Detection of mutated primers and impact on targeted metagenomics results

Aymeric Antoine-Lorquin, Frédéric Mahé, Micah Dunthorn, Catherine Belleannée

Detection of mutated primers and impact on targeted metagenomics results

Aymeric Antoine-Lorquin[1], Frédéric Mahé[2], Micah Dunthorn[3] and Catherine Belleannée[1]

1- *Dyliss - Dynamics, Logics and Inference for biological Systems and Sequences, Inria, Rennes University, FRANCE*
2- LSTM - Laboratoire des symbioses tropicales et méditerranéennes, CIRAD, FRANCE
3- Technische Universität Kaiserslautern, GERMANY

1 Introduction

High-throughout sequencing platforms are widely used in metabarcoding studies of environmental microbial diversity as they can quickly produce millions of reads (e.g., [de Vargas et al. 2015, Lax et al. 2014, Tedersoo et al. 2014]). Resulting reads in these studies are clustered into molecular operational taxonomic units (OTUs) and compared statistically. Before the raw reads can be clustered and compared, though, they are passed through various cleaning and removal steps. One of these cleaning steps is the identification of primer sequences: if the primer cannot be detected the read is usually removed.

Some previous studies, particularly by [Huse et al. 2007], have shown that eliminating some suspects reads, such as those containing ambigous bases ("N"), inexact primers or with anormal length, could improve the overall quality of a sample . Then, in order to limit the number of spurious OTUs retrieved from samples, eliminating the reads with mutated primers has become the norm. This process has the advantage of not requiring the use of complex tools, since it is possible to search for exact primers with simple regular expressions natively supported by many programming languages (python, perl, ruby, etc.). However, this strategy may also eliminate correct sequences and this practice raises questions: Does the removal of all reads with mutated primers cause information loss? Or, in a more practical perspective : now there are tools to reject the less reliable sequences in clustering as SWARM [Mahe et al. 2015b], is there an interest to seek mutated primers? Can it bring new sequences? Are these sequences relevant? Can it contribute to detect more species?

Such are the questions that we try to answer in this article, through a metagenomic analysis that estimates eukaryotic soil biodiversity.

To try to answer these questions, we analysed data on tropical soils to study the impact on metabarcoding results of keeping reads with mutated primers. The study shows that keeping such reads allows identifying more sequences. The majority of the new sequences are quite similar to sequences with exact primers. A minority of the new sequences contributes to validate new clusters as potential species signatures.

2 Materials and methods

2.1 High-throughout sequencing datasets

We used data from nine soil samples collected in a Neotropical rainforest by [Mahe et al. 2015a]: Sample 1 (L020, Lat. 10.408167, Long. -84.019564), Sample 2 (L030, 10.418269,-84.011220),

Sample 3 (L040, 10.424040,-84.006255), Sample 4 (L050, 10.420614,-84.009819), Sample 5 (L060, 10.432789,-84.010707), Sample 6 (L070, 10.422997,-84.021248), Sample 7 (L080, 10.414865,-84.026901), Sample 8 (L090, 10.417751,-84.025404), and Sample 9 (L100, 10.424392,-84.039239). A tenth sample was not used during analysis because it contains a majority of sequences with mutated primers. We think that this sample form a too much particular case that reflect not the common reality.

Full details of sampling and sequencing can be found in [Mahe et al. 2015a]. In brief, each sample was amplified using [Stoeck et al. 2010] general eukaryotic primers for the V4 hyper-variable region of the 18S rRNA (TAReuk454FWD1 and TAReuk-Rev3). Each amplification was then sequenced with both Roche/454 GS FLX+ with Titanium chemistry and Illumina MiSeq with v3 chemistry. For Roche/454 reads, Sffinfo was used to demultiplex and convert flowgram files to fna and qual files. For Illumina MiSeq reads, PEAR v0.9.0 [Zhang et al. 2014] was used with default parameter to assemble fastq files, which were then converted to fasta format.

The 454/Roche sequencing results of the 9 samples are merged into one dataset called "454/Roche dataset", for a total of 310,375 sequences. The Illumina/MiSeq sequencing results of the 9 samples are merged into one dataset called "Illumina/MiSeq dataset", for a total of 5,223,138 sequences.

2.2 Exact and mutated primer detection

2.2.1 Exact V4 primers

The universal forward and reverse primers for the V4 region are given by the regular expressions (the variable nucleotides are indicated in parentheses):

Exact_V4F: CCAGCA[GC]C[CT]GCGGTAATTCC
Exact_V4R: CTTTCGTTCTTGAT[CT][AG]A

Fig M3: exact models for V4F and V4R primers.

The V4F_exact model recognizes the following fourth sequences CCAGCA**G**C**C**GCGGTAATTCC, CCAGCA**G**C**T**GCGGTAATTCC, CCAGCA**C**C**C**GCGGTAATTCC, CCAGCA**C**C**T**GCGGTAATTCC. The V4R_exact model recognizes the following fourth sequences CTTTCGTTCTTGAT**CA**A, CTTTCGTTCTTGAT**CG**A, CTTTCGTTCTTGAT**TA**A, CTTTCGTTCTTGAT**TG**A.

The search for exact primers is done using a Python script. It takes as inputs the regular expression of the primer, a direction of research (5' for the forwad primer, 3' for the reverse primer) and a set of sequences. It gives two sets as output: a first one with the sequences without the targeted primer, a second one with the sequences with the targeted primer. In this set, the first hit found on each sequence is trimmed.

The Python script is launched twice: a first one to search and trim the forward primer on the set of sequences, and a second one for the reverse primer.

The Python script is given in supplementary data (cf supplementary data #1).

2.2.2 Mutated V4 primers

By observing the reads from the Neotropical rainforest soils, common variants of the primers in high-throughput sequencing data were identified (such as CCAGCA**G**C**C**ACGGTAATTCC, CCAGCA**G**C**C**GCG-TAATTCC…).

These observations were used to develop the following mutation templates for the primers:

Mutated_V4F: CCAGCA[GC]C[CT]GCGGTAATTCC up to 2 substitutions OR up to 1 insertion / deletion
Mutated_V4R: CTTTCGTTCTTGAT[CT][AG]A (up to 2 substitutions OR up to 1 insertion / deletion) and (up to 2 truncated nucleotides in 3 ')

Fig M4: mutated models for V4F and V4R primers.

The search for mutated primers is done using a grammatical pattern matching tool, Logol [Belleannee et al. 2014] which allows a full control on the specifics of the model. The Logol grammars for each primer are given in supplementary data (cf supplementary data #5).

Actually, the grammar for the mutated model looks for either a pattern with a maximum of 2 substitutions or a pattern with a maximum of 1 insertion/deletion.

2.3 Workflow for exact and mutated primer detection

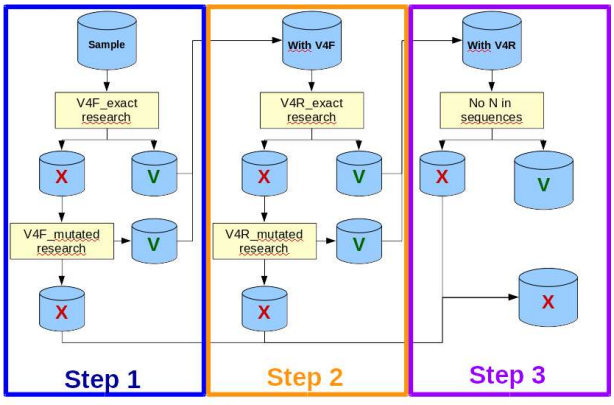

Fig M1: Exact and mutated primers detection workflow. Step 1 detects and trims the V4F primer, then Step 2 detects and trims the V4R primer, finally, Step 3 discards reads with ambiguous nculeotides

For each technology, all nine samples are merged for the analysis.

Exact and mutated primers are searched in reads using the following workflow:

- **Step1: V4F detection**
  Reads are analysed with the V4F_exact model using regular expression in order to detect and trim an exact V4F primer. Sequences without exact V4F primer are analysed with

V4F_mutated model using Logol in order to detect and trim a mutated V4F primer. Sequences with a V4F primer go to the step 2 while sequences without V4F form the group of rejected sequences.

- **Step2: V4R detection**
  Reads are analysed with the V4R_exact model using regular expression in order to detect and trim an exact V4R primer. Sequences without exact V4R primer are analysed with V4R_mutated model using Logol in order to detect and trim a mutated V4R primer. Sequences with a V4R primer go to the step 3 while sequences without V4F join the group of rejected sequences.
- **Step3: 'N' clearing**
  Reads which contain ambiguous nucleotides ('N') join the group of rejected sequences. Others form the group of valid sequences. This verification takes place at the end of the process in order to preserve the reads which would have 'N' only positionned in primers (i.e. primers with ambigous nucleotides can be detected by V4_mutated models).

At the end of the process, we consider two sets: the "**valid sequences**", with both V4F and V4R primers detected and the "**rejected sequences**", with at least one undetected primer or with internal 'N'.
The internal part of a valid sequence, after primer trimming, will be now called **amplicon**.
The "valid sequence" dataset can be splitted into two subdatasets: the "**previous amplicons**", with both exact V4F and V4R primers detected, and the "**new amplicons**", with at least one primer detected using mutated models.

2.4 Method for comparing two populations of sequences: Bray-Curtis dissimilarity test

The Bray-Curtis dissimilarity test is used to view on a graph the similarity of two sets of sequences. It is based on a side-by-side comparison of sequence profiles of two sets of identical size.
Applied to our case-study, this test allows estimating wether the newly detected amplicons are broadly similar to the previous ones. That is to say, if the set of new amplicons are biologically plausible in relation to the set of previous sequences or if it is composed of very distant sequences and therefore probably biologically false.

The new amplicons being far outnumbered by new sequences, the final value of dissimilarity was obtained by averaging 10,000 dissimilarity calculations between new amplicons and a random subsample (of same quantity than new amplicons) of previous amplicons. These calculations are made independently for each biological sample.

2.5 Clustering similar sequences using SWARM: OTUs detection

A clustering was made with the tool SWARM v2.1.6 [Mahe et al. 2015b], using the default options and enabling the -f fastidious option, on the valid sequence dataset.

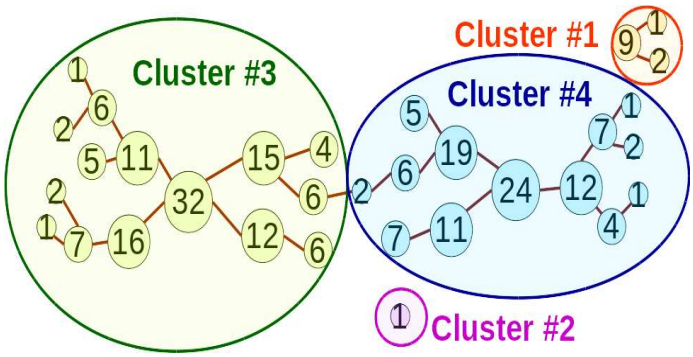

Fig M2: SWARM clustering. Each node represents a sequence whose quantity is indicated in the node. Each edge represents a distance of one mutation between two sequences. SWARM builds clusters according the proximity between sequences. If the path between two nodes of a cluster passes through a lower abundance node than the two nodes, SWARM splits the cluster in two clusters (cf the path between node '32' and node '24' is decreasing until '2' before rising).

Each cluster made by SWARM contains sequences near each other to a few substitions.

According to previous recommandations [Huse et al. 2010], the **validation rule for a cluster** is as follows: to be preserved, a cluster must gather a minimum of 3 sequences or at least 2 sequences from two different samples. Each valid cluster forms an **OTU** (Operationnal Taxonomic Unit) usually considered as a good candidate to represent a biological species.

Applied to our case-study, this test allows to valide new amplicons that are clustered with previous amplicons in a valid cluster.

2.6 New OTUs validation

Among SWARM clusters, some are "new OTUs" composed with only new amplicons. So, these amplicons were not at all detected with the research of exact primers. These amplicons raise questions: are they the mark of yet undetected species (as the previous conventionally obtained OTUs) or are they erroneous data ? Two different methods was used to estimate their credibility.

2.6.1 Validation by comparison with public dataset

To check if some of these amplicons are already known ones, each new OTU was tested against the BLAST database [Altschul et al. 1990], using the default options. We select the matches with the best recovery rate of alignment and at least 90\% of recovery, and we chose among them the best match having at least 80\% of similarity.

2.6.2 Validation by cross-clustering using SWARM

To check if some of these amplicons were already found with the other technology, the representative amplicons of new OTUs from 454 technology were mixed with previous amplicons from Illumina technology before being clustered using the tool SWARM, using the default options. This clustering can check whether a new 454 amplicon is similar enough to previous Illumina amplicons to cluster with them.

The cross-validation of the new Illumina amplicons was done in the same way using the previous 454 amplicons.

2.7 Seeking even more amplicons: fishing amplicons into rejected sequences

In order to recover mutated primers not taken into account by the initial mutation pattern, we tried to identify already known amplicons in rejected sequences.

Each rejected sequence was scanned in order to detect into the read the exact presence of a known amplicon (i.e. an amplicon detected in the sample by the search of both primers, exact or mutated). The upstream extremity of the rejected sequence forms the forward extremity and the downstream extremity of the rejected sequence forms the reverse extremity. The forward extremity is aligned with each of the four sequences of the V4F_exact model, and the reverse extremity is aligned with the V4R_exact model. The alignment is made using the logiciel LALIGN [Pearson et al. 1988].

In case of good alignment, the mutated primer is kept aside for further analysis (cf [Results and discussion, partie 5]).

3 Results and discussion

3.1 Looking for mutated primers increases the number of recovered sequences

The 310,375 reads of 454/Roche dataset are filtered using the workflow presented in [Materials and methods:PrimerFinder workflow]. Looking for the exact primers helps recover 90.2\% of total sequences (cf fig R1). The search for mutated primers captures 8.3\% of additional sequences (+25,619).

The 5,223,138 reads of Illumina dataset are filtered in the same way. Looking for the exact primers helps recover 82.7\% of total sequences (cf fig R1). The search for mutated primers captures 7.1\% of additional sequences (+368,270).

They are now two subdatasets from each technologies: sequences with both exact primers, that were already detected using only regex search (the "previous amplicons") and sequence with at least one mutated primers, that can not be detected using only regex search (the "new amplicons").

	454 dataset	Illumina dataset
Initial sequences	**310,375** séquences	**5,223,138** séquences
Sequences with V4F_exact	*302,5*05	4,592,349
Sequences with V4F_mutated	7,001	109,792
Sequences without V4F	*869*	*520,997*
Total sequences with V4F	**309,506 (99.72\%)**	**4,702,141 (90.03\%)**
Sequences with V4R_exact	287,360	4,412,349
Sequences with V4R_mutated	19,311	273,185
Sequences without V4R	*2,835*	*16,522*

Total sequences with V4R	**306,671 (98.81\%)**	**4,685,619 (89.71\%)**
Sequences with N	*978*	*44*
Total sequences without N	**305,693 (98.49\%)**	**4,685,575 (89.71\%)**
Sequences found with only Regex search Added sequences with Logol search	*280,074 (90.24\%)* *25,619 (8.25\%)*	*4,317,315 (82.66\%)* *368,270 (7.05\%)*
Rejected sequences	4,682 (1.51\%)	537,563 (10.29\%)

Fig R1: Recall in 454/Roche and Illumina/MiSeq datasets at each workflow step. Percentages are calculated based on the amount of initial sequences in the data sets.

Search for mutated primer allow finding new amplicons. But one wonders what these new sequences are worth: are they like normal sequences?

3.2 Looking for mutateds primers detects sequence similar to sequences with exact primers

3.2.1 Bray-Curtis test: globally new amplicon set and previous amplicon set are similar

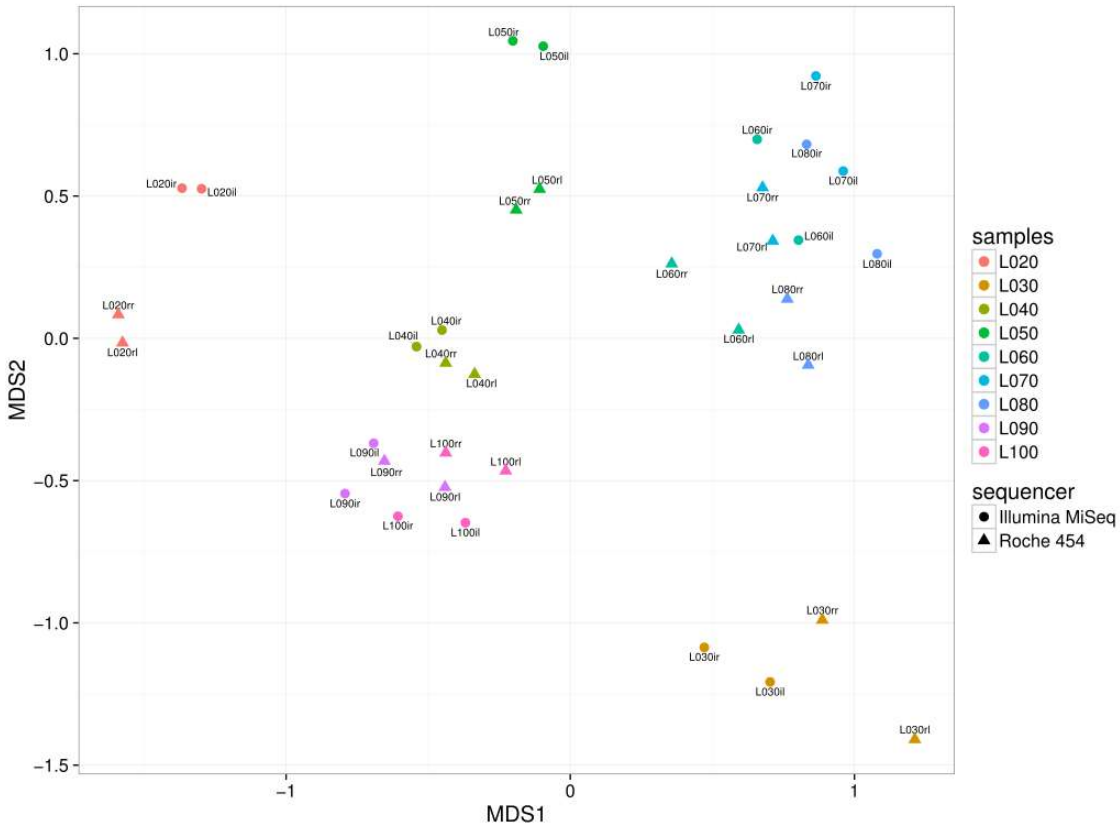

Fig R6: Bray-Curtis dissimilarity results for previous and new dataset. Dataset Id are builded following "LXab" pattern, where "LX" is the sample Id, "a" the technology ("i" for Illumina, "r" for 454) and "b" for the detection logiciel ("r" for regex (previous dataset) and "l" for Logol (new dataset)).

In order to evaluate the overall resemblance between amplicons bordered by exact primers and those bordered by mutated primers, both sets were compared using the Bray-Curtis test.

The result firstly shows that new amplicons are very close to previous ones. Secondly, it shows that new and previous sequences obtained via the same sequencing technology are closer together than their counterpart of the alternative sequencing technology (e.g., regex/Illumina amplicons are closer to Logol/Illumina amplicons than to regex/454 sequences).

So amplicons with mutated primers are globally similar to amplicons bordered by exact primers.

3.2.2 SWARM clustering: a new amplicon is not significantly more isolated than a previous amplicon

Are new amplicons mainly noise or interesting data?
An invalid cluster being a cluster with only one or two sequences (cf [materials and method:Clustering similar sequences using SWARM: OTUs detection]), its sequences reflect sequences isolated from the rest of the dataset, considered as noise (cf [Huse_ironning_2010]). So to answer the question we count how many amplicons of each category (new and previous) are in invalid clusters. For that, all amplicons (new and previous) from a technology have been clustered with SWARM.

454/Roche technology	**Previous amplicons**	**New amplicons**	**New/Previous ratio**
Before clustering	280,074	25,619	9.15\%
Clustered sequences *Rejected sequences*	266,988 (95.33\%) *13,086 (4.67\%)*	23,310 (90.99\%) *2,309 (9.01\%)*	8.73\%
Illumina/MiSeq technology	**Previous amplicons**	**New amplicons**	**New/Previous ratio**
Before clustering	4,317,315	368,247	8.53\%
Clustered sequences *Rejected sequences*	3,770,004 (87.32\%) *547,311 (12.68\%)*	279,318 (75.85\%) *88,929 (24.15\%)*	7.4\%

Fig R7: Proportion of amplicons validated by SWARM clustering in previous and new amplicons (for the 454/Roche data and Illumina/MiSeq data).

New amplicons (i.e. with at least one mutated primer) contain a bit less validated amplicons than the previous amplicons (i.e. with bot exact primers), but they are in a large majority valid sequences (only 9\% of rejected amplicons in 454/Roche new amplicons, and 24\% of rejected amplicons in Illumina/MiSeq new amplicons, cf R7).

3.3 Looking for mutated primers increases the number of OTUs

Looking for mutated primers allows obtaining new amplicons, but do they contribute to detect new OTUs?

To this end, we focus on the localisation of these new sequences after the clusterisation step.

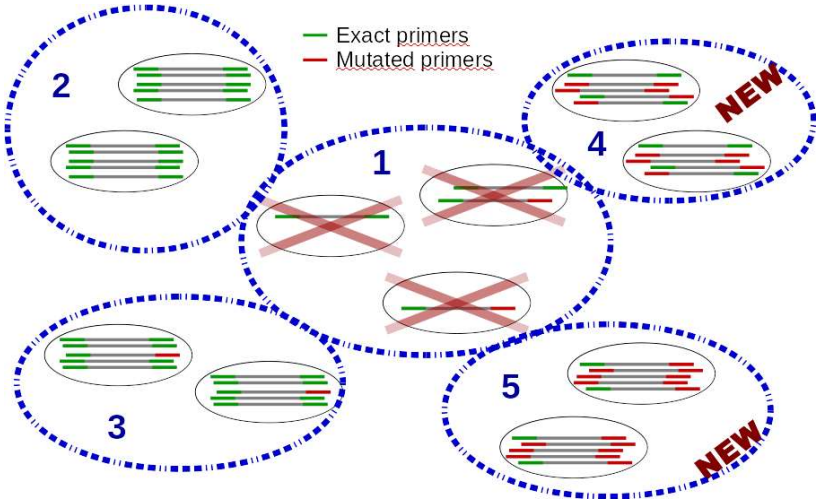

Fig R5: Illustration of the five types of cluster after SWARM clustering

We consider five different types of clusters at the end of the SWARM clustering:

1. **Invalid cluster**: the quantity of sequences in the cluster is too low to valid the cluster (validation threshold: at least 3 sequences or 2 sequences from 2 different samples).
2. **Unchanged previous OTUs**: the cluster contains only previous amplicons. Looking for sequences with mutated primers has not modified the composition of this cluster.
3. **Mixed previous OTUs**: the cluster contains both previous amplicons and new amplicons. Looking for sequences with mutated primers modify the composition of the cluster but not its detection. Previous amplicons are in sufficient quantity to valid the cluster by themselves.
4. **Mixed new OTUs**: the cluster contains both previous amplicons and new amplicons. Looking for sequences with mutated primers modify the composition of the cluster and its detection: previous amplicons are not in sufficient quantity to valid the cluster by themselves and adding the new amplicons allow validating the cluster.
5. **Completely new OTUs**: the cluster contains only new amplicons. This cluster was not detected at all using only sequences with exact primers.

For Illumina/MiSeq technology, the majority of new amplicons (75.85\%) are clustered in valid cluster (case 2 to 5, cf Fig R2), and the majority (70.03\%) are clustered in "mixed previous OTUs" (case 3), so they join pre-existant OTUs. Theses clusters regroup the majority of previous amplicons (78.46\%) (cf Fig R2).
A few parts of new amplicons (0.39\%) are clustered in "mixed new OTUs" (case 4), i.e. these new amplicons are similar to previous amplicons that have not suffisant quantity to form a valid the cluster. Without new amplicons, the cluster can not be valid.
Another few parts of new amplicons (5.62\%) are clustered in "completely new OTUs" (case 5), i.e. such new amplicons do not resemble to any previous amplicon, but it occurs in suffisant quantity to form a valid cluster.

For 454/Roche technology, the majority of new amplicons (90.99\%) are clustered in valid cluster (case 2 to 5, cf Fig R2), and the majority (80.63\%) are clustered in "mixed previous OTUs" (case 3), so they join pre-existant OTUs. Theses clusters regroup the majority of previous amplicons (89.89\%) (cf Fig R2).

A few parts of new amplicons (1.31\%) are clustered in "mixed new OTUs" (case 4), i.e. these new amplicons are similar to previous amplicons that have not suffisant quantity to form a valid the cluster. Without new amplicons, the cluster can not be valid.
Another few parts of new amplicons (9.04\%) are clustered in "completely new OTUs" (case 5), i.e. such new amplicons do not resemble to any previous amplicon, but it occurs in suffisant quantity to form a valid cluster.

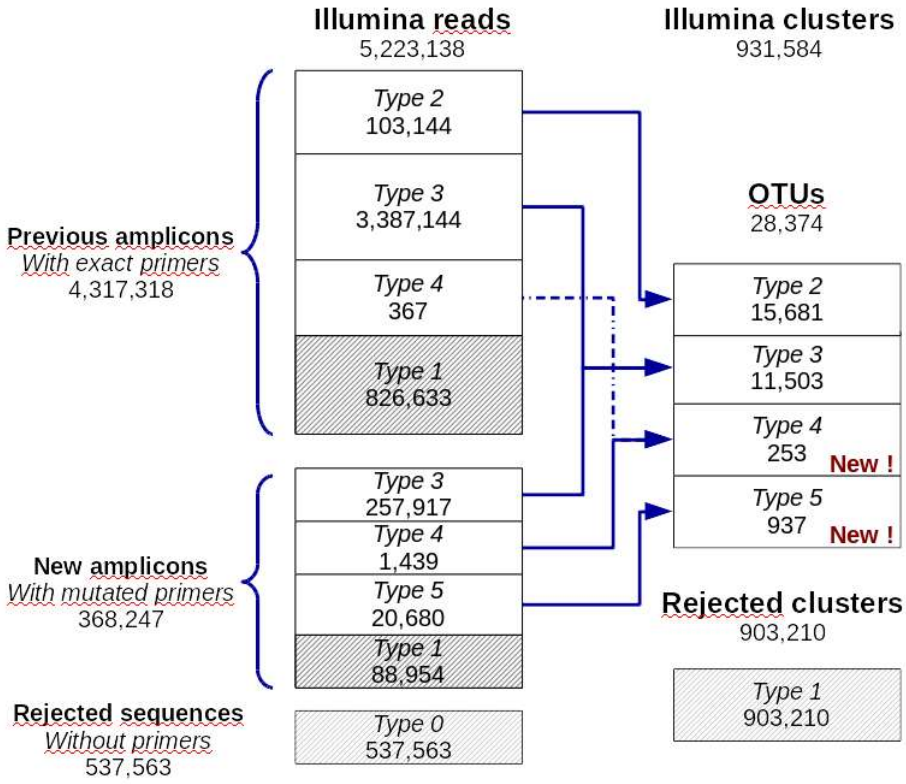

Fig R11: Distribution of Illumina reads through the primer search workflow.

To resume, here are our main observations on SWARM data:

- The search for mutated primers makes it possible to increase the number of obtained clusters (cf Fig R3).
- The majority of new amplicons are clustered with the majority of previous amplicons, into clusters that was already detected looking only for exact primers, so: a mutated primer does not necessarily involve a mutated amplicon; some sequences with mutated primer contain an amplicon highly similar to an amplicon of sequence with exact primers.
- In the same way, the new amplicons involved in mixed new cluster join amplicons detected by exact primers research, so these new amplicons seem relevant. Even if the cluster was not valid using only exact primer research, the amount of both sequences in the cluster is suffisant to pass the threshold of validation. The probability that two sequences with high sequencing errors are suffisantly close to form a homogenous cluster is very low according [Huse et al. 2010].
- A minority of new amplicons does not cluster with previous amplicons. Those are the amplicons that raise more interrogations: it is interesting to know whether these new groups bring information or add noise (cf [partie3.4]).

Find below the repartition of amplicons in SWARM clusters.

	454 / Roche		**Illumina / MiSeq**	
Sequence type	Previous	New	Previous	New
Sequence quantity	280,074	25,619	4,317,315	368,270
Rejected amplicons	4.67%	9.01%	19.15%	24.15%
Unchanged previous OTUs	5.38%	Ø	2.39%	Ø
Mixed previous OTUs	89.89%	80.63%	78.46%	70.03%
Mixed new OTUs	0.05%	1.31%	0.01%	0.39%
Completely new OTUs	Ø	9.04%	Ø	5.62%

Fig R2: Becoming of new and previous amplicons after the clusterisation with SWARM. Unchanged previous OTUs contain only previous amplicons. Mixed previous OTUs contain both new and previous amplicons, but previous amplicons are in suffisant quantity to valid the cluster by themselves. Mixed new OTUs contain both new and previous amplicons, but added new amplicons increase quantity enough in order to valid the cluster. Completely new OTUs contains only new amplicons.

	454 / Roche		**Illumina / MiSeq**	
Total OTU	4,435		28,374	
Unchanged previous OTU	2,375	53.55\%	15,681	55.27\%
Mixed previous OTU	1,756	39.59\%	11,503	40.54\%
Mixed new OTU	99	2.23\%	253	0.89\%
Completely new OTU	205	4.62\%	937	3.3\%

Fig R3: Proportion of new OTU added by new sequences after the clusterisation with SWARM. Unchanged previous OTUs contain only previous amplicons. Mixed previous OTUs contain both new and previous amplicons, but previous amplicons are in suffisant quantity to valid the cluster by themselves. Mixed new OTUs contain both new and previous amplicons, but added new amplicons increase quantity enough in order to valid the cluster. Completely new OTUs contains only new amplicons.

3.4 Some completely new OTUs are biologically relevant

The completely new OTUs contain only new amplicons but in suffisant quantity to validate the cluster. In order to verify the relevance of these new amplicons, we have checked if some of them are similar to 18S sequences already present in database or if some are already found in samples with another technology (cf [Materials and method,2.6]).

	454 / Roche		**Illumina / MiSeq**	
Completely new OTUs	205		937	
SWARM cross-validation	6	2.93\%	27	2.88\%
BLAST validation	3	1.46\%	76	8.11\%
Both validations	1	0.49\%	2	0.21\%
No validation	195	95.1\%	832	88.8\%
Total validation	**10**	**4.88\%**	**105**	**11.2\%**

Fig R8: Validation of completely new OTUs for 454/Roche and Illumina/MiSeq technologies.

Only a few sequences can be identified by comparison against public databank (2\% in 454/Roche, 8\% in Illumina/MiSeq), which is not very surprising because of the lack of information on eukaryote tropical data soils species. Indeed, these species are massively unknown: for example, BLAST validation used on previous dataset (i.e. sequences with exact primers) allows the identification of only 4.6\% in 454/Roche and 1.6\% in Illumina/MiSeq.

Although we can clearly not assert that all new OTUs are valid, we can see that some of these completely new OTUs have a biological validity.

3.5 Analyzing the mutated primers gives rise to a new mutation model

Even by using the model of mutations, there are sequences were at least one primer is not detected. In order to improve the model, we have analysed sequences that are rejected by the workflow for the sample L020.

3.5.1 Illumina/MiSeq sample

With Illumina/MiSeq technology, there is 10\% of rejected reads (32,923 reads).

Using the fishing amplicon method presented in [Materials and methods,2.7], we have find 436 exact amplicons present in these reads (1,3\%, we will name these reads "**recovered reads**").
Into this 436 recovered reads, we found that 16 \% (70 reads) did not have at least one primer: the read is beginning or ending directly by the amplicon. So, the reads will never be detected by a primer search workflow.

Looking for definitively rejected sequences: not recovered sequences are analysed using BLAST, in order to see if they are close or not of already known 18S sequences. 56\% (18,358) of not recovered sequences have 100\% of identity with the phage PhiX sequences. These reads are used to improve the sensitivity of the sequencing: the sample was not perfectly cleaned before analysis.

Primers of recovered reads (not detected by the mutated_V4 models) was merged with mutated primers of new amplicons, in order to build a panorama of mutation type present in the sample.

Mutated primer	**V4F**	**V4R**
Covered by the mutation model	87.03\%	93.42\%
• Most important features	• 37.92\% - 1 Substitution • 34.99\% - 1 Deletion • 5.4\% - 2 Substitutions	• 74.91\% - 1 Substitution • 7.01\% - 2 Substitutions • 6.2\% - 1 Deletion • 3.41\% - 1 Insertion
Without at least one primer	0.82\%	0.06\%
Important features not covered by the model	7.09\% - 1 deletion at 5' extremity 1.45\% - 2 Deletion	4.8\% - 2 Deletion
Total	96.39\%	98.28\%

Fig R4: Principal models of mutation present in mutated primers for Illumina/MiSeq technology. The missing percentages correspond to heavily mutated primers (>50% mutations)

3.5.2 454 / Roche sample

With 454/Roche technologies, there is 1\% of rejected reads (431 reads).

Using the fishing amplicon method presented in [Materials and methods,2.7], we have find 133 exact amplicon present in these reads (30.86\%, we will name these reads "recovered reads").

Primers of recovered reads (not detected by the mutated_V4 models) was merged with mutated primers of new amplicons, in order to build a panorama of mutation type present in the sample.

Mutated primer	**V4F**	**V4R**
Covered by the mutation model	78.24\%	94.36\%
• Most important features	• 34.25\% - 1 Deletion • 30.7\% - 1 Substitution • 12.14\% - 1 Insertion	• 27.37\% - 1 Insertion • 26.76\% - 1 Deletion • 23.72\% - 2 Deletion at 3' • 10.64\% - 1 Substitution
Without at least one primer	9.28\%	0%
Important features not covered by the model	5.96\% - 1 deletion at 5' extremity 3.09\% - 2 Deletions	0.81\% - 2 Insertions 0.76\% - 2 Deletions
Total	96.57\%	95.93\%

Fig R9: Principal models of mutation present in mutated primers for 454/Roche technology. The missing percentages correspond to heavily mutated primers (>50% mutations)

3.5.3 Final mutated V4 models

The mutation pattern used to detect mutated primers find the majority of mutated primers (cf Fig R4 and R9). This model can be improved to take into account some mutation patterns not covered, such

as the possibility to allow up 2 deletions.

Final_mutated_V4F: CCAGCA[GC]C[CT]GCGGTAATTCC up to 2 mutations
Final_mutated_V4R: CTTTCGTTCTTGAT[CT][AG]A up to 2 mutations

Fig R10: New mutated models for V4F and V4R primers.

More broadly, contrarly to the new mutated_V4F and V4R models, the new model does not require considering separately for substitution and indel counts: to allow a global value of two mutations lead to a better recall of "new sequences" and "recovered reads" while facilitating its implementation through more known software (such as CutAdapt [Martin 2011]).

4 Conclusion

We have shown that finding mutated primers in the metagenomic sequencing data makes it possible to increase the number of amplicons detected, i.e. the number of reads retained at the end of the primer detection workflow. We called them new amplicons. We have also shown that new amplicons obtained are very close to amplicons found with both exact primers. The majority of these new amplicons are similar to amplicons with exact primers: the presence of mutated primers in reads therefore does not necessarily imply a more mutated amplicon than normal. Finally, we showed that some new OTUs obtained via the new amplicons were not detected by a standard workflow looking for exact primers: thus integrating amplicons from reads with mutated primers allows the validation of +4\% of OTUs (+1 190 OTUs) in Illumina/MiSeq in our tropical soils study. In addition, some of these new OTUs could be identified as the signature of known species (11% in Illumina/MiSeq) and thus allows the detection of new species present in the samples. Same analyse was done on 454/Roche technology with similar results.
Thus, the search for mutated primers makes it possible to exploit a sample more completely, which can be useful when the samples can not be duplicated. Nevertheless, in order to conclude more precisely, it would be necessary to repeat the study on populations of species better known than those of tropical soils in order to really be able to demonstrate whether the new detected OTUs add predominantly information or noise.

5 Supplementary Data

Supplementary data #1: Python regex script
Supplementary data #2: PrimerFinder workflow script (zip workflow F.M?)
Supplementary data #3: fasta of mutated primers
Supplementary data #4: List of mutated primers of recovered reads
Supplementary data #5: Logol models

6 Bibliography

[Altschul et al. 1990] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. « Basic local alignment search tool. Journal of Molecular Biology », 215(3):403-410, Oct. 1990.

[Belleannee et al. 2014] C. Belleannée, O. Sallou, and J. Nicolas. « Logol: Ex-

pressive Pattern Matching in Sequences. Application to Ribosomal Frameshift Modeling ». In M. Comin, L. Kall, E. Marchiori, A. Ngom, and J. Rajapakse, editors, Pattern Recognition in Bioinformatics, number 8626 in LNCS, pages 34-47. Springer International Publishing, Aug. 2014.

[de Vargas et al. 2015] C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahe, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Charon, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horak, O. Jaillon, G. Lima-Mendez, J. Lukes, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, and E. Karsenti. « Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean ». Science (New York, N.Y.), 348(6237):1261605, May 2015.

[Huse et al. 2007] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch. « Accuracy and quality of massively parallel DNA pyrosequencing ». Genome Biology, 8(7):R143, 2007.

[Huse et al. 2010] S. M. Huse, D. M. Welch, H. G. Morrison, and M. L. Sogin. « Ironing out the wrinkles in the rare biosphere through improved OTU clustering ». Environmental Microbiology, 12(7):1889{1898, July 2010.

[Lax et al. 2014] S. Lax, D. P. Smith, J. Hampton-Marcell, S. M. Owens, K. M. Handley, N. M. Scott, S. M. Gibbons, P. Larsen, B. D. Shogan, S.Weiss, J. L. Metcalf, L. K. Ursell, Y. Vazquez-Baeza, W. Van Treuren, N. A. Hasan, M. K. Gibson, R. Colwell, G. Dantas, R. Knight, and J. A. Gilbert. « Longitudinal analysis of microbial interaction between humans and the indoor environment ». Science (New York, N.Y.),345(6200):1048-1052, Aug. 2014.

[Mahe et al. 2015a] F. Mahe, J. Mayor, J. Bunge, J. Chi, T. Siemensmeyer, T. Stoeck, B. Wahl, T. Paprotka, S. Filker, and M. Dunthorn. « Comparing High-throughput Platforms for Sequencing the V4 Region of SSU-rDNA in Environmental Microbial Eukaryotic Diversity Surveys ». Journal of Eukaryotic Microbiology, 62(3):338-345, May 2015.

[Mahe et al. 2015b] F. Mahe, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. « Swarm v2: highly-scalable and high-resolution amplicon clustering ». PeerJ, 3:e1420, 2015.

[Martin 2011] M. Martin. « Cutadapt removes adapter sequences from high-throughput sequencing reads ». EMBnet.journal, 17(1):pp. 10{12, May 2011.

[Stoeck et al. 2010] T. Stoeck, D. Bass, M. Nebel, R. Christen, M. D. M. Jones, H.-W. Breiner, and T. A. Richards. « Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water ». Molecular Ecology, 19 Suppl 1:21-31, Mar. 2010.

[Pearson et al. 1988] W. R. Pearson and D. J. Lipman. « Improved tools for biological sequence comparison ». Proceedings of the National Academy of Sciences of the USA, 85(8):2444-2448, Apr. 1988.

[Tedersoo et al. 2014] L. Tedersoo, M. Bahram, S. Polme, U. Koljalg, N. S. Yorou, R. Wijesundera, L. V. Ruiz, A. M. Vasco-Palacios, P. Q. Thu, A. Suija, M. E. Smith,

C. Sharp, E. Saluveer, A. Saitta, M. Rosas,T. Riit, D. Ratkowsky, K. Pritsch, K. Poldmaa, M. Piepenbring, C. Phosri, M. Peterson, K. Parts, K. Partel, E. Otsing, E. Nouhra, A. L.Njouonkou, R. H. Nilsson, L. N. Morgado, J. Mayor, T. W. May, L. Majuakim, D. J. Lodge, S. S. Lee, K.-H. Larsson, P. Kohout, K. Hosaka, I. Hiiesalu, T. W. Henkel, H. Harend, L.-d. Guo, A. Greslebin, G. Grelet, J. Geml, G. Gates, W. Dunstan, C. Dunk, R. Drenkhan, J. Dearnaley, A. D. Kesel, T. Dang, X. Chen, F. Buegger, F. Q. Brearley, G. Bonito, S. Anslan, S. Abell, and K. Abarenkov. « Global diversity and geography of soil fungi”. Science, 346(6213):1256688, Nov. 2014.