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

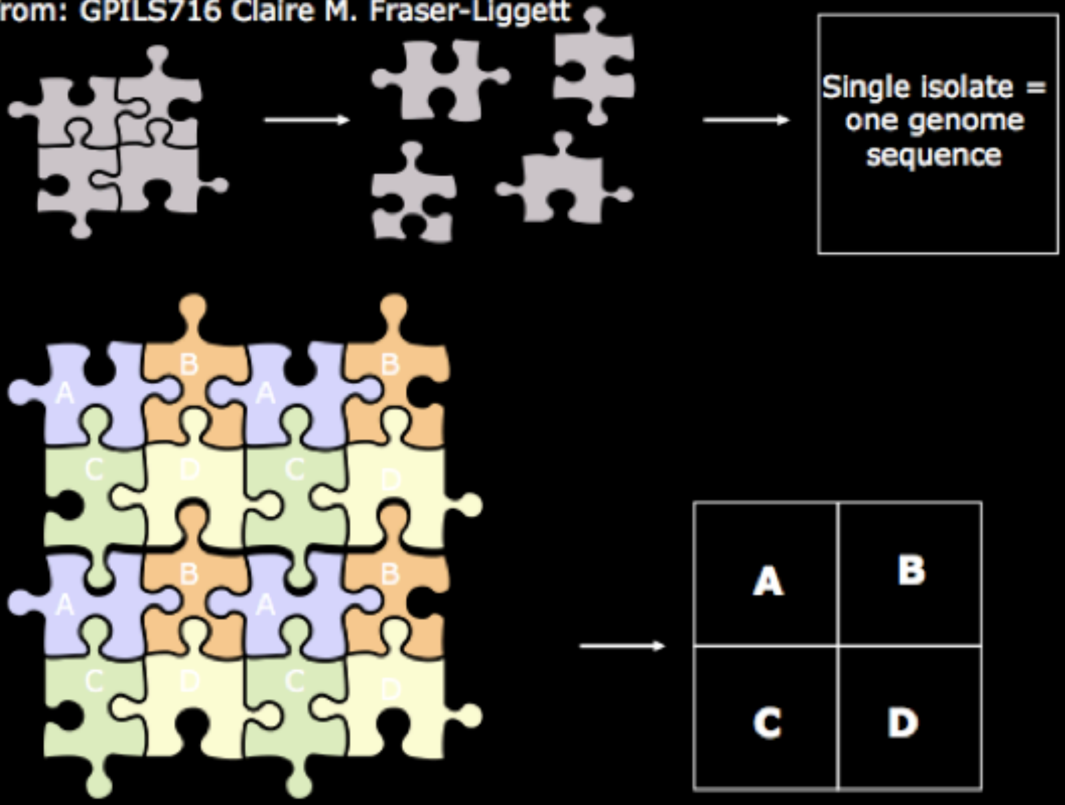
Abstract:
Here we present major improvements to Metastats software and underlying statistical methods.

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?

From: GPILS716 Claire M. Fraser-Liggett



Single isolate = one genome sequence

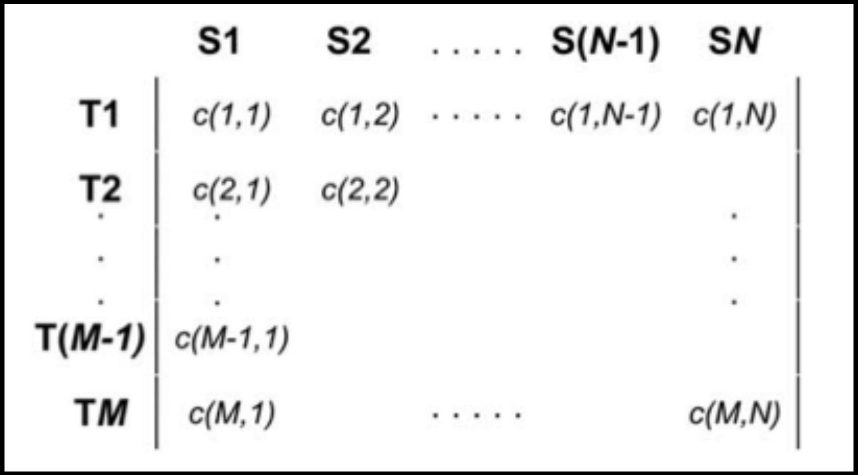
Environmental sample – multiple sources of DNA

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

illustration: Don Smith



	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*)

Hypothesis

$$H_0 := \mu_1 - \mu_2 = 0$$

$$H_1 := \mu_1 \neq \mu_2$$

$$P_{H_0}(t \notin A_\alpha) \leq \alpha$$

- Pvalues
  - P-value is the probability that one observing a test statistic the same or more extreme than what was observed (under H_0)
  - (probability of rejecting hypothesis when it's true)
  - We will reject our null hypothesis when our p-value is less than our significance level (alpha). Ie. significant

Hypothesis

$$H_0 := \mu_1 - \mu_2 = 0$$

$$H_1 := \mu_1 \neq \mu_2$$

$$P_{H_0}(t \notin A_\alpha) \leq \alpha$$

- Pvalues
  - P-value is the probability that one observing a test statistic the same or more extreme than what was observed (under H_0)
  - (probability of rejecting hypothesis when it's true)
  - We will reject our null hypothesis when our p-value is less than our significance level (alpha). Ie. significant

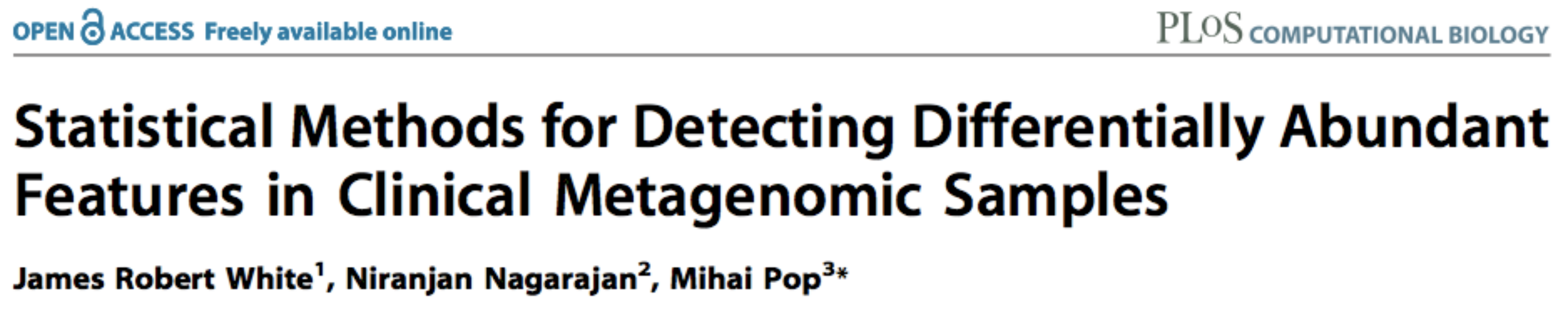
Hypothesis

$$H_0 := \mu_1 - \mu_2 = 0$$

$$H_1 := \mu_1 \neq \mu_2$$

$$P_{H_0}(t \notin A_\alpha) \leq \alpha$$

- Pvalues
  - P-value is the probability that one observing a test statistic the same or more extreme than what was observed (under H_0)
  - (probability of rejecting hypothesis when it's true)
  - We will reject our null hypothesis when our p-value is less than our significance level (alpha). Ie. significant



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples

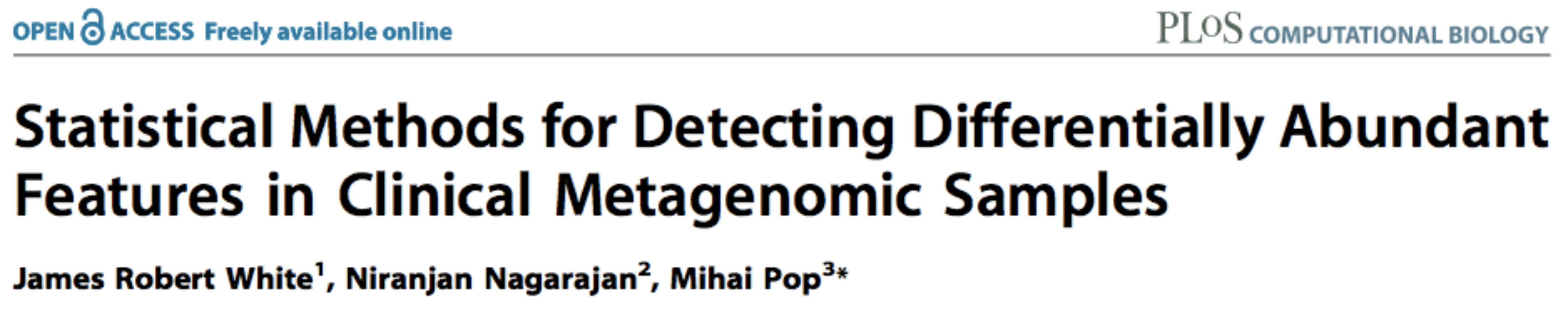
James Robert White[1], Niranjan Nagarajan[2], Mihai Pop[3]*

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

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

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

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



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

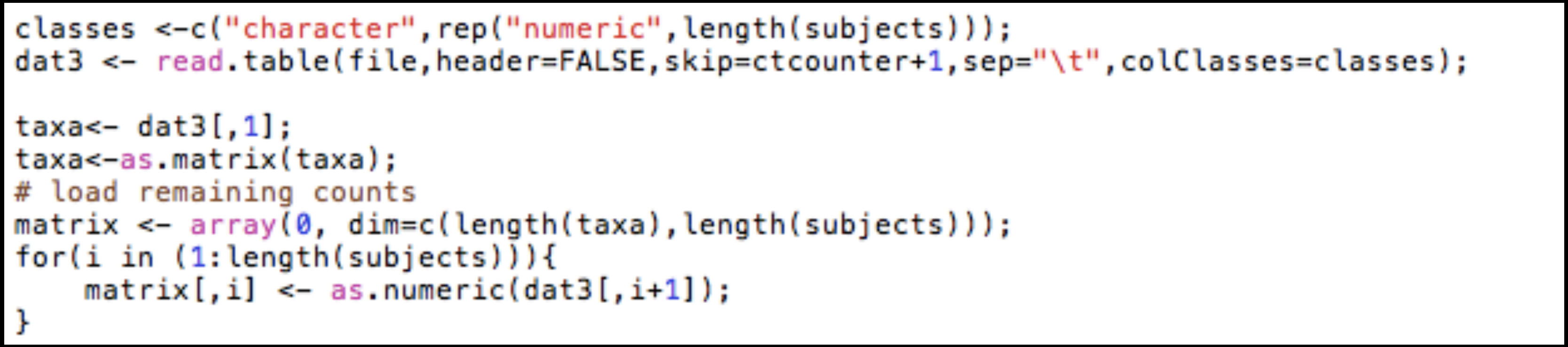
Many “spurious” zeros

Normalization induces spurious correlations

important in time series analyses

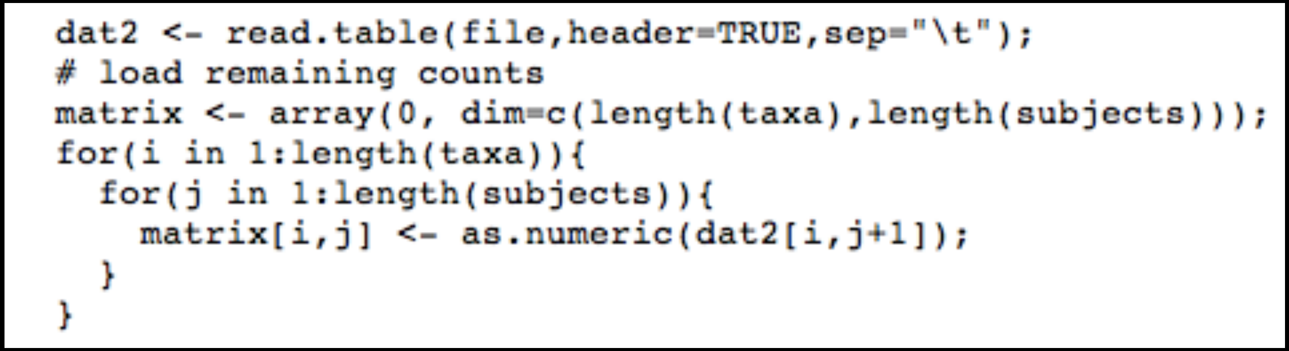
Loading data

- New

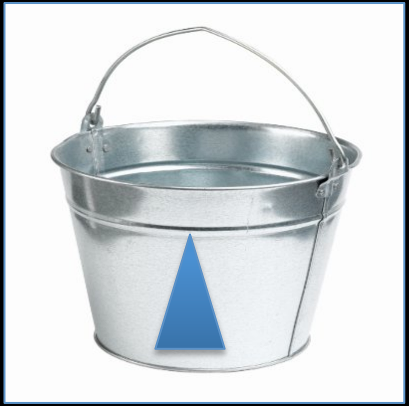
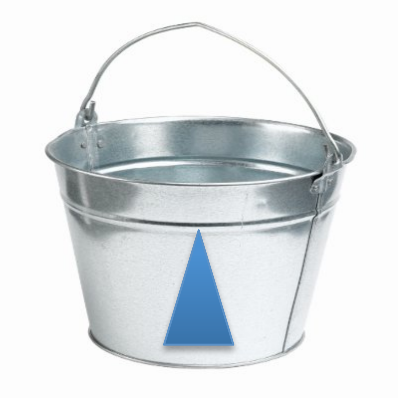
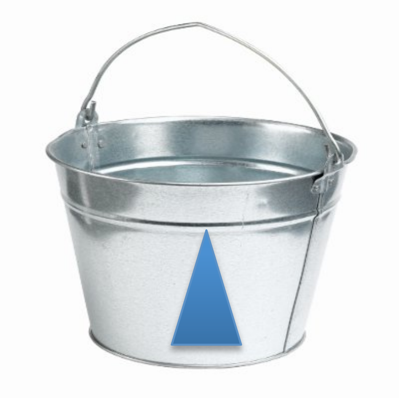
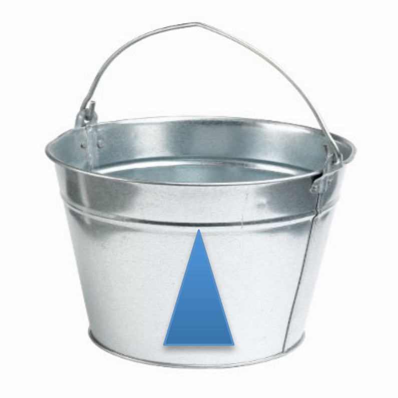
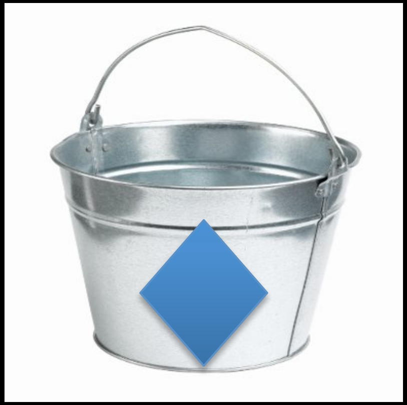
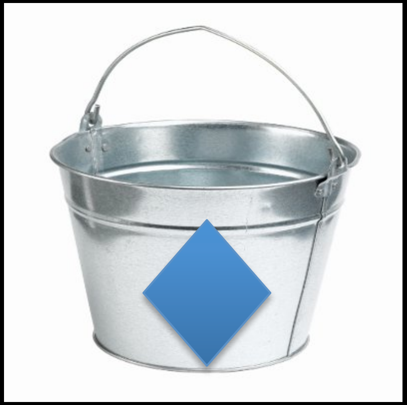
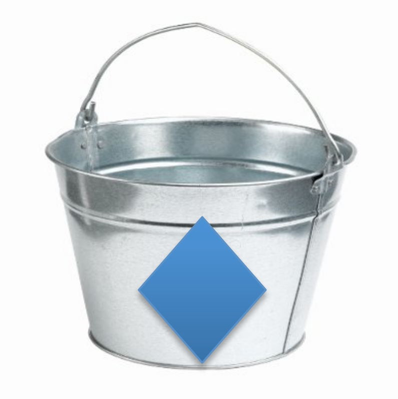
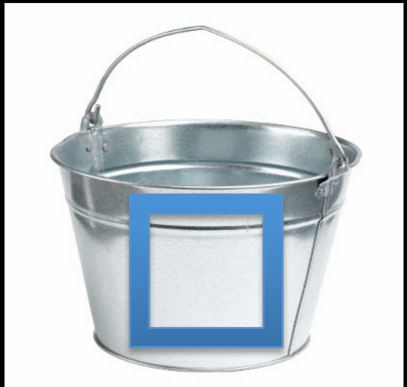
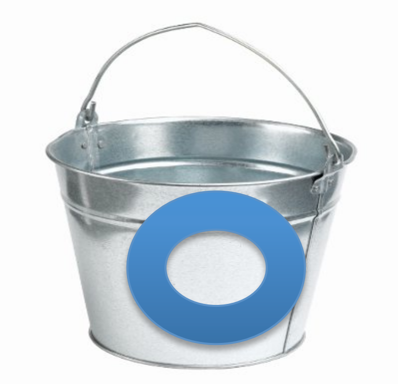
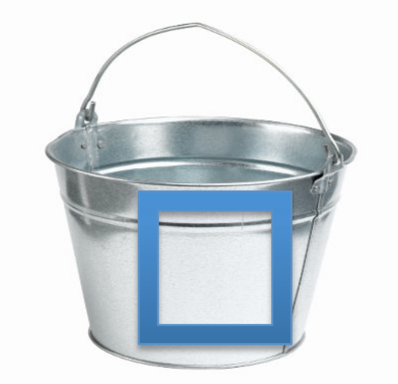


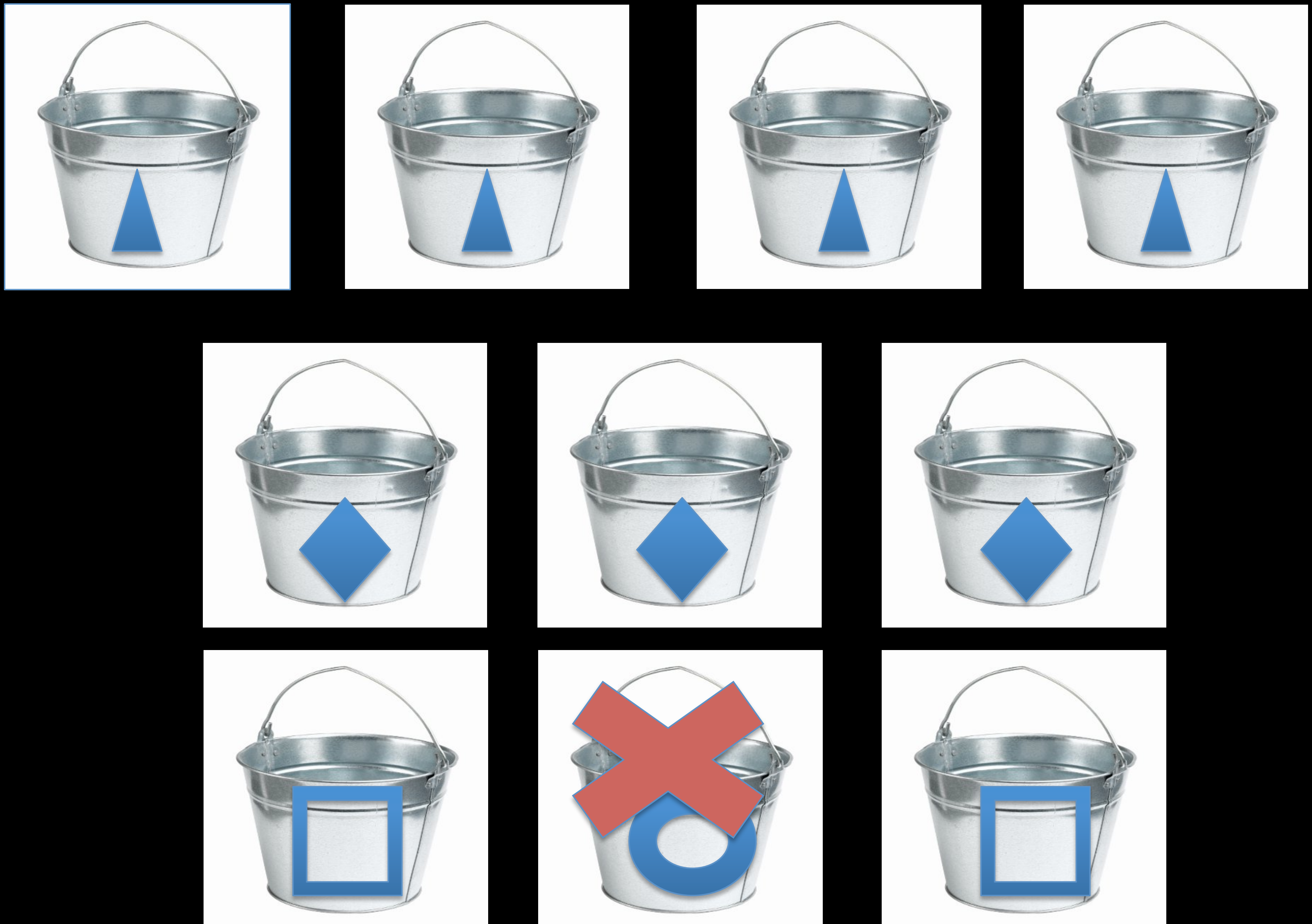
```
classes <-c("character",rep("numeric",length(subjects)));
dat3 <- read.table(file,header=FALSE,skip=ctcounter+1,sep="\t",colClasses=classes);

taxa<- dat3[,1];
taxa<-as.matrix(taxa);
# load remaining counts
matrix <- array(0, dim=c(length(taxa),length(subjects)));
for(i in (1:length(subjects))){
    matrix[,i] <- as.numeric(dat3[,i+1]);
}
```

- Old



```
dat2 <- read.table(file,header=TRUE,sep="\t");
# load remaining counts
matrix <- array(0, dim=c(length(taxa),length(subjects)));
for(i in 1:length(taxa)){
  for(j in 1:length(subjects)){
    matrix[i,j] <- as.numeric(dat2[i,j+1]);
  }
}
```



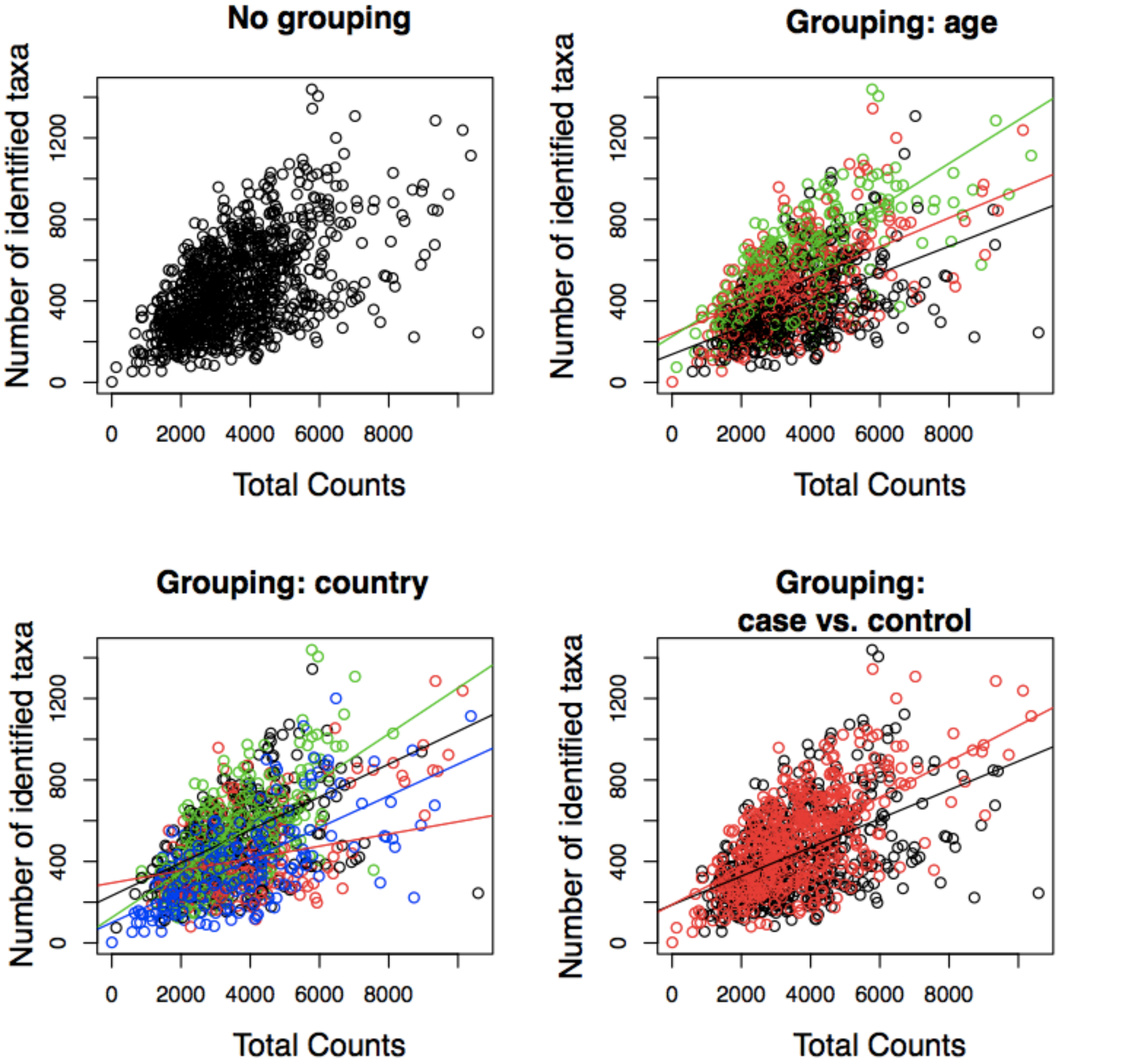
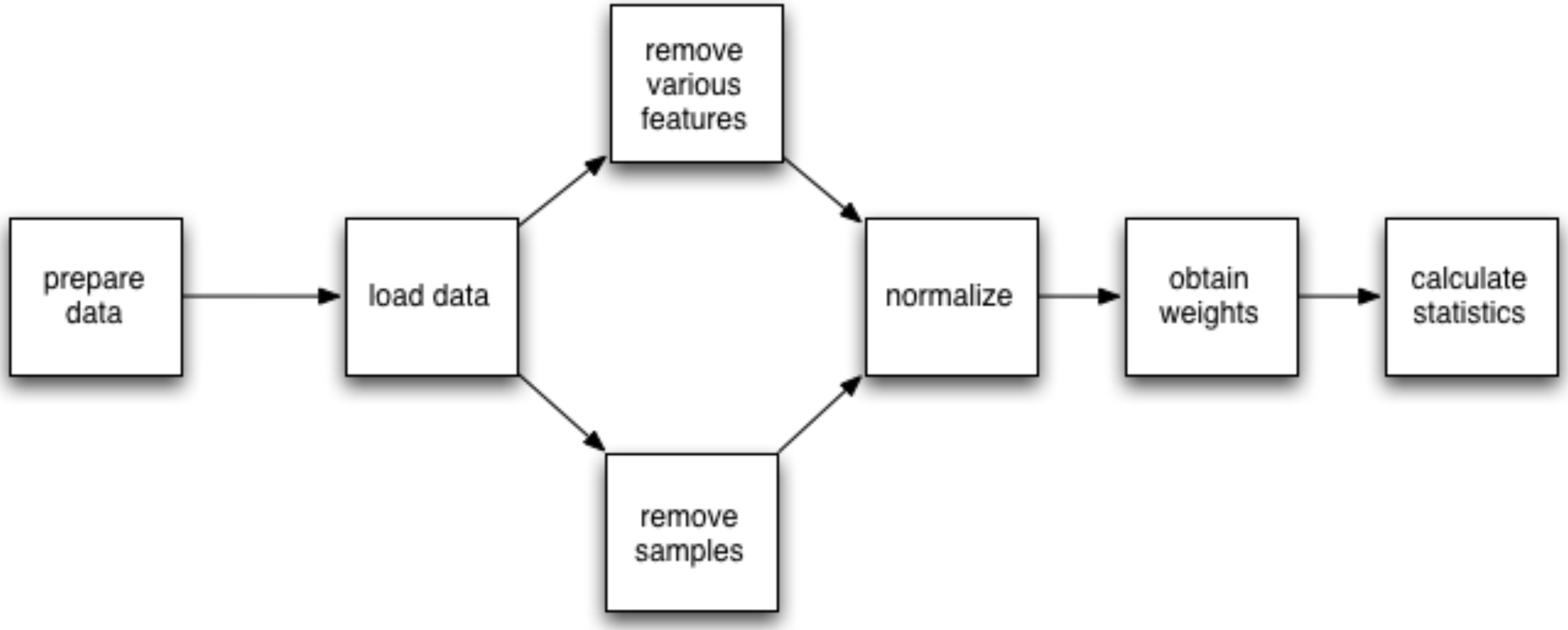


FIG B:
BLACK = AGE 0
RED = AGE 1
GREEN = AGE 2

FIG C:
BLACK =
COUNTRY 0
RED =
COUNTRY 1
GREEN =
COUNTRY 2
BLUE =
COUNTRY 3

FIG D:
BLACK = CASE
RED = CONTROL



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]

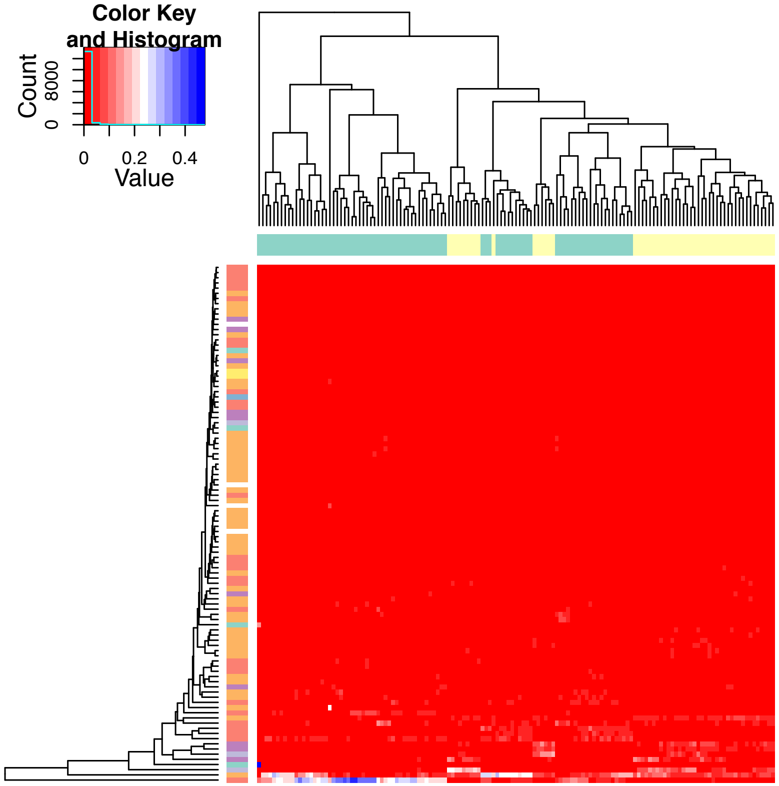
[1] Pearson, Mathematical Contributions to the Theory of Evolution. On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs

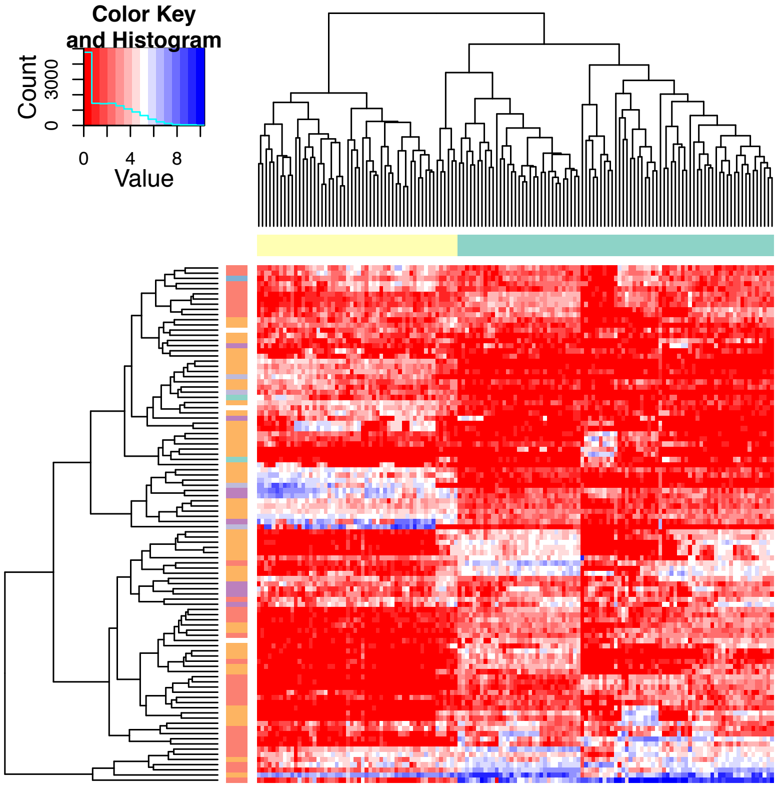
[2] Bullard et. al., Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, BMC Bioinformatics, 2010

Normalization

1. Cumulative Distribution Normalization
    1. Followed by the old method for testing, a

2. Cumulative Sum Normalization
    1. Followed by EM-algorithm

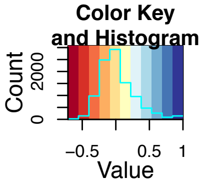


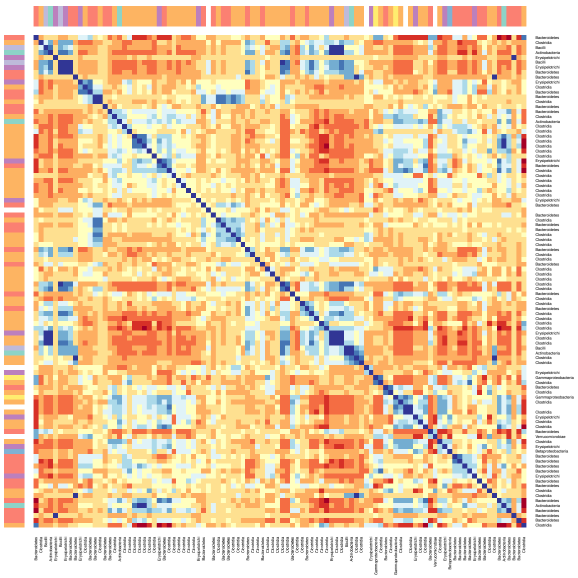


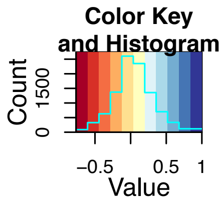
Normalization

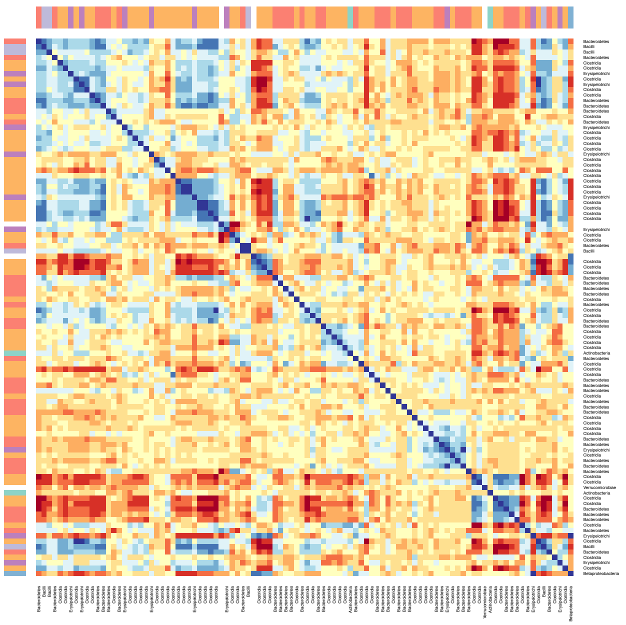
1. Cumulative Distribution Normalization
   1. Followed by the old method for testing, a

2. Cumulative Sum Normalization
   1. Followed by EM-algorithm





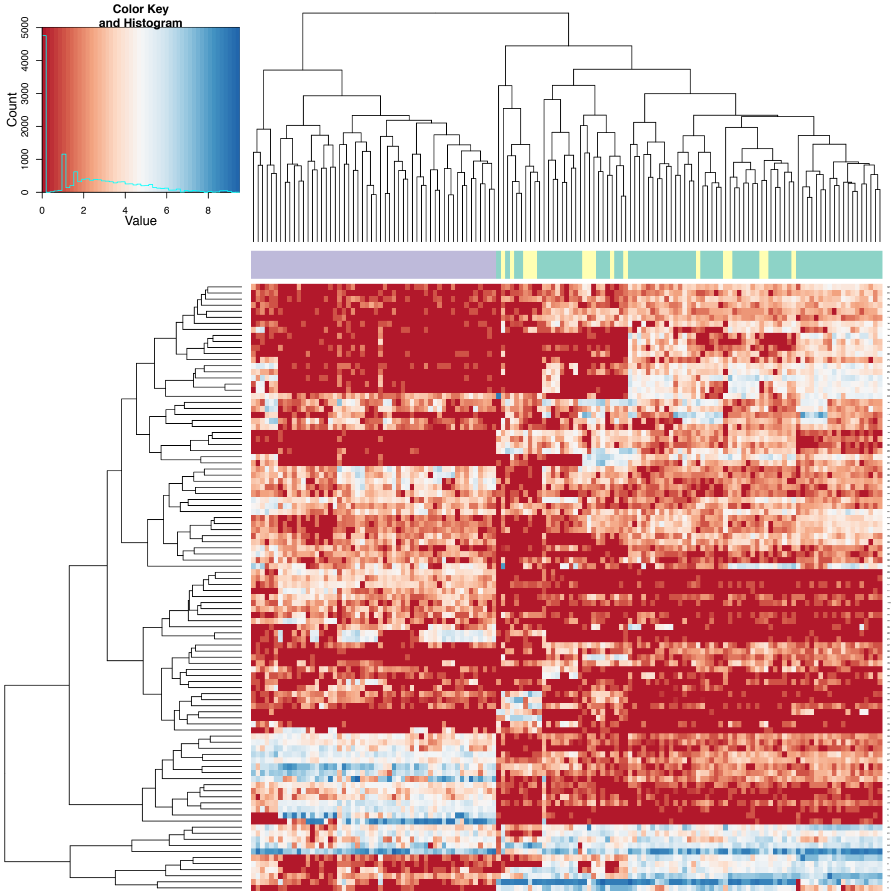




Normalization

1. Cumulative Distribution Normalization
    1. Followed by the old method for testing, a

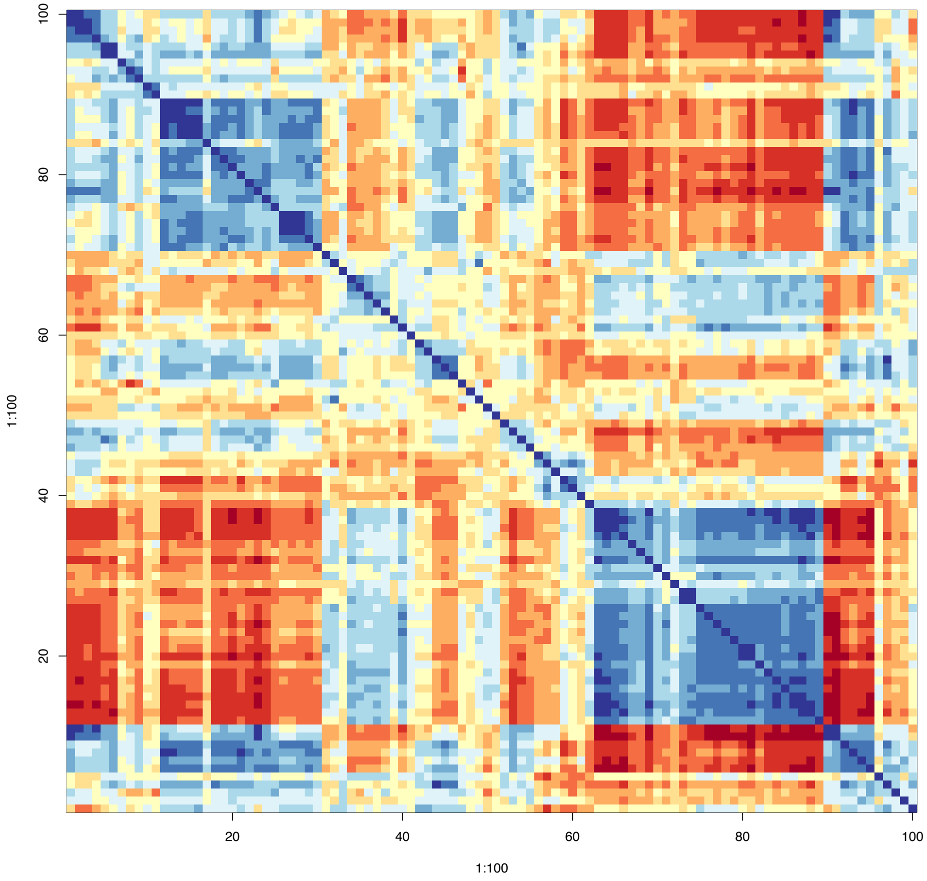
2. Cumulative Sum Normalization
    1. Followed by EM-algorithm



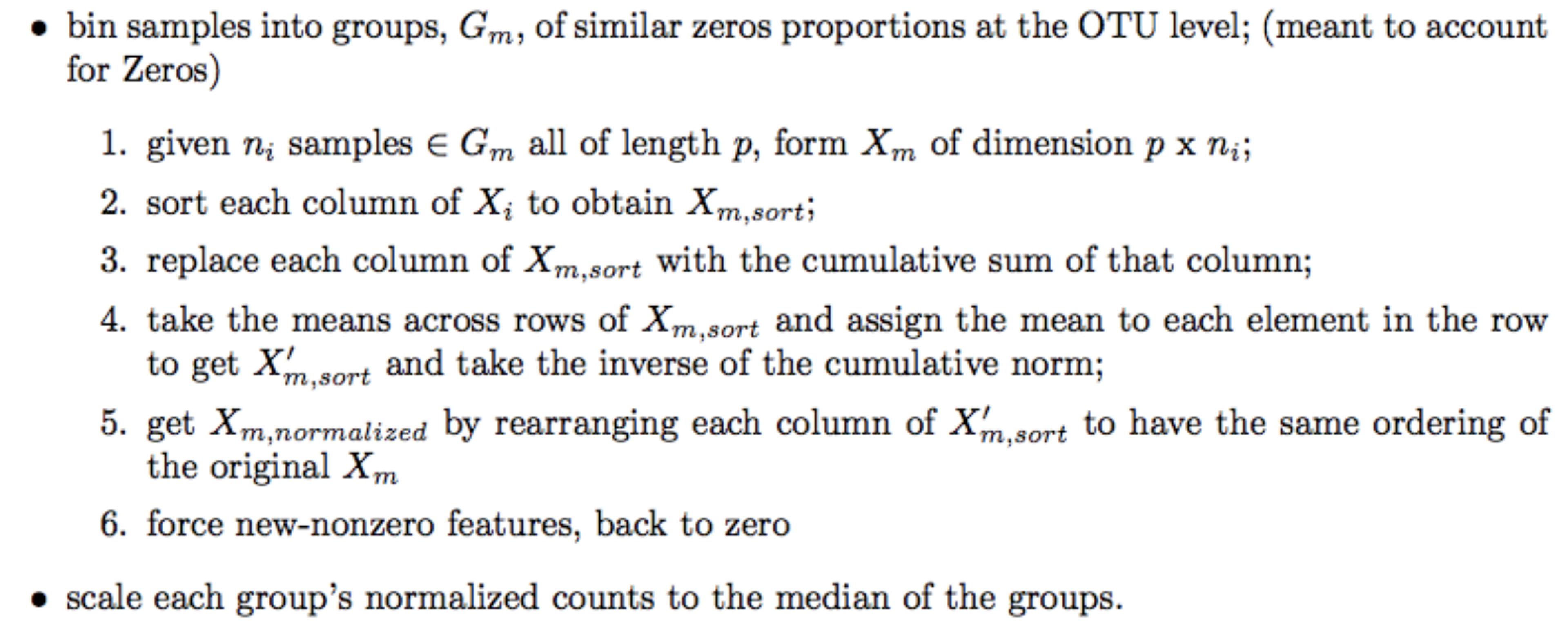
Normalization

1. Cumulative Distribution Normalization
   1. Followed by the old method for testing, a

2. Cumulative Sum Normalization
   1. Followed by EM-algorithm



Cumulative Distribution Normalization



- bin samples into groups, G_m, of similar zeros proportions at the OTU level; (meant to account for Zeros)

  1. given n_i samples $\in G_m$ all of length p, form X_m of dimension p x n_i;
  2. sort each column of X_i to obtain $X_{m,sort}$;
  3. replace each column of $X_{m,sort}$ with the cumulative sum of that column;
  4. take the means across rows of $X_{m,sort}$ and assign the mean to each element in the row to get $X'_{m,sort}$ and take the inverse of the cumulative norm;
  5. get $X_{m,normalized}$ by rearranging each column of $X'_{m,sort}$ to have the same ordering of the original X_m
  6. force new-nonzero features, back to zero

- scale each group's normalized counts to the median of the groups.

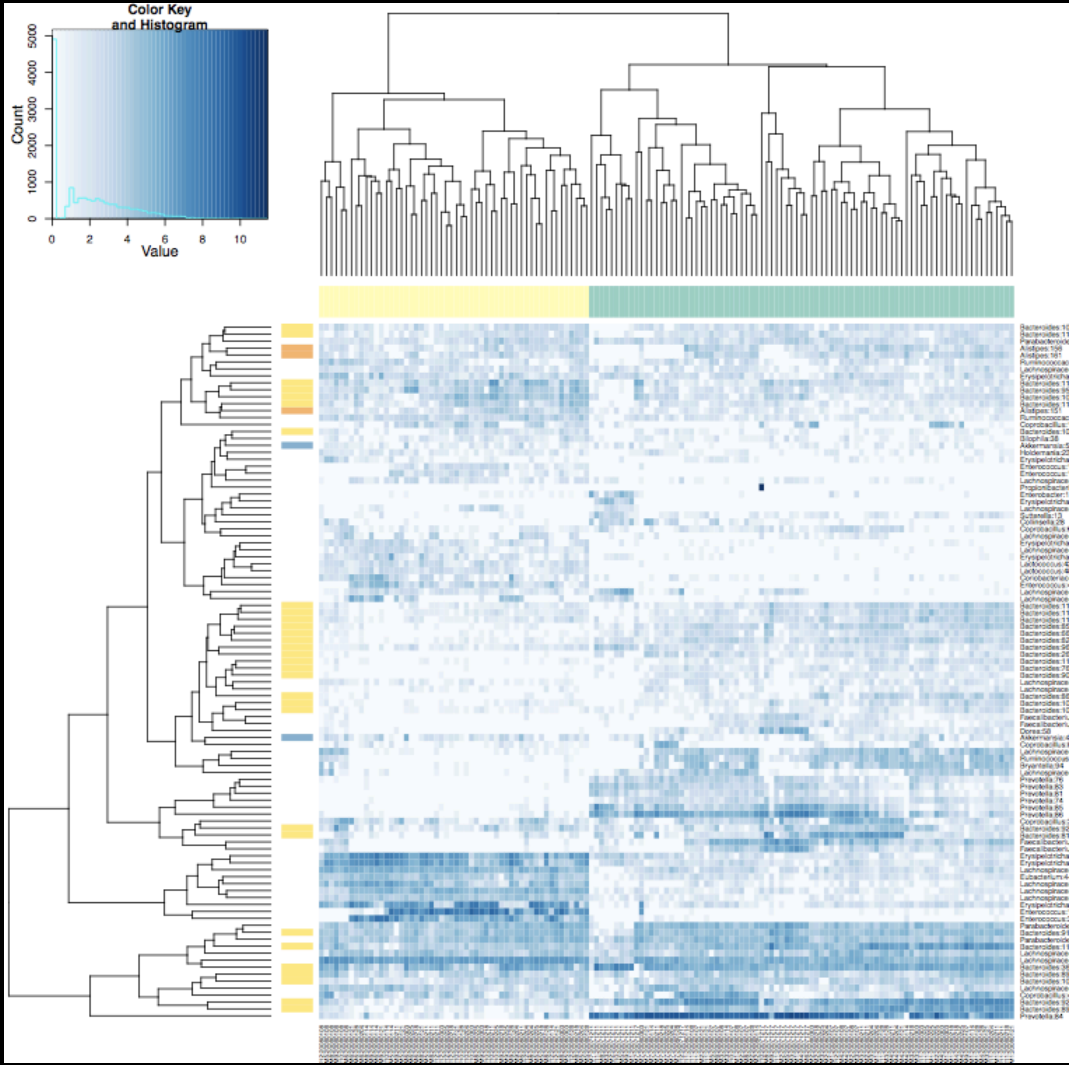
Genes are sampled preferentially as sequencing yield increases (# PCR cycles biases as well).

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

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

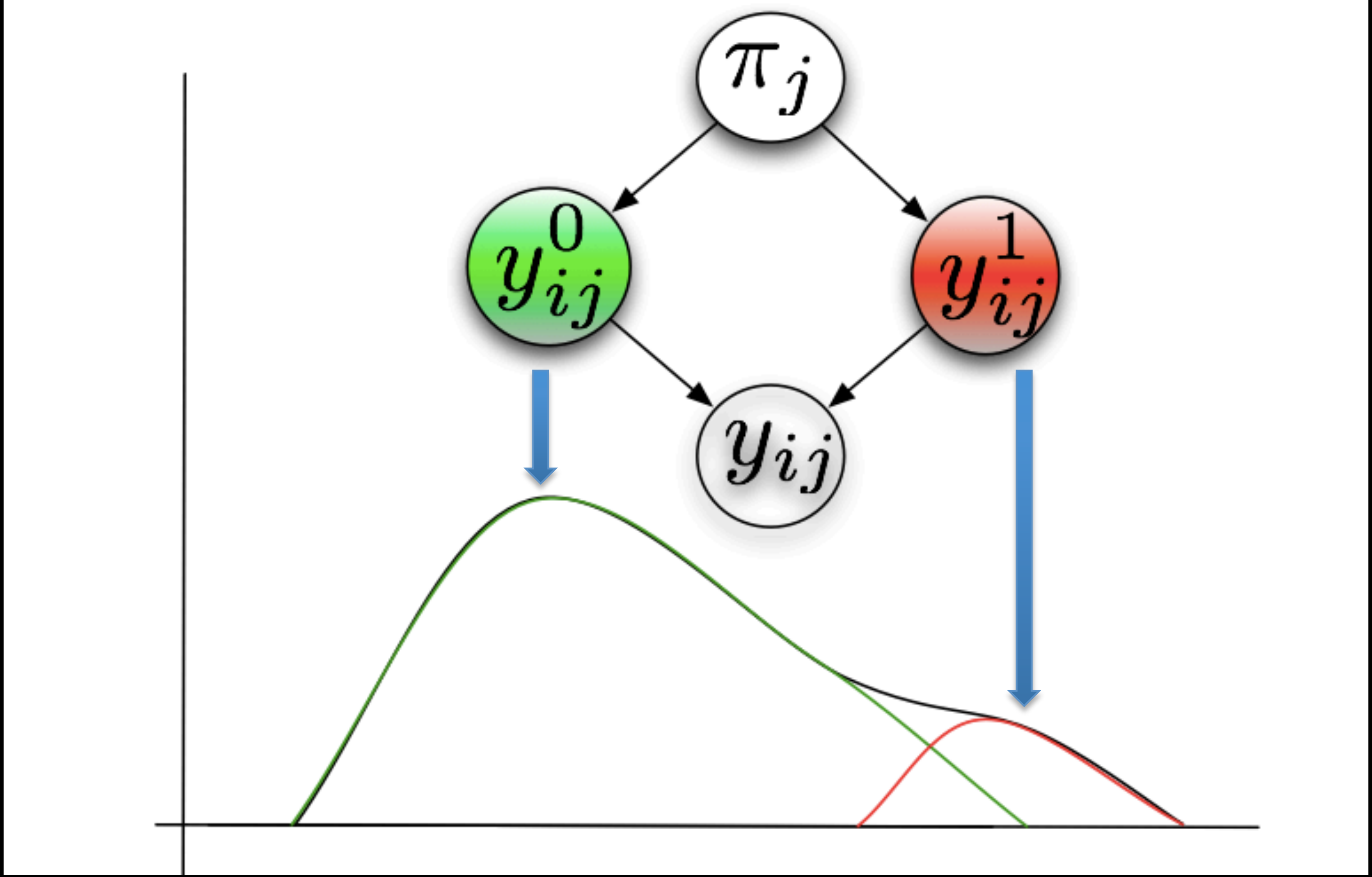
This procedure addresses the issues:

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



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

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



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 $f_{count}(y;\mu,\sigma^2)$~$N(\mu,\sigma^2)$
- 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

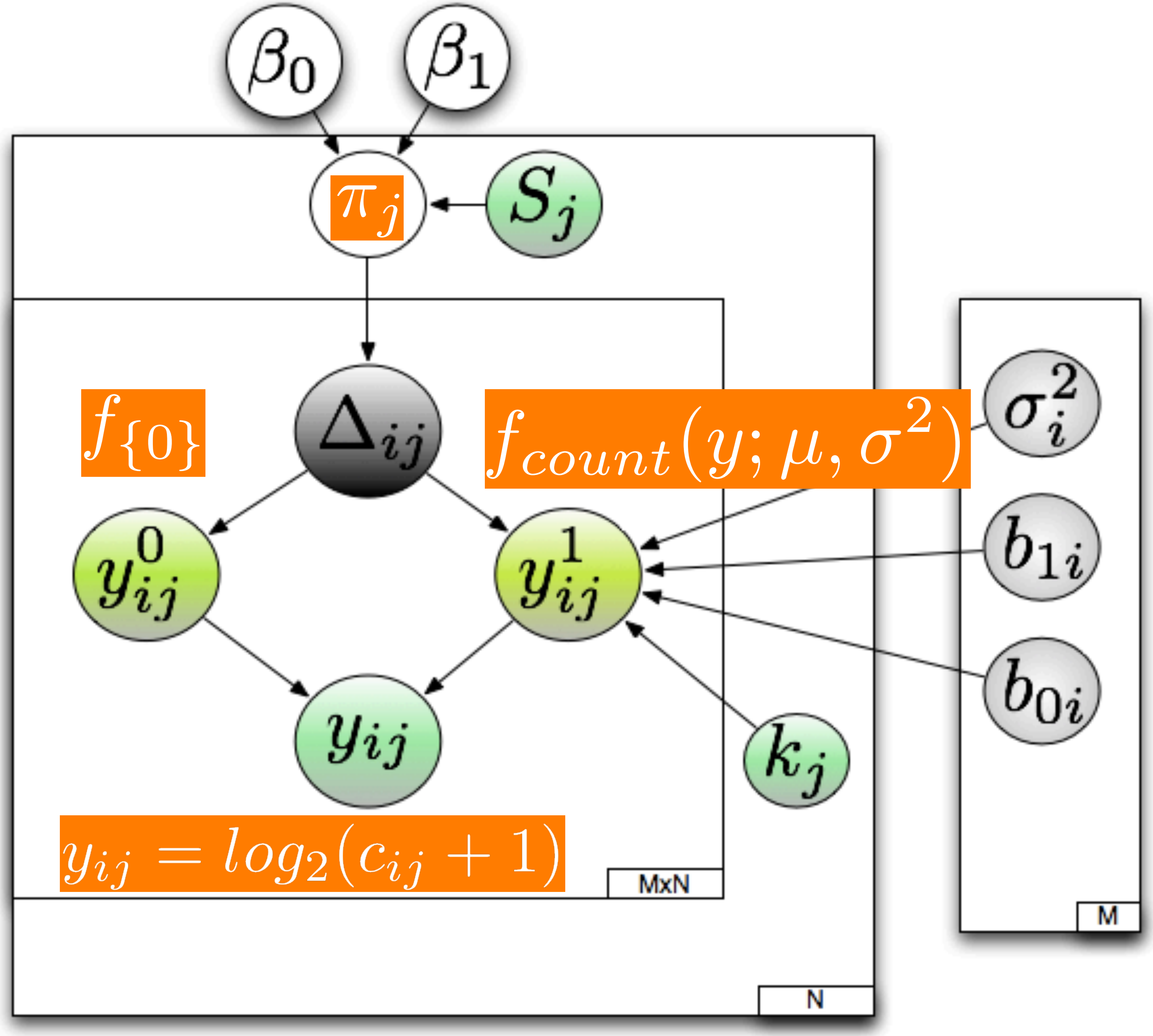
- And a mean specified as:

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

Or $y_{ij} = log_2(c_{ij} + 1)$

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

- 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. 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:

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

Algorithm continued

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

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

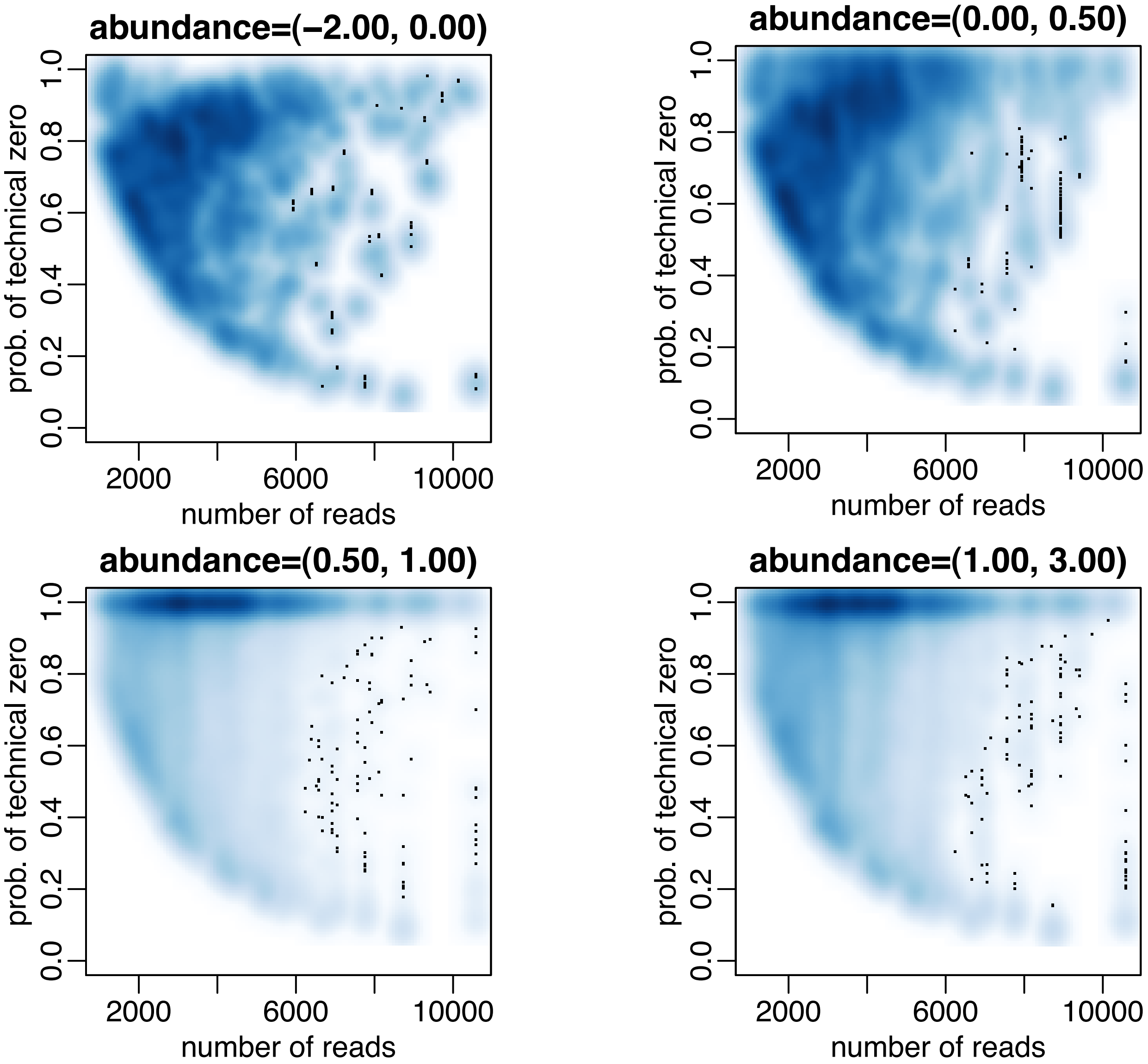
Validation

- For normalization methods it was always checked by hand that the proper normalization was calculated.

- Ensured that data is loaded properly, etc.

- Next up is to compare non-zero matrix results with another method, the log model fit, to ensure exact same results.

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


abundance=(−2.00, 0.00)
abundance=(0.00, 0.50)
abundance=(0.50, 1.00)
abundance=(1.00, 3.00)
prob. of technical zero
0.0
0.2
0.4
0.6
0.8
1.0
number of reads
2000
6000
10000

Eta

No eta

h memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

Project Schedule

- November 30:
  - Preprocessing data
  - Finish normalization codes
  - **Finished**

- December 15:
  - Continue reading
  - Finish Zig model
  - Midyear report
  - **Finished** (except report)

Project Schedule

- Done up to now:
  - Wrote cleanup scripts
  - Wrote cumulative sum normalization scripts
  - Wrote cumulative distribution normalization script
  - Wrote EM algorithm subroutines
  - Prepared scripts to compare various methods
  - Validated by hand loading scripts
  - Validated normalization scripts
  - Validated EM algorithm with non-zero matrix

  - Produced heatmaps of normalized data
  - Produced smoothed scatterplots of the probabilities of weights

Project Schedule

- To do:
  - Finish validating EM Algorithm
  - Check robustness of normalization method by FDR methods
    - Permute counts (within features) …
  - Compare calculated p-values, t-statistics, fold changes to:
    - Old metastats, log, log with eta parameter, Zig no eta parameter
  - Testing of method with simulated data:
    - Compare to Kruskal-Wallis, old method, etc (ROC Curves)
  - Testing and analysis of various datasets including:
    - Gnotobiotic mice
    - Gates dyssentery data
  - Parallelize (if necessary)

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.