Statistical challenges in analyzing 16S microbiome data

An application to the identification of microbe-regulated pathways in allergy and auto-immunity

Marine Jeanmougin

Institut Curie, U932 - Immunity and cancer

Journées MAS, August 28th, 2014

Outline

1. Introduction
   - The MAARS project
   - Microbiome data production and features

2. Normalisation of 16S data
   - Motivations
   - State of the art
   - Evaluation of current methods

3. Preliminary results
   - Exploratory analysis
   - Integration of microbiome and transcriptome data

Outline

1 Introduction

- The MAARS project
- Microbiome data production and features

2 Normalisation of 16S data

- Motivations
- State of the art
- Evaluation of current methods

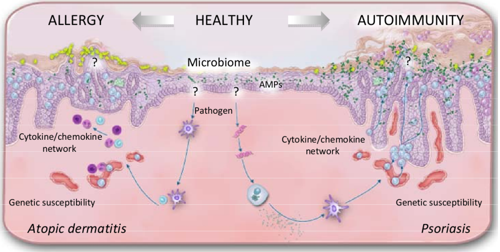
3 Preliminary results

- Exploratory analysis
- Integration of microbiome and transcriptome data

The MAARS Project

Goal

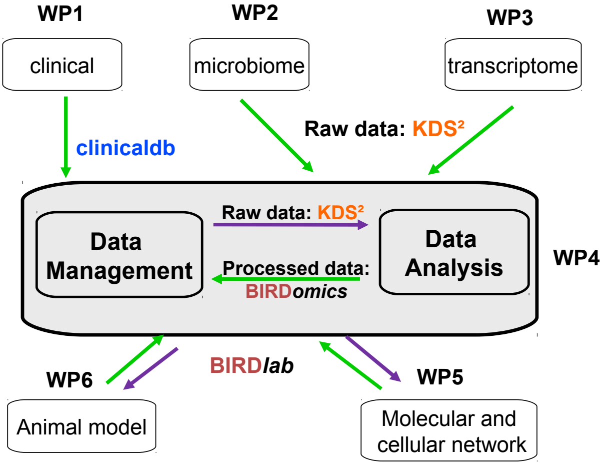
⤳ Unravel the **inflammatory pathways** during the **host-pathogen interactions** which may trigger allergic or autoimmune inflammation



Clinical impact

⤳ Identify key microbes and molecular targets to develop **novel intervention strategies**

WP4: data management and analysis



King's college
- Sophia Tsoka
- Gareth Muirhead

FIOH
- Dario Greco

Karolinska
- Juha Kere
- Shintaro Katayama

Fios Genomics
- Varrie Ogilvie
- Sarah Lynagh
- Max Bylesjo

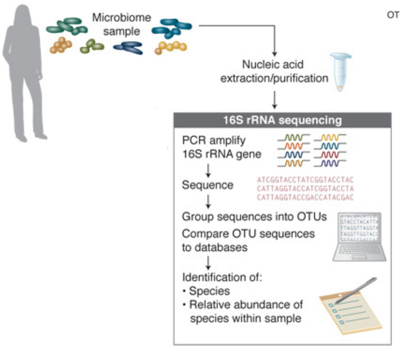
Institut Curie
- Vassili Soumelis
- Philippe Hupé
- Gerome Jules-Clément
- Marine Jeanmougin

Knowledge and **D**ata **S**haring **S**ystem
Clinical **D**ata**B**ase
BIological **R**esult **D**atabase

16S data production

The skin microbiome

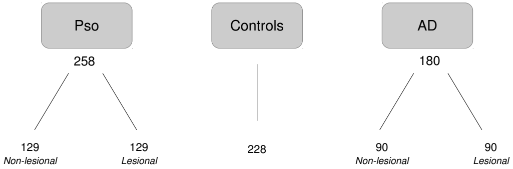
- Ecosystem of microbes that live on the skin
- Culture independent microbiome research:
  - total microbiome DNA sequencing
  - 16S rRNA sequencing



OTU = Operational Taxonomic Unit, a group of very similar 16S sequences

16S data features

- **Discrete counts** of sequence reads: number of time each OTU was found in a sample
- **Large-scale** data: ~ 17000 OTUs × 666 samples



- **Heterogeneneous** data due to:
  - biological phenomena: some species are found in only a small % of samples
  - technical reasons: others are not detected (insufficient seq depth)

→ **Library size** (total reads per sample) vary by orders of magnitude

→ **Sparsity**: *i.e.* most OTUs are rare (98% of sparsity in raw data)

→ **Overdispersion**: variance grows faster than the mean

Outline

1. Introduction
   - The MAARS project
   - Microbiome data production and features
2. Normalisation of 16S data
   - Motivations
   - State of the art
   - Evaluation of current methods
3. Preliminary results
   - Exploratory analysis
   - Integration of microbiome and transcriptome data

Normalisation: motivations

Comparison across samples with different library sizes may induce biases in the downstream analysis

- **Differential analysis**: the higher sequencing depth, the higher counts
- **Diversity/richness estimation**: rarefaction phenomenon
  "The number of taxonomic features detected in a sample depends on the amount of sequencing performed"

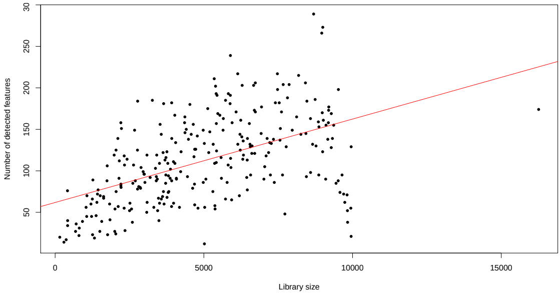


Figure: Illustration of the rarefaction phenomenon on MAARS data

Normalisation: current practices

Rarefying

Random subsampling of each sample to a common depth:

- Omission of available data: **add artificial uncertainty**
- Inflate the variance and induce a **loss of power** in differential analysis

Total-sum scaling (TSS): proportional abundance of species

Divide read counts by the total number of reads in each sample:

$$\widetilde{c_{ij}} = \frac{c_{ij}}{s_j}$$

where:

- c_{ij} is the number of times taxonomic feature i was observed in sample j
- $s_j = \sum_i c_{ij}$, sum of counts for sample i

In practice...

- Does not account for **heteroscedasticity**
- Dillies et al. demonstrated biais in RNA-seq data: undue influence of **high-count genes** on normalized counts
  - ↗ FPR when **differences in library composition**

Normalisation: alternative approaches

Methods derived from the field of RNA-seq data analysis:

1. **Quantile** (Q): Quantiles of the count distributions are matched between samples
2. **Upper-Quartile** (UQ): scale factors are calculated from the 75% quantile of the counts for each library
3. **Relative Log Expression** (RLE) - DESeq (Anders & Huber 2010):

$$\hat{s}_j = median_i\left(\frac{c_{ij}}{(\pi_{v=1}^{n} c_{iv})^{1/n}}\right)$$

   where n is the sample size.

4. **Trimmed Mean of M-values** (TMM) - EdgeR (Robinson et al. 2010)
   Trim data by log-fold-changes M_i and absolute intensity A_i:

$$M_i = \log_2 \frac{c_{ij}/s_j}{c_{ij'}/s_{j'}}; \qquad A_i = \frac{1}{2}\log_2(c_{ij}/s_j \times c_{ij'}/s_{j'});$$

   ▷ Scaling factor: trimmed mean of the log-abundance ratios

5. **Voom** (Law et al. 2014)
   Log-counts per million (log-cpm) value:

$$y_{ij} = \log_2\left(\frac{c_{ij}+0.5}{s_j+1} \times 10^6\right)$$

   The library size is offset by 1 to ensure that $0 < \frac{c_{ij}+0.5}{s_j+1} < 1$

Normalisation: alternative approaches

Methods derived from the field of RNA-seq data analysis:

1. **Quantile** (Q): Quantiles of the count distributions are matched between samples
2. **Upper-Quartile** (UQ): scale factors are calculated from the 75% quantile of the counts for each library
3. **Relative Log Expression** (RLE) - DESeq (Anders & Huber 2010):

   $$\hat{s}_j = median_i\left(\frac{c_{ij}}{(\pi_{v=1}^{n} c_{iv})^{1/n}}\right)$$

   where n is the sample size.
4. **Trimmed Mean of M-values** (TMM) - EdgeR (Robinson et al. 2010)
   Trim data by log-fold-changes M_i and absolute intensity A_i:

   $$M_i = \log_2 \frac{c_{ij}/s_j}{c_{ij'}/s_{j'}}; \qquad A_i = \frac{1}{2}\log_2(c_{ij}/s_j \times c_{ij'}/s_{j'});$$

   ▷ Scaling factor: trimmed mean of the log-abundance ratios
5. **Voom** (Law et al. 2014)
   Log-counts per million (log-cpm) value:

   $$y_{ij} = \log_2\left(\frac{c_{ij} + 0.5}{s_j + 1} \times 10^6\right)$$

   The library size is offset by 1 to ensure that $0 < \frac{c_{ij}+0.5}{s_j+1} < 1$

Normalisation: alternative approaches

Methods derived from the field of RNA-seq data analysis:

1. **Quantile** (Q): Quantiles of the count distributions are matched between samples
2. **Upper-Quartile** (UQ): scale factors are calculated from the 75% quantile of the counts for each library
3. **Relative Log Expression** (RLE) - DESeq (Anders & Huber 2010):

$$\hat{s}_j = median_i\left(\frac{c_{ij}}{(\pi_{v=1}^{n} c_{iv})^{1/n}}\right)$$

   where n is the sample size.

4. **Trimmed Mean of M-values** (TMM) - EdgeR (Robinson et al. 2010)
   Trim data by log-fold-changes M_i and absolute intensity A_i:

$$M_i = \log_2 \frac{c_{ij}/s_j}{c_{ij'}/s_{j'}}; \qquad A_i = \frac{1}{2}\log_2(c_{ij}/s_j \times c_{ij'}/s_{j'});$$

   ▷ Scaling factor: trimmed mean of the log-abundance ratios

5. **Voom** (Law et al. 2014)
   Log-counts per million (log-cpm) value:

$$y_{ij} = \log_2\left(\frac{c_{ij}+0.5}{s_j+1} \times 10^6\right)$$

   The library size is offset by 1 to ensure that $0 < \frac{c_{ij}+0.5}{s_j+1} < 1$

Normalisation: alternative approaches

Methods derived from the field of RNA-seq data analysis:

1. **Quantile** (Q): Quantiles of the count distributions are matched between samples
2. **Upper-Quartile** (UQ): scale factors are calculated from the 75% quantile of the counts for each library
3. **Relative Log Expression** (RLE) - DESeq (Anders & Huber 2010):

   $$\hat{s}_j = median_i\left(\frac{c_{ij}}{(\pi_{v=1}^{n} c_{iv})^{1/n}}\right)$$

   where n is the sample size.
4. **Trimmed Mean of M-values** (TMM) - EdgeR (Robinson et al. 2010)
   Trim data by log-fold-changes M_i and absolute intensity A_i:

   $$M_i = \log_2 \frac{c_{ij}/s_j}{c_{ij'}/s_{j'}}; \qquad A_i = \frac{1}{2}\log_2(c_{ij}/s_j \times c_{ij'}/s_{j'});$$

   ▷ Scaling factor: trimmed mean of the log-abundance ratios
5. **Voom** (Law et al. 2014)
   Log-counts per million (log-cpm) value:

   $$y_{ij} = \log_2\left(\frac{c_{ij}+0.5}{s_j+1} \times 10^6\right)$$

   The library size is offset by 1 to ensure that $0 < \frac{c_{ij}+0.5}{s_j+1} < 1$

Normalisation: alternative approaches

Methods derived from the field of RNA-seq data analysis:

1. **Quantile** (Q): Quantiles of the count distributions are matched between samples
2. **Upper-Quartile** (UQ): scale factors are calculated from the 75% quantile of the counts for each library
3. **Relative Log Expression** (RLE) - DESeq (Anders & Huber 2010):

   $$\hat{s}_j = median_i\left(\frac{c_{ij}}{(\pi_{v=1}^{n} c_{iv})^{1/n}}\right)$$

   where n is the sample size.
4. **Trimmed Mean of M-values** (TMM) - EdgeR (Robinson et al. 2010)
   Trim data by log-fold-changes M_i and absolute intensity A_i:

   $$M_i = \log_2 \frac{c_{ij}/s_j}{c_{ij'}/s_{j'}}; \qquad A_i = \frac{1}{2}\log_2(c_{ij}/s_j \times c_{ij'}/s_{j'});$$

   ▷ Scaling factor: trimmed mean of the log-abundance ratios
5. **Voom** (Law et al. 2014)
   Log-counts per million (log-cpm) value:

   $$y_{ij} = \log_2\left(\frac{c_{ij} + 0.5}{s_j + 1} \times 10^6\right)$$

   The library size is offset by 1 to ensure that $0 < \frac{c_{ij}+0.5}{s_j+1} < 1$

Normalisation : Cumulative-Sum Scaling (CSS)

CSS strategy

Paulson, J. *et al.* (2013), Nature Methods

- q_j^l: lth quantile of sample j
- $s_j^l = \sum_{i|c_{ij} \leq q_j^l} c_{ij}$
- N: normalization constant (ex: the $\text{med}_j(s_j^l)$)

$$\widetilde{c_{ij}} = \frac{c_{ij}}{s_j^l} N$$

▸ avoid placing undue influence on high-count features

Selection of the appropriate quantile

- $\bar{q}^l = \text{med}_j(q_j^l)$, median lth quantile across samples
- $d_l = med_j|q_j^l - \bar{q}^l|$, **median absolute deviation** of sample-specific quantiles
- $\hat{l}$: smallest value for which high instability is detected

Except for voom, all approaches decrease the range of library sizes

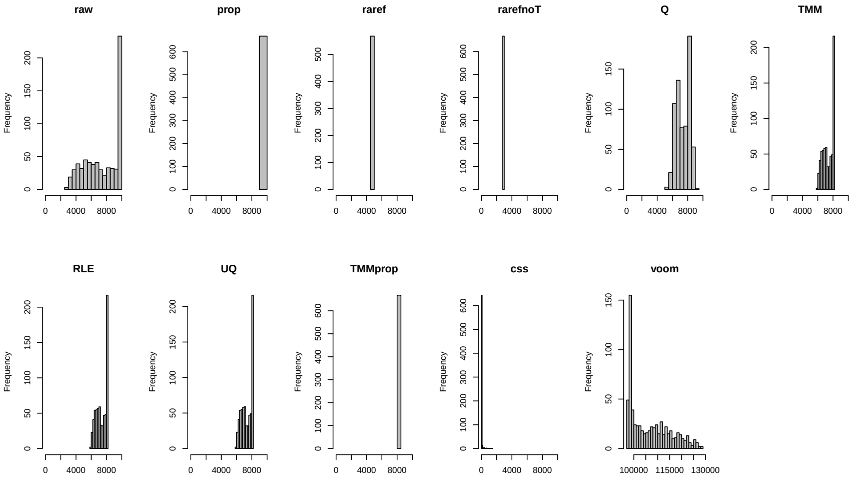


Figure: Distribution of library sizes across normalisation approaches

All the normalisation methods improve the homogeneity between technical replicates

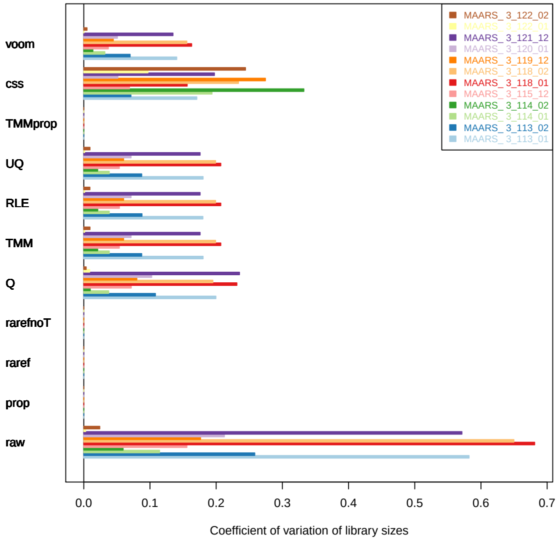


Figure: Homogeneity of library sizes between technical replicates

Voom increases the distance between technical replicates

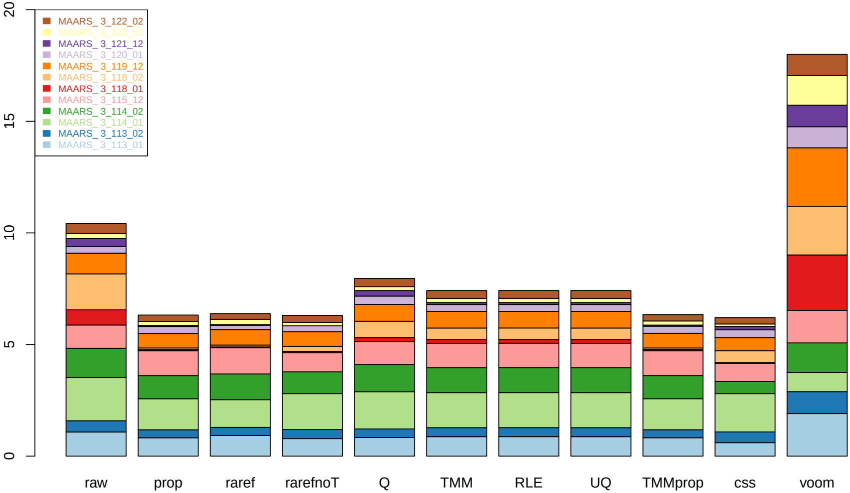


Figure: Distances between technical replicates

Rarefying decreases the biological signal

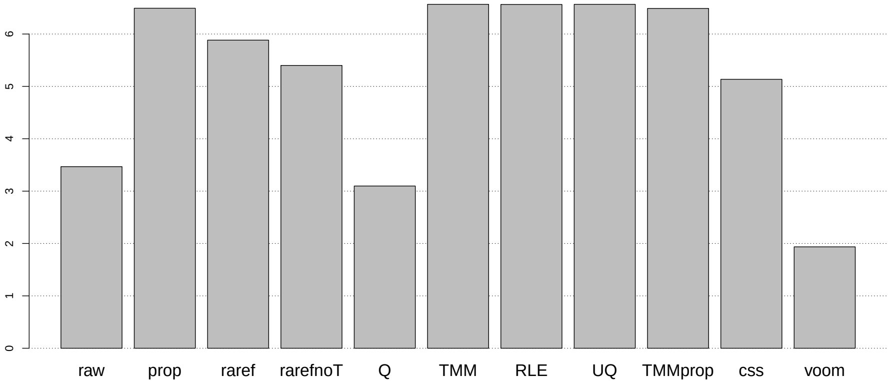


Figure: Ratio of distances between clinical groups and technical replicates

Proportion and rarefying approaches show high FPR

McMurdie, P.J. and Holmes, S. (2014), PLOS Comp. Biol.

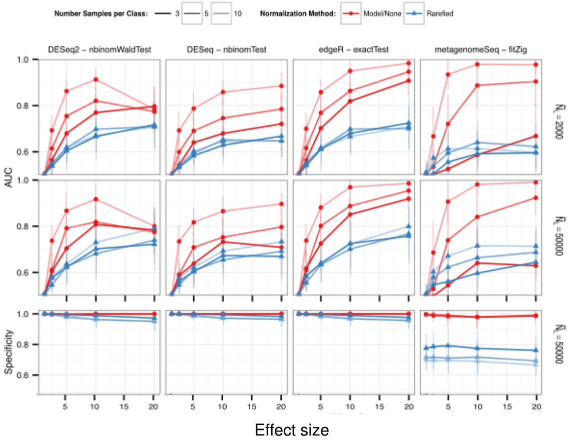


Figure: Performance of differential abundance detection on simulated data

Preliminary results on permuted data show that proportions and rarefying exhibit a FPR of 30%

Conclusions on normalisation approaches

- TMM and RLE are the best compromises :
  - show good results on simulated data (McMurdie 2014)
  - reduce the heterogeneity in library sizes
  - lower the distances between technical replicates
  - do not degrade the biological signal
- UQ performs well but need to be tested on simulated data
- Voom, Q and CSS normalisation approaches to be proscribed
- Perspectives for differential abundance testing: zero-inflated negative binomial model

Outline

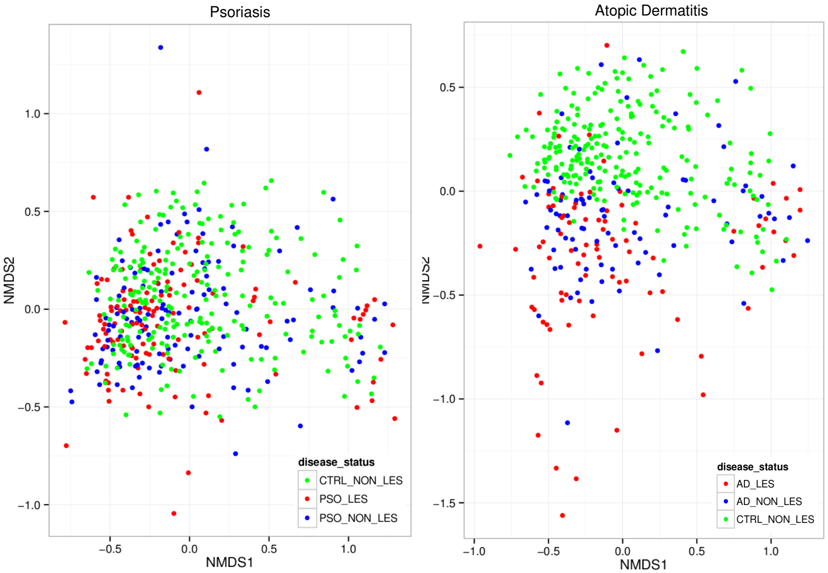
1 Introduction
- The MAARS project
- Microbiome data production and features

2 Normalisation of 16S data
- Motivations
- State of the art
- Evaluation of current methods

3 Preliminary results
- Exploratory analysis
- Integration of microbiome and transcriptome data

Microbiome data enable to discriminate AD samples from controls

Non-metric MultiDimensional Scaling



Integration of microbiome and transcriptome data

▷ Unravel the interdependencies between skin microbiome and transcriptome

Univariate analysis

- Associate the presence of a given microbe with different transcriptome profiles

Multivariate exploratory analysis

Canonical Correlation Analysis:

⇝ identify largest correlations between linear combinations of transcriptome and OTU profiles

Let us consider two matrices X and Y of order $n \times p$ and $n \times q$ respectively, with $p \leq q$.

For $S = 1, \ldots, p$, find $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_p$ such as:

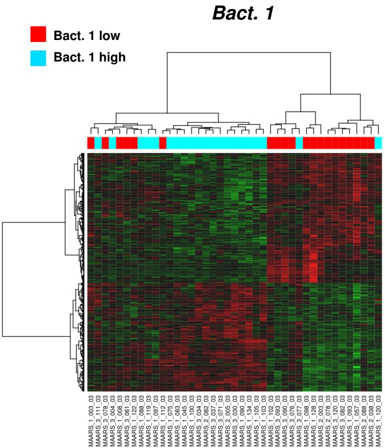
$$\begin{aligned} \rho_s &= \max_{a^S, b^S} cor(Xa^S, Yb^S) \qquad (1) \\ &= cor(U^S, V^S) \qquad (2) \end{aligned}$$

with $cor(U^S, U^K) = cor(V^S, V^K) = 0$ for $S \neq K$.

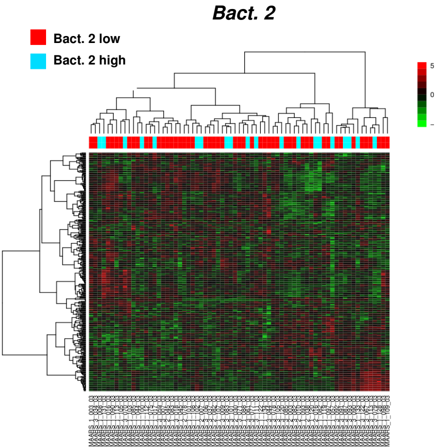
- U^S and V^S: canonical variates
- ρ_S: canonical correlations

Univariate approach: preliminary results

Abundance of given bacteria in AD is associated with different transcriptome profiles

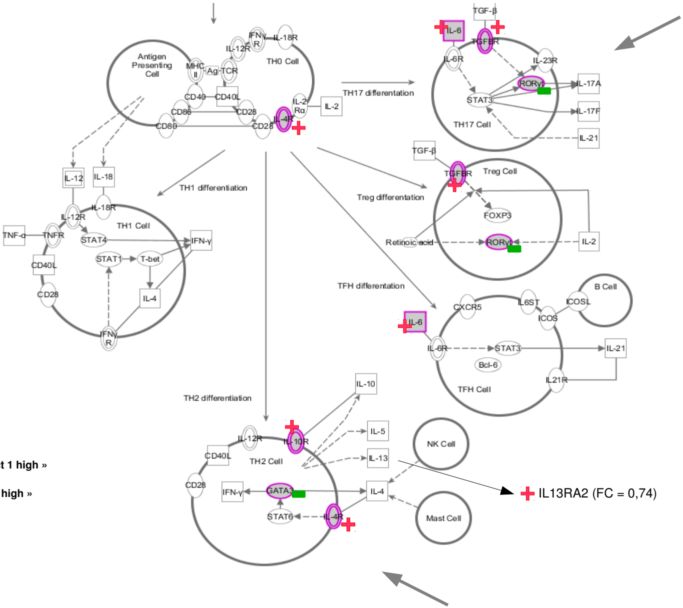


Heatmap of differentially expressed genes in bact. 1 low/high patients



Heatmap of the 400 « most significant » genes in *bact 2.* low/high patients

High abundance of bact. 1 is related to a dysregulated T-helper cell differentiation pathway



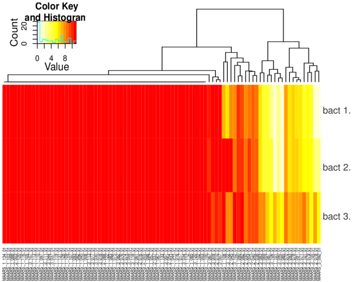
✚ IL13RA2 (FC = 0,74)

🟩 **Down-regulation in « bact 1 high »**

✚ **Up-regulation in « bact 1 high »**

Conclusions and perspectives

- 16S data: large-scale count data
  - similar features than RNA-seq data
  - BUT with a higher level of sparsity
- Normalization methods used in RNA-seq analysis
  - perform well on 16S data
  - should be transferred to microbiome research (instead of rarefying)
- No consensus for differential analysis
- Investigate co-occurences/co-exclusions of microbes



The MAARS consortium

Thanks !

Alix, Mahé, Paula, Sol, Caro, Max, Gérôme, Maude, Phil, Lucia, Irit, Vassili, Salvo, Sofia, Anto, Colline, Aurore.

Thank you !