Comparative pangenomics unveils distinct host adaptation levels and conserved biosynthetic potential in microbiome *Clostridia*

Running title: Pangenomics of host adaptation in *Clostridia*

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

Bacterial species adopt various lifestyles to thrive in diverse ecological niches, ranging from living freely in the soil to being part of various human and animal microflora. However, the impact of these adaptation processes on their genomes and metabolisms remains largely unexplored beyond the genus level. Investigating these evolutionary dynamics for higher taxonomic levels can enhance our understanding of the relationship between host adaptation and their functional capabilities. In this regard, we examined the evolutionary trajectories and metabolic capabilities of the *Clostridia* class, which display a variety of lifestyles and are of high importance for industry, medicine and microbiome research. First, we uncover that the clostridial orders have significantly different adaptation rates. Second, we show that the *Oscillospirales* order has undergone extensive genomic and functional specialisation toward a host-associated lifestyle, while the *Lachnospirales* order tends to be at a lower level of host association, retaining a remarkably high number of free-living trait genes and a high degree of metabolic versatility. Third, we reveal substantial differences in genomic architecture and metabolic versatility between the clostridial orders and link these to the progressing stages of host adaptation. Additionally, we identify widely conserved biosynthetic gene clusters, highlighting untapped biosynthetic potential of evolutionary significance. Hence, the beyond-genus level analyses in this study provide valuable new insights with implications for the biological sciences and biotechnology.

Keywords

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

Introduction

Microorganisms constantly interact with their environment, adapting not only through short-term regulatory and metabolic responses, but also through long-term genomic changes. The continuous acquisition, mutation and deletion of genes is closely tied to environmental selection pressures, giving rise to various lifestyles that vary in their dependence on external factors. These selection pressures drive evolutionary processes, such as host adaptation and genome streamlining, which can comprehensively reshape microbial genomes. This, in turn, has a profound impact on their functional and metabolic versatility, and hence their ability to survive in different ecological niches (Dufresne, Garczarek and Partensky 2005, Siguier, Gourbeyre and Chandler 2014, Marais, Calteau and Tenaillon 2008).

Species in the *Clostridia* class exhibit 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, exemplified by industrially relevant species like *C. acetobutylicum* and *C. beijerinckii* (Birgen et al. 2021, Patakova et al. 2019, Roberts et al. 2010) and notorious pathogens like *C. botulinum*, *C. tetani* and *C. difficile* (Cohen et al. 2017, Smith, Hill and Raphael 2015, Sandhu and McBride 2018). Moreover, *Clostridium* species were the first anaerobes discovered to produce biosynthetic gene cluster (BGC)-encoded secondary metabolites with potent biological activities (Pahalagedara et al. 2020, Letzel, Pidot and Hertweck 2013, Behnken and Hertweck 2012, Sugimoto et al. 2019).

Despite the abundance and relevance of clostridial species, only few comparative studies have explored their evolutionary history and ecological adaptations beyond the species level. Previous higher-level comparative studies focused at most on a few individual species or a single genus. For example, a study of *Clostridium sensu stricto* found evidence of a highly plastic genome, harbouring a relatively large metabolic core and many specialised genetic capabilities that underpin the mainly free-living nature of this genus (Udaondo, Duque and Ramos 2017). Another study mainly focusing on *Clostridium sensu stricto* identified divergent evolution of the shikimate pathway in pathogenic clostridia and extensive gene mobility among their toxin genes (Cruz-Morales et al. 2019). A last study focusing on the gut-associated genus *Blautia* revealed

that many gene exchange events are associated with metabolic functions, likely driven by the dynamically changing host environment. Moreover, horizontal gene transfer may have caused a metabolic dependence on other gut bacteria (Maturana and Cárdenas 2021).

In this work, we present a comprehensive study of the interplay between host adaptation and genomic architecture for the taxonomic class of *Clostridia*. By combining pangenomic and phylogenetic analyses, we characterised the evolutionary dynamics within different orders and inferred lifestyles for each order's ancestors. In parallel, we assessed the metabolic versatility of each species and linked this with their evolutionary history. We also screened all genomes for genomic hallmarks and mobile genetic elements (MGEs) that are related with progressing host adaptation. Finally, we zoomed in on secondary metabolite BGCs, assessing their conservation and evolutionary patterns.

Materials and methods

Data gathering, QC and filtering

All genome assemblies were retrieved from RefSeq using the search query "Clostridia" in the NCBI Datasets CLI (v16.2.0) (O'Leary et al. 2024), excluding MAGs and atypical assemblies (# = 10599 as of 22/05/2025). The resulting set was filtered by number of contigs (<= 200), N50 (>= 50 kb), completeness (>= 90%) and contamination (<= 10%), and dereplicated at species level using skDER (v1.3.1) (Salamzade and Kalan 2023) (ANI cutoff: 96%, AF cutoff: > 50%; **Add. File 2: Figure S1**), resulting in 1010 representative assemblies. For consistency, all assemblies were reannotated with Bakta (v1.9.1) (Schwengers et al. 2021) and taxonomically verified using GTDB-tk (v2.4.1) (Chaumeil et al. 2022). To ensure sufficient statistical power in downstream analyses, only taxonomic orders with at least 30 representative species were retained.

Metabolic versatility analysis

The metabolic versatility of all taxa was assessed by KEGG module completeness as determined by the MicrobeAnnotator (v2.0.5) pipeline (Ruiz-Perez, Conrad and Konstantinidis 2021) in light mode at default settings. The module completenesses were clustered using Ward's hierarchical clustering and plotted using in-house Python scripts. Optimal leaf ordering was enabled while plotting the clustermap.

90 ***Pangenome construction and analysis***

91 A pangenome was constructed for the full genome set using PPanGGOLiN (v2.0.5) (Gautreau et al. 2020) with
92 gene defragmentation enabled. To account for the high diversity in this class-level genome set, clustering was
93 carried out with a relaxed sequence identity threshold of 40% and an alignment coverage cutoff of 40%, while
94 the pangenome was partitioned using a core genome threshold of 90% and an accessory genome threshold
95 of 0.099% (or 1/1010), as appeared appropriate from our parameter benchmark (**Add. File 3**).
96 Presence/absence matrices for the different clostridial orders were extracted from the full class-level matrix
97 by selecting for gene families present in at least one taxon from that order. The same pangenome partitioning
98 thresholds were applied for consistency.

99 ***Phylogeny construction***

100 A phylogenetic tree for the full clostridial genome set was constructed with PhyloPhlAn (v3.0.67) (Asnicar et
101 al. 2020) with a custom gene marker database derived from the representative protein sequences of the
102 earlier identified core gene families. The tree was rooted using *B. subtilis* 168 as outgroup (RefSeq:
103 GCF_000009045.1). The phylogeny was computed at medium diversity setting using DIAMOND (Buchfink,
104 Reuter and Drost 2021) as the mapping algorithm, MAFFT (Katoh and Standley 2013) as the alignment
105 algorithm, Trimal (Capella-Gutiérrez, Silla-Martínez and Gabaldón 2009) as the alignment trimmer, and IQ-
106 TREE2 (Minh et al. 2020) as the tree constructor. 10.000 ultrafast bootstraps (Minh, Nguyen and von Haeseler
107 2013) were requested by manually adding the "-bb 10000" flag to the IQ-TREE entry in the PhyloPhlAn
108 configuration file. The outgroup was trimmed from the rooted phylogeny using the R package *ape*, after which
109 the tree was visualised with iTol (v6.9) (Letunic and Bork 2024) and ggtree (v3.11.1) (Yu et al. 2017).

110 ***Functional annotation of the pangenome***

111 The protein family reference sequences of the pangenome were functionally annotated using eggNOG-
112 mapper (v2.1.12) (Cantalapiedra et al. 2021). COG category counts and fractions were calculated using pandas
113 (v2.2.1) and plotted using ggplot2 (v3.5.1).

114 ***Genomic divergence rate analysis***

115 A genomic divergence rate analysis estimates the rate of gene exchange within a clade and statistically
116 compares these rates for significant differences. It requires the phylogenies of each examined clade to have
117 an identical time scale. Therefore, a subtree was extracted for each order from the full phylogeny using *ape*.
118 Gene exchange (GE) models were inferred for each order using *Panstripe* (Tonkin-Hill et al. 2023). All options
119 were left at their defaults, except that Gaussian distributions were employed to fit the underlying generalised
120 linear models (GLMs). Cumulative GE plots and statistical comparisons of the GE models were carried out
121 using the appropriate *Panstripe* functions.

122 ***Ancestral state reconstruction***

123 To track the evolutionary origin of gene exchanges, the ancestral states of all gene families were reconstructed
124 from the full core genome phylogeny and the full pangenome presence/absence matrix. This was done
125 according to Wagner parsimony with equal penalty scores for both gene gains and losses using Count (v10.4)
126 (Csurös 2010). KOG annotations of the genes exchanged at the tree nodes of interest were retrieved from the
127 earlier made eggNOG annotations and visualised using in-house R scripts.

128 ***Mobilome analysis***

129 Pseudogenes were detected using Pseudofinder (v1.1.0) (Syberg-Olsen et al. 2022) at default settings. The
130 NCBI Protein database filtered for taxon ID 186801 (the *Clostridia* class) was used as reference database and
131 indexed using DIAMOND (v2.0.14). Insertion elements (ISEs) were inferred using ISEscan (v1.7.2.3) (Xie and
132 Tang 2017) at default settings. Biosynthetic gene clusters (BGCs) were identified using antiSMASH (v8.0.0)
133 (Blin et al. 2025) at default settings. Prophages were found by running PHASTEST (Wishart et al. 2023) in lite
134 mode using the image from Docker Hub (digest 2477a2f3b6b8). Phage infectivity graphs were made using
135 *supervenn* (v0.5.0). All mobile element counts were normalised by dividing by the genome length and
136 statistically compared using the Brunner-Munzel test ($\alpha = 0.05$) in scipy. Significance annotations were added
137 using *statannotations* (v0.6).

138 ***BGC clustering analysis***

139 All BGCs identified during the mobilome analysis were clustered using BiG-SCAPE (v1.1.8) (Navarro-Muñoz et
140 al. 2020), adding MIBiG reference BGCs (flag '--mibig'), mixing all BGC classes (flag '--mix'), and including
141 singletons (flag '--include_singletons'). A BiG-SCAPE distance cutoff value of 0.3 was applied as appeared
142 appropriate from parameter benchmarking (**Add. File 4**) and the resulting similarity networks. To integrate
143 phylogroup metadata in CytoScape (v3.10.2), the unidirectional network links in the output network files of
144 BiG-SCAPE were annotated by order or MIBiG and converted into bidirectional links by adding the reverse
145 link. BGC gene alignments were generated using clinker (v0.0.28) (Gilchrist and Chooi 2021).

146 # Results

147 ***Composing the genome set***

148 We downloaded all available *Clostridia* genomes from NCBI. After applying quality controls and species
149 dereplication, we had a collection of 1010 genomes. These sequences have been collected from various
150 isolation sources, including both open ecosystems as well as host-associated systems like human or animal
151 microbiomes. After taxonomic classification using GTDB-tk, we selected four orders that together cover a
152 large share of the *Clostridia* class, while being sufficiently large for strong statistical power (# >= 30) (**Add. File**
153 **2: Figure S2**): *Clostridiales*, *Lachnospirales*, *Oscillospirales* and *Peptostreptococcales*. The *Clostridiales*
154 comprise mainly free-living species from *Clostridium sensu stricto*. The *Lachnospirales* and the *Oscillospirales*
155 contain prevalent constituents of the human gut microbiome (the *Clostridium XIVa* and *Clostridium IV* species
156 groups, resp.). The *Peptostreptococcales* are commonly found as part of various microflora in the human body
157 as well (e.g. *C. difficile*).

158 ***Pangenome partition sizes and functional annotation indicate high genome plasticity***

159 To investigate the evolutionary dynamics, we first captured the gene conservation in all genomes through the
160 construction of a pangenome for our full genome set. Using a carefully selected set of clustering and
161 partitioning thresholds (**Add. File 3**), the pangenomes for each assessed group of taxa were constructed and
162 partitioned (**Table 1**). The recurringly large shell and cloud partitions suggest that there is a high level of
163 diversity among all orders, meaning that each genome is likely to harbour unique gene families. This finding

164 highlights the high level of genomic plasticity that is typical for species living in highly dynamic environments

165 like open ecosystems or human-associated microflora.

Table 1. Number of gene families in the pangenome partitions for each examined order, both in absolute numbers and referenced to the total number of gene families in each order's pangenome.

	All genomes (# = 1010)	***Clostridiales* (# = 209)**	***Lachnospirales* (# = 349)**	***Oscillospirales* (# = 164)**	***Peptostreptococcales* (# = 94)**
Core (> 90%)	353 0.09 %	690 0.79 %	574 0.36 %	465 0.59 %	444 0.81 %
Shell (0.1 % - 90%)	158,627 39.78 %	39,179 44.64 %	63,580 40.06 %	30,325 38.43 %	21,451 39.04 %
Cloud (< 0.1%)	239,797 60.16 %	47,903 54.58 %	94,546 59.58 %	48,109 60.98 %	33,053 60.15 %

166 After establishing the pangenome partitions, we took a closer look at the functional differences

167 between the orders (**Figure 1**). The functional profile is similar for most core genomes and is dominated by

168 the ribosomal proteins (COG category J). Nevertheless, a notable difference is the larger share of signal

169 transduction mechanisms in the core genome of *Clostridiales* (COG category T). Such mechanisms are a

170 hallmark of a free-living lifestyle as they facilitate more complex sensing of the environment, enabling more

171 specialised responses. In almost all orders, the majority of the shell and cloud gene families have no known

172 function (COG categories S and -), suggesting the presence of mobile genetic elements (MGEs), highly

173 specialised functions, and/or rapidly evolving proteins that fall outside the scope of our current

174 understanding.

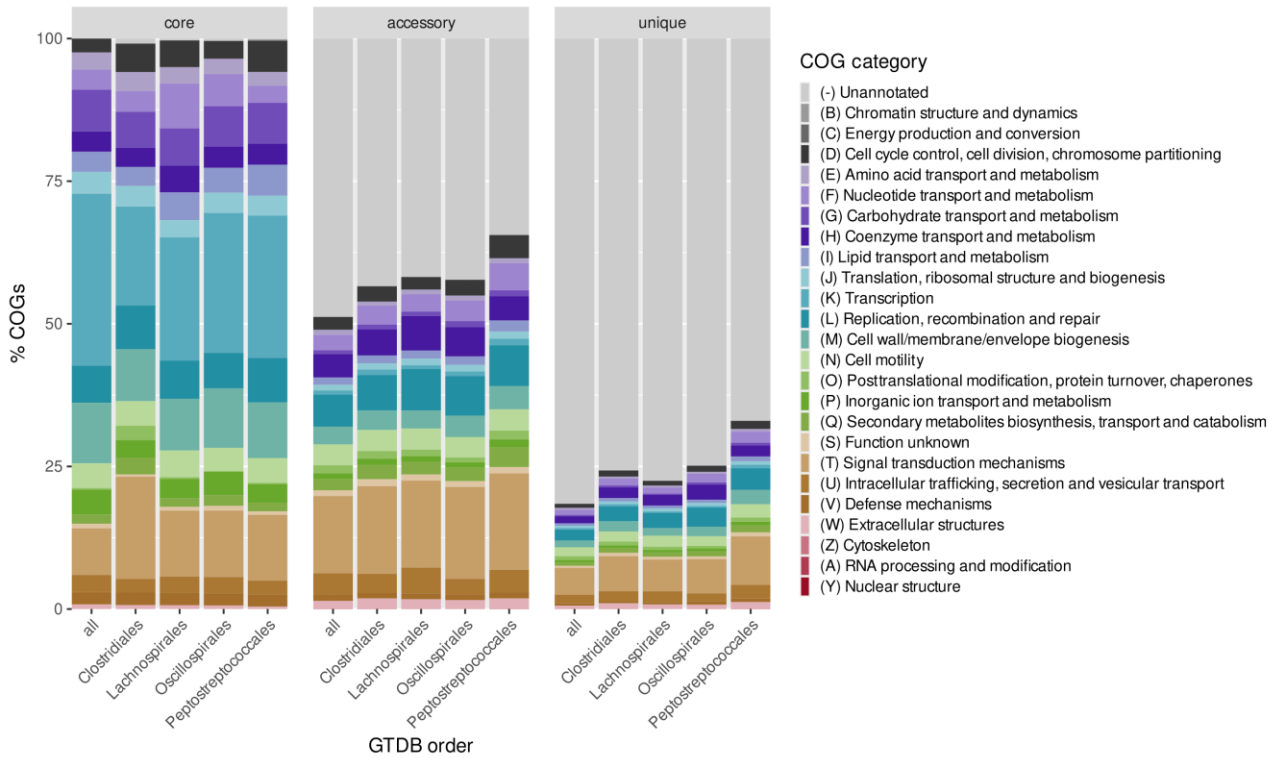

Figure 1. COG functional annotation profiles for all pangenome partitions of each examined clostridial order. Fractions are referenced to the number of gene families in a partition. The functional profiles are similar, except for a larger share of signal transduction genes in the core genome of the *Clostridiales* order, which indicates a free-living lifestyle.

175 ***KEGG pathway completeness analysis uncovers two distinct levels of metabolic versatility***

176 A KEGG pathway completeness analysis gauges the metabolic versatility of an organism by assessing the
177 completeness of metabolic pathway modules as documented in the KEGG Pathway database, based on the
178 genomic presence of the module's constituent genes. Remarkably, two distinct levels of pathway
179 completeness were observed for our full genome set. 394 out of 1010 organisms have visibly more complete
180 pathway modules, indicating a richer metabolism (**Figure 2**). This discrepancy is not driven by a limited
181 number of pathways but is the result of a genome-wide trend, as multiple modules tend to be more complete.
182 This aligns with earlier observations of a remarkably high metabolic versatility within the *Bacillota* phylum,
183 especially within the *Clostridium* genus (Pascal Andreu et al. 2023, Vieira-Silva et al. 2016).

184 Interestingly, the *Lachnospirales* order has a higher fraction of versatile organisms than the other
185 orders (55% vs. 25% - 35%; **Add. File 2: Supp. Fig. S3**). Moreover, only this order comprises genera and
186 lineages that are (almost) all metabolically versatile (*Enterocloster, Lacrimispora Butyvibrio,* one of two
187 *Blautia* lineages) Also, this versatility appears independent of genome size, except for the *Clostridiales* order

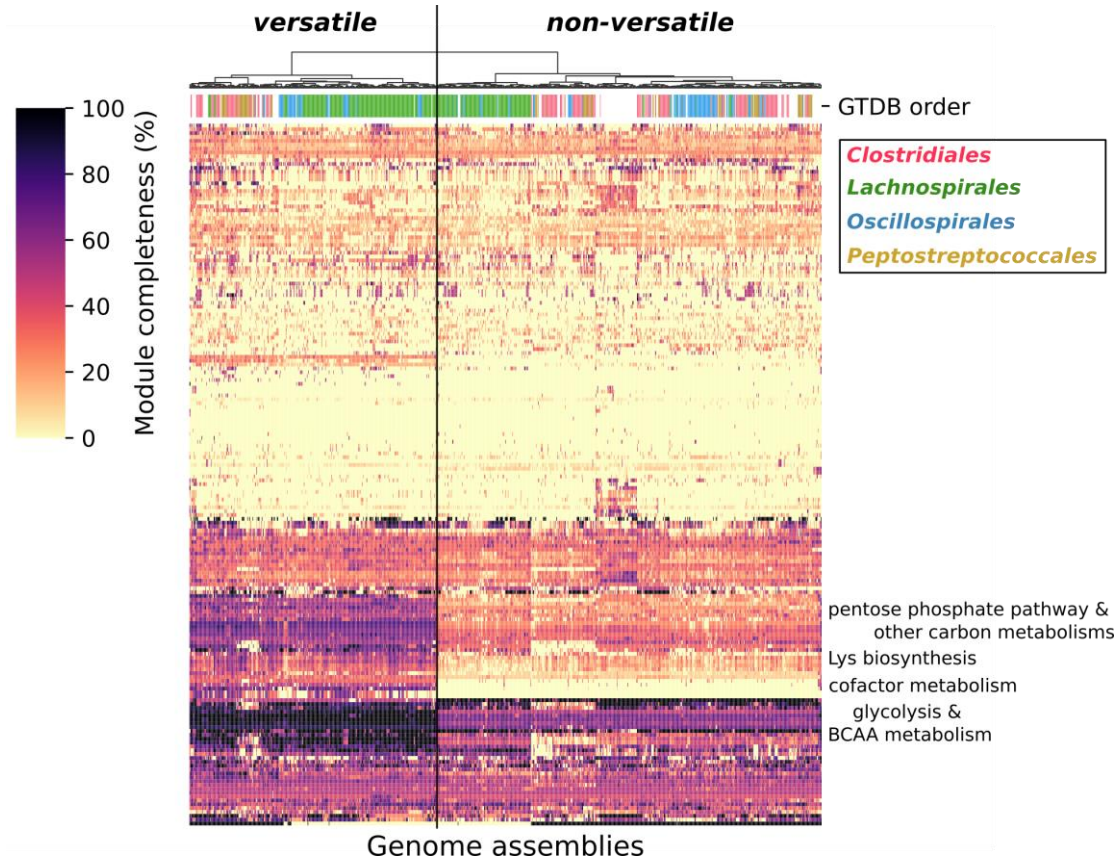

Figure 2. Ward's clustered heatmap of KEGG module completeness, annotated by taxonomic order. Metabolically versatile genomes recurringly show a general trend in higher module completenesses. Summarising, the key differing modules comprise glycolysis-related and BCAA biosynthesis pathways, as well as Lys biosynthesis, the pentose phosphate pathway and other carbon metabolisms. For a figure with all KEGG module labels, see Add. File 2: Fig. S5.

188 (**Add. File 2: Supp. Fig. S4**). Multiple clostridial genera contain both versatile and non-versatile species,
189 confirming an intra-genus divide that has previously been reported (Vieira-Silva et al. 2016).

190 Notably, the most prominent differences in module completeness are found in core metabolic
191 processes: the central carbon metabolism (glycolysis, pentose phosphate pathway) and branched-chain
192 amino acid and Lys biosynthesis. A versatile *Lachnospirales* species thus has a higher metabolic independence
193 than one would expect from a host-associated bacterium. In turn, this more developed primary metabolism
194 may support a more extensive secondary metabolism, providing additional advantages for the microbe.

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

196 The higher metabolic versatility of the *Lachnospirales* suggests that this order has a divergent evolutionary
197 trajectory than its other host-associated counterparts *Oscillospirales* and *Peptostreptococcales*. To assess

198 whether this discrepancy indeed may result from distinct evolutionary dynamics, gene exchange (GE) models
199 were fitted for each order separately and statistically compared to each other using *Panstripe*.

200 *Panstripe* uses a compound generalised linear model (GLM) to relate phylogenetic branch lengths
201 with the number of genes exchanged at those branches. Using an additional GLM term, internal nodes of the
202 tree can be differentiated from terminal ones, yielding core and tip gene exchange rates, respectively.
203 Exchange rates significantly different from 0 signify an association between branch length and the number of
204 gene exchange events, and thus active gene exchange. The core rate is to be interpreted as the overall gene
205 exchange rate underpinning the general core structure of the phylogeny, while the tip rate relates to short-
206 term gene mobility processes such as transposon or phage activity. However, in highly diverse genome sets
207 like the one in this work, an orthologous gene family may have been split into multiple gene families during
208 pangenome construction, artificially inflating both core and tip gene exchange rates. Consequently, *Panstripe*
209 is modelling here a more general genomic divergence (GD) rate, capturing both intrinsic gene evolution as
210 well as gene exchanges, rather than only the latter. Nevertheless, we argue that this more general genomic
211 divergence rate is still sufficient to address our research question, as it provides a measure of the pace at
212 which genomes have been evolving.

213 This GD modelling revealed several interesting findings about the evolutionary dynamics of the four
214 orders (**Table 2**). First, the *Oscillospirales* have the lowest core divergence rate, which is congruent with the
215 observation that the streamlined genomes of host-associated species evolve at a reduced rate (Dufresne,
216 Garczarek and Partensky 2005, Marais, Calteau and Tenaillon 2008). In contrast, the *Lachnospirales* have
217 significantly higher core and tip divergence rates, indicating a higher level of both long-term gene divergence
218 and short-term gene mobility, respectively. Strikingly, this contradicts its host-associated status since higher
219 levels of gene divergence and mobility imply still ongoing host adaptation (see also **Figure G**) (Siguier,
220 Gourbeyre and Chandler 2014). Furthermore, the tip divergence rate of the *Peptostreptococcales* is higher as
221 well, however, the core rates are not. One possible explanation is that this concerns increased gene mobility
222 at the onset of host adaptation. Another possibility is that the *Peptostreptococcales* are reaching the end of

223 the host conversion process, which reflects in dropping adaptation rates and the last to be eliminated gene
224 mobility.

Table 2. Fitted core and tip divergence rates of the GE models for each order, including p-values of all pairwise rate comparisons. *Oscillospirales* have a significantly lower divergence rate, consistent with its host-associated nature, whereas *Lachnospirales* display a high divergence rate inconsistent with its presumed host-associated status. This inconsistency can be explained by *Lachnospirales* tending to be in the earlier stages of host adaptation

		Tip divergence rate			
		Clostridiales 363.39 ± 42.28	*Lachnospirales* 621.03± 35.88	*Oscillospirales* 445.80 ± 35.63	*Peptostreptococcales* 537.80 ± 67.47
Core divergence rate	*Clostridiales* 5140.29 ± 603.74	/	$9.8 \cdot 10^{-6}$ ****	0.150	0.0255 *
	Lachnospirales 5995.25 ± 547.81	0.314	/	$3.37 \cdot 10^{-3}$ **	0.306
	Oscillospirales 4218.27 ± 412.98	0.210	0.0199 *	/	0.186
	Peptostreptococcales 5222.06 ± 684.72	0.928	0.399	0.177	/

225 ***Ancestral state reconstruction uncovers different ancestral lifestyles for the host-associated orders***

226 An ancestral state reconstruction (ASR) infers the clade-specific genes that have been passed on from the
227 common ancestor to its descendants. Based on the annotation of these reconstructed ancestral genes, we
228 can infer the lifestyle of this common ancestor, pinpointing the historic origin of lifestyle differences. The ASR
229 of this phylogeny shows that a relatively high number of genes have been gained and lost in the vicinity of
230 the root node of each order (**Figure**).

231 796 genes appear to have been present since the clostridial last common ancestor (LCA). These are
232 active in several central subsystems like glycolysis, tRNA biosynthesis and the ribosomes. In addition, they
233 include genes for flagellar assembly, two-component signalling (2CS) systems and DNA repair systems. Hence,
234 the last common ancestor of the clostridial lineage appears a well-equipped free-living microbe.

235 Regarding the most recent common ancestors of each order, some interesting contrasts can be
236 observed in both the number and functional annotation of the exchanged genes (**Add. File 5**). The MRCA of
237 *Clostridiales* gained 88 genes. Only about half of these families is annotated, which are mostly associated with
238 metabolism, signal transduction and cell motility, aligning with a free-living lifestyle. In contrast, it lost only
239 19 genes, of which the annotated share is again mostly metabolism-related. Notably, the ancestor of a large
240 subcluster containing the *Clostridium sensu stricto* clade has gained 182 genes, which are mostly associated

241 with metabolism, and DNA replication and repair, indicating that this subcluster even further refined and
242 expanded its ancestral free-living capacities.

243 The *Oscillospirales* MRCA gained 42 genes and lost 38, mostly related to metabolism in both cases.
244 Interestingly, the direct ancestor of this node has lost a considerable number of gene families. These 143 lost
245 genes were involved in cell motility, cofactor metabolism and signal transduction pathways. These losses align
246 with the reduction in free-living capabilities that typically occurs when a microorganism transitions to a host-
247 associated lifestyle.

248 In contrast, the *Lachnospirales* MRCA has gained 62 genes and lost 40. The gained genes are
249 associated with amino acid and lipid metabolism, and DNA replication and repair, while the lost ones are
250 mostly unannotated. Also here, there is a descendant that has exchanged a considerable number of genes:
251 the MRCA of the *Clostridium XIVa* clade. This descendant gained 133 genes and lost 121. Gained
252 functionalities affect several metabolic systems, including secondary metabolite biosynthesis, while lost
253 functionalities are linked with cell motility and signal transduction. Both the number of exchanged genes and
254 the combination of gaining metabolic capabilities with losing free-living traits suggest that the *Lachnospirales*
255 are at a different level of host association than the *Oscillospirales*.

256 Lastly, the *Peptostreptococcales* ancestral lineages exchanged relatively few genes. For example, the
257 MRCA gained 49 genes and lost only 7. The gained genes are involved in several metabolic subsystems, such
258 as amino acid metabolism and glycan biosynthesis. This low level of gene exchange may indicate that the
259 *Peptostreptococcales* are at the end of the host conversion process or only have just started it.

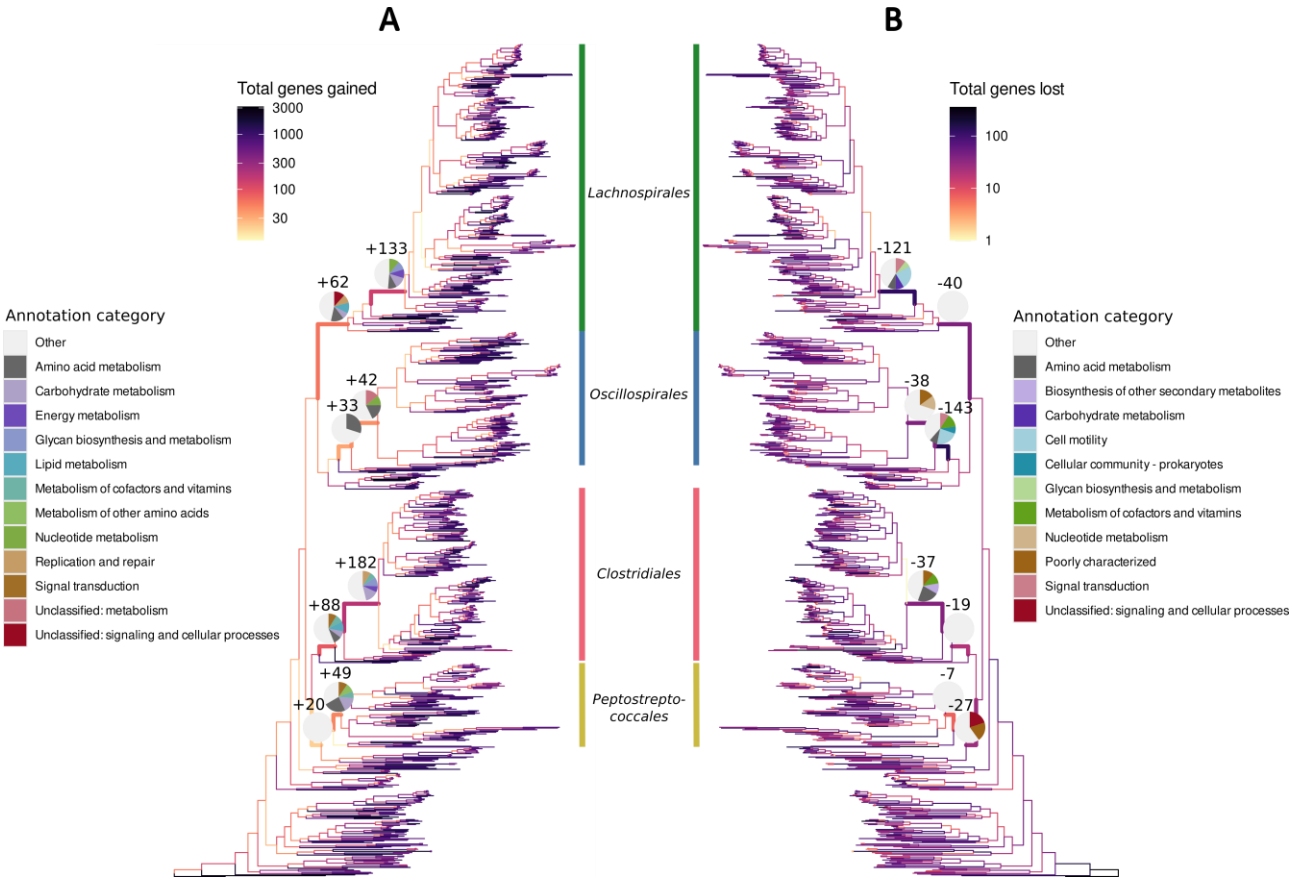

Figure 3 Wagner parsimony gene gain (A) and loss (B) profiles for the full genome set (coloured branches), and absolute number and KEGG BRITE level B annotation profile of exchanged genes for a selection of ancestral nodes (inset pie charts).

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

261 Both the GD and the ASR analyses indicate varying levels of host association among the clostridial orders.
262 Moreover, the higher observed metabolic versatility of *Lachnospirales* may be related with its lower levels of
263 host association. To further support our identification of these varying stages of host adaptation, we screened
264 our entire genome set for the abundance of a set of genomic hallmarks tied with progressing host adaptation.

265 More specifically, the transition from a free-living to a host-associated lifestyle in bacteria is thought
266 to be coupled with a genome streamlining process driven by insertion sequences (ISes) and intrinsic
267 mutations (Siguier, Gourbeyre and Chandler 2014). ISes are a type of intracellular mobile genetic elements
268 (MGEs) that employ transposases for their mobility. When bacteria move from a nutritionally poor free-living

269 environment to a nutritionally rich host-associated setting, many genes lose their essentiality due to the
270 constant availability of metabolites. This causes a reduced selection pressure on non-essential genes,
271 enabling higher mutation rates as well as IS insertions into these genes. Both these phenomena drive
272 pseudogene formation (Marais, Calteau and Tenaillon 2008). Over time, the accumulation and multiplication
273 of IS regions promotes genome rearrangements and deletions, ultimately yielding smaller streamlined
274 genomes tailored to their new host-associated lifestyle (**Figure G**) (Siguier, Gourbeyre and Chandler 2014).

275 To explore this hypothesis, we screened our full clostridial genome set for the abundance of the
276 following MGEs and genomic features: (1) genome sizes, (2) ISes, (3) pseudogenes, (4) prophage regions, and
277 (5) biosynthetic gene clusters. (1), (2) and (3) are key hallmarks of the genome streamlining process. (4) may
278 reflect a shift in environmental niche. (5) may provide an indication of the metabolic versatility of the
279 organism, as the necessary precursor metabolites must be available.

280 The results align strikingly well with our previous interpretations. First, *Oscillospirales* and
281 *Peptostreptococcales* genomes are significantly smaller than *Clostridiales* and *Lachnospirales* (**Figure A**).

282 Second, *Lachnospirales* and *Peptostreptococcales* genomes harbour significantly more ISes than
283 *Clostridiales*, which, in turn, harbour more than *Oscillospirales* genomes (**Figure B**). This observation aligns
284 with our hypothesis that *Lachnospirales* is in the early stages of host adaptation and thus is undergoing an IS
285 expansion. Furthermore, the smaller genomes and lower number of ISes of the *Peptostreptococcales* indicate
286 that this latter group is slightly more advanced in this transition process than the *Lachnospirales*.

287 A deeper analysis of the IS families revealed some substantial differences in their distribution among
288 the species orders (**Add. File 2: Figure S6**). For example, family IS1634 is only present in *Lachnospirales* and
289 *Peptostreptococcales*, while family IS5 is only present in *Oscillospirales*. Notably, families IS110, IS66 and ISL3
290 are more prevalent in *Lachnospirales* species. These IS families do not have a specific insertion site preference
291 (Siguier, Gourbeyre and Chandler 2014, Siguier et al. 2015), so their abundances may merely be the result of
292 IS expansion during the host conversion process.

293 Third, pseudogenes are significantly less abundant in *Clostridiales* than in *Peptostreptococcales,*
294 which at their turn harbour less than *Oscillospirales* and *Lachnospirales.* (**Figure C**). This implies that both gut-
295 associated clusters have undergone or are undergoing more intense pseudogene formation.

296 Fourth, the number of BGCs is significantly higher in *Clostridiales* and *Peptostreptococcales*, with
297 *Lachnospirales* and *Oscillospirales* at comparable lower levels (**Figure D**). This hints that *Clostridiales* and
298 *Peptostreptococcales* species maintain a greater capacity for specialised metabolism, whereas the other two
299 orders seem to have eliminated a substantial share of their BGCs. Given that other genomic features indicate
300 that *Lachnospirales* is only in the early stages of host adaptation, it may seem that BGCs are among the first
301 to be eliminated under altered selection pressures despite the maintenance of non-BGC metabolic versatility.
302 However, this may also reflect our current knowledge gap for BGCs from gut-associated species. Regarding
303 the BGC classes, the most substantial differences in abundance are observed in the cyclic lactone autoinducer
304 (CLA), NRPS(-like), and RiPP-like classes. In addition, there is a consistent presence of ranthipeptides and
305 terpene precursors across all orders (**Add. File 2: Figure S7**).

306 Fifth, *Lachnospirales* have significantly less prophages than the other orders (**Figure E**). Regarding
307 infection targets, *Lachnospirales* are infected by the highest number of prophage types, while
308 *Peptostreptococcales* by the lowest. Interestingly, about 75% of the detected prophage types infect clostridial
309 species from multiple orders, with 39 even infecting species from all orders (**Figure F**). Besides, some
310 prophages carry passenger genes for methyltransferases, restriction enzymes, toxin(-antitoxin system)s and
311 transporters (including beta-lactam efflux pumps and other resistance proteins), highlighting a mechanism by
312 which clostridial species may acquire new genes (**Add. File 6**). Hence, while there is a clear difference in the
313 prophages present, it remains unclear whether this merely reflects a niche shift or that these prophages carry
314 genes that have become essential for the bacterium's survival due to the host adaptation.

315 To conclude, the abundances of these genomic features show again that the clostridial species orders
316 are in various stages of the host adaptation process (**Figure G**).

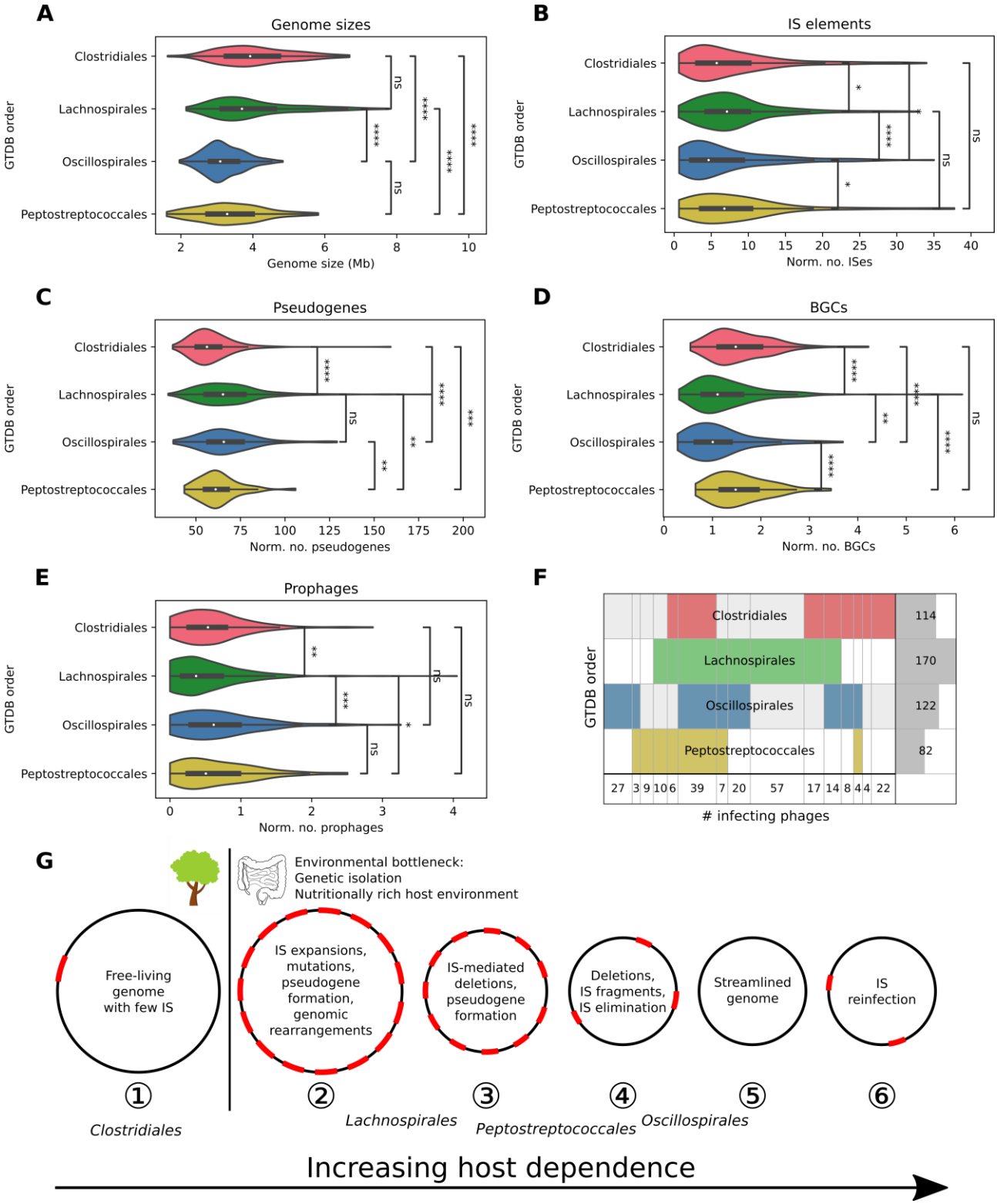

Figure 4. **Screening of the genomic hallmarks of host dependence. (A) – (F)** Comparison of the abundance of a genomic feature across all orders. Feature counts are normalised by dividing by the genome length in Mb. The applied statistical test is the Brunner-Munzel test (α = 0.05). (A) genome size, (B) insertion sequence elements, (C) pseudogenes, (D) biosynthetic gene clusters, (E) prophage regions, (F) number of infecting phages by host order. **(G) Schematical overview of the impact of the genome streamlining process on the bacterial genomic architecture**, including the inferred position of each clostridial order in this process. Figure inspired on (Siguier, Gourbeyre and Chandler 2014). (1) The (ancestral) free-living genome has only few ISes. (2) As a result of an environmental bottleneck, IS expansion occurs, leading to an increased abundance of ISes, pseudogenes and mutations. Genomic rearrangements occur. (3) With time, the repetitive regions of the ISes facilitate the deletion of non-essential genomic regions. (4) Eventually, extensive deletion leads to the elimination of the ISEs. (5) Finally, the genome reaches its streamlined form, with a gene content tailored for life in its host environment. (6) Upon contact with IS-carrying strains or phages, ISes may be reintroduced.

317

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

319 The species in our clostridial dataset have faced varying selection pressures. Hence, any BGC that is conserved
320 despite these varying pressures, is likely to provide potent properties that offer an evolutionary advantage
321 large enough to justify its conservation. Such high-potential BGCs should emerge in a BGC clustering analysis.

322 Our clustering revealed that the five largest BGC groups are two terpene-precursor groups and three
323 ranthipeptide groups. Both *Clostridiales* and *Lachnospirales* harbour one terpene-precursor and one or two
324 ranthipeptide BGC groups (**Figure**). Alignment of gene clusters from each of these three largest ranthipeptide
325 groups shows that the core biosynthetic genes, the SCIFF precursor and the RRE gene, are conserved across
326 all groups (**Add. File 2: Figure S8)**. Moreover, these ranthipeptide groups are well-distributed across the full
327 phylogeny, with every species coding maximum one ranthipeptide (**Add. File 2: Figure S9**). Interestingly, two
328 of these ranthipeptide BGCs have both been experimentally shown to participate in quorum sensing activities
329 (Chen et al. 2020). The fact that these two BGCs belong to the same ranthipeptide group in our analysis
330 suggests that similar quorum sensing functions may be associated with this entire ranthipeptide group.

331 The smaller identified BGC groups are cyclic lactone autoinducers and deazapurines, next to more
332 smaller ranthipeptides and terpene precursor groups. We coclustered our BGC set with the MIBiG database,
333 a highly curated and standardised repository of known BGCs (Zdouc et al. 2025). Interestingly, very few MIBiG
334 BGCs cocluster with our clostridial BGCs, underscoring the novelty of these conserved ranthipeptides and
335 terpene precursors, and their high potential for new functional significance. We also observed a large group
336 of singleton BGCs (**Add. File 7**), which may comprise BGCs acquired more recently by horizontal gene transfer
337 or gene rearrangement. While these singletons have not undergone sufficient evolutionary selection to refine
338 their functional significance, any potent and biologically relevant activity cannot be excluded.

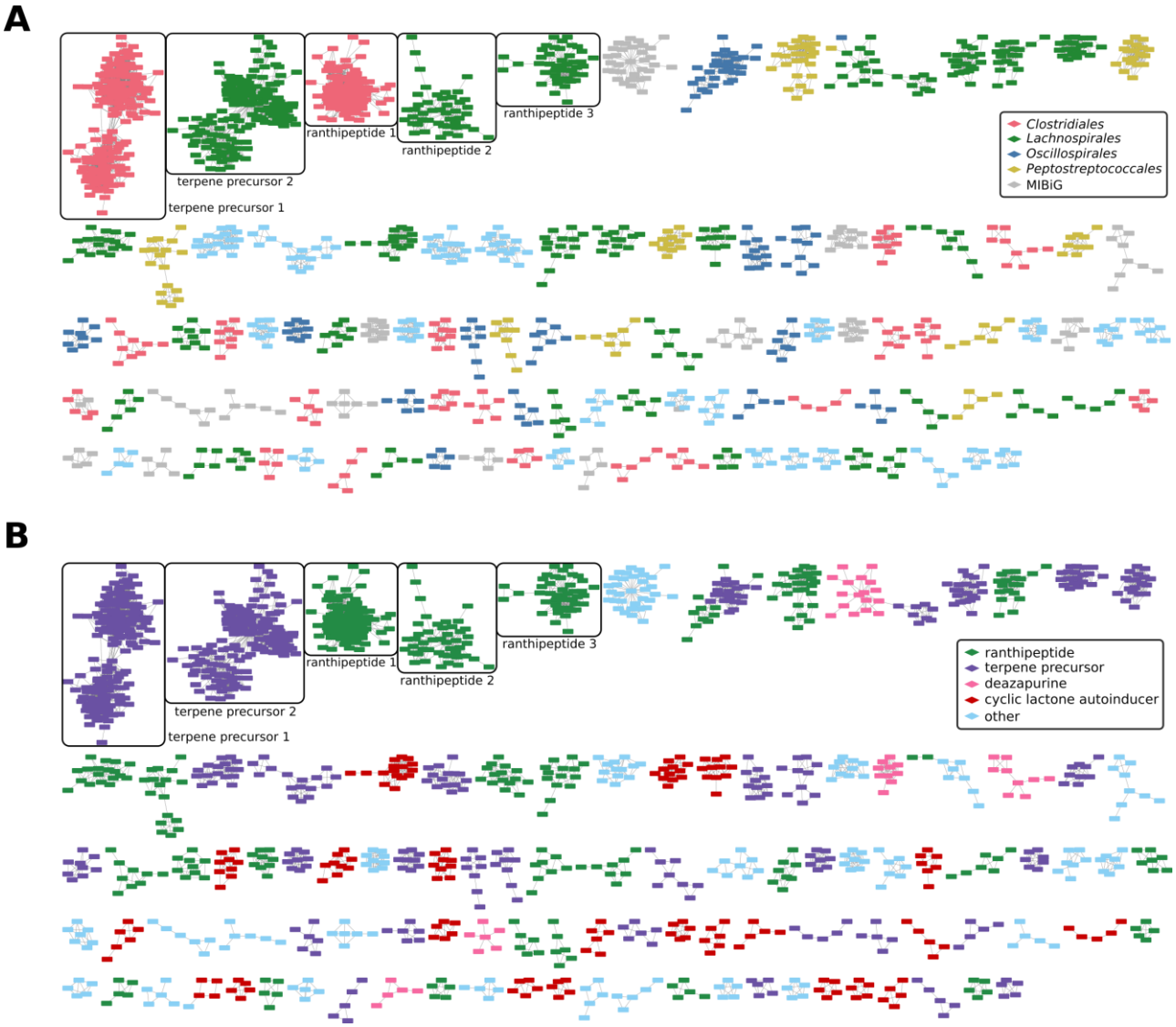

Figure 5. BiG-SCAPE similarity networks of all BGCs detected in our full genome set (threshold 0.3), annotated (A) by clostridial order or as MIBiG entry, and (B) by BGC type. The five largest groups on the top left are either terpene precursors or ranthipeptides. Only groups with at least five members are shown.

Discussion

This work presents the first comparative pangenomics analysis of multiple orders of the *Clostridia* class and is the first to report about their intriguingly different host adaptation tendencies. Moreover, it uncovered a remarkably large repertoire of novel BGCs, including several highly conserved ranthipeptides.

A major finding of this study is the varying stages in host adaptation in each taxonomic order. By using multiple independent approaches in parallel, we were able to position each order on the host association spectrum. In particular, the early-stage host association trend of the *Lachnospirales* highlights a new important nuance of the current host-associated status of this prevalent clade of gut symbionts. Much of the genomic and metabolic diversity of this clade may have been lumped by the generic label “host-associated”.

348 Bacteria in the early stages of host association likely still retain part of their free-living gene repertoire. Indeed,
349 the KEGG pathway completeness assessment has shown that the *Lachnospirales* are more likely to be
350 metabolically versatile. Due to their higher innate capacity to produce more complex metabolites, these
351 species would be excellent starting points for the development of anaerobic chassis organisms for
352 heterologous expression of potent BGCs. Moreover, they have a high potential as therapeutic delivery
353 vehicles in gut environments, as they are adapting to exactly those environments (Kubiak and Minton 2015).

354 Another key outcome of this study is the comprehensive identification of clostridial BGCs, providing
355 the first comparative view of the natural product repertoire of clostridial bacteria. Despite extensive host
356 adaptation and a long track record of exposure to varying selection pressures, several ranthipeptide and
357 terpene precursor BGCs have been widely conserved. The lack of MIBiG entries coclustering with these BGCs
358 underscores their novelty, tapping further into the unexplored richness of BGCs from the anaerobic realm.
359 These novel conserved BGCs are prime targets for further experimental characterisation.

360 Nevertheless, there are some limitations to be considered for this study. First, since we conducted a
361 high-level analysis averaging counts from species orders, the conclusions presented here are to be considered
362 general tendencies of each order and should not be extrapolated to individual species without further
363 analysis. For example, it may be possible that a *Lachnospirales* species is already highly adapted to its host
364 because of higher selection pressures, or because it started the host adaptation process early compared to
365 its close relatives. Second, we dereplicated our dataset to ensure an equal species representation in the
366 pangenome. Non-representative genomes may harbour even more novel diversity, both at the pangenome
367 level as well as at the BGC level.

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

Data availability

All genome assemblies analysed in this study were retrieved from NCBI RefSeq. Accession codes and relevant metadata can be found in Add. File 1.

Code availability

The full analysis pipeline has been deposited at GitHub (https://github.com/LucoDevro/closHomer) and at Zenodo (https://doi.org/10.5281/zenodo.14719962).

Author contribution statement

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 wrote the article, which was revised by LDV, JA and JM.

Acknowledgements

We thank Jan Aerts for his helpful advice on improving our data visualisations.

Funding statement

This work was funded by the European Commission through the European Research Council (MiStiC, 101078461).

Competing interest statement

We declare no competing interests.

References

Asnicar, Francesco, Thomas, Andrew Maltez, Beghini, Francesco, Mengoni, Claudia, Manara, Serena, Manghi, Paolo, et al., 'Precise Phylogenetic Analysis of Microbial Isolates and Genomes from Metagenomes Using PhyloPhlAn 3.0', *Nature Communications*, 11/1 (2020), 2500

Behnken, Swantje, and Hertweck, Christian, 'Anaerobic Bacteria as Producers of Antibiotics', *Applied Microbiology and Biotechnology*, 96/1 (2012), 61–67

Birgen, Cansu, Degnes, Kristin F., Markussen, Sidsel, Wentzel, Alexander, and Sletta, Håvard, 'Butanol Production from Lignocellulosic Sugars by Clostridium Beijerinckii in Microbioreactors', *Biotechnology for Biofuels*, 14/1 (2021)

Blin, Kai, Shaw, Simon, Vader, Lisa, Szenei, Judit, Reitz, Zachary L, Augustijn, Hannah E, et al., 'AntiSMASH 8.0: Extended Gene Cluster Detection Capabilities and Analyses of Chemistry, Enzymology, and Regulation', *Nucleic Acids Research*, 2025

Buchfink, Benjamin, Reuter, Klaus, and Drost, Hajk-Georg, 'Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND', *Nature Methods*, 18/4 (2021), 366–68

Cantalapiedra, Carlos P, Hernández-Plaza, Ana, Letunic, Ivica, Bork, Peer, and Huerta-Cepas, Jaime, 'EggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale', *Molecular Biology and Evolution*, 38/12 (2021), 5825–29

Capella-Gutiérrez, Salvador, Silla-Martínez, José M., and Gabaldón, Toni, 'TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses', *Bioinformatics*, 25/15 (2009), 1972–73

Chaumeil, Pierre-Alain, Mussig, Aaron J, Hugenholtz, Philip, and Parks, Donovan H, 'GTDB-Tk v2: Memory Friendly Classification with the Genome Taxonomy Database', *Bioinformatics*, 38/23 (2022), 5315–16

Chen, Yunliang, Yang, Yunpeng, Ji, Xinjian, Zhao, Ran, Li, Guoquan, Gu, Yang, et al., 'The SCIFF-Derived Ranthipeptides Participate in Quorum Sensing in Solventogenic Clostridia', *Biotechnology Journal*, 15/10 (2020)

Cohen, Jonathan E., Wang, Rong, Shen, Rong-Fong, Wu, Wells W., and Keller, James E., 'Comparative Pathogenomics of Clostridium Tetani', *PLOS ONE*, 12/8 (2017), e0182909

Cruz-Morales, Pablo, Orellana, Camila A, Moutafis, George, Moonen, Glenn, Rincon, Gonzalo, Nielsen, Lars K, et al., 'Revisiting the Evolution and Taxonomy of Clostridia, a Phylogenomic Update', *Genome Biology and Evolution*, 11/7 (2019), 2035–44

Csurös, Miklós, 'Count: Evolutionary Analysis of Phylogenetic Profiles with Parsimony and Likelihood', *Bioinformatics*, 26/15 (2010), 1910–12

Dufresne, Alexis, Garczarek, Laurence, and Partensky, Frédéric, 'Accelerated Evolution Associated with Genome Reduction in a Free-Living Prokaryote', *Genome Biology*, 6/2 (2005), R14

Gautreau, Guillaume, Bazin, Adelme, Gachet, Mathieu, Planel, Rémi, Burlot, Laura, Dubois, Mathieu, et al., 'PPanGGOLiN: Depicting Microbial Diversity via a Partitioned Pangenome Graph', *PLOS Computational Biology*, 16/3 (2020), e1007732

Gilchrist, Cameron L.M., and Chooi, Yit Heng, 'Clinker & Clustermap.Js: Automatic Generation of Gene Cluster Comparison Figures', *Bioinformatics*, 37/16 (2021), 2473–75

Katoh, K., and Standley, D. M., 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30/4 (2013), 772–80

Kubiak, Aleksandra M., and Minton, Nigel P., 'The Potential of Clostridial Spores as Therapeutic Delivery Vehicles in Tumour Therapy', *Research in Microbiology*, 166/4 (2015), 244–54

Letunic, Ivica, and Bork, Peer, 'Interactive Tree of Life (ITOL) v6: Recent Updates to the Phylogenetic Tree Display and Annotation Tool', *Nucleic Acids Research*, 52/W1 (2024), W78–82

Letzel, Anne Catrin, Pidot, Sacha J., and Hertweck, Christian, 'A Genomic Approach to the Cryptic Secondary Metabolome of the Anaerobic World', *Natural Product Reports*, 30/3 (2013), 392–428

Marais, Gabriel A.B., Calteau, Alexandra, and Tenaillon, Olivier, 'Mutation Rate and Genome Reduction in Endosymbiotic and Free-Living Bacteria', *Genetica*, 134/2 (2008), 205–10

Maturana, José Luis, and Cárdenas, Juan P., 'Insights on the Evolutionary Genomics of the Blautia Genus: Potential New Species and Genetic Content Among Lineages', *Frontiers in Microbiology*, 12 (2021)

Minh, B. Q., Nguyen, M. A. T., and von Haeseler, A., 'Ultrafast Approximation for Phylogenetic Bootstrap', *Molecular Biology and Evolution*, 30/5 (2013), 1188–95

Minh, Bui Quang, Schmidt, Heiko A, Chernomor, Olga, Schrempf, Dominik, Woodhams, Michael D, von Haeseler, Arndt, et al., 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37/5 (2020), 1530–34

Navarro-Muñoz, Jorge C., Selem-Mojica, Nelly, Mullowney, Michael W., Kautsar, Satria A., Tryon, James H., Parkinson, Elizabeth I., et al., 'A Computational Framework to Explore Large-Scale Biosynthetic Diversity', *Nature Chemical Biology*, 16/1 (2020), 60–68

O'Leary, Nuala A., Cox, Eric, Holmes, J. Bradley, Anderson, W. Ray, Falk, Robert, Hem, Vichet, et al., 'Exploring and Retrieving Sequence and Metadata for Species across the Tree of Life with NCBI Datasets', *Scientific Data*, 11/1 (2024), 732

Pahalagedara, Amila Srilal Nawarathna Weligala, Flint, Steve, Palmer, Jon, Brightwell, Gale, and Gupta, Tanushree Barua, 'Antimicrobial Production by Strictly Anaerobic Clostridium Spp.', *International Journal of Antimicrobial Agents*, 55/5 (2020)

457 Pascal Andreu, Victòria, Augustijn, Hannah E., Chen, Lianmin, Zhernakova, Alexandra, Fu, Jingyuan,
458 Fischbach, Michael A., et al., 'GutSMASH Predicts Specialized Primary Metabolic Pathways from the
459 Human Gut Microbiota', *Nature Biotechnology*, 41/10 (2023), 1416–23
460 Patakova, Petra, Branska, Barbora, Sedlar, Karel, Vasylkivska, Maryna, Jureckova, Katerina, Kolek, Jan, et al.,
461 'Acidogenesis, Solventogenesis, Metabolic Stress Response and Life Cycle Changes in Clostridium
462 Beijerinckii NRRL B-598 at the Transcriptomic Level', *Scientific Reports*, 9/1 (2019)
463 Roberts, Seth B, Gowen, Christopher M, Brooks, J Paul, and Fong, Stephen S, 'Genome-Scale Metabolic
464 Analysis of Clostridium Thermocellum for Bioethanol Production', *BMC Systems Biology*, 4/1 (2010), 31
465 <https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-4-31>
466 Ruiz-Perez, Carlos A., Conrad, Roth E., and Konstantinidis, Konstantinos T., 'MicrobeAnnotator: A User-
467 Friendly, Comprehensive Functional Annotation Pipeline for Microbial Genomes', *BMC Bioinformatics*,
468 22/1 (2021), 11
469 Salamzade, Rauf, and Kalan, Lindsay R., 'SkDER: Microbial Genome Dereplication Approaches for
470 Comparative and Metagenomic Applications', *BioRxiv*, 2023
471 Sandhu, Brindar K., and McBride, Shonna M., 'Clostridioides Difficile', *Trends in Microbiology*, 26/12 (2018),
472 1049–50
473 Schwengers, Oliver, Jelonek, Lukas, Dieckmann, Marius Alfred, Beyvers, Sebastian, Blom, Jochen, and
474 Goesmann, Alexander, 'Bakta: Rapid and Standardized Annotation of Bacterial Genomes via
475 Alignment-Free Sequence Identification', *Microbial Genomics*, 7/11 (2021)
476 Siguier, Patricia, Gourbeyre, Edith, and Chandler, Mick, 'Bacterial Insertion Sequences: Their Genomic
477 Impact and Diversity', *FEMS Microbiology Reviews*, 38/5 (2014), 865–91
478 Siguier, Patricia, Gourbeyre, Edith, Varani, Alessandro, Ton-Hoang, Bao, and Chandler, Mick, 'Everyman's
479 Guide to Bacterial Insertion Sequences', *Microbiology Spectrum*, 3/2 (2015)
480 Smith, Theresa J., Hill, Karen K., and Raphael, Brian H., 'Historical and Current Perspectives on Clostridium
481 Botulinum Diversity', *Research in Microbiology*, 166/4 (2015), 290–302
482 Sugimoto, Yuki, Camacho, Francine R., Wang, Shuo, Chankhamjon, Pranatchareeya, Odabas, Arman, Biswas,
483 Abhishek, et al., 'A Metagenomic Strategy for Harnessing the Chemical Repertoire of the Human
484 Microbiome', *Science*, 366/6471 (2019)
485 Syberg-Olsen, Mitchell J, Garber, Arkadiy I, Keeling, Patrick J, McCutcheon, John P, and Husnik, Filip,
486 'Pseudofinder: Detection of Pseudogenes in Prokaryotic Genomes', *Molecular Biology and Evolution*,
487 39/7 (2022)
488 Tonkin-Hill, Gerry, Gladstone, Rebecca A., Pöntinen, Anna K., Arredondo-Alonso, Sergio, Bentley, Stephen D.,
489 and Corander, Jukka, 'Robust Analysis of Prokaryotic Pangenome Gene Gain and Loss Rates with
490 Panstripe', *Genome Research*, 33/1 (2023), 129–40
491 Udaondo, Zulema, Duque, Estrella, and Ramos, Juan Luis, 'The Pangenome of the Genus Clostridium',
492 *Environmental Microbiology*, 19/7 (2017), 2588–2603
493 Vieira-Silva, Sara, Falony, Gwen, Darzi, Youssef, Lima-Mendez, Gipsi, Garcia Yunta, Roberto, Okuda, Shujiro,
494 et al., 'Species-Function Relationships Shape Ecological Properties of the Human Gut Microbiome',
495 *Nature Microbiology*, 1/8 (2016)
496 Wishart, David S, Han, Scott, Saha, Sukanta, Oler, Eponine, Peters, Harrison, Grant, Jason R, et al.,
497 'PHASTEST: Faster than PHASTER, Better than PHAST', *Nucleic Acids Research*, 51/W1 (2023), W443–50
498 Xie, Zhiqun, and Tang, Haixu, 'ISEScan: Automated Identification of Insertion Sequence Elements in
499 Prokaryotic Genomes', *Bioinformatics*, 33/21 (2017), 3340–47
500 Yu, Guangchuang, Smith, David K., Zhu, Huachen, Guan, Yi, and Lam, Tommy Tsan-Yuk, 'Ggtree: An R
501 Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other
502 Associated Data', *Methods in Ecology and Evolution*, 8/1 (2017), 28–36
503 Zdouc, Mitja M, Blin, Kai, Louwen, Nico L L, Navarro, Jorge, Loureiro, Catarina, Bader, Chantal D, et al.,
504 'MIBiG 4.0: Advancing Biosynthetic Gene Cluster Curation through Global Collaboration', *Nucleic Acids*
505 *Research*, 53/D1 (2025), D678–90
506