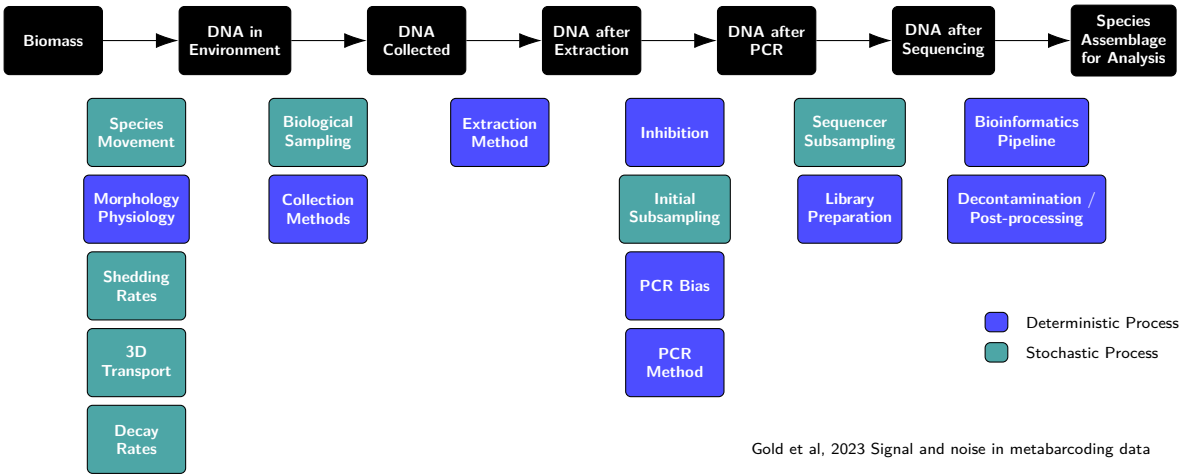
Non-Permanent Seminar

Analysis of variance on genomic abundance data

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

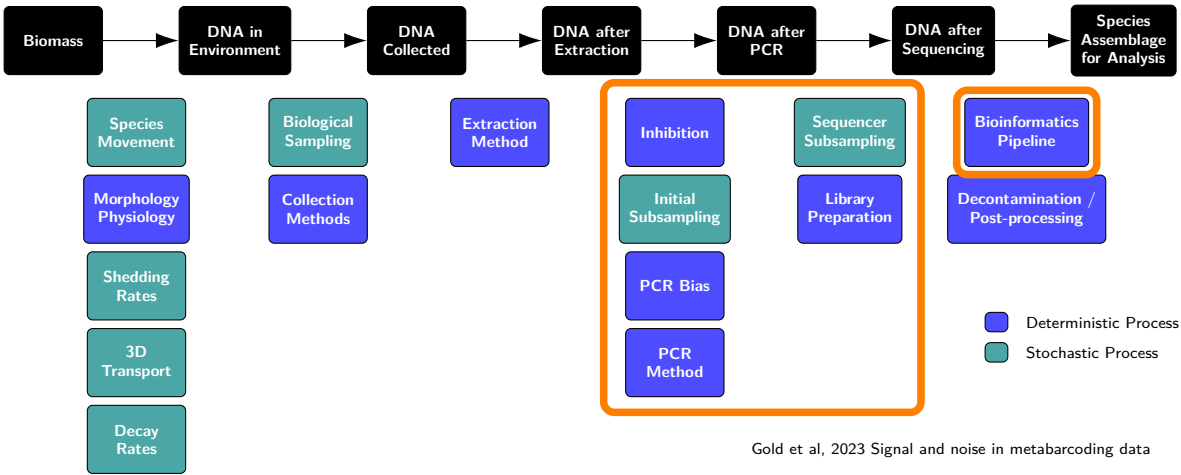
What metabarcoding & amplicon sequencing are used for

Processes of metabarcoding/amplicon sequences and bias



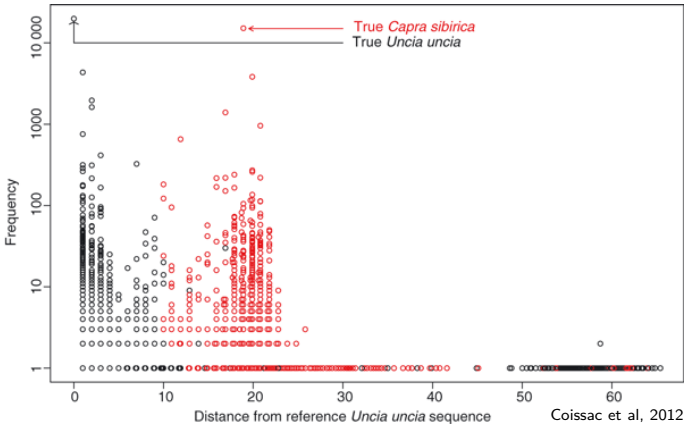
Gold et al, 2023 Signal and noise in metabarcoding data

Processes of metabarcoding/amplicon sequences and bias



Gold et al, 2023 Signal and noise in metabarcoding data

Lots of noise



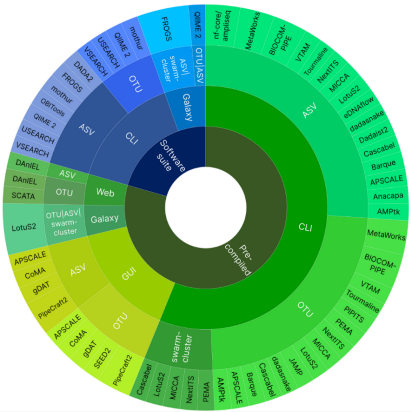
Coissac et al, 2012

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

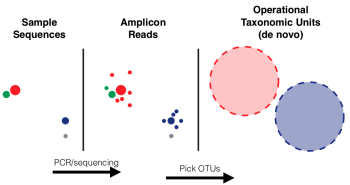
From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?

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:

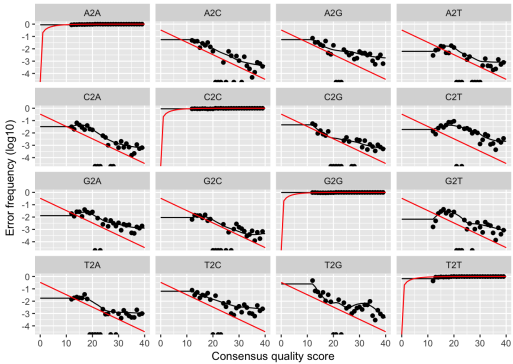
$$\lambda = N_s \cdot P(r \mid s)$$

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

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

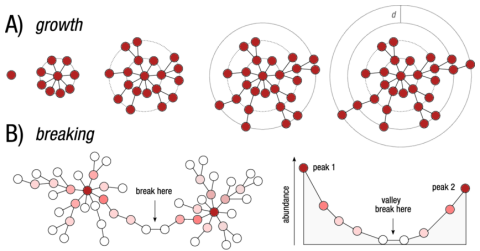
- Most abundant sequence → accepted as true
- Each rarer sequence is aligned to accepted ones
- If sequence is:
  - within small distance ($d = 1$–2)
  - and abundance ratio is low

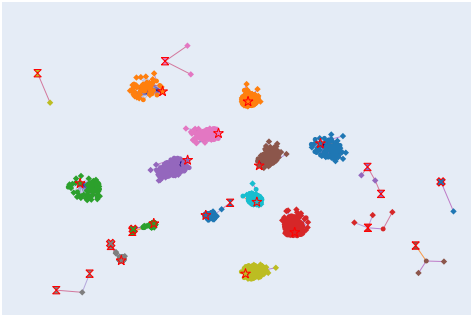
$$\frac{a_r}{a_s} \leqslant \alpha(d), \qquad \alpha(d) \approx 10^{-d}$$

⇒ 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.

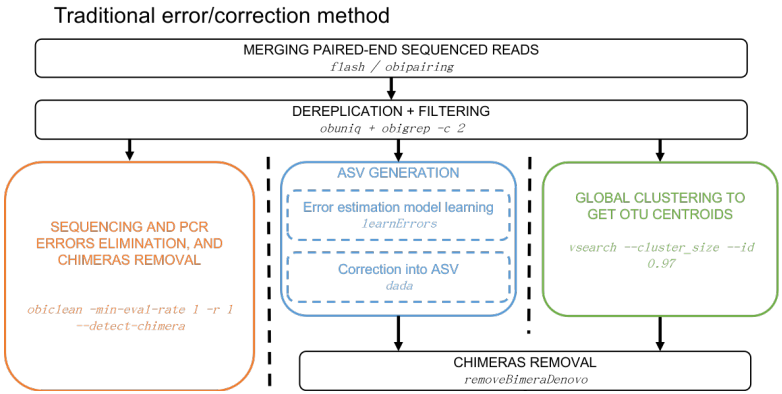
Classify sequences as:

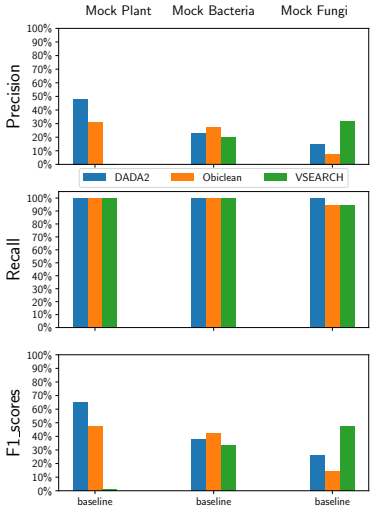
Head : most abundant in its neighborhood
Internal : connected to a more abundant sequence
Singleton : no neighbor

- 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





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.

We want to model and compare the within-sample variance of each sequence across replicates and the between-sample variance of each sequence across samples, taking into account the compositionality, zero inflation, heteroscedasticity and correlation between samples as it is expected to have lower variability between non-independent samples.

Statistical modelling

We define the vector of non-zero counts in replicate r of sample s, where $n_{s,r} = \sum_{i \in I_s} x_{i,s,r}$ and estimate $\mathbf{p}_{i,s,r}$ by the observed proportions $\hat{\mathbf{p}}_{s,r} = \frac{\mathbf{x}_{s,r}}{n_{s,r}}$.

We then apply the log transformation on the proportions to obtain the log-ratio transformed data to deal with the compositionality (abundance data are in a simplex):

We model the log-ratio transformed data $\phi_{i,s,r}$ as follows :

where ϕ_i is the overall mean of sequence i across samples and replicates, estimated as :

$b_{i,s}$ is the sample-level random effect for sequence i in sample s :

And $\epsilon_{i,s,r}$ is the replicate-level random effect for sequence i in replicate r of sample s :

$$(x_{i,s,r})_{i \in I_s} = \mathbf{X}_{s,r} \sim \text{Multinomial}(n_{s,r}, \mathbf{p}_{s,r}) \quad (1)$$

$$\phi_{i,s,r} = \log(\frac{x_{i,s,r}}{n_{s,r}}) \quad (2)$$

$$\phi_{i,s,r} = \phi_i + b_{i,s} + \epsilon_{i,s,r} \quad (3)$$

$$\hat{\phi}_i = \mathbb{E}_{S_i}[\mathbb{E}_r[\phi_{i,s,r}]] = \frac{1}{|S_i|} \sum_{s \in S_i} \frac{1}{R_s} \sum_{r=1}^{R_s} \phi_{i,s,r} \quad (4)$$

$$b_{i,s} = (B_i)_s \; B_i \sim \mathcal{N}(\mu_{|S_i}, \Sigma_{|S_i}) \quad (5)$$

$$\epsilon_{i,s,r} \sim \mathcal{N}(0, \sigma^2_{i,s}) \quad (6)$$

Within-sample variance

For each sample s, we model $\phi_{i,s,r}$ considering $\phi_{i,s} = \phi_i + b_{i,s}$, the mean of sequence i in sample s, so that :

$$\phi_{i,s,r} = \phi_{i,s} + \epsilon_{i,s,r} \quad \hat{\phi_{i,s}} = \mathbb{E}_r[\phi_{i,s,r}] = \frac{1}{R_s}\sum_{r=1}^{R_s} \phi_{i,s,r} \tag{7}$$

We modelize the heteroscedasticity at sample level, so for each sample s, we model the variance $\sigma^2_{i,s}$ as a function of the mean $\phi_{i,s}$ and adopt a log-linear variance function:

$$\log \sigma^2_{i,s} = \beta_{0_s} + \beta_{1_s}\phi_{i,s}. \tag{8}$$

On the sample s the observed within-sample variance for sequence i is estimated as:

$$s^2_{w\_i,s} = \frac{1}{R_s - 1}\sum_{r=1}^{R_s}(\phi_{i,s,r} - \hat{\phi}_{i,s})^2 \tag{9}$$

following the distribution:

$$s^2_{w\_i,s} \sim S^2_{w\_i,s} = \frac{\sigma^2_{i,s}}{R_s - 1}\chi^2_{R_s-1} \tag{10}$$

For the global within-sample variance across samples for sequence i, we define $U_i \sim \mathcal{U}\{1, \dots, |S_i|\}$ a uniform discrete random variable over the samples and define $S^2_{w_i}$ as:

$$S^2_{w\_i,s} = S^2_{w_i} | U_i = s \tag{11}$$

So that $S^2_{w_i}$ is a uniformly distributed mixture of the $S^2_{w\_i,s}$ across samples, and we can estimate it by the mean of the observed within-sample variances across samples:

$$S^2_{w\_i} = \frac{1}{|S_i|}\sum_{s \in S_i} S^2_{w\_i,s} \tag{12}$$

Between-sample variance

To modelize the distribution of the between-sample variance, we can see the log-ratio transformed data as a global S-variate normal distribution $\Phi \sim \mathcal{N}(\mu, \Sigma)$. For each pair of samples (s, s'), we denote $I_{s,s'} = I_s \cap I_{s'}$

$$\mu_s = \mathbb{E}_{I_s}[\mathbb{E}_r[\phi_{i,s,r}]] = \frac{1}{|I_s|R_s} \sum_{i \in I_s} \sum_{r=1}^{R_s} \phi_{i,s,r} \tag{13}$$

$$\text{Cov}(\Phi_s, \Phi_{s'}) = \frac{1}{|I_{s,s'}|R_sR_{s'}} \sum_{i \in I_{s,s'}} \sum_{r=1}^{R_s} \sum_{r'=1}^{R_{s'}} (\phi_{i,s,r} - \phi_s^{(ss')})(\phi_{i,s',r'} - \phi_{s'}^{(ss')}) \quad \phi_s^{(ss')} = \frac{1}{|I_{s,s'}|R_sR_{s'}} \sum_{i \in I_{s,s'}} \sum_{r=1}^{R_s} \sum_{r'=1}^{R_{s'}} \phi_{i,s,r} \tag{14}$$

To study the between-sample variance of sequence i, we consider the restricted multivariate normal distribution $\Phi_i \sim \mathcal{N}(\mu_{|S_i}, \Sigma_{|S_i})$ where $\mu_{|S_i}$ and $\Sigma_{|S_i}$ are the mean vector and covariance matrix restricted to the samples where sequence i is observed.
Moreover, we want to reduce the correlation between samples if few sequences are shared between them. Thus, we regularize the covariance matrix $\Sigma_{|S_i}$ through a shrinkage approach :

$$\Sigma^*_{|S_i} = (1 - \alpha)\Sigma_{|S_i} + \alpha \text{diag}(\Sigma_{|S_i}) \tag{15}$$

Finally, the between-sample variance of sequence i is modeled as the variance of the multivariate normal distribution $\Phi_i \sim \mathcal{N}(\mu_{|S_i}, \Sigma^*_{|S_i})$.

Between-sample variance

The between-sample variance across samples for sequence i, $S^2_{b\_i}$ is defined as:

$$S^2_{b\_i} = \mathrm{Var}(\Phi_i) = \mathbb{E}[(\Phi_i - \mathbb{E}[\Phi_i])^2] = \frac{1}{|S_i|}\|\Phi_i - P_i\Phi_i\|^2 = \frac{1}{|S_i|}\|(I - P_i)\Phi_i\|^2 \tag{16}$$

where $P_i = \left(\frac{1}{|S_i|}\right)_{|S_i|\times|S_i|}$, so $(I - P_i)\Phi_i$ correspond to Φ_i centered by its mean across samples.

We denote $Y_i = (I - P_i)\Phi_i$, so $Y_i \sim \mathcal{N}\left((I - P_i)\mu_{|i}, (I - P_i)\Sigma^*_{|i}(I - P_i)^\top\right)$ and $S^2_{b\_i} = \frac{1}{|S_i|}\|Y_i\|^2$ follow a generalized chi-squared distribution. To express this distribution, we can perform the spectral decomposition of the covariance matrix of Y_i : $(I - P_i)\Sigma^*_{|i}(I - P_i)^\top = Q\Lambda Q^\top$ and we denote $Z_i = \sqrt{\Lambda^{-1}}Q^\top Y_i \sim \mathcal{N}(\sqrt{\Lambda^{-1}}Q^\top(I - P_i)\mu_{|i}, I)$ the isotropic multivariate normal distribution. So finally :

$$\begin{aligned} S^2_{b\_i} &= \frac{1}{|S_i|}\|Y_i\|^2 = \frac{1}{|S_i|}Y_i^\top Y_i = \frac{1}{|S_i|}Z_i^\top \Lambda Z_i = \sum_{k=1}^{|S_i|}\frac{\lambda_k}{|S_i|}Z^2_{i,k} \\ S^2_{b\_i} &\sim \sum_{k=1}^{|S_i|}\frac{\lambda_k}{|S_i|}\chi^2_1(\delta_k) \quad \text{where } \delta_k = (\sqrt{\Lambda^{-1}}Q^\top(I - P_i)\mu_{|i})^2_k \end{aligned} \tag{17}$$

λ_k are the eigenvalues of the covariance matrix of Y_i and δ_k are the non-centrality parameters of the chi-squared distribution.

Comparison of variance

We can compute several quantities of interest, that could be expressed as quadratic forms such as :

- $P(S^2_{b\_i} < S^2_{w\_i})$: the probability that the between-sample variance is lower than the within-sample variance for sequence i, which is a measure of how likely this sequence is to be signal rather than noise.

$$P(S^2_{b\_i} < S^2_{w\_i}) = \frac{1}{|S_i|} \sum_{s \in S_i} P\left(\sum_{k=1}^{|S_i|} \frac{\lambda_k}{|S_i|} \chi^2_1(\delta_k) - \frac{\sigma^2_{i,s}}{R_s - 1} \chi^2_{R_s - 1} < 0 \right) \tag{18}$$

- $P(S^2_{w\_i} > s^2_{b\_i})$: the probability that the within-sample variance is greater than the observed between-sample variance, which is a measure of how likely this sequence is to be noise rather than signal.

$$P(S^2_{w\_i} > s^2_{b\_i}) = \frac{1}{|S_i|} \sum_{s \in S_i} P\left(\frac{\sigma^2_{i,s}}{R_s - 1} \chi^2_{R_s - 1} > s^2_{b\_i} \right) \tag{19}$$

- $P(S^2_{b\_i} < s^2_{w\_i})$: the probability that the between-sample variance is less than the observed within-sample variance, which is another measure of how likely this sequence is to be noise rather than signal.

$$P(S^2_{b\_i} < s^2_{w\_i}) = \frac{1}{|S_i|} \sum_{s \in S_i} P\left(\sum_{k=1}^{|S_i|} \frac{\lambda_k}{|S_i|} \chi^2_1(\delta_k) < s^2_{w\_i,s} \right) \tag{20}$$

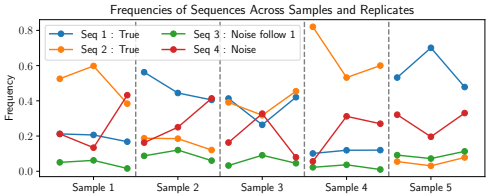
Pairwise comparison of variance

We can also perform pairwise comparisons of variance between sequences i and j to study the relative abundance of these two sequences across samples in a way that if the ratio have a low variance, the two sequences are linked as they have a similar profile across samples, so one can be an error of the other.
Studing :

$$Var(\log(\frac{p_{i,s}}{p_{j,s}})) = Var(\phi_{i,s} - \phi_{j,s}) \tag{21}$$

Toy example

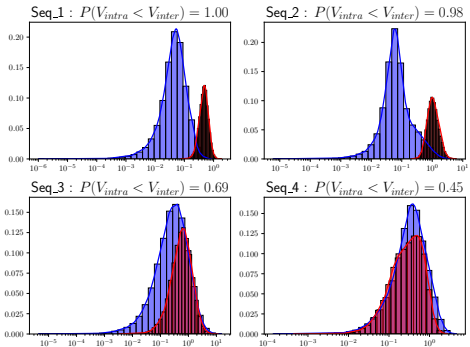
	S1			S2			S3			S4			S5		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Seq 1	23	20	21	45	47	47	38	29	38	9	13	12	58	67	56
Seq 2	56	58	48	15	20	14	36	35	41	73	58	62	6	3	9
Seq 3	6	6	2	7	13	7	3	10	4	2	4	1	10	7	13
Seq 4	23	13	54	13	26	48	15	36	7	5	34	28	35	19	39

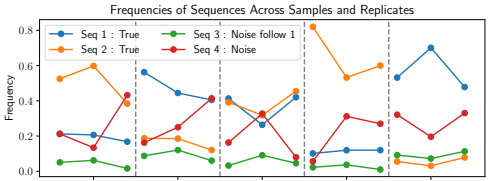


Toy example

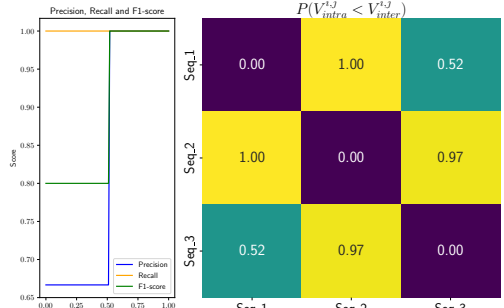
	S1			S2			S3			S4			S5		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
Seq 1	23	20	21	45	47	47	38	29	38	9	13	12	58	67	56
Seq 2	56	58	48	15	20	14	36	35	41	73	58	62	6	3	9
Seq 3	6	6	2	7	13	7	3	10	4	2	4	1	10	7	13
Seq 4	23	13	54	13	26	48	15	36	7	5	34	28	35	19	39

First step :



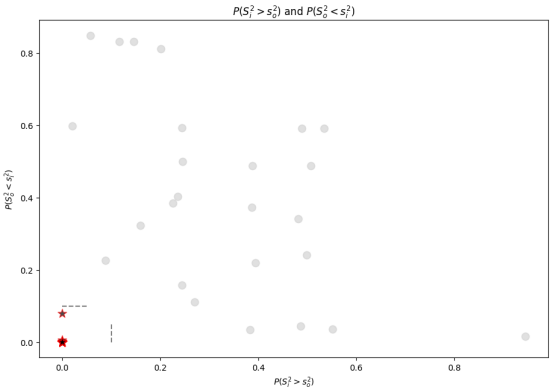


Second step :

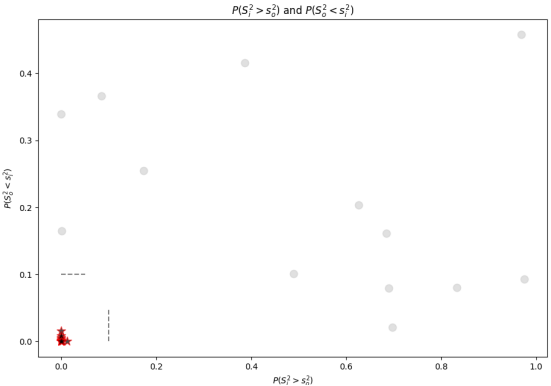


Results on mock community data

Example on bacterial mock :



Example on plant mock :

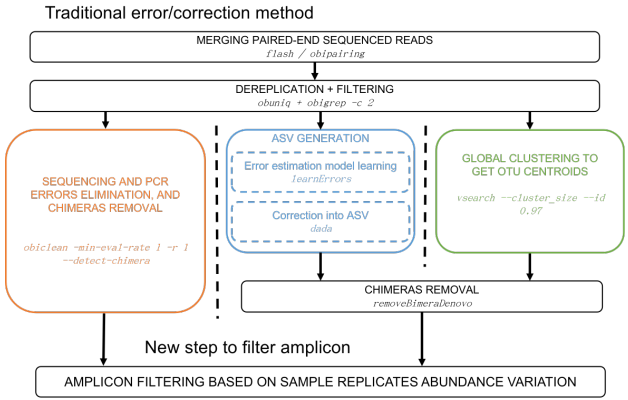


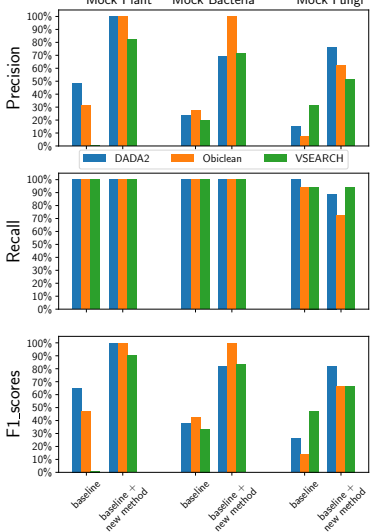
Test on mock community data

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

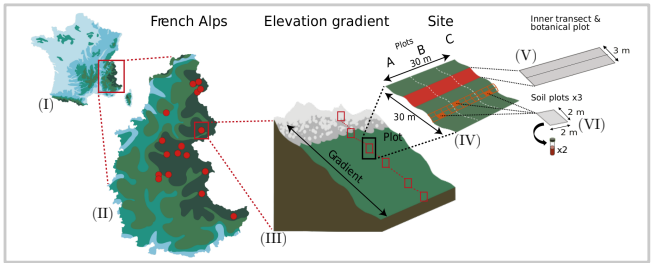
Mock Bacteria,
Gevers & HMP Consortium 2012
primer : 16S rRNA V4
4 samples, 2 replicates
20 species

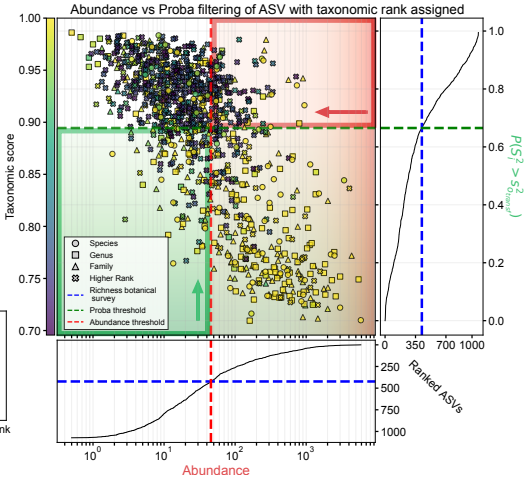
Mock Fungi
Bakker et al. 2018
primer : ITS1 rDNA
8 samples, 3 replicates
19 species

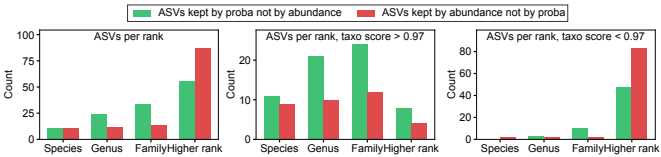




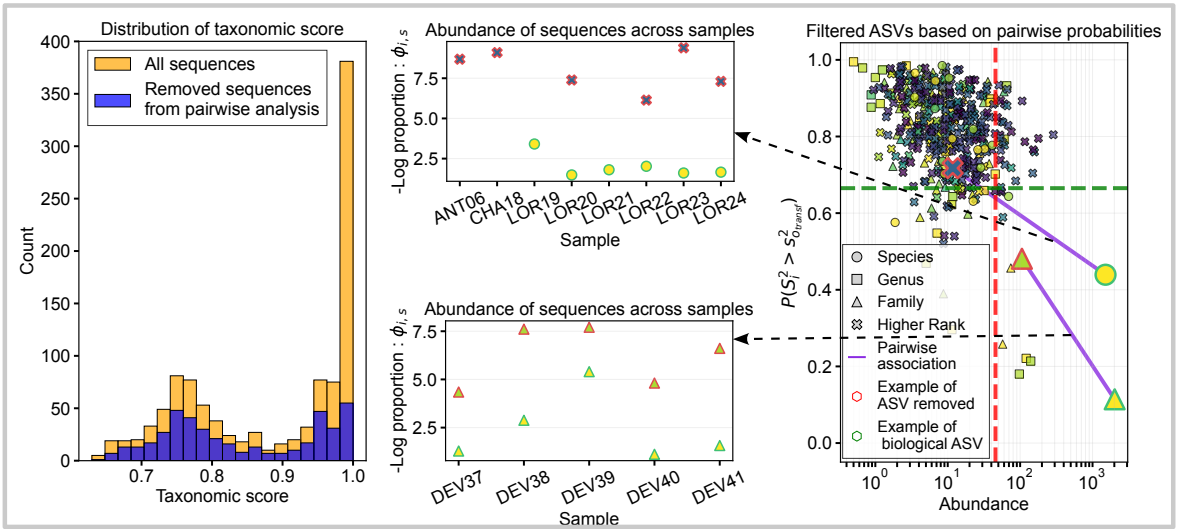
Results on fieldwork campaign data



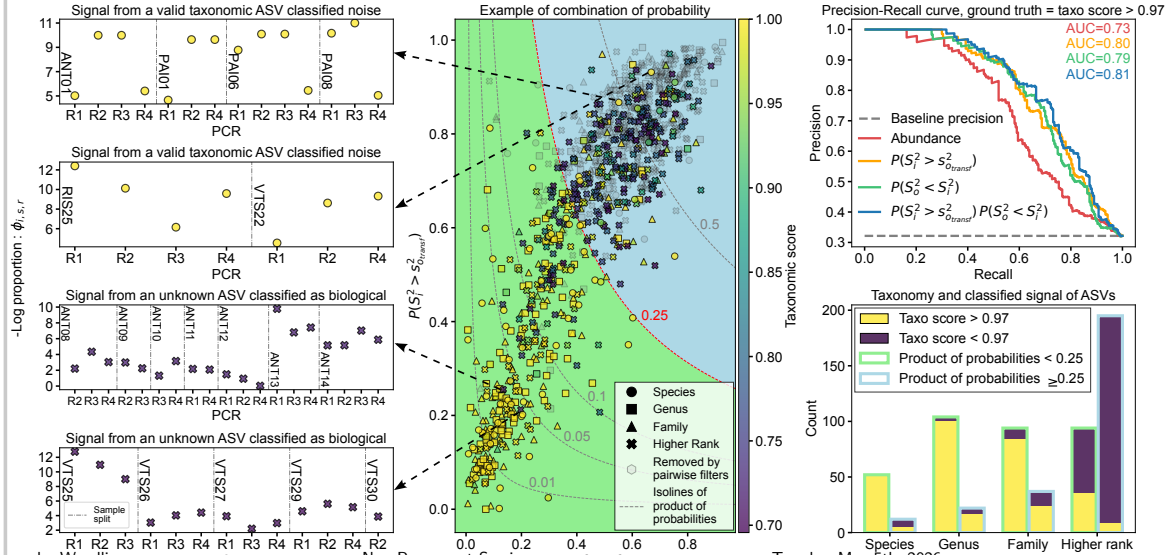




Results on fieldwork campaign data



Results on fieldwork campaign data



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