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@enterome.com
MP: matthieu.pichaud@gmail.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 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.

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

140 number of mapped reads across samples (co-abundance) and should be consistently observed
141 in samples if sequencing depth allows (co-occurrence). Remarkably, core genes and an
142 accessory gene should have directly proportional counts only in the subset of samples where
143 they are both detected (Figure 1). To group the core genes of a species and then identify its
144 accessory genes, we developed a measure that evaluates proportionality between gene
145 counts using samples where the number of mapped reads is high enough (see Methods).
146
147 To evaluate this new measure of proportionality, we generated an abundance table that
148 simulates the counts of genes from a single virtual species across 300 samples (see Methods).
149 We considered that each sample was carrying a unique strain of the species with specific gene
150 content. Genes present in all the samples were labeled as core and those detected in a subset
151 as accessory. We used this dataset to compare the performance of the Pearson correlation
152 coefficient, the Spearman correlation coefficient and the proposed measure of proportionality
153 for detecting a relation between the abundance profile of the species core genome and all its
154 genes including accessories (Figure 2). Pearson and Spearman correlation coefficients
155 decrease all the more as the prevalence of a tested gene decreases while the proposed
156 measure remains high, as only samples where both the species core and its accessory gene
157 are detected are used for calculation. Therefore, the association between core genes and
158 many accessory genes will be missed using the correlation coefficients. However, accessory
159 genes observed in similar subsets of samples may be grouped into small distinct clusters as
160 their abundance profiles should be correlated.
161
162 Finally, we derived a robust version of the measure to identify associated genes despite the
163 presence of samples with inconsistent counts named hereafter outliers. We evaluated this
164 robust measure of proportionality against the non-robust version described above by adding
165 an increasing percentage of outliers to the genes abundance profiles. For a given percentage
166 of outliers, each of these genes was compared to the outlier-free abundance profile of the
167 species core genome. This simulation showed that the non-robust measure of proportionality
168 decreases all the more as the percentage of outliers increases whereas the robust measure
169 remains high; demonstrating that proportionality is still detected (Figure 3). However, the
170 robust measure decreases significantly when the percentage of outliers is high and the gene
171 prevalence is low.

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

173 We developed MSPminer, a program that uses measures of proportionality to group co-
174 abundant genes into Metagenomic Species Pan-genomes (MSPs, Figure 4). MSPminer
175 empirically distinguishes core from accessory genes based on their presence absence patterns
176 (see Methods) and tags genes observed in samples where the core is not detected as shared
177 (Figure 5).

178

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

187 *Taxonomy*

188 By considering the lowest taxonomic rank assigned to MSPs, 278 (16.6%) were annotated at
189 species level, 85 (5%) at genus level, 119 (7.1%) at phylum level and the remaining 1 213 MSPs
190 (71.5%) could not be annotated (Supplementary Figure 1) indicating that a clear majority of
191 MSPs correspond to species not represented in reference genomes databases. Only 4 MSPs
192 were annotated as Eukaryotes and corresponded to intestinal parasites of the *Blastocystis*
193 genus. All others MSPs were assigned to prokaryotic species of which only 3 were Archaea and
194 corresponded to the species *Methanobrevibacter smithii*, *Candidatus Methanomethylophilus*
195 *alvus* and *Methanosphaera stadtmanae*. Among the MSPs annotated as Bacteria, the phyla
196 *Firmicutes* (268 MSPs), *Bacteroidetes* (118 MSPs), *Proteobacteria* (50 MSPs) and
197 *Actinobacteria* (29 MSPs) were the most represented (Supplementary Figure 2) as expected
198 from human gut metagenomes [43]. Some species were represented by multiple MSPs such
199 as *Faecalibacterium prausnitzii* (5 MSPs), *Bacteroides fragilis* (2 MSPs), *Methanobrevibacter*
200 *smithii* (2 MSPs) or *Hungatella hathewayi* (2 MSPs) suggesting a high nucleotide and gene
201 content variability between strains attributed so far to the same species. Conversely, some
202 MSPs had their core genes attributed to multiple species, usually from the same genus

203 (Supplementary Table 3). Although some inconsistencies in the taxonomy assignment have
204 previously been reported [22,44], it is possible that several MSPs regroup genes from highly
205 related species with high average nucleotide identity and similar gene content. MSPs
206 annotated at species level had a consistent taxonomical annotation as 97% of the core genes
207 (median) were assigned to the same species (Supplementary Figure 3). Consistency of the
208 taxonomic annotation was lower for accessory genes (62%) caused mainly by unannotated
209 genes.

210 *MSP content*

211 Most MSPs were small (median number of genes = 1 784) even if 53 had more than 5 000
212 genes (Supplementary Figure 4). As expected, a strong positive correlation (Pearson's r = 0.8)
213 between the total number of genes in a MSP and its number of accessory genes was observed
214 (Supplementary Figure 5). Interestingly, four outliers corresponding to the intestinal parasites
215 previously described had a high number of core genes and few accessory genes. This suggests
216 that Eukaryotic genomes have a larger number of genes and a lower gene content variability
217 than Prokaryotes. Among the MSPs with the most accessory genes (Supplementary Table 2),
218 many corresponded to species reported as highly variable such as *Escherichia coli* [45],
219 *Klebsiella pneumoniae* [46] or *Clostridium bolteae* [47]. As previously observed in population
220 genomics studies comparing multiple strains of the same species [48,49], the prevalence of
221 accessory genes in MSPs often follows a bimodal distribution (Supplementary Figure 6)
222 showing either a high or low prevalence but rarely intermediate. Thus, the number of
223 accessory genes in a MSP is correlated (Spearman's rho = 0.86) with its prevalence
224 (Supplementary Figure 7). Indeed, the more a MSP is detected in many samples, the more
225 exhaustively MSPminer will recover its accessory genes, especially the rare ones. Many MSPs
226 annotated at species level had accessory genes previously unobserved in available genomes
227 (Supplementary Table 2)

228 *Prevalence*

229 As for the genes in the catalog, most MSPs were detected in very few samples (Supplementary
230 Figure 8). Only 40 MSPs were detected in at least 70% of the samples showing that the
231 common microbial core of the human gut microbiota is limited to a few dozen species
232 (Supplementary Table 2). No clear relation between the prevalence of the MSPs and their

233 mean abundance was found (Supplementary Figure 9). However, 2 MSPs corresponding to
234 *Bacteroides vulgatus* and *Bacteroides uniformis* were both very prevalent (detected in 97.6%
235 and 94.6% of the samples respectively) and very abundant (mean relative abundance of 7.8%
236 and 4.4% respectively). Interestingly, many rare MSPs were abundant in the few samples
237 which carried them. Many of these MSPs were annotated as bacteria of the *Lactobacillus*
238 genus most likely consumed as probiotics. However, some others correspond to known
239 invasive species associated with severe dysbiosis such as *Fusobacterium nucleatum* [50] or
240 *Clostridium clostridioforme* [51].

241 *Census of universal single copy marker genes*

242 To check that MSPs correspond to real microbial species and evaluate the completeness of
243 their set of core genes, we identified in each of them 40 universal single copy marker genes
244 (SCM) [44]. 878 MSPs (54%) had at least 30 SCM and 403 (24%) had all of them (Supplementary
245 Figure 10 A). As housekeeping genes, SCMs are essential to the microbe survival and should
246 be found among core genes. Indeed, 92% of the SCMs were core genes in their respective MSP
247 and the rest was mainly high prevalent accessory genes (Supplementary Figure 10 B). This
248 shows the classification of genes as core or accessory performed by MSPminer is reliable.

249 *Comparison to sequenced genomes*

250 We compared the MSPs to 642 sequenced genomes for which at least 10% of their constituent
251 genes were detected in the Integrated Genes Catalog of the human gut microbiome [12]
252 (Supplementary Table 4). In total, these genomes covered 398 species representing 114
253 different genera. 624 (97.1%) were unambiguously assigned to 281 different MSPs. By keeping
254 only one representative per species, 47.1% (resp. 60.4%) of the genomes had at least 75% of
255 their genes grouped in their corresponding MSP considering either all their genes or only those
256 that were in the catalog (Supplementary Figure 11). In compliance with the results of the
257 taxonomic analysis, highly related species were assigned to the same MSP such as *Escherichia*
258 *coli*, *Escherichia fergusonii* and all genomes of the *Shigella* genus. Conversely, some species
259 grouping highly divergent strains were represented by several MSPs including
260 *Faecalibacterium prausnitzii* or *Bacteroides fragilis.*

261 **Comparison to the Canopy clustering algorithm**

262 The Canopy clustering algorithm [38] was compared to MSPminer by applying both tools to
263 the metagenomic dataset described above. In total, Canopy grouped 2 691 408 genes into 3
264 463 Co-Abundance gene Groups (CAGs) while MSPminer grouped 3 267 132 genes (+17.6%)
265 into 1 677 MSPs (~ two-fold less objects).
266
267 178 MSPs encompassing 154 617 genes had no equivalent among the CAGs. Most MSPs were
268 rare as 75% were detected in fewer than 5 samples but had a significant size as 50% were
269 composed of at least 700 genes (Supplementary Figure 12). Remarkably, for 75% of them, the
270 3 samples with the highest counts represented at least 90% of the sum of MSP abundance on
271 all the samples. By default, Canopy discards such cases to avoid detection of spurious
272 correlations but MSPminer limits this risk by applying a variance-stabilizing transformation
273 and a stringent association criterion. In addition, Canopy grouped most core genes of a MSP
274 into a single CAG while many accessory genes were missed or assigned to small separate CAGs
275 (Figure 6). In agreement with the results of the simulation, most of the missed accessory genes
276 had a medium or low prevalence. As they contained many unexpected zeros, the correlations
277 with the core of their respective species were below the limit set in Canopy (Figure 7).

278 **MSPs quantification for biomarkers discovery**

279 To demonstrate that MSPminer was useful for biomarkers discovery, we first looked for
280 differentially abundant MSPs according to the geographical origin of samples. We discovered
281 94 MSPs differentially abundant between Westerners and Chinese (q-value < 10^{-3}, log2 fold
282 change ≥ 1) including 72 more abundant in Westerners and 22 in Chinese (Supplementary
283 Figure 13 and Supplementary Table 5). Among the discriminant MSPs, all those assigned to
284 the *Proteobacteria* phylum (*Klebsiella pneumoniae*, *Escherichia coli*, *Bilophila wadsworthia*)
285 were more abundant in Chinese which is consistent with previously published results [12].
286 Interestingly, two MSPs assigned to *Faecalibacterium prausnitzii* were significant but one was
287 more abundant in Westerners and the other in Chinese. This shows that some strains of this
288 species are associated with geographical origin of samples. In addition, we discovered 75 MSPs
289 differentially abundant between Europeans and Americans (q-value < 10^{-3}, log2 fold change ≥
290 1) of which 70 were more abundant among Europeans. (Supplementary Figure 14 and

291 Supplementary Table 5). This result is consistent with previous studies showing lower gut
292 microbiota diversity among Americans compared to Europeans [33].
293
294 Secondly, we checked if some accessory genes in the MSPs were more prevalent in samples
295 of a given geographical origin. By way of example, we found in 297 accessory genes associated
296 with geographical origin (chi-squared test, p-value < 10^{-10}) in the msp_0039 corresponding to
297 *Prevotella copri* (Supplementary Figure 15 and Supplementary Table 6). Remarkably, 98% of
298 the significant genes were more prevalent among Westerners or Chinese although European
299 and American samples were separated in the analysis. This result indicates that strains of
300 *Prevotella copri* carried by Chinese are distinct from those of Westerners.

301 Discussion

302 Direct proportionality hypothesis and limits

303 MSPminer relies on a new robust measure to detect genes with directly proportional counts.
304 Even if this relation is more stringent than those assessed by Pearson's or Spearman's
305 correlation coefficients, it was successfully used to reconstitute Metagenomic Species Pan-
306 genomes of the human gut microbiota. In fact, most genes from sequenced genomes were
307 grouped into a single MSP showing that direct proportionality is the most common relation
308 between genes from the same biological entity. MSPminer misses some genes for which
309 counts are not ruled by this relation. Indeed, proportionality is disrupted when gene copy
310 number varies across samples [14] (Supplementary Figure 16), when a sample contains
311 multiples strains [19,20] and when a gene is shared between several MSPs (Supplementary
312 Figure 17) because of horizontal gene transfer [52] or grouping of highly similar orthologs.
313 Nevertheless, the first two cases have most likely a limited impact as the majority of strains
314 tend to have the same gene copy numbers [14] and samples often carry a dominant strain
315 [20]. Regarding shared genes, their signals are a linear combination of the MSPs that carry
316 them. Thus, they will be identified only if these MSPs are mostly detected in separate sets of
317 samples.

318 Computing performance

319 MSPminer can process large datasets made up of thousands of samples and millions of genes
320 in just a few hours on a regular single node server. The program is only limited by the quantity

321 of RAM available on the machine on which it is executed as the input count matrix must be
322 fully loaded into memory. Consequently, RAM consumption grows linearly according to the
323 number of samples and genes in the count matrix (Supplementary Figure 18).
324 MSPminer achieves good parallel efficiency (Supplementary Figure 19) through two
325 parallelization strategies. First, a novel Map/Reduce programming model assigns genes to as
326 many subsets as the number of available samples. Genes with greatest counts in the same
327 sample are first compared, which not only decreases the number of comparisons to perform
328 but increases the probability that related genes are placed in the same bin compared to
329 random assignment (Supplementary Figure 20). Each subset of genes is processed in parallel
330 and synchronization is only required before the reduction/merging step (see Methods). To
331 avoid comparisons of all pairs of genes, others used an iterative algorithm where a random
332 gene is compared to all the others until a significant proportion of genes is clustered [38]. This
333 method allows fast identification of big clusters but struggles with the large number of small
334 clusters and singleton genes. Furthermore, parallelism implementation is more complex as
335 synchronization is required to detect duplicate clusters. In a second phase, MSPminer
336 performs pairwise comparison of clusters to detect those corresponding to MSPs core gene
337 sets. Then, full MSPs including accessory and shared genes are retrieved in parallel from the
338 signal of their respective core genes. Here, no synchronization is required as core genes sets
339 are supposedly independent.

340 **Quality of MSPs**

341 The quality of the MSPs is impacted by all the upstream steps required for generating the
342 count matrix, as well as with the biological and ecological characteristics of the dataset. At the
343 sequencing level, the number of reads (sequencing depth) generated for each sample impacts
344 the detection and coverage of subdominant species, while reads length affects the quality of
345 the assembly and the ability to assign a read to a gene without ambiguity. At the
346 bioinformatics level, assembly, gene prediction, gene redundancy removal, mapping and
347 counting require expertise to select the most appropriate strategies, tools and parameters.
348 Indeed, assemblers returning chimeric contigs which combine sequences from highly related
349 species, inaccurate predictors generating truncated or merged genes, redundancy removal
350 with a common threshold for all genes (95% of nucleotide identity) lead to genes of variable
351 quality in catalogues. Then, when quantifying genes, keeping only uniquely mapped reads

352 underestimates the abundance of some genes whereas considering shared reads can generate
353 false positives. Finally, at the biology level, a high number of samples with varied phenotypes
354 will improve the comprehensiveness and quality of MSPs. Indeed, as the number of samples
355 grows, MSPminer will identify rare species and will extend the list of accessory genes of the
356 MSPs corresponding to species with an open pan-genome. In addition, highly prevalent
357 accessory genes will be reclassified from core to accessory as observed while sequencing an
358 increasing number of strains of a species [49].

359 Applications

360 As illustrated in this paper, MSPminer supports the analysis of metagenomic data at species-
361 level by identifying and quantifying the MSPs present in samples. Subsequently, MSPs
362 associated with a given phenotype (e.g. the geographic origin) can be investigated both
363 quantitatively and qualitatively. Here, information from unknown or non-sequenced species
364 can be exploited. Compared to methods relying on marker genes [33,34], MSPminer improves
365 the estimation of species abundance by automatically detecting among core genes those with
366 the highest specificity, the highest counts and lowest dispersion.
367 Moreover, in each MSP, genes or modules of accessory genes associated with the tested
368 phenotype can be explored opening the way to a strain-level analysis. Thus, biomarkers
369 corresponding to functional traits specific to certain strains can be discovered.
370
371 MSPminer also provides microbial population genetics from large cohorts which can support
372 culture-dependent methods by identifying species of particular interest, such as those with no
373 reference genome available or with reference genomes distant from the strains actually
374 present. Reciprocally, MSPminer will benefit from advances in culture-dependent methods
375 which provide reference genomes of low abundance species detectable by shotgun
376 sequencing but difficult to assemble [23,24].

377 Further developments

378 Several improvements of MSPminer are considered. Algorithms that could identify relevant
379 associations currently missed by MSPminer will be evaluated. For instance, deconvolution
380 algorithms [14] may discover genes shared between several MSPs while kernel density
381 estimators [53] 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 [54] 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 [55].

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 21). Therefore, instead of relying on genes length, α was robustly estimated as follow:

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

A square root transformation was applied to stabilize variance as suggested by several authors for count data [56,57]. For a comparison of some data transformations, refer to Supplementary Figure 22. 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 5) whereas those above were classified as structural (red points in Figure 5).

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:

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

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

Non-robust version

The relationship of proportionality between two genes was evaluated by the concordance correlation coefficient [58] computed on samples that can be used for quantification.

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

$$\frac{2 \cdot \alpha \cdot s_{xy}}{\alpha^2 \cdot s_x + (\alpha \cdot \overline{x} - y^2)}$$

where $x = \sqrt{g_x}$, $y = \sqrt{g_y}$, $\overline{x}$ 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:

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

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 5).

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 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.

486 The theoretical number of counts attributed to each gene was calculated according to its
487 length and its presence or not in the strain. The observed gene counts were generated by
488 using a negative binomial distribution (mean=theoretical gene count; overdispersion=0.05) to
489 approach real metagenomic data.
490 Outliers were added in each gene by selecting 10%, 20% and 30% of their non-null samples
491 and multiplying each observed count by either ¼, ⅓, 2 ,3 or 4.

492 **Metagenomic Species Pan-genome generation**

493 The Supplementary Figure 23 gives an overview of the MSPminer workflow.

494 *Input data*

495 MSPminer processes raw gene counts tables. In this study, we used the table of the integrated
496 catalog of reference genes in the human gut microbiome available at the following URL:
497 ftp://climb.genomics.cn/pub/10.5524/100001_101000/100064/2.ProfileTables/1267sample.
498 gene.pairNum.table.gz

499 *Data filtering*

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

502 *Data transformation*

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

505 *Genes binning*

506 Genes with the highest counts in the same sample were binned. To limit bias due to variable
507 sequencing depth, raw read counts were normalized by the number of mapped reads prior to
508 bin assignment (Supplementary Figure 24). Note that normalized counts were used in this step
509 only.

510 *Seeds creation*

511 This step identifies sets of co-abundant and co-occurring genes called *seeds* hereafter.
512 Seeds were created in parallel in each bin by a greedy approach. First, genes were compared
513 pairwise. All pairs of genes with a non-robust measure of proportionality of at least 0.8 and

514 no structural zeros were saved in a list. Then, the list was sorted by decreasing measure of
515 proportionality. The pair of genes with the highest measure of proportionality was selected as
516 a centroid. Genes related to one of the centroid genes were grouped together in a new seed.

517 *Seed representative*

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

523 *Seeds merging*

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

532 *Core seeds identification*

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

541 *Metagenomic Species Pan-genome generation*

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

546

547 Let g_x be the median vector of the number of mapped reads on the core seed and g_y the
548 vector of the number of mapped reads on a gene related to the core seed.

549 The related gene was assigned to one of the 4 following categories:

550 1. core genes were detected in the same samples as the core seed (Figure 5.A)

551 $$(g_y \infty g_x)\ and\ (\forall s \in [1,m] c_{xs} \geq t_x \rightarrow\ c_{ys} \neq 0\ and\ c_{ys} \geq t_y \rightarrow\ c_{xs} \neq 0)$$

552 2. accessory genes were detected in a subset of samples where the core seed was
553 detected (Figure 5.B)

554 $$(g_y \infty g_x)\ and\ (\exists s \in [1,m] c_{xs} \geq t_x and\ c_{ys} = 0)\ and\ (\forall s \in [1,m]\ c_{ys} \geq t_y \rightarrow\ c_{xs} \neq 0)$$

555 3. shared core genes were detected in all the samples where the core seed was detected
556 plus some samples where it was not (Figure 5.C).

557 $$(g_y \infty g_x)\ and\ (\forall s \in [1,m]\ c_{xs} \geq t_x \rightarrow\ c_{ys} \neq 0\ )\ and\ (\exists s \in [1,m] c_{ys} \geq t_y and\ c_{xs} = 0)$$

558 4. shared accessory genes were detected in a subset of samples where the core seed was
559 detected plus some samples where it was not (Figure 5.D).

560 $$(g_y \infty g_x)\ and\ (\exists s \in [1,m] c_{xs} \geq t_x and\ c_{ys} = 0)\ and\ (\exists s \in [1,m] c_{ys} \geq t_y and\ c_{xs} = 0)$$

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

564 Comparison to the Canopy clustering algorithm

565 Raw gene counts table was used as input for MSPminer.

566 The implementation of the Canopy clustering algorithm was downloaded at
567 https://www.cbs.dtu.dk/projects/CAG/Supplementary_Software_canopy_clustering.zip. The
568 gene count table normalized following the procedure described in [3] was taken as an input.
569 Default parameters were used.

570 MSPs were projected on Co-Abundance gene Groups (CAGs) and reciprocally with an in-house
571 script.

572 **Biomarkers discovery**

573 *Identification of MSPs associated with geographical origin*

574 A two-tailed Wilcoxon rank-sum test was used on relative median abundance of the 30 best
575 representative core genes of each MSP (1 696 tested variables). The obtained p-values were
576 adjusted by the Benjamini-Hochberg procedure. In addition, a $\log_2$ ratio was computed
577 between the median abundances of the MSP in the two populations tested. MSPs with an
578 adjusted p-value inferior to 10^{-2} and a $\log_2$ ratio superior to 1 were considered significant.

579 *Identification of accessory genes associated with geographical origin*

580 For each accessory gene of a MSP, a 2x2 contingency table counting in both populations the
581 number of samples where the gene was present or absent was built. Only samples where MSP
582 core genome was detected were kept. A gene was considered as present in a sample if at least
583 two reads were mapped on it. Then, a chi-squared test was performed on each contingency
584 table. Accessory genes with a p-value inferior to 10^{-10} were considered significant.

585 *Taxonomic annotation*

586 Genes were aligned at the nucleotide level using BLASTn [59] (version 2.6.0) against KEGG
587 GENOME [60] (Release 82.0, April 2017) and RefSeq [61] (Release 81, March 2017). Hits that
588 covered less than 80% of the query gene or with a e-value superior to 0.01 were discarded.
589 Thresholds of 95%, 80% and 65% of nucleotide identity were respectively used for taxonomic
590 annotation at species, genus and phylum level.
591 At a given taxonomic level, "no consensus" was reported if the selected hits did not share the
592 same annotation. Finally, results from KEGG were preferred to those from RefSeq.
593
594 MSPs were assigned to the lowest level taxon representing more than 50% of the annotations
595 of their core genes. Taxonomic annotations of MSPs with at least 50% of their core genes
596 annotated at species level (including "no consensus") and less than 80% assigned to the most
597 represented species were considered ambiguous (c.f. Supplementary Table 3)

598 *Functional annotation of the gene catalog*

599 Translated genes were aligned using BLASTp [59] (version 2.6.0) against KEGG [60] (Release
600 81.0, April 2017). Hits with a bitscore inferior to 60 or with an e-value superior to 0.01 were

601 discarded. Each gene was finally assigned to the functional group (KEGG Orthology or Enzyme
602 Commission) associated with the best significant hit
603
604 The 40 universal single copy marker genes were discovered using fetchMG v1.0 [62]

605 *Comparison of sequenced genomes to MSPs*

606 Genomes used to build the integrated catalog of the human gut microbiome [12], HMP
607 reference genomes [63] and genomes from species detected while performing the taxonomic
608 annotation of the MSPs were downloaded from GenBank [64]. Genes from the reference
609 catalog were aligned against these genomes with BLASTn [59] (version 2.6.0; arguments: -
610 perc_identity 95 -ungapped). Alignments of less than 100 nucleotides were discarded. Hits
611 found for a single gene at neighboring positions on the target genome were merged. 642
612 genomes with less than 10% of their constituent genes detected in the reference catalog were
613 kept. Genes were annotated with the related MSP information when available. Hits from most
614 abundant MSP were kept and overlapping hits from less abundant MSPs were discarded. The
615 local GC-content of genome was computed using a sliding window of approximately 100
616 nucleotides. Finally these data were plotted using Circos [65].

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, prevalence, 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 MSPs associated with the geographic origin of samples.

Supplementary Table 6: XLS file listing the MSPs associated with the geographic origin of samples.

638 # References

639 1. O'Hara AM, Shanahan F. The gut flora as a forgotten organ. EMBO Rep. [Internet].
640 2006;7:688–93. Available from:
641 http://embor.embopress.org/cgi/doi/10.1038/sj.embor.7400731

642 2. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward D V, et al. Dysfunction of the
643 intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. [Internet].
644 2012;13:R79. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-
645 2012-13-9-r79

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

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

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

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

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

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

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

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

666 10. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome:
667 Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys.
668 Res. Commun. 2016;469:967–77.

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

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

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

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

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

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

688 17. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, et al. A Culture-
689 Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of
690 Shiga-Toxigenic Escherichia coli O104:H4. JAMA [Internet]. 2013;309:1502. Available from:

691 http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2013.3231

692 18. Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang Y-F. Analysis of Ultra Low Genome
693 Conservation in Clostridium difficile. Horsburgh MJ, editor. PLoS One [Internet].
694 2010;5:e15147. Available from: http://dx.plos.org/10.1371/journal.pone.0015147

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

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

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

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

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

712 24. Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously
713 uncultured members of the human gut microbiota by culturomics. Nat. Microbiol. [Internet].
714 2016;1:16203. Available from: http://www.nature.com/articles/nmicrobiol2016203

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

717 26. Pop M. Genome assembly reborn: Recent computational challenges. Brief. Bioinform.

718 2009;10:354–66.

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

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

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

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

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

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

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

742 34. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for
743 enhanced metagenomic taxonomic profiling. Nat. Methods [Internet]. 2015;12:902–3.
744 Available from: http://www.nature.com/doifinder/10.1038/nmeth.3589

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

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

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

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

759 39. Huson LW. Performance of Some Correlation Coefficients When Applied to Zero-Clustered
760 Data. J. Mod. Appl. Stat. Methods [Internet]. 2007;6:530–6. Available from:
761 http://digitalcommons.wayne.edu/jmasm%5Cnhttp://digitalcommons.wayne.edu/jmasm/vo
762 l6/iss2/17

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

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

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

770 43. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of
771 the human gut microbiome. Nature [Internet]. 2011;474:666–666. Available from:

772 http://www.nature.com/doifinder/10.1038/nature10187

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

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

779 46. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis
780 of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella*
781 *pneumoniae* , an urgent threat to public health. Proc. Natl. Acad. Sci. [Internet].
782 2015;112:E3574–81. Available from:
783 http://www.pnas.org/lookup/doi/10.1073/pnas.1501049112

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

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

793 49. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised Genome
794 Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. Casades?s J,
795 editor. PLoS Genet. [Internet]. 2009;5:e1000344. Available from:
796 http://dx.plos.org/10.1371/journal.pgen.1000344

797 50. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al.
798 Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res.
799 2012;22:299–306.

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

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

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

808 54. Maronna RA, Martin RD, Yohai VJ. Robust Statistics: Theory and Methods. Ann. Stat.
809 2006;30:17–23.

810 55. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation
811 around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol.
812 2013;49:764–6.

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

815 57. Osborne J. Notes on the use of data transformations. Pract. Assess. [Internet]. 2002;8:4–
816 13. Available from:
817 c:%5CUsers%5CTeresa%5CDocuments%5CTeseFinal%5C2010%5Cserinus2010%5CPDF'sPara
818 Organizar%5CPDF's2010%5Ctransformation.pdf

819 58. Lin LI-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics
820 [Internet]. 1989;45:255. Available from:
821 http://www.jstor.org/stable/2532051?origin=crossref

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

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

827 61. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-
828 redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res.
829 2007;35.

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

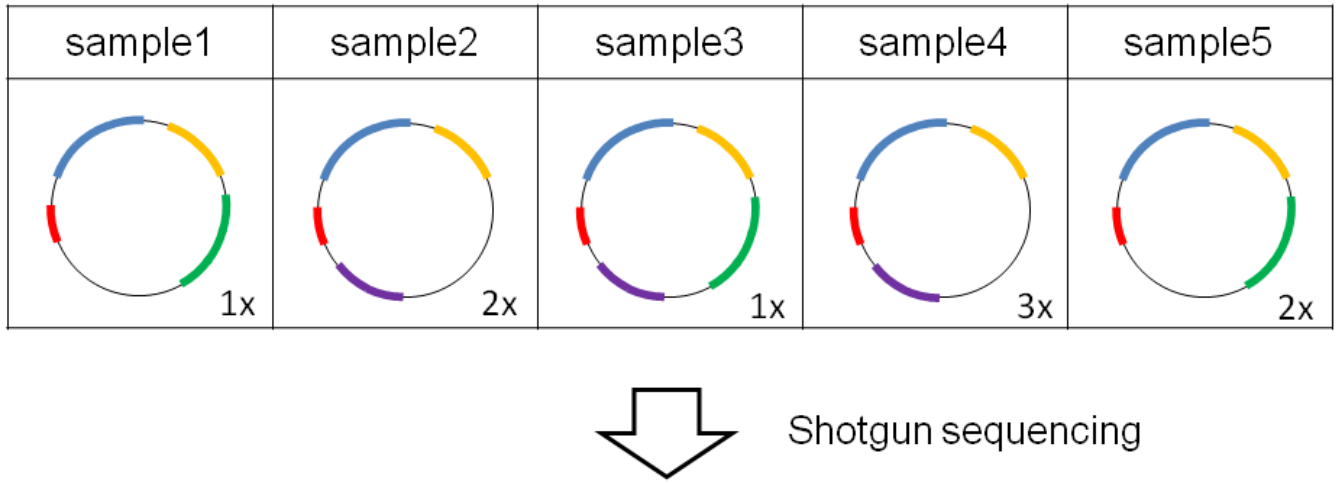
832 63. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, et al. A
833 Catalog of Reference Genomes from the Human Microbiome. Science (80-.). [Internet].
834 2010;328:994–9. Available from:
835 http://www.sciencemag.org/cgi/doi/10.1126/science.1183605

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

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

841

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**	

Figure 1

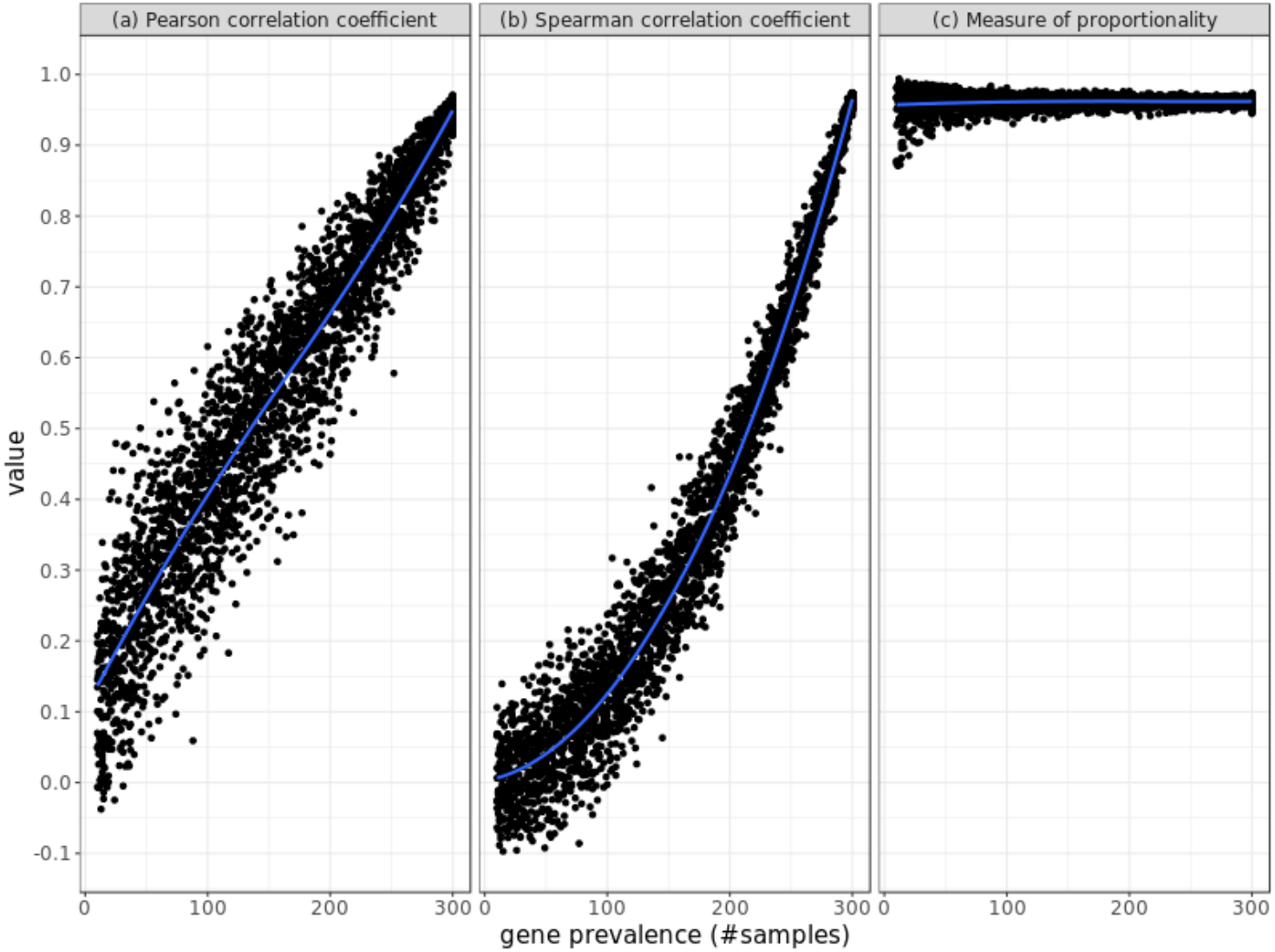

(a) Pearson correlation coefficient
(b) Spearman correlation coefficient
(c) Measure of proportionality
value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0
100
200
300
gene prevalence (#samples)

Figure 2

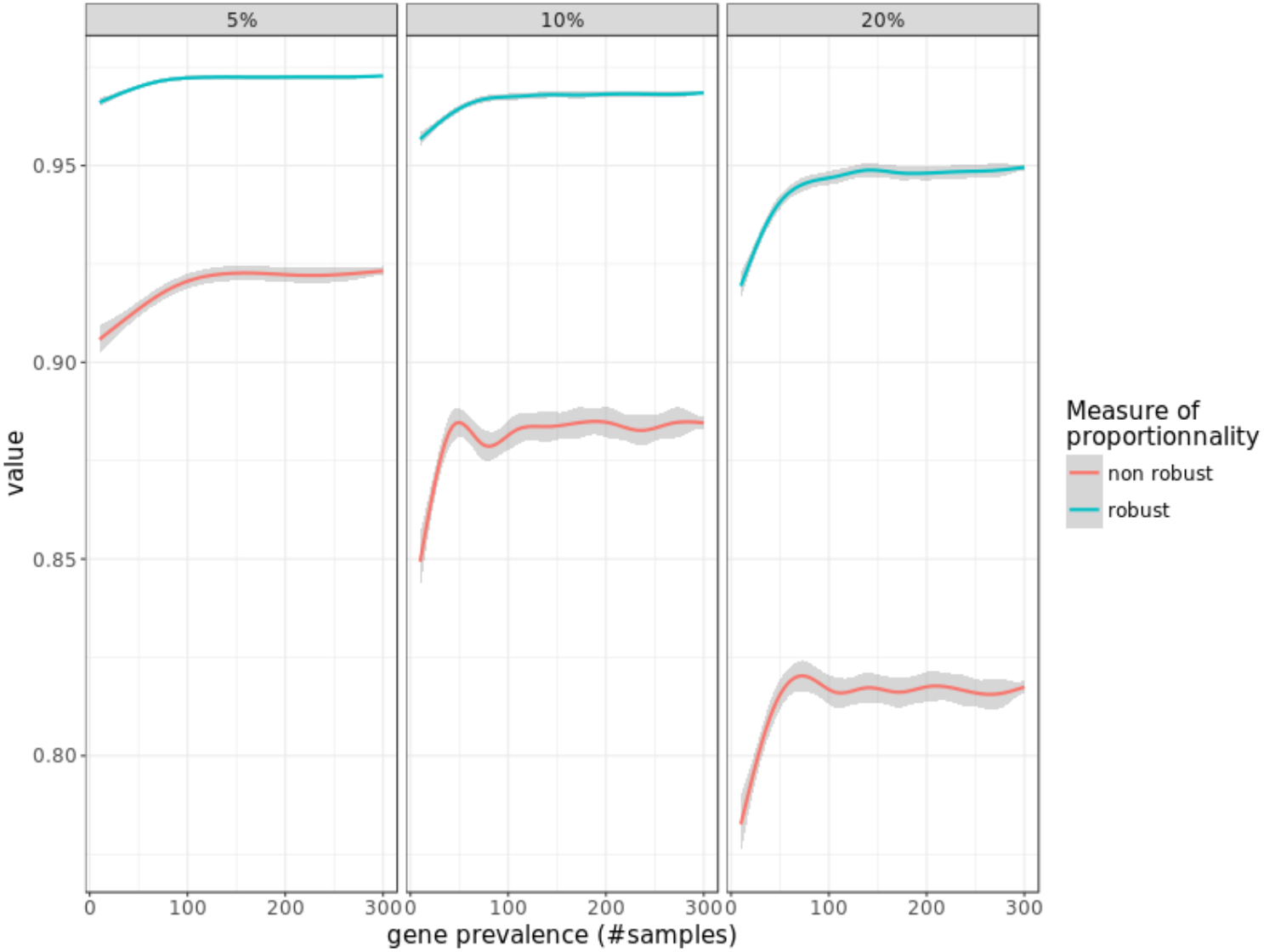

5%
10%
20%
0.95
0.90
0.85
0.80
value
0
100
200
300
gene prevalence (#samples)
Measure of proportionnality
non robust
robust

Figure 3

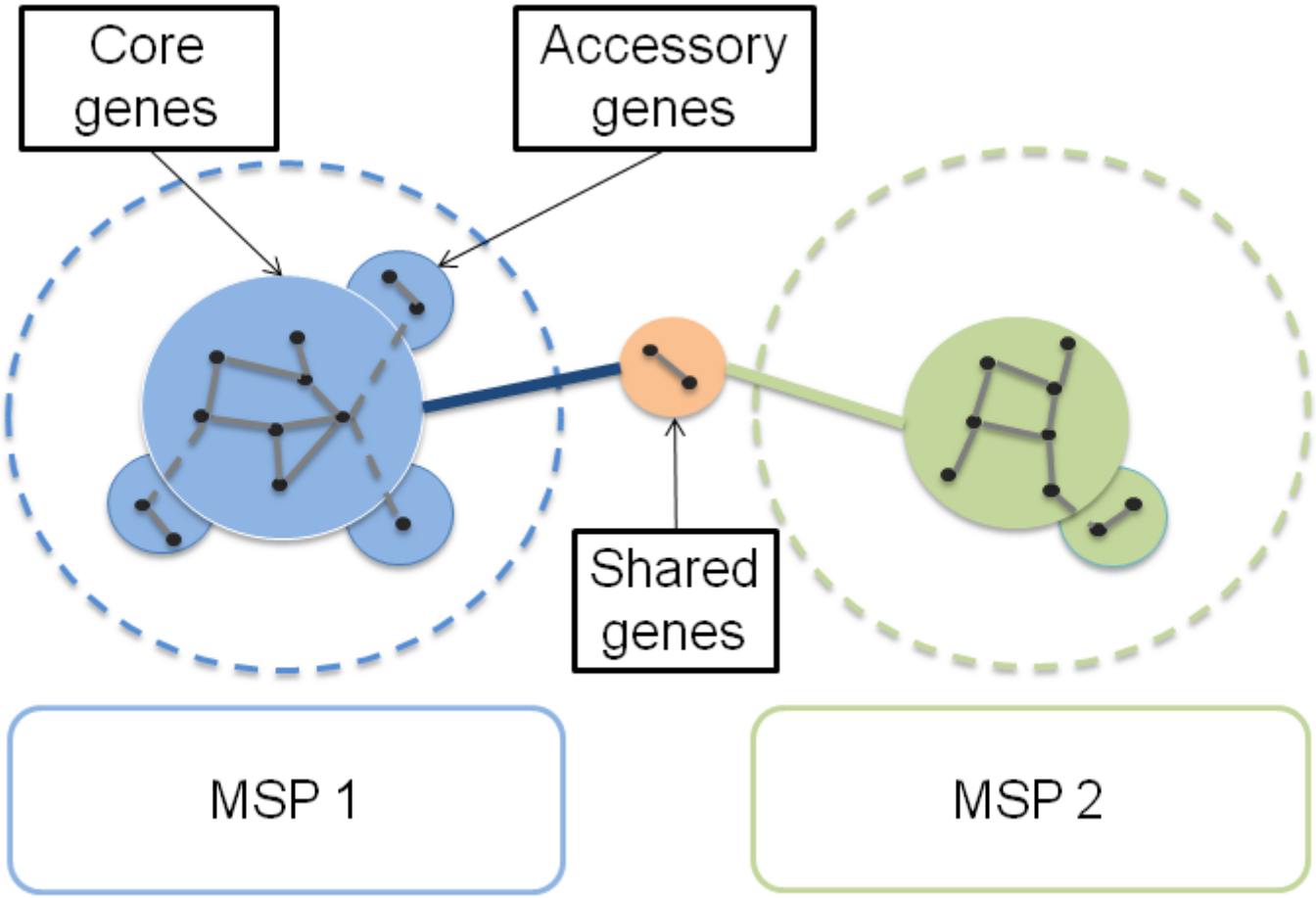
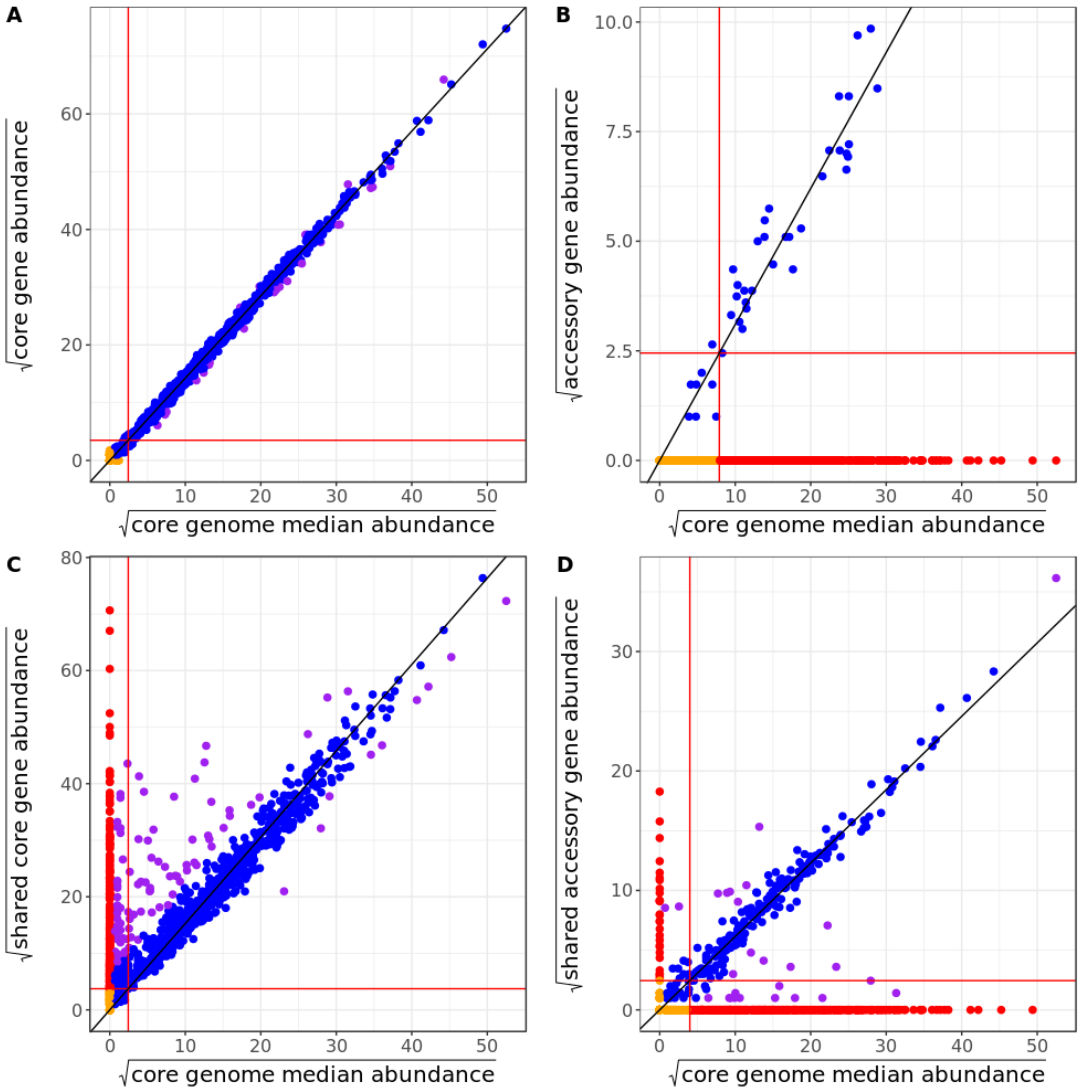
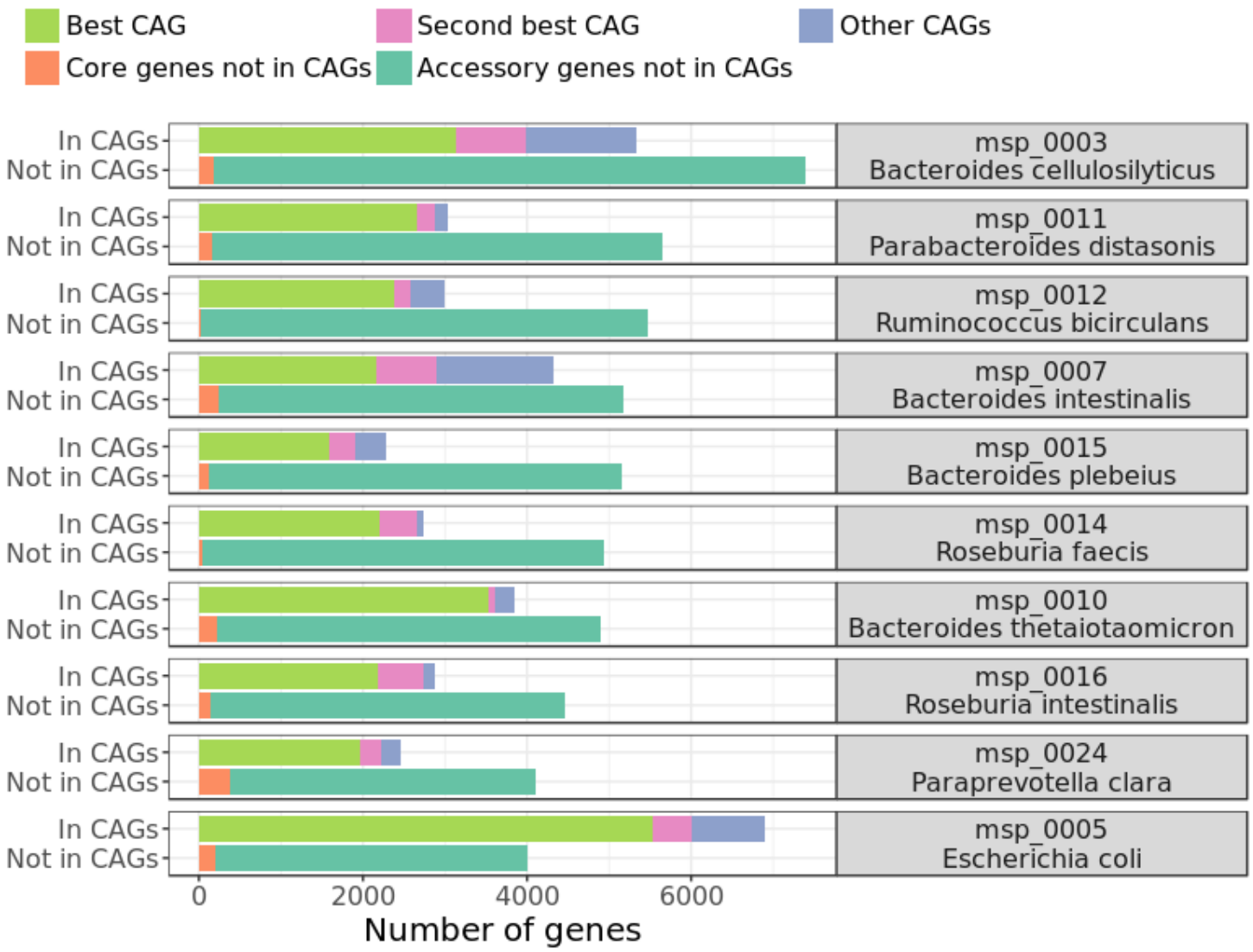


Figure 4



852

853 Figure 5



854

855 Figure 6

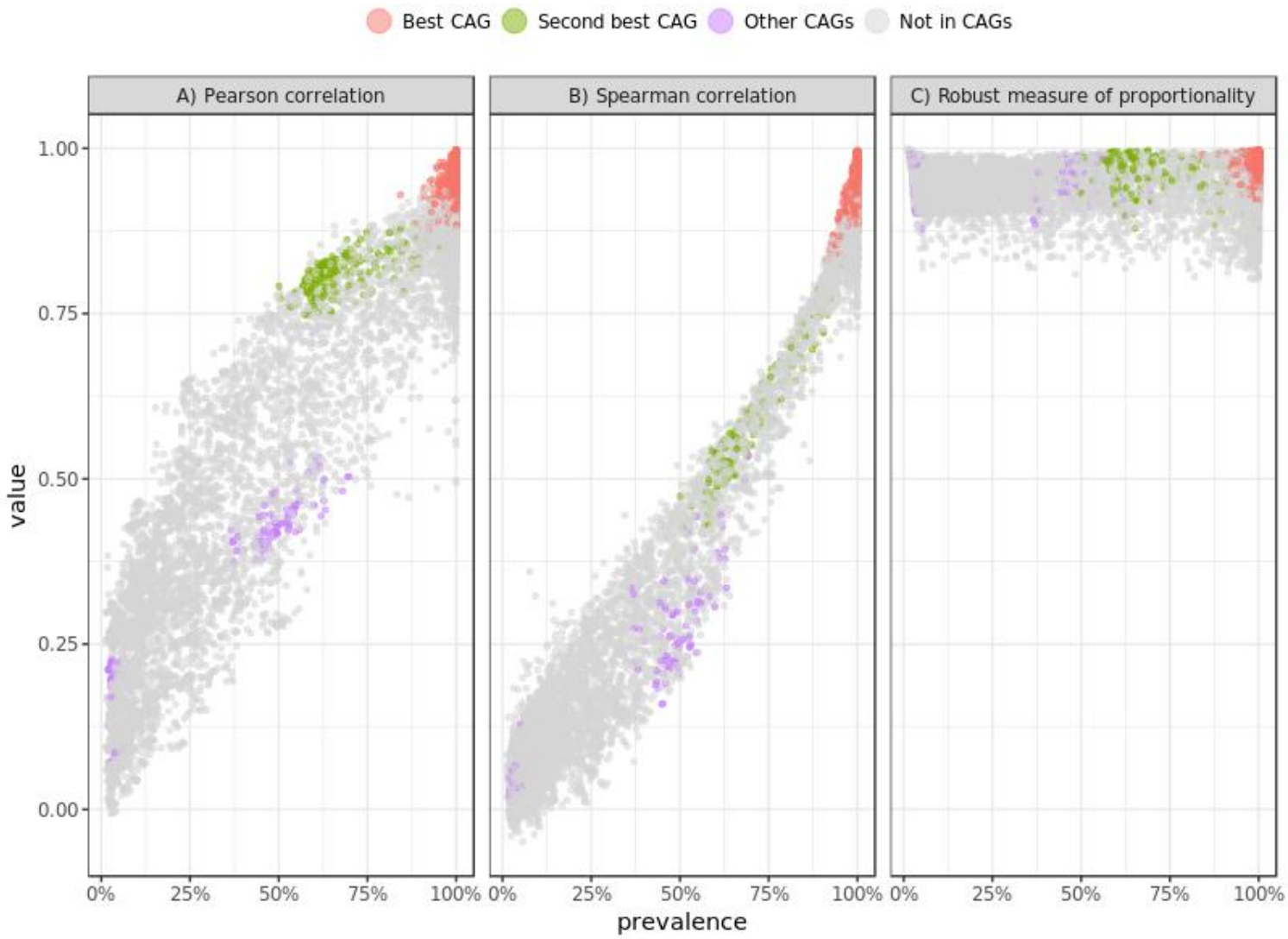

Best CAG
Second best CAG
Other CAGs
Not in CAGs
A) Pearson correlation
B) Spearman correlation
C) Robust measure of proportionality
value
1.00
0.75
0.50
0.25
0.00
0%
25%
50%
75%
100%
prevalence

Figure 7

858 # Figures legends

859 **Figure 1 : Simple model illustrating the rationale behind the method**

860 5 samples carry different strains from the same species. Three core genes (red, blue, orange)
861 are present in all the strains. Two accessory genes (green, purple) are present only in some
862 strains. The abundance of the species in each sample ranges from 1 to 3 copies.
863 After shotgun sequencing, a raw gene abundance matrix is built.
864 A strict proportionality relationship is expected between two core genes, the proportionality
865 coefficient being equal to the ratio of their length. In contrast, such relationship between a
866 core and an accessory gene should be observed only in the subset of samples where the
867 accessory gene is present.

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

870 using:

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

874 **Figure 3 : Comparison of the noisy genes abundance profiles of the virtual species**
875 **to the original median signal of its core genome**

876 using either the non-robust or robust measure of proportionality with an increasing
877 percentage of outliers (5%, 10% and 20%).

878 **Figure 4 : Structure of Metagenomic Species Pan-genomes (MSPs)**

879 **Figure 5 : The four types of genes in a MSP**

880 Comparison of the core genome median abundance of the msp_0043 (*Ruminococcus bromii*)
881 to:

882 A. the gene MH0003_GL0010264 classified as core. The gene is detected in all the
883 samples where MSP is detected.
884 B. the gene MH0025_GL0082295 classified as accessory. The gene is missing in 516
885 samples where the MSP is detected.

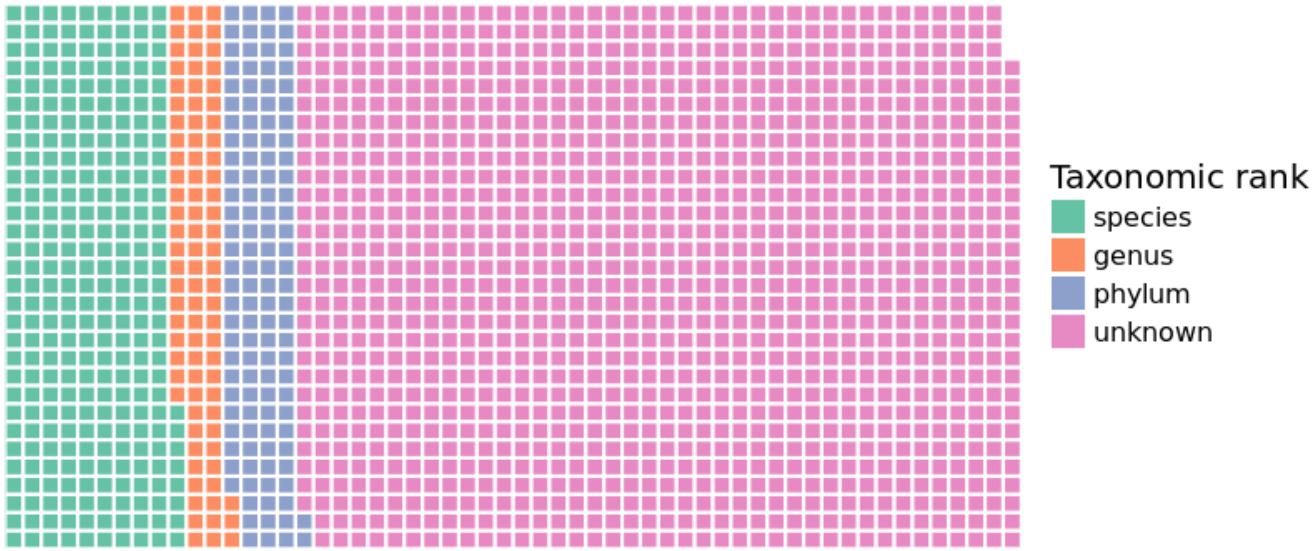
886 C. the gene 657321.RBR_R_22270 classified as its shared core. The gene is present in all
887 the samples where MSP is detected but also in 286 samples where the MSP is not.
888 D. the gene MH0205_GL0102923 classified as shared accessory. The gene is missing in
889 454 samples where the MSP is detected but present in 28 samples where the MSP is
890 not.

891 **Figure 6: Comparison of CAGs and MSPs**

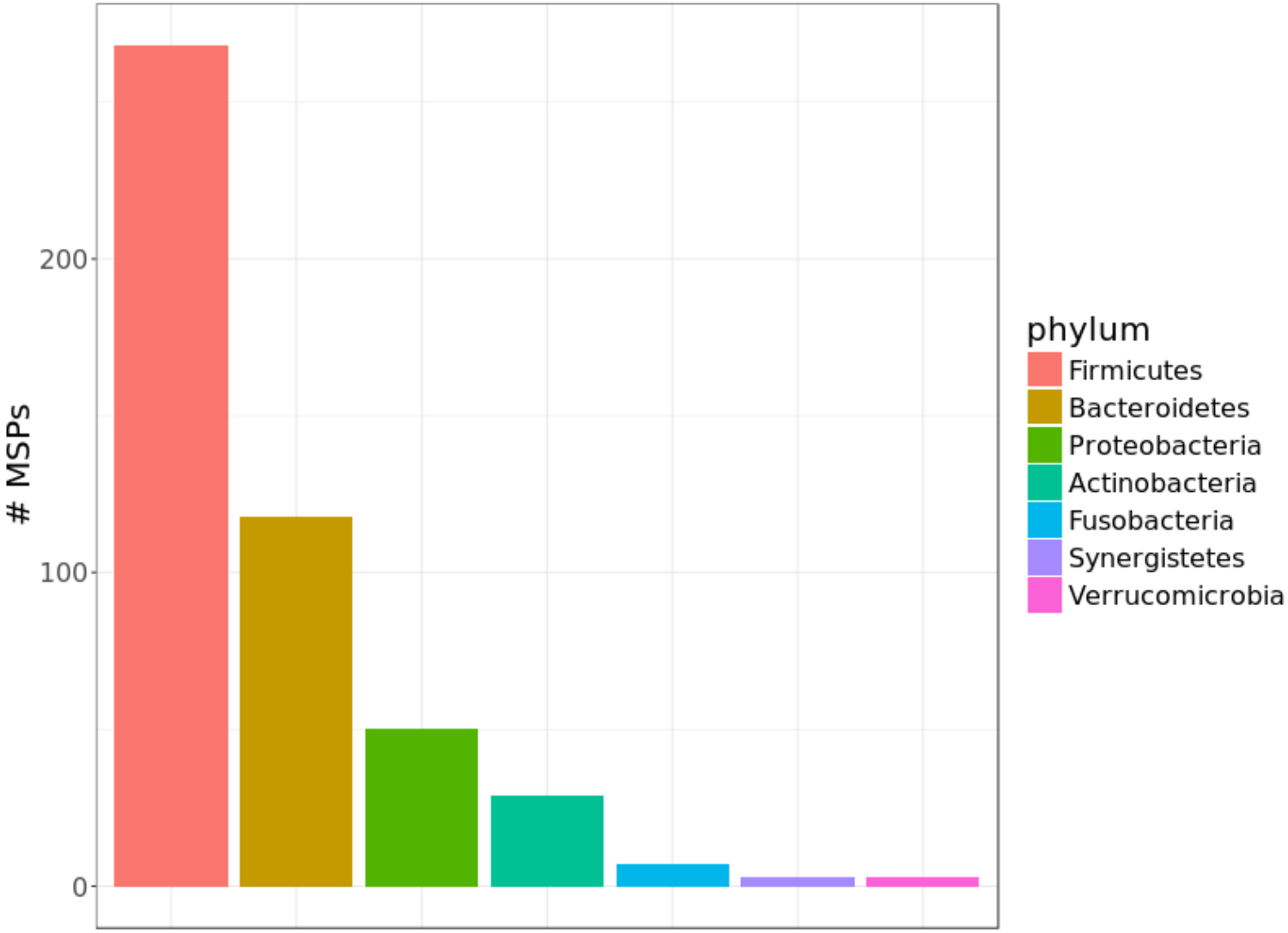
892 **Figure 7: Comparison of the genes abundance profiles of the msp_0011**
893 **(*Parabacteroides distasonis*) to the median signal of its core genome.**

894 Three measures are compared:
895 A. the Pearson correlation coefficient
896 B. the Spearman correlation coefficient
897 C. the measure of proportionality.
898 Grey points correspond to genes unclassified by Canopy whereas those colored were grouped
899 in CAGs.
900

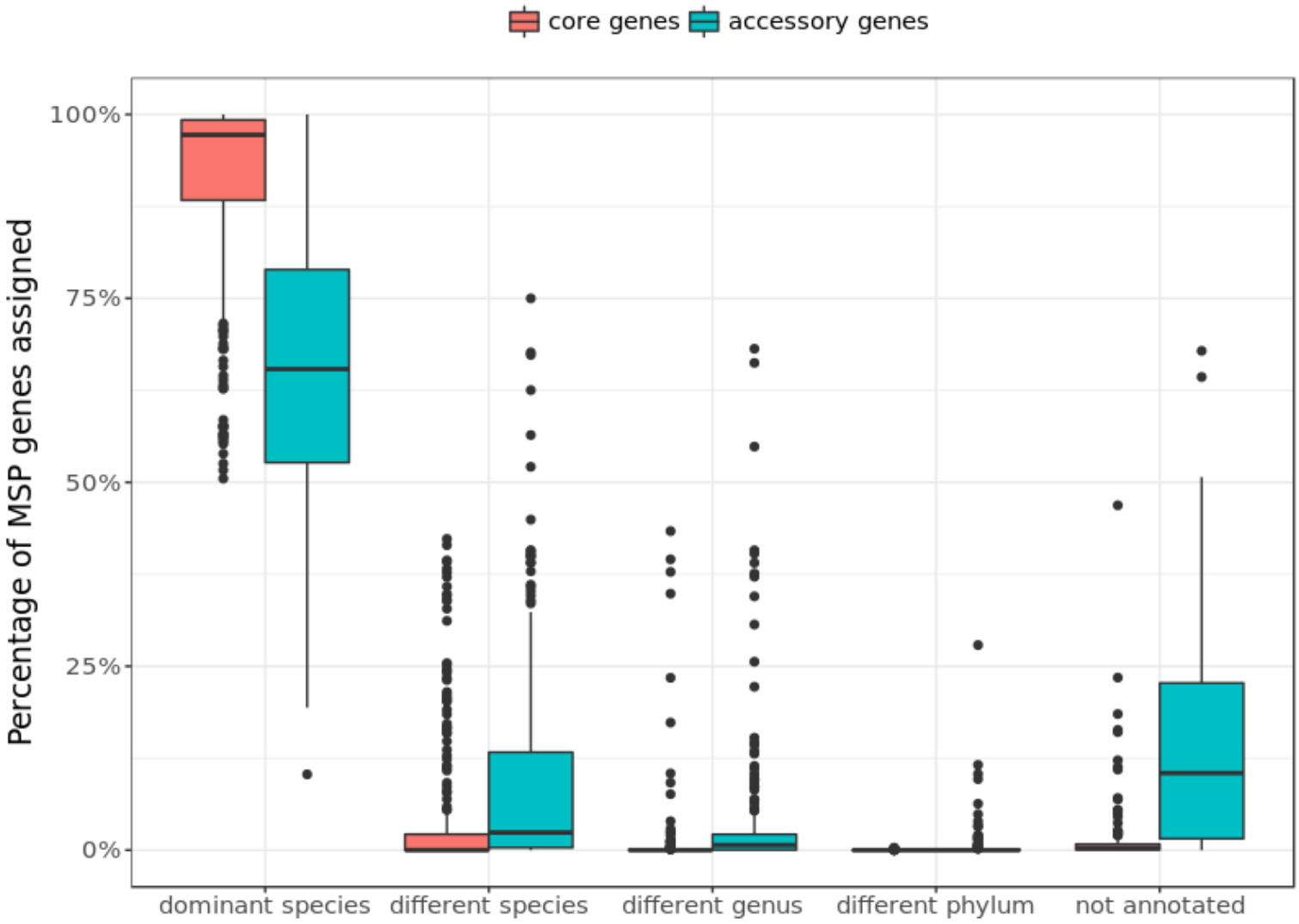
Supplementary Figures



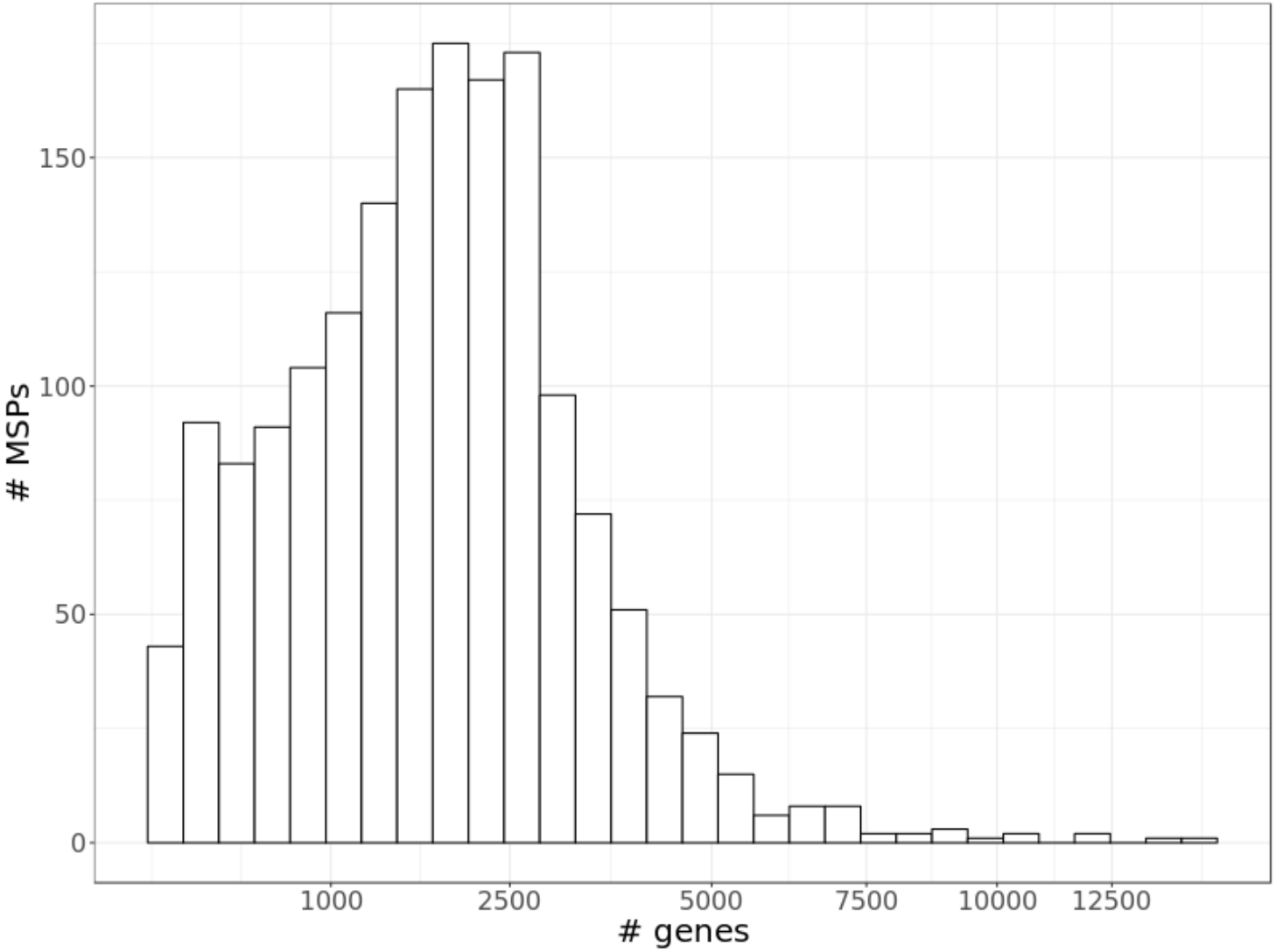
Supplementary Figure 1



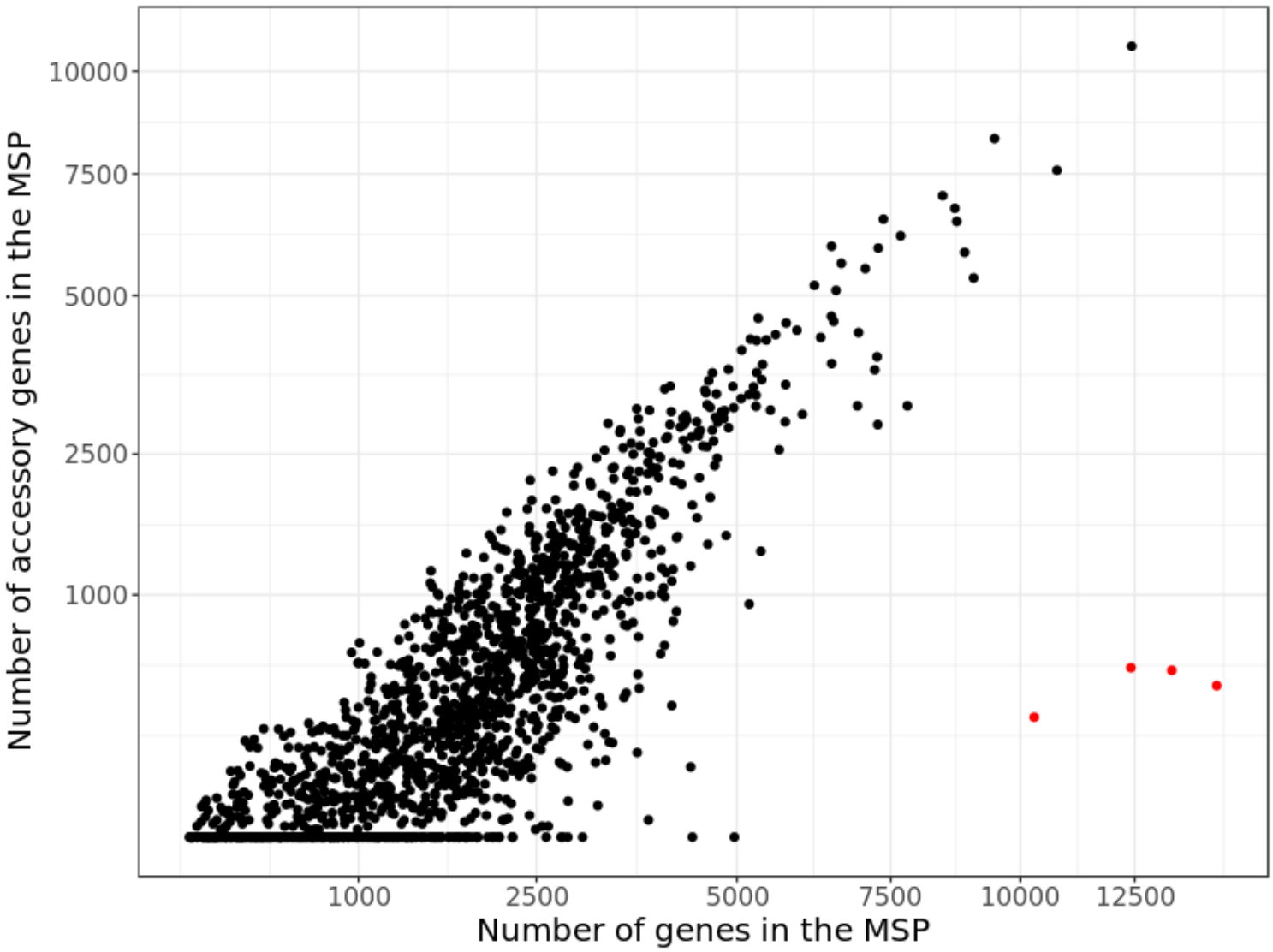
Supplementary Figure 2



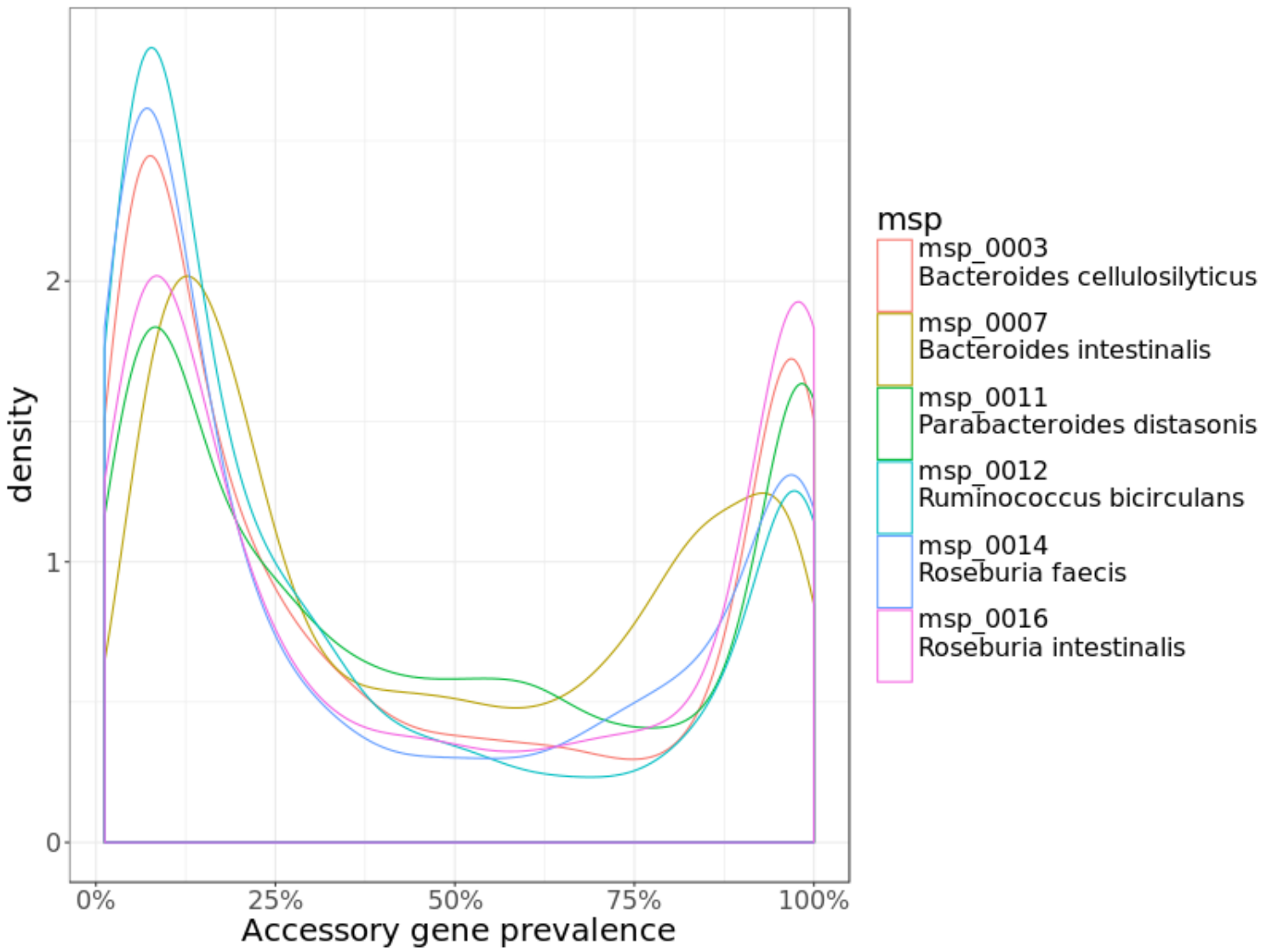
Supplementary Figure 3



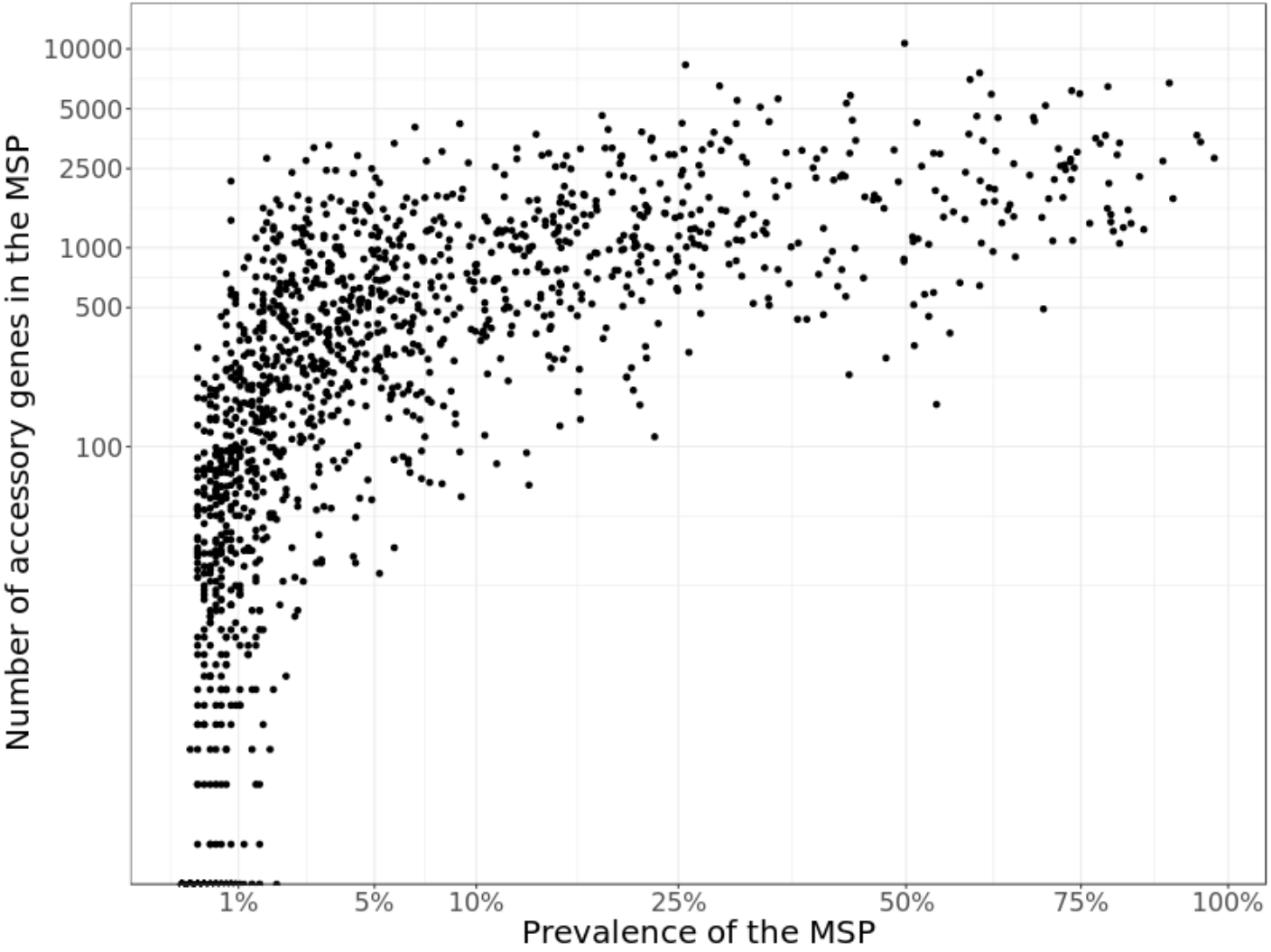
Supplementary Figure 4



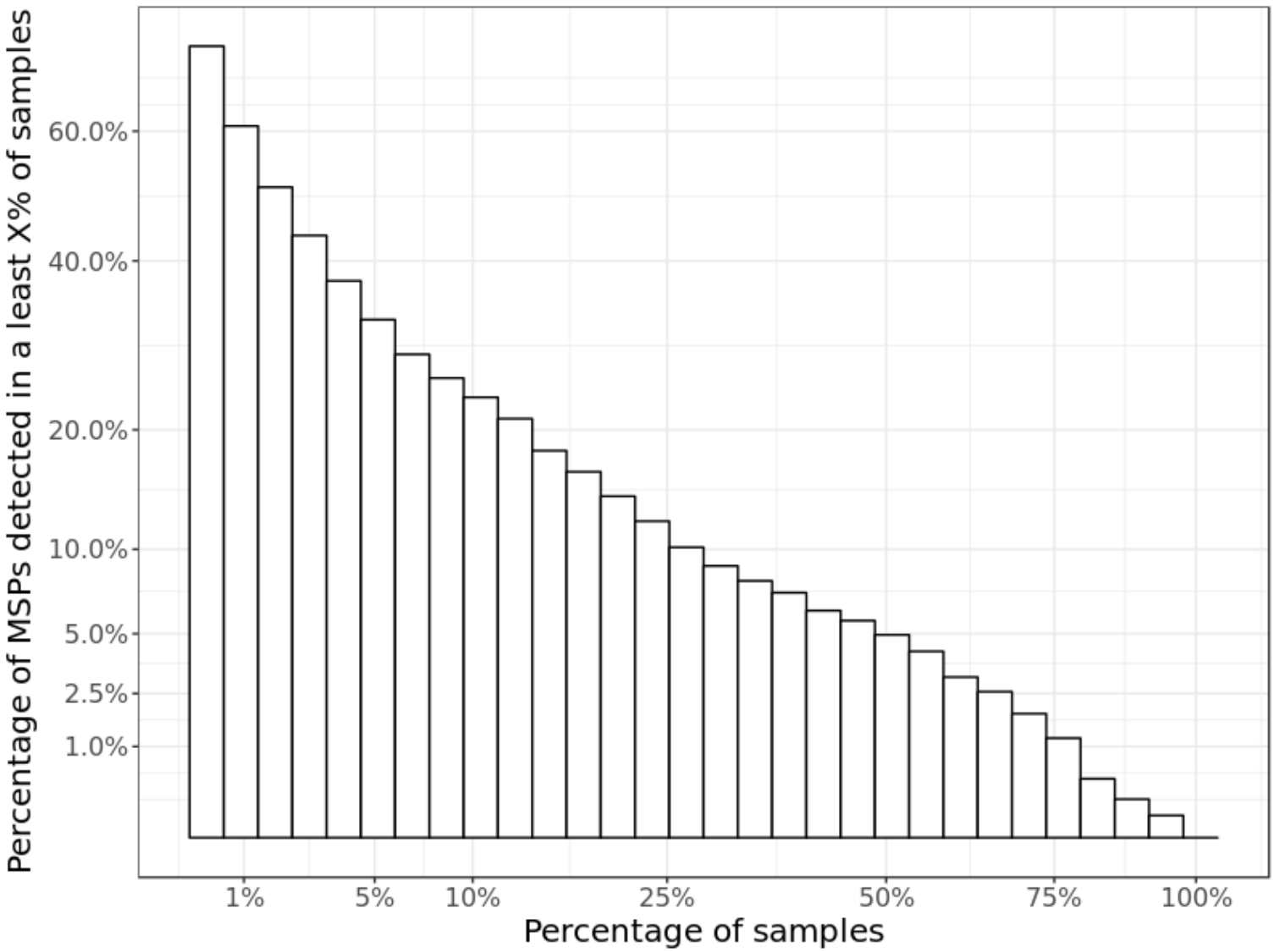
Supplementary Figure 5



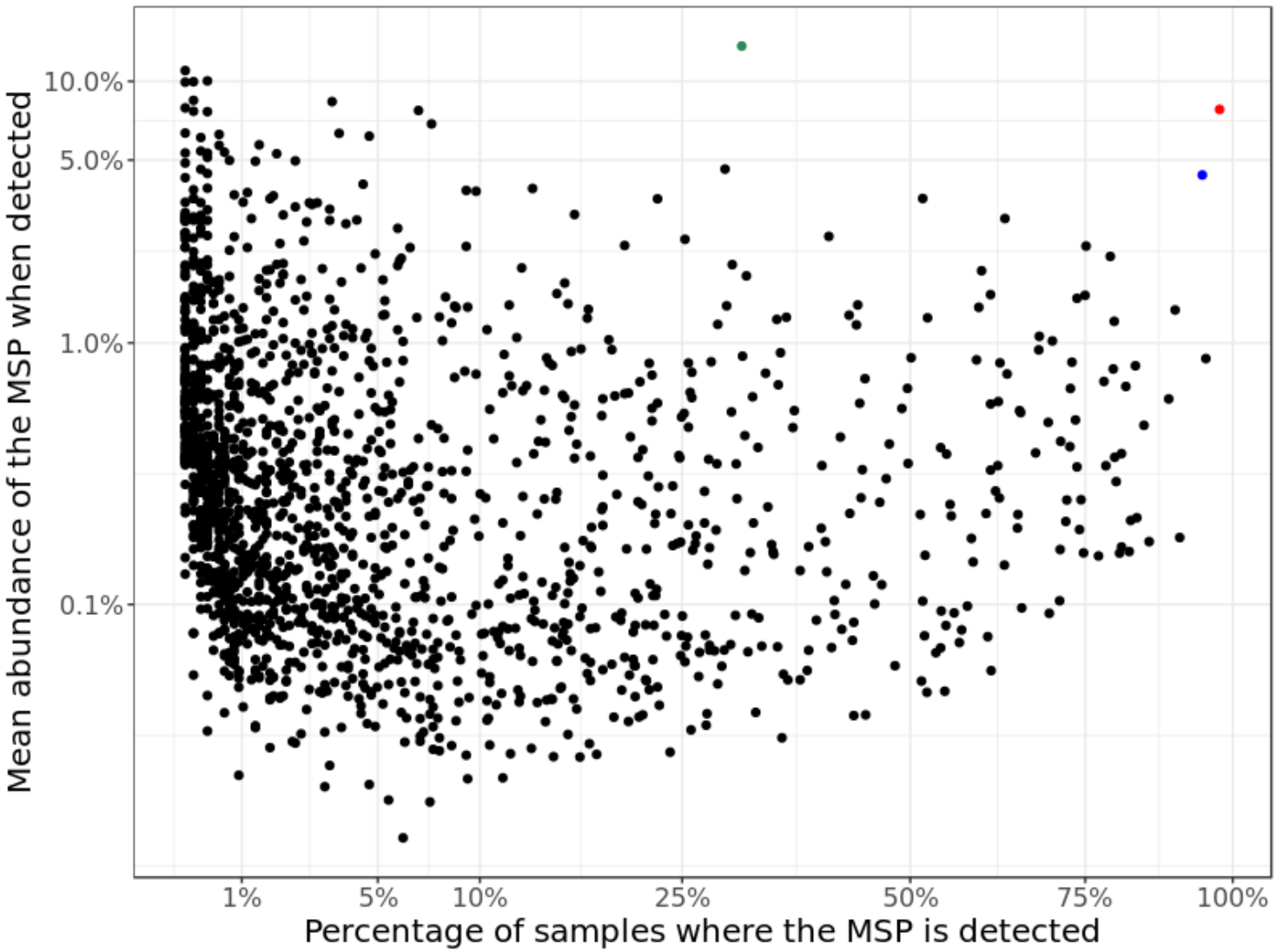
Supplementary Figure 6



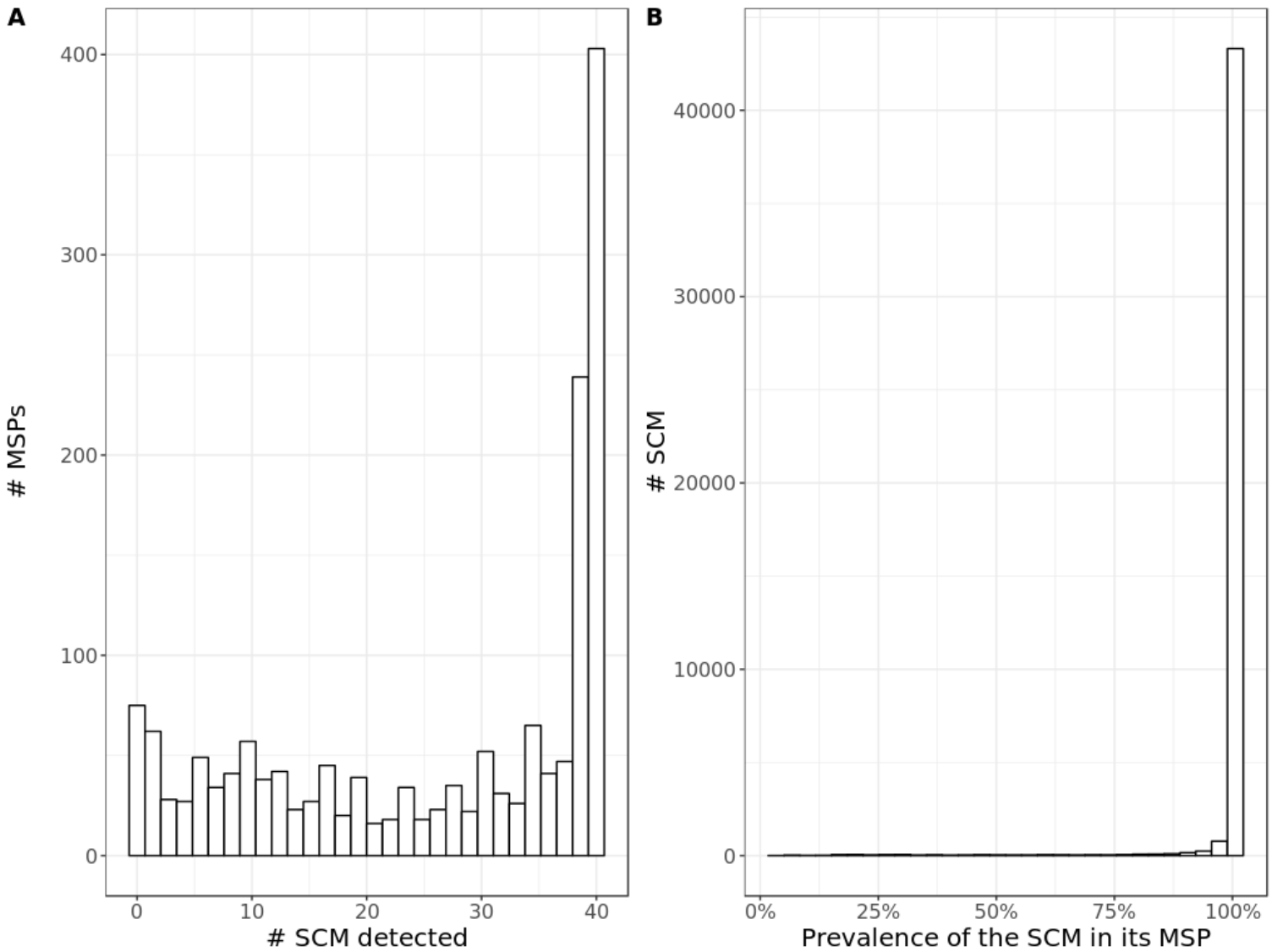
Supplementary Figure 7



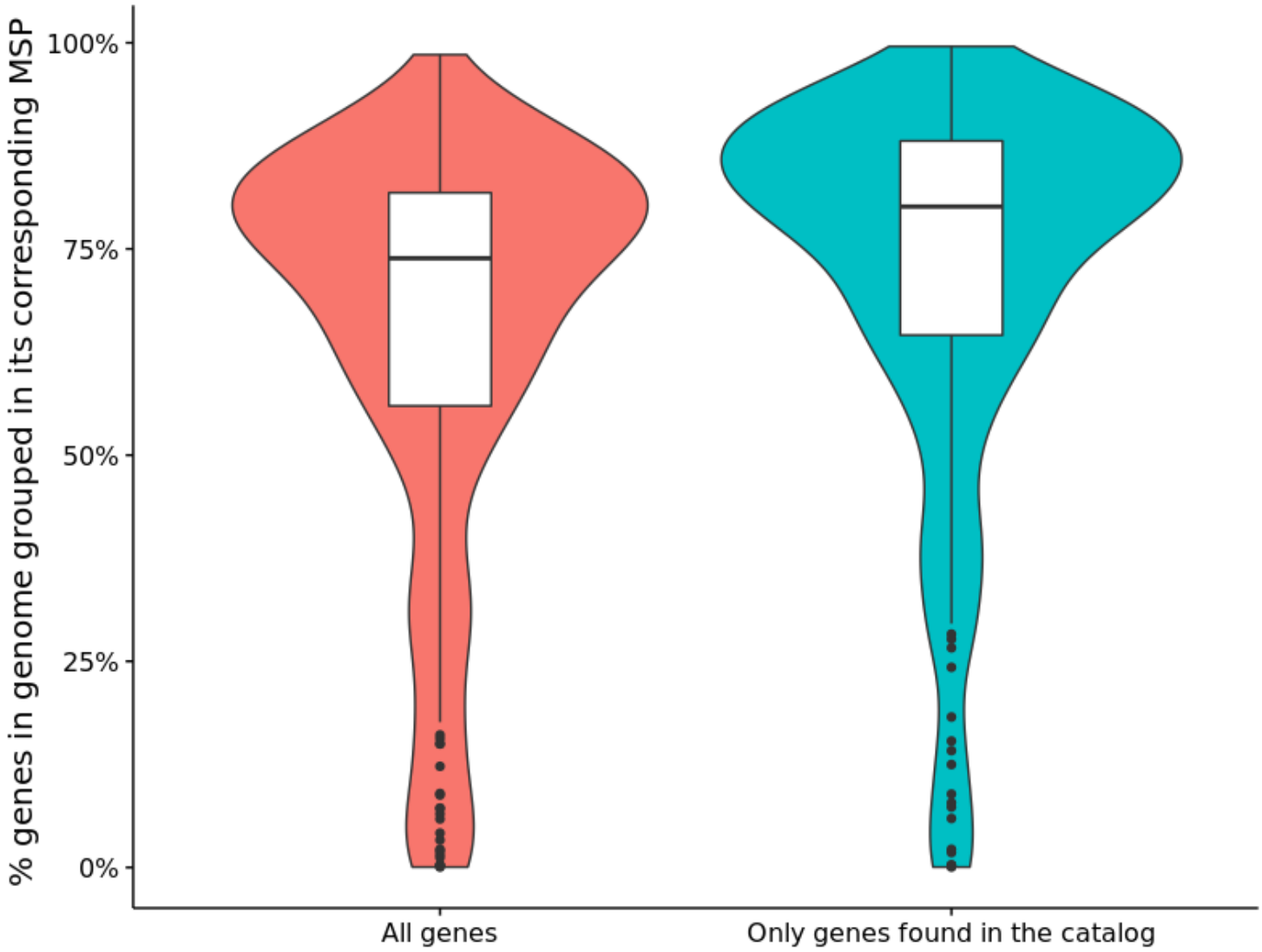
Supplementary Figure 8



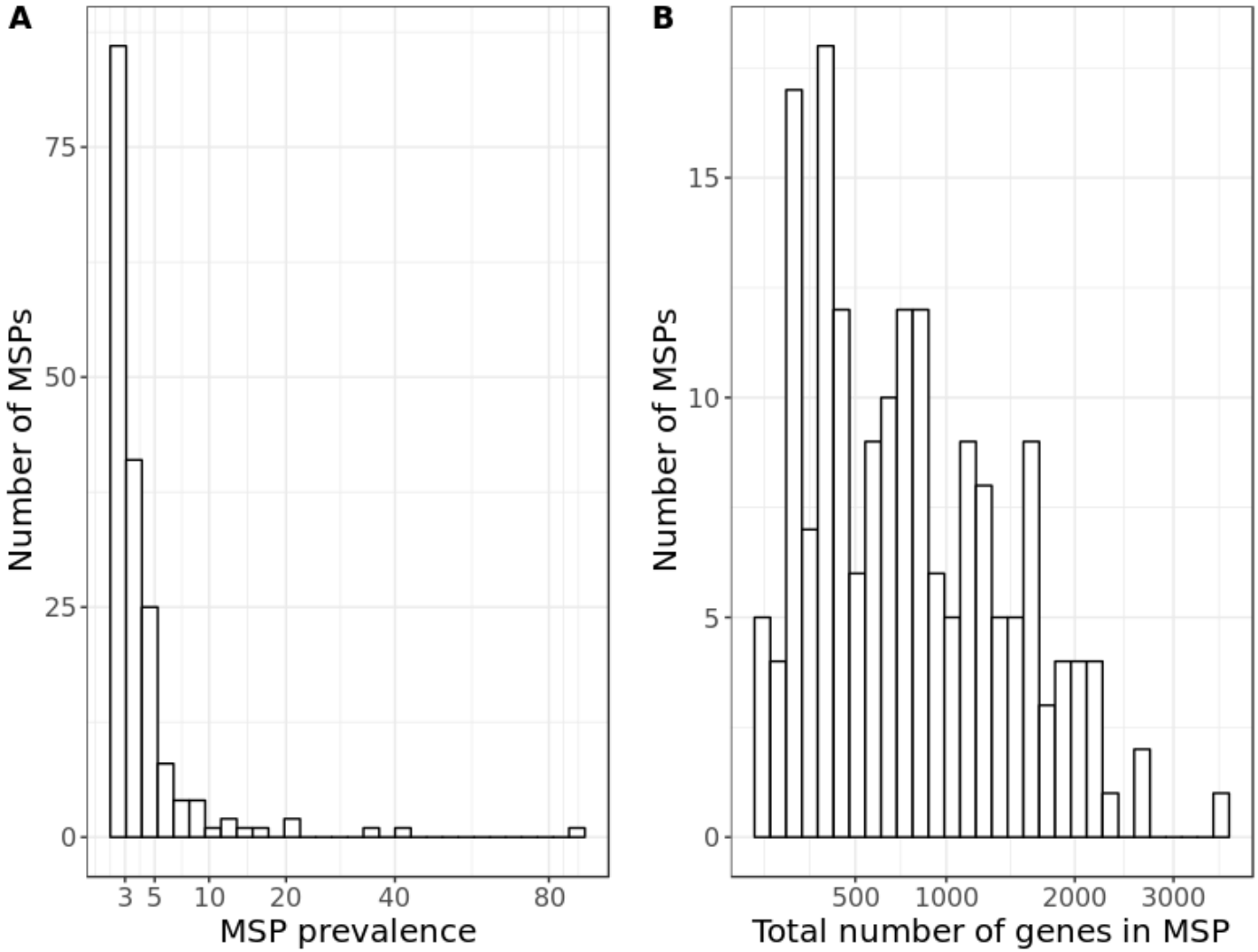
Supplementary Figure 9



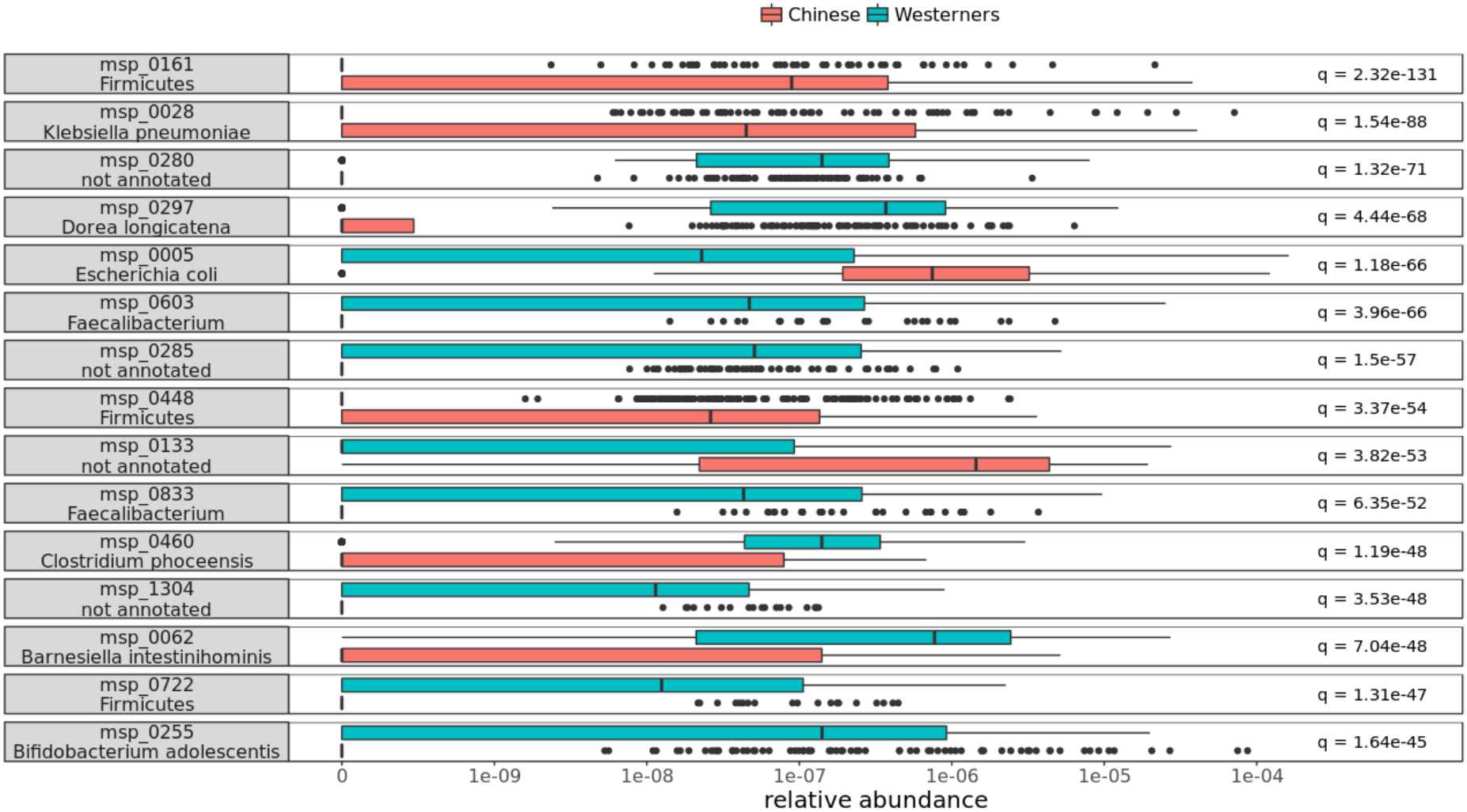
Supplementary Figure 10



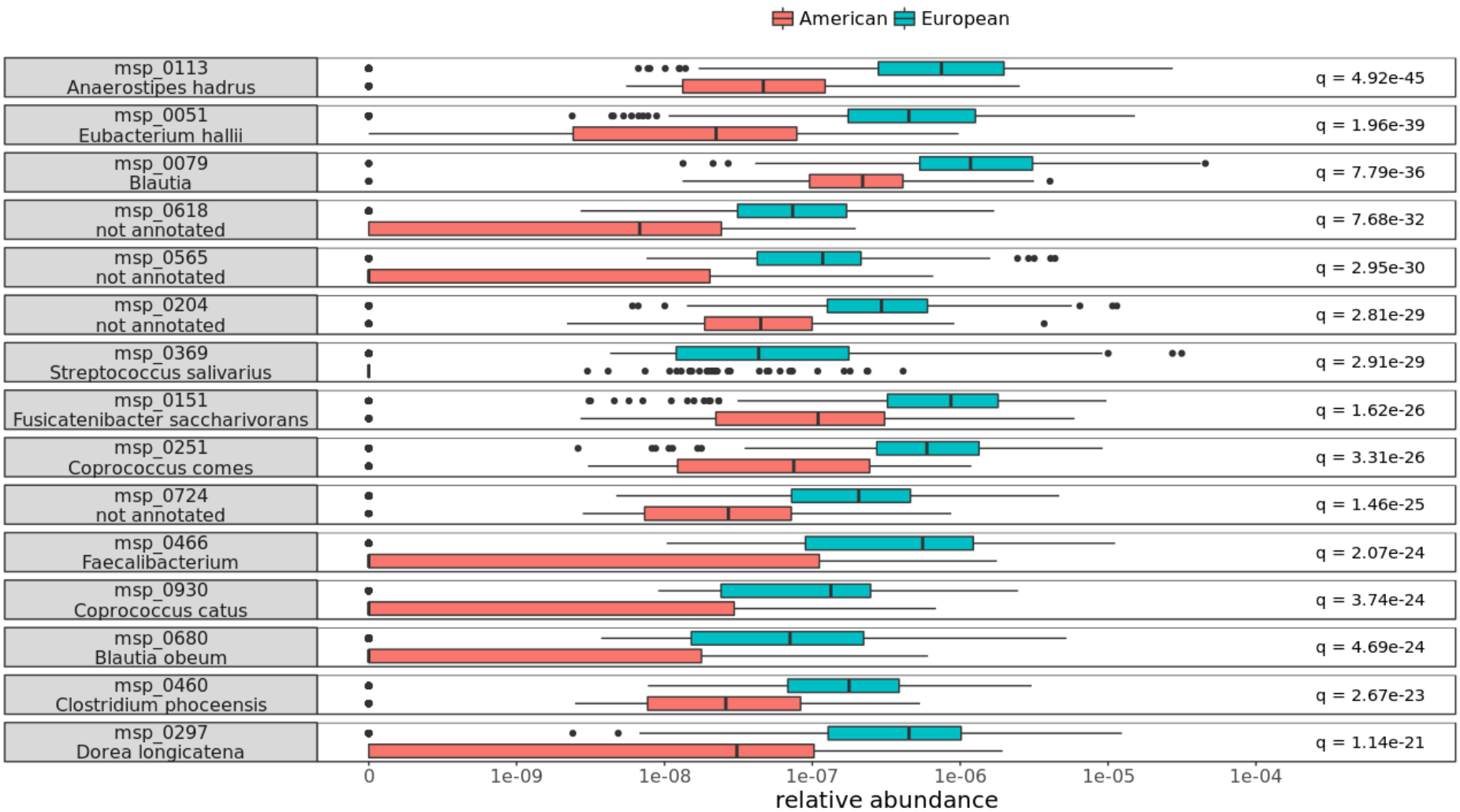
Supplementary Figure 11



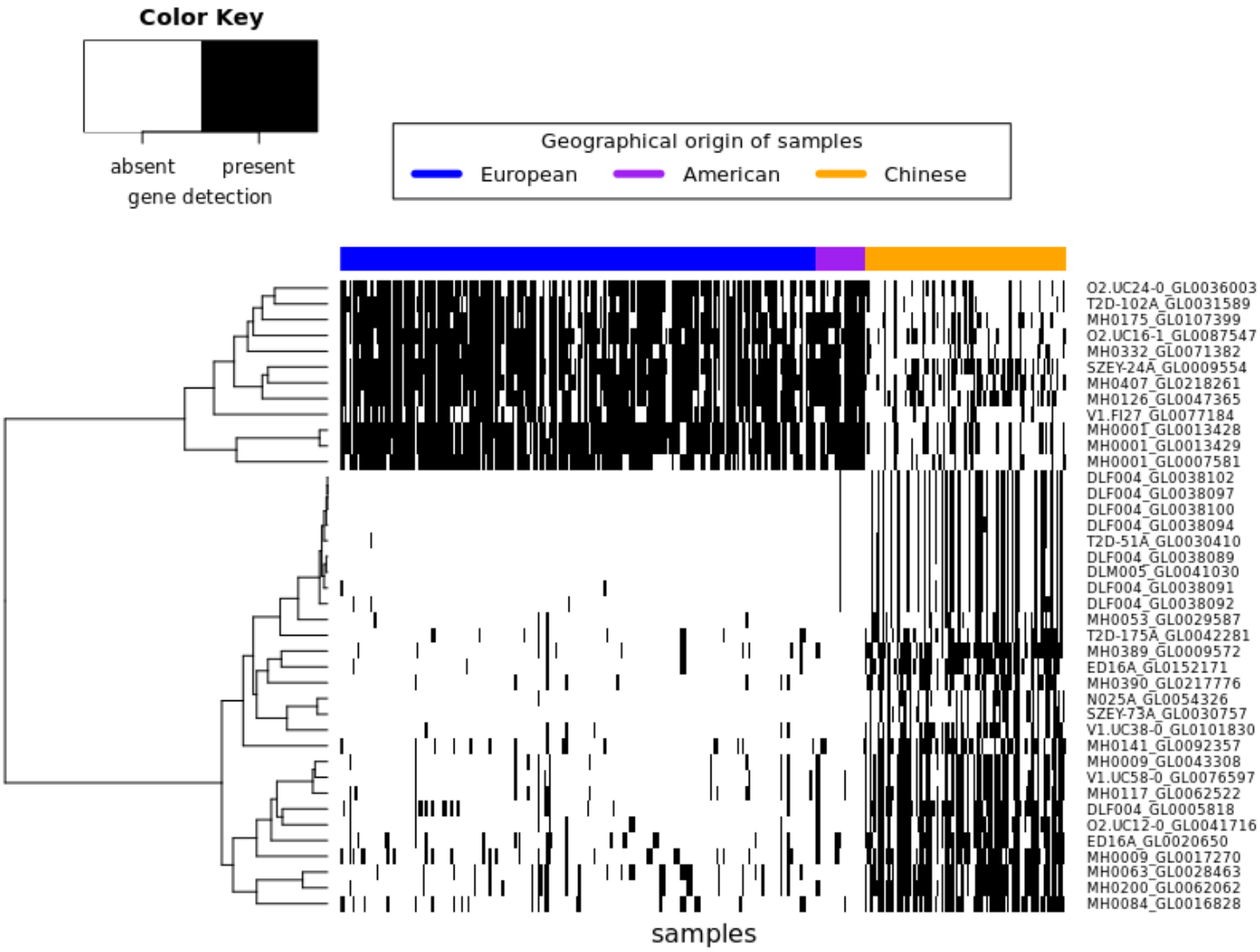
Supplementary Figure 12



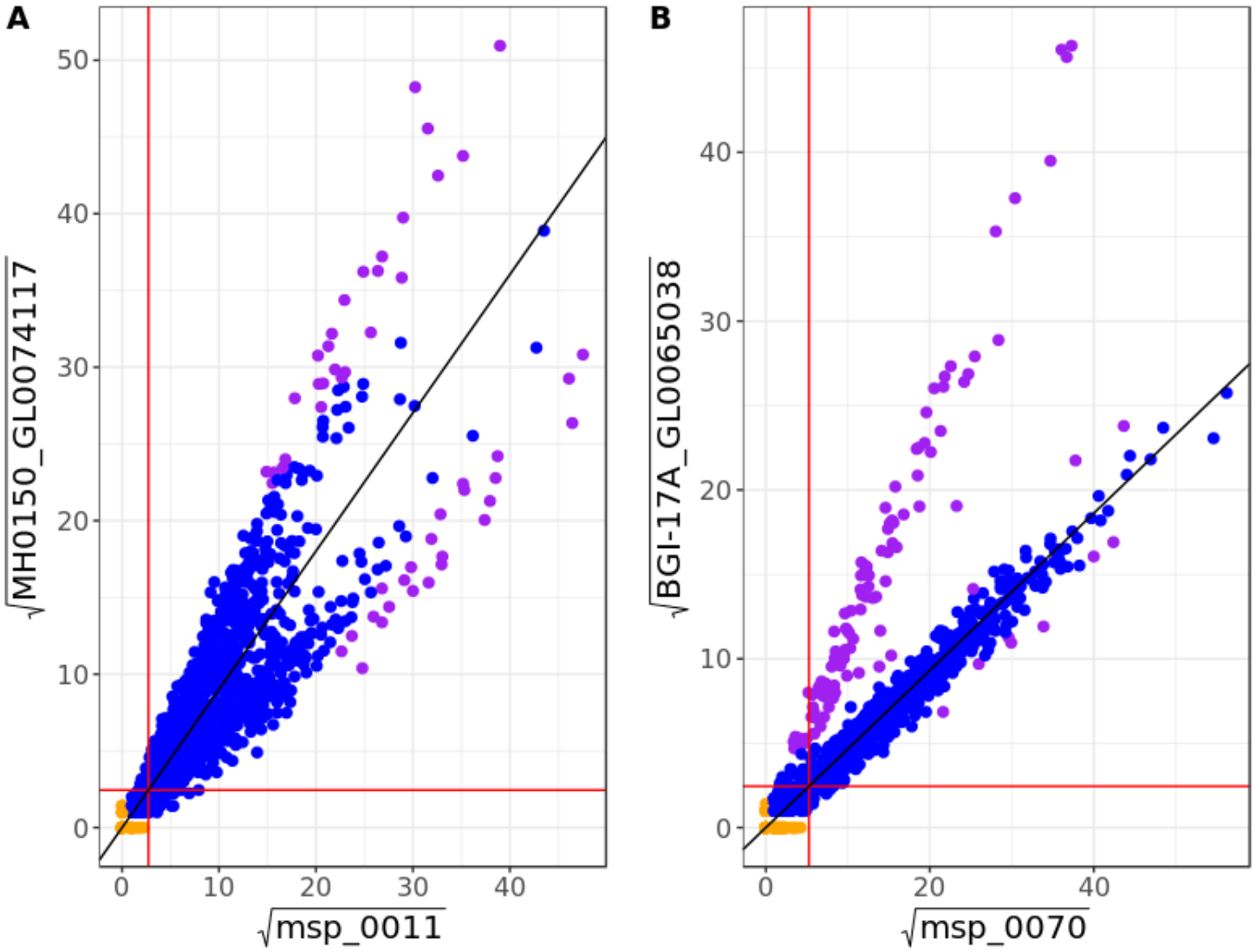
Supplementary Figure 13



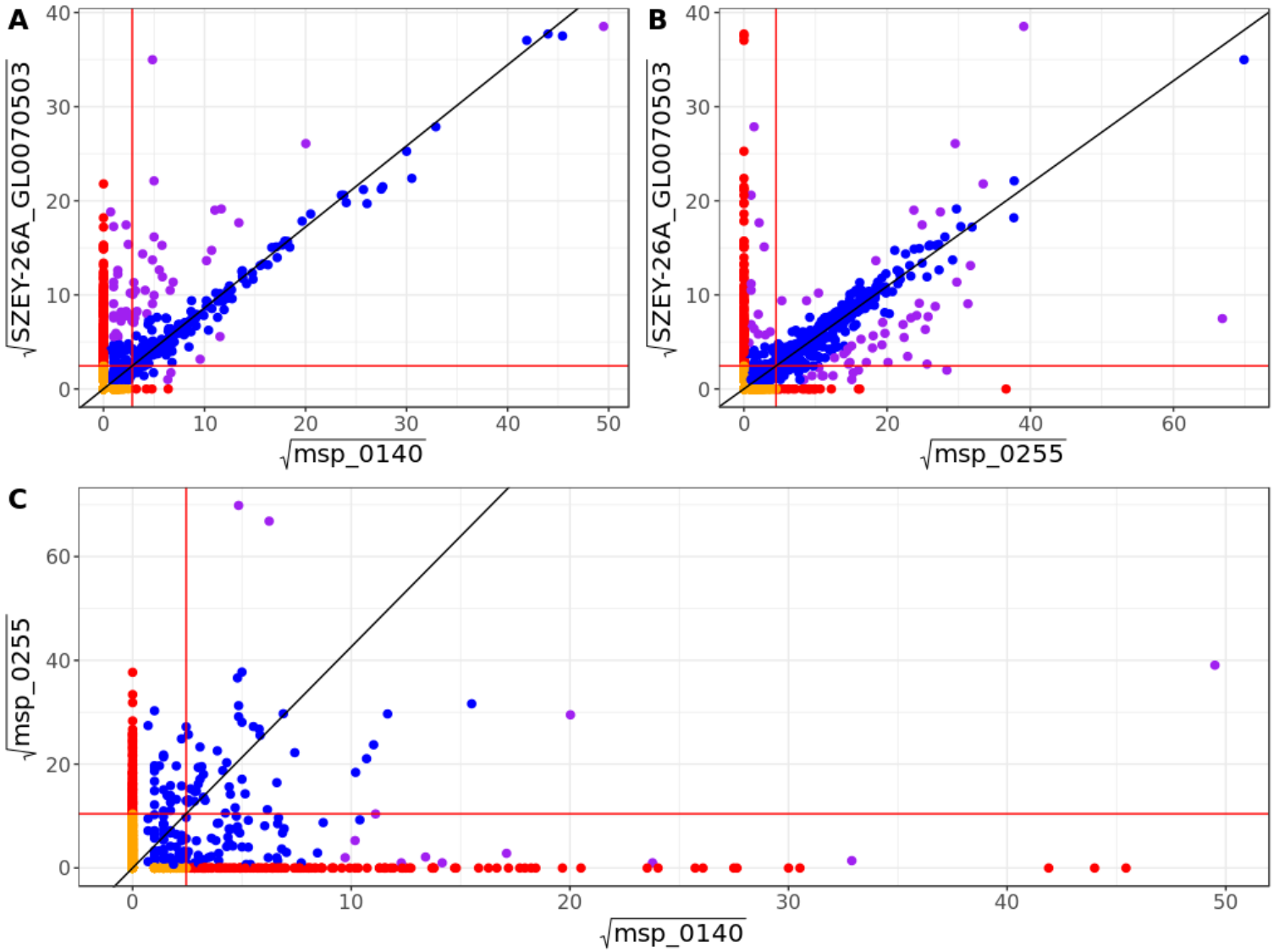
Supplementary Figure 14



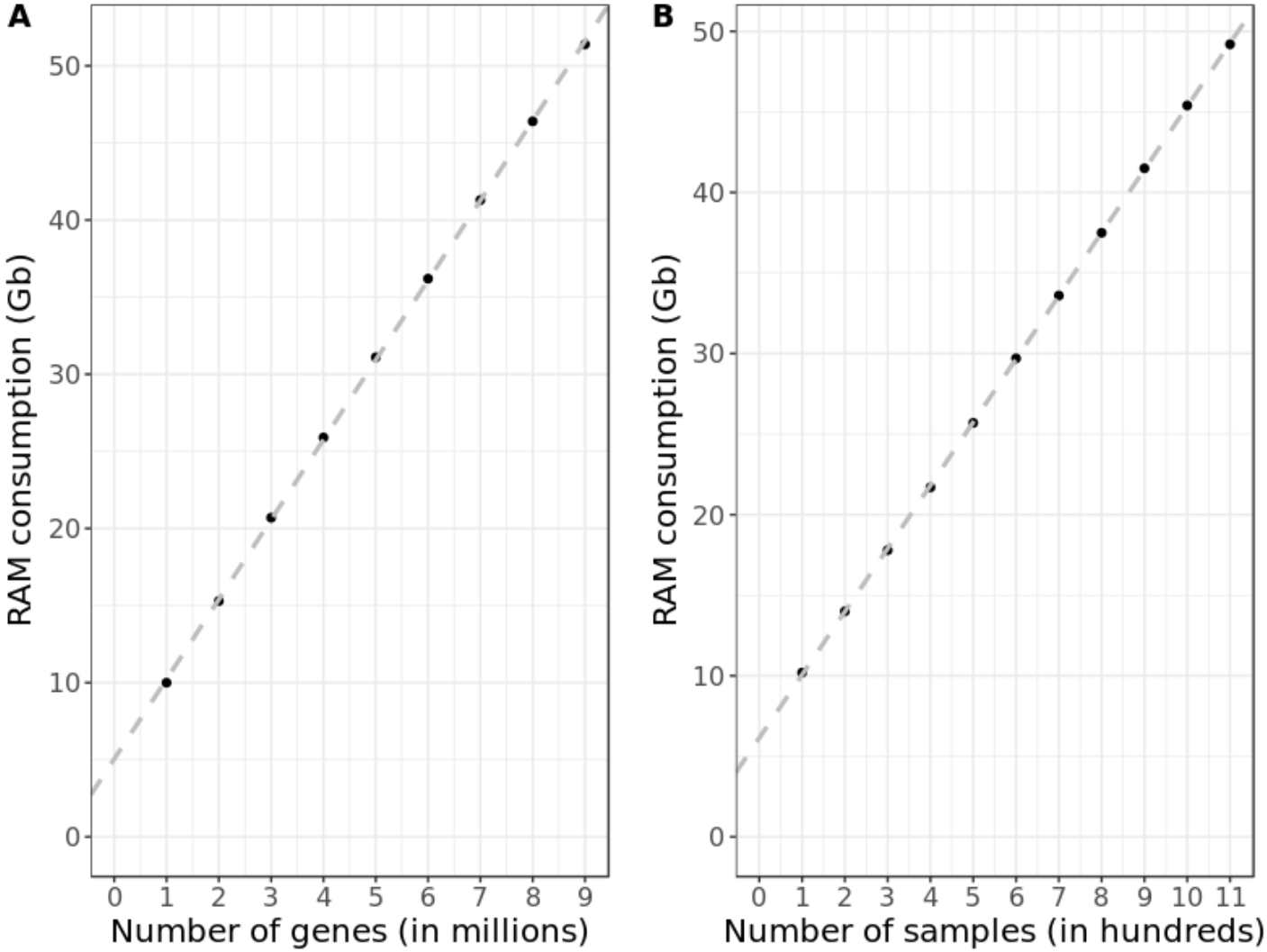
Supplementary Figure 15



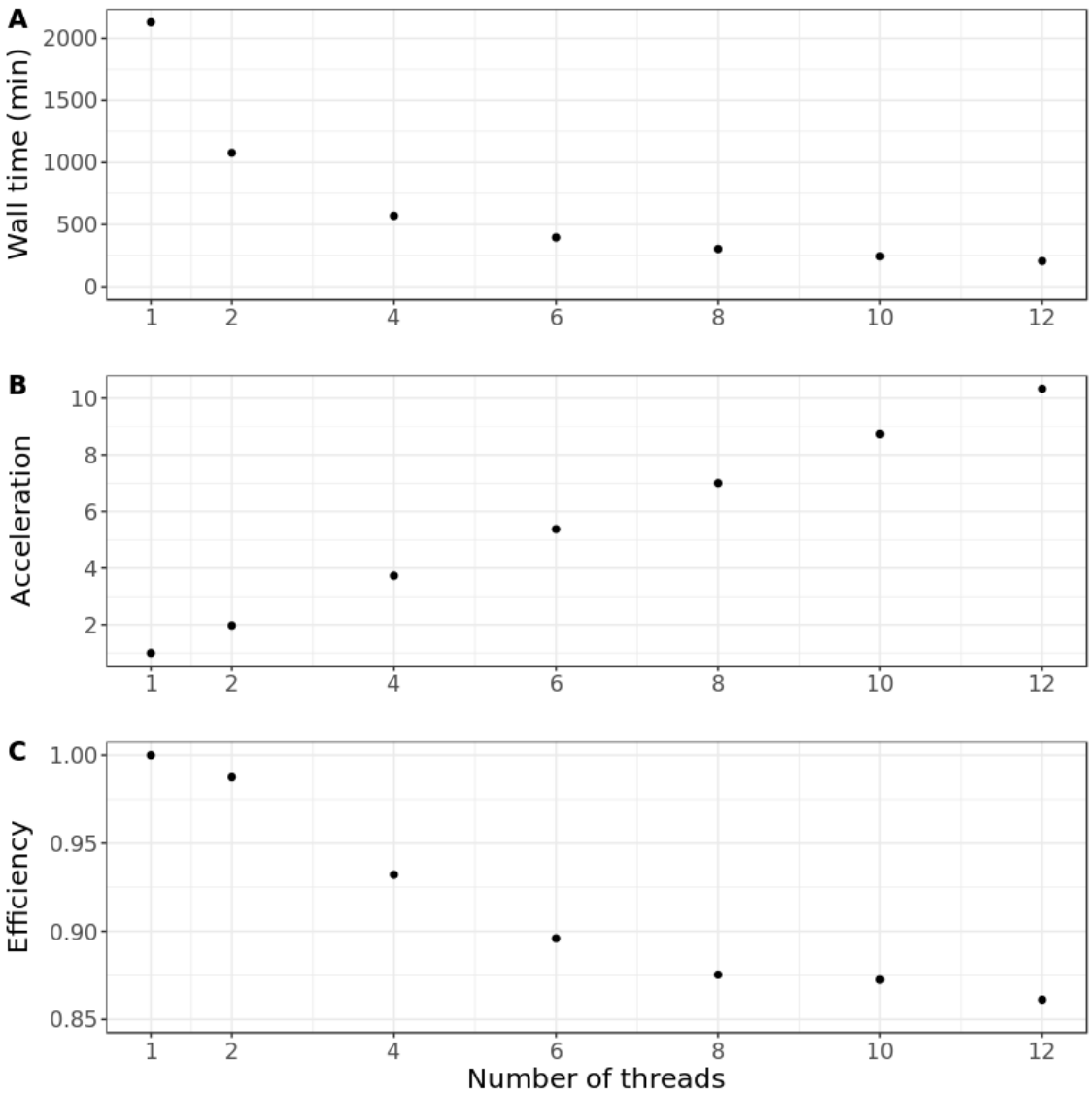
Supplementary Figure 16



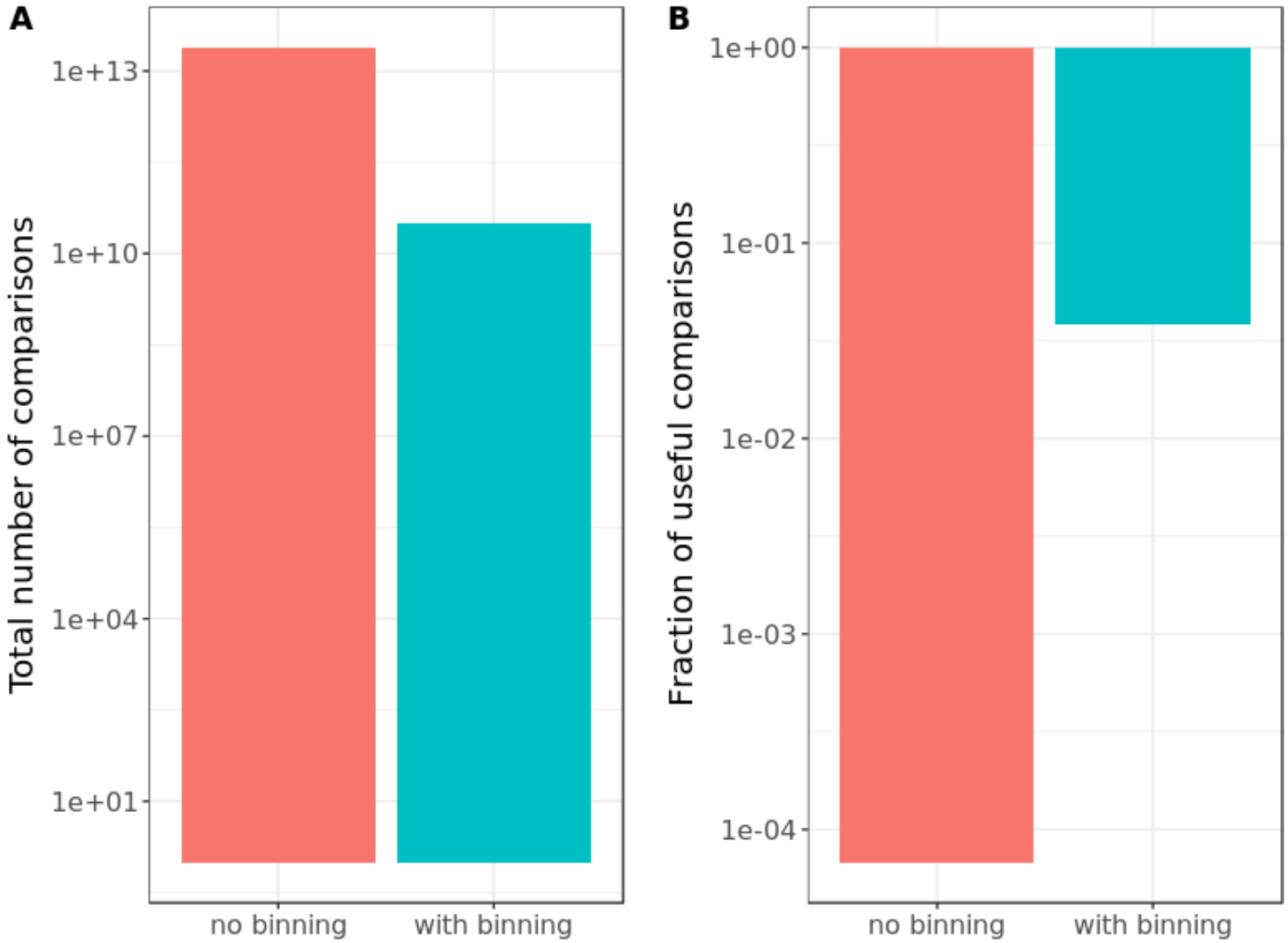
Supplementary Figure 17



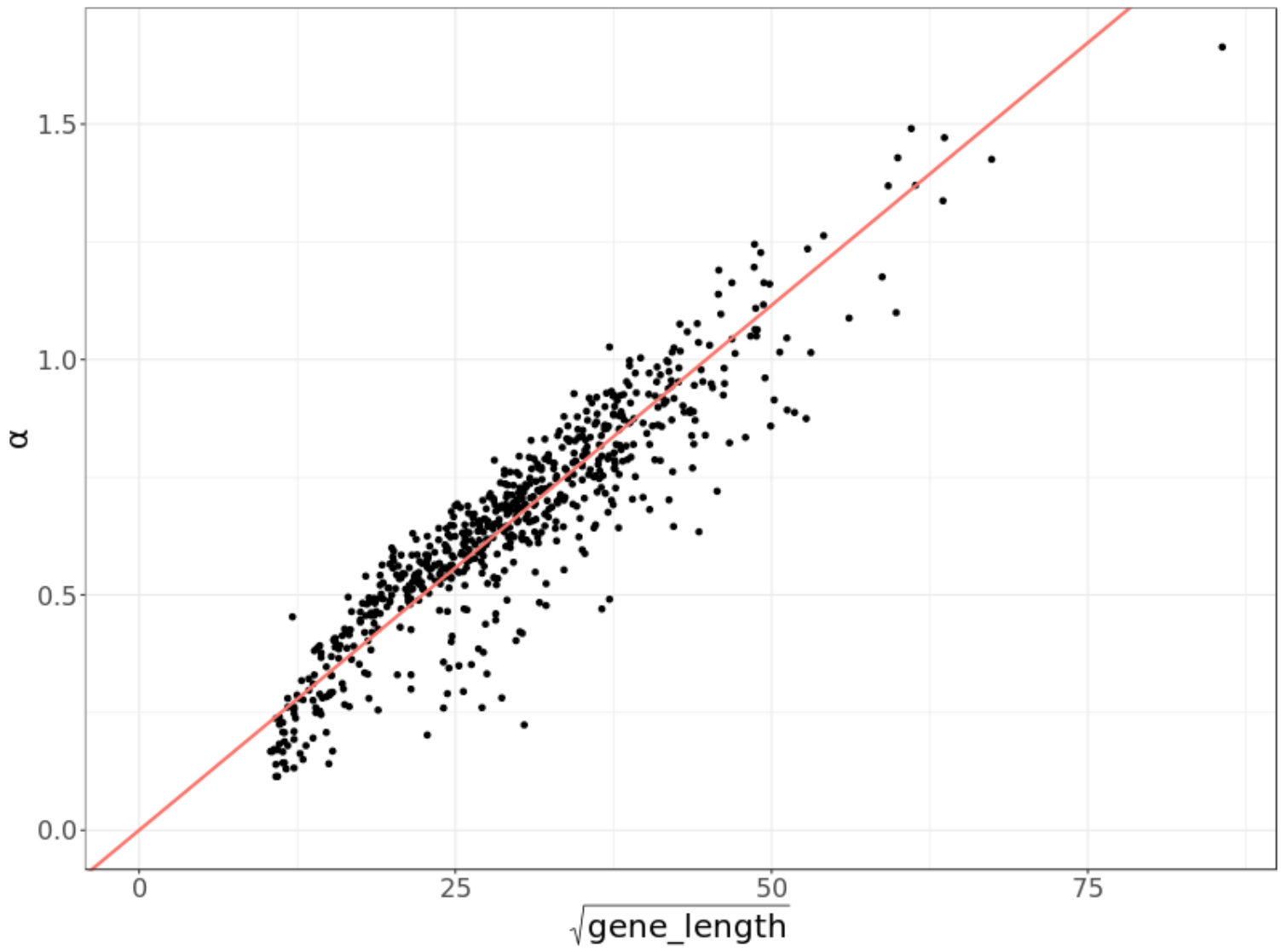
Supplementary Figure 18



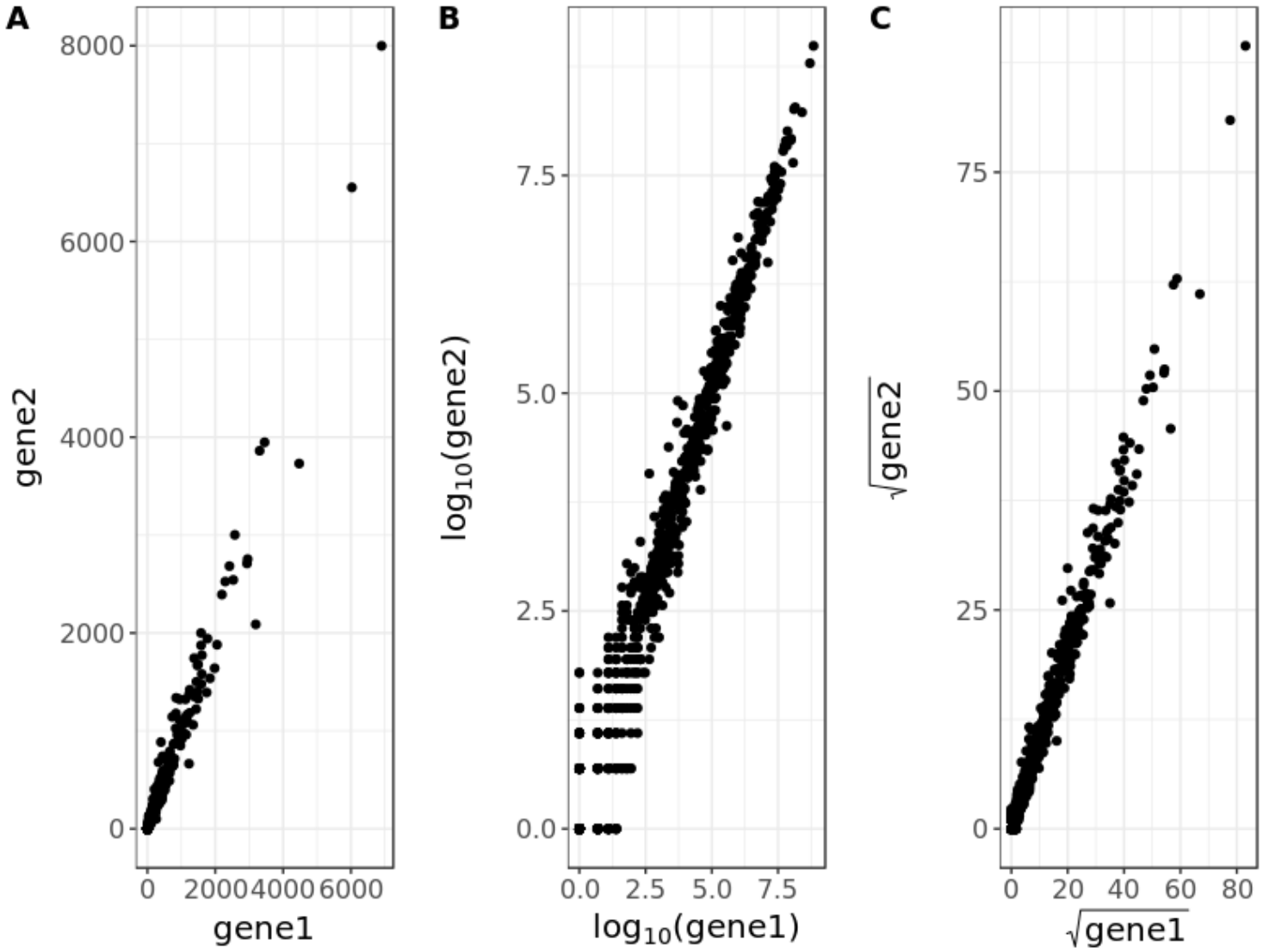
Supplementary Figure 19



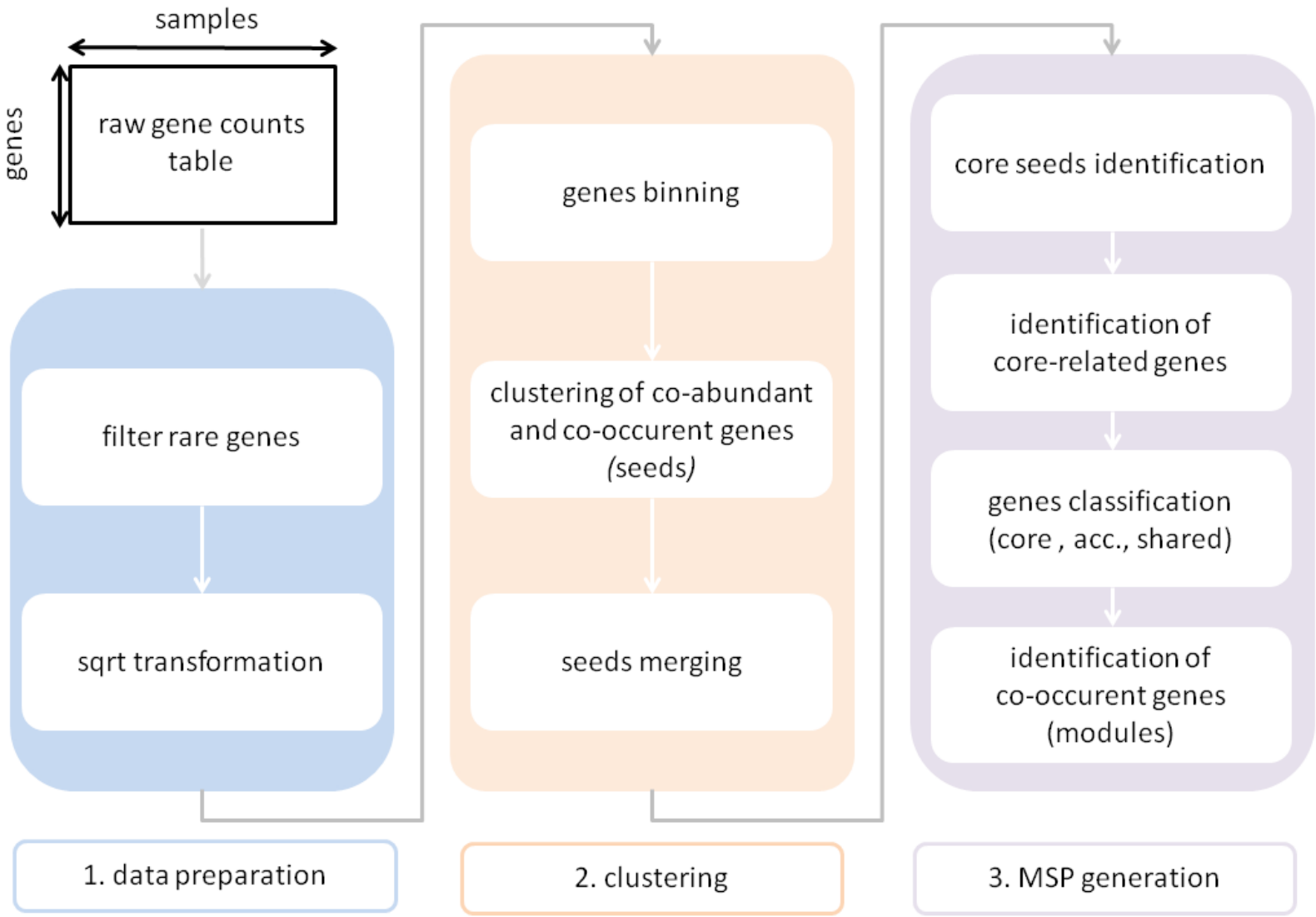
Supplementary Figure 20



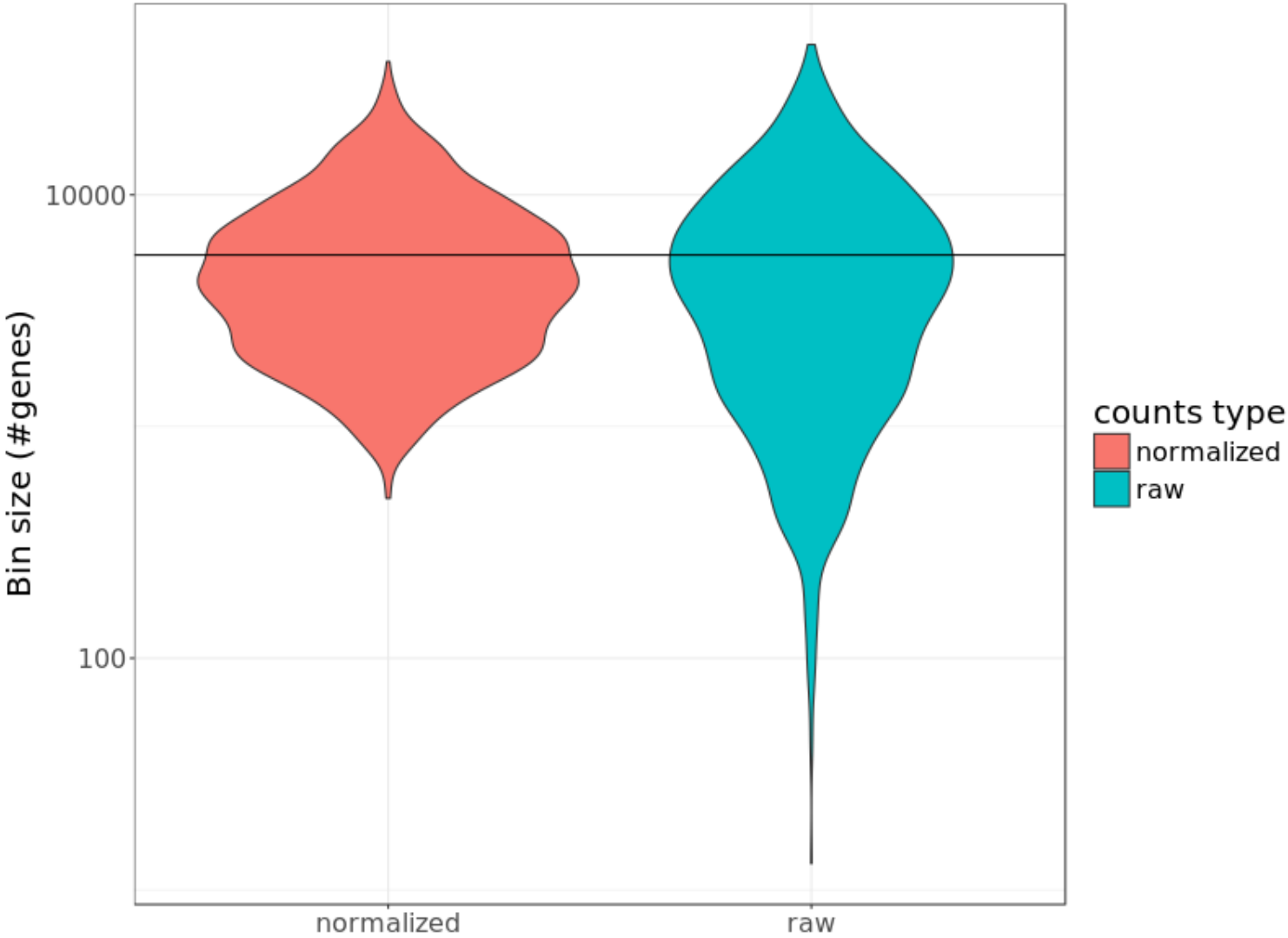
Supplementary Figure 21



Supplementary Figure 22



Supplementary Figure 23



Supplementary Figure 24

Supplementary Figures legends

Supplementary Figure 1 : Lowest taxonomic rank assigned to MSPs

(1 square = 1 MSP)

Supplementary Figure 2 : Most represented phyla among MSPs

Supplementary Figure 3 : Taxonomic homogeneity of MSPs

Only MSPs annotated at species level were kept.

Supplementary Figure 4 : Histogram of the total number of genes in MSPs

Supplementary Figure 5 : Comparison of the number of genes in a MSP and its number of accessory genes.

The outliers in red correspond to eukaryotic intestinal parasites of the *Blastocystis* genus.

Supplementary Figure 6 : Prevalence of the accessory genes of some MSPs among samples where their respective core genes are detected.

Supplementary Figure 7 : Comparison of the number of samples where a MSPs is detected and its number of accessory genes.

Supplementary Figure 8 : Prevalence of MSPs among the 1267 samples.

Supplementary Figure 9 : Comparison of the percentage of samples where a MSPs is detected and its mean abundance.

The red, blue and green points correspond respectively to the MSPs associated to *Bacteroides vulgatus*, *Bacteroides uniformis* and *Prevotella copri*

Supplementary Figure 10: 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.

978 **Supplementary Figure 11: Percentage of genes from reference genomes grouped in**
979 **their best corresponding MSPs.**

980 **Supplementary Figure 12: Description of the MSPs with no equivalent among CAGs.**

981 A. Histogram of the prevalence of the MSPs with no equivalent among CAGs.
982 B. Histogram of the number of genes in the MSPs with no equivalent among CAGs.

983 **Supplementary Figure 13 : Relative abundance of the top 10 discriminant MSPs**
984 **between chinese and westerners.**

985 **Supplementary Figure 14 : Relative abundance of the top 10 discriminant MSPs**
986 **between americans and europeans.**

987 **Supplementary Figure 15 : The 40 accessory genes (in column) of the msp_0039**
988 **(Prevotella copri) most associated with geographical origin of samples (in row).**
989

990 **Supplementary Figure 16 : Example of genes subject to copy-number variation.**

991 A. Comparison of the core genome median abundance of the msp_0011
992 (*Parabacteroides distasonis*) to one of its core genes (MH0150_GL0074117) subject to
993 Copy-Number Variation. Here, MSPminer wrongly estimates the coefficient of
994 proportionality as there is no dominant gene copy number among the samples.
995 B. Comparison of the core genome median abundance of the msp_0070 (*Eubacterium*
996 *rectale*) to one of its core genes (BGI-17A_GL0065038) subject to Copy-Number
997 Variation. Here, MSPminer correctly estimates the coefficient of proportionality for
998 the majority of samples (blue points) that have the same number of copies of the gene.
999 The others are classified as outliers (purple points).

1000 **Supplementary Figure 17 : Example of genes shared between several MSPs.**

1001 The gene SZEY-26A_GL0070503 is classified as shared accessory in the msp_0140
1002 (*Bifidobacterium pseudocatenulatum*) and the msp_0255 (*Bifidobacterium adolescentis*)

1003 A. Comparison of the core genome median abundance of the msp_0140 to the gene
1004 SZEY-26A_GL0070503. Despite the presence of outliers, a direct proportionality

1005 relationship is observed in many samples where both the msp and the gene are
1006 present.

1007 B. Comparison of the core genome median abundance of the msp_0255 to the gene
1008 SZEY-26A_GL0070503. Despite the presence of outliers, a direct proportionality
1009 relationship is observed in many samples where both the msp and the gene are
1010 present.

1011 C. Comparison of the core genome median abundance of the msp_0140 to the core
1012 genome median abundance of the msp_0255. Although the two MSPs share a gene,
1013 they have distinct signals.

1014 **Supplementary Figure 18 : Memory usage of MSPminer.**

1015 A. RAM consumption as a function of the number of genes The number of samples was
1016 set at 1267. The grey line represents the best linear fit (intercept=5.05, slope=5.17).

1017 B. RAM consumption as a function of the number of samples. The number of genes was
1018 set at 9 879 896. The grey line represents the best linear fit (intercept=6.14,
1019 slope=3.92).

1020 **Supplementary Figure 19 : Parallel performance of MSPminer**

1021 The tests were performed on a server with an Intel® Xeon® processor E5-2630 (6 physical
1022 cores with hyperthreading) and 128Gb of RAM.

1023 **Supplementary Figure 20 : Performance of MSPminer with or without genes**
1024 **binning.**

1025 A. Total number of comparisons performed by comparing all genes pairwise with or
1026 without binning. Here, genes binning divides the number of comparisons performed
1027 by 763.

1028 B. Fraction of useful comparisons for MSP cores creation with or without binning. The
1029 fraction without binning was estimated by drawing randomly one billion pairs of genes.
1030 Here, genes binning multiplies the fraction of useful comparisons by 562.

1031 **Supplementary Figure 21 : Comparison of the genes length and their respective**
1032 **coefficient of proportionality.**

1033 Core genes abundance profiles of the msp_0043 (*Ruminococcus bromii*) were compared to its
1034 core genome median abundance. For each gene, the estimated coefficient of proportionality
1035 α was compared to its length. Then, a robust linear regression with a null slope was performed
1036 (red line). As expected, the coefficient of proportionality α is proportional to the gene length.
1037 However, some genes have a coefficient of proportionality lower than expected (points below
1038 the line) and may correspond to genes not uniformly covered.

1039 **Supplementary Figure 22 : Impact of some data transformations on gene counts.**

1040 Raw (A), log-transformed (B) and square root transformed (C) abundance profiles of two core
1041 genes (MH0062_GL0071339 and MH0011_GL0069836) of the msp_0003 (*Bacteroides*
1042 *cellulosilyticus*) were compared.

1043 A. Raw data is not appropriate when the order of magnitude of counts is highly variable
1044 as it gives a too important weight to high counts.
1045 B. Log-transformation "spreads"the counts. However, it requires the introduction of a
1046 pseudo-count for zeros and variance tends to be proportional to the counts.
1047 C. Sqrt-transformation "spreads" the counts and stabilizes variance.

1048 **Supplementary Figure 23 : MSPminer workflow**

1049 **Supplementary Figure 24 : Number of genes per bin on the Integrated Gut Catalog**
1050 **with or without normalization.**

1051 Normalization consists in dividing gene counts by number of mapped reads per sample. The
1052 black line represents the optimal bin size (6 971 229/1 267=5 502 genes)