Supplementary material 2

Test of HMP Mock community samples on ASaiM Galaxy instance and comparison with *EBI metagenomics* results

Bérénice Batut[1], Eric Peyretaillade[1], Jean-François Brugère[1], Pierre Peyret[1]

[1] EA-4678 CIDAM, Clermont Université, Université d'Auvergne, Clermont-Ferrand, France

The HMP metagenomes mock pilot is a project with metagenomic shotgun sequences from a controlled microbiota community (with 22 known microbial species). Two datasets are available: even and staggered mock communities. These controlled datasets are available in *EBI metagenomics* database. We analyzed these datasets with the workflow available with ASaiM Galaxy instance and compared taxonomic and functional results with the ones obtained with *EBI metagenomics* pipeline (version 1.0). Details about these analyses (workflows, scripts, results, parameters, …) are available on a dedicated GitHub repository.

1 Data

Two datasets are available for this project. The first dataset (SRR072232) contains a genomic mixture of 22 microbial species (Table 1) where the ribosomal RNA operon counts vary by up to four orders of magnitude per organism (Table 1). The second dataset (SRR072233) is a genomic mixture from 22 same microbial species (Table 1) containing equimolar ribosomal RNA operon counts per organism. After shotgun sequencing, first dataset (SRR072232) is constitued of 1,225,169 metagenomic sequences and the second dataset of 1,386,198 metagenomic sequences.

Taxonomy									Abundances	
Domain	Kingdom	Phylum	Class	Order	Family	Genus	Species	Strains	SRR072232	SRR072233
Archaea	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	*Methanobrevibacter*	*Methanobrevibacter smithii*	ATCC 35061	1,000,000	100,000
Bacteria	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*Actinomyces*	*Actinomyces odontolyticus*	ATCC 17982	1,000	100,000
					Propionibacteriaceae	*Propionibacterium*	*Propionibacterium acnes*	DSM 16379	10,000	100,000
		Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	*Bacteroides*	*Bacteroides vulgatus*	ATCC 8482	1,000	100,000
		Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*	*Deinococcus radiodurans*	DSM 20539	1,000	100,000
		Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	*Bacillus cereus thuringiensis*	ATCC 10987	100,000	100,000
					Listeriaceae	*Listeria*	*Listeria monocytogenes*	ATCC BAA-679	10,000	100,000
					Staphylococcaceae	*Staphylococcus*	*Staphylococcus aureus*	ATCC BAA-1718	100,000	100,000
							Staphylococcus epidermidis	ATCC 12228	1,000,000	100,000
				Lactobacillales	Enterococcaceae	*Enterococcus*	*Enterococcus faecalis*	ATCC 47077	1,000	100,000
					Lactobacillaceae	*Lactobacillus*	*Lactobacillus gasseri*	DSM 20243	10,000	100,000
					Streptococcaceae	*Streptococcus*	*Streptococcus agalactiae*	ATCC BAA-611	100,000	100,000
							Streptococcus mutans	ATCC 700610	1,000,000	100,000
							Streptococcus mitis oralis pneumoniae	ATCC BAA-334	1,000	100,000
			Clostridia	Clostridiales	Clostridiaceae	*Clostridium*	*Clostridium beijerinckii*	ATCC 51743	100,000	100,000
		Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Rhodobacter*	*Rhodobacter sphaeroides*	ATCC 17023	1,000,000	100,000
			Betaproteobacteria	Neisseriales	Neisseriaceae	*Neisseria*	*Neisseria meningitidis*	ATCC BAA-335	10,000	100,000
			Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	*Helicobacter*	*Helicobacter pylori*	ATCC 700392	10,000	100,000
			Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	*Acinetobacter baumannii*	ATCC 17978	10,000	100,000
					Pseudomonadaceae	*Pseudomonas*	*Pseudomonas aeruginosa*	ATCC 47085	100,000	100,000
				Enterobacteriales	Enterobacteriaceae	*Escherichia*	*Escherichia coli*	ATCC 70096	1,000,000	100,000
Eukaryotes	Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	Debaryomycetaceae	*Candida*	*Candida albicans*	SC5314	1,000	100,000
								Total	5,566,000	2,200,000

Table 1. Expected species, their taxonomy and their abundances on both samples (SRR072232 and SRR072233)

2 Methods

2.1 Analyses from *EBI Metagenomics*

Both datasets have been analysed with *EBI metagenomics* pipeline (Version 1.0) (Figure 1).

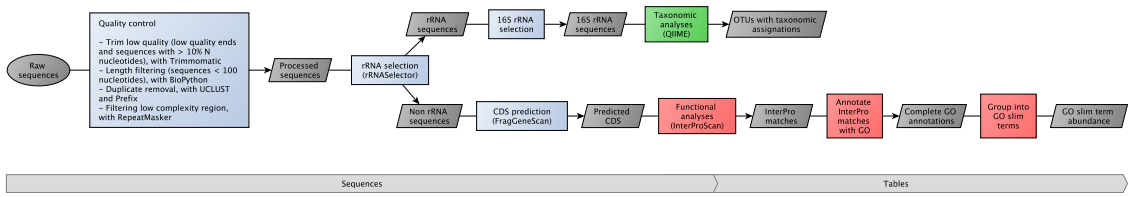

Fig. 1. EBI metagenomics pipeline (version 1.0). The grey boxes correspond to data, the blue boxes to pretreatment steps, the red boxes to functional analysis steps and the green boxes to taxonomic analysis steps.

The results are available on *EBI metagenomics* database, which have been downloaded and formatted.

From OTUs with taxonomic assignation, abundances of each assigned clade are extracted and several relative abundance measures are computed: relative abundances of clades for all OTUs and relative abundances of clades for OTUs with complete taxonomic assignation from kingdom to family. Percentage of unassigned clades (without complete taxonomic assignation) is also computed at all taxonomic levels.

For functional analysis, *EBI metagenomics* pipeline (Figure 1) offers 3 types of results: matches with InterPro, complete GO annotations and GO slim annotations. Here, we focus on GO slim annotations for easy comparison with ASaiM workflow results (Figure 2). The annotations are formatted to extract relative abundances (in percentage) of GO slim annotations inside three groups (cellular components, biological processes and molecular functions).

2.2 Analyses with ASaiM workflow

Both datasets are analyzed using ASaiM workflow dedicated to single-end microbiota data (Figure 2).

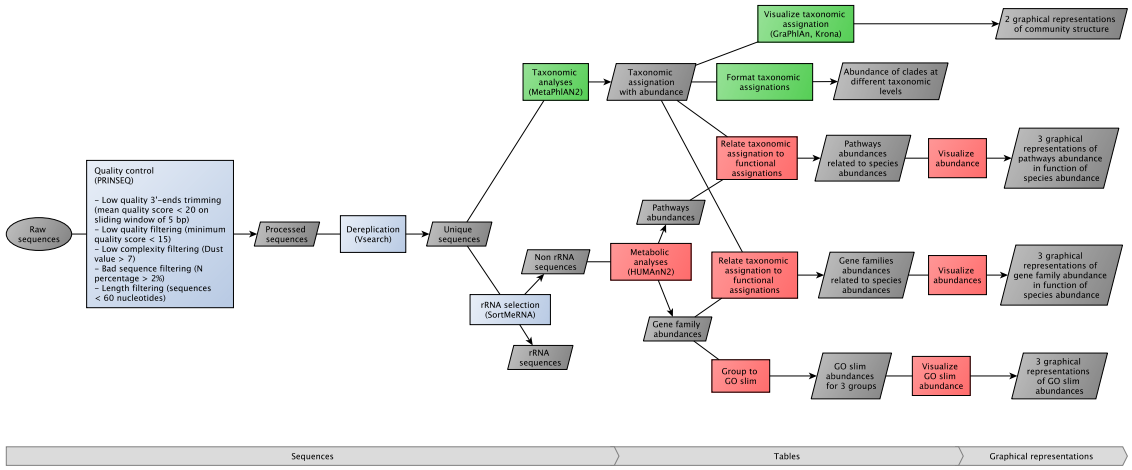

Fig. 2. ASaiM workflow for analysis of raw single-end microbiota sequences. This workflow is available with ASaiM Galaxy instance and used to analyze both datasets. The grey boxes correspond to data, the blue boxes to pretreatment steps, the red boxes to functional analysis steps and the green boxes to taxonomic analysis steps.

This workflow is available with ASaiM Galaxy instance. For this analysis, the ASaiM Galaxy instance is deployed on a Debian GNU/Linux System with 8 cores Intel(R) Xeon(R) at 2.40GHz and with 32 Go of RAM. Several statistics are followed during workflow execution (Table 2).

Statistics		SRR072232	SRR072233
Execution time	Whole workflow	4h44	5h22
	PRINSEQ	0h38	0h44
	Vsearch	16s	19s
	SortMeRNA	0h55	0h58
	MetaPhlAN2	0h09	0h10
	HUMAnN2	3h01	3h26
%CPU used	Min	4.8%	4.8%
	Mean	4.8%	4.8%
	Max	4.8%	4.8%
Size of the process in memory (kb)	Min	1,515,732	1,515,732
	Mean	1,515,744	1,515,743
	Max	1,515,768	1,515,764

Table 2. Computation statistics on ASaiM for both samples (SRR072233 and SRR072233)

Once ASaiM Galaxy instance is deployed (a task that can take several hours), datasets analyses inside the workflow are relatively fast: < 5h and < 5h30 for datasets with 1,225,169 and 1,386,198 sequences respectively (Table 2). The main time consuming step is the functional assignation with *HUMAnN2* (Abubucker *et al.*, 2012) which last $\simeq$ 64% of overall time execution (Table 2). The percentage of used CPU is stable over workflow execution, just like the size of the process in memory (variability inferior to 40 kb) (Table 2).

After workflow execution, taxonomic results are formatted to extract the percentage of unassigned clades at different taxonomic levels (clades without more accurate taxonomic assignation).

No further formatting step is needed for functional results (relative abundance of gene families, pathways with and without species relation and GO slim terms) of one sample. To compare functional results (gene families and pathways) between both samples (SRR072232 and SRR072233), a workflow is developed and executed (Figure 3).

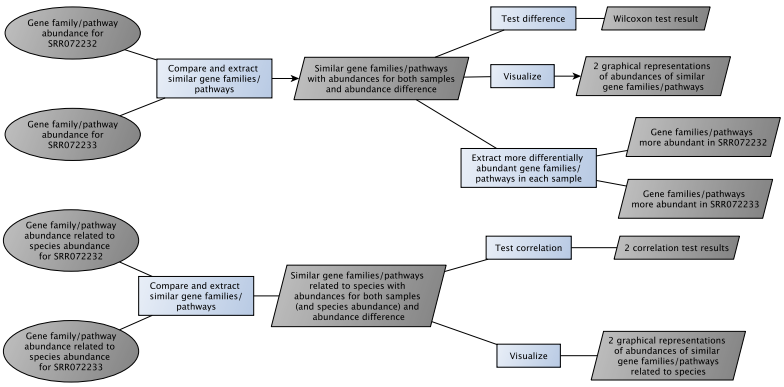

Fig. 3. Workflow to compare ASaiM functional results (gene families or pathways) between both samples. This workflow is available with ASaiM Galaxy instance. The grey boxes correspond to data, the blue boxes to processing steps.

2.3 Comparison of results from *EBI metagenomics* and ASaiM

Results from *EBI metagenomics* results and the ones from ASaiM are not directly comparable. Several processing steps are then needed.

With *MetaPhlAn* in ASaiM workflow, relative abundance of clades is computed on assigned reads. No count is made of non assigned reads. To compare relative abundances between *EBI metagenomics* and *ASaiM*, we focus on relative abundances computed on OTUS or reads with a complete taxonomic assignation from kingdom to family. These results are also compared to expected relative abundances obtained from sample descriptions (Table 1).

In both *EBI metagenomics* and ASaiM workflows (Figures 1 and 2), functional matches are grouped into GO slims terms. These terms are a subset of the terms in the whole Gene Ontology with a focus on microbial metabolic functions. They give a broad overview of the ontology content. To compare *EBI metagenomics* and ASaiM results, relative abundance of GO slim terms for both samples and both workflows are concatenated and compared, given the workflow depicted in Figure 4.

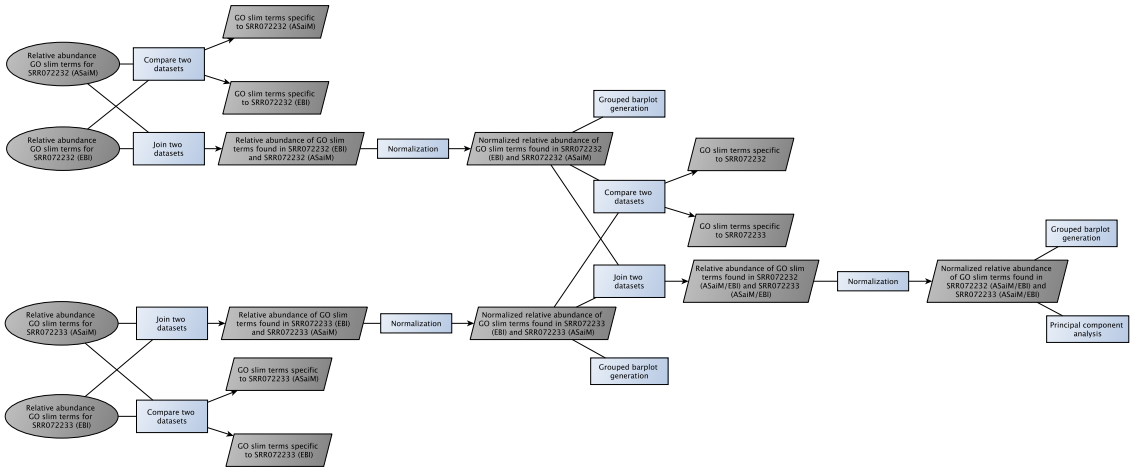

Fig. 4. Workflow to compare GO slim annotation abundances between samples (SRR072232, SRR072233) and workflows (EBI metagenomics, ASaiM). This workflow is available with ASaiM Galaxy instance. The grey boxes correspond to data, the blue boxes to processing steps.

3 Results

3.1 Pretreatments

In both workflows (Figures 1 and 2), raw sequences are pre-processed before any taxonomic or functional analysis. These preprocessing steps include quality control to remove low quality, small or duplicated sequences and also rRNA sorting to sort rRNA sequences from non rRNA sequences (Figures 1 and 2). The used tools and parameters for these pretreatments are different between *EBI metagenomics* pipeline (Figure 1) and ASaiM workflow (Figure 2). Even with similar raw sequences, pretreatment outputs are different (Table 3).

	SRR072232				SRR072233			
Sequences	EBI		ASaiM		EBI		ASaiM	
Raw sequences	1,225,169				1,386,198			
Sequences after quality control and dereplication	997,622	81.4%	1,175,853	96%	1,197,748	86.4%	1,343,451	96.9%
rRNA sequences	8,910	0.9%	16,016	1.4%	9,214	0.8%	13,850	1%
non rRNA sequences	988,712	99.1%	1,159,837	98.6%	1,188,534	99.2%	1,329,601	99%

Table 3. Statistics of pretreatments for EBI and ASaiM on both samples (SRR072233 and SRR072233)

The first interesting point in pretreatment results is the difference in sequence number after quality control and dereplication (Table 3). With ASaiM, more sequences (> 96 %) are conserved during these first steps of quality control and dereplication than with *EBI*

metagenomics (< 87 %, Table 3). This difference may be explained by threshold differences for minimum length. In *EBI metagenomics* pipeline, sequences with less than 100 nucleotides are removed (Figure 1), while in ASaiM the threshold is fixed to 60 nucleotides (Figure 2). However, this threshold difference does not explain all the observed difference in sequence number after quality control and dereplication. Indeed, when in ASaiM workflow PRINSEQ (Schmieder and Edwards, 2011) is run with exactly same parameters but filtering of sequences with less than 100 nucleotides, 1,135,008 (92.6%) and 1,304,023 (94.1%) sequences are conserved for SRR072232 and SRR072233 respectively after quality control and dereplication. These proportion are still higher than the one observed with *EBI metagenomics* pipeline (Table 3). Smaller length threshold with ASaiM does not then explain all difference in sequence number after quality control and dereplication.

In both datasets and with both workflows, few rRNA sequences are found in datasets (Table 3). Indeed, these datasets are metagenomic datasets and then focus on gene sequences. Few copies of rRNA genes are found in organisms (bacteria, archeae or eukaryotes) and are then expected in metagenomic sequences. Despite small number of sequences, a difference of rRNA sequence number is observed between *EBI metagonomics* and ASaiM workflows (Table 3). Higher proportions of rRNA sequences are systematically found with ASaiM workflow. Indeed, in *EBI metagenomics* pipeline (Figure 1), *rRNASelector* (Lee *et al.*, 2011) is used to select rRNA bacterial and archaeal sequences (no eukaryotes sequences). In ASaiM workflow (Figure 2), rRNA sequences are sorted using *SortMeRNA* (Kopylova *et al.*, 2012) and 8 databases for bacteria, archaea and also eukaryotes rRNA. < 5% of all sequences are matched against databases dedicated to eukaryotes rRNA sequences, but it does not explain all differences of rRNA sequence proportions between *EBI metagenomics* and ASaiM. This difference may be due to completness of the databases: databases used by *rRNASelector* (Lee *et al.*, 2011) are older and probably less complete than databases used by *SortMeRNA* (Kopylova *et al.*, 2012).

After pretreatments, more sequences are conserved for taxonomic and functional analyses in ASaiM workflow than in *EBI metagenomics* pipeline, for both samples (Table 3).

3.2 Taxonomic analyses

3.2.1 Raw ASaiM results

In ASaiM workflow (Figure 2), taxonomic analysis is made using *MetaPhlAN* (2.0) (Truong *et al.*, 2015; Segata *et al.*, 2012) on sequences after pretreatments. *MetaPhlAn* profiles the microbial community structure using a database of unique clade-specific marker genes identified from 17,000 reference genomes. This step of taxonomic assignation with *MetaPhlAn* is fast in ASaiM Galaxy instance (less than 10 minutes for > 1,100,000 sequences, Tables 2 and 3).

Raw *MetaPhlAn* results consist in a plain text file with relative abundance of clades at different taxonomic levels. Visualisation tools help to represent *MetaPhlAn* results. In ASaiM, two such tools are used: *Krona* (Ondov *et al.*, 2011) for an interactive representation of taxonomic assignation (SRR072232 and SRR072233) and *GraPhlan* for a static representation (Figures 5 and 6).

Despite same expected species, the taxonomic diversity in SRR072232 dataset (Figure 5) is reduced compared to the one in SRR072233 dataset (Figure 6). Less taxons are found for each taxonomic levels. From the 22 expected species (Table 1), 17 are found for SRR072232 and 20 for SRR072233 (Figure 7). The 2 expected species (*Candidata albicans* and *Lactobacillus gasseri*) missing in SRR072233 dataset are also missing in SRR072232 dataset (Figure 7). This may be due to a lack of phylogenetic markers for these species in the database used in *MetaPhlAn*.

Except *Staphylococcus*, the observed relative abundances of species for SRR072232 follows the expected ones, with some variability (Figure 7): smaller for small expected abundances and higher for high expected abundances. For SRR072233, same abundance is expected for all species, but a high variability is observed (Figure 7).

One species is interesting: *Deinococcus radiodurans*. In both samples, this species is found at abundances $\simeq$ 9 times higher than expected (Figure 7). This over-abundance in both samples may be explained by over-abundance in reference database and also by the high resistance of this particuliar bacteria.

3.2.2 Comparison with EBI results and expected taxonomy

After these first comparisons between ASaiM results taxonomic and expected ones, we compare ASaiM taxonomic results and *EBI metagenomics* taxonomic results.

In *EBI metagenomics* pipeline (Figure 1), *QIIME* (Caporaso *et al.*, 2010) is used on 16S sequences to identify OTUs and taxonomic assignation for these OTUs. In ASaiM (Figure 2), *MetaPhlAn* (Truong *et al.*, 2015; Segata *et al.*, 2012) is computed on sequences after quality control and dereplication, without any sorting step. *MetaPhlAn* (Truong *et al.*, 2015; Segata *et al.*, 2012) searches diverse phylogenetic markers, and not only 16S ones as *QIIME* (Caporaso *et al.*, 2010) does, on all sequence types (rRNA, non rRNA, …). Inside so different sequences, the proportion of sequences with phylogenetic markers is smaller. Indeed, with *MetaPhlAn* (Truong *et al.*,

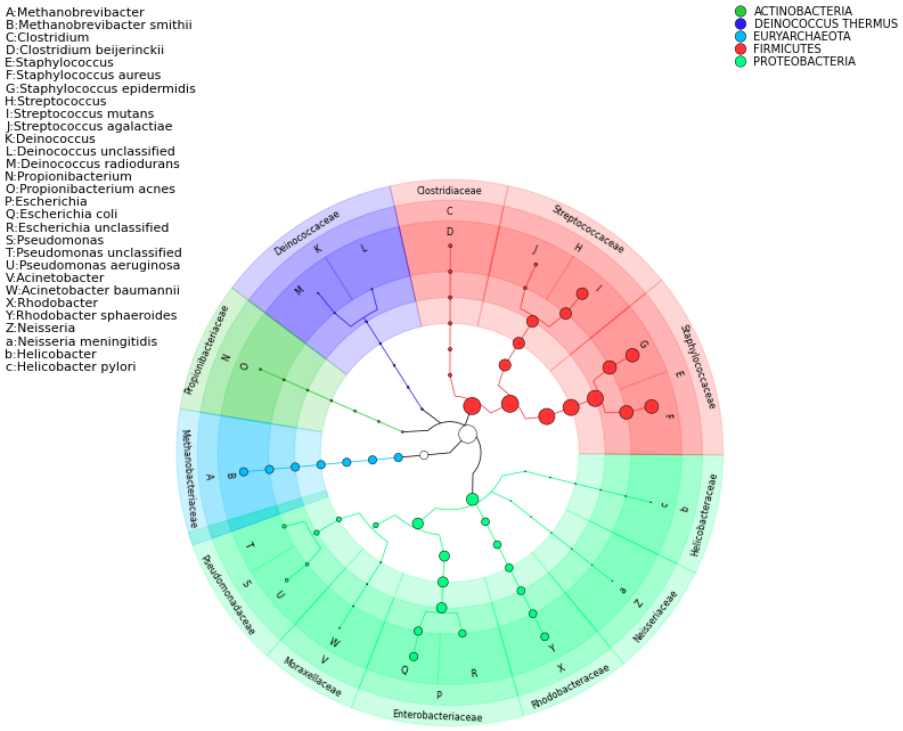

Fig. 5. GraPhlAn representation of taxonomic assignation obtained for SRR072232 with ASaiM

2015; Segata *et al.*, 2012), the percentage of unassigned reads is $\simeq 9$ times higher than with *QIIME* (Caporaso *et al.*, 2010) (SRR072232: 6.4% with *EBI metagenomics* against 62.61% with ASaiM; SRR072233: 13% with *EBI metagenomics* against 53.93% with ASaiM). However, with *MetaPhlAn* (Truong *et al.*, 2015; Segata *et al.*, 2012), the taxonomic assignations are more accurate, complete (until species level) and statistically supported (based on more sequences). Moreover, they do not focus only on bacteria or archea.

With both *EBI metagenomics* and ASaiM, some observed taxonomic assignations are unexpected (Tables 4 and 5). For ASaiM, 3 species in each sample are identified as "unclassified" (Table 4). They are affiliated to the correct genus but not to the species. These unclassified sequences may be due to incomplete annotations in reference database, because expected species are known and observed. These expected species are observed in lower abundance than expected (Figure 7) and would be closer to expected abundances with correct annotation of unclassified species.

Similarly, some observed clades and their sub-clades are unexpected in *EBI metagenomics* taxonomic results (Table 5). The taxonomic levels of these unexpected clades are higher (class, order and family) than unexpected taxonomic level in ASaiM (species, Table 6). Taxonomic assignations with *MetaPhlAN* (Truong *et al.*, 2015; Segata *et al.*, 2012) are then more accurate and precise.

Interestingly, for both workflows (*EBI metagenomics* and ASaiM), the proportion of unexpected clades is higher for SRR072232 than for SRR072233 (Table 6). We do not have a good explanation for this phenomenon.

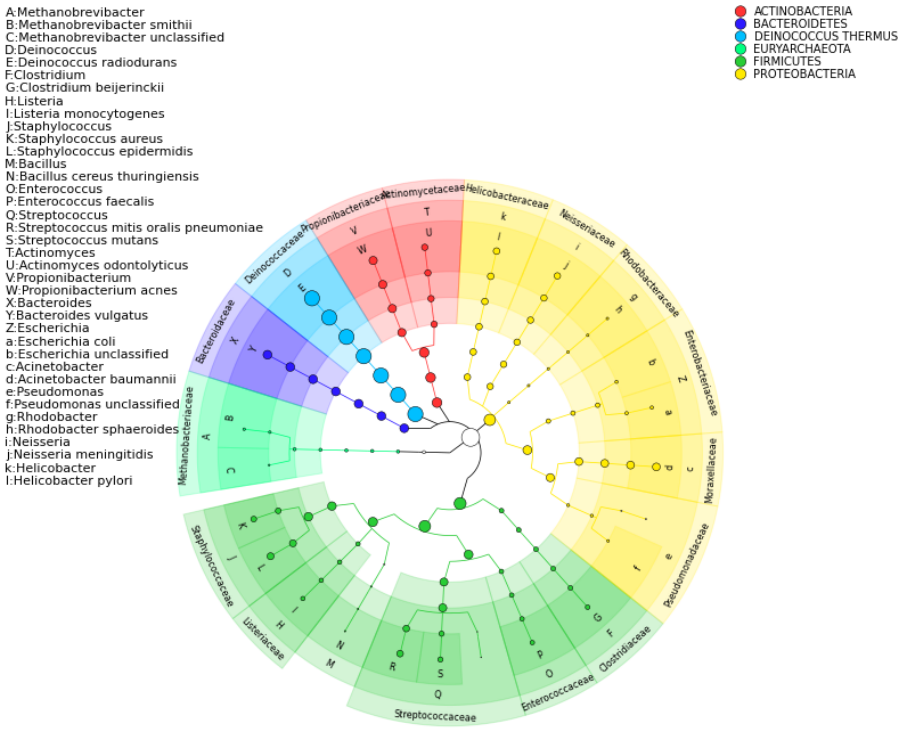

Fig. 6. GraPhlAn representation of taxonomic assignation obtained for SRR072233 with ASaiM

Species	SRR072232	SRR072233
Escherichia unclassified	4.85%	0.8%
Pseudomonas unclassified	1.12%	0.56%
Methanobrevibacter unclassified	-	0.24%
Deinococcus unclassified	0.16%	-

Table 4. Relative abundances of unclassified species in ASaiM taxonomic results for both samples (SRR072233 and SRR072233)

Taxonomic results obtained with *EBI metagenomics* pipeline are less precise than the one obtained with ASaiM workflow. Indeed, the most precise taxonomic level is family for *EBI metagenomics* and species for ASaiM. Then, to compare taxonomic results, we focus on family level (Figure 8).

Except Listeriacecae and Bacillaceae, all families found with *EBI metagenomics* are found with ASaiM (Figure 8). Some families are not found with both workflows. There may be two possibles explanation for these missing families.

The first reason relies on incompletness of reference databases used to assign taxonomy in workflows. Indeed, if sequences of some expected families do no match with any sequence in references databases, the corresponding families and the corresponding taxonomy

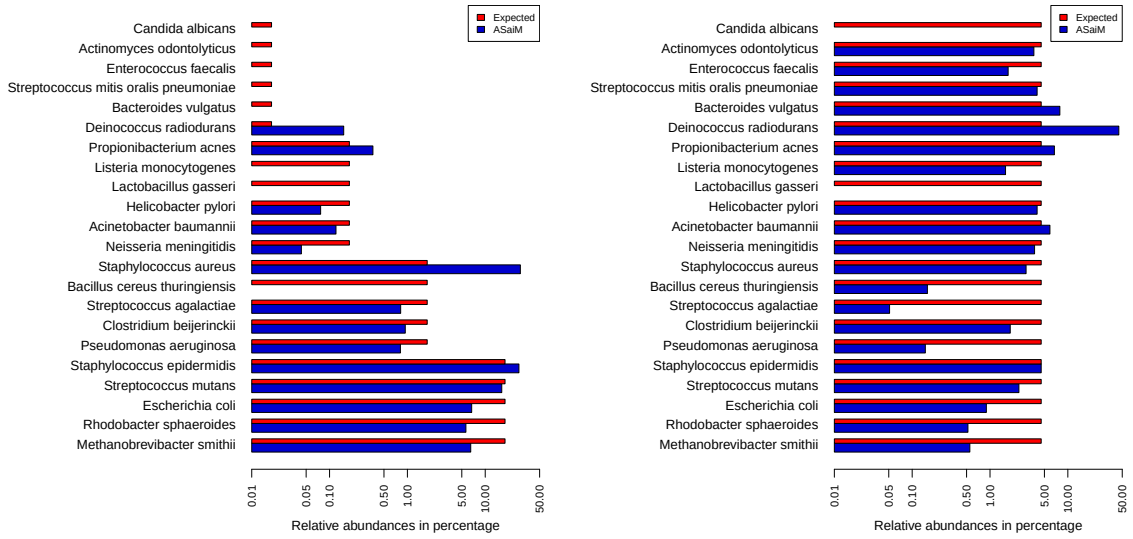

Fig. 7. Relative abundances (percentage in log scale) of expected species for SRR072232 (left) and SRR072232 (right) with comparison between expected abundances (red thin bars) and abundances obtained with ASaiM (blue wide bars)

Clade	Taxonomic level	SRR072232	SRR072233
Methanopyri	Class	0.09%	0.21%
Rickettsiales	Order	5.71%	1.43%
Methanopyrales	Order	0.09%	0.21%
Rickettsiales mitochondria	Family	5.71%	1.43%
Methanopyraceae	Family	0.09%	0.21%
Paraprevotellaceae	Family	-	0.09%
Cryptosporangiaceae	Family	-	0.5%

Table 5. Relative abundances of unexpected clades and their sub-clades in EBI metagenomics taxonomic results for both samples (SRR072233 and SRR072233)

	SRR072232		SRR072233	
Taxonomic level	EBI	ASaiM	EBI	ASaiM
Domain	-	-	-	-
Kingdom	-	-	-	-
Phylum	-	-	-	-
Class	0.09%	-	0.21%	-
Order	5.71%	-	1.64%	-
Family	5.71%	-	2.23%	-
Genus	*No information*	-	*No information*	-
Species	*No information*	6.13%	*No information*	1.6%

Table 6. Relative abundances of unexpected clades at different taxonomic levels in taxonomic results of EBI metagenomics and ASaiM for both samples (SRR072233 and SRR072233)

will not be found. It can be the case of Listeriaceae: this family is found with correct abundance with *EBI metagenomics* pipeline, but with ASaiM, this family is not found for SRR072232 and found in under-abundance for SRR072233 (Figure 8). Then, sequences

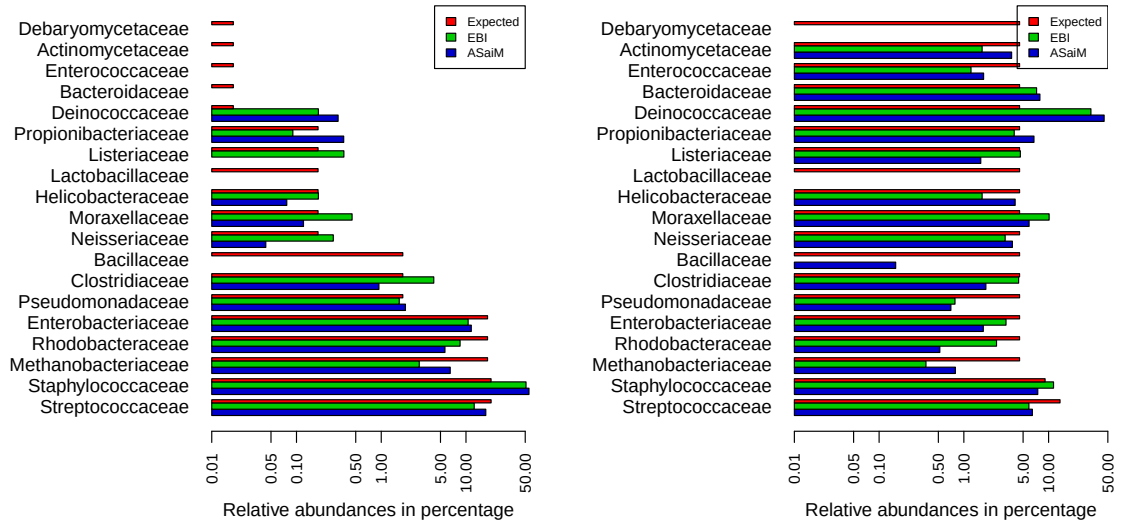

Fig. 8. Relative abundances of expected families for SRR072232 (left) and SRR072233 (right) with comparison between expected abundances (red thin bars), abundances obtained with EBI metagenomics (green wide bars) and abundances obtained with ASaiM (blue wide bars) and

corresponding to this family match some sequences in *MetaPhlAn* reference database, but this database is incomplete to match all expected sequences and correctly estimate abundance.

However, this incompletness of reference databases can not explain missing families in both *EBI metagenomics* and ASaiM (different tools and reference databases). An other explanation for these missing families can be proposed: too few sequences corresponding to the expected families are in datasets despite the expected abundance. This phenomenon may be due to experimental or sequencing errors. Sequences corresponding to expected families are then underrepresented in overall sequences and can not be detected in taxonomic analyses. Bacillaceae family seems a good example of this phenomenon. This family is not found neither by *EBI metagenomics* taxonomic analyses for both samples nor by ASaiM taxonomic analyses for SRR072232 (Figure 8), but it is found in low abundance by ASaiM taxonomic analyses for SRR072233 (abundance 29 times smaller than expected). The taxonomic signal of this family is really low, too low to be detected by *EBI metagnomics*. This under abundance of sequences corresponding to some families is a good explanation for families not found by the two methods, particularly when the expected abundance for these families is low (*e.g* Debaryomycetaceae, Actinomycetaceae, Bacteroidaceae, Enterococcaceae, Figure 8).

In these families with expected low abundance, one family is an exception: Deinococcaceae (Figure 8). In SRR072232 sample, unlike other families with low expected abundance, this family is found and with an abundance > 10 times higher than expected. In SRR072233 sample, the observed abundance is > 7 times higher than expected with both methods (Figure 8). In this family, one species is expected: *Deinococcus radiodurans*. An over-abundance of this species has already been observed in ASaiM taxonomic results (Figure 7).

To get a broader comparison of expected family abundances and *EBI metagenomics* and ASaiM observed family abundances, one principal component analysis (PCA) is made, for each sample, on observed families (for which abundance is not null in *EBI metagenomics* or ASaiM results). First axis of these analyses explains most of data variability (84% and 65% for SRR072232 and SRR072233 respectively, Figure 9) and is highly correlated (r^2 = 0.997 and r^2 = 0.994 for SRR072232 and SRR072233 respectively) with total abundance of families (sum of expected, *EBI metagenomics* and ASaiM abundances for each family).

In both samples, observed results (with *EBI metagenomics* and ASaiM) are grouped together and are orthogonal to expected results (Figure 9). And none of the workflows (*EBI metagenomics* one or ASaiM one) is closer to expected results. Both workflows have then similar results concerning family abundances.

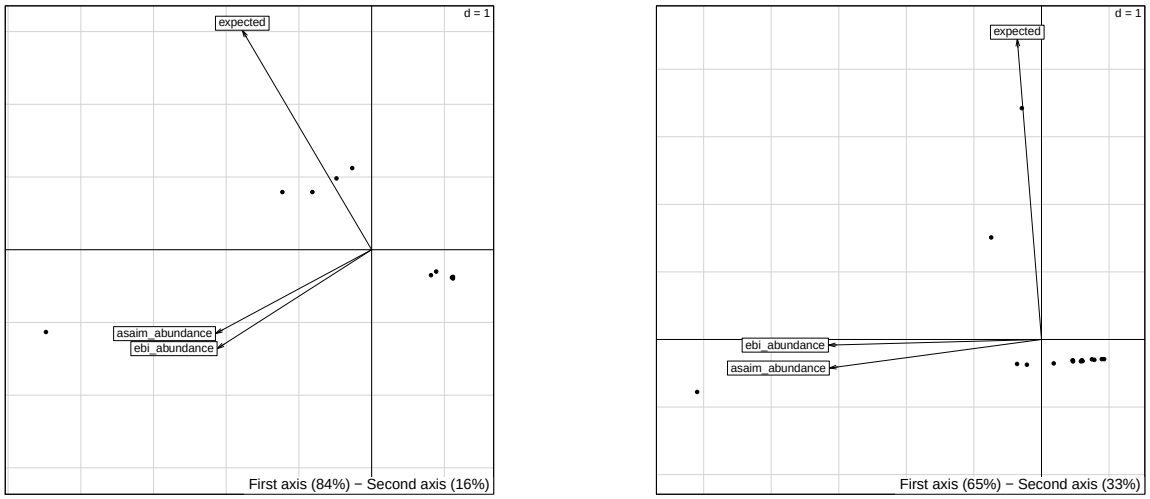

Fig. 9. Scatter diagram of principal component analysis of the relative abundances (in percentage) of observed families for SRR072232 (in left) and SRR072233(in right). Only observed families in EBI metagenomics or ASaiM are used in these analyses.

3.3 Functional analyses

3.3.1 Raw ASaiM results

In ASaiM workflow (Figure 2), functional analyses is made using *HUMAnN*2 (Abubucker *et al.*, 2012). This tool profiles presence/absence and abundance of UniRef50 gene families and MetaCyc pathways from metagenomic/metatranscriptomic datasets. It is helpful to describe the metabolic profil of a microbial community. This step of functional profiling with *HUMAnN*2 is the longest step in ASaiM workflow (Table 2).

*HUMAnN*2 generates three outputs: UniRef50 gene families, coverage and abundance of MetaCyc pathways. In both samples, > 90,000 UniRef50 gene families and > 480 MetaCyc pathways (Table 7) are reconstructed from > 1,100,000 non rRNA sequences (Table 3). More gene families and pathways are found for SRR072233 than for SRR072232 (Table 7) but the values are similar.

44,933 gene families are found in both samples (Table 7). Even if less than 50% of gene families are similar, similar gene families represent more than 50% of relative abundance in both samples (Table 7). Inside similar gene families, relative abundance of gene families in both samples is different (Figure 10): the median value of normalized abundance of similar gene families is smaller for SRR072232 (significant *p-value* for Wilcoxon test, Table 7).

Similarly, a high proportion (> 95 %) of pathways are found in both samples (Table 7) and they represents nearly all abundance (> 99.5 %, Table 7). Unlike gene families, relative abundance of pathways are similar in both samples (non-significant *p-value* for Wilcoxon test, Table 7 and Figure 10).

In *HUMAnN2* results, abundances of gene families and pathways are stratified at the community level. Contribution of identified species for < 35% of gene families and > 80% pathways are then accessible and they represent > 90% of relative abundance for gene families and > 50% for pathways (Table 7). This taxonomic information (relation between species and gene families or pathways) can be related to taxonomic information and species abundances from *MetaPhlAN2* (Figure 11).

For both samples, relative abundances of gene families and pathways are highly correlated to observed relative abundance of corresponding species (Figure 11 and Table 8). This relation is expected for gene families: a species is more abundant if more sequences corresponding to this species are found. The relative abundance of these sequences is then high. It applies to all sequences, particularly sequences corresponding to gene families. For pathways, the relation is more complex: a pathway is identified if a high proportion of the gene families expected for this pathway is found. And the abundance of a pathway is proportional to the number of complete "copies" of this pathway in the species. Then, a pathway is abundant if its parts are all found in numerous copies. The high correlations between species relative abundance and mean relative pathway abundance (Figure 11, Table 8) confirm correct pathway reconstructions in our datasets.

		UniRef50 gene families		MetaCyc pathways	
		SRR072232	SRR072233	SRR072232	SRR072233
All	Number	98,569	129,691	487	500
	Similar	44,933		475	
	% of similar inside all	45.59%	34.65%	97.54%	95%
	Relative abundance (%)	89.16%	50.67%	99.85%	99.53%
	p-value of Wilcoxon test on normalized relative abundance	$1.31 \cdot 10^{-14}$ (***)		0.24	
Associated to a species	Number	26,219	41,005	402	400
	% of associated to a species inside all	26.60%	31.62%	82.56%	80%
	Relative abundance (%)	93.40%	90.24%	61.08%	51.52%
	Similar	19,815		363	
	% of similar inside associated to a species	68.02%	48.32%	90.30%	90.75%
	Relative abundance of similar inside associated to a species (%)	89.17%	44.75%	91.87%	42.70%

Table 7. Global information about UniRef50 gene families and MetaCyc pathways obtained with HUMAnN2 for both samples (SRR072233 and SRR072233). For each characteristics (gene families and pathways), several information is extracted: all number, number percentage and relative abundance (%) of similar characteristics and p-value of Wilcoxon test on relative abundance normalized by the sum of relative abundance for all similar characteristics.

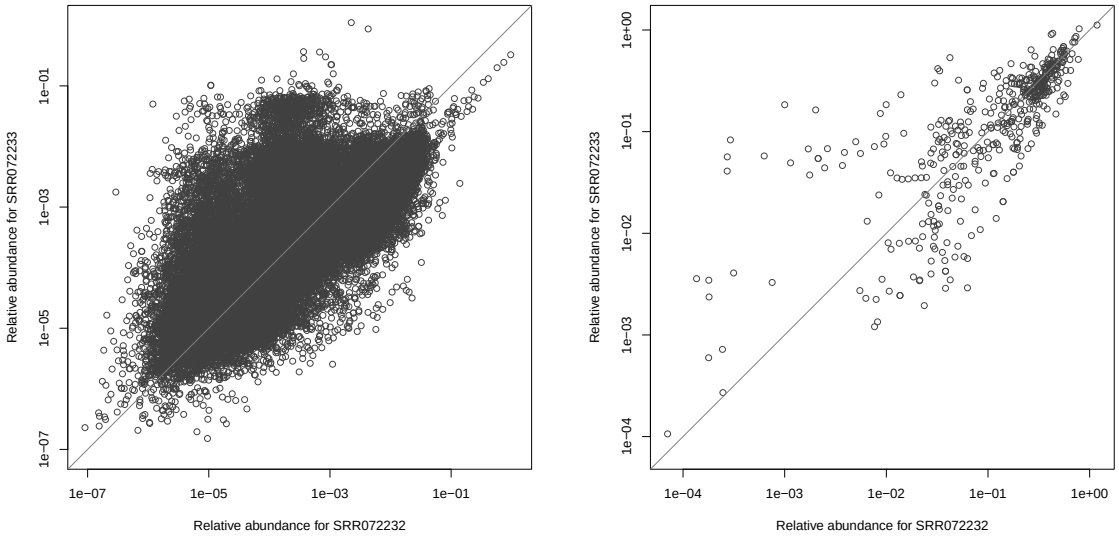

Fig. 10. Normalized relative abundances (%) for similar UniRef50 gene families (left graphics) and MetaCyc pathways (right graphics) for both samples (SRR072233 and SRR072233). The relative abundances of each similar characteristics (gene families or pathways) is computed with HUMAnN2 and normalized by the sum of relative abundance for all similar characteristics.

Similar relations between species abundances and gene family or pathway mean abundances are found in both samples (Table 8). Indeed, differences in gene families and pathways abundances between both samples are mostly explained by differences in abundance of corresponding species (correlation with coefficient > 0.80 and with significant *p-values*, Figure 12, Table 8).

Unlike mean abundance, number of different gene families for each species is not correlated to species abundances (Figure 14, Table 8). Then, a highly abundant species is abundant because its gene families are in numerous copies but not necessarily because different gene families are numerous. The number of different gene families for each species is more correlated with the median number of

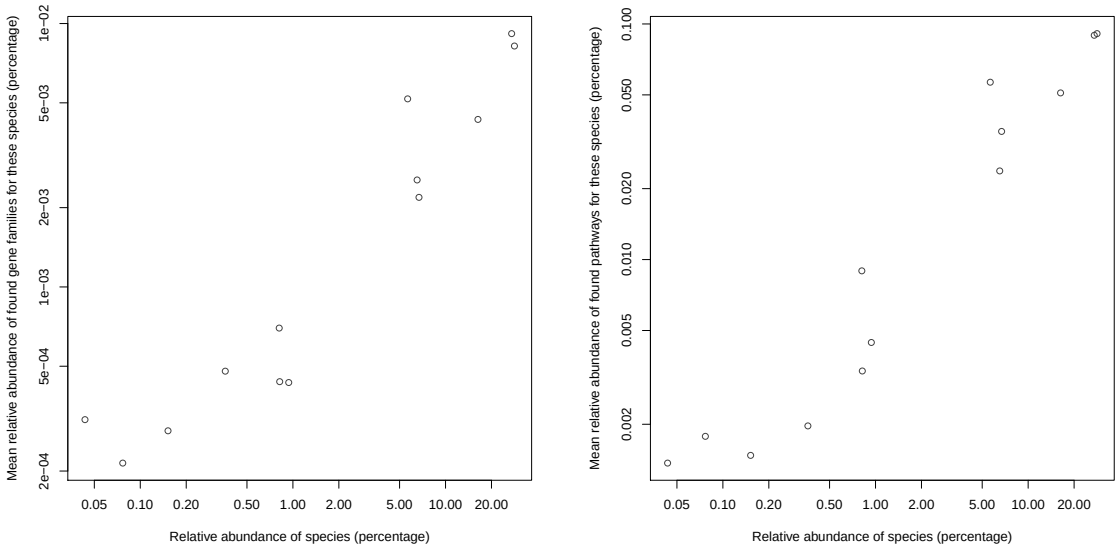

Fig. 11. Mean abundances of gene families (left) and pathways (right) in fonction of related species abundance for SRR072232 (log scale). Correlation coefficients and p-values are detailed in Table 8

		UniRef50 gene families		MetaCyc pathways	
		SRR072232	SRR072233	SRR072232	SRR072233
Mean abundance (Figure 11)					
Correlation with species abundance	r^2	0.91	0.98	0.90	0.93
	p-value	$1.51 \cdot 10^{-7}$	$< 2.2 \cdot 10^{-16}$	$1.91 \cdot 10^{-7}$	$5.88 \cdot 10{-}12$
Difference of mean abundance between SRR072233 and SRR072232 (Figure 12)					
Correlation with difference of species abundance	r^2	0.89		0.84	
	p-value	$4.121 \cdot 10^{-7}$		$4.651 \cdot 10^{-6}$	
Number (Figure 14)					
Correlation with species abundance	r^2	0.1	0.04		
	p-value	0.27	0.39		
Correlation with species median protein number	r^2	0.56	0.37		
	p-value	$2.027 \cdot 10^{-3}$	$4.28 \cdot 10^{-3}$		
Difference of number between SRR072233 and SRR072232					
Correlation with difference of species abundance	r^2	0.42		0.29	
	p-value	0.013		0.046	

Table 8. Correlation coefficients and p-values (Pearson's test) for UniRef50 gene families and MetaCyc pathways obtained with HUMAnN2 for both samples (SRR072233 and SRR072233).

protein for these species (Figure 14, Table 8). We expect also a correlation with the number of gene families corresponding to these species in reference database (UniRef50) of *HUMAnN2*. However, we do not have access to this information and we could not confirm this hypothesis.

In both samples, less abundant a species is, higher is the difference between number of observed gene families for this species and expected median protein number (Figure 14). Indeed, less abundant species have fewer sequences than more abundant species in overall

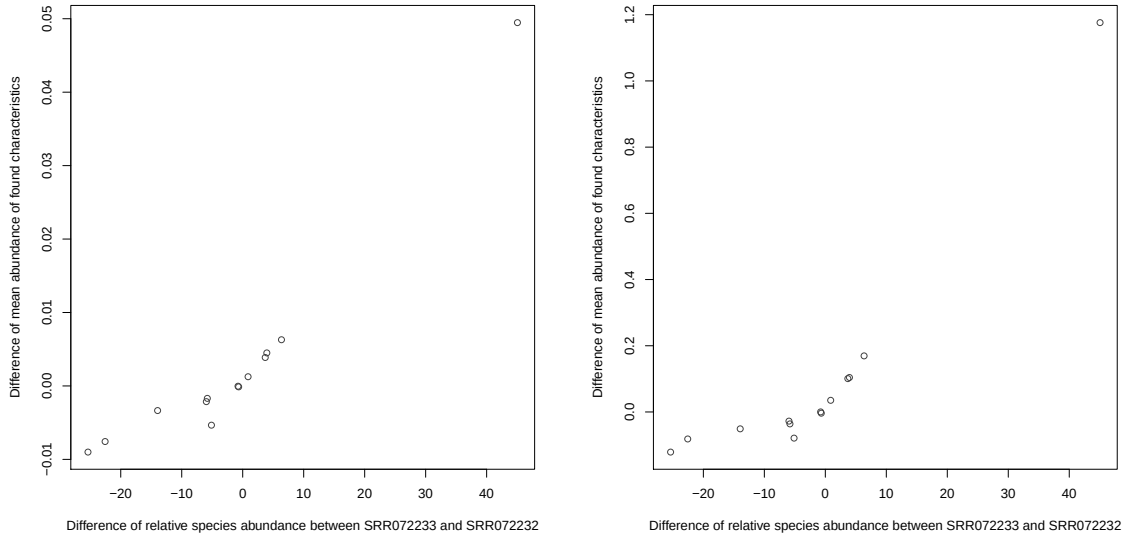

Fig. 12. Difference in mean abundances for gene families (left) and pathways (right) in fonction of difference of related species abundance between SRR072233 and SRR072232. Correlation coefficients and p-values are detailed in Table 8

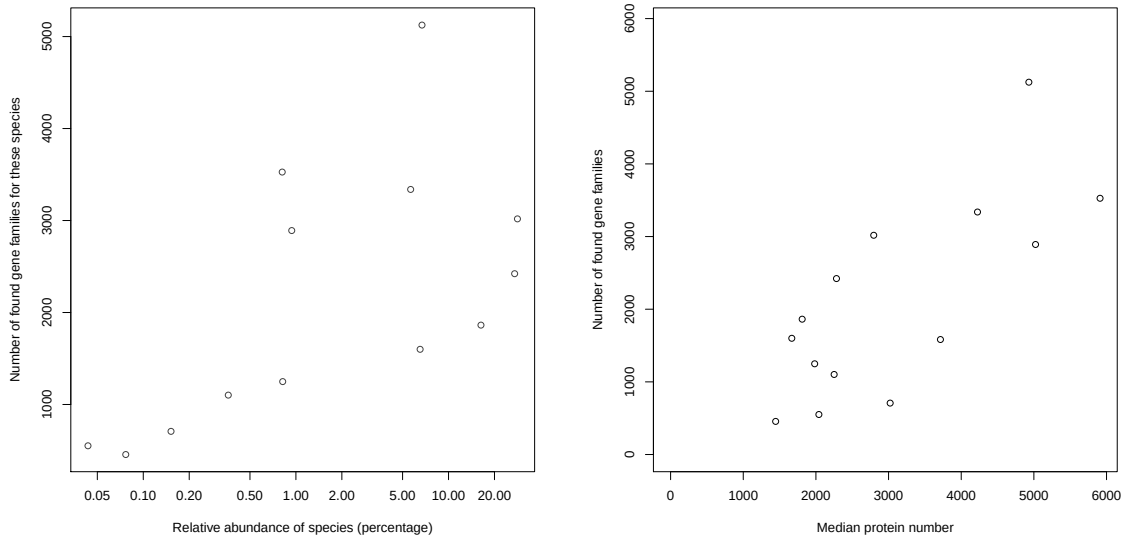

Fig. 13. Number of gene families in fonction of corresponding species abundance (left) and median protein number for species (right) for SRR072232. Correlation coefficients and p-values are detailed in Table 8. The median protein number for each species has been extracted from NCBI.

sequence dataset. These sequences are then diluted in overall sequences and the signal to identify these sequences and corresponding gene families is low and noisy. Lower proportion of gene families are then identified for less abundant species.

With more than 40,000 gene families and almost 500 pathways, it is difficult to get a broad overview of the metabolic profile of studied microbial community. Each gene family and pathway is precise and related to specific metabolic functions. This information is

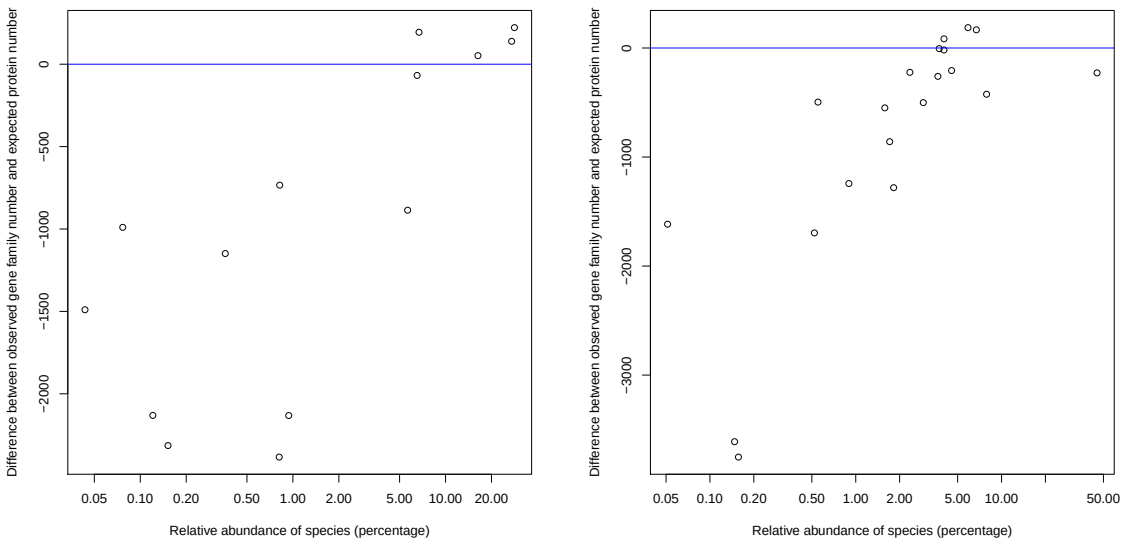

Fig. 14. Difference between observed number of different gene families and expected median protein number in function of relative abundance of corresponding species (log scale) for SRR072232 (left) and SRR072233 (right). The median protein number for each species has been extracted from NCBI.

interesting when you need detailed metabolic information and to go deeply inside metabolic profil. However, to get a general overview of the metabolic processes, UniRef50 gene families and even MetaCyc pathways are too numerous and too precise. UniRef50 gene families and their abundances can be grouped into slim Gene Ontology terms (Figure 15). These results are commented in relation with *EBI metagenomics* results in next section.

3.3.2 Comparison of *EBI metagenomics* and ASaiM results

In *EBI metagenomics* pipeline (Figure 1), functional analyses are based on InterPro and its identifiants. In ASaiM workflow (Figure 2), we have access to UniRef50 gene families and their abundances computed with *HUMAnN2*. These functional results are then not directly comparable. But, in both workflows, UniRef50 gene families and InterPro proteins are grouped into slim Gene Ontology terms to get a broad overview of functional profile of the community. The GO slim terms are grouped into 3 groups: cellular components, biological processes and molecular functions.

Few GO slim terms are not similar for *EBI metagenomics* and ASaiM (Table 9). They are negligible in term of relative abundance inside the three groups (< 20%).

Barplot representations of GO slim term abundances for both samples and both workflows can be difficult to interpret (*e.g* for the cellular component on Figure 16). We used then a principal component analysis (PCA) on normalized relative abundance of GO slim term abundance inside each group to simplify visualization and interpretation (Figures 16 and 17).

Scatter representation of first plan (constitued of first two axes) of the PCA is similar for the three groups (Figures 16 and 17). First axis explains most of data variability (between 64% and 87%) and is highly negatively correlated with total abundance on both samples and both workflows of GO slim terms (Table 10). Then GO slim terms found on left part of scatter representation (Figures 16 and 17) are highly abundant: cellular processes related to membrane and cytoplasm (Figure 16), biosynthetic processes, nitrogen compound metabolic process, small molecular metabolic process, transport and DNA metabolic process for biological processes (Figure 17) and nucleotide binding for molecular functions (Figure 17). This first axis does not discrimate samples or workflows (Figures 16 and 17). Previous conclusions can be then extrapolated for both workflows and both samples. Results from ASaiM workflow are then similar in term of GO slim term abundances to the one obtained with *EBI metagenomics* pipeline.

The discrimination between *EBI metagenomics* and ASaiM results appears with second axis (Table 10). This axis explains between 13% and 35% of overall data variability (Table 10). Some GO slim terms such as membrane, hydrolase activity or nitrogen compound

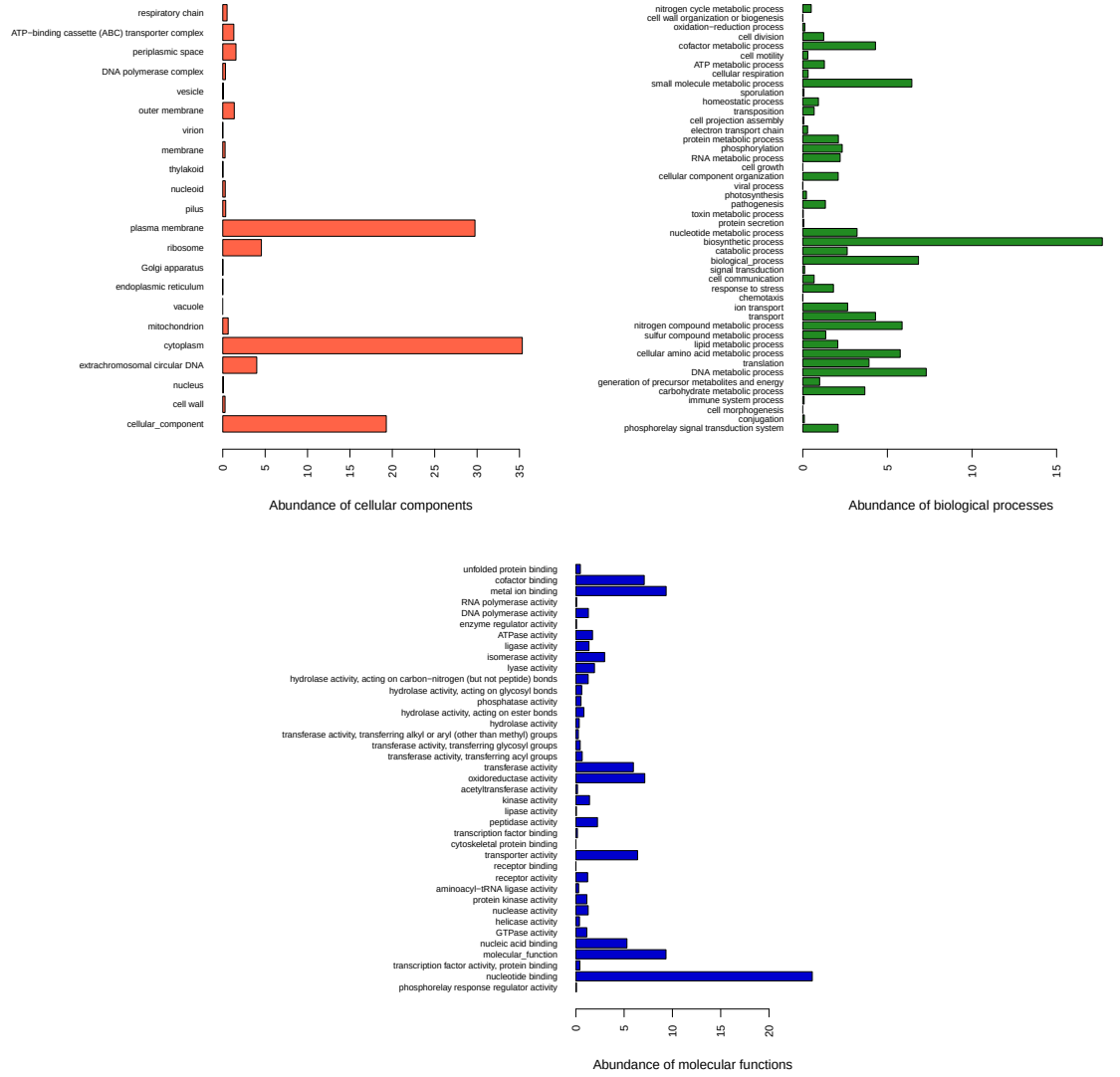

Fig. 15. Relative abundances of GO slim terms in SRR072232 for cellular components (top left), biological processes (top right) and molecular function (bottom)

metabolic process are found in higher proportion in *EBI metagenomics* results than in ASaiM and some like biosynthetic process, plasma membrane or nucleotide binding are in lower proportion (Figures 16 and 17).

None of the first two axes discriminates both samples. Variability between both samples seems then less important than variability between both workflows and mostly variability between GO slim terms.

EBI metagenomics and ASaiM functional results are similar in terms of GO slim terms abundance as the discrimination between both workflow results appears as a secondary explanation for variability of GO slim term abundances.

		SRR072232		SRR072233	
GO id	GO name	EBI	ASaiM	EBI	ASaiM
Cellular components					
GO:0031012	Extracellular matrix	$1.71 \cdot 10^{-2}$	-	$2.74 \cdot 10^{-2}$	$1.37 \cdot 10^{-5}$
GO:0005667	Transcription factor complex	0	-	$9.81 \cdot 10^{-3}$	-
GO:0005694	Chromosome	2.80	-	2.61	-
GO:0005856	Cytoskeleton	$2.23 \cdot 10^{-1}$	-	$8.44 \cdot 10^{-2}$	-
GO:0016469	Proton-transporting two-sector ATPase complex	1.34	-	1.44	-
GO:0019861	Flagellum	$9.78 \cdot 10^{-1}$	-	$6.24 \cdot 10^{-1}$	-
GO:0005575	Unknown cellular component	-	19.29	-	19.84
Biological processes					
GO:0006351	Transcription, DNA-dependent	3.27	-	3.06	-
GO:0044403	Symbiosis, encompassing mutualism through parasitism	$1.91 \cdot 10^{-2}$	-	$4.35 \cdot 10^{-3}$	-
GO:0046039	GTP metabolic process	$5.59 \cdot 10^{-2}$	-	$5.29 \cdot 10^{-2}$	-
GO:0008150	Unknown biological process	-	6.84	-	5.29
Molecular functions					
GO:0001071	Nucleic acid binding transcription factor activity	1.56	-	1.33	-
GO:0003774	Motor activity	$9.87 \cdot 10^{-2}$	-	$5.32 \cdot 10^{-2}$	-
GO:0045182	Translation regulator activity	$1.38 \cdot 10^{-3}$	-	0	-
GO:0003674	Unknown molecular function	-	9.34	-	10.88

Table 9. GO slim terms not found in both samples (SRR072232, SRR072233) and/or with both workflows (EBI metagenomics, ASaiM), with the relative abundance (in percentage) in GO slim groups (cellular components, biological processes and molecular functions)

		Cellular components	Biological processes	Molecular functions
First axis				
Explained variability		64%	87 %	85%
Correlation with total abundance	r^2	0.999	0.999	0.996
	p-value	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
Second axis				
Explained variability		35%	13 %	15%

Table 10. Principal component analysis (PCA) axes and correlations. Total abundance corresponds for each GO slim terms to the sum of abundance of this GO slim term for both samples and both workflows.

4 Conclusion

With ASaiM workflow, raw sequences from a metagenomic dataset are rapidly analyzed (in few hours in a standard computer). It makes it a powerful tools. Moreover, based on Galaxy, ASaiM workflow posseses all Galaxy's strength: accessibility, reproducibility and also modularity. The numerous results can also be accessed during workflow execution.

Taxonomic analysis is complete and accurate with *MetaPhlAn2*. From the expected taxonomy of both samples, few clades are not found or unexpected. The lowest accurate taxonomic level (species) is more precise than the one obtained with *EBI metagenomics*. With *HUMAnN2* results combined to *MetaPhlan2* results and GO slim term grouping, functional analyses are complete: precise functional assignations, broad overview of metabolic profile of studied microbial community and relation with observed community structure.

Many post-treatments are also possible in ASaiM Galaxy instance. For example, most of graphic representations and most statistical analyses of this report are done inside ASaiM Galaxy instance with available dedicated workflows.

ASaiM Galaxy instance with its workflows and tools is a then powerful framework to analyze shotgun raw sequence data from microbiota.

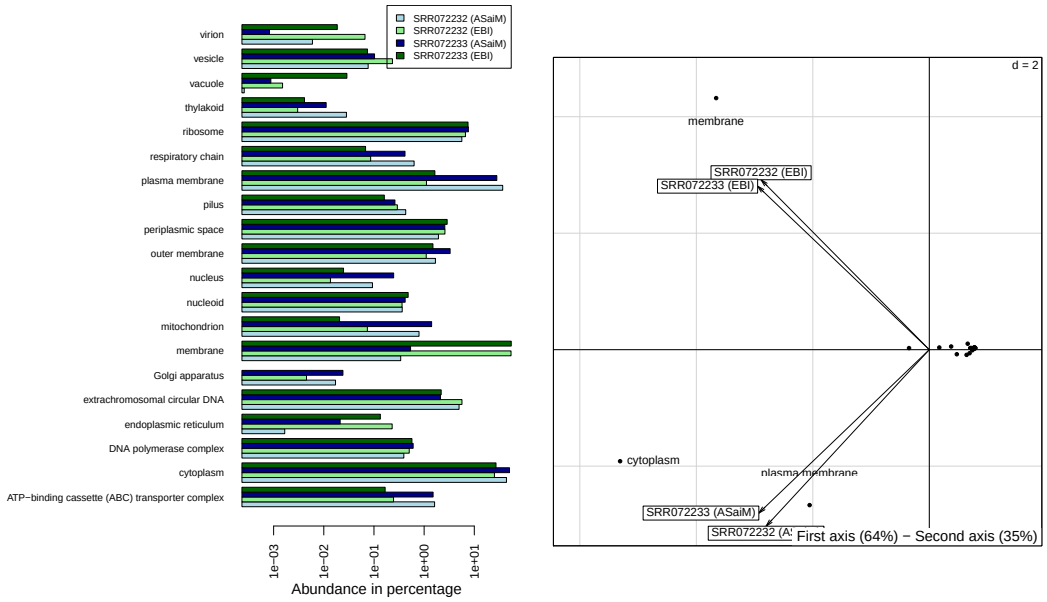

Fig. 16. Barplot representation (in left, logarithm scale) and scatter diagram of principal component analysis of the normalized relative abundances (in percentage) of the cellular component GO slim terms for both samples (SRR072233 and SRR072233) and both workflows (EBI metagenomics and ASaiM). The relative abundances of each GO slim terms is normalized by the sum of relative abundance for the found cellular component GO slim terms in both samples and with both workflows.

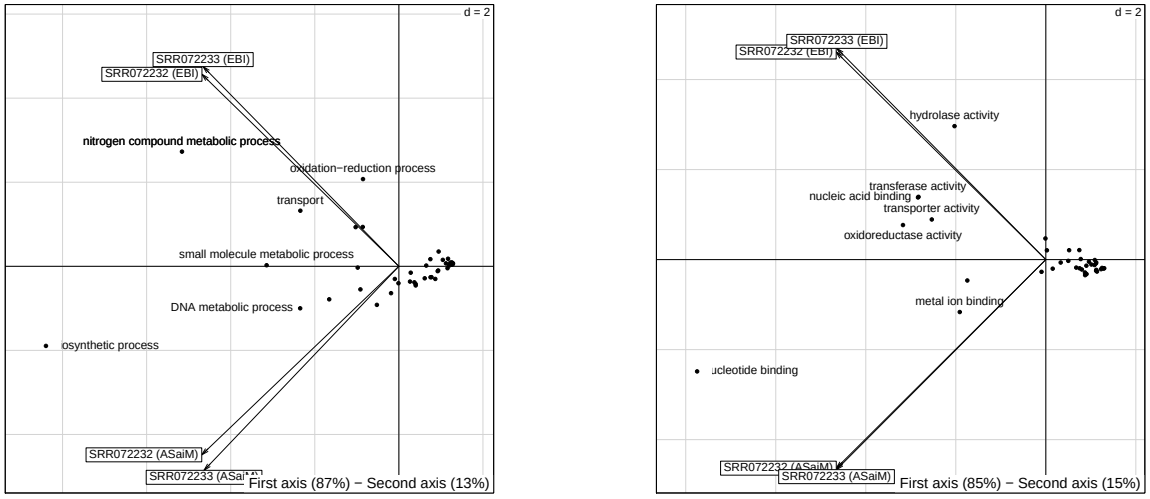

Fig. 17. Scatter diagram of principal component analysis of the normalized relative abundances (in percentage) of the biological process (in left) and of the molecular functions (in right) GO slim terms for both samples (SRR072233 and SRR072233) and both workflows (EBI metagenomics and ASaiM). The relative abundances of each GO slim terms is normalized by the sum of relative abundance for the found biological process GO slim terms in both samples and with both workflows.

References

Abubucker,S. *et al.* (2012) Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Comput Biol*, **8**, e1002358.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Kopylova,E. *et al.* (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.

Lee,J.-H. *et al.* (2011) rRNASelector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *J Microbiol.*, **49**, 689–691.

Ondov,B.D. *et al.* (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Segata,N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Meth*, **9**, 811–814.

Truong,D.T. *et al.* (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Meth*, **12**, 902–903.