Metastats 2.0

An improved method and software for analyzing metagenomic data

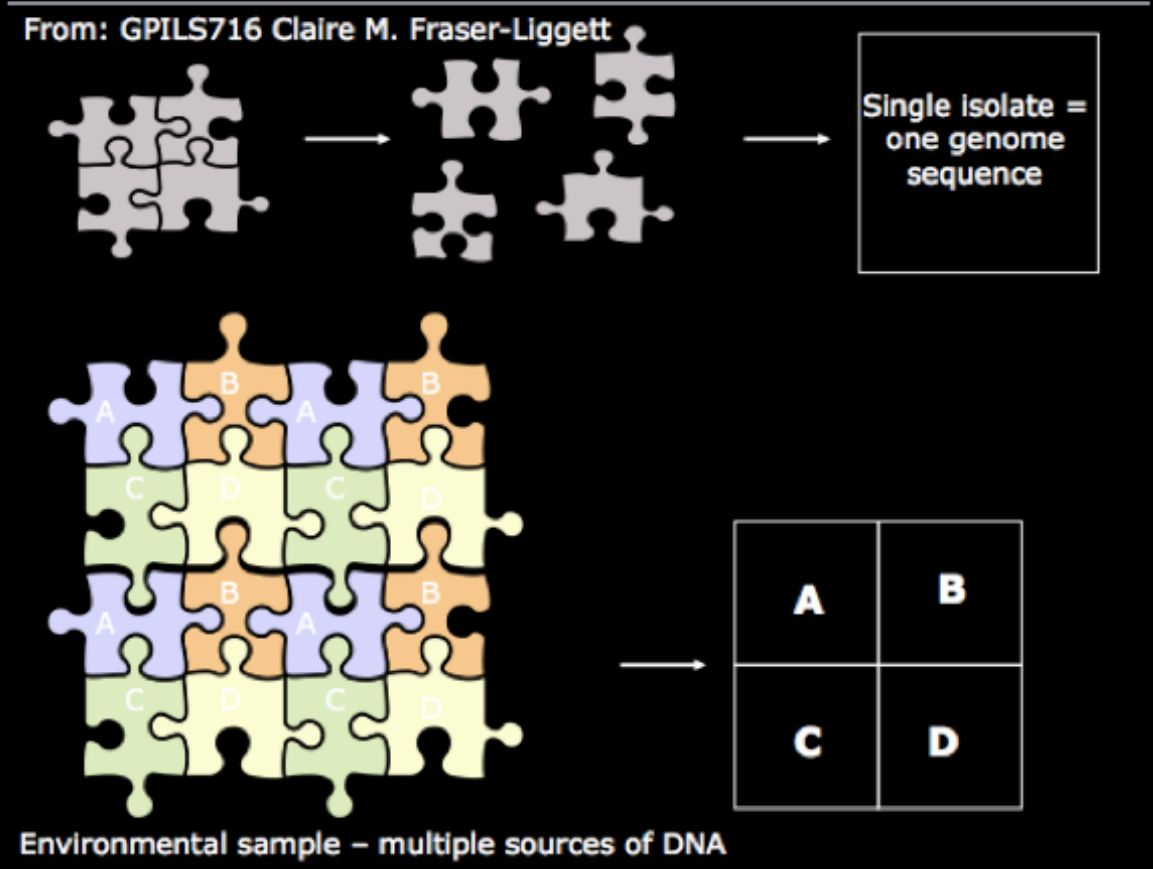
Joseph N. Paulson
jpaulson@umiacs.umd.edu
Mihai Pop
mpop@umiacs.umd.edu
Héctor Corrada Bravo
hcorrada@umiacs.umd.edu

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)	S*N*
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)				
T*M*	c(M,1)		. . . . .		c(M,N)

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]*

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

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

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

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

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!

Doesn't account for depth of coverage

Normalization induces spurious correlations









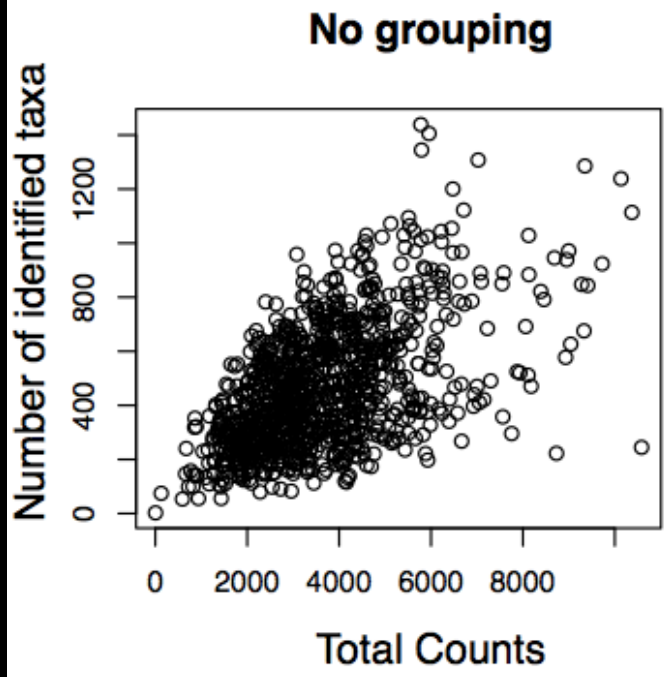


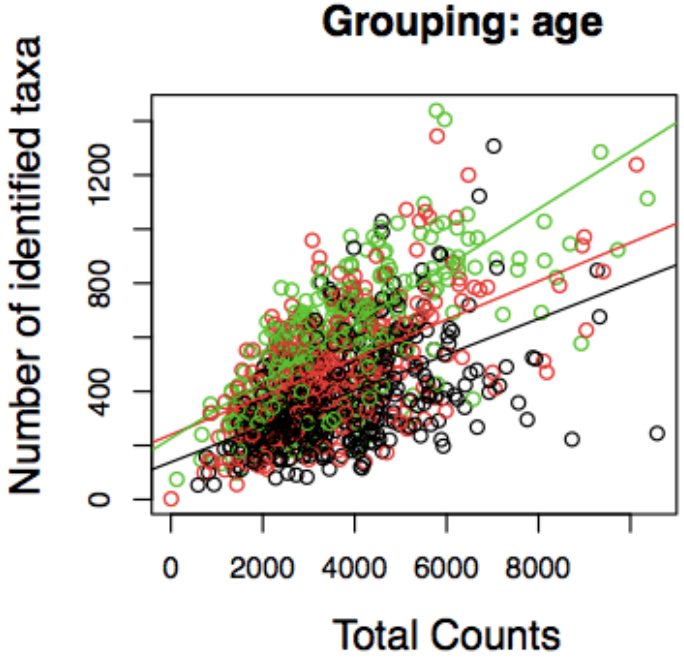


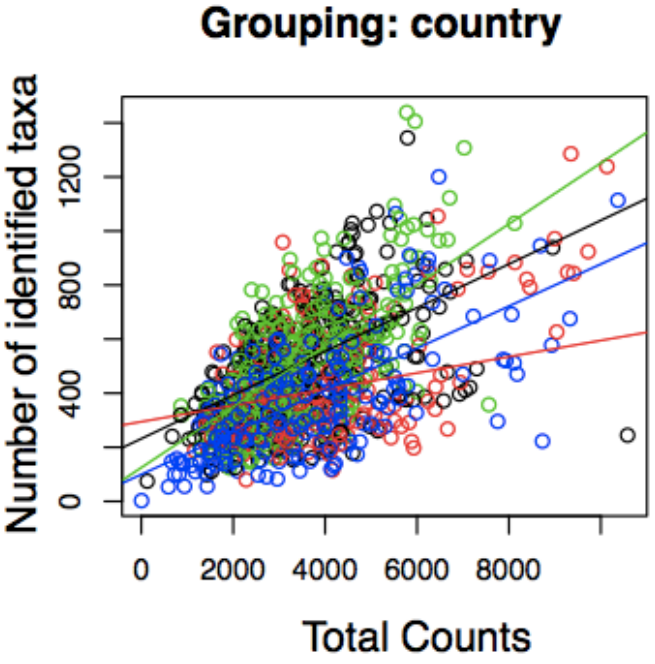


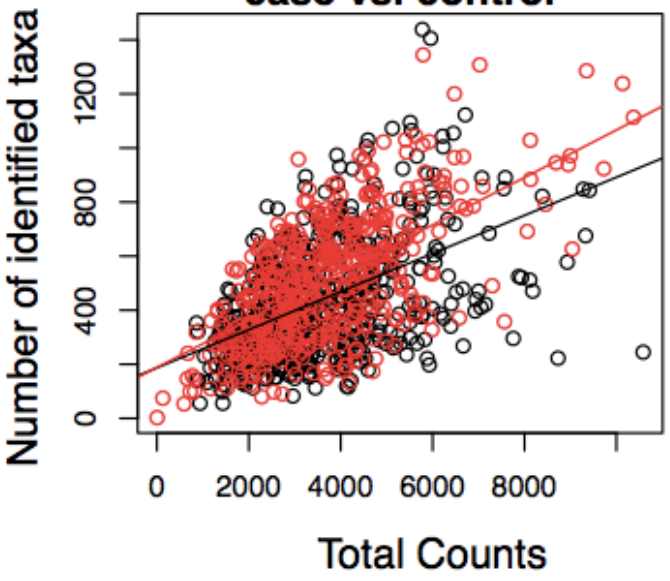




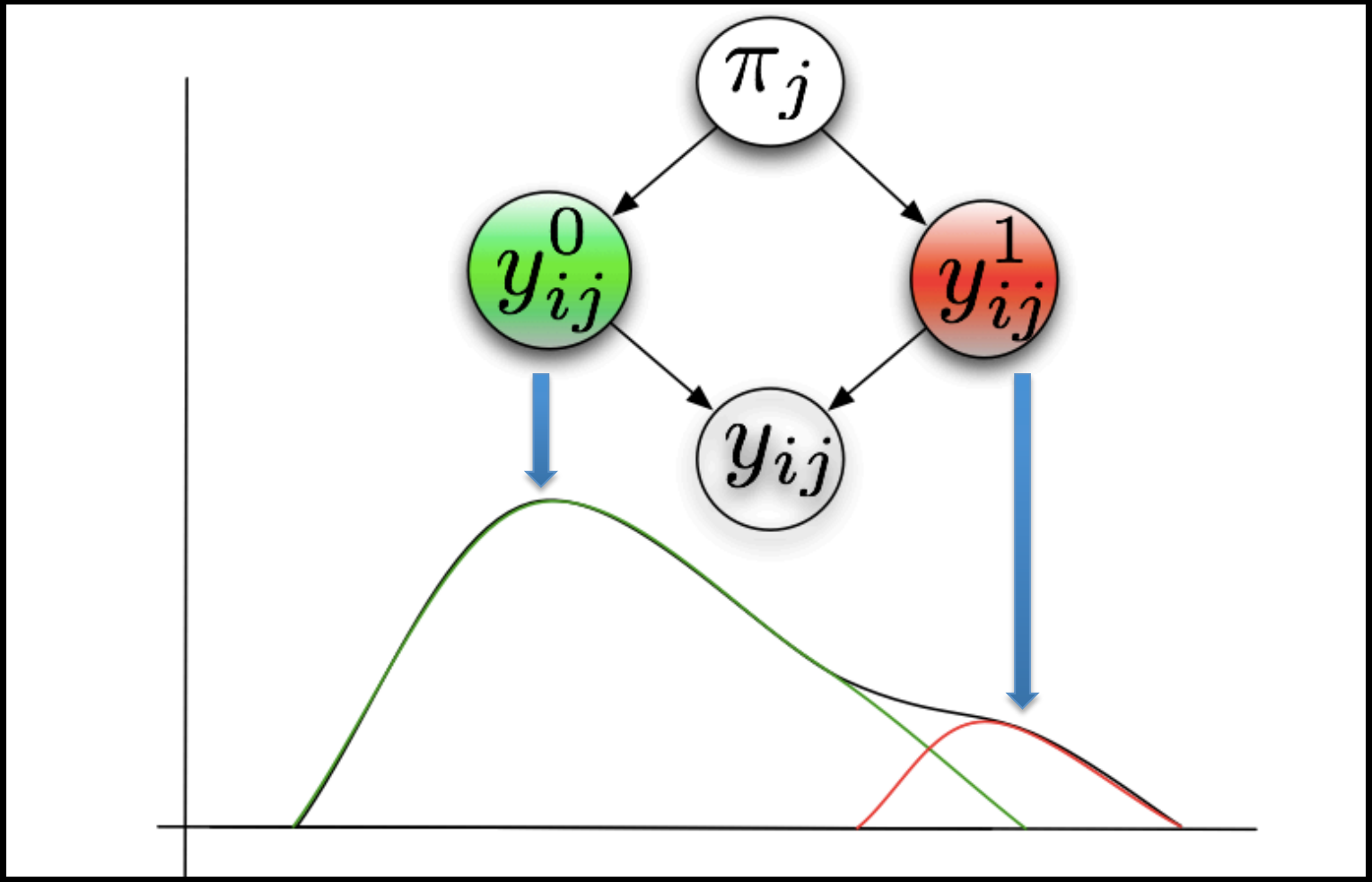

No grouping
Number of identified taxa
0
400
800
1200
0
2000
4000
6000
8000
Total Counts


Grouping: age
Number of identified taxa
0
400
800
1200
0
2000
4000
6000
8000
Total Counts


Grouping: country
Number of identified taxa
0
400
800
1200
0
2000
4000
6000
8000
Total Counts


Grouping:
case vs. control
Number of identified taxa
0
400
800
1200
0
2000
4000
6000
8000
Total Counts

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



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:

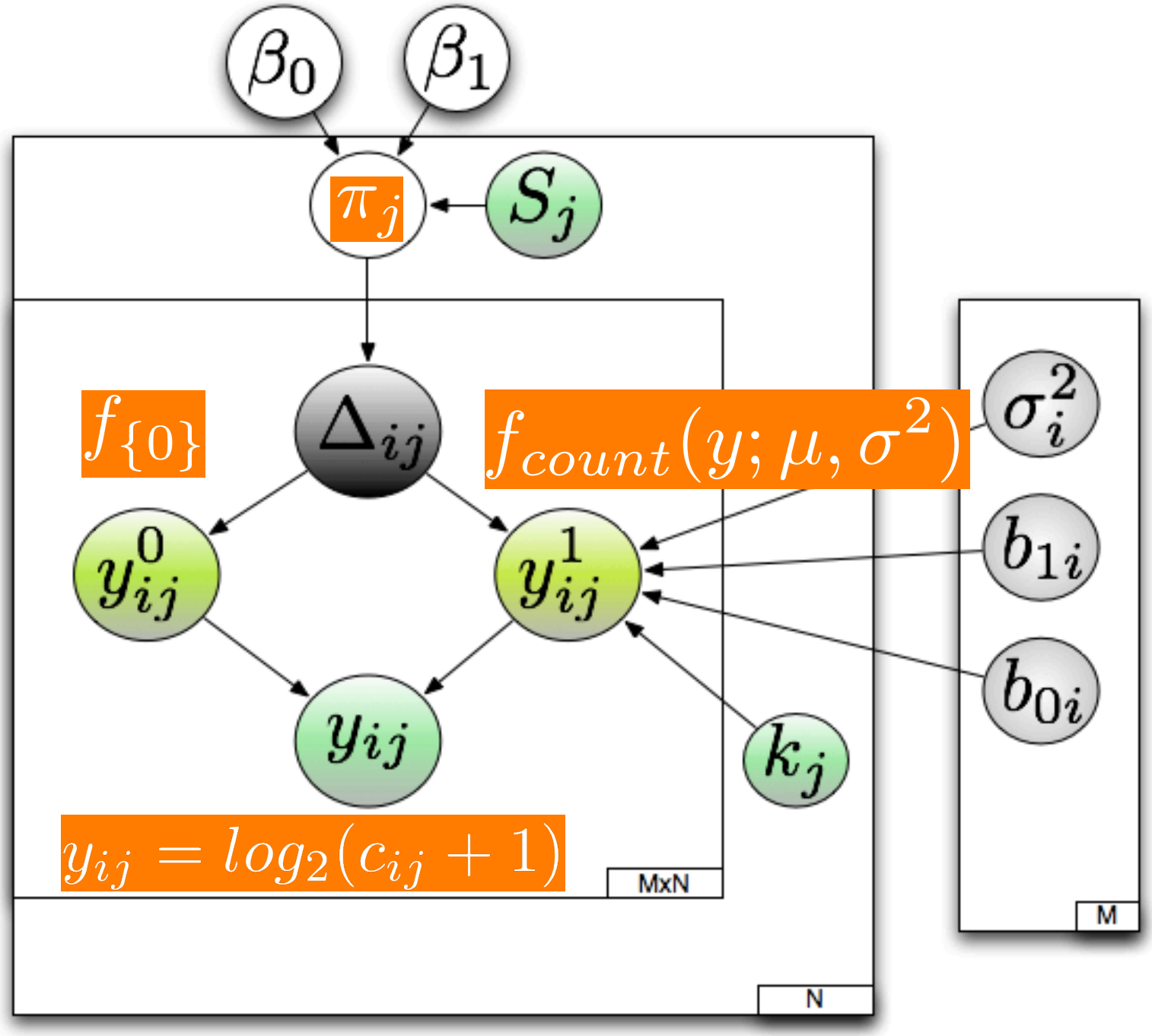
$$f_{zig}(y_{ij}; \theta) = \pi_j(S_j) \cdot f_{\{0\}}(y_{ij}) +$$
$$(1 - \pi_j(S_j)) \cdot f_{count}(y_{ij}; \mu_i, \sigma_i^2)$$

Zero-inflated Gaussian

- And a mean specified as:

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

- 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

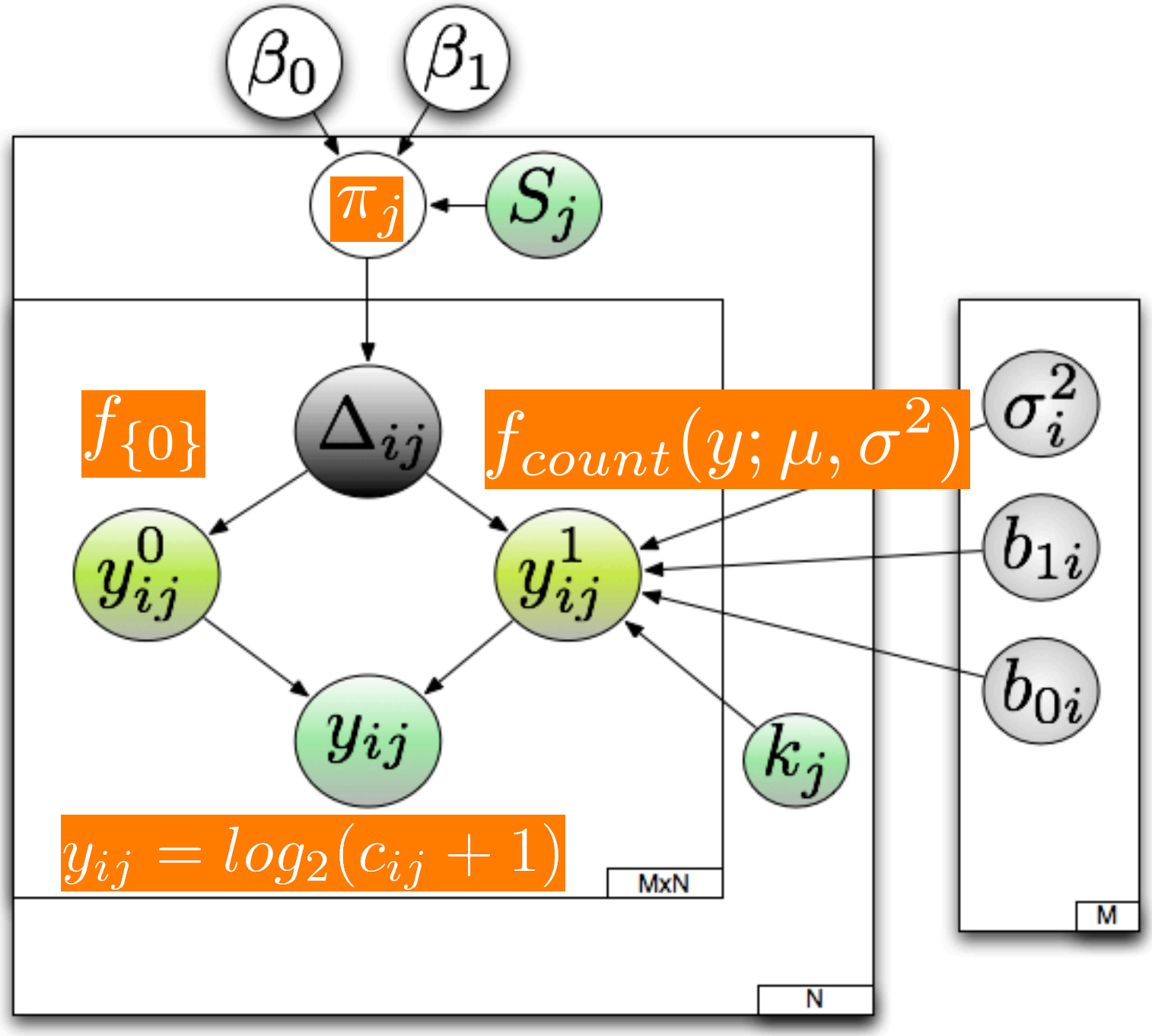
Mixture parameters

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:

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


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

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:

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. Permute class membership (labels)
8. Calculate new t-statistic, permute and calculate p-values

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 $\hat{z}_{ij}$.

Algorithm continued

- Permute the labels K_j b_{1i}
- Compute $t_i^{ob} = \frac{}{(\sigma_i^2/\Sigma(1-z_{ij}))^{.5}}$
- Divided by the newly weighted standard error.
- Calculate $p_i = \frac{\{|t_i^{ob}| \geq |t_i| b \in 1...B\}}{B}$

- Plan to add a few other tests.

Algorithm 2

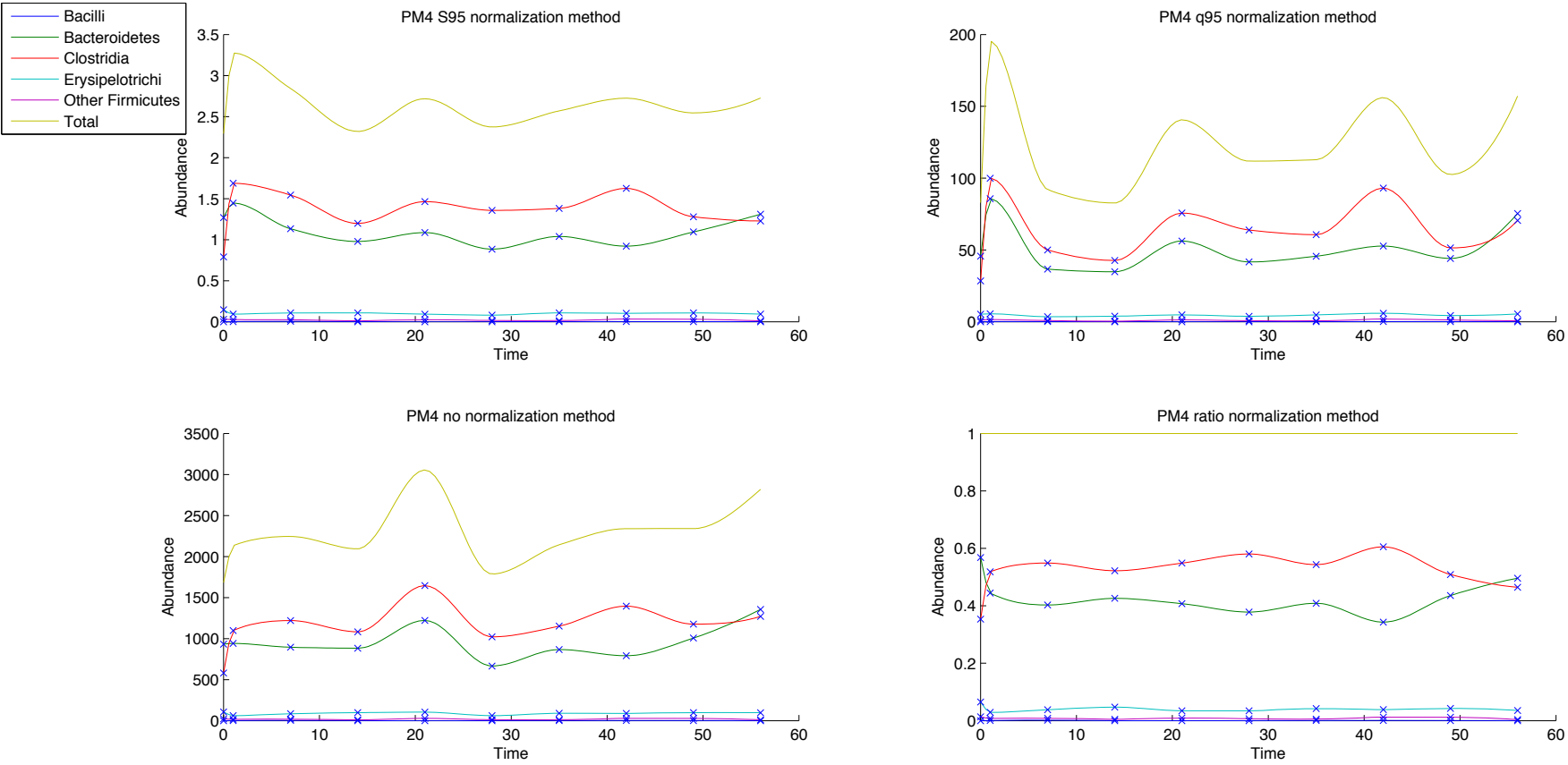
- Ratio Normalization:
  - What are the issues with it??

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

  - 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


Bacilli
Bacteroidetes
Clostridia
Erysipelotrichi
Other Firmicutes
Total
PM4 S95 normalization method
PM4 q95 normalization method
PM4 no normalization method
PM4 ratio normalization method
Abundance
Time

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:

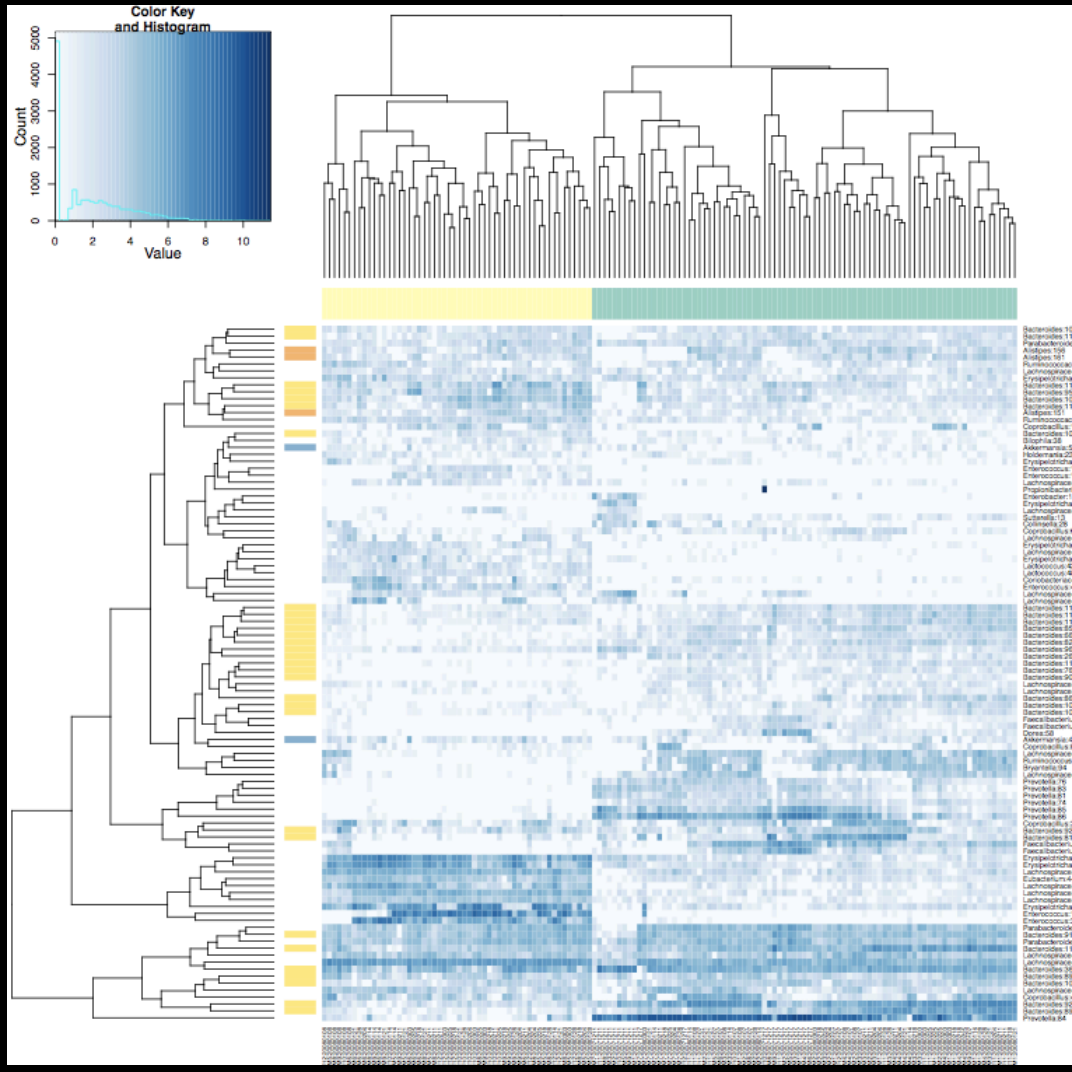
$$S_{95j} = \sum_i c_{ij} \leq q_{95j}$$

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: $y_{ij} = c_{ij} / q_{75j}$



Implementation

- Software:
  - R and possibly C
    - Make use of R and various R package functions
    - Make use of open MP (time permitting)
- Numerically, the bottleneck is the bootstrapping measure (fitting the weighted least squares).
  - Thankfully that step is trivially parallelizable.
- Hardware:
  - Develop on my Macbook Air
    - 1.6 core duo
    - 4 gigs of ram
  - Run on Ginkgo
    - 8 x Quad-core AMD Opteron™ Processor 8365 (2300MHz) (32 cores)
    - 256 GB Ram
    - RHEL5 x86_64

Databases

- Diseased and healthy dysentery data
- Oral microbiome
- Two diet groups of gnotobiotic mice

- Access others with more time from Genbank database.

Validation

- Compare non-zero matrix results with another method, the log model fit, to ensure exact same results.

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

- Simulate data for known quantities (known difference, small variance) and see how model reacts.

Testing

- Ensure that preprocessing of the data is handled correctly – biologically
- Compare to Metastats, Kruskal-Wallis (non-parametric test), etc.

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
- April 15:
  - Parallelize
- May 15:
  - Deliver all
  - Final report

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.