High-dimensional compositional microbiota data: state-of-the-art of methods and software implementations

Perrine Soret, Marta Fernandez Avalos, Soon Cheng, Rodolphe Thiebaut

High-dimensional compositional microbiota data: state-of-the-art of methods and software implementations

GDR "Statistique et Santé", 2017, Bordeaux

Perrine Soret[1,2,3] Marta Avalos[1,2] Cheng Soon Ong [5] Rodolphe Thiébaut[1,2,3,4]

[1]Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219

[2]INRIA SISTM Team, F-33405 Talence, France

[3]Vaccine Research Institute (VRI), F-94000 Créteil, France

[4]CHU Bordeaux, Department of Public Health, F-33000 Bordeaux, France

[5]Data61, CSIRO, 7 London Circuit, Canberra ACT 2601, Australia

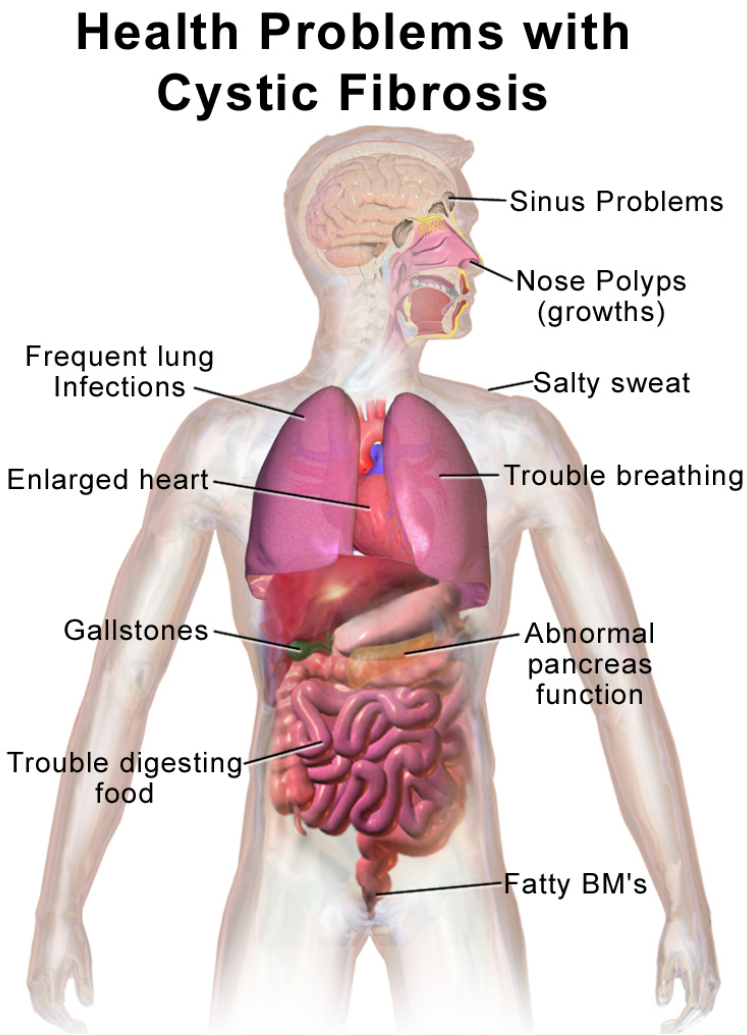
5 Octobre 2017

Human Microbiota

- Aggregate of microorganisms
- Include bacteria, archaea, protists, fungi and viruses
- Body = 90% bacterial and 10% human
- 10× more than human cells
- Microbiota, the key of our health
- Genetic code = Microbiome
- High-Throughput sequencing technologies

Cystic Fibrosis

- Genetic disorder that affects mostly the lungs
- Main symptoms: frequent chest infections and coughing or shortness of breath
- Life expectancy: around 40 years



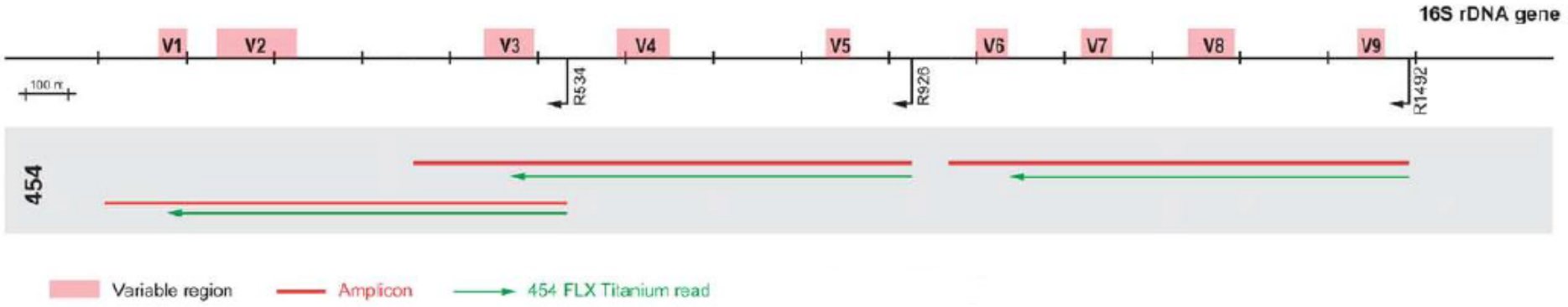
MucoFong cohort

- 3-year multi-center (Angers, Bordeaux, Dijon, Dunkerque, Grenoble, Lille and Rouen) prospective observational study from 2007 to 2010
- Supported by the national clinical research program "PHRC"
- 300 patients with following Criteria inclusion:
  - (i) Patient with a well-documented diagnosis of CF validated with the criteria in force (Rosenstein, 1998)
  - (ii) Belonging to one of the participating centers and older than six years
  - (iii) Undergoing a mycological analysis of sputum sample as part of either the annual microbiological checkup or the clinical management of an acute pulmonary exacerbation
  - (iv) Written informed consent form endorsed.
  - (v) Pulmonary transplant recipients were excluded at baseline.

Subsample at the inclusion:

- 37 sputum samples was sequencing using NGS
- 75 bacteria and 80 fungal
- Composition and structure of the CF airways microbiota and mycobiota

Sequencing and phylogenetic assignation



V3-V4
Targeted, amplifed
and sequenced

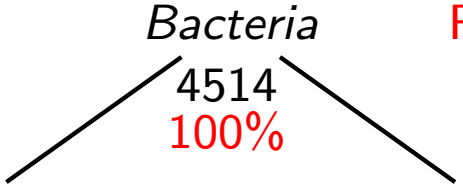
→ Clustering with
GreenGene base
(pyrosequence with 97%)

→ OTUs-based analysis

OTU: Operational Taxonomic Unit

Microbiota composition - Phylogenetic tree

Kingdom

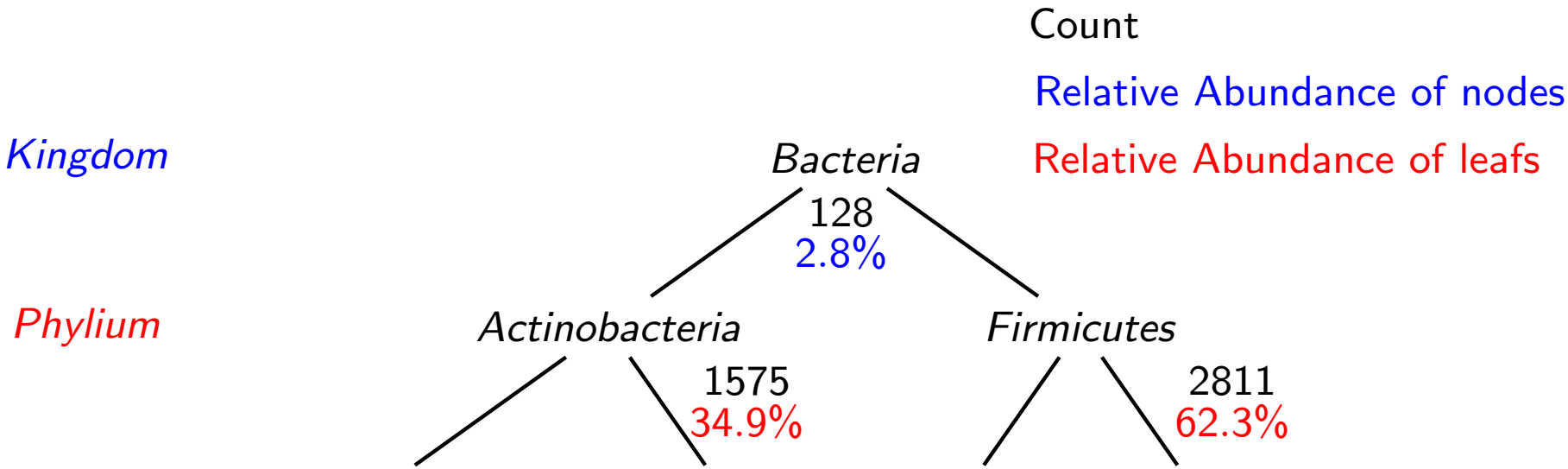


Count

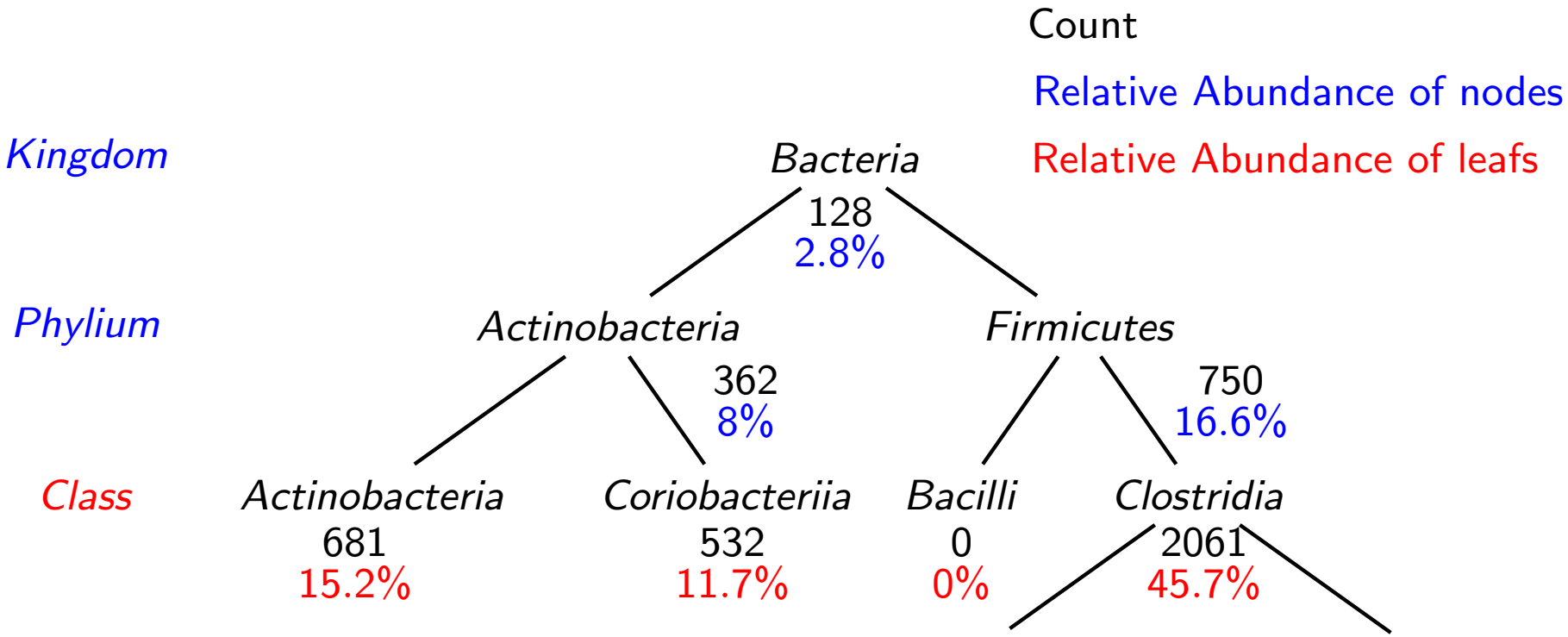
Relative Abundance of nodes

Relative Abundance of leafs

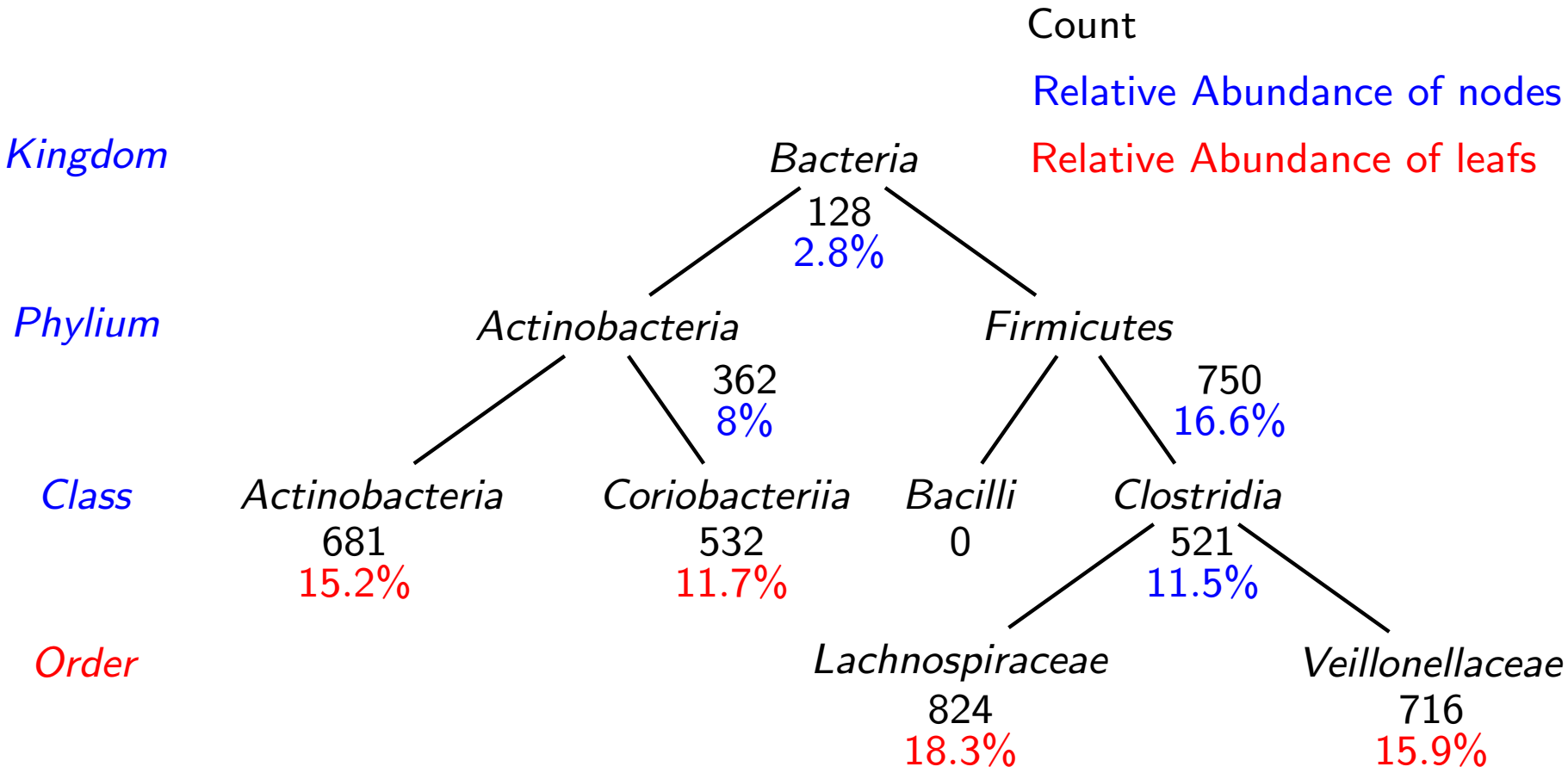
Microbiota composition - Phylogenetic tree



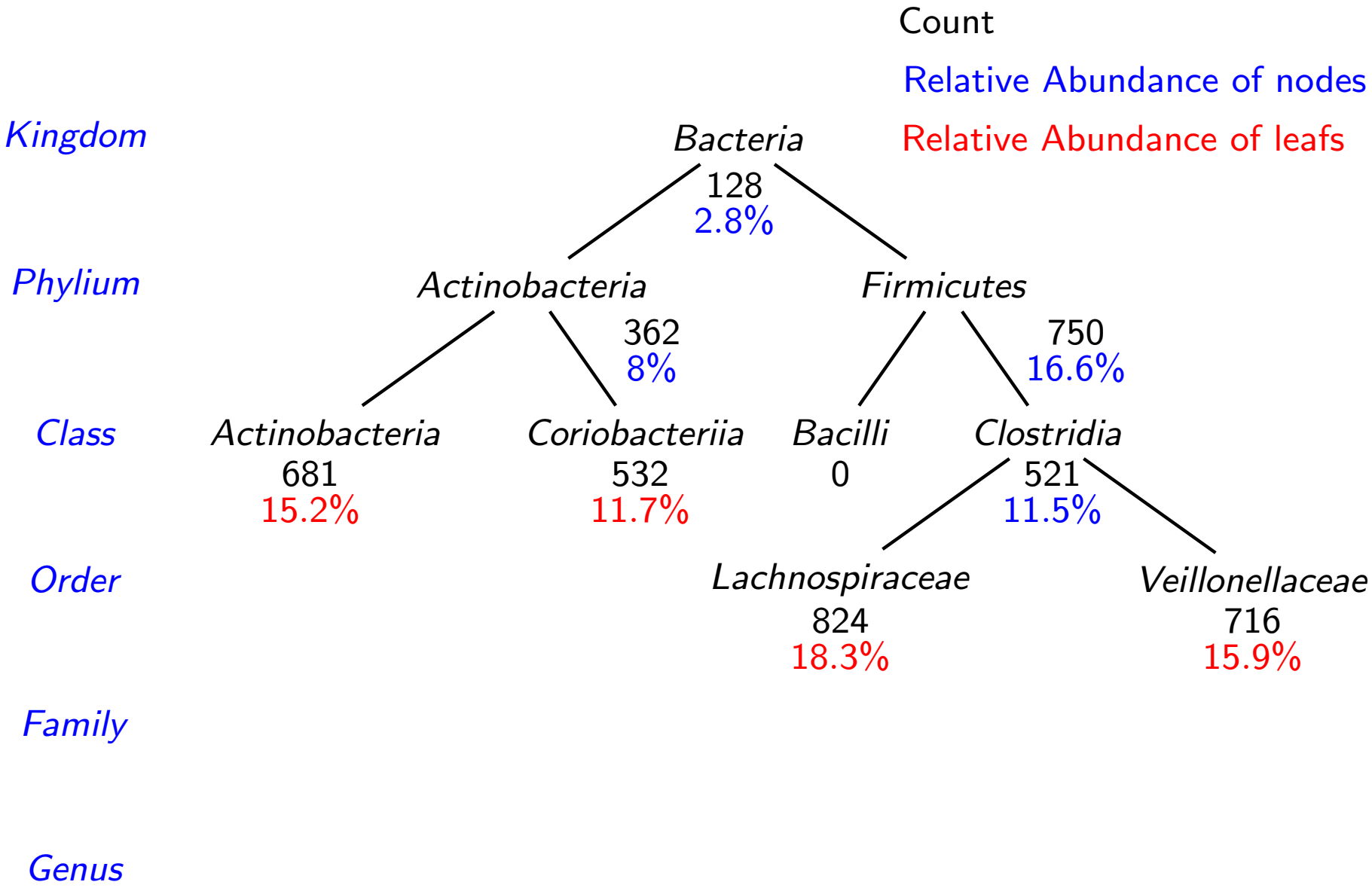
Microbiota composition - Phylogenetic tree



Microbiota composition - Phylogenetic tree



Microbiota composition - Phylogenetic tree

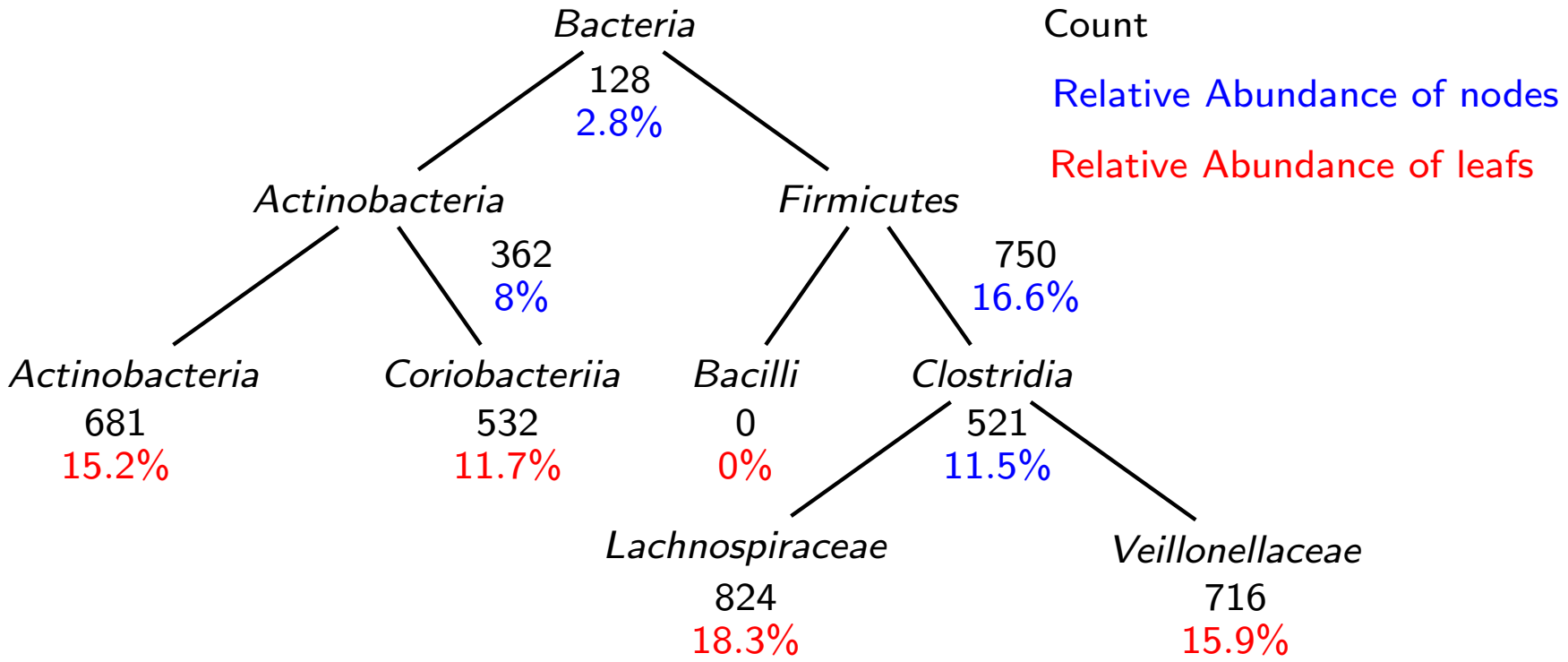


Basic concepts

- $\mathbf{z} = (z_1, \cdots, z_D)' \in \mathbb{R}_+^D \Rightarrow$ Count
- $\mathbf{x} = (x_1, \cdots, x_D)' \in \mathcal{S}^D \Rightarrow$ Relative abundance

$$\mathcal{C}(\mathbf{z}) = \left(\frac{z_1}{\sum_{i=1}^D z_i}, \cdots, \frac{z_D}{\sum_{i=1}^D z_i} \right) = (x_1, \cdots, x_D)'$$

- Simplex: $\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \cdots x_D)' : \; x_j > 0, \; j = 1, \cdots, D; \; \sum_{j=1}^D x_i = 1 \right\}$



Ternary diagram

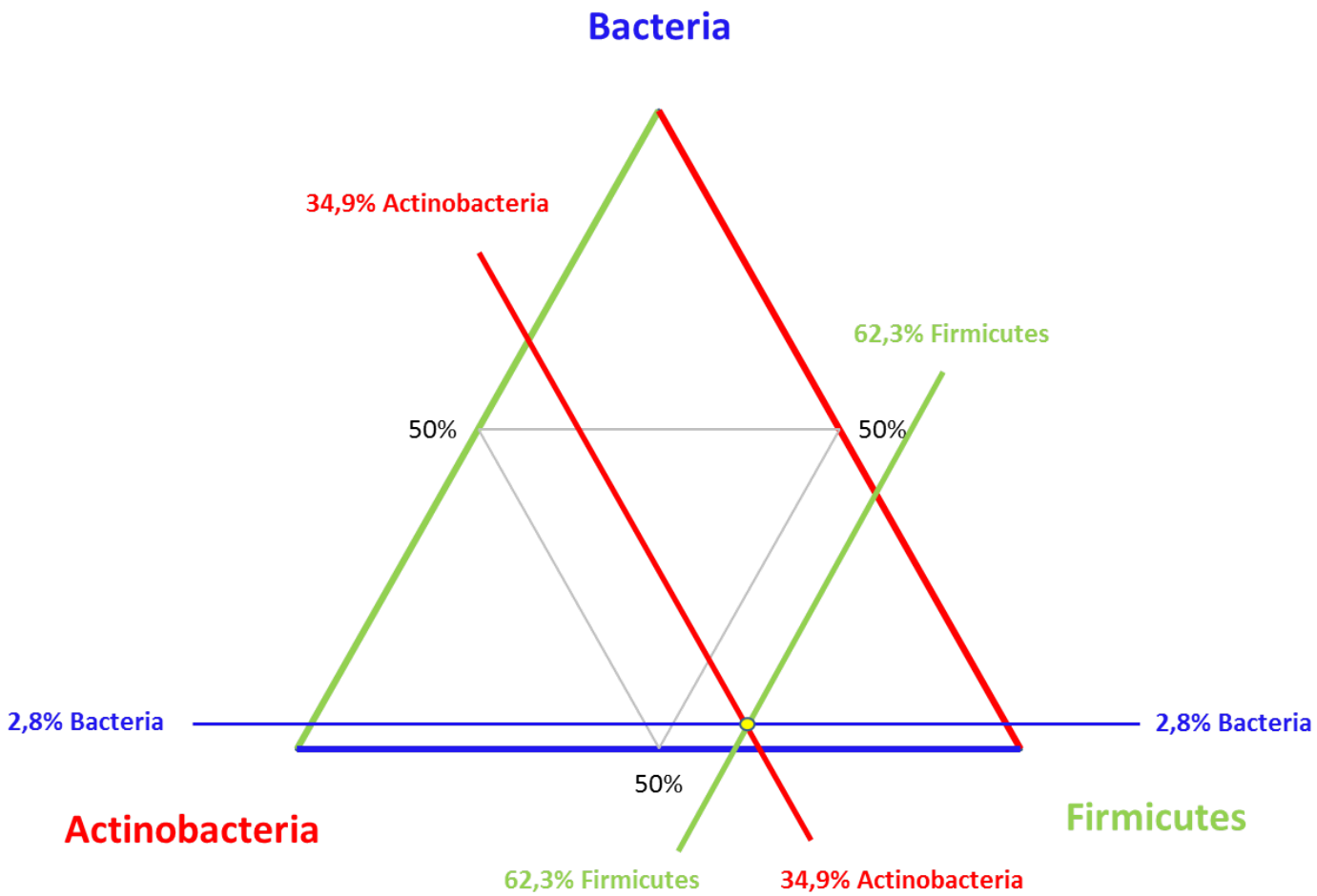


Figure 1: Ternary diagram with Bacteria, Actinobacteria and Firmicutes

Conditions to apply statistical methods

- Permutation-invariant
- Scale invariance
- Subcompositional coherence

- Log-ratio transformation
  - Additive log-ratio (alr) - Aitchison, 1986 [1]
  - Centered log-ratio (clr) - Aitchison, 1986 [1]
  - Isometric log-ratio (ilr) - Egozcue, 2003 [4]

Example: Compositional of 3 OTUs

$$[2.8,\ 34.9,\ 62.3] = [34.9,\ 62.3,\ 2.8] = [62.3,\ 2.8,\ 34.9]$$

Conditions to apply statistical methods

- Permutation-invariant
- Scale invariance
- Subcompositional coherence

- Log-ratio transformation
  - Additive log-ratio (alr) - Aitchison, 1986 [1]
  - Centered log-ratio (clr) - Aitchison, 1986 [1]
  - Isometric log-ratio (ilr) - Egozcue, 2003 [4]

Example: Compositional of 3 OTUs

	Children [Bact, Acti, Firmi]	Adult [Bact, Acti, Firmi]
Number of OTUs	$[53, 76, 14]$	$[28, 41, 8]$
Abundance relative (%)	$[37, 53, 10]$	$[37, 53, 10]$

Conditions to apply statistical methods

- Permutation-invariant
- Scale invariance
- Subcompositional coherence

- Log-ratio transformation
  - Additive log-ratio (alr) - Aitchison, 1986 [1]
  - Centered log-ratio (clr) - Aitchison, 1986 [1]
  - Isometric log-ratio (ilr) - Egozcue, 2003 [4]

Example: Compositional of 3 OTUs

	Children [Bact, Acti, Firmi]	Adult [Bact, Acti, Firmi]
Number of OTUs	$[66, 152, 28]$	$[56, 81, 16]$
Abundance relative (%)	$[37, 53, 10]$	$[37, 53, 10]$

Conditions to apply statistical methods

- Permutation-invariant
- Scale invariance
- Subcompositional coherence

- Log-ratio transformation
  - Additive log-ratio (alr) - Aitchison, 1986 [1]
  - Centered log-ratio (clr) - Aitchison, 1986 [1]
  - Isometric log-ratio (ilr) - Egozcue, 2003 [4]

Example: Compositional of 3 OTUs

We consider the composition of 2 patients P_1 and P_2 and $d(P_1, P_2)$ the distance between both compositions.

$$
\begin{aligned}
& d(P_1, P_2)\,[\text{Phyilum level}] \\
&= d(P_1, P_2)\,[\text{Class level}] \\
&= d(P_1, P_2)\,[\text{Order level}] \\
&= d(P_1, P_2)\,[\text{Family level}] \\
&= d(P_1, P_2)\,[\text{Genus level}]
\end{aligned}
$$

Objective of this study

Clinical goal: Analysis the association between micromycetes (bacterias or fungus) in the degradation of pulmonary function in patients with cystic fibrosis disease (Mucofong Data).

Statistical challenge:

$$\text{Compositional Data} + \text{High-dimension}$$

Objective of this study

Clinical goal: Analysis the association between micromycetes (bacterias or fungus) in the degradation of pulmonary function in patients with cystic fibrosis disease (Mucofong Data).

Statistical challenge:

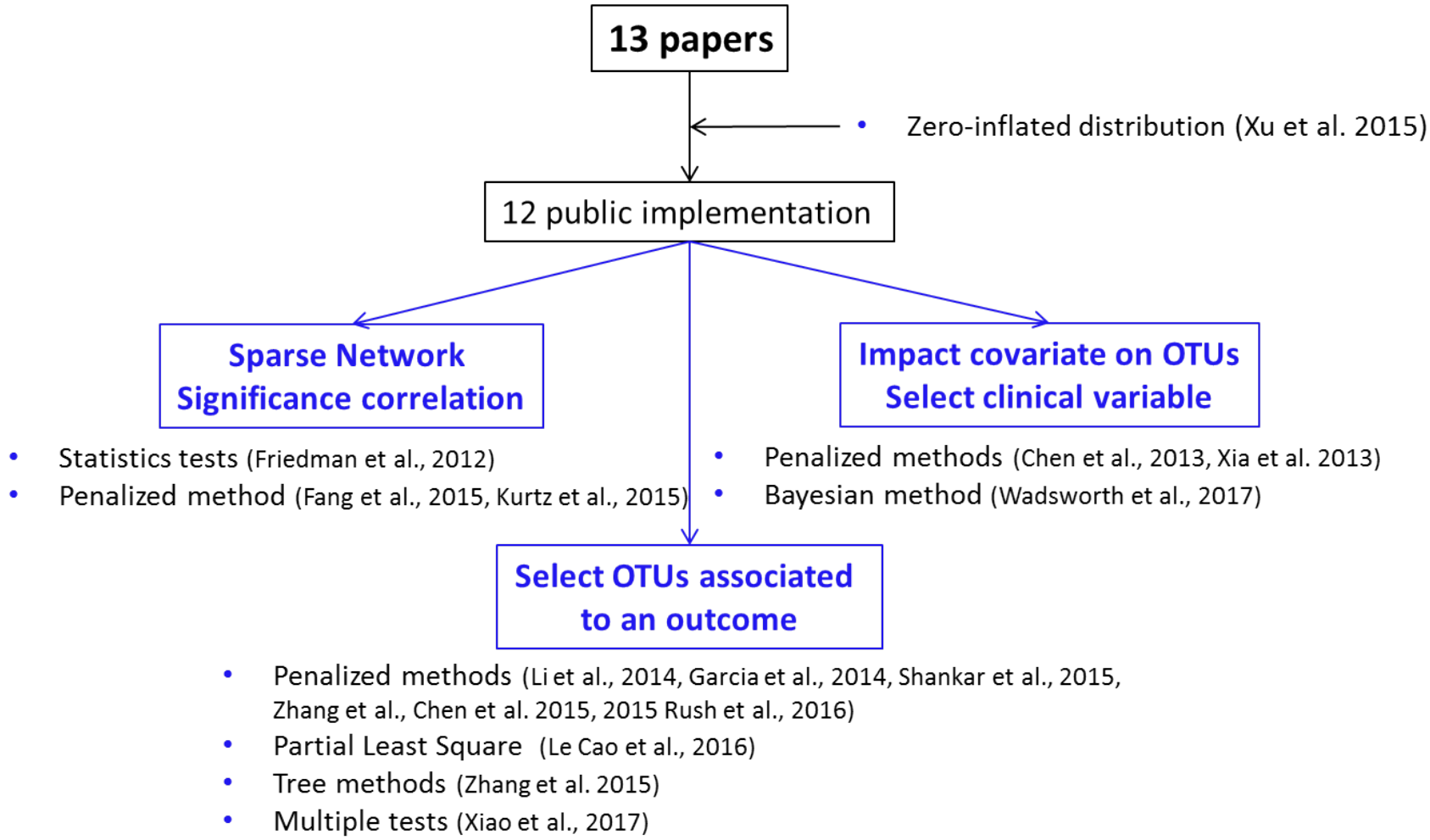
$$\text{Compositional Data} + \text{High-dimension}$$

In this presentation:

- Review of statistical methods for high-dimensional compositional data
- To compare existing methods adapted to our problem

(Not exhaustive) State-of-the-art:

Key-words in PubMed (Title/Abstract): ("High-dimensional" or "variable selection" or "OTU selection") and "microbiome"



Compared Methods

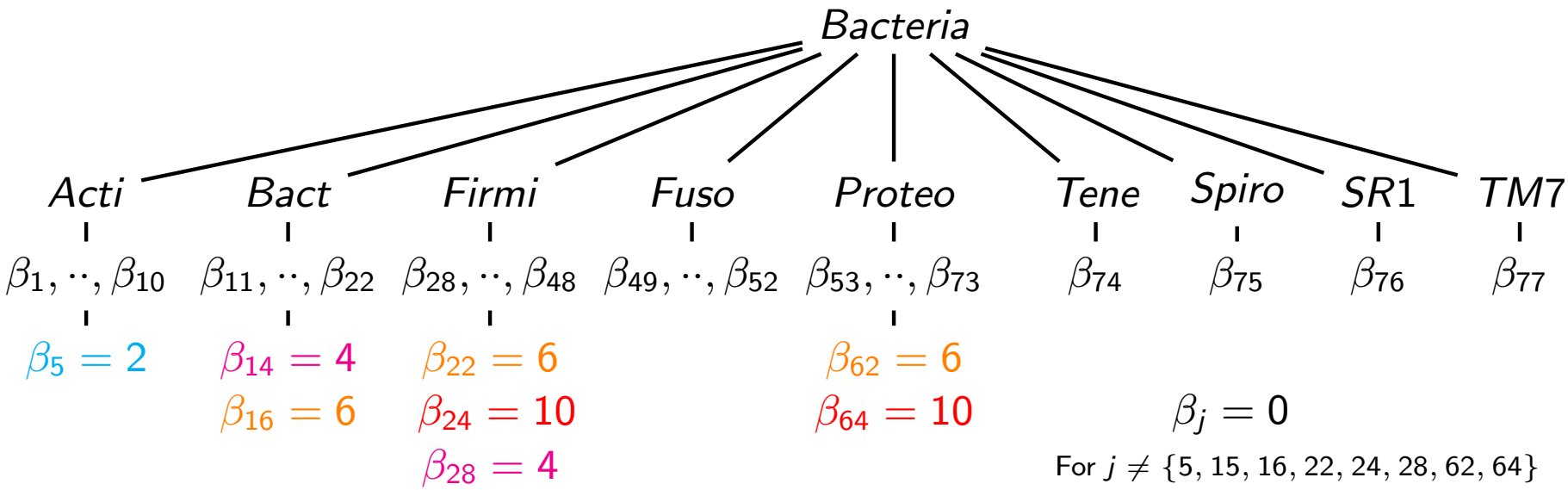
Article	Description of method
Log-Contrast (Lin et al., Biometrika, 2014 [9])	• $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|$ $\quad \sum_{j=1}^{p} \beta_j = 0$ Regression without hierarchical structure included
SGSL (Garcia et al., Bioinformatics, 2014 [6])	• $\mathbf{X} = \left(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(k)}, \cdots, \mathbf{X}^{(L)}\right)$ with $\mathbf{X}^{(k)} \in \mathcal{M}_{(n \times p_k)}$ • $\mathbf{X}^{(k)} = \left(\mathbf{X}^{(k,1)}, \cdots, \mathbf{X}^{(k,m)}, \cdots, \mathbf{X}^{(k,M_k)}\right)$ with $\mathbf{X}^{(k,m)} \in \mathcal{M}_{(n \times p_{k,m})}$ • $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \sum_{k=1}^{L} \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\| + \alpha_1\lambda \sum_{k=1}^{L} \sqrt{p_k}\|\boldsymbol{\beta}^{(k)}\|_2 +$ $\alpha_2\lambda \sum_{k=1}^{L}\sum_{m=1}^{M_k} \sqrt{p_{k,m}}\|\boldsymbol{\beta}^{(k,m)}\|_2 + (1-\alpha_1-\alpha_2)\lambda\|\boldsymbol{\beta}\|_1$ Regression with 3 hierarchical level included
Phy-Lasso (Rush et al., preprint, 2016 [11])	• $\mathbf{X} = \left(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(k)}, \cdots, \mathbf{X}^{(L)}\right)$ with $\mathbf{X}^{(k)} \in \mathcal{M}_{(n \times p_k)}$ • $t = 1, \cdots, T+1$ • $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \sum_{k=1}^{L} \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}\| - \sum_{t=1}^{T}\sum_{k=1}^{L} \alpha_k^t - \lambda\|\boldsymbol{\beta}\|_1$ Regression with hierarchical structure included
StructFDR (Xao et al., Bioinformatics, 2017 [15])	• H_{0j}: the jth OTU is not associated with $\mathbf{Y}$ • Hierarchical model • Permutation-based FDR control algorithm Multiple test with hierarchical structure included

Simulation data from MucoFong

- FEV_1 (Force Expired Volume) is generated from a Gaussian distribution
- 100 data sets simulated of size $n = 33$ from:

$$FEV_i = \beta_0 + \text{clr}\left(\mathbf{OTU}\right)_i \boldsymbol{\beta} + \varepsilon_i \qquad i = 1, \cdots, n \tag{1}$$

- $\mathbf{OTU}_{(n \times p)} \sim \mathcal{DM}(\boldsymbol{\pi}, \theta)$ where π and θ estimated by real data with *dirmult()* function
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{77})$
- $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)$



Tuning parameter selection and comparison criteria

- Data $\mathbf{D}$ is randomly chunked into K disjoint blocks
- $\mathbf{D}_k$ is the learning data, used to estimate coefficients
- $\mathbf{D}_{\setminus k}$ is the test data used to evaluate the loss function L

Cross-Validation : 10 fold-CV

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{K}\frac{1}{n_k}\sum_{i\in\mathbf{D}_k}\left(Y_i - \mathbf{X}_i\hat{\beta}(\lambda)_{\mathbf{D}_{\setminus k}}\right)^2, \tag{2}$$

Goal: Select OTUs associated to the outcome

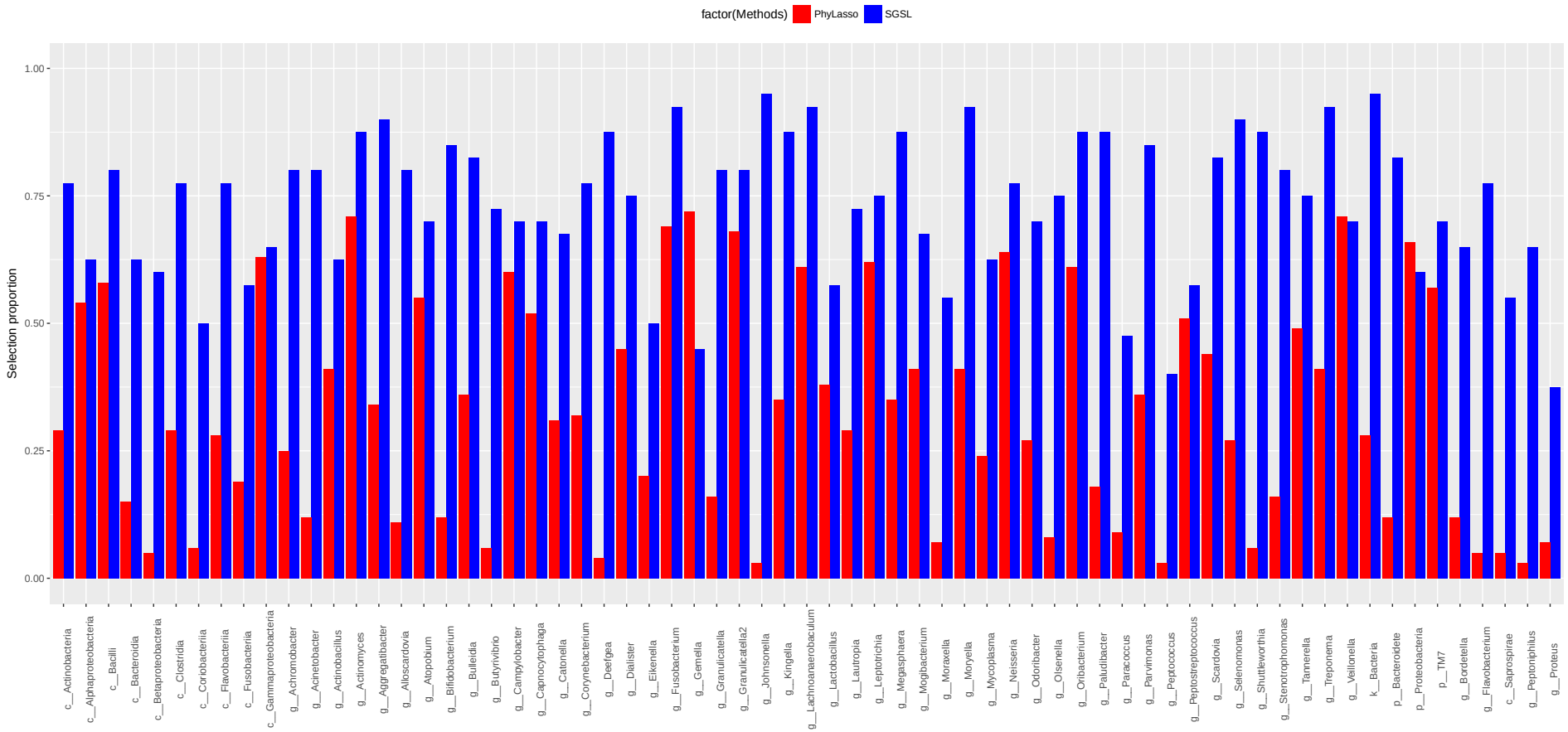
- 100 simulations ($S = 100$)
- Selection proportion (SP)

$$SP\left(\hat{\beta}_j\right) = \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}_{\hat{\beta}_{j,s}\neq 0} \quad \in \quad [0,1] \tag{3}$$

Selection proportion of significant OTU

OTUs	Mean relative abundance (%)	True β	LogContrast	SGSL	PhyLasso	StructFDR
OTU_{24}	0.5	10	0	0.35	0.23	0.01
OTU_{64}	13.1	10	0.48	0.90	0.92	0.34
OTU_{16}	4.9	6	0.04	0.90	0.83	0.01
OTU_{22}	1.8	6	0.01	0.70	0.62	0
OTU_{62}	3.4	6	0.01	0.77	0.77	0
OTU_{14}	1.9	4	0.01	0.70	0.65	0
OTU_{28}	8.4	4	0.04	0.67	0.86	0
OTU_{5}	4.7	2	0.01	0.92	0.78	0

Selection proportion of no significant OTU: PhyLasso vs SGSL



- Microbiota $\Rightarrow$ Topic in progress
- Compositional data $\Rightarrow$ Specific
- In high-dimensional $\Rightarrow$ Recent research

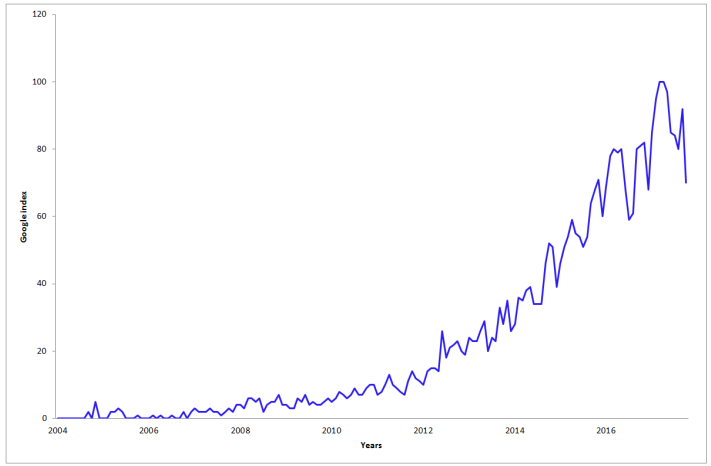


Figure 2: Trend for the key word "Microbiota" on Google Trends

In this presentation

- Review statistical method
- Comparison methods on simulation based on real data (MucoFong cohort) $\Rightarrow$ Phy-Lasso

Analysis of MucoFong Data

- *Characterization of the respiratory mycobiome and microbiome in cystic fibrosis: first evidence for inter-kingdom cross-talk during acute pulmonary exacerbation*, Soret et al. (in progress)

THANK YOU FOR YOUR ATTENTION

Références I

J. Aitchison.
The statistical analysis of compositional data.
1986.

J. Chen and H. Li.
Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis.
The annals of applied statistics, 7(1), 2013.

L. Chen, H. Liu, J.-P. A. Kocher, H. Li, and J. Chen.
glmgraph: an r package for variable selection and predictive modeling of structured genomic data.
Bioinformatics, 31(24):3991–3993, 2015.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barcelo-Vidal.
Isometric logratio transformations for compositional data analysis.
Mathematical Geology, 35(3):279–300, 2003.

J. Friedman, E. J. Alm, and C. von Mering.
Inferring correlation networks from genomic survey data.
PLoS Computational Biology, 8, 9 2012.

T. P. Garcia, S. Müller, R. J. Carroll, and R. L. Walzem.
Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data.
Bioinformatics, 30(6):831–837, 2014.

F. H., H. C., Z. H., and D. M.
Cclasso: correlation inference for compositional data through lasso.
Bioinformatics, 31(19):3172, 2015.

Références II

Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau.

Sparse and compositionally robust inference of microbial ecological networks.
PLOS Computational Biology, 11(5):1–25, 05 2015.

W. Lin, P. Shi, R. Feng, H. Li, et al.

Variable selection in regression with compositional covariates.
Biometrika, 101(4):785–797, 2014.

K.-A. Lê Cao, M.-E. Costello, V. A. Lakis, F. Bartolo, X.-Y. Chua, R. Brazeilles, and P. Rondeau.

Mixmc: A multivariate statistical framework to gain insight into microbial communities.
PLOS ONE, 11(8):1–21, 08 2016.

S. T. Rush, C. H. Lee, W. Mio, and P. T. Kim.

The phylogenetic lasso and the microbiome.
arXiv preprint arXiv:1607.08877, 2016.

J. Shankar, S. Szpakowski, N. V. Solis, S. Mounaud, H. Liu, L. Losada, W. C. Nierman, and S. G. Filler.

A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses.
BMC bioinformatics, 16(1):31, 1 Feb. 2015.
regeval repository:`http://github.com/openpencil/regeval`.

W. D. Wadsworth, R. Argiento, M. Guindani, J. Galloway-Pena, S. A. Shelburne, and M. Vannucci.

An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data.
BMC Bioinformatics, 18(1):94, Feb 2017.

Références III

F. Xia, J. Chen, W. K. Fung, and H. Li.
A logistic normal multinomial regression model for microbiome compositional data analysis.
Biometrics, 69(4):1053–1063, 2013.

J. Xiao, H. Cao, and J. Chen.
False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing.
Bioinformatics, 2017.

L. Xu, A. D. Paterson, W. Turpin, and W. Xu.
Assessment and selection of competing models for zero-inflated microbiome data.
PLOS ONE, 10(7):1–30, 07 2015.

Q. Zhang, H. Abel, A. Wells, P. Lenzini, F. Gomez, M. A. Province, A. A. Templeton, G. M. Weinstock, N. H. Salzman, and I. B. Borecki.
Selection of models for the analysis of risk-factor trees: leveraging biological knowledge to mine large sets of risk factors with application to microbiome data.
Bioinformatics, 31(10):1607–1613, 2015.