bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Florian Plaza Oñate[1,2], Alessandra C. L. Cervino[1], Frédéric Magoulès[3], S. Dusko Ehrlich[2] & Matthieu Pichaud[1]§

[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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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].

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [42]. 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 1 267 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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [43]. 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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Methanomethylophilus*

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [44]. 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 [22,45], 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 more accessory genes (Supplementary Table 2), many corresponded to species reported as highly variable such as *Escherichia coli* [46], *Klebsiella pneumoniae* [47] or *Clostridium bolteae* [48]. As previously observed in population genomics studies comparing multiple strains of the same species [49,50], 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 MSPminer will recover its accessory genes, especially the rare ones. Many MSPs

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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* [51] or *Clostridium clostridioforme* [52].

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) [45]. 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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [53] to its corresponding MSP (Figure 7). Among the 3 850 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).

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 6A). 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 6B). 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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [54] 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).

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [50].

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

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

algorithms [14] may discover genes shared between several MSPs while kernel density estimators [55] 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 [56] 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 [57].

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 $g_i = (c_{i1}, c_{i2}, \dots, c_{im})$ $i \in [1, n]$ 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 ($x, y \in [1, n]$ 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 $g_x \infty g_y$) which accounts for the points mentioned above.

Let α be the coefficient of proportionality between g_x and g_y. α 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, α was robustly estimated as follow:

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

A square root transformation was applied to stabilize variance as suggested by several authors for count data [58,59]. For a comparison of some data transformations, refer to Supplementary Figure 28. The median was used to tolerate some outliers.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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:

$$\alpha \geq 1 \rightarrow (t_x = t \; and \; t_y = \alpha \cdot t)$$

$$\alpha < 1 \rightarrow \left(t_x = \frac{t}{\alpha} \; and \; t_y = t\right)$$

Non-robust version

The relationship of proportionality between the two genes was evaluated by a modified version of the Lin's concordance correlation coefficient [60] by considering only samples were both genes had non-null counts.

The Lin's concordance correlation coefficient originally designed to assess relationships of the type $y = x$ was defined as:

$$\frac{2 \cdot cov(x, y)}{\sigma_x^2 + \sigma_y^2 + (\overline{x} - \overline{y})^2}$$

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\bar{x}$ and $\bar{y}$ are the means, σ_x^2 and σ_y^2 are the variances and $cov(x, y)$ is the covariance of the variables x and y.

To assess relationships of the type $y = kx$ where k is a constant, the variable x was substituted by kx in the formula hereabove:

$$\frac{2k \cdot cov(x, y)}{k^2 \cdot \sigma_x^2 + \sigma_y^2 + (k \cdot \overline{x} - \overline{y})^2}$$

After substituting the variables x and y by $\sqrt{g_x}$ and $\sqrt{g_y}$ respectively and k by $\sqrt{\alpha}$, the final formula was:

$$\frac{2\sqrt{\alpha} \cdot cov(\sqrt{g_x}, \sqrt{g_y})}{\alpha \cdot \sigma^2_{\sqrt{g_x}} + \sigma^2_{\sqrt{g_y}} + \left(\sqrt{\alpha} \cdot \overline{\sqrt{g_x}} - \overline{\sqrt{g_y}}\right)^2}$$

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:

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

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 $Q_3 + 1.5 \cdot \mathrm{IQR}$ 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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 2 000 core genes and 2 900 accessory genes of variable length. Genes length was generated from a negative binomial (mean = 1 000; overdispersion = 0.3) and values below 100 and above 5 000 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 (mean=log(1 000 000), sd=1) and values below 100 000 and above 10 000 000 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, 2 ,3 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 [61].

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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene 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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 4A).

$$(g_y \infty g_x)\ and\ (\forall s \in [1,m] c_{xs} \geq t_x \rightarrow\ c_{ys} \neq 0\ and\ c_{ys} \geq t_y \rightarrow\ c_{xs} \neq 0)$$

2. accessory genes were detected in a subset of samples where the core seed was detected (Figure 4B).

$$(g_y \infty g_x)\ and\ (\exists s \in [1,m] c_{xs} \geq t_x and\ c_{ys} = 0)\ and\ (\forall s \in [1,m]\ c_{ys} \geq t_y \rightarrow\ c_{xs} \neq 0)$$

3. shared core genes were detected in all the samples where the core seed was detected plus some samples where it was not (Figure 4C).

$$(g_y \infty g_x)\ and\ (\forall s \in [1,m]\ c_{xs} \geq t_x \rightarrow\ c_{ys} \neq 0\ )\ and\ (\exists s \in [1,m] c_{ys} \geq t_y and\ c_{xs} = 0)$$

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 4D).

$$(g_y \infty g_x)\ and\ (\exists s \in [1,m] c_{xs} \geq t_x and\ c_{ys} = 0)\ and\ (\exists s \in [1,m] c_{ys} \geq t_y and\ c_{xs} = 0)$$

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 $\log_2$ 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 $\log_2$ 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. Log_2 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 [62] (version 2.6.0) against KEGG GENOME [63] (Release 82.0, April 2017) and RefSeq [64] (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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [65] (version 0.12.7) based on eggNOG orthology assignments [66]. Sequence similarity searches were performed using HMMER [67].

The 40 universal single copy marker genes were discovered using fetchMG v1.0 [68]

Comparison of the MSPs to sequenced genomes

Genomes used to build the integrated catalog of the human gut microbiome [12], HMP reference genomes [69] and genomes from species detected while performing the taxonomic annotation of the MSPs were downloaded from GenBank [70]. When not provided, CDS were predicted with Prodigal [71]. Genes from the reference catalog were aligned against the genomes with BLASTn [62] (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 [72].

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Declarations

Funding

This work was funded by Enterome, the ANRT (Association Nationale de la Recherche et de la Technologie) 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.

Competing interests

The authors declare that they have no competing interests.

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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. O'Hara AM, Shanahan F. The gut flora as a forgotten organ. EMBO Rep. [Internet]. 2006;7:688–93. Available from: http://embor.embopress.org/cgi/doi/10.1038/sj.embor.7400731

2. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward D V, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. [Internet]. 2012;13:R79. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-r79

3. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature [Internet]. 2012;490:55–60. Available from: http://www.nature.com/doifinder/10.1038/nature11450

4. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature [Internet]. 2013;500:541–6. Available from: http://www.nature.com/doifinder/10.1038/nature12506

5. Kostic AD, Xavier RJ, Gevers D. The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead. Gastroenterology [Internet]. 2014;146:1489–99. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0016508514002200

6. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, et al. Alterations of the human gut microbiome in liver cirrhosis. Nature [Internet]. 2014;513:59–64. Available from: http://www.nature.com/doifinder/10.1038/nature13568

7. Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, et al. Gut microbiome development along the colorectal adenoma–carcinoma sequence. Nat. Commun. [Internet]. 2015;6:6528. Available from: http://www.nature.com/doifinder/10.1038/ncomms7528

8. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS One. 2013;8.

9. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, et al. The truth

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. [Internet]. 2015;15:66. Available from: http://www.biomedcentral.com/1471-2180/15/66

10. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. [Internet]. Elsevier Ltd; 2016;469:967–77. Available from: http://dx.doi.org/10.1016/j.bbrc.2015.12.083

11. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front. Microbiol. [Internet]. 2016;7. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2016.00459/abstract

12. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat. Biotechnol. [Internet]. 2014;32:834–41. Available from: http://www.nature.com/doifinder/10.1038/nbt.2942

13. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature [Internet]. Nature Publishing Group; 2012;493:45–50. Available from: http://www.nature.com/doifinder/10.1038/nature11711

14. Greenblum S, Carr R, Borenstein E. Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. Cell [Internet]. 2015;160:583–94. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867415000136

15. Zhu A, Sunagawa S, Mende DR, Bork P. Inter-individual differences in the gene content of human gut bacterial species. Genome Biol. [Internet]. 2015;16:82. Available from: http://genomebiology.com/2015/16/1/82

16. Larsbrink J, Rogers TE, Hemsworth GR, McKee LS, Tauzin AS, Spadiut O, et al. A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. Nature [Internet]. 2014;506:498–502. Available from: http://www.nature.com/doifinder/10.1038/nature12907

17. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, et al. A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Shiga-Toxigenic Escherichia coli O104:H4. JAMA [Internet]. 2013;309:1502. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2013.3231

18. Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang Y-F. Analysis of Ultra Low Genome Conservation in Clostridium difficile. Horsburgh MJ, editor. PLoS One [Internet]. 2010;5:e15147. Available from: http://dx.plos.org/10.1371/journal.pone.0015147

19. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nat. Biotechnol. [Internet]. 2015;33:1045–52. Available from: http://www.nature.com/doifinder/10.1038/nbt.3319

20. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure & genetic diversity from metagenomes. Genome Res. 2017;27:626–38.

21. Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat. Methods [Internet]. 2016;13:435–8. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3802

22. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. [Internet]. 2016;26:1612–25. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.201863.115

23. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of ?unculturable? human microbiota reveals novel taxa and extensive sporulation. Nature [Internet]. 2016;533:543–6. Available from: http://www.nature.com/doifinder/10.1038/nature17645

24. Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. Nat. Microbiol. [Internet]. 2016;1:16203. Available from: http://www.nature.com/articles/nmicrobiol2016203

25. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26. Pop M. Genome assembly reborn: Recent computational challenges. Brief. Bioinform. 2009;10:354–66.

27. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. [Internet]. 2012;13:R122. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-12-r122

28. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.

29. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics [Internet]. 2015;31:1674–6. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv033

30. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. [Internet]. 2017; Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.213959.116

31. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front. Plant Sci. [Internet]. 2014 [cited 2017 Jan 9];5. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2014.00209/abstract

32. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature [Internet]. 2010;464:59–65. Available from: http://www.nature.com/doifinder/10.1038/nature08821

33. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods [Internet]. 2013;10:1196–9. Available from: http://www.nature.com/doifinder/10.1038/nmeth.2693

34. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat. Methods [Internet]. 2015;12:902–3.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Available from: http://www.nature.com/doifinder/10.1038/nmeth.3589

35. Wang J, Jia H. Metagenome-wide association studies: fine-mining the microbiome. Nat. Rev. Microbiol. [Internet]. 2016;14:508–22. Available from: http://www.nature.com/doifinder/10.1038/nrmicro.2016.83

36. Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. Biometrika [Internet]. 2011;98:199–214. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asq075

37. Karlsson FH, Tremaroli V, Nookaew I, Bergstr?m G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature [Internet]. 2013;498:99–103. Available from: http://www.nature.com/doifinder/10.1038/nature12198

38. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. [Internet]. 2014 [cited 2014 Jul 9];32:822–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24997787

39. Huson LW. Performance of Some Correlation Coefficients When Applied to Zero-Clustered Data. J. Mod. Appl. Stat. Methods [Internet]. 2007;6:530–6. Available from: http://digitalcommons.wayne.edu/jmasm%5Cnhttp://digitalcommons.wayne.edu/jmasm/vol6/iss2/17

40. Kowalski CJ. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. Appl. Stat. [Internet]. 1972;21:1. Available from: http://www.jstor.org/stable/10.2307/2346598?origin=crossref

41. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Pract. Assessment, Res. Eval. 2004;9:1–8.

42. Almeida M, Pop M, Le Chatelier E, Prifti E, Pons N, Ghozlane A, et al. Capturing the most wanted taxa through cross-sample correlations. ISME J. [Internet]. 2016;10:2459–67. Available from: http://www.nature.com/doifinder/10.1038/ismej.2016.35

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

43. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr. Opin. Genet. Dev. 2005. p. 589–94.

44. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature [Internet]. 2011;474:666–666. Available from: http://www.nature.com/doifinder/10.1038/nature10187

45. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nat. Methods [Internet]. 2013;10:881–4. Available from: http://www.nature.com/doifinder/10.1038/nmeth.2575

46. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nat. Rev. Microbiol. [Internet]. 2010;8:207–17. Available from: http://www.nature.com/doifinder/10.1038/nrmicro2298

47. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae* , an urgent threat to public health. Proc. Natl. Acad. Sci. [Internet]. 2015;112:E3574–81. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1501049112

48. Dehoux P, Marvaud JC, Abouelleil A, Earl AM, Lambert T, Dauga C. Comparative genomics of Clostridium bolteae and Clostridium clostridioforme reveals species-specific genomic properties and numerous putative antibiotic resistance determinants. BMC Genomics [Internet]. 2016;17:819. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3152-x

49. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, et al. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. Genome Biol. [Internet]. 2010;11:R107. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r107

50. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. Casades?s J, editor. PLoS Genet. [Internet]. 2009;5:e1000344. Available from:

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

http://dx.plos.org/10.1371/journal.pgen.1000344

51. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. 2012;22:299–306.

52. Finegold SM, Song Y, Liu C, Hecht DW, Summanen P, Könönen E, et al. Clostridium clostridioforme: A mixture of three clinically important species. Eur. J. Clin. Microbiol. Infect. Dis. 2005;24:319–24.

53. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, et al. Evolution of Symbiotic Bacteria in the Distal Human Intestine. PLoS Biol. 2007;5:1574–86.

54. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc. Natl. Acad. Sci. [Internet]. 2008;105:10039–44. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0800679105

55. Silverman BW. Using Kernel Density Estimates to Investigate Multimodality. J. R. Stat. Soc. Ser. B. 1981;43:97–9.

56. Maronna RA, Martin RD, Yohai VJ. Robust Statistics: Theory and Methods. Ann. Stat. 2006;30:17–23.

57. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 2013;49:764–6.

58. Bland JM, Altman DG. Statistics Notes: Transforming data. Bmj [Internet]. 1996;312:770–770. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.312.7033.770

59. 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

60. Lin LI-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[Internet]. 1989;45:255. Available from: http://www.jstor.org/stable/2532051?origin=crossref

61. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. Supporting data for the paper: "An integrated catalog of reference genes in the human gut microbiome" [Internet]. GigaScience Database; 2014. Available from: http://gigadb.org/dataset/100064

62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. [Internet]. 1990;215:403–10. Available from: http://www.sciencedirect.com/science/article/pii/S0022283605803602

63. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 1999. p. 29–34.

64. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35.

65. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol. Biol. Evol. [Internet]. 2017;278:631–7. Available from: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/mbe/PAP/10.1093_molbev_msx148/3/msx148.pdf?Expires=1499795422&Signature=W7WZu1nfufNHcrf1QFCH~13BjwFo82LccucPBoCvqGvLaClaQFIzzY6iT3cfR4A1bFPxYnhBaz6D-wBoorppxD5SfKpgd63KRkl2HHDcb2BEBArfT4f

66. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93.

67. Eddy SR. Accelerated profile HMM searches. PLoS Comput. Biol. 2011;7.

68. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. PLoS One. 2012;7.

69. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, et al. A Catalog of Reference Genomes from the Human Microbiome. Science (80-.). [Internet].

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

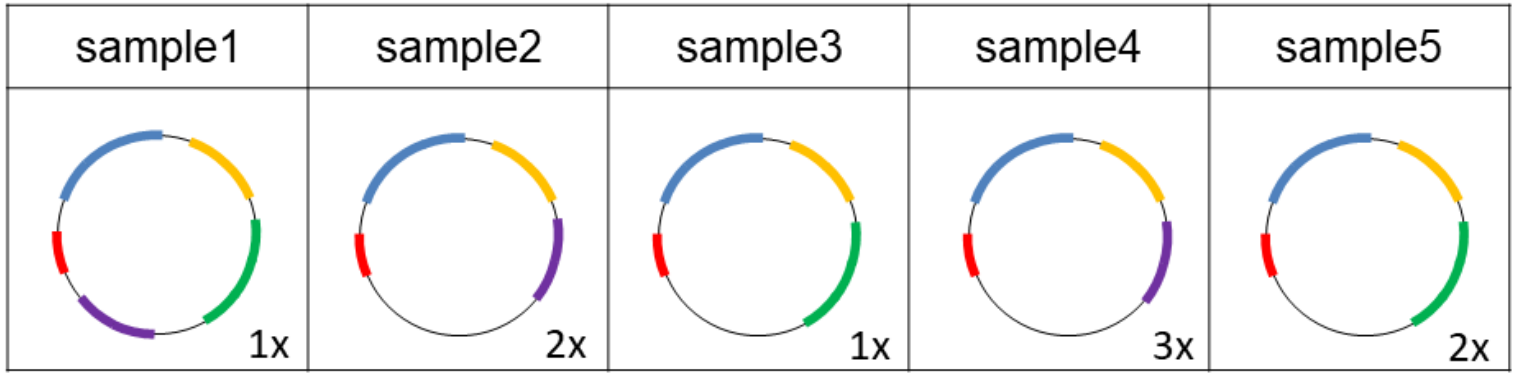
2010;328:994–9. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1183605

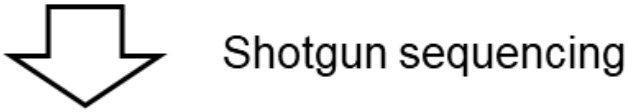
70. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41.

71. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

72. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. [Internet]. 2009;19:1639–45. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.092759.109

904 **Figures**





		sample1	sample2	sample3	sample4	sample5	
core gene 1		1	2	1	3	2	Co-abundant genes
core gene 2		3	6	3	9	6	
core gene 3		2	4	2	6	4	
accessory gene 1		3	**0**	3	**0**	6	**Partially** co-abundant with core genes
accessory gene2		**0**	4	2	6	**0**	

905

906 Figure 1

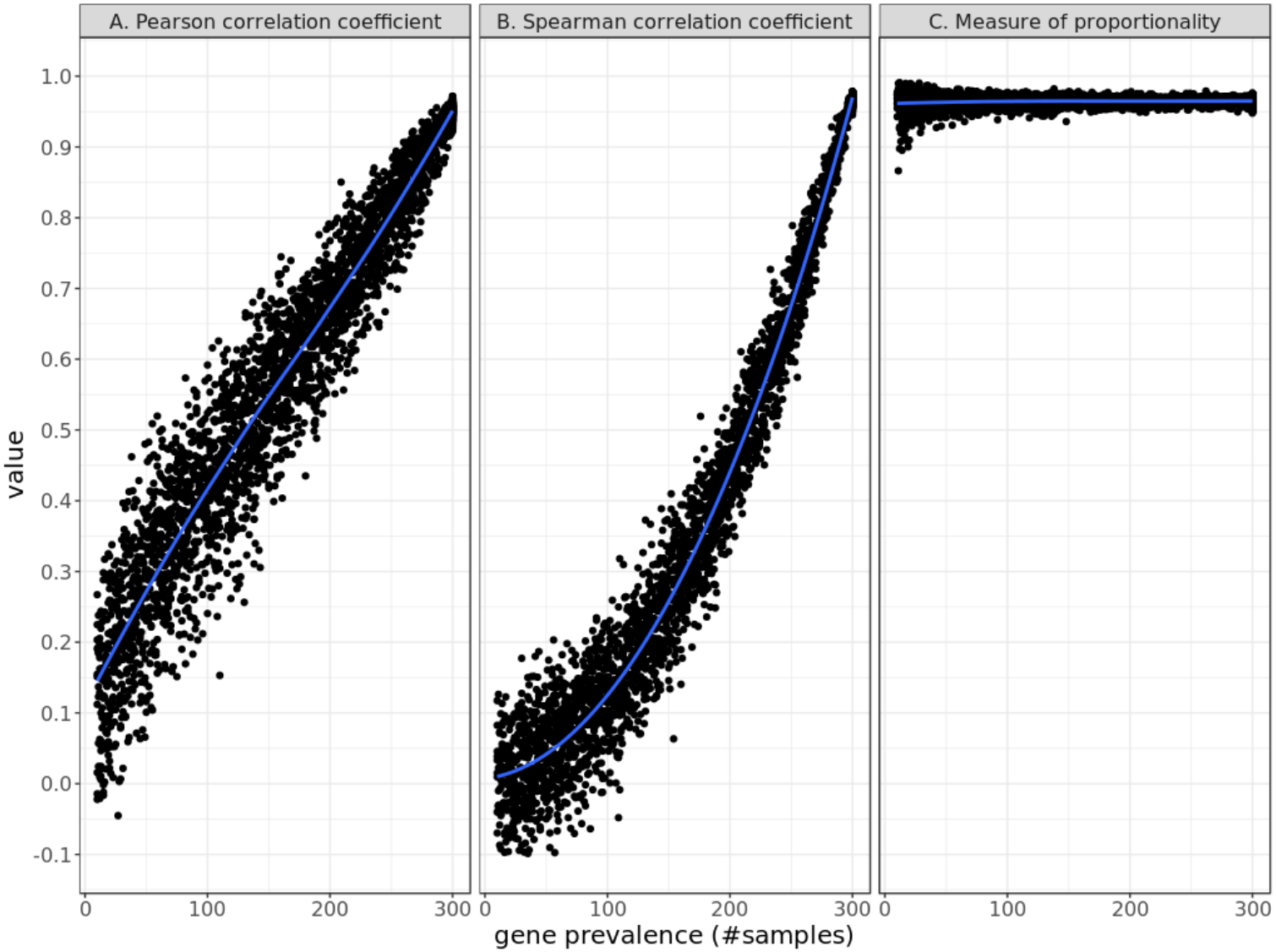

A. Pearson correlation coefficient
B. Spearman correlation coefficient
C. Measure of proportionality
value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0
100
200
300
gene prevalence (#samples)

Figure 2

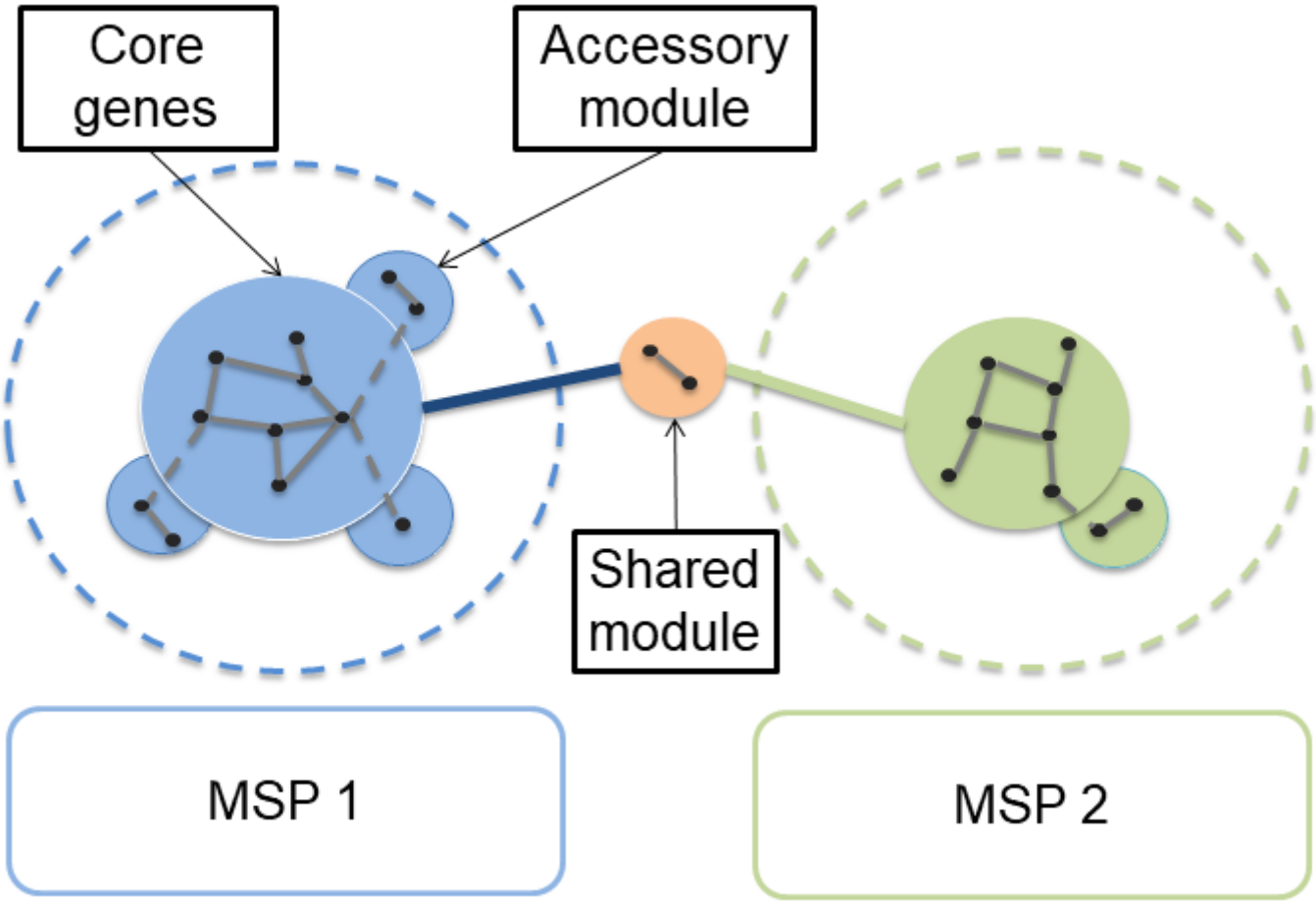


Figure 3

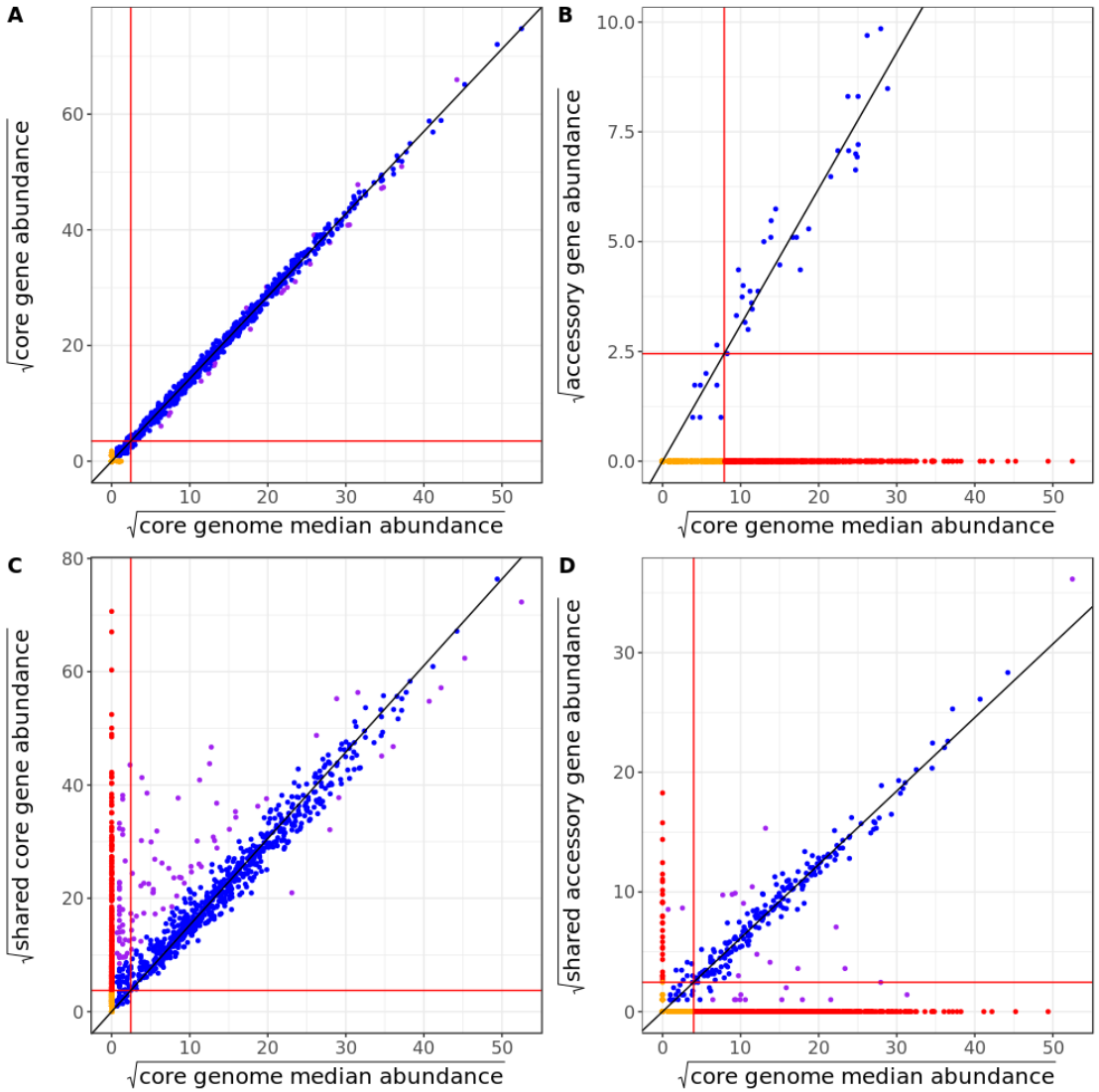

A
B
C
D
√core gene abundance
√accessory gene abundance
√shared core gene abundance
√shared accessory gene abundance
√core genome median abundance

Figure 4

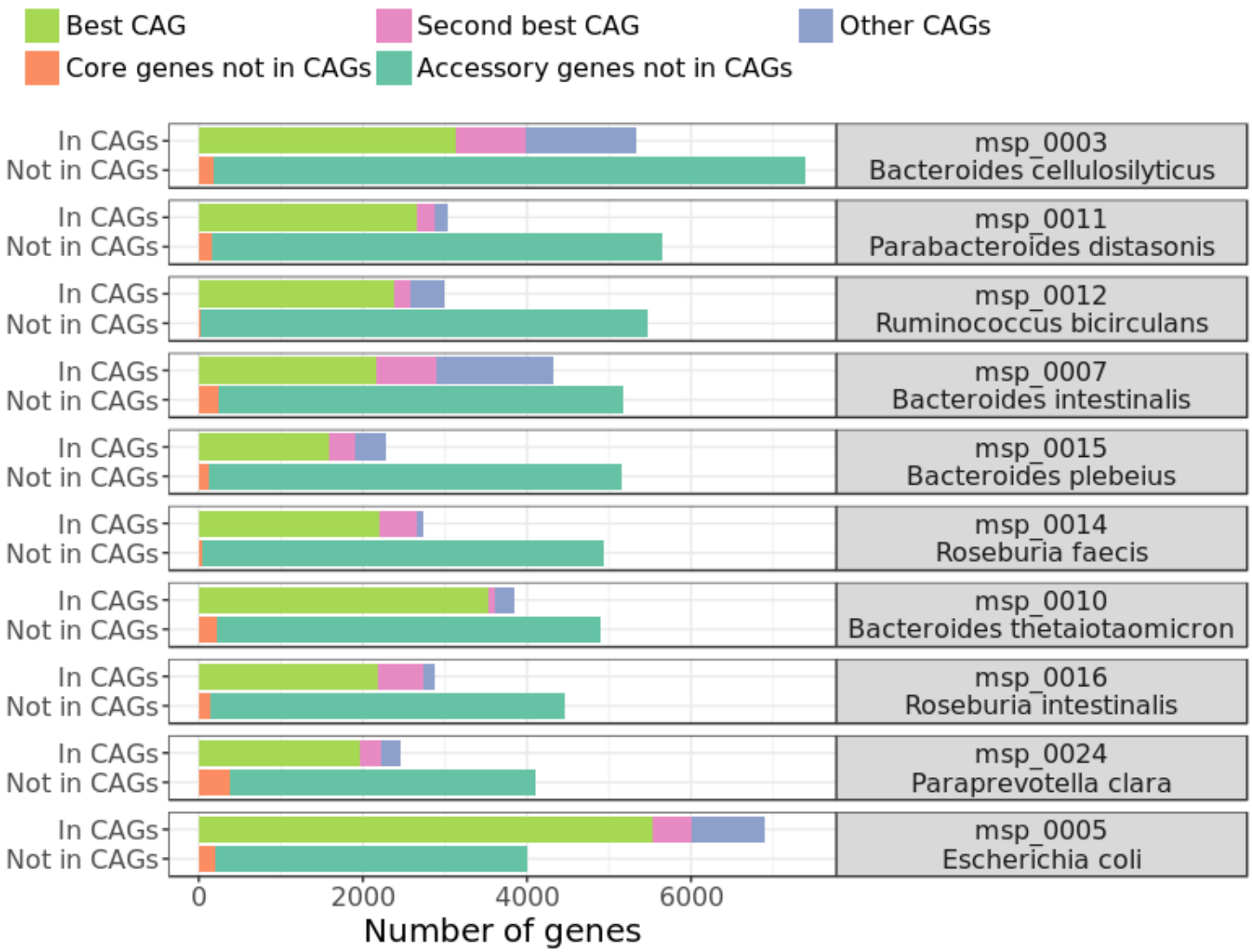

Best CAG
Second best CAG
Other CAGs
Core genes not in CAGs
Accessory genes not in CAGs
In CAGs
Not in CAGs
msp_0003
Bacteroides cellulosilyticus
msp_0011
Parabacteroides distasonis
msp_0012
Ruminococcus bicirculans
msp_0007
Bacteroides intestinalis
msp_0015
Bacteroides plebeius
msp_0014
Roseburia faecis
msp_0010
Bacteroides thetaiotaomicron
msp_0016
Roseburia intestinalis
msp_0024
Paraprevotella clara
msp_0005
Escherichia coli
0
2000
4000
6000
Number of genes

Figure 5

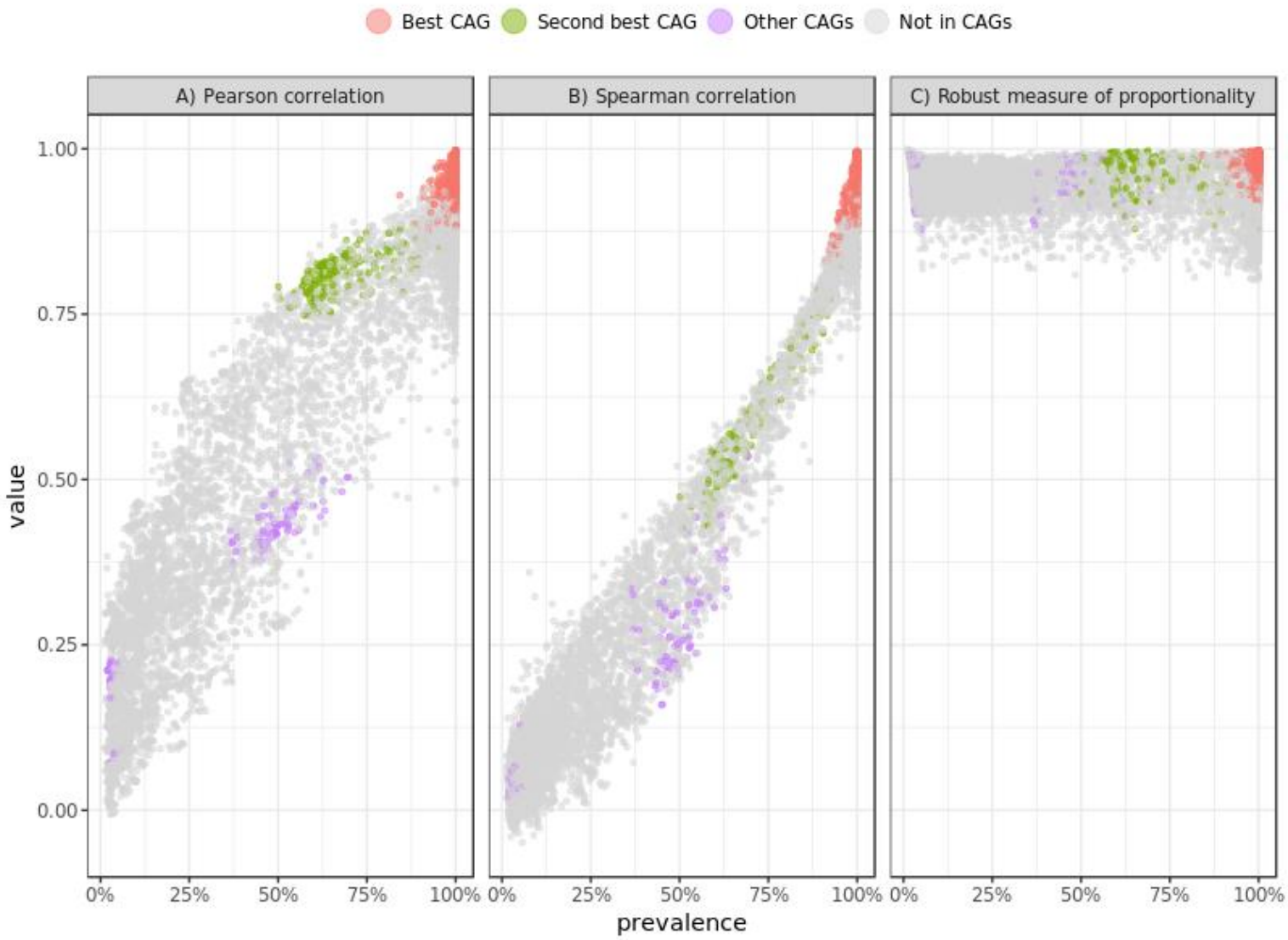

Best CAG
Second best CAG
Other CAGs
Not in CAGs
A) Pearson correlation
B) Spearman correlation
C) Robust measure of proportionality
value
1.00
0.75
0.50
0.25
0.00
0%
25%
50%
75%
100%
prevalence

Figure 6

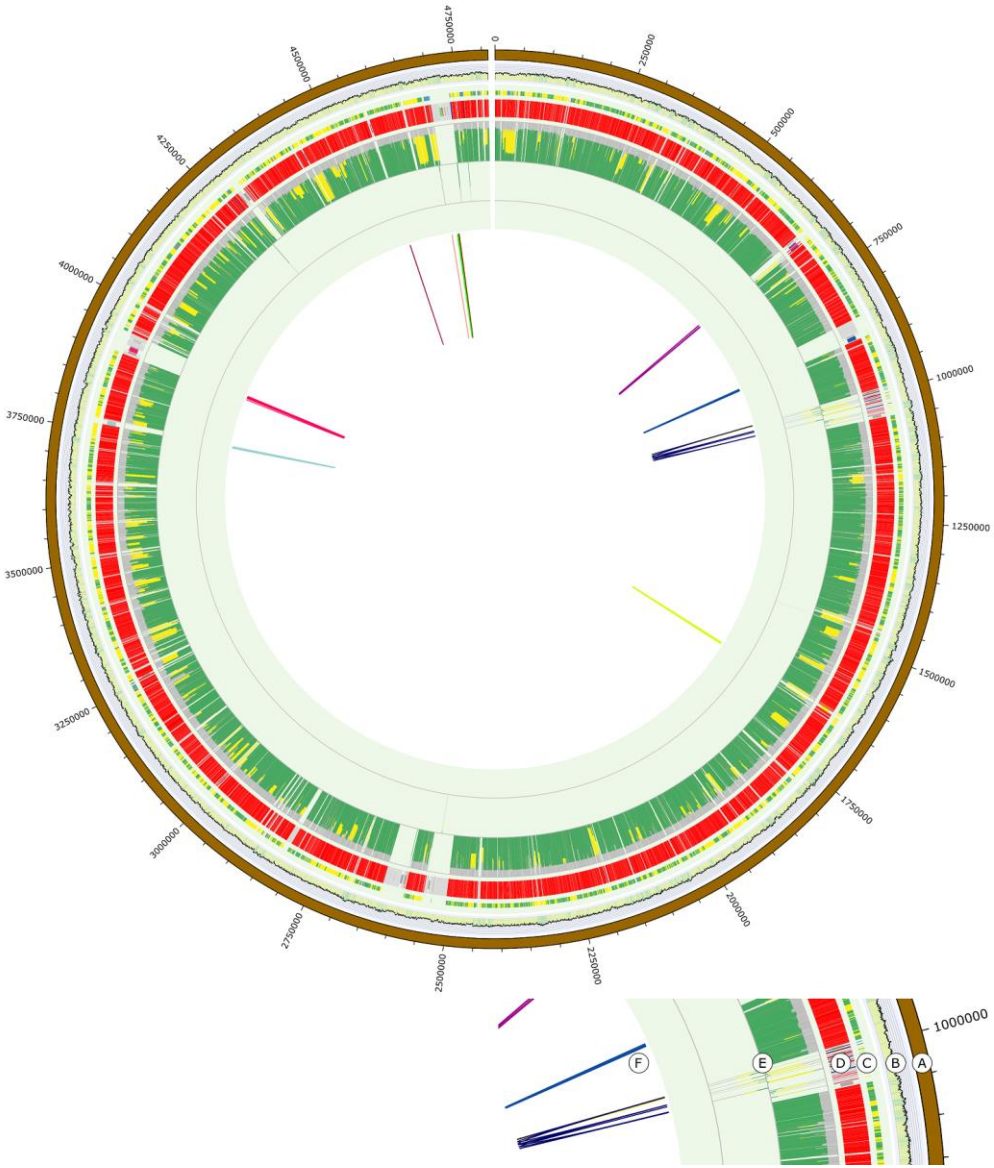

0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
2500000
2750000
3000000
3250000
3500000
3750000
4000000
4250000
4500000
4750000
1000000
A
B
C
D
E
F

Figure 7

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/173203; this version posted October 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- 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

F. relation between genes associated to a MSP different from the most represented (format: edges). Genes from the same alien MSP are linked by edges.