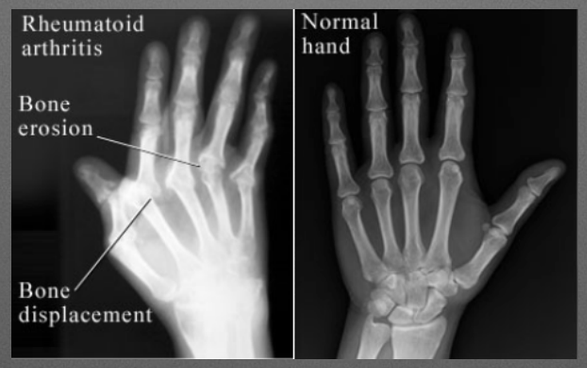
Travaux & problématique en métagénomique au CBiB

Groupe de travail « France Génomique Métagénomique »

BARRE Aurélien - Paris - 30 juin 2015

Virus de plantes



Santé

Paléo métagenomique

Questions :

- Connaitre la liste des espèces présentes
- Comparer des échantillons
- Analyser fonctionnellement le milieu

Classification génomique bactérienne

2013-2015

Polyathrite rhumatoide

- Etude de faisabilité avec un séquençage de reads courts (75 nt) avec la technologie Illumina
- Gestion des ambiguïtés résultant du mapping multiple des reads sur les séquences 16 S
- Prise en compte de la phylogénie des bactéries
- Recherche de lien entre la composition en bactéries & le développement de la maladie

Etapes du projet

- **Phase 1 : Etude de faisabilité (3 échantillons)**
- **Phase 2 : Analyse comparative entre patients sains et malades en aveugle (11 échantillons)**
- Phase 3 : Analyse de variation de la composition au cours du traitement & de la prédisposition à la réponse au traitement (12 échantillons)

Les taxonomies disponibles

- NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/)
- Greengenes (http://greengenes.lbl.gov/)
- Ribosomal Dataset Project (http://rdp.cme.msu.edu/)

Formulation du problème

In A genomic reference ***S*** (set of sequences)
A taxonomic reference ***T*** (tree)
A leaf set ***L***, s.t. $\forall Li \in L$ has an associated sequence of *S*
A set ***R*** of sequence reads

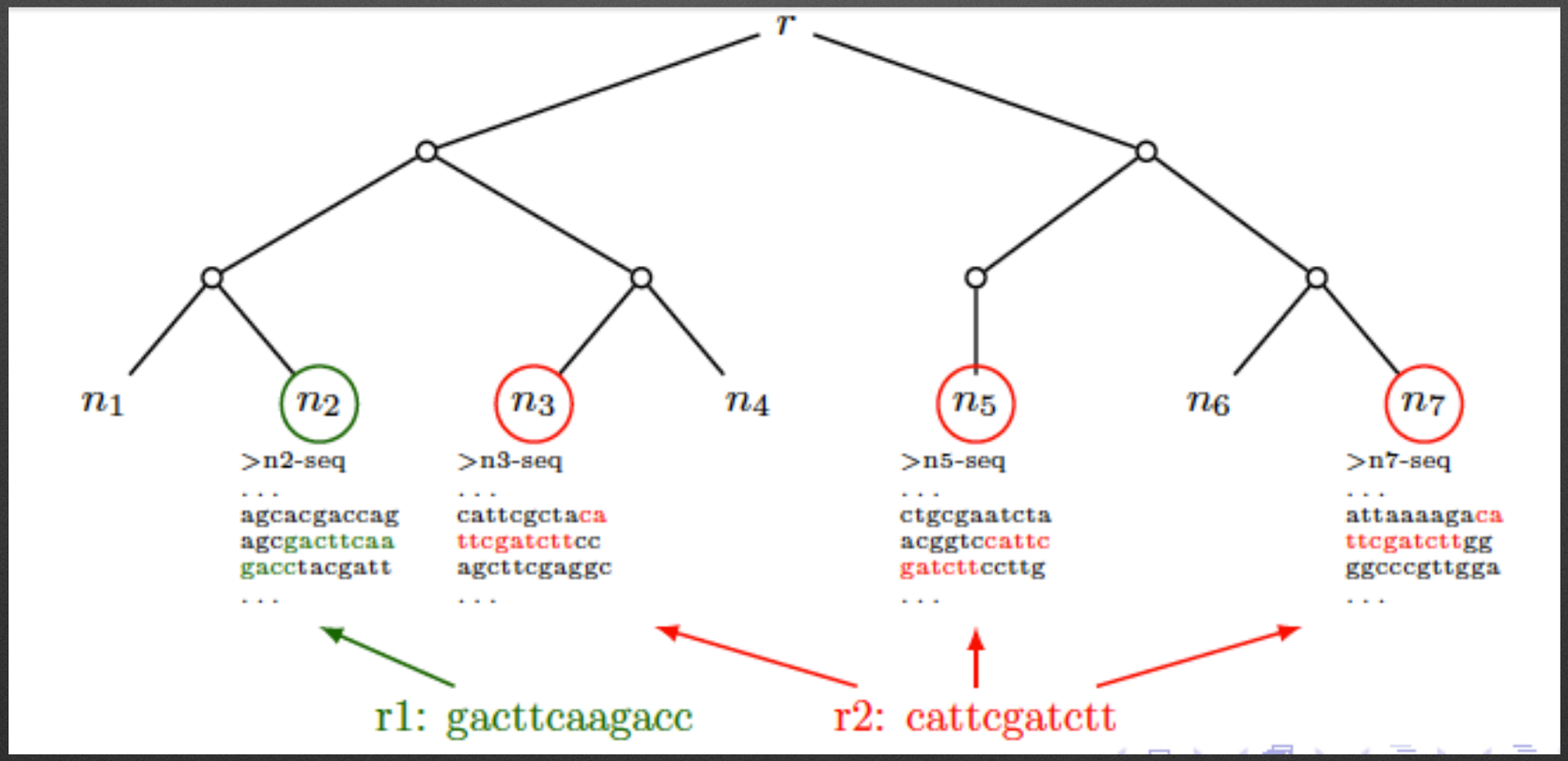
Out For each read $Ri \in R$, a node in *T* that represents a “good” subset of matches $\boldsymbol{M}i \in L$

D. Alonso-Alemany, A. Barre, S. Beretta, P. Bonizzoni, M. Nikolski, G. Valiente,
Further Steps in TANGO: Improved Taxonomic Assignment in Metagenomics, Bioinformatics (2013)

1 - Mapping

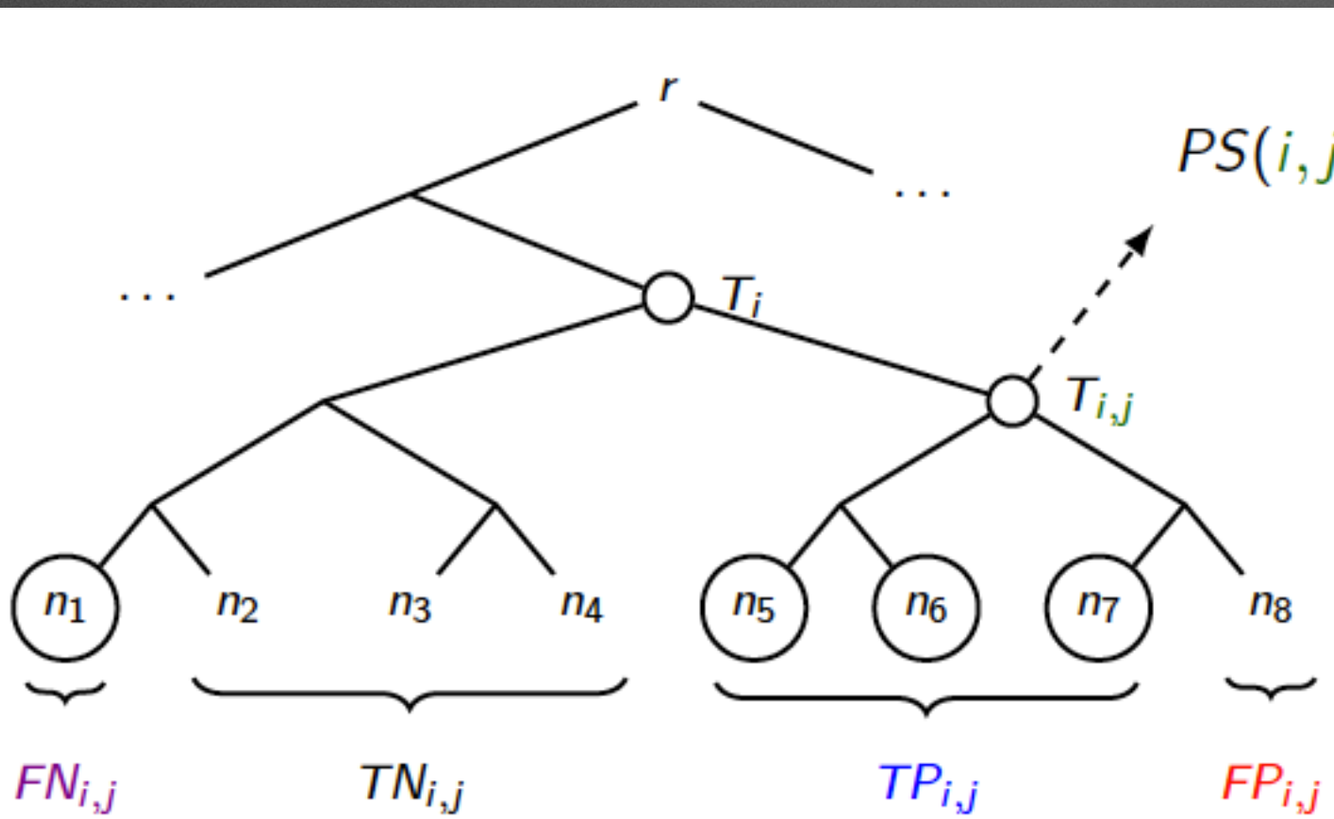
Alignement des reads contre les séquences de référence

- single match -> read non ambigu
- multiple matches -> read ambigu



2 - Assignation taxonomique

Recherche de la meilleure assignation basée sur le calcul d'un *Penalty Score* (PS)



$$PS(i,j) = q \cdot \frac{|FN_{i,j}|}{|TP_{i,j}|} + (1-q) \cdot \frac{|FP_{i,j}|}{|TP_{i,j}|}$$

3 - Taxonomy equalizing

Input:

- Two taxonomic trees $T1, T2$
- Leaf mapping $\varnothing: L(T1) \rightarrow L(T2)$

Output:

- Node mapping $\varnothing: T1 \rightarrow T2$

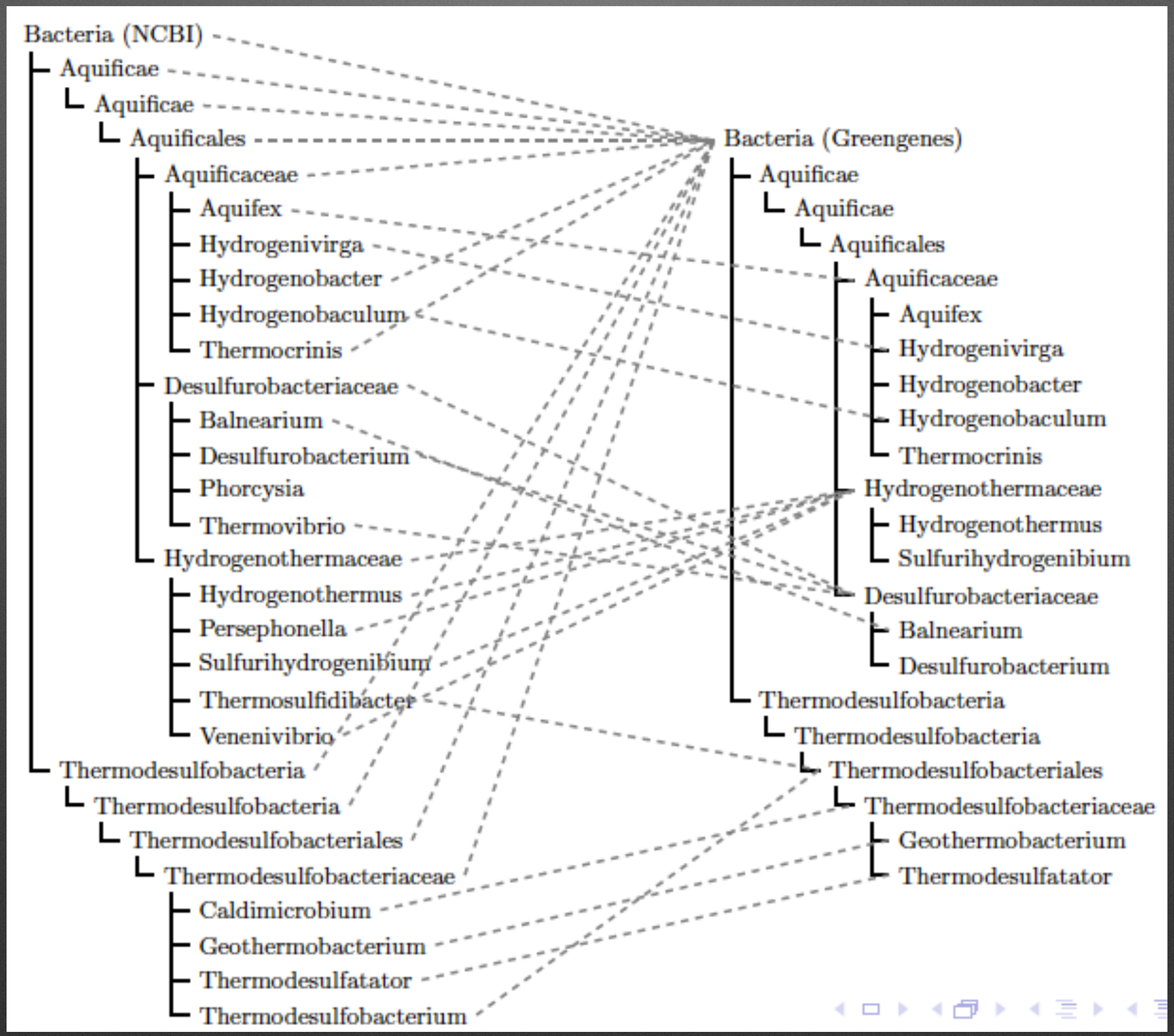
Algorithmic Solution

- *post-order* traversal of the tree $T1$
- for each node $ni \in T1$:

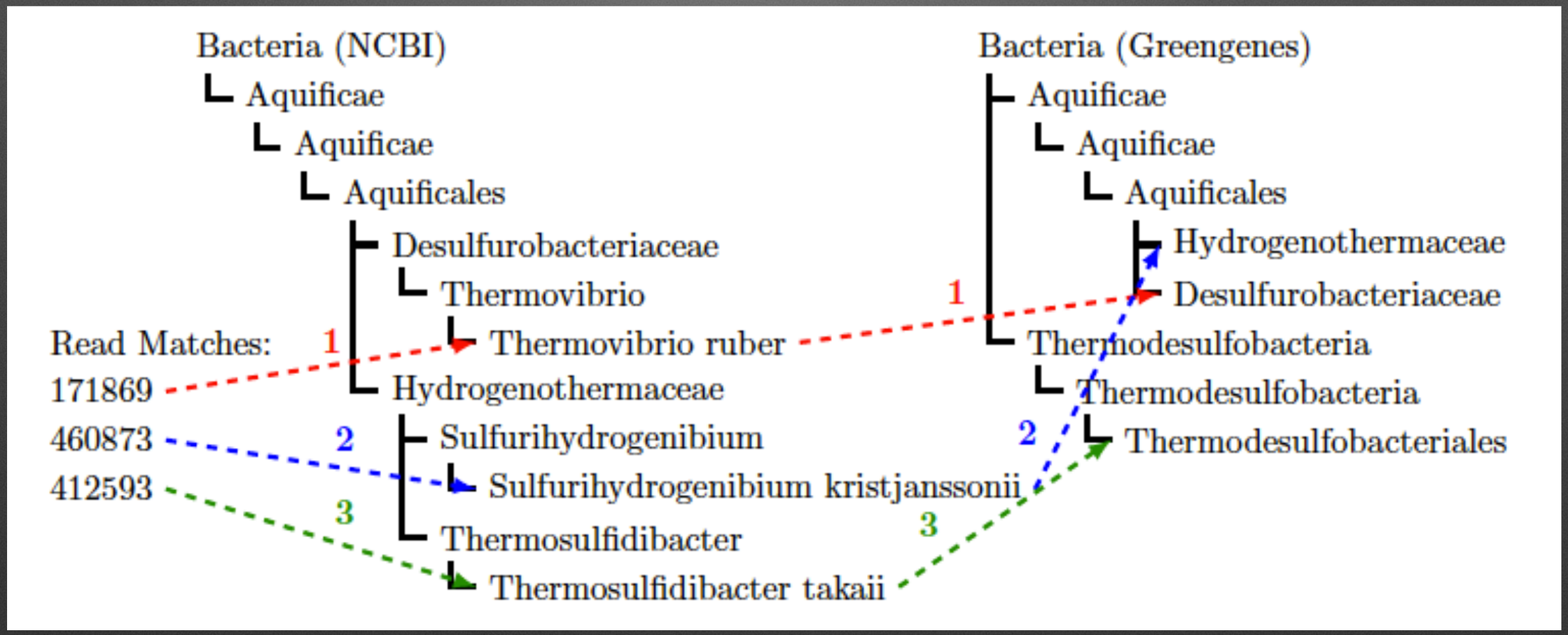
$$\varnothing(ni) = LCA(\{\varnothing(nk) \mid nk \in desc(ni)\})$$

where:

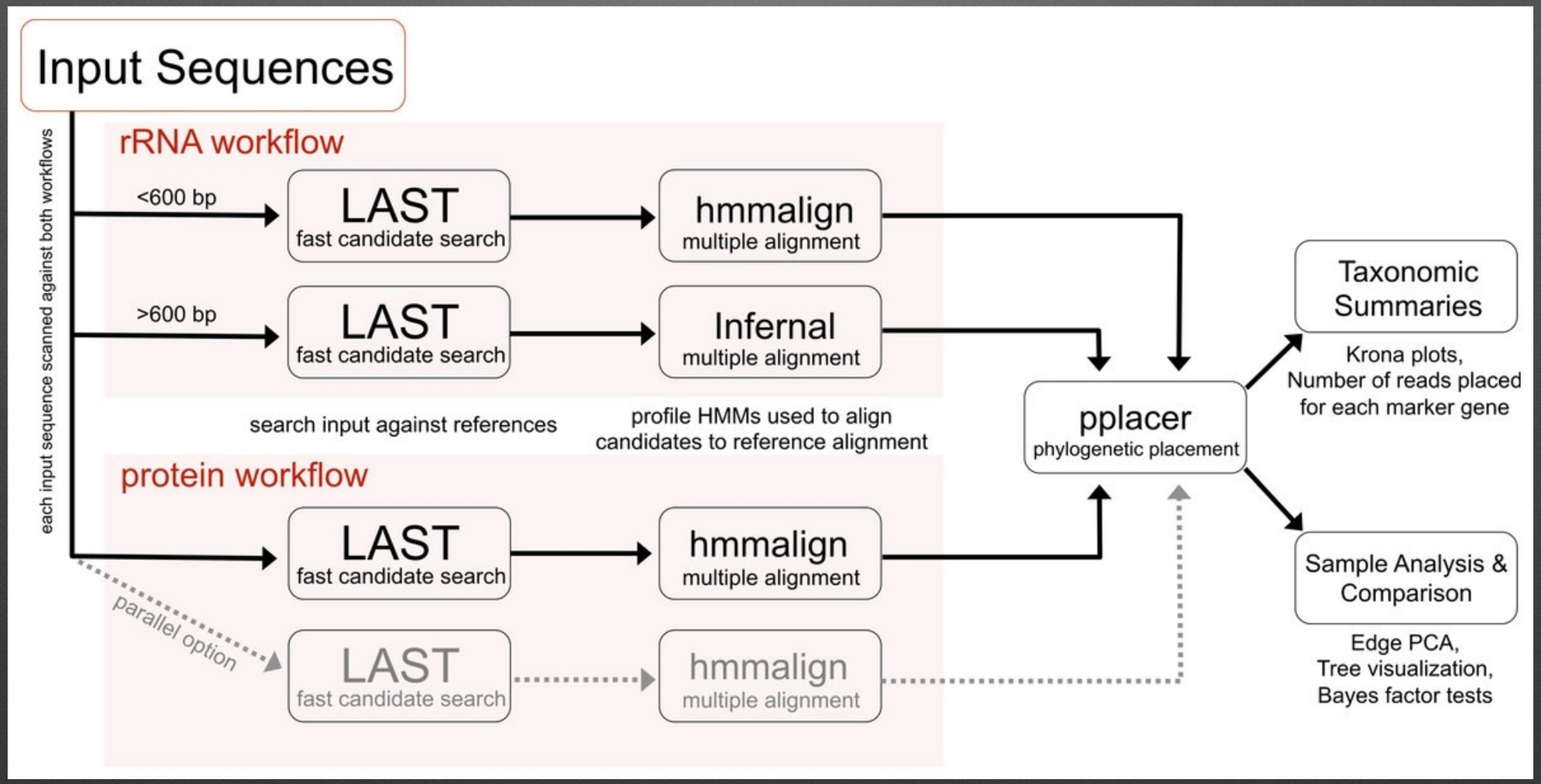
- $LCA(X)$ is the *Lowest Common Ancestor* of nodes $xi \in X$
- $desc(x)$ is the set of descendants of the node x


Bacteria (NCBI)
Aquificae
Aquificae
Aquificales
Aquificaceae
Aquifex
Hydrogenivirga
Hydrogenobacter
Hydrogenobaculum
Thermocrinis
Desulfurobacteriaceae
Balnearium
Desulfurobacterium
Phorcysia
Thermovibrio
Hydrogenothermaceae
Hydrogenothermus
Persephonella
Sulfurihydrogenibium
Thermosulfidibacter
Venenivibrio
Thermodesulfobacteria
Thermodesulfobacteria
Thermodesulfobacteriales
Thermodesulfobacteriaceae
Caldimicrobium
Geothermobacterium
Thermodesulfatator
Thermodesulfobacterium
Bacteria (Greengenes)
Aquificae
Aquificae
Aquificales
Aquificaceae
Aquifex
Hydrogenivirga
Hydrogenobacter
Hydrogenobaculum
Thermocrinis
Hydrogenothermaceae
Hydrogenothermus
Sulfurihydrogenibium
Desulfurobacteriaceae
Balnearium
Desulfurobacterium
Thermodesulfobacteria
Thermodesulfobacteria
Thermodesulfobacteriales
Thermodesulfobacteriaceae
Geothermobacterium
Thermodesulfatator

Reads relabeling



Alternative pour l’assignation taxonomique : Phylosift



Pipeline

Phases 1 et 2

Alignement avec GEM
contre GreenGenes

puis

Assignation taxonomique
avec Tango

Phases 3

Alignement avec BWA
contre GreenGenes

puis

Assignation taxonomique
avec Tango et confirmation
par Phylosift

Comparaison

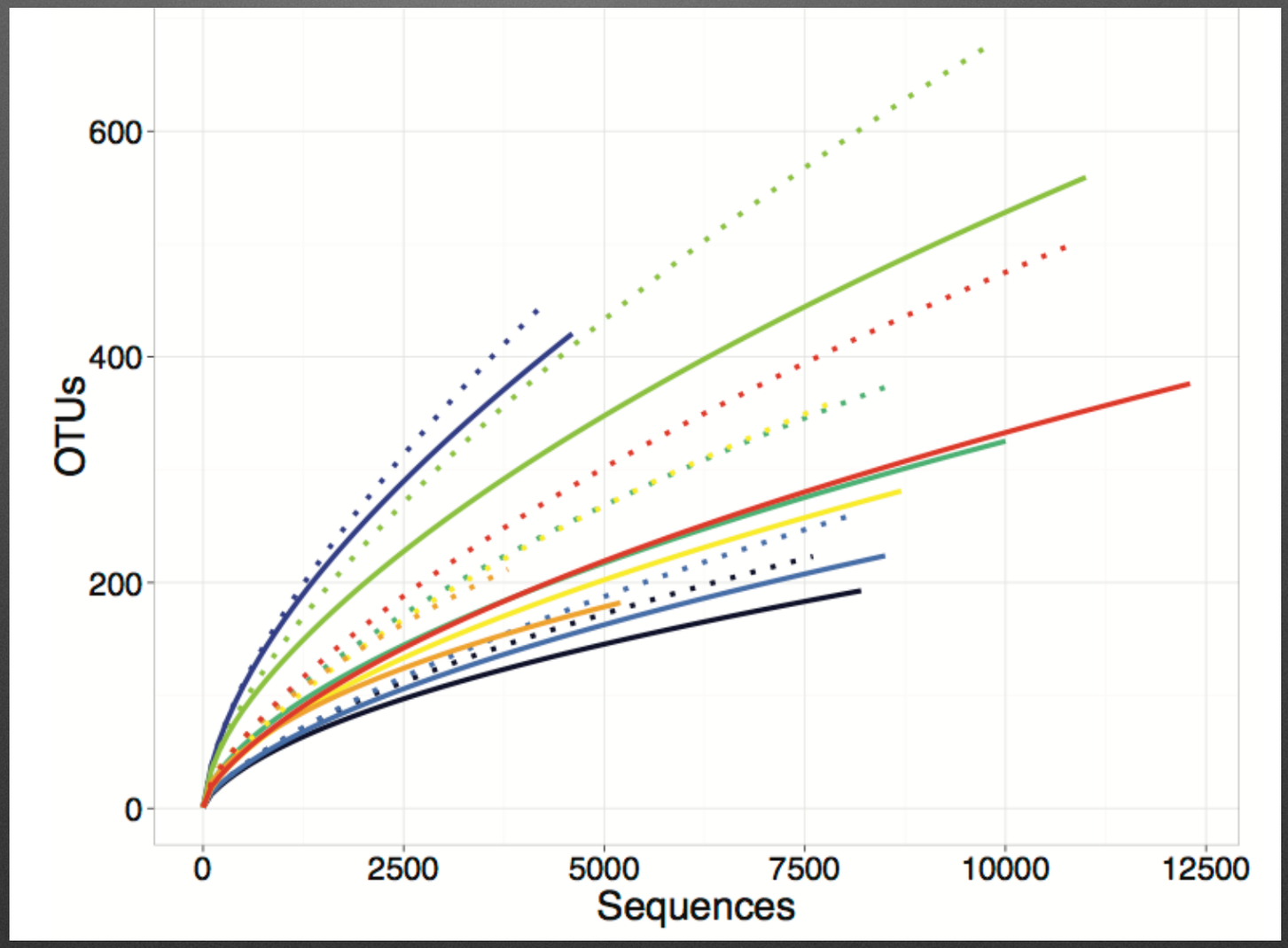
- Normalisation basée sur les k-mer (logiciel khmer) et sur le nombre de reads produits dans chaque échantillon

- Estimation de la variabilité intra-condition et inter-condition

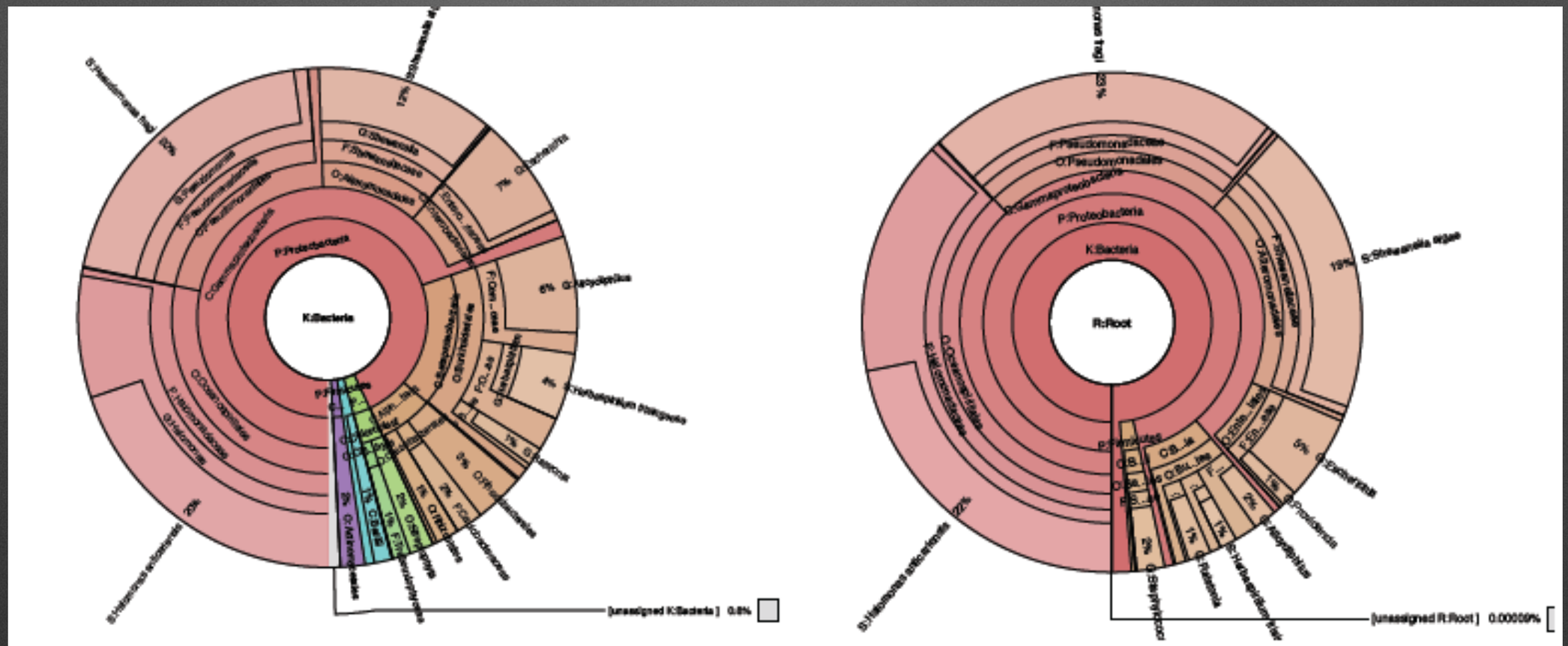
- Analyse statistique classique des noeuds basée sur une version améliorée du package bioconductor « metagenomeseq »

- Représentation graphique des résultats

Les données de séquençage fournies pour l'analyse ne sont malheureusement pas toujours optimales et suffisantes



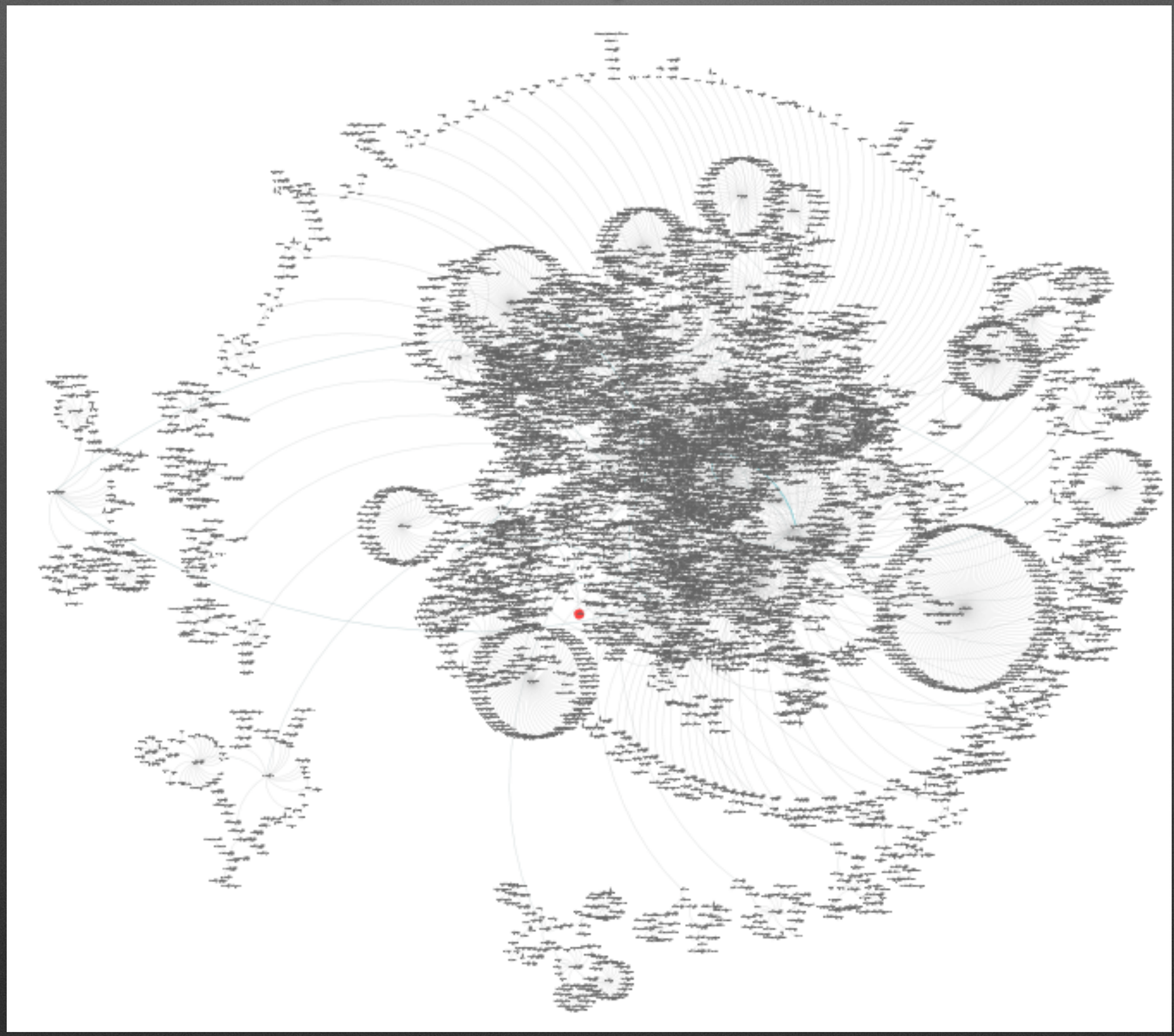
Représentation type Krona produit par Phylosift et Tango



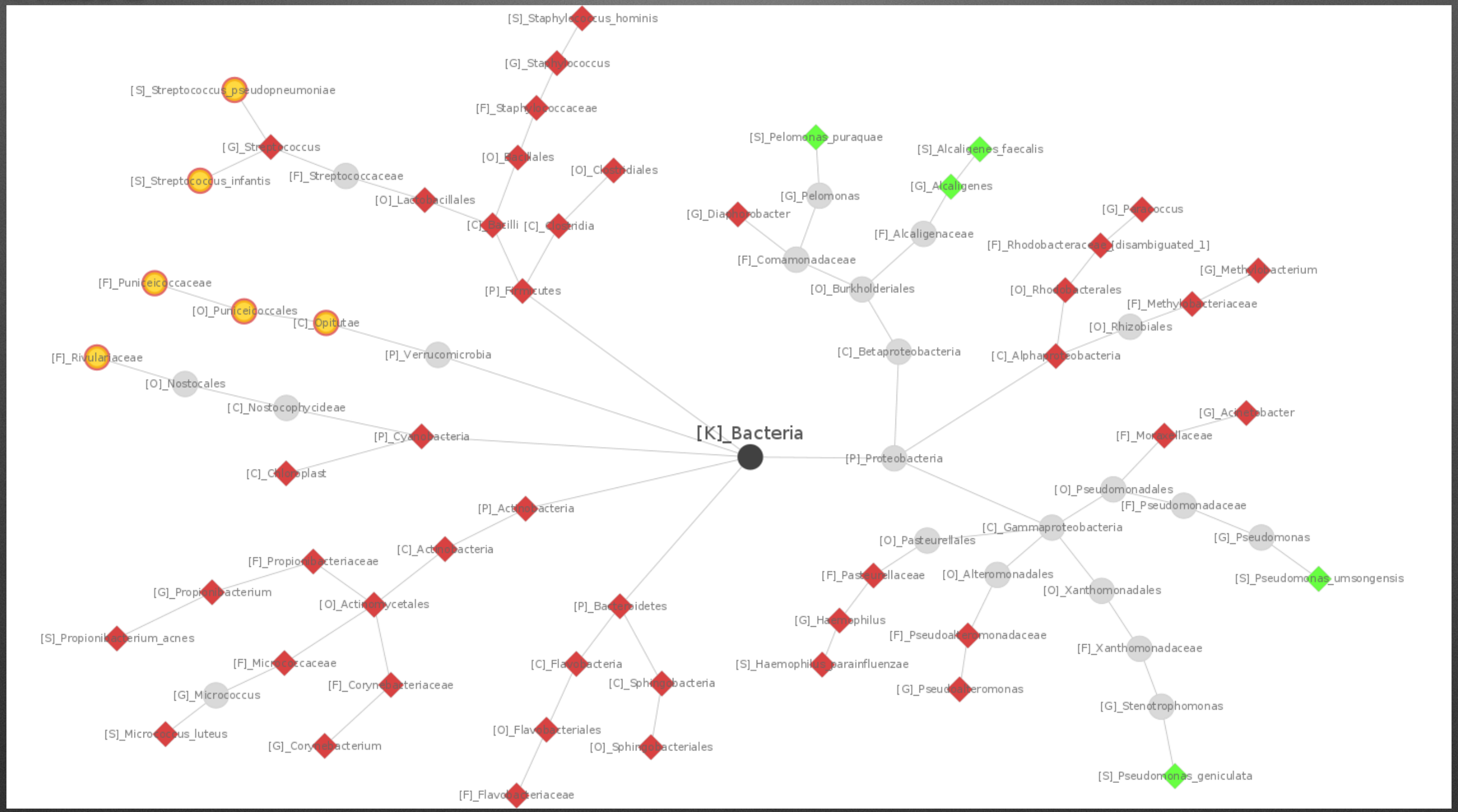
Healthy

Rhumatoid arthritis

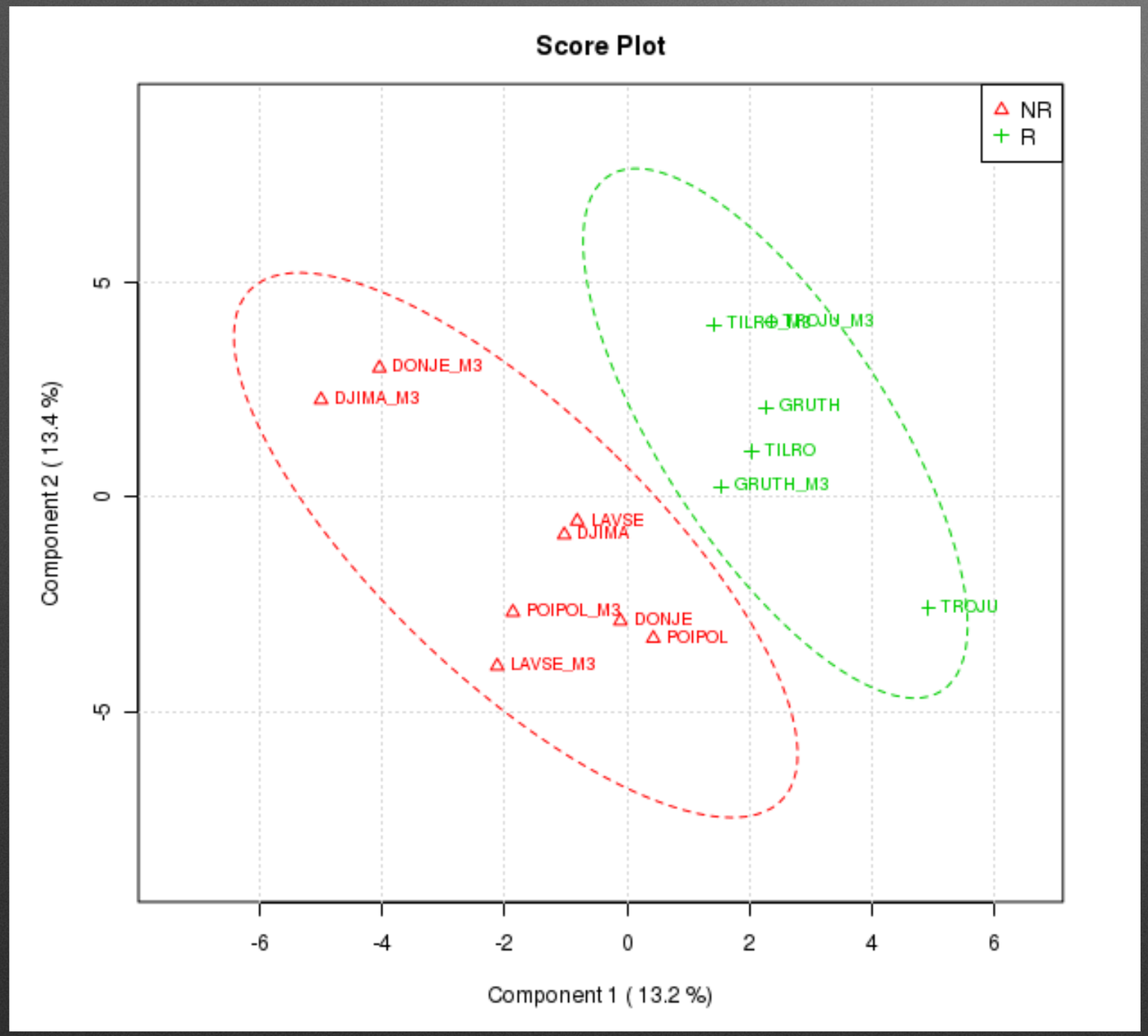
GreenGenes DB (450.000 nodes)

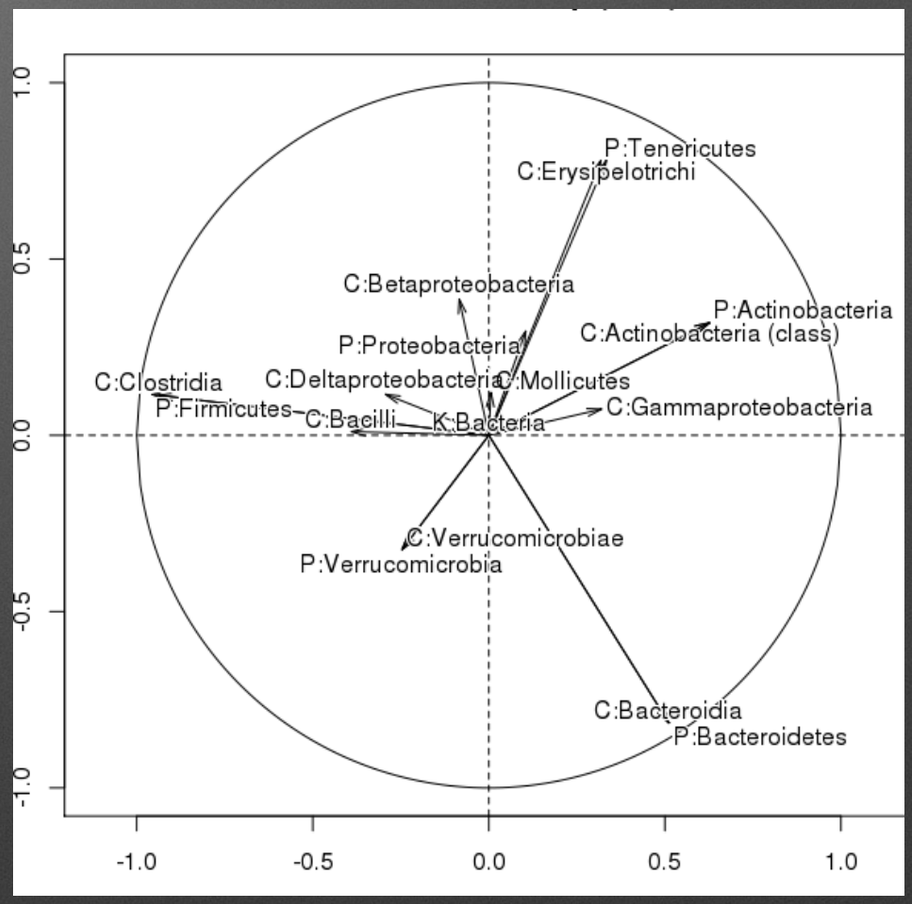


Phase 2


[S]_Staphylococcus_hominis
[G]_Staphylococcus
[S]_Streptococcus_pseudopneumoniae
[F]_Staphylococcaceae
[G]_Streptococcus
[O]_Bacillales
[S]_Streptococcus_infantis
[F]_Streptococcaceae
[O]_Clostridiales
[O]_Lactobacillales
[C]_Bacilli
[C]_Clostridia
[P]_Firmicutes
[F]_Puniceicoccaceae
[O]_Puniceicoccales
[C]_Opitutae
[F]_Rivulariaceae
[P]_Verrucomicrobia
[O]_Nostocales
[C]_Nostocophycideae
[P]_Cyanobacteria
[K]_Bacteria
[C]_Chloroplast
[P]_Actinobacteria
[C]_Actinobacteria
[F]_Propionibacteriaceae
[G]_Propionibacterium
[O]_Actinomycetales
[S]_Propionibacterium_acnes
[P]_Bacteroidetes
[F]_Micrococcaceae
[C]_Flavobacteria
[C]_Sphingobacteria
[F]_Corynebacteriaceae
[G]_Micrococcus
[O]_Flavobacteriales
[O]_Sphingobacteriales
[S]_Micrococcus_luteus
[G]_Corynebacterium
[F]_Flavobacteriaceae
[S]_Pelomonas_puraquae
[S]_Alcaligenes_faecalis
[G]_Alcaligenes
[G]_Pelomonas
[G]_Diaphorobacter
[F]_Alcaligenaceae
[G]_Paracoccus
[F]_Comamonadaceae
[F]_Rhodobacteraceae_(disambiguated_1)
[O]_Burkholderiales
[G]_Methylobacterium
[O]_Rhodobacterales
[F]_Methylobacteriaceae
[O]_Rhizobiales
[C]_Betaproteobacteria
[C]_Alphaproteobacteria
[G]_Acinetobacter
[F]_Moraxellaceae
[P]_Proteobacteria
[O]_Pseudomonadales
[F]_Pseudomonadaceae
[C]_Gammaproteobacteria
[G]_Pseudomonas
[O]_Pasteurellales
[S]_Pseudomonas_umsongensis
[F]_Pasteurellaceae
[O]_Alteromonadales
[O]_Xanthomonadales
[G]_Haemophilus
[F]_Pseudoalteromonadaceae
[F]_Xanthomonadaceae
[S]_Haemophilus_parainfluenzae
[G]_Pseudoalteromonas
[G]_Stenotrophomonas
[S]_Pseudomonas_geniculata

Phase 3


Score Plot
NR
R
DONJE_M3
DJIMA_M3
TILRO_M3
TROJU_M3
GRUTH
TILRO
GRUTH_M3
LAVSE
DJIMA
POIPOL_M3
DONJE
POIPOL
LAVSE_M3
TROJU
Component 2 (13.4 %)
Component 1 (13.2 %)
5
0
-5
-6
-4
-2
0
2
4
6


P:Tenericutes
C:Erysipelotrichi
C:Betaproteobacteria
P:Actinobacteria
C:Actinobacteria (class)
P:Proteobacteria
C:Clostridia
C:Deltaproteobacteria
C:Mollicutes
P:Firmicutes
C:Bacilli
K:Bacteria
C:Gammaproteobacteria
C:Verrucomicrobiae
P:Verrucomicrobia
C:Bacteroidia
P:Bacteroidetes
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0

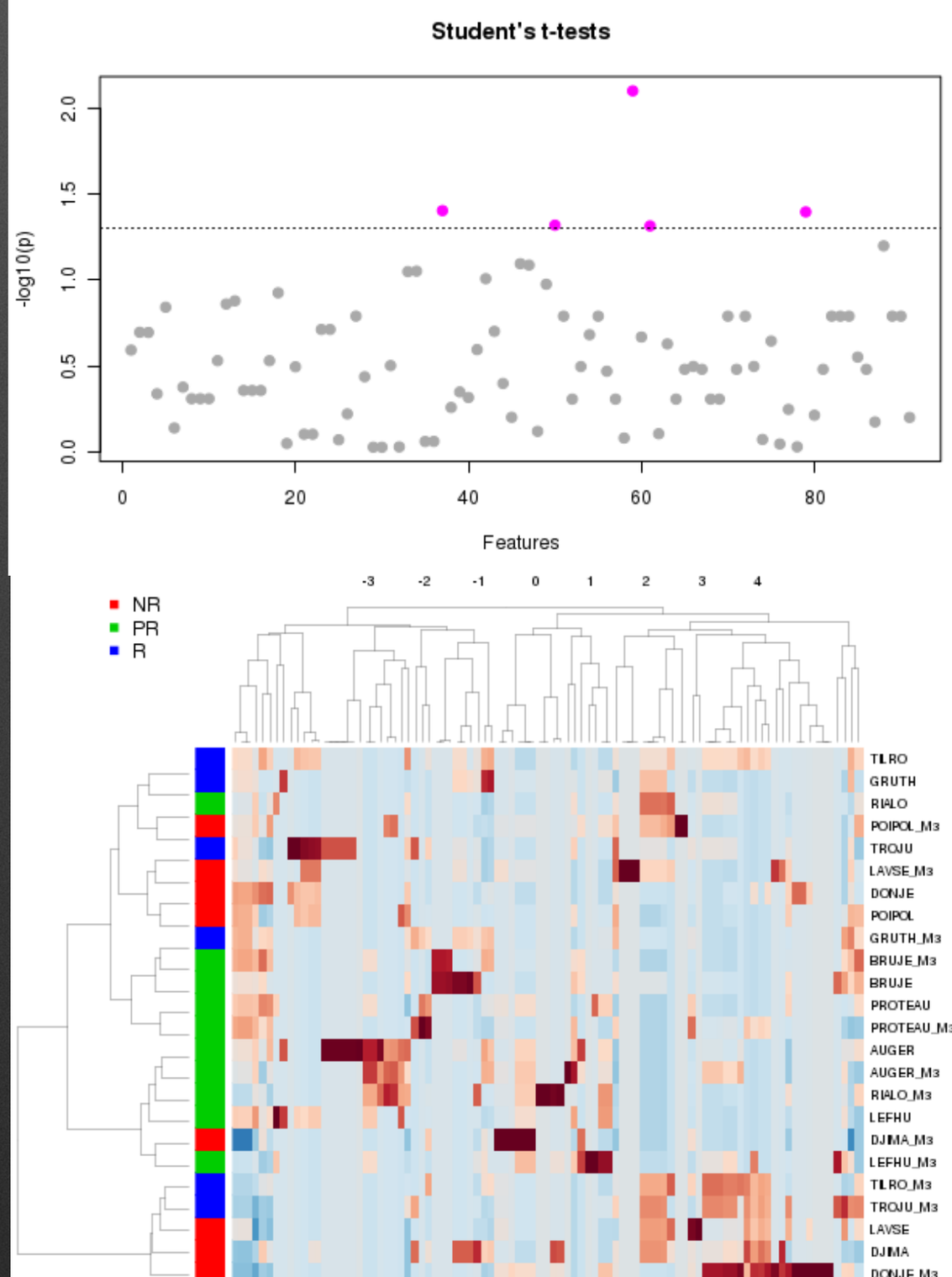
- Ensemble de méthodes pour l'importation, le normalisation, l'analyse et la visualisation des données métagénomiques

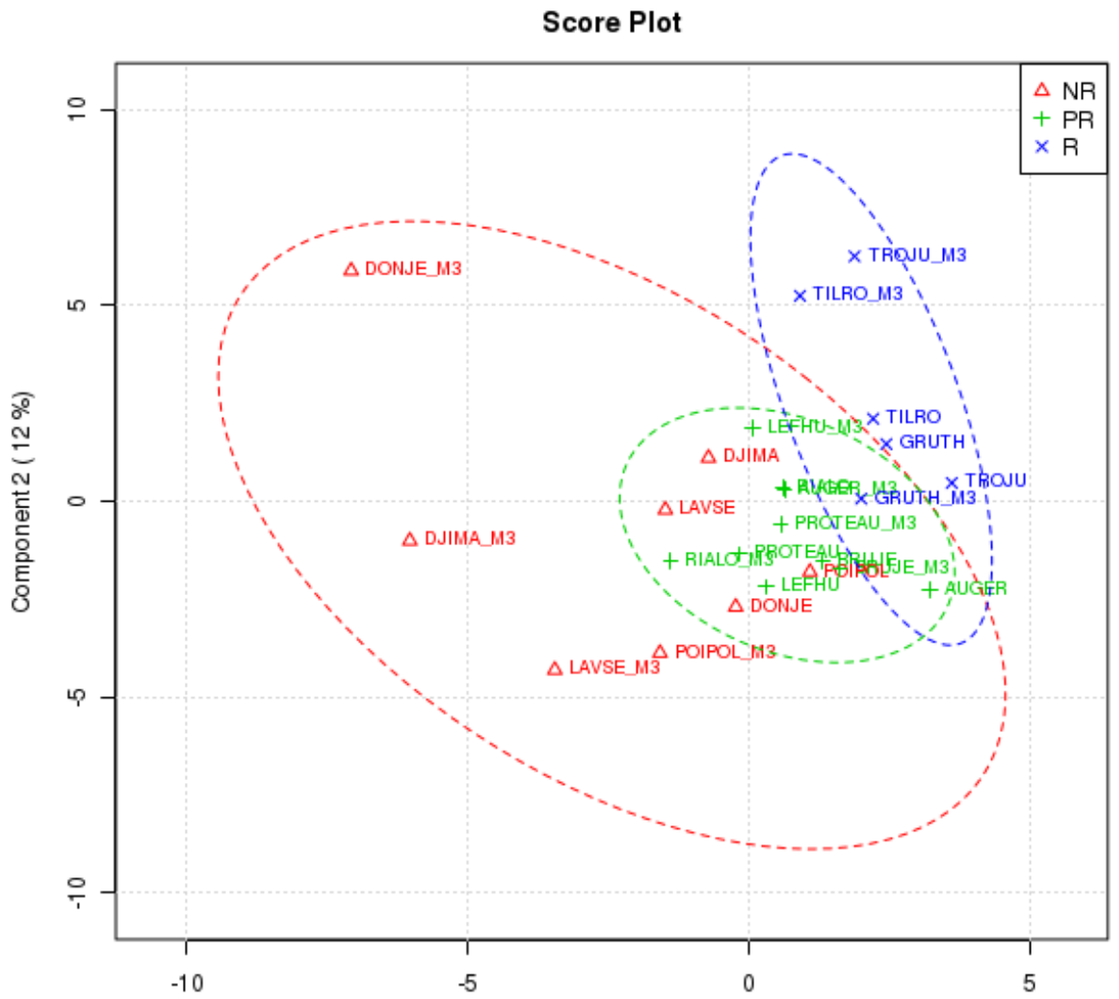
- Workflow très flexible

- Post traitement possible avec R (Rdata format)

METAGENassist

statistical tools for comparative metagenomics





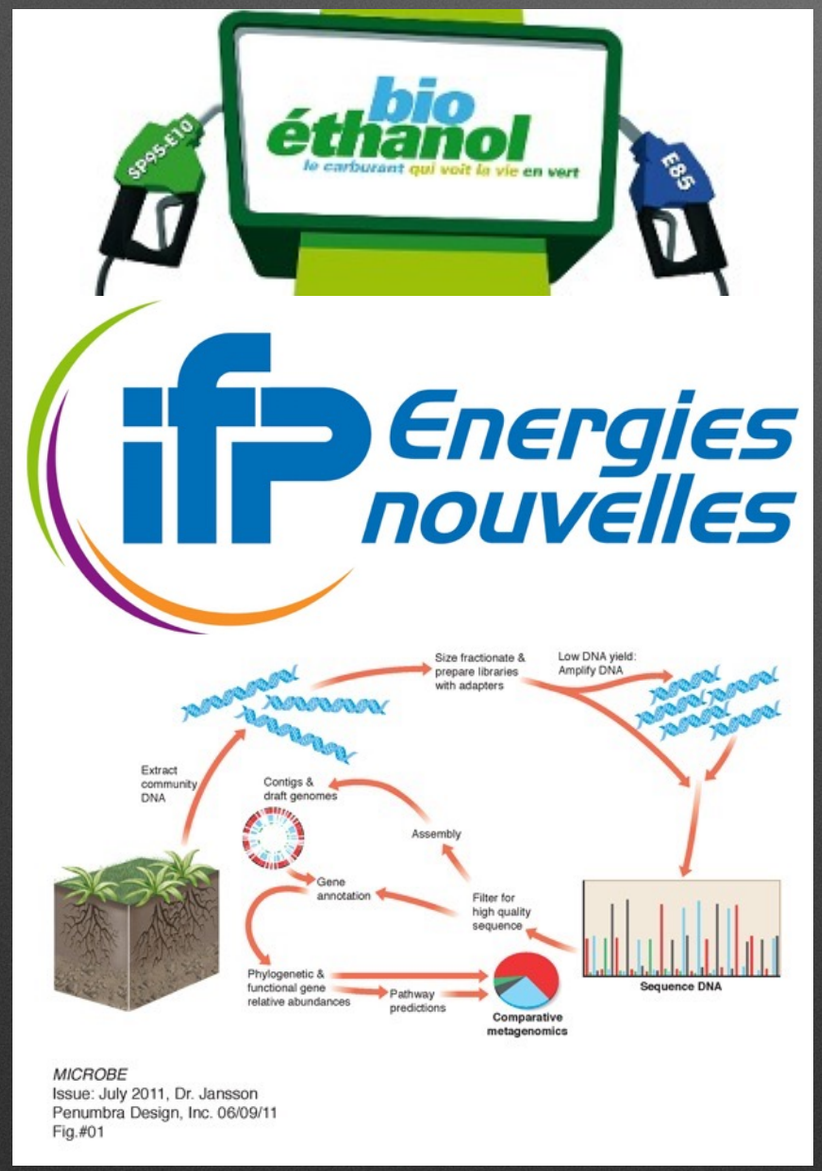
Développement en cours d'un workflow dans l'esprit de MetagenAssist mais qui intègre des étapes en amont (assignation taxonomique) & des étapes en aval (visualisation des résultats sur une phylogénie)

Intégration possible dans Galaxy

Métagénomique fonctionnelle

2014-2015

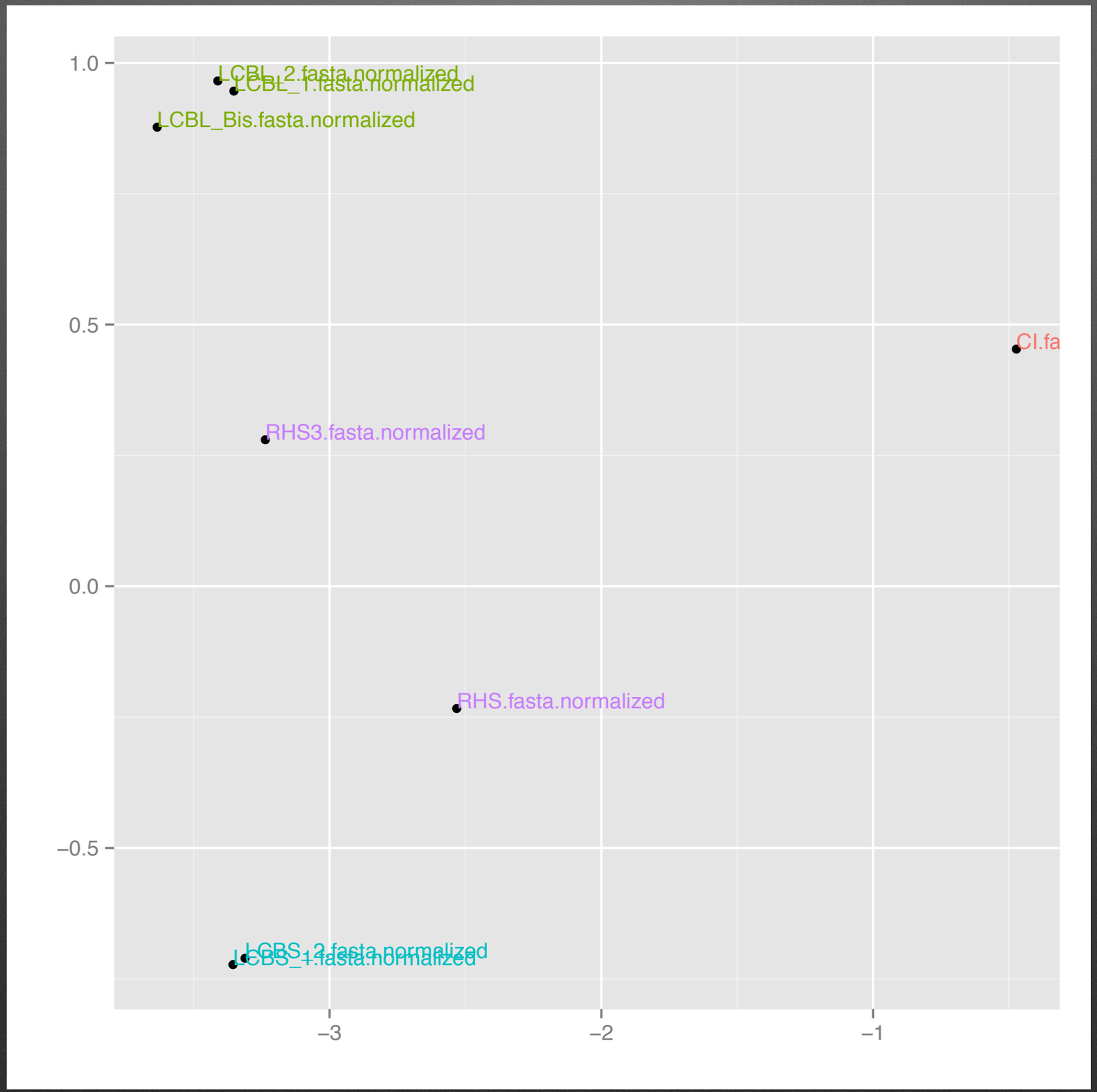
Projet Biomines

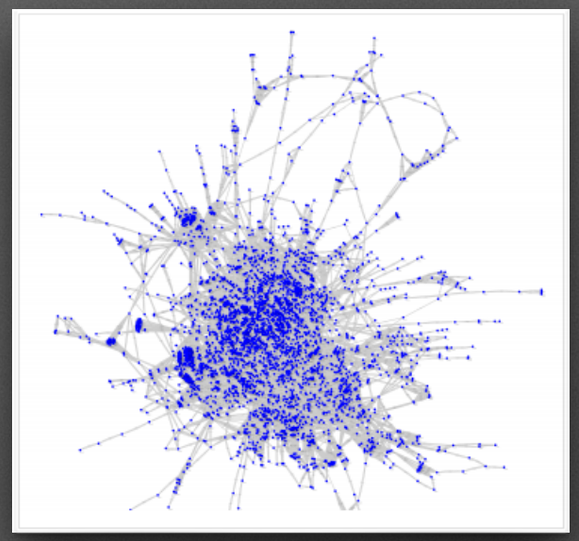


- Déterminer un nouveau bioprocess pour produire bio-ethanol
- Etude de compost pour en déterminer la composition en micro organismes et le potentiel enzymatique de 8 échantillons
- Séquençage principalement en 454
- Analyse DNA-seq et RNA-seq

- Assemblage des reads avec Abyss décevant
- Analyse de la composition des compost avec Phylosift : séquençage suffisant & résultats de clustering très probants ..... sauf pour le RNA-seq
- Binning de mélanges des fosmides avec Concoct en se basant sur le degré de couverture des contigs et la fréquence en k-mer

EPCA


1.0
0.5
0.0
−0.5
−3
−2
−1
LCBL_2.fasta.normalized
LCBL_1.fasta.normalized
LCBL_Bis.fasta.normalized
CI.fa
RHS3.fasta.normalized
RHS.fasta.normalized
LCBS_2.fasta.normalized
LCBS_1.fasta.normalized



- Test du package « mmnet » (microbiome metabolic network) qui intègre les données du KEGG et des prédictions enzymatiques issues de MGRast pour construire des réseaux qui peuvent être ensuite comparés.

- Annotation fonctionnelle des compost par recherche d’ORF sur les reads avec FragGeneScan puis aggrégation d’information (tigrfam et blast)
  + Etape de réduction des données avec uclust

Paleo - metagenomic

- Analyse d’échantillons d’ADN ancien de tombes
- Recherche de la composition en espèce de ces échantillons afin de déterminer les rites funéraires
- Etape de filtrage pour dissocier l’ADN ancien des contamination contemporaines grace aux dégradation spécifiques (substitution G-> A)

Partenaires

M. Nikolski
A. Groppi

G. Valiente
D. Alonso-Alemany
S. Beretta

Mesocentre de Calcul Intensif Aquitain (MCIA)