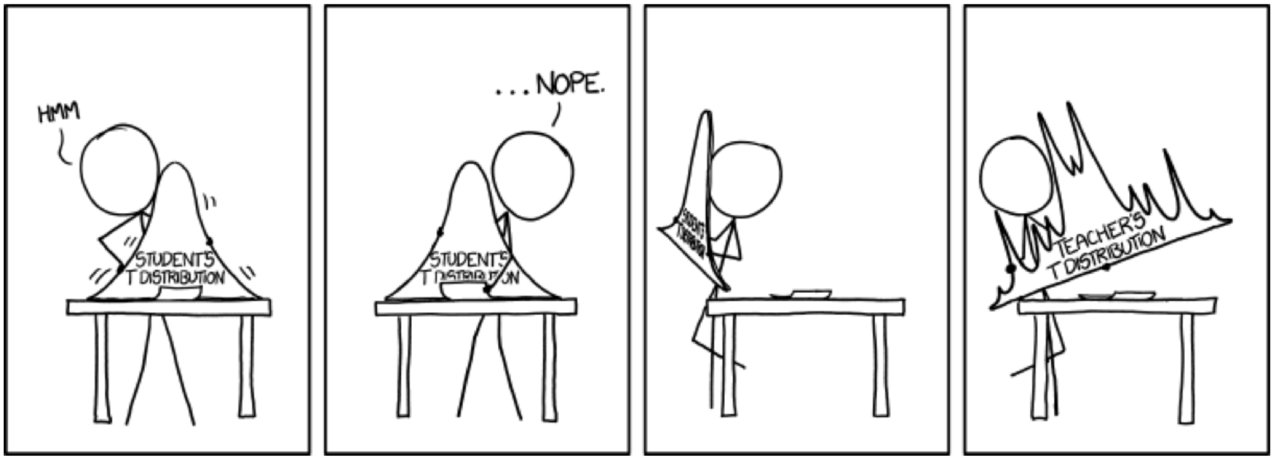
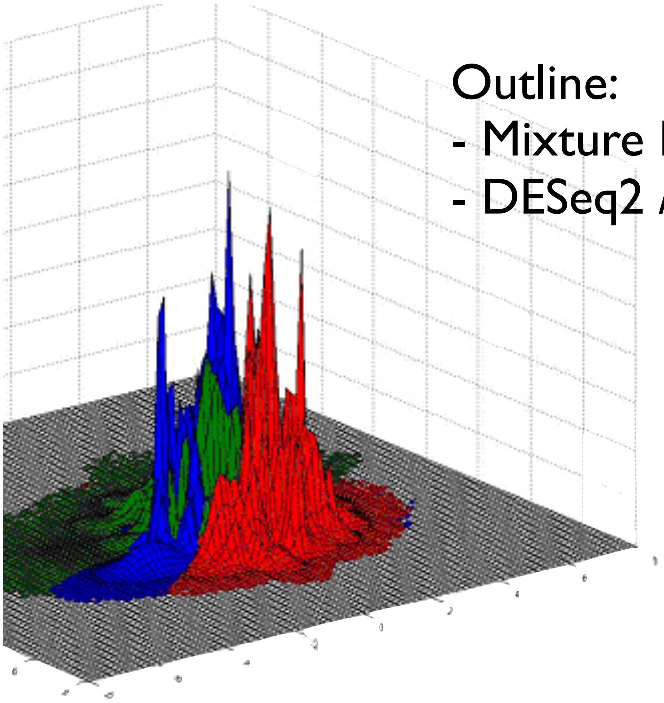
Lecture 3:
Mixture Models for Microbiome data



Lecture 3:
Mixture Models for Microbiome data



Outline:
- Mixture Models (Negative Binomial)
- DESeq2 / Don't Rarefy. Ever.

Hypothesis Tests - reminder

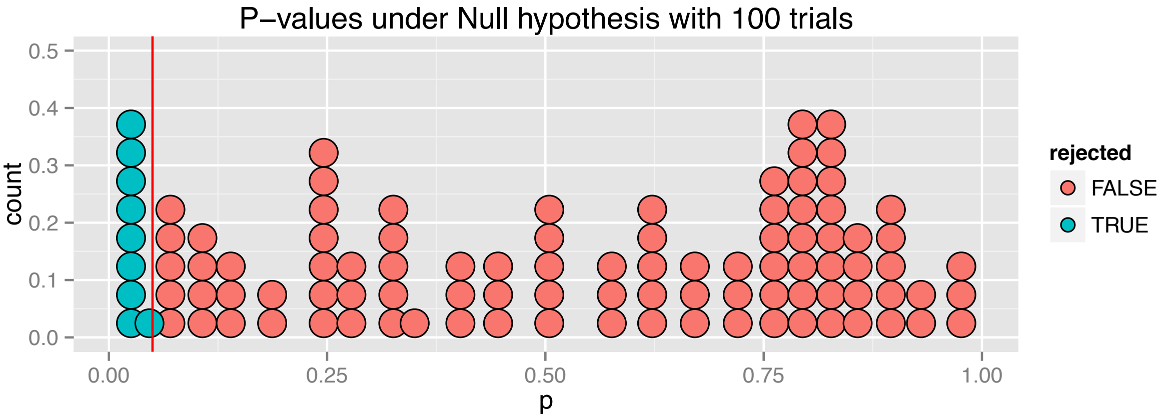
- A hypothesis is a precise disprovable statement.
- “Null hypothesis” - the default position. “Nothing special”
- Alternative/Rejection: Evidence disagrees with the Null
- Null hypothesis cannot be *confirmed* by the data.

Hypothesis Tests - some examples

test	**R function**
t-test	`t.test`
Mann-Whitney U-test	`wilcox.test`
correlation test	`cor.test`
Chi-Square test	`chisq.test`
Neg-Binom Wald test	`DESeq2::nbinomWaldTest`

Multiple Testing

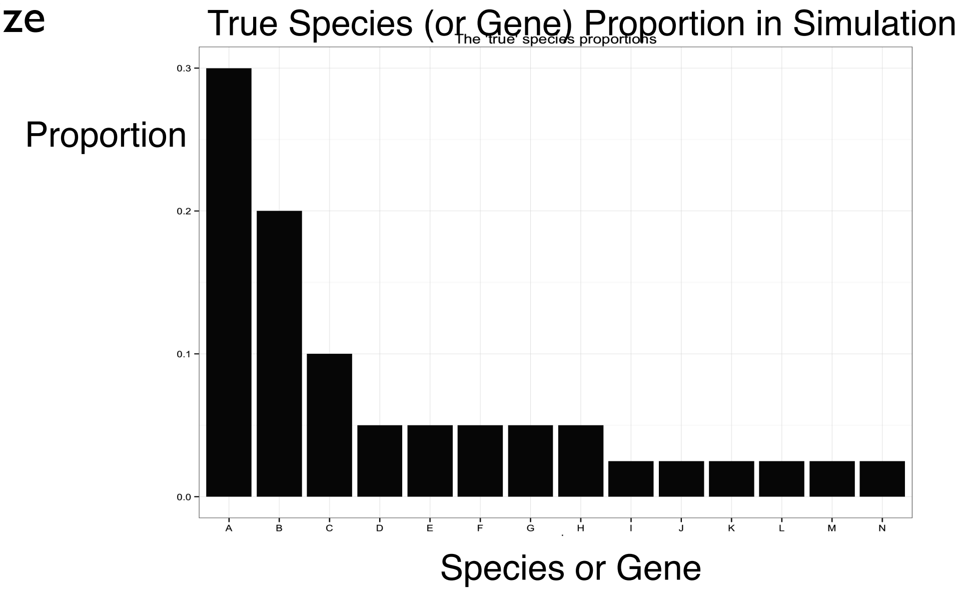
- In "Big Data", we often want to test many hypotheses in one batch.
- p-values are distributed uniformly when null hypothesis is true
- The expected number of rejections **by chance** is $m*\alpha$



Model Uncertainty in NGS Count Data

- Uncertainty Depends on Library Size

Poisson-only Count Simulation



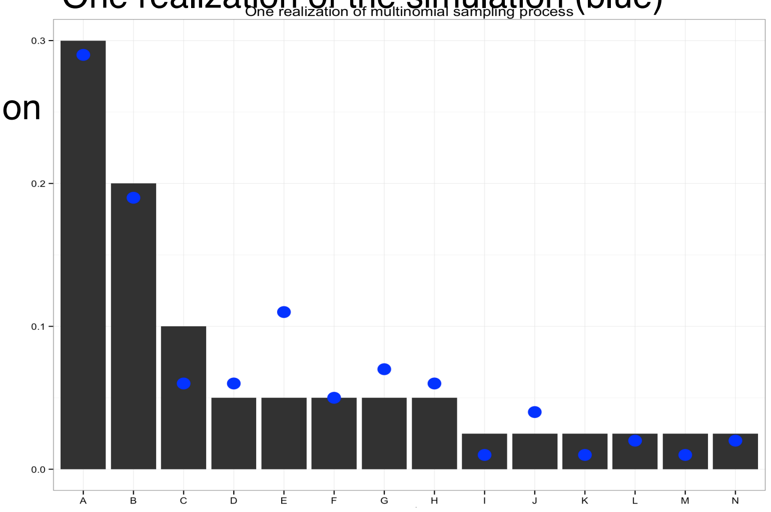
Model Uncertainty in NGS Count Data

- Uncertainty Depends on Library Size

Poisson-only Count Simulation

One realization of the simulation (blue)

Proportion

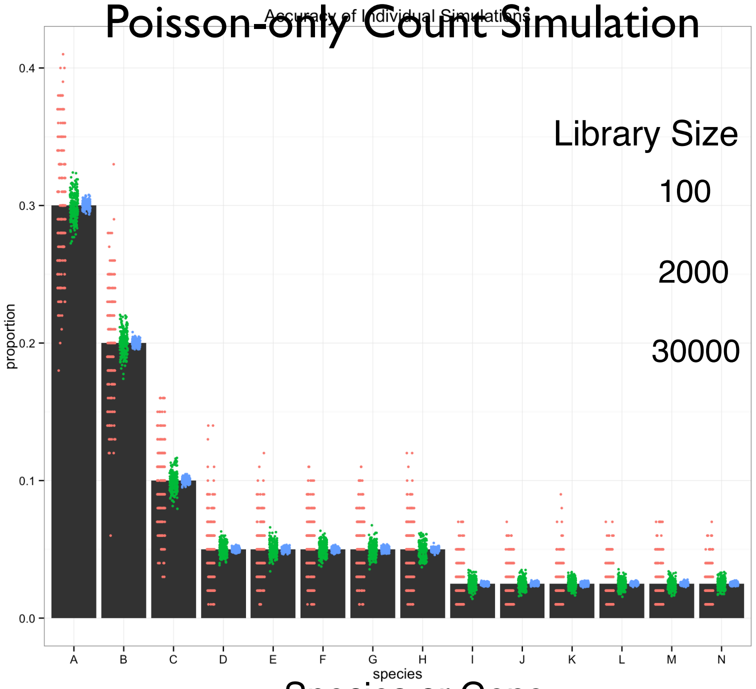


Species or Gene

Model Uncertainty in NGS Count Data

- Uncertainty Depends on Library Size
- Repeat simulation (resampling) many times and different library sizes

Poisson-only Count Simulation



Library Size

100

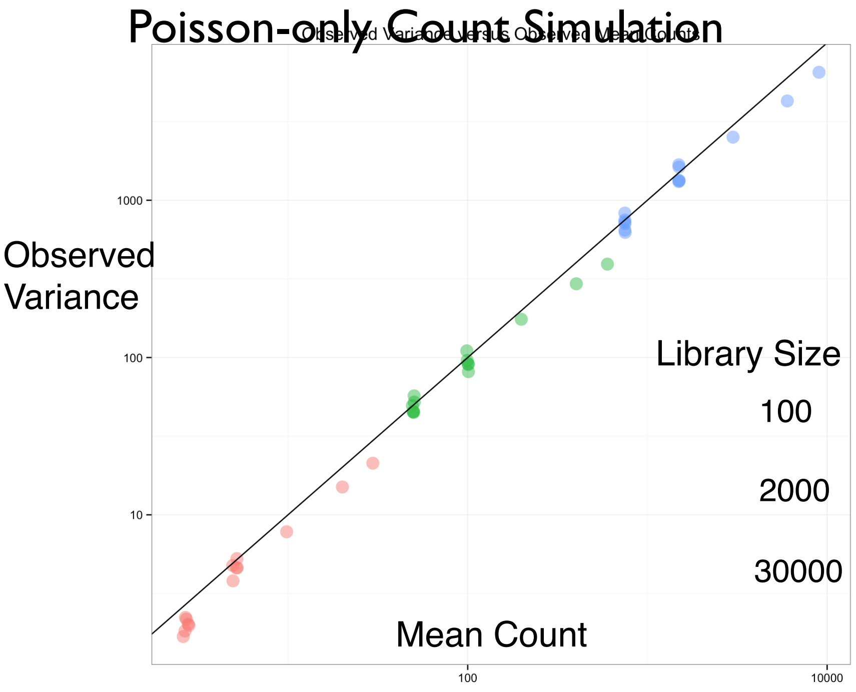
2000

30000

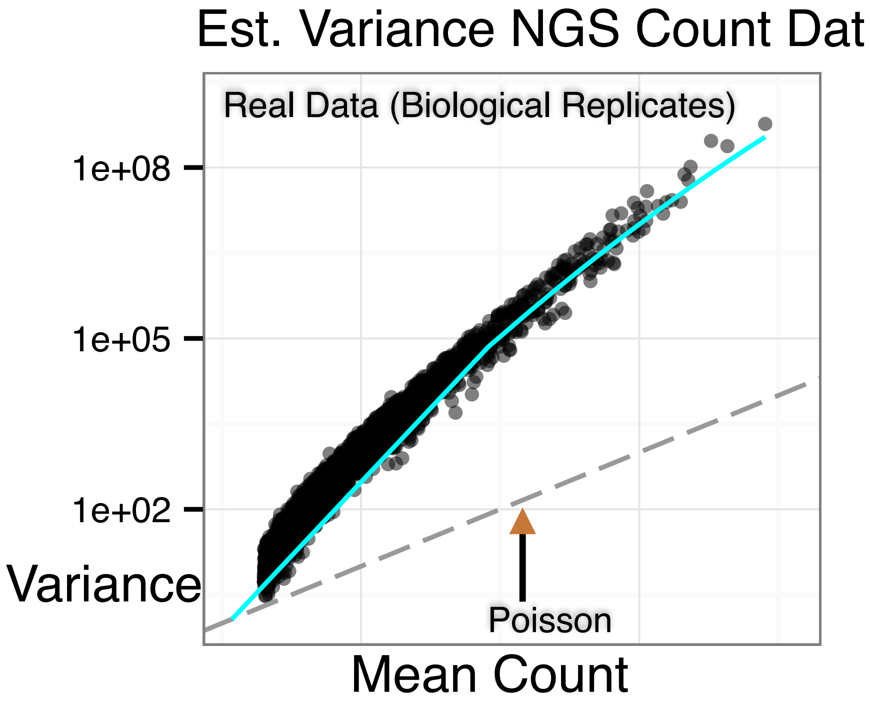
Species or Gene

Model Uncertainty in NGS Count Data

- Uncertainty Depends on Library Size
- Repeat simulation many times and different library sizes
- This turns out to describe technical sequencing replicates



Model Uncertainty in NGS Count Data

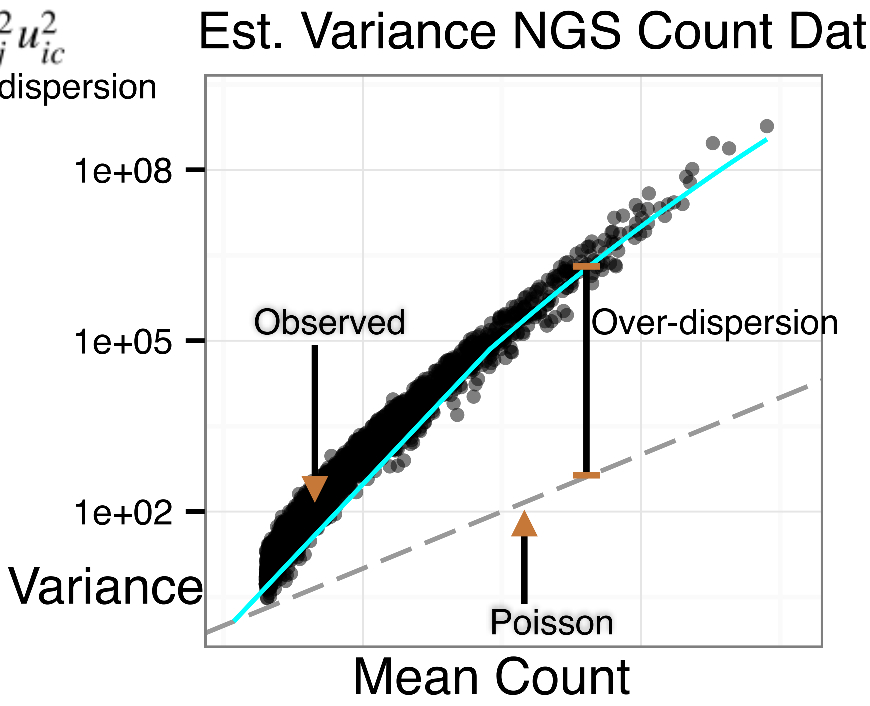


Model Uncertainty in NGS Count Data

Negative Binomial

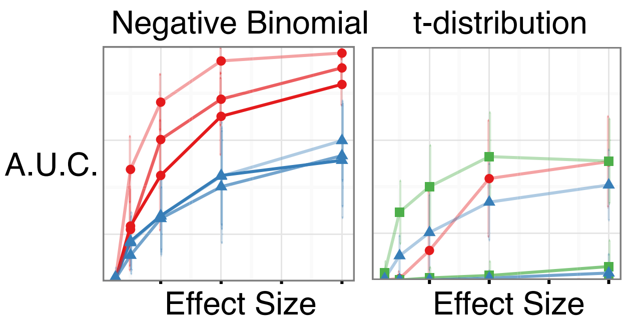
Variance = $\underset{\text{Poisson}}{u_{ic}s_j} + \underset{\text{Overdispersion}}{\phi_{ic}s_j^2u_{ic}^2}$

- Over-dispersion
- Strong Function of Mean
- Share Information Across Genes to Improve Fit (Performance)



Model Uncertainty in NGS Count Data

- Negative Binomial is an infinite mixture of Poisson R.V.
- Intuition: relevant when we have (almost) as many different distributions (poisson means) as observations
- Borrow from RNA-Seq analysis implementations? (Yes)



McMurdie & Holmes (2014). PLoS Computational Biology

- Robinson, Oshlack (2010). A scaling normalization… RNA-Seq data. *Genome Biology*
- *Anders, & Huber (2010).* Differential expression … sequence count data*. Genome Biology*

Transition: Mixture Models

Technical details in:
mixture-model-Holmes-mathy-details.pdf

Finite Mixture Model

Example: Finite mixture of two normals

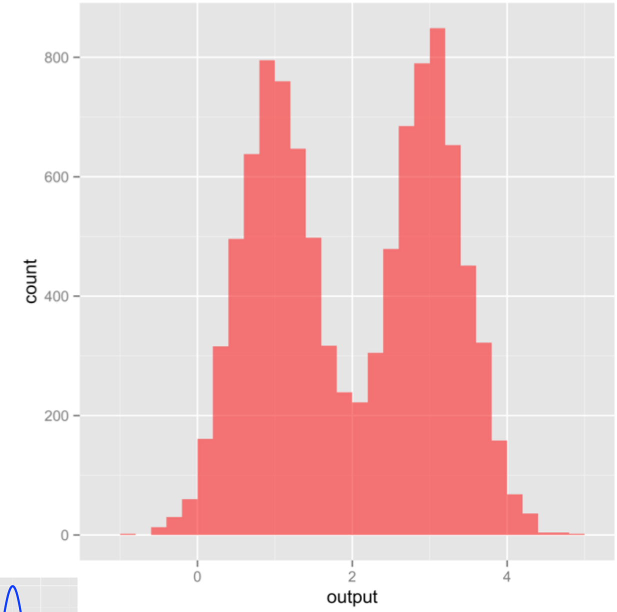
Flip a fair coin.

If it comes up heads

Generate a random number from a Normal with mean 1 and variance 0.25. R: `rnorm` function.

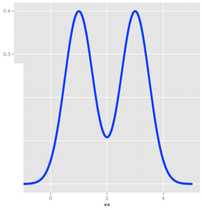
If it comes up tails

Generate a random number from a Normal with mean 2 and variance 0.25.



This is what the resulting histogram would look like if we did this 10,000 times.

$$f(x) = \frac{1}{2}\,\phi_1(x) + \frac{1}{2}\,\phi_2(x)$$



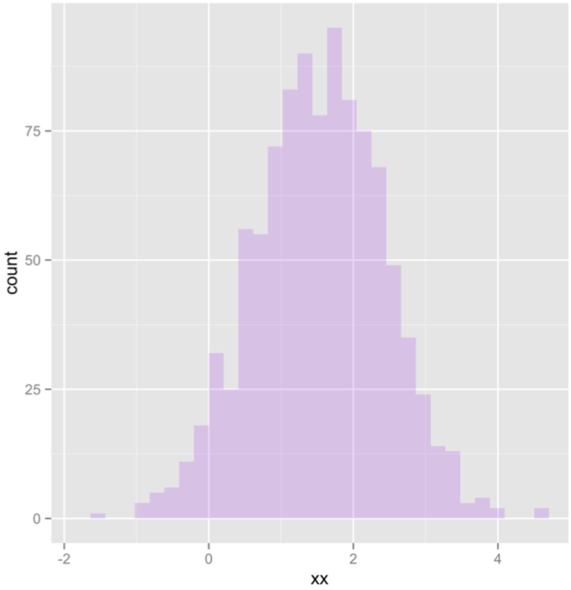
Finite Mixture Model

Example: Finite mixture of two normals

However in many cases the separation is not so clear.

Challenge: Here is a histogram generated by two Normals with the same variances.

Can you guess the two parameters for these two Normals?



$$f(x) = \frac{1}{2}\,\phi_1(x) + \frac{1}{2}\,\phi_2(x)$$

Finite Mixture Model

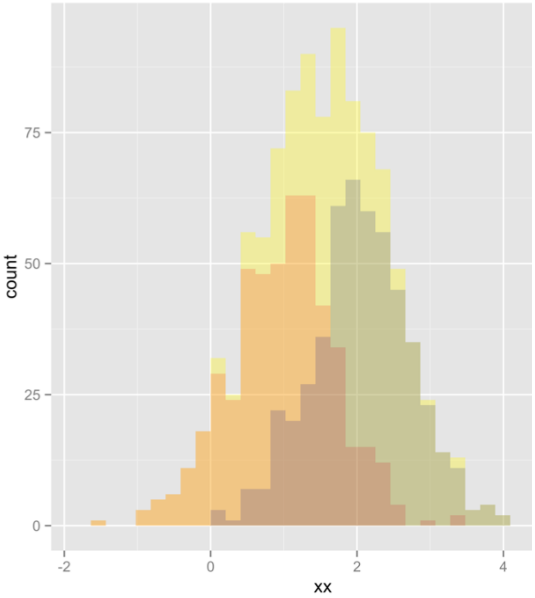
Here we knew the answer

(the *source* every data point)

In practice, this information is usually missing, and we call it a *latent* variable

Discovering the hidden class: EM

For simple parametric components, can use **EM (Expectation-Maximization)** algorithm to infer the value of the hidden variable.



$$f(x) = \frac{1}{2}\,\phi_1(x) + \frac{1}{2}\,\phi_2(x)$$

Expectation Maximization (EM)

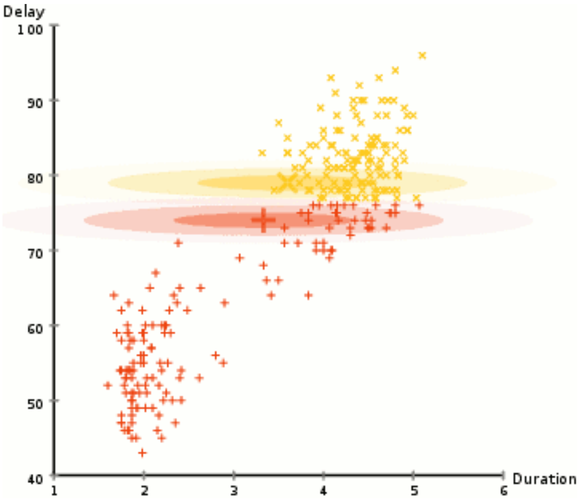
Very popular iterative procedure

Lots of implementations. E.g. FlexMix

http://cran.r-project.org/web/views/Cluster.html

http://cran.r-project.org/web/packages/flexmix/index.html

1. First, initialize θ to some random values.
2.Compute best value for U.
3. Use the just-computed values of U to compute a better estimate for θ. Parameters associated with a particular value of U only use data points whose associated latent variable has that value.
4. Iterate steps 2 and 3 until convergence



http://en.wikipedia.org/wiki/Expectation–maximization_algorithm

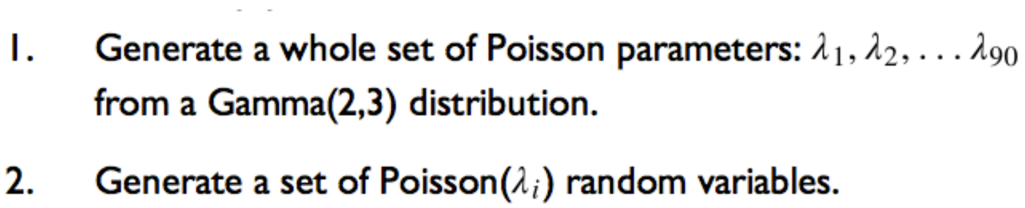
Infinite Mixture Model

Sometimes mixtures can be useful without us having to find who came from which distribution.

This is especially the case when we have (almost) as many different distributions as observations.

In some cases the total distribution can still be studied, even if we don't know the source of each component distribution.

e.g. Gamma-Poisson a.k.a. Negative Binomial

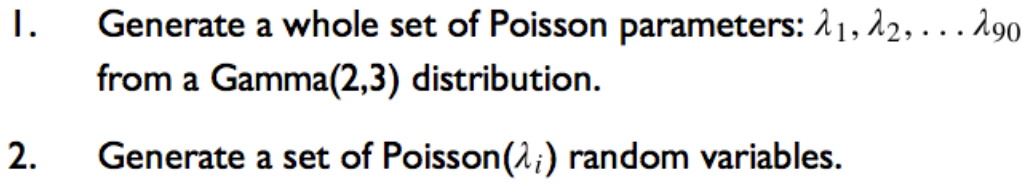


1. Generate a whole set of Poisson parameters: $\lambda_1, \lambda_2, \ldots \lambda_{90}$ from a Gamma(2,3) distribution.

2. Generate a set of Poisson(λ_i) random variables.

Infinite Mixture Model - N.B.

Generative Description:



1. **Generate a whole set of Poisson parameters:** $\lambda_1, \lambda_2, \ldots \lambda_{90}$ **from a Gamma(2,3) distribution.**

2. **Generate a set of Poisson(**λ_i**) random variables.**

Summarized Mathematically:

variance: $u_{ic}s_j + \phi_{ic}s_j^2u_{ic}^2.$

Poisson Overdispersion

Negative Binomial is useful for modeling:

- Overdispersion (in Ecology)
- Simplest Mixture Model for Counts
- Different evolutionary mutation rates
- Throughout Bioinformatics and Bayesian Statistics
- Abundance data

Summary of Mixture Models

Finite Mixture Models

Mixture of Normals with different means and variances.

Mixtures of multivariate Normals with different means and covariance matrices

Decomposing the mixtures using the EM algorithm.

Common Infinite Mixture Models

Gamma-Poisson for read counts
Dirichlet-Multinomial (Birthday problem and the Bayesian setting).

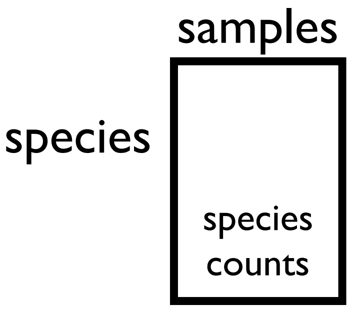
Inefficient Normalization by “rarefying”

& applicability of Negative Binomial Mixture Model

- Modern sequencing creates libraries of unequal sizes
- Early analyses focused on library-wise distances:

  paradigm: rarefy - UniFrac - PCoA - Write Paper

- This approach has “leaked” into formal settings, standard normalization method is “rarefying”



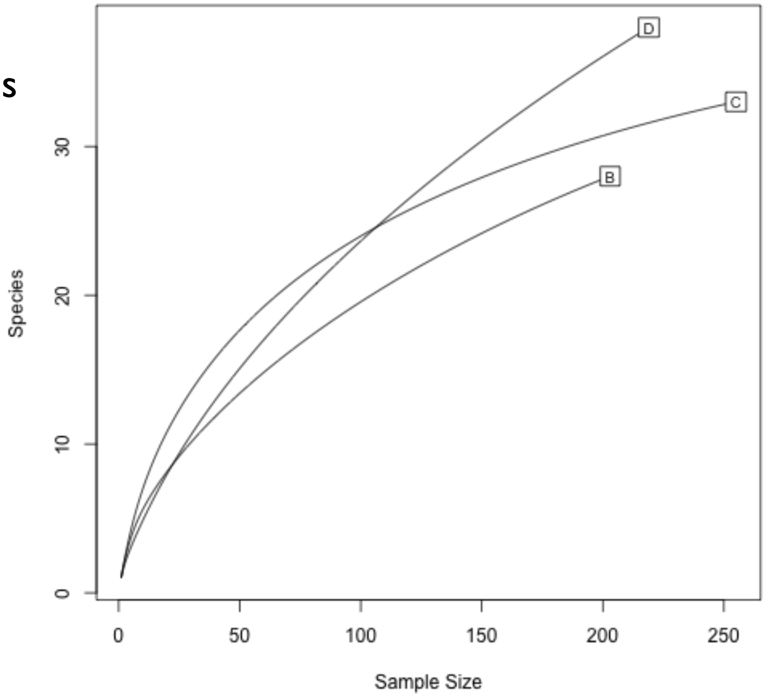
Inefficient Normalization by “rarefying”

the original idea…

- Sanders 1968
- non-parametric richness
- estimate coverage
- Normalize? - No.

Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*

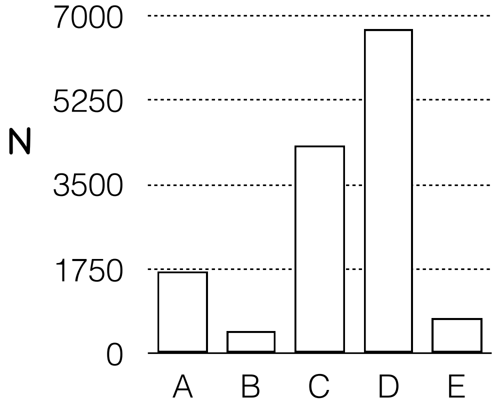
rarefaction curves



Inefficient Normalization by “rarefying”

1. Select a minimum library size $N_{L,min}$
2. Discard libraries (samples) that are smaller than $N_{L,min}$
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$

Library Sizes (column sums)



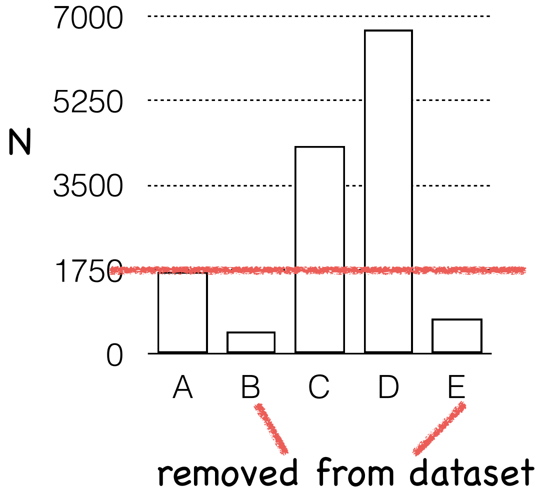
Hughes & Hellmann (2005) *Methods in Enzymology*

Gotelli, & Colwell (2001) *Ecology Letters*

Inefficient Normalization by “rarefying”

1. Select a minimum library size $N_{L,min}$
2. Discard libraries (samples) that are smaller than $N_{L,min}$
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$

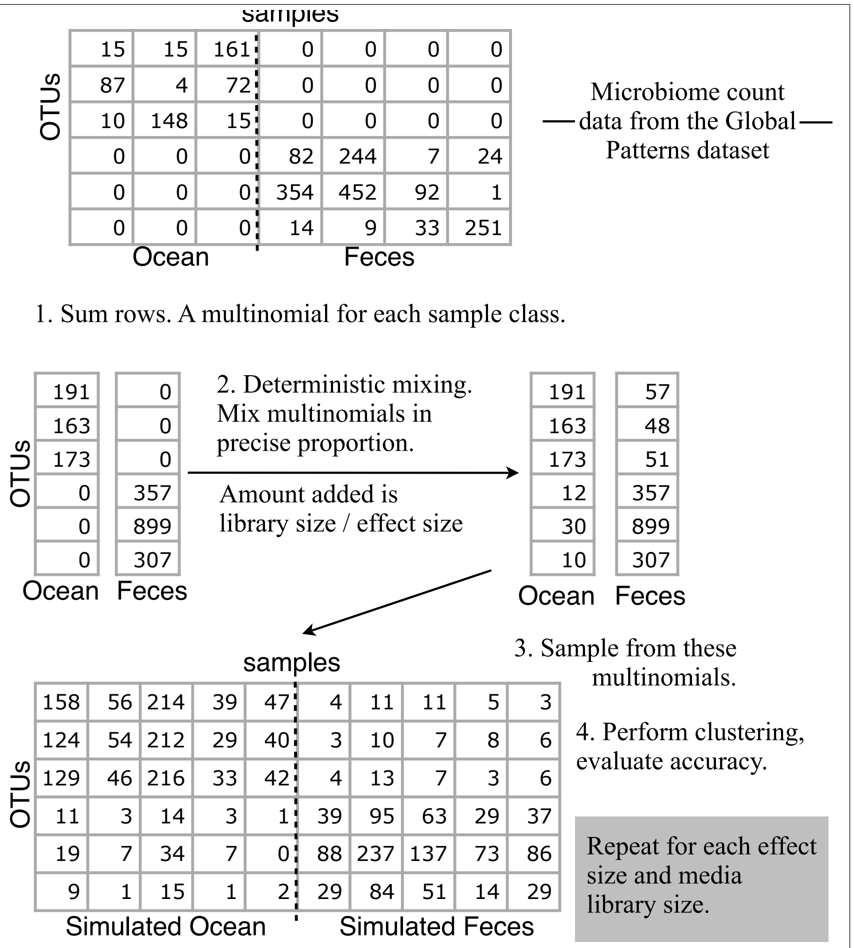
Library Sizes (column sums)



Hughes & Hellmann (2005) *Methods in Enzymology*

Gotelli, & Colwell (2001) *Ecology Letters*

Microbiome Clustering Simulation



samples

OTUs

15	15	161	0	0	0	0
87	4	72	0	0	0	0
10	148	15	0	0	0	0
0	0	0	82	244	7	24
0	0	0	354	452	92	1
0	0	0	14	9	33	251
	Ocean			Feces		

Microbiome count data from the Global Patterns dataset

1. Sum rows. A multinomial for each sample class.

OTUs

Ocean	Feces
191	0
163	0
173	0
0	357
0	899
0	307

2. Deterministic mixing. Mix multinomials in precise proportion.

Amount added is library size / effect size

Ocean	Feces
191	57
163	48
173	51
12	357
30	899
10	307

3. Sample from these multinomials.

samples

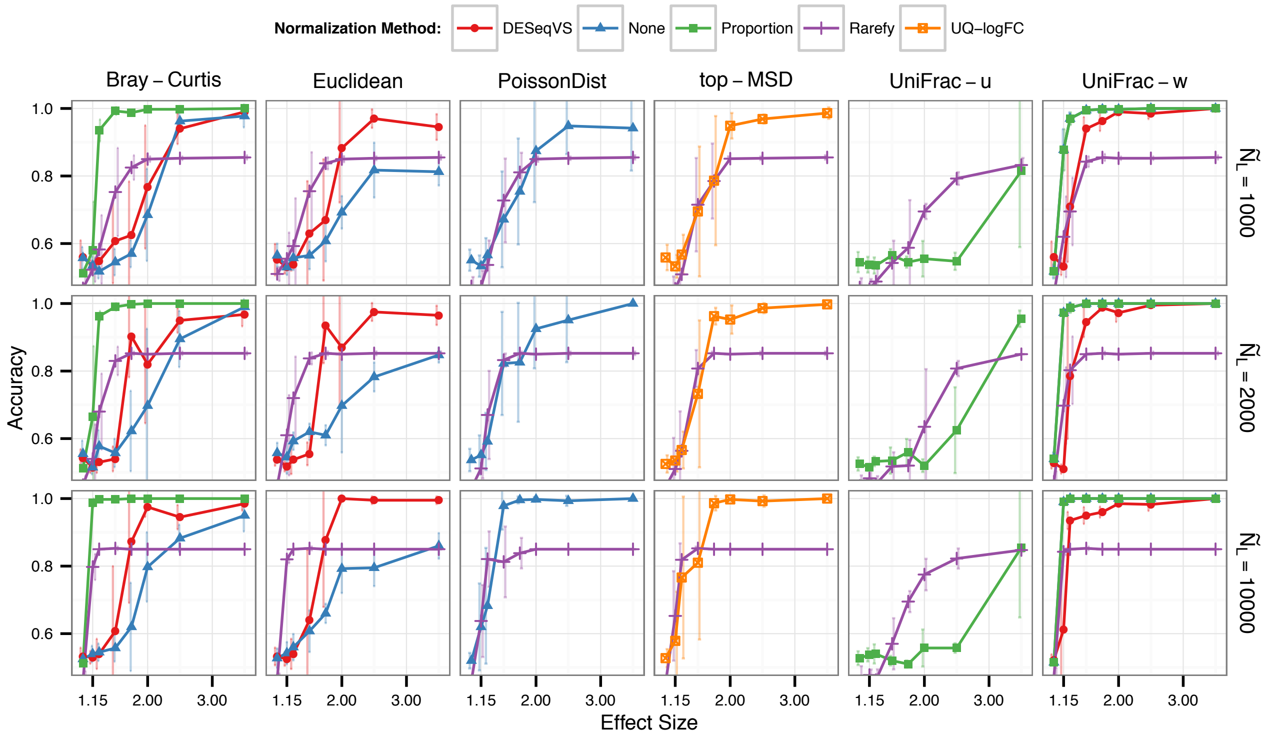
OTUs

158	56	214	39	47	4	11	11	5	3
124	54	212	29	40	3	10	7	8	6
129	46	216	33	42	4	13	7	3	6
11	3	14	3	1	39	95	63	29	37
19	7	34	7	0	88	237	137	73	86
9	1	15	1	2	29	84	51	14	29
		Simulated Ocean					Simulated Feces		

4. Perform clustering, evaluate accuracy.

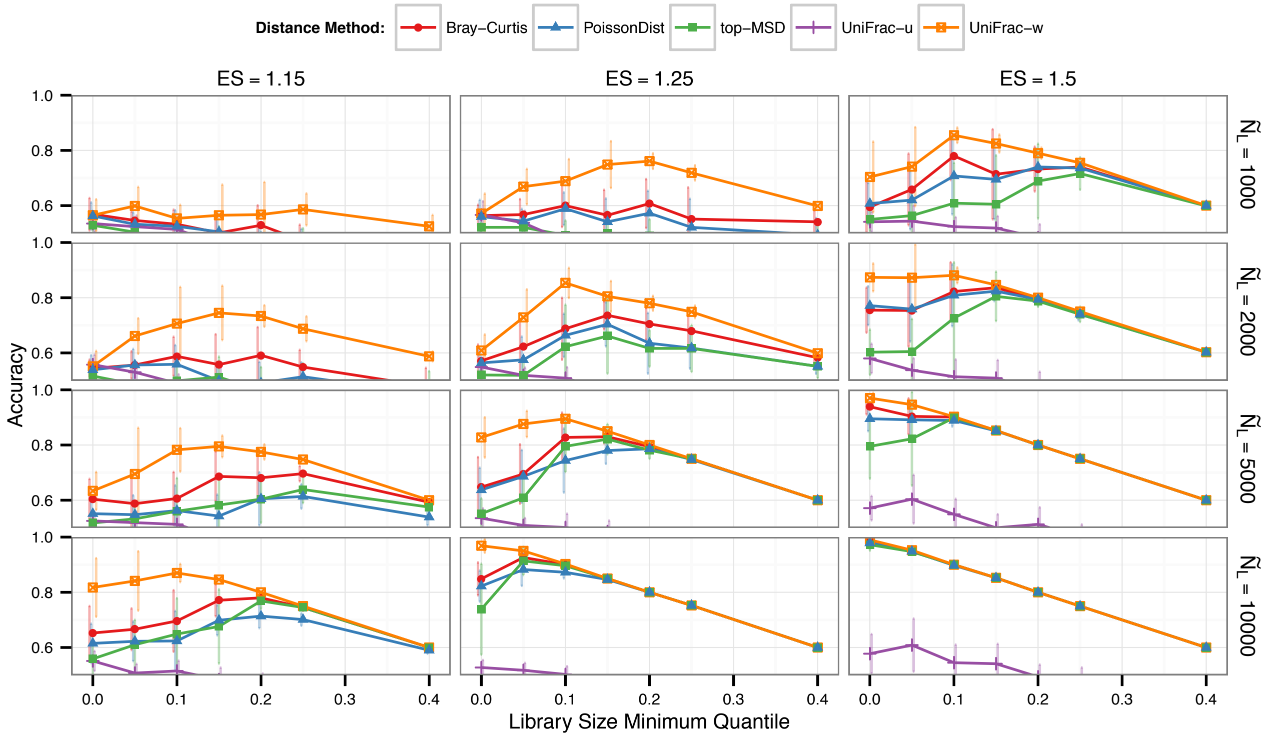
Repeat for each effect size and media library size.

Microbiome Clustering - Simulation



Microbiome Clustering - Simulation

Performance Depends on $\tilde{N}_L$

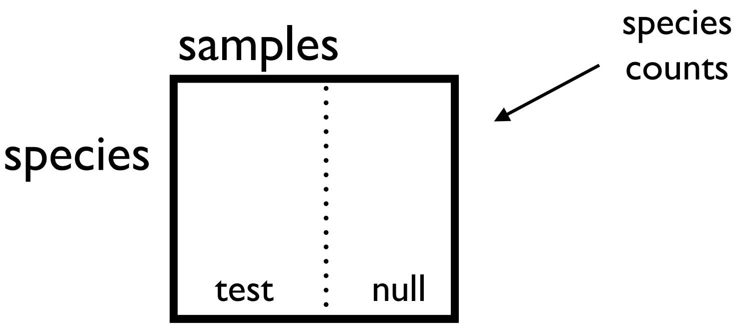


Issues with rarefying — clustering

- **Loss of Power**:
  1. Microbiome samples that cannot be classified because they were discarded (< $N_{L,min}$).
  2. Samples that are poorly distinguishable because of the discarded fraction of the original library.

- **Arbitrary threshold**:
  1. Choice clearly affects performance
  2. Optimum value, $^{*}N_{L,\,min}$, can't be known in practice

Differential Abundance



samples

species

species counts

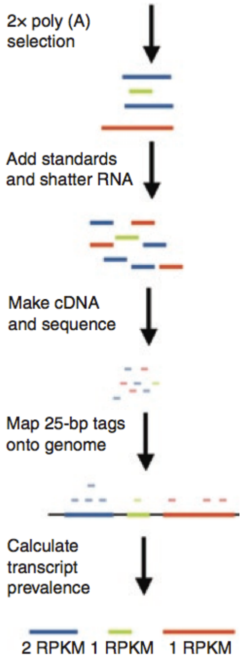
test

null

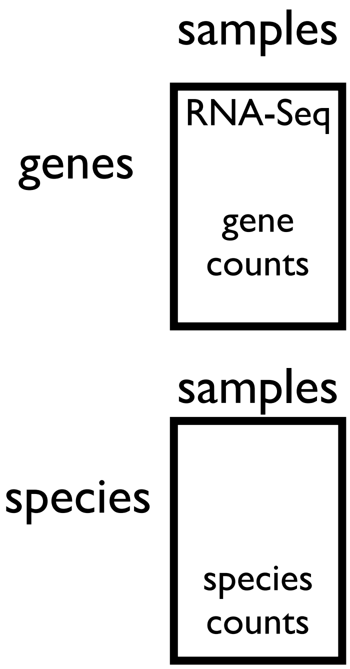
Which species have proportions that are different between the sample classes?

Differential Abundance

What about NB Mixture Model?



2× poly (A) selection
Add standards and shatter RNA
Make cDNA and sequence
Map 25-bp tags onto genome
Calculate transcript prevalence
2 RPKM
1 RPKM
1 RPKM



samples

genes

RNA-Seq

gene counts

samples

species

species counts

Mortazavi, et al (2008). Mapping & quantifying … transcriptomes by RNA-Seq. *Nature Methods*

Differential Abundance

What about NB Mixture Model?

Is Negative Binomial effective for this data?

1. Is there appreciable overdispersion?
2. Is there a useful across-species trend?

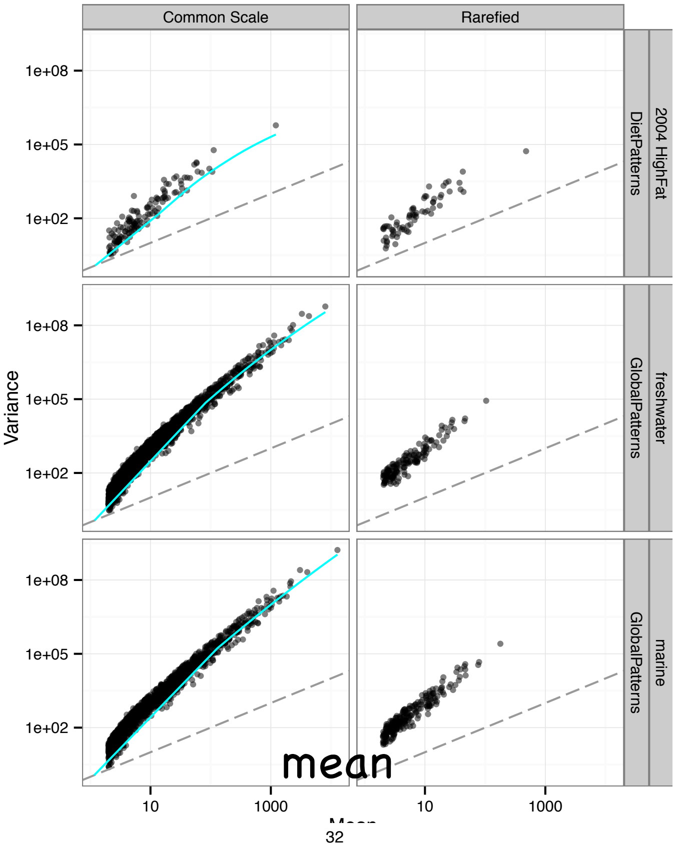
$$K_{ij} \sim NB(s_j\mu_i, \phi_i)$$

$$\nu_i \;=\; s_j\mu_i + \phi_i s_j\mu_i^2$$

- Robinson, Oshlack (2010). A scaling normalization… RNA-Seq data. *Genome Biology*
- *Anders, & Huber (2010).* Differential expression … sequence count data*. Genome Biology*

Differential Abundance

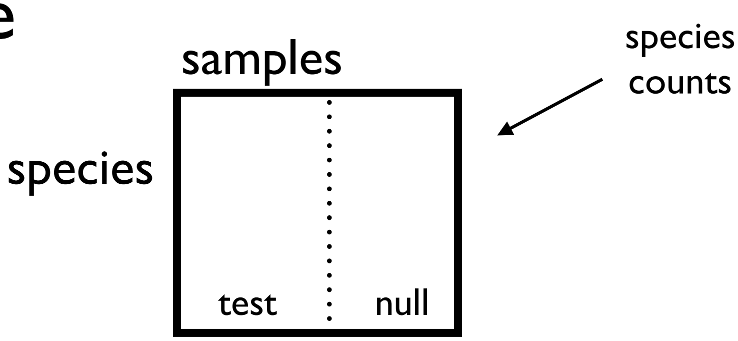
variance



mean

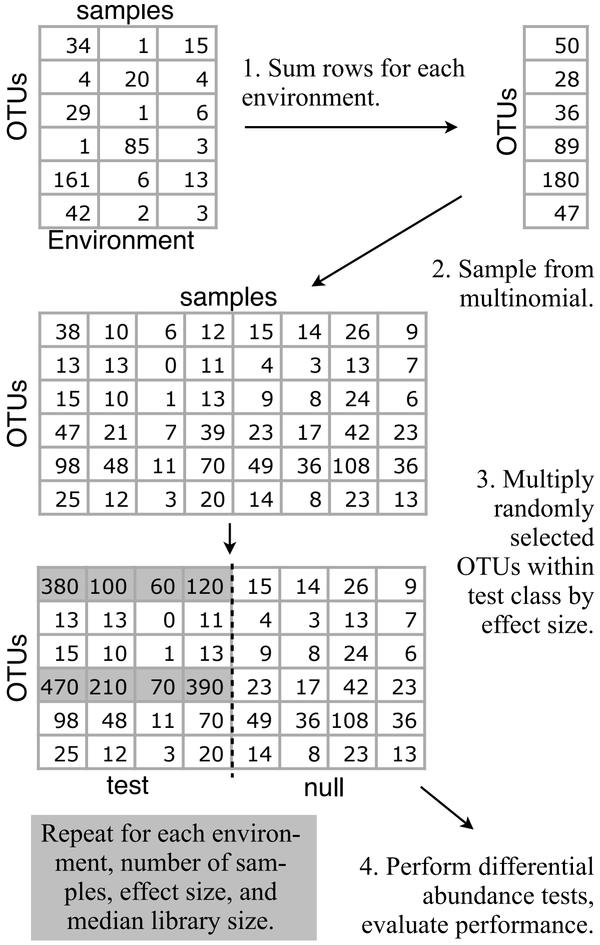
Microbiome Survey Biological Replicates

Differential Abundance Simulation



Which species have proportions that are different between the sample classes?

Differential Abundance Simulation



34	1	15
4	20	4
29	1	6
1	85	3
161	6	13
42	2	3

1. Sum rows for each environment.

50
28
36
89
180
47

2. Sample from multinomial.

38	10	6	12	15	14	26	9
13	13	0	11	4	3	13	7
15	10	1	13	9	8	24	6
47	21	7	39	23	17	42	23
98	48	11	70	49	36	108	36
25	12	3	20	14	8	23	13

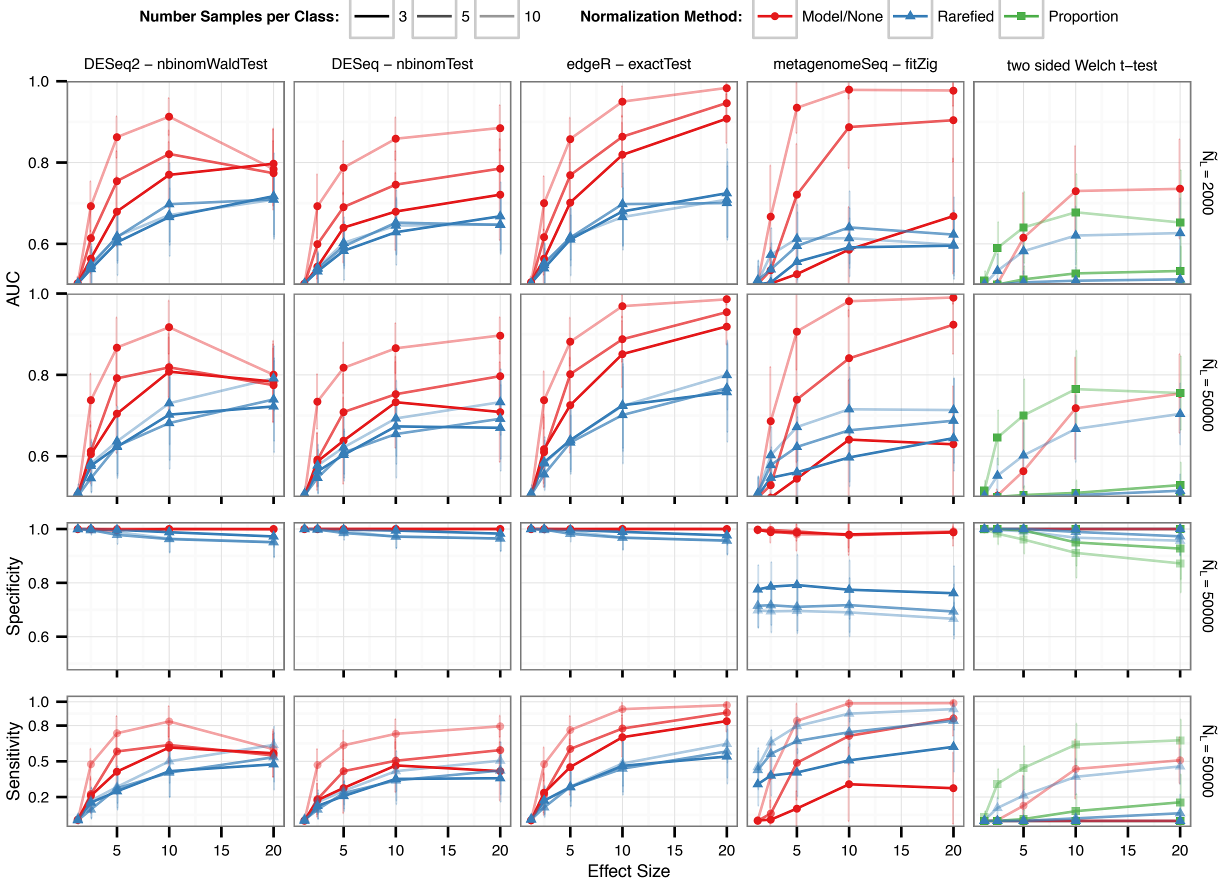
3. Multiply randomly selected OTUs within test class by effect size.

380	100	60	120	15	14	26	9
13	13	0	11	4	3	13	7
15	10	1	13	9	8	24	6
470	210	70	390	23	17	42	23
98	48	11	70	49	36	108	36
25	12	3	20	14	8	23	13

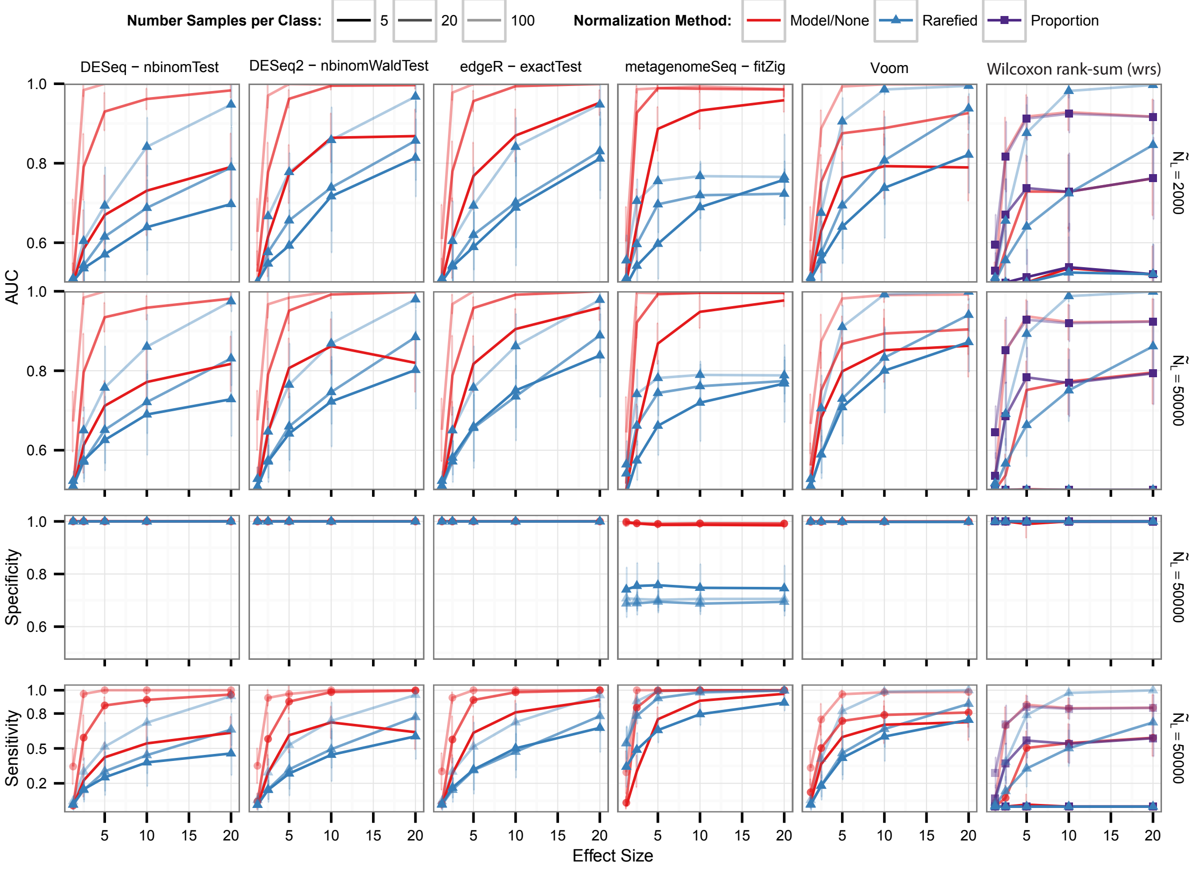
Repeat for each environment, number of samples, effect size, and median library size.

4. Perform differential abundance tests, evaluate performance.

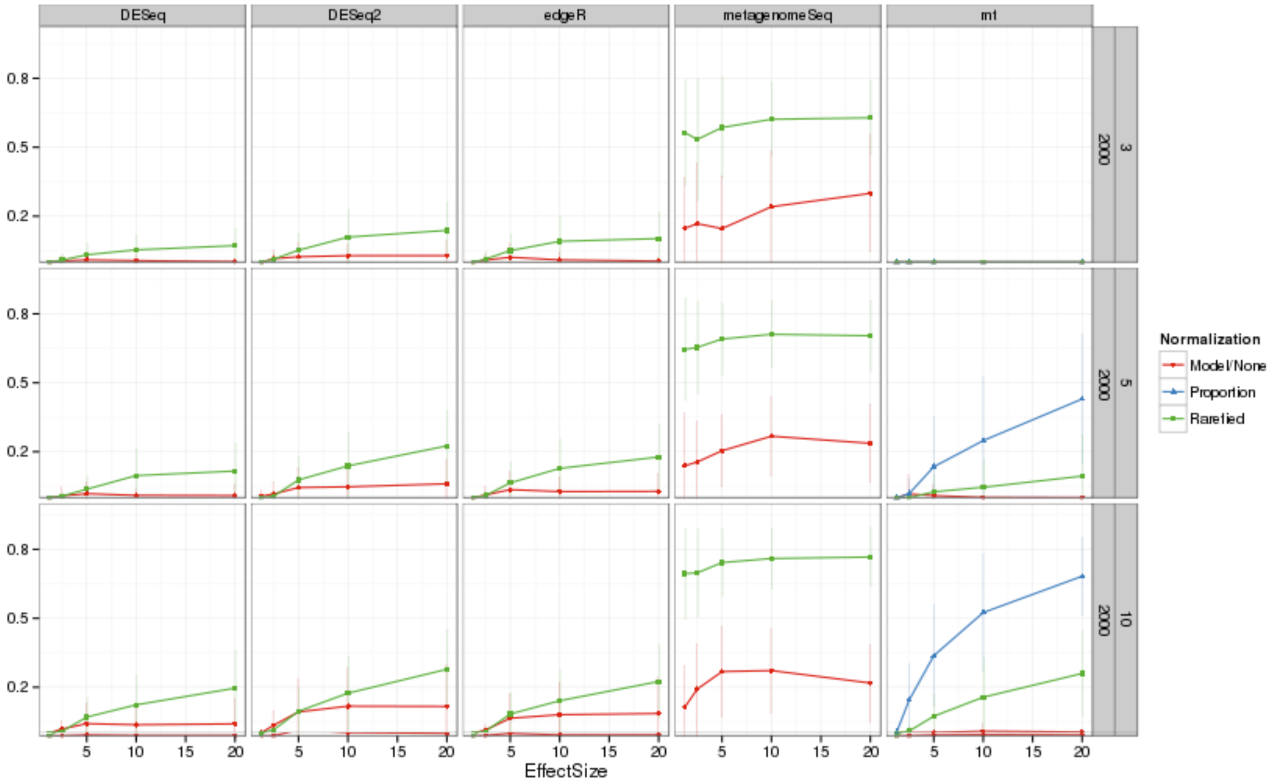
Differential Abundance - Simulation



Differential Abundance - Alt Simulation (Courtesy: Sophie Weiss, UC Boulder)



Differential Abundance - Simulation — False Positive Rates



Issues with rarefying — Differential Abundance

1. Rarefied counts worse sensitivity in every analysis method we attempted.

2. Rarefied counts also worse specificity (high FPs)

   - No accounting for overdispersion

   - Added noise from subsampling step

Transition: Lab 3

Negative Binomial mixture model for differential abundance multiple testing