Supplementary material 2

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

Bérénice Batut[1], Clémence Defois[1], Kévin Gravouil[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 and its workflows are tested and validated on two mock metagenomic datasets of a controlled microbiota community (with 22 known microbial strains). These datasets are available on *EBI metagenomics* database.

Taxonomic and functional results produced by ASaiM framework are extensively analyzed and compared with results obtained with the *EBI metagenomics* pipeline (Hunter *et al.*, 2014). Details about these analyses (workflows, scripts) are available on a dedicated GitHub repository and results with all intermediary and final files on Zenodo.

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 commodity computer. Hence, these analyses validate ASaiM framework and its main workflow.

1 Data

Two mock community samples for Human Microbiome Project (HMP) are available on *EBI metagenomics* database. Both samples contain a genomic mixture of 22 microbial strains (Table 1) with differences in abundances of the strains. In first sample (SRR072232), the rRNA operon counts vary by up to four orders of magnitude per strains (Table 1), whereas second sample (SRR072233) contains equimolar ribosomal rRNA operon counts per strain (Table 1).

Both samples were sequenced using 454 GS FLX Titanium to get 1,225,169 raw metagenomic sequences for first dataset and 1,386,198 raw metagenomic sequences for second dataset.

2 Methods

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

2.1 Abundance computation using mapping on reference genomes

Targeted abundances of strains are based on rRNA operon counts added for each strains to simulate the mock community. To get "real" abundances of expected strains in metagenomic datasets, raw reads are mapped on reference genomes of expected strains using BWA (Li and Durbin, 2009, 2010) (using default parameters). Relative abundances based on mapping results are compared to relative abundances based on rRNA operon counts as described in sample metadata (Figure 1).

Taxonomy									Targeted abundances (%)	
Domain	Kingdom	Phylum	Class	Order	Family	Genus	Species	Strains	SRR072232	SRR072233
Archaea	Archaea	Euryarchaeota	Methanobacteria	Methanobacteriales	Methanobacteriaceae	*Methanobrevibacter*	*Methanobrevibacter smithii*	ATCC 35061	$1.797 \cdot 10^{1}$	4.545
Bacteria	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	*Actinomyces*	*Actinomyces odontolyticus*	ATCC 17982	$1.797 \cdot 10^{-2}$	4.545
					Propionibacteriaceae	*Propionibacterium*	*Propionibacterium acnes*	DSM 16379	$1.797 \cdot 10^{-1}$	4.545
		Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	*Bacteroides*	*Bacteroides vulgatus*	ATCC 8482	$1.797 \cdot 10^{-2}$	4.545
		Deinococcus-Thermus	Deinococci	Deinococcales	Deinococcaceae	*Deinococcus*	*Deinococcus radiodurans*	DSM 20539	$1.797 \cdot 10^{-2}$	4.545
		Firmicutes	Bacilli	Bacillales	Bacillaceae	*Bacillus*	*Bacillus cereus thuringiensis*	ATCC 10987	1.797	4.545
					Listeriaceae	*Listeria*	*Listeria monocytogenes*	ATCC BAA-679	$1.797 \cdot 10^{-1}$	4.545
					Staphylococcaceae	*Staphylococcus*	*Staphylococcus aureus*	ATCC BAA-1718	1.797	4.545
							Staphylococcus epidermidis	ATCC 12228	$1.797 \cdot 10^{1}$	4.545
				Lactobacillales	Enterococcaceae	*Enterococcus*	*Enterococcus faecalis*	ATCC 47077	$1.797 \cdot 10^{-2}$	4.545
					Lactobacillaceae	*Lactobacillus*	*Lactobacillus gasseri*	DSM 20243	$1.797 \cdot 10^{-2}$	4.545
					Streptococcaceae	*Streptococcus*	*Streptococcus agalactiae*	ATCC BAA-611	1.797	4.545
							Streptococcus mutans	ATCC 700610	$1.797 \cdot 10^{1}$	4.545
							Streptococcus mitis oralis pneumoniae	ATCC BAA-334	$1.797 \cdot 10^{-2}$	4.545
			Clostridia	Clostridiales	Clostridiaceae	*Clostridium*	*Clostridium beijerinckii*	ATCC 51743	1.797	4.545
		Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	*Rhodobacter*	*Rhodobacter sphaeroides*	ATCC 17023	$1.797 \cdot 10^{1}$	4.545
			Betaproteobacteria	Neisseriales	Neisseriaceae	*Neisseria*	*Neisseria meningitidis*	ATCC BAA-335	$1.797 \cdot 10^{-1}$	4.545
			Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	*Helicobacter*	*Helicobacter pylori*	ATCC 700392	$1.797 \cdot 10^{-1}$	4.545
			Gammaproteobacteria	Pseudomonadales	Moraxellaceae	*Acinetobacter*	*Acinetobacter baumannii*	ATCC 17978	$1.797 \cdot 10^{-1}$	4.545
					Pseudomonadaceae	*Pseudomonas*	*Pseudomonas aeruginosa*	ATCC 47085	1.797	4.545
				Enterobacteriales	Enterobacteriaceae	*Escherichia*	*Escherichia coli*	ATCC 70096	$1.797 \cdot 10^{1}$	4.545
Eukaryotes	Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	Debaryomycetaceae	*Candida*	*Candida albicans*	SC5314	$1.797 \cdot 10^{-2}$	4.545

Table 1. Expected strains, their taxonomy and their targeted relative abundance (percentage) based on ribosomal RNA operon counts (abundance, from metadata on EBI metagenomics database) on both samples (SRR072232 and SRR072233)

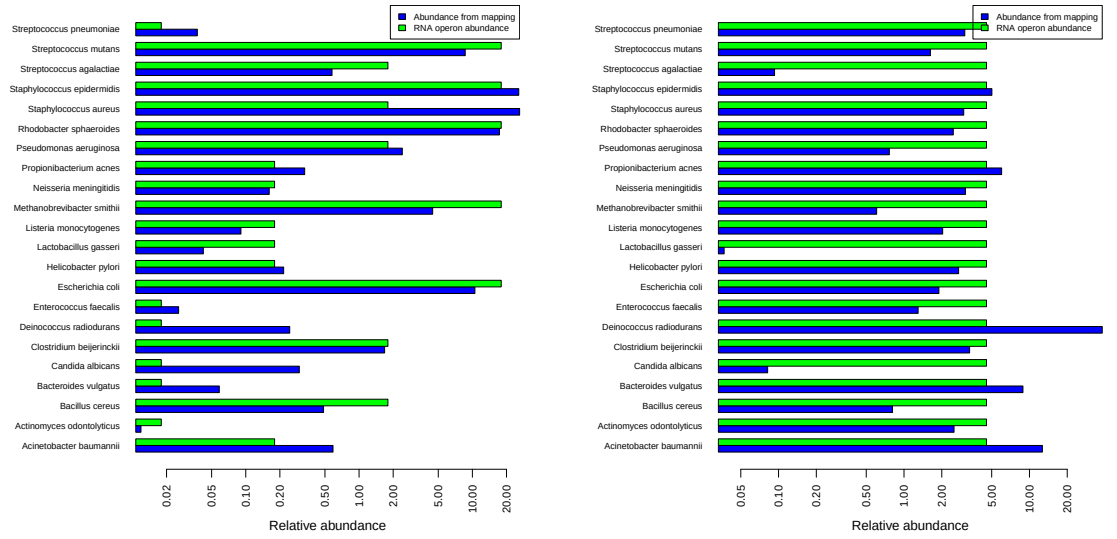

Fig. 1. Comparison of relative abundances (percentage, in log scale) between expectation given the ribosomal RNA operon counts (green, Table 1) and mapping against reference genomes for both samples (SRR072232 on left, SRR072233 on right)

For SRR02232, similar variations in abundances between species are observed using mapping and rRNA operon count (Figure 1). Observations are different for SRR072233 (Figure 1): expected abundances (based on RNA operon abundances) are identical for all species, but unexpected variations are observed for mapping based abundances. These differences between expected abundances (from RNA operon counts) and mapping-based abundances may be due to bias induced during biological manipulations or sequencing. The observed differences can also be explained by copy number variation of targeted rRNA operon between species. Indeed, the targeted abundances are based on rRNA operon count. The number of rRNA operon counts is not identical for all targeted strains. Hence, even with identical targeted abundances of rRNA operon (*e.g.* for SRR072233), the abundance of strains may differ. It will then induce a difference between relative abundance based on mapping reads on whole genome and the targeted relative abundance based on rRNA operon counts because of rRNA copy variation number.

As taxonomic analyses in *EBI metagenomics* and ASaiM workflows are executed on metagenomic sequences, mapping based abundances are used as expected abundances, instead of relative abundances based on rRNA operon counts from metadata.

2.2 Analyses using *EBI Metagenomics*

Both datasets have been analysed with *EBI metagenomics* pipeline (Version 1.0) (Figure 2). Results are downloaded and formatted to allow comparisons with ASaiM results.

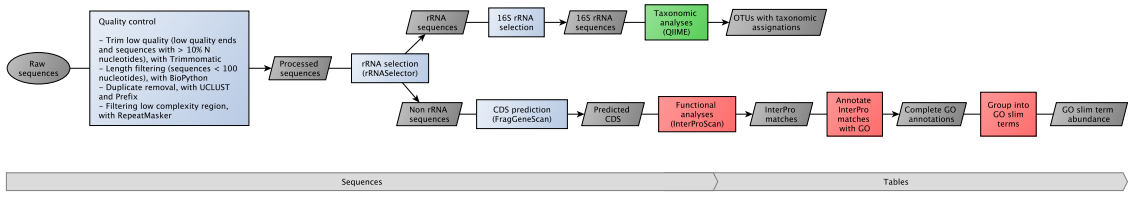

Fig. 2. EBI metagenomics pipeline (version 1.0). The grey boxes correspond to data, the blue boxes to pretreatment steps, the red boxes to functional analysis steps and the green boxes to taxonomic analysis steps.

OTUs with taxonomic assignation are extracted and aggregated to compute relative abundances of each clade at all taxononic levels.

For functional analysis, 3 types of results are generated with *EBI metagenomics* pipeline (Figure 2): matches with InterPro, complete GO annotations and GO slim annotations. Here for comparison purpose (Figure 3), we focus on GO slim annotations. Annotations are formatted to extract relative abundances (in percentage) of GO slim term annotations inside each GO slim term categories (cellular components, biological processes and molecular functions).

2.3 Analyses using ASaiM framework

Main workflow (Figure 3) of ASaiM framework is used to analyze both datasets

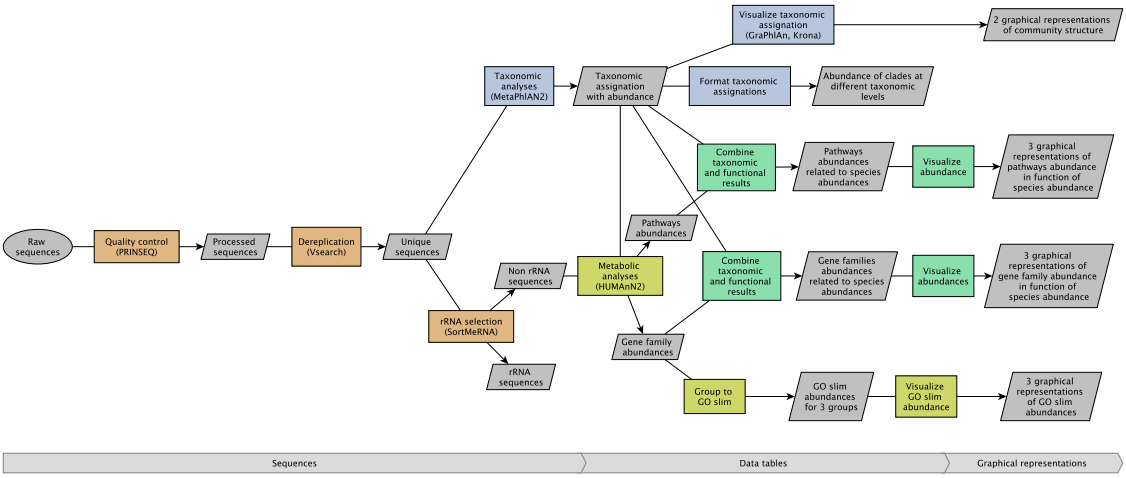

Fig. 3. ASaiM workflow for analysis of raw single-end microbiota sequences. This workflow is available with ASaiM Galaxy instance and used to analyze both datasets. The grey boxes correspond to data, the blue boxes to preprocessing steps, the red boxes to functional analysis steps and the green boxes 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 2.40 GHz and 32 Go of RAM. During workflow execution, size of used memory and execution time are checked (Table 2). Workflow execution is relatively fast: < 5h and < 5h30 for datasets with 1,225,169 and 1,386,198 sequences respectively (Table 2). The most time-consuming step is functional assignation with *HUMAnN2* (Abubucker *et al.*, 2012) which last $\simeq$ 64% of overall time execution (Table 2). 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)

Comparative analysis workflows available with ASaiM framework are used to compare taxonomic and functional results of both datasets.

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 relative abundances computed using mapping of raw reads on reference genomes.

In both *EBI metagenomics* and ASaiM workflows (Figures 2 and 3), 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 workflow depicted in Figure 4.

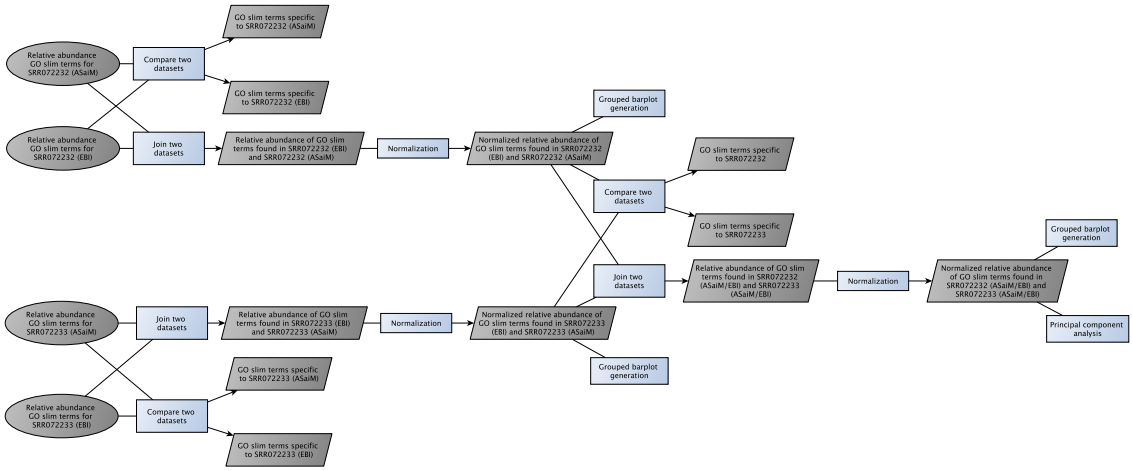

Fig. 4. Workflow to compare GO slim annotation abundances between samples (SRR072232, SRR072233) and workflows (EBI metagenomics, ASaiM). This workflow is available with ASaiM Galaxy instance. The grey boxes correspond to data, the blue boxes to processing steps.

3 Results

3.1 Results of preprocessing in pipelines

In both workflows (Figures 2 and 3), 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 a step to sort rRNA/rDNA sequences from non rRNA/rDNA sequences (Figures 2 and 3). The used tools and parameters for these pretreatments are different between *EBI metagenomics* pipeline (Figure 2) and ASaiM workflow (Figure 3), inducing different pretreatment outputs (Table 3).

Sequence number after quality control and dereplication is different in *EBI metagenomics* and ASaiM workflow results (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 differences in threshold for minimum length. In *EBI metagenomics* pipeline, sequences with less than 100 nucleotides are removed, while in ASaiM the threshold is fixed to 60 nucleotides. 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

	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)

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. Differences are induced by used tools and their underlying algorithms and implementations.

In both datasets and with both workflows, few rDNA sequences are found in datasets (Table 3). These datasets are metagenomic datasets and then focus on whole genome 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. *EBI metagenomics* pipeline (Figure 2) uses *rRNASelector* (Lee *et al.*, 2011) to select rDNA bacterial and archaeal sequences (no eukaryotes sequences). In ASaiM workflow (Figure 3), 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* are older and probably less complete than SILVA (v119) databases used by *SortMeRNA*.

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

3.2 Taxonomic analyses

The mock metagenomic datasets contain sequences of 22 known microbial strains. The expected taxonomy inside the datasets is then known as well as the expected relative abundances (based on mapping on reference genomes, Figure 1). We can then use this information (Figure 5) to analyze ASaiM framework taxonomic results and compare them to *EBI metagenomics* pipeline taxonomic results.

3.2.1 ASaiM taxonomic results

In ASaiM workflow (Figure 3), *MetaPhlAN* (2.0) (Truong *et al.*, 2015; Segata *et al.*, 2012) is used for taxonomic analyses on sequences after preprocessing. *MetaPhlAn* profiles the microbial community structure using a database of unique clade-specific marker genes identified from 17,000 reference genomes. *MetaPhlAn* execution is fast in ASaiM workflow (less than 10 minutes for > 1,100,000 sequences, Table 2).

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 the ASaiM framework, two tools are available: *Krona* (Ondov *et al.*, 2011) for interactive representations of taxonomic assignation and *GraPhlan* for static representations. Original static representations are modified (*e.g.* colors, legend) to help comparison with expected taxonomy (Figure 6).

Despite same expected species, taxonomic diversity in SRR072232 dataset is reduced compared to the one in SRR072233 dataset (Figure 6). Less taxons are found for each taxonomic levels. And, from 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). Phylogenetic markers for these species may be missing in the database used by *MetaPhlAn*. On the other hand, few sequences of these species in SRR072233 are found using mapping on expected species genomes. The phylogenetic signal may be too low to detect these species. Hence, all species with mapping-based abundance smaller than 0.1% are not found using ASaiM for both datasets (Figure 7).

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

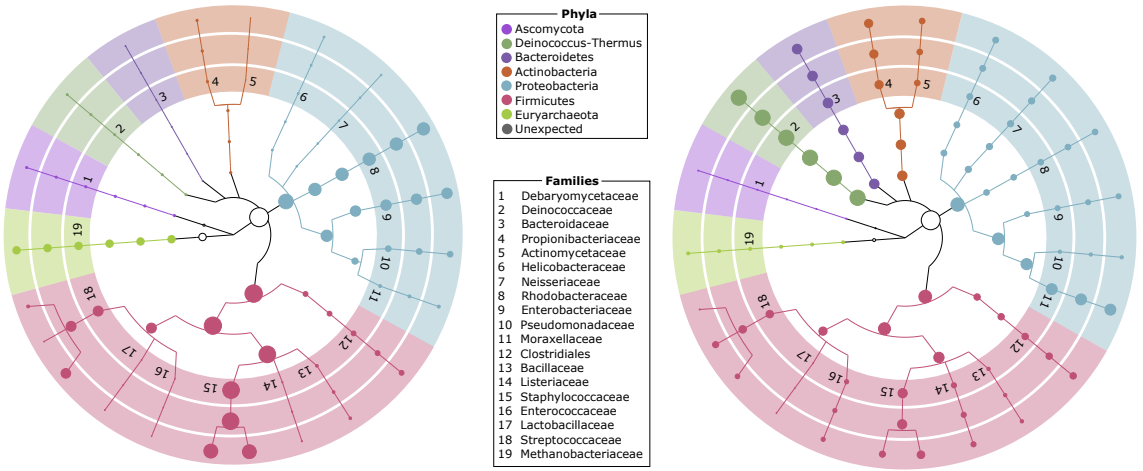

Fig. 5. Expected taxonomy for SRR072232 (left) and SRR072233 (right) from domains to species. Circle diameters at each taxonomic levels are proportional to mapping-based relative abundance of corresponding taxon.

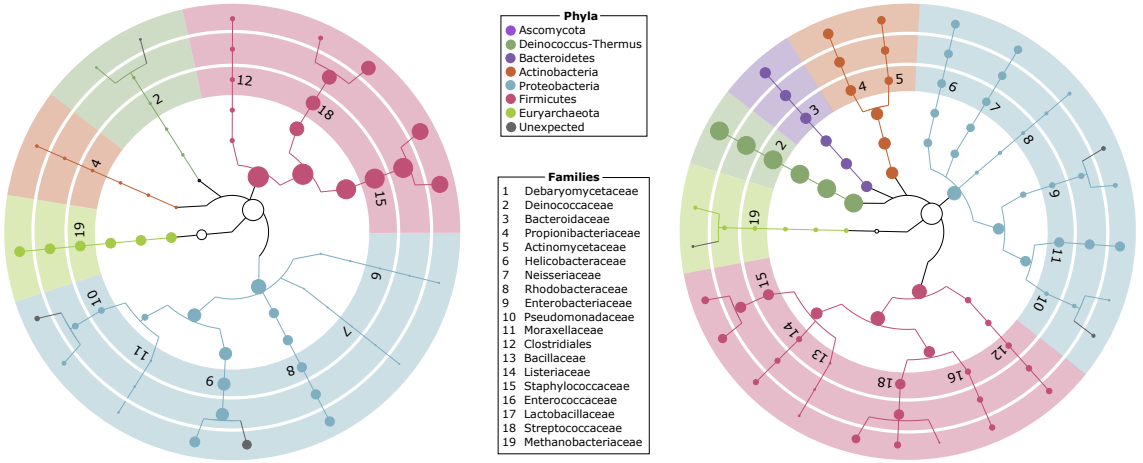

Fig. 6. Taxonomy for SRR072232 (left) and SRR072233 (right) from domains to species, found with ASaiM framework. Circle diameters at each taxonomic levels are proportional to relative abundance of corresponding taxon. Colors and family numbers are the same as the ones used in Figure 5. Gray circles and lines represent unexpected lineages.

may be missing in *MetaPhlAn2* database. As the second species is found with ASaiM in SRR072333, same explanation based on incompletness of reference database.

3.2.2 Comparison of ASaiM taxonomic results with EBI metagenomics taxonomic results

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

In *EBI metagenomics* pipeline (Figure 2), *QIIME* (Caporaso *et al.*, 2010) is used on 16S sequences to identify OTUs and taxonomic assignation for these OTUs. In ASaiM (Figure 3), *MetaPhlAn* is executed on sequences after quality control and dereplication, without

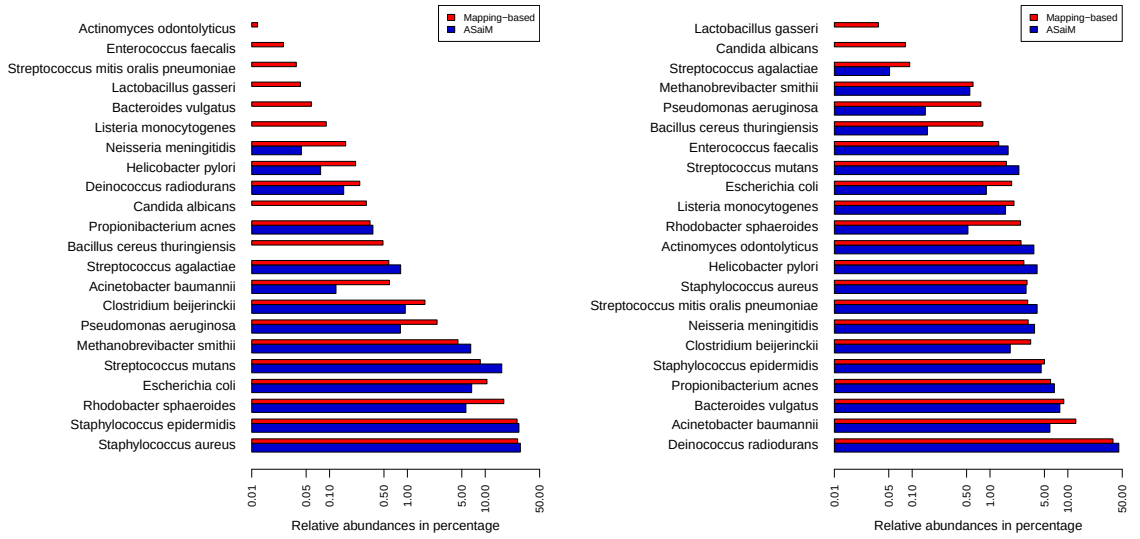

Fig. 7. Relative abundances (percentage in log scale) of expected species for SRR072232 (left) and SRR072232 (right) with comparison between expected abundances (red thin bars) and abundances obtained with ASaiM (blue wide bars)

any sorting step. *MetaPhlAn* searches diverse phylogenetic markers on all sequence types (rDNA, non rDNA, …), not only 16S ones as *QIIME* (Caporaso *et al.*, 2010) does. Among so different sequences, the proportion of sequences with precise and "unique" phylogenetic markers is smaller. With *MetaPhlAn*, the percentage of unassigned reads is $\simeq$ 9 times higher than with *QIIME* (SRR072232: 6.4% with *EBI metagenomics* against 62.61% with ASaiM; SRR072233: 13% with *EBI metagenomics* against 53.93% with ASaiM). On the other hand, the raw number of assigned sequences using ASaiM framework is >52 times higher than the raw number of assigned sequences using *EBI metagenomics* pipeline. Taxonomic assignations within ASaiM framework are then based on more sequences and more statistically supported than taxonomic assignations from *EBI metagenomics* pipeline.

Moreover, taxonomic lineages from *EBI metagenomics* are limited to family level (Figure 8), while they go to species level with ASaiM (Figure 6). *MetaPhlAn* gives taxonomic assignations which are more complete and statistically supported.

With both *EBI metagenomics* pipeline and ASaiM framework, some observed taxonomic assignations are unexpected (Table 4, Figures 6 and 8). For ASaiM framework, 3 species in each sample are identified as "unclassified" (Table 4): they are affiliated to the correct genus but not to correct species. Corresponding sequences may be then incompletely annotated and affiliated. The expected species (*Escherichia* unclassified, *Pseudomonas* unclassified, *Methanobrevibacter* unclassified, *Deinococcus* unclassified) are observed in datasets (Figure 6), but in lower abundance than expected. If sequences corresponding to close unclassified species are correctly affiliated, these species would have observed relative abundances closer to mapping-based abundances.

With *EBI metagenomics*, taxonomic levels of unexpected clades are higher (class, order and family) than taxonomic level of unexpected clades in ASaiM framework (species, Table 4, Figure 8). Taxonomic assignations with *MetaPhlAN* are then more accurate and precise. As the most precise taxomic level for *EBI metagenomics* is family (Figure 8), further comparisons focus on this level (Figure 9).

Similarly to previous observations on raw ASaiM results, species with mapping-based abundance smaller than 0.1% are found neither with ASaiM nor with *EBI metagenomics* (Figure 9). Nonetheless, the detection threshold seems slightly smaller for *EBI metagenomics*: for SRR072232, Listeriaceae family is detected with *EBI metagenomics* and not with ASaiM (Figure 9). On the other hand, Bacillaceae and Debaryomycetaceae families are not found with *EBI metagenomics* for both datasets (Figure 9), despite mapping-based abundance higher than 0.1%. Used databases may be then incomplete regarding some phylogenetic markers, particularly the ones corresponding to missing families.

Variations in observed abundances for *EBI metagenomics* or ASaiM correspond to variations in mapping-based abundances (Figure 7): small observed abundances for small mapping-based abundances and high observed abundances for high mapping-based abundances.

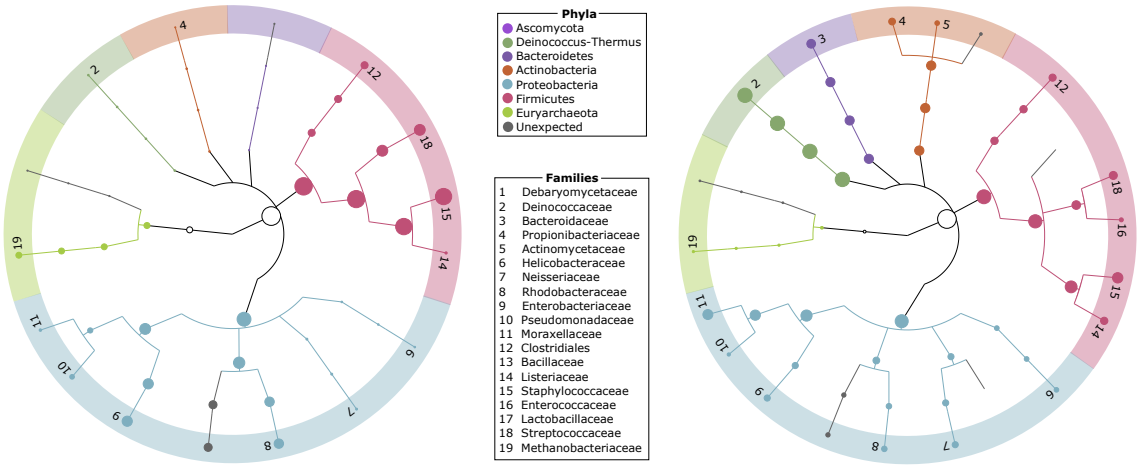

Fig. 8. Taxonomy for SRR072232 (left) and SRR072233 (right) from domains to families, found with EBI metagenomics pipeline. Circle diameters at each taxonomic levels are proportional to relative abundance of corresponding taxon. Colors and family numbers are the same as the ones used in Figure 5. Gray circles and lines represent unexpected lineages.

		SRR072232		SRR072233	
Taxonomic level	Clade	EBI	ASaiM	EBI	ASaiM
Class					
	Methanopyri	0.09%		0.21%	
Order					
	Rickettsiales	5.71%		1.43%	
	Methanopyrales	0.09%		0.21%	
Family					
	Rickettsiales mitochondria	5.71%		1.43%	
	Methanopyraceae	0.09%		0.21%	
	Paraprevotellaceae			0.09%	
	Cryptosporangiaceae			0.5%	
Genus		*No information*		*No information*	
Species		*No information*		*No information*	
	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 unexpected clades at different taxonomic levels in taxonomic results of EBI metagenomics and ASaiM framework for both samples (SRR072233 and SRR072233)

For a broader comparison, a principal coordinate analysis (PCoA) is computed, for each sample, on Bray-Curtis distances computed on relative abundances of families (Figure 10).

First axis of these analyses explains most of data variability (98% for both datasets, Figure 10). Mapping-based, *EBI metagenomics* and ASaiM abundances are not discriminated on this first axis (Figure 10), 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,

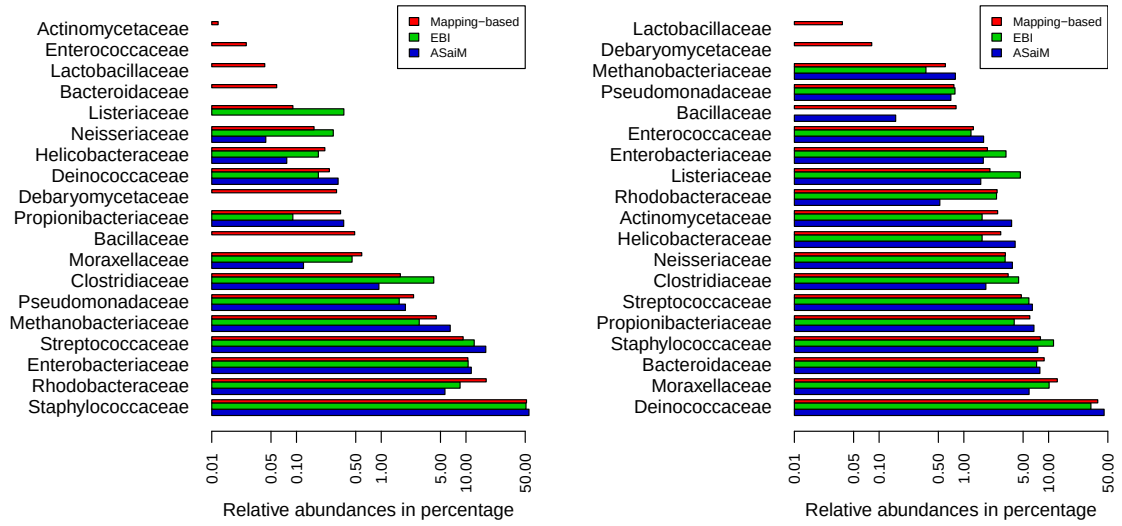

Fig. 9. Relative abundances (percentage, log scale) of expected families for SRR072232 (left) and SRR072233 (right) with comparison between mapping-based relative abundances (red thin bars), abundances obtained with EBI metagenomics (green wide bars) and abundances obtained with ASaiM (blue wide bars).

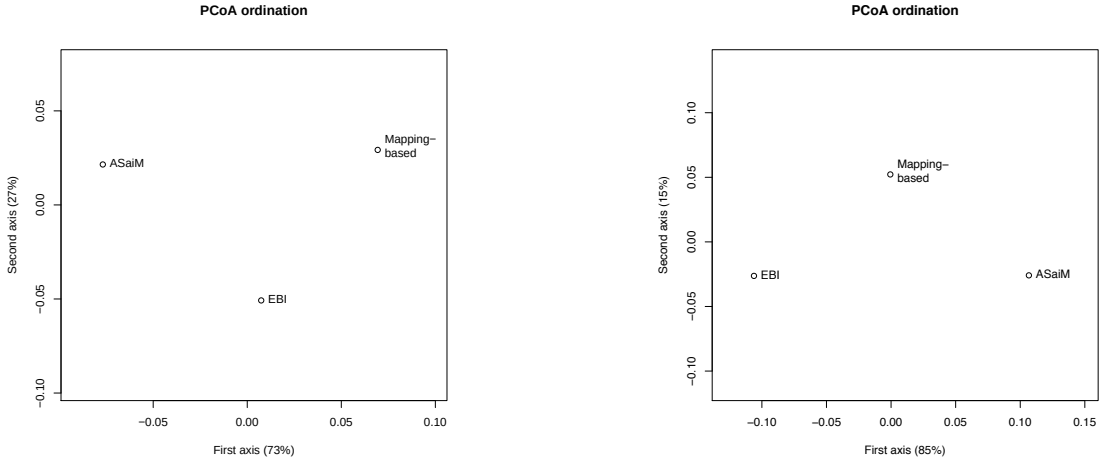

Fig. 10. Scatter diagram of principal coordinate analysis of the Bray-Curtis distances computed on relative abundances of families for SRR072232 (left) and SRR072233 (right).

similar abundances for observed families are obtained for *EBI metagenomics* and ASaiM and these abundances are close to abundances computed using mapping on expected species genomes.

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

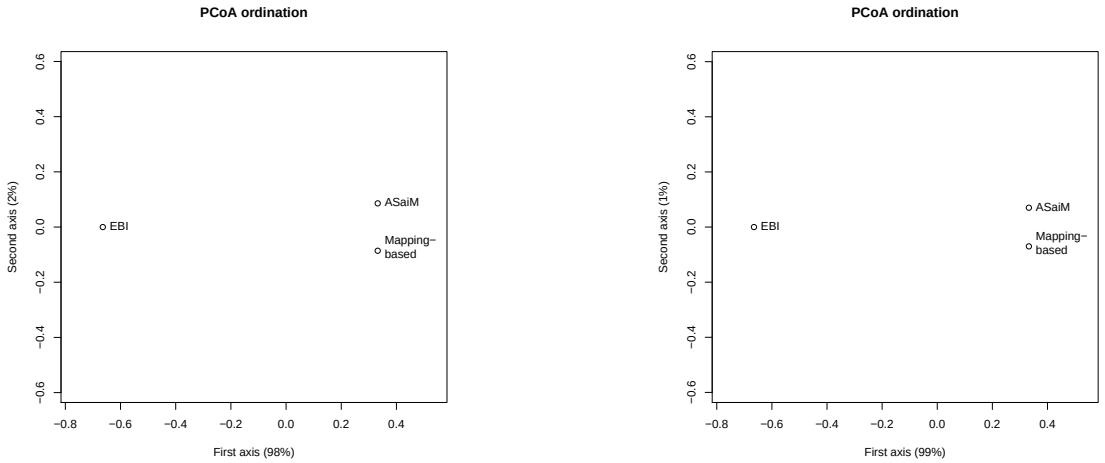

Fig. 11. Scatter diagram of principal coordinate analysis of the Bray-Curtis distances computed on relative abundances of species for SRR072232 (left) and SRR072233 (right).

3.3 Functional analyses

3.3.1 Raw ASaiM results

In ASaiM framework (Figure 3), *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 5) are reconstructed from > 1,100,000 non rDNA sequences (Table 3).

		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 5. 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.

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

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 5). But different gene families have a limited impact on overall metabolism (< 50% of relative abundance, Table 5). They may correspond to gene families which are difficult to detect because they are not abundant or they are "done" by organisms in low abundances. Global metabolism functions such as pathways are similar in both datasets (> 95%

of similar pathways representing > 99.5% of overall abundance, Table 5). Hence, the unexpected observed differences are limited and may be due to bias induced by low abundances and detection threshold.

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

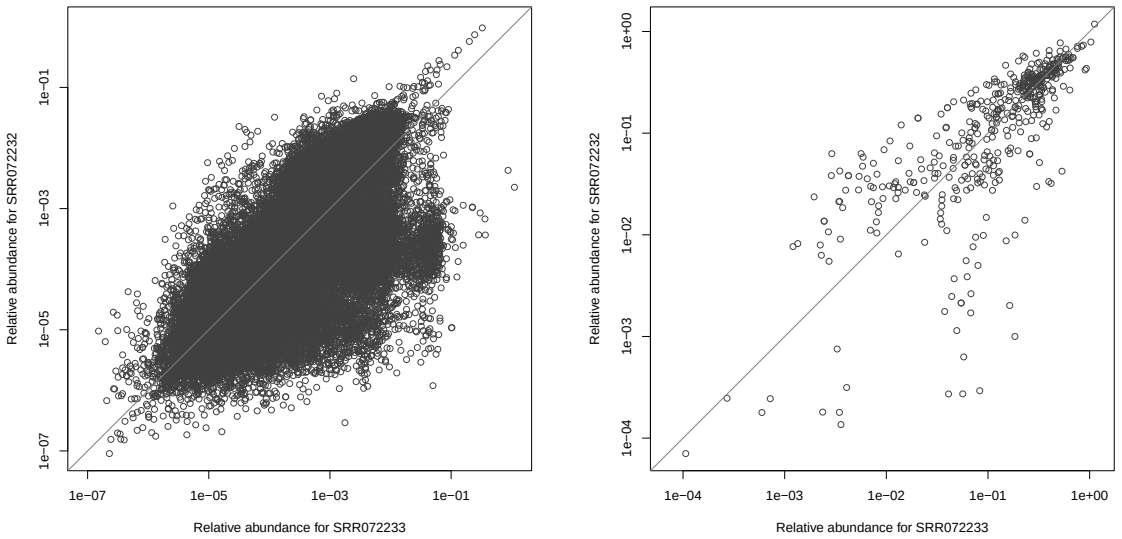

Fig. 12. Normalized relative abundances (%) for similar UniRef50 gene families (left graphics) and MetaCyc pathways (right graphics) for both samples (SRR072233 and SRR072233). The relative abundances of each similar characteristics (gene families or pathways) is computed with HUMAnN2 and normalized 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 13). For both datasets, we observe similar profiles of GO slim terms (Figure 13).

Both communities, with same expected strains but in different abundances, have different metabolic profiles. As expected, the difference are not metabolic functions, but more on abundances of these functions.

3.3.2 Comparison of *EBI metagenomics* and ASaiM results

In ASaiM workflow (Figure 3), UniRef50 gene families and their abundances are computed with *HUMAnN2*. In *EBI metagenomics* pipeline (Figure 2), functional analyses are based on InterPro and its identifiants. These functional results are then not 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.

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). Pprincipal component analyses (PCA) are then computed 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 high proportion of data variability (between 64% and 87%, Table 6). On left part of scatter representation (Figures 14 and 15), GO slim terms are highly abundant with 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).

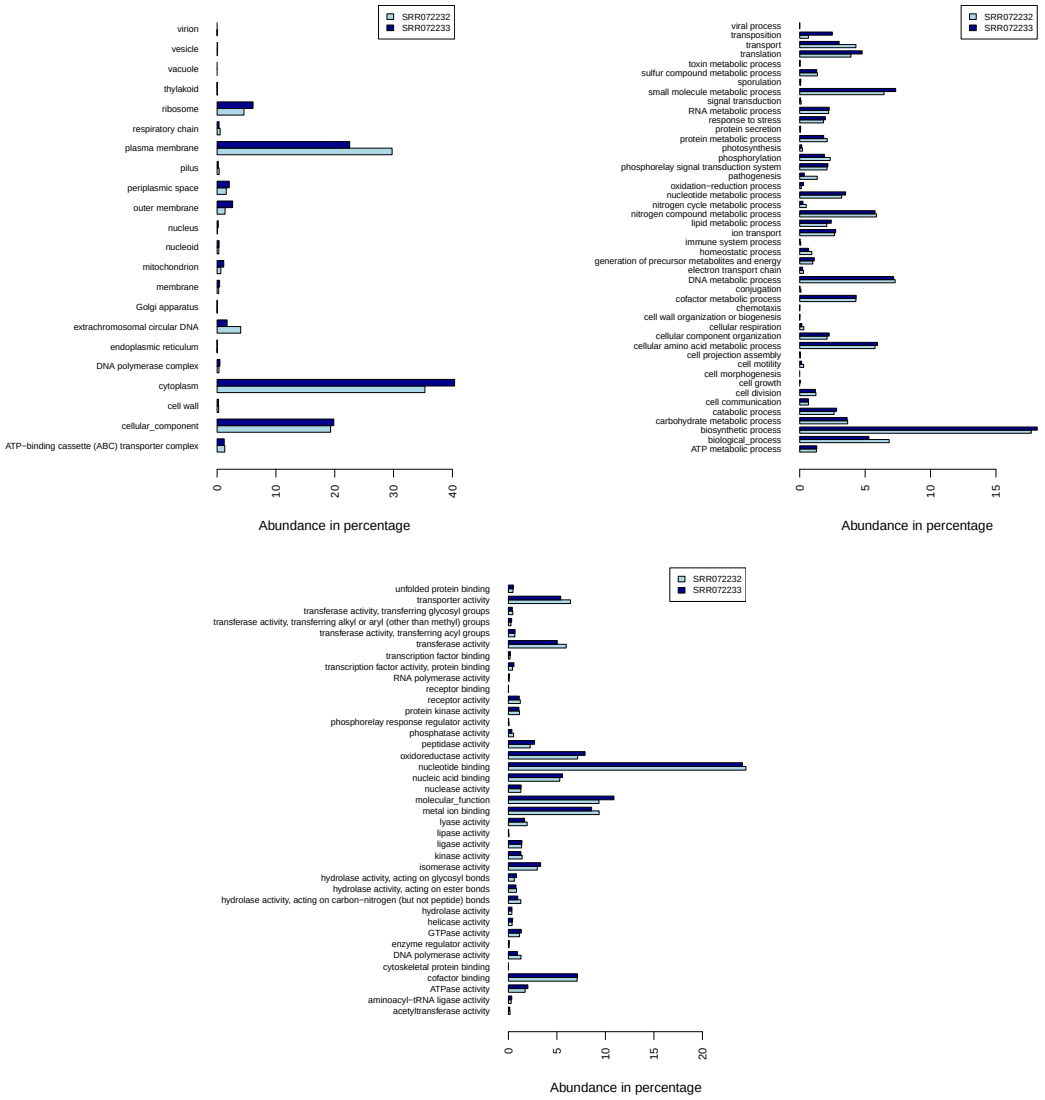

Fig. 13. Relative abundances of GO slim terms in SRR072232 and SRR072233 for cellular components (top left), biological processes (top right) and molecular function (bottom)

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

Table 6. 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 6), explaining between 13% and 35% of overall data variability (Table 6). Some GO slim terms such as membrane, hydrolase activity or nitrogen compound metabolic

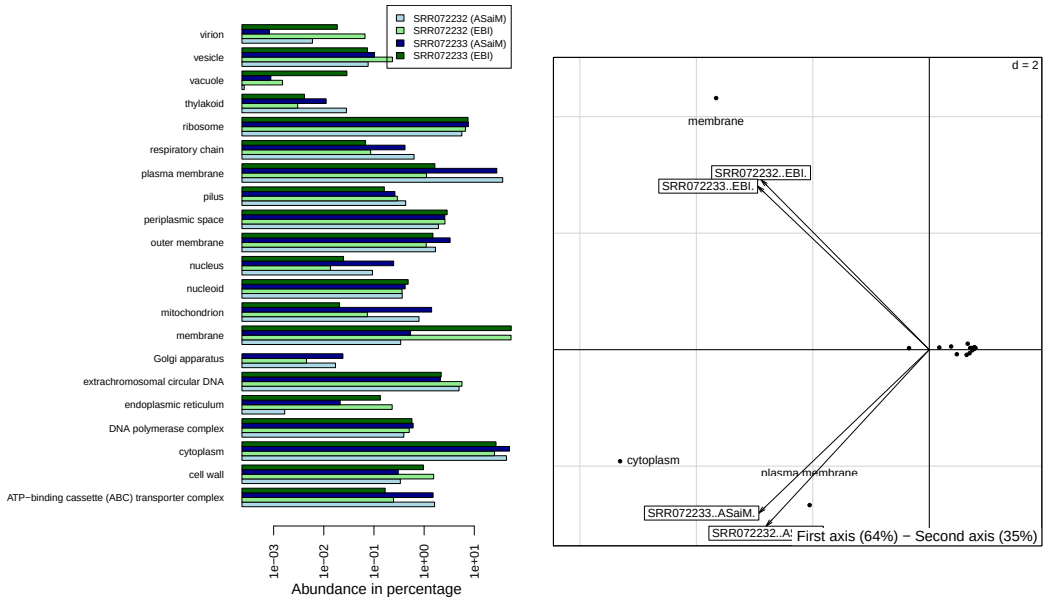

Fig. 14. Barplot representation (in left, logarithm scale) and scatter diagram of principal component analysis of the normalized relative abundances (in percentage) of the cellular component GO slim terms for both samples (SRR072233 and SRR072233) and both workflows (EBI metagenomics and ASaiM). The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

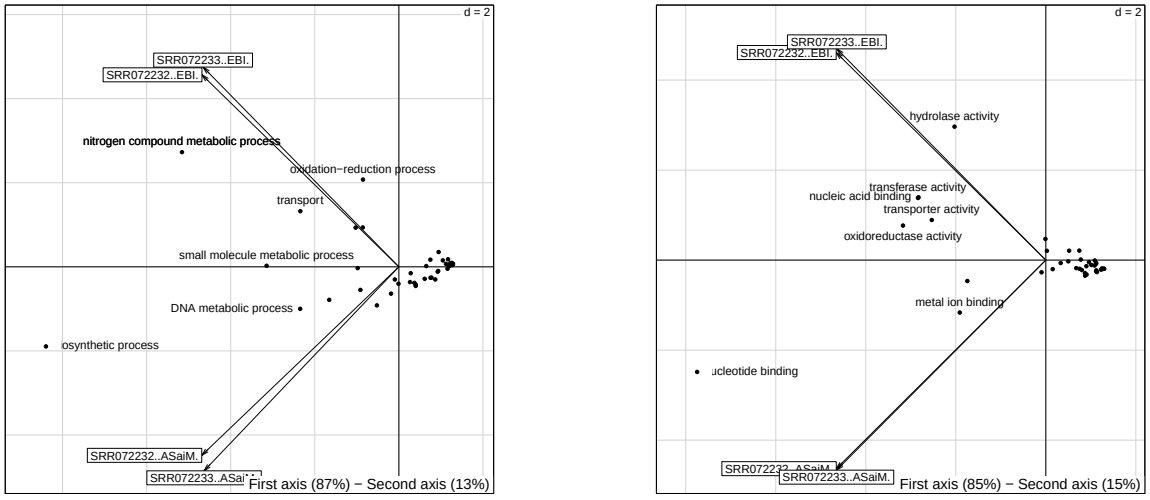

Fig. 15. Scatter diagram of principal component analysis of the normalized relative abundances (in percentage) of the biological process (in left) and of the molecular functions (in right) GO slim terms for both samples (SRR072233 and SRR072233) and both workflows (EBI metagenomics and ASaiM). The relative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

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.

Similar function results are obtained for *EBI metagenomics* and ASaiM workflows when we look at GO slim terms abundance: the discrimination between both workflow results appears only as a secondary explanation for variability in GO slim term abundances.

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 results and answer questions such as "Which species contribute to which metabolic functions? And, in which proportion?". < 35% of gene families (> 90% of relative abundance) and > 80% pathways (> 50% of relative abundance) can be then related to the community structure (species and their abundance, Table 7).

		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 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.

For both samples, a significant correlation is observed between CDS number in species (data from GenBank) and number of gene families found for these species (Table 8). 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 great.

		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 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). 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 8). Sequences of an abundant species in a community are supposed to be abundant in metagenomic sequences of the community. This relation holds for 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 8) 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 8).

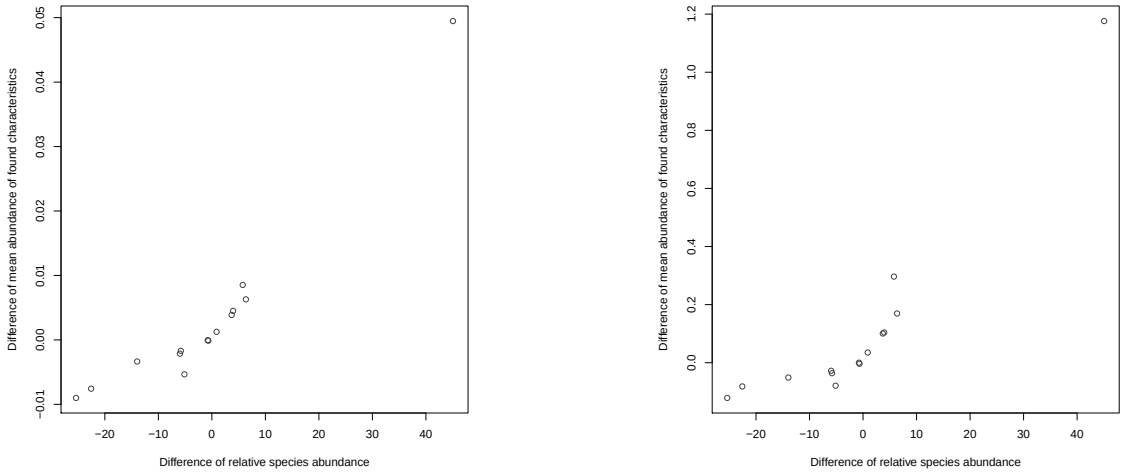

Fig. 16. Difference in mean abundances for gene families (left) and pathways (right) in function of difference of related species abundance between both samples. Correlation coefficients and p-values are detailed in Table 8

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 with complete, accurate and statistically supported information. With *HUMAnN2* 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.