Statistical challenges from the analysis of NGS-Metagenomics experiments

Jean-Jacques Daudin & Sébastien Li-Thiao-Té & Emilie Lebarbier
UMR AgroParisTech-INRA
ANR Project *Computational Biology for Metagenomics Experiments* (CBME)

NGS, 24 mars 2010

Outline

Life on Earth (from Wooley et al. PLOS 2010)

- 5.10^{30} prokariotic cells: largest reservoir on Earth, $350-550.10^{15}$g Carbon, $85-130.10^{15}$g Nitrogen, $9-14.10^{15}$g Phosphorous.
- More bacterial cells in one human body (10^{14}) than our own cells (10^{13})
- Bacteria, Archaea and microeukariotes dominate Earth's habitats

- Only a small percentage of microbes can be cultured and sequenced as a sole organism
- Species interact between them in their habitat in real life
- Metagenomics: obtain genomic information directly from microbial communities in their natural habitats

What is Metagenomics?

Genomic study of uncultured microorganisms sampled from their habitats

(A) Sampling from habitat
(B) filtering particles
(C) DNA extraction and lysis
(D) cloning and library
(E) sequence the clones
(F) sequence assembly.

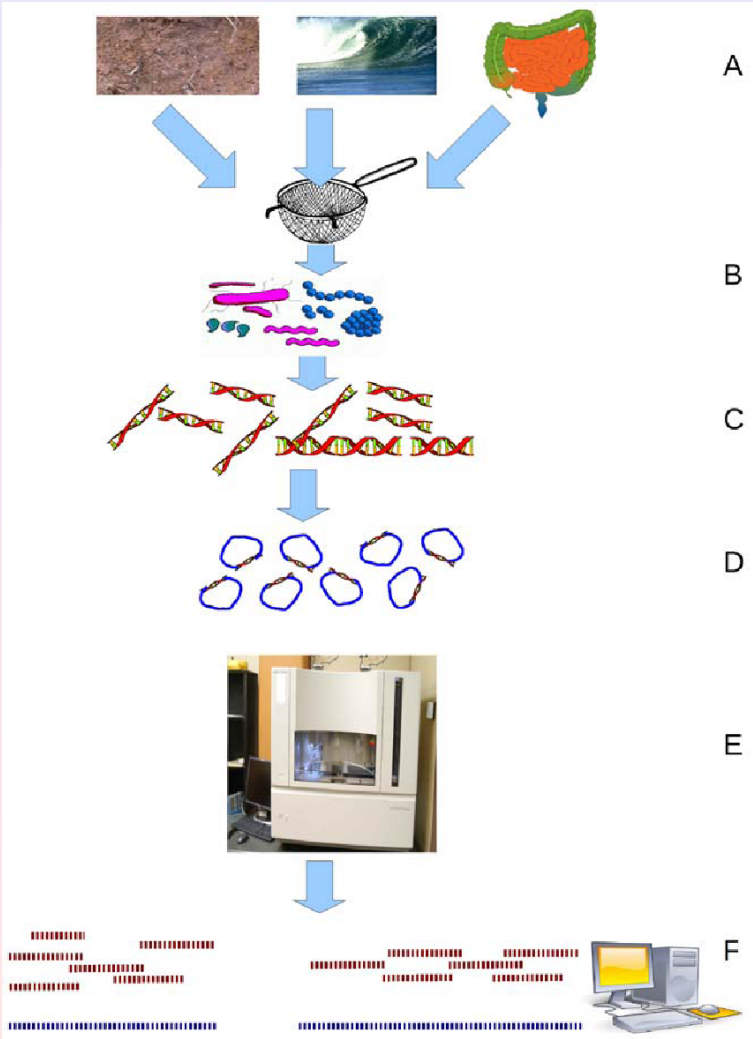


Figure from *A Primer on Metagenomics* John C. Wooley et al. Plos Computational Biology 2010

Applications

- Sequence the genome of all the life on earth (soil, sea, air, life)
- Discovery of new genes, enzymes, functions → fine chemicals, agrochemicals and pharmaceuticals
- Monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments.
- Human Microbiome: understand the changes in the human microbiome that can be correlated with human health
- Understand microbes communities, measure biodiversity...

Bioinformatics challenges

- In cultured microbes, the genomic data come from a single clone, making sequence assembly and annotation tractable.
- In metagenomics, the data come from heterogeneous microbial communities (10 to 10,000 species), with the sequence data being noisy and partial.

From sampling, to assembly, to gene calling and function prediction, bioinformatics faces new demands in interpreting voluminous, noisy, and often partial sequence data.

Statistical challenges

- Normalization of NGS data (see Dudoit et al.)
- How many species (genes) ?
- Repeatability and Comparative Metagenomics
- Clustering/Binning the reads

From the reads to the dataset: Mapping or Binning ?

Definition of the counted objects (Species, OTU, genes)

Reads give counts. What do we count exactly ?

1. Mapping: alignment to a reference set (reference genome or 16S rDNA or 18S rDNA banks). Issues: incomplete, depends on updates and databases, variable number of repetitions of 16/18S rDNA.
2. Binning or clustering using similarities between reads. Issues: needs long reads (100bp), the objects are not well defined.

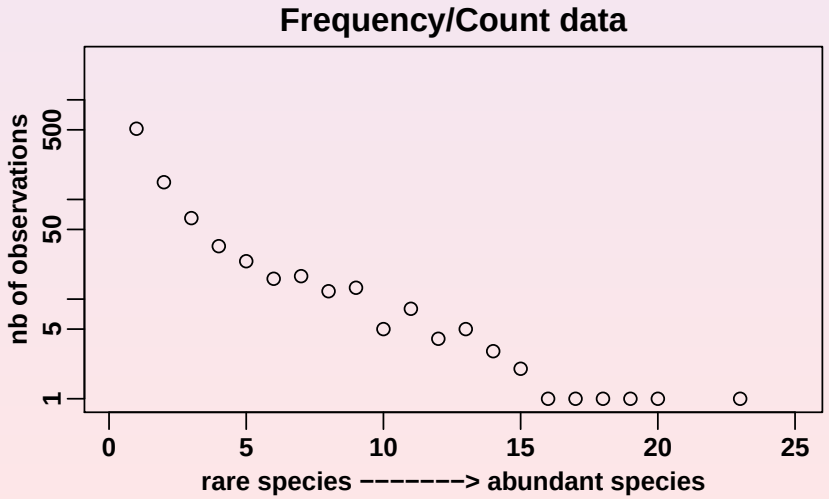
Challenge: estimate the number of unseen objects

Data obtained after mapping / binning:

Species	A	B	C	D	E	...
Nb times seen	10	430	10	289	3	...

Frequency / Count data:

Nb of occurences	0	1	2	3	4	5	...
Nb of species	?	513	149	65	34	24	...



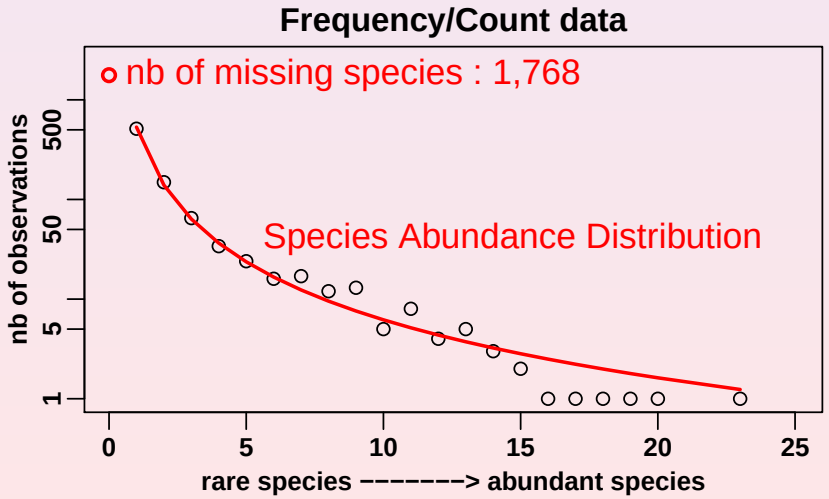
Challenge: estimate the number of unseen objects

Data obtained after mapping / binning:

Species	A	B	C	D	E	...
Nb times seen	10	430	10	289	3	...

Frequency / Count data:

Nb of occurences	0	1	2	3	4	5	...
Nb of species	?	513	149	65	34	24	...



An old question, see J. Bunge(1993) for a review

- butterflies (Fisher 1943)
- number of words in a language (Efron 1976)
- number of coins, cholera epidemic, number of drug users
- sampling fish from a lake, insects from a forest

Objectives:

- study the relative distribution of species
- estimate the biodiversity
- plan the number of experiments (sequencing effort)

An empirical tool: Rarefaction curve

- Green: most or all species have been sampled
- blue: this habitat has not been exhaustively sampled
- red: species rich habitat, only a small fraction has been sampled.

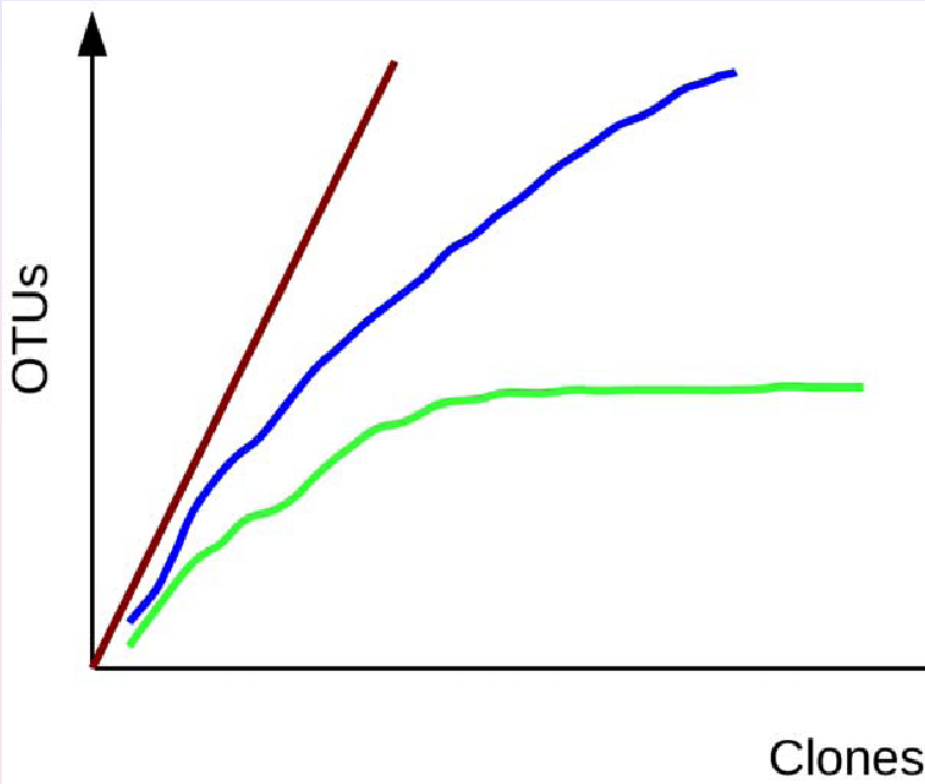


Figure from *A Primer on Metagenomics* John C. Wooley et al. Plos Computational Biology 2010

Standard sampling model

C: total number of species.
Each species i contributes X_i observed individuals.
$X_i \sim \mathscr{P}(\lambda_i)$
Species Abundance Distribution (SAD): distribution f of X.
The SAD can be modelled :

- directly, e.g. lognormal, inverse Gaussian, Pareto
- via a model for λ_i. In the usual model, λ_i are iid samples of Λ with distribution f_Λ.

$$f(x) = \int \frac{e^{-\lambda}\lambda^x}{x!} f_\Lambda(\lambda)\mathrm{d}\lambda$$

Models for the Species Abundance Distribution

f_Λ	f
Dirac	Poisson
	Lognormal
Exponential	Geometric
Gamma	Negative Binomial
	inverse Gaussian
	Pareto

Mixture models,

$$f_\Lambda(\lambda,\theta) = \sum_{j=1}^{g} \pi_j f_{\Lambda j}(\lambda,\theta)$$

$$\mathscr{P}[X = x|\theta] = f(x,\theta) = \sum_{j=1}^{g} \pi_j f_j(x,\theta)$$

$$\text{where } f_j(x,\theta) = \int \frac{e^{-\lambda}\lambda^x}{x!} f_j(\lambda)\mathrm{d}\lambda$$

Truncated observations

Species with no observed individuals do not appear in the dataset

X_i is not observed but rather "$Y_i = X_i | X_i > 0$"

$$f(y) = \frac{f(x)}{1 - f(0)}$$

Parametric / frequentist approaches

Algorithm

- fit a zero-truncated parametric model to the data,
- compute the coverage $1 - \hat{p}_0$ based on the parameters,
- deduce $\hat{C} = \frac{n}{1 - \hat{p}_0}$

Features:

- finite mixture of exponentials, lognormal, inverse Gaussian are popular
- asymptotic covariance matrix is known

Issues:

- model selection
- actual variance and confidence intervals not explored

Non-parametric / frequentist approaches

Chao-type estimators:

- Chao1: $n + f(1)/(2f(2))$
- ACE: $\frac{n}{1-f(1)/n} + \frac{f(1)}{1-f(1)/n}\gamma^2$

Features:

- computationally simple
- asymptotic variance formula available

Issues:

- choice of γ
- large variance

Comparison

Lepidoptera	$\tau = 10$	
Method	Estimate (SE)	Interval
Parametric-Bayesian		
Objective priors	266 (20.7)[a]	(247,313)
Parametric-frequentist, MLE		
Bunge & Barger (2008)	266 (9.8)[a]	(252,293)
Nonparametric, coverage-based		
Chao and Lee (1992)	263 (8.4)[d]	(252,286)
Nonparametric, MLE		
Böhning (2005)	335 (523.7)	(256,1405)

Questions

Statistical Questions

- Confidence intervals need improvement
- Model selection
- Robustness:
  Poorly behaved with respect to outliers.
  Little work on sensitivity/robustness to model choice.
- Modelling covariate information

Definition of the counted objects (Species, OTU, genes)

From which object comes a read ?

- Alignment to a reference set (reference genome or 16S rDNA or 18S rDNA banks): incomplete, repetition
- clustering of OTU using BLAST similarities between reads

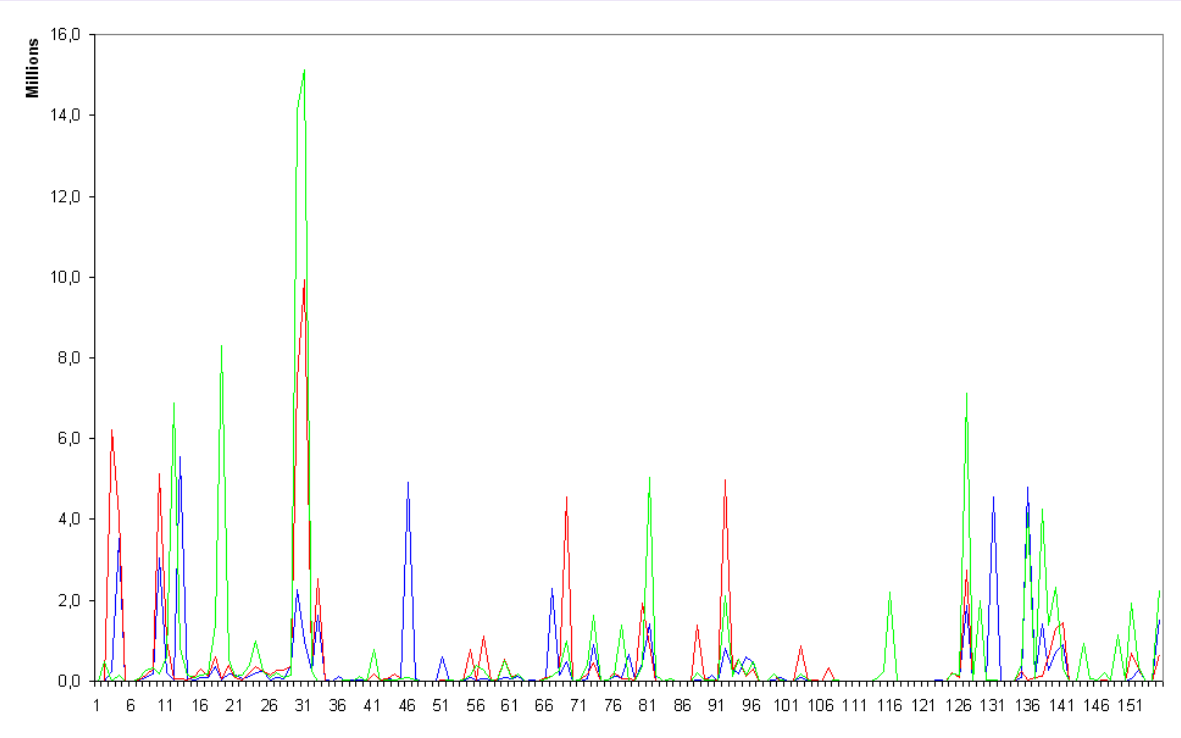
Comparative Metagenomics: data set

Assume a reference genome is available.

Sample	X_1	...	X_p	Y_1	...	...	Y_q
1	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
n	...	...	...	...	...	...	...

- X_1, ...X_p: continuous/discrete covariates describing each sample. In MetaHit study (Nature 2010), $p = 5$: gender, desease (Crohn/Ulcerative Colitis/Healthy), age, bmi, country
- Y_1, ...Y_q: variables describing the reads for each sample. In MetaHit study , $q = 155$, Y_j : sum of reads lenght aligning on species j, $j = 1, q$.

Repeatability of samples: high biological variability in many Metagenomics Experiments



x-axis : species (from 1 to 155). y-axis number of reads.

Profile of 3 samples with the same covariates in a gut-metagenomic experiment.

Repeatability of samples

Consequences of the variability between samples with the same values for covariables

- Repetitions within treatment are essential
- Low power of the tests
- The experimental design is crucial for Metagenomics Experiments with comparative purpose
- The statistical model must take this variability into account

Statistical Models

Y_{sij} = nb of reads of sample j from condition i, aligned on the species s. ($i = 1,2$, $j = 1, n_i$ and $s = 1,155$ for the gut exp.)

Models or Tests

for each s (excepted 4 and 6), $Y_{ij} \sim \mathscr{P}(\lambda_{ij})$ and $\lambda_{ij} \sim \mathscr{L}(\mu_i, \sigma_i)$

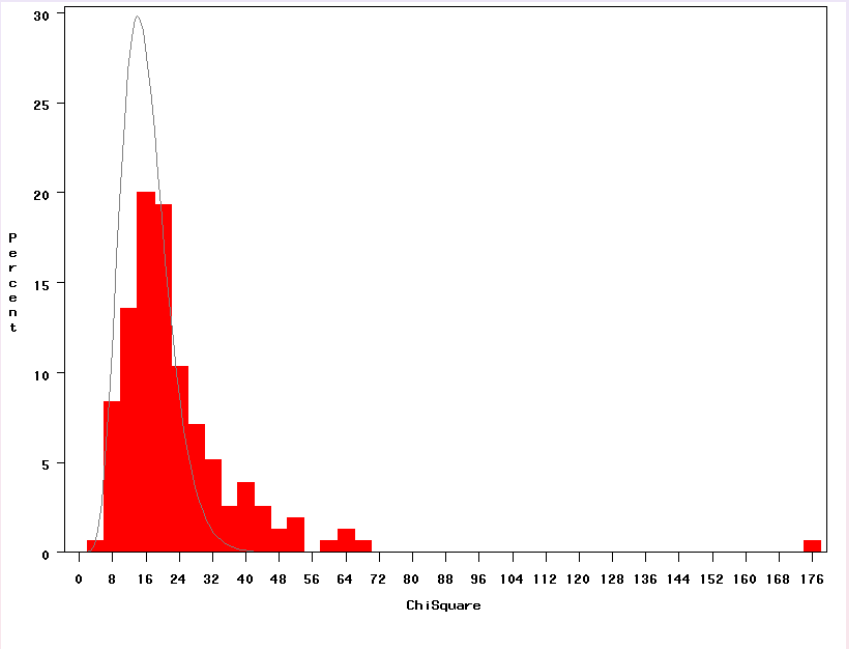
1. $\mathscr{L}(\mu_i, \sigma_i) = \delta_{\mu_i}$ standard Poisson regression,
2. $\mathscr{L}(\mu_i, \sigma_i) = \delta_{\mu_i}$ standard Poisson regression with a scale factor
3. $\log \lambda_{ij} \sim \mathscr{N}(\mu_i, \sigma)$ Poisson regression with random sample effect and log-link
4. $\log Y_{ij} = \mu + \alpha_i + E_{ij}$ with $E_{ij} \sim \mathscr{N}(0, \sigma)$ GLM after log-transformation
5. $\mathscr{L}(\mu_i, \sigma_i) = \Gamma(\mu_i, \sigma_i) \Leftrightarrow Y_{ij} \sim$ Negative Binomial
6. non-parametric Wilcoxon test for comparing the 2 conditions

Comparison of Statistical Models

1. Standard Poisson regression (with or without scale factor) Should not be used in Metagenomics experiments: does not properly integrates the biological variance between samples. This is also true for the Fisher exact test.
2. Poisson regression with random sample effect and log-link and GLM after log-transformation (GLM-Log) give very similar results, although the former does not converge in few cases.
3. Negative Binomial model gives results different from (GLM-Log). Seem to be less robust to extreme values. Gives more false positive (White et al. PLOS CB,2009).
4. Wilcoxon test is concordant with GLM-Log. However P-Values may be too high for using multiple-test control (i.e. FDR) with low repetition numbers. Ties.

Fit of the models

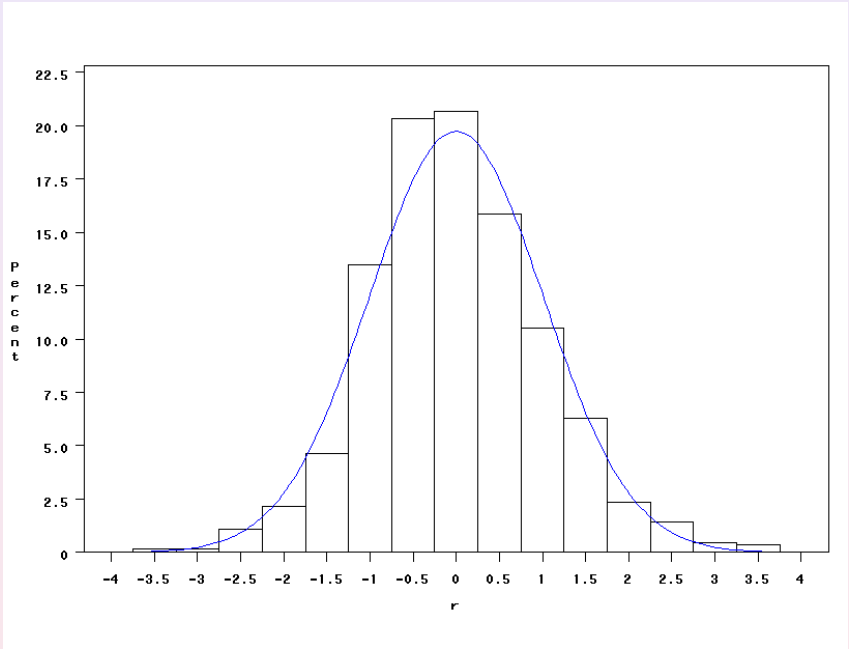
BN Chi-Square Statistics



x-axis:Fit Statistic for BN model, y-axis:frequency

curve: ChiSquare pdf

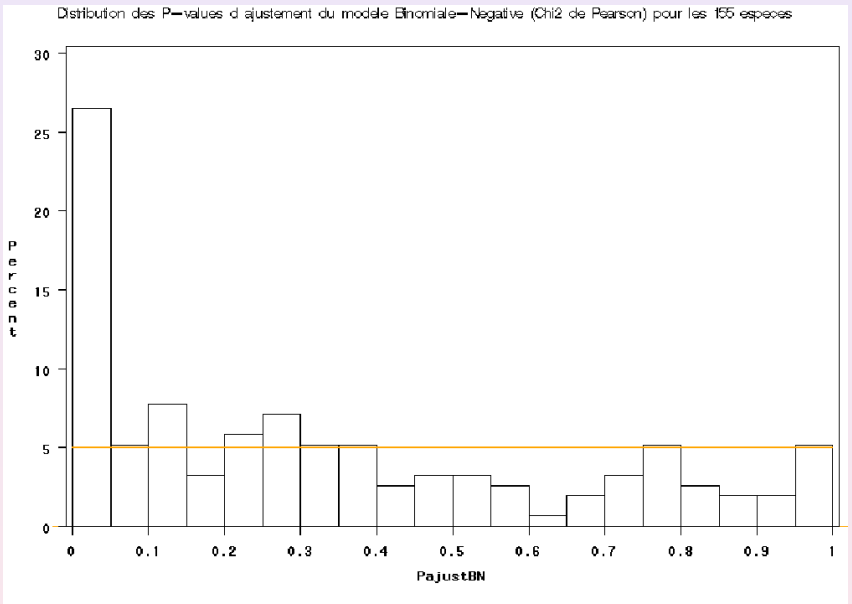
GLM residuals



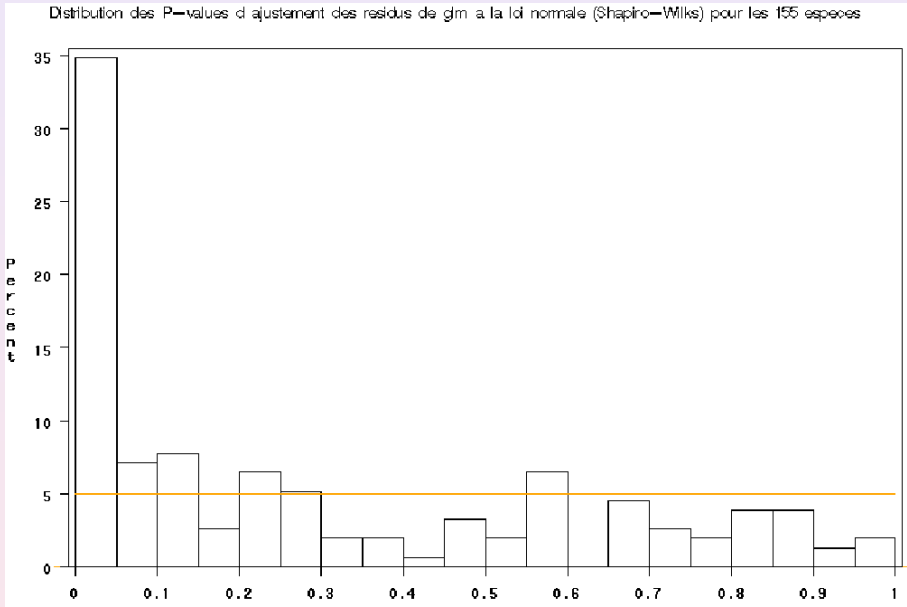
x-axis:GLM Residual, y-axis:frequency

curve: $\mathcal{N}(0,1)$ pdf

Fit of the models(2)

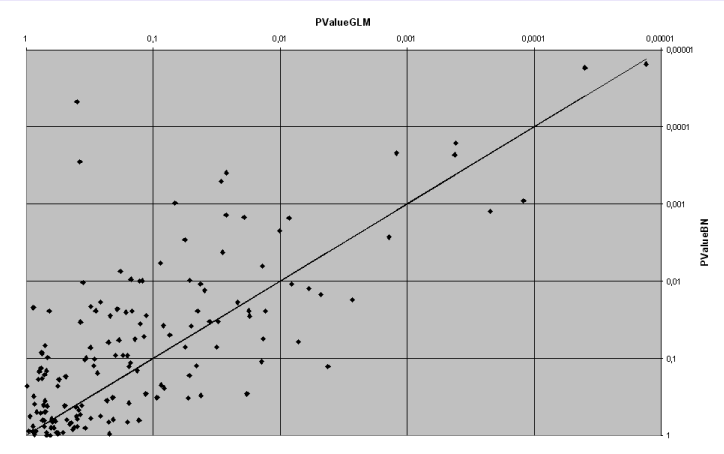


x-axis:PvaluesFitBN, y-axis:frequency

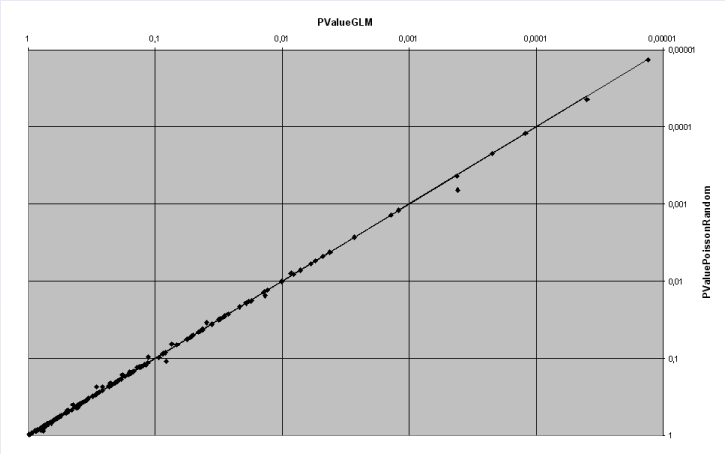


x-axis:PvaluesFitGLM, y-axis:frequency

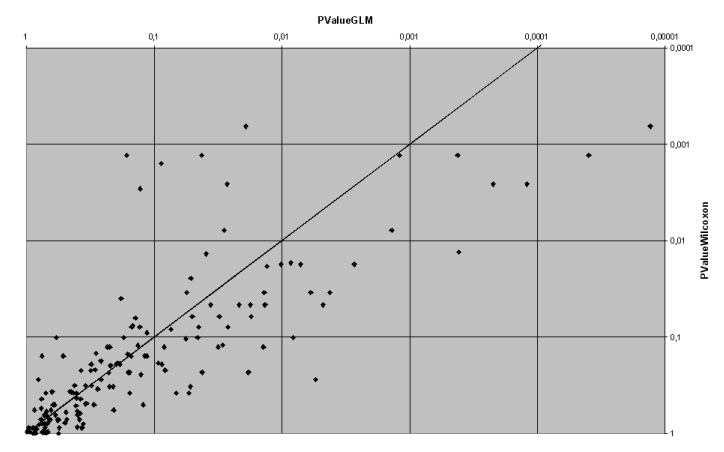
Comparison of PValues for different methods



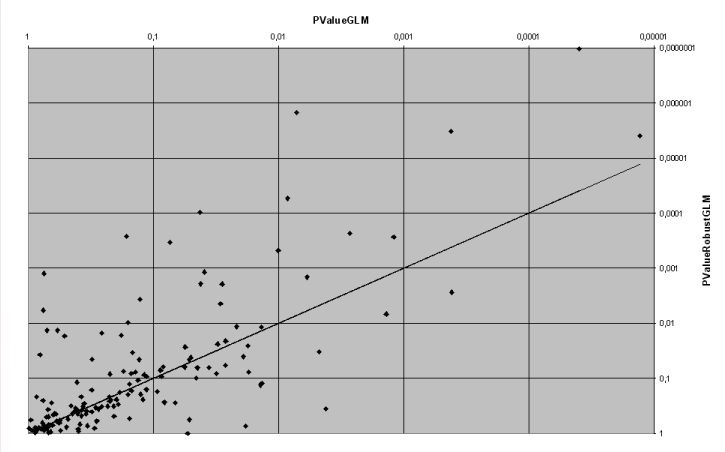
x-axis:PValuesGLM, y-axis:PValuesBN



x-axis:PValuesGLM, y-axis:PValues Mixed-Poisson



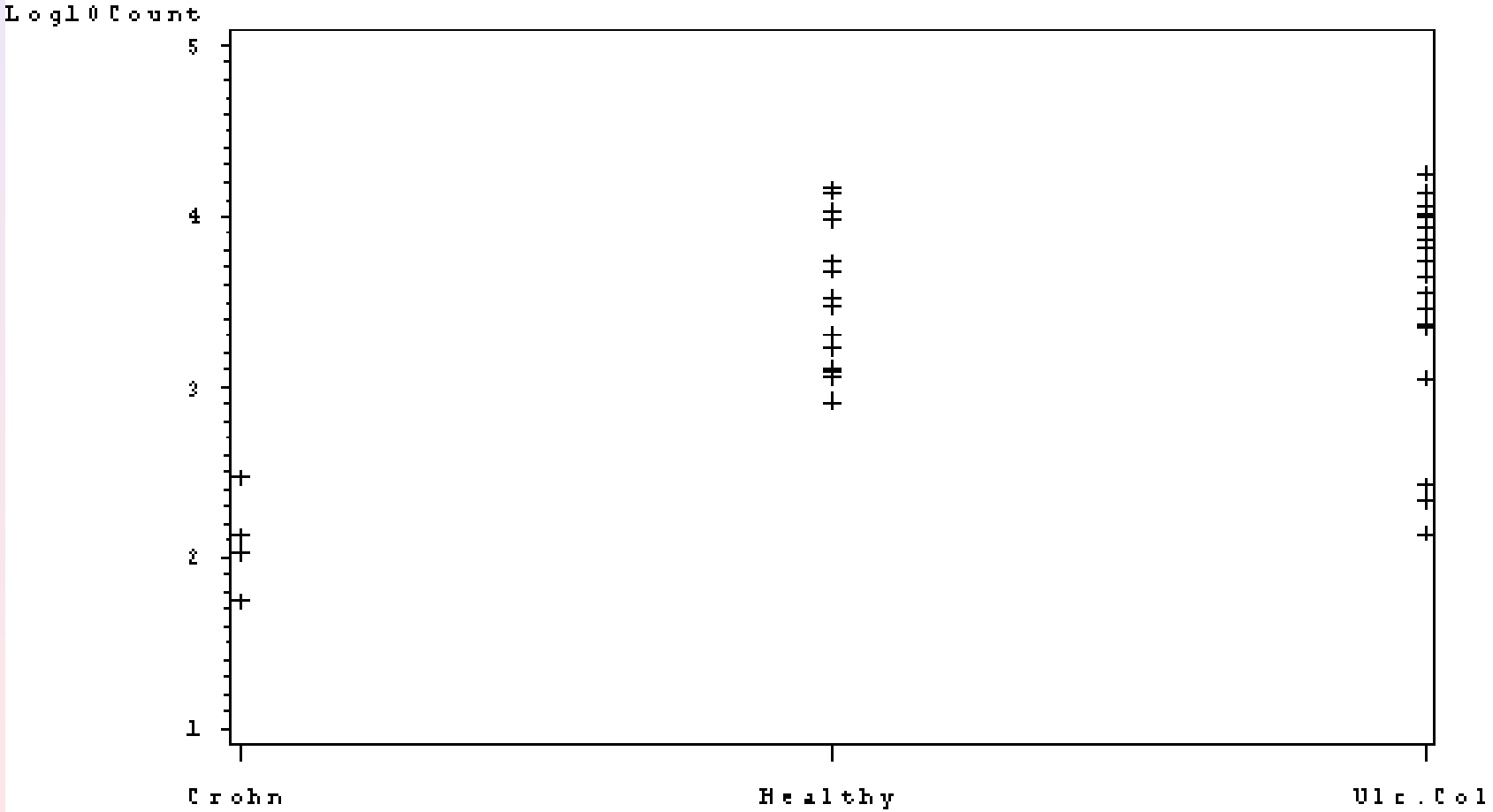
x-axis:PValuesGLM, y-axis:PValuesWilcoxon



x-axis:PValuesGLM, y-axis:PValues RobustGLM

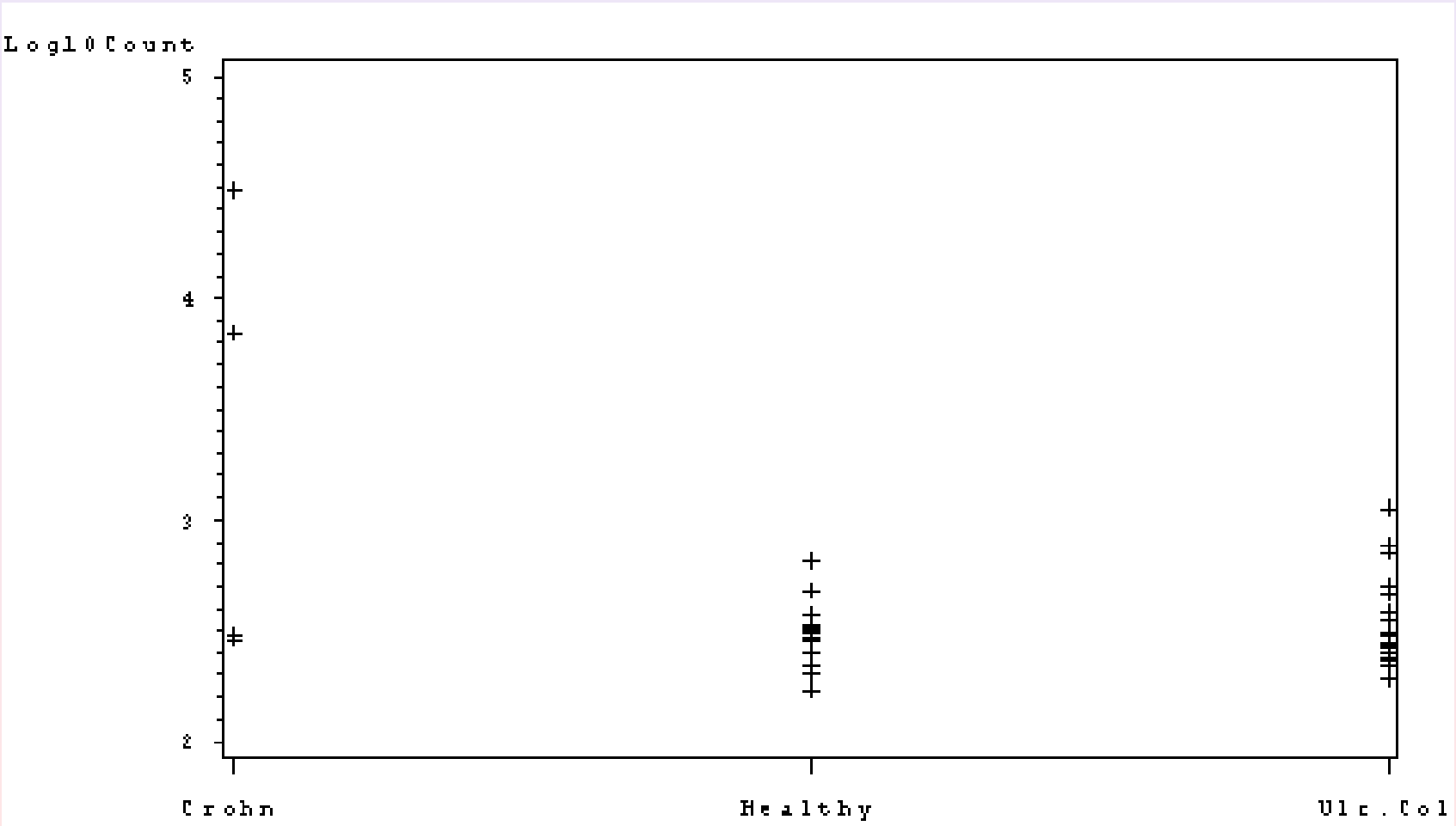
Example of a Species with a clear difference between Crohn and Healthy

The difference is statistically significant (FDR=5%) for all methods

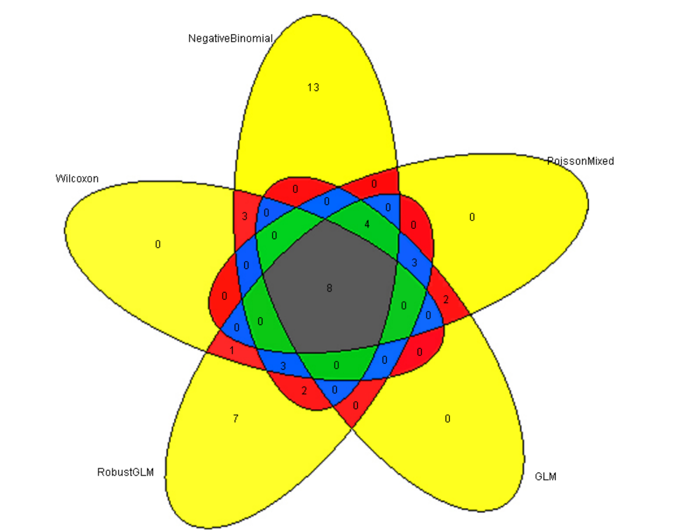


Example of a Species with a less clear difference between Crohn and Healthy

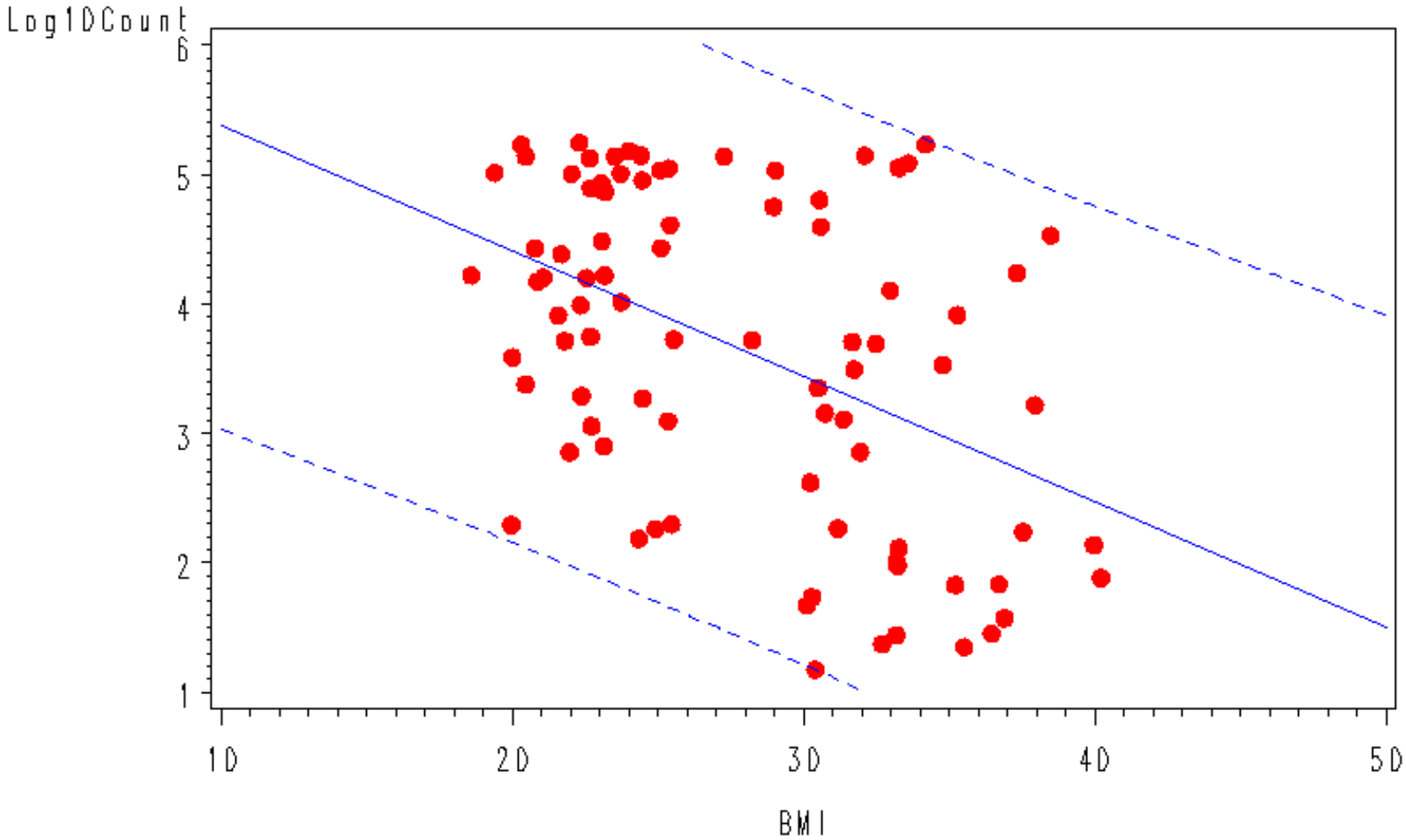
The difference is statistically significant (FDR=5%) for BN only



Comparison of the lists of the different methods(FDR=5%)



Other methods for comparative metagenomics: Regression



Other methods for comparative metagenomics: CCA

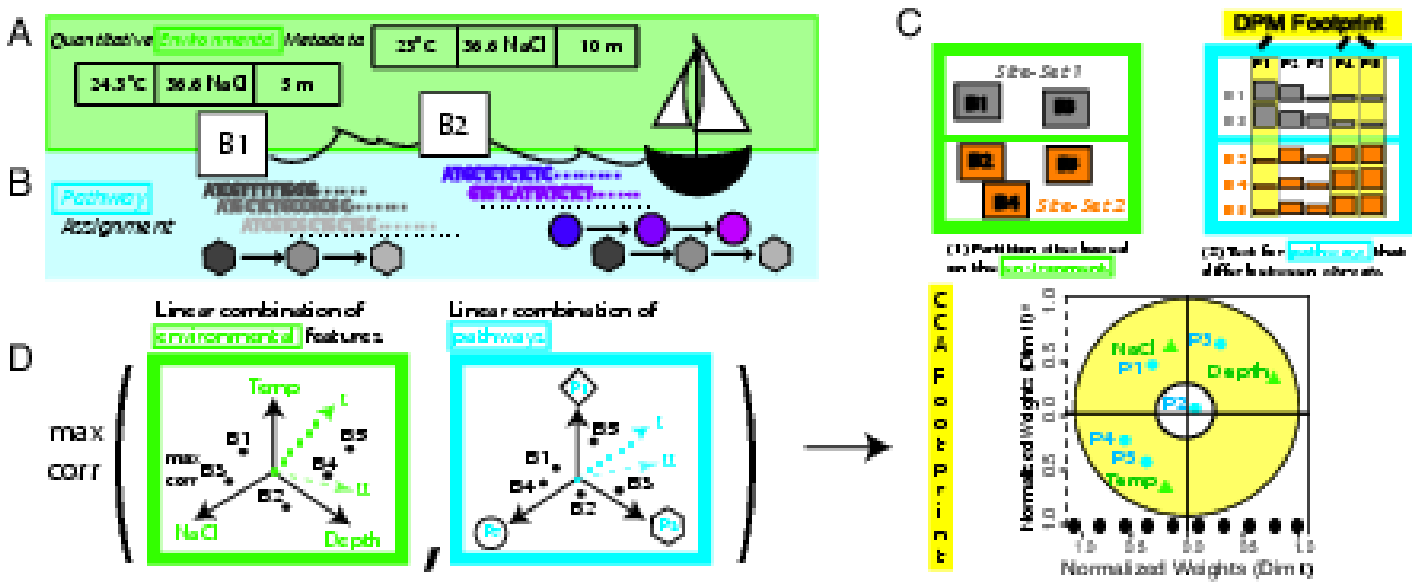


Fig. 1. Schematic representation of approach. The large squares labeled B1, B2, etc. represent the geographic sites (buckets). Each bucket has sequence and environmental feature data associated with it. (*A*) Mapping quantitative environmental features [salinity (ppt), sample depth (position in water column from which the sample was collected), water column depth (measured from surface to floor), and chlorophyll]. (*B*) Metagenome-derived metabolism at different levels of resolution (see *Materials and Methods*). Reads are color-coded according to their corresponding pathway elements (shapes). Different pathways are represented by different shapes (square, circle, etc.). All of the instances of a particular pathway are summed and normalized to compute the pathway score. (*C*) Schematic representation of DPM (see details in text). (*D*) Schematic representation of CCA (see details in text).

Regularized Canonical Correlation Analysis (Gianoulis et al. PNAS 2009)

Criticism against alignment methods

- Many (10% to 70%) reads cannot be aligned on any bank
- The banks are not reliable and change over time
- The results of alignments depends on the bank and the tuning (% identity...)
- The results of the the analysis of a metagenomic experiment depends on external factors...including time!

→ classification of reads in groups (unsupervised classification, binning) based on

- k-mer frequency of DNA sequences
- Distance based on an alignement score between reads

General view for long reads > 100bp

Method	Similarity	Type	Authors
TETRA	4-mers	binning	Teeling et al. (2004)
PhyloPythia	k-mers	classification	McHardy et al. (2007)
MEGAN	hits BLAST	clustering	Huson et al. (2007)
CARMA	Pfam	clustering	Krause et al. (2008)
S-GSOM	SOM	binning	Chan et al. (2008)
MG-DOTUR	BLAST	clustering	Schloss/Handelsman
LikelyBin	k-mers/MCMC	binning	Kislyuk et al. (2009)
Phymm	IMM	classification	Brady Salzberg (2009)
TACOA	k-NN	s-supervised	Diaz et al. (2009)

Table modified from G. Perriere (Univ. Lyon1)

Methods for short reads < 100 bp

- Correspondance Analysis (G. Perriere et al.)
- Mixture model for Markov Chain (E. lebarbier et al.)

Mixture model for Markov Chain

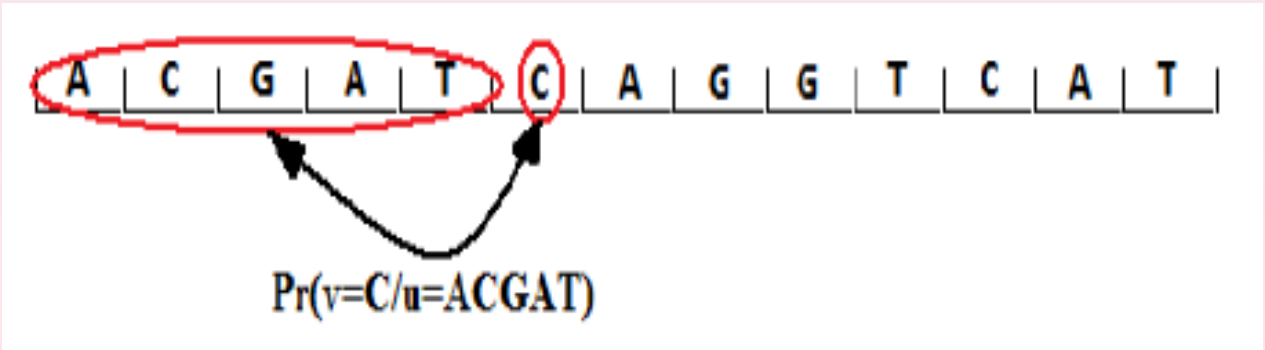
Motivation. Clustering the sequences according to their word composition.

Data. Sample of n sequences $\{S^1, \ldots, S^n\}$:

$$S^i = (S^i_1, \ldots, S^i_t, \ldots, S^i_{\ell_i}), \qquad \ell_i = \text{length of } S^i.$$

Markov chains. A sequence S is generated according to MC(ϕ) iff

$$\Pr\{S_t = s_t | S_{t-m} = s_{t-m}, \ldots, S_{t-1} = s_{t-1}\} = \phi(s_{t-m}, \ldots s_{t-1}; s_t).$$



Mixture model for Markov Chain

- m is the order of the Markov chain.
- $\phi(\cdot;\cdot)$ are the transition probabilities. They fit the frequencies of the $(m+1)$-mers:

$$\widehat{\phi}(\mathtt{at};\mathtt{c}) = \frac{\mathrm{N}(\mathtt{atc})}{\mathrm{N}(\mathtt{at})}.$$

Interpretation. Markov chains of order m account for the sequence contents in $(m+1)$-mers, e.g

- M0 is fitted to the nucleotide frequencies;
- M2 is fitted to the codon frequencies;
- M5 is fitted to the di-codon frequencies.

Mixture model for Markov Chain

Mixture of Markov chains.

- The n sequences are spread into K groups:

$$Z_{ik} = 1 \text{ if } i \in k, \quad 0 \text{ otherwise.}$$

- Each sequence i belongs to group k ($k = 1, \ldots K$) with probability π_k:

$$\Pr\{i \in k\} = \Pr\{Z_{ik} = 1\} = \pi_k$$

interpreted as the prior probability to belong to group k.

- Provided sequence i belongs to group k, it is generated according to a Markov chain with parameter ϕ_k:

$$(S^i | Z_{ik} = 1) \sim MC(\phi_k)$$

Statistical inference for Mixture model for Markov Chain

Each group k is characterized by

- π_k the proportion of sequences that belong to group k
- ϕ_k the transitions of the Markov Chain in this group

Mixture models are incomplete data models since we miss the group to which each sequence belongs.

E-M algorithm: provides maximum likelihood estimates.

- E-step: estimates the probability for each sequence to belong to each group:

$$\tau_{ik} = \Pr\{Z_{ik} = 1 | S^i\} = \frac{\pi_k \Pr(S^i \mid \phi_k)}{\sum_{k'} \pi_{k'} \Pr(S^i \mid \phi_{k'})}$$

interpreted as the posterior probability to belong to group k.

Statistical inference

- M-step: estimates the transition probabilities and the proportions

$$\widehat{\phi}_k(s_{t-m},\dots s_{t-1};s_t) = \frac{\widehat{\mathrm{N}}_k(s_{t-m},\dots s_{t-1}s_t)}{\widehat{\mathrm{N}}_k(s_{t-m},\dots s_{t-1})}\ ,\ \widehat{\pi}_k = \frac{\sum_{i=1}^n \widehat{\tau}_{ik}^{(h)}}{n}$$

where $\widehat{\mathrm{N}}_k(gca) = \sum_i \widehat{\tau}_{ik}\mathrm{N}_k(gca)$.

Classification. The $\widehat{\tau}_{ik}$ can be used to perform 'maximum a posteriori' (MAP) classification:

$$\mathrm{MAP}_i = \arg\max_k \widehat{\tau}_{ik}.$$

Choice of m and K

The order m of the Markov chain and the number of groups K have to be chosen in some way.

- The order m can be fixed according to biological considerations (see M0, M2, M5).
- The number of group K can be chosen by using criterion

$$\text{BIC(K)} = \log \widehat{\Pr}(\text{S}|\text{K}) - \frac{1}{2} \log(\#\text{ data}) \times (\#\text{ parameters}).$$

Some partial conclusions(1)

Metagenomics experiments may be classified by the number of unknown species (genes), US

- few US (less than 20%, mine, cheese, gut) → reads aligned on a reference genome → focus on comparative metagenomics
- many US (10 to 10000, soil, sea)
  - estimation of the number of species is a scientific question
  - impossible to build a reference genome → binning or clustering reads may be usefull

Some partial conclusions(2)

- Comparative Metagenomics: counts of reads may be analyzed using standard statistical tools after log-transformation, excepted for low counts (but...are low counts really interesting?)
- Many species (genes) → FDR, regularized methods.
- Biological variability is high → many biological replicates are necessary.
- Experimental design is an issue. No experiment with only one replicate should be published.

References

- J. C. Wooley, A. Godzik and I. Friedberg (2010) *A Primer on Metagenomics*, PLoS Computational Biology.
- MetaHIT Consortium (2010) *A human gut microbial gene catalogue established by metagenomic sequencing*, Nature in Press.
- Bunge, J. and Fitzpatrick, M.(1993) *Estimating the number of species: a review* JASA, 88, 421
- White, J.R., Nagarajan, N. and Pop, M.(2009) *Statistical methods for detecting Differentially Abundant features in clinical metagenomic samples* PLOS Computational Biology, Vol 5, issue 4
- Patrick D Schloss and Jo Handelsman (2008) *A statistical toolbox for metagenomics: assessing functional diversity in microbial communities*, BMC Bioinformatics.