Metastats 2.0

An improved method and software for analyzing metagenomic data

Joseph N. Paulson
jpaulson@umiacs.umd.edu
Mihai Pop
mpop@umiacs.umd.edu
Héctor Corrada Bravo
hcorrada@umiacs.umd.edu

Abstract:
Here we present major improvements to Metastats software and underlying statistical methods.

1) A mixed-model zero-inflated Gaussian distribution.
2) A novel normalization method.

Application Background

- What is metagenomics?
- Why is it important?
- What do I hope to do?

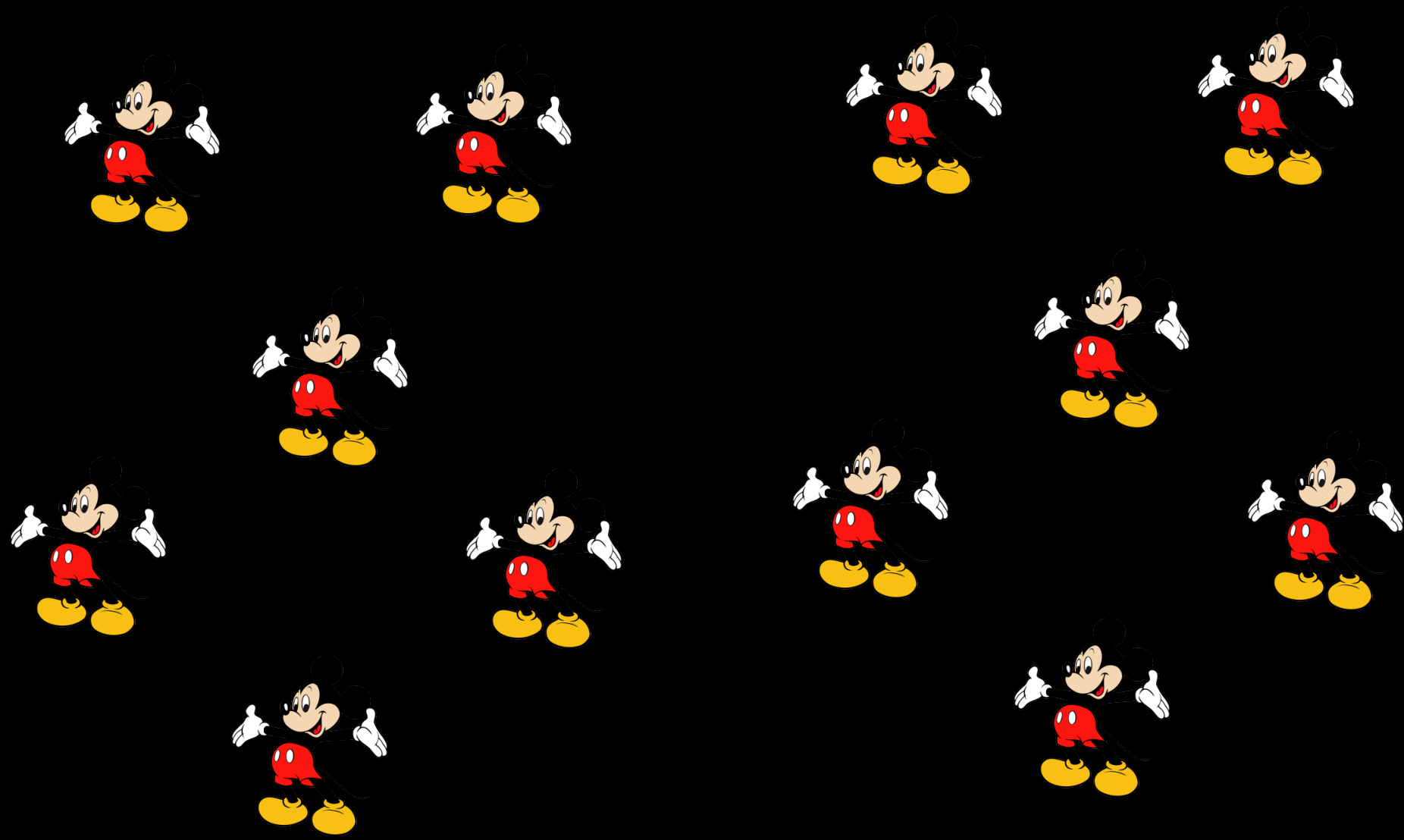
Application Background

Detection of differential abundance!

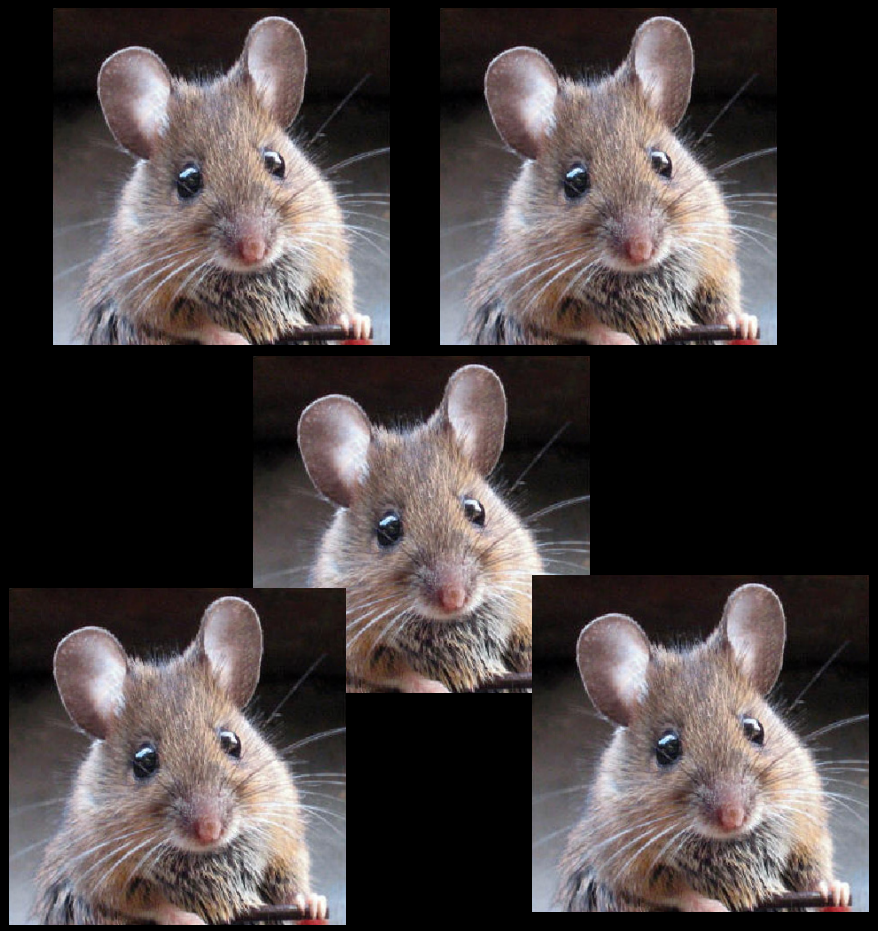
Definition: A count, c_ij is the number of reads annotated as a particular taxa i for the jth sample

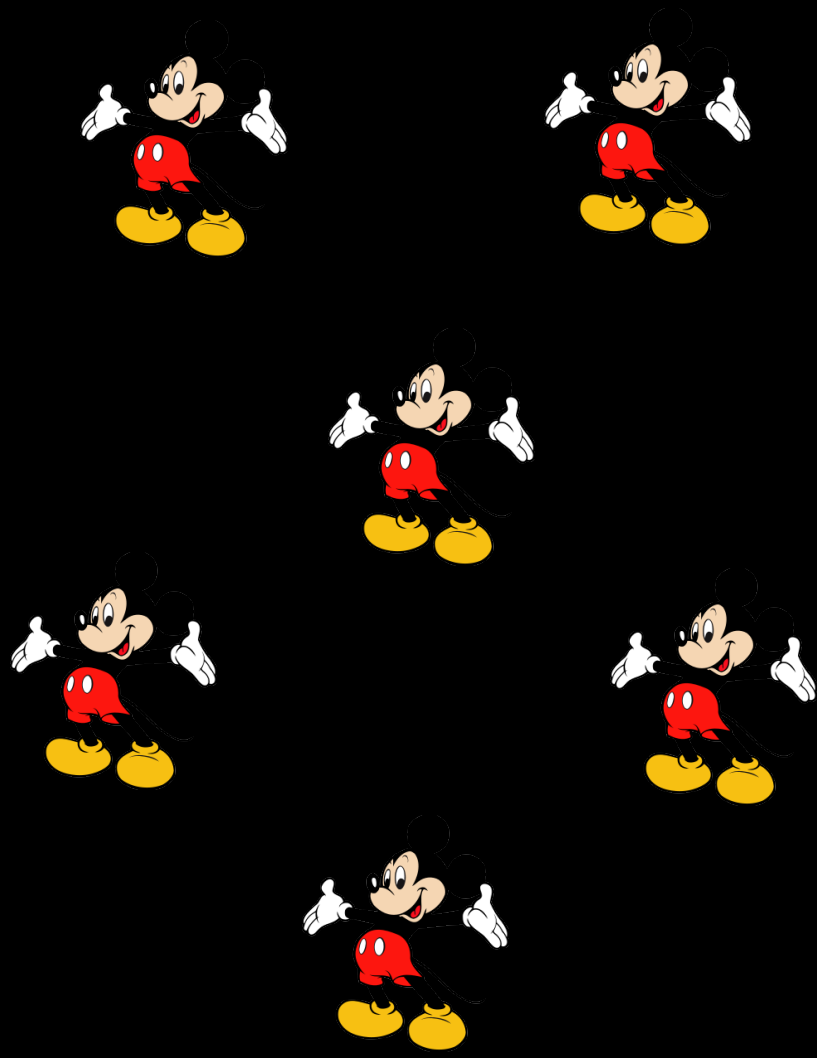
	S1	S2	.....	S(N-1)	SN
T1	c(1,1)	c(1,2)	.....	c(1,N-1)	c(1,N)
T2	c(2,1)	c(2,2)			
.	.				.
.	.				.
.	.				.
T(M-1)	c(M-1,1)				
TM	c(M,1)		.....		c(M,N)

Mouse diet



Mouse diet





Datasets

- Gates – matched study
  - 1016 samples
  - 4 Countries
    - Gambia
    - Mali
    - Bangladesh
    - Kenya
  - Monthly age metadata
  - Half controls, half cases

BILL & MELINDA GATES foundation

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples

James Robert White[1], Niranjan Nagarajan[2], Mihai Pop[3]*

$$\bar{X}_{it} = \frac{1}{n_t} \sum_{j \in treatment\ t} f_{ij}$$

$$s_{it}^2 = \frac{1}{n_t - 1} \sum_{j \in treatment\ t} (f_{ij} - \bar{X}_{it})^2$$

$$t_i = \frac{\bar{X}_{i1} - \bar{X}_{i2}}{(s_{i1}^2/n_1 + s_{i2}^2/n_2)^{.5}}$$

$$p_i = \frac{\{|t_i^{ob}| \geq |t_i| b \in 1 \ldots B\}}{B}$$

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples

James Robert White[1], Niranjan Nagarajan[2], Mihai Pop[3]*

Too slow! Can't handle large datasets

- More and more data coming daily!
- Lots of for loops
- Error

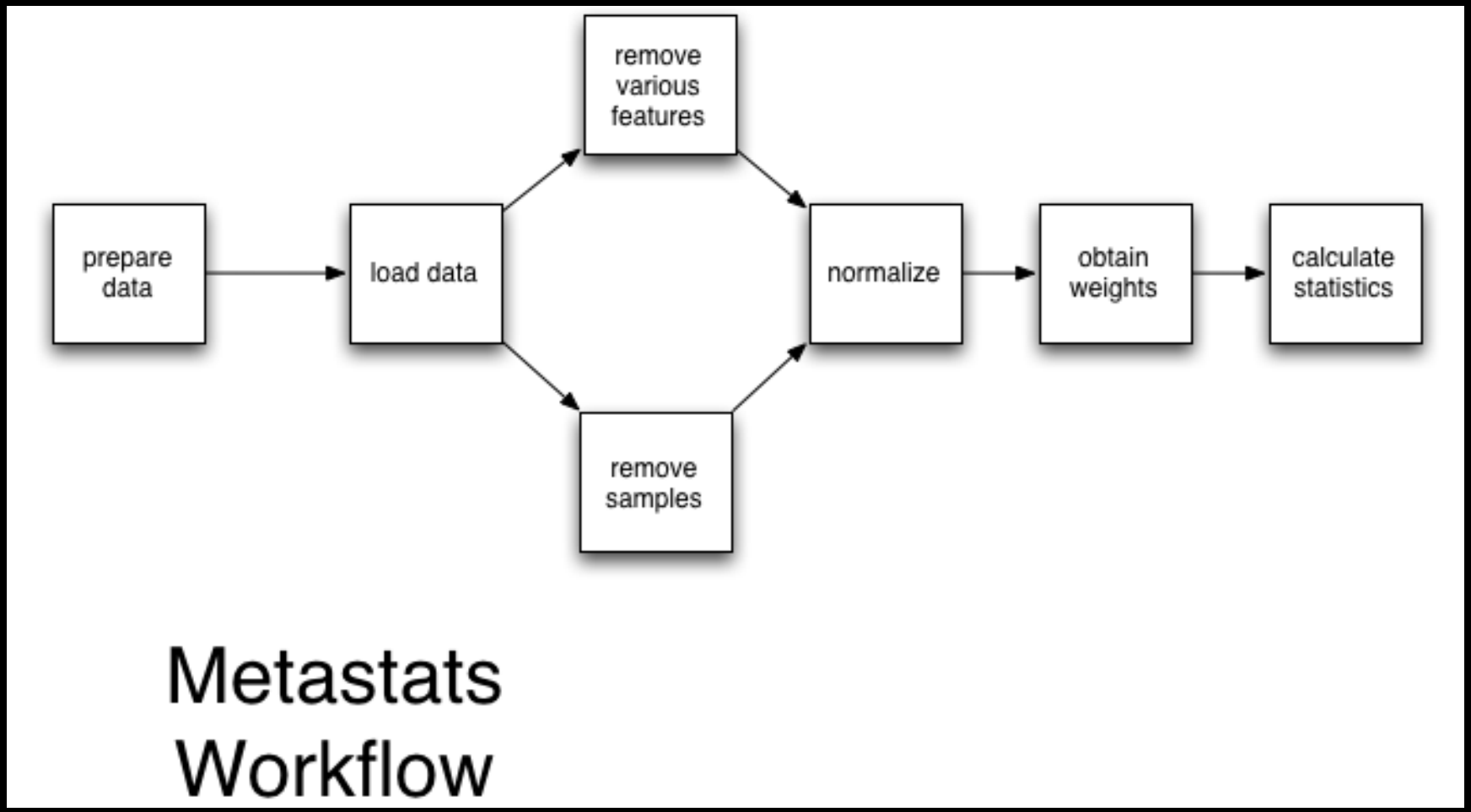
Doesn't account for depth of coverage

Many "spurious" zeros

Normalization induces spurious correlations

important in time series analyses

Metastats workflow



Normalization

- Ratio Normalization:
  - What are the issues with it??

$$y_{Aj} = c_{Aj}/(c_{1j} + ... + c_{Aj} + c_{Bj} + ...c_{Mj})$$

  - Spurious correlation [1]
  - False negatives [2]
  - False positives [2]

[1] Pearson, Mathematical Contributions to the Theory of Evolution. On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs

[2] Bullard et. al., Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, BMC Bioinformatics, 2010

Example of cumulative distribution normalization

Original

2	4	4
5	4	14
4	6	8
3	5	8
3	3	9

→

Ordered

2	3	4
3	4	8
3	4	8
4	5	9
5	6	14

→

Cumulative Sum

2	3	4
5	7	12
8	11	20
12	16	29
17	32	43

→

Averaged

3	3	3
8	8	8
13	13	13
19	19	19
30	30	30

Example of cumulative distribution normalization

Averaged

3	3	3
8	8	8
13	13	13
19	19	19
30	30	30

→

First difference

3	3	3
5	5	5
5	5	5
6	6	6
11	11	11

→

Re-ordered

3	5	3
11	5	11
6	11	5
5	6	5
5	3	6

Cumulative scaling normalization

Genes are sampled preferentially as sequencing yield increases (# PCR cycles biases as well).

Unlike RNA-seq data[c], we assume finite capacity in metagenomic communities:

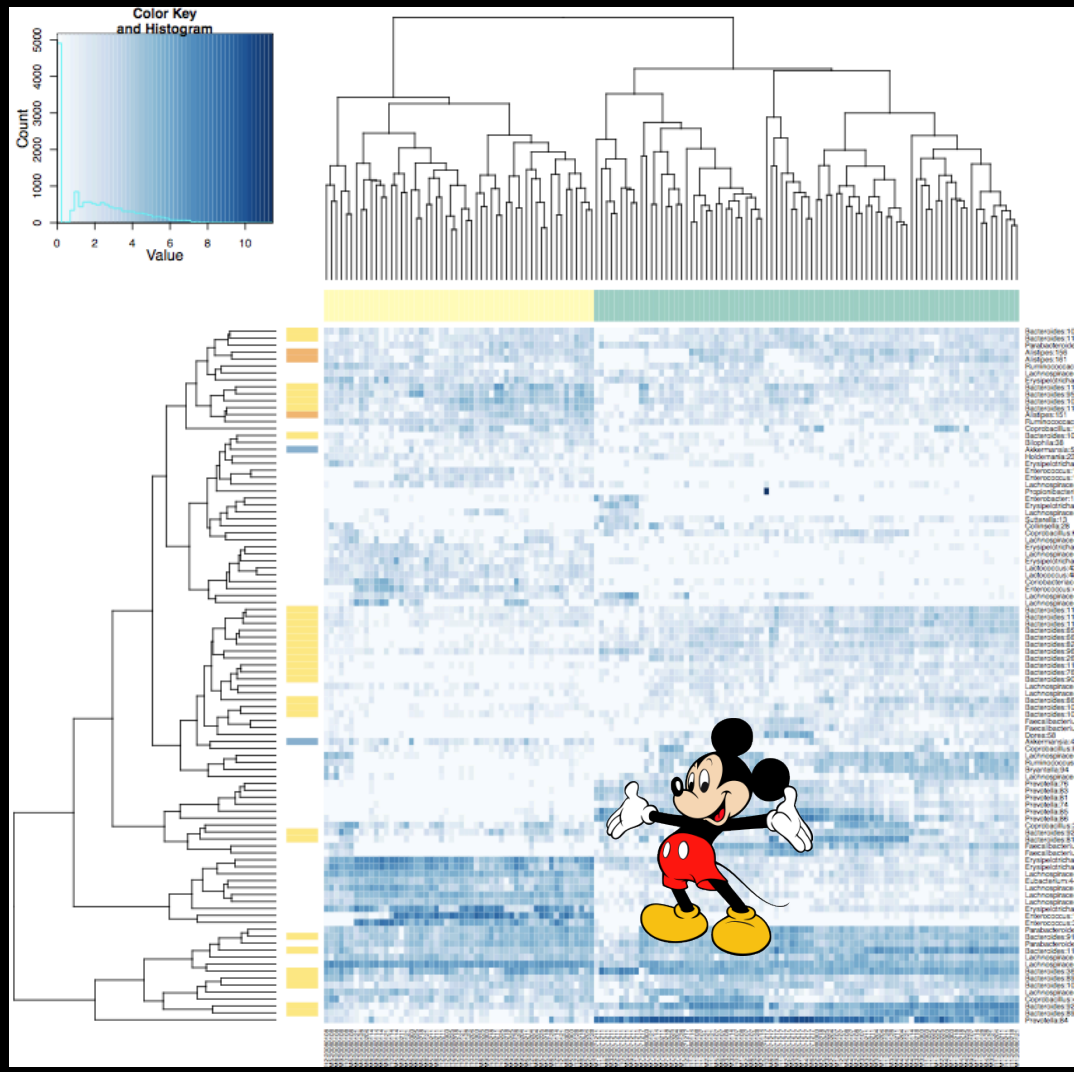
$$S_{95j} = \sum_i c_{ij} \leq q_{95j}$$

This procedure addresses the issues:

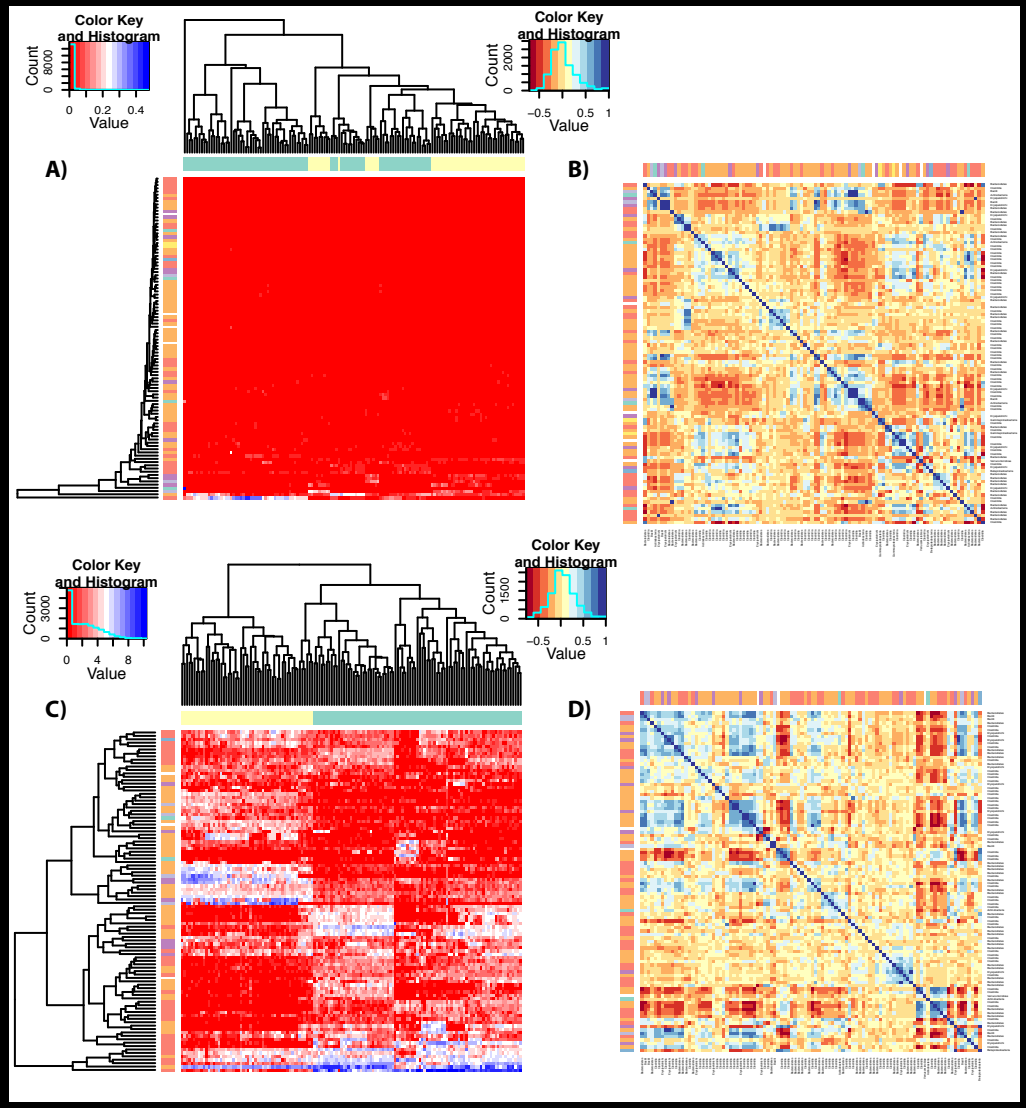
- constraints communities with respect to a total capacity
- No undue influence on features that are preferentially sampled.

---

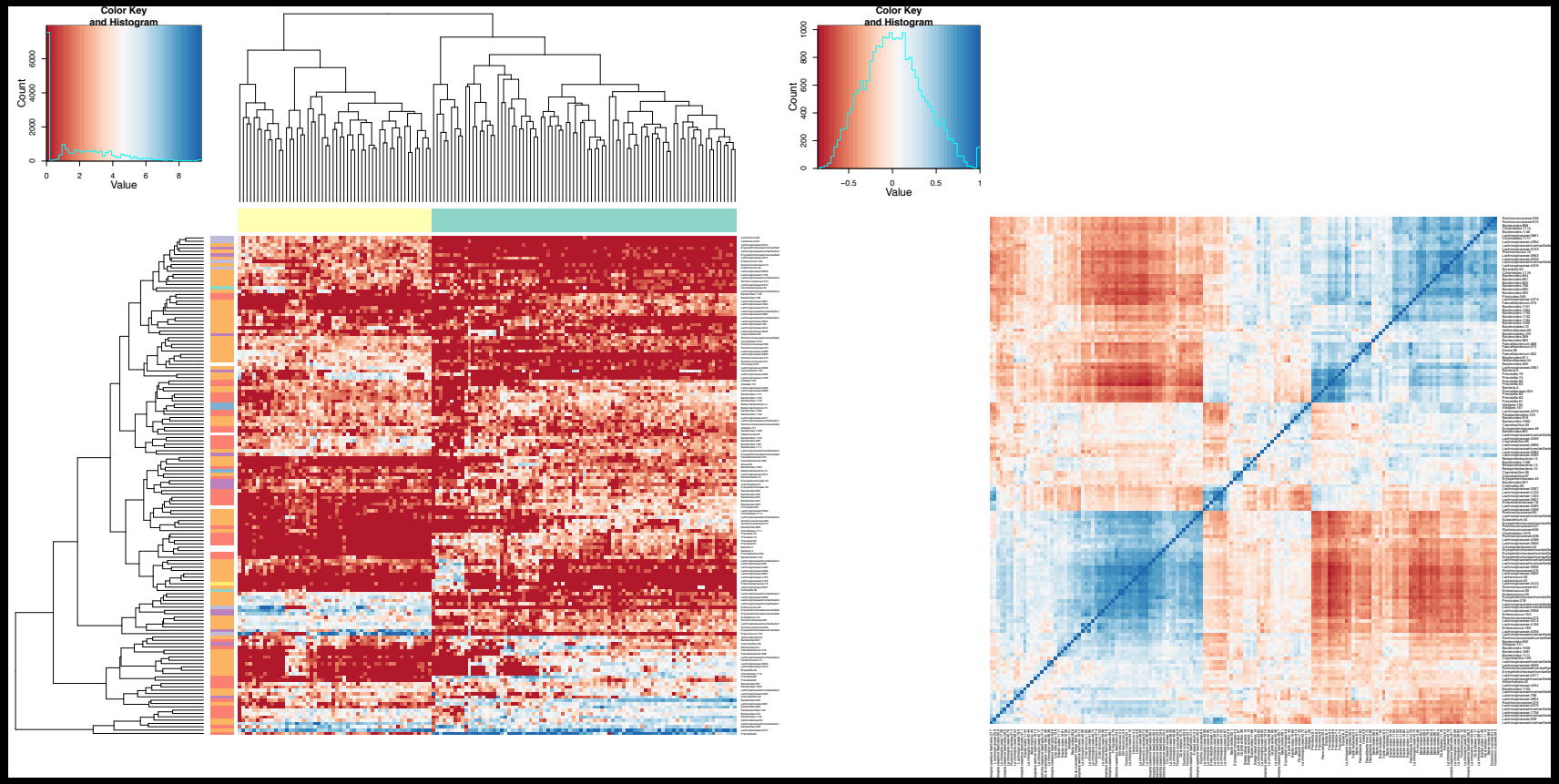
[c] RNA-seq data normalization: $y_{ij} = c_{ij} / q_{75j}$



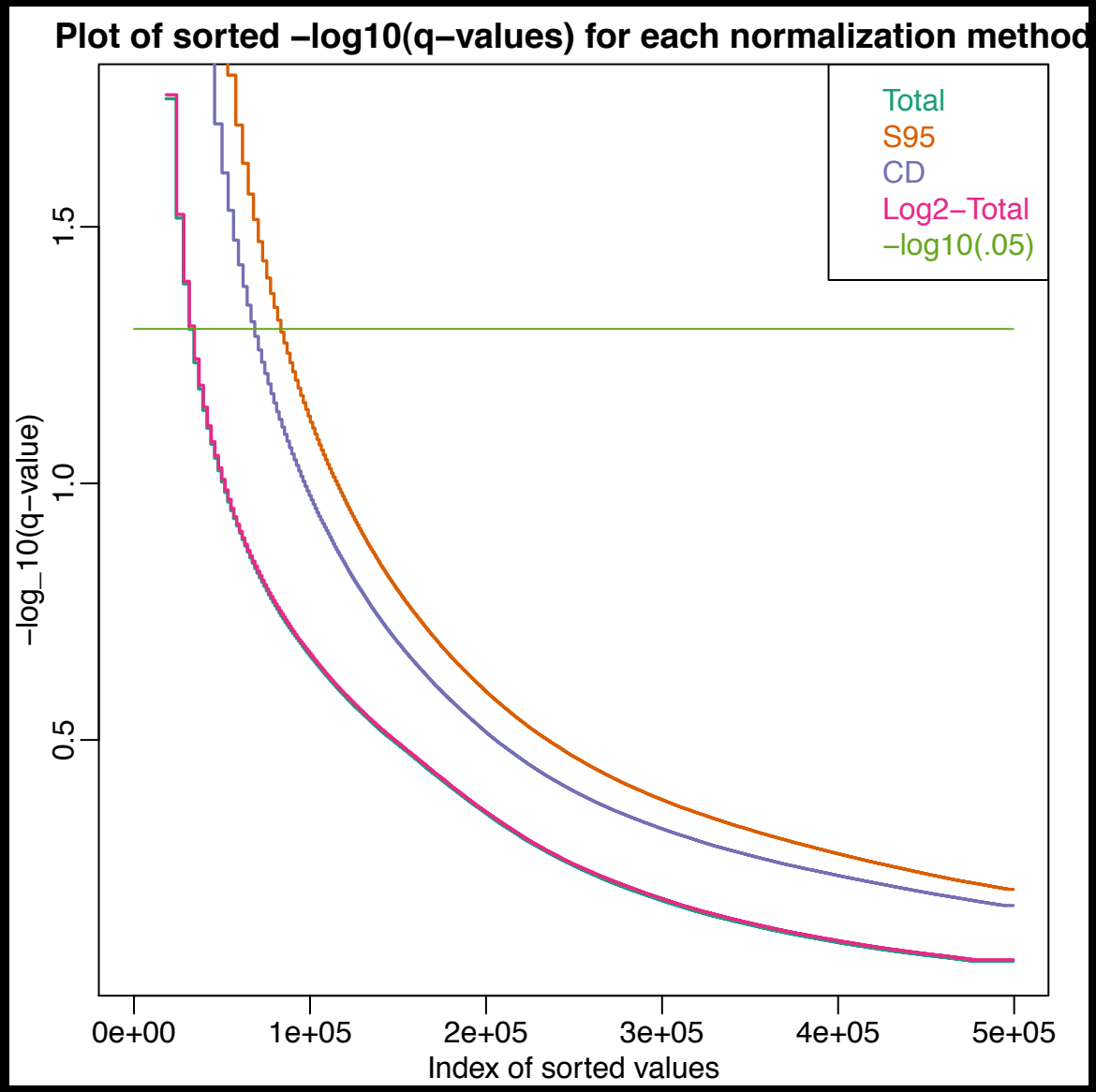
Normalization



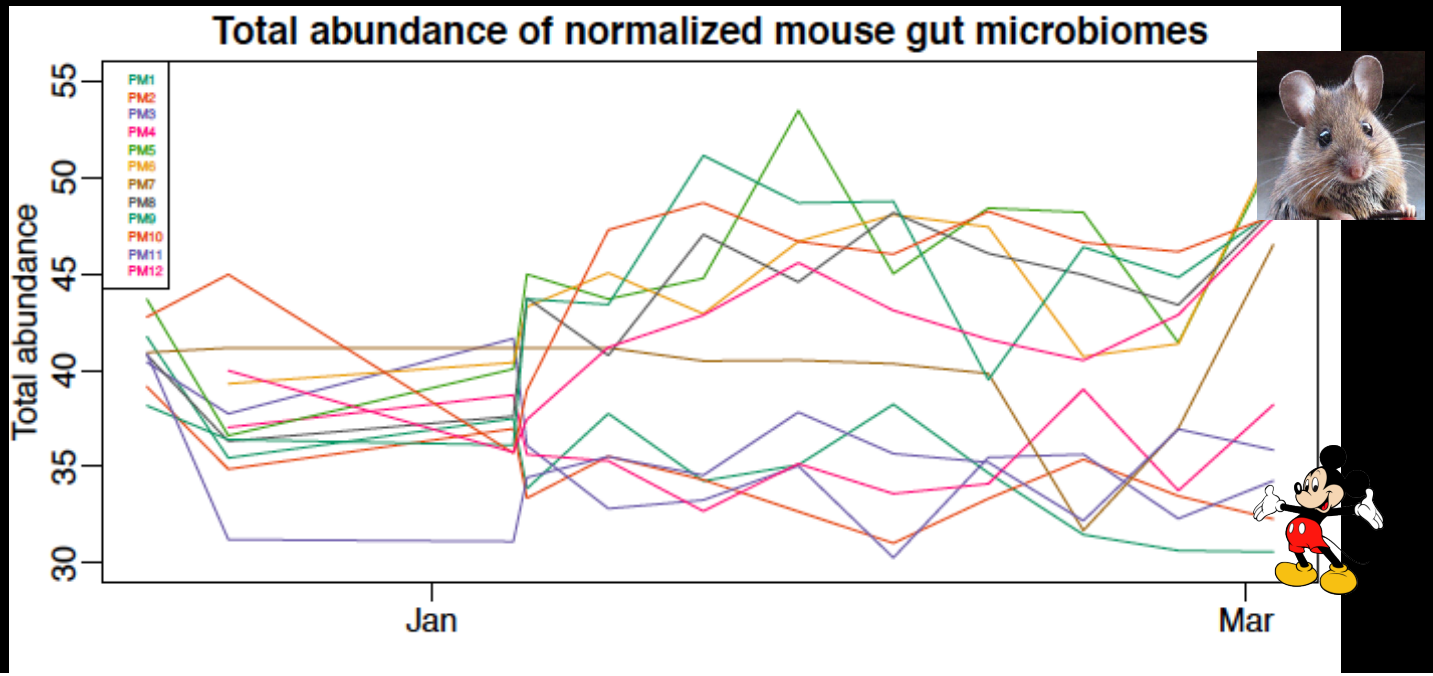
Normalization

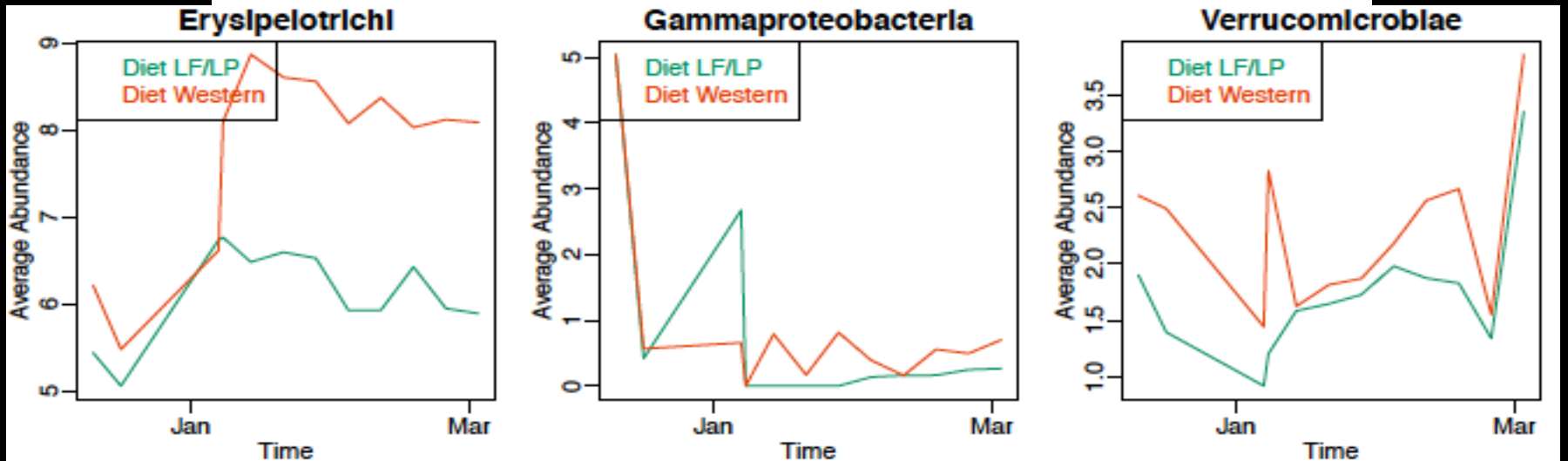


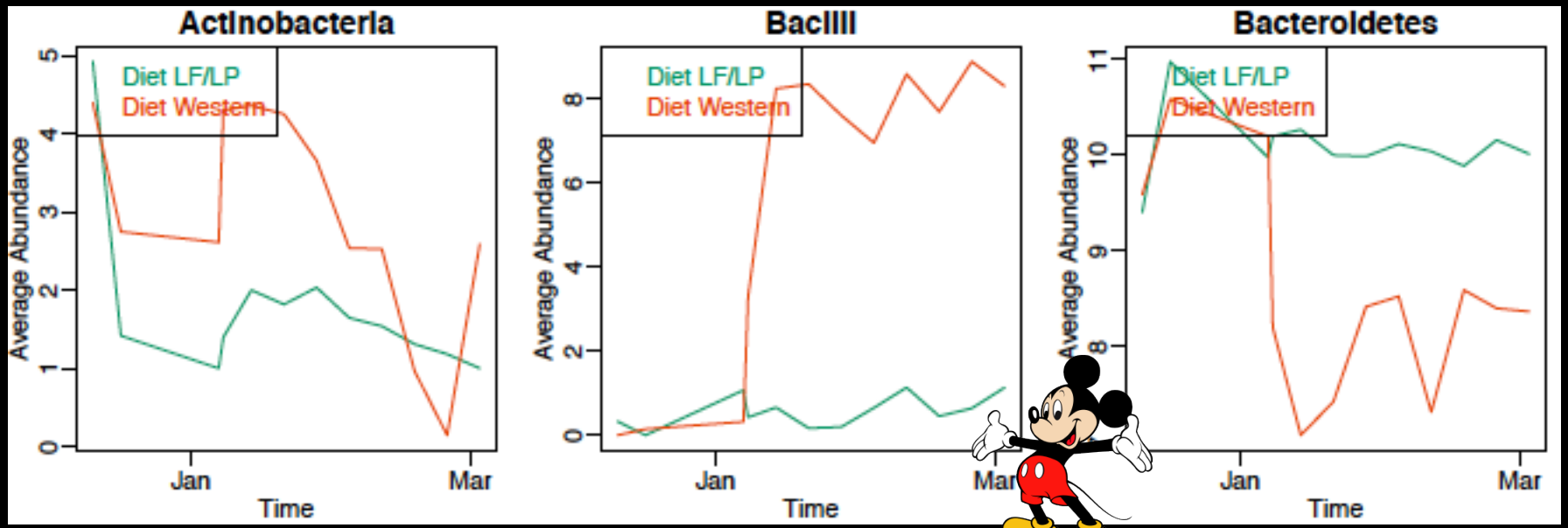
Comparison of methods

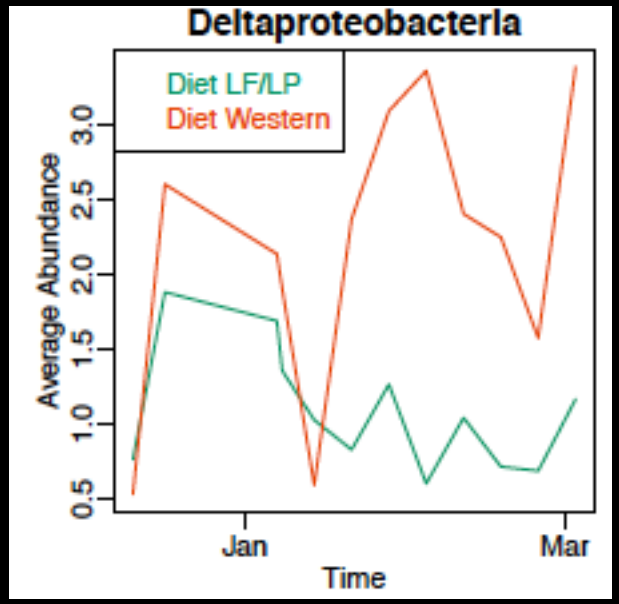


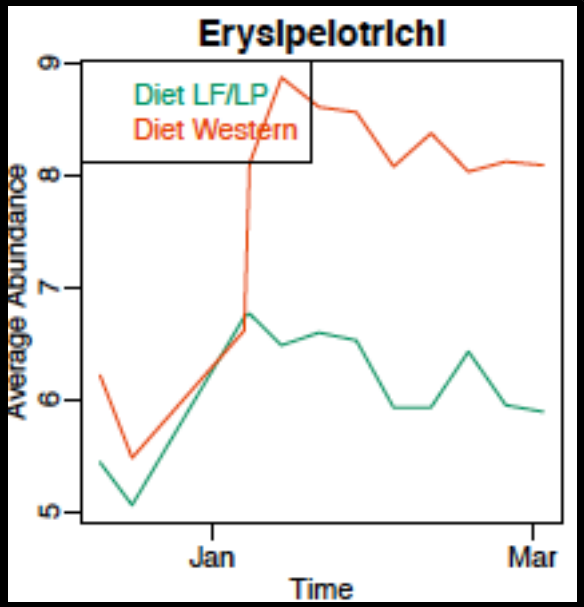
Future work


Total abundance of normalized mouse gut microbiomes
Total abundance
30
35
40
45
50
55
PM1
PM2
PM3
PM4
PM5
PM6
PM7
PM8
PM9
PM10
PM11
PM12
Jan
Mar


Erysipelotrichi
Gammaproteobacteria
Verrucomicrobiae
Diet LF/LP
Diet Western
Average Abundance
Time
Jan
Mar



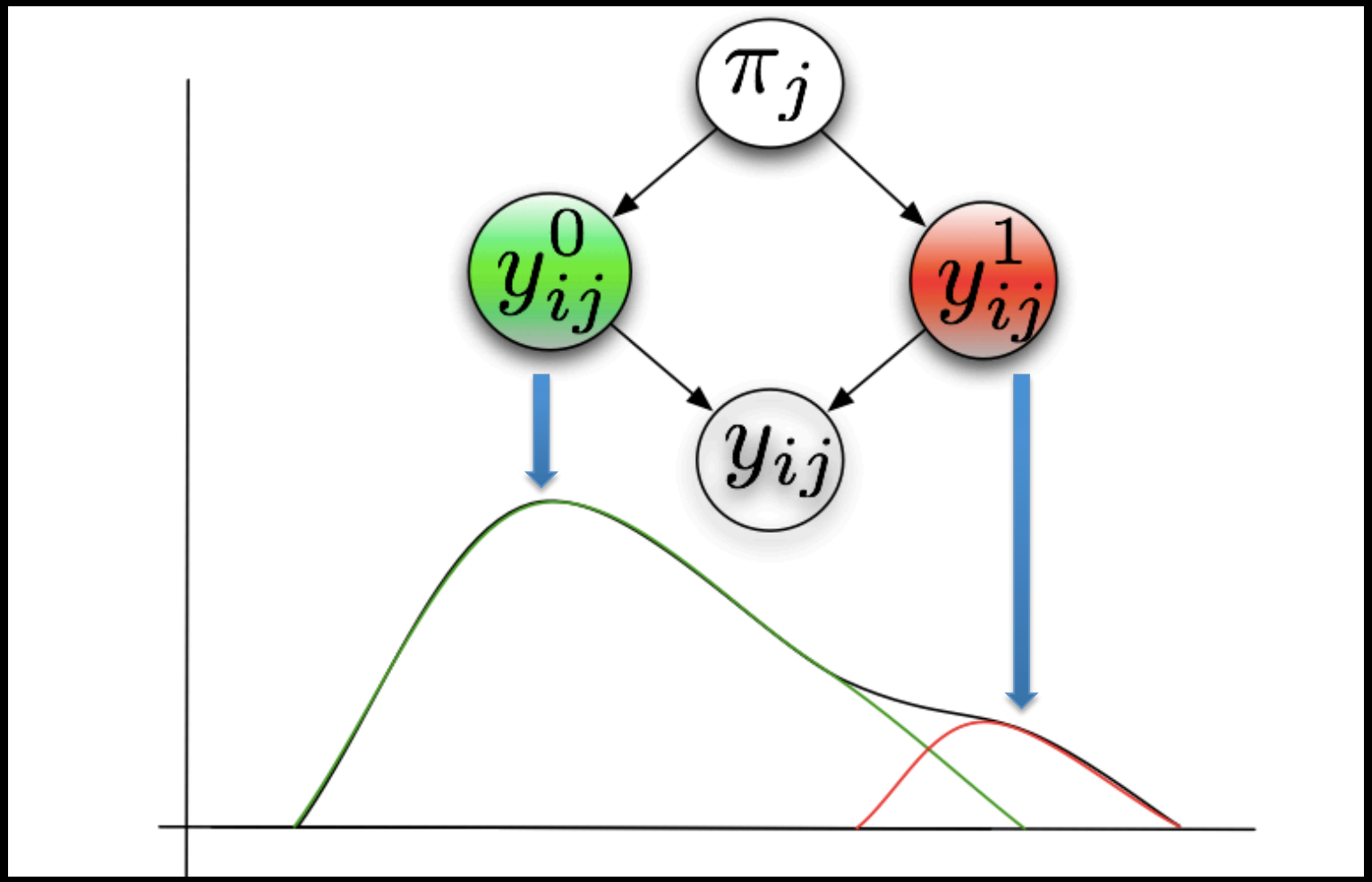


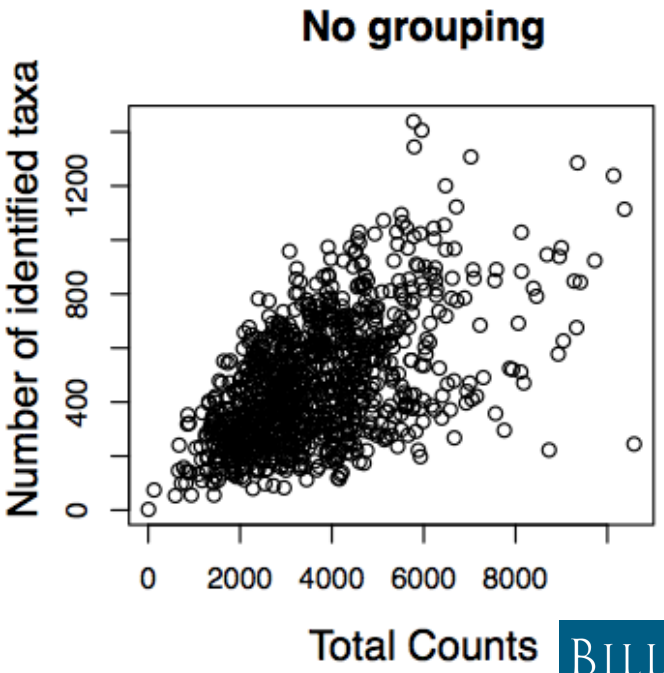


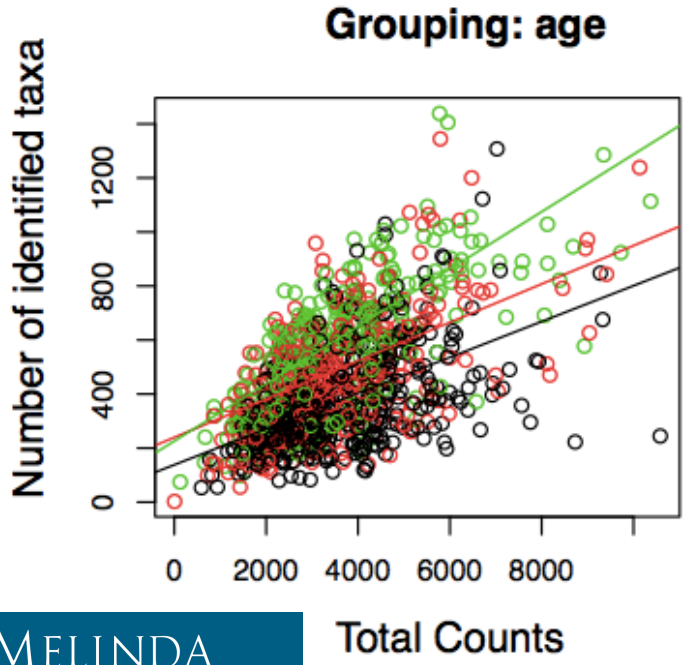
Bacteria	T-Test
Actino.	.723901
Bacilli*	8.5228
Bactero.*	-5.061
Deltapro.	1.6532
Eryspel.*	5.1369

Model to account for zeros

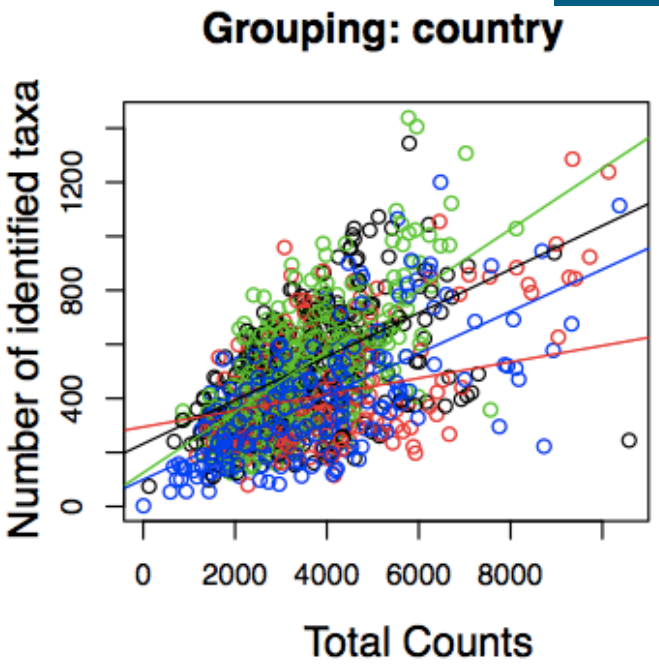
$$f_{total}(y_{ij};\theta) = \pi \cdot f_0(y_{ij}) + (1-\pi) \cdot f_1(y_{ij})$$







BILL & MELINDA GATES foundation



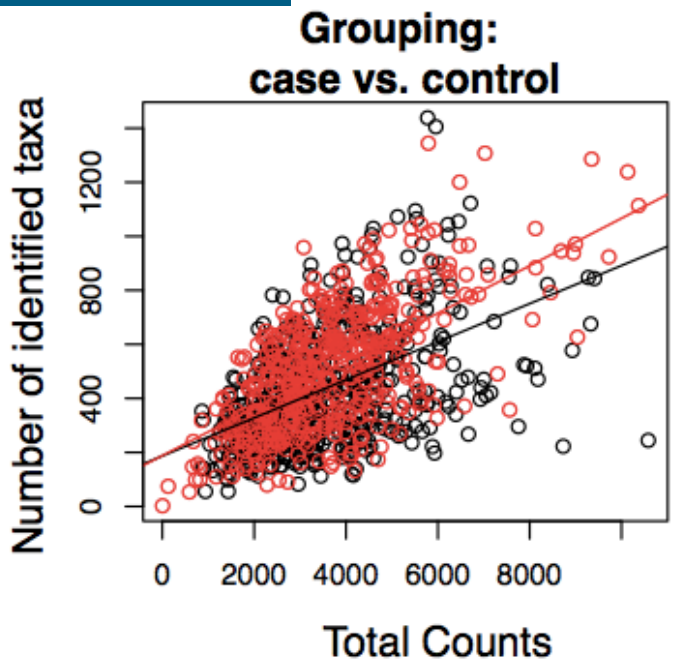


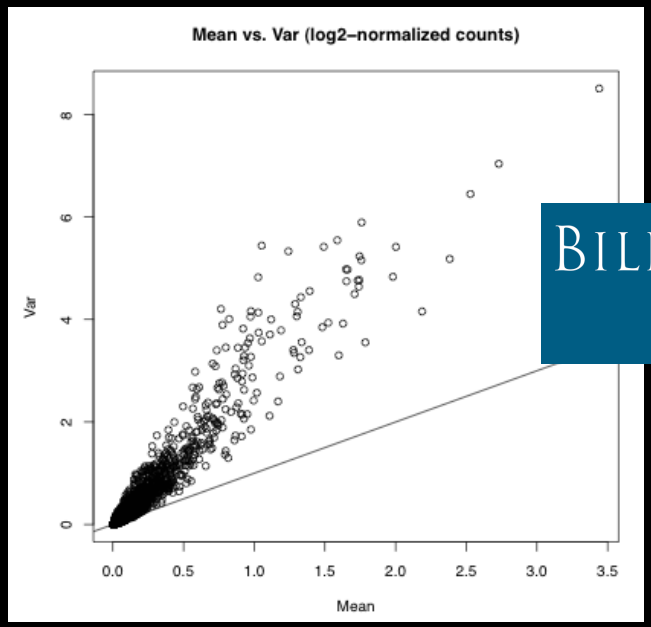
FIG B:
BLACK = AGE 0
RED = AGE 1
GREEN = AGE 2

FIG C:
BLACK = COUNTRY 1
RED = COUNTRY 2
GREEN = COUNTRY 4
BLUE = COUNTRY 6

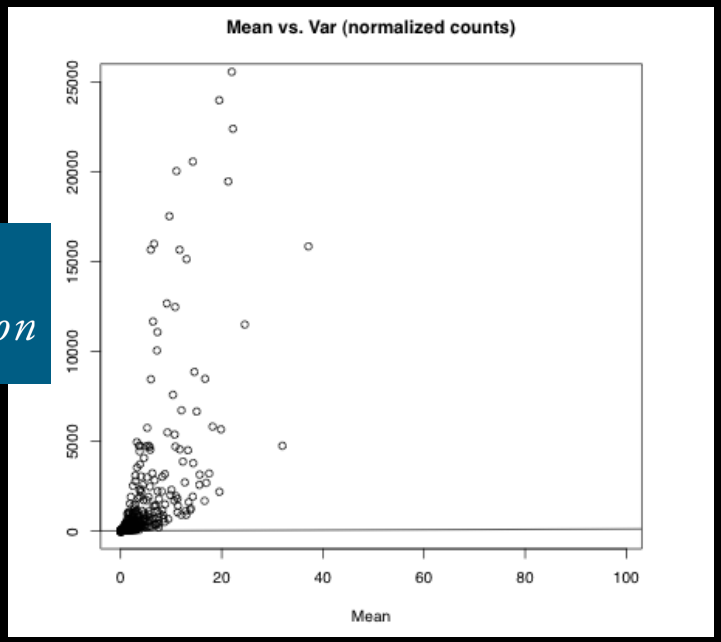
FIG D:
BLACK = CASE
RED = CONTROL

Approach: Zero-inflated Gaussian

- Very sparse data – likely technical error
- In the log space we can control for variance



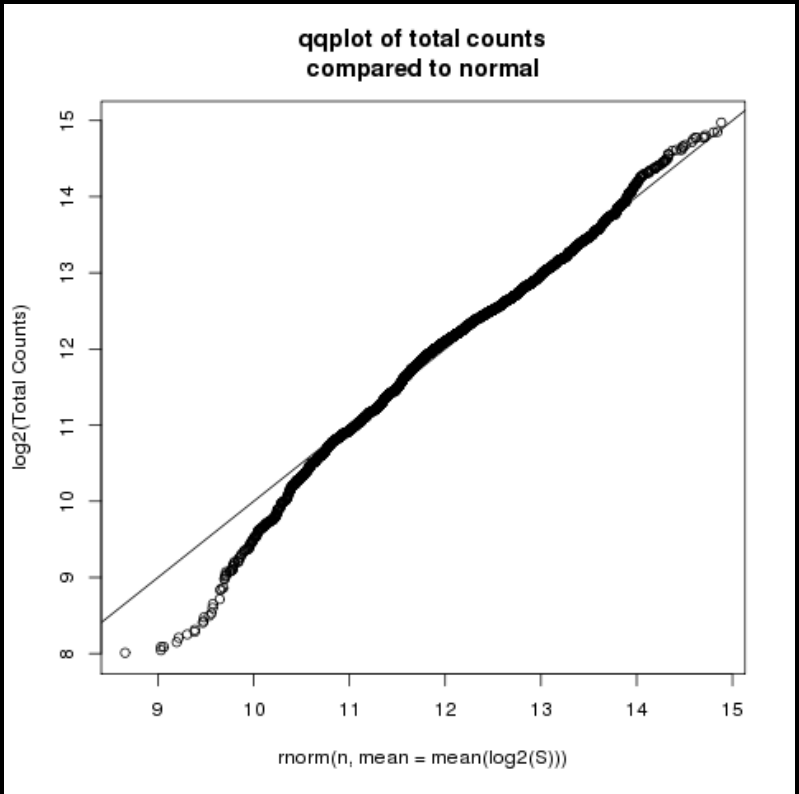
BILL & MELINDA GATES foundation



Approach: Zero-inflated Gaussian

- Total abundance follows a log-normal distribution

http://www.hmpdacc.org/HMQCP/



Approach: Zero-inflated Gaussian

- Counts are log transformed as: $y_{ij} = log_2(c_{ij} + 1)$
- Mixture of point mass, $f_{\{0\}}$, at zero and a count distribution $f_{count}(y; \mu, \sigma^2) \sim N(\mu, \sigma^2)$
- Mixture parameter π_j
- Values $\theta = \{S_j, \beta_0, \beta_1, \mu_i, \sigma_i^2\}$
- Density is:

$$f_{zig}(y_{ij}; \theta) = \pi_j(S_j) \cdot f_{\{0\}}(y_{ij}) +$$
$$(1 - \pi_j(S_j)) \cdot f_{count}(y_{ij}; \mu_i, \sigma_i^2)$$

Zero-inflated Gaussian

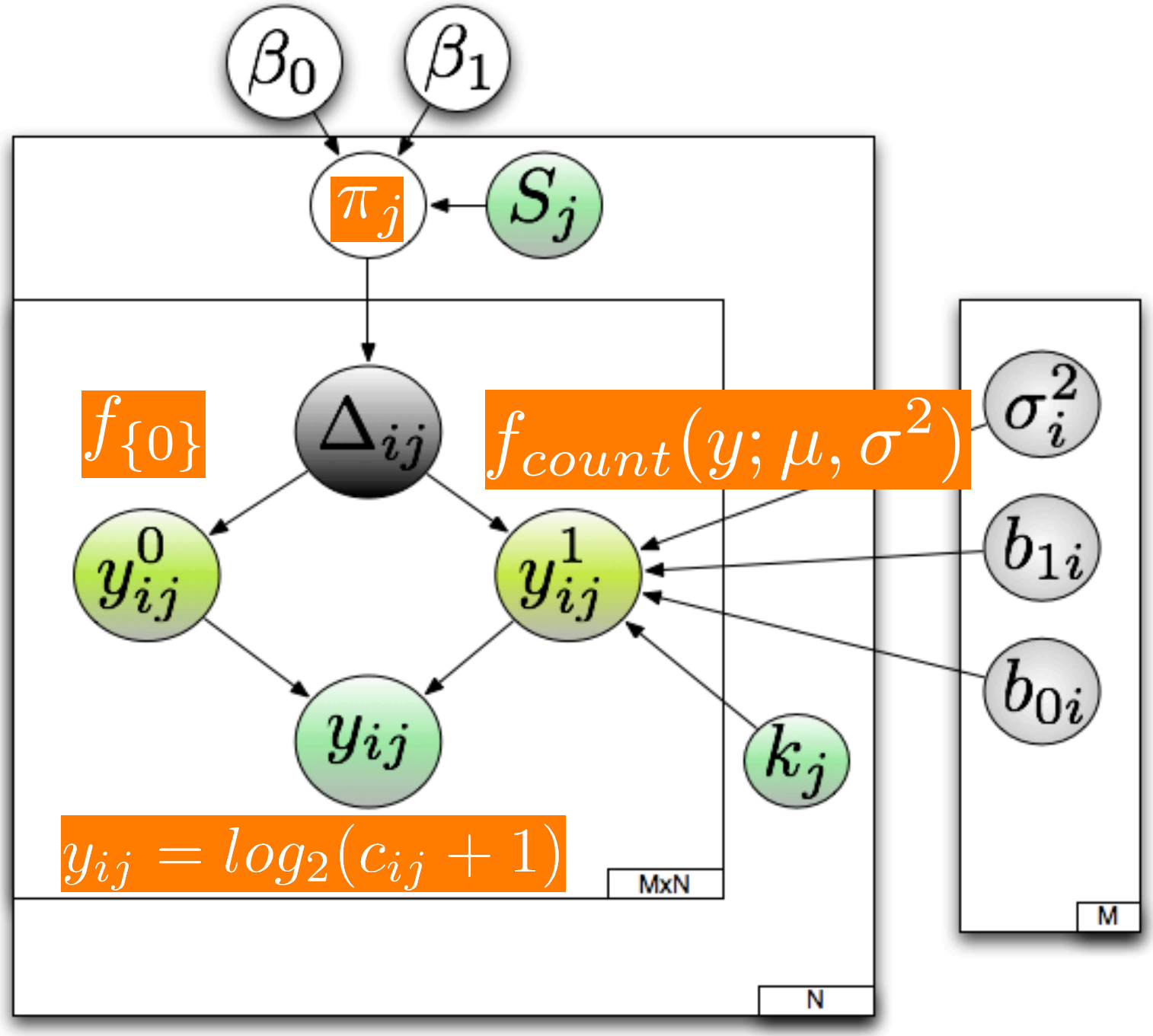
- And a mean specified as:

$$E(y_{ij}|k(j)) = \pi_j \cdot 0 + (1 - \pi_j) \cdot (b_{i0} + b_{i1} \cdot k(j))$$

Or

$$E(y_{ij}|k(j)) = \pi_j \cdot 0 + (1 - \pi_j) \cdot (b_{i0} + b_{i1} \cdot k(j) + \eta_i log_2(s95_j))$$

- Where k_j is our class label


β_0
β_1
π_j
S_j
$f_{\{0\}}$
Δ_{ij}
$f_{count}(y; \mu, \sigma^2)$
σ_i^2
y_{ij}^0
y_{ij}^1
b_{1i}
b_{0i}
y_{ij}
k_j
$y_{ij} = log_2(c_{ij} + 1)$
MxN
M
N

Algorithm:

1. Preprocess Data
2. Take initial guesses for the expected value of the latent indicator variables.
   - ij positions with counts > 0, the value is 0, else .5

For i in 1…..M:

3. Expectation
4. Maximize
5. Calculate negative log-likelihoods for each feature

Repeat

7. Generate moderated t-statistic using an empirical bayes method

Validation

- Procedure
  - We simulated data using chosen parameters from our model.
  - We spuriously induced zeros in the data.
  - We then checked the posterior probabilities z_ij to see where they converged.
- Success!!
  - The spuriously induced zeros' z_ijs converged to 1.
  - The non-zero's z_ijs remained 0.

$$z_{ij} = Pr(\Delta_{ij} = 1 | \hat{\theta}, y_{ij}) = E(\Delta_{ij} | \hat{\theta}, y_{ij})$$

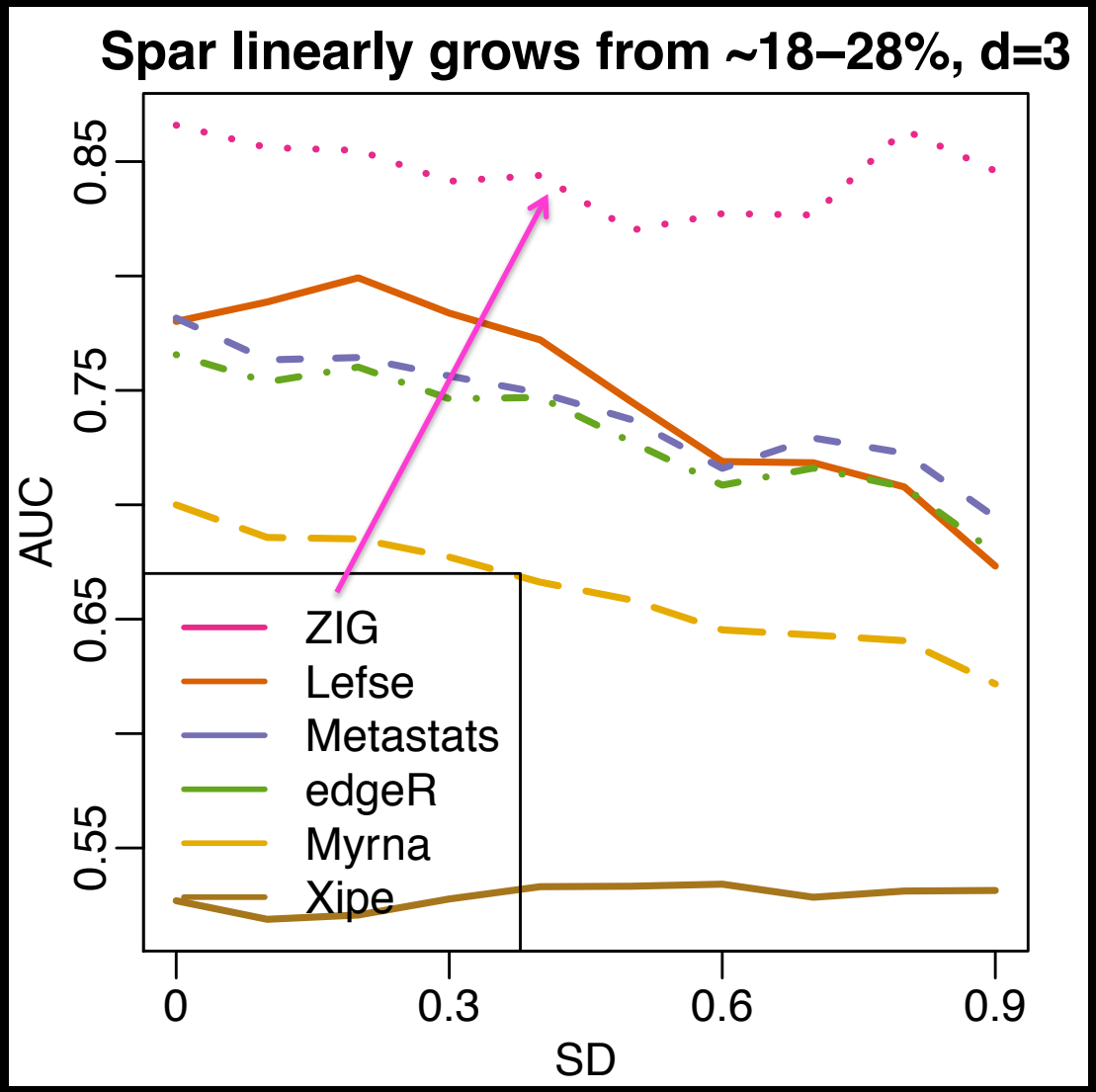
Validation


Prob. of technical zero
0.0
0.4
0.8
3200
3400
3600
3800
4000
4200
4400
Number of reads
SUCCESS!

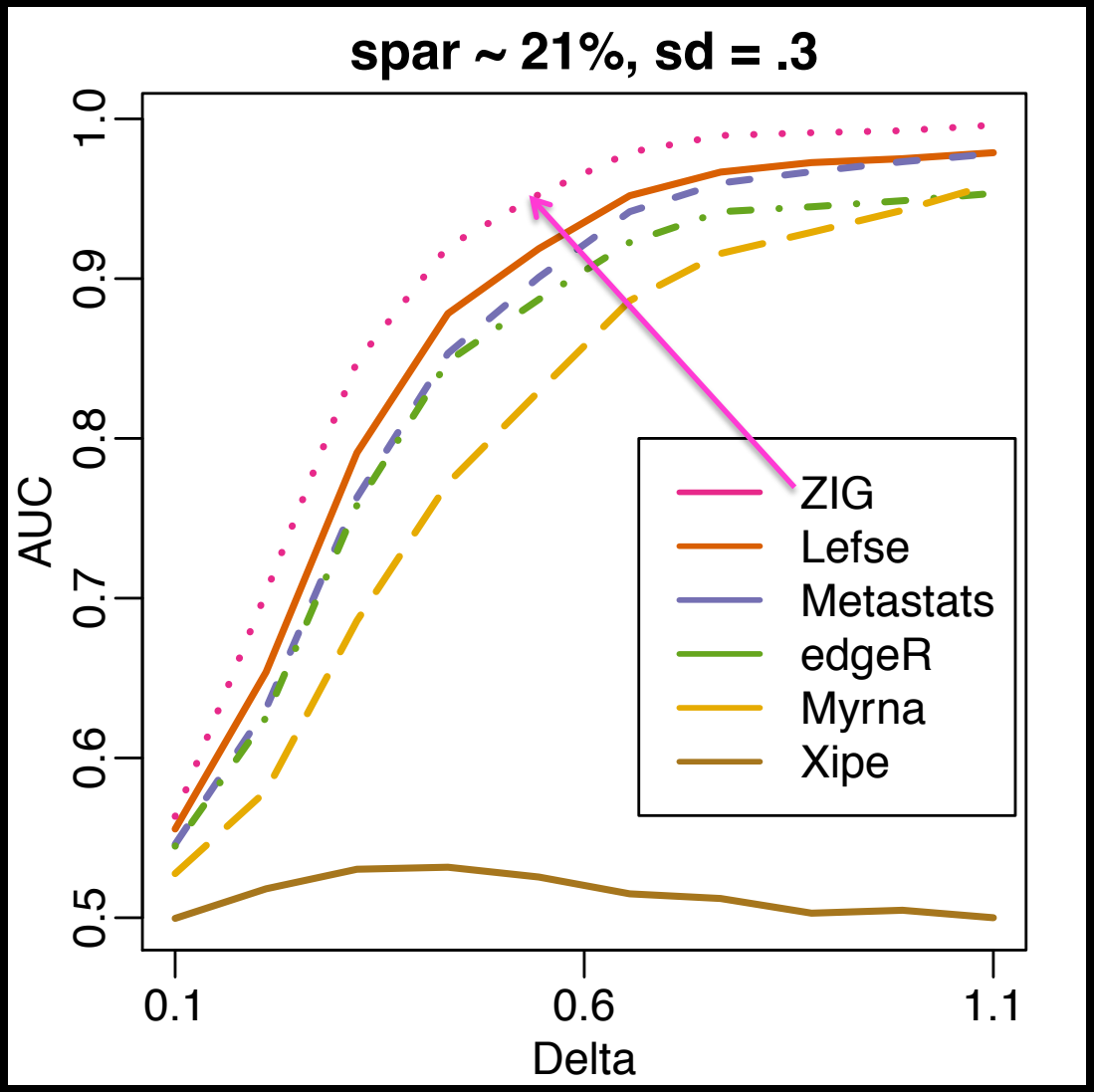
Testing

- Goal – Simulate real datasets
  - Same as validation, but we add random noise from U{-1,1}, change sparsity levels, and an additional 5% of the data was randomly extra counts.

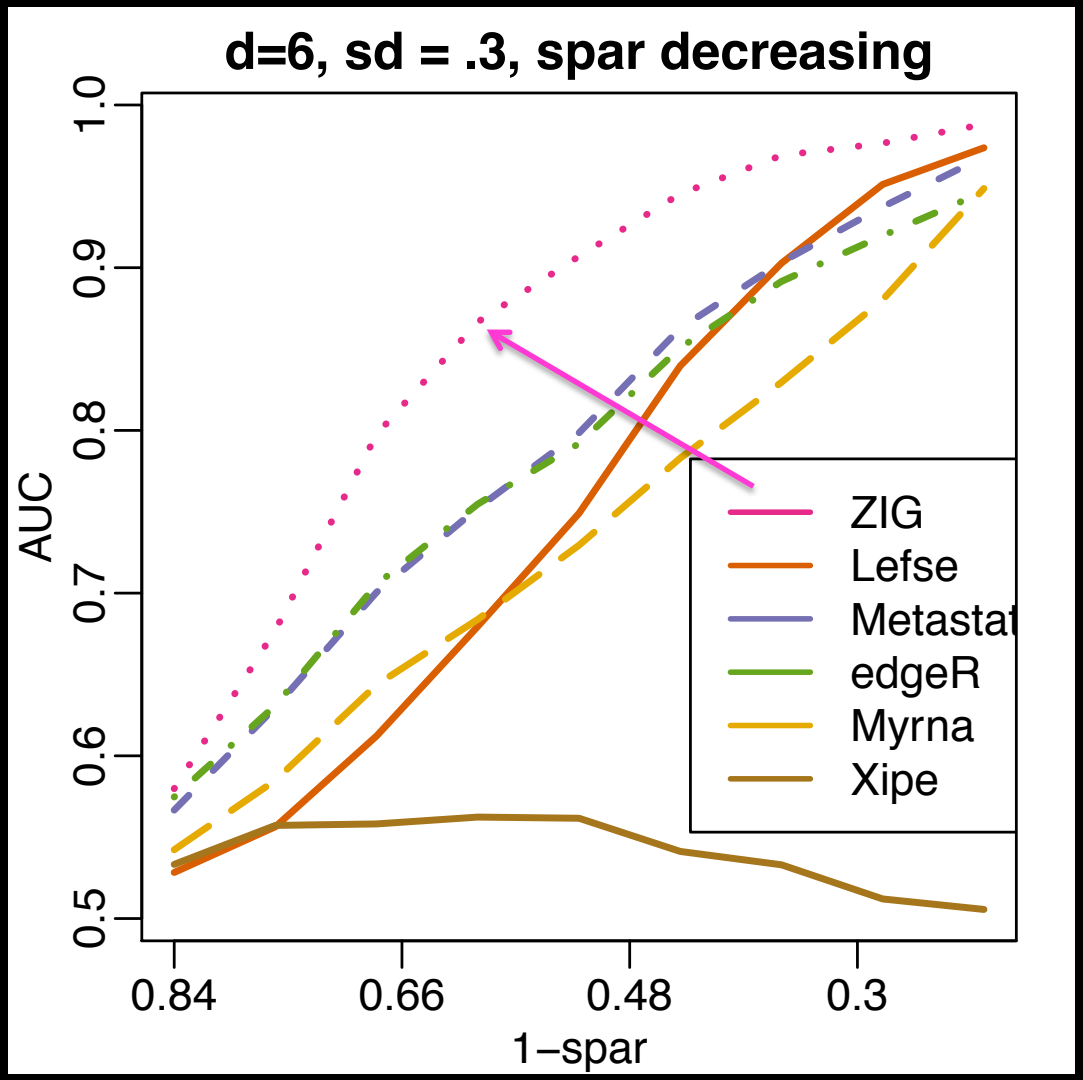
Testing


Spar linearly grows from ~18–28%, d=3
AUC
0.55
0.65
0.75
0.85
SD
0
0.3
0.6
0.9
ZIG
Lefse
Metastats
edgeR
Myrna
Xipe

Testing

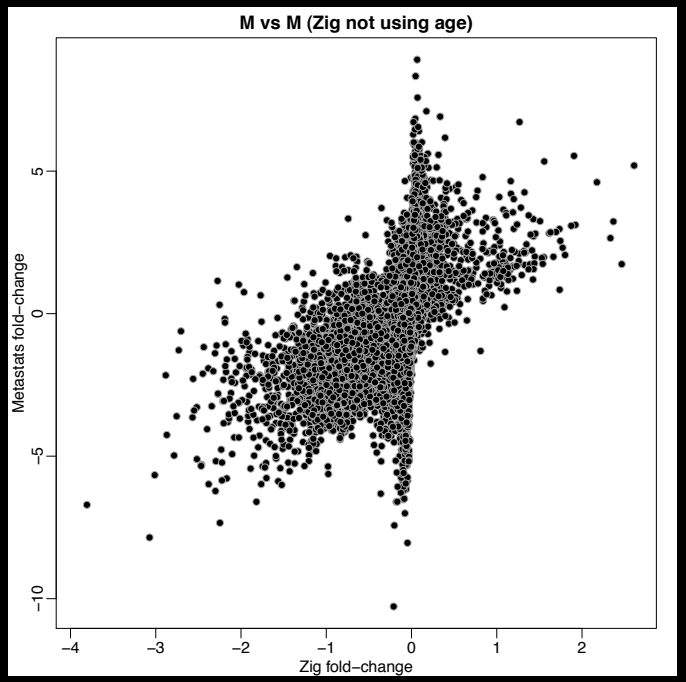

spar ~ 21%, sd = .3
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.1
0.6
1.1
Delta
ZIG
Lefse
Metastats
edgeR
Myrna
Xipe

Testing


d=6, sd = .3, spar decreasing
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.84
0.66
0.48
0.3
1−spar
ZIG
Lefse
Metastat
edgeR
Myrna
Xipe

Comparison with Metastats

$$E(y|k(j)=0)=\beta_0$$
$$E(y|k(j)=1)=\beta_0+\beta_1$$
$$\log_2(\frac{\mu_1}{\mu_0})=\log_2(\frac{2^{\beta_0+\beta_1}}{2^{\beta_0}})=\log_2(2^{\beta_1})=\beta_1$$



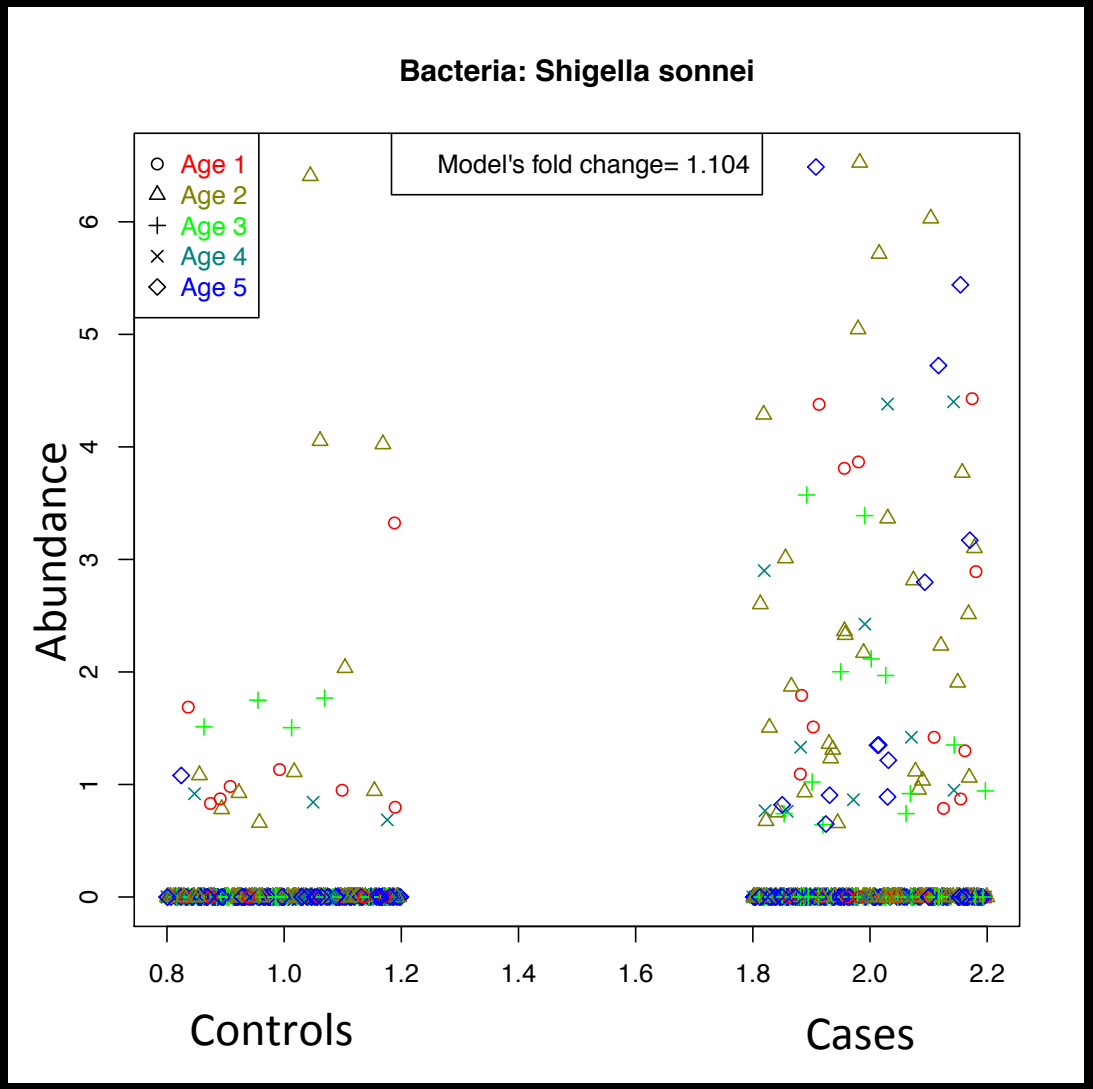
Comparison of log2 fold-change estimates between original Metastats and the zero-inflated model. Fold-change estimates are consistent between the two methods. However, the original Metastats method estimated large fold-changes for OTUs with small overall abundance that were driven by a small number of non-zero counts. The zero-inflated model is able to control estimates for these OTUs, thereby reducing false discoveries.

Gates results

BILL & MELINDA GATES foundation

Taxa
nomatch
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Prevotella;Prevotella sp. DJF_B116
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium;Clostridium difficile
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides sp. CJ78
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis;Bacteroides fragilis YCH46
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis;Bacteroides fragilis YCH46
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;Bacteroides fragilis
Bacteroidetes/Chlorobi group;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Prevotella;Prevotella sp. BI-42
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Granulicatella;environmental samples;Granulicatella sp. oral clone ASCG05
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Proteobacteria;delta/epsilon subdivisions;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Campylobacter;Campylobacter jejuni;Campylobacter jejuni subsp. jejuni
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;environmental samples;Streptococcus sp. oral clone ASCC04
Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus sp. C101
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Granulicatella;Granulicatella adiacens
Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;environmental samples;Streptococcus sp. oral clone ASCE09
Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus mitis
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia;Escherichia coli
Proteobacteria;Gammaproteobacteria;Pasteurellales;Pasteurellaceae;Haemophilus;Haemophilus haemolyticus
Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Shigella;Shigella sonnei
nomatch
Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;Lactobacillus fermentum
Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus pasteurianus

Gates results



BILL & MELINDA GATES foundation

Conclusion

- Developed a ton of useful tools for biologists to analyze their data in an R framework!
- Time to go and analyze some cool things with these better fold change estimates!

- Future directions could possibly include:
  - Develop a method to calculate p-values in a bootstrapped fashion that is computationally feasible.

Project Schedule

- November 30:
  - Preprocessing data
  - Finish normalization codes
- December 15:
  - Continue reading
  - Finish Zig model
  - Midyear report
- January 15:
  - Continue reading
  - Validation of methods
- February 15:
  - Finish a comparison of normalization methods
  - Package, comment, etc.
- March 15:
  - Analyze various datasets
- May 15:
  - Deliver all
  - Final report

Deliverables

- Code for cumulative sum normalization
- Code for cumulative scaling normalization
- Code for the Expectation Maximization algorithm
- Code for calling the E-M algorithm
- Code to load data
- Code for simulation study
- Final presentation
- Final report
- Mouse diet data
  - Script to format mouse data and calculate p-values/FDR

Bibliography

- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. **The Elements of Statistical Learning.** Dordrecht: Springer, 2009. Print.

- McCulloch, Charles E., S. R. Searle, and John M. Neuhaus. **Generalized, Linear, and Mixed Models.** Hoboken, NJ: Wiley, 2008. Print.

- White, James Robert, Niranjan Nagarajan, and Mihai Pop. **"Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples."** Ed. Christos A. Ouzounis. PLoS Computational Biology 5.4 (2009): E1000352. Print.

- Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.

- Efron B, Tibshirani R (1993) An introduction to the bootstrap. New York: Chapman & Hall.

- Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.

Expectation-Maximization

E-step:
Estimates responsibilities,

$$z_{ij} = Pr(\Delta_{ij} = 1 | \hat{\theta}, y_{ij}) = E(\Delta_{ij} | \hat{\theta}, y_{ij})$$

as:

$$\hat{z}_{ij} = \frac{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij})}{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij}) + (1 - \hat{\pi}_j) \cdot f_{count}(y_{ij}; \hat{\theta}_{ij})}$$

Expectation-*Maximization*

M-step:
Estimate parameters $\hat{\theta}_{ij} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{b}_{0i}, \hat{b}_{1i}\}$
given current estimates of $\hat{z}_{ij}$.

Current mixture parameters are estimated as:

$$\hat{\pi}_j = \sum_{i=1}^{M} \frac{1}{M} \hat{z}_{ij}$$

Parameters for the count distribution are estimated using weighted least squares where the weights are $\hat{z}_{ij}$.

Mixture parameters

Zero-valued features depend on a sample's total number of counts, S_j
They follow a binomial distribution.

We model the linear effect with our mixture parameter π_j

via linear regression with a transformation function:

$$log \frac{\pi_j}{1 - \pi_j} = \beta_0 + \beta_1 \cdot log(S_j)$$

Log-likelihood

We can get the maximum-likelihood estimates using the Expectation-Maximization algorithm, where we treat mixture membership Δ_{ij} = 1 if y_ij comes from the zero point mass as a latent indicator variable.

Denote the full set of estimates as $\theta_{ij} = \{\beta_0, \beta_1, b_{0i}, b_{1i}\}$

$$l(\theta_{ij}; y_{ij}, S_j) = (1 - \Delta_{ij}) \log f_{count}(y; \mu_i, \sigma_i^2) + \Delta_{ij} \log \pi_j(S_j) + (1 - \Delta_{ij}) \log(1 - \pi_j(S_j))$$

Algorithm continued

$$t_i^{ob} = \frac{b_{1i}}{(\sigma_i^2/\Sigma(1 - z_{ij}))^{.5}}$$

- We use the Empirical Bayes method to construct a moderated t-statistic and use a parametric t-distribution.