SMPGD Satellite Day

Detection of exact sequence variants in metabarcoding by PCR abundance signal clustering.

Alexandre Wendling - PhD student alexandre.wendling@univ-grenoble-alpes.fr
Director : Clovis Galiez clovis.galiez@univ-grenoble-alpes.fr

LABORATOIRE JEAN KUNTZMANN
MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

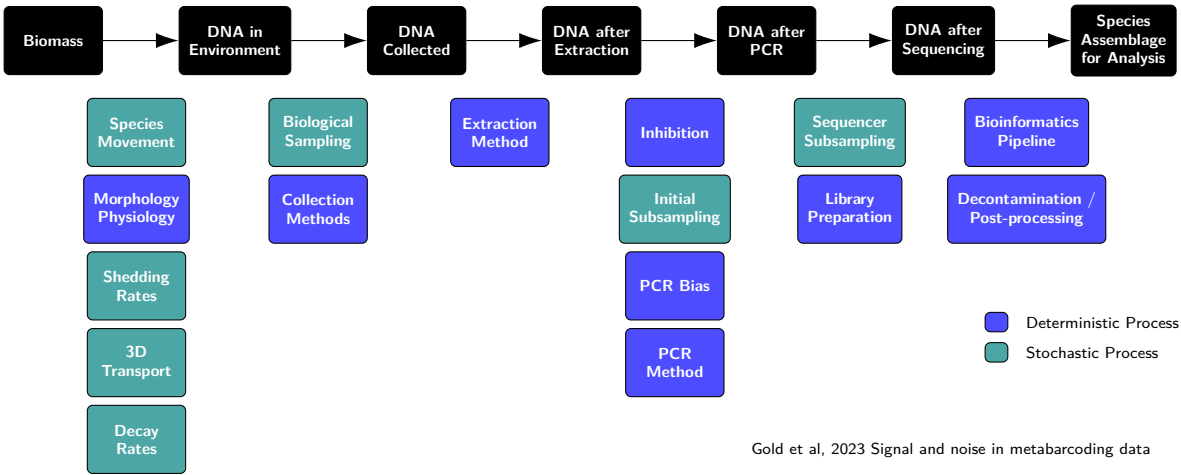
StatOmique

cnrs

GDR Groupement de recherche
BIMMM Bio-Informatique Moléculaire : Modélisation et Méthodologie

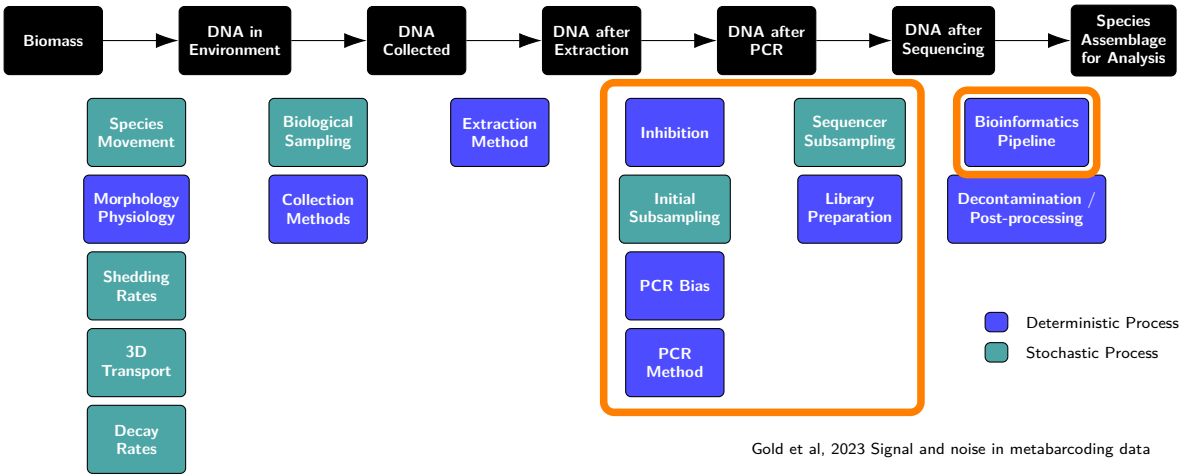
What metabarcoding & amplicon sequencing are used for

Processes of metabarcoding/amplicon sequences and bias



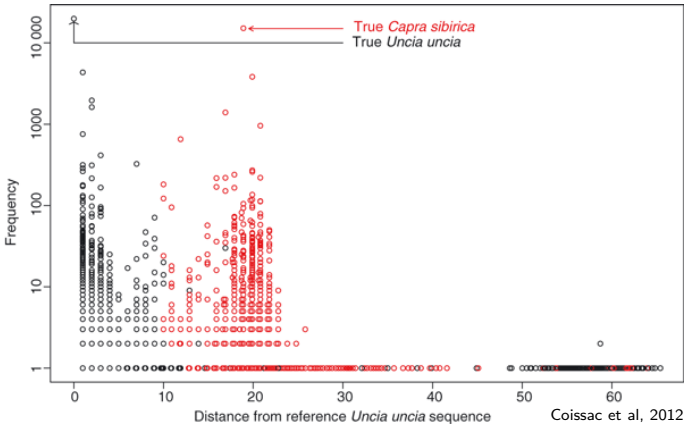
Gold et al, 2023 Signal and noise in metabarcoding data

Processes of metabarcoding/amplicon sequences and bias



Gold et al, 2023 Signal and noise in metabarcoding data

Lots of noise



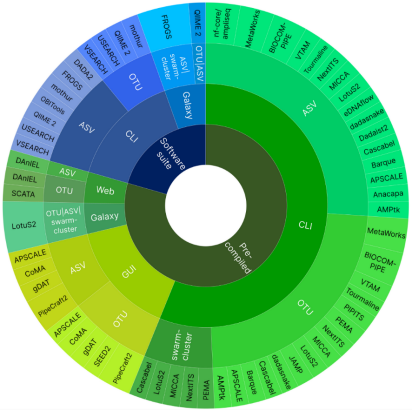
Coissac et al, 2012

- What are the true biological sequences ?
- Impact on scientific conclusions

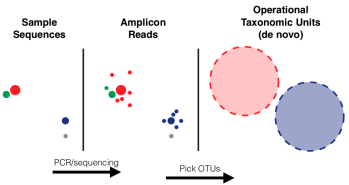
From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?

Irene Calderón-Sanou ✉ Tamara Münkemüller, Frédéric Boyer, Lucie Zinger, Wilfried Thuiller

ASV *vs* OTU



Hakimzadeh et al, 2023 : A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses



	ASVs	De novo	Closed-ref
Precise	✓	~	~
Tractable	✓	~	✓
Reproducible	✓	X	✓
Comprehensive	✓	✓	X

DADA2: Substitution Error Model (Callahan et al, 2016)

Core assumptions: each observed read is assumed to be a noisy version of a true sequence Each observed read $r = (r_1, \ldots, r_L)$ is generated from a true sequence $s = (s_1, \ldots, s_L)$ with quality scores $q = (q_1, \ldots, q_L)$:

$$P(r \mid s, q) = \prod_{i=1}^{L} p(r_i \mid s_i, q_i)$$

Abundance-based statistical test

For a candidate sequence r derived from a true sequence s:

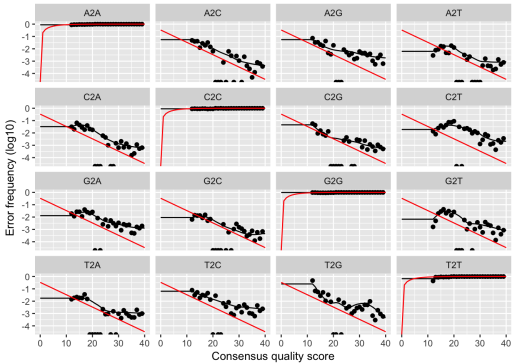
$$\lambda = N_s \cdot P(r \mid s)$$

$$n_r \sim \text{Poisson}(\lambda)$$

Sequences inconsistent with the error model are retained as ASVs.

Substitution error model

For each Phred score q, DADA2 learns an empirical substitution matrix:



(pooling between samples)

UNOISE / USEARCH: Abundance-Based Denoising (Edgar 2016)

Core assumption: true biological sequences are more abundant than their errors.

Dereplicated sequences (global pool)

Sequence	Abundance
S_1	12 450
S_2	3 120
S_3	640
S_4	58
S_5	7

Inference

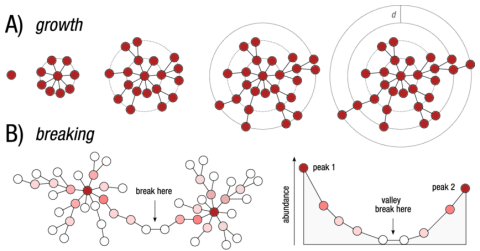
- Most abundant sequence → accepted as true
- Each rarer sequence is aligned to accepted ones
- If sequence is:
  - within small distance ($d = 1$–2)
  - and abundance ratio is low

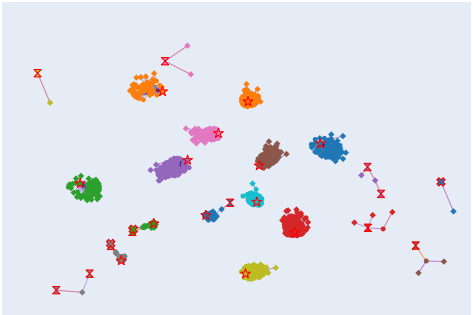
$$\frac{a_r}{a_s} \leqslant \alpha(d), \qquad \alpha(d) \approx 10^{-d}$$

⇒ rare sequence classified as sequencing error

Obiclean/Swarm : Graph of sequences (Boyer et al,2016/Mahé et al,2015)

Idea: PCR errors are close in sequence space to their source sequences and less abundant.





Graph of sequences connected by one insertion, deletion, or substitution, reasoning locally by PCR.

Classify sequences as:

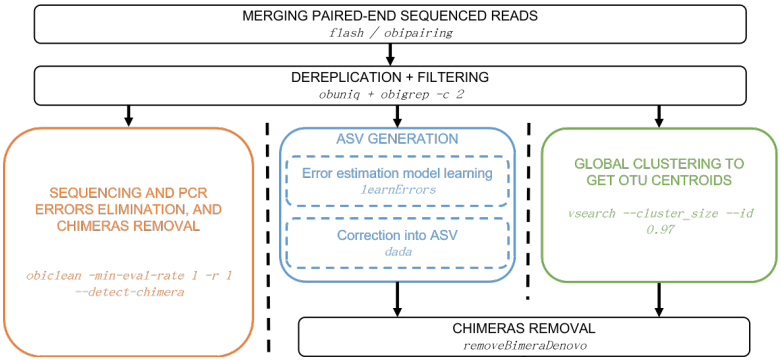
Head : most abundant in its neighborhood
Internal : connected to a more abundant sequence
Singleton : no neighbor

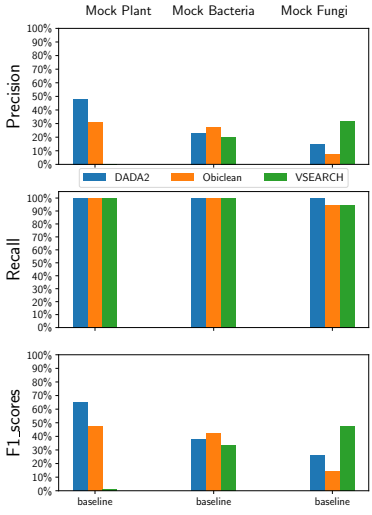
- Retain heads (and possibly singletons) as biological sequences

Test on mock community data

Mock Plants, Moinard 2023	Mock Bacteria, Gevers & HMP Consortium 2012	Mock Fungi Bakker et al. 2018
primer : chloroplast DNA	primer : 16S rRNA V4	primer : ITS1 rDNA
12 samples, 20 replicates	4 samples, 2 replicates	8 samples, 3 replicates
14 species	20 species	19 species

Traditional error/correction method





New idea to filter ASV/OTUs based on PCR replicates

Main assumptions: there are more biological signals than PCR errors.

- The variation in abundance of a biological sequence between samples is greater than its variation in abundance between PCR replicates of the same sample
- The abundance of a PCR error is determined by the abundance of the source biological sequence from which it originates, such that the variance in the error/source sequence abundance ratio is of the same order of magnitude between samples and between PCR replicates of the same sample.

Distribution of variance

Comparison of abundance variance between samples and between PCR replicates of the same sample.

The count $(\mathbf{X}_{s,r})_i$ of sequence i in sample s and replicate r is modelled by a multinomial distribution whose conjugate prior is a Dirichlet distribution.

$$\mathbf{X}_{s,r} \sim \text{Multinomial}(N_{s,r}, \mathbf{p}_{s,r}) \qquad \mathbf{p}_{s,r} \sim \text{Dirichlet}(\mathbf{X}_{s,r})$$

- First step: standalone comparison

  Intra sample,

  $$V^i_{intra} = \oplus_s \text{Var}(\log(\mathbf{p}_{s,\bullet,i}))$$

- Second step: Pairwise comparison

  Intra sample,

  $$V^{i,j}_{intra} = \oplus_s \text{Var}(\log(\frac{\mathbf{p}_{s,\bullet,i}}{\mathbf{p}_{s,\bullet,j}}))$$

Inter sample,

$$V^i_{inter} = \text{Var}(\log(\mathbf{p}_{\bullet,r*,i}))$$

$r*$ random replicate, var on 2 axes of 4D tensor

Inter sample,

$$V^{i,j}_{inter} = \text{Var}(\log(\frac{\mathbf{p}_{\bullet,r*,i}}{\mathbf{p}_{\bullet,r*,j}}))$$

$r*$ random replicate, var on 2 axes of 4D tensor

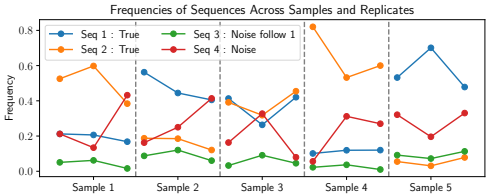
Sequence classification

- First step :
  - **Biological** : if $P(V^i_{intra} < V^i_{inter}) \geqslant 0.95$
  - **Need to check** : if $0.6 < P(V^i_{intra} < V^i_{inter}) < 0.95$
  - **Noise** : if $P(V^i_{intra} < V^i_{inter}) \leqslant 0.6$

- Second step (pairwise comparison for "Biological" and "Need to check") :
  - **Biological** : if $P(V^{i,j}_{intra} < V^{i,j}_{inter}) \geqslant \tau$
  - **Noise** : if $P(V^{i,j}_{intra} < V^{i,j}_{inter}) < \tau$
  - with τ an adjustable threshold depending on the desired sensitivity/specificity trade-off

Toy example

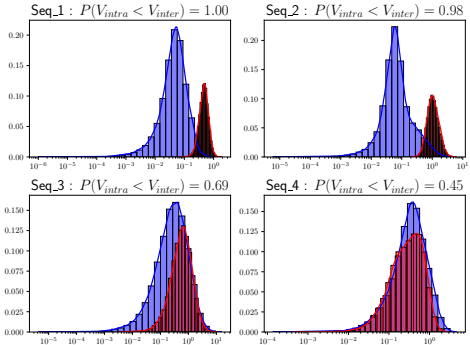
	S1			S2			S3			S4			S5		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Seq 1	23	20	21	45	47	47	38	29	38	9	13	12	58	67	56
Seq 2	56	58	48	15	20	14	36	35	41	73	58	62	6	3	9
Seq 3	6	6	2	7	13	7	3	10	4	2	4	1	10	7	13
Seq 4	23	13	54	13	26	48	15	36	7	5	34	28	35	19	39

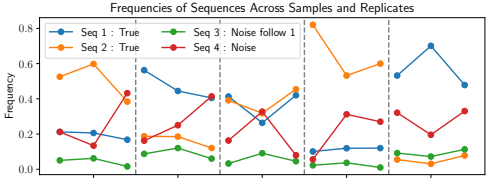


Toy example

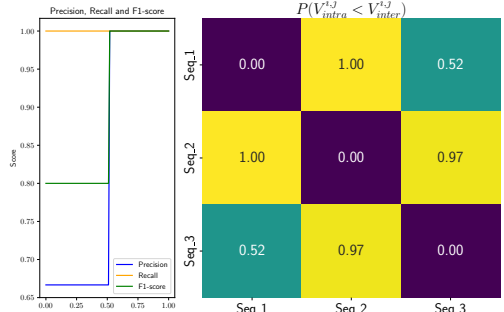
	S1			S2			S3			S4			S5		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Seq 1	23	20	21	45	47	47	38	29	38	9	13	12	58	67	56
Seq 2	56	58	48	15	20	14	36	35	41	73	58	62	6	3	9
Seq 3	6	6	2	7	13	7	3	10	4	2	4	1	10	7	13
Seq 4	23	13	54	13	26	48	15	36	7	5	34	28	35	19	39

First step :



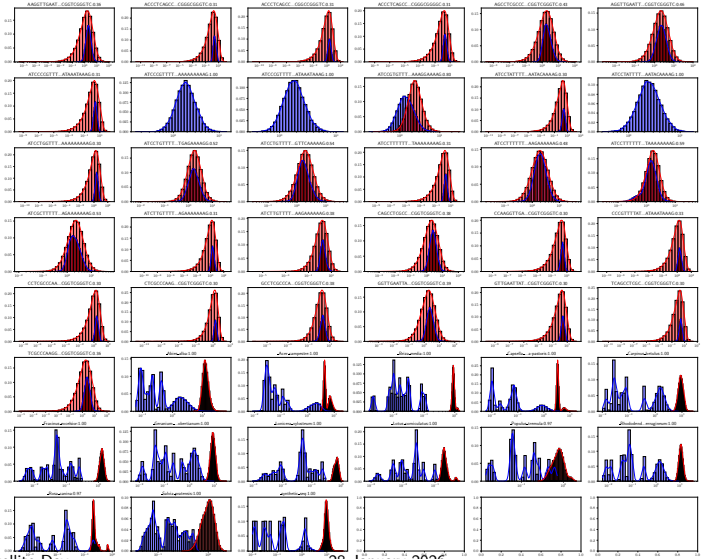


Second step :

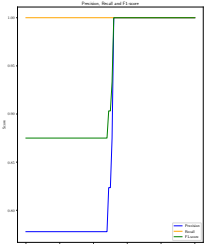


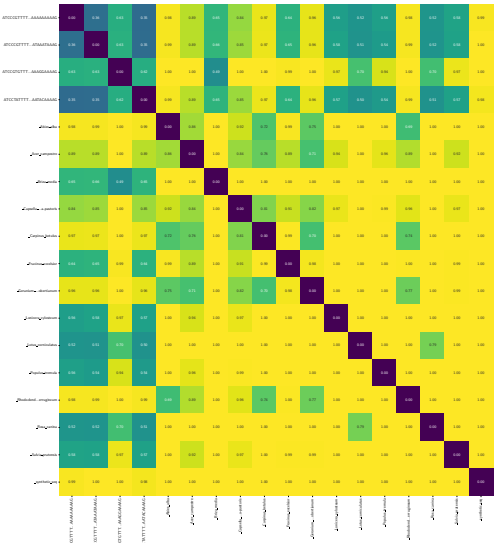
Mock Plants

Number of reads	10 529 455
Number of sequences	91 859
Number of sequences with count > 1	28 035
Number of heads	661
Number of heads after chimera	184
Number of heads in at least 2 replicates	45
Number of sequences after dada2 pipeline	56
Number of true biological sequences	14



Mock Plants





Check noise association through sequence identity

ATCCGTGTTTT...CAAAGGAAAAG: (Briza media, 0.96), (Rhododendron ferrugineum, 0.72), (Geranium robertianum, 0.71)

ATCCCGTTTTA...AAAAAAAAAAG: (Rosa canina, 0.91), (Salvia pratensis, 0.85), (Geranium robertianum, 0.76)

ATCCCGTTTTA...AATAAATAAAG: (Rosa canina, 0.91), (Carpinus betulus, 0.85), (Geranium robertianum, 0.77)

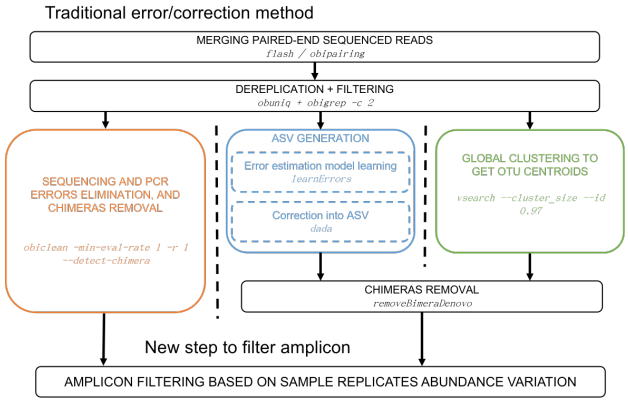
ATCCTATTTTC...GAATACAAAAG: (Populus tremula, 0.96), (Carpinus betulus, 0.79), (Acer campestre, 0.66)

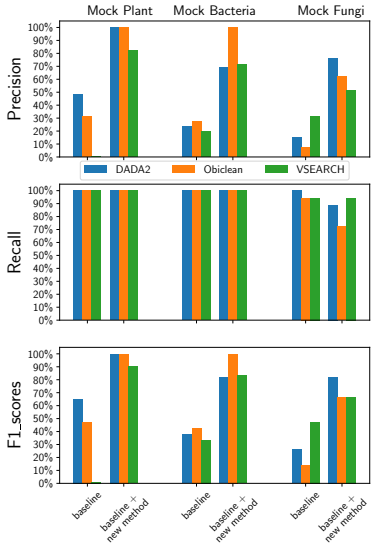
Test on mock community data

Mock Plants,
Moinard 2023
primer : chloroplast DNA
12 samples, 20 replicates
14 species

Mock Bacteria,
Gevers & HMP Consortium 2012
primer : 16S rRNA V4
4 samples, 2 replicates
20 species

Mock Fungi
Bakker et al. 2018
primer : ITS1 rDNA
8 samples, 3 replicates
19 species





Current and Ongoing work

Conclusion:

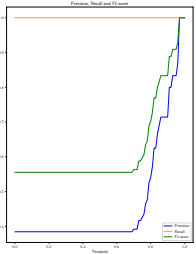
- New approach using PCR replicate information
- Better description of species richness
- Complementary to taxonomic assignment, where there is a bias towards exhaustiveness and choice of sequence identity threshold
- Possibility of finding 'unknown' biological sequences

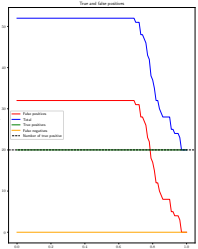
Bootstraping approch to simulate distributions of variance is cumbersome

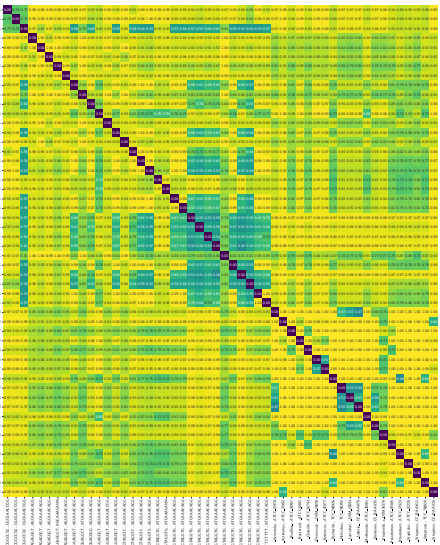
Ongoing work:

- Modeling statisticaly the distributions and take into account:
- Compositional nature of the data,
- Heteroscedasticity (variance magnitude impact by the mean)
- Correlations between samples

Mock Bacteria, pairwise variance example







Taxonomy	Even1	Even2	Staggered1	Staggered2
Acinetobacter_baumannii_ATCC_17978	0.047619	0.047619	0.002144	0.002144
Actinomyces_odontolyticus_ATCC_17982	0.047619	0.047619	0.000214	0.000214
Bacillus_cereus_ATCC_10987	0.047619	0.047619	0.021436	0.021436
Bacteroides_vulgatus_ATCC_8482	0.047619	0.047619	0.000214	0.000214
Clostridium_beijerinckii_ATCC_51743	0.047619	0.047619	0.021436	0.021436
Deinococcus_radiodurans_DSM_20539_	0.047619	0.047619	0.000214	0.000214
Enterococcus_faecalis_ATCC_47077	0.047619	0.047619	0.000214	0.000214
Escherichia_coli_ATCC_700926	0.047619	0.047619	0.214362	0.214362
Helicobacter_pylori_ATCC_700392	0.047619	0.047619	0.002144	0.002144
Lactobacillus_gasseri_DSM_20243	0.047619	0.047619	0.002144	0.002144
Listeria_monocytogenes_ATCC_BAA-679	0.047619	0.047619	0.002144	0.002144
Methanobrevibacter_smithii_ATCC_35061	0.047619	0.047619	0.214362	0.214362
Neisseria_meningitidis_ATCC_BAA-335	0.047619	0.047619	0.002144	0.002144
Propionibacterium_acnes_DSM16379	0.047619	0.047619	0.002144	0.002144
Pseudomonas_aeruginosa_ATCC_47085	0.047619	0.047619	0.021436	0.021436
Rhodobacter_sphaeroides_ATCC_17023	0.047619	0.047619	0.021436	0.021436
Staphylococcus_aureus_ATCC_BAA-1718	0.047619	0.047619	0.214362	0.214362
Staphylococcus_epidermidis_ATCC_12228	0.047619	0.047619	0.021436	0.021436
Streptococcus_agalactiae_ATCC_BAA-611	0.047619	0.047619	0.214362	0.214362
Streptococcus_mutans_ATCC_700610	0.047619	0.047619	0.000214	0.000214
Streptococcus_pneumoniae_ATCC_BAA-334	0.047619	0.047619	0.021436	0.021436