

RESEARCH

Open Access



Comparative population genomics reveal the genetic features associated with the plant host adaptation of *Clostridium butyricum*

Cong-Jian Li^{1†}, Yu Zhao^{1†}, Mei Tang^{1†}, Xiao Chu², Peng-Chao Zhan¹, Xing-Wang Jiang¹, Jin-Yu Tian¹, Xuan Hai¹, Yi-Fei Lu¹, Ling-Ling Yang^{1*} and Xiao-Yang Zhi^{1*}

Abstract

Background Plants are increasingly considered as secondary reservoirs for enterics. However, little is known about their population dynamics and the genetic mechanisms during plant colonization. *Clostridium butyricum* is a gut symbiont of humans and animals and, rarely, a pathogen. Here, 55 strains of *C. butyricum* isolated from the roots of *Paris polyphylla* var. *yunnanensis* provided a new model for understanding plant-host adaptation of enterics.

Results These strains, along with 67 non-endophytic *C. butyricum* strains (nECB), were examined for population structure, revealing that they diverged into four well-defined lineages, whereas endophytic *C. butyricum* strains (ECB) from different sources were scattered across two lineages. The population diversity estimate confirmed the genetic distinctiveness among four lineages and uncovered distinct evolutionary processes that might drive the divergence of ECB-related lineages. Frequent gene flow between ECB and nECB suggested that plant-host colonization does not lead to genetic isolation. Extensive recombinations within and between lineages demonstrated the major role of recombination in shaping population genetic structure and diversification in *C. butyricum*. Additionally, the endophytic variance analysis identified several genes associated with CRISPR, defense systems, and metabolism that contribute to endophytic colonization by *C. butyricum*.

Conclusion This study provides novel insights into the ongoing adaptation of *C. butyricum* to plant hosts and illuminates the genetic mechanisms underlying this host transition. By elucidating population structure, gene flow, recombination patterns, and candidate adaptive genes, our findings advance the understanding of host-associated evolution in enteric bacteria.

Keywords *Clostridium butyricum*, Population genomics, Plant host adaptation, Recombination

[†]Cong-Jian Li, Yu Zhao and Mei Tang contributed equally to the work.

*Correspondence:

Ling-Ling Yang
yangll@ynu.edu.cn
Xiao-Yang Zhi
xyzhi@ynu.edu.cn

¹Yunnan Institute of Microbiology, Key Laboratory of Microbial Diversity in Southwest China of Ministry of Education, School of Life Sciences, Yunnan University, Kunming 650091, People's Republic of China

²Institute of Biomedical Research, Yunnan University, Kunming 650091, People's Republic of China



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

The gut microbiota and its effect on host health are increasingly relevant and supported by many investigations in model systems and proof-of-concept clinical trials. Recently, fresh produce-associated gastrointestinal disease outbreaks have remained a pressing issue for consumers and researchers [1]. Despite increased awareness among producers and consumers and advances in sanitation protocols and technology, these outbreaks still recur [2, 3]. One explanation is that some enteric pathogens, such as nontyphoidal *Salmonella* spp. and enterovirulent *Escherichia coli*, can persist on and, in some cases, colonize plant tissues, thereby facilitating transmission via raw produce. However, host switching is not necessarily unidirectional: plant- and soil-associated bacteria may also acquire the capacity to colonize animal intestines [4]. Recent comparative and functional genomic studies reveal that certain common metabolic capabilities, such as nitrogen-fixing genes, contribute to the fitness of enteric bacteria in plants [5, 6]. While horizontal gene acquisition from endophytes may boost plant-colonization abilities in specific isolates, these adaptations are not consistently present across all genotypes associated with plant environments [7]. Besides these enteric pathogens, various human and animal gut commensal bacteria can colonize plants as alternative hosts [7–9]. Therefore, uncovering the genetic mechanisms that underpin host switching from different perspectives will advance our understanding of microbial adaptive evolution.

Clostridium butyricum is a strictly anaerobic, Gram-positive, spore-forming bacillus known for its high butyric acid production [10]. Studies have shown that *C. butyricum* is present in 10–20% of human fecal samples and is among the earliest colonizers in infants, highlighting its role as an essential human gut symbiont [11, 12]. Several strains have excellent potential to benefit host health through several mechanisms [13]. One crucial mechanism for conferring health benefits is the short-chain fatty acids (SCFAs)-mediated homeostasis of the gut microbiota, although the precise pathways underlying these benefits are not fully understood [14]. Consequently, this bacterium has been used as a probiotic in Asia (mainly Japan, Korea, and China) for decades [15]. On the other hand, some *C. butyricum* strains detected in stool samples have also been implicated in pathological conditions such as botulism in infants and necrotizing enterocolitis in preterm neonates [16, 17]. In addition, distinct toxigenic *C. butyricum* strains have been implicated in foodborne botulism, where toxin production occurs in contaminated foods rather than via gut colonization. Recently, several studies have also explored differences among *C. butyricum* strains from the standpoint of functional genes associated with the benefit and toxic traits at the molecular level [18, 19]. Besides humans,

C. butyricum has been associated with the guts of various animals and is frequently found in the environment, including soil, fermentation products, freshwater, and marine sediments [20, 21].

Molecular phylogenetics based on the DNA sequences of bacterial strains found in different hosts is a typical way to analyze host-adaptive population structure. Traditionally, the clustering of clades corresponding to hosts was explained by natural selection in a given environment (ecotype), whereas the opposite was genetic drift (neutral diversification) [22]. However, genome sequence-based population genetics highlight the limitations of simple theoretical models and provide more complex evolutionary models to explain host-associated population structure, including neutral diversification with and without transmission barriers, ecotype model, divided genomes, introgression between strains, and negative frequency-dependent selection [23]. Furthermore, mutation and recombination could lead to DNA sequence variations that affect gene expression during transcription and translation processes [24]. Ultimately, these changes have implications for the ability of the host-associated organism to adapt. Thus, genome-wide population-level studies are powerful tools for revealing the genetic basis of host adaptation.

Despite the recent increase in reports of plant colonization of human gut microbes, the evolutionary mechanisms underlying the plant-associated lifestyles remain poorly understood. In this work, we isolated and sequenced a batch of endophytic *C. butyricum* strains (ECB) and collected 67 genome sequences of strains from other habitats (nECB) to reveal the evolutionary trajectory of its adaptation to the plant host. We performed the population genomic analysis to investigate the population structure and genetic diversity of these *C. butyricum* strains. Four genetically distinct lineages were detected, two of which were associated with endophytic *C. butyricum* strains and exhibited distinct evolutionary patterns. We explored the gene flow pattern of *C. butyricum* strains, suggesting unrestricted gene exchanges between ECB and nECB. We also highlighted recombination as a primary driver in shaping the population dynamics and diversification of *C. butyricum* strains. Furthermore, comparative genomics identified genes and pathways that may underlie habitat adaptation and link genomic and ecological variation.

Methods

Strain isolation and identification

Paris polyphylla var. *yunnanensis* is a critical medicinal plant species valued for its hemostatic, analgesic, and anti-inflammatory properties and is in colossal demand, mainly from the Chinese and Indian pharmaceutical and traditional medicine industries. The present healthy roots

of *P. polyphylla* var. *yunnanensis* were sampled from two different herbal medicine planting bases in Chenggong and Fengyi counties (Yunnan Province, southwest China). The anaerobic endophytes were isolated from surface-sterilized root tissues using the procedure previously described by Li et al. [25]. Genomic DNAs were extracted from endophytic strains using the Genomic DNA Rapid Isolation Kit for bacterial cells (BioDev-Tech, China) following the manufacturer's instructions. The 16S rRNA genes were amplified using the universal bacterial PCR primers 27F (5'-AGAGTTTGATCCTG GCT-3') and 1492R (5'-GGTTACCTTGTTACGACT T-3'). Sequencing of the 16S rRNA gene amplicons was performed using the BigDye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems) on an automated DNA analyzer system (PRISM 3730XL DNA analyzer, Applied Biosystems) at SinoGenoMax Corporation (Beijing, China). The 16S rRNA gene sequences were compared against those of recognized type strains using the EzTaxon server (<http://eztaxon-e.ezbiocloud.net/>) [26]. 55 strains identified as *C. butyricum* were selected for subsequent analysis.

Genome sequencing, assembly, and annotation

Purified genomic DNA was sent to Wuhan Benagen (Wuhan, China) for genome sequencing using an Illumina MiSeq system. Briefly, genomic DNA was randomly sheared using a Bioruptor ultrasonicator to construct libraries for paired-end (2×150 bp) sequencing. Low-quality reads were filtered from the sequencing datasets using Fastp v0.22.0 [27]. The remaining reads were de novo assembled using SPAdes v3.13.0 [28], and the contigs or scaffolds shorter than 1000 bp were excluded from the assembly. The 55 assembled genomes, together with an additional 67 genomes available in the NCBI GenBank Assembly database, were selected based on genome quality to obtain strain-level *C. butyricum* representatives (Table S1). We used CheckM v1.2.0 [29] and FastANI v1.33 [30] to estimate genome quality and select genomes with completeness ≥ 95%, contamination ≤ 5%, and average ANI values (relative to type strain DSM 10702^T) ≥ 95%. Gene prediction and functional annotation were conducted using Prokka v1.13 [31] with default options. Protein-coding genes were assigned Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) and Clusters of Orthologous Genes (COG) category by using EggNOG-mapper v2.1.8 [32] with “-evaluate 1e-5” option (eggNOG DB version: 5.0.2). The mobile genetic elements (plasmids and prophages) in genomes were identified using geNomad v1.8.0 [33].

Phylogenomic analysis and lineage definition

Predicted proteomes for all *C. butyricum* strains included in this study were inferred for orthology using

OrthoFinder v2.2.2 [34] with default parameters. The 1,383 single-copy orthogroups were aligned individually using Clustal Omega v1.2.4 [35]. Then, the resultant alignments were concatenated, and ambiguous sites were trimmed by trimAL v1.4.rev22 [36]. A maximum likelihood (ML) species tree was constructed of a concatenation using IQ-TREE v2.1.2 [37] with model selection by ModelFinder [38] and visualized by FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>) and iTOL (<https://itol.embl.de/>).

A genome-relatedness network was constructed from pairwise genomic ANI values, with genomes represented as nodes and the genome-relatedness as edges. Accordingly, the genome-relatedness between two genomes was quantified as the average ANI value between them and used as the edge weight in the network. After filtering edges with ANI values below 98.5%, four connected components were detected in the network, of which each was defined as a *C. butyricum* lineage. The network was visualized by Cytoscape v3.9.0 [39].

Uniform manifold approximation and projection (UMAP) ordination analysis was conducted to determine the grouping of *C. butyricum* isolates based on flexible gene sets (inferred by ROARY v3.12.0) [40]. In addition, using the pipelines, the phylogenomic tree of the genus *Clostridium* was constructed to examine the phylogenetic status of the 122 strains [41]. Briefly, all genomes of the type strains of the genus *Clostridium* by the NCBI assembly database (accession date, 24 March 2024) were downloaded from the NCBI genome FTP site (<https://www.ncbi.nlm.nih.gov/genomes>). After quality control, 108 genomes, together with the 122 *C. butyricum* genomes, were subjected to the PhyloPhlAn workflow v3.0.67 [42] to identify the set of 400 marker genes, which was then used to generate the phylogeny of the genus *Clostridium*.

Population diversity analysis

Demographic changes within each *C. butyricum* lineage were explored by calculating neutrality statistics. Tajima's D (TD) neutrality statistic was assessed using the package PopGenome v2.7.5 [43] in R. We used three genetic parameters to estimate the intra- and interpopulation genetic differentiation of *C. butyricum* lineages. We calculated nucleotide diversity (π), the fixation index (F_{ST}), and absolute divergence (D_{XY}) using the PopGenome package. All statistics were estimated from whole-genome alignment in overlapping 5 kb sliding windows with a 2.5 kb step.

Gene flow pattern and recombination estimation

All assembled *C. butyricum* genomes were provided as input to PopCOGenT [44] to determine population structure based on the gene flow. Herein, “gene flow” is inferred as homologous recombination events resulting

from horizontal genetic transfer (HGT) between strains, as opposed to vertical inheritance. It identifies gene flow between each pair of genomes by detecting regions of higher-than-expected similarity (termed length bias) under a null model of clonal descent [44]. Then, genomes connected by gene flow are grouped into populations, and clusters within populations are defined by genomes that share relatively higher gene flow [44]. Meanwhile, genomes with nucleotide diversity below 0.0355% should be classified into a single clonal complex, as insufficient mutations prevent the identification of gene flow. Genomic recombinations in *C. butyricum* isolates were analyzed using Gubbins v2.3.4 [45]. SNPs (single nucleotide polymorphisms) in whole-genome alignment were identified using the reference-based method Snippy v3.1 (<https://github.com/tseemann/snippy>), with the complete genome of strain JKY6D1 as the reference.

Endophytic relevant variants analysis

A comparative genome analysis was conducted between endophytic *C. butyricum* strains (ECB) and non-endophytic *C. butyricum* strains (nECB) to uncover the genomic variation of *C. butyricum* strains for endophytic adaptation. The annotated KOs (KEGG orthology) for all genomes were mapped to the corresponding prokaryotic pathways as defined by the KEGG database. Mann–Whitney *U* tests were used to assess enrichment and depletion of KEGG pathways and COG categories across the whole genome of ECB and nECB strains. The number of genes associated with a specific pathway or category was converted to the relative abundance. Notably, in the whole-genome comparison, pathways and COG categories were found to be significantly enriched in either ECB or nECB strains when the Benjamini–Hochberg-adjusted *P* value in the Mann–Whitney *U* test was less than 0.05, thereby ensuring robust statistical validation of these functional distinctions.

Pan-GWAS was performed using a presence/absence matrix of orthologous gene clusters generated by ROARY, where zero denoted absent gene clusters and one denoted present gene clusters. The association analyses were calculated using SCOARY v1.6.9 [46], supplied with the phylogenetic tree generated in this study. Associations were considered significant if the Bonferroni adjusted *P* < 0.05 and sensitivity and specificity were greater than 80%. All parameters for the prior bioinformatics software were set to default unless otherwise stated.

Results

Plant host-associated *C. butyricum* were mainly distributed in specific lineages

The genomes of 55 endophytic *C. butyricum* strains (ECB; isolates recovered from surface-sterilized internal root tissues were defined as the endophytes) isolated

from the root of *Paris polyphylla* var. *yunnanensis* were sequenced and subjected to phylogenomic analysis with 67 non-endophytic *C. butyricum* (nECB) genomes available from the public database. Detailed information for these 122 genomes is provided in Table S1. On average, the genomes of *C. butyricum* were 4.55 Mb in length and contained 4,104 predicted protein-coding genes. We also summarized mobile genetic element-associated sequences (putative plasmid contigs and predicted integrated prophage regions) across all genomes. The mean pairwise genome-wide average nucleotide identity (ANI) among 122 strains was 98.29%, with the lowest ANI of 97.14% observed between strain AGR2140 (lineage L1) and strain INCQS635 (lineage L4). The type strain, DSM 10702^T of *C. butyricum*, shared an average ANI of 98.74% with the other 121 strains. All these ANI values were greater than 95–96%, the threshold for demarcating prokaryotic species [47]. Furthermore, these strains formed a stable monophyletic cluster in the phylogenomic tree of the genus *Clostridium* (Fig. S1) and shared ANIs < 95% with other *Clostridium* species (data not shown). These results support the idea that the 122 strains form a cohesive unit descending from a last common ancestor.

To construct the phylogeny of *C. butyricum*, a concatenated alignment involving 1,383 single-copy core genes and covering 401,442 unambiguous amino acid sites was used to generate a maximum likelihood phylogenetic tree. Meanwhile, we devised a simple network clustering algorithm that identified well-delineated phylogenetic clusters of *C. butyricum* based on the ANI value (see Methods, Table S2). This algorithm extracted the connected components from the network where edges represent ANI values greater than or equal to 98.5%, and nodes represent each genome. Four robust lineages (L1–L4) were identified by both phylogenomic analysis (Fig. 1A) and network clustering analysis (Fig. 1B) despite the high genetic relatedness of *C. butyricum* strains. Moreover, examination of the flexible gene content (UMAP methods to cluster *C. butyricum* strains based on gene presence and absence) shows a pattern similar to the phylogeny of core genes (Fig. S2). The majority of novel endophytic isolates distributed into lineages L1 (32/37, 86.5%) and L2 (22/26, 85.6%), whereas non-endophytes mainly exhibited in lineages L3 (5/5, 100%) and L4 (53/54, 96.4%). Remarkably, the isolates from samples collected at two locations (Chenggong and Fengyi) were distributed across the three major phylogenetic branches, which contained only endophytes. These observations indicated that *C. butyricum* contained at least two phylogenetically close lineages associated with plant endophytic habitats, suggesting that its adaptation to plant hosts was a continuous evolutionary process rather than an isolated event.

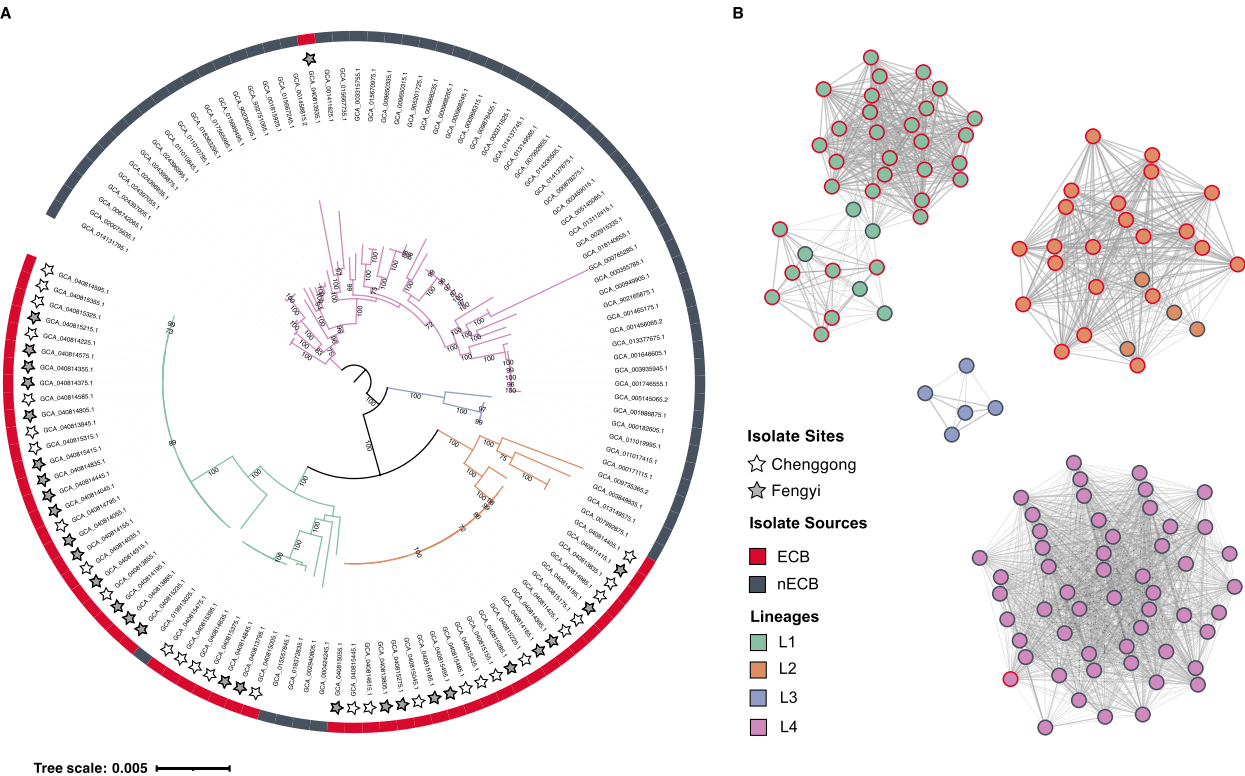


Fig. 1 Phylogenetic diversity of *Clostridium butyricum* strains. **A** The maximum-likelihood phylogeny of 122 assembled *C. butyricum* genomes based on concatenated alignments of 1,383 core genes. Bootstrap values above 70% were reported. Colored stars and rings outside the tree depict the isolation sites and sources. Colored corresponding branches on the trees indicated the division of lineages. **B** The genome network of 122 *C. butyricum* strains. Genomes are represented as nodes, with edges indicating the pairwise average nucleotide identity (ANI) between genomes (edges with ANI < 98.5% were filtered). The source data are provided in Table S2

Two ECB-related lineages experienced distinct evolutionary processes

To assess genetic diversity at the lineage level, we calculate the genome-wide nucleotide diversity (π), Tajima's D (TD), fixation index (F_{ST}), and absolute nucleotide diversity (D_{XY}) based on whole genome alignment. All four lineages showed some genetic divergence among individuals within each lineage. The estimates of mean nucleotide diversity (π) of each lineage range from 0.00094 (L2) to 0.00268 (L3) (Table S3). Furthermore, there were significant disparities (Wilcoxon test, $P < 2e-16$) in genetic diversity between L1 and L2, where ECBs are prevalent. L2 showed the lowest nucleotide diversity (mean $\pi = 0.00094$); L1 exhibited more than two times these levels (mean $\pi = 0.00202$). In addition, L1 and L4, which exhibited higher levels of intra-lineage nucleotide diversity, showed nearly zero TD values (-0.00496 for L1 and -0.10226 for L4). However, in L2 and L3 with lower intra-lineage nucleotide diversity, we observed the negative TD values (-1.43045 for L2 and -1.15956 for L3). Because these statistics are estimated at the lineage level and lineages include isolates from multiple sources, the observed patterns are compatible with differences in demographic history and population structure (e.g.,

recent expansion after a bottleneck and/or increased clonality), but they do not uniquely distinguish selection from demography.

The mean fixation indices (F_{ST} values) in all inter-lineage comparisons spanned 0.25 to 0.53 (Fig. 2). The wide range of calculated F_{ST} values implied that lineage pairs represent distinct stages across the divergence continuum of *C. butyricum*. The high genome-wide F_{ST} (> 0.25) values in all inter-lineage comparisons indicated discernible divergence between any two lineages. Notably, about 18% (268/1526) of regions with F_{ST} values equal to zero were identified in all lineage pairs. The SNP distribution pattern analysis revealed that the sites of these regions were invariant across all *C. butyricum* strains. Comparison of two ECB-related lineages (L1 and L2) revealed the overall pattern of high heterogeneity along the genome and less evident peaks of elevated divergence in the genome, suggesting strong genetic differentiation (Fig. 2). Moreover, compared to intra-lineage nucleotide diversity, absolute nucleotide diversity at the inter-lineage level (D_{XY}) between any two lineages was higher (Table S3), highlighting the lineage differentiation between *C. butyricum*.

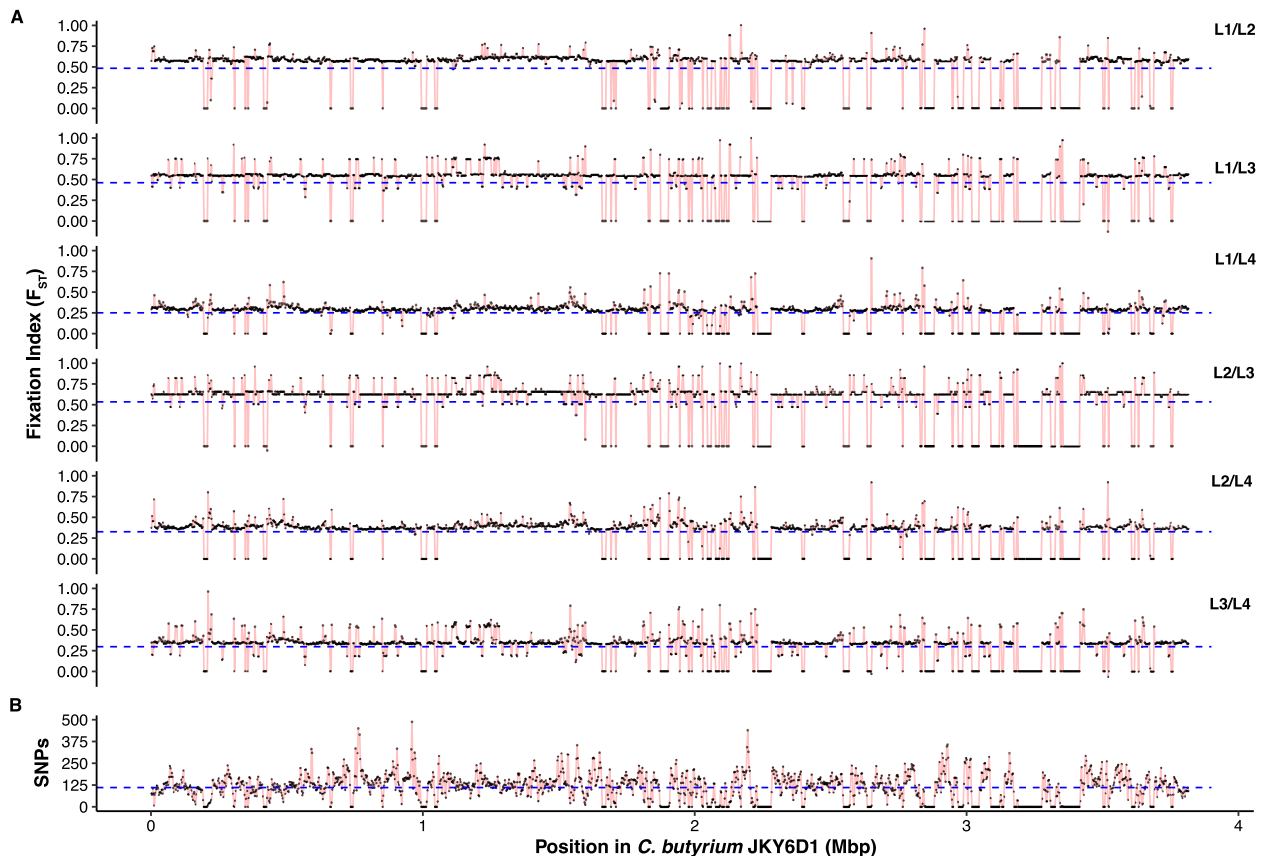


Fig. 2 The values of F_{ST} (A) and number of SNPs (B) within a 5-kb sliding window with a 2.5-kb overlap along the genome. The blue broken lines show the average number of SNPs or F_{ST} values per 5 kb

In summary, the two ECB-enriched lineages (L1 and L2) show markedly different diversity patterns. L1 has higher nucleotide diversity and Tajima's D close to zero, whereas L2 has lower diversity and a more negative Tajima's D (Table S3). These lineage-level statistics are consistent with distinct demographic histories and population structures (e.g., differing degrees of recent clonal expansion or bottlenecks), but they do not uniquely distinguish selection from demography given the mixed sampling sources within each lineage (Table S1). Moreover, the comparison of D_{XY} and F_{ST} across the whole genome highlighted the divergence between L1 and L2. Meanwhile, there was a higher divergence between ECB-related and nECB-related lineages. These observations implied that the two ECB-related lineages might have undergone different evolutionary patterns.

Evident gene flow existed between clonal complexes of ECB and nECB

To reveal more details about the differentiation of ECB-related and nECB-related lineages, we investigated gene flow patterns among *C. butyrium* strains. We employed PopCOGenT to characterize gene-flow discontinuities and delineate gene-flow-based clusters (genetically

isolated subpopulations). Based on 55 ECB and 67 nECB genomes, this approach identified four genetically isolated subpopulations (labeled from I to IV, Fig. 3A). In our dataset, lineage L1 is split into subpopulations I and II, whereas lineages L3 and L4 are grouped into subpopulation IV, reflecting stronger within-cluster gene-flow connectivity than between clusters. This pattern reflects PopCOGenT's gene-flow-based clustering, in which genomes show denser inferred recent homologous exchange within subpopulations than across subpopulation boundaries. Based on the identities among individual genomes, all 122 genomes were grouped into 45 clonal complexes (CCs; each CC represents a set of near-identical genomes collapsed for network construction) (Table S1). Of these 45 CCs, 33 were singletons (involved only one genome). Four subpopulations accommodated 4, 4, 5, and 32 CCs.

The division of these subpopulations basically corresponded to four phylogenetic lineages. Lineage L1 corresponded to subpopulations I and II; lineage L2 corresponded to subpopulation III; and lineages L3 and L4 constituted subpopulation IV. The structural differences between subpopulations L1 and L2 were consistent with their differences in genetic diversity and phylogenetic

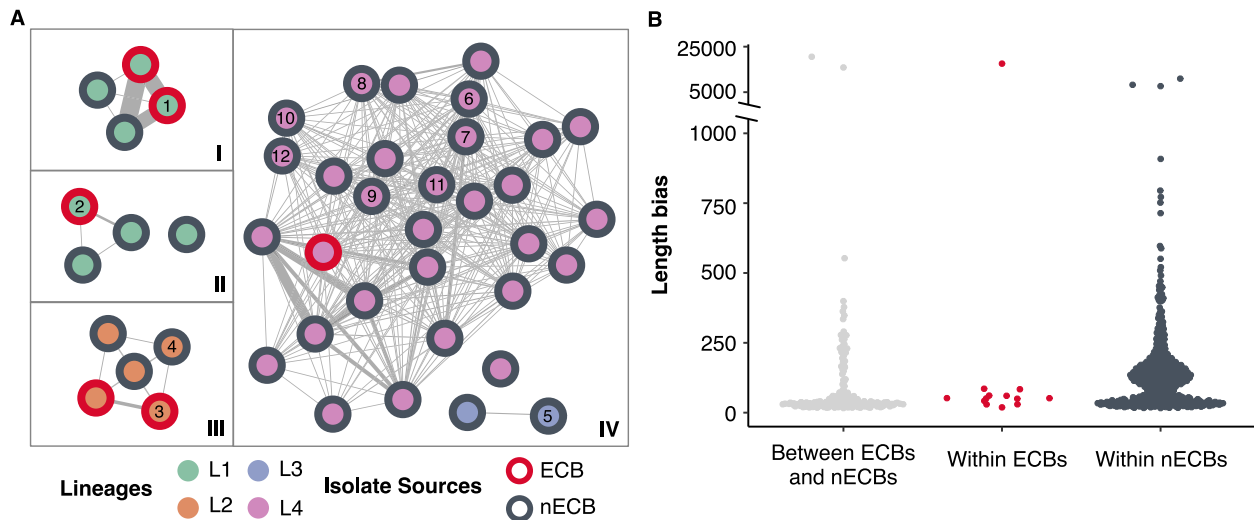


Fig. 3 Gene flow network and population structure identified by PopCOGenT. **A** Gene flow network. Edges indicate evidence of recent homologous DNA exchange (gene flow, consistent with homologous recombination) between any two genomes, quantified by the length-bias statistic (observed length-bias fraction); thicker edges indicate stronger inferred connectivity. Edges with a length bias fraction < 100 were filtered out. Genomes belonging to the same clonal complex (clonal complex consists of redundant members that do not contribute to the gene pool of a population) are collapsed into a single node. Node outlines are colored by isolation source, and nodes are colored by lineage assignment. The Arabic numerals labeled in the node represent the serial number of the clonal complex that contains multiple genomes. **B** The distribution of length bias fraction within ECB and nECB, as well as between ECB and nECB

divergence (Fig. 3). Moreover, the endophytic isolates from *Paris polyphylla* var. *yunnanensis* were mainly distributed in subpopulations I, II, and III. Because the endophytic genomes analyzed here are derived from a single host species and include clonal expansions, we interpret habitat–subpopulation patterns within the limits of this sampling. Frequent genetic exchanges have been observed between ECBs and nECBs (Fig. 3B). Especially for subpopulation I, the level of gene flow between ECB and nECB was even higher than that within the ECBs. In subpopulation IV, although there was only one ECB strain (YIM B08210), it still had a high rate of gene flow with other nECB members (Fig. 3A). Therefore, the rampant genetic fluidity between ECB and nECB genomes demonstrated that no gene flow boundaries existed between nECB and ECB strains, thus indicating the habitat switch of *C. butyricum* occurred recently. In other words, *C. butyricum* has not formed a distinct and universally unique ecotype within the plant endophytic environment. This suggests that its endophytic adaptation may stem from pre-existing genetic variation in the soil environment, with the plant host primarily acting to selectively expand these already-adapted clones rather than driving entirely new genetic differentiation.

Recombination played a crucial role in the population differentiation of *C. butyricum*

Frequent gene flow across the genomes of *C. butyricum* strains reveals evolutionary processes in the presence of recombination. Therefore, to conduct an overall and

contemporary recombination analysis, we employed Gubbins to infer recent (which affected only a subset of recipient lineages after lineage differentiation) and ancestral (which were shared by all strains within lineages before lineage differentiation) recombination events. Notably, we identified the 826 ancestral recombination events that contribute significantly to the genesis of four lineages (Fig. 4A). Out of these recombinations, 386 have occurred at the last common ancestors (LCAs) of each lineage, with the number of events being 67 (L1), 94 (L2), 102 (L3) and 126 (L4), respectively. The total recombinant fragment lengths of the core genome were 419 kb (~11%, L1), 498 kb (~13%, L2), 251 kb (~6%, L3), and 420 kb (~11%, L4), respectively (Table S4). In addition, more ancient recombinations were detected in the LCA of L3 and L4 (246) as well as in the LCA of L2, L3, and L4 (191) (Fig. 4A), with recombinant fragment lengths being 252 kb (~6%) and 334 kb (~9%). Ancestral recombinations occurred more frequently in lineages L3 and L4, where the nECB strains were dominant. However, this pattern should be interpreted cautiously for lineage L3, given its small number of non-redundant genomes and narrow sampling.

Recent recombination was a common phenomenon across four lineages and was much more frequent than ancestral recombination (Fig. 4A). Lineage L4 exhibited over 4000 recent recombination events. For 54 strains of L4, the average number of recent recombination events was 406 (one recent recombination event involving multiple strains will be counted multiple times when we

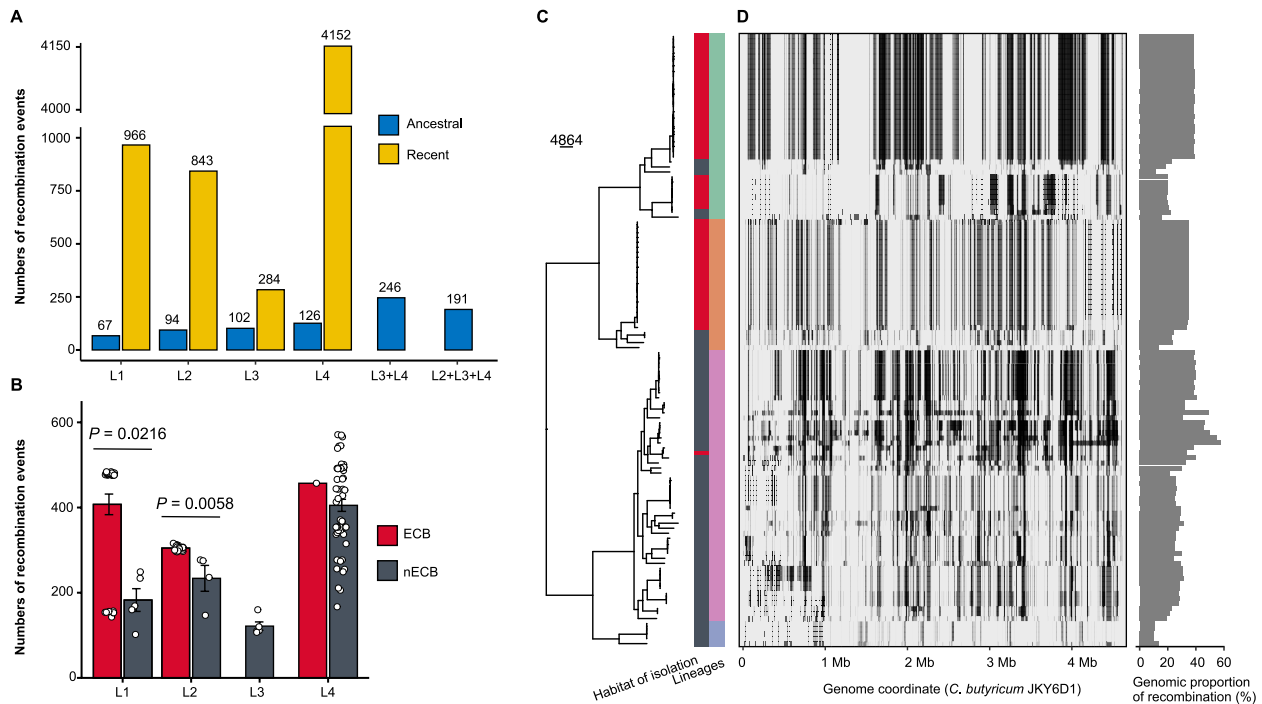


Fig. 4 Genomic recombination and its contribution to the divergence of *C. butyricum*. **A** Total frequencies of recent and ancestral recombinations within individual lineages and the common ancestors of corresponding lineages. **B** The frequency of recent recombinations between ECBs and nECBs. **C** The maximum-likelihood phylogeny of *C. butyricum* was inferred by the core genome alignments in which the recombinant mutations were excluded. The color strips outside the tree indicate the isolate source and lineages, as in Fig. 1. **D** Genomic pattern and proportion of recent recombinations for individual strains

calculate the average number of recent recombination events that occurred for each strain) (Table S4). However, the total (284) and average (122) number of recent recombination events in lineage L3, which exclusively consists of nECB strains, was much lower than that of L4 (Fig. 4B and Table S4). Within lineages L1 and L2, recent recombination events occurred frequently, with a total (average) number of 966 (377) and 843 (294), respectively. The frequency of recombination events that occurred in ECB strains was significantly higher than that in nECB strains in both lineages L1 and L2 (Fig. 4B). Furthermore, these recent recombinations had led to highly mosaic genomes, with the overall genomic fraction of recombinant segments varying from 10 to 57% amongst strains of different lineages (Fig. 4C and D). Interestingly, in each lineage, both unique recombinant sequences were seen in only one isolate, and shared recombinant sequences were seen in multiple isolates. It indicated that recombination events might occur at different stages during the diversification of lineages (Fig. 4D). Remarkably, removing the SNPs present in the recent recombinant regions significantly decreased the genetic distance between *C. butyricum* strains, demonstrating the importance of recombination in the diversification of *C. butyricum* (Fig. S3, Table S5).

Endophytic *C. butyricum* exhibited signs of functional divergence

To uncover the molecular mechanisms driving adaptation to the endophytic environment in *C. butyricum*, we evaluated the differentiation of functional gene composition between ECB and nECB strains. Analysis using the COG database revealed that the categories involved information storage and processing of translation, ribosomal structure and biogenesis (COG J), transcription (COG K), and metabolism of nucleotide transport and metabolism (COG F) were enriched in ECB strains (Fig. 5A and Table S6). Conversely, none of the COG categories were significantly higher in nECB strains than in ECB strains. On the other hand, KEGG annotation results showed that pathways related to metabolism (amino acid metabolism, glycerophospholipid metabolism, carbohydrate metabolism, and energy metabolism) and biofilm formation were significantly enriched in ECB strains (Fig. 5B and Table S7).

We also perform pangenome-wide association studies (pan-GWAS) using the Scoary algorithm to identify variants associated with endophytes, further revealing potential mechanisms of adaptation of *C. butyricum* strains to endophytic environments. This approach identified 72 genes that were positively or negatively associated with endophytic habitat, with adjusted *P* values < 0.05;

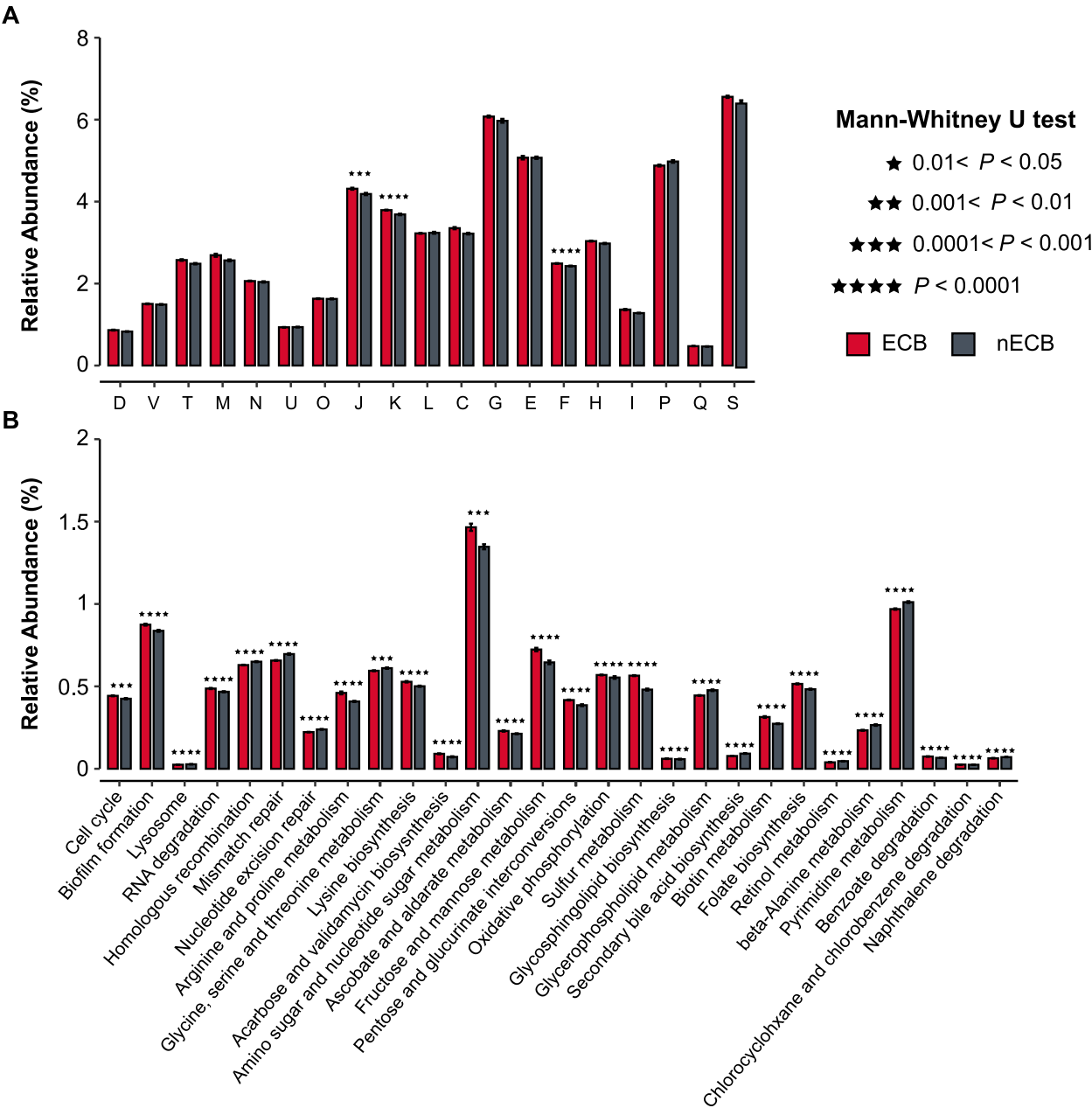


Fig. 5 Functional comparison among ECB and nECB strain genomes using COG categories (A) and KEGG pathways (B). The bar charts in the plot represent the relative abundance of COG categories or KEGG pathways. Bars represent mean ± SD. Statistical significance was evaluated using two-sided Mann–Whitney *U* tests on per-genome values with Benjamini–Hochberg correction (FDR)

47 were enriched in ECB strains, and the other 25 were enriched in nECB strains (Fig. 6A and Table S8). Most of the enriched genes in ECBs were undocumented. The remaining genes were distributed across 14 KEGG pathways, including defense systems, transporters, bacterial chemotaxis, bacterial motility, and the metabolism of folate, glycan, and nucleotide (Fig. 6A). Interestingly, two CRISPR-associated cas proteins (cas 5 and cas 8) were found to be associated with ECB strains. Moreover, the total length of the integrated prophages differed

significantly between the nECB and ECB genomes but not the plasmid (Fig. 6B). Given that CRISPR-associated proteins are under intense pressure in plant-associated microorganisms [48], it is suggested that bacterium–bacteriophage, interbacterial, and host–bacterium interactions might govern plant colonization of ECBs. Consistent with previous studies of plant-associated bacteria [49, 50], we found ECB-associated genes involved in environmental sensing and host/environment navigation (transporters, two-component systems,

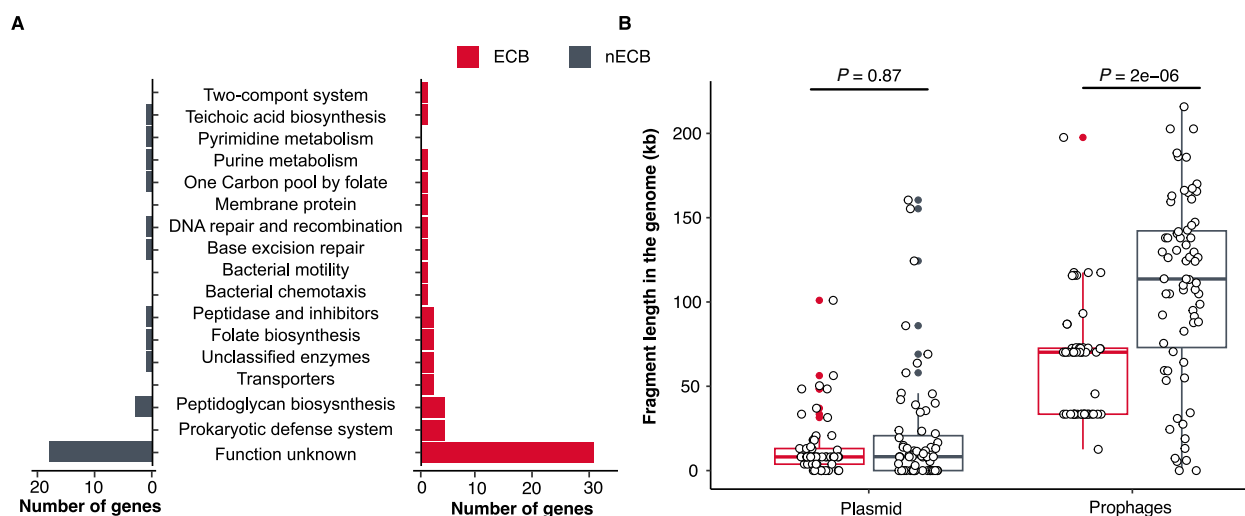


Fig. 6 Enrichment analysis of functional genes associated with ECB. **A** Pan-GWAS analysis results showing typical genes associated with endophytic environments; **B** The detection of mobile genetic elements among genomes of 122 *C. butyricum* strains

chemotaxis, and motility), which may contribute to recognizing root-derived cues and accessing microhabitats during colonization.

Discussion

Clostridium butyricum is widely distributed in soil and other environmental reservoirs and has been reported from diverse host- and food-associated settings, including silage, vegetables, cultured milk products, and human and animal gut [51]. Regarding plant-associated *C. butyricum*, an investigative study using pure culture techniques conducted in the United Kingdom found a high detection rate of *C. butyricum* in fresh and canned vegetables, especially potato skins [52]. Combining the currently available genomic data of *C. butyricum* with its anaerobic physiology, the studied isolates are predominantly from habitats associated with human and animal intestinal tracts. In this work, we examined the population genomics of *C. butyricum* isolated from the root of *Paris polyphylla* var. *yunnanensis* and compared them to public genomes from other environments. Four reliable lineages (L1–L4) of *C. butyricum* strains were identified based on core and accessory genome analysis and were broadly associated with habitats. The endophytic strains were mainly distributed in lineages L1 and L2. These two lineages contained plant endophytic strains and strains derived from other habitats, including human and animal gastrointestinal tracts (e.g., infant, poultry, swine), various environmental sources (e.g., soil, sludge, sediments), and industrial fermentation ecosystems (e.g., liquor pit mud, ABE fermenters). Given that these medicinal planting bases typically employ organic fertilization practices, the soil microbial community, including *C. butyricum*, may have been influenced by such amendments. It implied that these endophytic strains isolated

from the roots of *P. polyphylla* var. *yunnanensis* probably originated from the soil treated with fertilizer. However, the exact environmental source cannot be definitively determined from our current data. In addition, the relevant root samples were obtained from two geographically distinct planting sites. The phylogeny also indicates that the endophytic collections in L1 and L2 are dominated by one to two clonal expansions, consistent with local proliferation of a limited number of genotypes after entry into the plant environment. This pattern suggests that the divergence between L1 and L2 likely predates the endophytic lifestyle and may have arisen in a broader environmental reservoir such as agricultural soil, where opportunities for recombination and gene exchange are expected to be higher than within the endophytic niche.

Genetic diversity analysis revealed that four *C. butyricum* lineages were genetically distinct, and two ECB-enriched (L1 and L2) show contrasting lineage-level diversity patterns. At the lineage level, L1 displays higher nucleotide diversity (π) and Tajima's D close to zero, whereas L2 shows markedly lower π and a more negative Tajima's D. There are compatible with differences in demographic history and population structure—such as differing degrees of recent clonal expansion and/or bottlenecks—rather than uniquely indicating selection. Lineage L3, which contained toxigenic strains with infant botulism cases, showed lower π and a more negative Tajima's D. However, given the sampling bias, the genetic diversity in lineages L2 and L3 may not be fully represented, and thus diversity and recombination estimates for L3 are expected to be low and should be interpreted cautiously. More generally, “gut-associated” labels in available genomes largely reflect isolation source, and functional assays would be required to test whether gut environments impose consistent adaptive differences

relative to the endophytic niche. Despite these limitations, our data suggest that most genetic diversity in *C. butyricum* is generated in broader environmental reservoirs (e.g., soil and other mixed habitats), whereas the endophytic niche is characterized by the local expansion of a limited number of genotypes rather than the emergence of a deeply diverged, plant-restricted ecotype. In line with this, the ECB and nECB genomes remain connected by frequent gene flow signals, indicating limited long-term genetic isolation (Fig. 3).

Relative rates of mutation and recombination shaped the population genetic structure of bacteria [53, 54]. We found extensive recent and ancestral recombination events among strains of *C. butyricum*, thereby suggesting recombination as a major driver of population structures and subsequent diversification in *C. butyricum* (Fig. 4). Lineage L4 with higher genetic diversity ($\pi = 0.00268$) displayed more frequent recent recombinations (N_R , the average number of recent recombination events per genome, was 406); whereas lineage L3 with lower genetic diversity ($\pi = 0.00148$) displayed less frequent recent recombinations (N_R was 122). A higher frequency of recent recombination leads to greater genetic diversity. However, the frequency of recent recombination differed between lines L2 and L3, which also show low genetic diversity. In addition, the recent recombination occurred with relatively high frequency in lineages L1 ($N_R = 377$) and L2 ($N_R = 294$), but they showed a remarkable disparity in genetic diversity ($\pi = 0.00203$ in L1; $\pi = 0.00094$ in L2) (Figs. 2 and 4). It can be seen that the effect of recent recombination on genomic variation varies in different lineages. When we focused on endophytic clonal complexes within these lineages, we found that recombination frequencies varied, and the frequency of recent recombination among them differed significantly. Within lineage L1, the strains of CC2 exhibited the highest frequency of recombinations (mean number of recombination events = 478), with ~50% traced to their last common ancestor. In contrast, strains of CC1 showed the lowest frequency of recombinations (mean number of recombination events = 154), with only 12% derived from their LCA (Table S9). CC3, another clonal complex of ECBs in lineage L2, displayed 305 recombinations with ~30% derived from its LCA. Recombination had different effects on genome reshaping in other lineages and clone complexes under similar environmental pressures (Fig. 4D).

Plant-associated bacteria have evolved genes that allow them to thrive in plant environments [55]. To understand how genetic mechanisms have contributed to host adaptation, it is essential to identify adaptation features in bacterial genome sequences and to test their functional consequences [23]. In this study, by using the pan-GWAS analysis, several candidate genes (47

genes) predominantly in the ECB strains were identified to be significantly associated with *C. butyricum* endophytic lifestyle and perhaps contributed to their success in colonizing the root of *P. polyphylla* var. *yunnanensis*. Unfortunately, most of them (63.83%, 30 genes) were uncharacterized. The remaining well-documented genes were scattered in 14 pathways (Fig. 6A, Table S8), with the annotations related to the prokaryotic defense system, primarily modules involved in phage and mobile genetic element defense, whereas a small subset of hits mapped to peptidoglycan biosynthesis pathways (4 genes), likely reflecting variation in specific genes rather than differential presence of this conserved pathway. It was inconsistent with a recent large-scale comparative genomics analysis, which showed that plant-associated bacteria encode more carbohydrate-metabolism functions and fewer mobile elements than related non-plant-associated genomes [55]. This difference may reflect that our comparison is within a single species and based on root endophytes from one host plant, whereas the previous study summarizes broader taxa and plant-associated lifestyles. In addition, several ABC transporters, which play an essential role in bacterial immunity [56], were identified as positively associated with endophytes. Combined with the lower sensitivity to bacteriophages (Fig. 6B), it demonstrated that the ECB possesses uncommon mechanisms for endophytic adaptation and that the entire defense system may play a critical role. While our comparative and population genomic analyses identify candidate genetic features associated with the endophytic lifestyle, in planta colonization and competition experiments will be required to determine whether ECB strains indeed show a colonization advantage over nECB strains in *Paris polyphylla* var. *yunnanensis*.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-026-12876-8>.

Supplementary Material 1.

Authors' contributions

Cong-Jian Li: Methodology; investigation; data curation; writing – original draft; writing – review and editing; visualization; formal analysis. Yu Zhao: Data curation; writing – review and editing; visualization; formal analysis. Mei Tang: Methodology; investigation; formal analysis; writing – review and editing. Xiao Chu: Investigation; writing – review and editing. Peng-Chao Zhan: Investigation; formal analysis. Xing-Wang Jiang: Investigation; formal analysis. Jin-Yu Tian: Investigation. Xuan Hai: Investigation. Yi-Fei Lu: Investigation. Ling-Ling Yang: Conceptualization; supervision; writing – review and editing. Xiao-Yang Zhi: Conceptualization; supervision; writing – review and editing; project administration.

Funding

This research was supported by the National Natural Science Foundation of China (Grant Number: 32560005); Program for Excellent Young Talents, Yunnan University.

Data availability

All 55 endophytic *C. butyricum* genomes sequenced by this study have been submitted to the NCBI Assembly database under the bioproject No. PRJNA1136167. The accession numbers of 122 *C. butyricum* genomes (including 67 reference genomes) are listed in Supplementary Table S1. The analysis scripts supporting the findings of this study have been deposited in a public GitHub repository (<https://github.com/MSE-Lab/ECBscripts.git>) to ensure reproducibility.

Declarations

Competing interests

The authors declare no competing interests.

Received: 26 October 2025 / Accepted: 16 April 2026

Published online: 28 April 2026

References

- Brandl MT, Cox CE, Teplitski M. *Salmonella* interactions with plants and their associated microbiota. *Phytopathology*. 2013;103(4):316–25.
- Greene SK, Daly ER, Talbot EA, Demma LJ, Holzbauer S, Patel NJ, et al. Recurrent multistate outbreak of *Salmonella* Newport associated with tomatoes from contaminated fields, 2005. *Epidemiol Infect*. 2008;136(2):157–65.
- Teplitski M, de Moraes M. Of mice and men...and plants: comparative genomics of the dual lifestyles of enteric pathogens. *Trends Microbiol*. 2018;26(9):748–54.
- Baumler A, Fang FC. Host specificity of bacterial pathogens. *Cold Spring Harb Perspect Med*. 2013;3(12):a010041.
- Hazen TH, Zhao LC, Sahl JW, Robinson G, Harris AD, Rasko DA, et al. Characterization of *Klebsiella* sp. Strain 10982, a colonizer of humans that contains novel antibiotic resistance alleles and exhibits genetic similarities to plant and clinical *Klebsiella* isolates. *Antimicrob Agents Chemother*. 2014;58(4):1879–88.
- Fouts DE, Tyler HL, Deboy RT, Daugherty S, Ren QH, Badger JH, et al. Complete genome sequence of the N2-fixing broad host range endophyte *Klebsiella pneumoniae* 342 and virulence predictions verified in mice. *Plos Genet*. 2008;4(7):e1000141.
- Kirzinger MWB, Butz CJ, Stavrinides J. Inheritance of *Pantoea* type III secretion systems through both vertical and horizontal transfer. *Mol Genet Genomics*. 2015;290(6):2075–88.
- De Maayer P, Chan WY, Rubagotti E, Venter SN, Toth IK, Birch PRJ, et al. Analysis of the *Pantoea ananatis* pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. *BMC Genomics*. 2014;15:404.
- Vicente CSL, Nascimento FX, Barbosa P, Ke HM, Tsai IJ, Hirao T, et al. Evidence for an opportunistic and endophytic lifestyle of the *Bursaphelenchus xylophilus*-associated bacteria *Serratia marcescens* PWN146 isolated from wilting *Pinus pinaster*. *Microb Ecol*. 2016;72(3):669–81.
- Tipton L, Darcy JL, Hynson NA. A developing symbiosis: enabling cross-talk between ecologists and microbiome scientists. *Front Microbiol*. 2019;10:10.
- Finegold SM, Sutter VL, Mathisen GE. Normal indigenous intestinal flora. In: Hentges DJ, editor. Human intestinal microflora in health and disease. New York: Academic Press; 1983. p. 3–31.
- Mountzouris KC, McCartney AL, Gibson GR. Intestinal microflora of human infants and current trends for its nutritional modulation. *Brit J Nutr*. 2002;87(5):405–20.
- Guo PT, Zhang K, Ma X, He PL. *Clostridium* species as probiotics: potentials and challenges. *J Anim Sci Biotechnol*. 2020;11(20):24.
- Pan LL, Niu WY, Fang X, Liang WJ, Li HL, Chen W, et al. *Clostridium butyricum* strains suppress experimental acute pancreatitis by maintaining intestinal homeostasis. *Mol Nutr Food Res*. 2019;63(13):e1801419.
- Seki H, Shiohara M, Matsumura T, Miyagawa N, Tanaka M, Komiyama A, et al. Prevention of antibiotic-associated diarrhea in children by *Clostridium butyricum* MIYAIRI. *Pediatr Int*. 2003;45(1):86–90.
- Fenicia L, Anniballi F, Aureli P. Intestinal toxemia botulism in Italy, 1984–2005. *Eur J Clin Microbiol*. 2007;26(6):385–94.
- Howard FM, Flynn DM, Bradley JM, Noone P, Szawatkowski M. Outbreak of necrotising enterocolitis caused by *Clostridium butyricum*. *Lancet*. 1977;2(8048):1099–102.
- Pei Z, Liu Y, Yi Z, Liao J, Wang H, Zhang H, et al. Diversity within the species *Clostridium butyricum*: pan-genome, phylogeny, prophage, carbohydrate utilization, and antibiotic resistance. *J Appl Microbiol*. 2023. <https://doi.org/10.1093/jambio/xxad127>.
- Yang Y, Shao Y, Pei C, Liu Y, Zhang M, Zhu X, et al. Pangenome analyses of *Clostridium butyricum* provide insights into its genetic characteristics and industrial application. *Genomics*. 2024;116(3):110855.
- Cassin N, Benamar S, La Scola B. *Clostridium butyricum*: from beneficial to a new emerging pathogen. *Clin Microbiol Infect*. 2016;22(1):37–45.
- Yasueda A, Mizushima T, Nezu R, Sumi R, Tanaka M, Nishimura J, et al. The effect of *Clostridium butyricum* MIYAIRI on the prevention of pouchitis and alteration of the microbiota profile in patients with ulcerative colitis. *Surg Today*. 2016;46(8):939–49.
- Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, et al. Population genomics of early events in the ecological differentiation of bacteria. *Science*. 2012;336(6077):48–51.
- Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. *Nat Rev Genet*. 2018;19(9):549–65.
- Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. 2022;20(4):206–18.
- Li CJ, Zhang Z, Zhan PC, Lv AP, Li PP, Liu L, et al. Comparative genomic analysis and proposal of *Clostridium yunnanense* sp. nov., *Clostridium rhizosphaerae* sp. nov., and *Clostridium paridis* sp. nov., three novel *Clostridium sensu stricto* endophytes with diverse capabilities of acetic acid and ethanol production. *Anaerobe*. 2023;79:102686.
- Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Micro*. 2017;67(5):1613–7.
- Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–90.
- Bankovich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455–77.
- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25(7):1043–55.
- Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9(1):5114.
- Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068–9.
- Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;47(D1):D309–14.
- Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol*. 2023. <https://doi.org/10.1038/s41587-023-01953-y>.
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20(1):238.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972–3.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era (vol 37, pg 1530, 2020). *Mol Biol Evol*. 2020;37(8):2461–2461.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587–9.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–504.
- Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691–3.
- Li Y, Wang M, Sun ZZ, Xie BB. Comparative genomic insights into the taxonomic classification, diversity, and secondary metabolic potentials

- of *Kitasatospora*, a genus closely related to *Streptomyces*. *Front Microbiol.* 2021;12:683814.
42. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun.* 2020;11(1):2500.
 43. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol.* 2014;31(7):1929–36.
 44. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. A reverse ecology approach based on a biological definition of microbial populations. *Cell.* 2019;178(4):820–34.
 45. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 2015;43(3):e15.
 46. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 2016;17(1):238.
 47. Riesco R, Trujillo ME. Update on the proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Micr.* 2024;74(3):006300.
 48. Bulgarelli D, Garrido-Oter R, Munch PC, Weiman A, Droge J, Pan Y, et al. Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe.* 2015;17(3):392–403.
 49. Xu J, Zhang Y, Zhang P, Trivedi P, Riera N, Wang Y, et al. The structure and function of the global citrus rhizosphere microbiome. *Nat Commun.* 2018;9(1):4894.
 50. Knief C, Delmotte N, Chaffron S, Stark M, Innerebner G, Wassmann R, et al. Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J.* 2012;6(7):1378–90.
 51. Stoeva MK, Garcia-So J, Justice N, Myers J, Tyagi S, Nemchek M, et al. Butyrate-producing human gut symbiont, *Clostridium butyricum*, and its role in health and disease. *Gut Microbes.* 2021;13(1):1–28.
 52. Ghoddusi HB, Sherburn R. Preliminary study on the isolation of *Clostridium butyricum* strains from natural sources in the UK and screening the isolates for presence of the type E botulinum toxin gene. *Int J Food Microbiol.* 2010;142(1–2):202–6.
 53. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* 2009;3(2):199–208.
 54. Guttman DS, Dykhuizen DE. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science.* 1994;266(5189):1380–3.
 55. Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, et al. Genomic features of bacterial adaptation to plants. *Nat Genet.* 2017;50(1):138–50.
 56. Zheng G, Hehn R, Zuber P. Mutational analysis of the *sbo-alb* locus of *Bacillus subtilis*: identification of genes required for subtilisin production and immunity. *J Bacteriol.* 2000;182(11):3266–73.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.