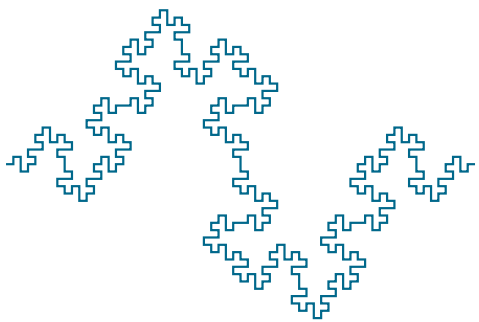
High-Throughput Sequencing Course

Microbiome Data Analysis and Compositional Data

Biostatistics and Bioinformatics



Summer 2019

Duke University
School of Medicine

DUKE BIOSTATISTICS & BIOINFORMATICS

Section 1

Introduction

Human Microbiomial Community

- Complex microbial community (microbiome): prokaryotes (bacteria), archaea, fungi, and viruses.
- Number of microbial cells: about 10 times the total number of human cells.
- Microorganisms rarely live alone; they function as integrated communities.
- The collective genomes of these microbes constitue an extended human genome that encodes genetic and metabolic capabilities that humans do not inherently possess.

Microbiome: An Important Contributor to Human Health

- The composition of the microbiome varies based on diet, health, and environment.
- New evidence show microbiome may play a role in complex diseases, including obesity, cardiovascular diseases, and type II diabetes, etc.
- Commensal bacteria can control the response of cancer to therapy by modulating the tumor microenvironment.
- Gut microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders via circulating metabolites.

Section 2

Taxa Composition Sequencing

Sequencing Strategy

- The 16 S ribosomal RNA (rRNA) gene: ubiquitous in the bacterial kingdom and contains highly variable regions. Each bacterial cell is assumed to have the same number of copies of this gene.
  1. isolate form all bacteria the DNA strands corresponding to some variable region of the gene.
  2. count the different versions of the sequences.
  3. call the bacteria to which the versions correspond.

Pros of Taxa Composition Sequencing

- This approach can reveal the phylogenetic structure of a microbial community, which is very helpful for downstream analysis.
- rRNA makes up 80% of total bacterial RNA, this approach allows for detecting of rare members with high sensitivity.

Phylogenetic Tree

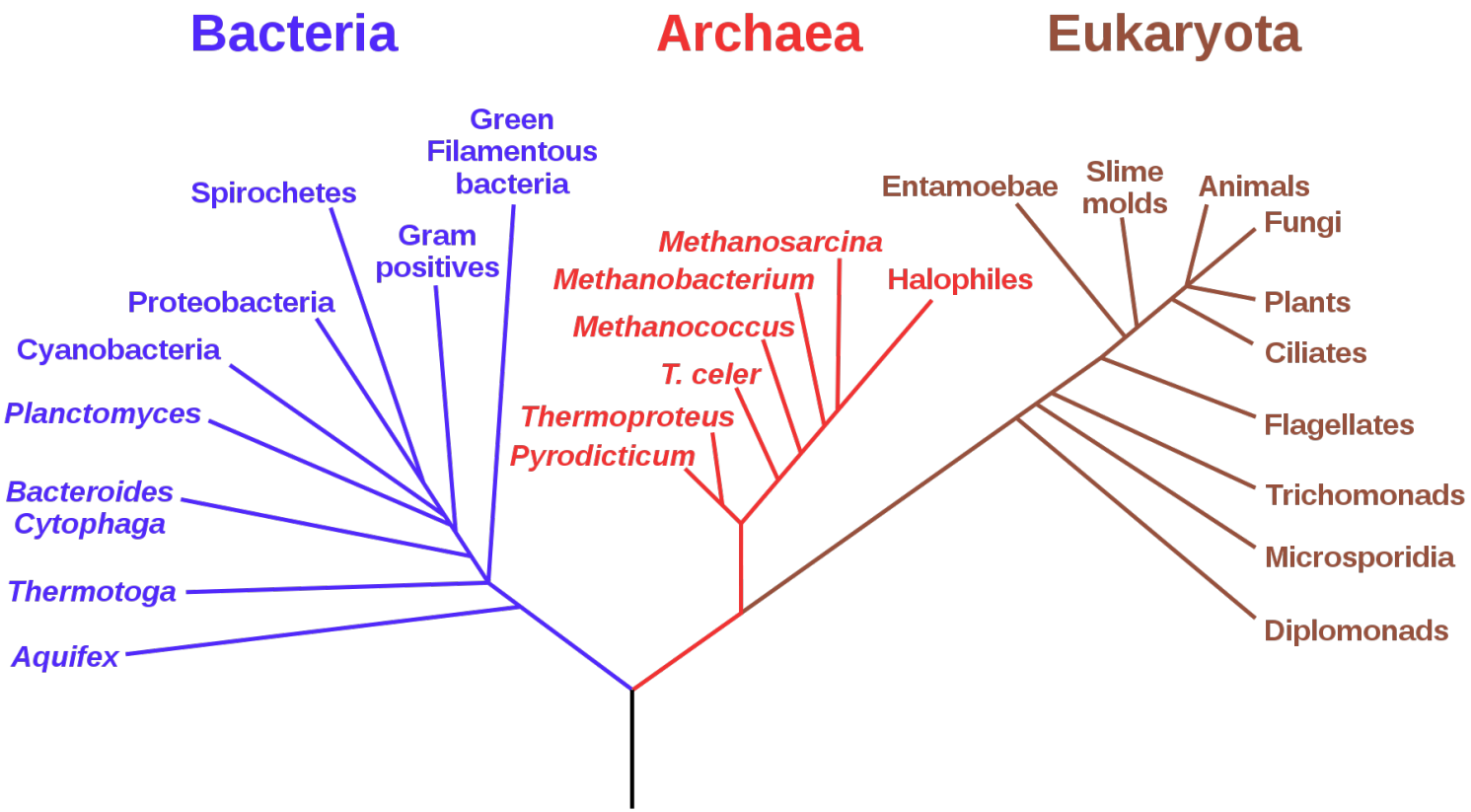


Figure: A speculatively rooted tree for rRNA genes, showing the three life domains: bacteria, archaea, and eukaryota. The black trunk at the bottom of the tree links the three branches of living organisms to the last universal common ancestor. Cited from https://en.wikipedia.org/wiki/Phylogenetic_tree.

Cons of Taxa Composition Sequencing

- 16S data do not provide any information about bacterial gene inventory and functionality.
- 16S data do not provide high sensitivity in identifying bacteria at the species for strain level.

Summarizing to Compositional Data

- Two approaches
  - mapping to an existing phylogenetic tree
  - clustering into operational taxonomic units (OTUs) at a certain similarity level in a taxonomic-independent way.

Mapping to an existing phylogenetic tree

Given a reference phylogenetic tree, we can use software Pplacer to get the compositional data.
https://github.com/matsen/pplacer

- uses a linear time maximum likelihood and Bayesian phylogenetic placement to assign each read to an edge of the tree.
- calculates the posterior probability of a read placement on an edge so that it can quantify uncertainty on an edge-by-edge basis.
- output: a file of counts of reads for each of the interval edges.

Clustering into OTUs

16S rRNA sequence reads can be clustered into OTUs at a certain (say 97%) similarity level based on the pairwise Hamming distances.

- These OTUs can be used to approximate the taxonomic rank species, although they do not precisely represent bacterial species.
- Each OTU is characterized by a representative DNA sequence and can be assigned a taxonomic lineage by comparing it with a known bacterial 16S rRNA database.
- We can further aggregate OTUs from the same genus and analyze the abundances at the level of genus, families, order, classes, or phyla; more robust
- Most OTUs are extremely low in abundance, and a large proportion are found only once (possibly as a result of sequencing error).

Combo Data Set Example

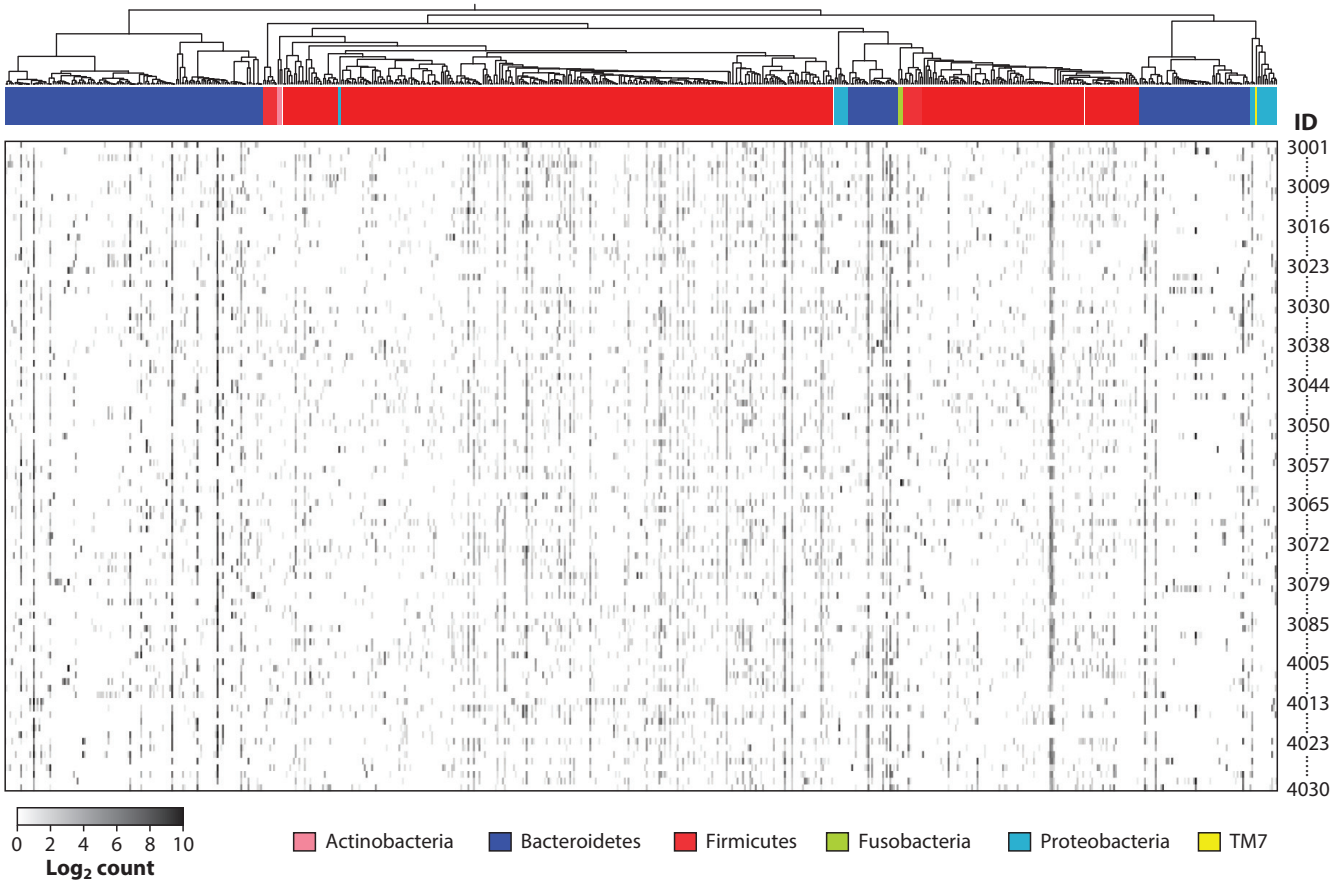


Figure: OTU abundances for the gut microbiome for the COMBO data set (Wu et al., 2011). Rows represent individuals, column represent OTUs, and each entry is the base-2 logarithm of the observed read counts. The phylogenetic tree is plotted on the top of the figure. The plot is cited from Li (2015).

Section 3

Shotgun Metagenomic Sequencing

Sequencing Strategy

- DNA extracted from samples.
- DNA randomly shared into smaller fragments.
- The fragments are sequenced to get reads.
- Assembly (reference-based assembly or *de-novo* assembly), binning, annotation, etc. Reference: Torsten Thomas, Jack Gilbert, and Folker Meyer (Feb. 2012). "Metagenomics - a guide from sampling to data analysis". In: *Microb Inform Exp* 2.1, p. 3. DOI: 10.1186/2042-5783-2-3

Pros and Cons of Shotgun Metagenomic Sequencing

Pros:

- The data provide functional and biological process-level characterization of microbial communities.
- The data also allow the reconstruction of draft genome sequences for individual community members.
- This approach makes possible the detection of new species and new genes.

Cons:

- To achieve the same level of sensitivity in detecting rare taxa as 16S sequencing, much deeper sequencing is required.
- Rich data impose challenges in computation and statistical analysis.

Alignment to Complete Genomic Sequences

Florent E Angly et al. (Dec. 2009). "The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes". In: *PLoS Comput Biol* 5.12, e1000593. DOI: 10.1371/journal.pcbi.1000593

- Score: $F_{ij} = m_i t_j 2^{-S_{ij}}$
- m_i is the effective query sequence length
- t_j is the effective length of the target genome j
- S_{ij} is the high-scoring segment pair (HSP) bit score.
- Converting to weights: $w_{ij} = (1/F_{ij})/(\sum_j 1/F_{ij})$.
- Final relative abundance for species j is $X_j = (W_j/t_j)/(\sum_j W_j/t_j)$.

Alignment to Marker Genes

- MetaPhlAn (Metagenomic phylogenetic analysis): uses clade-specific marker genes to unambiguously assign reads to microbial clades and to quantify species abundances based on reads aligned to these marker genes.
- GSMer: identifies genome-specific k-mer from currently sequenced microbial genomes, and the resulting k-mers are then used for strain- or species-level identification in metagenomes.
- Both assume the reads are uniformly distributed across different marker genes.
- Modified method described in Li (2015).

Composition-Based Approach

- Clustering based on k-mers, where k is usually small.
- Rationale: k-mer frequencies reflect organism-specific characteristics.
- Read s can be mapped to a high dimensional space of nucleotide patterns $o = \{o_1, \ldots, o_P\}$, where each o_i is defined by its pattern length k and the number of literals l. In this space, read s is represented by the compositional vector $\mathbf{a} = \{a_1, \ldots, a_p\}$, where a_i is the frequency of pattern o_i in s.

Section 4

Divergence

Phylogenetic Tree Terminologies

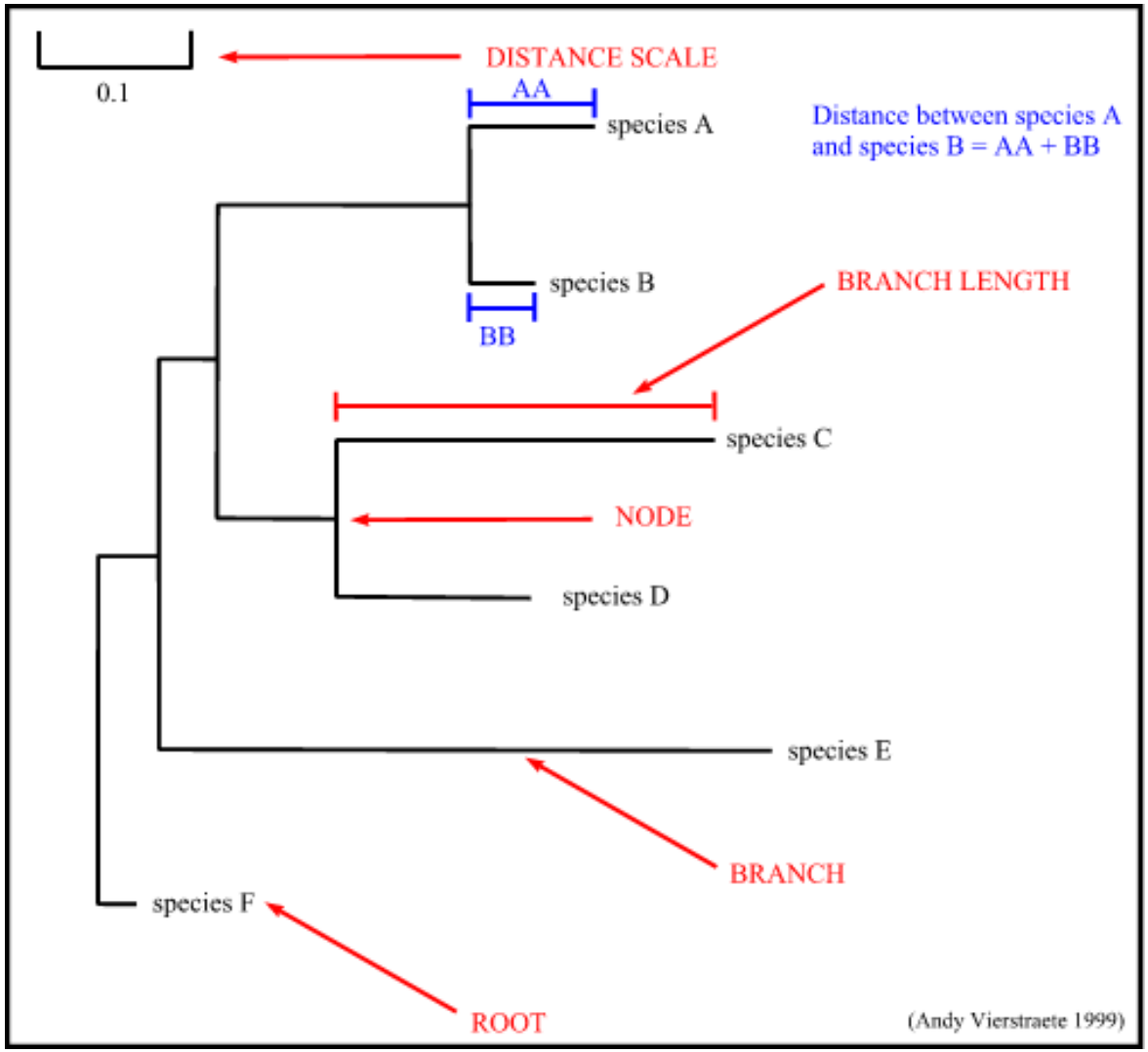


Figure: The tree terminology

Phylogenetic α-Diversity

Phylogenetic Diversity (one-parameter definition)

$$PD = \sum_i l_i g_\theta(D(i))$$

- l_i is the branch length of the ith edge
- $D(i)$ is the fraction of reads in the sample that are in the leaves on the distal side (*i.e.*, away from the root of the tree) of edge i
- $g_\theta(x) = [2\min(x, 1-x)]^\theta$.
- $\theta = 0.25$ or $\theta = 0.5$ ofter gave a better predictor of dysbiosis than did using $\theta = 0$ or $\theta = 1$.

Bray-Curtis Distance

One community-level measure of the distance between two microbiome samples A and B

$$d_{AB} = \sum_{j=1}^{p} |n_{Aj} - n_{Bj}|/(n_{A+} + n_{B+})$$

- n_{Aj} (n_{Bj}) is the count of taxa j in sample A $((B))$ and n_{A+} (n_{B+}) is the total count of all taxa.

Unweighted UniFrac Distance

$$d^U = \sum_{i=1}^{n} \frac{l_i |I(p_i^A > 0) - I(p_i^B > 0)|}{\sum_{i=1}^{n} l_i}.$$

- $I(\cdot)$ is the indicator function (= 1 is the condition is true, and = 0 is condition is false)
- l_i is the branch length of edge i
- p_i^A (p_i^B) is the taxa proportions descending from branch i for community A (community B).
- Incorporating phylogenetic tree information.

Weighted UniFrac Distance

$$d^W = \frac{\sum_{i=1}^{n} l_i |p_i^A - p_i^B|}{\sum_{i=1}^{n} l_i (p_i^A + p_i^B)} = \frac{\sum_{i=1}^{n} l_i (p_i^A + p_i^B) \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} l_i (p_i^A + p_i^B)}.$$

- Incorporating both phylogenetic tree information and taxa abundances

One-parameter generalized UniFrac (gUniFrac) distance

$$d^{(\theta)} = \frac{\sum_{i=1}^{n} l_i (p_i^A + p_i^B)^\theta \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{n} l_i (p_i^A + p_i^B)^\theta}.$$

- $\theta \in [0, 1]$ controls the contribution from high-abundance branches.
- Chen et al. (2012) showed that using $\theta = 0.25$ or $\theta = 0.5$ usually provides better power in detecting community compositional differences than does using other values of θ.

Section 5

Regression Modeling for Microbiome Data

Multinomial Distribution

Multinomial distribution:

$$f_{\mathrm{M}}(n_1, \ldots, n_p; \phi) = \binom{n_+}{\boldsymbol{n}} \prod_{j=1}^{p} \phi_j^{n_j}.$$

- $\boldsymbol{n} = (n_1, \ldots, n_p)'$ is the taxa read counts.
- $n_+ = \sum_{j=1}^{p} n_j$.
- $(\phi) = (\phi_1, \ldots, \phi_p)'$ are underlying species proportions for which $\sum_{j=1}^{p} \phi_j$.
- Remark: The observed variation from the heterogeneity of the microbiome samples is usually large than the variation modeled by the multinomial distribution.

Dirichlet-Multinomial (DM) Distribution

$$f_{\mathrm{DM}}\left(n_1, n_2, \ldots, n_p; \boldsymbol{\gamma}\right) = \binom{n_+}{\mathbf{y}} \frac{\Gamma\left(n_+ + 1\right)\Gamma\left(\gamma_+\right)}{\Gamma\left(n_+ + \gamma_+\right)} \prod_{j=1}^{p} \frac{\Gamma\left(n_j + \gamma_j\right)}{\Gamma\left(\gamma_j\right)\Gamma\left(n_j + 1\right)}$$

- This assumes $(\phi_1, \ldots, \phi_p)$ follows a Dirichlet distribution $\mathrm{Dir}(\gamma_1, \ldots, \gamma_p)$.
- $\gamma_+ = \sum_{j=1}^{p} \phi_j$.

Regression based on DM Distribution

Research Question: Are the taxa proportions associated with covariates?

- Covariates $\boldsymbol{Z} = (Z_1, \ldots, Z_q)'$, such as gender, treatment, disease status, etc..
- log-linear model

$$\gamma_j(\boldsymbol{Z}) = \exp\left(\alpha_j + \sum_{k=1}^{q} \beta_{jk} z_k\right).$$

- β_{jk} measures the effect on the jth taxon of the kth covariate.
- MLE works for small number of p (the number of taxa) and q (the number of covariates).

High Dimensional Regression Model

$$Y = \boldsymbol{Z}^p \boldsymbol{\beta}_{\setminus p} + \epsilon.$$

- $\boldsymbol{Z}^p = \{\log(x_{ij}/x_{ip})\} \in \mathbb{R}^{n\times(p-1)}$ for which the pth taxon serves as a reference. x_{ij} is the taxa proportion of subject i in taxon j. Note that $\sum_{j=1}^{p} x_{ij} = 1$.
- $\boldsymbol{\beta}_{\setminus p} = (\beta_1, \ldots, \beta_p)'$ is a regression coefficient
- ϵ is an independent noise term distributed as $N(0, \sigma^2)$.

Lasso on High Dimensional Regression Model

$$\hat{\boldsymbol{\beta}}_{\setminus p} = \arg\min_{\boldsymbol{\beta}_{\setminus p}} \left(\frac{1}{2n} \|y - Z\boldsymbol{\beta}_{\setminus p}\|_2^2 + \lambda \|\boldsymbol{\beta}_{\setminus p}\|_1 \right).$$

- Lasso (ℓ_1) penalty shrink some coordinates of $\boldsymbol{\beta}_{\setminus p}$ to be zero.
- λ is the tuning parameter for controlling the proportion of zeros.
- Flaw: the shrinkage is NOT invariant w.r.t. the reference taxon.

Constrained Regression Model

$$Y = \boldsymbol{Z}\boldsymbol{\beta} + \epsilon, \quad \mathbf{1}'\boldsymbol{\beta} = 0.$$

- $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^p$.
- $\boldsymbol{Z} = (Z_1, \ldots, Z_p)' = (\log x_{ij}) \in \mathbb{R}^{n \times p}$, where n is the number of subjects, p is the number of taxa.
- Constrained convex optimization:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|y - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right) \quad \text{subject to } \mathbf{1}'\boldsymbol{\beta} = 0$$

Section 6

References

- Angly, Florent E et al. (Dec. 2009). "The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes". In: *PLoS Comput Biol* 5.12, e1000593. DOI: 10.1371/journal.pcbi.1000593.
- Chen, Jun et al. (Aug. 2012). "Associating microbiome composition with environmental covariates using generalized UniFrac distances". In: *Bioinformatics* 28.16, pp. 2106–13. DOI: 10.1093/bioinformatics/bts342.
- Li, Hongzhe (Apr. 2015). "Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis". In: *Annual Review of Statistics and Its Application* 2.1, pp. 73–94. ISSN: 2326-831X. DOI: 10.1146/annurev-statistics-010814-020351.
- Thomas, Torsten, Jack Gilbert, and Folker Meyer (Feb. 2012). "Metagenomics - a guide from sampling to data analysis". In: *Microb Inform Exp* 2.1, p. 3. DOI: 10.1186/2042-5783-2-3.


Wu, Gary D et al. (Oct. 2011). "Linking long-term dietary patterns with gut microbial enterotypes". In: *Science* 334.6052, pp. 105–8. DOI: 10.1126/science.1208344.