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

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 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

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 (Supplementary Figure 1).
170 However, the robust measure decreases significantly when the percentage of outliers is high
171 and the gene 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 3). 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 4). Finally, non-core genes observed in the same subset of samples are grouped into
178 modules of co-occurring genes.
179
180 We applied MSPminer to the largest publicly available gene abundance table provided with
181 the Integrated Gene Catalog of the human gut microbiome [12]. In this table, 9 879 896 genes
182 are quantified across 1 267 stool samples from individuals of various geographical origin
183 (Europe, USA and China) and diverse health status (healthy, obese, diabetic, with
184 inflammatory bowel disease etc.). 6 971 229 (70.6%) genes with counts greater than 6 in at
185 least 3 samples were kept. Among these, 3 262 914 (46.8%) were organized into 1 677
186 Metagenomic Species Pan-Genomes (MSPs) with at least 200 core genes (Supplementary
187 Table 1).

188 *Taxonomy*

189 By considering the lowest taxonomic rank assigned to MSPs, 278 (16.6%) were annotated at
190 species level, 85 (5%) at genus level, 119 (7.1%) at phylum level and the remaining 1 213 MSPs
191 (71.5%) could not be annotated (Supplementary Figure 2) indicating that a clear majority of
192 MSPs correspond to species not represented in reference genomes databases. Only 4 MSPs
193 were annotated as Eukaryotes and corresponded to intestinal parasites of the *Blastocystis*
194 genus. All others MSPs were assigned to prokaryotic species of which only 3 were Archaea and
195 corresponded to the species *Methanobrevibacter smithii*, *Candidatus Methanomethylophilus*

196 *alvus* and *Methanosphaera stadtmanae*. Among the MSPs annotated as Bacteria, the phyla
197 *Firmicutes* (268 MSPs), *Bacteroidetes* (118 MSPs), *Proteobacteria* (50 MSPs) and
198 *Actinobacteria* (29 MSPs) were the most represented (Supplementary Figure 3) as expected
199 from human gut metagenomes [43]. Some species were represented by multiple MSPs such
200 as *Faecalibacterium prausnitzii* (5 MSPs), *Bacteroides fragilis* (2 MSPs), *Methanobrevibacter*
201 *smithii* (2 MSPs) or *Hungatella hathewayi* (2 MSPs) suggesting a high nucleotide and gene
202 content variability between strains attributed so far to the same species. Conversely, some
203 MSPs had their core genes attributed to multiple species, usually from the same genus
204 (Supplementary Table 3). Although some inconsistencies in the taxonomy assignment have
205 previously been reported [22,44], it is possible that several MSPs regroup genes from highly
206 related species with high average nucleotide identity and similar gene content. MSPs
207 annotated at species level had a consistent taxonomical annotation as 97% of the core genes
208 (median) were assigned to the same species (Supplementary Figure 4). Consistency of the
209 taxonomic annotation was lower for accessory genes (62%) caused mainly by unannotated
210 genes.

211 *MSP content*

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

227 annotated at species level had accessory genes previously unobserved in available genomes
228 (Supplementary Table 2)

229 *Prevalence*

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

242 *Census of universal single copy marker genes*

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

251 *Comparison to sequenced genomes*

252 We compared the MSPs to 642 sequenced genomes for which at least 10% of their constituent
253 genes were detected in the Integrated Genes Catalog of the human gut microbiome [12]
254 (Supplementary Table 4). In total, these genomes covered 398 species representing 114
255 different genera. 624 (97.1%) were unambiguously assigned to 281 different MSPs. By keeping
256 only one representative per species, 47.1% (resp. 60.4%) of the genomes had at least 75% of

257 their genes grouped in their corresponding MSP considering either all their genes or only those
258 that were in the catalog (Supplementary Figure 12). In compliance with the results of the
259 taxonomic analysis, highly related species were assigned to the same MSP such as *Escherichia*
260 *coli*, *Escherichia fergusonii* and all genomes of the *Shigella* genus. Conversely, some species
261 grouping highly divergent strains were represented by several MSPs including
262 *Faecalibacterium prausnitzii* or *Bacteroides fragilis.*
263
264 To give another perspective on the MSPs, we compared the complete genome of
265 *Parabacteroides distasonis* ATCC 8503 [52] to its corresponding MSP (Figure 7). Among the 3
266 850 genes predicted in the genome, 3 781 (98%) had at least a close homolog in the Integrated
267 Gene Catalogue and 3 442 (89%) were found in the msp_0011. As expected, almost all the
268 core genes from the MSP were found in the genome (1 867 / 1 921, 97%), as well as accessory
269 genes with a prevalence higher than 80% (522 / 599, 87%). Only a small fraction of less
270 prevalent accessory genes was found in the genome (1371 / 5 090, 27%) (Supplementary
271 Figure 13). Genes grouped in the same modules tended to be physically close, in coherence
272 with genome organization of prokaryotes (Supplementary Figure 14). Remarkably, some
273 singleton genes were surrounded by genes from the same module, which shows that the
274 stringent grouping criteria used by MSPminer may split genes with slightly different
275 presence/absence patterns into different modules while they could have been grouped
276 (Supplementary Figure 15). Finally, few genomic regions contained genes that were not
277 assigned to the MSP (Supplementary Table 5A). Interestingly, some of these regions were
278 annotated as mobile elements (Supplementary Table 5B). Although some could be false
279 negatives, many were appropriately excluded as they were observed in too few samples or
280 their counts did not meet the proportionality-based grouping criterion (Supplementary Figure
281 16).

282 **Comparison to the Canopy clustering algorithm**

283 The Canopy clustering algorithm [38] was compared to MSPminer by applying both tools to
284 the metagenomic dataset described above. In total, Canopy grouped 2 691 408 genes into 3
285 463 Co-Abundance gene Groups (CAGs) while MSPminer grouped 3 267 132 genes (+17.6%)
286 into 1 677 MSPs (~ two-fold less objects).
287

288 178 MSPs encompassing 154 617 genes had no equivalent among the CAGs. Most MSPs were
289 rare as 75% were detected in fewer than 5 samples but had a significant size as 50% were
290 composed of at least 700 genes (Supplementary Figure 17). Remarkably, for 75% of them, the
291 3 samples with the highest counts represented at least 90% of the sum of MSP abundance on
292 all the samples. By default, Canopy discards such cases to avoid detection of spurious
293 correlations but MSPminer limits this risk by applying a variance-stabilizing transformation
294 and a stringent association criterion. In addition, Canopy grouped most core genes of a MSP
295 into a single CAG while many accessory genes were missed or assigned to small separate CAGs
296 (Figure 5). In agreement with the results of the simulation, most of the missed accessory genes
297 had a medium or low prevalence. As they contained many unexpected zeros, the correlations
298 with the core of their respective species were below the limit set in Canopy (Figure 6).

299 **MSPs quantification for biomarkers discovery**

300 To demonstrate that MSPminer was useful for biomarkers discovery, we first looked for
301 differentially abundant MSPs according to the geographical origin of samples. We discovered
302 94 MSPs differentially abundant between Westerners and Chinese (q-value < 10^{-3}, log2 fold
303 change ≥ 1) including 72 more abundant in Westerners and 22 in Chinese (Supplementary
304 Figure 18 and Supplementary Table 6). Among the discriminant MSPs, all those assigned to
305 the *Proteobacteria* phylum (*Klebsiella pneumoniae*, *Escherichia coli* and *Bilophila*
306 *wadsworthia*) were more abundant in Chinese which is consistent with previously published
307 results [12]. Interestingly, two MSPs assigned to *Faecalibacterium prausnitzii* were significant
308 but one was more abundant in Westerners and the other in Chinese. This shows that some
309 strains of this species are associated with geographical origin of samples. In addition, we
310 discovered 75 MSPs differentially abundant between Europeans and Americans (q-value <
311 10^{-3}, log2 fold change ≥ 1) of which 70 were more abundant among Europeans.
312 (Supplementary Figure 19 and Supplementary Table 6). This result is consistent with previous
313 studies showing lower gut microbiota diversity among Americans compared to Europeans
314 [33].
315
316 Secondly, we checked if MSPs could be used to perform strain-level analysis. To do this, we
317 tested if some accessory genes in the MSPs were more prevalent in samples of a given
318 geographical origin. By way of example, we found 680 accessory genes associated with

319 geographical origin (chi-squared test, p-value < 10^{-10}) in the msp_0011 corresponding to
320 *Parabacteroides distasonis* (Supplementary Figure 20 and Supplementary Table 7).
321 Remarkably, genes involved in cell filamentation (V1.UC58-4_GL0042624 and V1.UC18-
322 0_GL0014340) were more prevalent in Chinese than Westerners. More generally, many
323 significant genes were more prevalent among Westerners or Chinese although European and
324 American samples were separated in the analysis. This result suggests that strains of
325 *Parabacteroides distasonis* carried by Chinese are distinct from those of Westerners.

326 Discussion

327 Direct proportionality hypothesis and limits

328 MSPminer relies on a new robust measure to detect genes with directly proportional counts.
329 Even if this relation is more stringent than those assessed by Pearson's or Spearman's
330 correlation coefficients, it was successfully used to reconstitute Metagenomic Species Pan-
331 genomes of the human gut microbiota. In fact, most genes from sequenced genomes were
332 grouped into a single MSP showing that direct proportionality is the most common relation
333 between genes from the same biological entity. MSPminer misses some genes for which
334 counts are not ruled by this relation. Indeed, proportionality is disrupted when gene copy
335 number varies across samples [14] (Supplementary Figure 21), when a sample contains
336 multiples strains [19,20] and when a gene is shared between several MSPs (Supplementary
337 Figure 22) because of horizontal gene transfer [53] or grouping of highly similar orthologs.
338 Nevertheless, the first two cases have most likely a limited impact as the majority of strains
339 tend to have the same gene copy numbers [14] and samples often carry a dominant strain
340 [20]. Regarding shared genes, their signals are a linear combination of the MSPs that carry
341 them. Thus, they will be identified only if these MSPs are mostly detected in separate sets of
342 samples.

343 Computing performance

344 MSPminer can process large datasets made up of thousands of samples and millions of genes
345 in just a few hours on a regular single node server. The program is only limited by the amount
346 of RAM available on the machine on which it is executed as the input count matrix must be
347 fully loaded into memory. Consequently, RAM consumption grows linearly according to the
348 number of samples and genes in the count matrix (Supplementary Figure 23).

349 MSPminer achieves good parallel efficiency (Supplementary Figure 24) through two
350 parallelization strategies. First, a novel Map/Reduce programming model assigns genes to as
351 many subsets as the number of available samples. Genes with greatest counts in the same
352 sample are first compared, which not only decreases the number of comparisons to perform
353 but increases the probability that related genes are placed in the same bin compared to
354 random assignment (Supplementary Figure 25). Each subset of genes is processed in parallel
355 and synchronization is only required before the reduction/merging step (see Methods). To
356 avoid comparisons of all pairs of genes, others used an iterative algorithm where a random
357 gene is compared to all the others until a significant proportion of genes is clustered [38]. This
358 method allows fast identification of big clusters but struggles with the large number of small
359 clusters and singleton genes. Furthermore, parallelism implementation is more complex as
360 synchronization is required to detect duplicate clusters. In a second phase, MSPminer
361 performs pairwise comparison of clusters to detect those corresponding to MSPs core gene
362 sets. Then, full MSPs including accessory and shared genes are retrieved in parallel from the
363 signal of their respective core genes. Here, no synchronization is required as core genes sets
364 are supposedly independent.

365 **Quality of MSPs**

366 The quality of the MSPs is impacted by all the upstream steps required for generating the
367 count matrix, as well as with the biological and ecological characteristics of the dataset. At the
368 sequencing level, the number of reads (sequencing depth) generated for each sample impacts
369 the detection and coverage of subdominant species, while reads length affects the quality of
370 the assembly and the ability to assign a read to a gene without ambiguity. At the
371 bioinformatics level, assembly, gene prediction, gene redundancy removal, mapping and
372 counting require expertise to select the most appropriate strategies, tools and parameters.
373 Indeed, assemblers returning chimeric contigs which combine sequences from highly related
374 species, inaccurate predictors generating truncated or merged genes, redundancy removal
375 with a common threshold for all genes (95% of nucleotide identity) lead to genes of variable
376 quality in catalogues. Then, when quantifying genes, keeping only uniquely mapped reads
377 underestimates the abundance of some genes whereas considering shared reads can generate
378 false positives. Genes grouped in MSPs were significantly longer than those that were not
379 (median length of 780 bp vs 498 bp, Wilcoxon rank-sum test p-value= 0) (Supplementary

380 Figure 26) as longer genes have a higher and less dispersed counts. Nevertheless, end-to-end
381 mapping probably plays a role as it may fail to fully align a read whose size is approximately
382 the same as the target gene. Finally, at the biology level, a high number of samples with varied
383 phenotypes will improve the comprehensiveness and quality of MSPs. Indeed, as the number
384 of samples grows, MSPminer will identify rare species and will extend the list of accessory
385 genes of the MSPs corresponding to species with an open pan-genome. In addition, highly
386 prevalent accessory genes will be reclassified from core to accessory as observed while
387 sequencing an increasing number of strains of a species [49].

388 **Applications**

389 As illustrated in this paper, MSPminer supports the analysis of metagenomic data at species-
390 level by identifying and quantifying the MSPs present in samples. Subsequently, MSPs
391 associated with a given phenotype (e.g. the geographical origin) can be investigated both
392 quantitatively and qualitatively. Here, information from unknown or non-sequenced species
393 can be exploited. Compared to methods relying on marker genes [33,34], MSPminer improves
394 the estimation of species abundance by automatically detecting among core genes those with
395 the highest specificity, the highest counts and lowest dispersion.
396 Moreover, in each MSP, genes or modules of accessory genes associated with the tested
397 phenotype can be explored opening the way to a strain-level analysis. Thus, biomarkers
398 corresponding to functional traits specific to certain strains can be discovered.
399
400 MSPminer also provides microbial population genetics from large cohorts which can support
401 culture-dependent methods by identifying species of particular interest, such as those with no
402 reference genome available or with reference genomes distant from the strains actually
403 present. Reciprocally, MSPminer will benefit from advances in culture-dependent methods
404 which provide reference genomes of low abundance species detectable by shotgun
405 sequencing but difficult to assemble [23,24].

406 **Further developments**

407 Several improvements of MSPminer are considered. Algorithms that could identify relevant
408 associations currently missed by MSPminer will be evaluated. For instance, deconvolution

409 algorithms [14] may discover genes shared between several MSPs while kernel density
410 estimators [54] may be useful for detecting genes with highly variable copy numbers.
411 To increase MSPminer specificity, a statistical test determining how much an association
412 between the core genome of a MSP and a gene is unexpected would be highly useful. Such a
413 statistical test could assess both co-occurrence with a Fisher exact test and co-abundance with
414 correlation test or equivalent. Currently, zero counts are either classified as structural or
415 undetermined while comparing genes abundance profiles. However, a statistical model
416 determining the probability that a zero is structural would allow overcoming threshold effects
417 and classifying with more accuracy a gene as core, accessory or shared.
418 Alternatives to the median for computing the representative of a MSP are envisaged. The sum
419 which cumulates counts from multiple genes is a prime candidate, as clusters would be
420 quantified with higher accuracy and increased quantification range particularly in samples
421 where its abundance is low. However, one should carefully account for outliers
422 Finally, the robustness of the measure of proportionality could be improved. For instance, a
423 robust linear regression [55] may replace the median for estimating the coefficient of
424 proportionality between the abundance profiles of two genes while the median absolute
425 deviation (MAD) could improve the detection of outliers [56].

Methods

Measure of proportionality between two genes

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

Let $g_i = (c_{i1}, c_{i2}, \dots, c_{im})$ $i \in [1, n]$ 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:

$\sqrt{\alpha} = median\left(\frac{\sqrt{c_{ys}}}{\sqrt{c_{xs}}}\right)$ $\forall s \in [1, m]$ 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 [57,58]. For a comparison of some data transformations, refer to Supplementary Figure 28. The median was used to tolerate some outliers.

To estimate the coefficient of proportionality, only samples were both genes counts were above a threshold t were kept. This has the following advantages:

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

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

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

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

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

Non-robust version

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

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

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

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 4).

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

Simulation

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

The species pan-genome consisted of 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.

516 The theoretical number of counts attributed to each gene was calculated according to its
517 length and its presence or not in the strain. The observed gene counts were generated by
518 using a negative binomial distribution (mean=theoretical gene count; overdispersion=0.05) to
519 approach real metagenomic data.
520 Outliers were added in each gene by selecting 5%, 10% and 20% of their non-null samples and
521 multiplying each observed count by either ¼, ⅓, 2 ,3 or 4.

522 **Metagenomic Species Pan-genome generation**

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

524 *Input data*

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

528 *Data filtering*

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

531 *Data transformation*

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

534 *Genes binning*

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

539 *Seeds creation*

540 This step identifies sets of co-abundant and co-occurring genes called *seeds* hereafter.
541 Seeds were created in parallel in each bin by a greedy approach. First, genes were compared
542 pairwise. All pairs of genes with a non-robust measure of proportionality of at least 0.8 and
543 no structural zeros were saved in a list. Then, the list was sorted by decreasing measure of

544 proportionality. The pair of genes with the highest measure of proportionality was selected as
545 a centroid. Genes related to one of the centroid genes were grouped together in a new seed.

546 *Seed representative*

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

552 *Seeds merging*

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

561 *Core seeds identification*

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

570 *Metagenomic Species Pan-genome generation*

571 The representatives of each core seed were compared to all the genes. Because core seeds
572 were identified all at once in the previous step, the MSPs generation was run in parallel. Genes

with a robust measure of proportionality of at least 0.8 were considered as related to the core seed.

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

1. core genes were detected in the same samples as the core seed (Figure 4.A).

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

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

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

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

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

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 4.D).

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

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.

601 **Biomarkers discovery**

602 *Identification of MSPs associated with geographical origin*

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

608 *Identification of accessory genes associated with geographical origin*

609 For each accessory gene of a MSP, a 2x2 contingency table counting in both populations the
610 number of samples where the gene was present or absent was built. Only samples where the
611 MSP core genome was detected were kept. A gene was considered as present in a sample if
612 at least two reads were mapped on it. Then, a chi-squared test was performed on each
613 contingency table. Accessory genes with a p-value inferior to 10^{-10} were considered
614 significant. Log_2 presence ratios equal to +infinity or −infinity were replaced by +10 or −10
615 respectively.

616 **Taxonomic annotation of the gene catalog**

617 Genes were aligned at the nucleotide level using BLASTn [61] (version 2.6.0) against KEGG
618 GENOME [62] (Release 82.0, April 2017) and RefSeq [63] (Release 81, March 2017). Hits that
619 covered less than 80% of the query gene or with a e-value superior to 0.01 were discarded.
620 Thresholds of 95%, 80% and 65% of nucleotide identity were respectively used for taxonomic
621 annotation at species, genus and phylum level.
622 At a given taxonomic level, “no consensus” was reported if the selected hits did not share the
623 same annotation. Finally, results from KEGG were preferred to those from RefSeq.
624
625 MSPs were assigned to the lowest level taxon representing more than 50% of the annotations
626 of their core genes. Taxonomic annotations of MSPs with at least 50% of their core genes
627 annotated at species level (including “no consensus”) and less than 80% assigned to the most
628 represented species were considered ambiguous (c.f. Supplementary Table 3)

629 **Functional annotation of the gene catalog**

630 Translated genes were annotated with eggnog-mapper [64] (version 0.12.7) based on eggNOG
631 orthology assignments [65]. Sequence similarity searches were performed using HMMER [66].
632
633 The 40 universal single copy marker genes were discovered using fetchMG v1.0 [67]

634 **Comparison of the MSPs to sequenced genomes**

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

Declarations

Funding

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

Authors' contributions

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

Additional files

Supplementary Table 1: Tab-separated file listing the genes and modules in the MSPs

Supplementary Table 2: XLS file describing the MSPs (taxonomic annotation, number of genes, number of universal marker genes, prevalence and abundance)

Supplementary Table 3: XLS file listing the MSPs with ambiguous annotation at species level.

Supplementary Table 4: XLS file summarizing the comparison of the MSPs to 642 sequenced genomes.

Supplementary Table 5: XLS file listing the genomic regions of *Parabacteroides distasonis* ATCC 8503 containing genes not assigned to the msp_0011.

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

Supplementary Table 7: XLS file listing the accessory genes of the msp_0011 (*Parabacteroides distasonis*) associated with the geographic origin of samples.

References

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

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

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

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

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

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

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

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

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

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

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

725 18. Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang Y-F. Analysis of Ultra Low Genome
726 Conservation in Clostridium difficile. Horsburgh MJ, editor. PLoS One [Internet].
727 2010;5:e15147. Available from: http://dx.plos.org/10.1371/journal.pone.0015147

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

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

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

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

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

745 24. Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously
746 uncultured members of the human gut microbiota by culturomics. Nat. Microbiol. [Internet].
747 2016;1:16203. Available from: http://www.nature.com/articles/nmicrobiol2016203

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

750 26. Pop M. Genome assembly reborn: Recent computational challenges. Brief. Bioinform.
751 2009;10:354–66.

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

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

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

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

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

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

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

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

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

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

804 the human gut microbiome. Nature [Internet]. 2011;474:666–666. Available from:
805 http://www.nature.com/doifinder/10.1038/nature10187

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

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

812 46. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis
813 of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella*
814 *pneumoniae* , an urgent threat to public health. Proc. Natl. Acad. Sci. [Internet].
815 2015;112:E3574–81. Available from:
816 http://www.pnas.org/lookup/doi/10.1073/pnas.1501049112

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

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

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

830 50. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al.
831 Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res.

2012;22:299–306.

51. 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.

52. 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.

53. 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

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

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

56. 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.

57. 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

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

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

60. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. Supporting data for the paper: "An

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

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

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

865 63. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-
866 redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res.
867 2007;35.

868 64. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast
869 Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper.
870 Mol. Biol. Evol. [Internet]. 2017;278:631–7. Available from: https://oup.silverchair-
871 cdn.com/oup/backfile/Content_public/Journal/mbe/PAP/10.1093_molbev_msx148/3/msx1
872 48.pdf?Expires=1499795422&Signature=W7WZu1nfufNHcrf1QFCH~13BjwFo82LccucPBoCvq
873 GvLaClaQFIzzY6iT3cfR4A1bFPxYnhBaz6D-wBoorppxD5SfKpgd63KRkl2HHDcb2BEBArfT4f

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

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

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

880 68. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, et al. A
881 Catalog of Reference Genomes from the Human Microbiome. Science (80-.). [Internet].
882 2010;328:994–9. Available from:
883 http://www.sciencemag.org/cgi/doi/10.1126/science.1183605

884 69. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank.

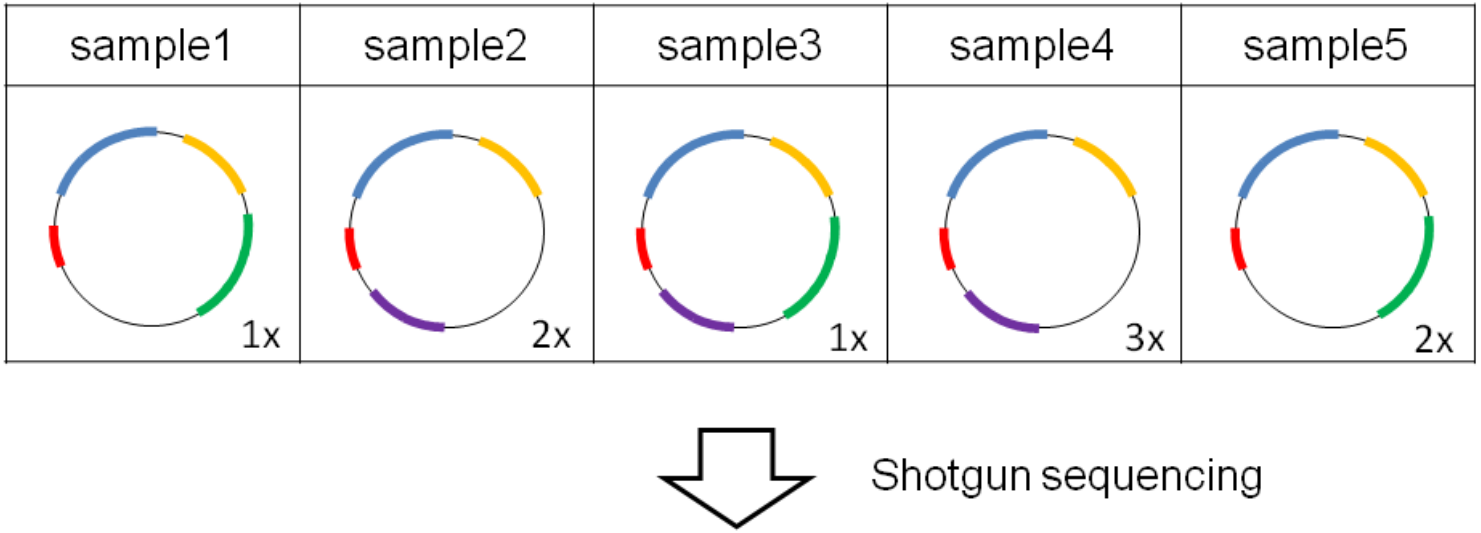
885 Nucleic Acids Res. 2013;41.

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

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

891

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

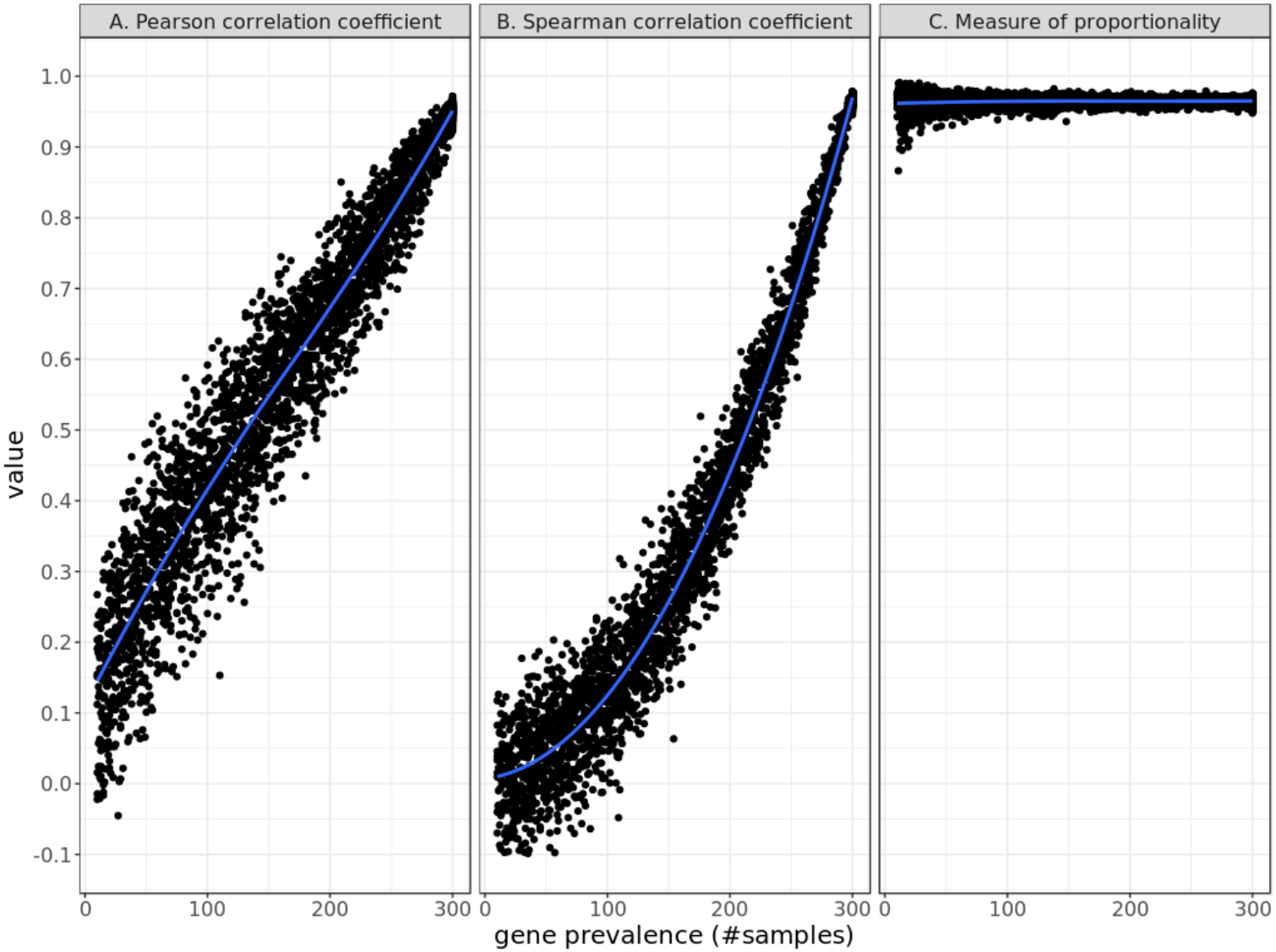

A. Pearson correlation coefficient
B. Spearman correlation coefficient
C. Measure of proportionality
value
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0
100
200
300
gene prevalence (#samples)

Figure 2

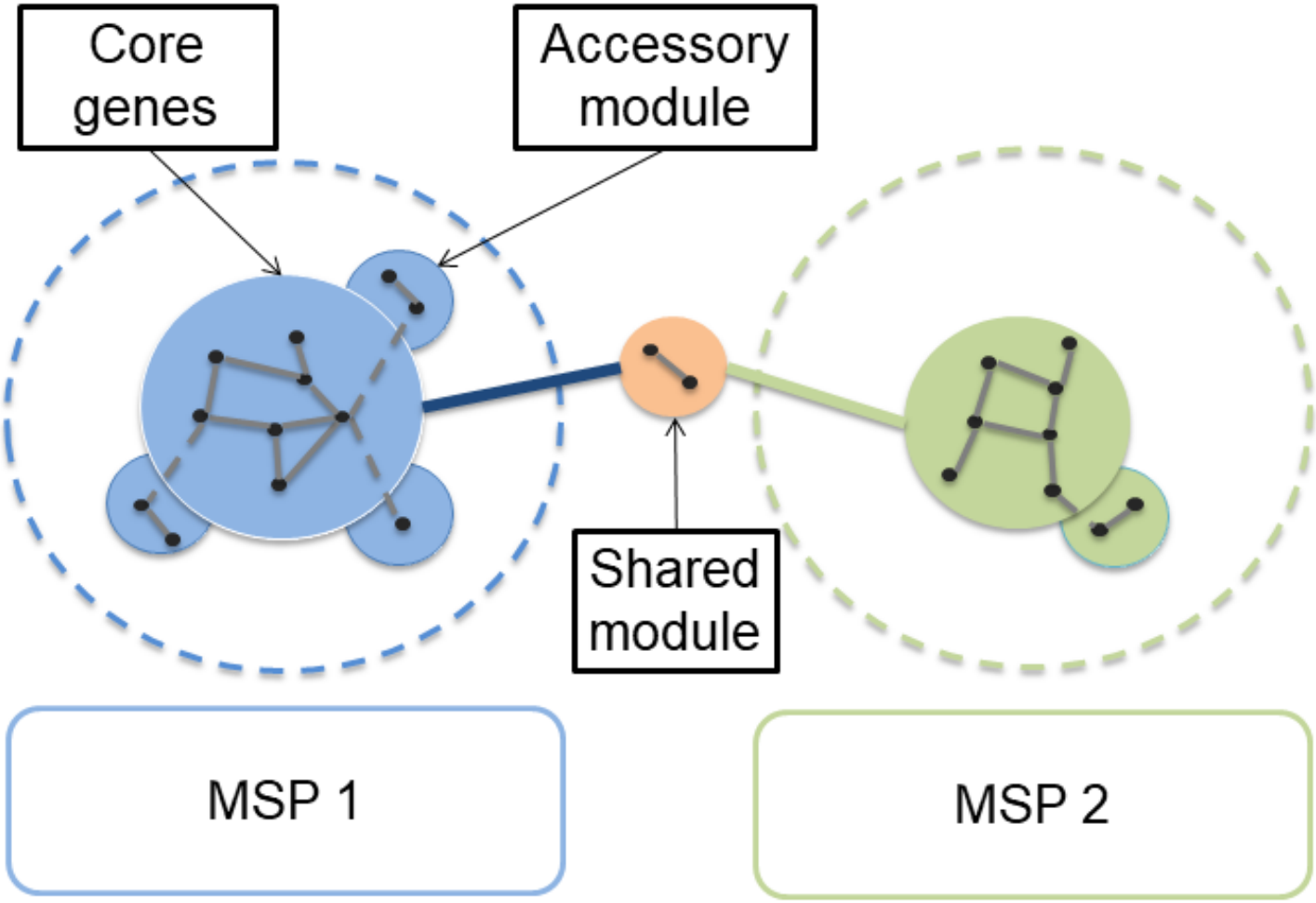


Figure 3

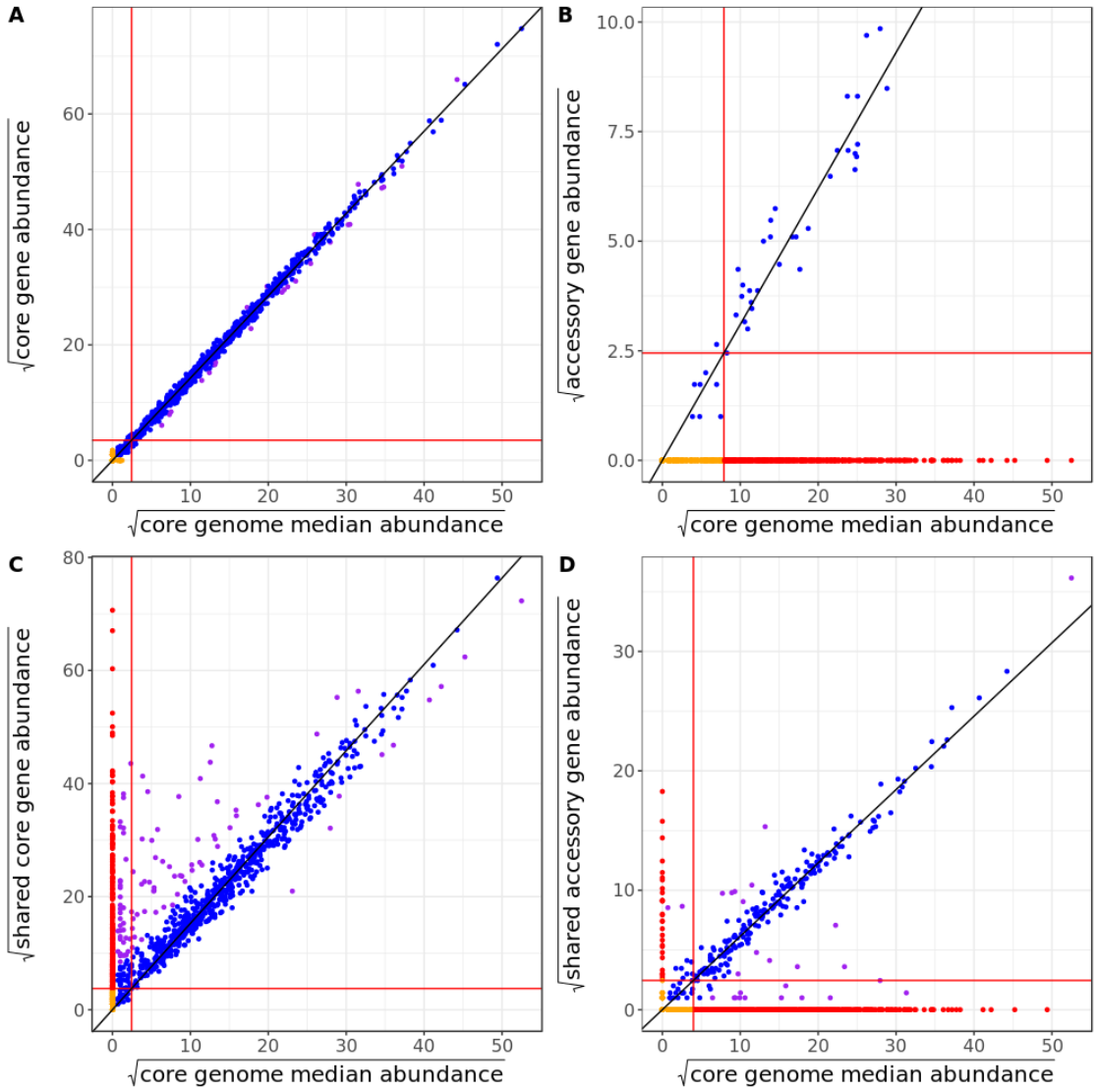

A
B
C
D
√core gene abundance
√accessory gene abundance
√shared core gene abundance
√shared accessory gene abundance
√core genome median abundance

Figure 4

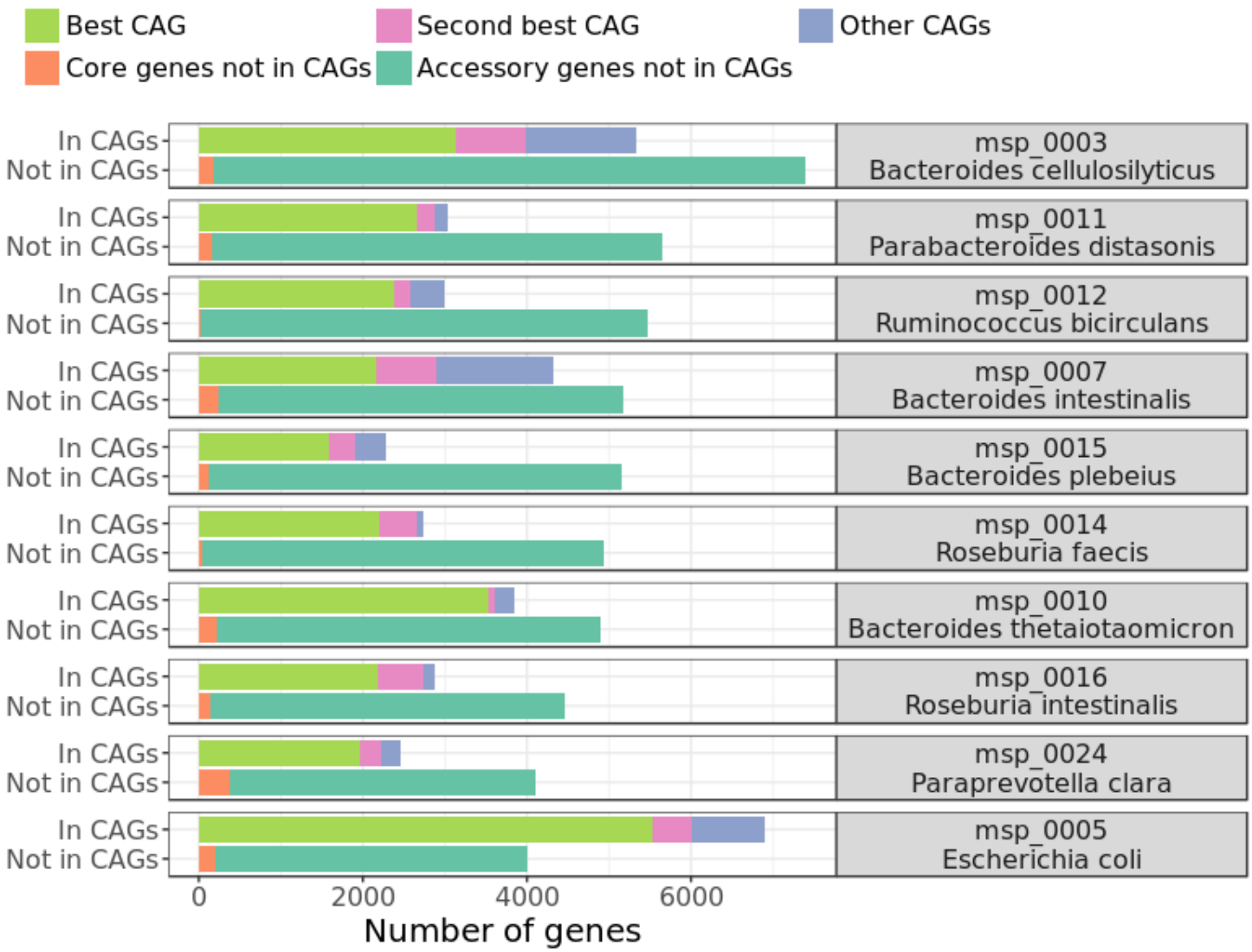

Best CAG
Second best CAG
Other CAGs
Core genes not in CAGs
Accessory genes not in CAGs
In CAGs
Not in CAGs
msp_0003
Bacteroides cellulosilyticus
msp_0011
Parabacteroides distasonis
msp_0012
Ruminococcus bicirculans
msp_0007
Bacteroides intestinalis
msp_0015
Bacteroides plebeius
msp_0014
Roseburia faecis
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

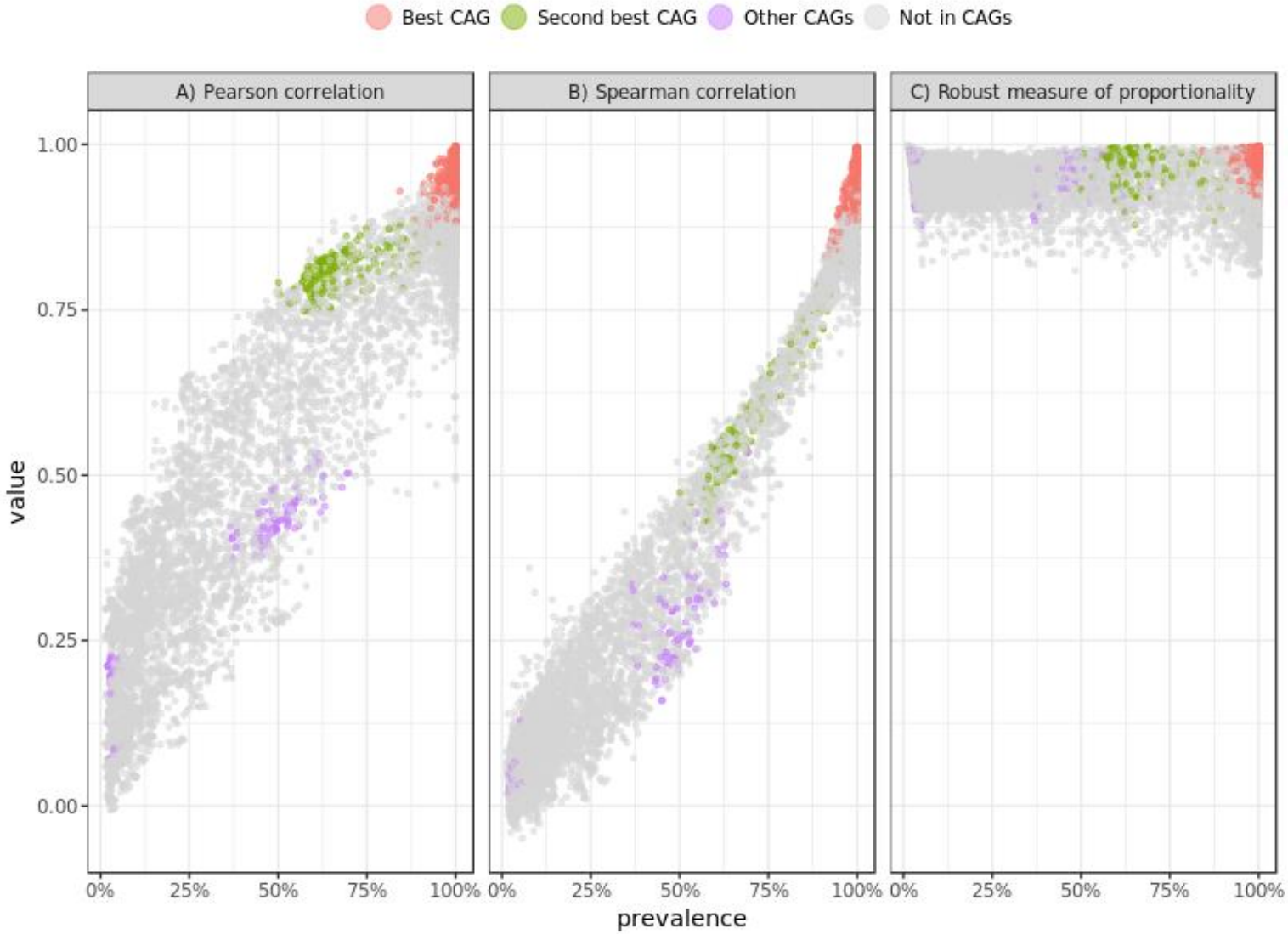

Best CAG
Second best CAG
Other CAGs
Not in CAGs
A) Pearson correlation
B) Spearman correlation
C) Robust measure of proportionality
value
1.00
0.75
0.50
0.25
0.00
0%
25%
50%
75%
100%
prevalence

Figure 6

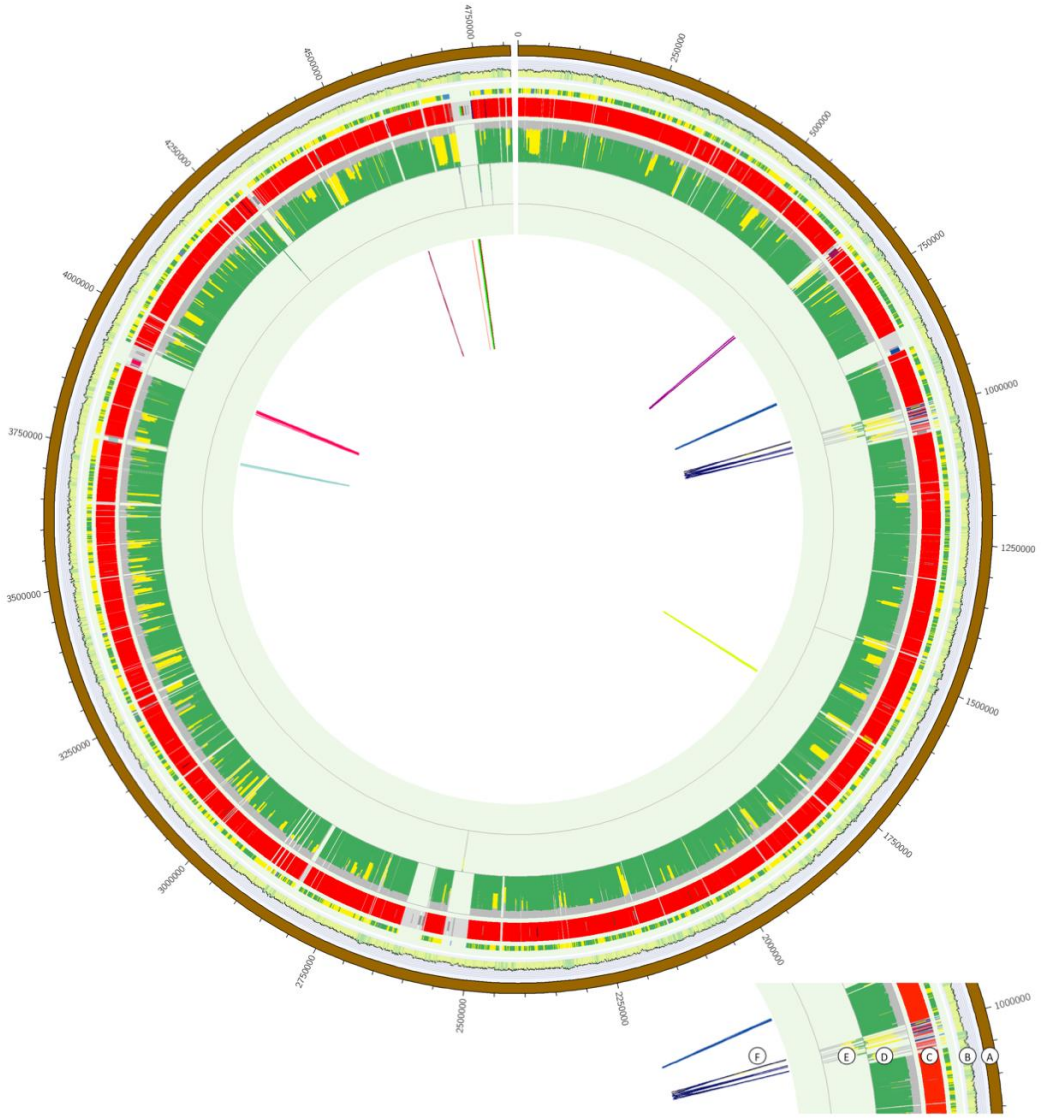

0
250000
500000
750000
1000000
1250000
1500000
1750000
2000000
2250000
2500000
2750000
3000000
3250000
3500000
3750000
4000000
4250000
4500000
4750000
1000000
F
E
D
C
B
A

Figure 7

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.

936 **Figure 5: Comparison of the gene content of some MSPs and their corresponding**
937 **CAGs**

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

940 Three measures are compared:

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

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

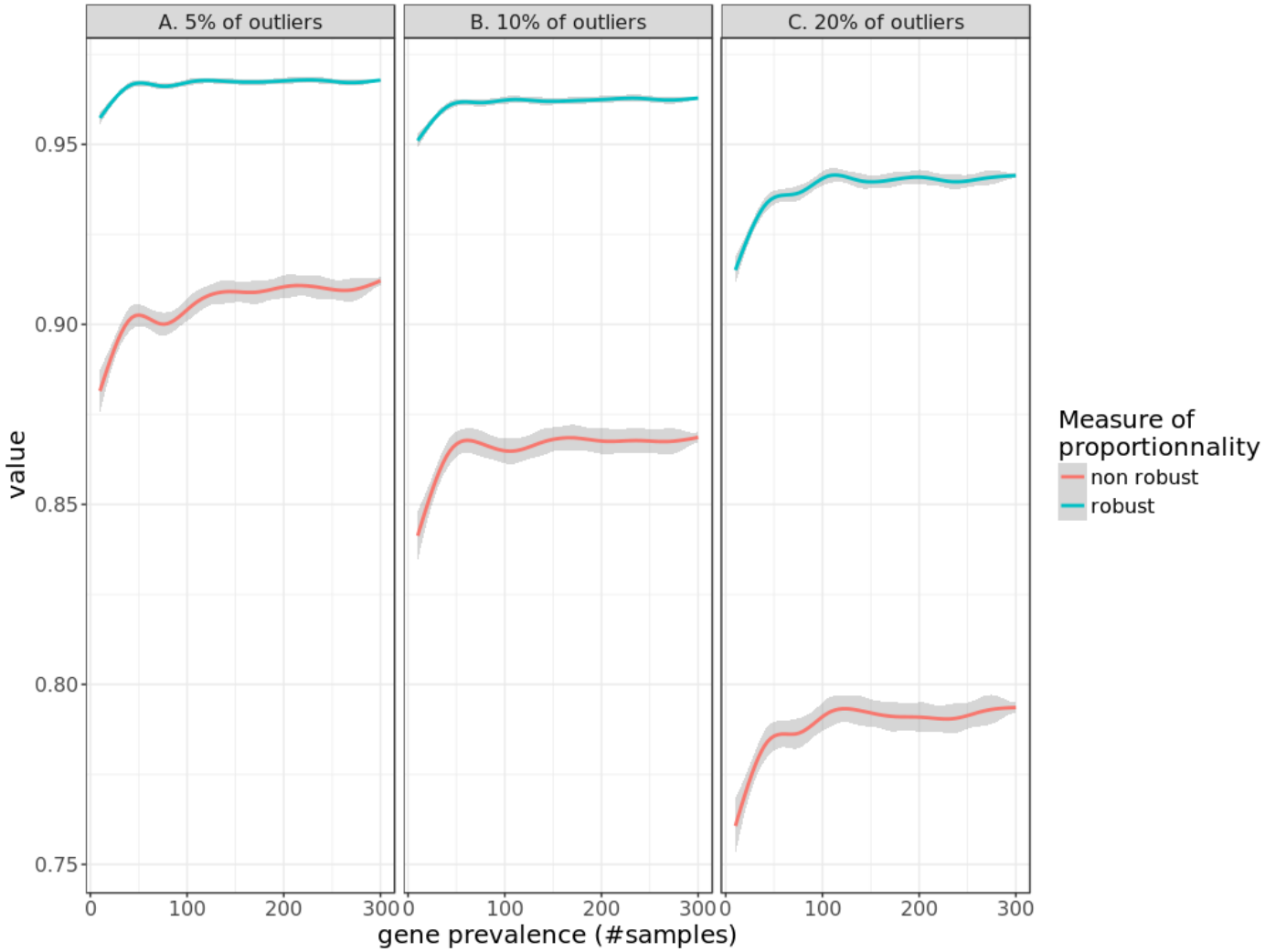
946 **Figure 7: Circos representation of the mapping of the msp_0011 on the genome of**
947 ***P. distasonis strain* ATCC 8503**

948 Description of layers from outside to inside:

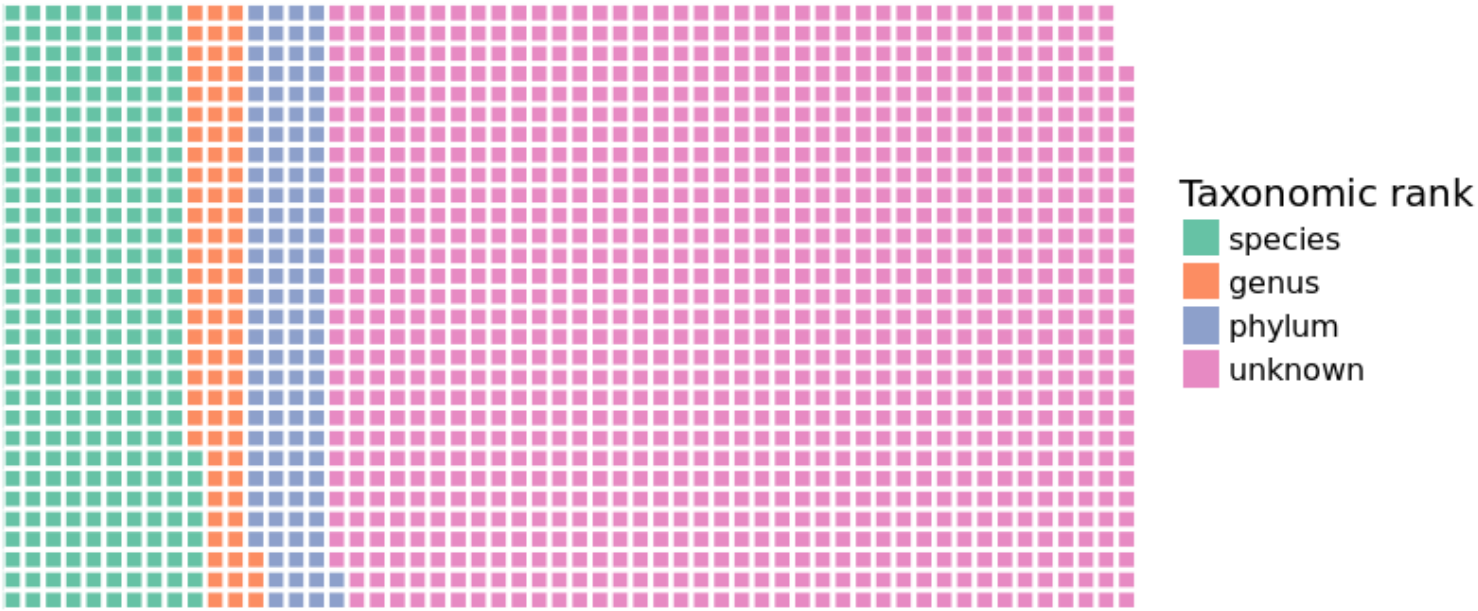
949 A. Position on chromosome
950 B. GC-content (format: histogram)
951 C. gene or module type (format: highlight):
952 - green: core
953 - yellow: accessory
954 - blue: shared core
955 - purple: shared accessory
956 D. MSP (format: highlight):
957 - Bandwidth:
958 · wide: gene grouped in a MSP
959 · narrow: gene grouped in a seed
960 - color code:
961 · red: gene grouped in the most represented MSP
962 · other color + grey: gene grouped in another MSP or a seed
963 E. Sample assignment (format: histogram):
964 - facing outwards if the gene is related to the most represented MSP, facing
965 inwards otherwise.

966 - color code:
967 - grey: samples where the MSP module is not detected
968 - green: samples where the MSP core and the gene are detected
969 - yellow: samples where the MSP core is detected but not the gene
970 - purple: samples where the gene is detected but not the MSP core
971

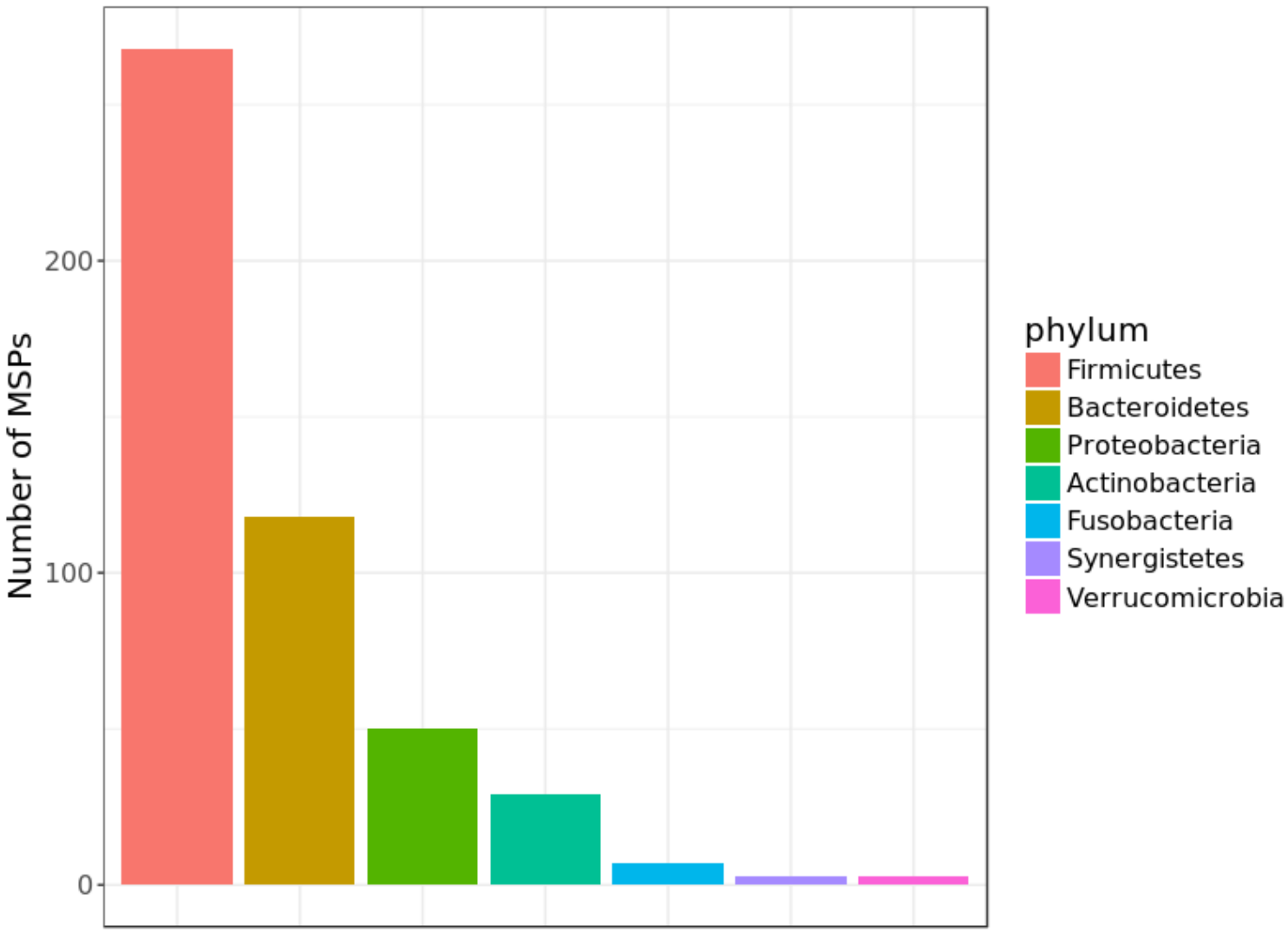
Supplementary Figures



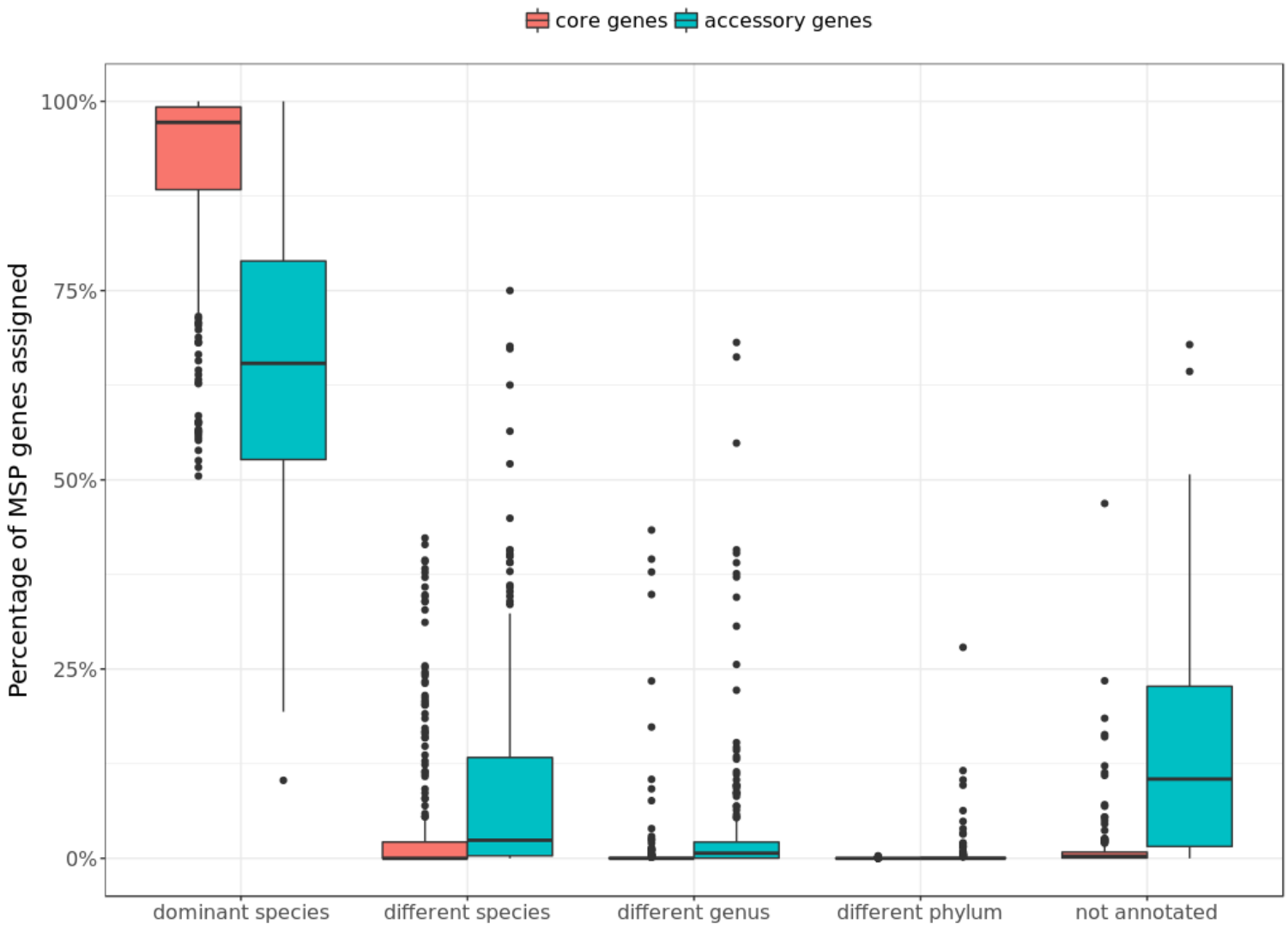
Supplementary Figure 1



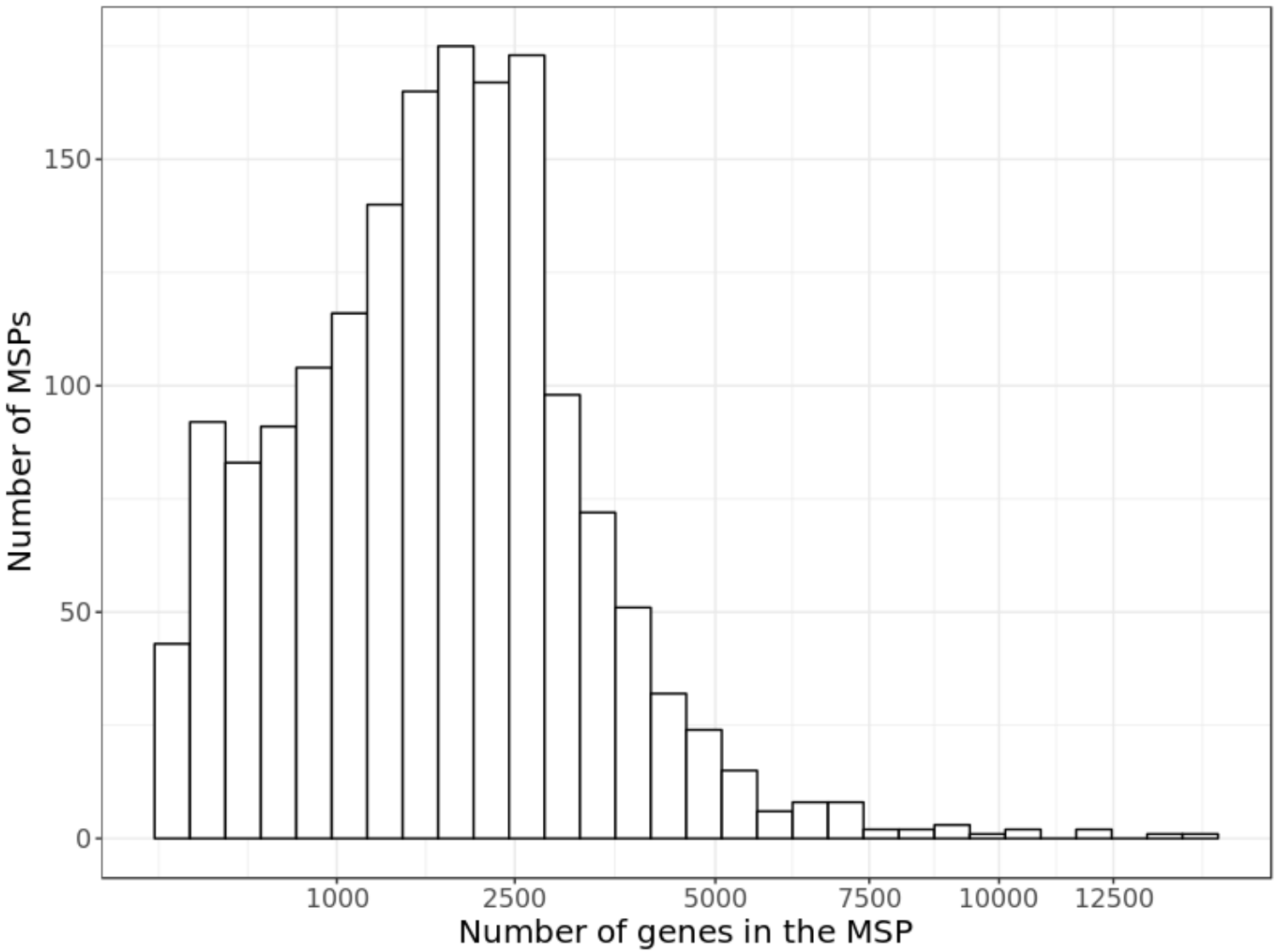
Supplementary Figure 2



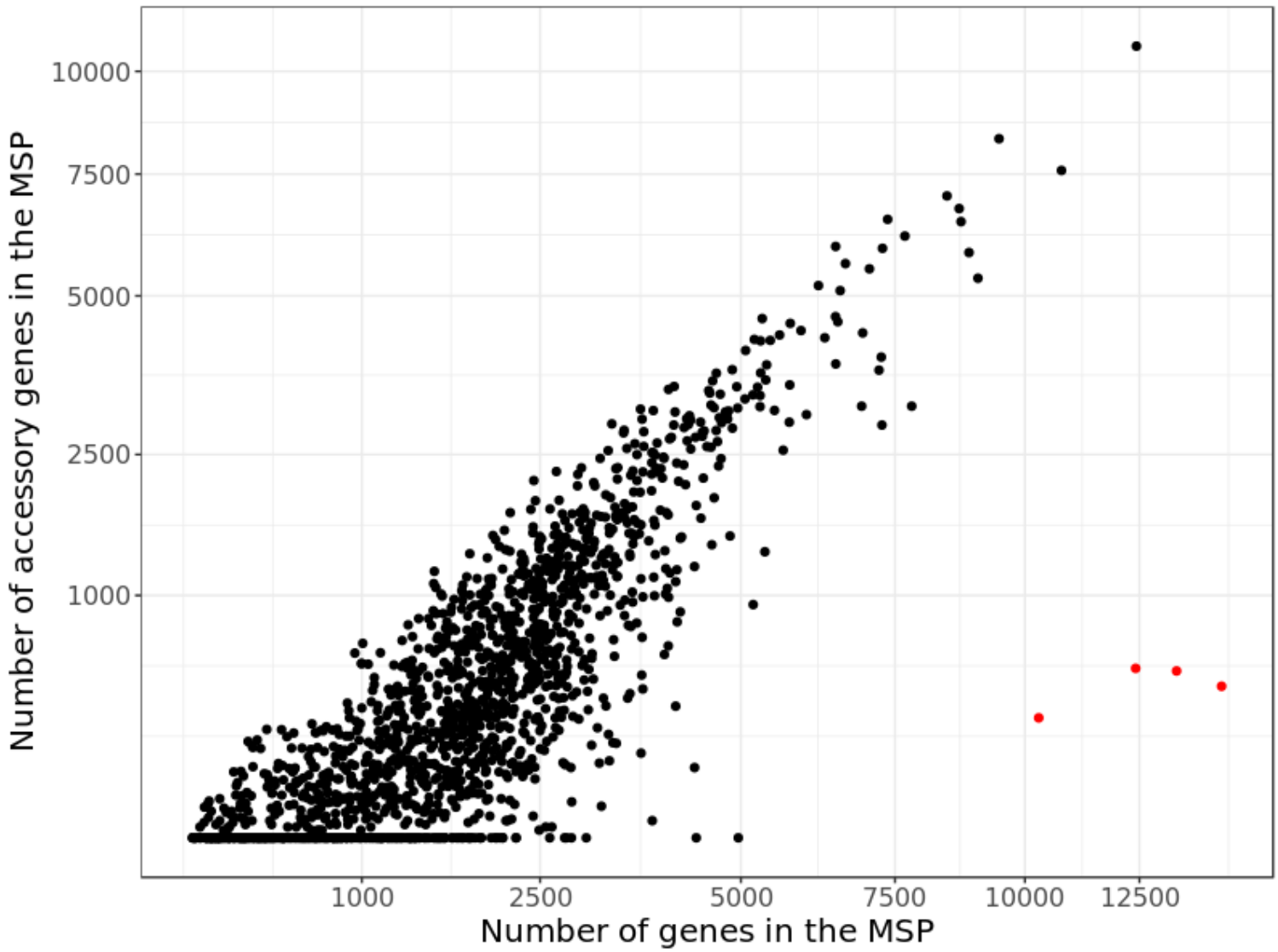
Supplementary Figure 3



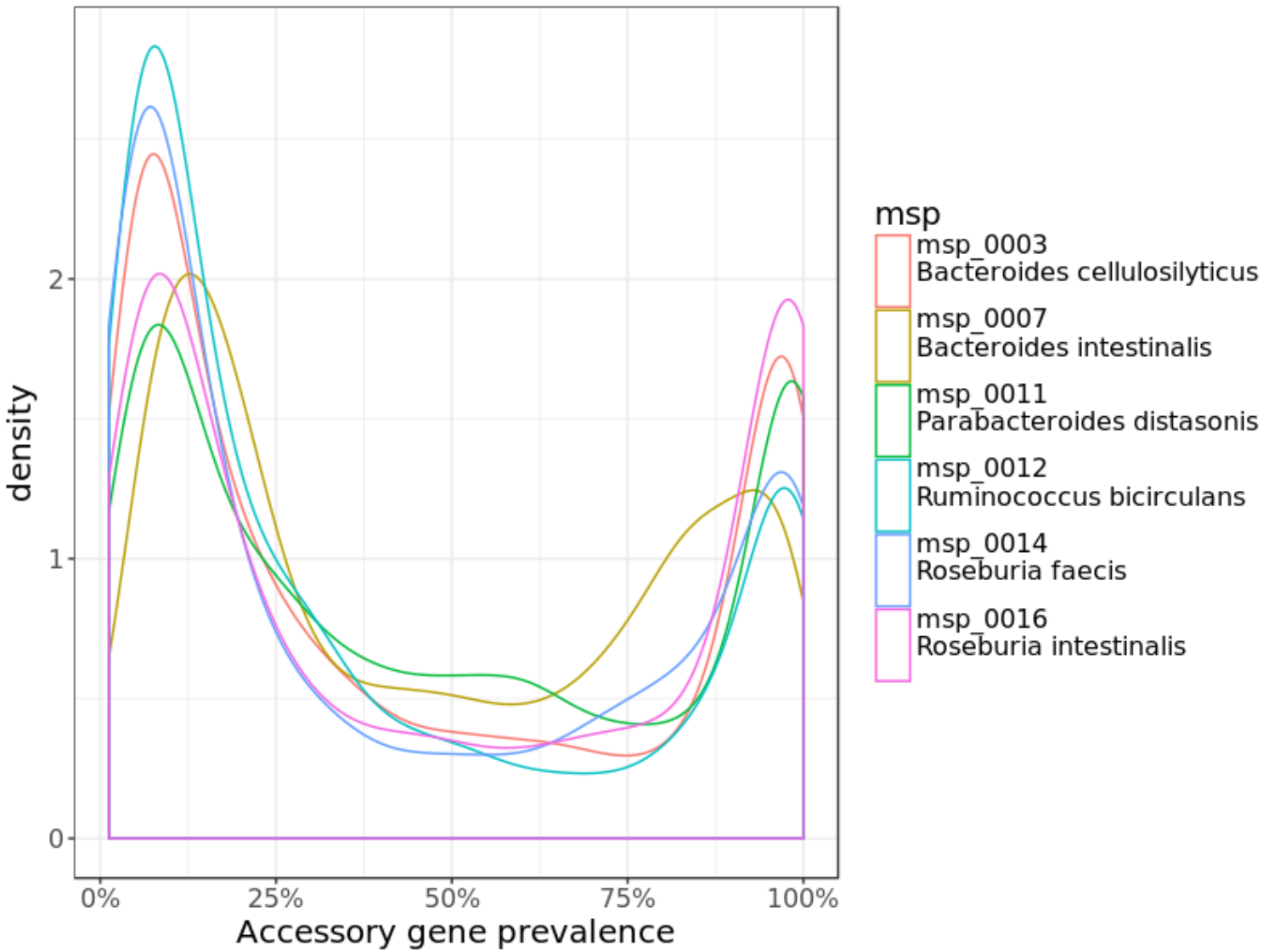
Supplementary Figure 4



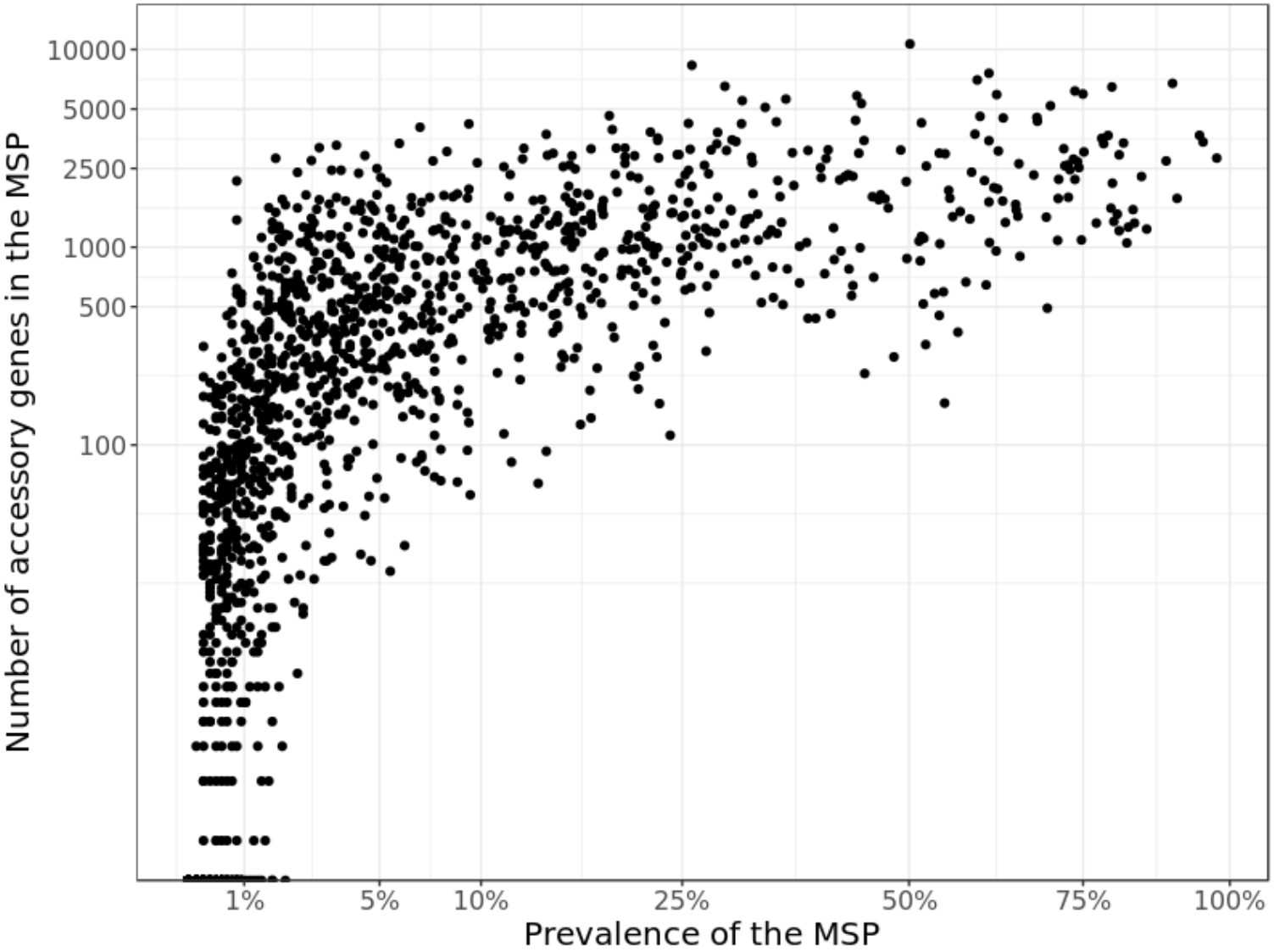
Supplementary Figure 5



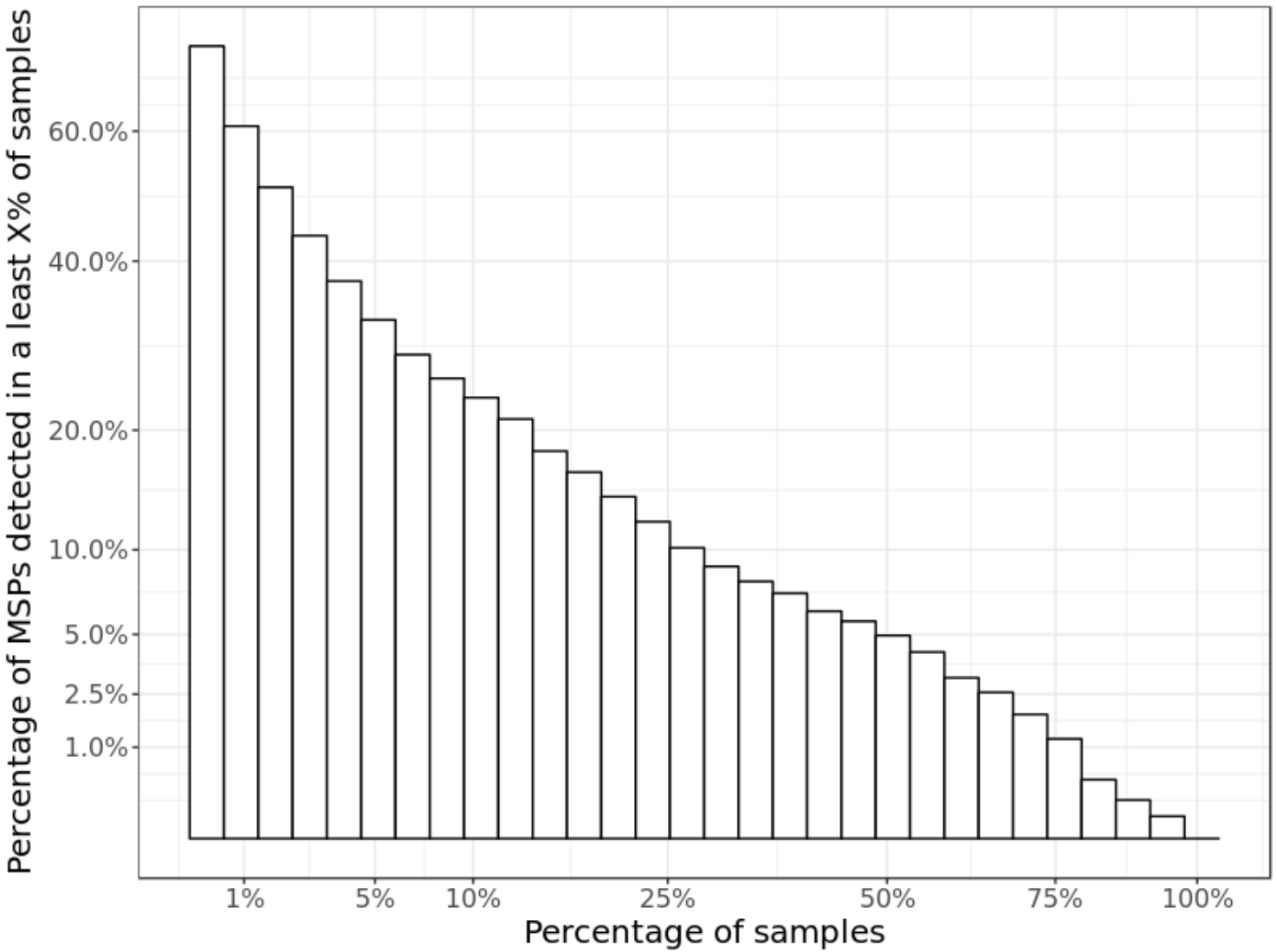
Supplementary Figure 6



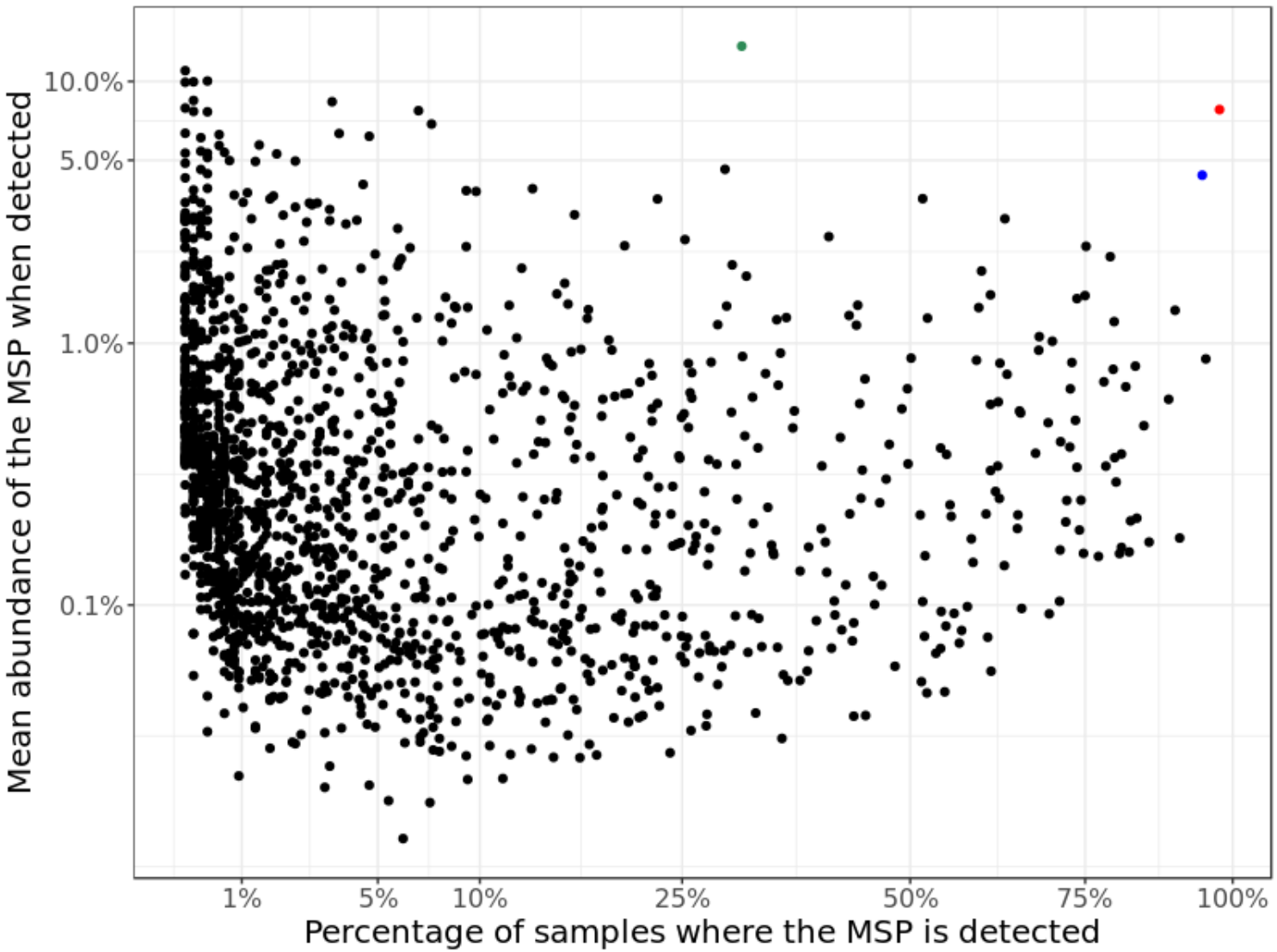
Supplementary Figure 7



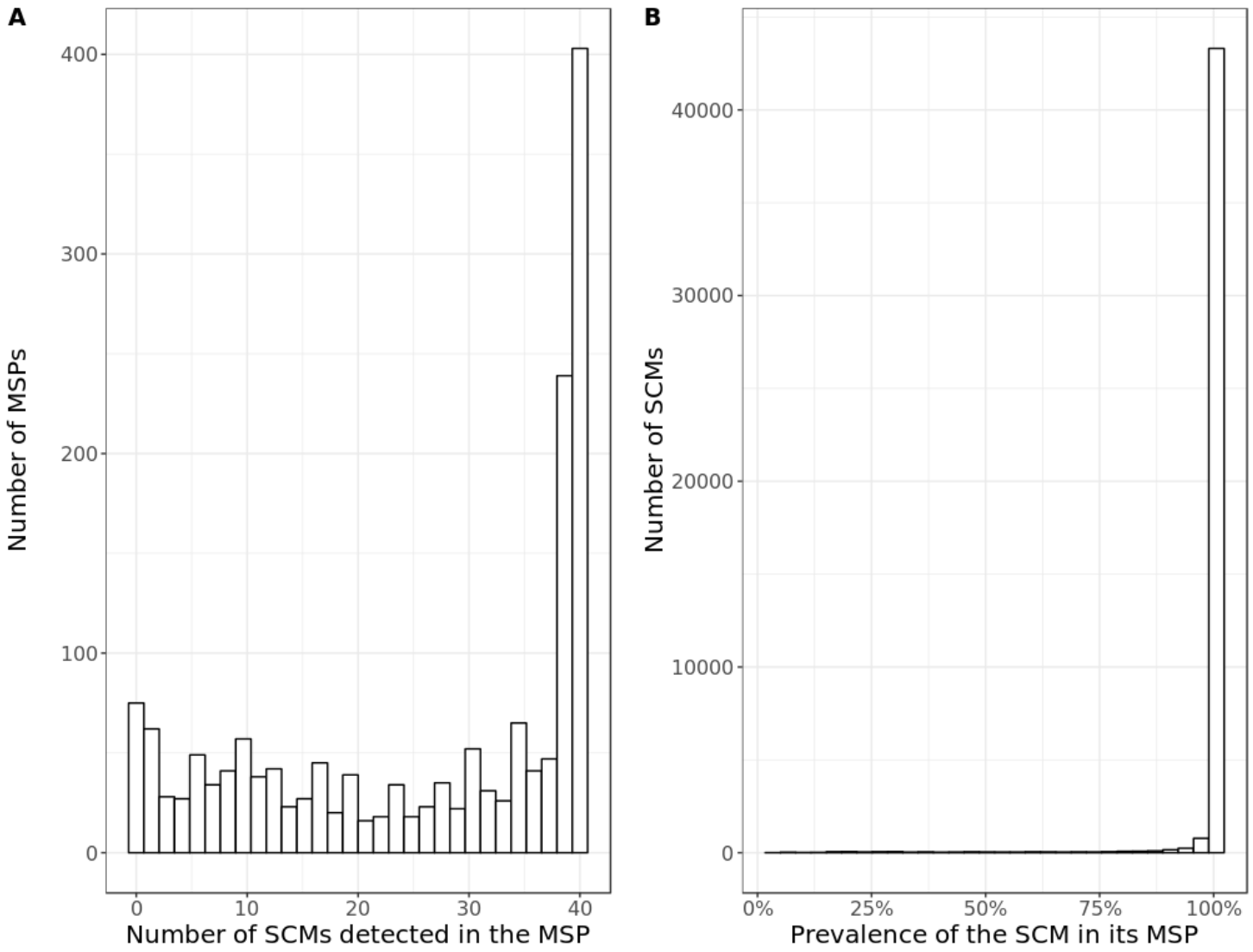
Supplementary Figure 8



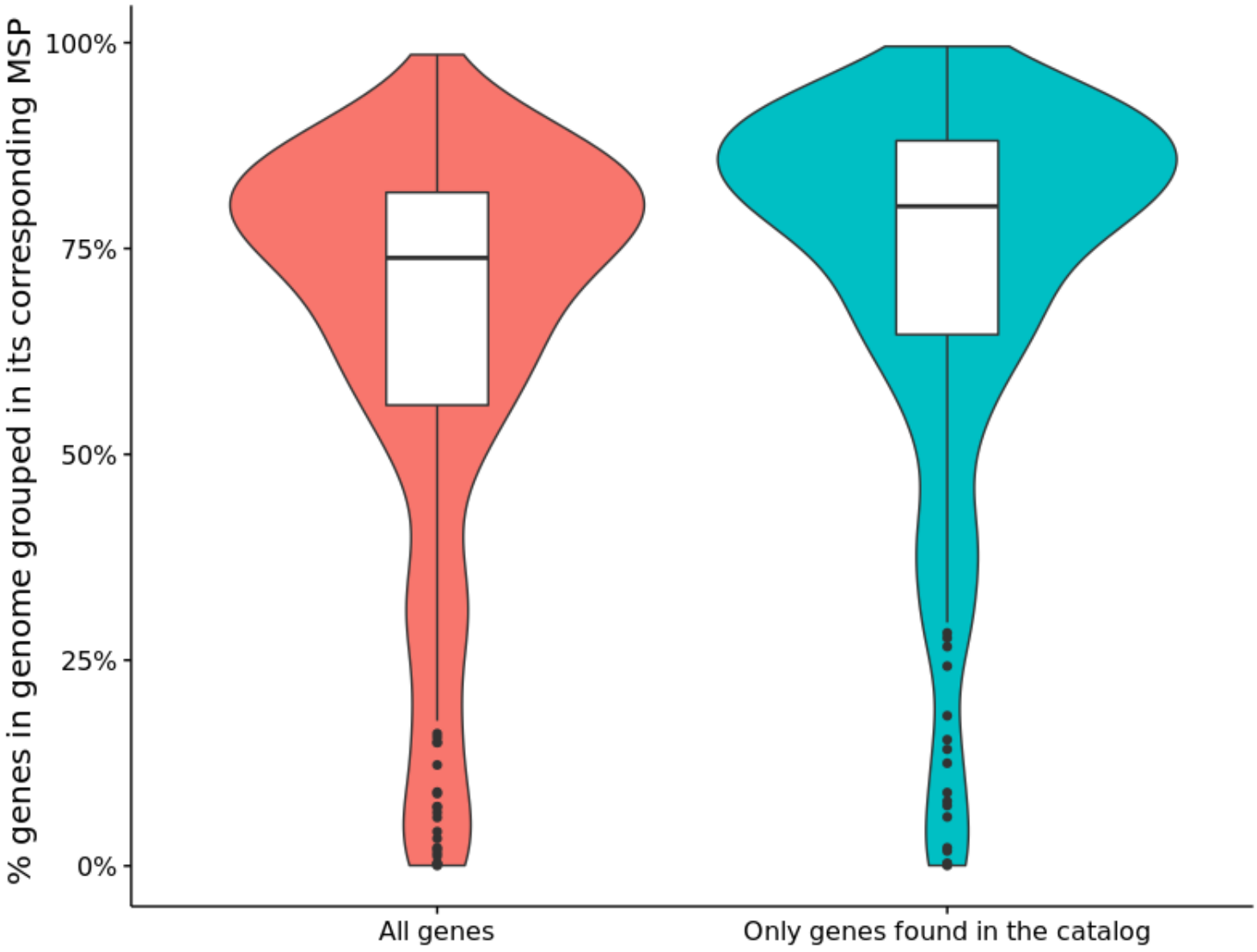
Supplementary Figure 9



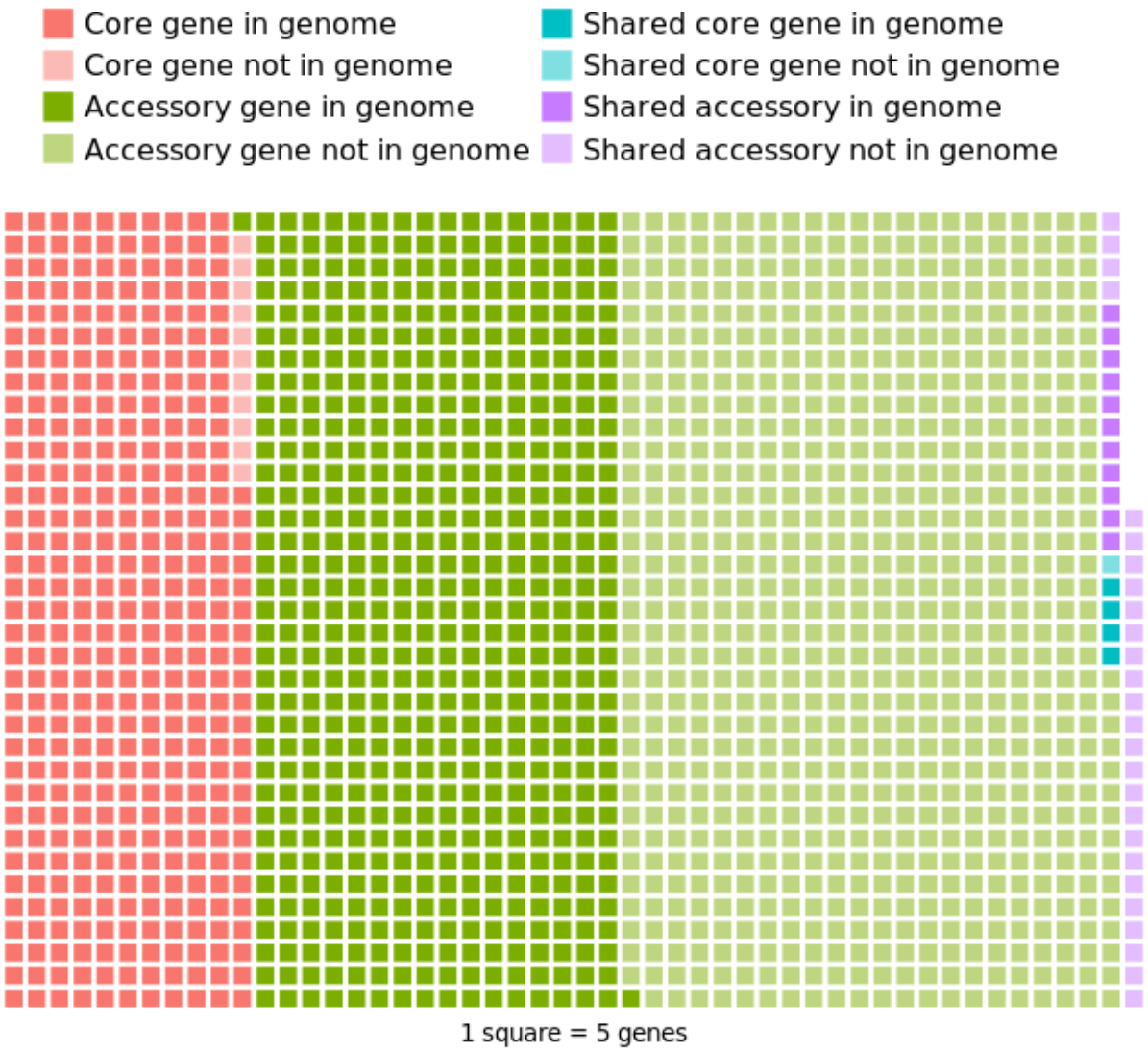
Supplementary Figure 10



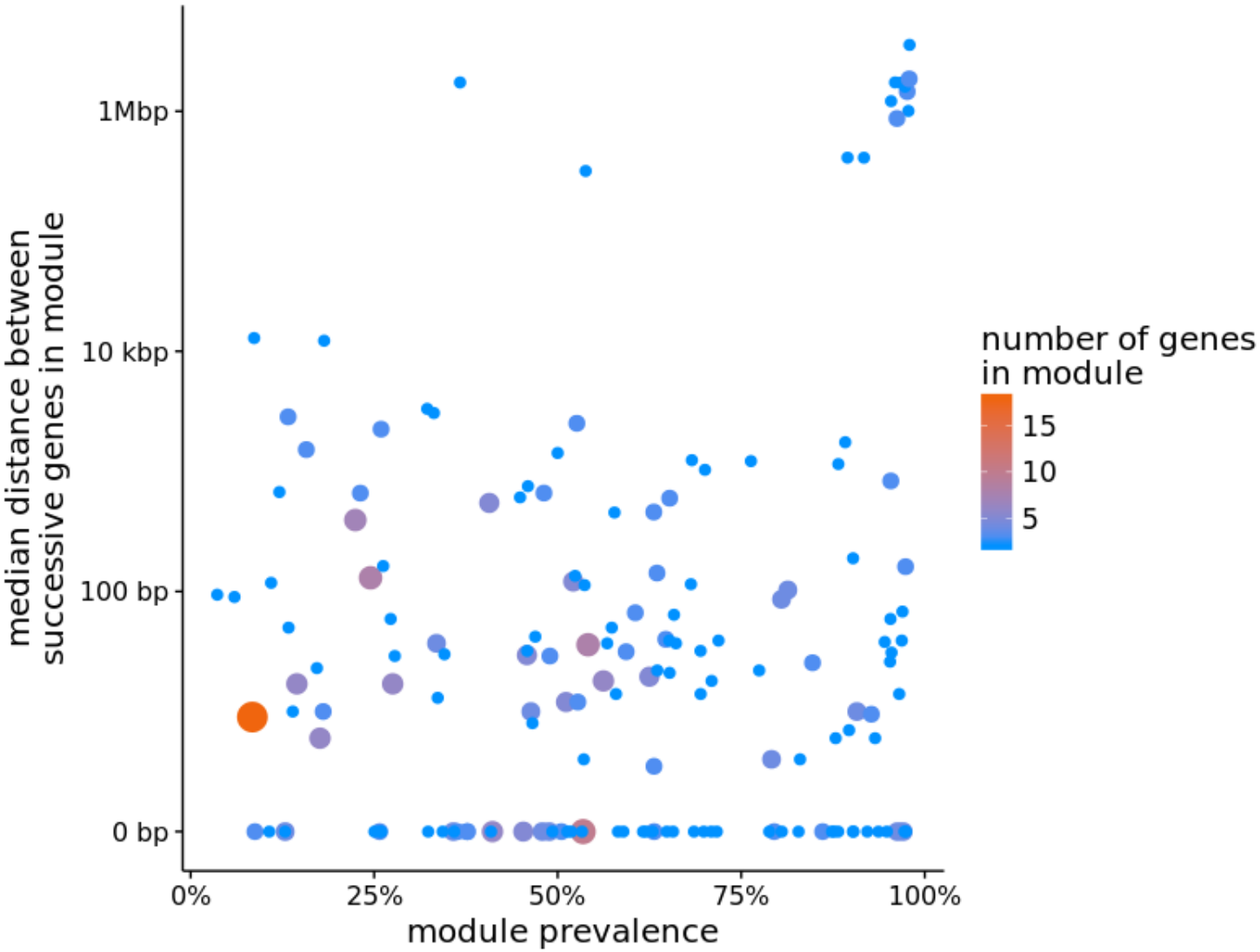
Supplementary Figure 11



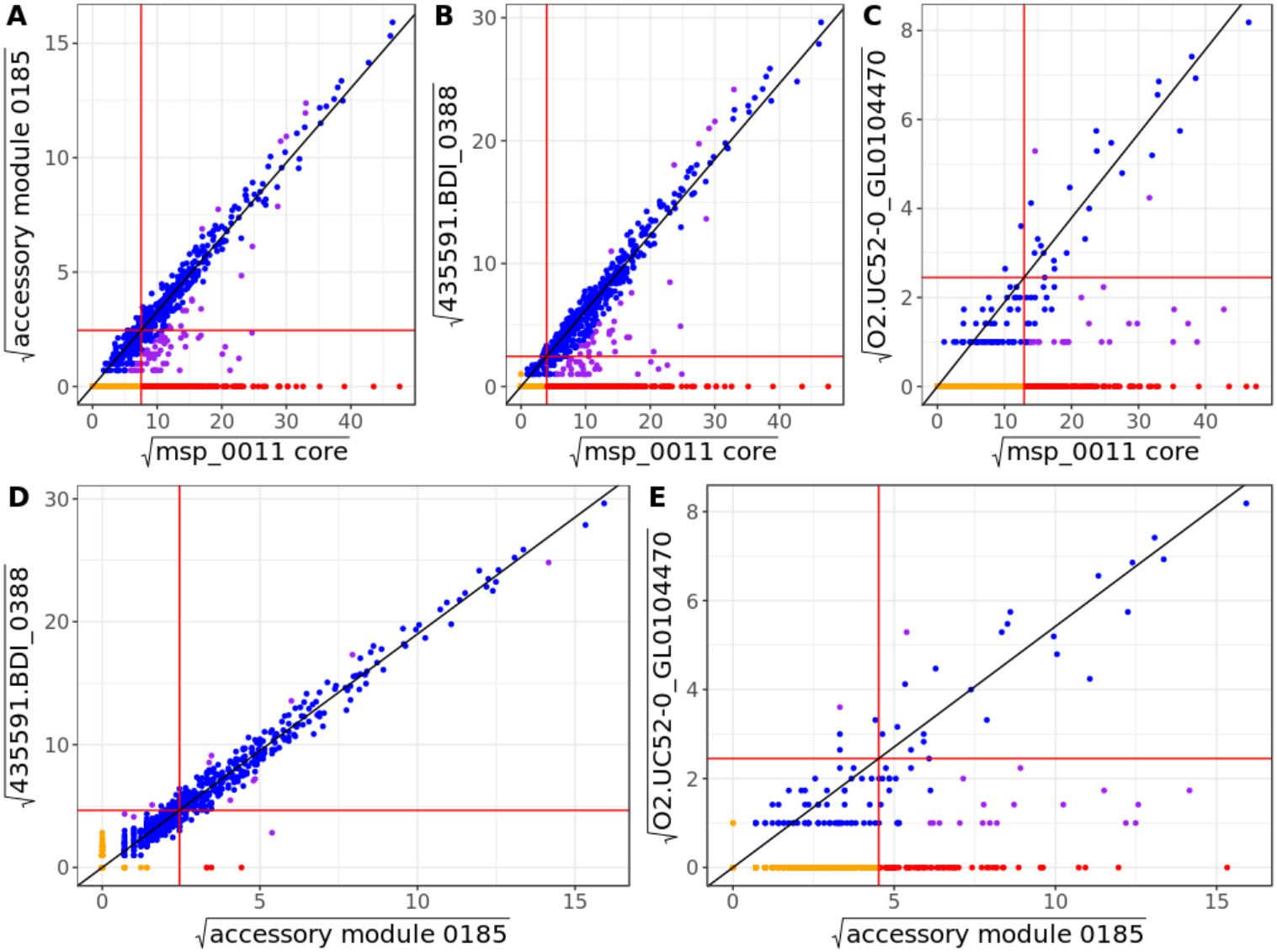
Supplementary Figure 12



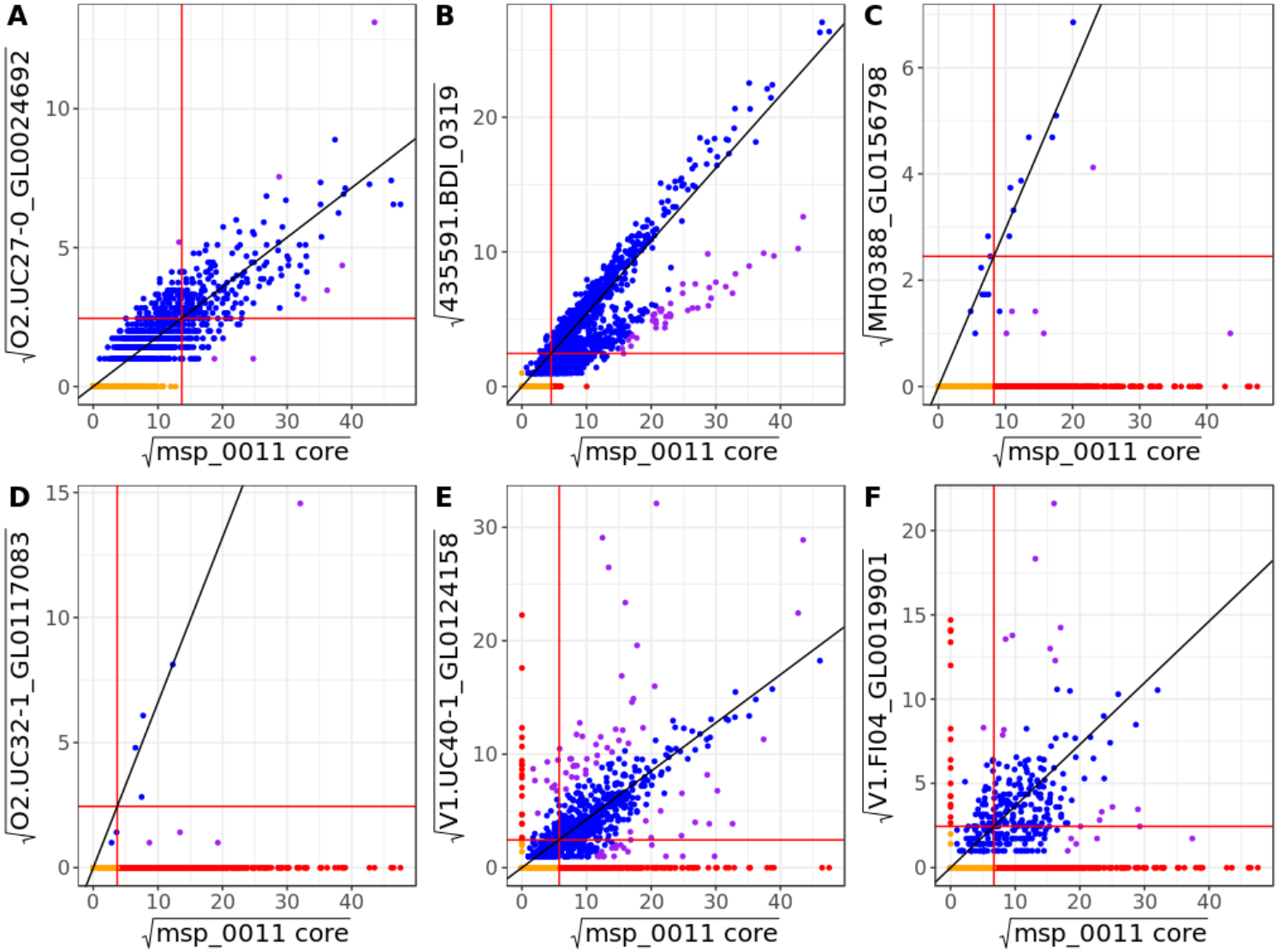
Supplementary Figure 13



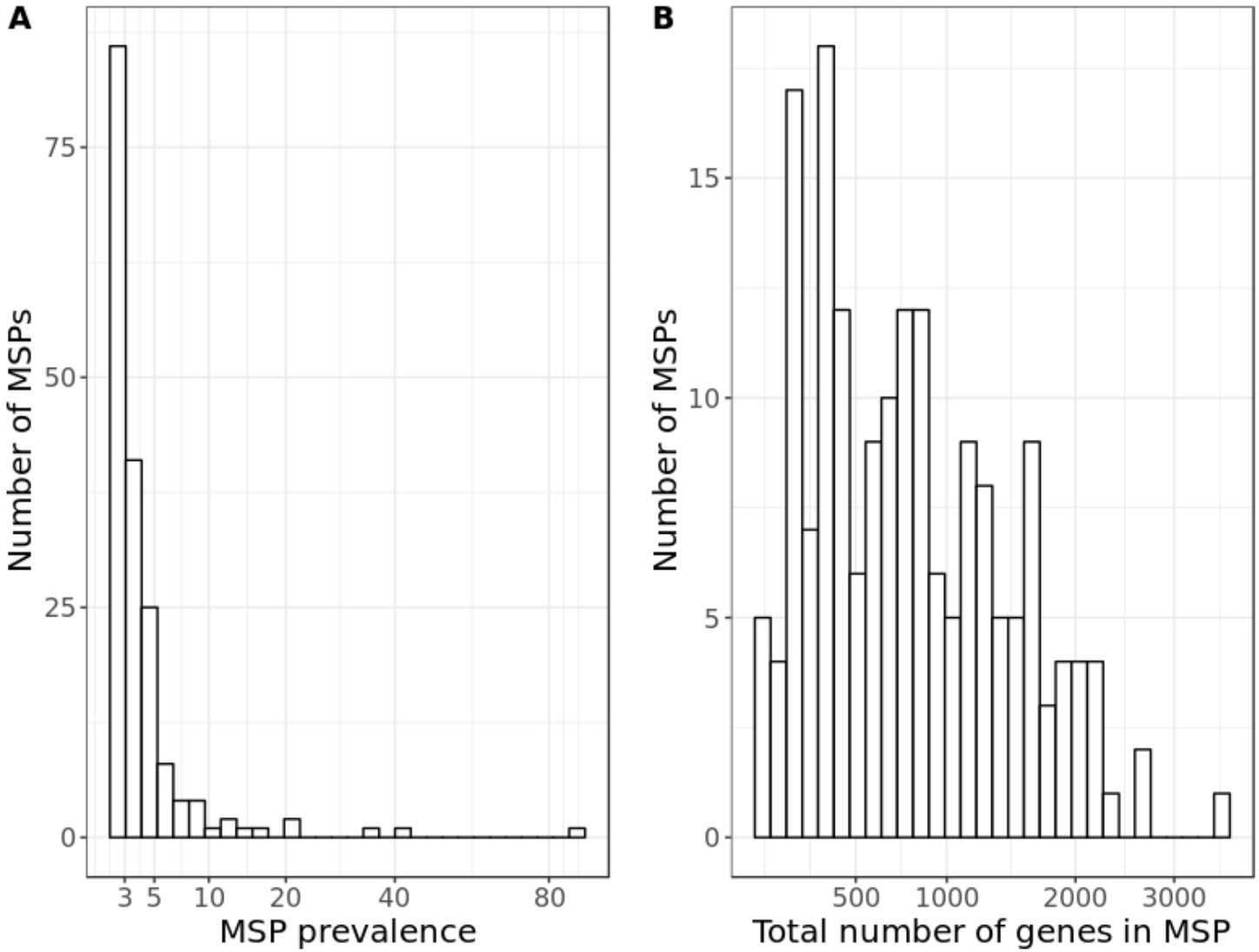
Supplementary Figure 14



Supplementary Figure 15

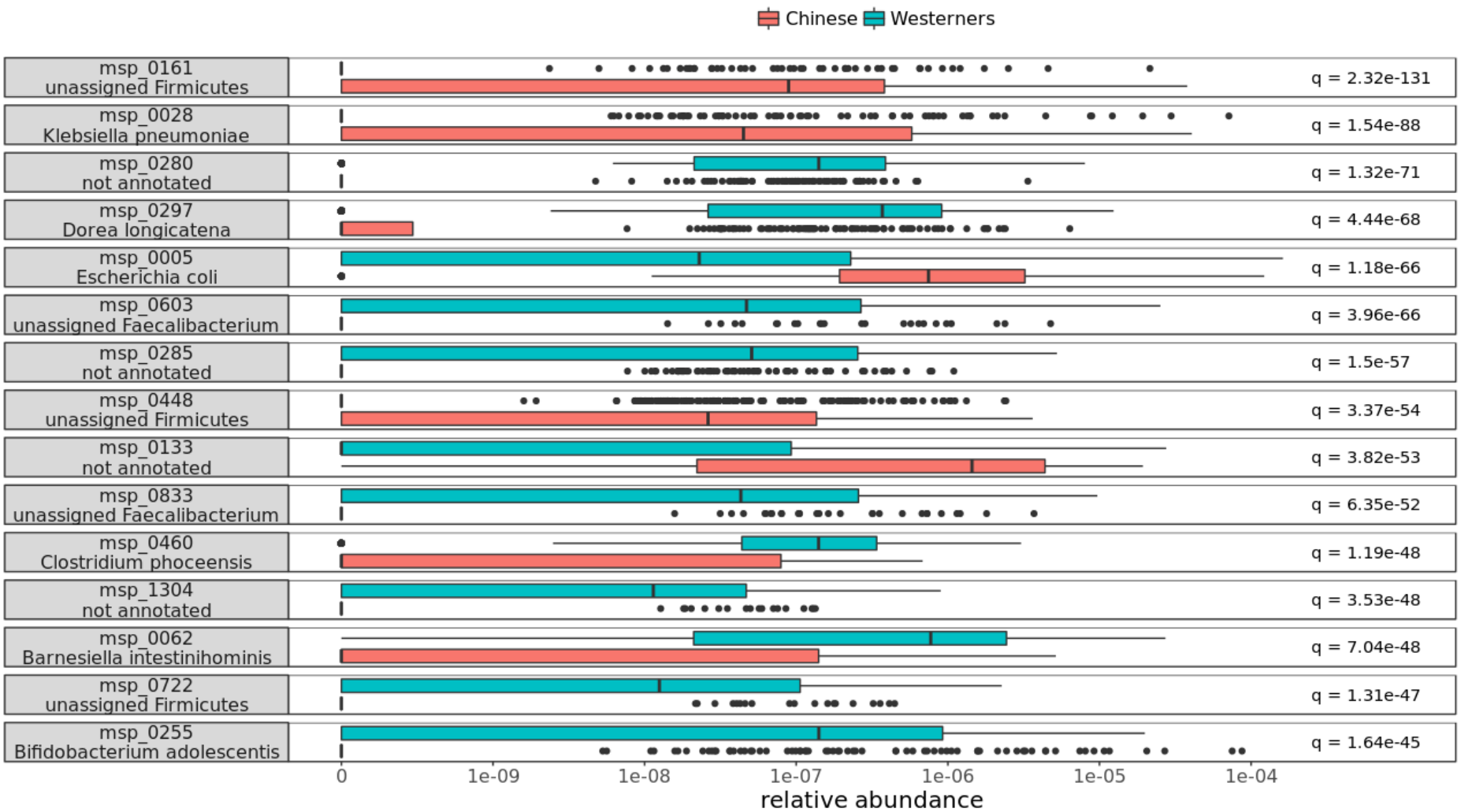


Supplementary Figure 16

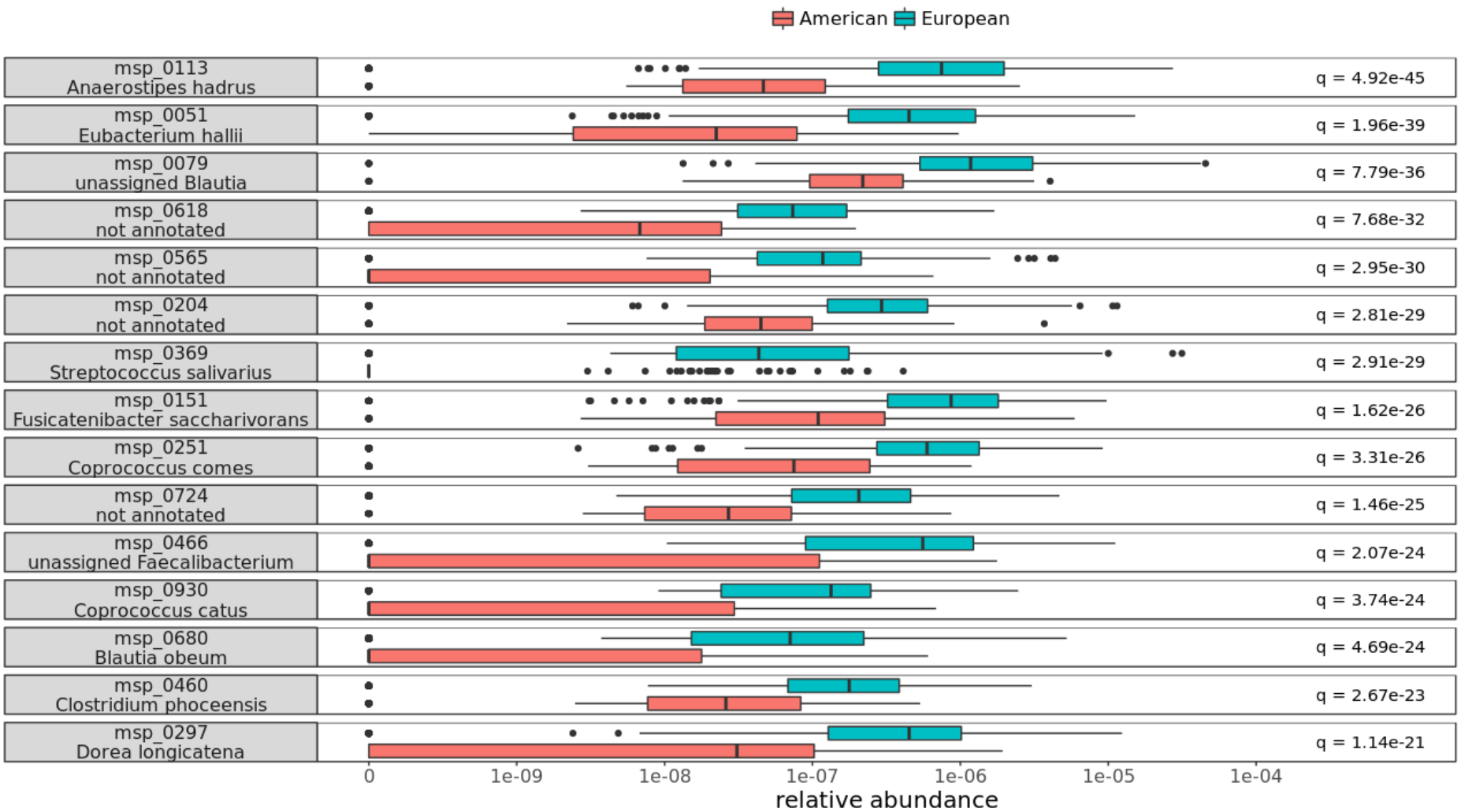


1005

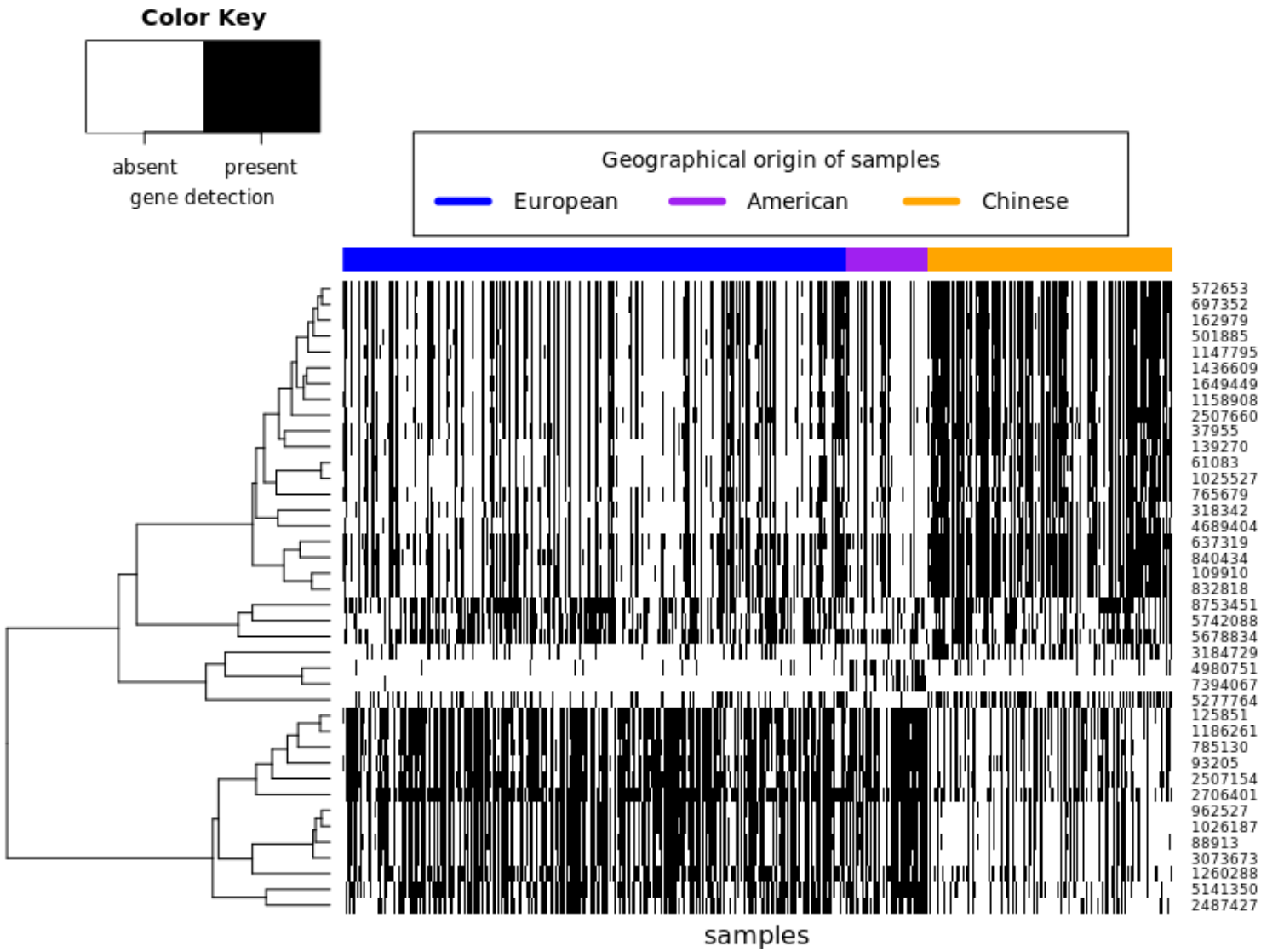
1006 Supplementary Figure 17



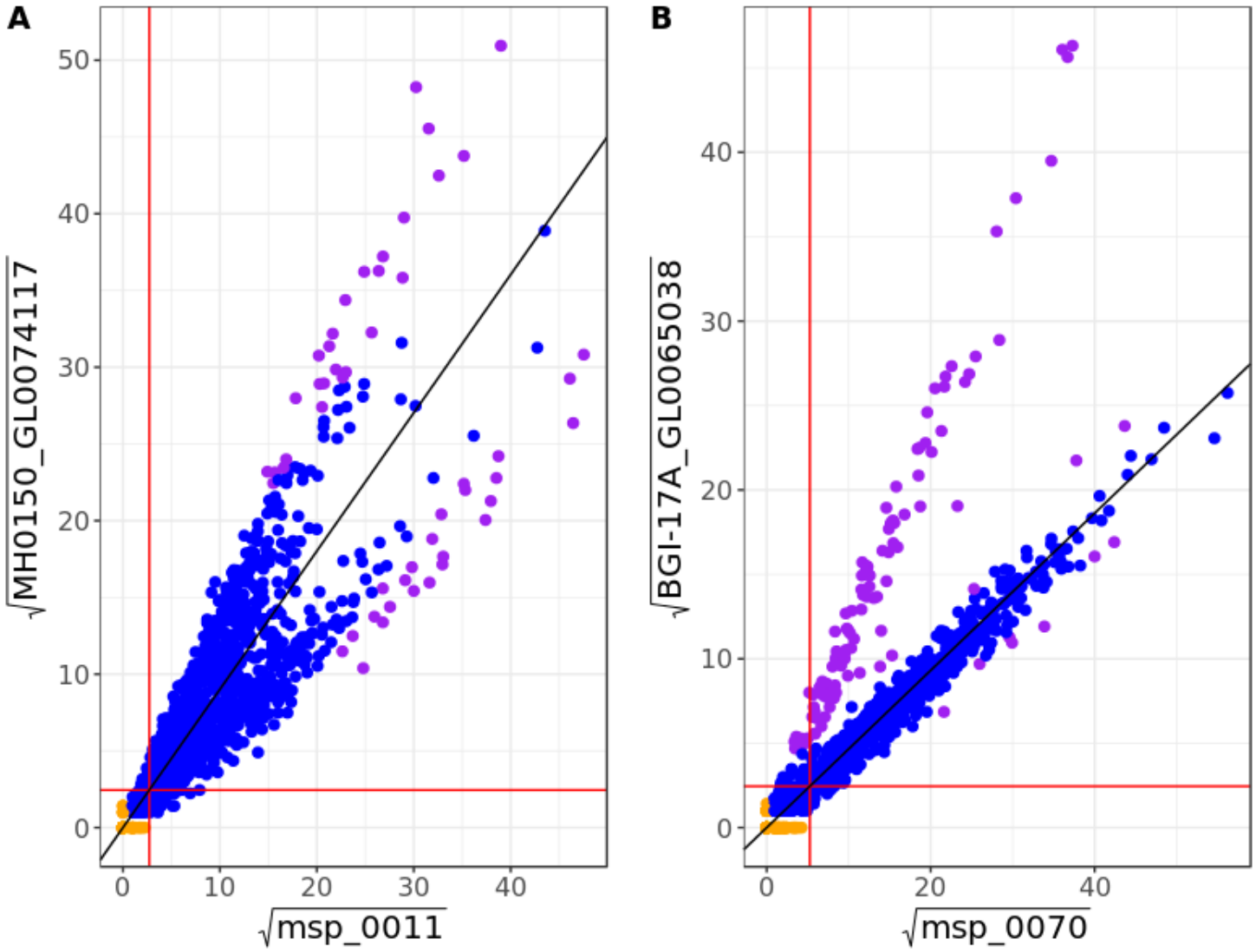
Supplementary Figure 18



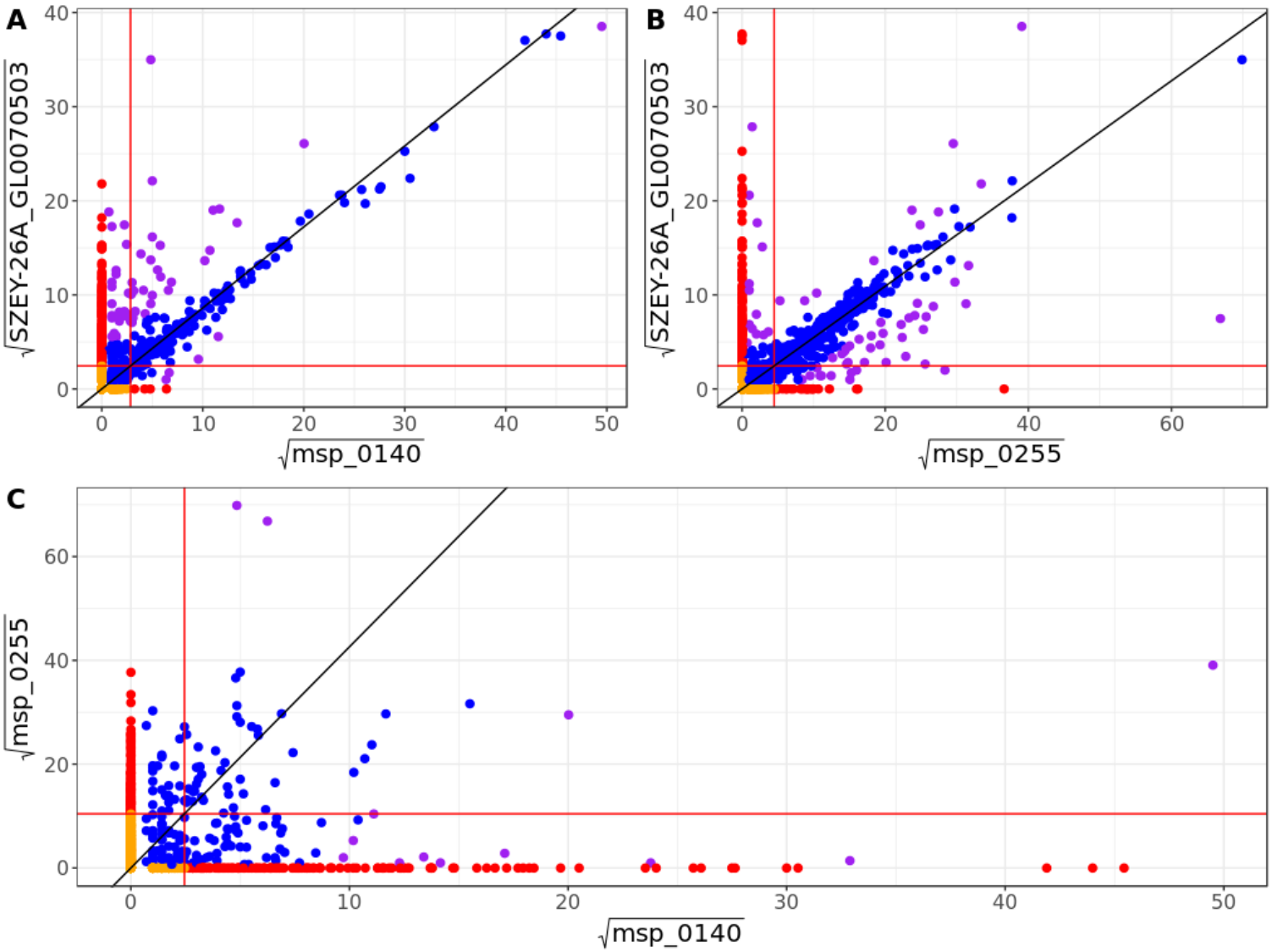
Supplementary Figure 19



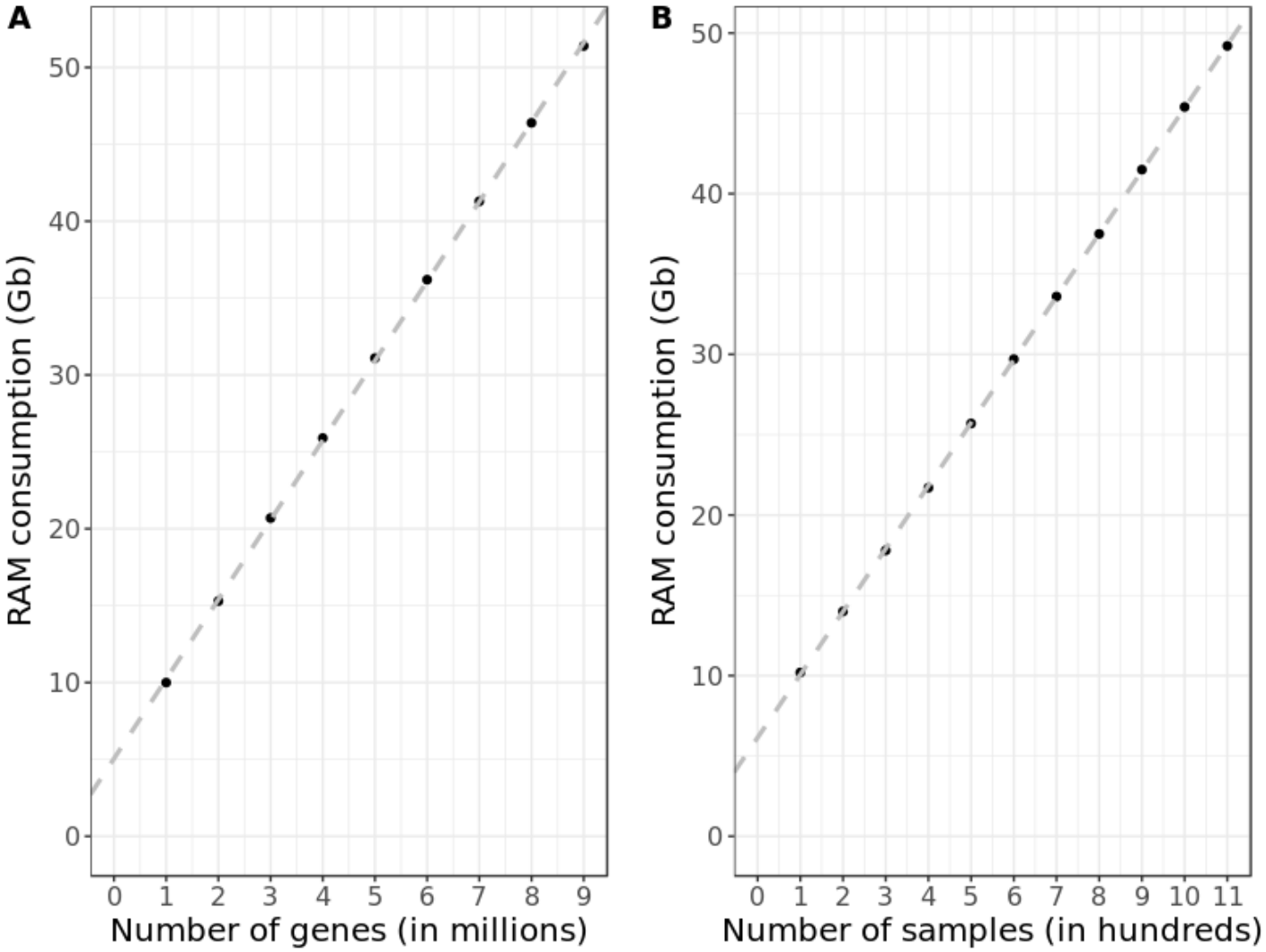
Supplementary Figure 20



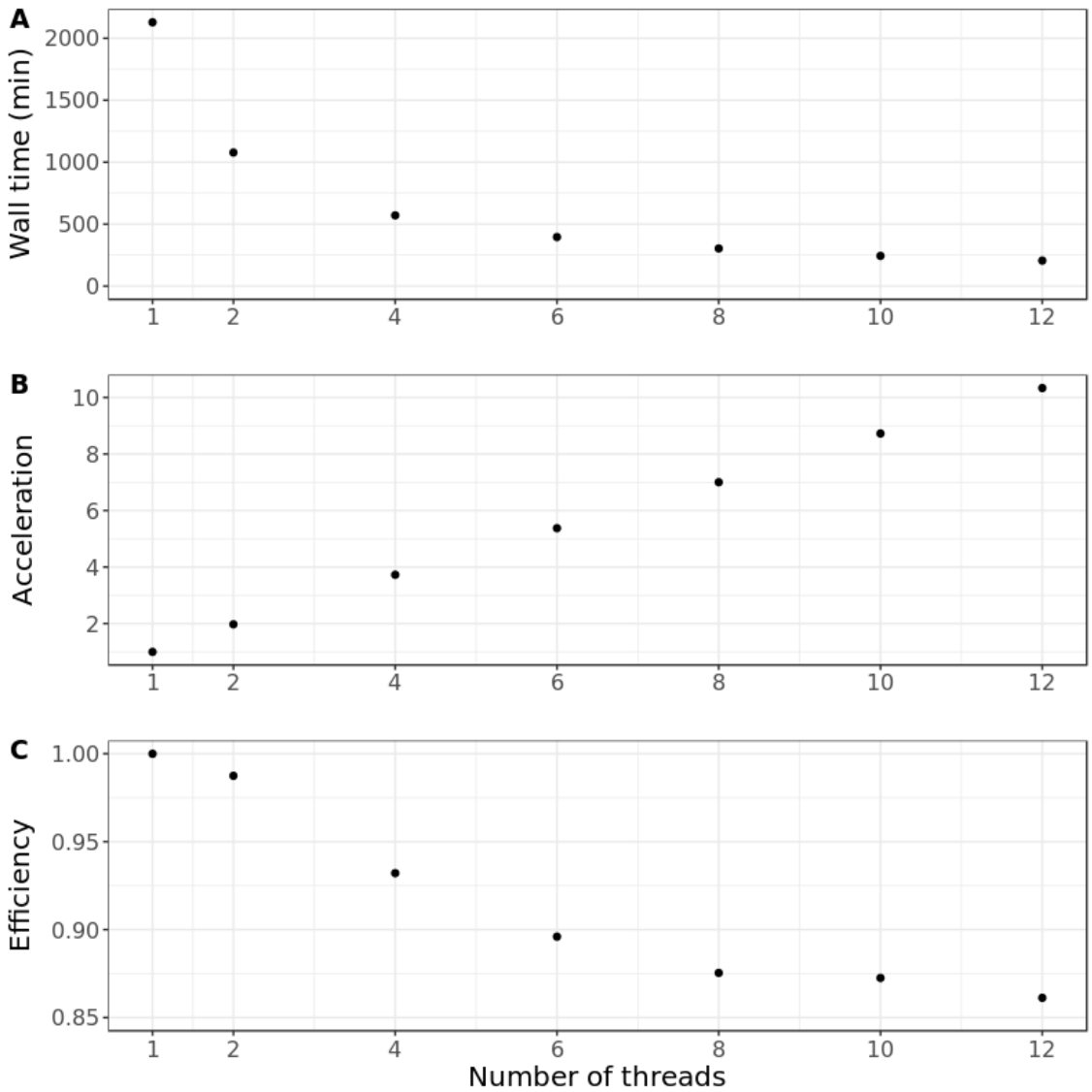
Supplementary Figure 21



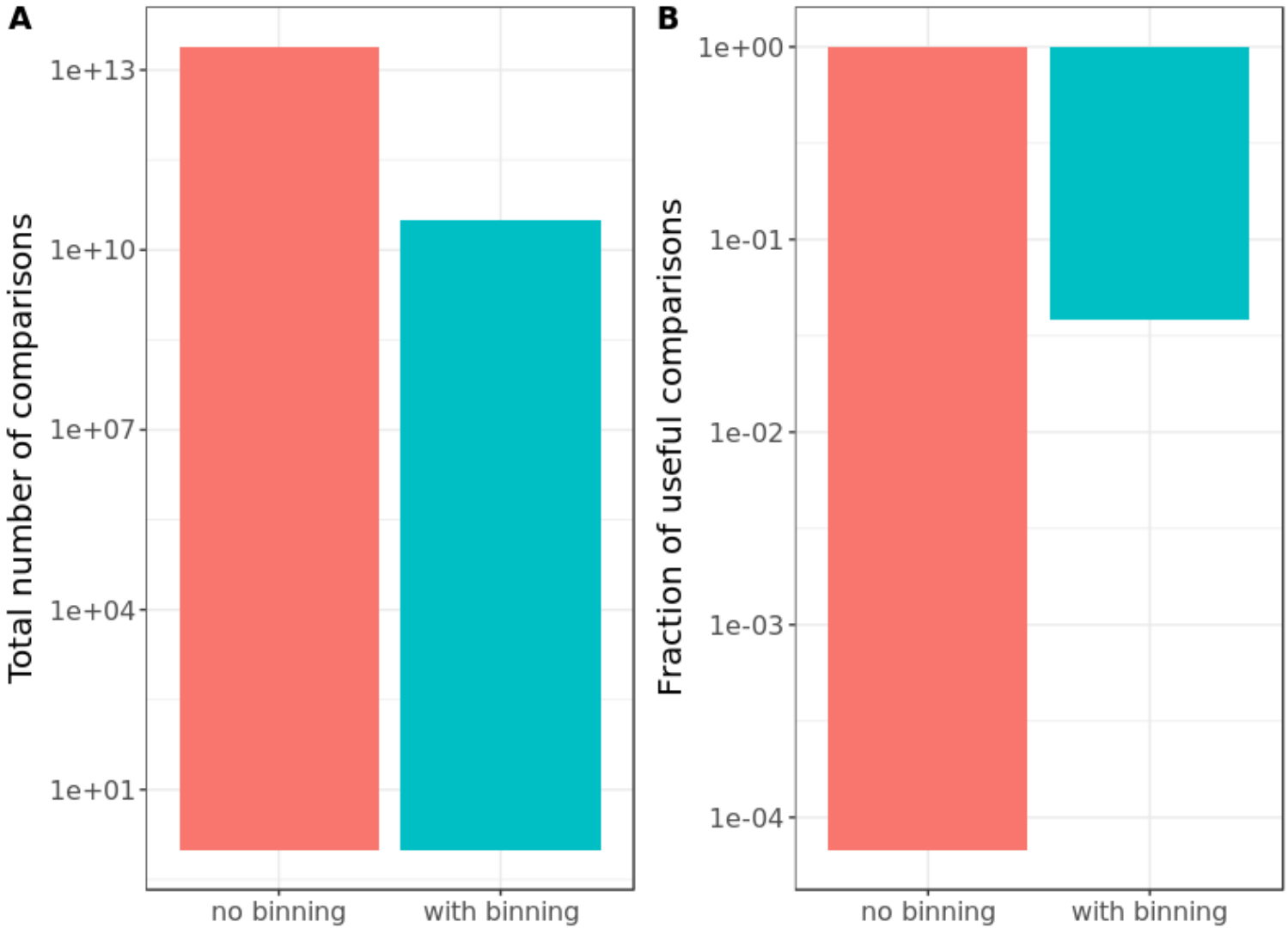
Supplementary Figure 22



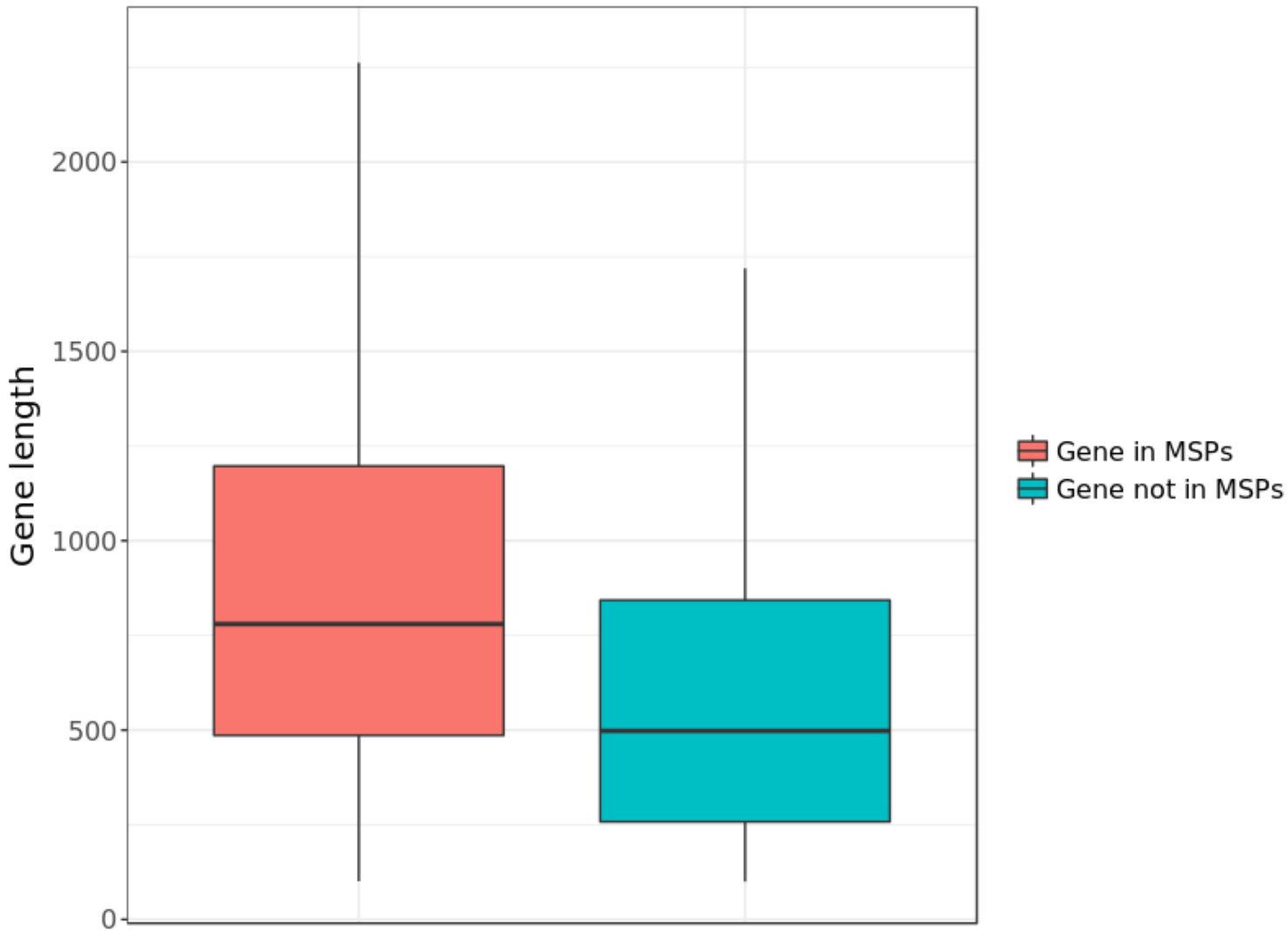
Supplementary Figure 23



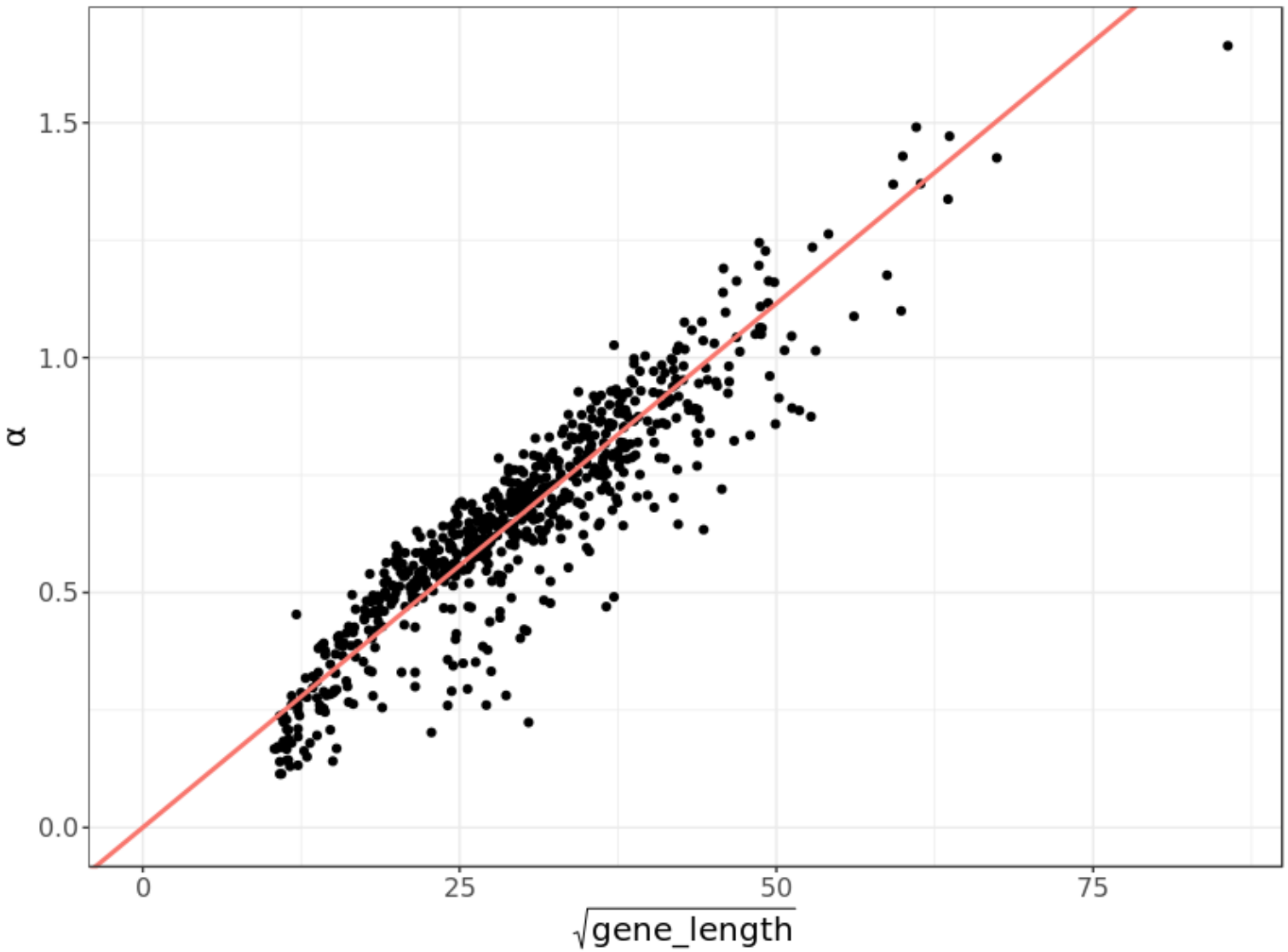
Supplementary Figure 24



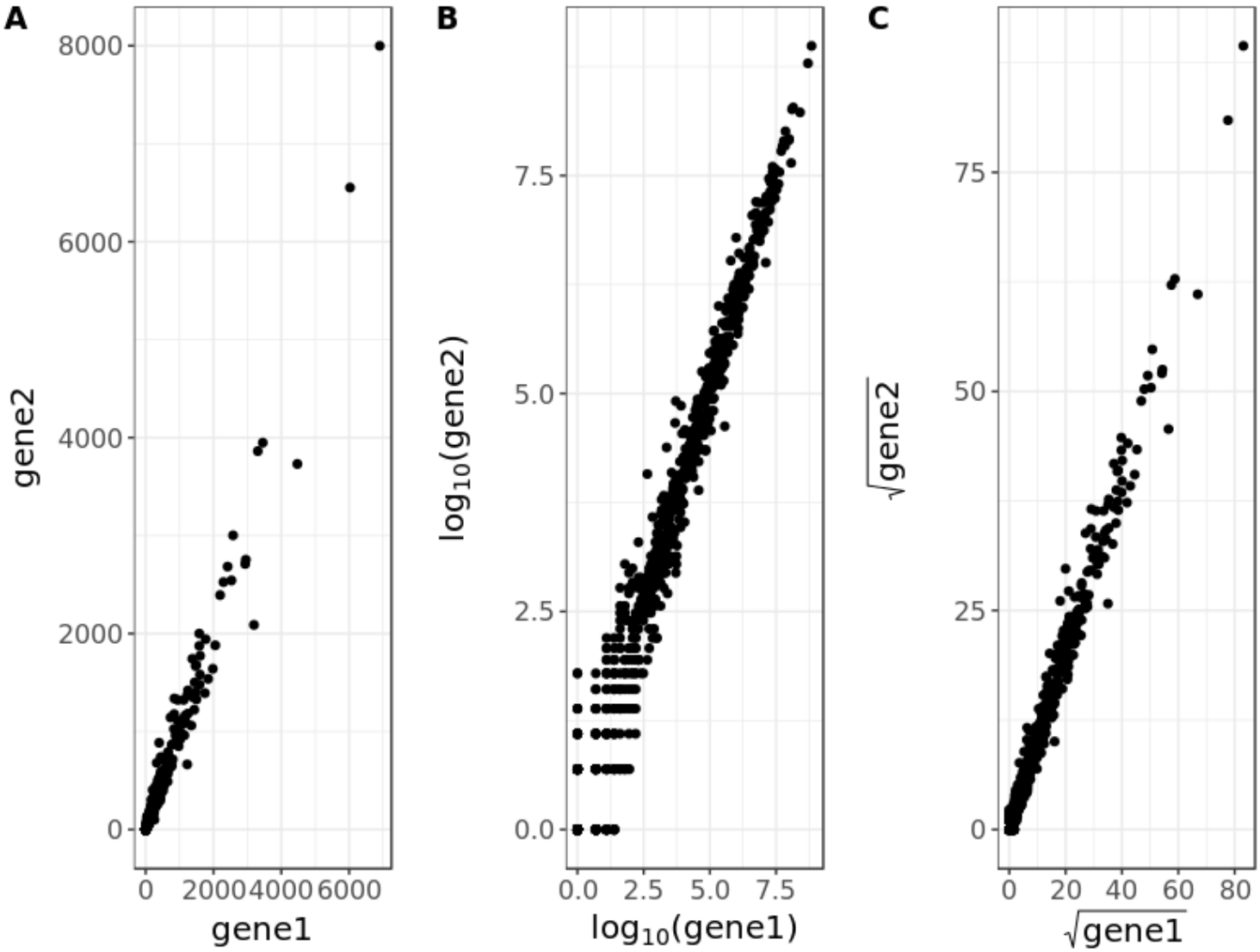
Supplementary Figure 25



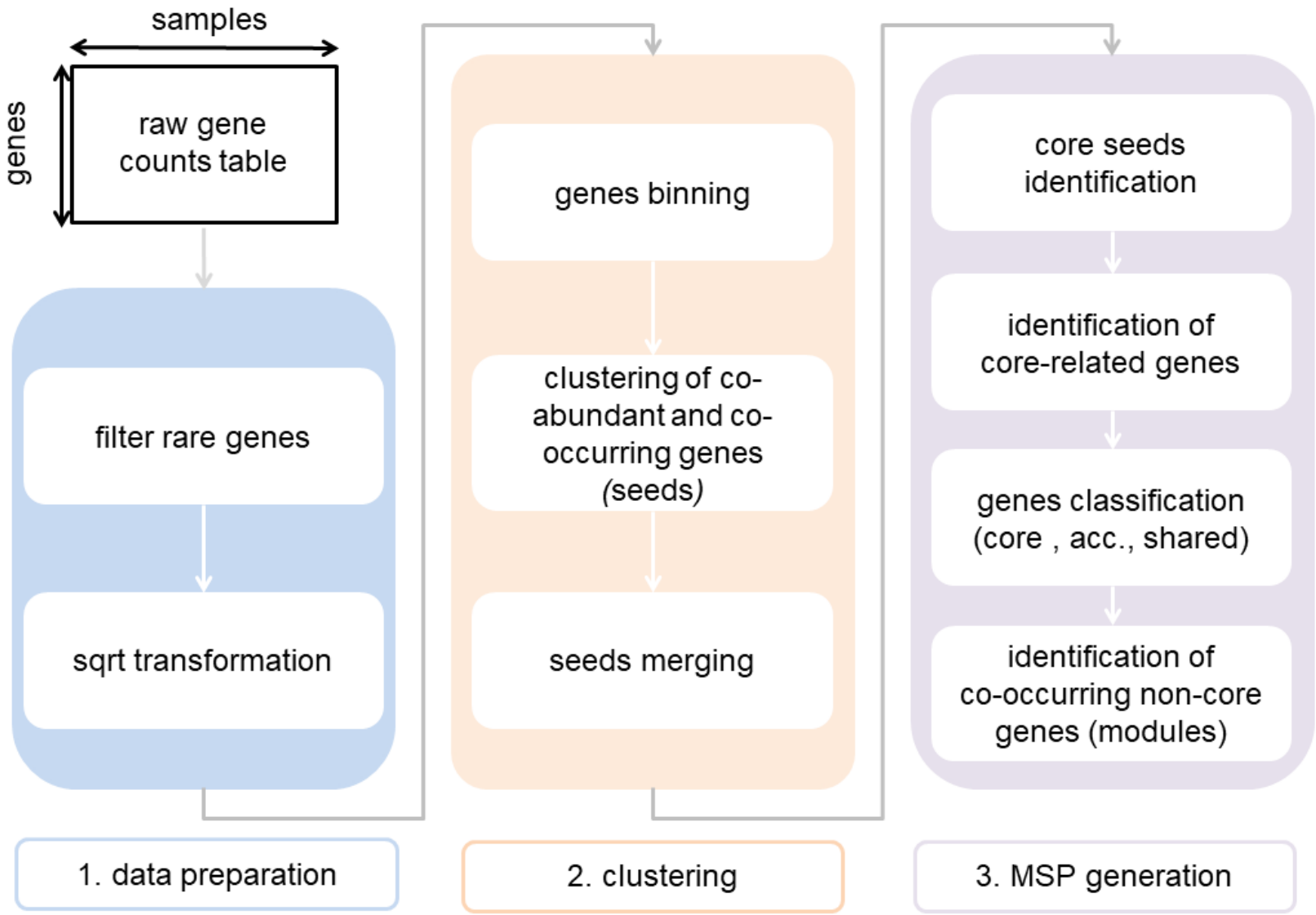
Supplementary Figure 26



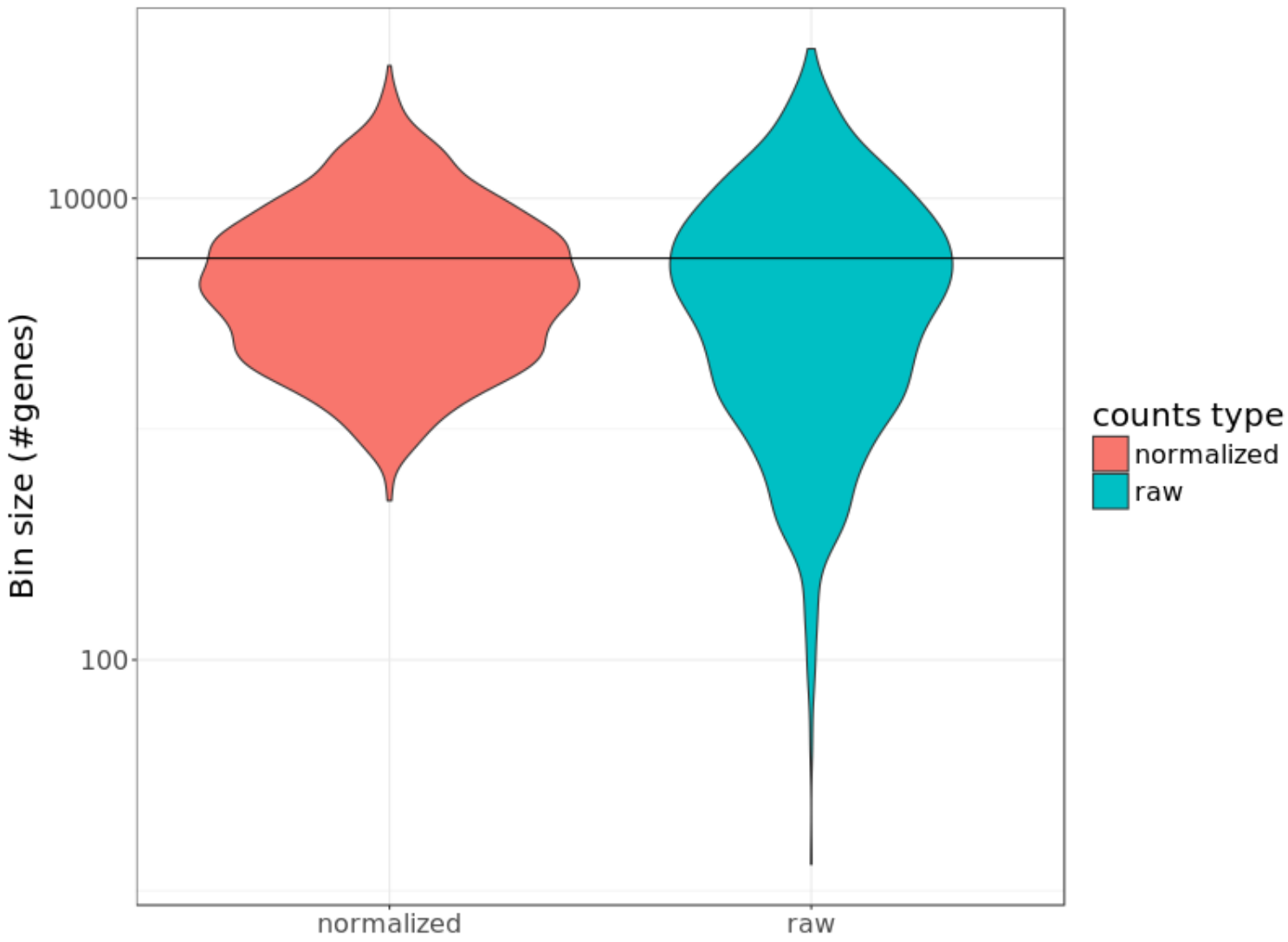
Supplementary Figure 27



Supplementary Figure 28



Supplementary Figure 29



Supplementary Figure 30

1033 # Supplementary Figures legends

1034 **Supplementary Figure 1: Comparison of the noisy genes abundance profiles of the**
1035 **virtual species to the original median signal of its core genome**

1036 using either the non-robust or robust measure of proportionality with an increasing
1037 percentage of outliers (5%, 10% and 20%).

1038 **Supplementary Figure 2: Lowest taxonomic rank assigned to MSPs**

1039 (1 square = 1 MSP)

1040 **Supplementary Figure 3: Most represented phyla among MSPs**

1041 **Supplementary Figure 4: Taxonomic homogeneity of MSPs**

1042 Only MSPs annotated at species level were kept.

1043 **Supplementary Figure 5: Histogram of the total number of genes in MSPs**

1044 **Supplementary Figure 6: Comparison of the number of genes in a MSP and its**
1045 **number of accessory genes.**

1046 The four outliers in red correspond to eukaryotic intestinal parasites of the *Blastocystis* genus
1047 (msp_0001, msp_0002, msp_0004 and msp_0006).

1048 **Supplementary Figure 7: Prevalence of the accessory genes of some MSPs among**
1049 **samples where their respective core genes are detected.**

1050 **Supplementary Figure 8: Comparison of the number of samples where a MSPs is**
1051 **detected and its number of accessory genes.**

1052 **Supplementary Figure 9: Prevalence of MSPs among the 1267 samples.**

1053 **Supplementary Figure 10: Comparison of the percentage of samples where a MSPs**
1054 **is detected and its mean abundance.**

1055 The red, blue and green points correspond to the MSPs associated to *Bacteroides vulgatus*,
1056 *Bacteroides uniformis* and *Prevotella copri* respectively.

1057 **Supplementary Figure 11: Single Copy Marker genes in MSPs**

1058 A. Number of universal single copy marker genes (SCM) in MSPs

1059 B. Prevalence of the universal single copy marker genes (SCM) in the MSPs they belong

1060 to.

1061 **Supplementary Figure 12: Percentage of genes from reference genomes grouped in**

1062 **their best corresponding MSPs.**

1063 **Supplementary Figure 13: Comparison of the genes detected in *P. distasonis* ATCC**

1064 **8503 with those of its corresponding MSP.**

1065 Each square represents 5 genes. Dark colors correspond to genes detected both in the genome

1066 and the MSP while light colors correspond to genes not found in the genome.

1067 **Supplementary Figure 14: Median distance between consecutive genes in modules**

1068 **of the msp_0011 detected in P. distasonis ATCC 8503, as a function of module**

1069 **prevalence and module size.**

1070 The median distance between successive genes in a module was below 150 bp in 75% of the

1071 cases.

1072 Median distances of 0 base pairs correspond to modules with overlapping genes.

1073 **Supplementary Figure 15: Example of genes found within a genomic region of *P.***

1074 ***distasonis* ATCC 8503 delimited by a module**

1075 The 4 genes of the accessory module 0185 are detected in the genome of *P. distasonis* ATCC

1076 8503 between positions 441 398 - 442 762.

1077 The genes 435591.BDI_0388 and O2.UC52-0_GL0104470 are detected in this region between

1078 positions 441903 -442697 and 442472 - 442697 respectively but were not grouped in the

1079 module.

1080

1081 3.A, 3.B and 3.C: The association of the core of the msp_0011 with the module and the two

1082 singletons is clear.

1083 3.D: The gene 435591.BDI_038 was not grouped in the module because it was missing in 3

1084 samples where the module was detected. Here, one may consider that the clustering criteria

1085 was too stringent.

1086 3.E: The gene O2.UC52-0_GL0104470 was not grouped in the module because it was missing
1087 in 77 samples where the module was detected. Here, excluding the module gene seems
1088 relevant.

1089 **Supplementary Figure 16: Example of genes of *P. distasonis* ATCC 8503 not found in**
1090 **the msp_0011**

1091 4.A: O2.UC27-0_GL0024692 (alien region 4) corresponds to a false negative with
1092 overdispersed counts. The robust measure of proportionality was above the threshold set by
1093 MSPMiner.
1094 4.B: 435591.BDI_0319 (alien region 3) corresponds to a false negative subject to copy-number
1095 variation. The robust measure of proportionality was above the threshold set by MSPMiner.
1096 4.C: MH0388_GL0156798 (alien region 10) probably corresponds to a false negative. The
1097 number of observations was too low to detect an association.
1098 4.D: O2.UC32-1_GL0117083 (alien region 52) might correspond to a false negative. The
1099 number of observations was too low to detect an association.
1100 4.E V1.UC40-1_GL0124158 (alien region 86) corresponds to a true negative as there is no clear
1101 association. Remarkably, this gene was annotated as a mobile element (conjugative
1102 transposon protein).
1103 4.F V1.FI04_GL0019901 (alien region 21) corresponds to a true negative as there is no clear
1104 association. Interestingly, this gene was classified as shared accessory in the msp_0025
1105 corresponding to *Parabacteroides merdae*.

1106 **Supplementary Figure 17: Description of the MSPs with no equivalent among CAGs.**

1107 A. Histogram of the prevalence of the MSPs with no equivalent among CAGs.
1108 B. Histogram of the number of genes in the MSPs with no equivalent among CAGs.

1109 **Supplementary Figure 18: Relative abundance of the top 10 discriminant MSPs**
1110 **between Chinese and Westerners.**

1111 **Supplementary Figure 19: Relative abundance of the top 10 discriminant MSPs**
1112 **between Americans and Europeans.**

1113 **Supplementary Figure 20: The 40 accessory genes (in column) of the msp_0011**
1114 **(*Parabacteroides distasonis*) most associated with the geographical origin of**
1115 **samples (in row).**

1116 **Supplementary Figure 21: Example of genes subject to copy-number variation.**

1117 A. Comparison of the core genome median abundance of the msp_0011
1118 (*Parabacteroides distasonis*) to one of its core genes (MH0150_GL0074117) subject to
1119 Copy-Number Variation. Here, MSPminer wrongly estimates the coefficient of
1120 proportionality as there is no dominant gene copy number among the samples.
1121 B. Comparison of the core genome median abundance of the msp_0070 (*Eubacterium*
1122 *rectale*) to one of its core genes (BGI-17A_GL0065038) subject to Copy-Number
1123 Variation. Here, MSPminer correctly estimates the coefficient of proportionality for
1124 the majority of samples (blue points) that have the same number of copies of the gene.
1125 The others are classified as outliers (purple points).

1126 **Supplementary Figure 22: Example of genes shared between several MSPs.**

1127 The gene SZEY-26A_GL0070503 is classified as shared accessory in the msp_0140
1128 (*Bifidobacterium pseudocatenulatum*) and the msp_0255 (*Bifidobacterium adolescentis*)

1129 A. Comparison of the core genome median abundance of the msp_0140 to the gene
1130 SZEY-26A_GL0070503. Despite the presence of outliers, a direct proportionality
1131 relationship is observed in many samples where both the msp and the gene are
1132 present.
1133 B. Comparison of the core genome median abundance of the msp_0255 to the gene
1134 SZEY-26A_GL0070503. Despite the presence of outliers, a direct proportionality
1135 relationship is observed in many samples where both the msp and the gene are
1136 present.

1137 C. Comparison of the core genome median abundance of the msp_0140 to the core
1138 genome median abundance of the msp_0255. Although the two MSPs share a gene,
1139 they have distinct signals.

1140 **Supplementary Figure 23: Memory usage of MSPminer.**

1141 A. RAM consumption as a function of the number of genes. The number of samples was
1142 set at 1267. The grey line represents the best linear fit (intercept=5.05, slope=5.17).
1143 B. RAM consumption as a function of the number of samples. The number of genes was
1144 set at 9 879 896. The grey line represents the best linear fit (intercept=6.14,
1145 slope=3.92).

1146 **Supplementary Figure 24: Parallel performance of MSPminer**

1147 The tests were performed on a server with an Intel® Xeon® processor E5-2630 (6 physical
1148 cores with hyperthreading) and 128Gb of RAM.

1149 **Supplementary Figure 25: Performance of MSPminer with or without genes**
1150 **binning.**

1151 A. Total number of comparisons performed by comparing all genes pairwise with or
1152 without binning. Here, genes binning divides the number of comparisons performed
1153 by 763.
1154 B. Fraction of useful comparisons for MSP cores creation with or without binning. The
1155 fraction without binning was estimated by drawing randomly one billion pairs of genes.
1156 Here, genes binning multiplies the fraction of useful comparisons by 562.

1157 **Supplementary Figure 26: Comparison of the length of the genes grouped in the**
1158 **MSPs with those that were not.**

1159 **Supplementary Figure 27: Comparison of the genes length and their respective**
1160 **coefficient of proportionality.**

1161 Core genes abundance profiles of the msp_0043 (*Ruminococcus bromii*) were compared to its
1162 core genome median abundance. For each gene, the estimated coefficient of proportionality
1163 α was compared to its length. Then, a robust linear regression with a null slope was performed
1164 (red line). As expected, the coefficient of proportionality α is proportional to the gene length.

1165 However, some genes have a coefficient of proportionality lower than expected (points below
1166 the line) and may correspond to genes not uniformly covered.

1167 **Supplementary Figure 28: Impact of some data transformations on gene counts.**

1168 Raw (A), log-transformed (B) and square root transformed (C) abundance profiles of two core
1169 genes (MH0062_GL0071339 and MH0011_GL0069836) of the msp_0003 (*Bacteroides*
1170 *cellulosilyticus*) are compared.

1171 A. Raw data is not appropriate when the order of magnitude of counts is highly variable
1172 as it gives a too important weight to high counts.
1173 B. Log-transformation "spreads"the counts. However, it requires the introduction of a
1174 pseudo-count for zeros and variance tends to be proportional to the counts.
1175 C. Sqrt-transformation "spreads" the counts and stabilizes variance.

1176 **Supplementary Figure 29: MSPminer workflow**

1177 **Supplementary Figure 30 Number of genes per bin on the Integrated Gut Catalog**
1178 **with or without normalization.**

1179 Normalization consists in dividing gene counts by number of mapped reads per sample. The
1180 black line represents the optimal bin size (6 971 229/1 267=5 502 genes)