cdot}}_{\text{True relative abundance profile of sample } i} \circ \underbrace{\mathbf{exp}(\boldsymbol{\beta})}_{\text{Taxon-specific “efficiencies”}}\Big) \cdot \underbrace{k_i}_{\text{Proportionality term}}$$



Simulated Sequencing Data on a Simple Community

Consider a specimen

- containing taxa A, B, and C
  - in relative abundances 0.5, 0.25, and 0.25, respectively

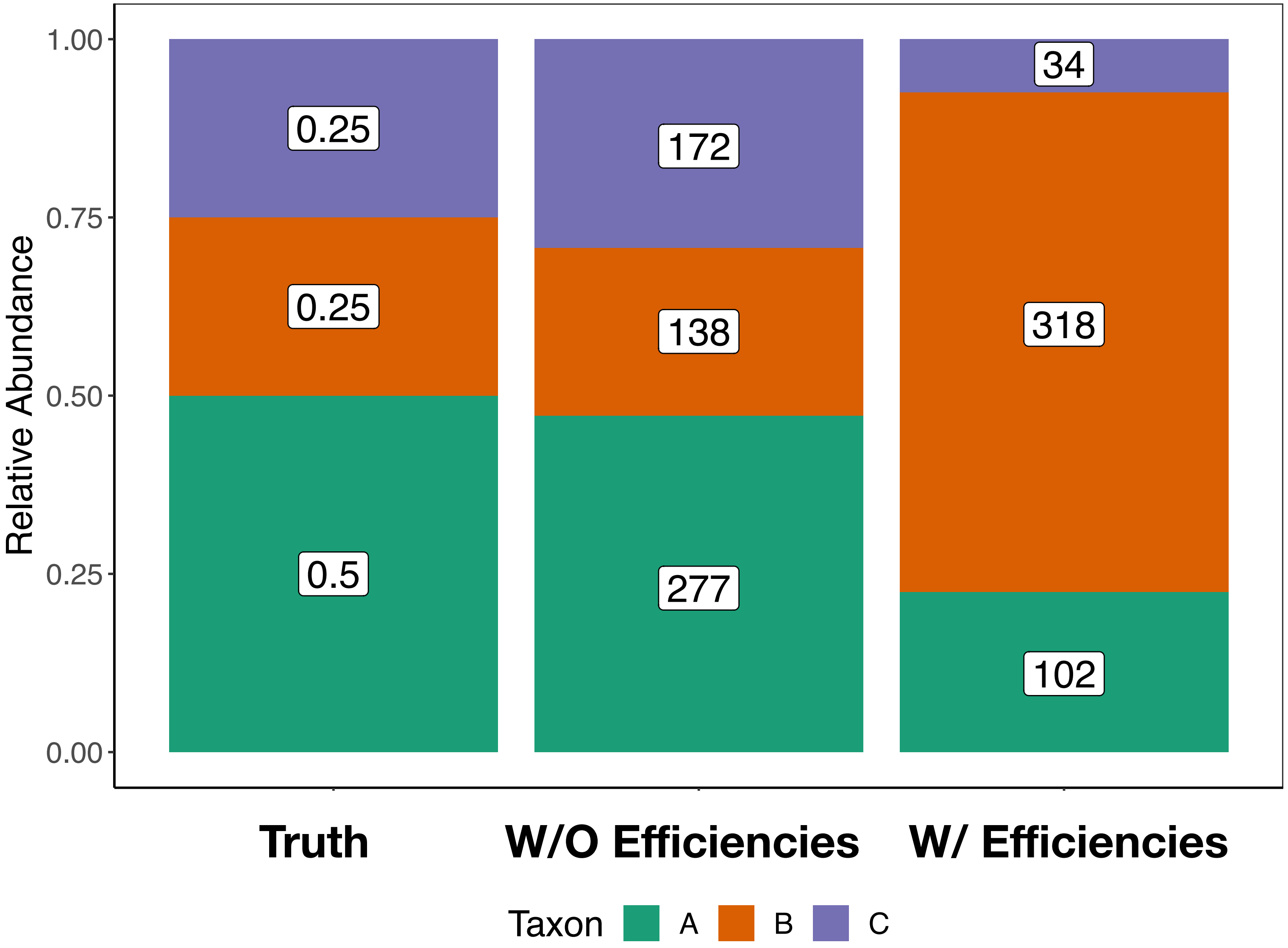
We can simulate* sequencing data under

- assumption $\mathbb{E}[W_{i\cdot}] \propto \rho_{i\cdot}$

* First setting (with same efficiencies across taxa): each count simulated as a negative binomial with mean $\mu_j = 500 * \rho_j$ and size parameter $s = 5$ (s.t. Var$(W_{ij}) = \mu_j + \mu_j^2/s$)

Efficiencies: An Example

$$\boxed{\underbrace{\mathbb{E}[W_{i\cdot}]}_{\text{Expected counts in sample } i} = \Big(\underbrace{\rho_{i\cdot}}_{\text{True relative abundance profile of sample } i} \circ \underbrace{\exp(\boldsymbol{\beta})}_{\text{Taxon-specific “efficiencies”}}\Big) \cdot \underbrace{k_i}_{\text{Proportionality term}}}$$


Simulated Sequencing Data on a Simple Community

Consider a specimen

- containing taxa A, B, and C
  - in relative abundances 0.5, 0.25, and 0.25, respectively

We can simulate* sequencing data under

- assumption $\mathbb{E}[W_{i\cdot}] \propto \rho_{i\cdot}$
- assumption $\mathbb{E}[W_{i\cdot}] \propto \exp(\boldsymbol{\beta}) \circ \rho_{i\cdot}$
  - with $\exp(\boldsymbol{\beta}) = (2,8,1)$

* First setting (with same efficiencies across taxa): each count simulated as a negative binomial with mean $\mu_j = 500 * \rho_j$ and size parameter $s = 5$ (s.t. Var$(W_{ij}) = \mu_j + \mu_j^2/s$)

* Second setting (differing efficiencies): each count simulated as a negative binomial with $\mu_j = [500/\exp(\bar{\boldsymbol{\beta}})] * \rho_{ij}$ and size parameter $s = 5$ (s.t. Var$(W_{ij}) = \mu_j + \mu_j^2/s$)

Efficiencies: An Example

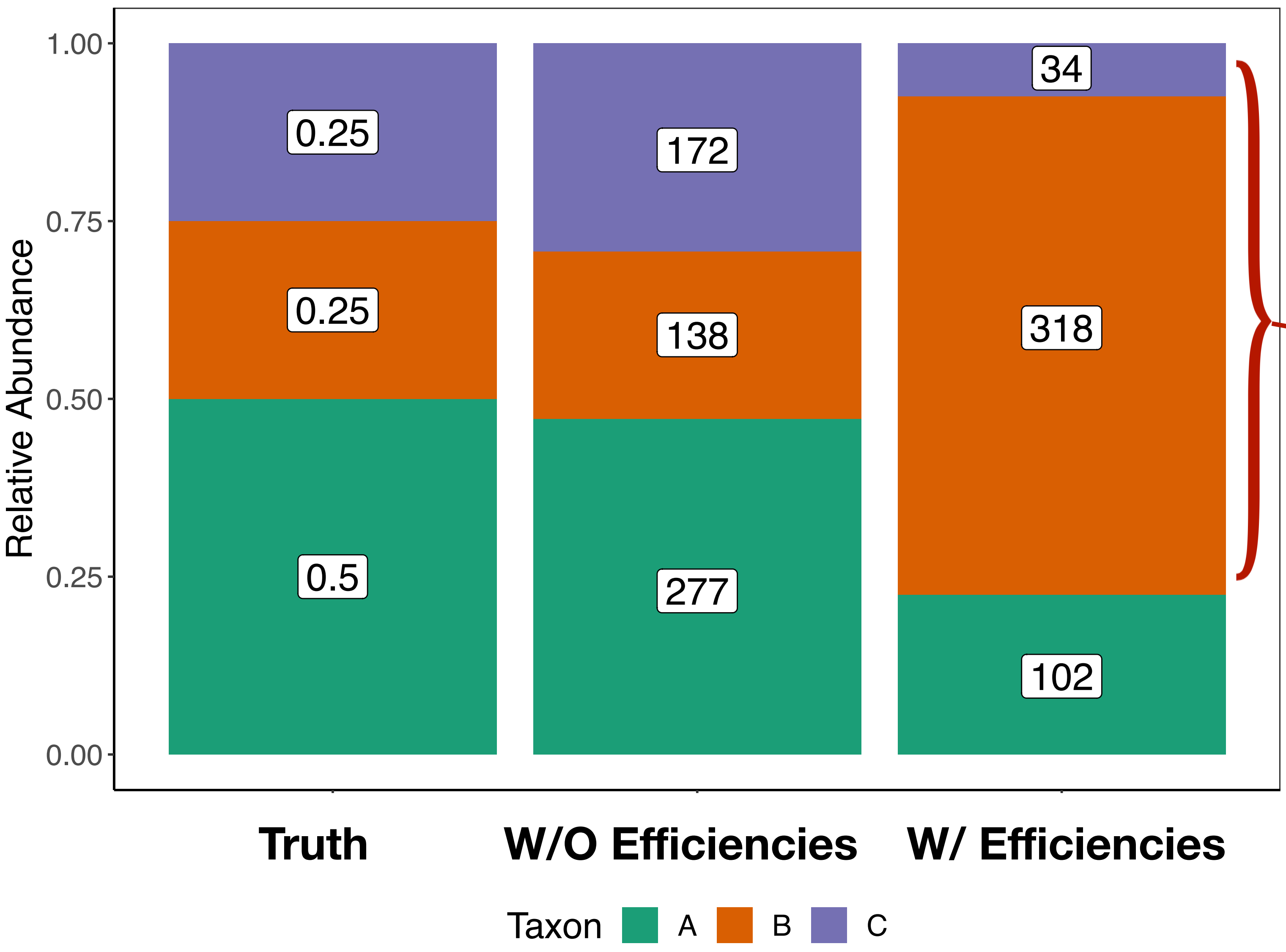
$$\underbrace{\mathbb{E}[W_{i\cdot}]}_{\text{Expected counts in sample } i} = \Big(\underbrace{\rho_{i\cdot}}_{\text{True relative abundance profile of sample } i} \circ \underbrace{\exp(\boldsymbol{\beta})}_{\text{Taxon-specific “efficiencies”}}\Big) \cdot \underbrace{k_i}_{\text{Proportionality term}}$$

Detection efficiency of taxon B relative to taxon C $\exp(\beta_2) = 8$

↓

Ratio of counts in taxon B to counts in taxon C ≈ 8 times too large

Simulated Sequencing Data on a Simple Community



Consider a specimen

- containing taxa A, B, and C
  - in relative abundances 0.5, 0.25, and 0.25, respectively

We can simulate* sequencing data under

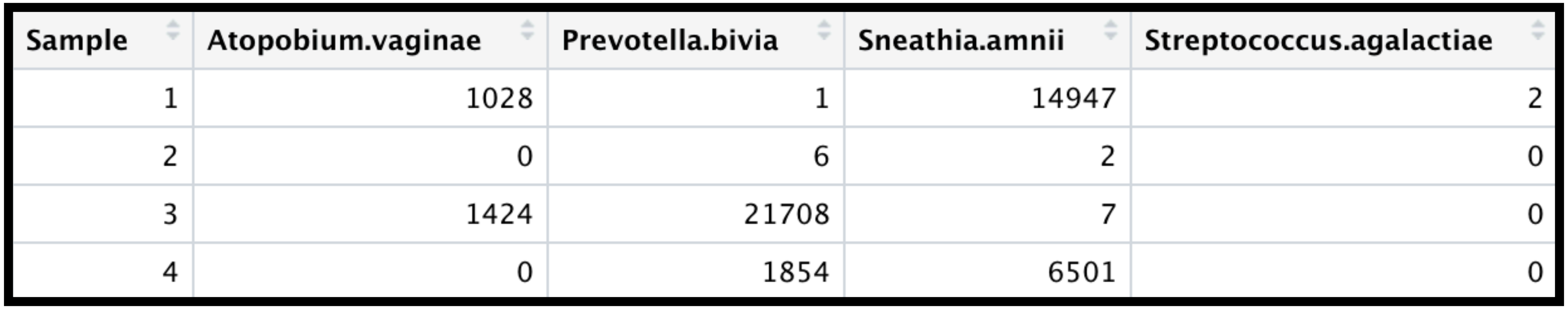
- assumption $\mathbb{E}[W_{i\cdot}] \propto \rho_{i\cdot}$
- assumption $\mathbb{E}[W_{i\cdot}] \propto \exp(\boldsymbol{\beta}) \circ \rho_{i\cdot}$
  - with $\exp(\boldsymbol{\beta}) = (2{,}8{,}1)$

* First setting (with same efficiencies across taxa): each count simulated as a negative binomial with mean $\mu_j = 500 * \rho_j$ and size parameter $s = 5$ (s.t. $\text{Var}(W_{ij}) = \mu_j + \mu_j^2/s$)

* Second setting (differing efficiencies): each count simulated as a negative binomial with $\mu_j = [500/\exp(\bar{\boldsymbol{\beta}})] * \rho_{ij}$ and size parameter $s = 5$ (s.t. $\text{Var}(W_{ij}) = \mu_j + \mu_j^2/s$)

Estimating Relative Abundance in Presence of Efficiencies

- McLaren et al. (2019)
  - Method for estimating p_{ij} and e_j via a centered log-ratio transformation of counts W_{ij}
    - Need to know presence/absence in advance
    - Zero counts, spurious counts an issue

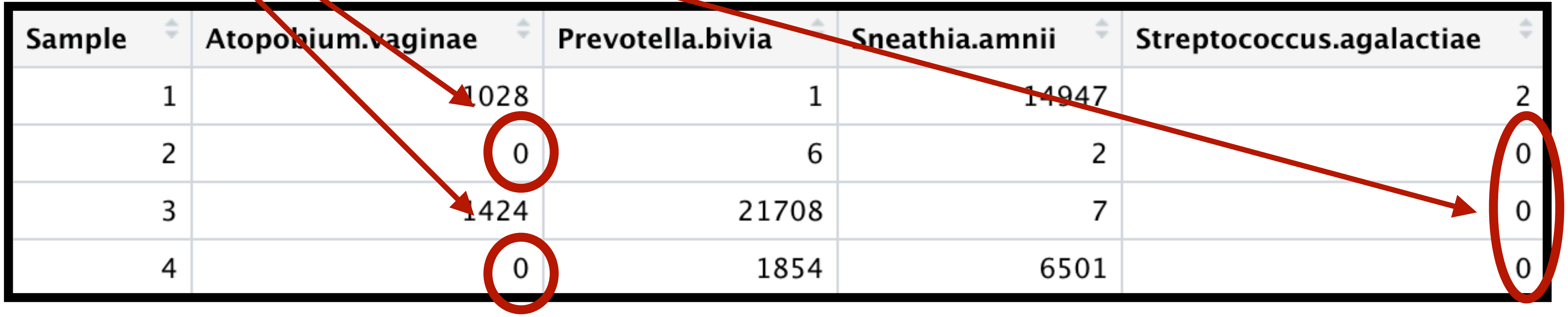


Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

Estimating Relative Abundance in Presence of Efficiencies

- McLaren et al. (2019)
  - Method for estimating p_{ij} and e_j via a centered log-ratio transformation of counts W_{ij}
    - Need to know presence/absence in advance
    - Zero counts, spurious counts an issue

Log(0) undefined



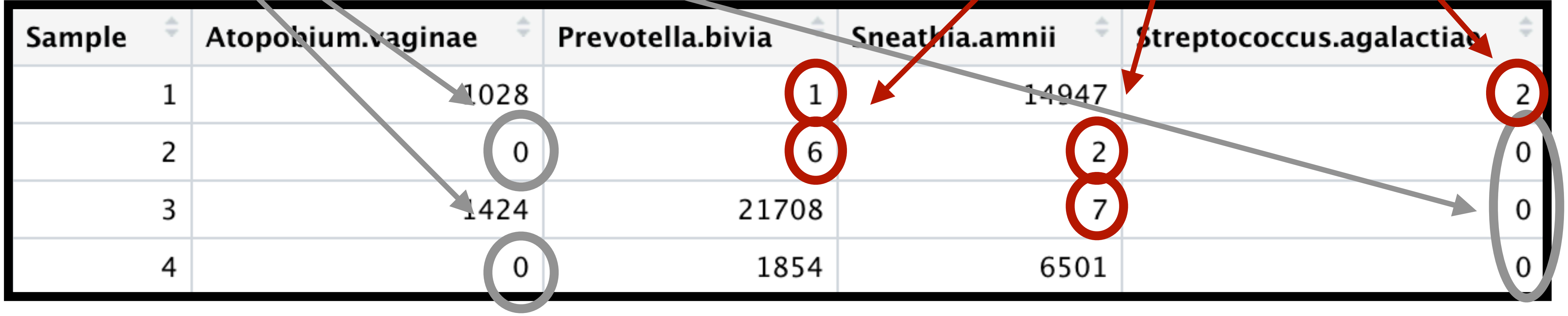
Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

Estimating Relative Abundance in Presence of Efficiencies

- McLaren et al. (2019)
  - Method for estimating p_{ij} and e_j via a centered log-ratio transformation of counts W_{ij}
  - Need to know presence/absence in advance
  - Zero counts, spurious counts an issue

Log(0) undefined

$W_{ij} > 0$ **when** $p_{ij} = 0$



Sample	Atopobium.vaginae	Prevotella.bivia	Sneathia.amnii	Streptococcus.agalactiae
1	1028	1	14947	2
2	0	6	2	0
3	1424	21708	7	0
4	0	1854	6501	0

Generalizing McLaren et al.

- Clausen-Willis approach: model counts W directly
  - Attempt to model **spurious reads** (due to, e.g., contamination) in addition to detection efficiencies
  - Mean model for a count W_{ij}

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Mean Model Details

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Mean Model Details

True relative abundance of taxon j in sample I

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Mean Model Details

True relative abundance of taxon j in sample I

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Proportionality constant

Mean Model Details

True relative abundance of taxon j in sample I

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Proportionality constant

Log efficiency e_j

Mean Model Details

True relative abundance of taxon j in sample I

$$\mu_{ij} := \mathbb{E}[W_{ij}|\boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

(Unknown) relative abundance profile of spurious read source

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Proportionality constant

Log efficiency e_j

Mean Model Details

True relative abundance of taxon j in sample I

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

(Unknown) relative abundance profile of spurious read source

$$= \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} + \underbrace{\tilde{p}_{ij}\,\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

Proportionality constant

Log efficiency e_j

Intensity of spurious reads

Mean Model Details: A Bit More Generality

$$\mu_{ij} := \ \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \quad \underbrace{p_{ij} exp(\gamma_i + \beta_j)}_{\text{contribution of sample}} \quad + \quad \underbrace{\tilde{p}_{ij}\text{exp}(\tilde{\gamma})}_{\text{" " spurious read sources}}$$

$$= \quad \underbrace{Z_i p^j exp(\gamma_i + X_i \beta^j)}_{\text{contribution of sample}} \quad + \quad \underbrace{\tilde{Z}_i(\tilde{p}^j \circ \text{exp}(\tilde{\gamma}))}_{\text{" " spurious read sources}}$$

$$= \quad \underbrace{\Big[(\mathbf{Zp} \circ \text{exp}(\boldsymbol{\gamma}\mathbf{1}_J^T + \mathbf{X}\boldsymbol{\beta})}_{\text{contribution of samples}} \quad + \quad \underbrace{\tilde{\mathbf{Z}}[\tilde{\mathbf{p}} \circ \text{exp}(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^T + \tilde{\mathbf{X}}\boldsymbol{\beta})]\Big]_{ij}}_{\text{contribution of spurious read sources}}$$

Mean Model Details: A Bit More Generality

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

$$= \Big[\underbrace{(\mathbf{Z}\mathbf{p} \circ \text{exp}(\boldsymbol{\gamma}\mathbf{1}_J^T + \mathbf{X}\boldsymbol{\beta})}_{\text{contribution of samples}} + \underbrace{\tilde{\mathbf{Z}}[\tilde{\mathbf{p}} \circ \text{exp}(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^T + \tilde{\mathbf{X}}\boldsymbol{\beta})] \Big]_{ij}}_{\text{contribution of spurious read sources}}$$

More general form allows us to

- Easily incorporate technical replicates
- Model differing efficiencies across samples
  - E.g., due to different protocols in different batches
- Model multiple sources of spurious reads
- And more

More details in supplemental slides if you're interested

Defining an Estimator

We estimate unknown parameters in mean model

$$\mu_{ij} := \mathbb{E}[W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}}]$$

by modeling

$$W_{ij} | \boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\gamma}, \tilde{\mathbf{p}}, \tilde{\boldsymbol{\gamma}} \sim \text{Poisson}(\mu_{ij})$$

and estimate parameters via maximum likelihood

An Applied Example

- Data from Brooks et al. (2015)
  - 40 whole-cell samples of known composition prepped and sequenced (via 16S) together
    - All specimens composed of some combination of 7 common bacterial species in the vaginal microbiome
  - We observe some spurious reads (nonzero number of reads in taxa known to be absent in a particular sample)
  - Probably reasonable to model a single detection efficiency for each taxon

An Applied Example (cont.)

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n\times J}\,|\,\mathbf{p},\boldsymbol{\gamma},\boldsymbol{\beta},\tilde{\mathbf{p}},\tilde{\boldsymbol{\gamma}}] =$$

$$(\mathbf{p} \circ \text{exp}(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \text{exp}(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

- **n = 40 samples sequenced**
- **K = 40 unique specimens**
- **J = 7 taxa considered**
- **p = 1 ($1 \times J$) efficiency effect**
- **$\tilde{n}$ = 1 spurious read source**

An Applied Example (cont.)

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n\times J} \,|\, \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] =$$

$$(\mathbf{p} \circ \text{exp}(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \text{exp}(\tilde{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

$n \times J = 40 \times 7$ **matrix of true relative abundances** p_{ij} **for** $i = 1,\ldots,40$ **and** $j = 1,\ldots,7$

- n **= 40 samples sequenced**
- K **= 40 unique specimens**
- J **= 7 taxa considered**
- p **= 1 (**$1 \times J$**) efficiency effect**
- $\tilde{n}$ **= 1 spurious read source**

An Applied Example (cont.)

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n\times J} \,|\, \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] =$$

$$(\mathbf{p} \circ \exp(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \exp(\tilde{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

$n \times J = 40 \times 7$ **matrix of true relative abundances** p_{ij} **for** $i = 1,\ldots,40$ **and** $j = 1,\ldots,7$

$n \times 1 = 40 \times 1$ **matrix of read depth values for samples** $i = 1,\ldots,40$

- n **= 40 samples sequenced**
- K **= 40 unique specimens**
- J **= 7 taxa considered**
- p **= 1 (**$1 \times J$**) efficiency effect**
- $\tilde{n}$ **= 1 spurious read source**

An Applied Example (cont.)

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_7]^T$, constrain $\beta_7 = 0$; β_j has interp. log relative eff. of taxon j rel. to taxon 7 (for $j = 1,\ldots,6$)

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n\times J} \,|\, \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] =$$

$$(\mathbf{p} \circ \exp(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \exp(\tilde{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

$n \times J = 40 \times 7$ matrix of true relative abundances p_{ij} for $i = 1,\ldots,40$ and $j = 1,\ldots,7$

$n \times 1 = 40 \times 1$ matrix of read depth values for samples $i = 1,\ldots,40$

- n = 40 samples sequenced
- K = 40 unique specimens
- J = 7 taxa considered
- p = 1 ($1 \times J$) efficiency effect
- $\tilde{n}$ = 1 spurious read source

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_7]^T$, constrain $\beta_7 = 0$; β_j has interp. log relative eff. of taxon j rel. to taxon 7 (for $j = 1,\ldots,6$)

An Applied Example (cont.)

Model all samples as having on average same abundance of spurious reads

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n\times J}\,|\,\mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] =$$

$$(\mathbf{p} \circ \exp(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \exp(\tilde{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

$n \times J = 40 \times 7$ matrix of true relative abundances p_{ij} for $i = 1,\ldots,40$ and $j = 1,\ldots,7$

$n \times 1 = 40 \times 1$ matrix of read depth values for samples $i = 1,\ldots,40$

- **n = 40 samples sequenced**
- **K = 40 unique specimens**
- **J = 7 taxa considered**
- **p = 1 ($1 \times J$) efficiency effect**
- **$\tilde{n}$ = 1 spurious read source**

Mean Model Example (cont.)

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_7]^T$, constrain $\beta_7 = 0$; β_j has interp. log relative eff. of taxon j rel. to taxon 7 (for $j = 1,\ldots,6$)

Model all samples as having on average same abundance of spurious reads

Single source of spurious reads $[\tilde{p}_{11}, \ldots, \tilde{p}_{1J}]$

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n \times J} \,|\, \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] =$$

$$(\mathbf{p} \circ \exp(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \exp(\tilde{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

$n \times J = 40 \times 7$ matrix of true relative abundances p_{ij} for $i = 1,\ldots,40$ and $j = 1,\ldots,7$

$n \times 1 = 40 \times 1$ matrix of read depth values for samples $i = 1,\ldots,40$

- n = 40 samples sequenced
- K = 40 unique specimens
- J = 7 taxa considered
- p = 1 ($1 \times J$) efficiency effect
- $\tilde{n}$ = 1 spurious read source

Mean Model Example (cont.)

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_7]^T$, constrain $\beta_7 = 0$; β_j has interp. log relative eff. of taxon j rel. to taxon 7 (for $j = 1,\ldots,6$)

Model all samples as having on average same abundance of spurious reads

Single source of spurious reads $[\tilde{p}_{11}, \ldots, \tilde{p}_{1J}]$

Single intensity of spurious reads $\tilde{\gamma}$

Proposed mean model:

$$\mathbf{E}[\mathbf{W}_{n \times J} \,|\, \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] =$$

$$(\mathbf{p} \circ \exp(\boldsymbol{\gamma} \mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n [\tilde{\mathbf{p}} \circ \exp(\tilde{\gamma} \mathbf{1}_J^T + \boldsymbol{\beta})]$$

$n \times J = 40 \times 7$ matrix of true relative abundances p_{ij} for $i = 1,\ldots,40$ and $j = 1,\ldots,7$

$n \times 1 = 40 \times 1$ matrix of read depth values for samples $i = 1,\ldots,40$

- n = 40 samples sequenced
- K = 40 unique specimens
- J = 7 taxa considered
- p = 1 ($1 \times J$) efficiency effect
- $\tilde{n}$ = 1 spurious read source

Performance on Brooks (2015) Data

Fit mean model

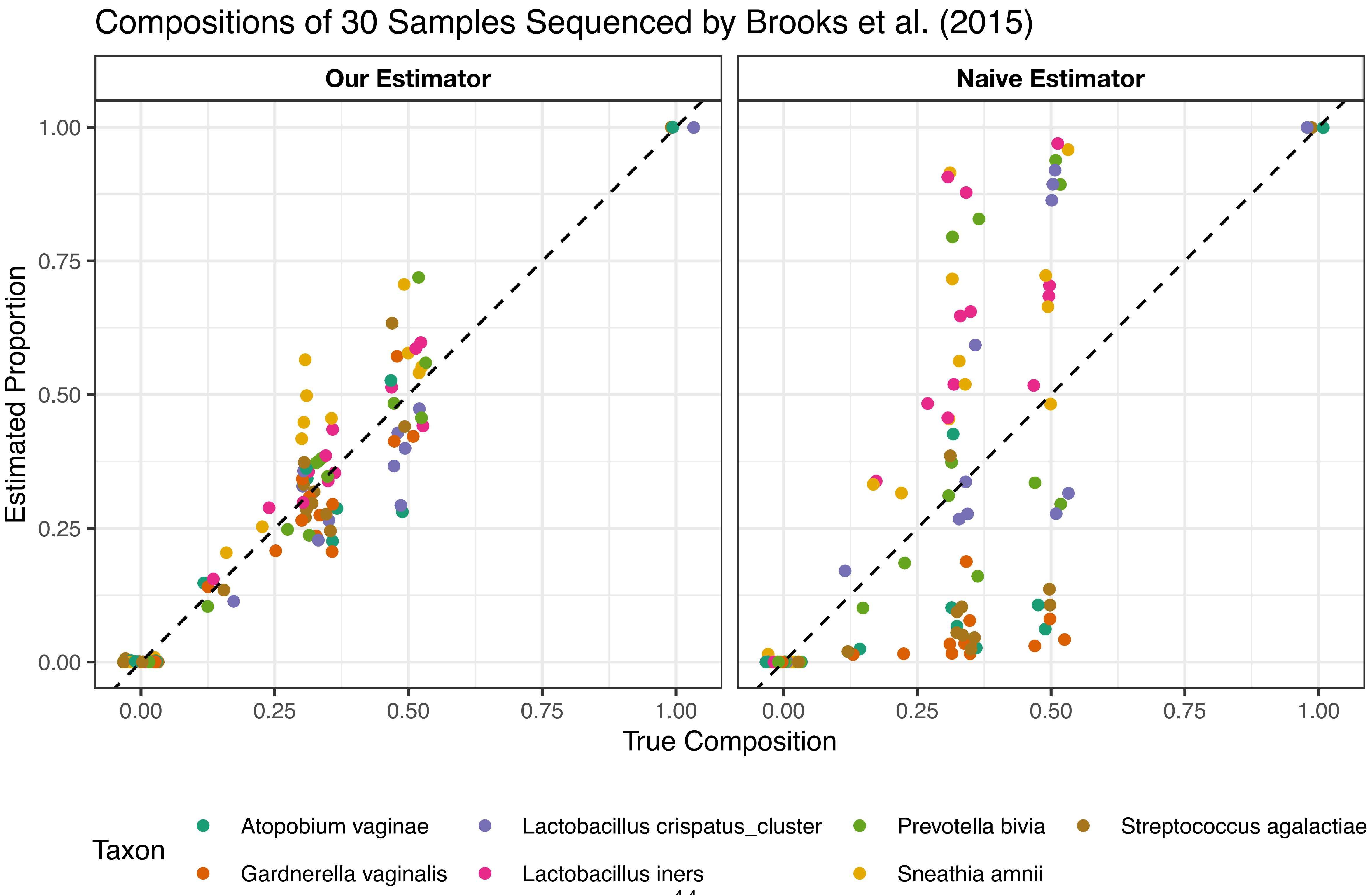
$$\mathbf{E}[\mathbf{W}_{n\times J} \,|\, \mathbf{p}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \tilde{\mathbf{p}}, \tilde{\gamma}] = (\mathbf{p} \circ \text{exp}(\boldsymbol{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta}) + \mathbf{1}_n[\tilde{\mathbf{p}} \circ \text{exp}(\tilde{\gamma}\mathbf{1}_J^T + \boldsymbol{\beta})]$$

- One sample per unique specimen
- One set of log efficiencies $\boldsymbol{\beta}$
- One source of spurious reads $\tilde{\mathbf{p}}$ modeled as having same read abundance across samples

To data from 40 samples sequenced together

- Use known true compositions of first 10 samples
- All other compositions estimated from data

Performance on Brooks (2015) Data



Performance on Brooks (2015) Data

- Additionally, this model fit estimated 6 relative abundances with spurious counts to be zero
  - $\approx 10\,\%$ of taxon-sample pairs with spurious reads (with true relative abundance zero)
  - Better choice of $\tilde{Z}$ might perform better

Future Work

- Inference via a modified bootstrap
- Predicting efficiencies in taxa not present in specimens of known composition
- Investigating use of covariates for spurious reads (e.g., DNA concentration)

Acknowledgments

This work is joint with my fabulous advisor, Dr. Amy D. Willis, Assistant Professor, UW Department of Biostatistics, and it is the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM133420

statistical
diversity
lab

I am grateful as well for the invaluable feedback my colleagues in the Statistical Diversity Lab have provided in the development of this model and presentation

Bryan Martin
(UW Statistics)

Cecilia Shi
(UW Statistics)

Pauline Trinh
(UW DEOHS)

Maria Valdez Cabrera
(UW Biostatistics)

Sarah Teichman
(UW Statistics)

Thank You!

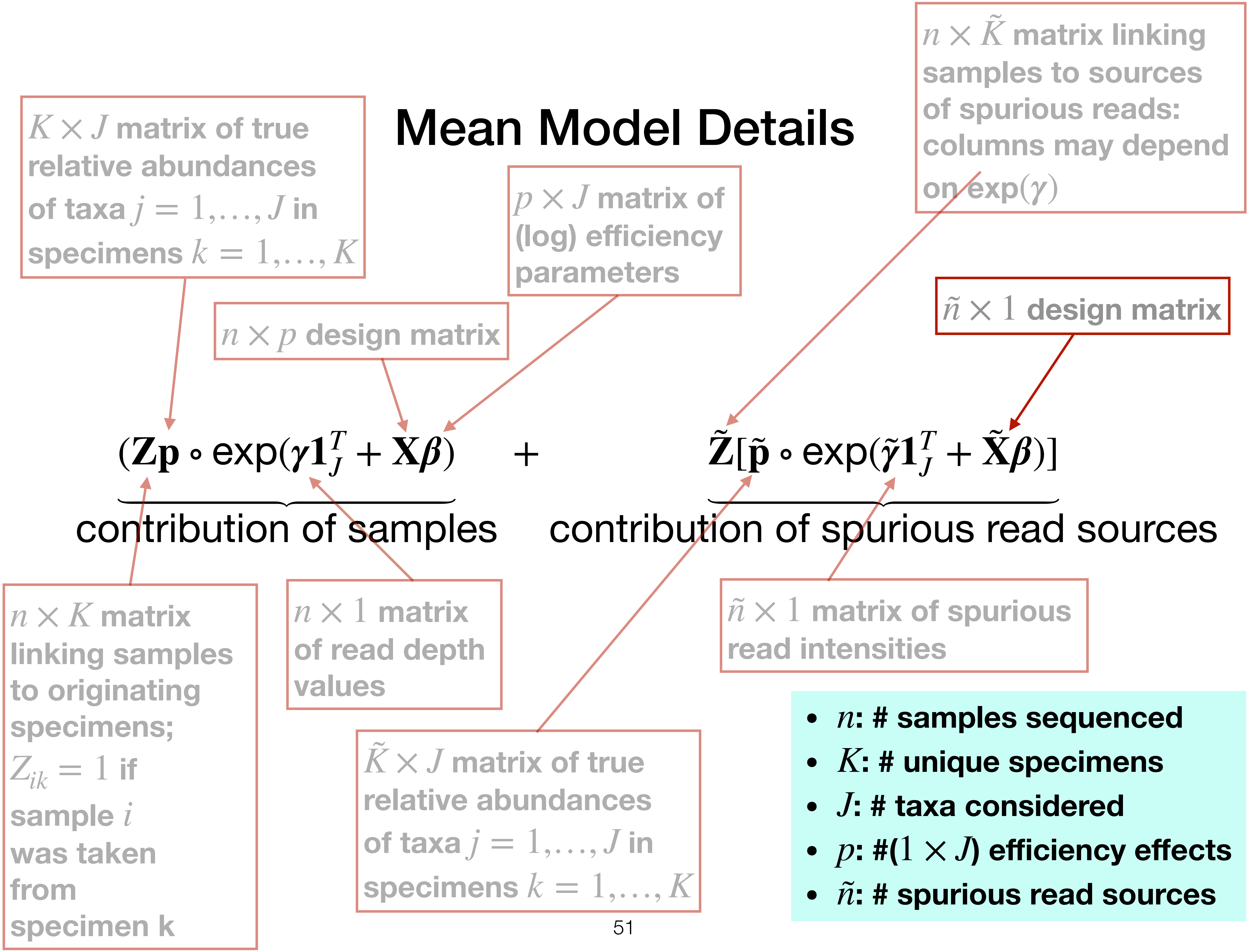
References

Brooks, J. Paul, David J. Edwards, Michael D. Harwich, Maria C. Rivera, Jennifer M. Fettweis, Myrna G. Serrano, Robert A. Reris et al. "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies." *BMC microbiology* 15, no. 1 (2015): 1-14.

McLaren, Michael R., Amy D. Willis, and Benjamin J. Callahan. "Consistent and correctable bias in metagenomic sequencing experiments." *elife* 8 (2019): e46923.

Supplemental slides

Mean Model Details



$K \times J$ matrix of true relative abundances of taxa $j = 1, \ldots, J$ in specimens $k = 1, \ldots, K$

$n \times p$ design matrix

$p \times J$ matrix of (log) efficiency parameters

$n \times \tilde{K}$ matrix linking samples to sources of spurious reads: columns may depend on $\exp(\gamma)$

$\tilde{n} \times 1$ design matrix

$$\underbrace{(\mathbf{Zp} \circ \exp(\boldsymbol{\gamma}\mathbf{1}_J^T + \mathbf{X}\boldsymbol{\beta})}_{\text{contribution of samples}} + \underbrace{\tilde{\mathbf{Z}}[\tilde{\mathbf{p}} \circ \exp(\tilde{\boldsymbol{\gamma}}\mathbf{1}_J^T + \tilde{\mathbf{X}}\boldsymbol{\beta})]}_{\text{contribution of spurious read sources}}$$

$n \times K$ matrix linking samples to originating specimens; $Z_{ik} = 1$ if sample i was taken from specimen k

$n \times 1$ matrix of read depth values

$\tilde{K} \times J$ matrix of true relative abundances of taxa $j = 1, \ldots, J$ in specimens $k = 1, \ldots, K$

$\tilde{n} \times 1$ matrix of spurious read intensities

- n: # samples sequenced
- K: # unique specimens
- J: # taxa considered
- p: #($1 \times J$) efficiency effects
- $\tilde{n}$: # spurious read sources