Comparative pangenome and mobilome analysis of clostridial species clusters reveals different levels of host adaptation and novel conserved biosynthetic potential

Lucas De Vrieze[1,2], Jan Aerts[3], Joleen Masschelein[1,2, #]

[1] Laboratory for Biomolecular Discovery & Engineering, Department of Biology, KU Leuven, Heverlee, Belgium

[2] VIB-KU Leuven Center for Microbiology, Heverlee, Belgium

[3] Augmented Intelligence for Data Analytics, Department of Biosystems, KU Leuven, Heverlee, Belgium

[#]Corresponding author, joleen.masschelein@kuleuven.be

Email addresses:

Lucas De Vrieze: lucas.devrieze@kuleuven.be

Jan Aerts: jan.aerts@kuleuven.be

Joleen Masschelein: joleen.masschelein@kuleuven.be

Abstract

Background

The class *Clostridia* comprises a diverse and industrially relevant group of anaerobes, ranging from free-living soil inhabitants to beneficial human gut symbionts. Analysing their genomic history and ecological adaptation dynamics beyond the species level allows us to gain deeper insights into their evolutionary traits and understand the impact of host adaptation on their innate functional capabilities.

Results

We constructed pangenomes and built core genome phylogenies for four ecologically distinct clostridial species clusters (rRNA clusters I, IV, XIVa and XIVb), analysing 492 clostridial genomes in total. All pangenomes were functionally annotated and fitted into a genomic divergence rate model, revealing a striking contrast in metabolic versatility and evolutionary dynamics among the host-associated clusters, IV and XIVa. Based on these findings, we hypothesized that cluster XIVa species are in the earlier stages of a host adaptation process compared to their cluster IV relatives. Targeted follow-up analyses confirmed this and yielded two additional key findings. First, cluster XIVa species are comparably less dependent on their host metabolically, and possess a more extensive secondary metabolism than anticipated for a host-associated clade. Second, a remarkable conservation of ranthipeptide biosynthetic gene clusters reveals underexplored biosynthetic potential with significant evolutionary advantages.

Conclusions

These results show that comparing genomic patterns beyond the species level can uncover distinct ecological trends, such as host and niche adaptation, across entire species groups. Understanding these trends provides valuable insights into the metabolic versatility and biosynthetic potential of these groups.

Keywords

Pangenome analysis, *Clostridia*, Host adaptation, Phylogenetic analysis, Secondary metabolism, Biosynthetic Gene Clusters, Mobilome analysis

Background

In 1880, Adam Prazmowski proposed a new genus, *Clostridium,* with *Clostridium butyricum*, originally described by Louis Pasteur, as the type strain [1]. Over time, this genus became a dumping ground for many Gram-positive anaerobic bacteria, resulting in a taxonomically incoherent compendium of species [2]. To address this, Collins et al. introduced a 19-cluster classification of *Clostridium* species in 1994, based on 16S rRNA analysis [3]. After several reassessments and expansions, this grouping still stands today [4, 5]. Nevertheless, ongoing reclassification efforts have further refined this framework, resulting in the establishment of the parent class *Clostridia*, the establishment of new genera, such as *Blautia* and *Hathewaya*, and the recognition of 16S rRNA cluster I as *Clostridium sensu stricto* [2, 6–8].

Species in the *Clostridia* class display a remarkable variety of lifestyles and phenotypes, ranging from thermophiles and psychrophiles to acidophiles and pathogens. They occupy diverse ecological niches, including soil, sediments and the intestinal tracts of humans and animals. This versatility underlies their increasing industrial and medical relevance, particularly in three key areas. First, species like *C. acetobutylicum* and *C. beijerinckii* stand out for their high innate metabolic flexibility. Advances in understanding their genetics have enabled their transformation into industry-grade microbial factories for the breakdown of complex lignocellulosic polysaccharides [9–11]. Moreover, they are industrially employed for the production of solvents, organic acids, alcohols and H_2, e.g. through the acetone-butanol-ethanol (ABE) fermentation process [12–16]. Second, *Clostridium* species were the first anaerobes discovered to produce biosynthetic gene cluster (BGC)-encoded secondary metabolites with potent biological activities (e.g. antibacterial) [17–24]. Recent bioinformatic analyses have revealed that they harbour an even greater repertoire of BGCs than initially anticipated, substantially expanding secondary metabolite research into the anaerobic realm [25, 26]. Third, species like *C. botulinum*, *C. perfringens*, *C. noyvi* and *C. tetani* are notorious human pathogens and have been studied extensively in medical contexts due to their virulence and toxin production [27–30]. Furthermore, species of rRNA clusters IV and XIVa are prominent gut commensals, making up 10-40% of the healthy intestinal microbiome in humans and animals [31–33]. These species play

68 an important role in maintaining gut homeostasis and overall health by producing beneficial substances, such
69 as short-chain fatty acids, and preventing pathogen colonisation [31, 34].
70 Despite the abundance and relevance of clostridial species, few comparative studies have explored their
71 mutual evolutionary history and ecological adaptations beyond the species level. Such research could uncover
72 shared genetic features or evolutionary trajectories that enhance our understanding of host adaptation and
73 the genetic traits defining distinct rRNA species clusters. This, in turn, might help reveal hidden genomic
74 patterns and facilitate the identification of new probiotics.
75 Previous higher-level comparative studies have focused on individual rRNA clusters or genera. For example, a
76 study on *Clostridium sensu stricto* (or rRNA cluster I) found evidence of a highly plastic genome, harbouring a
77 relatively large metabolic core and many specialised genetic capabilities that underpin the mainly free-living
78 nature of this cluster [35]. Another study focusing on the gut-associated *Blautia* genus of rRNA cluster XIVa
79 revealed that many gene exchange events are associated with metabolic functions, likely driven by the
80 dynamically changing host environment. Moreover, horizontal gene transfer may have caused a metabolic
81 dependence on other gut bacteria [36].
82 In this work, we present a seminal study of the relationship between genome dynamics and host or niche
83 adaptation at the rRNA species cluster level of the *Clostridia* class. We examined a selection of four clostridial
84 rRNA clusters with distinct ecological properties. Cluster I was included because of its metabolically versatile
85 free-living species (e.g. *C. acetobutylicum*) and notorious human pathogens (e.g. *C. botulinum*). Clusters IV
86 and XIVa were selected for their host-associated and commensalistic nature, likely supported by substantially
87 different genetic bases (e.g. *Ruminococcus*, *Blautia*). Lastly, cluster XIVb was chosen for its remarkable mix of
88 free-living extremophiles (e.g. deep-sea halophiles) and microbiome-associated species. First, we carried out
89 a pangenome and a phylogenetic analysis to map the genetic diversity and evolutionary relationships of these
90 clusters. Next, we analysed the gene exchange dynamics and reconstructed ancestral gene states to
91 understand the spread and origin of gene families. We also screened all genomes for mobile genetic elements
92 (MGEs) that can be linked to increasing host adaptation. Finally, we zoomed in on secondary metabolite BGCs,

assessing their conservation and evolutionary patterns. Together, these analyses provide intriguing insights into the evolutionary dynamics and genetic basis of host and niche adaptation in clostridial species.

Results

ANI similarity analysis suggests different levels of genomic heterogeneity within rRNA clusters

As a first step to assess the genomic diversity among the rRNA clusters, pairwise average nucleotide identity (ANI) values were calculated and hierarchically clustered for our entire clostridial genome dataset (**Add. File 1)**. This analysis revealed four high-level ANI groups, three of which almost exclusively consist of a subset of the rRNA clusters (**Figure 1**). ANI group 1 almost completely identifies with rRNA cluster I and shows no detectable genome similarity with genomes from other rRNA clusters, except for a few from cluster XIVb. ANI groups 2 and 3 mostly comprise rRNA cluster XIVa and IV species, respectively. In contrast, ANI group 4 is much less coherent, lacking a high average intra-group similarity. Furthermore, it contains species belonging to clusters IV, XIVa as well as XIVb. Hence, this mixed ANI group suggests greater genomic heterogeneity within parts of these three rRNA clusters. Additionally, 24 species pairs had an ANI value exceeding the maximum threshold of 96% typically used to delineate different species (**Add. File 2: Table S1**), raising questions about whether these closely related species truly represent two separate ones. Notably, all threshold-exceeding taxa pairs are located adjacent to each other on a phylogenetic tree of all clostridial reference genomes (**Add. File 3: Figure S1**).

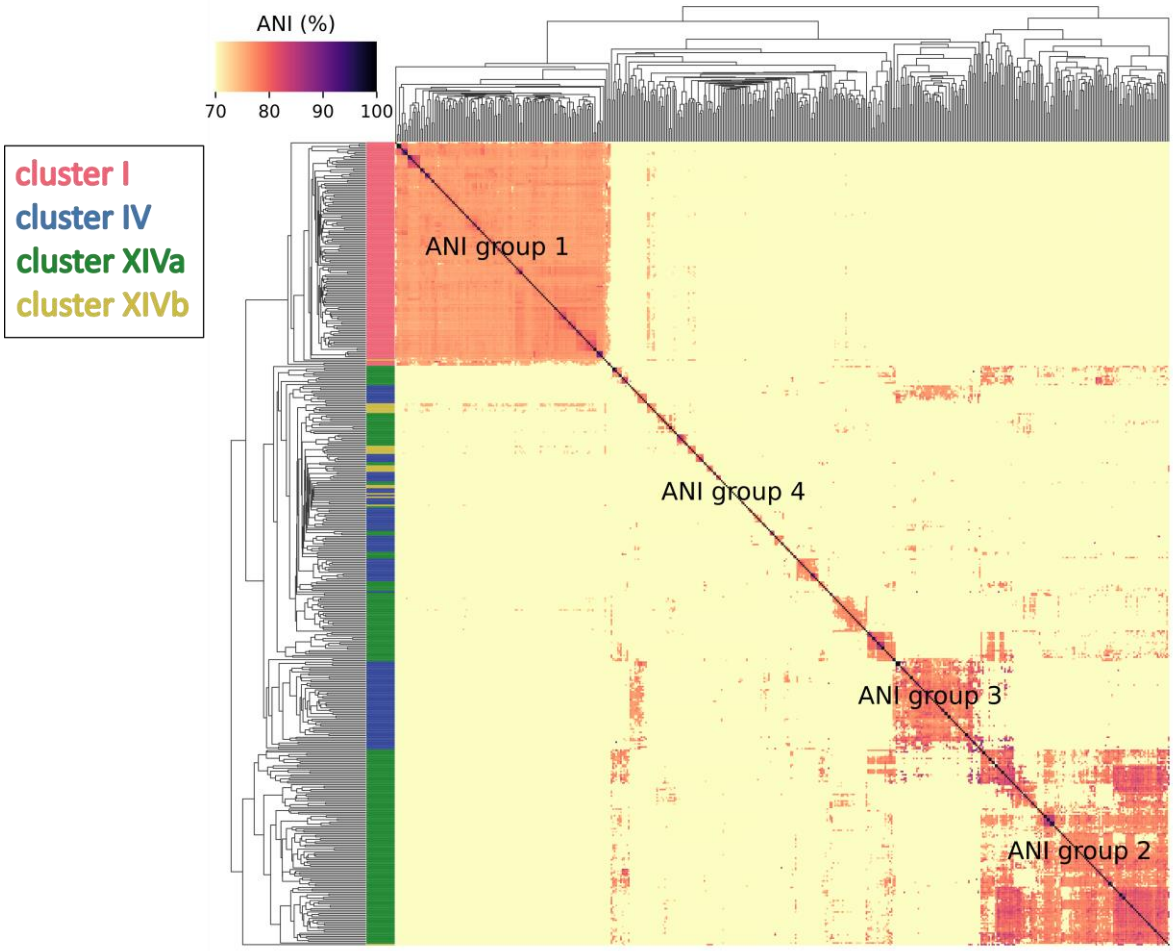

Figure 1: Hierarchically clustered heatmap of pairwise ANI values for all genomes in the full *Clostridia* set. In case no similarity was detected, a default value of 70% was imputed. rRNA cluster annotations are indicated on the left. Four high-level ANI groups have been identified, of which group 4 is more heterogeneous, as illustrated by the absence of a large high ANI block.

110 ***Construction of pangenomes for each rRNA cluster***

111 Next, a pangenome was constructed for each rRNA cluster to identify their conserved genetic base (core
112 partition) and determine which gene families are shared only partially (shell partition) or are unique (cloud
113 partition). This analysis revealed 378, 104, 211, 248 and 104 core gene families for rRNA clusters I, IV, XIVa,
114 XIVb and our full clostridial set, respectively. These core genomes are very small, representing only 0.03 % -
115 0.67 % of the pangenomes (**Table 1**). Interestingly, the core genome of cluster IV contains substantially fewer
116 genes in absolute numbers and is at par with the core genome of the full set. Its cloud is of comparable size
117 to those of the other clusters, suggesting a higher level of intra-cluster genomic heterogeneity. Such a
118 heterogeneity aligns with the observations from the ANI analysis.

Table 1: Number of gene families in the pangenome partitions of each examined genome set, in absolute numbers and referenced to the total number of gene families in each genome set.

	Cluster I (# = 136)	**Cluster IV (# = 123)**	**Cluster XIVa (# = 211)**	**Cluster XIVb (# = 22)**	**Full set (# = 492)**
Core (> 99 %)	378 0.37 %	104 0.09 %	211 0.12 %	248 0.67 %	104 0.028 %
Shell (15 % - 99 %)	3526 3.49 %	2151 1.92 %	3013 1.76 %	3198 8.64 %	2843 0.76 %
Cloud (< 15 %)	97,116 96.14 %	109,665 97.99 %	168,295 98.12 %	33,547 90.68 %	371,219 99.21 %

The computed accumulation curves indicate that the pangenomes of all rRNA clusters are open, meaning that each genome is likely to harbour unique gene families. This finding highlights the high level of genomic plasticity that is typical for species living in highly dynamic environments like open ecosystems or the gastrointestinal tract. Furthermore, the gene frequency U-shape plots (**Figure 2**, **Add. File 3: Figures S2-S5**) show that most genes are only conserved on the sequence level among a limited number of closely related species, often fewer than ten. This is consistent with the high genomic diversity at the class taxonomic level, which often pushes the sequence similarity between homologous genes beyond the clustering identity threshold.

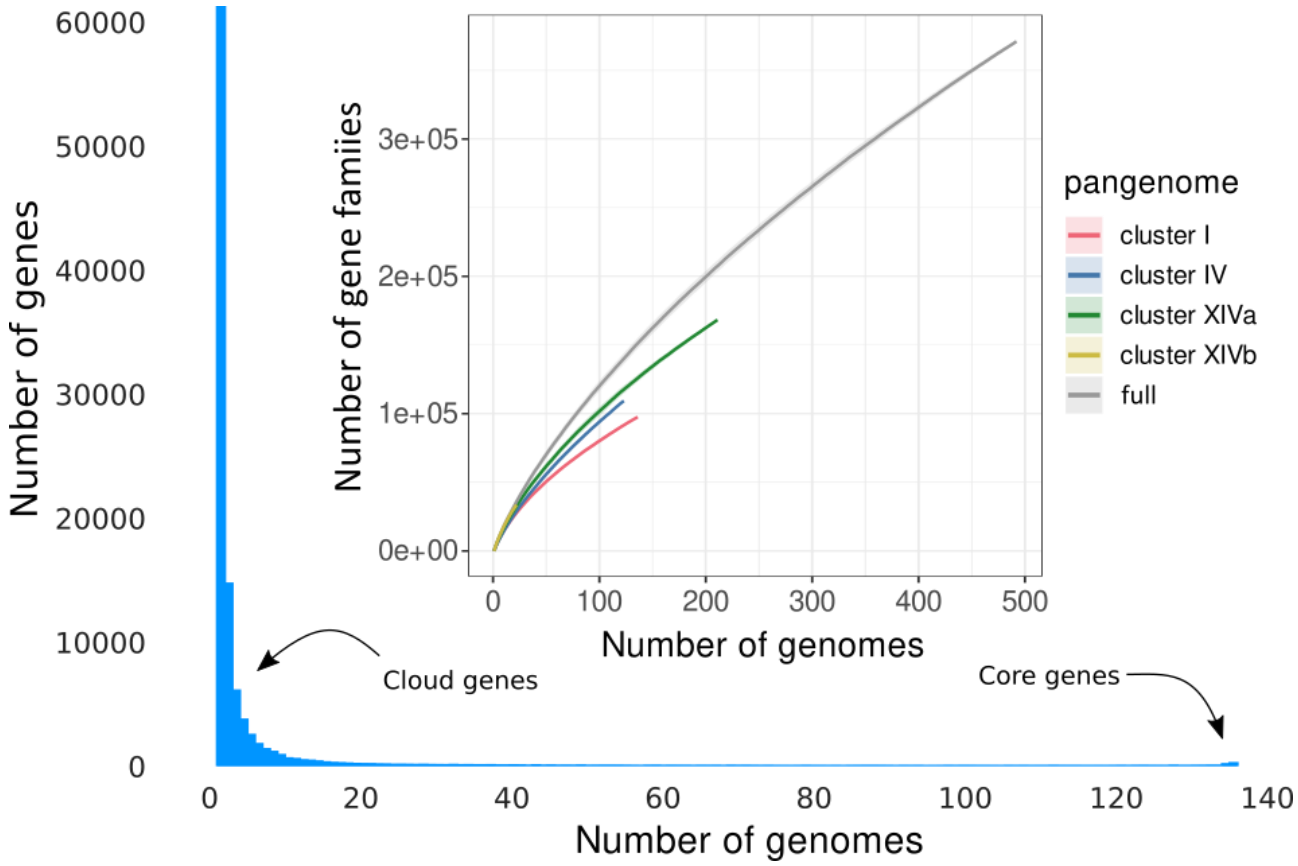

Figure 2: U-shape plot of the distribution of the 101,020 gene families in the cluster I pangenome. (Inset) Gene family accumulation curves for all examined genome sets. All pangenomes are open, indicating highly plastic genomes, which is typical for species living in highly dynamic environments.

Core genome phylogeny confirms rRNA-based clades and exposes different trends in genome size

To establish an evolutionary baseline, a phylogeny was constructed for the full clostridial set (**Figure 3**). The resulting phylogenetic tree is split into three main branches corresponding to rRNA clusters I, IV and XIVa. Bootstrap support for the deepest tree nodes is high, corroborating the rRNA-based grouping. The internal structure of cluster IV, however, is not as robust, as exemplified by the very low 56 % bootstrap support at one of its internal nodes. Remarkably, the taxa of cluster XIVb do not group together in one clade but instead appear as small, distantly related groups within cluster XIVa. Hence, it might not be appropriate to refer to XIVb as a distinct 'cluster'. Since they are confidently separated according to the bootstrap values, cluster XIVb may instead comprise species from different undersampled clostridial species clusters.

The rRNA cluster I lineage comprises several species of *Clostridium sensu stricto*, such as solvent producers *C. acetobutylicum*, *C. beijerinckii* and *C. kluyveri,* as well as renowned pathogens like *C. tetani*, *C. botulinum* and *C. novyi*. The average genome size within this clade is 4.09 Mb (+/- 1.01 Mb), a broad range that likely reflects

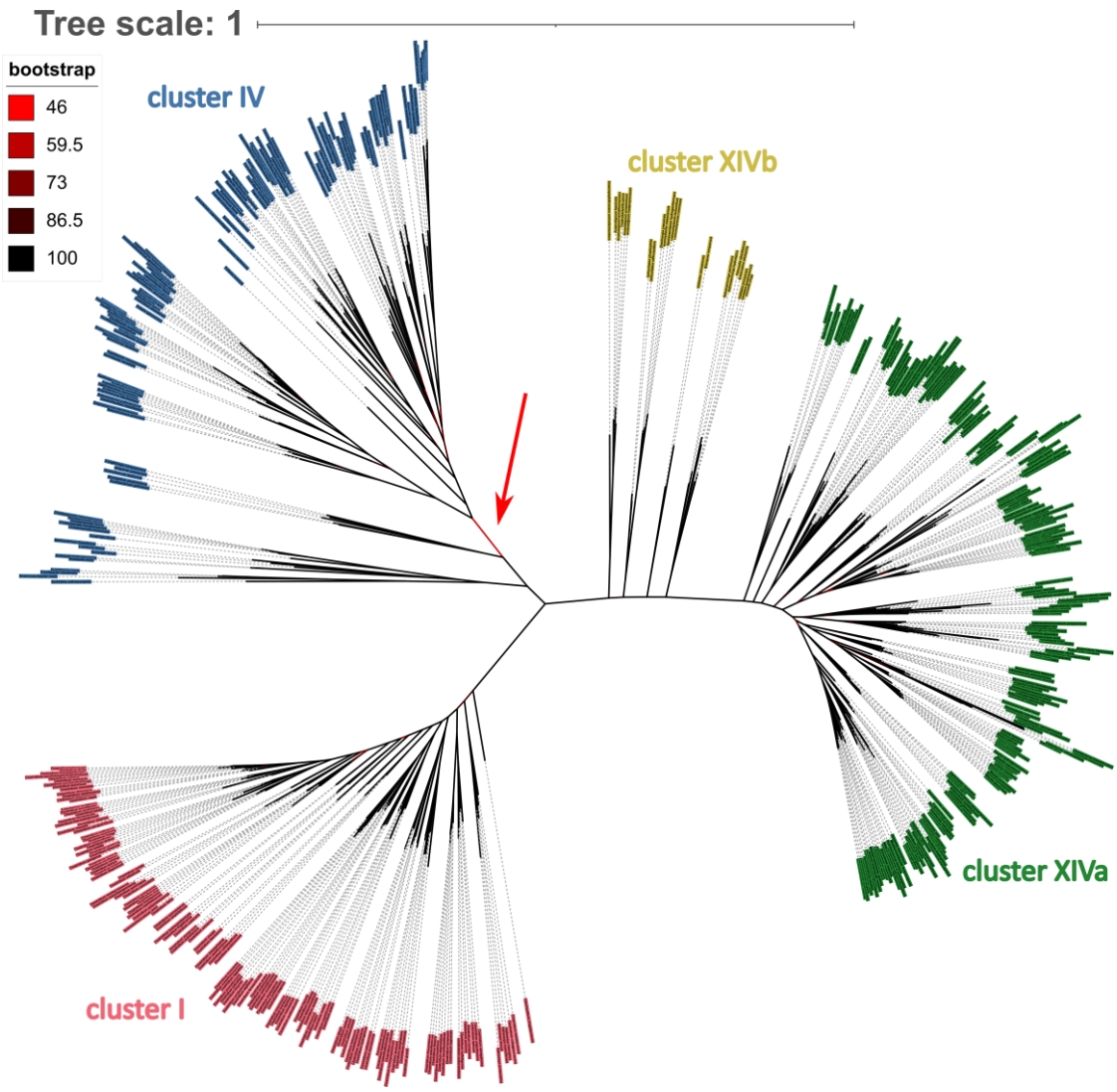

Figure 3: Unrooted phylogenetic tree of the full *Clostridia* set. Bootstrap support is visualized at the tree branches, while the rRNA cluster annotations are visible as background colour of the leafs. Note the low bootstrap at the lower branches of cluster IV, indicated with a red arrow.

139 the lifestyle of the species. Pathogens and other host-associated bacteria tend to have smaller genomes
140 tailored to their host environments, while the more versatile free-living bacteria often possess large genomes
141 harbouring many specialised abilities [37]. The genomes in this lineage seem to confirm this trend, as the
142 smallest genomes can be found in pathogens like *C. novyi* and relatives, while the largest genomes belong to
143 metabolically versatile species like *C. beijerinckii* and *C. saccharoperbutylacetonicum*.
144 The rRNA cluster IV lineage consists of at least four distinct subgroups, as indicated by the phylogenetic tree
145 (**Figure 3**) and the UMAP plot (**Add. File 3: Figure S6**). The average genome size of cluster IV species is 3.00
146 Mb (+/- 0.52 Mb), which is remarkably smaller and less variable than that of cluster I species. This can be
147 explained by the fact that many species in this lineage are members of the gut microbiome, of which the

148 genomes are typically reduced in size due to host adaptation [38, 39]. Unlike cluster I, which mostly contains
149 *Clostridium* species, cluster IV contains many different genera, that sometimes do not cluster well or appear
150 closely related to other genera. The most striking example in this regard is the *Ruminococcus* genus, which is
151 part of two separate subgroups within this lineage, both with high bootstrap support. Similarly, the genus
152 *Faecalispora* is closely related to two *Clostridium* species, *C. merdae* and *C. minihomine*. The fact that these
153 species bear the *Clostridium* label suggests that they have yet to be reclassified. Their close phylogenetic
154 relationship with *Faecalispora* provides an argument for their reclassification into this genus.
155 The lineage of cluster XIVa includes prominent gut microbiome species, such as the *Blautia* genus and several
156 *Ruminococcus* species. The mean genome size of this lineage is 3.97 Mb (+/- 1.18 Mb). Interestingly, while
157 the average size is comparable to that of cluster I, the standard deviation is notably larger. This is remarkable
158 for a clade consisting of many presumed gut-associated species. This lineage also encompasses a large
159 number of genera that appear to have undergone different levels of taxonomic reclassification. The most
160 striking example of inconsistency is again the *Ruminococcus* genus, which is situated in two distinct branches
161 of this lineage. Taken together with its appearance in cluster IV, this genus is clearly in need of a taxonomic
162 review. Another example is the *Lacrimispora* genus, which is interspersed with *Clostridium* spp. like *C. indolis*
163 and *C. aerotolerans*, posing a strong argument for reclassifying these spp. as *Lacrimispora*.
164 Lastly, for cluster XIVb, the average genome size is 3.86 Mb (+/- 1.13 Mb), which is comparable to that of
165 cluster XIVa. An examination of the isolation sources in the NCBI BioSample database revealed that this
166 lineage contains a remarkable mix of both extremophile free-living bacteria, such as halophiles and deep-sea
167 sediment isolates, as well as presumed gut-associated species isolated from sources like faeces or rumen.
168 This large ecological diversity raises questions about the appropriateness of grouping these species in a single
169 taxonomic cluster.
170 ***Functional annotation confirms different levels of intra-cluster heterogeneity***
171 A functional annotation serves to grasp the high-level function of all gene families in each partition, revealing
172 which functions are conserved across all genomes. The core genome consists for a large part of ribosomal
173 proteins, as illustrated by COG category J and BRITE level A category Genetic Information Processing.

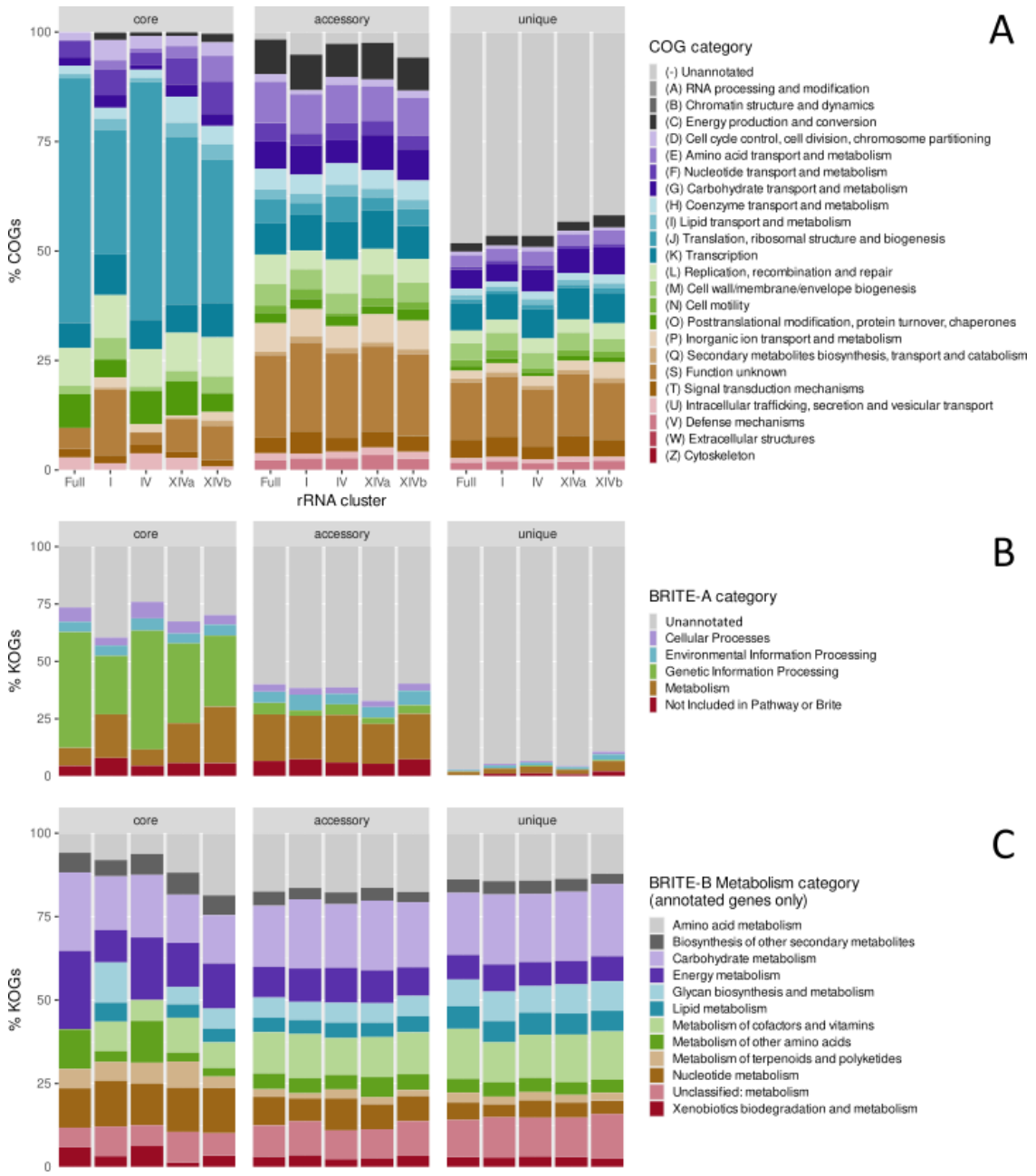

Figure 4: Functional annotation of all pangenome partitions of all examined genome sets by COG terms (A) and by KOG/BRITE terms, both at level A (B), and at level B for A = Metabolism (C). Percentages are referenced to the full gene set (A and B), or to the set of annotated genes (C).

174 Interestingly, the proportion of ribosomal proteins is larger for the more heterogeneous cluster IV, probably
175 due to a lower number of other conserved genes. In addition, cluster IV species do not share a large
176 metabolism of cofactors and vitamins, and lack a common metabolism for glycans and lipids (see **Figure 4C**).
177 On average, about 10% of the core genes were not annotated by COG (categories - and S), and around 20%
178 did not receive an annotation in the BRITE system.

179 For the annotated fractions of the accessory and unique partitions of the pangenomes, the differences are
180 less pronounced. However, the unannotated fractions are much larger (for COG, about 25% on average in the
181 accessory genome and about 60% in the unique; for BRITE, about 65% in the accessory and about 95% in the
182 unique). Most unannotated proteins were part of the unique partition, suggesting the existence of many
183 highly specialised functions or rapidly evolved proteins that fall outside the scope of our current
184 understanding.

185 ***KEGG pathway completeness analysis uncovers two levels of metabolic versatility***

186 A KEGG pathway completeness analysis gauges the metabolic versatility of an organism by assessing the
187 completeness of metabolic pathway modules as documented in the KEGG Pathway database, based on the
188 genomic presence of their constituent genes. Remarkably, two distinct levels of pathway completeness were
189 observed within the full clostridial genome set. 191 out of 492 organisms have more complete pathway

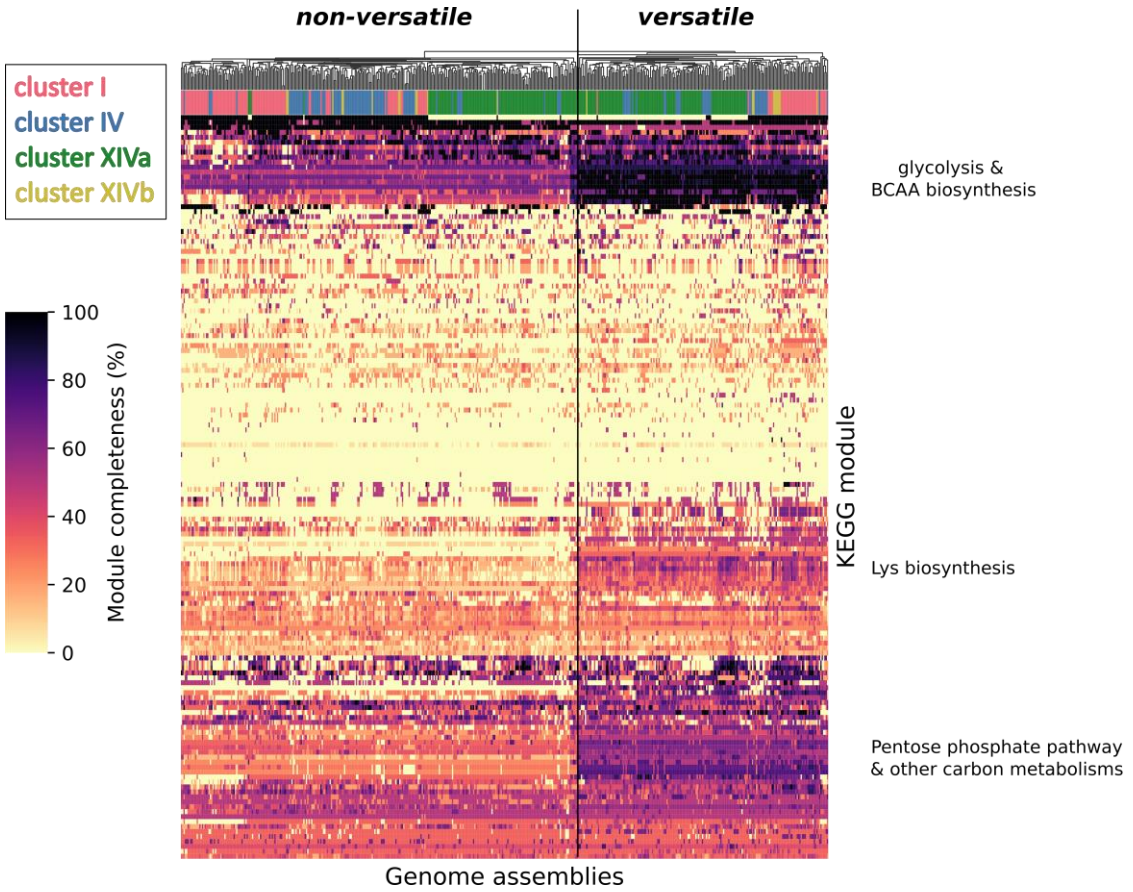

Figure 5: Hierarchically clustered heatmap of KEGG module completeness, annotated by rRNA cluster. The recurringly higher completeness for multiple modules shows that this is a genome-wide trend. For a figure with all KEGG module labels, see Add. File 1: Fig. S7. Summarising, the upper key differing module block comprises glycolysis-related and BCAA biosynthesis pathways, while the lower two blocks encompass Lys biosynthesis, and the pentose phosphate pathway and other carbon metabolisms, resp.

190 modules, suggesting a richer and thus more versatile metabolism (**Figure 5**). This discrepancy is not driven by
191 a limited number of pathways but represents a genome-wide trend, as many modules tend to be more
192 complete. This aligns with earlier observations of a remarkably high metabolic versatility within the *Bacillota*
193 phylum, especially within the *Clostridium* genus [40, 41].
194 Interestingly, clusters XIVa and XIVb tend to host a higher fraction of versatile organisms than clusters I and
195 IV (50% and 45% vs. 28% and 29%, resp.). Also, this versatility appears independent of genome size (**Figure**
196 **6**). Multiple clostridial genera contain both versatile and non-versatile species, underscoring an intra-genus
197 diversity that has previously been reported [41]. Nevertheless, some genera, such as *Enterocloster* and
198 *Lacrimispora*, and one lineage of *Blautia*, are entirely composed of metabolically versatile species.

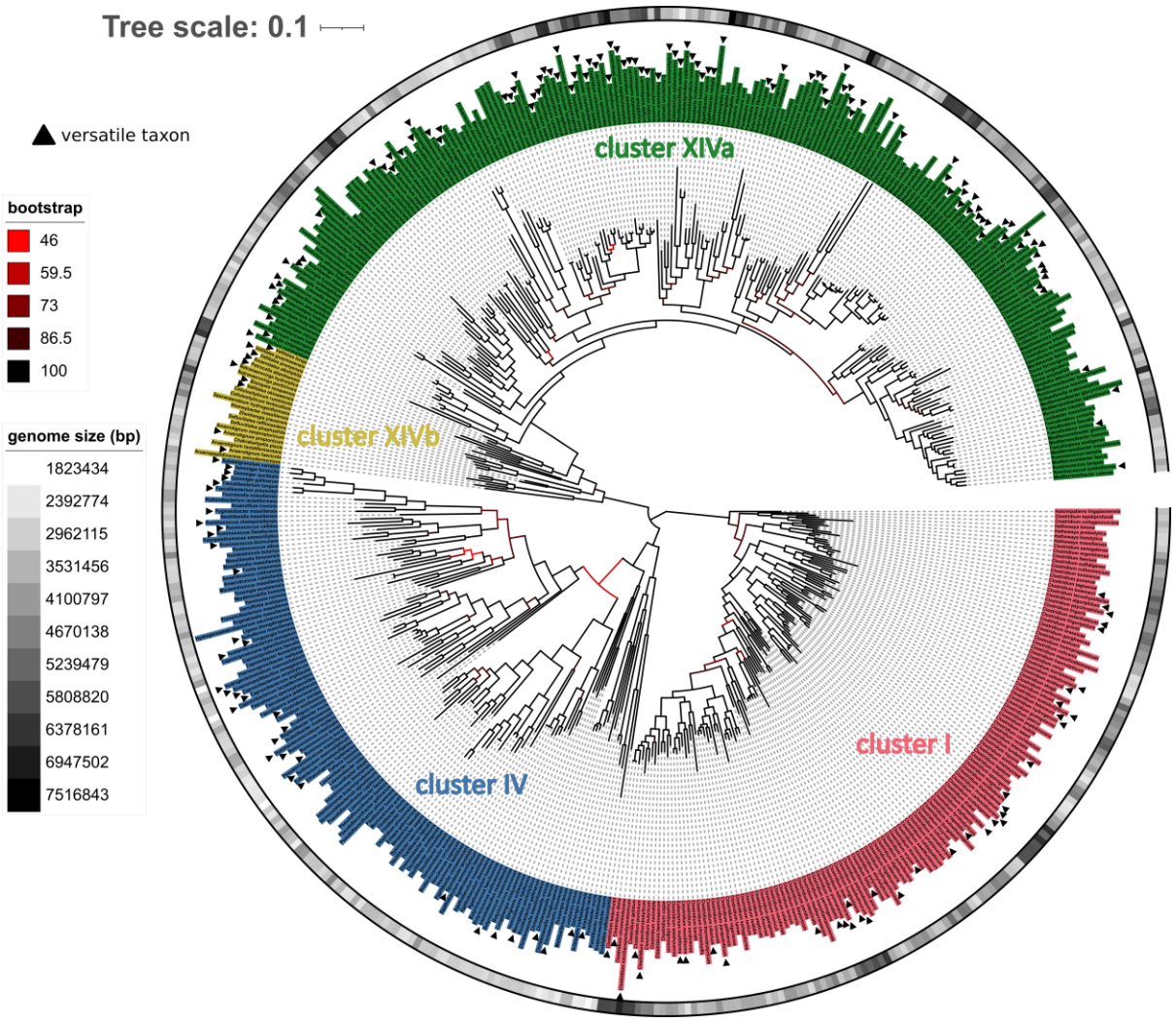

Figure 6: Rooted phylogeny annotated by rRNA cluster (label colour), bootstrap support (branch colour), genome size (colour gradient ring) and high metabolic versatility (black triangular markers). Metabolically versatility does not cluster well with phylogenetic descent nor with genome size, yet some lineages are consistently versatile (cfr. *Blautia, Enterocloster-Lacrimispora*).

199 The most prominent differences in metabolic module completeness are observed in core metabolic
200 processes: the central carbon metabolism (glycolysis, pentose phosphate pathway, citrate cycle) and amino
201 acid biosynthesis (Pro and Arg biosynthesis, Val, Ile and Leu biosynthesis, Lys biosynthesis). This indicates that
202 a higher metabolic versatility translates in an ability to synthesise more metabolites from less nutrients,
203 underpinning a higher metabolic independence than one would expect from a host-associated bacterium.
204 Furthermore, a more diverse primary metabolism may support the development and maintenance of a more
205 extensive secondary metabolism, providing further functional and adaptive advantages.

206 ***Genomic divergence rate analysis reveals significantly different evolutionary dynamics***

207 The ANI analysis and pangenome annotation revealed that the genomic diversity within cluster IV is higher
208 than within other clusters, suggesting a divergent evolutionary trajectory. This difference may lie, for example,
209 in the intrinsic mutation rate and/or different gene gain/loss dynamics.

210 To assess whether this genomic heterogeneity indeed results from distinct evolutionary dynamics, gene
211 exchange (GE) models were fitted for each rRNA cluster separately and statistically compared to each other.
212 These models were constructed using the Panstripe algorithm, which takes the core genome phylogeny and
213 the binary gene family presence/absence data as inputs. It fits a compound generalised linear model (GLM)
214 to relate phylogenetic branch lengths with the number of genes exchanged at those branches. Using an
215 additional GLM term, it differentiates between internal nodes of the tree and terminal ones, yielding core
216 and tip gene exchange rates, respectively. Exchange rates significantly different from 0 signify an association
217 between branch length and the number of gene exchange events, and thus an ongoing gene exchange
218 process. The core rate can be interpreted as the overall gene exchange rate underpinning the general
219 structure of the phylogeny, while the tip rate can be considered a snapshot of short-term gene exchange
220 driven by gene mobility processes. However, in highly diverse genome sets like the one in this work, an
221 orthologous gene family may be falsely split into multiple gene families during pangenome construction,
222 artificially inflating both core and tip gene exchange rates. Consequently, Panstripe models a more general
223 genomic divergence, capturing both intrinsic gene evolution as well as gene exchanges, rather than only the

224 latter. Nevertheless, this generalised genomic divergence rate is sufficient for addressing our research
225 question, as it provides a measure of the speed at which clostridial genomes have evolved.
226 The GE modelling revealed several interesting findings (**Figure 7**) (for fitted models and p-values of the model
227 comparison: see **Add. File 2: Tables S2-S3**). First, the core divergence rate of cluster IV is significantly lower
228 than those of clusters I and XIVa. This aligns well with its presumed host-associated status, as host-adapted
229 genomes tend to evolve at a slower rate [42, 43]. In sharp contrast, cluster XIVa, the other clade of host-
230 associated species, does not display a lower divergence rate. Instead, its divergence rate is comparable to
231 that of cluster I, a clade mostly composed of free-living species. Lastly, cluster XIVb appears to align more
232 closely with clusters I and XIVa in terms of divergence rate, but the comparisons lack statistical power due to
233 the low number of species in cluster XIVb.

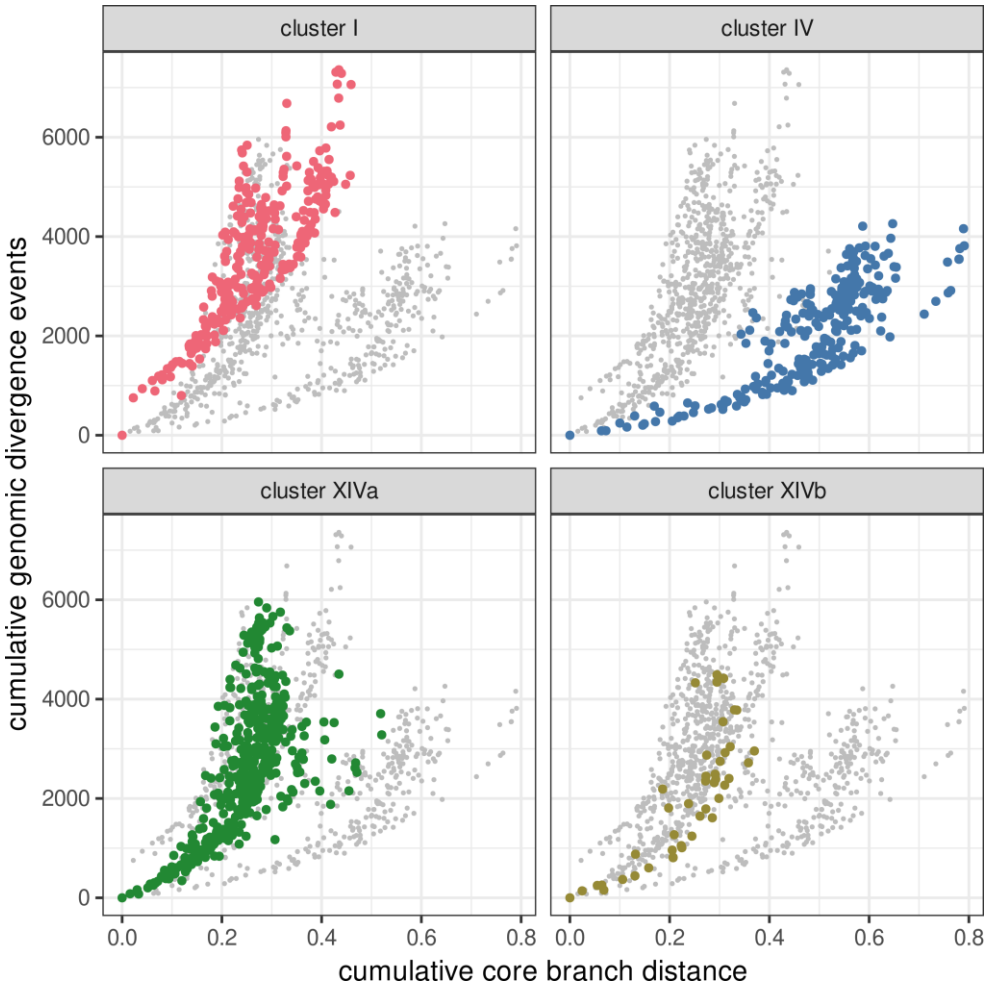

Figure 7: Cumulative number of genomic divergence events (a.o. gene exchange events) versus the cumulative branch length starting from the root of the subtree of each rRNA cluster genome set. Cluster IV has a significantly lower divergence rate consistent with its host-associated nature, whereas cluster XIVa displays a high divergence rate inconsistent with its presumed host-associated nature. Cluster XIVb visibly sides with cluster XIVa but is too small to draw statistically sound conclusions.

234 ***Ancestral state reconstruction uncovers different genetic bases for the host-associated clades***

235 An ancestral state reconstruction (ASR) enables the identification of genes that likely constitute the inherited
236 genetic base of a certain clade within a phylogeny, allowing clade-specific traits to be inferred. Indeed, the
237 ASR of this phylogeny shows that a noticeable number of genes have been gained and lost in the root node
238 of each rRNA cluster (**Figure 8**). Additionally, the number of genes exchanged in a node tends to increase
239 closer to a terminal node, with up to 4078 gained and 365 lost genes. This may reflect the more mobile genes
240 of the accessory and cloud pangenome that are present in a more limited subset of the examined species.

241 At the root of the tree, the model predicted 492 gene families, active in central subsystems like glycolysis,
242 tRNA biosynthesis and the ribosomes (BRITE level C annotations) **(Add. File 3: Figures S8-S9, Add. File 4**). As
243 these predictions also include genes for flagellar assembly, two-component signalling (2CS) systems and DNA
244 repair systems, the last common ancestor of the clostridial lineage may have been a well-equipped free-living
245 microbe.

246 Regarding the most recent common ancestor (MRCA) of the rRNA clades, some interesting contrasts can be
247 observed in both the number and functional annotation of the exchanged genes (**Add. File 3: Figure S7**). The
248 MRCA of cluster I lost no genes and gained 603, half of which have been annotated. The annotated functions
249 are related to cell motility mechanisms (flagellar assembly, chemotaxis), coenzyme biogenesis (porphyrin),
250 DNA repair systems, and signal transduction (2CS) (**Add. File 3: Figures S8-S9, Add. File 4**). This suggests that
251 the MRCA of cluster I retained and further expanded the ancestral free-living capacities, likely to adapt to
252 new dynamic environments.

253 The MRCA of cluster IV only gained 22 genes, but lost 117. The gained genes are linked to amino acid and
254 other carbon metabolisms, while the lost genes are associated with nucleotide metabolism, signalling
255 systems (2CS), cell motility (flagellar assembly, chemotaxis), riboflavin synthesis and DNA repair systems (**Add.
256 File 3: Figures S8-S9, Add. File 4**). These losses align with the reduction in free-living capabilities that typically
257 occurs when a microorganism transitions to a host-associated lifestyle. The loss of DNA repair systems is a
258 well-documented characteristic of this lifestyle shift [44]. This indicates that the MRCA of cluster IV may have
259 gone through this lifestyle conversion.

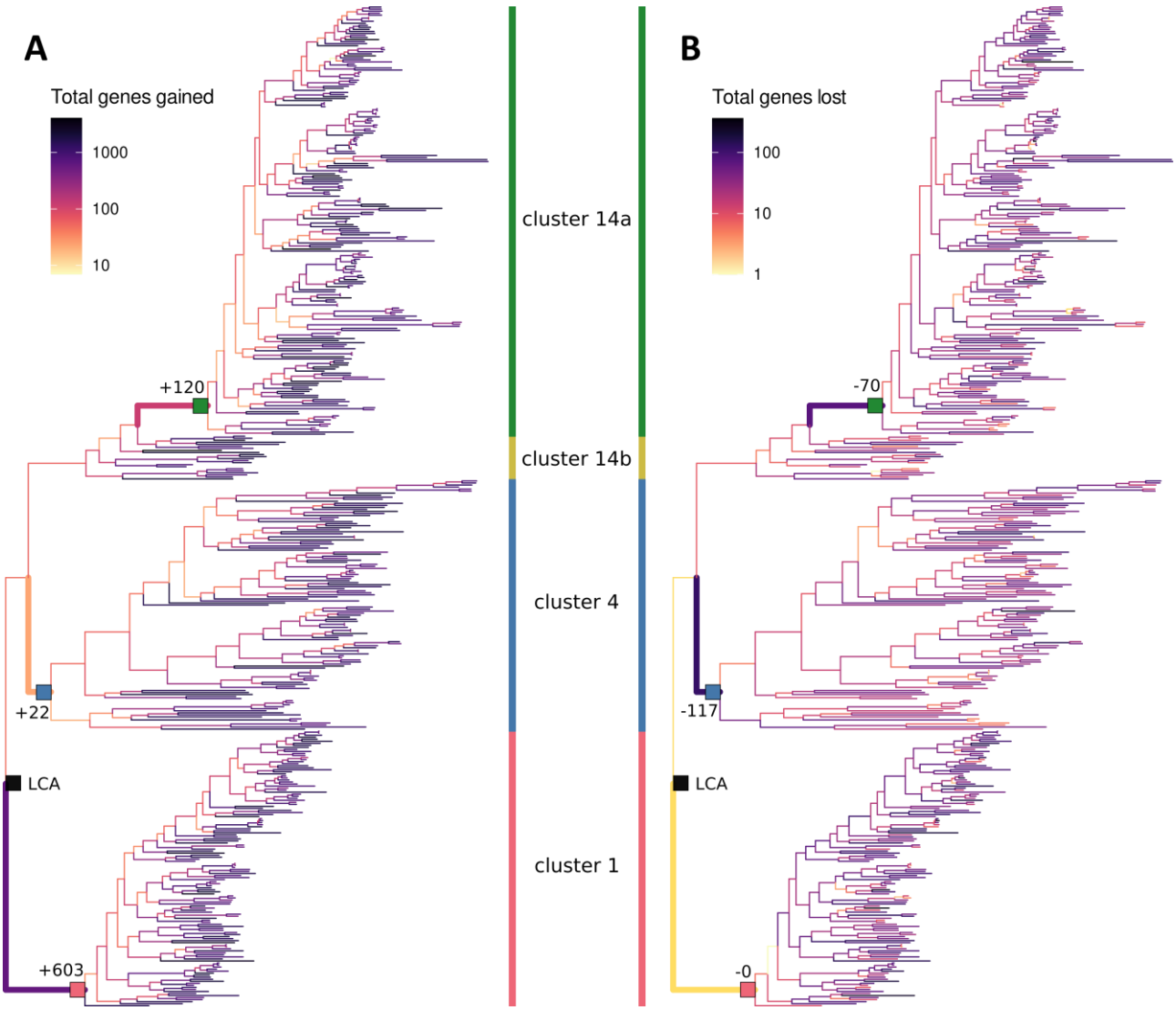

Figure 8: Wagner parsimony gene gain (A) and loss (B) profiles over the core gene phylogeny of the full *Clostridia* set. In the root nodes of each examined rRNA cluster (square markers on top of thickened branch), a notable number of genes has been exchanged (numbers on top of markers). The number of exchanged genes at a node tends to increase the closer it is to a terminal node.

In contrast, the MRCA of cluster XIVa gained 120 genes and lost 70. Gained functionalities include amino acid cata- and anabolisms, butanoate metabolism and, interestingly, pathways for synthesizing secondary metabolites, such as streptomycin, polyketide sugar units, amino and nucleotide sugars, and acarbose and validamycin. This suggests that the MRCA of cluster XIVa was well equipped for secondary metabolite synthesis, especially for polyketides and aminoglycosides. The lost genes include those related to cell motility systems, 2CS and amino acid biosynthesis, consistent with an increasingly host-associated lifestyle (**Add File 1: Figures S8-S9, Add. File 4**). The combination of gained metabolic capabilities and the loss of free-living pathways suggests that cluster XIVa has a different level of host association than its cluster IV relatives.

268 ***Focused mobilome screen confirms different levels of host adaptation***

269 Both the GE and the ASR analyses indicate that cluster IV and XIVa have a different level of host adaptation.
270 Moreover, the KEGG pathway completeness assessment shows that species within cluster XIVa are more likely
271 to be metabolically versatile. Due to its broader metabolic repertoire and higher genomic divergence rate, we
272 hypothesise that cluster XIVa is less dependent on its host than cluster IV because it is in an earlier stage of
273 the host adaptation process. To support this hypothesis, we screened our entire genome set for the
274 abundance of a set of genomic hallmarks tied with increasing levels of host adaptation.

275 More specifically, the transition from a free-living to a host-associated lifestyle in bacteria is thought to be
276 coupled with a genome streamlining process driven by insertion sequences (ISes) and intrinsic mutations [45].
277 ISes are a type of intracellular mobile genetic elements (MGEs) that employ transposases for their mobility.
278 When bacteria move from a nutritionally poor free-living environment to a nutritionally rich host-associated
279 setting, many genes lose their essentiality due to the constant availability of metabolites. This causes a
280 reduced selection pressure on non-essential genes, enabling higher mutation rates as well as IS insertions
281 into these genes. Both these phenomena drive pseudogene formation [46]. Over time, the accumulation and
282 multiplication of IS regions promotes genome rearrangements and deletions, ultimately yielding smaller
283 streamlined genomes tailored to their new host-associated lifestyle [45].

284 To explore this hypothesis, we screened the full clostridial genome set for the abundance of the following
285 selection of MGEs and related genomic features: (1) genome sizes, (2) IS elements, (3) pseudogenes, (4)
286 prophage regions, and (5) biosynthetic gene clusters. (1), (2) and (3) are key hallmarks of the genome
287 streamlining process. (4) may reflect a shift in environmental niche as such a change influences the type and
288 abundance of infecting phages. (5) gives an indication of the metabolic versatility of the organism, as the
289 necessary precursor metabolites must be available.

290 The results show interesting parallels. First, cluster IV genomes are significantly smaller than those of cluster
291 XIVa, which, in turn, are significantly smaller than cluster I genomes (**Figure 9A**). Cluster XIVb genomes are
292 significantly larger than those of cluster IV, but the number of genomes in this group is too small to make
293 further sound statistical conclusions.

294 Second, regarding IS elements, cluster XIVa genomes harbour significantly more IS elements than cluster I
295 genomes, which, in turn, count significantly more IS elements than cluster IV genomes (**Figure 9B**).
296 Interestingly, this observation aligns with the hypothesis that cluster XIVa is undergoing an IS expansion and
297 is only in the early stages of host adaptation. On the other hand, cluster IV species appear to be in the later
298 stages of genome streamlining, where both IS sequences and genomic regions have been deleted. Again, the
299 small sample size of cluster XIVb genomes precludes sound statistical analysis.
300 A deeper analysis of IS families revealed some substantial differences in their distribution among the rRNA
301 clusters (**Add. File 3: Figure S10**). For example, families IS110, IS66 and ISL3 are more prevalent in cluster XIVa
302 species. These IS families do not have a specific insertion site preference [45, 47], so higher abundances of
303 these ISes again indicates more IS replications in cluster XIVa. On the other hand, families IS481 and IS5 are
304 more abundant among cluster IV species. Intriguingly, there is documented evidence that IS481 has played a
305 fundamental role in the IS expansion and genome reduction process in *Bordetella pertussis*, a human
306 pathogen causing whooping cough [48]. Hence, the increased abundance of this IS family in cluster IV species
307 supports the notion that an IS expansion process may have taken place in cluster IV species.
308 Third, pseudogenes are significantly less abundant in cluster I than in clusters IV and XIVa (**Figure 9C**). This
309 corroborates the hypothesis that both gut-associated clusters have undergone or are undergoing a phase of
310 more intense pseudogene formation. While cluster XIVb contains significantly fewer pseudogenes than
311 cluster XIVa, the limited sample size prevents definitive conclusions.
312 Fourth, the number of BGCs is significantly higher in clusters I and XIVb, with cluster XIVa at similar lower
313 levels as cluster IV (**Figure 9D**). This suggests that cluster I and cluster XIVb species maintain a greater capacity
314 for specialised metabolism, whereas the other two clusters seem to have eliminated a substantial share of
315 their BGCs. Given that other genomic features indicate that cluster XIVa is only in the early stages of host
316 adaptation, it seems that biosynthetic gene clusters are among the first to be eliminated under altered
317 selection pressures, yet this is independent from non-BGC metabolic versatility. In terms of BGC classes, the
318 most substantial differences in abundance are observed in the cyclic lactone autoinducer (CLA), NRPS(-like),

319 and RiPP-like classes. Except for the CLA class, the abundance follows a consistent pattern: IV < XIVa < XIVb =
320 I. In addition, there is a consistent presence of ranthipeptides across all clusters (**Add. File 3: Figure S11**).
321 Fifth, there is no significant difference in the number of prophages among the clusters (**Figure 9E**). However,
322 the prophage types are consistent with the clusters: *Clostridium* phages mainly infect the *Clostridium* species
323 of cluster I, while *Faecalibacterium* phages target bacteria in cluster IV, which includes *Faecalibacteria* (**Add.**
324 **File 3: Figure S12**). Regarding infection targets, cluster XIVa is infected by the highest number of prophage
325 types, while cluster XIVb by the lowest. Interestingly, about two thirds of the detected prophage types infect
326 clostridial species of multiple rRNA clusters, with 19 even infecting species from all classes (**Figure 9F**). Hence,
327 while there is a clear shift in the prophages present, it remains unclear whether this is due to host association
328 or simply because of changes in the environmental niche. Besides, it should be noted that some prophages
329 carry passenger genes for methyltransferases, toxin(-antitoxin system)s and transporters (including beta-
330 lactam efflux pumps and other resistance proteins), or superinfection protection systems (**Add. File 2: Table**
331 **S4**). This highlights a potential mechanism by which clostridial species acquire new genes.
332 In summary, these genomic feature analyses support the hypothesis that cluster XIVa species tend to be in
333 earlier stages of a host adaptation process compared to their cluster IV relatives.

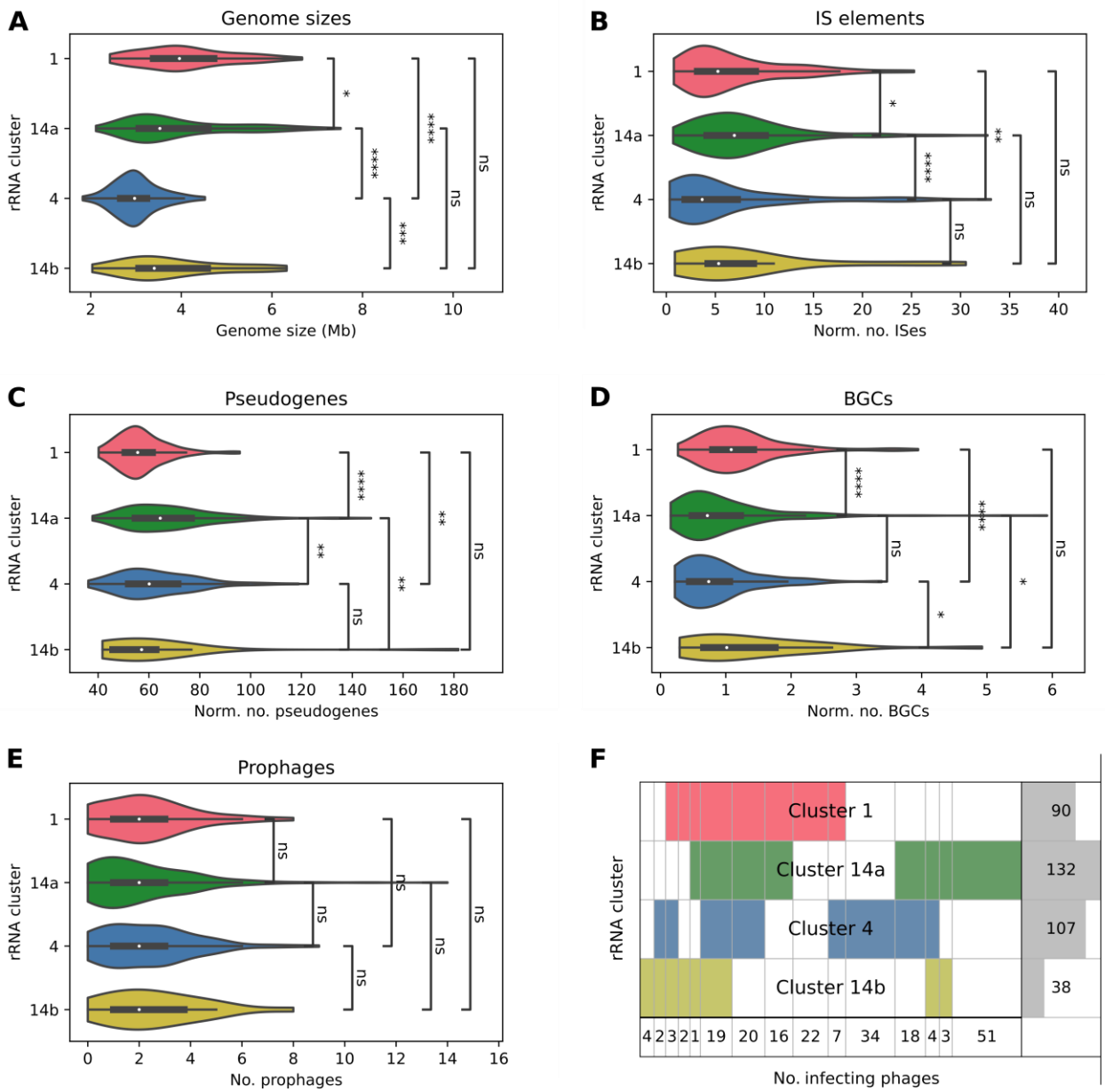

Figure 9: Comparative violinplots of each examined genomic property or feature in all rRNA-clustered genome subsets: (A) genome size, (B) insertion sequence elements, (C) pseudogenes, (D) biosynthetic gene clusters, (E) prophage regions, (F) number of infecting phages by rRNA cluster. Count normalisation was done by dividing by the genome length in Mb.

334 ***BGC clustering analysis reveals strongly conserved BGCs***

335 The species in our clostridial dataset have faced varying selection pressures, especially those in rRNA cluster
336 IV, which are in the later stages of a lifestyle conversion. Hence, any BGC that is conserved despite these
337 varying pressures, is very likely to provide potent properties that offer an evolutionary advantage large
338 enough to justify its conservation. Such high-potential BGCs should emerge in a BGC clustering analysis.

339 Our analysis revealed that the five largest BGC groups consist of ranthipeptide BGCs, two of which are carried
340 by cluster I species, one by cluster IV species, and another two by cluster XIVa species. Additionally, a few
341 cluster XIVb BGCs are linked to one of the XIVa ranthipeptide BGC groups (**Figure 10**). Alignment of the gene
342 clusters within each of these five largest ranthipeptide groups shows that the core biosynthetic genes, the
343 SCIFF precursor and the RRE gene, are conserved across all groups (**Add. File 3: Figures S13-S18**). The denser
344 groups have a higher number of conserved genes, although many remain unannotated. Moreover, these
345 ranthipeptide groups are well-distributed across the full phylogeny, with no species harbouring ranthipeptide
346 BGCs from multiple groups (**Add. File 3: Figure S19**). Interestingly, two of these ranthipeptide BGCs have both
347 been experimentally shown to participate in quorum sensing activities [49]. The fact that these two BGCs
348 belong to different ranthipeptide groups in Figure 10 (m.s. group 1 and 5) suggests that similar quorum
349 sensing functions may be associated with all these ranthipeptide BGCs.
350 The smaller BGC groups identified are CLA-NRPS and CLA-phosphonate hybrids, as well as RiPPs for cluster I.
351 Cluster XIVa has conserved CLA, CLA-NRPS hybrids, RRE-containing peptides and lanthipeptide class I BGCs.
352 When we expanded our dataset with the MIBiG database, a highly curated and standardised repository of

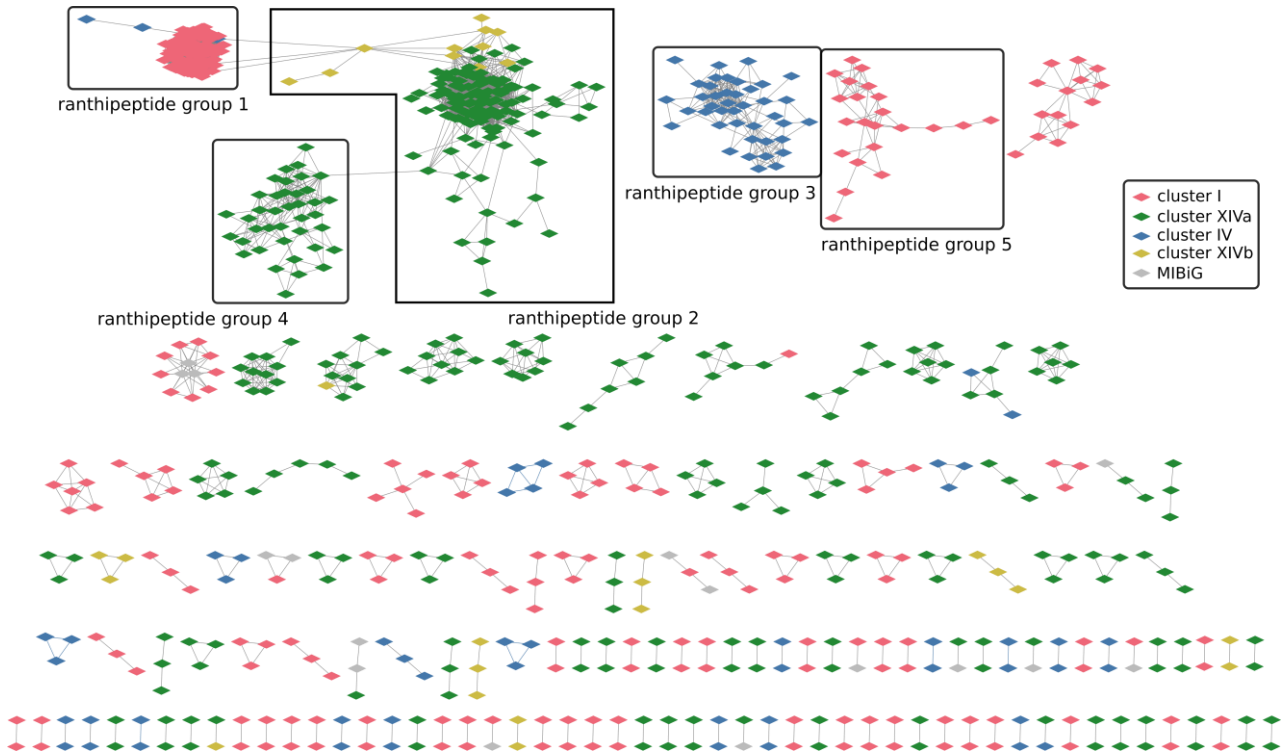

Figure 10: BiG-SCAPE similarity network of all BGCs detected in the full *Clostridia* genome set, annotated by rRNA cluster or as MIBiG entry. The five largest clusters on the top left are all ranthipeptides. Singleton entries have been omitted.

353 known BGCs [50], we found that very few MIBiG BGCs cocluster with these clostridial BGCs. This observation
354 underscores the novelty of these conserved ranthipeptides and their high potential for functional
355 significance. We also observed a large group of singleton BGCs (**Add. File 5**), which may comprise BGCs
356 acquired more recently by horizontal gene transfer. While these singletons have not undergone sufficient
357 evolutionary selection to confirm their functional significance, it is possible that some of them harbour potent
358 and biologically relevant activities.

359 Discussion

360 This work presents the first comparative genomics analysis of species from multiple rRNA species clusters of
361 the *Clostridia* class and is the first to report about the different yet intriguing host adaptation tendencies of
362 these rRNA clusters. Moreover, it uncovered a remarkably large repertoire of novel BGCs, including several
363 highly conserved ranthipeptides.

364 As anticipated, the core genome of this diverse class-level set of genomes is very small (about 0.5%).
365 Nevertheless, it includes an interesting number of highly conserved genes active in different core
366 metabolisms and gene expression systems. We have also shown the presence of continuous genomic
367 divergence dynamics over the entire phylogeny, as well as an open pangenome – a hallmark of species
368 capable of colonising diverse environments and exchanging genetic material through multiple mechanisms.
369 The large representation of various metabolic pathways in both the accessory and the unique pangenome
370 further highlights the extent of this ongoing genomic diversification in both free-living and host-associated
371 species.

372 Another major finding of this study is the varying levels of host adaptation observed in each rRNA cluster and
373 how it seemingly contradicts with their presumed lifestyle. Cluster I and cluster XIVa have similar genomic
374 divergence rates despite their different general lifestyle tendencies (free-living vs. host-associated). We
375 hypothesised that cluster XIVa is indeed host-dependent, but tends to be in the earlier stages of a host
376 adaptation process compared to the more specialized host-associated cluster IV. This hypothesis is
377 corroborated by the abundance of several genomic hallmarks associated with host adaptation. As cluster XIVa
378 appears to be less host-dependent, it likely retains part of its free-living gene repertoire. Indeed, the KEGG

pathway completeness assessment has shown that cluster XIVa has a higher fraction of metabolically versatile species. Hence, cluster XIVa species seem to have a higher innate capacity to produce more complex metabolites, which may have translated in a higher tendency to maintain a more extensive secondary metabolism. As a result, cluster XIVa species are better starting points for the development of anaerobic chassis organisms for heterologous expression of BGCs. Moreover, they may have a higher potential to serve as therapeutic delivery vehicles in gut environments [51].

A final key outcome of this study is the identification of BGCs, providing the first comparative view of the natural product repertoire of clostridial bacteria. Despite extensive host adaptation and a long track record of exposure to varying selection pressures, several ranthipeptide BGCs have been widely conserved. The lack of MIBiG entries in these clusters signals that these natural products are novel, tapping further into the unexplored richness of BGCs from the anaerobic realm. These novel conserved BGCs are prime targets for further experimental characterisation.

Nevertheless, there are some limitations that need to be considered for this study. First, since we conducted a high-level analysis averaging counts from rRNA clades, the conclusions presented here are to be considered general tendencies of each rRNA cluster and should not be extrapolated to individual species without further analysis. For example, it may be possible that a cluster XIVa species is already highly adapted to its host because it started the host adaptation process early compared to its close relatives. Second, we only included reference genomes in our dataset to ensure an equal species representation for the pangenome analysis. Non-reference genomes may harbour even more novel diversity, both at the pangenome level as well as at the BGC level.

Conclusions

Overall, the combined application of comparative genomics, phylogenetic techniques and narrow-focus genomic feature screens at higher taxonomic levels offers a flexible approach to investigate host adaptation trends. We uncovered counterintuitively different levels of host adaptation in clostridial species clusters that translate into different levels of metabolic versatility. Taken together, the findings of this study are an encouraging starter for the systematic investigation of host adaptation levels in host-associated bacteria

classes like *Clostridia*. Moreover, in the light of natural product research, our findings pave the way for (1) establishing clostridial species as new anaerobic host organisms for a diverse range of applications, as well as for (2) expanding the natural product repertoire further into the anaerobic realm.

Methods

Data gathering

Due to the absence of a clear complete listing of *Clostridia* species by rRNA cluster, the available cluster assignments were assembled from multiple sources. For clusters I and XIVb, some previous taxonomic reviews provided assignments for 58 and 5 species, respectively [2, 3, 5]. For clusters IV and XIVa, assignments were retrieved from EzBioCloud for the 28 and 34 included type species, respectively [52]. Following preliminary analyses, *C. gauvreauii* was found to be misannotated in EzBioCloud as a cluster IV species and reassigned to cluster XIVa. This resulted in a preliminary set of 125 species. To identify which species belong to which rRNA cluster, an unrooted phylogeny of the full *Clostridia* class was constructed with PhyloPhlAn using the default marker database (other settings as described in the dedicated section further below). For this, all 896 *Clostridia* reference genomes (as of 24/04/2024) were retrieved from RefSeq using the NCBI-Datasets CLI (v16.2.0) [53] with flags "--reference" and "taxon 186801". Then, each rRNA cluster of interest was defined as all taxa descending from the Most Recent Common Ancestor (MRCA) of all species in the preliminary set that were assigned to that rRNA cluster. Taxa assigned to another rRNA cluster were left out (**Add. File 3: Figure S1**). MRCA identification and clade extraction were done with the R package *ape* (v5.8) [54]. This resulted in a first set of 541 clostridial genomes (144 cluster I, 128 cluster IV, 244 cluster XIVa and 25 cluster XIVb).

Data selection and QC

To ensure a sufficient genome quality, all genomes in this first set were checked for completion and contamination levels using CheckM [55]. Only genomes with a completion level higher than 90% and a contamination level lower than 10% were retained. Additional QC measures were executed using the QC module of Panaroo (v1.4.2) [56]. This facilitates identifying genomes with an outlier number of genes or contigs, and genomes that do not cluster well using an MDS plot of pairwise Mash distances [57]. Using a

431 Mash reference database, this preprocessing script also allows to identify possible sources of contamination.
432 The final set comprised 492 clostridial genomes and is referenced to in this work as “the full clostridial set”.
433 For consistency, all selected genomes were reannotated with Bakta v1.9.3 [58].

434 ***ANI clustering analysis***

435 Similarities between the genomes in the full clostridial set were assessed by calculating the pairwise ANI
436 (average nucleotide identity) values using fastANI (v1.33) [59] with a minimum shared genome fraction of
437 5%. The unidirectional ANI values of both comparison senses were averaged into a symmetrical matrix of
438 bidirectional ANI values using pandas (v2.2.1). Missing values were filled by a default ANI value of 70%. Next,
439 an ANI clustermap was generated using scipy (v1.11.2) and seaborn (v0.12.2). Optimal leaf ordering was
440 enabled.

441 ***Pangenome construction and analysis***

442 Pangenomes were computed for the genomes of each rRNA cluster separately, as well as for the full genome
443 set. Clustering was performed using MMseqs2 [60] via the interface of PPanGGOLiN (v2.0.5) [61] to benefit
444 from its user-friendly annotation data integration. To account for the high diversity in this class-level genome
445 set, clustering was carried out with a relaxed sequence identity threshold of 50%, and an alignment coverage
446 cutoff of 50%. These thresholds were chosen because they yielded a comparable number of core genes as in
447 an earlier study of rRNA species cluster I [35]. Gene defragmentation by PPanGGOLiN was enabled. To allow
448 for genes lost during the pangenome construction, the pangenome partitions were assigned using the
449 commonly used relaxed core genome threshold of 99% and accessory genome threshold of 15%. Gene family
450 accumulation curves and UMAP plots (no. neighbours: 15; no. epochs: 200; Euclidian metric) were generated
451 using the R package *Panstripe* (v0.3.0) [62]. The gene family frequency U-shape plots were obtained using a
452 modified version of the *roary_plots.py* script of Marco Galardini, which is available as a supplementary script
453 for the *roary* pangenome tool [63].

454 ***Phylogeny construction***

455 A phylogenetic tree for the full clostridial set was constructed with PhyloPhlAn (v3.0.67) [64] with a custom
456 gene marker database constructed from the representative protein sequences of the earlier identified core

457 gene families. The tree was rooted using *B. subtilis* 168 as outgroup (RefSeq: GCF_000009045.1). The
458 phylogeny was computed at medium diversity setting using DIAMOND [65] as the mapping algorithm, MAFFT
459 [66] as the alignment algorithm, Trimal [67] as the alignment trimmer, and IQ-TREE2 [68] as the tree
460 constructor. 10.000 ultrafast bootstraps [69] were requested by manually adding the "-bb 10000" flag to the
461 IQ-TREE entry in the PhyloPhlAn configuration file. The outgroup was trimmed from the rooted phylogeny
462 using *ape*, after which the tree was visualised with iTol (v6.9) [70] and the R package *ggtree* (v3.11.1) [71].

463 ***Functional annotation of the pangenome***

464 All protein family reference sequences of each pangenome were functionally annotated using both COG
465 categories and hierarchical KEGG BRITE categories. COG assignments were obtained by running a local
466 instance of eggNOG-mapper (v2.1.12) [72] with settings mirroring the webserver's defaults. KEGG BRITE
467 annotations were carried out by the KAAS webserver [73] using the BBH method in GHOSTX mode and the
468 prokaryotic representative species set. Category counts and fractions were calculated using pandas and
469 plotted using ggplot2 (v3.5.1). KEGG module completeness was assessed using the MicrobeAnnotator (v2.0.5)
470 pipeline [74] in light mode at default settings and in-house processing and plotting scripts. Optimal leaf
471 ordering was enabled for plotting the clustermap.

472 ***Genomic divergence rate analysis***

473 A genomic divergence rate analysis estimates the rate of gene exchange within a clade and statistically
474 compares these rates for significant differences. It requires the phylogenies of each examined clade to be at
475 exactly the same time scale. Therefore, separate subtrees by rRNA cluster were extracted from the earlier
476 computed full phylogeny using the *ape* R package. Gene exchange (GE) models were inferred for each
477 separate rRNA cluster separately using the R package *Panstripe*, each time leveraging the respective
478 pangenome presence/absence matrix and phylogenetic subtree. All options were left at their defaults, except
479 that Gaussian distributions were used for fitting the underlying generalised linear models (GLMs). Cumulative
480 GE plots and a statistical comparison of the GE models were obtained using appropriate *Panstripe* functions.

481 ***Ancestral state reconstruction***

482 To track the evolutionary origin of gene exchanges, the ancestral states of the pangenome were reconstructed
483 from the full core genome phylogeny, and the presence/absence data of the full pangenome. This gene
484 gain/loss (GGL) model was inferred according to Wagner parsimony with equal penalty scores for both gene
485 gains and losses using the Count tool (v10.4) [75]. The annotation of genes exchanged at tree nodes of interest
486 were retrieved and visualised using in-house R scripts leveraging the earlier made eggNOG and KEGG BRITE
487 annotations. Cluster XIVb was left out of this analysis as it does not have a single MRCA.

488 ***Mobilome analysis***

489 Pseudogenes were detected using Pseudofinder (v1.1.0) [76] at default settings. The NCBI Protein database
490 sliced for taxon ID 186801 (the *Clostridia* class) was used as reference database and indexed using DIAMOND
491 (v2.0.14). Insertion elements (ISEs) were inferred using ISEscan (v1.7.2.3) [77] at default settings. Biosynthetic
492 gene clusters (BGCs) were identified using a local instance of antiSMASH (v7.0.0) [26] with settings mirroring
493 the webserver's defaults and Prodigal as the gene finding tool. Prophages were found by running PHASTEST
494 [78] locally in lite mode using the multi-container Docker image from Docker Hub (digest 2477a2f3b6b8).
495 Infectivity graphs were made using *supervenn* (v0.5.0). All mobile element counts were normalised by dividing
496 by the genome length and statistically compared using the Mann-Whitney U-test ($\alpha = 0.05$) in scipy.
497 Significance annotations were added using *statannotations* (v0.6).

498 ***BGC clustering analysis***

499 All BGCs identified during the mobilome analysis were clustered using BiG-SCAPE (v1.1.8) [79], adding MIBiG
500 reference BGCs (flag '--mibig'), mixing all BGC classes (flag '--mix'), and including singletons (flag '--
501 include_singletons'). BiG-SCAPE uses a custom distance metric combining protein domain content, order,
502 copy number and sequence identity. A BiG-SCAPE distance cutoff value of 0.5 was applied as appeared
503 appropriate from parameter benchmarking (see also **Add. File 3: Figure S20**) and as it also clustered well on
504 rRNA species cluster level. All other settings were left at their defaults. To integrate the rRNA cluster metadata
505 in CytoScape (v3.10.2), the unidirectional network links in the output network files of BiG-SCAPE were
506 annotated by rRNA cluster or MIBiG and converted into bidirectional links by adding the reverse link. BGC

gene alignments were generated for a representative selection of BGCs of each ranthipeptide group using clinker (v0.0.28) [80].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data is available as supplementary materials in Additional Files 1 to 5. Additional File 1 contains the accessions of all genomes used in this study. All code and conda environments of the analysis workflows have been deposited at GitHub (https://github.com/LucoDevro/closHomer) and at Zenodo (https://doi.org/10.5281/zenodo.14719962).

Competing interests

We declare no competing interests.

Funding

This work was funded by the European Union (ERC, MiStiC, 101078461).

Author's contributions

LDV and JM conceived this study. LDV and JM designed the research. LDV performed the data analysis, implemented the workflow, and gathered the results. LDV analysed and interpreted the data, with assistance from JA and JM. LDV, JA and JM wrote and revised the article.

Acknowledgements

Not applicable

References

[1] A. Prazmowski, 'Untersuchungen über die entwickelungsgeschichte und fermentwirkung einiger bacterien-arten', Leipzig, 1880.

[2] P. A. Lawson and F. A. Rainey, 'Proposal to restrict the genus Clostridium prazmowski to Clostridium butyricum and related species', *Int J Syst Evol Microbiol*, vol. 66, no. 2, pp. 1009–1016, 2016, doi: 10.1099/ijsem.0.000824.

[3] M. D. Collins *et al.*, 'The Phylogeny of the Genus Clostridium: Proposal of Five New Genera and Eleven New Species Combinations', *Int J Syst Bacteriol*, vol. 44, no. 4, pp. 812–826, Oct. 1994, doi: 10.1099/00207713-44-4-812.

[4] E. Stackebrandt, I. Kramer, J. Swiderski, and H. Hippe, ' Phylogenetic basis for a taxonomic dissection of the genus Clostridium ', *FEMS Immunol Med Microbiol*, vol. 24, no. 3, pp. 253–258, Jul. 1999, doi: 10.1111/j.1574-695x.1999.tb01291.x.

[5] J. Wiegel, R. S. Tanner, and F. A. Rainey, 'An introduction to the family Clostridiaceae', 3rd ed., M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, Eds., New York, NY: Springer US, 2006. doi: 10.1007/0-387-30744-3.

[6] C. Liu, S. M. Finegold, Y. Song, and P. A. Lawson, 'Reclassification of Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus and Ruminococcus schinkii as Blautia coccoides gen. nov., comb. nov., Blautia hansenii comb. nov., Blautia hydrogenotrophica comb. nov., Blautia luti comb. nov., Blautia producta comb. nov., Blautia schinkii comb. nov. and description of Blautia wexlerae ...', *Int J Syst Evol Microbiol*, vol. 58, no. 8, pp. 1896–1902, Aug. 2008, doi: 10.1099/ijs.0.65208-0.

[7] S. Kaur, M. Yawar, P. A. Kumar, and K. Suresh, 'Hungatella effluvii gen. nov., sp. nov., an obligately anaerobic bacterium isolated from an effluent treatment plant, and reclassification of Clostridium hathewayi as Hungatella hathewayi gen. nov., comb. nov', *Int J Syst Evol Microbiol*, vol. 64, no. PART 3, pp. 710–718, 2014, doi: 10.1099/ijs.0.056986-0.

[8] R. S. Gupta and B. Gao, 'Phylogenomic analyses of clostridia and identification of novel protein signatures that are specific to the genus Clostridium sensu stricto (cluster I)', *Int J Syst Evol Microbiol*, vol. 59, no. 2, pp. 285–294, 2009, doi: 10.1099/ijs.0.001792-0.

[9] C. Birgen, K. F. Degnes, S. Markussen, A. Wentzel, and H. Sletta, 'Butanol production from lignocellulosic sugars by Clostridium beijerinckii in microbioreactors', *Biotechnol Biofuels*, vol. 14, no. 1, Dec. 2021, doi: 10.1186/s13068-021-01886-1.

[10] S. B. Roberts, C. M. Gowen, J. P. Brooks, and S. S. Fong, 'Genome-scale metabolic analysis of Clostridium thermocellum for bioethanol production', *BMC Syst Biol*, vol. 4, no. 1, p. 31, Dec. 2010, doi: 10.1186/1752-0509-4-31.

[11] W. L. Schroeder, T. Kuil, A. J. A. van Maris, D. G. Olson, L. R. Lynd, and C. D. Maranas, 'A detailed genome-scale metabolic model of Clostridium thermocellum investigates sources of pyrophosphate for driving glycolysis', *Metab Eng*, vol. 77, pp. 306–322, May 2023, doi: 10.1016/j.ymben.2023.04.003.

[12] Y. Song *et al.*, 'Functional cooperation of the glycine synthase-reductase and Wood–Ljungdahl pathways for autotrophic growth of *Clostridium drakei*', *Proceedings of the National Academy of Sciences*, vol. 117, no. 13, pp. 7516–7523, Mar. 2020, doi: 10.1073/pnas.1912289117.

[13] C. B. Milne *et al.*, 'Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052', *BMC Syst Biol*, vol. 5, Aug. 2011, doi: 10.1186/1752-0509-5-130.

[14] M. L. Dahle, E. T. Papoutsakis, and M. R. Antoniewicz, '13C-metabolic flux analysis of Clostridium ljungdahlii illuminates its core metabolism under mixotrophic culture conditions', *Metab Eng*, vol. 72, pp. 161–170, Jul. 2022, doi: 10.1016/j.ymben.2022.03.011.

[15] P. Patakova *et al.*, 'Acidogenesis, solventogenesis, metabolic stress response and life cycle changes in Clostridium beijerinckii NRRL B-598 at the transcriptomic level', *Sci Rep*, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-018-37679-0.

[16] R. de Souza Pinto Lemgruber *et al.*, 'Systems-level engineering and characterisation of Clostridium autoethanogenum through heterologous production of poly-3-hydroxybutyrate (PHB)', *Metab Eng*, vol. 53, pp. 14–23, May 2019, doi: 10.1016/j.ymben.2019.01.003.

[17] J. S. Li *et al.*, 'Discovery and Biosynthesis of Clostyrylpyrones from the Obligate Anaerobe Clostridium roseum', *Org Lett*, vol. 22, no. 21, pp. 8204–8209, Nov. 2020, doi: 10.1021/acs.orglett.0c02656.

[18] A. S. N. W. Pahalagedara, S. Flint, J. Palmer, G. Brightwell, and T. B. Gupta, 'Antimicrobial production by strictly anaerobic Clostridium spp.', *Int J Antimicrob Agents*, vol. 55, no. 5, May 2020, doi: 10.1016/j.ijantimicag.2020.105910.

[19] S. Pidot, K. Ishida, M. Cyrulies, and C. Hertweck, 'Discovery of clostrubin, an exceptional polyphenolic polyketide antibiotic from a strictly anaerobic bacterium', *Angewandte Chemie - International Edition*, vol. 53, no. 30, pp. 7856–7859, Jul. 2014, doi: 10.1002/anie.201402632.

[20] A. C. Letzel, S. J. Pidot, and C. Hertweck, 'A genomic approach to the cryptic secondary metabolome of the anaerobic world', *Nat Prod Rep*, vol. 30, no. 3, pp. 392–428, Mar. 2013, doi: 10.1039/c2np20103h.

[21] S. Behnken and C. Hertweck, 'Anaerobic bacteria as producers of antibiotics', *Appl Microbiol Biotechnol*, vol. 96, no. 1, pp. 61–67, Oct. 2012, doi: 10.1007/s00253-012-4285-8.

[22] T. Lincke, S. Behnken, K. Ishida, M. Roth, and C. Hertweck, ' Closthioamide: An Unprecedented Polythioamide Antibiotic from the Strictly Anaerobic Bacterium Clostridium cellulolyticum ', *Angewandte Chemie*, vol. 122, no. 11, pp. 2055–2057, Mar. 2010, doi: 10.1002/ange.200906114.

[23] Y. Sugimoto *et al.*, 'A metagenomic strategy for harnessing the chemical repertoire of the human microbiome', *Science (1979)*, vol. 366, no. 6471, Dec. 2019, doi: 10.1126/science.aax9176.

[24] G. Shabuer, K. Ishida, S. J. Pidot, M. Roth, H.-M. Dahse, and C. Hertweck, 'Plant pathogenic anaerobic bacteria use aromatic polyketides to access aerobic territory', *Science (1979)*, vol. 350, no. 6261, pp. 670–674, Nov. 2015, doi: 10.1126/science.aac9990.

[25] K. Scherlach and C. Hertweck, 'Mining and unearthing hidden biosynthetic potential', *Nat Commun*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/s41467-021-24133-5.

[26] K. Blin *et al.*, 'AntiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation', *Nucleic Acids Res*, vol. 51, no. W1, pp. W46–W50, Jul. 2023, doi: 10.1093/nar/gkad344.

[27] D. M. Aronoff, 'Clostridium novyi, sordellii, and tetani: Mechanisms of disease', *Anaerobe*, vol. 24, pp. 98–101, Dec. 2013, doi: 10.1016/j.anaerobe.2013.08.009.

[28] J. E. Cohen, R. Wang, R.-F. Shen, W. W. Wu, and J. E. Keller, 'Comparative pathogenomics of Clostridium tetani', *PLoS One*, vol. 12, no. 8, p. e0182909, Aug. 2017, doi: 10.1371/journal.pone.0182909.

[29] J. A. Lacey *et al.*, 'Whole genome analysis reveals the diversity and evolutionary relationships between necrotic enteritis-causing strains of Clostridium perfringens', *BMC Genomics*, vol. 19, no. 1, p. 379, Dec. 2018, doi: 10.1186/s12864-018-4771-1.

[30] T. J. Smith, K. K. Hill, and B. H. Raphael, 'Historical and current perspectives on Clostridium botulinum diversity', *Res Microbiol*, vol. 166, no. 4, pp. 290–302, May 2015, doi: 10.1016/j.resmic.2014.09.007.

[31] T. Grenda, A. Grenda, P. Domaradzki, P. Krawczyk, and K. Kwiatek, 'Probiotic Potential of Clostridium spp.—Advantages and Doubts', *Curr Issues Mol Biol*, vol. 44, no. 7, pp. 3118–3130, 2022, doi: 10.3390/cimb44070215.

[32] L. R. Lopetuso, F. Scaldaferri, V. Petito, and A. Gasbarrini, 'Commensal Clostridia: leading players in the maintenance of gut homeostasis', *Gut Pathog*, vol. 5, no. 1, p. 23, 2013, doi: 10.1186/1757-4749-5-23.

[33] G. L. Hold, S. E. Pryde, V. J. Russell, E. Furrie, and H. J. Flint, 'Assessment of microbial diversity in human colonic samples by 16S rDNA sequence analysis', *FEMS Microbiol Ecol*, vol. 39, no. 1, pp. 33–39, Jan. 2002, doi: 10.1111/j.1574-6941.2002.tb00904.x.

[34] A. E. Livanos *et al.*, 'Rapid gastrointestinal loss of Clostridial Clusters IV and XIVa in the ICU associates with an expansion of gut pathogens', *PLoS One*, vol. 13, no. 8, p. e0200322, Aug. 2018, doi: 10.1371/journal.pone.0200322.

[35] Z. Udaondo, E. Duque, and J. L. Ramos, 'The pangenome of the genus Clostridium', *Environ Microbiol*, vol. 19, no. 7, pp. 2588–2603, Jul. 2017, doi: 10.1111/1462-2920.13732.

[36] J. L. Maturana and J. P. Cárdenas, 'Insights on the Evolutionary Genomics of the Blautia Genus: Potential New Species and Genetic Content Among Lineages', *Front Microbiol*, vol. 12, Apr. 2021, doi: 10.3389/fmicb.2021.660920.

[37] N. A. Moran, 'Minireview Microbial Minimalism: Genome Reduction in Bacterial Pathogens', *Cell*, vol. 108, pp. 583–586, 2002.

[38] L. Klasson, 'Evolution of minimal-gene-sets in host-dependent bacteria', *Trends Microbiol*, vol. 12, no. 1, pp. 37–43, Jan. 2004, doi: 10.1016/j.tim.2003.11.006.

[39] J. P. McCutcheon and N. A. Moran, 'Extreme genome reduction in symbiotic bacteria', *Nat Rev Microbiol*, vol. 10, no. 1, pp. 13–26, Jan. 2012, doi: 10.1038/nrmicro2670.
[40] V. Pascal Andreu *et al.*, 'gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota', *Nat Biotechnol*, vol. 41, no. 10, pp. 1416–1423, Oct. 2023, doi: 10.1038/s41587-023-01675-1.
[41] S. Vieira-Silva *et al.*, 'Species-function relationships shape ecological properties of the human gut microbiome', *Nat Microbiol*, vol. 1, no. 8, Jun. 2016, doi: 10.1038/nmicrobiol.2016.88.
[42] A. Dufresne, L. Garczarek, and F. Partensky, 'Accelerated evolution associated with genome reduction in a free-living prokaryote', *Genome Biol*, vol. 6, no. 2, p. R14, 2005, doi: 10.1186/gb-2005-6-2-r14.
[43] G. A. B. Marais, A. Calteau, and O. Tenaillon, 'Mutation rate and genome reduction in endosymbiotic and free-living bacteria', *Genetica*, vol. 134, no. 2, pp. 205–210, 2008, doi: 10.1007/s10709-007-9226-6.
[44] N. A. Moran, J. P. McCutcheon, and A. Nakabachi, 'Genomics and Evolution of Heritable Bacterial Symbionts', *Annu Rev Genet*, vol. 42, no. 1, pp. 165–190, Dec. 2008, doi: 10.1146/annurev.genet.41.110306.130119.
[45] P. Siguier, E. Gourbeyre, and M. Chandler, 'Bacterial insertion sequences: Their genomic impact and diversity', *FEMS Microbiol Rev*, vol. 38, no. 5, pp. 865–891, Sep. 2014, doi: 10.1111/1574-6976.12067.
[46] G. A. B. Marais, A. Calteau, and O. Tenaillon, 'Mutation rate and genome reduction in endosymbiotic and free-living bacteria', *Genetica*, vol. 134, no. 2, pp. 205–210, Oct. 2008, doi: 10.1007/s10709-007-9226-6.
[47] P. Siguier, E. Gourbeyre, A. Varani, B. Ton-Hoang, and M. Chandler, 'Everyman's Guide to Bacterial Insertion Sequences', *Microbiol Spectr*, vol. 3, no. 2, Apr. 2015, doi: 10.1128/microbiolspec.mdna3-0030-2014.
[48] J. Parkhill *et al.*, 'Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica', *Nat Genet*, vol. 35, no. 1, pp. 32–40, Sep. 2003, doi: 10.1038/ng1227.
[49] Y. Chen *et al.*, 'The SCIFF-Derived Ranthipeptides Participate in Quorum Sensing in Solventogenic Clostridia', *Biotechnol J*, vol. 15, no. 10, Oct. 2020, doi: 10.1002/biot.202000136.
[50] B. R. Terlouw *et al.*, 'MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters', *Nucleic Acids Res*, vol. 51, no. D1, pp. D603–D610, Jan. 2023, doi: 10.1093/nar/gkac1049.
[51] A. M. Kubiak and N. P. Minton, 'The potential of clostridial spores as therapeutic delivery vehicles in tumour therapy', *Res Microbiol*, vol. 166, no. 4, pp. 244–254, May 2015, doi: 10.1016/j.resmic.2014.12.006.
[52] J. Jongsik Chun, ' Taxonomy of Clostridium clusters XIVa and IV'. Accessed: Nov. 25, 2024. [Online]. Available: https://help.ezbiocloud.net/taxonomy-of-clostridium-cluster-xiva-iv/
[53] N. A. O'Leary *et al.*, 'Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets', *Sci Data*, vol. 11, no. 1, p. 732, Jul. 2024, doi: 10.1038/s41597-024-03571-y.
[54] E. Paradis and K. Schliep, 'ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R', *Bioinformatics*, vol. 35, no. 3, pp. 526–528, Feb. 2019, doi: 10.1093/bioinformatics/bty633.
[55] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Res*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.
[56] G. Tonkin-Hill *et al.*, 'Producing polished prokaryotic pangenomes with the Panaroo pipeline', *Genome Biol*, vol. 21, no. 1, Jul. 2020, doi: 10.1186/s13059-020-02090-4.
[57] B. D. Ondov *et al.*, 'Mash: fast genome and metagenome distance estimation using MinHash', *Genome Biol*, vol. 17, no. 1, p. 132, Dec. 2016, doi: 10.1186/s13059-016-0997-x.
[58] O. Schwengers, L. Jelonek, M. A. Dieckmann, S. Beyvers, J. Blom, and A. Goesmann, 'Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification', *Microb Genom*, vol. 7, no. 11, Nov. 2021, doi: 10.1099/mgen.0.000685.

[59] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nat Commun*, vol. 9, no. 1, p. 5114, Nov. 2018, doi: 10.1038/s41467-018-07641-9.

[60] M. Steinegger and J. Söding, 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nat Biotechnol*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, doi: 10.1038/nbt.3988.

[61] G. Gautreau *et al.*, 'PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph', *PLoS Comput Biol*, vol. 16, no. 3, p. e1007732, Mar. 2020, doi: 10.1371/journal.pcbi.1007732.

[62] G. Tonkin-Hill, R. A. Gladstone, A. K. Pöntinen, S. Arredondo-Alonso, S. D. Bentley, and J. Corander, 'Robust analysis of prokaryotic pangenome gene gain and loss rates with Panstripe', *Genome Res*, vol. 33, no. 1, pp. 129–140, Jan. 2023, doi: 10.1101/gr.277340.122.

[63] A. J. Page *et al.*, 'Roary: Rapid large-scale prokaryote pan genome analysis', *Bioinformatics*, vol. 31, no. 22, pp. 3691–3693, May 2015, doi: 10.1093/bioinformatics/btv421.

[64] F. Asnicar *et al.*, 'Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0', *Nat Commun*, vol. 11, no. 1, p. 2500, May 2020, doi: 10.1038/s41467-020-16366-7.

[65] B. Buchfink, K. Reuter, and H.-G. Drost, 'Sensitive protein alignments at tree-of-life scale using DIAMOND', *Nat Methods*, vol. 18, no. 4, pp. 366–368, Apr. 2021, doi: 10.1038/s41592-021-01101-x.

[66] K. Katoh and D. M. Standley, 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Mol Biol Evol*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.

[67] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics*, vol. 25, no. 15, pp. 1972–1973, Aug. 2009, doi: 10.1093/bioinformatics/btp348.

[68] B. Q. Minh *et al.*, 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Mol Biol Evol*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi: 10.1093/molbev/msaa015.

[69] B. Q. Minh, M. A. T. Nguyen, and A. von Haeseler, 'Ultrafast Approximation for Phylogenetic Bootstrap', *Mol Biol Evol*, vol. 30, no. 5, pp. 1188–1195, May 2013, doi: 10.1093/molbev/mst024.

[70] I. Letunic and P. Bork, 'Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool', *Nucleic Acids Res*, vol. 52, no. W1, pp. W78–W82, Jul. 2024, doi: 10.1093/nar/gkae268.

[71] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T. Lam, 'ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data', *Methods Ecol Evol*, vol. 8, no. 1, pp. 28–36, Jan. 2017, doi: 10.1111/2041-210X.12628.

[72] C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas, 'eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale', *Mol Biol Evol*, vol. 38, no. 12, pp. 5825–5829, Dec. 2021, doi: 10.1093/molbev/msab293.

[73] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, 'KAAS: an automatic genome annotation and pathway reconstruction server', *Nucleic Acids Res*, vol. 35, no. Web Server, pp. W182–W185, May 2007, doi: 10.1093/nar/gkm321.

[74] C. A. Ruiz-Perez, R. E. Conrad, and K. T. Konstantinidis, 'MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes', *BMC Bioinformatics*, vol. 22, no. 1, p. 11, Dec. 2021, doi: 10.1186/s12859-020-03940-5.

[75] M. Csurös, 'Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood', *Bioinformatics*, vol. 26, no. 15, pp. 1910–1912, Jun. 2010, doi: 10.1093/bioinformatics/btq315.

[76] M. J. Syberg-Olsen, A. I. Garber, P. J. Keeling, J. P. McCutcheon, and F. Husnik, 'Pseudofinder: Detection of Pseudogenes in Prokaryotic Genomes', *Mol Biol Evol*, vol. 39, no. 7, Jul. 2022, doi: 10.1093/molbev/msac153.

[77] Z. Xie and H. Tang, 'ISEScan: automated identification of insertion sequence elements in prokaryotic genomes', *Bioinformatics*, vol. 33, no. 21, pp. 3340–3347, Nov. 2017, doi: 10.1093/bioinformatics/btx433.

[78] D. S. Wishart *et al.*, 'PHASTEST: faster than PHASTER, better than PHAST', *Nucleic Acids Res*, vol. 51, no. W1, pp. W443–W450, Jul. 2023, doi: 10.1093/nar/gkad382.
[79] J. C. Navarro-Muñoz *et al.*, 'A computational framework to explore large-scale biosynthetic diversity', *Nat Chem Biol*, vol. 16, no. 1, pp. 60–68, Jan. 2020, doi: 10.1038/s41589-019-0400-9.
[80] C. L. M. Gilchrist and Y. H. Chooi, 'Clinker & clustermap.js: Automatic generation of gene cluster comparison figures', *Bioinformatics*, vol. 37, no. 16, pp. 2473–2475, Aug. 2021, doi: 10.1093/bioinformatics/btab007.

Additional files

additional_file_1.xlsx

Additional file 1. Metadata tables of the genomes used. List of 541 genomes used at the start of the study with information on the species, NCBI assembly accession, clostridial rRNA cluster, assembly level, assembly size and number of genes, checkM metrics, and QC failure statuses

additional_file_2.docx

Additional file 2. Supplementary Tables S1-S4. Includes both supplementary tables and the legends.

additional_file_3.docx

Additional file 3. Supplementary Figures S1-S20. Includes both supplementary figures and the legends.

additional_file_4.xlsx

Additional file 4. Tables containing absolute and relative frequences of KEGG BRITE level C annotations for the gained and lost gene families of the LCA and the MRCA of each rRNA cluster, if any.

additional_file_5.zip

Additional file 5. Cytoscape session file of the BiG-SCAPE similarity network for all BGCs detected by antiSMASH in the clostridial genome dataset. The network includes MIBiG entries. Edges are included based on BiG-SCAPE similarity. Nodes are coloured by rRNA species cluster membership.