Abundance-based reconstitution of microbial pan-genomes from whole-metagenome shotgun sequencing data

Florian Plaza Oñate¹,², Alessandra C. L. Cervino¹, Frédéric Magoulès³, S. Dusko Ehrlich² & Matthieu Pichaud¹

¹ Enterome, 94-96 Avenue Ledru Rollin, 75011 Paris, France
² MGP MetaGénoPolis, INRA, Université Paris-Saclay, 78350 Jouy en Josas, France
³ CentraleSupélec, Université Paris-Saclay, 92290 Châtenay-Malabry, France

§ Corresponding author

Email addresses:
FPO: fplaza-onate [at] enterome [dot] com
MP: matthieu.pichaud [at] gmail [dot] com
Abstract

Analysis toolkits for whole-metagenome shotgun sequencing data achieved strain-level characterization of complex microbial communities by capturing intra-species gene content variation. Yet, these tools are hampered by the extent of reference genomes that are far from covering all microbial variability, as many species are still not sequenced or have only few strains available.

Binning co-abundant genes obtained from de novo assembly is a powerful reference-free technique for discovering and reconstituting gene repertoire of microbial species. While current methods accurately identify species core genes, they miss many accessory genes or split them in small separated clusters.

We introduce MSPminer, a computationally efficient software tool that reconstitutes Metagenomic Species Pan-genomes (MSPs) by binning co-abundant genes across large-scale metagenomic datasets. MSPminer relies on a new robust measure for grouping not only species core genes but accessory genes also. In MSPs, an empirical classifier distinguishes core from accessory and shared genes.

We applied MSPminer to the largest publicly available gene abundance table which is composed of 9.9M genes quantified in 1,267 stool samples. We show that MSPminer successfully reconstitutes in a matter of several hours gene repertoire of >1600 microbial species (some hitherto unknown) and detects many more accessory genes than existing tools. By compiling the information from thousands of samples, species gene content variability is better accounted for and their quantification is subsequently more precise.
Introduction

Metagenomics has revolutionized microbiology by allowing culture-independent characterization of microbial communities. Its advent has allowed an unprecedented genetic characterization of the human gut microbiota and emphasized its fundamental role in health and disease [1–7]. Shotgun metagenomics where whole-community DNA is randomly sequenced bypasses the biases and limitations of 16S rRNA sequencing [8,9] by providing high resolution taxonomic profiling as well as insights into the diverse physiological roles and the metabolic potential of the community [10,11].

The analysis of large cohorts revealed a substantial inter-individual microbial gene content variability [12] and nucleotide polymorphism [13] which reflects that individuals are not only carriers of various species, but also of different strains of the same species [14,15]. The characterization of the accessory genes found in individual strains is crucial in many contexts as they can provide functional advantages such as complex carbohydrates metabolism [16], antibiotic resistance or pathogenicity [17,18].

Recent analysis toolkits for shotgun metagenomics data achieved strain-level resolution when coverage is sufficient. To this end, they either capture intra-species single-nucleotide polymorphisms (SNPs) in pre-identified marker genes [19,20], gene content variation [21] or both [22]. However, these tools are hampered by the extent of the reference genomes.

Indeed, the microbial variability extends far beyond the content of reference genomes making metagenomic samples an untapped reservoir of information. First, it has been estimated that on average 50% of the species present in the human gut microbiota of western individuals lack reference genome and this proportion rises to 85% in individuals with traditional lifestyles [22]. Even if recent advancements of culture-based methods have proven that a substantial proportion of these species are actually culturable [23,24], the number of unknown species is probably still important. In addition, these techniques remain laborious and time consuming. Second, although species of public health interest (e.g. Escherichia coli, Salmonella enterica or Clostridium difficile) are represented by hundreds or even thousands of strains in genome databases [25], only few strains are available for the great majority of commensal species.
Consequently, accessory genes associated with microbial phenotypic traits may be missing in gene repertoires constructed from reference genomes.

Metagenomic assembly where overlapping reads are merged into longer sequences called contigs, is a powerful reference-free technique for overcoming the limitations of reference-based methods. Indeed, it allows the genetic survey of non-sequenced species and strains with previously unknown accessory genes. However, assembly remains a computationally challenging task [26] and despite the many dedicated tools proposed [27–30] the process only recovers incomplete genomes scattered in multiple contigs. To retrieve an exhaustive set of references, metagenomic assembly is performed on multiple samples. Then, non-redundant reference gene catalogs [31] are created and used as proxy for disease-related analyses [3,6] or descriptive purposes [12,32].

Metagenome-wide association studies have successfully identified associations between the gut microbiome and disease based on universal marker genes [33], clade specific marker genes [34] or whole gene sets [35]. The latter is the only method that allows an extensive investigation of microorganism core, accessory genes and mobile elements. However, testing millions of genes is biased and lacks statistical power. Testing all the variables is not adapted to situations where some of them are highly correlated [36], which is expected for genes coming from the same biological entity. The resulting list of significant genes will be biased towards organisms with the most genes in the pool as they have more chances of being picked up.

Considering that the genes originating from the same biological entity should have proportional abundances across samples, binning co-abundant genes has been proposed. However, clustering millions of genes is a very computationally intensive task and pairwise comparison of all gene abundance profiles is not feasible. To reduce the number of comparisons, some authors have performed binning on the subset of genes that were statistically significant [3,4,6] which restricts the analysis to the genes that are significant by themselves and does not improve the statistical power of the analysis. Others have proposed methods to perform the clustering of complete gene references based either on the Markov cluster algorithm [37] or a variant of the Canopy clustering algorithm [38].
Although direct proportionality is expected between co-abundant genes, these methods rely either on Pearson’s or Spearman’s correlation coefficients which respectively assess a linear association with a potentially non-null intercept or any monotonic association. Thus, these coefficients are too loose and spurious associations can be discovered. In addition, they are biased by sparse genes with many null counts [39], non-normal gene counts distributions [40] and presence of outliers [41].

Current clustering strategies group species core genes and highly prevalent accessory genes into the same cluster, but miss accessory genes of medium and low prevalence or assign them to small separate clusters. Dependency between clusters of essential genes and accessory clusters can be evaluated downstream using the Fisher’s exact test [38], which compares their presence/absence patterns across samples. Yet, this strategy does not account for the co-abundance of genes and is poorly discriminative when considering accessory clusters that are rare or associated with very prevalent species. In addition, it is not suitable for detecting clusters shared between several species.

To overcome these limitations, we have developed MSPminer, the first tool that discovers, delineates and structures Metagenomic Species Pan-genomes (MSPs) from large-scale shotgun metagenomics datasets without referring to genomes from isolated strains. MSPminer presents several significant improvements over existing methods. First, it relies on a new robust measure of proportionality for detection of co-abundant but not necessarily co-occurring genes as expected for non-core genes. Second, genes grouped in a MSP are empirically classified as core, accessory and shared genes.

To illustrate its usefulness, we applied MSPminer to the largest publicly available gene abundance table which is composed of 9.9M genes quantified in 1267 samples [12]. We show that MSPminer successfully groups genes from the same species and identifies additional genes. Gene variability of microbial species is better captured and their quantification is subsequently more precise. MSPminer is a computationally efficient multithreaded program implemented in C++ that can process large datasets with millions of genes and thousands of samples in just a few hours on a single node server.
Results

New measures of proportionality

The gene repertoire of microbial species is composed of core genes present in all strains and accessory genes present in only some of them [42]. In a shotgun metagenomic sequencing context, we assumed that core genes of a microbial species should have a directly proportional number of mapped reads across samples (co-abundance) and should be consistently observed in samples if sequencing depth allows (co-occurrence). Remarkably, core genes and an accessory gene should have directly proportional counts only in the subset of samples where they are both detected (Figure 1). To group the core genes of a species and then identify its accessory genes, we developed a measure that evaluates proportionality between gene counts using samples where the number of mapped reads is high enough (see Methods).

To evaluate this new measure of proportionality, we generated an abundance table that simulates the counts of genes from a single virtual species across 300 samples (see Methods). We considered that each sample was carrying a unique strain of the species with specific gene content. Genes present in all the samples were labeled as core and those detected in a subset as accessory. We used this dataset to compare the performance of the Pearson correlation coefficient, the Spearman correlation coefficient and the proposed measure of proportionality for detecting a relation between the abundance profile of the species core genome and all its genes including accessories (Figure 2). Pearson and Spearman correlation coefficients decrease all the more as the prevalence of a tested gene decreases while the proposed measure remains high, as only samples where both the species core and its accessory gene are detected are used for calculation. Therefore, the association between core genes and many accessory genes will be missed using the correlation coefficients. However, accessory genes observed in similar subsets of samples may be grouped into small distinct clusters as their abundance profiles should be correlated.

Finally, we derived a robust version of the measure to identify associated genes despite the presence of samples with inconsistent counts named hereafter outliers. We evaluated this robust measure of proportionality against the non-robust version described above by adding an increasing percentage of outliers to the genes abundance profiles. For a given percentage
of outliers, each of these genes was compared to the outlier-free abundance profile of the species core genome. This simulation showed that the non-robust measure of proportionality decreases all the more as the percentage of outliers increases whereas the robust measure remains high; demonstrating that proportionality is still detected (Supplementary Figure 1). However, the robust measure decreases significantly when the percentage of outliers is high and the gene prevalence is low.

Reconstitution of Metagenomic Species Pan-genomes of the human gut microbiota

We developed MSPminer, a program that uses measures of proportionality to group co-abundant genes into Metagenomic Species Pan-genomes (MSPs, Figure 3). MSPminer empirically distinguishes core from accessory genes based on their presence absence patterns (see Methods) and tags genes observed in samples where the core is not detected as shared (Figure 4). Finally, non-core genes observed in the same subset of samples are grouped into modules of co-occurring genes.

We applied MSPminer to the largest publicly available gene abundance table provided with the Integrated Gene Catalog of the human gut microbiome [12]. In this table, 9 879 896 genes are quantified across 1 267 stool samples from individuals of various geographical origin (Europe, USA and China) and diverse health status (healthy, obese, diabetic, with inflammatory bowel disease etc.). 6 971 229 (70.6%) genes with counts greater than 6 in at least 3 samples were kept. Among these, 3 262 914 (46.8%) were organized into 1 677 Metagenomic Species Pan-Genomes (MSPs) with at least 200 core genes (Supplementary Table 1).

Taxonomy

By considering the lowest taxonomic rank assigned to MSPs, 278 (16.6%) were annotated at species level, 85 (5%) at genus level, 119 (7.1%) at phylum level and the remaining 1 213 MSPs (71.5%) could not be annotated (Supplementary Figure 2) indicating that a clear majority of MSPs correspond to species not represented in reference genomes databases. Only 4 MSPs were annotated as Eukaryotes and corresponded to intestinal parasites of the Blastocystis genus. All others MSPs were assigned to prokaryotic species of which only 3 were Archaea and corresponded to the species Methanobrevibacter smithii, Candidatus Methanomethyphilus
alvus and Methanosphaera stadtmanae. Among the MSPs annotated as Bacteria, the phyla Firmicutes (268 MSPs), Bacteroidetes (118 MSPs), Proteobacteria (50 MSPs) and Actinobacteria (29 MSPs) were the most represented (Supplementary Figure 3) as expected from human gut metagenomes [43]. Some species were represented by multiple MSPs such as Faecalibacterium prausnitzii (5 MSPs), Bacteroides fragilis (2 MSPs), Methanobrevibacter smithii (2 MSPs) or Hungatella hathewayi (2 MSPs) suggesting a high nucleotide and gene content variability between strains attributed so far to the same species. Conversely, some MSPs had their core genes attributed to multiple species, usually from the same genus (Supplementary Table 3). Although some inconsistencies in the taxonomy assignment have previously been reported [22,44], it is possible that several MSPs regroup genes from highly related species with high average nucleotide identity and similar gene content. MSPs annotated at species level had a consistent taxonomical annotation as 97% of the core genes (median) were assigned to the same species (Supplementary Figure 4). Consistency of the taxonomic annotation was lower for accessory genes (62%) caused mainly by unannotated genes.

MSP content

Most MSPs were small (median number of genes = 1 784) even if 53 had more than 5 000 genes (Supplementary Figure 5). As expected, a strong positive correlation (Pearson’s r = 0.8) between the total number of genes in a MSP and its number of accessory genes was observed (Supplementary Figure 6). Interestingly, four outliers corresponding to the intestinal parasites previously described had a high number of core genes and few accessory genes. This suggests that Eukaryotic genomes have a larger number of genes and a lower gene content variability than Prokaryotes. Among the MSPs with the most accessory genes (Supplementary Table 2), many corresponded to species reported as highly variable such as Escherichia coli [45], Klebsiella pneumoniae [46] or Clostridium bolteae [47]. As previously observed in population genomics studies comparing multiple strains of the same species [48,49], the prevalence of accessory genes in MSPs often follows a bimodal distribution (Supplementary Figure 7) showing either a high or low prevalence but rarely intermediate. Thus, the number of accessory genes in a MSP is correlated (Spearman’s rho = 0.86) with its prevalence (Supplementary Figure 8). Indeed, the more a MSP is detected in many samples, the more exhaustively MSPminerv will recover its accessory genes, especially the rare ones. Many MSPs
annotated at species level had accessory genes previously unobserved in available genomes

(Supplementary Table 2)

**Prevalence**

As for the genes in the catalog, most MSPs were detected in very few samples (Supplementary Figure 9). Only 40 MSPs were detected in at least 70% of the samples showing that the common microbial core of the human gut microbiota is limited to a few dozen species (Supplementary Table 2). No clear relation between the prevalence of the MSPs and their mean abundance was found (Supplementary Figure 10). However, 2 MSPs corresponding to *Bacteroides vulgatus* and *Bacteroides uniformis* were both very prevalent (detected in 97.6% and 94.6% of the samples respectively) and very abundant (mean relative abundance of 7.8% and 4.4% respectively). Interestingly, many rare MSPs were abundant in the few samples which carried them. Many of these MSPs were annotated as bacteria of the *Lactobacillus* genus most likely consumed as probiotics. However, some others correspond to known invasive species associated with severe dysbiosis such as *Fusobacterium nucleatum* [50] or *Clostridium clostridioforme* [51].

**Census of universal single copy marker genes**

To check that MSPs correspond to real microbial species and evaluate the completeness of their set of core genes, we identified in each of them 40 universal single copy marker genes (SCM) [44]. 878 MSPs (54%) had at least 30 SCM and 403 (24%) had all of them (Supplementary Figure 11 A and Supplementary Table 2). As housekeeping genes, SCMs are essential to the microbe survival and should be found among core genes. Indeed, 92% of the SCMs were core genes in their respective MSP and the rest was mainly high prevalent accessory genes (Supplementary Figure 11 B). This shows the classification of genes as core or accessory performed by MSPminer is reliable.

**Comparison to sequenced genomes**

We compared the MSPs to 642 sequenced genomes for which at least 10% of their constituent genes were detected in the Integrated Genes Catalog of the human gut microbiome [12] (Supplementary Table 4). In total, these genomes covered 398 species representing 114 different genera. 624 (97.1%) were unambiguously assigned to 281 different MSPs. By keeping only one representative per species, 47.1% (resp. 60.4%) of the genomes had at least 75% of
their genes grouped in their corresponding MSP considering either all their genes or only those that were in the catalog (Supplementary Figure 12). In compliance with the results of the taxonomic analysis, highly related species were assigned to the same MSP such as *Escherichia coli*, *Escherichia fergusonii* and all genomes of the *Shigella* genus. Conversely, some species grouping highly divergent strains were represented by several MSPs including *Faecalibacterium prausnitzii* or *Bacteroides fragilis*.

To give another perspective on the MSPs, we compared the complete genome of *Parabacteroides distasonis* ATCC 8503 [52] to its corresponding MSP (Figure 7). Among the 850 genes predicted in the genome, 3 781 (98%) had at least a close homolog in the Integrated Gene Catalogue and 3 442 (89%) were found in the msp_0011. As expected, almost all the core genes from the MSP were found in the genome (1 867 / 1 921, 97%), as well as accessory genes with a prevalence higher than 80% (522 / 599, 87%). Only a small fraction of less prevalent accessory genes was found in the genome (1371 / 5 090, 27%) (Supplementary Figure 13). Genes grouped in the same modules tended to be physically close, in coherence with genome organization of prokaryotes (Supplementary Figure 14). Remarkably, some singleton genes were surrounded by genes from the same module, which shows that the stringent grouping criteria used by MSPminer may split genes with slightly different presence/absence patterns into different modules while they could have been grouped (Supplementary Figure 15). Finally, few genomic regions contained genes that were not assigned to the MSP (Supplementary Table 5A). Interestingly, some of these regions were annotated as mobile elements (Supplementary Table 5B). Although some could be false negatives, many were appropriately excluded as they were observed in too few samples or their counts did not meet the proportionality-based grouping criterion (Supplementary Figure 16).

**Comparison to the Canopy clustering algorithm**

The Canopy clustering algorithm [38] was compared to MSPminer by applying both tools to the metagenomic dataset described above. In total, Canopy grouped 2 691 408 genes into 3 463 Co-Abundance gene Groups (CAGs) while MSPminer grouped 3 267 132 genes (+17.6%) into 1 677 MSPs (~ two-fold less objects).
178 MSPs encompassing 154,617 genes had no equivalent among the CAGs. Most MSPs were rare as 75% were detected in fewer than 5 samples but had a significant size as 50% were composed of at least 700 genes (Supplementary Figure 17). Remarkably, for 75% of them, the 3 samples with the highest counts represented at least 90% of the sum of MSP abundance on all the samples. By default, Canopy discards such cases to avoid detection of spurious correlations but MSPminer limits this risk by applying a variance-stabilizing transformation and a stringent association criterion. In addition, Canopy grouped most core genes of a MSP into a single CAG while many accessory genes were missed or assigned to small separate CAGs (Figure 5). In agreement with the results of the simulation, most of the missed accessory genes had a medium or low prevalence. As they contained many unexpected zeros, the correlations with the core of their respective species were below the limit set in Canopy (Figure 6).

**MSPs quantification for biomarkers discovery**

To demonstrate that MSPminer was useful for biomarkers discovery, we first looked for differentially abundant MSPs according to the geographical origin of samples. We discovered 94 MSPs differentially abundant between Westerners and Chinese (q-value < $10^{-3}$, log2 fold change ≥ 1) including 72 more abundant in Westerners and 22 in Chinese (Supplementary Figure 18 and Supplementary Table 6). Among the discriminant MSPs, all those assigned to the *Proteobacteria* phylum (*Klebsiella pneumoniae*, *Escherichia coli* and *Bilophila wadsworthia*) were more abundant in Chinese which is consistent with previously published results [12]. Interestingly, two MSPs assigned to *Faecalibacterium prausnitzii* were significant but one was more abundant in Westerners and the other in Chinese. This shows that some strains of this species are associated with geographical origin of samples. In addition, we discovered 75 MSPs differentially abundant between Europeans and Americans (q-value < $10^{-3}$, log2 fold change ≥ 1) of which 70 were more abundant among Europeans (Supplementary Figure 19 and Supplementary Table 6). This result is consistent with previous studies showing lower gut microbiota diversity among Americans compared to Europeans [33].

Secondly, we checked if MSPs could be used to perform strain-level analysis. To do this, we tested if some accessory genes in the MSPs were more prevalent in samples of a given geographical origin. By way of example, we found 680 accessory genes associated with
geographical origin (chi-squared test, p-value < 10^{-10}) in the msp_0011 corresponding to *Parabacteroides distasonis* (Supplementary Figure 20 and Supplementary Table 7). Remarkably, genes involved in cell filamentation (V1.UC58-4_GL0042624 and V1.UC18-0_GL0014340) were more prevalent in Chinese than Westerners. More generally, many significant genes were more prevalent among Westerners or Chinese although European and American samples were separated in the analysis. This result suggests that strains of *Parabacteroides distasonis* carried by Chinese are distinct from those of Westerners.

**Discussion**

**Direct proportionality hypothesis and limits**

MSPminer relies on a new robust measure to detect genes with directly proportional counts. Even if this relation is more stringent than those assessed by Pearson’s or Spearman’s correlation coefficients, it was successfully used to reconstitute Metagenomic Species Pan-genomes of the human gut microbiota. In fact, most genes from sequenced genomes were grouped into a single MSP showing that direct proportionality is the most common relation between genes from the same biological entity. MSPminer misses some genes for which counts are not ruled by this relation. Indeed, proportionality is disrupted when gene copy number varies across samples [14] (Supplementary Figure 21), when a sample contains multiples strains [19,20] and when a gene is shared between several MSPs (Supplementary Figure 22) because of horizontal gene transfer [53] or grouping of highly similar orthologs. Nevertheless, the first two cases have most likely a limited impact as the majority of strains tend to have the same gene copy numbers [14] and samples often carry a dominant strain [20]. Regarding shared genes, their signals are a linear combination of the MSPs that carry them. Thus, they will be identified only if these MSPs are mostly detected in separate sets of samples.

**Computing performance**

MSPminer can process large datasets made up of thousands of samples and millions of genes in just a few hours on a regular single node server. The program is only limited by the amount of RAM available on the machine on which it is executed as the input count matrix must be fully loaded into memory. Consequently, RAM consumption grows linearly according to the number of samples and genes in the count matrix (Supplementary Figure 23).
MSPminer achieves good parallel efficiency (Supplementary Figure 24) through two parallelization strategies. First, a novel Map/Reduce programming model assigns genes to as many subsets as the number of available samples. Genes with greatest counts in the same sample are first compared, which not only decreases the number of comparisons to perform but increases the probability that related genes are placed in the same bin compared to random assignment (Supplementary Figure 25). Each subset of genes is processed in parallel and synchronization is only required before the reduction/merging step (see Methods). To avoid comparisons of all pairs of genes, others used an iterative algorithm where a random gene is compared to all the others until a significant proportion of genes is clustered [38]. This method allows fast identification of big clusters but struggles with the large number of small clusters and singleton genes. Furthermore, parallelism implementation is more complex as synchronization is required to detect duplicate clusters. In a second phase, MSPminer performs pairwise comparison of clusters to detect those corresponding to MSPs core gene sets. Then, full MSPs including accessory and shared genes are retrieved in parallel from the signal of their respective core genes. Here, no synchronization is required as core genes sets are supposedly independent.

Quality of MSPs

The quality of the MSPs is impacted by all the upstream steps required for generating the count matrix, as well as with the biological and ecological characteristics of the dataset. At the sequencing level, the number of reads (sequencing depth) generated for each sample impacts the detection and coverage of subdominant species, while reads length affects the quality of the assembly and the ability to assign a read to a gene without ambiguity. At the bioinformatics level, assembly, gene prediction, gene redundancy removal, mapping and counting require expertise to select the most appropriate strategies, tools and parameters. Indeed, assemblers returning chimeric contigs which combine sequences from highly related species, inaccurate predictors generating truncated or merged genes, redundancy removal with a common threshold for all genes (95% of nucleotide identity) lead to genes of variable quality in catalogues. Then, when quantifying genes, keeping only uniquely mapped reads underestimates the abundance of some genes whereas considering shared reads can generate false positives. Genes grouped in MSPs were significantly longer than those that were not (median length of 780 bp vs 498 bp, Wilcoxon rank-sum test p-value= 0) (Supplementary
Figure 26) as longer genes have a higher and less dispersed counts. Nevertheless, end-to-end mapping probably plays a role as it may fail to fully align a read whose size is approximately the same as the target gene. Finally, at the biology level, a high number of samples with varied phenotypes will improve the comprehensiveness and quality of MSPs. Indeed, as the number of samples grows, MSPminer will identify rare species and will extend the list of accessory genes of the MSPs corresponding to species with an open pan-genome. In addition, highly prevalent accessory genes will be reclassified from core to accessory as observed while sequencing an increasing number of strains of a species [49].

Applications

As illustrated in this paper, MSPminer supports the analysis of metagenomic data at species-level by identifying and quantifying the MSPs present in samples. Subsequently, MSPs associated with a given phenotype (e.g. the geographical origin) can be investigated both quantitatively and qualitatively. Here, information from unknown or non-sequenced species can be exploited. Compared to methods relying on marker genes [33,34], MSPminer improves the estimation of species abundance by automatically detecting among core genes those with the highest specificity, the highest counts and lowest dispersion. Moreover, in each MSP, genes or modules of accessory genes associated with the tested phenotype can be explored opening the way to a strain-level analysis. Thus, biomarkers corresponding to functional traits specific to certain strains can be discovered.

MSPminer also provides microbial population genetics from large cohorts which can support culture-dependent methods by identifying species of particular interest, such as those with no reference genome available or with reference genomes distant from the strains actually present. Reciprocally, MSPminer will benefit from advances in culture-dependent methods which provide reference genomes of low abundance species detectable by shotgun sequencing but difficult to assemble [23,24].

Further developments

Several improvements of MSPminer are considered. Algorithms that could identify relevant associations currently missed by MSPminer will be evaluated. For instance, deconvolution
algorithms [14] may discover genes shared between several MSPs while kernel density
estimators [54] may be useful for detecting genes with highly variable copy numbers.

To increase MSPminer specificity, a statistical test determining how much an association
between the core genome of a MSP and a gene is unexpected would be highly useful. Such a
statistical test could assess both co-occurrence with a Fisher exact test and co-abundance with
correlation test or equivalent. Currently, zero counts are either classified as structural or
undetermined while comparing genes abundance profiles. However, a statistical model
determining the probability that a zero is structural would allow overcoming threshold effects
and classifying with more accuracy a gene as core, accessory or shared.

Alternatives to the median for computing the representative of a MSP are envisaged. The sum
which cumulates counts from multiple genes is a prime candidate, as clusters would be
quantified with higher accuracy and increased quantification range particularly in samples
where its abundance is low. However, one should carefully account for outliers

Finally, the robustness of the measure of proportionality could be improved. For instance, a
robust linear regression [55] may replace the median for estimating the coefficient of
proportionality between the abundance profiles of two genes while the median absolute
deviation (MAD) could improve the detection of outliers [56].
Methods

Measure of proportionality between two genes

Let $M$ be a $n \times m$ matrix where $n$ is the number of genes and $m$ the number of samples. $M$ is composed of counts $c_{ij}$ representing the number of reads mapped on gene $i$ in sample $j$.

Let $g_i = (c_{i1}, c_{i2}, ..., c_{im})$ be the vector of the number of mapped reads on the gene $i$ across the $m$ samples.

The distribution of count data $g_i$ has the following properties:

1. Variance tends to be proportional to the counts.
2. It ranges over several orders of magnitude due to uneven sequencing depth and variable relative abundance of the gene between samples.
3. It usually contains many zeros as the majority of genes are observed in a few samples.
   For instance, the count matrix used in this study contains 92% of zeros.
4. It is prone to outliers. (see Discussion)

Let $g_x$ and $g_y (x, y \in [1, n] \text{ and } x \neq y)$ denote the vectors of the number of mapped reads on two distinct genes. A robust measure is proposed to assess direct proportionality between $g_y$ and $g_x$ (formally written $g_x \propto g_y$) which accounts for the points mentioned above.

Let $\alpha$ be the coefficient of proportionality between $g_x$ and $g_y$. $\alpha$ is a strictly positive constant expected to be roughly equal to the ratio of $g_x$ and $g_y$ length. However, it can be impacted by other factors such as uneven coverage or gene duplication (Supplementary Figure 27). Therefore, instead of relying on genes length, $\alpha$ was robustly estimated as follow:

$$\sqrt{\alpha} = \text{median} \left( \frac{\sqrt{c_{ys}}}{\sqrt{c_{xs}}} \right) \forall s \in [1, m] \text{ such as } c_{ys} \geq t \text{ and } c_{xs} \geq t \text{ with } t = 6 \text{ by default}$$

A square root transformation was applied to stabilize variance as suggested by several authors for count data [57,58]. For a comparison of some data transformations, refer to Supplementary Figure 28. The median was used to tolerate some outliers.
To estimate the coefficient of proportionality, only samples were both genes counts were above a threshold $t$ were kept. This has the following advantages:

1. It discards samples were both genes are absent as they do not provide any quantitative information for the estimation.
2. It discards samples with overdispersed low counts which do not allow a precise estimation of the coefficient of proportionality.
3. It discards samples where only one gene has a null count. In such sample, the zero count can be either a sampling zero that corresponds to an undetected gene because of sampling or technical effects or a structural zero that corresponds to unobserved gene actually absent in the sample. Distinguish structural from sampling zeros is crucial to classify a gene as core or accessory. Here, zeros below the threshold $t$ were of an undetermined type (yellow points in Figure 4) whereas those above were classified as structural (red points in Figure 4).

When $\alpha > 1$, $g_y$ yields more counts than $g_x$. As a result, a null count from $g_x$ can be misclassified as a structural zero.

When $\alpha \neq 1$, different quantification thresholds for $g_x$ and $g_y$ respectively named $t_x$ and $t_y$ were used to reflect the different yields for $g_x$ and $g_y$:

$$\alpha \geq 1 \rightarrow (t_x = t \text{ and } t_y = \alpha \cdot t)$$
$$\alpha < 1 \rightarrow (t_x = \frac{t}{\alpha} \text{ and } t_y = t)$$

**Non-robust version**

The relationship of proportionality between two genes was evaluated by the Lin’s concordance correlation coefficient [59] computed on samples that can be used for quantification.

The Lin’s concordance correlation coefficient was modified to detect strictly proportional variables, and not relations of the type $y = x$ as originally designed:

$$\frac{2 \cdot \alpha \cdot s_{xy}}{\alpha^2 \cdot s_x + (\alpha \cdot \overline{x} - \overline{y}^2)}$$
where $x = \sqrt{g_x}$, $y = \sqrt{g_y}$, $\bar{x}$ is the mean of $x$, $s_x$ is the unbiased variance of $x$ and $s_{xy}$ is the biased covariance of $x$ and $y$.

Robust version

This measure was used to assess the proportionality between two genes in presence of few outliers. If not treated specifically, even a small number of outliers may decrease significantly the concordance coefficient calculated.

Residuals were computed in samples where both genes have non-null counts with the following formula:

$$abs(\sqrt{g_y} - \alpha \cdot \sqrt{g_x})$$

Let $Q_1$ and $Q_3$ be the first and third quartiles of the residuals. Let $IQR$ be the interquartile range defined by $IQR = Q_3 - Q_1$. Samples with residuals above $Q_3 + 1.5 \cdot IQR$ were considered as outliers (purple points in Figure 4).

Let $m$ be the number of samples for which residuals were computed. If there were more than $(m - 5) \cdot 0.3$ outliers (percentage of outliers asymptotically equal to 30%), the robust measure of proportionality was not computed. Otherwise, the concordance coefficient was calculated on non-outlier samples.

Simulation

A simulated gene abundance table which quantifies genes of a single virtual species across 300 virtual samples was generated.

The species pan-genome consisted of 2000 core genes and 2900 accessory genes of variable length. Genes length was generated from a negative binomial ($\text{mean} = 1000; \text{overdispersion} = 0.3$) and values below 100 and above 5000 were rejected.

Each sample was carrier of a strain with specific gene content. For each prevalence ranging from 10 to 299 samples, 10 accessory genes were considered. The subset of genes in which an accessory gene was detected was drawn randomly.

The number of mapped reads per sample was generated from a log normal law ($\text{mean}=\log(100000); \text{sd}=1$) and values below 100000 and above 1000000 were rejected.
The theoretical number of counts attributed to each gene was calculated according to its length and its presence or not in the strain. The observed gene counts were generated by using a negative binomial distribution (mean=theoretical gene count; overdispersion=0.05) to approach real metagenomic data. Outliers were added in each gene by selecting 5%, 10% and 20% of their non-null samples and multiplying each observed count by either ¼, ⅓, ⅔ or 4.

**Metagenomic Species Pan-genome generation**

The Supplementary Figure 29 gives an overview of the MSPminer workflow.

**Input data**

MSPminer processes raw gene counts tables. In this study, we used the table provided with the Integrated Gene Catalog of the human gut microbiome (1267sample.gene.pairNum.table) available on GigaDB [60].

**Data filtering**

Rare genes which do not support enough quantitative information for further processing were discarded. By default, genes with counts greater than 6 in at least 3 samples were kept.

**Data transformation**

A square root transformation was applied to gene counts. This transformation stabilizes gene counts variance and limits the skewness of gene counts distribution.

**Genes binning**

Genes with the highest counts in the same sample were binned. To limit bias due to variable sequencing depth, raw read counts were normalized by the number of mapped reads prior to bin assignment (Supplementary Figure 30). Note that normalized counts were used in this step only.

**Seeds creation**

This step identifies sets of co-abundant and co-occurring genes called seeds hereafter. Seeds were created in parallel in each bin by a greedy approach. First, genes were compared pairwise. All pairs of genes with a non-robust measure of proportionality of at least 0.8 and no structural zeros were saved in a list. Then, the list was sorted by decreasing measure of
proportionality. The pair of genes with the highest measure of proportionality was selected as a centroid. Genes related to one of the centroid genes were grouped together in a new seed.

**Seed representative**

For each seed, a pseudo gene referred as *representative* was computed as follow. First, the seed representative was defined as the median vector of the counts of all its genes. Then, each gene of the seed was compared to the seed representative using the measure proportionality. The final seed representative corresponded to the median vector of the counts of the 30 genes with the highest measure of proportionality.

**Seeds merging**

Some related genes may have been assigned to different bins, for instance, in a situation where samples with the highest counts had close values. Therefore, a merging step was performed. First, seeds from all the bins were pooled and sorted by decreasing size. Then, the representative of the largest seed was compared to the representatives of the other seeds. Seeds with a non-robust measure of proportionality of at least 0.8 and no structural zeros counts were merged with the largest seed to form the final seed. Merged seeds were removed from the list and the procedure was iterated until or there were no more seeds to process. After merging, seeds with less than 200 genes were discarded.

**Core seeds identification**

In this step, core seeds were identified among final seeds, based on the assumption that in a set of related seeds, the largest corresponds to a species core genome and the others are modules of either accessory or shared genes. First, seeds were sorted by decreasing number of genes. The largest seed was defined as a new core seed. Then, the representative of the core seed was compared to the representative of all remaining seeds. The seeds with a robust measure of proportionality of at least 0.8 were considered as related to the core seed and discarded from the list of potential cores. The procedure was iterated until there was no more seed to process.

**Metagenomic Species Pan-genome generation**

The representatives of each core seed were compared to all the genes. Because core seeds were identified all at once in the previous step, the MSPs generation was run in parallel. Genes
with a robust measure of proportionality of at least 0.8 were considered as related to the core seed.

Let $g_x$ be the median vector of the number of mapped reads on the core seed and $g_y$ the vector of the number of mapped reads on a gene related to the core seed. The related gene was assigned to one of the 4 following categories:

1. core genes were detected in the same samples as the core seed (Figure 4.A).
   $$(g_y \propto g_x) \text{ and } \forall s \in [1, m] c_{xs} \geq t_x \rightarrow c_{ys} \neq 0 \text{ and } c_{ys} \geq t_y \rightarrow c_{xs} \neq 0$$

2. accessory genes were detected in a subset of samples where the core seed was detected (Figure 4.B).
   $$(g_y \propto g_x) \text{ and } \exists s \in [1, m] c_{xs} \geq t_x \text{ and } c_{ys} = 0 \text{ and } \forall s \in [1, m] c_{ys} \geq t_y \rightarrow c_{xs} \neq 0$$

3. shared core genes were detected in all the samples where the core seed was detected plus some samples where it was not (Figure 4.C).
   $$(g_y \propto g_x) \text{ and } \forall s \in [1, m] c_{xs} \geq t_x \rightarrow c_{ys} \neq 0 \text{ and } \exists s \in [1, m] c_{ys} \geq t_y \text{ and } c_{xs} = 0$$

4. shared accessory genes were detected in a subset of samples where the core seed was detected plus some samples where it was not (Figure 4.D).
   $$(g_y \propto g_x) \text{ and } \exists s \in [1, m] c_{xs} \geq t_x \text{ and } c_{ys} = 0 \text{ and } \exists s \in [1, m] c_{ys} \geq t_y \text{ and } c_{xs} = 0$$

In each category, a clustering procedure similar to the one used to create seeds was run. It identified modules of co-occurring genes that may be interpreted as functional units, i.e. operons. Unclustered genes were saved as singleton modules.

**Comparison to the Canopy clustering algorithm**

The implementation of the Canopy clustering algorithm was downloaded at https://www.cbs.dtu.dk/projects/CAG/Supplementary_Software_canopy_clustering.zip. The gene count table normalized following the procedure described in [3] was taken as an input. Default parameters were used. MSPs were projected on Co-Abundance gene Groups (CAGs) and reciprocally with an in-house script.
Biomarkers discovery

Identification of MSPs associated with geographical origin

A two-tailed Wilcoxon rank-sum test was used on relative median abundance of the 30 best representative core genes of each MSP (1,696 tested variables). The obtained p-values were adjusted by the Benjamini-Hochberg procedure. In addition, a log₂ ratio was computed between the median abundances of the MSP in the two populations tested. MSPs with an adjusted p-value inferior to $10^{-2}$ and a log₂ ratio superior to 1 were considered significant.

Identification of accessory genes associated with geographical origin

For each accessory gene of a MSP, a 2x2 contingency table counting in both populations the number of samples where the gene was present or absent was built. Only samples where the MSP core genome was detected were kept. A gene was considered as present in a sample if at least two reads were mapped on it. Then, a chi-squared test was performed on each contingency table. Accessory genes with a p-value inferior to $10^{-10}$ were considered significant. Log₂ presence ratios equal to +infinity or −infinity were replaced by +10 or −10 respectively.

Taxonomic annotation of the gene catalog

Genes were aligned at the nucleotide level using BLASTn [61] (version 2.6.0) against KEGG GENOME [62] (Release 82.0, April 2017) and RefSeq [63] (Release 81, March 2017). Hits that covered less than 80% of the query gene or with a e-value superior to 0.01 were discarded. Thresholds of 95%, 80% and 65% of nucleotide identity were respectively used for taxonomic annotation at species, genus and phylum level. At a given taxonomic level, “no consensus” was reported if the selected hits did not share the same annotation. Finally, results from KEGG were preferred to those from RefSeq.

MSPs were assigned to the lowest level taxon representing more than 50% of the annotations of their core genes. Taxonomic annotations of MSPs with at least 50% of their core genes annotated at species level (including “no consensus”) and less than 80% assigned to the most represented species were considered ambiguous (c.f. Supplementary Table 3).
Functional annotation of the gene catalog

Translated genes were annotated with eggnog-mapper [64] (version 0.12.7) based on eggNOG orthology assignments [65]. Sequence similarity searches were performed using HMMER [66]. The 40 universal single copy marker genes were discovered using fetchMG v1.0 [67]

Comparison of the MSPs to sequenced genomes

Genomes used to build the integrated catalog of the human gut microbiome [12], HMP reference genomes [68] and genomes from species detected while performing the taxonomic annotation of the MSPs were downloaded from GenBank [69]. When not provided, CDS were predicted with Prodigal [70]. Genes from the reference catalog were aligned against the genomes with BLASTn [61] (version 2.6.0; arguments: -perc_identity 95 -ungapped). Alignments of less than 100 nucleotides were discarded. Hits found for a single gene at neighboring positions on the target genome were merged. 642 genomes with less than 10% of their constituent genes detected in the reference catalog were kept. Genes were annotated with the related MSP information when available. Hits from most abundant MSP were kept and overlapping hits from less abundant MSPs were discarded. The local GC-content of genome was computed using a sliding window of approximately 100 nucleotides. Finally these data were plotted using Circos [71].
Declarations

Funding
This work was funded by Enterome, the ANRT (Association Nationale de la Recherche Technique) via the grant CIFRE 2014/0057 and INRA MetaGenoPolis via the grant “Investissements d’avenir” ANR-11-DPBS-0001.

Authors’ contributions
FPO and MP designed the software, performed the analyses and wrote the manuscript. FPO implemented the software. FM, AC and SDE supervised the project and revised the manuscript.

Additional files
Supplementary Table 1: Tab-separated file listing the genes and modules in the MSPs
Supplementary Table 2: XLS file describing the MSPs (taxonomic annotation, number of genes, number of universal marker genes, prevalence and abundance)
Supplementary Table 3: XLS file listing the MSPs with ambiguous annotation at species level.
Supplementary Table 4: XLS file summarizing the comparison of the MSPs to 642 sequenced genomes.
Supplementary Table 5: XLS file listing the genomic regions of Parabacteroides distasonis ATCC 8503 containing genes not assigned to the msp_0011.
Supplementary Table 6: XLS file listing the MSPs associated with the geographic origin of samples.
Supplementary Table 7: XLS file listing the accessory genes of the msp_0011 (Parabacteroides distasonis) associated with the geographic origin of samples.
References


43. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of


58. Osborne J. Notes on the use of data transformations. Pract. Assess. [Internet]. 2002;8:4–13. Available from: c:%5CUsers%5CTeresa%5CDocuments%5CTeseFinal%5C2010%5Cserinus2010%5CPDF’sParaOrganizar%5CPDF’s2010%5Ctransformation.pdf


integrated catalog of reference genes in the human gut microbiome” [Internet]. GigaScience Database; 2014. Available from: http://gigadb.org/dataset/100064


48.pdf?Expires=1499795422&Signature=W7WZu1nfufNHcrfl1QFCH^13BjwFo82LccucPBoCvq

GvLaCiaQFlzzY6iT3cfR4A1bFPxYnhBaz6D-wBoorpxD5SfKpgd63KRkl2HHHdcb2BEBArfT4f


890  Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.092759.109
**Figures**

<table>
<thead>
<tr>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
<th>sample4</th>
<th>sample5</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="1x" /></td>
<td><img src="image2" alt="2x" /></td>
<td><img src="image3" alt="1x" /></td>
<td><img src="image4" alt="3x" /></td>
<td><img src="image5" alt="2x" /></td>
</tr>
</tbody>
</table>

↓ Shotgun sequencing

<table>
<thead>
<tr>
<th>sample1</th>
<th>sample2</th>
<th>sample3</th>
<th>sample4</th>
<th>sample5</th>
</tr>
</thead>
<tbody>
<tr>
<td>core gene 1</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>core gene 2</td>
<td></td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>core gene 3</td>
<td></td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>accessory gene 1</td>
<td></td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>accessory gene 2</td>
<td></td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

*Figure 1*
Figure 2
Figure 3: Core genes connected to Accessory module with a shared module linking MSP 1 to MSP 2.
Figure 4
Figure 5
Figure 6


Figures legends

Figure 1: Simple model illustrating the rationale behind the method
5 samples carry different strains from the same species. Three core genes (red, blue, orange) are present in all the strains. Two accessory genes (green, purple) are present only in some strains. The abundance of the species in each sample ranges from 1 to 3 copies.

After shotgun sequencing, a raw gene abundance matrix is built.

A strict proportionality relationship is expected between two core genes, the proportionality coefficient being equal to the ratio of their length. In contrast, such relationship between a core and an accessory gene should be observed only in the subset of samples where the accessory gene is present.

Figure 2: Comparison of the genes abundance profiles of the virtual species to the median signal of its core genome
using:

A. the Pearson correlation coefficient
B. the Spearman correlation coefficient
C. the measure of proportionality

Figure 3: Structure of Metagenomic Species Pan-genomes (MSPs)

Figure 4: Illustration of the four types of genes in a MSP
The core genome median abundance of the msp_0043 (*Ruminococcus bromii*) is compared to:

A. the gene MH0003/GL0010264 classified as core. The gene is detected in all the samples where MSP is detected.
B. the gene MH0025/GL0082295 classified as accessory. The gene is missing in 516 samples where the MSP is detected.
C. the gene 657321.RBR_R.22270 classified as its shared core. The gene is present in all the samples where MSP is detected but also in 286 samples where the MSP is not.
D. the gene MH0205/GL0102923 classified as shared accessory. The gene is missing in 454 samples where the MSP is detected but present in 28 samples where the MSP is not.
Figure 5: Comparison of the gene content of some MSPs and their corresponding CAGs

Figure 6: Comparison of the genes abundance profiles of the msp_0011 (Parabacteroides distasonis) to the median signal of its core genome.

Three measures are compared:

A. the Pearson correlation coefficient
B. the Spearman correlation coefficient
C. the measure of proportionality.

Grey points correspond to genes unclassified by Canopy whereas those colored were grouped in CAGs.

Figure 7: Circos representation of the mapping of the msp_0011 on the genome of P. distasonis strain ATCC 8503

Description of layers from outside to inside:

A. Position on chromosome
B. GC-content (format: histogram)
C. gene or module type (format: highlight):
   - green: core
   - yellow: accessory
   - blue: shared core
   - purple: shared accessory
D. MSP (format: highlight):
   - Bandwidth:
     - wide: gene grouped in a MSP
     - narrow: gene grouped in a seed
   - color code:
     - red: gene grouped in the most represented MSP
     - other color + grey: gene grouped in another MSP or a seed
E. Sample assignment (format: histogram):
   - facing outwards if the gene is related to the most represented MSP, facing inwards otherwise.
• color code:
  - grey: samples where the MSP module is not detected
  - green: samples where the MSP core and the gene are detected
  - yellow: samples where the MSP core is detected but not the gene
  - purple: samples where the gene is detected but not the MSP core
Supplementary Figures

Supplementary Figure 1

A. 5% of outliers

B. 10% of outliers

C. 20% of outliers

Measure of proportionnality
- non robust
- robust

Supplementary Figure 1
Supplementary Figure 3

- Firmicutes
- Bacteroidetes
- Proteobacteria
- Actinobacteria
- Fusobacteria
- Synergistetes
- Verrucomicrobia

Number of MSPs
Supplementary Figure 4

Percentage of MSP genes assigned

- core genes
- accessory genes

- dominant species
- different species
- different genus
- different phylum
- not annotated
Supplementary Figure 6
Supplementary Figure 7

The figure shows the density of accessory gene prevalence across different bacterial species. The x-axis represents accessory gene prevalence (%) ranging from 0% to 100%, and the y-axis represents density. Different bacterial species are color-coded as follows:
- **msp_0003** - *Bacteroides cellulosilyticus*
- **msp_0007** - *Bacteroides intestinalis*
- **msp_0011** - *Parabacteroides distasonis*
- **msp_0012** - *Ruminococcus bicirculans*
- **msp_0014** - *Roseburia faecis*
- **msp_0016** - *Roseburia intestinalis*
Supplementary Figure 8

The figure shows a scatter plot with the x-axis representing the prevalence of the MSP and the y-axis representing the number of accessory genes in the MSP. The data points are distributed across the plot, indicating a correlation between the two variables.
Supplementary Figure 9
Supplementary Figure 10

**Graph Description:**

The graph illustrates the relationship between the percentage of samples where the MSP is detected and the mean abundance of the MSP when detected. The x-axis represents the percentage of samples where the MSP is detected, ranging from 1% to 100%. The y-axis shows the mean abundance of the MSP when detected, ranging from 0.1% to 10.0%. The data points are distributed across the graph, indicating a trend that as the percentage of samples increases, the mean abundance decreases, although there are outliers that deviate from this trend.
Supplementary Figure 11

Panel A: Histogram showing the number of MSPs detected within a range of 0 to 40 SCMs.

Panel B: Bar graph illustrating the prevalence of SCMs within their corresponding MSPs, ranging from 0% to 100%.
Supplementary Figure 13
Supplementary Figure 15
Supplementary Figure 16
Supplementary Figure 17

(A) Distribution of MSP prevalence for different MSPs.

(B) Number of MSPs versus total number of genes in MSP.
Supplementary Figure 18
Supplementary Figure 19

---

*American* | *European*
--- | ---
msp_0113 Anaerostipes hadrus | ![Bar Chart](chart1.png)
msp_0051 Eubacterium hallii | ![Bar Chart](chart2.png)
msp_0079 unassigned Blautia | ![Bar Chart](chart3.png)
msp_0618 not annotated | ![Bar Chart](chart4.png)
msp_0565 not annotated | ![Bar Chart](chart5.png)
msp_0204 not annotated | ![Bar Chart](chart6.png)
msp_0369 Streptococcus salivarius | ![Bar Chart](chart7.png)
msp_0151 Fusicatenibacter saccharivorans | ![Bar Chart](chart8.png)
msp_0251 Coprococcus comes | ![Bar Chart](chart9.png)
msp_0724 not annotated | ![Bar Chart](chart10.png)
msp_0466 unassigned Faecalibacterium | ![Bar Chart](chart11.png)
msp_0930 Coprococcus catus | ![Bar Chart](chart12.png)
msp_0680 Blautia obeum | ![Bar Chart](chart13.png)
msp_0460 Clostridium phocaeensis | ![Bar Chart](chart14.png)
msp_0297 Dorea longicatenae | ![Bar Chart](chart15.png)

**q** values:
- msp_0113: 4.92e-45
- msp_0051: 1.96e-39
- msp_0079: 7.79e-36
- msp_0618: 7.68e-32
- msp_0565: 2.95e-30
- msp_0204: 2.81e-29
- msp_0369: 2.91e-29
- msp_0151: 1.62e-26
- msp_0251: 3.31e-26
- msp_0724: 1.46e-25
- msp_0466: 2.07e-24
- msp_0930: 3.74e-24
- msp_0680: 4.69e-24
- msp_0460: 2.67e-23
- msp_0297: 1.14e-21

---

CC-BY-NC-ND 4.0 International license available under a not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made this version posted August 17, 2017. ; https://doi.org/10.1101/173203

doi: bioRxiv preprint
Supplementary Figure 22

A

B

C

\sqrt{\text{msp}_{0140}}

\sqrt{\text{msp}_{0255}}

\sqrt{\text{msp}_{0255}}

\sqrt{\text{msp}_{0140}}
Supplementary Figure 23

(A) RAM consumption (Gb) vs. Number of genes (in millions)

(B) RAM consumption (Gb) vs. Number of samples (in hundreds)
Supplementary Figure 25

A

Total number of comparisons

B

Fraction of useful comparisons

- no binning
- with binning
Supplementary Figure 28

A  

B

C

gene1

log_{10}(gene2)

\sqrt{gene1}

gene2
Supplementary Figure 30

Counts type
- normalized
- raw

Bin size (#genes)

Supplementary Figure 30
Supplementary Figures legends

Supplementary Figure 1: Comparison of the noisy genes abundance profiles of the virtual species to the original median signal of its core genome using either the non-robust or robust measure of proportionality with an increasing percentage of outliers (5%, 10% and 20%).

Supplementary Figure 2: Lowest taxonomic rank assigned to MSPs
(1 square = 1 MSP)

Supplementary Figure 3: Most represented phyla among MSPs

Supplementary Figure 4: Taxonomic homogeneity of MSPs
Only MSPs annotated at species level were kept.

Supplementary Figure 5: Histogram of the total number of genes in MSPs

Supplementary Figure 6: Comparison of the number of genes in a MSP and its number of accessory genes.
The four outliers in red correspond to eukaryotic intestinal parasites of the *Blastocystis* genus (msp_0001, msp_0002, msp_0004 and msp_0006).

Supplementary Figure 7: Prevalence of the accessory genes of some MSPs among samples where their respective core genes are detected.

Supplementary Figure 8: Comparison of the number of samples where a MSPs is detected and its number of accessory genes.

Supplementary Figure 9: Prevalence of MSPs among the 1267 samples.

Supplementary Figure 10: Comparison of the percentage of samples where a MSPs is detected and its mean abundance.
The red, blue and green points correspond to the MSPs associated to *Bacteroides vulgatus*, *Bacteroides uniformis* and *Prevotella copri* respectively.
Supplementary Figure 11: Single Copy Marker genes in MSPs

A. Number of universal single copy marker genes (SCM) in MSPs

B. Prevalence of the universal single copy marker genes (SCM) in the MSPs they belong to.

Supplementary Figure 12: Percentage of genes from reference genomes grouped in their best corresponding MSPs.

Supplementary Figure 13: Comparison of the genes detected in P. distasonis ATCC 8503 with those of its corresponding MSP.

Each square represents 5 genes. Dark colors correspond to genes detected both in the genome and the MSP while light colors correspond to genes not found in the genome.

Supplementary Figure 14: Median distance between consecutive genes in modules of the msp_0011 detected in P. distasonis ATCC 8503, as a function of module prevalence and module size.

The median distance between successive genes in a module was below 150 bp in 75% of the cases. Median distances of 0 base pairs correspond to modules with overlapping genes.

Supplementary Figure 15: Example of genes found within a genomic region of P. distasonis ATCC 8503 delimited by a module

The 4 genes of the accessory module 0185 are detected in the genome of P. distasonis ATCC 8503 between positions 441 398 - 442 762.

The genes 435591.BDI_0388 and O2.UC52-0_GL0104470 are detected in this region between positions 441903 - 442697 and 442472 - 442697 respectively but were not grouped in the module.


3.D: The gene 435591.BDI_038 was not grouped in the module because it was missing in 3 samples where the module was detected. Here, one may consider that the clustering criteria was too stringent.
3.E: The gene O2.UC52-0_GL0104470 was not grouped in the module because it was missing in 77 samples where the module was detected. Here, excluding the module gene seems relevant.

**Supplementary Figure 16: Example of genes of *P. distasonis* ATCC 8503 not found in the msp_0011**

4.A: O2.UC27-0_GL0024692 (alien region 4) corresponds to a false negative with overdispersed counts. The robust measure of proportionality was above the threshold set by MSPMiner.

4.B: 435591.BDI_0319 (alien region 3) corresponds to a false negative subject to copy-number variation. The robust measure of proportionality was above the threshold set by MSPMiner.

4.C: MH0388_GL0156798 (alien region 10) probably corresponds to a false negative. The number of observations was too low to detect an association.

4.D: O2.UC32-1_GL0117083 (alien region 52) might correspond to a false negative. The number of observations was too low to detect an association.

4.E V1.UC40-1_GL0124158 (alien region 86) corresponds to a true negative as there is no clear association. Remarkably, this gene was annotated as a mobile element (conjugative transposon protein).

4.F V1.FI04_GL0019901 (alien region 21) corresponds to a true negative as there is no clear association. Interestingly, this gene was classified as shared accessory in the msp_0025 corresponding to *Parabacteroides merdae*.

**Supplementary Figure 17: Description of the MSPs with no equivalent among CAGs.**

A. Histogram of the prevalence of the MSPs with no equivalent among CAGs.

B. Histogram of the number of genes in the MSPs with no equivalent among CAGs.
Supplementary Figure 18: Relative abundance of the top 10 discriminant MSPs between Chinese and Westerners.

Supplementary Figure 19: Relative abundance of the top 10 discriminant MSPs between Americans and Europeans.

Supplementary Figure 20: The 40 accessory genes (in column) of the msp_0011 (Parabacteroides distasonis) most associated with the geographical origin of samples (in row).

Supplementary Figure 21: Example of genes subject to copy-number variation.

A. Comparison of the core genome median abundance of the msp_0011 (Parabacteroides distasonis) to one of its core genes (MH0150/GL0074117) subject to Copy-Number Variation. Here, MSPminer wrongly estimates the coefficient of proportionality as there is no dominant gene copy number among the samples.

B. Comparison of the core genome median abundance of the msp_0070 (Eubacterium rectale) to one of its core genes (BGI-17A/GL0065038) subject to Copy-Number Variation. Here, MSPminer correctly estimates the coefficient of proportionality for the majority of samples (blue points) that have the same number of copies of the gene. The others are classified as outliers (purple points).

Supplementary Figure 22: Example of genes shared between several MSPs.

The gene SZEY-26A/GL0070503 is classified as shared accessory in the msp_0140 (Bifidobacterium pseudocatenulatum) and the msp_0255 (Bifidobacterium adolescentis)

A. Comparison of the core genome median abundance of the msp_0140 to the gene SZEY-26A/GL0070503. Despite the presence of outliers, a direct proportionality relationship is observed in many samples where both the msp and the gene are present.

B. Comparison of the core genome median abundance of the msp_0255 to the gene SZEY-26A/GL0070503. Despite the presence of outliers, a direct proportionality relationship is observed in many samples where both the msp and the gene are present.
C. Comparison of the core genome median abundance of the msp_0140 to the core genome median abundance of the msp_0255. Although the two MSPs share a gene, they have distinct signals.

Supplementary Figure 23: Memory usage of MSPminer.
A. RAM consumption as a function of the number of genes. The number of samples was set at 1267. The grey line represents the best linear fit (intercept=5.05, slope=5.17).
B. RAM consumption as a function of the number of samples. The number of genes was set at 9,879,896. The grey line represents the best linear fit (intercept=6.14, slope=3.92).

Supplementary Figure 24: Parallel performance of MSPminer
The tests were performed on a server with an Intel® Xeon® processor E5-2630 (6 physical cores with hyperthreading) and 128Gb of RAM.

Supplementary Figure 25: Performance of MSPminer with or without genes binning.
A. Total number of comparisons performed by comparing all genes pairwise with or without binning. Here, genes binning divides the number of comparisons performed by 763.
B. Fraction of useful comparisons for MSP cores creation with or without binning. The fraction without binning was estimated by drawing randomly one billion pairs of genes. Here, genes binning multiplies the fraction of useful comparisons by 562.

Supplementary Figure 26: Comparison of the length of the genes grouped in the MSPs with those that were not.

Supplementary Figure 27: Comparison of the genes length and their respective coefficient of proportionality.
Core genes abundance profiles of the msp_0043 (Ruminococcus bromii) were compared to its core genome median abundance. For each gene, the estimated coefficient of proportionality $\alpha$ was compared to its length. Then, a robust linear regression with a null slope was performed (red line). As expected, the coefficient of proportionality $\alpha$ is proportional to the gene length.
However, some genes have a coefficient of proportionality lower than expected (points below the line) and may correspond to genes not uniformly covered.

**Supplementary Figure 28: Impact of some data transformations on gene counts.**

Raw (A), log-transformed (B) and square root transformed (C) abundance profiles of two core genes (MH0062_GL0071339 and MH0011_GL0069836) of the msp_0003 (*Bacteroides cellulosilyticus*) are compared.

A. Raw data is not appropriate when the order of magnitude of counts is highly variable as it gives too important weight to high counts.

B. Log-transformation “spreads” the counts. However, it requires the introduction of a pseudo-count for zeros and variance tends to be proportional to the counts.

C. Sqrt-transformation “spreads” the counts and stabilizes variance.

**Supplementary Figure 29: MSPminer workflow**

**Supplementary Figure 30  Number of genes per bin on the Integrated Gut Catalog with or without normalization.**

Normalization consists in dividing gene counts by number of mapped reads per sample. The black line represents the optimal bin size (6 971 229/1 267=5 502 genes)