PLOS ONE

RESEARCH ARTICLE

Comprehensive analysis of genomic variation, pan-genome and biosynthetic potential of *Corynebacterium glutamicum* strains

Md. Shahedur Rahman[1,2]*, **Md. Ebrahim Khalil Shimul**[1,2], **Md. Anowar Khasru Parvez**[3]

1 Department of Genetic Engineering and Biotechnology, Jashore University of Science and Technology, Jashore, Bangladesh, **2** Department of Genetic Engineering and Biotechnology, Bioinformatics and Microbial Biotechnology Laboratory, Jashore University of Science and Technology, Jashore, Bangladesh, **3** Department of Microbiology, Jahangirnagar University, Savar, Dhaka, Bangladesh

* ms.rahman@just.edu.bd

Citation: Rahman M.S, Shimul M.EK, Parvez M.AK (2024) Comprehensive analysis of genomic variation, pan-genome and biosynthetic potential of *Corynebacterium glutamicum* strains. PLoS ONE 19(5): e0299588. https://doi.org/10.1371/journal.pone.0299588

Editor: Paul Gladstone Livingstone, Cardiff's Metropolitan University: Cardiff Metropolitan University, UNITED KINGDOM

Received: May 3, 2023

Accepted: February 13, 2024

Published: May 8, 2024

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was financially supported by Information and Communication Technology Division (ICT Division) Bangladesh, Fund No.-24IF27334, received by SR. No additional external funding was received for this study. The funder had no role in study design, data collection and

Abstract

Corynebacterium glutamicum is a non-pathogenic species of the *Corynebacteriaceae* family. It has been broadly used in industrial biotechnology for the production of valuable products. Though it is widely accepted at the industrial level, knowledge about the genomic diversity of the strains is limited. Here, we investigated the comparative genomic features of the strains and pan-genomic characteristics. We also observed phylogenetic relationships among the strains based on average nucleotide identity (ANI). We found diversity between strains at the genomic and pan-genomic levels. Less than one-third of the *C. glutamicum* pan-genome consists of core genes and soft-core genes. Whereas, a large number of strain-specific genes covered about half of the total pan-genome. Besides, *C. glutamicum* pan-genome is open and expanding, which indicates the possible addition of new gene families to the pan-genome. We also investigated the distribution of biosynthetic gene clusters (BGCs) among the strains. We discovered slight variations of BGCs at the strain level. Several BGCs with the potential to express novel bioactive secondary metabolites have been identified. Therefore, by utilizing the characteristic advantages of *C. glutamicum*, different strains can be potential applicants for natural drug discovery.

1. Introduction

Corynebacterium glutamicum is a gram-positive, non-sporulating, non-pathogenic, and generally recognized as safe (GRAS) organism. It remains very robust against oxygen and substrate supply oscillation in the case of large-scale fermentations [1,2]. It is one of the most used microorganisms in industrial fermentation for producing amino acids, like lysine and glutamate, for decades [3,4]. *C. glutamicum* has undergone substantial modification to provide a wide range of beneficial products including chemicals, proteins, polymers, natural products, and biofuels [5–8]. Many studies of *C. glutamicum* have been published in the past decade [9], yet the genetic variations among the strains are unexplored.

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Whole genomes of closely related and geographically co-occurring microbial strains show enormous variation within species, resulting from allelic and gene content changes [10–13]. However, it is challenging to distinguish between two lineages that are thought to be the same species yet have significantly different gene contents using conventional taxonomic approaches [14–16]. Hence, a better understanding of the genomic characteristics of different *C. glutamicum* strains is required.

Genes for the production, control, and resistance of secondary metabolites are often grouped to create biosynthetic gene clusters (BGCs) in microbial genomes [17]. Utilization of bioinformatics tools for the analysis of microbial genome sequences reported that a single genome may include 20–80 distinct BGCs [18]. On the other hand, a microorganism may possess certain BGCs but it may not express them in laboratory conditions [19,20]. Research in this area will support wet lab methods development for natural product (NPs) producing strains that have greater potential to produce new compounds [18]. In 2017, Yang and Yang conducted a comparative analysis of *C. glutamicum* genomes, providing insights into the genetic diversity and evolutionary relationships within this significant industrial bacterium [21]. The research also pinpointed crucial mutations associated with amino acid production in various genetically engineered strains. However, certain limitations and challenges persist. Specifically, the pan-genome analysis was conducted with a relatively limited number of strains, potentially not encompassing the entire spectrum of species diversity. Furthermore, the identification of BGCs remains incomplete, highlighting areas for further investigation. So, it should be helpful to use functional genomic approaches to identify those unidentified BGCs at the genomic level. Therefore, the BGCs distribution and evolutionary connections among the *C. glutamicum* strains need to be explored. The primary aim of this study is to analyse pan-genomic variations within different strains and explore the distribution patterns of BGCs.

2. Materials and methods

2.1 Whole genome comparison

Genomic datasets of *C. glutamicum* strains were collected from National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/datasets, accessed on 30th May 2022). Initially, 65 complete genome sequences of *C. glutamicum* strains were retrieved in addition to the reference genome (In the NCBI database, *C. glutamicum* SCgG2 serves as the primary reference genome), all in FASTA format. The complete whole genome sequences of *C. glutamicum* were selected using the NCBI genome filter tool, and the assembly level was set to "complete". The choice of genomes was guided by contemporary research, emphasizing the pivotal role that high-quality genomes play in pangenome and genome mining analyses [22]. Consequently, this study excluded draft and scaffold level assemblies to ensure the integrity and reliability of the genomic data under examination. The use of complete genomes enhances the reliability and comprehensiveness of the study's findings, contributing to a more accurate understanding of the *C. glutamicum*'s genetic diversity, functional capabilities, and evolutionary insights. Then, whole genome comparisons were executed using OrthoANI v0.5.0 with default parameters, which uses an enhanced pairwise average nucleotide identity (ANI) algorithm [23]. After the comparison, we selected 30 complete genomes, other 35 were discarded due to 100% similarity match. The program was also used to clear species boundaries and to get diversity at the genetic level among whole genomes (Table 1). In this way, redundancy was avoided and the genetic diversity of *C. glutamicum* was ensured.

Table 1. List of *C. glutamicum* strains used in this study with their metadata.

Accession no.	Strains	Source	Location	Reference
CP012194.1	CP	air	China	[24]
CP018175.1	XV	soil	China	[25]
AP017557.2	AJ1511		Japan	[26]
AP022856.1	ATCC 21799		Japan	[27]
CP004062.1	ZL-6	soil	China	
CP010451.1	B253		China	[28]
CP053188.1	BE		South Korea	
CP022614.1	ATCC 14067	rotten onion	Guangzhou, Guangdong, China	
CP014984.1	YI	soil	China	[25]
NC_009342.1	R			[29]
CP004047.1	SCgG1	soil	China	
CP004048.1	SCgG2	soil	China	
CP013991.1	USDA-ARS-USMARC-56828	nasopharynx of calf	Tennessee, USA	
NZ_CP073911.1	CGMCC1.15647		Jiangsu, China	[30]
CP068290.1	ATCC 21573			
CP012297.1	B414	soil		
CP012298.1	CICC10064	soil		
CP005959.1	MB001			[31]
CP020658.1	TQ2223	soil	Tianjin, China	
CP025533.1	ATCC 13032		South Korea	[32]
CP025534.1	HA		South Korea	[32]
NZ_CP059382.1	BCA		Portugal	[33]
CP017995.1	C1	engineered derivative of C. glutamicum ATCC 13032	Germany	[34]
CP080542.1	CR101	lab strain	Bielefeld University, Germany	[35]
CP041729.1	JH41		Daejeon, South Korea	[36]
CP007724.1	AR1		Daejeon, South Korea	[37]
CP007722.1	ATCC 21831			[37]
CP020033.1	TCCC11822	soil	Tianjin, China	
CP016335.1	ATCC 13869	soil	Japan	[38]
NZ_CP022394.1	WM001	soil	Wuxi, Jiangsu, China	[39]

https://doi.org/10.1371/journal.pone.0299588.t001

2.2 Genome annotation

The process of locating and designating all the pertinent features on a genomic sequence is known as genome annotation [40]. Selected whole genome sequences were re-annotated using Prokka v1.14.6 with default parameters [41]. Prokka uses BLAST+ and identifies best match of annotated protein and candidate genes from various databases [41]. Prokka and FragGeneScan v1.31 were used with default parameters to identify the number of genes in each genome [42]. It uses a novel gene prediction technique and improved prediction of the protein-coding region in short reads by combining codon usages and sequencing error models in a Hidden Markov Model (HMM) [42].

2.3 Pan-genome analysis

Pan-genomic analysis was conducted utilizing Roary v3.11.2 (with default parameters), a robust computational tool specifically designed for such analyses. Roary classifies genes into distinct categories, including 'core genes', 'cloud genes', 'shell genes', and 'soft-core genes', employing a rigorous computational framework [43]. Bacterial Pan-genome Analysis tool

(BPGA v1.3) [44] was employed for the systematic classification of orthologous genes into core, accessory, and unique genomes. Subsequently, strains containing a relatively higher number of unique genes were subjected to annotation using the blast algorithm against the Clusters of Orthologous Genes (COG) database [45]. To gain in-depth insights into the functional aspects of these genes, further analyses were conducted utilizing the blast algorithm against both the COG and Kyoto Encyclopedia of Genes and Genomes (KEGG) database [46]. The estimation of the pan-genome and core genome was performed using the USEARCH v11.0.667 [47] program available in BPGA, employing a 50% sequence identity cut-off. The resulting data were then subjected to nonlinear fitting based on the model extrapolation of the pan-genome and core genome, ensuring a robust and comprehensive analysis of the bacterial genomic elements under investigation [44,48].

2.4 Phylogeny

FastTree v2.1.11 (with default parameters) was used to generate phylogenetic tree, which uses the maximum-likelihood method with generalized time-reversible (GTR) models of nucleotide evolution [49]. iTOL, an online platform was used to visualize the phylogenetic tree [50].

2.5 Identification of BGCs

We used three platforms to predict BGCs, which can accurately predict microbial secondary metabolite encoding regions by using sophisticated computer model services [51]. These are namely antiSMASH 6 (https://antismash.secondarymetabolites.org/, accessed on 9th, June 2022) [52], PRISM 4 (http://prism.adapsyn.com, accessed, accessed on 28th, June 2022) [53] and BAGEL4 (http://bagel4.molgenrug.nl, accessed on 29th, June 2022) [54]. BGC boundaries in this study was detected using antiSMASH 6, a computational tool that employs several techniques. Firstly, antiSMASH determines BGC boundaries based on the physical distance to core domains within the analyzed sequences [55]. It utilizes ClusterCompare output, conducting a search of all gene products against a database comprising highly conserved enzyme Hidden Markov Model (HMM) profiles indicative of specific BGC types [56]. The tool applies pre-defined cluster rules to identify individual protoclusters encoded in the genomic region. To standardize gene locations, antiSMASH employs a reference genome as a common coordinate system, allowing for the normalization of gene positions. Additionally, antiSMASH maps genomes of other strains containing the same or similar BGCs to the reference genome through alignment tools. This enables the identification and comparison of genomic regions corresponding to the BGCs across different strains in relation to the reference genome [52]. PRISM 4 predicts BGCs by analysing open reading frames from various databases [53]. BAGEL4 identifies ribosomally synthesized and post-translationally modified peptides (RiPPs), and Bacteriocin. It discovers gene clusters by using peptide database and/or through HMM motifs that are present in relevant contextual genes, augmented with literature references and links to UniProt and NCBI [54].

2.6 Genomic analysis and single nucleotide polymorphism identification

Genome comparisons among *C. glutamicum* strains were conducted using BLAST Ring Image Generator (BRIG-0.95-dist) with default settings. BRIG plays a pivotal role in facilitating the assessment of genotypic distinctions within closely related prokaryotic organisms [57]. In this study, we utilized the Mauve genome alignment system to analyze *C. glutamicum* strains [58]. Throughout evolution, microbial genomes can experience substantial mutations, including rearrangements and lateral transfers, leading to notable differences in gene order and content among closely related organisms. Mauve, a powerful tool, was employed to identify these

events, enabling comprehensive comparisons of multiple microbial genomes, even in the presence of high recombination rates. The Mauve system was configured with default settings, employing deed weight, full alignment, and iterative refinement techniques.

In our study, we utilized single nucleotide polymorphism (SNP) analysis as a methodology to discern genetic variations within the strains of *C. glutamicum*. The identification of variants among these strains was conducted through the implementation of Snippy v4.6.0 [59], with the reference sequence being *C. glutamicum* SCgG2. Notably, the prediction of Core SNPs was an additional aspect addressed in our analysis, employing the same Snippy tool for this specific task. This comprehensive approach allowed for a detailed exploration of genetic diversity and core variations within the *C. glutamicum* strains under investigation.

2.7 Identification of horizontal gene transfer

The prediction of horizontally transferred genes was carried out using HGTector v2.0b3 (with default settings) [60]. The analysis focused on identifying horizontal gene transfer (HGT) events within *C. glutamicum* AJ1511 and *C. glutamicum* AR1 genomes. A search was conducted utilizing the default remote database with stringent criteria, requiring a minimum identity and coverage of greater than 50%. The analysis was executed with default parameters to ensure comprehensive and accurate detection of potential HGT events in the studied strains.

2.8 Pathogenic and non-pathogenic properties and plasmid typing

The prediction of pathogenicity for the chosen strains was carried out using the PathogenFinder web tool [61]. This tool employs a predictive model that considers both the probability score and the resemblance to known pathogenic species in order to assess the likelihood of pathogenicity.

The plasmid sequences were obtained from the NCBI database. To ascertain the classification of plasmids, Plasmid Multi-Locus Sequence Typing (Plasmid MLST) was employed. Plasmid MLST is a molecular typing method that analyzes specific genetic markers across plasmids, providing insights into their type and lineage. This approach aids in categorizing plasmids based on their sequence diversity and assists in understanding the genetic variation and relationships among different plasmid strains.

3. Results

Demographic information about the strains used in this study are listed in **Table 1**. Among the strains, 11 were isolated from soil and others were isolated from air, mucus, rotten onion and lab strains. Among them 20 strains were isolated from Asian countries, 2 strains were isolated from Germany, and 1 strain from Portugal and United States of America each. Others origin are unknown.

3.1 Whole genome comparison

The degree of relatedness in the studied strains were identified by calculating ANI. ANI also clarifies whether the genomes reside in the same species by a cut-off values of $\geq$ 95% for same species. Our studied genomes have shown higher than 97% ANI values, confirming that all the genomes of the strains belong to the same species (**S1 Table**). A heat-map generated from the ANI scores have shown (**Fig 1**). There are five sub-groups in the heat-map and can be called as five clades. The clades were extracted from pairwise ANIs by using a hierarchical clustering algorithm with a cut-off value of 0.5. This means that strains with ANI values higher than 0.5

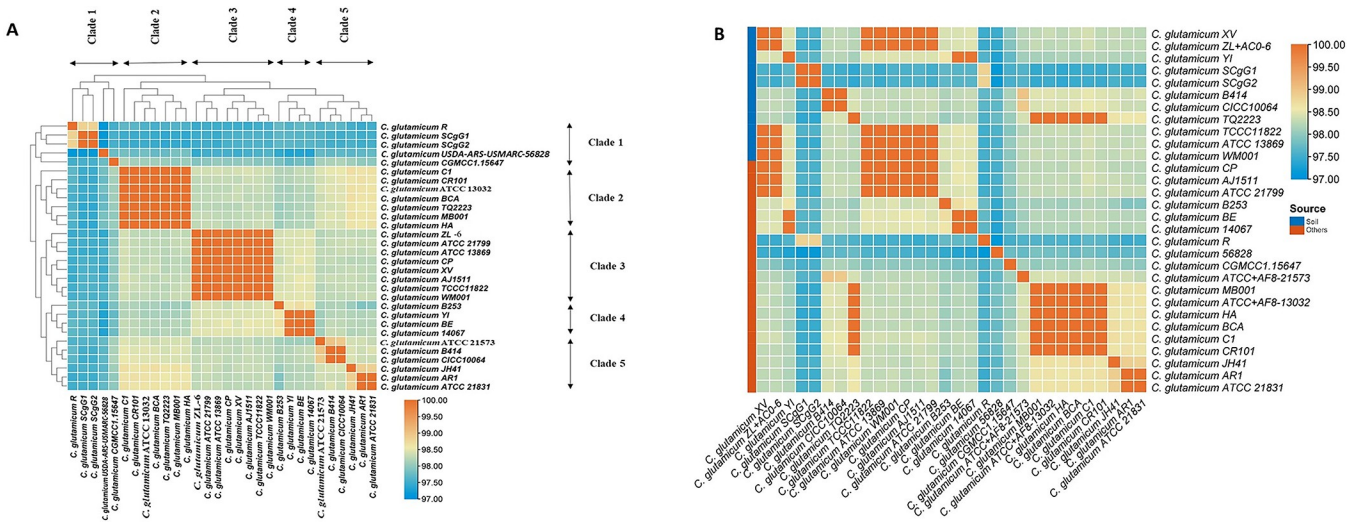

Fig 1. (A) ANI based whole genome comparison of *C. glutamicum* strains. The linkage method was average linkage, which calculates the average distance between all pairs of points in two clusters. The distance metric was Euclidean distance, which measures the straight-line distance between two points in a multidimensional space. The distance threshold was 0.5, which means that clusters with a distance less than or equal to 0.5 were merged together. This resulted in five clades, as shown by the horizontal dashed line in the plot. (B) ANI comparisons conducted among strains isolated from soil environments and strains isolated from both soil and non-soil environments within *C. glutamicum*.

https://doi.org/10.1371/journal.pone.0299588.g001

were grouped together in the same clade. The clades do not seem to have a strong correlation with the source and geographic location of the genomes. For example, clade 1 contains strains from soil and mucus sources, and from China and USA locations. Clade 2 contains strains from soil sources, and from Germany, China, South Korea, and Portugal locations. Clade 3 contains strains from soil, air and lab sources, and from China and Japan locations. Clade 4 contains strains from soil and rotten onion sources, and from China and South Korea locations. Clade 5 contains strains from soil sources, and from South Korea locations. Strains belonging to clade 1 (R, SCgG1, SCgG2) exhibit big genome size, a notable presence of multiple copies of NAPAA biosynthetic gene clusters (BGCs) and concurrently possess betalactone BGCs. This characteristic occurrence may contribute to their distinctiveness as outliers within the broader spectrum of analyzed genomes.

Fig 1B represents the ANI comparisons among various *C. glutamicum* strains isolated from different sources, including soil and non-soil environments. The ANI values revealed significant insights into the genetic relationships among these strains, shedding light on the impact of isolation sources on their genetic similarity. When we performed a more detailed ANI analysis, we observed that the strains isolated from soil environments, such as *C. glutamicum* XV, *C. glutamicum* ZL, and *C. glutamicum* YI, exhibited ANI values close to 98%, indicating a high genetic similarity. This suggests a common genetic background among these soil-isolated strains. On the other hand, when comparing soil-isolated strains with those from non-soil sources, the ANI values were notably lower, hovering around 97%. This discrepancy underscores the genetic divergence between strains from soil and non-soil origins. Such divergence could potentially be attributed to environmental factors and selective pressures specific to these habitats, leading to genetic adaptations unique to each niche.

3.2 Comparative genomic features of *C. glutamicum* strains

The average genome size of *C. glutamicum* strains was 3.24 Mbp (ranging from 2.84 Mbp to 3.36 Mbp) (**Fig 2A and S2 Table**). Coding sequence (CDS) count was predicted with the

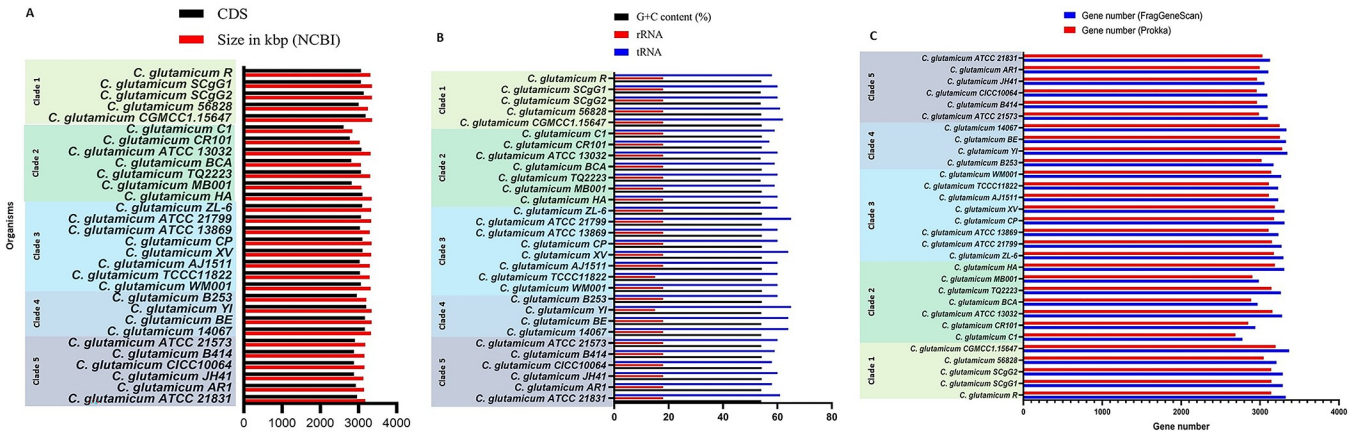

Fig 2. Overview of genomic features. (A) Genome size and CDS. (B) Genomic features (tRNA, rRNA and GC content (%)). (C) Gene number.

https://doi.org/10.1371/journal.pone.0299588.g002

highest 3200 CDS to lowest 2610 CDS with a mean of 3007 CDS among the whole genomes (Fig 2A **and** S2 Table). The average GC content was 54.15% among the genomes, and the approximate number of tRNA genes ranged from 57 to 65, while the predicted rRNA genes were 18 among all 28 strains excluding strain TCCC11822 and strain YI having 15 rRNA genes (Fig 2B **and** S2 Table).

The gene count, as determined by Prokka, displayed a range of 2688 to 3281 genes, with a mean value of 3078 genes per genome. In contrast, gene predictions through FragGeneScan exhibited a range of 2778 to 3369 genes, yielding a mean of 3197 genes per genome. It is noteworthy that Prokka's predictions resulted in a comparatively lower gene count than those obtained via FragGeneScan **(**Fig 2C **and** S2 Table).

3.3 Pan-genome analysis

Roary analysis predicted total 6854 protein-coding gene sequences. The number of core genes was 29% (99% < = strains < = 100%), the number of soft core genes was 2.86% (95% < = strains < 99%), the number of shell genes was 20.27% (15% < = strains < 95%), and the number of cloud genes was 47.78% (0% < = strains < 15%) (Fig 3A).

The high number of cloud genes exhibit significant variation and shows the 'open' nature of the *C. glutamicum* pan-genome (Fig 3B). The pan-genome of *C. glutamicum* was analysed using an empirical power law regression function based on the Allometric1 model ($f(x) = 3059.17x^{0.136303}$). The obtained parameter exponent (0.136303), falling between 0 and 1 and indicates that the pan-genome grows more slowly than other bacteria (possibly due to slower genetic diversification), but will grow indefinitely nonetheless (Fig 3C). In the context of Heaps' law, an 'open' pan-genome suggests the presence of a substantial and indeterminate number of additional genes, with its size potentially increasing boundlessly as more strains are included in the analysis [62–64]. *C. glutamicum* strains TQ2223, ATCC 13032, and HA exhibit a relatively low GC content coupled with a notable abundance of accessory genes (883, 925, and 924, respectively). Among these strains, *C. glutamicum* SCgG2 displays the lowest GC content and concurrently possesses the highest number of accessory genes (Fig 3D). In Fig 4A and 4B, the distribution of COG and KEGG categories for core, accessory, and unique genes is illustrated. Fig 4C displays the phylogenetic relationships among *C. glutamicum* strains based on core genes.

The core genome is primarily associated with essential biological functions such as amino acid transport and metabolism, translation, ribosomal structure and biogenesis, transcription,

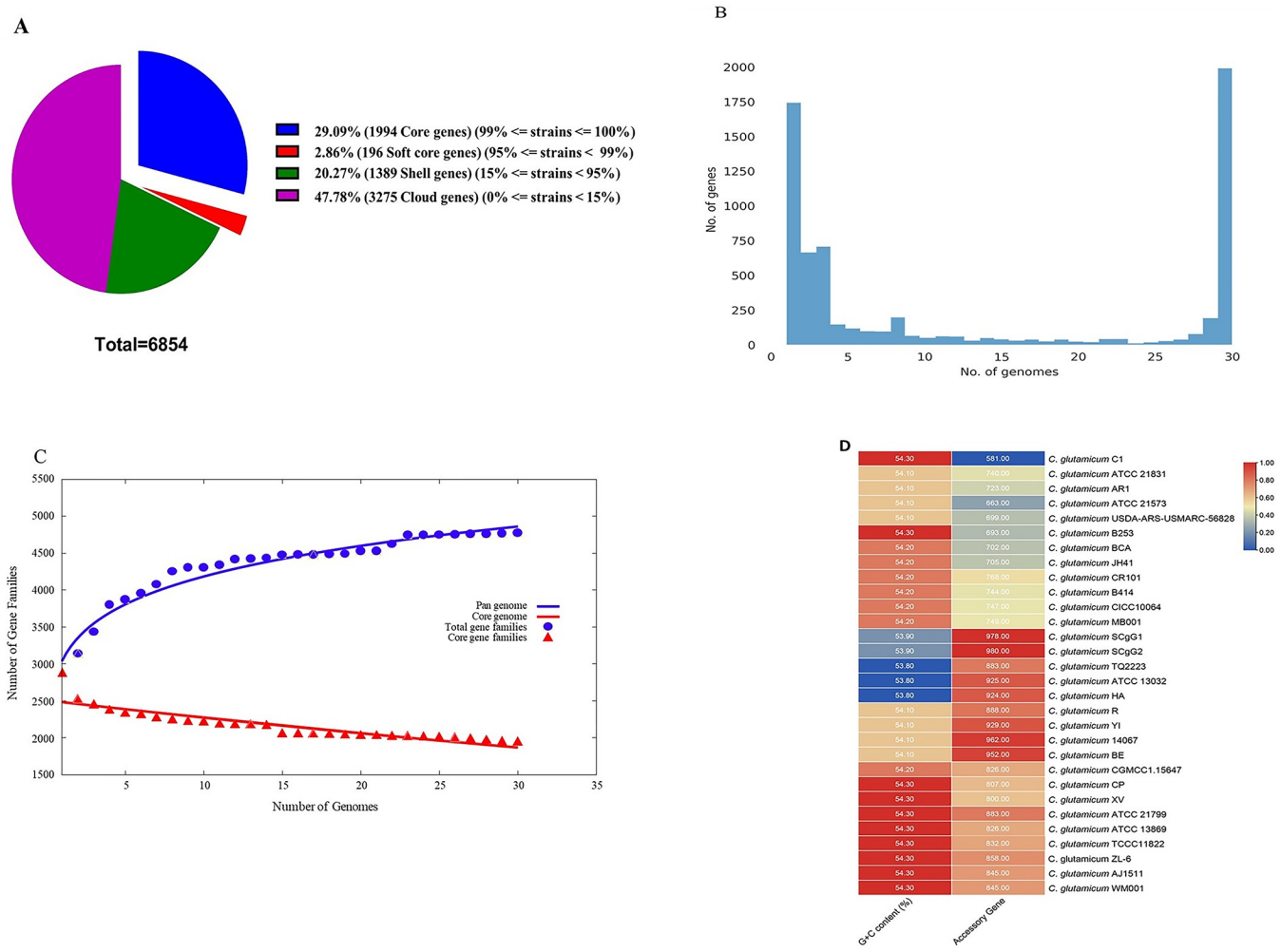

Fig 3. Pan-genome analysis of *C. glutamicum*. (A) The number of core genes, soft core genes, shell genes and cloud genes in the pan genome. (B) Gene frequency versus genomes number. (C) The pan genome profile trends obtained using BPGA v1.3. (D) Genomic G+C content (%) and accessory gene counts in various *C. glutamicum* strains.

https://doi.org/10.1371/journal.pone.0299588.g003

carbohydrate transport and metabolism, inorganic ion transport and metabolism, and post-translational modification, protein turnover, and chaperones. Simultaneously, the number of unique genes within *C. glutamicum* genomes varied significantly, indicating individual differences and a relatively high level of genomic diversity. This variability suggests their potential adaptation to diverse and extreme environments. Furthermore, KEGG pathway analysis revealed that these unique genes are involved in various biological processes related to metabolism, environmental information processing, and cellular processes.

3.4 Diversity and abundance of potential BGCs

AntiSMASH prediction identified six different classes of BGCs among the whole genomes. Identified BGCs include terpene, non-alpha poly-amino acids like e-polylysin (NAPAA), Beta-lactone, type 1 polyketide synthase (T1PKS), other unspecified ribosomally synthesized and post-translationally modified peptide product cluster (RiPP-like), and lanthipeptide class IV. Terpene synthesis BGCs were the most abundant BGCs in the genomes. NAPAA and T1PKS were the second most abundant BGCs, and collectively these 3 BGCS (Terpene, NAPAA, and

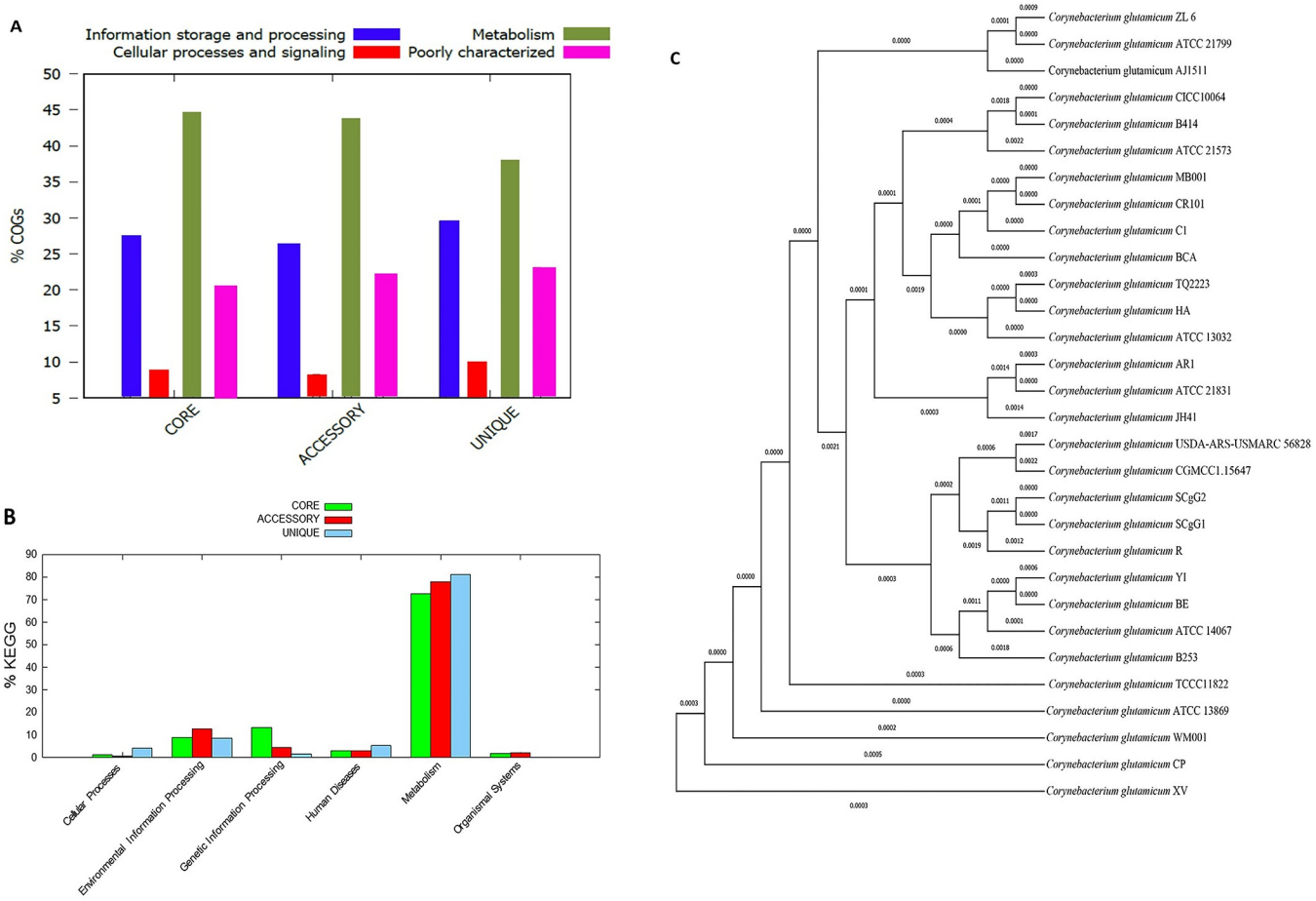

Fig 4. (A) COG distribution of core, accessory and unique genes. (B) KEGG distribution of core, accessory and unique genes. (C) Phylogenetic analysis of *C. glutamicum* strains based on core genes.

https://doi.org/10.1371/journal.pone.0299588.g004

T1PKS) comprised over 87% of the BGCs among the 30 strains of *C. glutamicum* (**Fig 5 and S3 Table**).

The strains harbour 2 to 6 BGCs and maximum 21 strains harbour 4 BGCs. The highest 6 BGCs were found in strain B253 and the lowest 2 BGCs in strain C1. Besides, 6 strains which are strain BE, 14067, YI, R, SCgG1, and SCgG2 contain 5 BGCs in their genomes (**Fig 5B and 5C**). Betalactone and lanthipeptide class IV BGCs were the most common BGCs. Betalactone BGCs were predicted in strain R, B253, SCgG1, and SCgG2, while lanthipeptide class IV BGCs were only found in strain B253 (**Fig 5C and S4 Table**). *C. glutamicum* strains have also been found harbouring double copies of same BGCs class in 20 strains. Terpene class BGCs were observed to be duplicated in up to 13 strains, exemplified by strain AR1, which possesses two distinct terpene class BGCs (identified as CGLAR1_11505 and CGLAR1_03580). Both clusters encode phytoene synthase, yet their gene products differ in size, with lengths of 287 and 304 amino acids, respectively (**Fig 5D**). In contrast, NAPAA class BGCs were identified as duplicated entities in 7 strains, as illustrated in Fig 5C and detailed in S4 Table. The antiSMASH analysis of *C. glutamicum* strains identified terpene BGCs across different clades, demonstrating considerable diversity in encoded compounds. In Clade 1, strains like CGMCC1.15647, USDA-ARS-USMARC-56828, R, SCgG2, and SCgG1 produced phytoene/squalene synthase. CGMCC1.15647 exhibited multiple copies of phytoene/squalene synthase genes in different

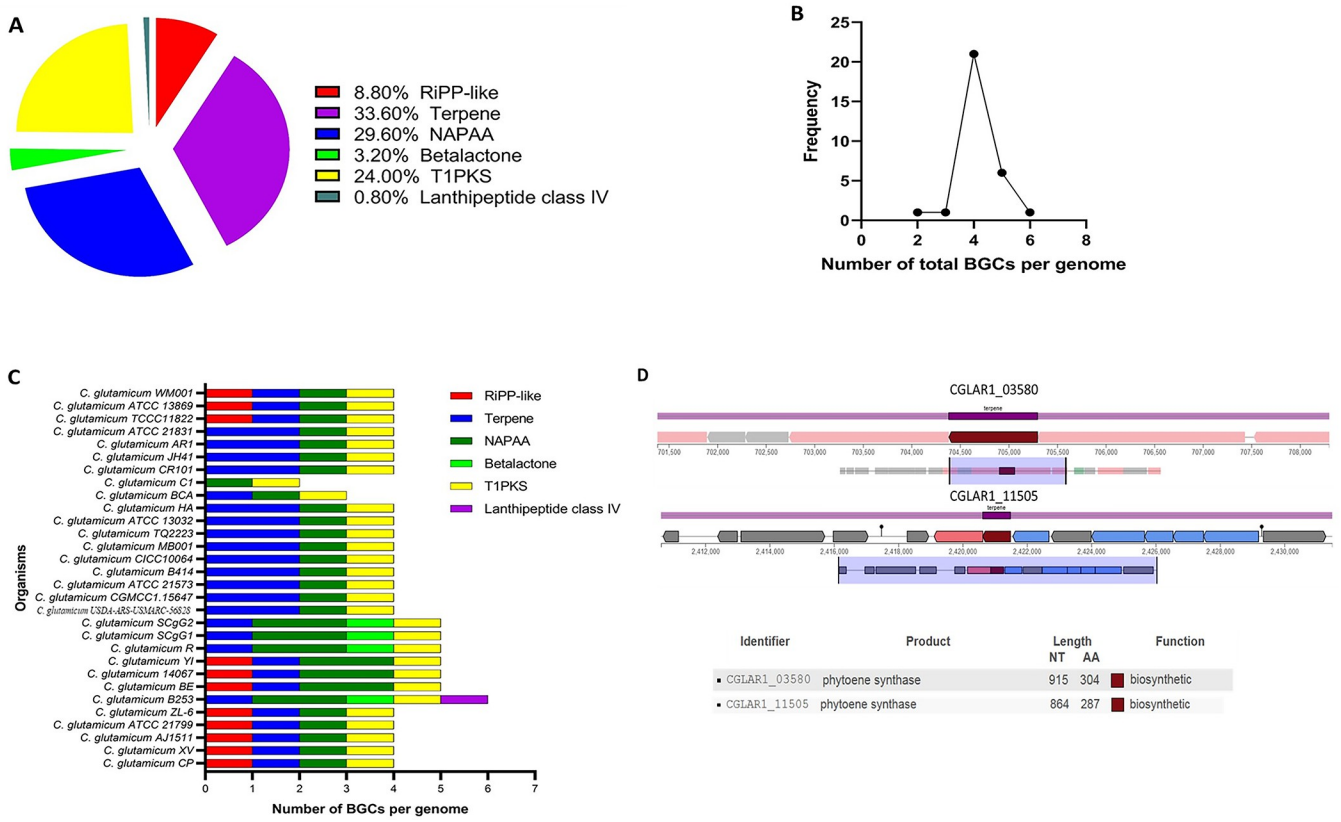

Fig 5. BGCs among *C. glutamicum* strains. (A) Distribution of different classes of BGCs among *C. glutamicum* strains. (B) BGCs frequency per genome. (C) Different classes of BGCs occurrence in the genomes. (D) BGCs of *C. glutamicum* AR1.

https://doi.org/10.1371/journal.pone.0299588.g005

genome regions, indicating intra-strain variability. Clade 2 strains, including ATCC_13032, HA, TQ2223, MB001, CR101, and BCA, showed varying similarity scores for phytoene synthase, highlighting potential enzyme differences. Clade 3 strains, except C1, produced phytoene synthase, with variations seen in strains JH41 and B414. ATCC_21573 produced phytoene/squalene synthase. These findings suggest nuanced terpene production within clades, with variability in gene length and location across strains, underscoring the intricate diversity in *C. glutamicum* terpene biosynthesis (Table 2).

The analysis conducted using BAGEL4 showcased the antimicrobial capabilities within *C. glutamicum* strains. Clade 1 and clade 2 strains were devoid of identifiable specific bacteriocin BGCs. Clade 3 and clade 4 strains exhibited a shared putative bacteriocin BGC named "Lactococcin_972," indicating potential similar antimicrobial characteristics. Conversely, Clade 5, akin to clade 1 and 2, did not demonstrate distinct bacteriocin BGCs (Table 3).

Correlation between BGCs number with genome size and total gene count indicates a moderate positive correlation ($R^2 = 0.349$ and $R^2 = 0.358$ respectively) (Fig 6). The diversity of BGCs among the strains with phylogenetic relationship were shown in five clades (Fig 7).

Additionally, strain B253, R, SCgG1, and SCgG2 contain hybrid BGCs. All four strains contained hybrid BGCs comprised with NAPAA and betalactone. But the locations of NAPAA-betalactone hybrid BGCs are different in the genomes. The locations are 256574–294301 base pairs in strain B253, 334064–369207 base pairs in strain R, 319,462–354,607 base pairs in strain SCgG1, and 319,463–354,608 base pairs in SCgG2 (Fig 8).

Table 2. Predicted terpene BGCs in different clades of *C. glutamicum* strains.

No of Clade	Name of Organism	Region	Location		Length		Compound			Similarity
			Start	End	NT	AA	Most Similar	Type	Product	
1	*C. glutamicum* CGMCC1.15647	Region 1	717,181	736,285	918	305	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
		Region 2	2,558,650	2,569,225	576	191	Carotenoid	Terpene	phytoene/squalene synthase family protein	-
			2,569,222	2,569,473	252	83	Carotenoid	Terpene	phytoene/squalene synthase family protein	-
	C. glutamicum USDA-ARS-USMARC-56828	Region 1	773,090	774,004	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,488,448	2,489,271	824	274	Carotenoid	Terpene	phytoene synthase	25%
	C. glutamicum R	Region 2	816,127	817,041	915	304	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
	C. glutamicum SCgG2	Region 2	828,029	828,946	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum SCgG1	Region 2	828,028	828,945	918	305	Carotenoid	Terpene	phytoene synthase	100%
2	*C. glutamicum* ATCC_13032	Region 1	636,907	637,821	915	304	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
		Region 2	2,581,072	2,581,980	909	302	Carotenoid	Terpene	phytoene/squalene synthase family protein	25%
	C. glutamicum HA	Region 1	45,599	46,507	909	302	Carotenoid	Terpene	phytoene/squalene synthase family protein	25%
		Region 4	1,416,327	1,417,241	915	304	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
	C. glutamicum TQ2223	Region 1	189,322	190,230	909	302	Carotenoid	Terpene	phytoene synthase	25%
		Region 2	2,098,009	2,098,923	915	304	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum MB001	Region 1	638,057	638,971	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,342,054	2,342,917	864	287	Carotenoid	Terpene	geranylgeranyl-diphosphategeranylgeranyltransferase	25%
	C. glutamicum C1	-	-	-	-	-	-	-	-	-
	C. glutamicum CR101	Region 1	631,448	632,362	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,315,455	2,316,318	864	287	Carotenoid	Terpene	geranylgeranyl-diphosphategeranylgeranyltransferase	25%
	C. glutamicum BCA	Region 1	2,325,423	2,326,286	864	287	Carotenoid	Terpene	phytoene/squalene synthase family protein	25%
3	*C. glutamicum* XV	Region 2	749,608	750,525	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum CP	Region 2	785,574	786,491	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum WM001	Region 1	295,442	296,359	918	305	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
	C. glutamicum ATCC_13869	Region 2	744,821	745,738	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum AJ1511	Region 1	746,437	747,354	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum ZL-6	Region 2	753,776	754,693	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum ATCC_21799	Region 4	2,581,219	2,582,136	918	305	Carotenoid	Terpene	phytoene synthase	87%
	C. glutamicum TCCC11822	Region 2	748,242	749,159	918	305	Carotenoid	Terpene	phytoene synthase	100%

(*Continued*)

Table 2. (Continued)

No of Clade	Name of Organism	Region	Location		Length		Compound			Similarity
			Start	End	NT	AA	Most Similar	Type	Product	
4	*C. glutamicum* B253	Region 1	697,068	697,985	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum BE	Region 1	294,621	295,538	918	305	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
	C. glutamicum YI	Region 3	704,032	704,949	918	305	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum ATCC_14067	Region 3	705,614	706,531	918	305	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
5	*C. glutamicum* AR1	Region 1	704,384	705,298	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,420,624	2,421,487	864	287	Carotenoid	Terpene	phytoene synthase	25%
	C. glutamicum ATCC_21831	Region 1	735,201	736,115	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,451,884	2,452,747	864	287	Carotenoid	Terpene	phytoene synthase	25%
	C. glutamicum JH41	Region 1	649,099	649,977	879	292	Carotenoid	Terpene	phytoene/squalene synthase family protein	25%
		Region 4	2,060,429	2,061,343	915	304	Carotenoid	Terpene	phytoene synthase	100%
	C. glutamicum B414	Region 1	723,465	724,379	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,428,684	2,429,508	825	274	Carotenoid	Terpene	phytoene synthase	25%
	C. glutamicum CICC10064	Region 1	725,672	726,586	915	304	Carotenoid	Terpene	phytoene synthase	100%
		Region 2	2,431,451	2,432,275	825	274	Carotenoid	Terpene	phytoene synthase	25%
	C. glutamicum ATCC_21573	Region 1	697,484	698,398	915	304	Carotenoid	Terpene	phytoene/squalene synthase family protein	100%
		Region 2	2,450,615	2,451,478	864	287	Carotenoid	Terpene	phytoene/squalene synthase family protein	25%

Similarity: Similarity percentage with the reference database.

Region: Identified BGCs location in genome.

https://doi.org/10.1371/journal.pone.0299588.t002

PRISM 4 identified 4 major classes of BGCs which were polyketide, nonribosomal peptide, dehydratase, class II/III confident bacteriocin. Polyketide and nonribosomal peptide BGCs were present in all strains, while dehydratase were found in 21 strains and class II/III confident bacteriocin were found in 12 strains of *C. glutamicum* (S5 Table).

Besides, genome mining by BAGEL4 revealed bacteriocin coding clusters among 12 strains (S5 Table). Our identified BGCs from different online platform for each strain is listed in Table 4.

3.5 Genomic and SNP analysis

In this study, we employed BRIG-0.95 for comprehensive genome comparisons among various strains of *C. glutamicum*. The reference genome, *C. glutamicum* SCgG2, was utilized as a baseline for these comparisons. Notably, a substantial portion of genes present in SCgG2 were

Table 3. Predicted bacteriocin BGC in *C. glutamicum* strains.

No of Clade	Name of Organism	Gene Length		Compound	
		Start	End	Type	Name
1	*C. glutamicum* CGMCC1.15647	-	-	-	-
	C. glutamicum USDA-ARS-USMARC-56828	-	-	-	-
	C. glutamicum R	-	-	-	-
	C. glutamicum SCgG2	-	-	-	-
	C. glutamicum SCgG1	-	-	-	-
2	*C. glutamicum* ATCC_13032	-	-	-	-
	C. glutamicum HA	-	-	-	-
	C. glutamicum TQ2223	-	-	-	-
	C. glutamicum MB001	-	-	-	-
	C. glutamicum C1	-	-	-	-
	C. glutamicum CR101	-	-	-	-
	C. glutamicum BCA	-	-	-	-
3	*C. glutamicum* XV	244312	244611	putative_bacteriocin	Lactococcin_972
	C. glutamicum CP	245624	245923	putative_bacteriocin	Lactococcin_972
	C. glutamicum WM001	3092978	3093184	putative_bacteriocin	Lactococcin_972
	C. glutamicum ATCC_13869	244322	244621	putative_bacteriocin	Lactococcin_972
	C. glutamicum AJ1511	245938	246237	putative_bacteriocin	Lactococcin_972
	C. glutamicum ZL-6	245649	245948	putative_bacteriocin	Lactococcin_972
	C. glutamicum ATCC_21799	2077312	2077611	putative_bacteriocin	Lactococcin_972
	C. glutamicum TCCC11822	244425	244631	putative_bacteriocin	Lactococcin_972
4	*C. glutamicum* B253	206829	207086	putative_bacteriocin	Lactococcin_972
	C. glutamicum BE	3167629	3167835	putative_bacteriocin	Lactococcin_972
	C. glutamicum YI	235165	235464	putative_bacteriocin	Lactococcin_972
	C. glutamicum ATCC_14067	236772	237071	putative_bacteriocin	Lactococcin_972
5	*C. glutamicum* AR1	-	-	-	-
	C. glutamicum ATCC_21831	-	-	-	-
	C. glutamicum JH41	-	-	-	-
	C. glutamicum B414	-	-	-	-
	C. glutamicum CICC10064	-	-	-	-
	C. glutamicum ATCC_21573	-	-	-	-

https://doi.org/10.1371/journal.pone.0299588.t003

found to be shared by the other strains, indicating a core genomic similarity among these strains.

However, a detailed examination of the genomic alignments revealed significant disparities between SCgG2 and other strains, as denoted by white gaps in Fig 9. These gaps signify regions where genes were absent in certain strains, indicating potential genetic variations. Such discrepancies could be attributed to the integration of mobile genetic elements, horizontal gene transfer events, or recombination phenomena. These mechanisms are known to drive genetic diversification in bacterial populations, leading to the acquisition or loss of specific genes over evolutionary time. The identification of these genomic variances underscores the dynamic nature of *C. glutamicum* genomes and highlights the genomic plasticity within this bacterial species.

Fig 10 illustrates the output from pairwise whole-genome Mauve alignments, confirming the presence of significant structural variations among the genomes of the analysed strains. In each comparison, matching coloured blocks and connecting lines delineate homologous genome sections between the compared pairs. Notably, strains TCCC11822, TQ2223, BCA,

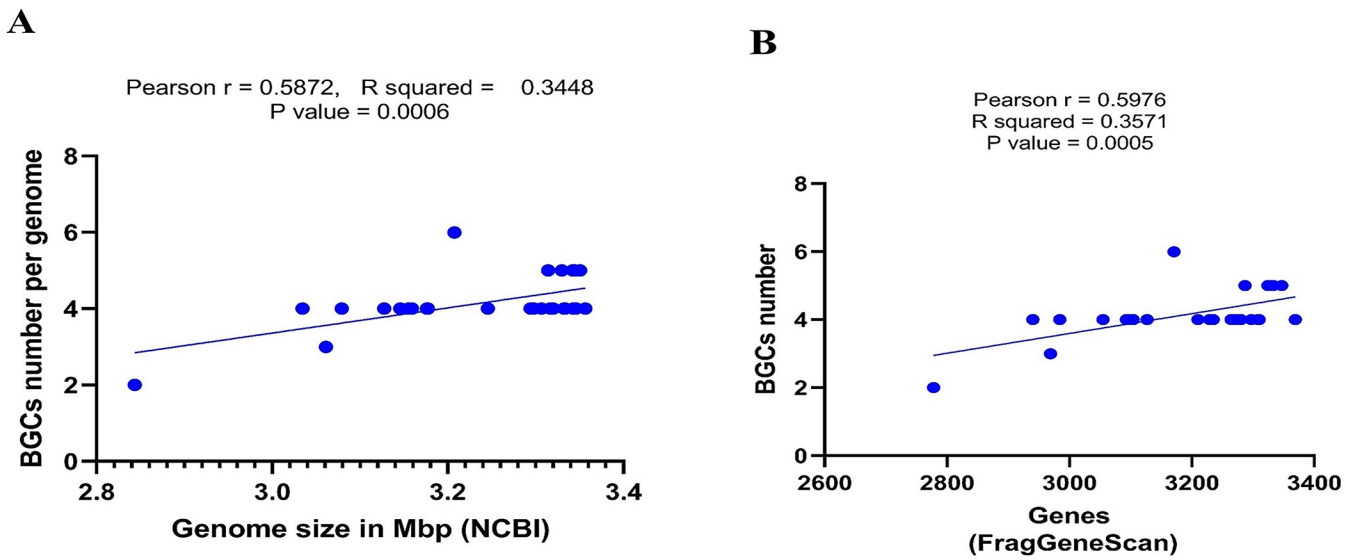

Fig 6. BGCs distribution in *C. glutamicum* strains. (A) Correlation of BGCs and genome size. (B) Correlation of BGCs and gene number.

https://doi.org/10.1371/journal.pone.0299588.g006

CR101, HA, and ATCC 21573 exhibited the most significant variations, indicating diverse genomic structures within these strains. These visual cues provide insights into the shared genomic regions and structural differences between the analysed strains.

SNPs analysis within the various *C. glutamicum* strains provided valuable insights into the genetic diversity of these strains. **Table 5** presents a comprehensive analysis of genetic variants among various strains of *C. glutamicum*. Notably, *C. glutamicum* USDA-ARS-US-MARC-56828 exhibited the highest number of total variants (41270), characterized by substantial counts in complex variants (7908) and SNPs (32960). This strain displayed a significant divergence compared to others. Conversely, strains like *C. glutamicum* SCgG1 showed minimal variants, with only 28 total variants. Several strains, such as *C. glutamicum* R, displayed a relatively low total variant count (20433) and a notable prevalence of deletions (141) and insertions (130). These findings underscore the genetic diversity within *C. glutamicum* strains, with certain strains exhibiting distinctive patterns of variation, potentially influencing their biological characteristics. The presence of unique SNPs in each strain suggests specific genomic changes, potentially influencing their functional attributes and ecological roles.

The phylogenetic tree, as illustrated in Fig 11 based on Core SNPs analysis, delineates the evolutionary relationships among the *C. glutamicum* strains. The tree is rooted with a reference strain (SCgG2). Noteworthy patterns emerge, revealing distinct clusters and branches that denote genetic proximity. For instance, strains like AJ1511, WM001, and TCCC11822 form a cluster, suggesting a shared genetic ancestry. Similarly, ZL-6 and ATCC 21799 exhibit close genetic relatedness. The tree also portrays a bifurcation between B253 and its cluster, including BE, ATCC 14067, and YI, reflecting their divergence. Further branching showcases the genetic relationships among diverse strains, emphasizing the intricate evolutionary dynamics within the *C. glutamicum* species. The placement of the reference strain in the analysis enables a comparative understanding of genetic variations, highlighting its pivotal role in contextualizing the evolutionary history of the examined strains. Overall, the phylogenetic tree provides a visual representation of the genetic distances and relationships, offering valuable insights into the evolutionary landscape of *C. glutamicum*.

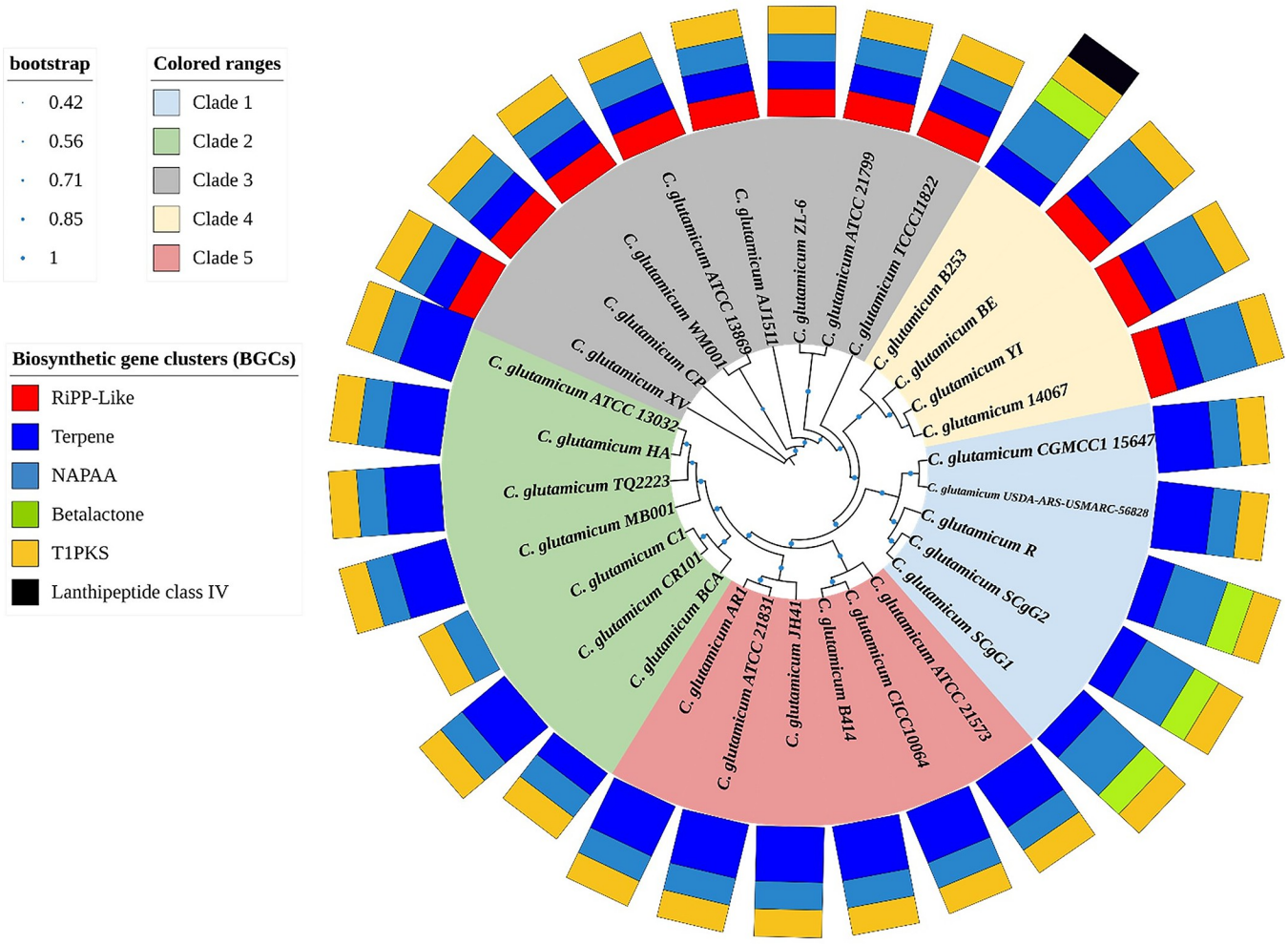

Fig 7. Major classes of BGCs in the genomes of *C. glutamicum* strains with phylogenetic distribution. These BGC classes are categorized into five clades, each delineated based on their specific biosynthetic gene content.

https://doi.org/10.1371/journal.pone.0299588.g007

3.6 Horizontal gene transfer

Utilizing the HGTector tool, an exhaustive analysis was performed, revealing a substantial number of HGT events within the genomes of *C. glutamicum* strains. Specifically, in the AJ1511 strain, 684 distinct HGT events were identified from a dataset of 3014 predicted proteins. These events were predominantly sourced from *Actinomycetes* (71%) and to a lesser extent, *Micrococcales* (21%). Similarly, in the AR1 strain, a total of 237 genes were predicted to have undergone HGT events out of 2759 proteins analysed. Notably, the majority of these events were attributed to *Actinomycetes* (73%), with a smaller fraction originating from *Micrococcales* (23%) as illustrated in Fig 12. Considering the prevalence of HGT events in AJ1511 and AR1, it is likely that other *C. glutamicum* strains, would reveal a mosaic of genetic origins. The genomic plasticity observed in these two strains is indicative of the adaptive strategies employed by *C. glutamicum* populations, emphasizing the role of HGT in shaping their genetic repertoire.

3.7 Pathogenicity, virulence properties and plasmid analysis

The investigation revealed that none of the strains belonging to *C. glutamicum* exhibited characteristics indicative of human pathogenicity. A detailed presentation of these findings is

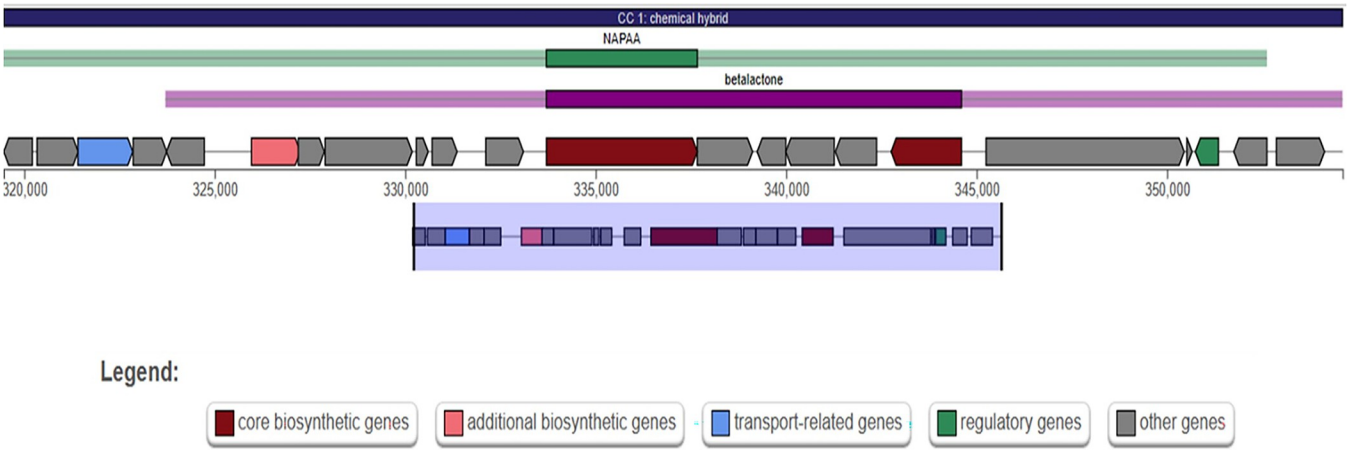

Fig 8. Hybrid BGCs structure *C. glutamicum* strains. Hybrid BGCs in four strains harbour same structure of NAPAA-betalactone. The different locations of NAPAA-betalactone are, 256574–294301 base pairs in strain B253, 334064–369207 base pairs in strain R, 319,462–354,607 base pairs in strain SCgG1, and 319,463–354,608 base pairs in SCgG2.

https://doi.org/10.1371/journal.pone.0299588.g008

Table 4. Different hits of BGCs from different genome mining tools using *C. glutamicum* genomes.

Species name	Isolation source	Genome size in Mbp	AntiSMASH Hit	Prism Hit	BAGEL Hit
C. glutamicum CP	air	3.34	4	4	1
C. glutamicum XV	soil	3.33	4	4	1
C. glutamicum AJ1511		3.3	4	4	1
C. glutamicum ATCC 21799		3.33	4	4	1
C. glutamicum ZL-6	soil	3.33	4	4	1
C. glutamicum B253		3.21	6	4	1
C. glutamicum BE		3.34	5	4	1
C. glutamicum 14067	rotten onion	3.33	5	4	1
C. glutamicum YI	soil	3.34	5	4	1
C. glutamicum R		3.31	5	3	0
C. glutamicum SCgG1	soil	3.35	5	3	0
C. glutamicum SCgG2	soil	3.35	5	3	0
C. glutamicum USDA-ARS-USMARC-56828	mucus of calf	3.25	4	3	0
C. glutamicum CGMCC1.15647		3.36	4	2	0
C. glutamicum ATCC 21573		3.18	4	3	0
C. glutamicum B414	soil	3.16	4	3	0
C. glutamicum CICC10064	soil	3.16	4	3	0
C. glutamicum MB001		3.08	4	2	0
C. glutamicum TQ2223	soil	3.31	4	2	0
C. glutamicum ATCC 13032		3.37	4	2	0
C. glutamicum HA		3.35	4	2	0
C. glutamicum BCA		3.06	3	2	0
C. glutamicum C1	derivative of ATCC 13032	2.84	2	2	0
C. glutamicum CR101	lab strain	3.03	4	2	0
C. glutamicum JH41		3.13	4	2	0
C. glutamicum AR1		3.15	4	3	0
C. glutamicum ATCC 21831		3.18	4	3	0
C. glutamicum TCCC11822	soil	3.3	4	4	1
C. glutamicum ATCC 13869	soil	3.3	4	4	1
C. glutamicum WM001	soil	3.32	4	4	1

https://doi.org/10.1371/journal.pone.0299588.t004

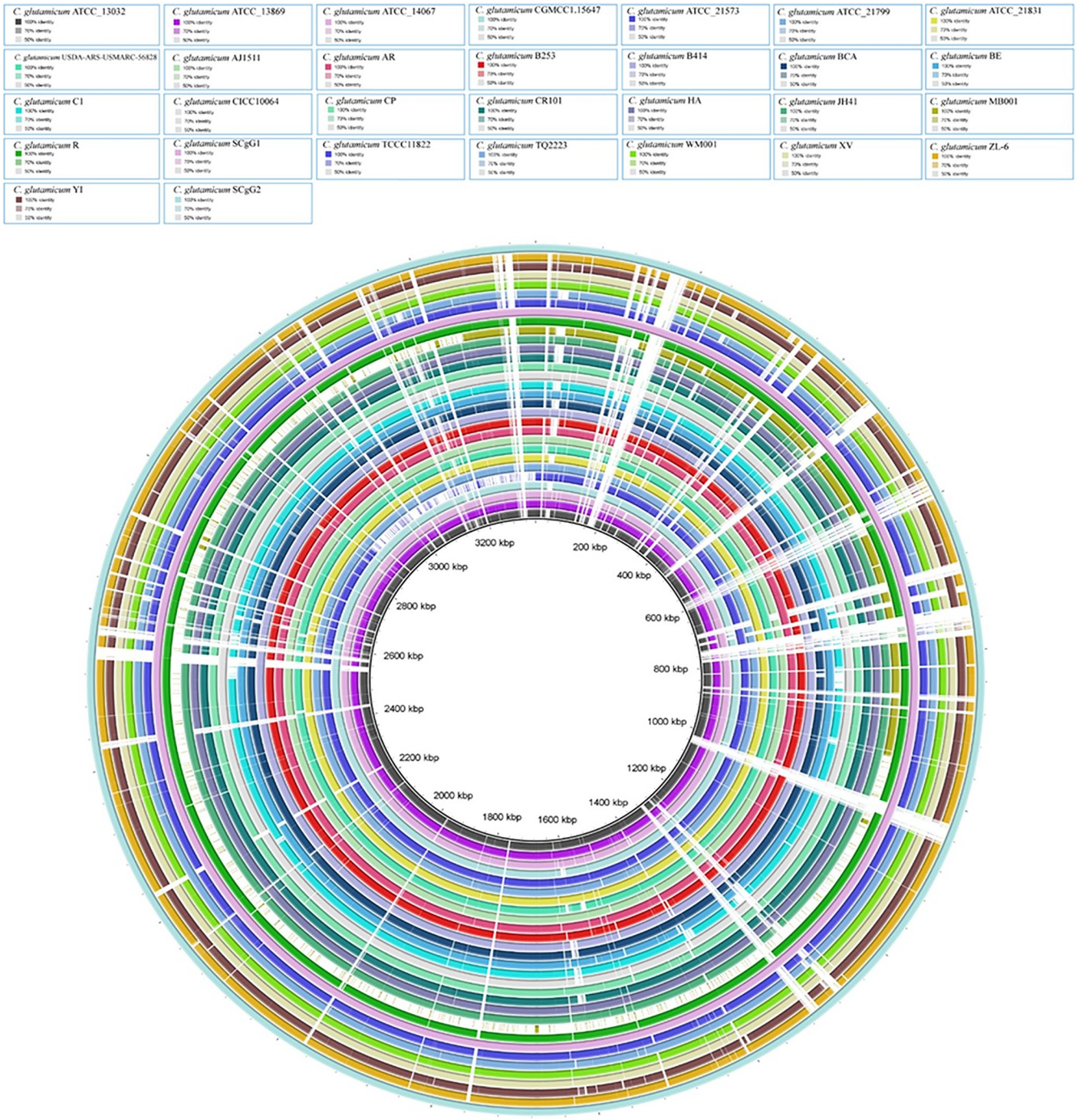

Fig 9. BRIG Diagram illustrating homologous chromosome segments of *C. glutamicum* strains using strain SCgG2 as the reference genome.

https://doi.org/10.1371/journal.pone.0299588.g009

encapsulated in the Table 6. This underscores the non-pathogenic nature of the examined *C. glutamicum* strains concerning human health. It is noteworthy that non-pathogenic bacteria lack the genetic elements associated with virulence, thereby affirming their incapacity to induce infections or diseases in humans.

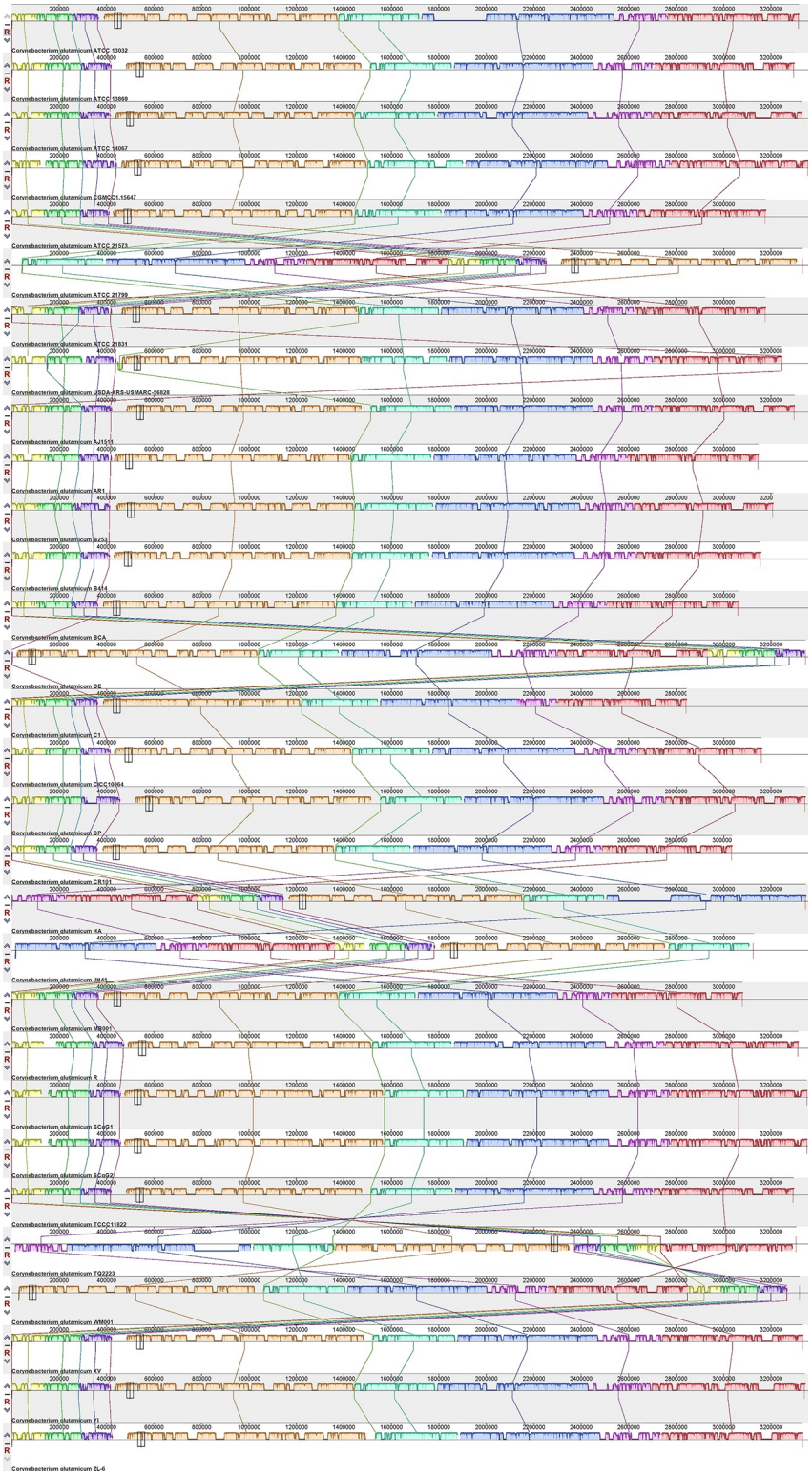

Fig 10. The pairwise whole-genome Mauve alignment analysis revealed substantial structural variations within the circular chromosomes of *C. glutamicum* strains.

https://doi.org/10.1371/journal.pone.0299588.g010

Table 5. SNPs detected in *C. glutamicum* strains.

Strain	Accession Number	Variant					
		Complex	Deletions	Insertions	MNPs	SNPs	Total Variant
***C. glutamicum* AJ1511**	AP017557.2	6934	187	206	-	30582	37909
***C. glutamicum* ATCC_21799**	AP022856.1	6904	190	204	-	30676	37974
***C. glutamicum* SCgG1**	CP004047.1	1	1	-	-	26	28
***C. glutamicum* SCgG2**	CP004048.1	Referance					
***C. glutamicum* ZL-6**	CP004062.1	7010	187	212	-	31300	38709
***C. glutamicum* MB001**	CP005959.1	6938	212	183	-	30316	37649
***C. glutamicum* ATCC_21831**	CP007722.1	7032	169	194	-	30780	38175
***C. glutamicum* AR1**	CP007724.1	7037	167	195	-	31104	38503
***C. glutamicum* B253**	CP010451.1	6646	205	213	-	30181	37245
***C. glutamicum* CP**	CP012194.1	6963	188	208	-	30985	38344
***C. glutamicum* B414**	CP012297.1	7380	183	188	-	30963	38714
***C. glutamicum* CICC10064**	CP012298.1	7389	181	190	-	30983	38743
***C. glutamicum* USDA-ARS-USMARC-56828**	CP013991.1	7908	197	205	-	32960	41270
***C. glutamicum* YI**	CP014984.1	6704	215	199	-	30639	37757
***C. glutamicum* ATCC_13869**	CP016335.1	6982	185	204	4	30676	38051
***C. glutamicum* C1**	CP017995.1	6387	200	179	1	28229	34996
***C. glutamicum* XV**	CP018175.1	6962	192	209	-	30990	38353
***C. glutamicum* TCCC11822**	CP020033.1	6934	185	208	-	30782	38109
***C. glutamicum* TQ2223**	CP020658.1	6985	235	183	-	30668	38071
***C. glutamicum* ATCC_14067**	CP022614.1	6649	212	201	-	29882	36944
***C. glutamicum* ATCC_13032**	CP025533.1	6986	217	186	2	30405	37796
***C. glutamicum* HA**	CP025534.1	7019	220	185	2	30376	37802
***C. glutamicum* JH41**	CP041729.1	6811	185	193	-	30173	37362
***C. glutamicum* BE**	CP053188.1	6640	208	201	-	29779	36828
***C. glutamicum* ATCC_21573**	CP068290.1	7170	190	195	-	31222	38777
***C. glutamicum* CR101**	CP080542.1	6941	212	193	-	30327	37673
***C. glutamicum* R**	NC_009342.1	3520	141	130	-	16642	20433
***C. glutamicum* WM001**	NZ_CP022394.1	6991	186	204	-	30871	38252
***C. glutamicum* BCA**	NZ_CP059382.1	6934	212	185	-	30146	37477
***C. glutamicum* CGMCC1.15647**	NZ_CP073911.1	7366	197	202	-	32021	39786

https://doi.org/10.1371/journal.pone.0299588.t005

The plasmid analysis across different strains of *C. glutamicum* revealed diverse characteristics. Strains CP, XV, B253, USDA-ARS-USMARC-56828, AR1, ATCC_21831, and ATCC_13869 were found to harbor IncA/C type plasmids, with varying lengths and GC content (Table 7). Notably, strains R and CGMCC1.15647 exhibited distinct plasmid types, namely IncI1 and IncHI1, respectively, and displayed substantial variations in plasmid sizes. The gene content of these plasmids varied among strains, encompassing differences in coding sequences (CDs), pseudo genes, CRISPR arrays, rRNAs, tRNAs, ncRNA, and frameshifted genes. Among the strains analyzed, 10 were reported to carry single plasmids, while C. glutamicum CGMCC1.15647 was unique with two plasmids. The prevalent IncA/C type plasmid, found in the majority of strains, is known for its role in modulating changes to bacterial host chromosomes. In contrast, *C. glutamicum* R carries an IncI1 type plasmid, responsible for encoding sex pili in bacteria. IncHI1 type plasmid is associated with antibiotic resistance. This comprehensive analysis underscores the diversity and functional significance of plasmids in *C. glutamicum* strains.

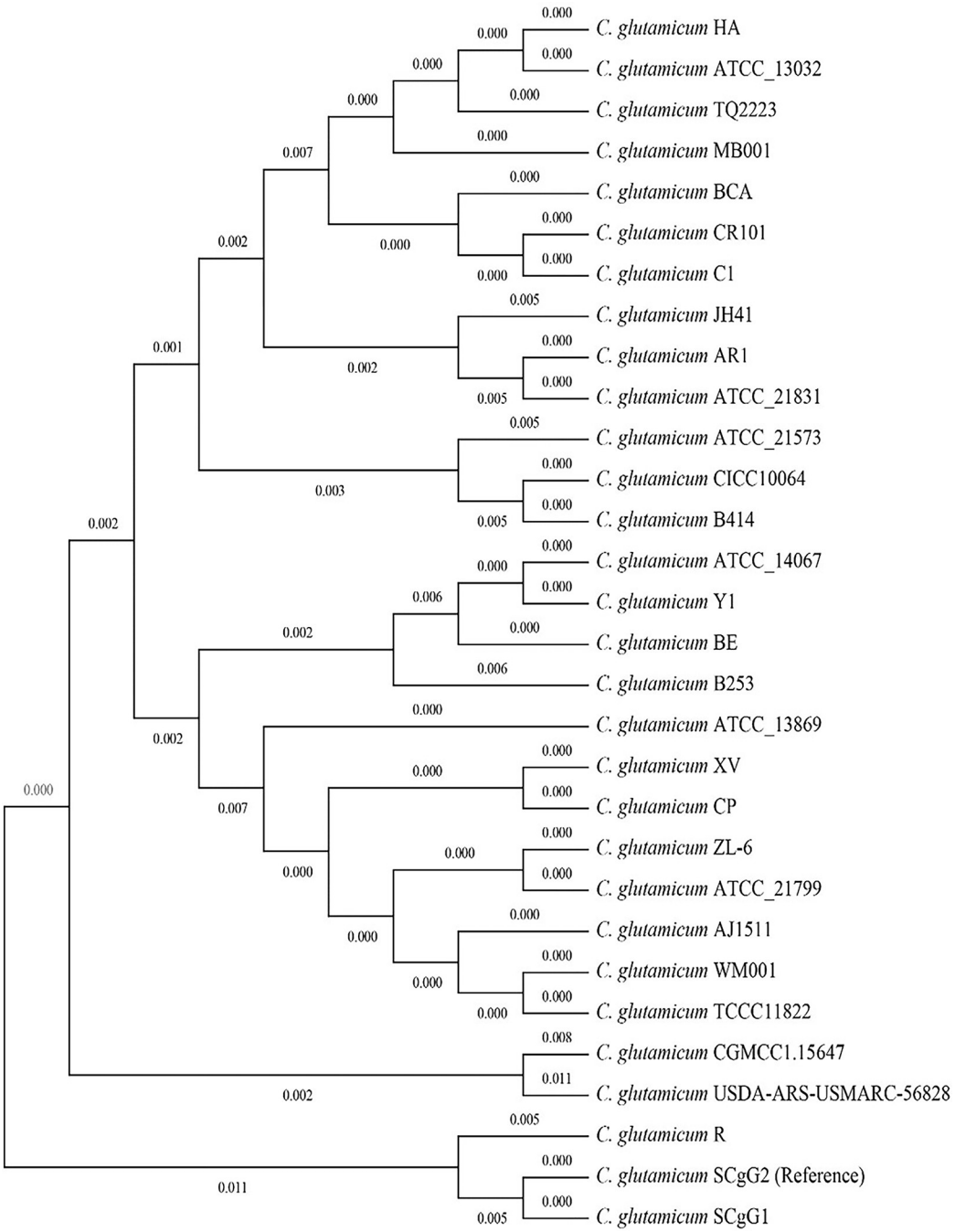

Fig 11. Visualizing the phylogeny of *C. glutamicum* strains based on core SNP genes.

https://doi.org/10.1371/journal.pone.0299588.g011

4. Discussion

Whole genome comparison by ANI calculation revealed high degree of relatedness between *C. glutamicum* strains. ANI computation with a higher than 97% score verifies that our studied genomes belong to the same species and are closely related. ANI comparison between *Corynebacterium cystitidis* strains showed a 95.1% score when isolated from the different hosts but

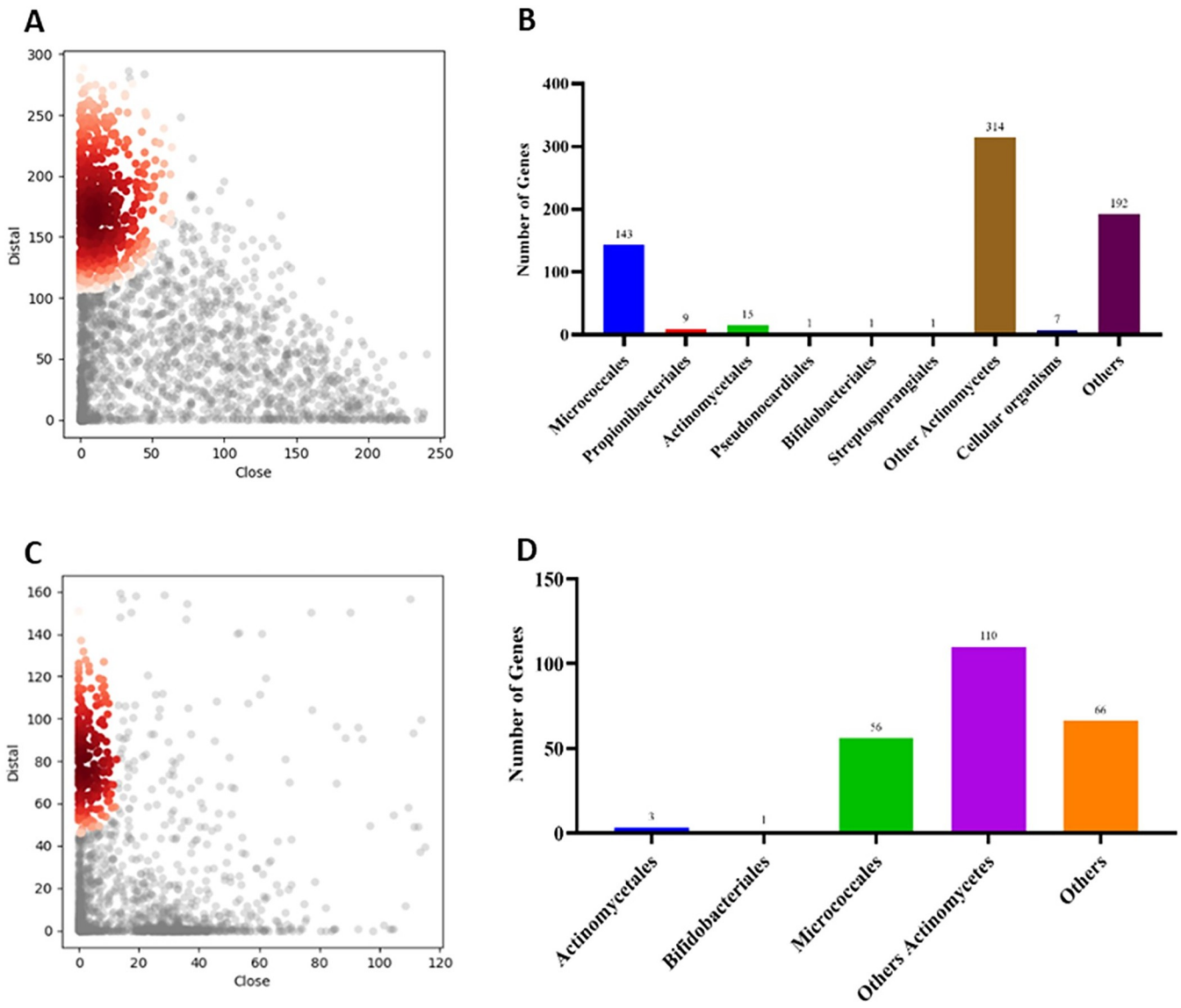

Fig 12. HGT events in *C. glutamicum* AJ1511 and AR1 Strains. (A) Scatter plot illustrating horizontally transferred genes in AJ1511 (Colour dots represents horizontally transferred genes and colourless dots represents native genes). (B) Distribution of donor organisms and the corresponding number of genes transferred in AJ1511. (C) Scatter plot showcasing horizontally transferred genes in AR1 (Colour dots represents horizontally transferred genes and colourless dots represents native genes). (D) Distribution of donor organisms and the corresponding number of genes transferred in AR1. (In the scatter plots, coloured dots represent genes transferred through HGT).

https://doi.org/10.1371/journal.pone.0299588.g012

showed a >99% score when isolated from the same host [65]. Our demographic data also support a >97% score since most of our strains are from soil sources.

The average genome size (3.24 Mbp) of the strains was slightly high, compared with non-pathogenic *C. casei* LMG S-19264 (3.11 Mbp) and *C. efficiens* YS-314 (3.15 Mbp) [66]. Moreover, the average number of genes (3197) was also higher than *C. casei* LMG S-19264 (2872) and *C. efficiens* YS-314 (3064) [66]. On the other hand, the average GC content was lower among *C. glutamicum* strains (54.15%) than other non-pathogenic *C. variabile* DSM 44702 (76.1%) and *C. efficiens* YS-314 (69.93%) [67]. We found variation in tRNA coding genes among the *C. glutamicum* strains, since the tRNA genes varied from 57 to 65 among the strains. Whereas, *C. variabile* DSM 44702 and *C. efficiens* YS-314 contains 59 and 56 tRNA

Table 6. Pathogenicity prediction results for various *C. glutamicum* strains.

Starin	Probability of being a human pathogen	Matched Pathogenic Families	Prediction
C. glutamicum AJ1511	0.258	0	Non-human pathogen
C. glutamicum ATCC_21799	0.258	0	Non-human pathogen
C. glutamicum SCgG1	0.265	0	Non-human pathogen
C. glutamicum SCgG2	0.265	0	Non-human pathogen
C. glutamicum ZL-6	0.262	0	Non-human pathogen
C. glutamicum MB001	0.257	0	Non-human pathogen
C. glutamicum ATCC_21831	0.262	0	Non-human pathogen
C. glutamicum AR1	0.261	0	Non-human pathogen
C. glutamicum B253	0.265	0	Non-human pathogen
C. glutamicum CP	0.257	0	Non-human pathogen
C. glutamicum B414	0.264	0	Non-human pathogen
C. glutamicum CICC10064	0.264	0	Non-human pathogen
C. glutamicum USDA-ARS-USMARC-56828	0.254	0	Non-human pathogen
C. glutamicum YI	0.261	0	Non-human pathogen
C. glutamicum ATCC_13869	0.258	0	Non-human pathogen
C. glutamicum C1	0.256	0	Non-human pathogen
C. glutamicum XV	0.256	0	Non-human pathogen
C. glutamicum TCCC11822	0.257	0	Non-human pathogen
C. glutamicum TQ2223	0.255	0	Non-human pathogen
C. glutamicum ATCC_14067	0.261	0	Non-human pathogen
C. glutamicum ATCC_13032	0.256	0	Non-human pathogen
C. glutamicum HA	0.257	0	Non-human pathogen
C. glutamicum JH41	0.26	0	Non-human pathogen
C. glutamicum BE	0.261	0	Non-human pathogen
C. glutamicum ATCC_21573	0.266	0	Non-human pathogen
C. glutamicum CR101	0.256	0	Non-human pathogen
C. glutamicum R	0.26	0	Non-human pathogen
C. glutamicum WM001	0.258	0	Non-human pathogen
C. glutamicum BCA	0.256	0	Non-human pathogen
C. glutamicum CGMCC1.15647	0.259	0	Non-human pathogen

https://doi.org/10.1371/journal.pone.0299588.t006

Table 7. Plasmid characteristics in different *C. glutamicum* strains.

Strain	Plasmid Accession Number	Type of Plasmid	Length	GC (%)	Genes	CDs	Pseudo Genes	CRISPR Arrays	rRNAs	tRNAs	ncRNA	Frameshifted Genes
C. glutamicum CP	CP012412.2	IncA/C	4447	53.5	3122	2898	145		6	60	1	62
C. glutamicum XV	CP018174.1	IncA/C	4447	53	3165	3079	193		6	65	3	
C. glutamicum B253	CP010452.1	-	2175	51	2984	2767	138	3	18	60	1	54
C. glutamicum R	AP009045.1	IncI1	49120	53.9								
C. glutamicum USDA-ARS-USMARC-56828	CP014329.1	IncA/C	48334	53.5	3062	2981	85	1	6	60	3	
C. glutamicum CGMCC1.15647	CP073912.1		47523	58								
	CP073913.1	IncHI1	24235	54								
C. glutamicum AR1	CP007725.1	IncA/C	16810	51.5	2901	2760	65		18	57	1	36
C. glutamicum ATCC_21831	CP007723.1	IncA/C	16810	51.5	2932	2785	68		18	60	1	42
C. glutamicum ATCC_13869	CP016336.1	IncA/C	4447	53.5	3075	2994	108		6	60	3	

https://doi.org/10.1371/journal.pone.0299588.t007

genes respectively [67]. Additionally, *C. glutamicum* strains possess more rRNA genes (15–18 rRNA genes) compared with other non-pathogenic *Brevibacterium auranticum* strains and *Brevibacterium linens* ATCC 19391 (12 rRNA genes) [68]. Besides, the average CDS among the strains was 3007, comparatively higher than *C. efficiens* YS-314 (2950 CDS) [69].

Pan-genome study of *Corynebacterium* at genus level showed very low number of core genes [66,67]. Analysis between 51 strains of various pathogenic and non-pathogenic species of *Corynebacterium* genus showed 8.69% of core genes [66]. Similarly, study of eleven *Corynebacterium* species showed 6.68% of core genes [67]. Contrary to genus level, we found core genes of 29.1% at sub-species level among *C. glutamicum* strains, which is somewhat higher than *C. pseudotuberculosis* core genes (26.1%) at sub-species level [70]. The number of cloud genes (strain-specific genes) was considerably large and covered 47.78% of the pan-genome, similar to *C. pseudotuberculosis* cloud genes (42.34%) [70]. The low percentage of core genes in *C. glutamicum* species likely results from a combination of factors such as horizontal gene transfer, adaptation to diverse environments, evolutionary divergence, and specialization. From an evolutionary perspective, this genetic diversity contributes to the species' ability to adapt, survive, and thrive in different ecological niches. Which strongly demonstrates the diversity among the strains. Large accessory genomes and a high number of strain-specific genes are frequently linked to horizontal gene transfer (HGT) in microorganisms [71]. Besides, we found low GC content of *C. glutamicum* strains with other non-pathogenic species of *Corynebacterium* genus. Our study also suggests a clear inverse relation between the abundance of accessory genes and the genomic GC content. Specifically, as the GC percentage increases, there is a notable decrease in the number of accessory genes observed. This finding supports the idea of possible relation of low GC content with horizontal gene transfer and codon reassignment of *C. glutamicum* [72–75].

Our study shows the open nature of the *C. glutamicum* pan-genome, which indicates that new gene families continuously will be added to the pan-genome. The open pan-genome of *Corynebacterium* at genus level was also reported by the pan-genomic analysis of 40 strains of eleven different *Corynebacterium* species [76]. Thus, the pan-genome of *C. glutamicum* indicates the diversity of the gene pool and the likeliness of increasing gene number.

Another objective of our study was to uncover the diversity and distribution of BGCs among the strains. Although BGCs producing metabolic products remained undocumented, predictions based on bioinformatics revealed that several of them might encode products with unique structures [77–79]. Thus, our computational approaches were to predict BGCs as a screening process for new bioactive compound production, which are to be effectively applied in the wet laboratories.

NAPAA of Nonribosomal peptide synthetases (NRPs) gene clusters and T1PKS of Polyketide synthases (PKSs) gene clusters were found in all the studied strains. Additionally, Terpene BGCs were found in 96.67% strains. T1PKS, Terpene, NAPAA and other NRPs were also most common in *Gordonia hongkongensis* EUFUS-Z298 [80], *Burkholderia* spp. [18], in activated sludge microbiome [81], and in *Ktedonobacteria* [82]. NAPAA, particularly e-poly-lysine, demonstrate notable antimicrobial efficacy, showcasing widespread utility in the food and pharmaceutical sectors. Conversely, T1PKS harbor the capability to biosynthesize peptides with antibiotic and antitumor properties. Terpenoids exhibit robust and specific biological activities, notably against diseases such as cancer and malaria. The consistency of limited number of BGCs among closely related bacterial population was previously reported [83]. Which indicates that BGCs 'fixation' can be occurred as a strong positive selection and to survive specific environment by the activity of encoded products [17]. The novel BCGs identified from the strains used for analysis include betalactone and lanthipeptide class IV BGCs. Betalactone BGCs were predicted in strain R, B253, SCgG1, and SCgG2, while lanthipeptide class IV BGCs

were only found in strain B253. Betalactones manifest noteworthy bioactivity against bacteria, fungi, and cancer cell lines. Lanthipeptides, belonging to the subclass of ribosomally-synthesized and posttranslationally-modified peptides (RiPPs), generally display feeble antibacterial activities, with Lenthipeptide-class-IV standing out as a noteworthy example. A study of *Bacillus cereus* strains identified different lanthipeptide classes, and concluded that several lanthipeptide classes can evolve independently, and most of the lanthipeptide BGCs can originated from intra-species horizontal gene transfer [84].

Additionally, PKS and NRPs BGCs which were most common in our studied genomes, are considered as representatives of two major classes of antibiotics [80]. Kalimantacin antibiotics with strong antistaphylococcal effect, from *Alcaligenes* species YL-02632S [85,86] and antibiotic batumin from *Pseudomonas batumici* have been produced utilizing these BGCs [87]. *C. glutamicum* is suitable for T1PKS and NRPs synthesis by heterologous expression since it possesses endogenous 4'-phosphopantetheinyl transferase (PPTase), PptAcg [88]. Roseoflavin, a broad-spectrum antibiotic was already produced using *C. glutamicum* via the heterologous expression of its BGCs [89]. We also found bacteriocin gene clusters in 12 strains of *C. glutamicum*. Bacteriocins have been seen as a feasible alternative to traditional antibiotics because of their distinct antibacterial processes. Besides, it can be used as innovative carrier molecules [90] and also as plant growth-promoting agent, antiviral agent, and anti-cancer agents [91].

Whole genome comparison based on ANI scores also revealed the phylogenetic relationship among the strains. We divided all 30 strains into five clades. Clade 1 with five strains, clade 2 with seven strains, clade 3 with eight strains, clade 4 with four strains, and clade 5 with six strains. We have seen diversity of the BGCs among clade 1, clade 2, and clade 4. Whereas members of clade 3 and clade 5 contain the same number of BGCs, but these two clades harbour different BGCs. We observed similar BGCs among the soil isolated strain CICC10064, B414, and TQ2223. Similarly, soil isolated strain XV, ZL-6, YI, TCCC11822, ATCC 13869, and WM001 have similar BGCs, where strain YI have gained extra NAPAA class. Soil isolated strain SCgG1 and ScgG2 have similar BGCs class with betalactone. On the other hand, strain C1, which is an engineered derivative of ATCC 13032 have lost double Terpene BGCs.

Additionally, we identified NAPAA-betalactone hybrid BGCs among strain B253, R, SCgG1 and SCgG2. Hybrid BGCs encodes genes that are responsible for multiple scaffold-synthesizing enzymes [92,93]. Occurrence of hybrid BGCs are common for some bacteria (98% occurrence in *Streptomyces*) [94], yet the exact roles of hybrid BGCs are not completely known [95,96]. It is noteworthy, that the specific locations of these hybrid BGCs within the genomes of these strains exhibit variation, as illustrated in Fig 8. This disparity implies that these hybrid BGCs might have undergone acquisition or rearrangement through horizontal gene transfer or recombination events, thereby contributing to genomic diversity across the strains. Consequently, our assertion of identifying hybrid BGCs is rooted in their gene content and functional characteristics, rather than their precise physical placement within the genomes.

We found that the number of BGCs is positively correlated with the genome size and the gene number of the strains. Strain SCgG1, ScgG2, BE, YI, 14067, and strain R with larger genome size and with high number of genes, each harbouring 5 BGCs in their genomes. Though, strain CGMCC1.15647 with the highest gene number and the largest genome size contains 4 BGCs. Thus, our correlation regression analysis shows that if the genome size and the gene number increase, the number of BGCs is more likely to increase. Generally, strains with larger genomes tend to exhibit a higher number of BGCs, a phenomenon attributed to the potential accumulation of accessory genes and genomic islands carrying BGCs [97].

The potential presence of sequencing errors within publicly available databases remains a notable concern. Only complete genome sequences of *C. glutamicum* strains, which enhance the reliability and comprehensiveness of the study's findings, were considered, addressing the

potential presence of sequencing errors within publicly available databases. Prokka and FragGeneScan were employed for genome annotation and gene prediction, representing widely used and validated tools for prokaryotic genomes. To ensure robust genome comparison and species delineation, OrthoANI, a pairwise average nucleotide identity (ANI) algorithm, more robust and accurate than traditional methods, was utilized. The pan-genome analysis employed Roary, BPGA, and USEARCH, utilizing rigorous computational frameworks and sequence identity cut-offs for gene classification and estimation. This comprehensive approach aimed to mitigate concerns related to sequencing errors, enhance reliability, and employ validated tools for effective genome annotation and pan-genome analysis in the study of *C. glutamicum* strains. Nevertheless, our investigation has revealed discernible diversity across various genomic features among the strains, along with variations in the abundance of biosynthetic gene clusters (BGCs) within their genomes. Virulence genes are pivotal elements that contribute to the pathogenicity of microorganisms, enabling them to induce diseases. In contrast, BGCs are typically responsible for encoding enzymes and proteins involved in synthesizing specific secondary metabolites, such as T1PKS, Terpene, NAPAA, betalactone, and lanthipeptide. The connection between BGCs and virulence is diverse, as certain secondary metabolites produced by BGCs can influence the virulence of microorganisms.

However, in our investigation, no identifiable secondary metabolites produced by BGCs were associated with virulence. Remarkably, all examined strains were found to be non-pathogenic. This suggests that there might be an absence of virulence genes located within the BGCs of these strains. The collective non-pathogenic nature of the strains reinforces the notion that the BGCs under scrutiny may not harbor genes contributing to virulence, further emphasizing the safety profile of these microorganisms in the context of human health.

While our study successfully identified numerous distinct polymorphic sites among the strains under investigation, it is crucial to acknowledge a limitation. The specific interaction or overlap between these polymorphic sites and BGCs in *C. glutamicum* has not been thoroughly explored within the scope of our research. This unexplored aspect represents a noteworthy limitation, suggesting a promising avenue for future investigation.

In all, we can say that strains of *C. glutamicum* can be a good candidate for engineering to produce various novel compound through BGCs expression. Also the strain may have potential to produce antibiotic, plant growth promoting agent, antiviral agent and anti-cancer agent.

5. Conclusions

Our objectives of the study were to elucidate the genetic variation, pan-genomic characteristics, and distribution of BGCs among 30 strains of *C. glutamicum*. We observed genetic variation and diversity in the BGCs distribution. Pan-genomic study of *C. glutamicum* strains revealed diversity at the sub-species level. We found a large number of strain-specific genes and the open nature of the *C. glutamicum* pan-genome. This study has yielded valuable insights into previously unexplored biosynthetic gene clusters (BGCs) that play a role in the production of betalactones, lanthipeptides, and NAPAA-betalactone hybrids. Thus, we conclude that various strains of *C. glutamicum* should be on focus for the discovery of natural drugs at the industrial level.

Supporting information

S1 Table. Average Nucleotide Identity (ANI) values for *C. glutamicum* strains.
(XLSX)

S2 Table. Genomic characteristics of various *C. glutamicum* strains.
(XLSX)

S3 Table. Biosynthetic Gene Clusters (BGCs) and corresponding hit counts.
(XLSX)

S4 Table. BGCs distribution across *C. glutamicum* strains.
(XLSX)

S5 Table. BGCs and genomic characteristics comparison of various *C. glutamicum* strains.
(XLSX)

Author Contributions

Conceptualization: Md. Shahedur Rahman, Md. Anowar Khasru Parvez.

Data curation: Md. Shahedur Rahman.

Formal analysis: Md. Ebrahim Khalil Shimul.

Investigation: Md. Shahedur Rahman, Md. Ebrahim Khalil Shimul.

Methodology: Md. Shahedur Rahman, Md. Anowar Khasru Parvez.

Software: Md. Shahedur Rahman, Md. Ebrahim Khalil Shimul.

Supervision: Md. Shahedur Rahman.

Validation: Md. Shahedur Rahman.

Visualization: Md. Shahedur Rahman.

Writing – original draft: Md. Ebrahim Khalil Shimul.

Writing – review & editing: Md. Shahedur Rahman, Md. Anowar Khasru Parvez.

References

1. Buchholz J., et al., CO2/HCO3− perturbations of simulated large scale gradients in a scale-down device cause fast transcriptional responses in Corynebacterium glutamicum. Applied microbiology and biotechnology, 2014. 98(20): p. 8563–8572.
2. Käß F., et al., Assessment of robustness against dissolved oxygen/substrate oscillations for C. glutamicum DM1933 in two-compartment bioreactor. Bioprocess and biosystems engineering, 2014. 37(6): p. 1151–1162. https://doi.org/10.1007/s00449-013-1086-0 PMID: 24218302
3. Krämer R., Secretion of amino acids by bacteria: physiology and mechanism. FEMS Microbiology Reviews, 1994. 13(1): p. 75–93.
4. Hermann T., Industrial production of amino acids by coryneform bacteria. Journal of biotechnology, 2003. 104(1–3): p. 155–172. https://doi.org/10.1016/s0168-1656(03)00149-4 PMID: 12948636
5. Becker J., Rohles C.M., and Wittmann C., Metabolically engineered Corynebacterium glutamicum for bio-based production of chemicals, fuels, materials, and healthcare products. Metabolic engineering, 2018. 50: p. 122–141. https://doi.org/10.1016/j.ymben.2018.07.008 PMID: 30031852
6. Shanmugam S., et al., High-efficient production of biobutanol by a novel Clostridium sp. strain WST with uncontrolled pH strategy. Bioresource technology, 2018. 256: p. 543–547. https://doi.org/10.1016/j.biortech.2018.02.077 PMID: 29486913
7. Shanmugam S., et al., Enhanced bioconversion of hemicellulosic biomass by microbial consortium for biobutanol production with bioaugmentation strategy. Bioresource technology, 2019. 279: p. 149–155. https://doi.org/10.1016/j.biortech.2019.01.121 PMID: 30716607
8. Wendisch V.F., Mindt M., and Pérez-García F., Biotechnological production of mono-and diamines using bacteria: recent progress, applications, and perspectives. Applied microbiology and biotechnology, 2018. 102(8): p. 3583–3594. https://doi.org/10.1007/s00253-018-8890-z PMID: 29520601
9. Lee J.-Y., et al., The actinobacterium Corynebacterium glutamicum, an industrial workhorse. 2016.

10. Croucher N.J., et al., Diversification of bacterial genome content through distinct mechanisms over different timescales. Nature communications, 2014. 5(1): p. 1–12. https://doi.org/10.1038/ncomms6471 PMID: 25407023

11. Zhu A., et al., Inter-individual differences in the gene content of human gut bacterial species. Genome biology, 2015. 16(1): p. 1–13.

12. Levade I., et al., Vibrio cholerae genomic diversity within and between patients. Microbial genomics, 2017. 3(12). https://doi.org/10.1099/mgen.0.000142 PMID: 29306353

13. Chang Q., et al., Genomic epidemiology of meticillin-resistant Staphylococcus aureus ST22 widespread in communities of the Gaza Strip, 2009. Eurosurveillance, 2018. 23(34): p. 1700592. https://doi.org/10.2807/1560-7917.ES.2018.23.34.1700592 PMID: 30153881

14. Jaspers E. and Overmann J.r., Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Applied and environmental microbiology, 2004. 70(8): p. 4831–4839. https://doi.org/10.1128/AEM.70.8.4831-4839.2004 PMID: 15294821

15. Segerman B., The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Frontiers in cellular and infection microbiology, 2012. 2: p. 116. https://doi.org/10.3389/fcimb.2012.00116 PMID: 22973561

16. Land M., et al., Insights from 20 years of bacterial genome sequencing. Functional & integrative genomics, 2015. 15(2): p. 141–161. https://doi.org/10.1007/s10142-015-0433-4 PMID: 25722247

17. Jensen P.R., Natural products and the gene cluster revolution. Trends in microbiology, 2016. 24(12): p. 968–977. https://doi.org/10.1016/j.tim.2016.07.006 PMID: 27491886

18. Alam K., et al., In silico genome mining of potential novel biosynthetic gene clusters for drug discovery from Burkholderia bacteria. Computers in Biology and Medicine, 2022. 140: p. 105046. https://doi.org/10.1016/j.compbiomed.2021.105046 PMID: 34864585

19. Xu F., et al., A genetics-free method for high-throughput discovery of cryptic microbial metabolites. Nature chemical biology, 2019. 15(2): p. 161–168. https://doi.org/10.1038/s41589-018-0193-2 PMID: 30617293

20. Baltz R.H., Gifted microbes for genome mining and natural product discovery. Journal of Industrial Microbiology and Biotechnology, 2017. 44(4–5): p. 573–588. https://doi.org/10.1007/s10295-016-1815-x PMID: 27520548

21. Yang J. and Yang S., Comparative analysis of Corynebacterium glutamicum genomes: a new perspective for the industrial production of amino acids. BMC Genomics, 2017. 18(1): p. 940. https://doi.org/10.1186/s12864-016-3255-4 PMID: 28198668

22. Dey S., et al., Unravelling the Evolutionary Dynamics of High-Risk Klebsiella pneumoniae ST147 Clones: Insights from Comparative Pangenome Analysis. Genes, 2023. 14(5): p. 1037. https://doi.org/10.3390/genes14051037 PMID: 37239397

23. Yoon S.-H., et al., A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie Van Leeuwenhoek, 2017. 110(10): p. 1281–1286. https://doi.org/10.1007/s10482-017-0844-4 PMID: 28204908

24. Gui Y., et al., Complete genome sequence of Corynebacterium glutamicum CP, a Chinese l-leucine producing strain. Journal of Biotechnology, 2016. 220: p. 64–65. https://doi.org/10.1016/j.jbiotec.2016.01.010 PMID: 26784991

25. Ma Y., et al., Comparative Genomic and Genetic Functional Analysis of Industrial L-Leucine–and L-Valine–Producing Corynebacterium glutamicum Strains. 2018.

26. Nishio Y., et al., Analysis of strain-specific genes in glutamic acid-producing Corynebacterium glutamicum ssp. lactofermentum AJ 1511. The Journal of General and Applied Microbiology, 2017. 63(3): p. 157–164. https://doi.org/10.2323/jgam.2016.09.004 PMID: 28392541

27. Kawaguchi H., Sazuka T., and Kondo A., Complete and draft genome sequences of amino acid-producing Corynebacterium glutamicum strains ATCC 21799 and ATCC 31831 and their genomic islands. Microbiology Resource Announcements, 2020. 9(32): p. e00430–20. https://doi.org/10.1128/MRA.00430-20 PMID: 32763926

28. Wu Y., et al., Complete genome sequence of Corynebacterium glutamicum B253, a Chinese lysine-producing strain. Journal of Biotechnology, 2015. 207: p. 10–11. https://doi.org/10.1016/j.jbiotec.2015.04.018 PMID: 25953304

29. Yukawa H., et al., Comparative analysis of the Corynebacterium glutamicum group and complete genome sequence of strain R. Microbiology, 2007. 153(4): p. 1042–1058. https://doi.org/10.1099/mic.0.2006/003657-0 PMID: 17379713

30. Meng L., et al., Enhancement of heterologous protein production in corynebacterium glutamicum via atmospheric and room temperature plasma mutagenesis and high-throughput screening. Journal of Biotechnology, 2021. 339: p. 22–31. https://doi.org/10.1016/j.jbiotec.2021.07.010 PMID: 34311028

31. Baumgart M., et al., Construction of a prophage-free variant of Corynebacterium glutamicum ATCC 13032 for use as a platform strain for basic research and industrial biotechnology. Applied and environmental microbiology, 2013. 79(19): p. 6006–6015. https://doi.org/10.1128/AEM.01634-13 PMID: 23892752

32. Lee J.-Y., et al., Adaptive evolution of Corynebacterium glutamicum resistant to oxidative stress and its global gene expression profiling. Biotechnology letters, 2013. 35(5): p. 709–717. https://doi.org/10.1007/s10529-012-1135-9 PMID: 23288296

33. Marques F., Luzhetskyy A., and Mendes M.V., Engineering Corynebacterium glutamicum with a comprehensive genomic library and phage-based vectors. Metabolic Engineering, 2020. 62: p. 221–234. https://doi.org/10.1016/j.ymben.2020.08.007 PMID: 32827704

34. Baumgart M., et al., Corynebacterium glutamicum chassis C1*: building and testing a novel platform host for synthetic biology and industrial biotechnology. ACS synthetic biology, 2018. 7(1): p. 132–144. https://doi.org/10.1021/acssynbio.7b00261 PMID: 28803482

35. Linder M., et al., Construction of an IS-Free Corynebacterium glutamicum ATCC 13 032 Chassis Strain and Random Mutagenesis Using the Endogenous ISCg1 Transposase. Frontiers in bioengineering and biotechnology, 2021. 9.

36. Park J., et al., Accelerated growth of Corynebacterium glutamicum by up-regulating stress-responsive genes based on transcriptome analysis of a fast-doubling evolved strain. 2020.

37. Park S.H., et al., Metabolic engineering of Corynebacterium glutamicum for L-arginine production. Nature communications, 2014. 5(1): p. 1–9. https://doi.org/10.1038/ncomms5618 PMID: 25091334

38. Yang J. and Yang S., Comparative analysis of Corynebacterium glutamicum genomes: a new perspective for the industrial production of amino acids. BMC genomics, 2017. 18(1): p. 1–13.

39. Ma W., et al., Poly (3-hydroxybutyrate-co-3-hydroxyvalerate) co-produced with l-isoleucine in Corynebacterium glutamicum WM001. Microbial cell factories, 2018. 17(1): p. 1–12.

40. Richardson E.J. and Watson M., The automatic annotation of bacterial genomes. Briefings in bioinformatics, 2013. 14(1): p. 1–12. https://doi.org/10.1093/bib/bbs007 PMID: 22408191

41. Seemann T., Prokka: rapid prokaryotic genome annotation. Bioinformatics, 2014. 30(14): p. 2068–2069. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

42. Rho M., Tang H., and Ye Y., FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research, 2010. 38(20): p. e191–e191. https://doi.org/10.1093/nar/gkq747 PMID: 20805240

43. Page A.J., et al., Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics, 2015. 31 (22): p. 3691–3693. https://doi.org/10.1093/bioinformatics/btv421 PMID: 26198102

44. Chaudhari N.M., Gupta V.K., and Dutta C., BPGA- an ultra-fast pan-genome analysis pipeline. Scientific Reports, 2016. 6(1): p. 24373. https://doi.org/10.1038/srep24373 PMID: 27071527

45. Tatusov R.L., et al., The COG database: an updated version includes eukaryotes. BMC bioinformatics, 2003. 4(1): p. 41. https://doi.org/10.1186/1471-2105-4-41 PMID: 12969510

46. Kanehisa M., et al., KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research, 2016. 45(D1): p. D353–D361. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662

47. Edgar R., Usearch, 2010, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

48. Liu R., et al., Comparative genomics reveals intraspecific divergence of Acidithiobacillus ferrooxidans: insights from evolutionary adaptation. Microbial Genomics, 2023. 9(6): p. 001038. https://doi.org/10.1099/mgen.0.001038 PMID: 37285209

49. Price M.N., Dehal P.S., and Arkin A.P., FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one, 2010. 5(3): p. e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

50. Letunic I. and Bork P., Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research, 2021. 49(W1): p. W293–W296. https://doi.org/10.1093/nar/gkab301 PMID: 33885785

51. Machado H., et al., Genome mining reveals unlocked bioactive potential of marine Gram-negative bacteria. BMC genomics, 2015. 16(1): p. 1–12.

52. Blin K., et al., antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research, 2021. 49(W1): p. W29–W35. https://doi.org/10.1093/nar/gkab335 PMID: 33978755

53. Skinnider M.A., et al., Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nature communications, 2020. 11(1): p. 1–9. https://doi.org/10.1038/s41467-019-13993-7

54. van Heel A.J., et al., BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic acids research, 2018. 46(W1): p. W278–W281. https://doi.org/10.1093/nar/gky383 PMID: 29788290

55. Salamzade R., et al., Evolutionary investigations of the biosynthetic diversity in the skin microbiome using lsaBGC. Microbial Genomics, 2023. 9(4). https://doi.org/10.1099/mgen.0.000988 PMID: 37115189

56. Eddy S.R., Profile hidden Markov models. Bioinformatics (Oxford, England), 1998. 14(9): p. 755–763. https://doi.org/10.1093/bioinformatics/14.9.755 PMID: 9918945

57. Alikhan N.-F., et al., BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics, 2011. 12(1): p. 402. https://doi.org/10.1186/1471-2164-12-402 PMID: 21824423

58. Darling A.C., et al., Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research, 2004. 14(7): p. 1394–1403. https://doi.org/10.1101/gr.2289704 PMID: 15231754

59. Seemann T., Snippy: rapid haploid variant calling and core SNP phylogeny. GitHub. Available at: github. com/tseemann/snippy, 2015.

60. Zhu Q., Kosoy M., and Dittmar K., HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. BMC Genomics, 2014. 15(1): p. 717. https://doi.org/10.1186/1471-2164-15-717 PMID: 25159222

61. Cosentino S., et al., PathogenFinder-distinguishing friend from foe using bacterial whole genome sequence data. PLOS ONE, 2013. 8(10): p. e77302. https://doi.org/10.1371/journal.pone.0077302 PMID: 24204795

62. Tettelin H., et al., Comparative genomics: the bacterial pan-genome. Current Opinion in Microbiology, 2008. 11(5): p. 472–477. https://doi.org/10.1016/j.mib.2008.09.006 PMID: 19086349

63. Hyun J.C., Monk J.M., and Palsson B.O., Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. BMC Genomics, 2022. 23(1): p. 7. https://doi.org/10.1186/s12864-021-08223-8 PMID: 34983386

64. Rajput A., et al., Pangenome analysis reveals the genetic basis for taxonomic classification of the Lactobacillaceae family. Food microbiology, 2023. 115: p. 104334. https://doi.org/10.1016/j.fm.2023.104334 PMID: 37567624

65. Elbir H., Almathen F., and Almuhasen F.M., Genomic differences among strains of Corynebacterium cystitidis isolated from uterus of camels. The Journal of Infection in Developing Countries, 2022. 16 (01): p. 134–146. https://doi.org/10.3855/jidc.15023 PMID: 35192531

66. Pal S., et al., Comparative evolutionary genomics of Corynebacterium with special reference to codon and amino acid usage diversities. Genetica, 2018. 146(1): p. 13–27. https://doi.org/10.1007/s10709-017-9986-6 PMID: 28921302

67. Ali A., et al., Microbial comparative genomics: an overview of tools and insights into the genus Corynebacterium. J Bacteriol Parasitol, 2013. 4(167): p. 2.

68. Levesque S., et al., Mobilome of Brevibacterium aurantiacum sheds light on its genetic diversity and its adaptation to smear-ripened cheeses. Frontiers in microbiology, 2019. 10: p. 1270. https://doi.org/10.3389/fmicb.2019.01270 PMID: 31244798

69. Brune I., et al., The individual and common repertoire of DNA-binding transcriptional regulators of Corynebacterium glutamicum, Corynebacterium efficiens, Corynebacterium diphtheriae and Corynebacterium jeikeium deduced from the complete genome sequences. BMC genomics, 2005. 6(1): p. 1–10. https://doi.org/10.1186/1471-2164-6-86 PMID: 15938759

70. Araújo C.L., et al., In silico functional prediction of hypothetical proteins from the core genome of Corynebacterium pseudotuberculosis biovar ovis. PeerJ, 2020. 8: p. e9643. https://doi.org/10.7717/peerj.9643 PMID: 32913672

71. Pohl S., et al., The extensive set of accessory Pseudomonas aeruginosa genomic components. FEMS microbiology letters, 2014. 356(2): p. 235–241. https://doi.org/10.1111/1574-6968.12445 PMID: 24766399

72. Santos M.A., et al., Driving change: the evolution of alternative genetic codes. TRENDS in Genetics, 2004. 20(2): p. 95–102. https://doi.org/10.1016/j.tig.2003.12.009 PMID: 14746991

73. Zhang R. and Zhang C.-T., A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. Bioinformatics, 2004. 20(5): p. 612–622. https://doi.org/10.1093/bioinformatics/btg453 PMID: 15033867

74. Osawa S. and Jukes T.H., Evolution of the genetic code as affected by anticodon content. Trends in Genetics, 1988. 4(7): p. 191–198. https://doi.org/10.1016/0168-9525(88)90075-3 PMID: 3070867

75. Osawa S., et al., Recent evidence for evolution of the genetic code. Microbiological reviews, 1992. 56 (1): p. 229–264. https://doi.org/10.1128/mr.56.1.229-264.1992 PMID: 1579111

76. Nasim F., Dey A., and Qureshi I.A., Comparative genome analysis of Corynebacterium species: The underestimated pathogens with high virulence potential. Infection, Genetics and Evolution, 2021. 93: p. 104928. https://doi.org/10.1016/j.meegid.2021.104928 PMID: 34022437

77. Bentley S.D., et al., Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). Nature, 2002. 417(6885): p. 141–147. https://doi.org/10.1038/417141a PMID: 12000953

78. Challis G.L. and Ravel J., Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. FEMS microbiology letters, 2000. 187(2): p. 111–114. https://doi.org/10.1111/j.1574-6968.2000.tb09145.x PMID: 10856642

79. Pawlik K., et al., A cryptic type I polyketide synthase (cpk) gene cluster in Streptomyces coelicolor A3 (2). Archives of microbiology, 2007. 187(2): p. 87–99. https://doi.org/10.1007/s00203-006-0176-7 PMID: 17009021

80. Mattheus W., et al., Isolation and purification of a new kalimantacin/batumin-related polyketide antibiotic and elucidation of its biosynthesis gene cluster. Chemistry & biology, 2010. 17(2): p. 149–159. https://doi.org/10.1016/j.chembiol.2010.01.014 PMID: 20189105

81. Liu L., et al., Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy. Microbiome, 2021. 9(1): p. 1–15.

82. Zheng Y., et al., Genome features and secondary metabolites biosynthetic potential of the class Ktedonobacteria. Frontiers in microbiology, 2019. 10: p. 893. https://doi.org/10.3389/fmicb.2019.00893 PMID: 31080444

83. Jensen P.R., et al., Species-specific secondary metabolite production in marine actinomycetes of the genus Salinispora. Applied and environmental microbiology, 2007. 73(4): p. 1146–1152. https://doi.org/10.1128/AEM.01891-06 PMID: 17158611

84. Xin B., et al., The Bacillus cereus group is an excellent reservoir of novel lanthipeptides. Applied and environmental microbiology, 2015. 81(5): p. 1765–1774. https://doi.org/10.1128/AEM.03758-14 PMID: 25548056

85. Kamigiri K., et al., Kalimantacins A, B and C, novel antibiotics from Alcaligenes sp. YL-02632S I. Taxonomy, fermentation, isolation and biological properties. The Journal of Antibiotics, 1996. 49(2): p. 136–139.

86. Tokunaga T., et al., Kalimantacin A, B, and C, novel antibiotics produced by Alcaligenes sp. YL-02632S II. Physico-chemical properties and structure elucidation. The Journal of Antibiotics, 1996. 49(2): p. 140–144.

87. Smirnov V., et al., Isolation of highly active strain producing the antistaphylococcal antibiotic batumin. Prikladnaia Biokhimiia i Mikrobiologiia, 2000. 36(1): p. 55–58.

88. Kallscheuer N., et al., Microbial synthesis of the type I polyketide 6-methylsalicylate with Corynebacterium glutamicum. Applied microbiology and biotechnology, 2019. 103(23): p. 9619–9631. https://doi.org/10.1007/s00253-019-10121-9 PMID: 31686146

89. Mora-Lugo R., Stegmüller J., and Mack M., Metabolic engineering of roseoflavin-overproducing microorganisms. Microbial cell factories, 2019. 18(1): p. 1–13.

90. Chikindas M.L., et al., Functions and emerging applications of bacteriocins. Current opinion in biotechnology, 2018. 49: p. 23–28. https://doi.org/10.1016/j.copbio.2017.07.011 PMID: 28787641

91. Drider D., et al., Bacteriocins: not only antibacterial agents. Probiotics and antimicrobial proteins, 2016. 8(4): p. 177–182. https://doi.org/10.1007/s12602-016-9223-0 PMID: 27481236

92. Zotchev S.B., Genomics-based insights into the evolution of secondary metabolite biosynthesis in actinomycete bacteria, in Evolutionary biology: genome evolution, speciation, coevolution and origin of life2014, Springer. p. 35–45.

93. Cimermancic P., et al., Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell, 2014. 158(2): p. 412–421. https://doi.org/10.1016/j.cell.2014.06.034 PMID: 25036635

94. Belknap K.C., et al., Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. Scientific reports, 2020. 10(1): p. 1–9.

95. Gallagher K.A. and Jensen P.R., Genomic insights into the evolution of hybrid isoprenoid biosynthetic gene clusters in the MAR4 marine streptomycete clade. BMC genomics, 2015. 16(1): p. 1–13.

96. Khaldi N., et al., Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome biology, 2008. 9(1): p. 1–10. https://doi.org/10.1186/gb-2008-9-1-r18 PMID: 18218086

97. Hollensteiner J., et al., Pan-genome analysis of six complete Paracoccus type strain genomes from hybrid next generation sequencing. bioRxiv, 2023: p. 2023.06. 19.545646.