Supplementary material 2

Validation of ASaiM framework and its workflows on HMP mock community smaples 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

ASaiM framework, particularly the main workflow, was tested and validated on two mock metagenomic datasets from a controlled microbiota community (with 22 known microbial strains). These datasets are available on *EBI metagenomics* dataset.

Taxonomic and functional results produced by ASaiM framework have been extensively analyzed and compared with results obtained with *EBI metagenomics* pipleine (https://www.ebi.ac.uk/metagenomics/pipelines/1.0) (Hunter *et al.*, 2014). Details about these analyses (workflows, scripts) are available on a dedicated GitHub repository and the results on Zenodo.

These analyses validate ASaiM framework and its main workflow. Hence, the main workflow produces accurate and precise taxonomic assignations, wide functional results (gene families, pathways, GO slim terms) and relations between taxonomic and functional results, in few hours on a standard computer.

1 Data

Two HMP mock community samples are available on EBI met. Both of them contain a genomic mixture of same 22 microbial strains (Table 1). The samples differ only by the abundances of the strains: in the first sample (SRR072232), the ribosomal RNA operon counts vary by up to four orders of magnitude per strains (Table 1), wherease the second sample (SRR072233) contains equimolar ribosomal RNA operon counts per strain (Table 1).

Both samples were sequenced using 454 GS FLX Titanium to get 1,22,169 raw metagenomic sequences for the first dataset and 1,386,198 raw metagenomic sequences for the second dataset.

2 Methods

Both datasets have been analyzed using ASaiM framework. The results extensively analyzed and compared to expected results and *EBI metagenomics* results. Details about these analyses (workflows, scripts) are available on a dedicated GitHub repository and the results on Zenodo.

2.1 Abundance computation using mapping on reference genomes

The expected abundances based on ribosomal RNA operon counts. During biological manipulations and sequencing, some bias may arise that modify the abundances of strains. Indeed, to get "real" abundances of expected strains, raw metagenomic sequences of both samples are mapped on genomes of expected strains using BWA (Li and Durbin, 2009, 2010).

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 strains, their taxonomy and their ribosomal RNA operon counts (abundance, from metadata on EBI metagenomics database) on both samples (SRR072232 and SRR072233)

2.2 Analyses using *EBI Metagenomics*

Both datasets have been analysed with *EBI metagenomics* pipeline (Version 1.0) (Figure 1). We downloaded the results and formatted them to help comparison with ASaiM results.

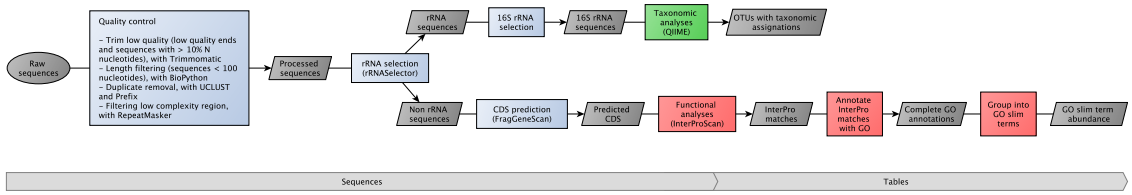

Fig. 1. EBI metagenomics pipeline (version 1.0). The grey boxes correspond to data, the blue boxes to pretreatment steps, the red boxes to functional analysis steps and the green boxes to taxonomic analysis steps.

First, from OTUs with taxonomic assignation, abundances of each assigned clade are extracted. 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 for each taxonomic level.

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 results (Figure 2). Annotations are formatted to extract relative abundances (in percentage) of GO slim term annotations inside each GO slim term group (cellular components, biological processes and molecular functions).

2.3 Analyses using ASaiM framework

Main workflow (Figure 2) of ASaiM framework is used to analyze both datasets

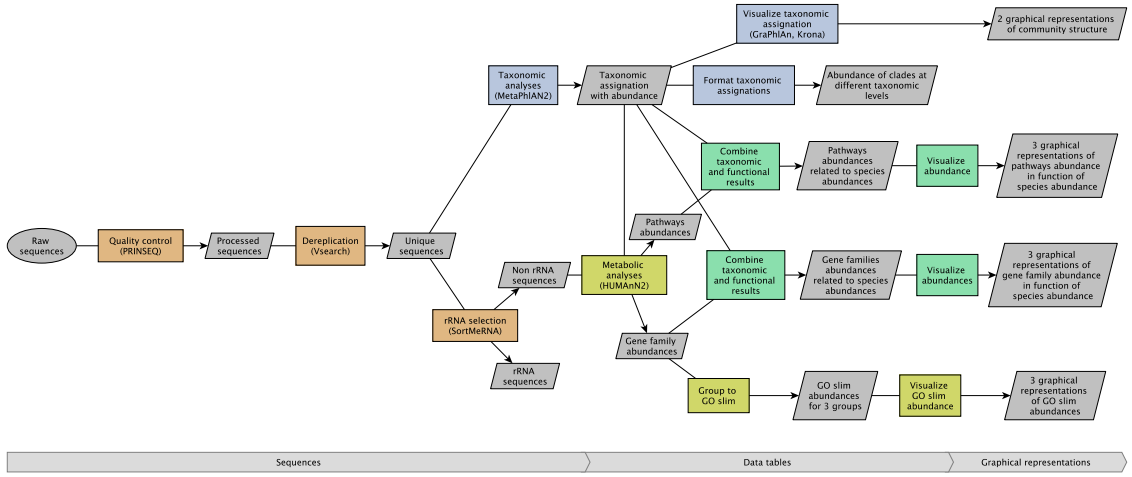

Fig. 2. ASaiM workflow for analysis of raw single-end microbiota sequences. This workflow is available with ASaiM Galaxy instance and used to analyze both datasets. The grey boxes correspond to data, the blue boxes to pretreatment steps, the red boxes to functional analysis steps and the green boxes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

For these analyses, ASaiM framework is deployed on a computer with Debian GNU/Linux System, 8 cores Intel(R) Xeon(R) at 2.40GHz and 32 Go of RAM. During workflow execution, we follow size of used memory and execution time (Table 2). Indeed,

workflow execution is 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). And, the size of the process in memory is stable over workflow execution (variability inferior to 40 kb) (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
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)

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).

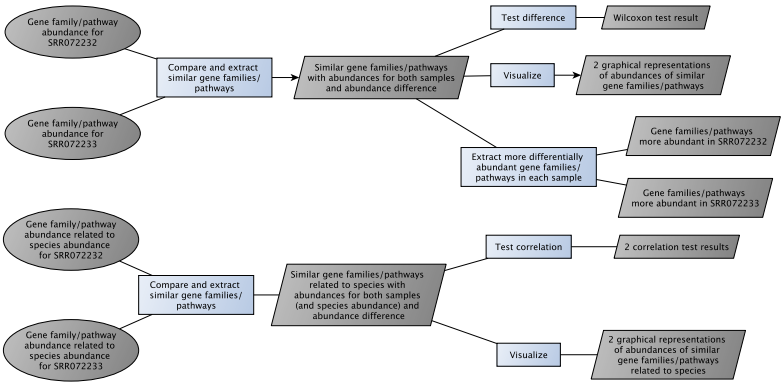

Fig. 3. Workflow to compare ASaiM functional results (gene families or pathways) between both samples. This workflow is available with ASaiM Galaxy instance. The grey boxes correspond to data, the blue boxes to processing steps.

2.4 Comparison of *EBI metagenomics* results and ASaiM results

EBI metagenomics results and ASaiM ones 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 both pipelines, 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.

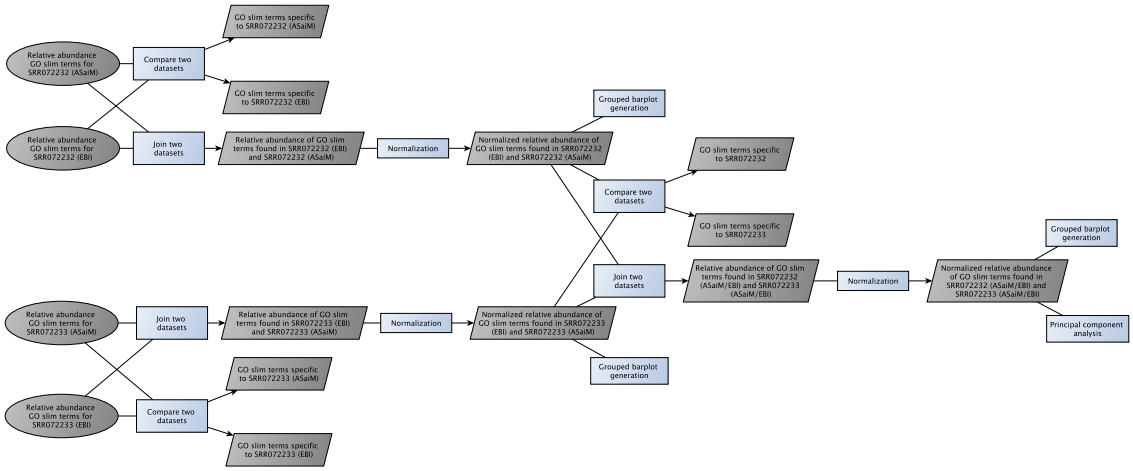

Fig. 4. Workflow to compare GO slim annotation abundances between samples (SRR072232, SRR072233) and workflows (EBI metagenomics, ASaiM). This workflow is available with ASaiM Galaxy instance. The grey boxes correspond to data, the blue boxes to processing steps.

3 Results

3.1 Results of pretreatments in pipelines

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/rDNA 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). So, even with similar raw input 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%
rDNA sequences	8,910	0.9%	16,016	1.4%	9,214	0.8%	13,850	1%
non rDNA 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 quality control with 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. The differences come from then from the used tools and their underlying algorithms and implementations.

In both datasets and with both workflows, few rDNA 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 rDNA sequences are systematically found with ASaiM workflow. Indeed, in *EBI metagenomics* pipeline (Figure 1), *rRNASelector* (Lee *et al.*, 2011) is used to select rDNA 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, and then 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 then 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 Abundances of expected strains and taxonomy

The expected abundances based on ribosomal RNA operon counts. During biological manipulations and sequencing, some bias may arise that modify the abundances of strains. Indeed, to get "real" abundances of expected strains, raw metagenomic sequences of both samples are mapped on genomes of expected strains.

For SRR02232, variations in abundances between species are similar using mapping than RNA operon count (Figure 5). Observations are different for SRR072233 (Figure 5): expected abundances (based on RNA operon abundances) are identical for all species, but unexpected variations exist for mapping based abundances.

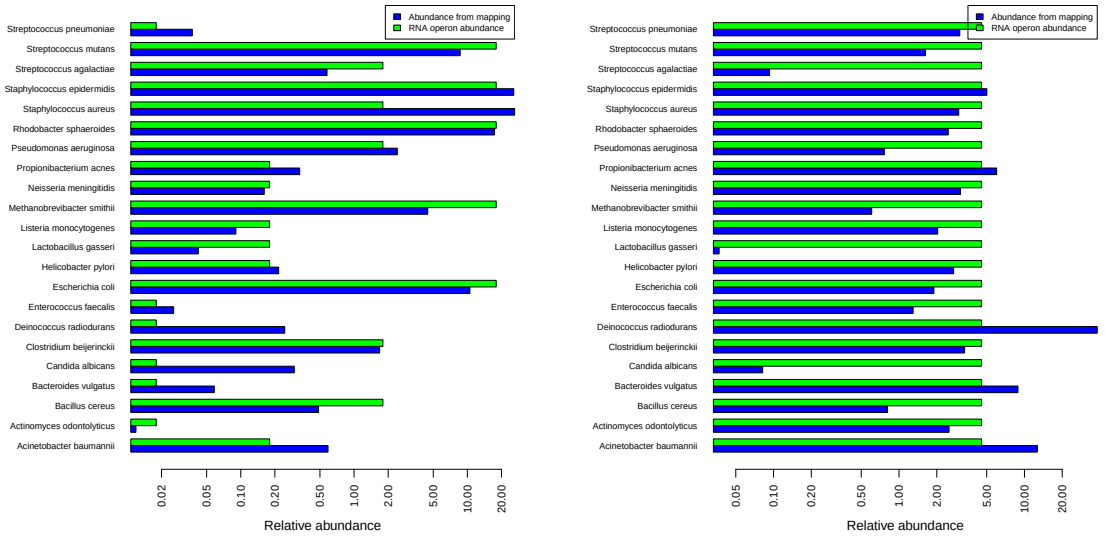

Fig. 5. Comparison of relative abundances between expectation given the ribosomal RNA operon counts (green, Table 1) and mapping against reference genomes for both samples (SRR072232 on left, SRR072233 on right)

These differences between expected abundances (from RNA operon counts) and mapping-based abundances may be due to bias induced during biological manipulations or sequencing. As next taxonomic analyses are based on raw metagenomic sequences, abundances based on mapping counts are used on further analyses instead of abundances based on ribosomal RNA operon counts from metadata (Figure 6).

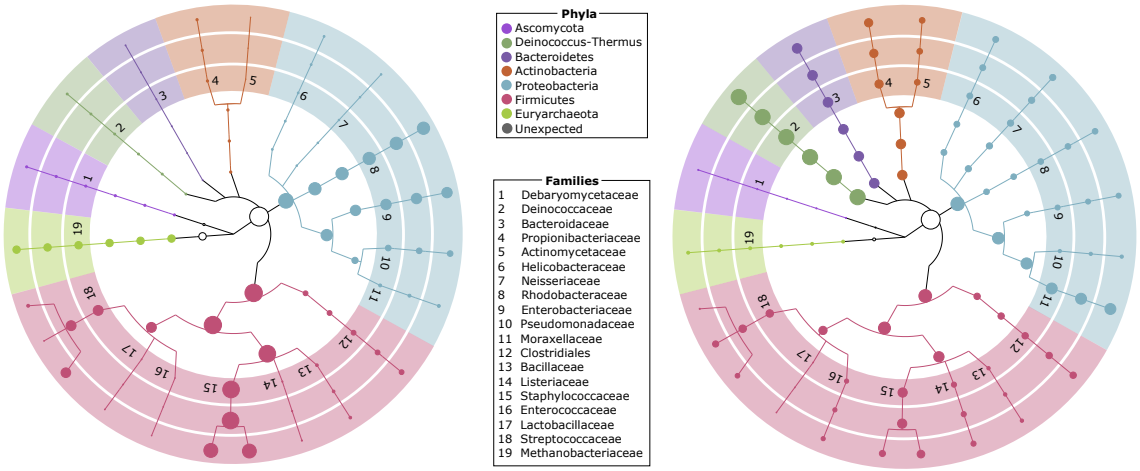

Fig. 6. Expected taxonomy for SRR072232 (left) and SRR072233 (right) from domains to species. Circle diameters at each taxonomic levels are proportional to mapping-based relative abundance of corresponding taxon.

3.2.2 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 interactive representations of taxonomic assignation and *GraPhlan* for static representations. These static representations are modified (legend *e.g.* colors, numbers for families) to help comparison with expected taxonomy (Figure 7).

Despite same expected species, taxonomic diversity in SRR072232 dataset is reduced compared to the one in SRR072233 dataset (Figure 7). Less taxons are found for each taxonomic levels. From 22 expected species (Table 1), 17 are found for SRR072232 and 20 for SRR072233 (Figure 8). The 2 expected species (*Candidata albicans* and *Lactobacillus gasseri*) missing in SRR072233 dataset are also missing in SRR072232 dataset (Figure 8). This may be due to a lack of phylogenetic markers for these species in the database used in *MetaPhlAn*. On the other hand, few sequences of these species in SRR072233 are found using mapping on expected species genomes. The signal may be too low to detect the species. Indeed, all species with mapping-based abundance smaller than 0.1% are not found using ASaiM for both datasets (Figure 8).

For SRR072232 datasets, two species with mapping-based abundance higher than 0.1% are not found: *Candida albicans* and *Bacillus cereus thuringiensis*. The first species is not found also with ASaiM in SRR072333, phylogenetic markers for this species may be lacking in *MetaPhlAn2* database. As the second species is found with ASaiM in SRR072333, same explanation based on lack on corresponding phylogenetic markers does not hold.

3.2.3 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.

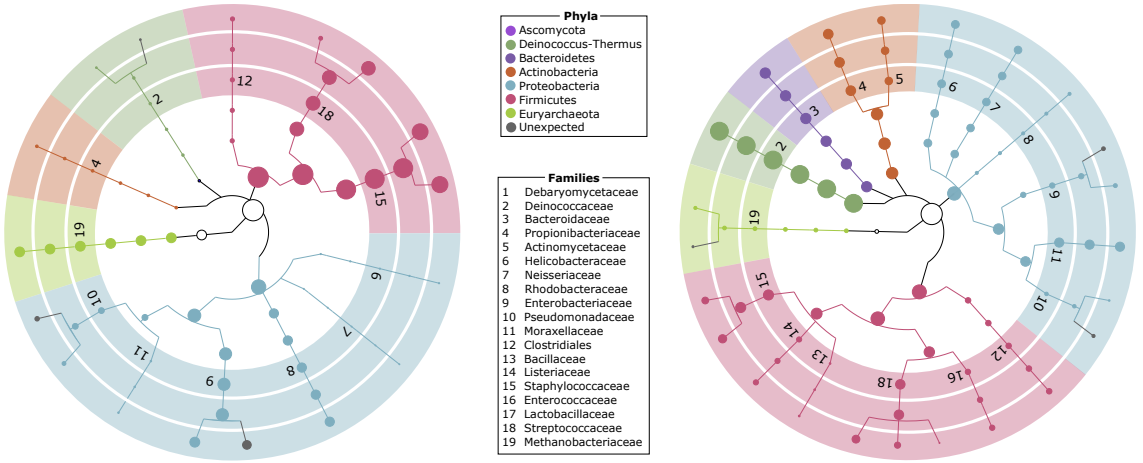

Fig. 7. Taxonomy for SRR072232 (left) and SRR072233 (right) from domains to species, found with ASaiM framework. Circle diameters at each taxonomic levels are proportional to mapping-based relative abundance of corresponding taxon. Colors and family numbers are the same as the ones used in Figure 6. Gray circles and lines represent unexpected lineages.

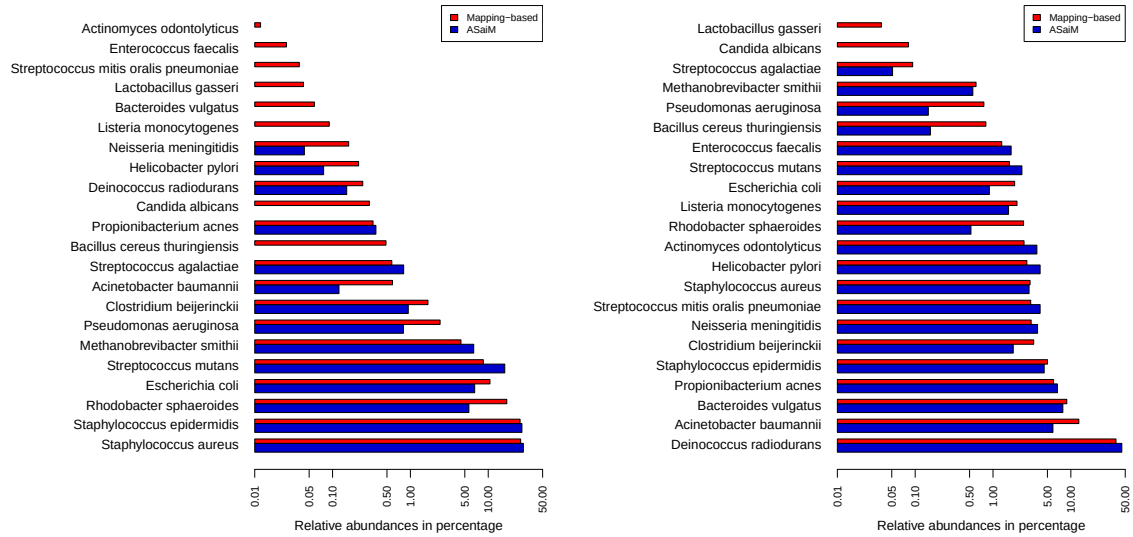

Fig. 8. Relative abundances (percentage in log scale) of expected species for SRR072232 (left) and SRR072232 (right) with comparison between expected abundances (red thin bars) and abundances obtained with ASaiM (blue wide bars)

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.*, 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). Nevertheless, taxonomic lineages from *EBI metagenomics* are limited to family level (Figure 9), while they go to species level with ASaiM (Figure 7). Hence, 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).

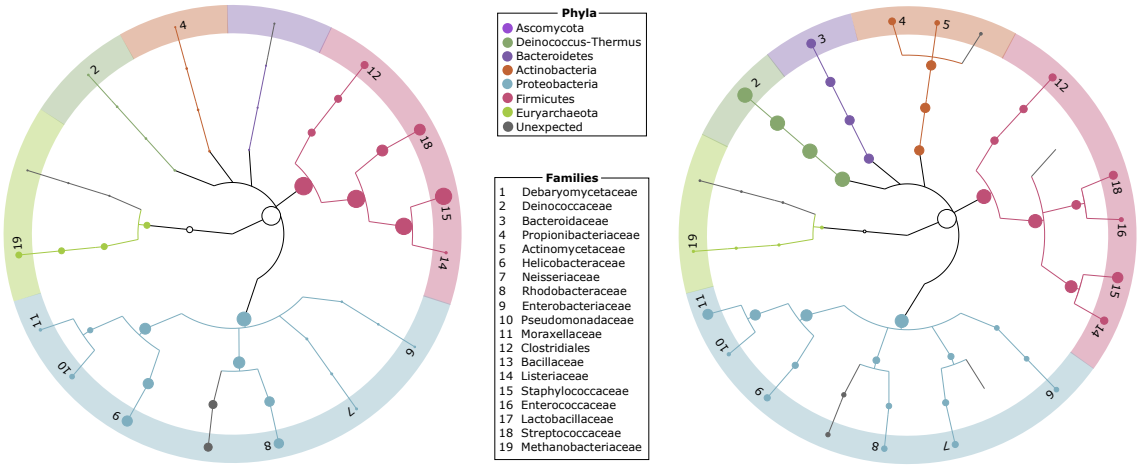

Fig. 9. Taxonomy for SRR072232 (left) and SRR072233 (right) from domains to families, found with EBI metagenomics pipeline. Circle diameters at each taxonomic levels are proportional to mapping-based relative abundance of corresponding taxon. Colors and family numbers are the same as the ones used in Figure 6. Gray circles and lines represent unexpected lineages.

With both *EBI metagenomics* and ASaiM, some observed taxonomic assignations are unexpected (Tables 4 and 5, Figures 7 and 9). 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 8) 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, Figure 9). Taxonomic assignations with *MetaPhlAN* (Truong *et al.*, 2015; Segata *et al.*, 2012) are then more accurate and precise.

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* (Figure 9) and species for ASaiM (Figure 7). Comparison of taxonomic results focus then on family level (Figure 10).

Similarly, to previous observations on raw ASaiM results, species with mapping-based abundance smaller than 0.1% are not found either with ASaiM or with *EBI metagenomics* (Figure 10). Nonetheless, the detection threshold seems slightly smaller for *EBI metagenomics*: for SRR072232, Listeriaceae family is detected with *EBI metagenomics* and not with ASaiM (Figure 10). On the other hand, Bacillaceae and Debaryomycetaceae families are not found with *EBI metagenomics* for both datasets (Figure 10), despite

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)

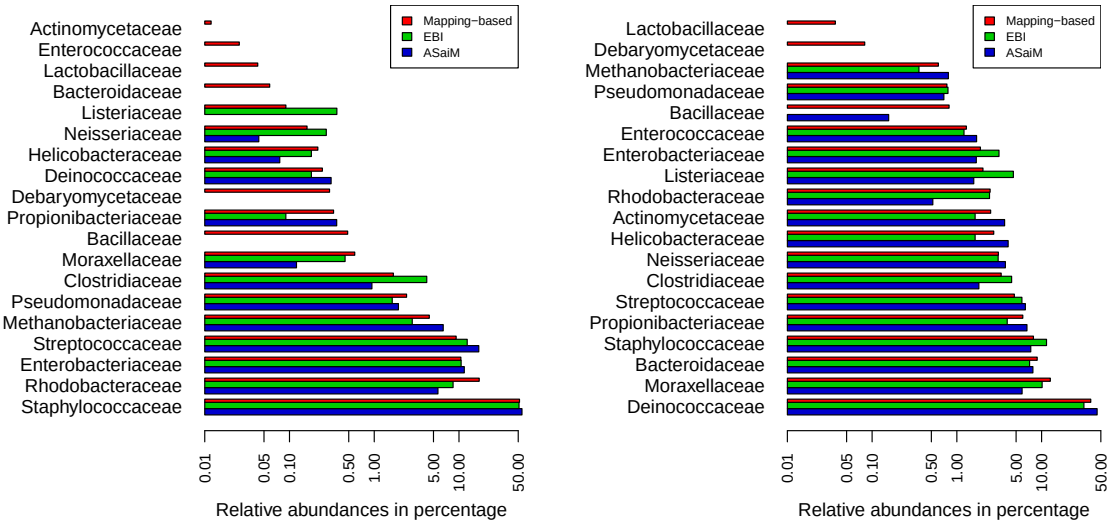

Fig. 10. Relative abundances of expected families for SRR072232 (left) and SRR072233 (right) with comparison between expected abundances (red thin bars), abundances obtained with EBI metagenomics (green wide bars) and abundances obtained with ASaiM (blue wide bars) and

mapping-based abundance higher than 0.1%. Used databases may be then incomplete regarding some phylogenetic markers, particularly the ones for missing families.

Except missing species, variations in observed abundances for *EBI metagenomics* or ASaiM correspond to variations in mapping-based abundances (Figure 8): small observed abundances for small mapping-based abundances and high observed abundances for high mapping-based abundances.

For a broader comparison, a principal component analysis (PCA) is runned, 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 (98% for both datasets, Figure 11). Mapping-based, *EBI metagenomics* and ASaiM abundances are not discriminated on this first axis (Figure 11), only on the second one which explains less than 2% of overall data variability. Differences between mapping-based, *EBI metagenomics* and ASaiM abundances are then reduced. Hence, similar abundances for observed families are then obtained for *EBI metagenomics* and ASaiM and these abundances are close to abundances computed using mapping to expected species.

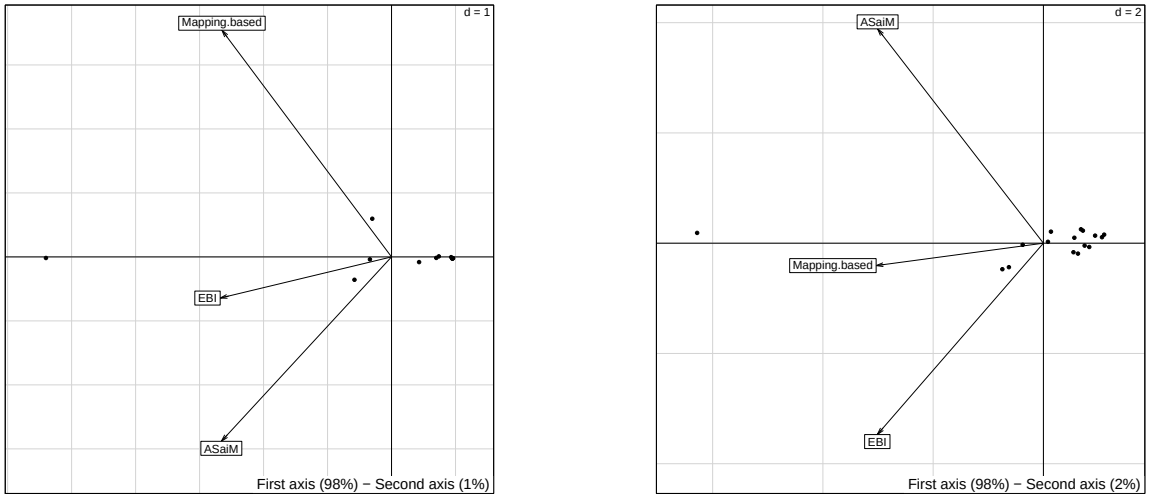

Fig. 11. Scatter diagram of principal component analysis of the relative abundances (in percentage) of families for SRR072232 (in left) and SRR072233 (in right). Only observed families in EBI metagenomics or ASaiM are used in these analyses.

ASaiM framework gives taxonomic results which are more accurate, complete (until species level) and statistically supported (based on more sequences) than *EBI metagenomics*. Moreover, community structure found with ASaiM framework is close to expected community structure of these mock datasets.

3.3 Functional analyses

3.3.1 Raw ASaiM results

In ASaiM framework (Figure 2), *HUMAnN2* (Abubucker *et al.*, 2012) is used for functional analyses. This tool profiles presence/absence and abundance of UniRef50 gene families and MetaCyc pathways from metagenomic/metatranscriptomic datasets. It then describes the metabolic profil of a microbial community.

HUMAnN2 generates three outputs: abundances of 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 rDNA sequences (Table 3).

Datasets are constitued of metagenomic sequences from genomic mixture of identical 22 microbioal strains (Table 1). Differences between datasets are on abundance of these strains. Similar metabolic functions made by same species are then supposed to be found in both datasets, but with different abundances.

However, differences of metabolic functions between both datasets are observed. Sets of gene families are different: 44,933 gene families are found in both samples (<46% for both samples, Table 7). However, different gene families have a limited impaxt on overall metabolism (< 50% of relative abundance, Table 7). Global metabolism functions such as pathways are similar in both datasets (> 95% of similar pathways representing > 99.5% of overall abundance, Table 7). Hence, the unexpected observed differences are limited and may be due to bias induced by biological manipulations or sequencing.

		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	

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.

On the other hand, abundances of similar metabolic functions are different (Figure 12), as expected.

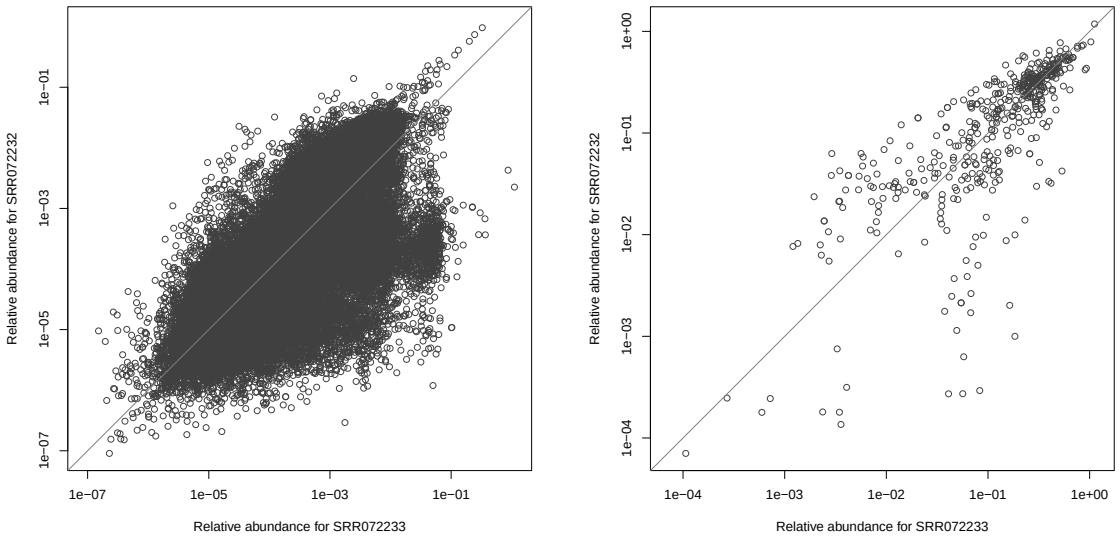

Fig. 12. Normalized relative abundances (%) for similar UniRef50 gene families (left graphics) and MetaCyc pathways (right graphics) for both samples (SRR072233 and SRR072233). The relative abundances of each similar characteristics (gene families or pathways) is computed with HUMAnN2 and normalized by the sum of relative abundance for all similar characteristics.

With more than 90,000 gene families and almost 500 pathways, the metabolic profile of studied microbial community is too large to get a broad overview. Each gene family and pathway is precise and related to specific metabolic functions. This information is interesting when you need detailed metabolic information and to go deeply inside metabolic profil. However, to get a broad overview of the metabolic processes, UniRef50 gene families and even MetaCyc pathways are too numerous and too precise. UniRef50 gene families and their abundances are then grouped into slim Gene Ontology terms (Figure **??**). Inside the 3 groups, the GO slim terms have similar abundances in both samples (Figure 13).

Both communities, with same expected strains but with different abundances of these species, are similarly doing metabolic tasks. Hence, functional results obtained with ASaiM fill the expectations.

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

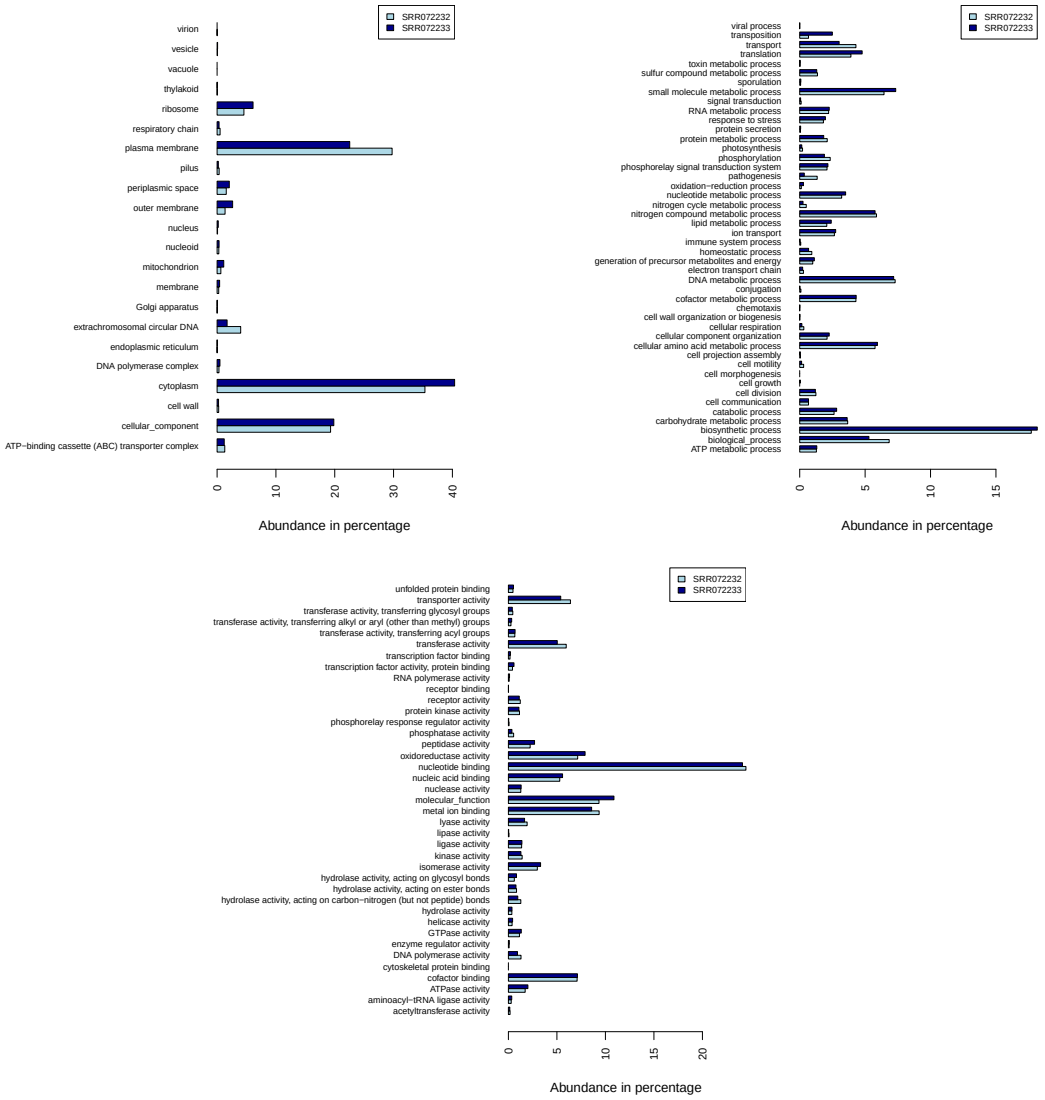

Fig. 13. Relative abundances of GO slim terms in SRR072232 and SRR072233 for cellular components (top left), biological processes (top right) and molecular function (bottom)

directly comparable. But, in both workflows, UniRef50 gene families and InterPro proteins are grouped into Gene Ontology slim terms to get a broad overview of functional profile of the community. These GO slim terms are grouped into 3 groups: cellular components, biological processes and molecular functions.

Few GO slim terms are different for *EBI metagenomics* and ASaiM (Table 8). They are negligible in term of relative abundance inside the three groups.

		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 8. 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)

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 14). 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 14 and 15).

Scatter representation of first plan (first two axes) of the PCA is similar for the three groups (Figures 14 and 15). First axis explains most data variability (between 64% and 87%, Table 9). GO slim terms found on left part of scatter representation (Figures 14 and 15) are highly abundant: cellular processes related to membrane and cytoplasm (Figure 14), biosynthetic processes, nitrogen compound metabolic process, small molecular metabolic process, transport and DNA metabolic process for biological processes (Figure 15) and nucleotide binding for molecular functions (Figure 15). This first axis does not discrimate samples or workflows (Figures 14 and 15). Results from ASaiM workflow are then similar in term of GO slim term abundances to the one obtained with *EBI metagenomics* pipeline.

	Cellular components	Biological processes	Molecular functions
First axis	64%	87 %	85%
Second axis	35%	13 %	15%

Table 9. Explained variability by axes of Principal component analysis (PCA) for GO slim terms of cellular components, biological processes and molecular functions

The discrimination between *EBI metagenomics* and ASaiM results appears with second axis (Table 9), explaining between 13% and 35% of overall data variability (Table 9). Some GO slim terms such as membrane, hydrolase activity or nitrogen compound metabolic process are found in higher proportion in *EBI metagenomics* results than in ASaiM and some like biosynthetic process, plasma membrane or nucleotide binding in lower proportion (Figures 14 and 15).

None of the first two axes discriminates between 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: the discrimination between both workflow results appears as a secondary explanation for variability in GO slim term abundances.

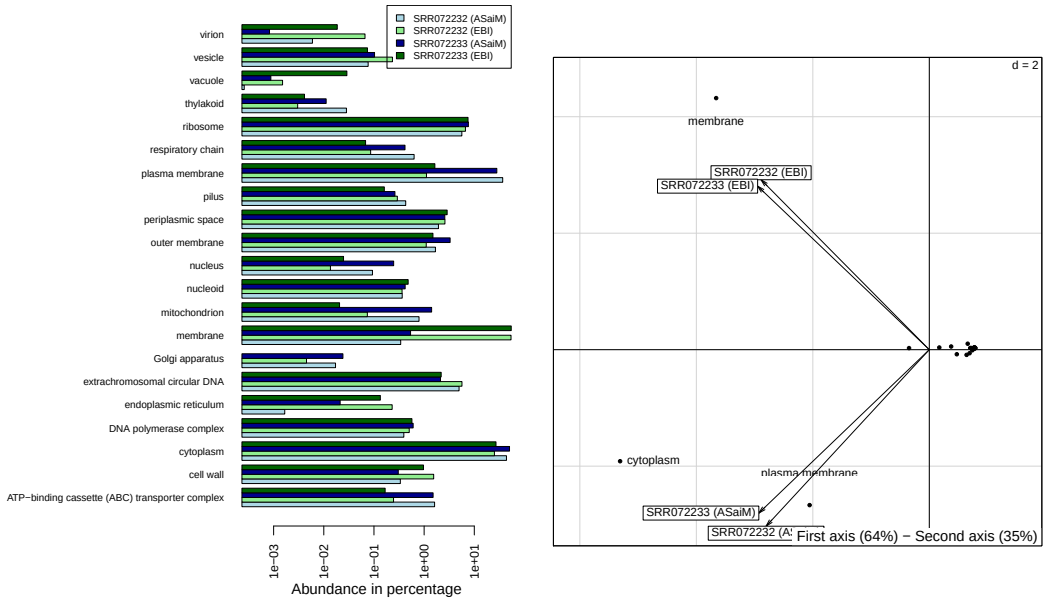

Fig. 14. Barplot representation (in left, logarithm scale) and scatter diagram of principal component analysis of the normalized relative abundances (in percentage) of the cellular component GO slim terms for both samples (SRR072233 and SRR072233) and both workflows (EBI metagenomics and ASaiM). The 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.

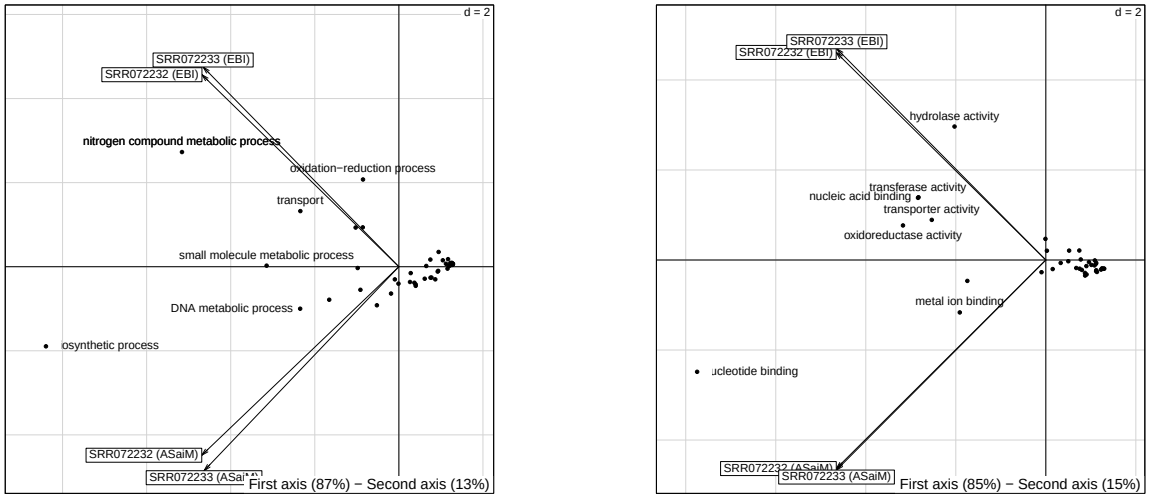

Fig. 15. Scatter diagram of principal component analysis of the normalized relative abundances (in percentage) of the biological process (in left) and of the molecular functions (in right) GO slim terms for both samples (SRR072233 and SRR072233) and both workflows (EBI metagenomics and ASaiM). The relative 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.

3.4 Taxonomically-related functional results

In *HUMAnN2* results, abundances of gene families and pathways are stratified at the community level. We can then relate functional results to taxonomic result and answer questions such as "Which species contribute to which metabolic functions? In which proportion?". < 35% of gene families (> 90% of relative abundance) and > 80% pathways (> 50% of relative abundance) can be related to the community structure (species and their abundance, Table 10).

		UniRef50 gene families		MetaCyc pathways	
		SRR072232	SRR072233	SRR072232	SRR072233
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 10. 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.

For both samples, we observed a significant correlation between CDS number in species (data from GenBank) and number of gene families found for these species (Table 11). The correlation, not so bad, is yet not perfect. Indeed, gene families have not a direct mapping to CDS (paralogs, duplications, …) and rely on exhaustivity of the reference database (UniRef) used by HUMAnN2. So, it may be interesting to investigate the relation between gene families corresponding to found species in UniRef and gene families found using HUMAnN2. This information is not available, but having a significant correlation between gene family number and CDS number is already a great point.

		UniRef50 gene families		MetaCyc pathways	
		SRR072232	SRR072233	SRR072232	SRR072233
Number					
Correlation with species CDS number	r^2	0.71	0.60		
	p-value	$4.67 \cdot 10^{-3}$	$5.09 \cdot 10^{-3}$		
Mean abundance (Figure 16)					
Correlation with species abundance	r^2	0.95	0.98	0.90	0.93
	p-value	$1.51 \cdot 10^{-7}$	$2.9 \cdot 10^{-13}$	$1.91 \cdot 10^{-7}$	$5.88 \cdot 10{-}12$
Difference of mean abundance					
Correlation with species abundance difference	r^2	0.89		0.84	
	p-value	$4.12 \cdot 10^{-7}$		$4.65 \cdot 10^{-6}$	

Table 11. Correlation coefficients and p-values (Pearson's test) for UniRef50 gene families and MetaCyc pathways obtained with HUMAnN2 for both samples (SRR072233 and SRR072233). CDS number for each strain has been extracted from GenBank given the links in Table 1

For both samples, relative abundances of gene families and pathways are highly correlated to observed relative abundance of involved species (Figure 16 and Table 11). Sequences of an abundant species in a community are supposed to be abundant in metagenomic sequences of the community. This relation concerns all sequences, particularly sequences corresponding to gene families. For pathways, the relation is more tricky: a pathway is identified if a high proportion of gene families involved in 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, leading to a tricky relation between species abundance and pathway abundance. But, the high correlations between species relative abundance and mean relative pathway abundance (Figure 16, Table 11) confirm good pathway reconstructions in our datasets, particularly for abundant species. To accentuate previous observations and conclusion, we also observe

a strong and significant correlation between species abundance difference and difference of gene family and pathway mean abundance between both samples (Table 11).

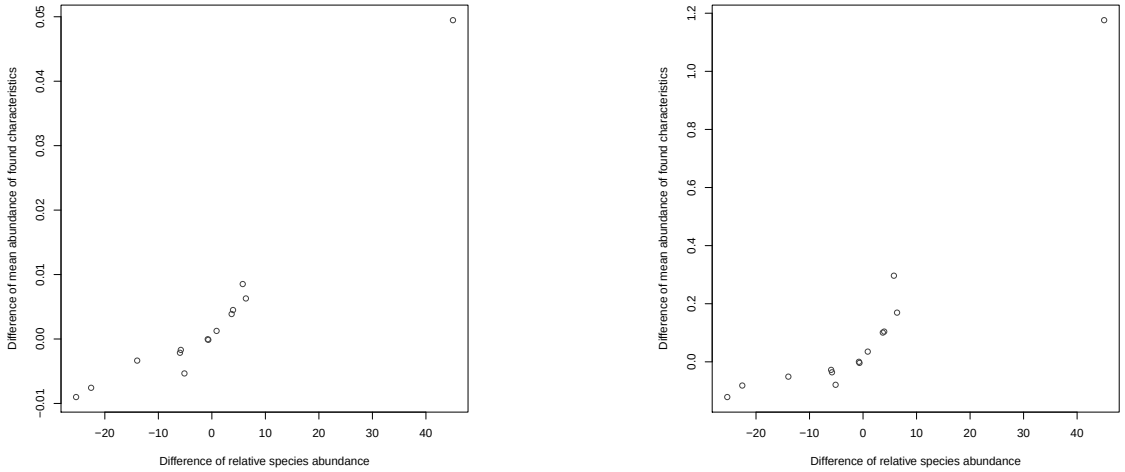

Fig. 16. Difference in mean abundances for gene families (left) and pathways (right) in function of difference of related species abundance between both samples. Correlation coefficients and p-values are detailed in Table 11

Hence, our approach based on MetaPhlAn2 and HUMAnN2 gives accurate and relevant taxonomically-related functional results.

4 Conclusion

With ASaiM framework, raw sequences from a metagenomic dataset are fast analyzed (in few hours in a standard computer). Moreover, based on Galaxy, ASaiM framework posseses all Galaxy's strength: accessibility, reproducibility and also modularity. Numerous intermediary results can also be accessed during whole workflow execution.

Taxonomic analysis using *MetaPhlAn2* gives a great insight on community structure wtih complete, accurate and statistically supported information. With *HUMAnN2* results and post-treatments on functional results, we get a broad overview of metabolic profile of studied microbial community. Furthermore, this metabolic profile is related to community structure to get information such as which species is involved in which metabolic function. This relation between function and taxonomy is really specific to ASaiM and not found in solutions like *EBI metagenomics*.

ASaiM framework based on Galaxy, numerous tools and workflows is a then powerful framework to analyze microbiota from shotgun raw sequence data.

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.

Hunter,S. *et al.* (2014) EBI metagenomics—a new resource for the analysis and archiving of metagenomic data. *Nucl. Acids Res.*, **42**, D600–D606.

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.

Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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.