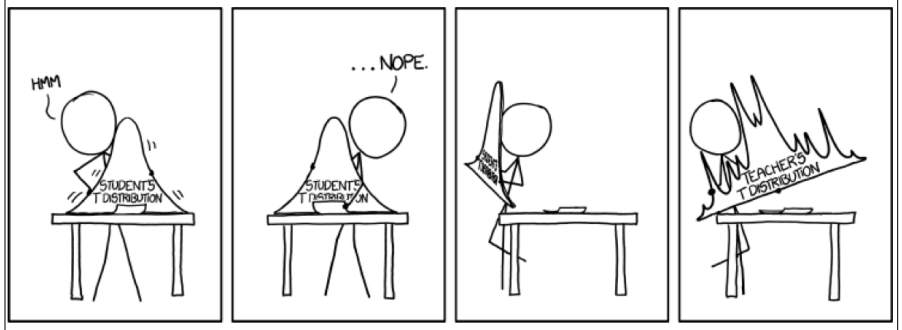
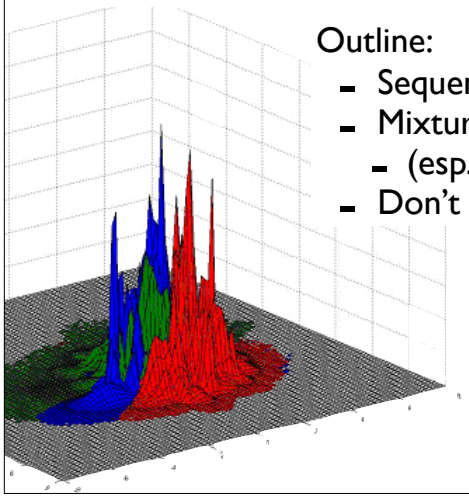
Lecture:
Mixture Models for Microbiome data



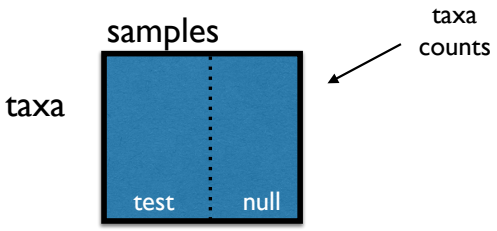
Lecture 3:
Mixture Models for Microbiome data



Outline:

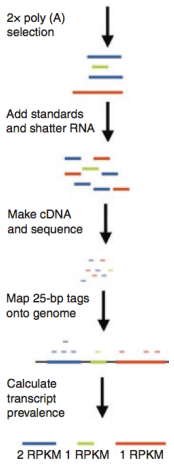
- Sequencing thought experiment
- Mixture Models (tangent)
  - (esp. Negative Binomial)
- Don't Rarefy. Ever.

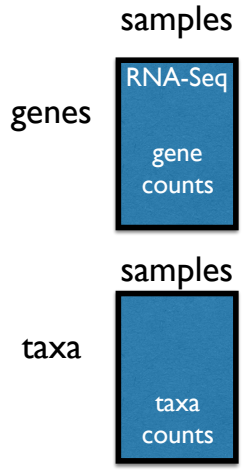
Differential Abundance



Our motivating scientific question:
Which taxa have proportions that are different between the sample classes?

Differential Abundance - analous to RNA-Seq



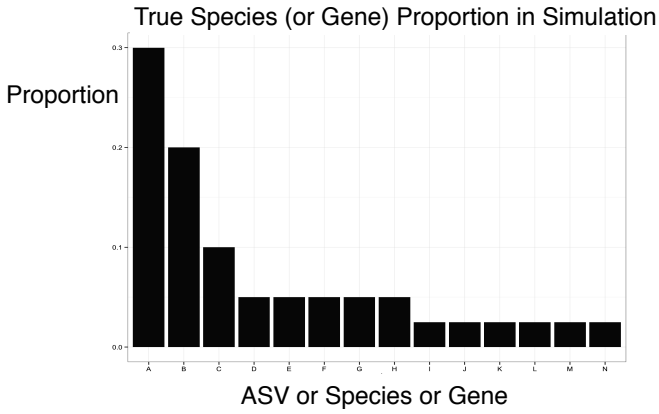


Mortazavi, et al (2008). Mapping & quantifying … transcriptomes by RNA-Seq. *Nature Methods*

Thought experiment for intuition building…

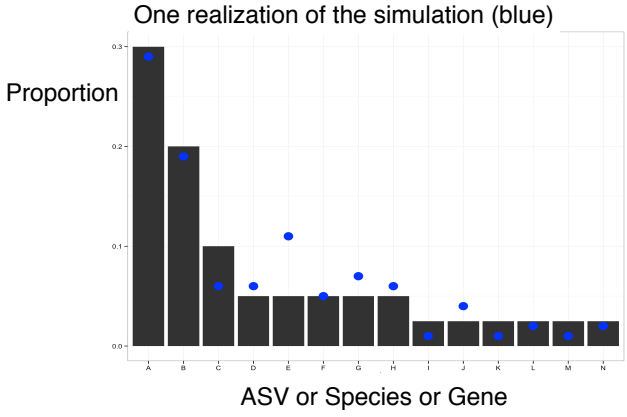
5

Model Uncertainty in NGS Count Data



6

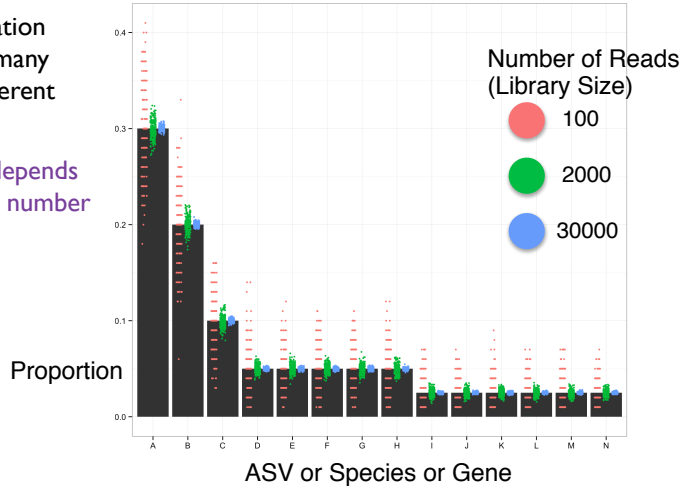
Model Uncertainty in NGS Count Data



7

Model Uncertainty in NGS Count Data

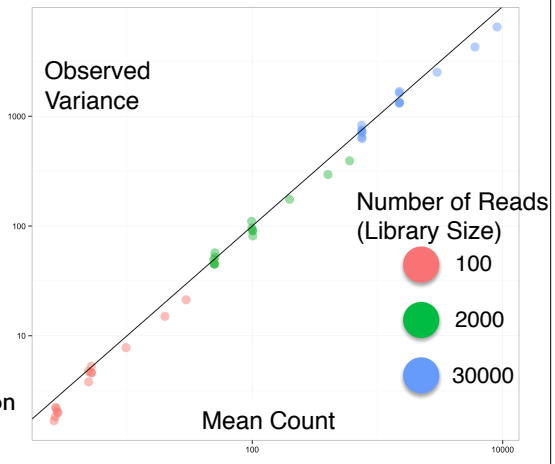
- Repeat simulation (resampling) many times and different library sizes
- Uncertainty depends (inversely) on number of reads



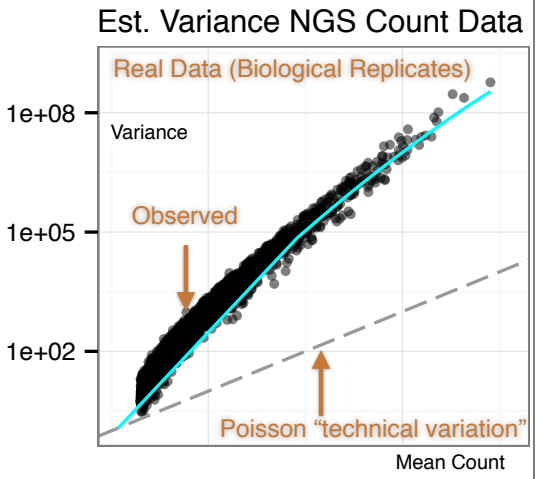
8

Model Uncertainty in NGS Count Data

- This simulation mirrors technical sequencing replicates well
- It is well characterized as a Poisson distribution
  - e.g. Variance == Mean
- Useful for intuition: re-sequencing from the same biological material on the same sequencer returns count data that looks Poisson
- What about biological replicates? How do you think that would look in this plot?

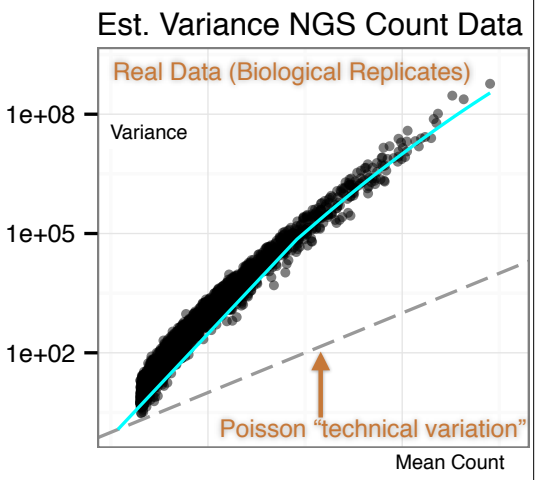


Model Uncertainty in NGS Count Data



Model Uncertainty in NGS Count Data

- The observed variance among biological replicates exceeds the mean (sometimes by a lot).
- The amount it exceeds the mean is usually still a strong smooth positive function of the mean, like the **light blue line**
- One way to model this is with the Negative Binomial distribution

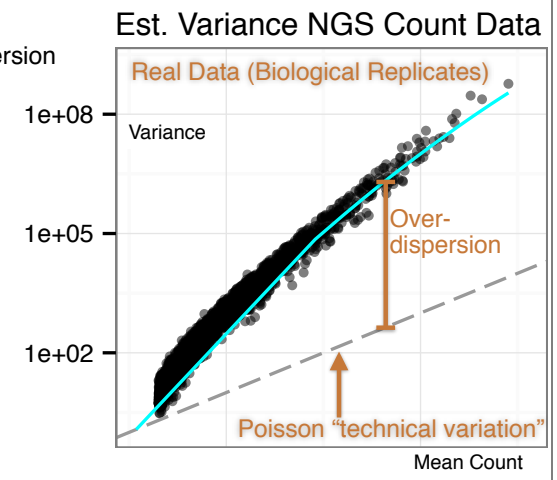


Model Uncertainty in NGS Count Data

Negative Binomial:

Variance = $u_{ic}s_j + \phi_{ic}s_j^2 u_{ic}^2$

Poisson Overdispersion



Model Uncertainty in NGS Count Data

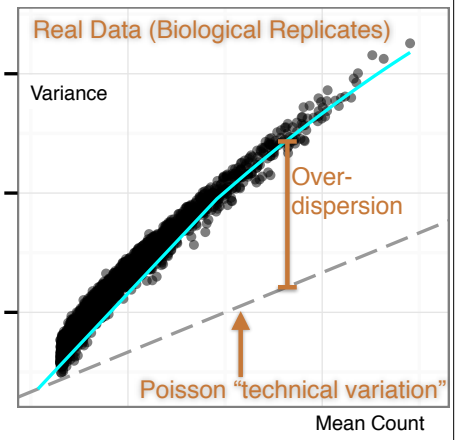
Negative Binomial

Variance = $u_{ic}s_j + \phi_{ic}s_j^2 u_{ic}^2$

Poisson Overdispersion

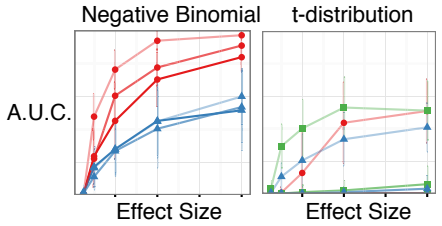
- How do you fit this many parameters?
- Share information across genes/features/ASVs in a joint inference of f ~ phi(count)
  - "fitting this curve is much easier than fitting a thousand phis"

Est. Variance NGS Count Data



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). Waste Not Want Not… *PLoS Computational Biology*
- Robinson, Oshlack (2010). A scaling normalization… RNA-Seq data. *Genome Biology*
- *Anders, & Huber (2010).* Differential expression … sequence count data. *Genome Biology*

Inefficient Normalization by "rarefying"

Inefficient Normalization by "rarefying"

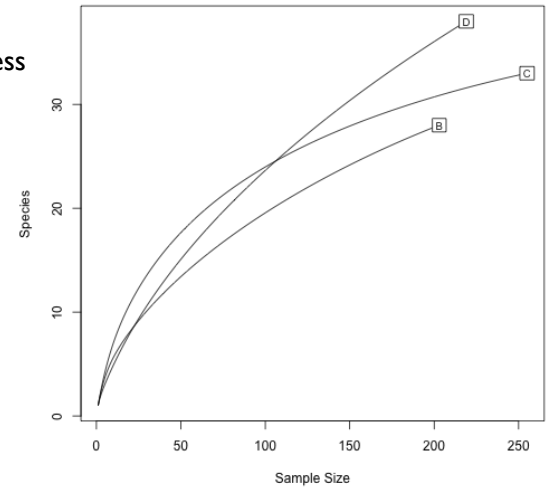
`rarefying ≠ rarefaction`

Inefficient Normalization by "rarefying"

the original idea…

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

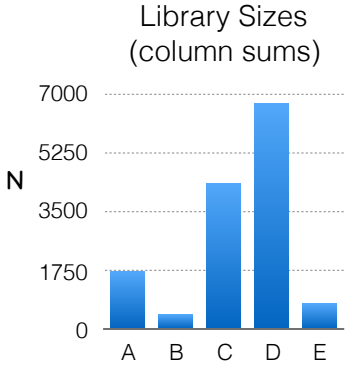
rarefaction curves



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

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 $N_{L,min}$

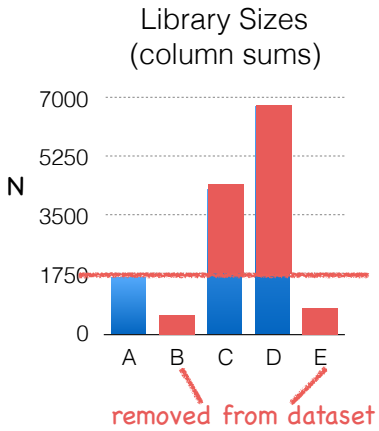


Hughes & Hellmann (2005) *Methods in Enzymology*

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 $N_{L,min}$



Hughes & Hellmann (2005) *Methods in Enzymology*

Gotelli, & Colwell (2001) *Ecology Letters*

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

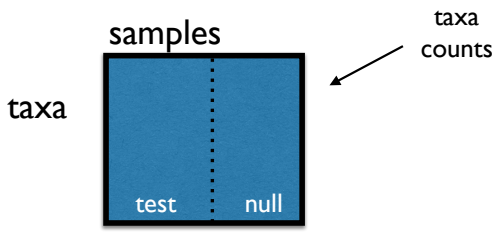
Issues with rarefying — clustering

- **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

Transition: Lab

Negative Binomial mixture model for differential abundance multiple testing using DESeq2, etc.

Differential Abundance



Scientific Question:
Which taxa have proportions that are different between the sample classes?

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.

Scientific Question:
Which taxa have proportions that are different between the sample classes?

Null Hypothesis:
The proportions of a taxa in the two sample classes are the same

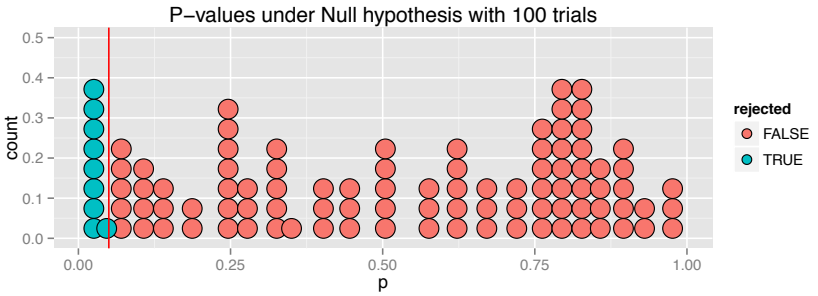
Hypothesis Tests - some examples

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`

There are obviously a lot more available in R…

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 $m*\alpha$



m = 100
alpha = 0.05
E(FP) = 5

Inefficient Normalization by "rarefying"

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

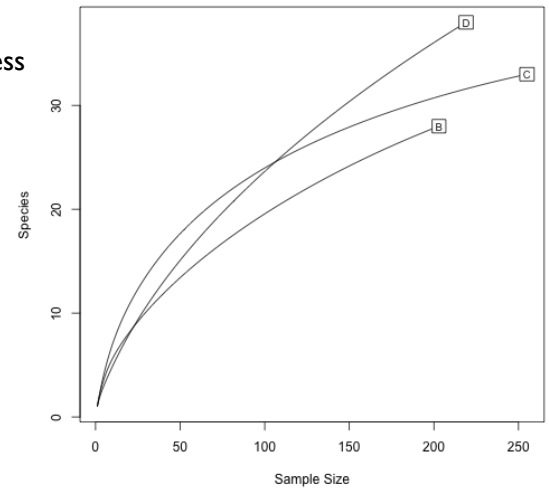
  paradigm: rarefy - UniFrac - PCoA - Write Paper
- This approach has "leaked" into formal settings, still quite a bit of inertia to maintain the practice

Inefficient Normalization by "rarefying"

the original idea…

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

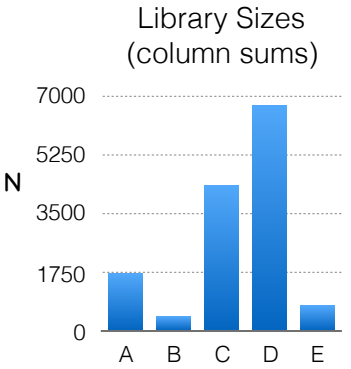
rarefaction curves



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

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 $N_{L,min}$

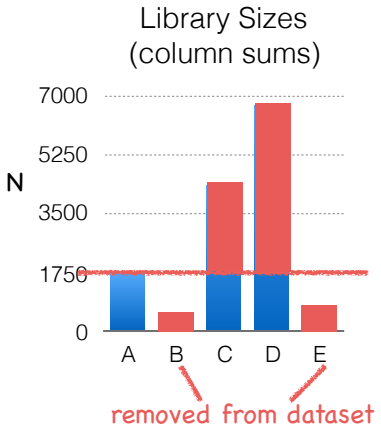


Hughes & Hellmann (2005) *Methods in Enzymology*

Gotelli, & Colwell (2001) *Ecology Letters*

Inefficient Normalization by "rarefying"

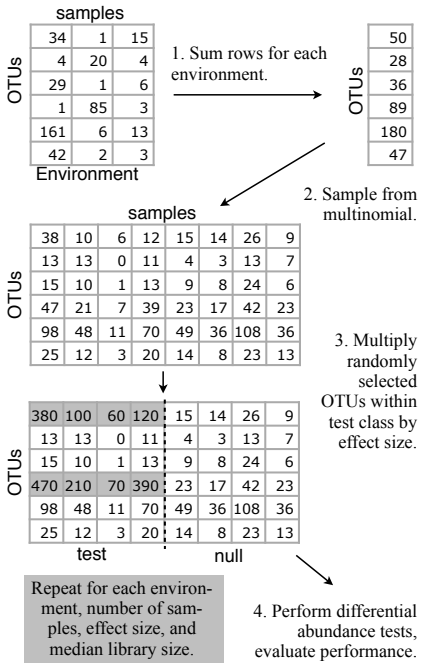
1. Select a minimum library size $N_{L,min}$
2. Discard libraries (samples) that are smaller than $N_{L,min}$
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$

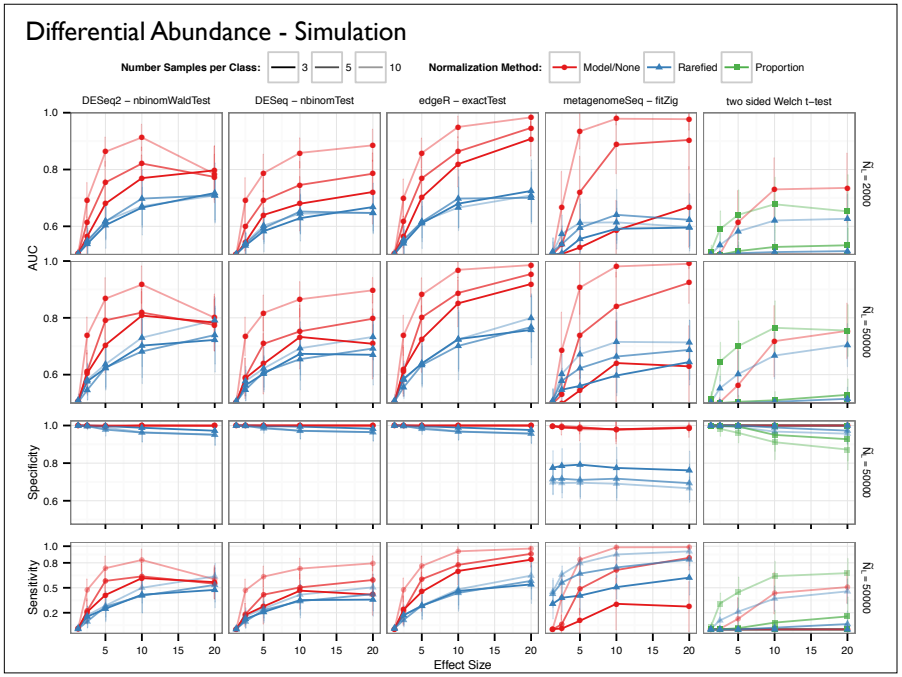


Hughes & Hellmann (2005) *Methods in Enzymology*

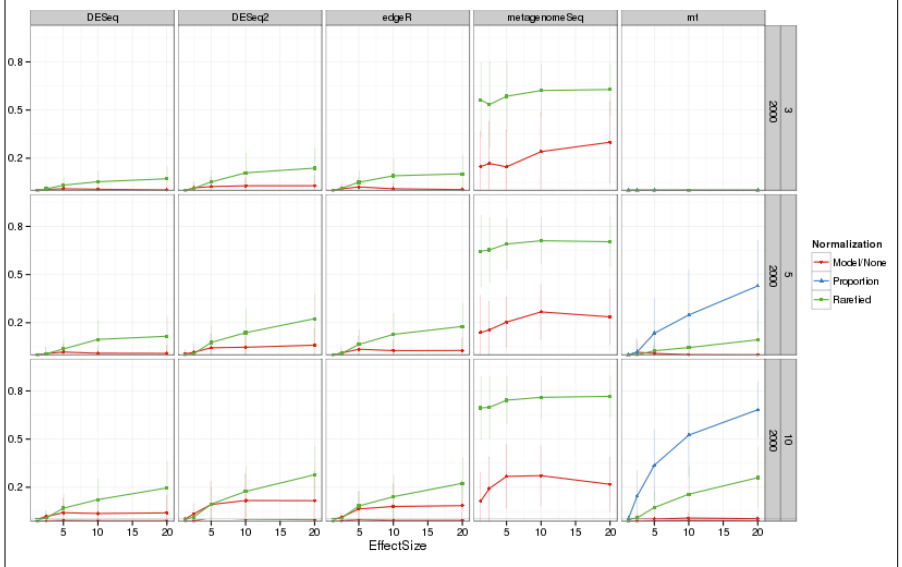
Gotelli, & Colwell (2001) *Ecology Letters*

Differential Abundance Simulation





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

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

End for now…

Further details
performance degradation of
clustering results by rarefying…

Inefficient Normalization by "rarefying"

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

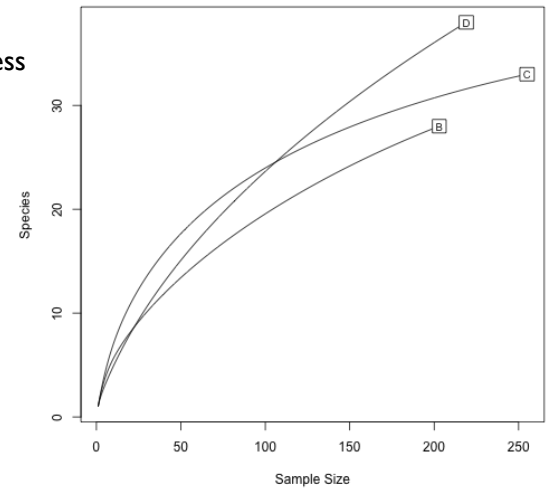
  paradigm: rarefy - UniFrac - PCoA - Write Paper
- This approach has "leaked" into formal settings, still quite a bit of inertia to maintain the practice

Inefficient Normalization by "rarefying"

the original idea…

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

rarefaction curves

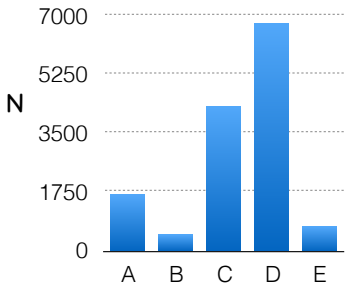


Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*

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)

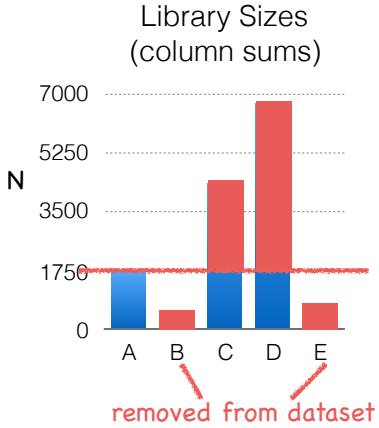


Hughes & Hellmann (2005) *Methods in Enzymology*

Gotelli, & Colwell (2001) *Ecology Letters*

Inefficient Normalization by "rarefying"

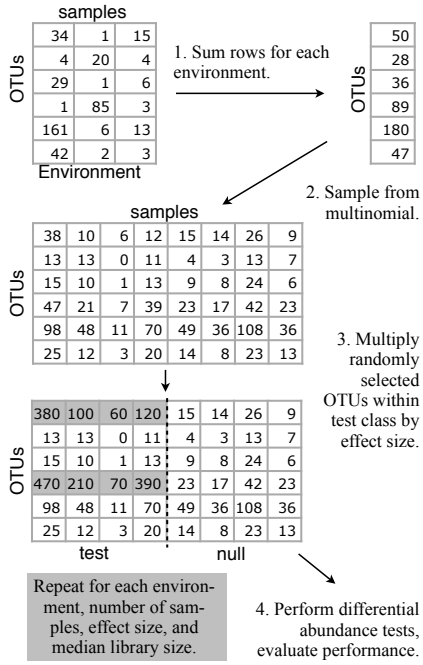
1. Select a minimum library size $N_{L,min}$
2. Discard libraries (samples) that are smaller than $N_{L,min}$
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$



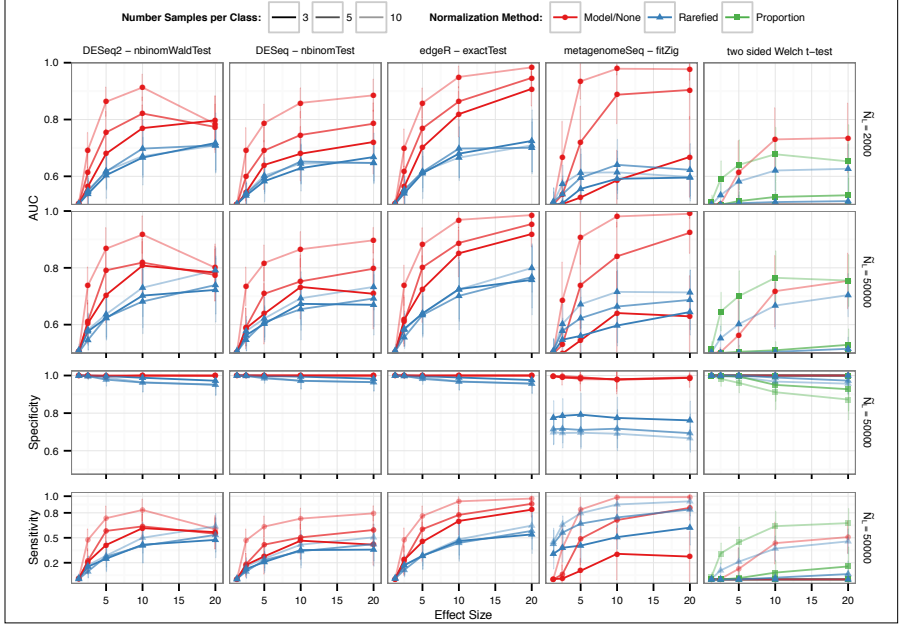
Hughes & Hellmann (2005) *Methods in Enzymology*

Gotelli, & Colwell (2001) *Ecology Letters*

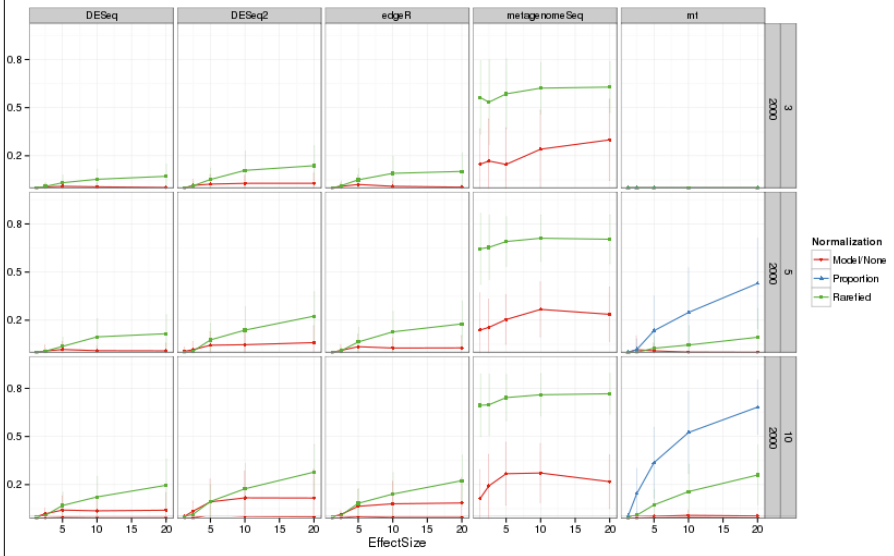
Differential Abundance Simulation



Differential Abundance - Simulation



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

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

Tangent: 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

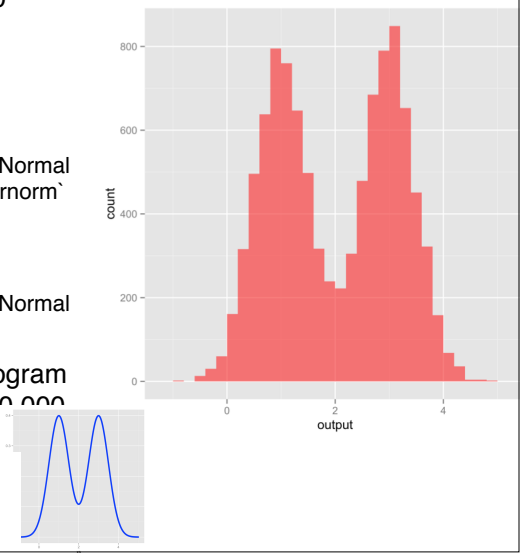
Generate a random number from a Normal with mean 1 and variance 0.25. R: `rnorm` function.

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.

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



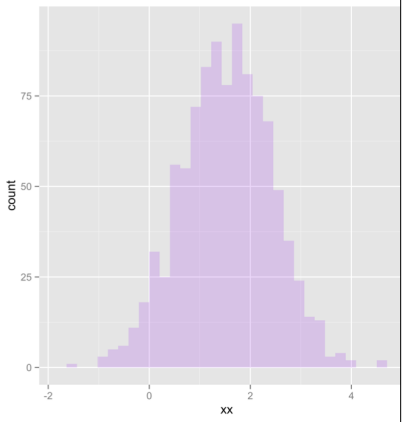
Finite Mixture Model

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

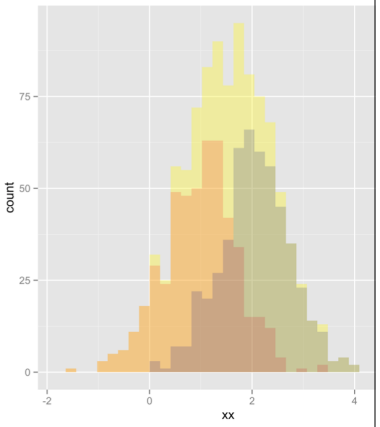
Here we knew the answer

(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)

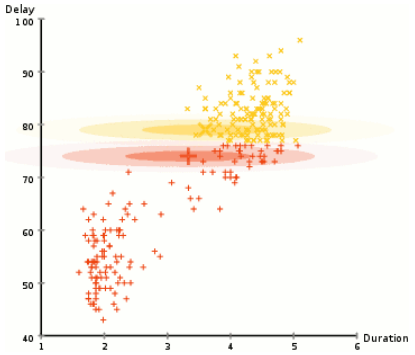
Very popular iterative procedure

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

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

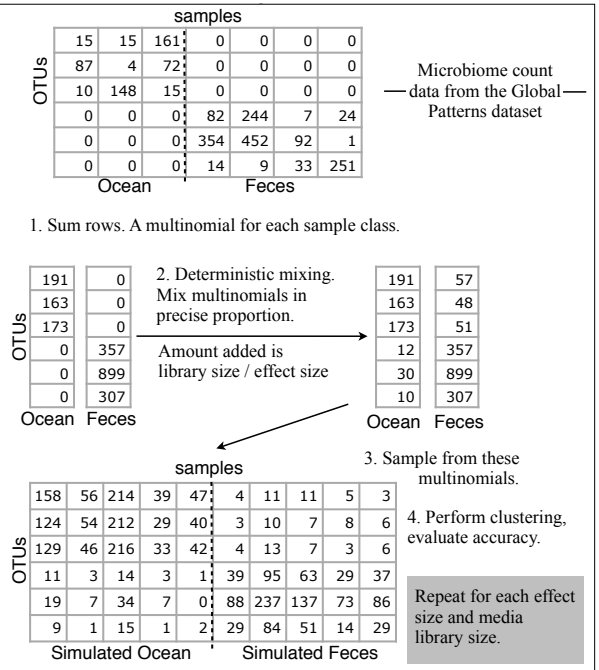
Decomposing the mixtures using the EM algorithm.

Common Infinite Mixture Models

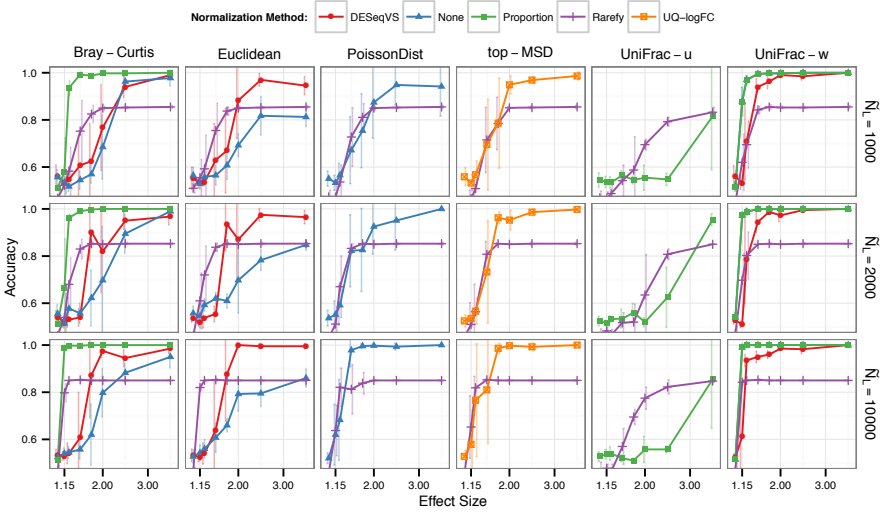
Gamma-Poisson (Negative Binomial) for read counts

Dirichlet-Multinomial (Birthday problem and the Bayesian setting).

Microbiome Clustering Simulation



Microbiome Clustering - Simulation



Microbiome Clustering - Simulation

Performance Depends on $\tilde{N}_L$