Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

Article

A genomic comparative analysis of *Bacillus coagulans* unravels the genetic potential of MA-13 strain for biotechnological applications

Martina Aulitto [1,2], Laura Martinez Alvarez[3], Gabriella Fiorentino [1,4,5], Danila Limauro[1,4,5], Xu Peng[3] and Patrizia Contursi [1,4,5]*

[1] Dipartimento di Biologia, Università degli Studi di Napoli Federico II, Naples, Italy,
[2] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States;
[3] Archaea Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[4] BAT Center—Interuniversity Center for Studies On Bioinspired Agro-Environmental Technology, University of Napoli Federico II, Portici, NA, Italy
[5] Task Force On Microbiome Studies, University of Naples Federico II, Naples, Italy
* Correspondence: contursi@unina.it ;

Abstract: The production of bio-chemicals requires the use of microbial strains with efficient substrate conversion and excellent environmental robustness, such as *Bacillus coagulans* spp. So far the genomes of about 50 strains have been sequenced. Herein, we report a comparative genomic analysis of nine strains on the full repertoire of CAZymes, secretion systems, and resistance mechanisms to environmental challenges. Moreover, *B. coagulans* Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) immune system along with CRISPR-associated Cas) genes, was also analysed. Overall, this study expands our understanding of the strains genomic diversity of *B. coagulans* to fully exploit its potential in biotechnological applications.

Keywords: *B. coagulans*; CAZymes; Bacteriocins; CRISPR-Cas systems

1. Introduction

Wide substrate efficiency and excellent environmental robustness are indispensable features for the microbial-based production of bio-chemicals [1,2]. For instance, bacterial fermentation is currently used as an eco-sustainable alternative to obtain lactic acid (LA) from raw materials [3]. So far several lactic-acid producer strains of *B. coagulans* have been used for the production of LA from lignocellulosic biomasses [4]. Interestingly, *B. coagulan*s grows at temperatures ranging from 50 to 55° and pH 5-6 that are chemico-physical conditions comparable to those used for the biomass saccharification [5–7]. Therefore *B. coagulans* represents an attractive biocatalyst for LA production because of its thermophilic nature that not only reduces the risk of microbial contamination but also makes it possible to develop efficient fermentation processes, i.e, simultaneous saccharification and co-fermentation (SSF) [6,8]. Recently, a novel *B. coagulans* strain, designed as MA-13 and isolated from canned beans manufacturing, has been tested in an SSF configuration for efficient production of LA from wheat straw [5]. Moreover, it turned out to be exceptionally resistant to extreme conditions, such as the presence of toxic compounds derived from the thermo-acidic treatment of lignocellulose pointing to this microbe as a suitable biological tool for the eco-friendly production of LA [6].

B. coagulans is a homofermentative microorganism and glucose is converted into lactate usually via the Embden–Meyerhof–Parnas (EMP) pathway and xylose via the pentose phosphate pathway (PPP) or the phosphoketolase pathway (PKP). Specifically, the yield of LA varies from 60% to 98% of the total fermentation products depending on whether xylose is metabolized through one or the other pathway [9,10]. Interestingly,

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

genomic comparison among different *B. coagulans* strains have revealed that some of them lack at least some of the genetic determinants of the pentose phosphate pathway, hinting at a significant metabolic and genomic diversity among the isolates [4,7,9]. On the other hand, knowledge of the genomic features of the different *B. coagulans* strains is essential to fully exploit their potential in white biotechnology applications [11]. Indeed, *B. coagulans* is also emerging as a promising GRAS (Generally Regarded As Safe) probiotic candidate, by sharing characteristics with both *Bacillaceae* and *Lactobacillaceae* [12]. Currently, *B. coagulans* strains are widely used as probiotics in marketed worldwide formulations namely, Sustenex, Lactospore, Neolactoflorene, etc [13] In a recent study it has been demonstrated that MA-13 over-produces under standard growth conditions α- and β-galactosidases that are key enzymes for improving the nutritional value of raffinose family oligosaccharides and lactose-containing food as well as for the production of galactooligosaccharides (GOS) that improve the digestibility and nutritional value of food. Moreover, the identification and comparative analysis of the full repertoire of Carbohydrate-Active enZymes (CAZymes) of other *B. coagulans* strains might help to shed light on the probiotic and prebiotic properties of the available strains [14].

Furthermore, *B. coagulans* has been exploited also in other diverse biotechnological applications in medicine, food, and chemical industry including the production of thermostable enzymes and antimicrobial peptides, such as coagulin [15].

So far many strains have been isolated and for most of them, the genome sequence is available [7,16–18]. To fully explore the biotechnological potential of MA-13, a comparative analysis among some strains has been carried out. Genome decoding is functional to unravel molecular mechanisms driving diversification, genetic variability, and host-pathogen interactions [19]. Previously, it has been reported comparative genomic analyses on a few *B. coagulans* strains with a special focus on carbohydrates catabolism and genetic determinants responsible for general defense systems from foreign DNA, have been reported [4]. In our analysis, we picked up the closest evolutionary related strains to MA and pinpoints the core and the dispensable genes that are present in some but not all the strains studied (non-core) as well as the strain-specific genes (singletons) related to CAZymes, protein secretion systems, antimicrobial peptides/proteins, secondary metabolites, and the defense systems protecting cells from foreign DNA. Moreover, we show that the genomic content varied among close individual strains, and such variations are traced back to the gain or loss of singleton, genomic islands, selfish DNA (plasmids, bacteriophages, integrative conjugative elements), and/or widespread horizontal gene transfer (HGT) [20].

2. Results and discussion

2.1. Genomic features of B. coagulans strains

A draft of the MA-13 genome was recently published [21] supplying a general overview of its genetic content. In this work, we provide a comparative analysis among MA-13 and other phylogenetically related strains, with the purpose to highlight the genomic features of MA-13 that point up its potential in different biotechnological fields. To this aim, the genome sequence of MA-13 was re-assembled using different tools (data not shown) and according to the quality check (QUAST), the best result was obtained using MEGAHIT. The updated genome assembly was also deposited in NCBI under the same accession number (WGS: SMSP02000001-SMSP02000116, ACCESSION SMSP00000000) and consists of 125 contigs, with a total length of 2,942,169 bp and an average G+C content of 47.2%. Compared to the previous work [21], the number of contigs was drastically reduced (from 1,653 to 125 contigs) and the total length adjusted from 3,237,270 to 2,942,169 bp. To date, forty-seven *B. coagulans* strains have been deposited in the NCBI database and their genomic features are summarized in Table S1. Only fourteen strains were reported to be sequenced as single replicons, while the others are deposited as draft genomes. In general *B. coagulans* genome size varies from 2.86 to 3.69 Mbp with an average size set around 3.0 Mbp as described for MA-13 as well as for H1, XZL4, AF24-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

21, AF24-19, MGYG-HGUT-00191, B4098, DSM1 strains (Table S1). The number of CDS ranges from 2,536 to 3,237 and the GC% content is around 46-47% (Table S1).

2.2. Phylogenetic and comparative genomic analysis

In microbial genomics, phylogenetic analysis is crucial not only to establish the genetic novelty and the genotype-phenotype relationships of the isolates but also to identify the closest relatives (micro)organisms within assembled genomes. With this aim, a species tree was generated using the genomes of nine *B. coagulans* strains available in the KBase system.

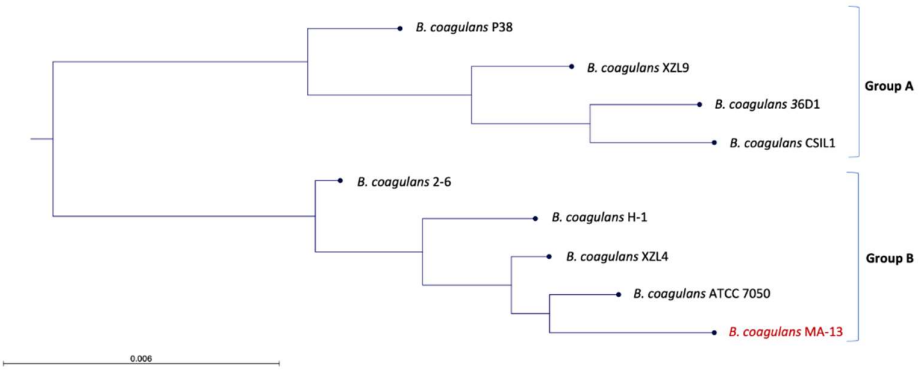

Figure 1. Phylogenetic tree based on the genome comparison of the nine strains of *B.coagulans*.

In our previous work, a phylogenetic tree was built based on 16S rRNA genes and two main groups were identified [5]. In this study, the comparative whole-genome sequence analysis highlighted the presence of two clades (named Group A and B) that underline the coexistence of two distinct evolutionary lineages. This hypothesis is supported by the similar genome size within each clade (~2.9 Mbp for Group A and ~3.4 Mbp for Group B, respectively) (Figure 1, Table S1). Moreover, the closest strain to MA-13 is DSM1 (accession number ALAS00000000) which was also isolated from agri-food wastes [18]. Then, we resolved to calculate the Average Nucleotide Identity (ANI) value of orthologous genes shared by the two genomes. The FastANI visualization of the reciprocal mappings shows the close evolutionary relationship between the two strains as clearly highlighted by the dark-red lines representing all the conserved regions shared by MA-13 and DSM1 which define an ANI value of 98.43%. Moreover, as further evidence of the evolutionary connection, we have calculated the DDH (DNA-DNA hybridization) values to demonstrate that the two microorganisms unambiguously belong to the same specie (DDH1=98.96%) but they are not members of the same sub-specie (DDH2= 60.49%).

To get a deeper insight into the genomic variation among the nine strains a comparison of all the genes grouped by sequence homology was carried out [22]. This analysis was performed using the phylogenetic tree (Figure 1) as input to build up a framework for exploring the genomic diversity and the common features through the analysis of non-core genes/singletons and core genes (Figure 2, Table S2). The formers are represented by all genes not universally conserved within a species, whilst the core is constituted by genes conserved across all the genomes [23].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

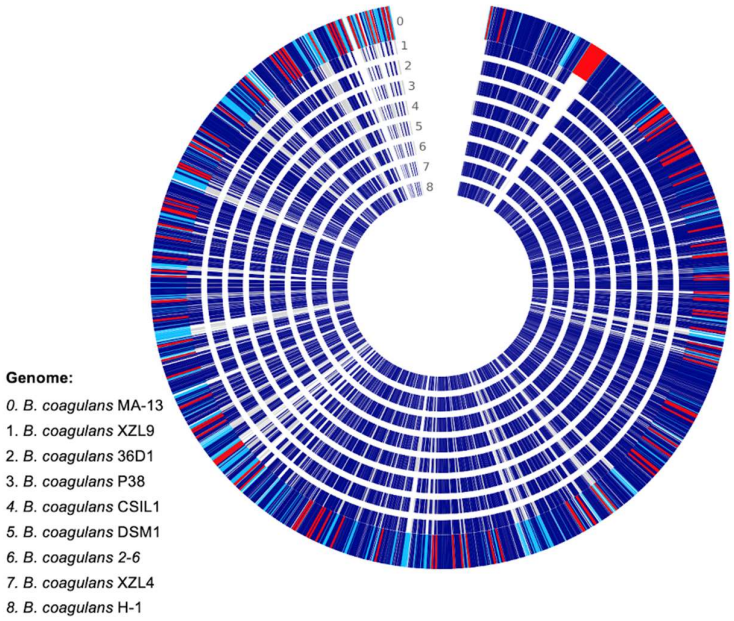

Figure 2. Comparative genomic analysis of *B. coagulans* MA-13 and the closest eight relatives. The circular plot includes white gray (non-core genes), dark blue (core genes), and singletons in red.

The genomic diversity and variability among the isolates are highlighted by the analysis of singletons and non-core genes that may be acquired from distal lineages through horizontal gene transfer and represent the genetic source for the emergence of novel functions (Figure 2, Table S2). To assess the functional features of singletons, TIGRFAM was used to cluster proteins in categories and sub-categories (Figure 3). Genes involved in central metabolism as well as several hypothetical genes and some with unknown functions are distributed over all the genomes. Moreover, a distinctive feature is the presence of prophage elements that are found only in MA-13, DSM1, and P38 (Figure 3). These strains are also characterized by singletons related to the functions of mobile genetic elements. This feature points to the availability of the genes attainable by the evolving microbial population and to the capability of microbial genomes to acquire new functions. The distribution of singletons in MA-13 strain revealed a significant proportion (~ 40%) of them falling within the hypothetical proteins category with unknown function thus indicating that a significant portion of them is still "terra incognita" and therefore the full understanding of its genomic potential needs to be still unveiled (Table S3). Moreover, transposable elements provide microorganisms with the ability to be responsive and susceptible to environmental changes by acquiring new genetic material and disseminating regulatory elements. The presence of singletons related to transposon activity, which is a distinctive feature of MA-13, points to significant genome plasticity, possibly to suit the requirements of environmental adaptation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

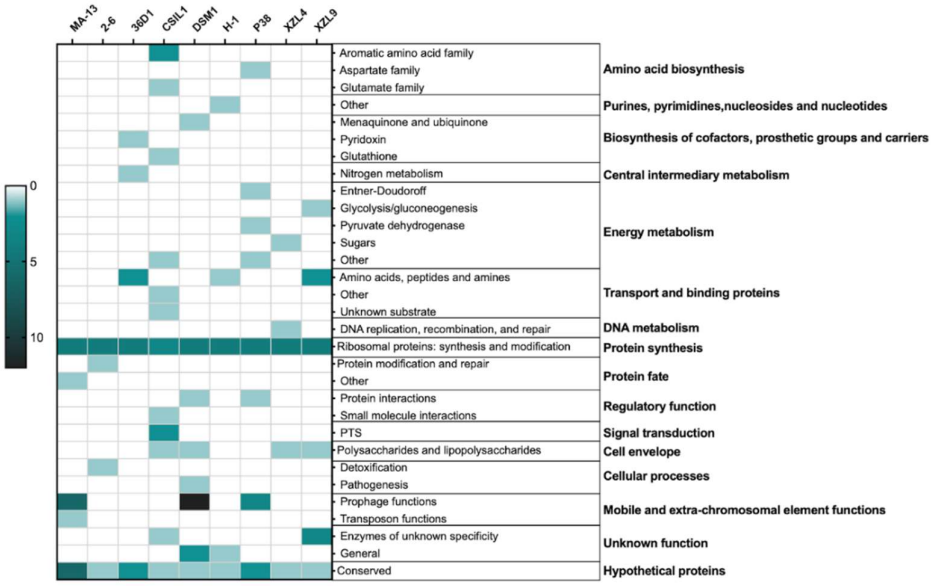

Figure 3. Heat-map of singletons.

2.3. CAZome of B. coagulans

CAZymes (Carbohydrate-Active enZYmes) repertoire covers diverse enzymatic functions related to the synthesis, modification, and breakdown of saccharides, thus exploiting a great potential in biotechnological applications. *B. coagulans* CAZome was explored through dbCAN HMMs and subsequently matched with HMMER 3 for functional annotation (Table 1) [24].

Table 1. CAZymes predicted in the genomes of *B.coagulans* strains. The total number of genes is based on the data obtained by NCBI* and dbCAN.

Strain	Total No. of genes*	No. of CAZymes	% CAZymes
MA-13	2689	35	1.30
DSM1	2766	37	1.34
36D1	3128	42	1.34
XZL9	3112	40	1.29
P38	3097	41	1.32
2-6	2743	33	1.20
H-1	2536	28	1.10
XZL4	2593	41	1.58
CSIL1	3165	47	1.48

The total amount of CAZymes in the nine genomes (344 genes in total) represents at least ~1% of all predicted coding sequences (Table 1). A deeper inspection of the CAZome is reported in Figure 4. The annotated genes were clustered in glycoside hydrolases (51.2 %), glycosyltransferases (23.2 %), carbohydrate esterases (17.1 %), and auxiliary activity (8.4%) families.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

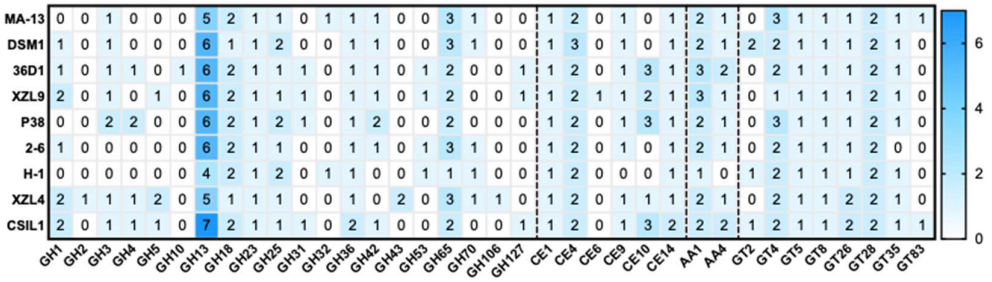

Figure 4. Heatmap showing the genome-wide distribution of CAZymes in the *B.coagulans* strains. A total of 36 families were identified.

Glycoside Hydrolases (GHs) catalyze the hydrolytic cleavage of the glycosidic bond and are involved in the degradation of agri-food and lignocellulose biomasses [25–29]. A total of 176 genes encoding GHs are clustered in 20 families, among which some are spread over all the compared genomes, i.e. GH13, GH18, GH23, GH25, GH36, GH65. The highest number of sequences (51 and 21, respectively) were found within GH13 and GH65 families whose members are mainly active on substrates containing α-glucoside linkages, such as starch [30]. GH42 representatives are found in almost all the strains except for H-1 and XZL4. Sequence analysis and experimental data indicate that the GH42 enzymes are mostly beta-galactosidases which are key providers of energy and carbons source through the breakdown of lactose to galactose and glucose [31]. GHs members belonging to families 18 might be involved in the sporulation pathway and this observation is in line with the lifestyle of *B. coagulans* members. Interestingly, four GH5 (in *B. coagulans* XZL4, XZL9, CSIL1) and one GH10 representative (in *B. coagulans* 36D1) were annotated. Both families include enzymes potentially acting on cellulose and hemicellulose, respectively [32,33]. In particular, the GH10 member was retrieved in the singletons list, suggesting that this enzyme was acquired through a 'genetic crosstalk' and/or horizontal gene transfer among bacterial genomes.

Glycosyl Transferases (GTs) are enzymes that catalyse the transfer of sugar moieties from activated donors to specific acceptor molecules, forming glycosidic bonds, and are involved in the synthesis of oligosaccharides, polysaccharides, and glycoconjugates [34]. Typically, the bacterial glycosyltransferases are poly-specific enzymes that mainly belong to families GT2 and GT4 and exploit the inverting and retaining catalytic mechanism, respectively. Except for GT2, GT35, and GT83, at least one member of the remaining identified GTs families is found in all the genomes. Moreover, GT4 and GT83 are the largest and the smallest groups, respectively. Noteworthy, the GT83 family which includes enzymes involved in lipopolysaccharide biosynthesis [35], typifies the GT repertoire of only two strains, i.e. MA-13 and CSIL1. Finally, family GT28 related to peptidoglycan biosynthesis, maintenance of the rod cell shape, and elongation of the lateral cell, is represented by two members for each genome.

Carboxyl Esterases (CEs) release acyl or alkyl groups attached by an ester linkage to carbohydrates [36]. Among the 18 CEs families, only six of them were annotated in this analysis. CE1, CE4, and CE14 are found in all the strains whilst a singleton encoding for a putative acetyl xylan esterase (CE6) was found in XZL9.

The remaining enzymes are categorized as Auxiliary Activities (AAs) in the CAZy database and comprise carbohydrate oxidases, redox enzymes involved in diverse functions such as lignin degradation and metabolism of lignin-derived compounds [37]. AA members (in total 29) were assigned to AA1 and AA4 families; the former are multicopper oxidases that use diphenols and related substances as donors and oxygen as acceptors [38], while the latter are potential vanillyl-alcohol oxidases [39].

The CAZome analysis along with the available biochemical and physiological data in the literature points to *B. coagulans* as model microorganism not only for carbohydrate degradation processes but also as a valuable source of novel enzymes for potential

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

applications especially in food and medical industries. Specifically, MA-13 is currently tested for its ability to use agri-food wastes as carbon and energy sources (manuscript in preparation), showing excellent degradative capability on diverse biomasses. The absence of enzymatic activity related to cellulose degradation can be used for applications aimed at the isolation of cellulose from agri-food wastes.

2.4. Protein secretion systems

Bacterial cells rely on a plethora of secretion systems to dynamically interact with their environment. Two main types of protein secretion systems, the Sec and Tat pathways are the most common ones and have been identified in all domains of life [40]. Both Tat and Sec systems are present in *B. coagulans* as shown by the core feature set that includes *tatC*/*tatA* and *secYEG* genes [4]. Moreover, the comparative analysis of non-core genes revealed that *B. coagulans* exploits other less widespread secretion systems (Table 2). Indeed, 36D1 and CSL1 bear the *hylD* gene that is a component of the *Escherichia coli* α-hemolysin secretion system (type I secretion system) which is otherwise made up also of *hlyA*, *hlyB,* and *tolC* [41].

Most but not all *B. coagulans* strains are featured by genes of the type VII secretion system (called T7SS) which exports proteins lacking a canonical cleavable signal peptide but bearing a WXG motif at the N-terminus [42]. EsxA and EsxB are molecular markers of the T7SS system and the relative genes (*esxA* and *esxB*) are usually part of gene clusters each harboring a membrane-associated ATPase of the FtsK–SpoIIIE protein family and other putative components of the secretion systems. Only P38 bears a complete set of the main components, whereas all the others show only some of them (Table 2).

Table 2. Non-core genes of *B.coagulans* secretion systems.

Accession number	*B.coagulans* strain	Function
WP_035181994.1	2-6	6 kDa early secretory antigenic target ESAT-6 (EsxA)
WP_041818879.1	2-6	Putative secretion accessory protein EsaA/YueB @ Bacteriophage SPP1 receptor
WP_013858856.1	2-6	FtsK/SpoIIIE family protein, putative EssC/YukB component of Type VII secretion system
WP_035181994.1	36D1	6 kDa early secretory antigenic target ESAT-6 (EsxA)
WP_014096028.1	36D1	secretion protein HlyD
WP_029141776.1	DSM_1	Putative secretion system component EssB/YukC
MBF8418457.1	MA-13	type VII secretion protein EssA
MBF8418459.1	MA-13	type VII secretion protein EssB
WP_026685044.1	CSIL1	secretion protein HlyD
WP_052123334.1	P38	FtsK/SpoIIIE family protein, putative EssC/YukB component of Type VII secretion system
WP_035190278.1	P38	FtsK/SpoIIIE family protein, putative EssC/YukB component of Type VII secretion system
WP_035190280.1	P38	Putative secretion system component EssB/YukC
WP_035190310.1	P38	Putative secretion accessory protein EsaA/YueB @ Bacteriophage SPP1 receptor
WP_035190312.1	P38	6 kDa early secretory antigenic target ESAT-6 (EsxA)

It is tempting to speculate that either some of the T7SS genes are dispensable for leader-less secretion systems or that single components might cross-function with Tat or Sec systems. Interestingly, MA-13 bears only *essA* and *essB*, encoding for membrane proteins usually working in association with *essC* and *esxA* that are instead missing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

Nevertheless, we have demonstrated in a previous work that WXG-proteins even not related to virulence factors are secreted in MA-13 through a leader-less mechanism [14]. In biotechnological applications, it would be useful to facilitate protein production or protein downstream processing using different secretion mechanisms for biotransformations and/or for screening enzyme variants in libraries. The study of the multiple secretion systems functioning in MA-13 along with its fermentation performance [5, 6] makes this host a suitable starting point for the development of a consolidated bioprocess for the production of recombinant proteins/enzymes of biotechnological interest.

2.5. Resistance mechanisms to environmental challenges

2.5.1. Toxin-antitoxin systems

Toxin–antitoxin systems (TAs) are ubiquitous among bacteria and play a crucial role in the dissemination and evolution of antibiotic resistance, for instance through stabilization of multi-resistant mobile genetic elements that inhibit or kill plasmid-free daughter cells [43]. Generally, toxins are proteins that affect metabolism and antitoxins are either RNA or proteins that counteract the effect of the toxin. Currently, TA systems are based on the coordinated action of a toxin which is always a stable protein, while the labile antitoxins can be either RNAs or proteins [44].

All *B. coagulans* strains are endowed with a common TA system retrieved from the core set and which relies on *ydcD* and *ydcE* genes, encoding for both an endoribonuclease that inactivates cellular mRNAs and its inhibitor, respectively. Moreover, we spotlighted in the MA-13 singletons set, the presence of a single gene (MBF8417901.1) encoding for the toxin of the PemK/MazF family [45]. The absence of the antitoxin component suggests that a functional TA system was lost during the evolution. It is tempting to speculate that the TA system based on *ydcD*/*ydcE* genes has been selected to face efficiently the challenge of an ever-changing environment. The TA systems of *B. coagulans* are a platform to discover novel antimicrobial targets, as well as to set up genetic reporter systems for plasmid maintenance and protein production.

2.5.2. Antimicrobial resistance mechanisms

Another strategy to withstand exposure to toxic compounds is antimicrobial resistance (AMR) which refers to the ability of microorganisms to cope with the effects of antibiotic treatments or antimicrobial peptides [46]. To advance phenotype prediction and to identify genomic regions related to AMR, an *in silico* analysis of the "resistome" of *B. coagulans* was performed using PATRIC software. Signatures of antimicrobial target genes were identified, among which those encoding for resistance to fluoroquinolones (*gyrA* and *gyrB*) and (*murA*) were found in MA-13, DSM1, and 2-6 (Table S4). Whilst those conferring resistance to sulfonamides (*folP*) and cycloserine (*Alr*) were present only in MA-13 and DSM1 strains. Interestingly, ORFs encoding for the primary target of the isoniazid (*inhA, kasA*) were only identified in MA-13, together with *gdpD* which is involved in phospholipid metabolism. In general, alteration in composition or distribution of phospholipids in the cell membrane is a common mechanism of resistance to lipopeptides in Gram-positive bacteria. Indeed, mutations in *Enterococcus faecalis gdpD* and *liaFSR* genes have been linked to the development of resistance to the cyclic lipopeptide daptomycin [47]. Then, the presence of the *gdpD* gene in MA-13 might represent a defense weapon against the action of antimicrobial peptides. In antibiotic-producing bacteria, it is known that biosynthetic operons of peptide antibiotics are frequently associated with membrane drug efflux transporters of the ABC family, which pump out the antibiotics for self-protection. Accordingly, all *B. coagulans* strains analyzed bear three genes encoding for efflux pumps (*bceA, bceB, bcrC*) [48]. Furthermore, MA-13 is the only strain with a complete genetic array responsible for resistance to bacitracin including the two-component signal transduction system (*bceRS*) which is adjacent to the efflux transporter genes (*bceA*/*B*) (Figure 5). This genetic proximity is common to other Gram-positive bacteria and the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

relative genes are located on contig k141_1005 in MA-13. In this system, BceR is a sensory transduction protein that responds to extracellular bacitracin stimulus and transmits the signal to the regulator BceS. This latter binds the promoter region upstream to *bceA/B* genes affecting their expression. Furthermore, MA-13 exhibits another genetic determinant of bacitracin resistance, i.e. *bcrC* encoding for a broad-specificity multidrug efflux pump which is located on a genomic region (contig k141_1) distant from the *bceRS/AB* genetic array. MA-13 BcrC protein is homolog (31.3% a.a. sequence identity) to BcrC of *Bacillus licheniformis*, a hydrophobic protein functioning as membrane component of the permease [49]. Differently from *Bacillus licheniformis* which bears a three-components operon (*bcrABC*), MA-13 possesses only the *bcrC* gene (Figure 5).

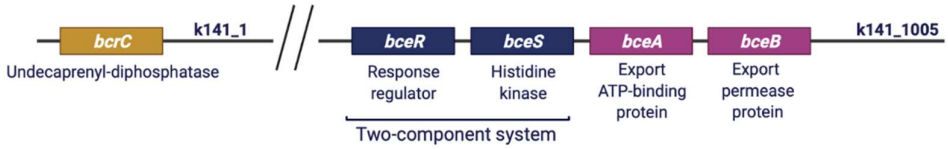

Figure 5. Genetic determinants of bacitracin resistance in MA-13.

Furthermore, the MA-13 genome bears genetic determinants for triclosan (TCS)-resistance, i.e. *fabL* and *fabI* genes, while a more complex cluster was identified in DSM1 including *fabD, fabH, fabI, fabL, fabF* (Table S4) [50]. TCS is a commonly used artificial antimicrobial agent widely spread in the environment. One of the major concerns about TCS is that resistance to this compound may contribute to reduced susceptibility to clinically important antimicrobials, due to either cross-resistance or co-resistance mechanisms. Indeed, bacteria exposed to continuously changing chemico-physical conditions are expected to be endowed with features of better adaptability and competitive fitness to overcome environmental challenges. The horizontal gene transfer (conjugation, transformation, and transduction) and the role played by mobile genetic elements (plasmids, transposons, insertion sequences, integrons, and integrative-conjugative elements) and the bacterial toxin-antitoxin system as well as the occurrence of genetic mutations, lead to a speedy bloom of antibiotic resistance amongst bacteria [20]. MA-13 which is an environmentally isolated strain is a multidrug resistance bacterium containing several diverse antibiotic resistance genes whose characterization might unravel novel mechanisms of resistance to organic compounds [6]. Our previous studies have highlighted the robustness of MA-13 toward biomass-derived inhibitors [5, 6] which allows its propagation in the presence of such toxic compounds without affecting LA production. It is tempting to speculate that MA-13 tolerates the presence of diverse classes of organic toxic compounds through highly efficient detoxification/extrusion systems and therefore is suitable for different biotechnological applications related to the food industry as well as to biomass-degrading processes.

2.5.3. Secondary metabolites: bacteriocins

The discovery of new natural compounds produced by microorganisms, the so-called secondary metabolites, is a hot topic for diverse biotechnological applications [51]. These bioactive molecules, such as pigments, alkaloids, toxins, antimicrobials, are not involved in the basal metabolism, rather they often perform ancillary functions [52]. The ability of *B. coagulans* to hinder bacterial growth and balance co-habitant microbiota populations by producing antimicrobial compounds was already demonstrated in several studies. For instance, the metabolic acidosis associated with the secretion of LA during fermentation is detrimental for acid-sensitive microorganisms.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

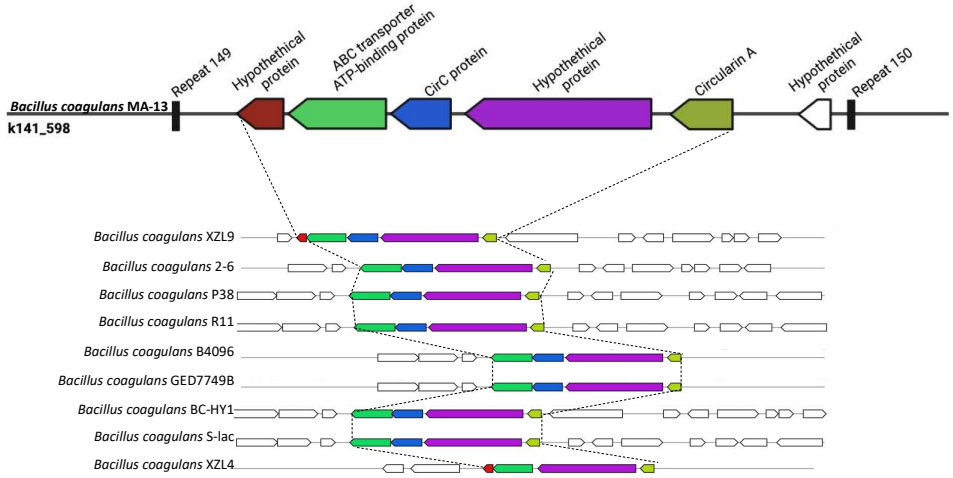

Figure 6. Graphical representation of circularin A gene cluster in diverse *B.coagulans* strains.

To understand the potential of MA-13 in the production of secondary metabolites, the genome was analysed using antiSMASH, a bioinformatic tool suitable to search for clusters involved in secondary metabolites synthesis (Figure 6). From this analysis, a gene encoding for a non-lanthionine-containing peptide (class II bacteriocins), the circularin A, was identified. Circularin A is usually produced as a pre-peptide that undergoes a proteolytic cleavage of the leader sequence followed by head-to-tail ligation between the N- and C-termini to produce a smaller circular antimicrobial peptide [53]. This bacteriocin is usually encoded by a gene cluster or in 4–10 genes containing operon. A genetic arrangement formed by ORFs encoding for circularin A, CircC (see below), ABC transporter, and two hypothetical proteins was found in MA-13 on contig k141_585. A similar cluster structure is shared among almost all the strains analysed (Figure 6), but in the case of MA-13, the region is delimited by two repeats. Moreover, given the presence of several surrounding mobile elements, among which a transposase IS4, it is likely that the acquisition of this genetic cassette occurred upon HGT in MA-13. The strain XZL4 is the only one missing the *circC* gene, whose product has known to have a central role in the maturation of circularin A. Bacteriocin production could be considered advantageous to the producer since these peptides can kill or inhibit bacteria by competing for the same ecological niche or the same nutrient pool. Moreover, probiotics that synthetize bacteriocins are of particular importance because these bacteria are widely used in dairy industries. In this context, MA-13 has been proven to exhibit probiotic and prebiotic properties, because of its ability to express enzymes catalysing the synthesis of GOS as well as to improve the digestibility and nutritional value of foods. These features along with the ability to produce bacteriocins make MA-13 an ideal candidate for exploitation in white biotechnology.

2.6. Innate and adaptive immunity

2.6.1. Innate immunity

Prokaryotic immunity is constituted by multiple systems to defend from mobile genetic elements (i.e. viruses and plasmids). Innate defense mechanisms do not require previous exposure to a pathogen. The examination of the different *B. coagulans* strains revealed the ubiquitous presence of the Wadjet defense immune system in the nine strains analysed (Figure 7 and Table S5) [54]. Wadjet cassettes are composed of the four genes *jetABCD*, where JetABC are homologs of the bacterial condensins MukF, MukE, and MukB, respectively. Bacterial condensins are involved in DNA condensation and segregation during bacterial replication. The fourth component of the operon, JetD, has a putative topoisomerase VI domain. Three types of Wadjet systems have been described, reflecting the domain organization of its components within the cassette [55]. *B. coagulans*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

strains predominantly contain type II Wadjet systems, although strains 36D1, CSIL1, and XZL9 harbor type III Wadjet operons (Figure 7). Wadjet system has been related to antiplasmid defense in the Gram-positives *Bacillus cereus* Q1 and *Mycobacterium smegmatis* [56] where it impairs plasmid transformation and plasmid segregation to daughter cells, respectively. It is tempting to speculate that the presence of Wadjet systems in *B. coagulans* is involved in preventing the acquisition of DNA from the environment. This would hamper the development of genetic systems for these strains, similarly to what occurs in *M. smegmatis,* where disruption of this system allowed transformation and maintenance of plasmid pAL5000 [56]. Other isolated defense systems are present in *B. coagulans* strains. Strain 2-6 harbors a single CBASS (cyclic-oligonucleotide-based anti-phage signaling systems) cassette in addition to a type II Wadjet system (Figure 7) [57]. CBASS systems are constituted by an oligonucleotide cyclase that produces cyclic nucleotide secondary messengers upon the encounter of a pathogen, and an effector protein. They are classified into four types according to their operon organization, type of effector, and cyclic nucleotide messenger. In particular, effectors of type II CBASS systems contain ubiquitin-associated domains and their mechanism of action is still unknown given that ubiquitin has not been found in bacteria. However, this system confers protection against phages in *E. coli* and *V. cholerae.* Strain XZL9 harbors two Wadjet cassettes of types II and III, plus two Gabija loci (Figure 7). The Gabija system is composed of proteins GajA and GajB where GajA is a sequence-specific DNA nicking endonuclease and GajB is predicted to be a UvrD-like helicase. It is hypothesized that GajA nuclease activity is activated by the change in the intracellular nucleotide levels resulting from virus replication and transcription, thus eliciting bacteriophage resistance [58].

Likewise many environmental strains, *B. coagulans* is not amenable to transformation with exogenous DNA and so far only 2 strains (DSM1 and P4-102B) are genetically tractable. However, setting up a reproducible genetic system is indispensable to apply metabolic engineering techniques to meet the requirements of commercial applications. The ubiquitous presence of the putative antiplasmid Wadjet immune system in *B. coagulans* may explain the recalcitrance of these strains to genetic tractability and constitutes a promising candidate for the development of strains amenable to genetic engineering.

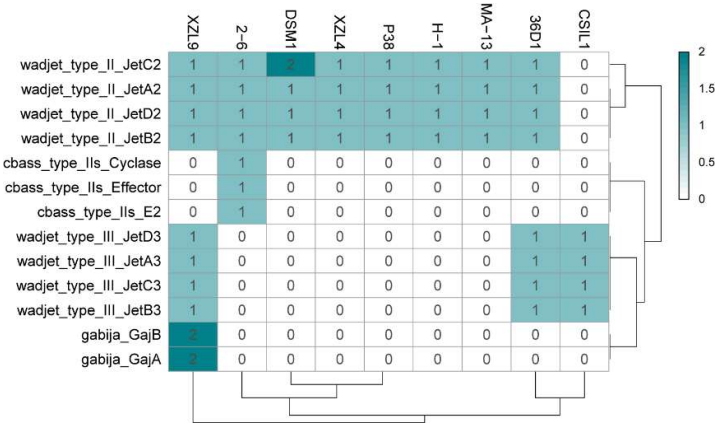

Figure 7. Heat map distribution of innate immunity systems in the *B. coagulans* strains.

2.6.2. Adaptive immunity: CRISPR-Cas system and mobile elements

CRISPR-Cas systems constitute an adaptive immune apparatus in a wide range of bacteria (~40%), and the majority of archaea (~ 90%); small guide RNAs (crRNAs) interfere with sequence-specifically nucleic acids neutralizing invading genetic elements [59]. A CRISPR locus consists of repeats (23-47 bp) separated by unique sequences (spacers) with similar length (21-72 bp), originating from mobile genetic elements such as bacteriophages, transposons, or plasmids. Moreover, the *cas* genes are generally closely

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

linked to CRISPR arrays and are diversified. Indeed, CRISPR-Cas systems are classified into two classes which are further subclassified into six types and several sub-types: in class I, the effector module consists of several Cas proteins whereas class II is defined by the presence of a single multidomain protein (e.g. Cas9). The CRISPR-Cas defense mechanism acts in a sequence-specific manner by recognizing and cleaving foreign DNA or RNA. The defence response can be divided into three stages: (i) adaptation or spacer acquisition, (ii) crRNA biogenesis, and (iii) target interference [60].

Figure 8. CRISPR-Cas cassettes present in *B. coagulans* genomes. The *cas* operons for each strain are depicted: A) MA-13, B) DSM1, C) 2-6, D) H-1, E) P38, F) XZL9, G) 36D1 and H) CSIL1. The headers above each operon refer to the type of *cas* operon, strain, contig, and location. The color of *cas* genes corresponds to their associated stage of the CRISPR-Cas response: blue – adaptation, yellow – interference, and red – crRNA maturation. CRISPR arrays are depicted as black, vertical lines.

MA-13 genome harbors 48 spacers, 51 repeats organized in at least two CRISPR arrays with the same direct repeat sequence. The exact number of arrays cannot be determined due to the fragmented nature of the MA-13 genomic assembly that includes small contigs containing partial, truncated CRISPR arrays. Additionally, these arrays are linked to several unidirectionally arranged *cas* genes organized in two loci (*cas1*, *cas2*, *cas3*, *cas4*, *cas5*, *cas7*, *cas8*) (Figure 8 and Table S6). Such genomic association is typical of Class I-C, characterized by the lack of *cas6*, encoding for an endoribonuclease employed by most type I systems for pre-crRNA processing and by the presence of a single protein encoded by the *cas8c* gene. The absence of *cas6* can be replaced by *cas5* whose product performs a similar catalytic reaction. In type I-C system, genes are typically encoded by a single (predicted) operon. Genes *cas 3*, *5*, *8c* and *7* of locus 1 share sequence identity below the 30% threshold with the corresponding genes located on the other locus. This indicates that these *cas* genes do not derive from gene duplication rather they have been acquired through independent events (see below). The comparative genomic analysis revealed that eight strains of *B. coagulans* possess *cas* operons of type I-B (2-6, 36D1, H-1, P-38, and XZL9), type I-C (CSIL1, DSM1, H-1, MA-13) or type IV (DSM1), mostly located in the vicinity of CRISPR arrays (Figure 8-9 and Table 3). Both type IV *cas* operons are located

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

within prophages in the DSM1 genome, suggesting they are of viral origin. Only strain XZL4 lacks a complete *cas* operon or CRISPR array.

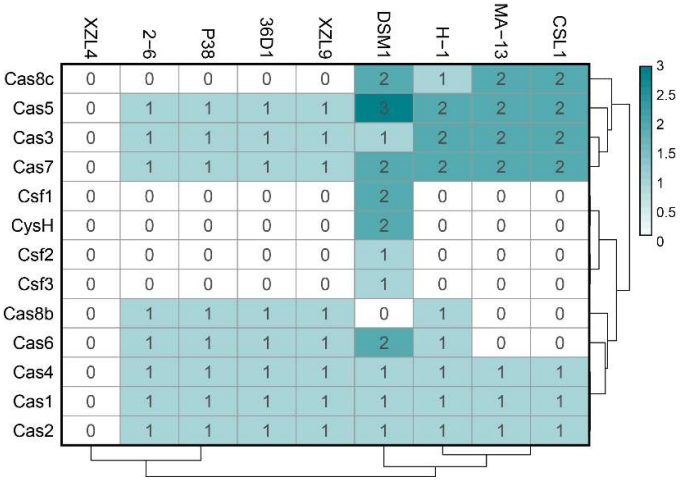

Figure 9. Heat map distribution of *cas* operon genes among *B. coagulans* strains.

Table 3. CRISPR arrays in the *B. coagulans* strains.

Strain	Contig	Prediction	Start	End	Consensus_repeat	Orientation	Cas_associated	N_repeats
2-6	2-6_NC_015634.1	I-B	2499794	2503038	GTTGAACTTTAACATTGGATGTATTTAAAT	R	yes	50
		I-B	2497017	2497373	GTTTCAATTCCTCATAGGTAAAATACAAAC	R	yes	6
36D1	36D1_NC_016023.1	Unknown	320637	321284	TTTTGAAGCCGTCAAAAGGACAAAA	F	orphan	13
		I-B	1096065	1097943	GTTAGTATTTTACCTATGAGGAATTGAAAC	R	orphan	29
		I-B	2117433	2121795	GTTTGTATTTTACCTATGAGGAATTGAAAC	F	yes	66
		I-B	2123872	2124703	GTTTGTATTTTACCTATGAGGAATTGAAAC	F	yes	13
		I-B	2126172	2127007	GTTTGTATTTTACCTATGAGGAATTGAAAC	F	yes	13
CSIL1	CSIL1_NZ_AXVW01000068.1_scaffold00063.63_C	I-C	13264	13753	GTCACACTCCTTGCGAGTGTGTGGATTGAAAT	F	yes	8
	CSIL1_NZ_AXVW01000098.1_scaffold00091.91_C	I-C	7	236	GTCGCTCCCTACATGGGGGCGTGGATTGAAATC	F	orphan	4
	CSIL1_NZ_KI519465.1_scaffold00026.26	I-B	19196	19356	GTTAGTATTTTACCTATGAGGAATTGAAA	F	orphan	3
DSM1	DSM1_NZ_CP009709.1	I-C	1536669	1541901	GTCGCTCCCTACGTGGGGGCGTGGATTGAAAT	R	yes	80
		I-B	2584186	2585866	GTTAGTATTTTACCTATGAGGAATTGAAAC	F	orphan	26
		I-C	2785593	2786680	GTCACACTCCTCGTGAGTGTGTGGATTGAAAT	F	yes	17
H-1	H-1_NZ_ANAQ01000049.1_000049	I-C	9436	9599	GTCACACTCCTCGTGAGTGTGTGAATTGAAAT	F	yes	3
	H-1_NZ_ANAQ01000116.1_000116	I-C	43	204	GTCACACTCCTCGTGAGTGTGTGAATTGAAAT	F	orphan	3
	H-1_NZ_ANAQ01000125.1_000125	I-C	3689	3983	GTCACACTCCTCGTGAGTGTGTGAATTGAAAT	R	orphan	5

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

	H-1_NZ_ANAQ01000203.1_000203	I-B	74	2139	GTTGAACTTTAACATTGGATGT ATTTAAAT	F	orphan	32
	H-1_NZ_ANAQ01000268.1_000268	I-B	17520	20714	GTTGAACTTTAACATTGGATGT ATTTAAAT	F	yes	49
	H-1_NZ_ANAQ01000293.1_000293	I-B	112	339	GTTGAACTTTAACATTGGATGT ATTTAAATT	R	orphan	4
MA-13	MA-13_NZ_SMSP02000017.1_69	I-C	32	1577	GTCGCTCCCTACATGGGGGCG TGGATTGAAAT	F	orphan	24
	MA-13_NZ_SMSP02000061.1_151	I-C	8671	10021	GTCGCTCCCTACATGGGGGCG TGGATTGAAAT	R	yes	21
	MA-13_NZ_SMSP02000062.1_309	I-C	52270	52630	GTCGCTCCCTACATGGGGGCG TGGATTGAAAT	F	yes	6
P38	P38_NZ_JSVI01000006.1_6	I-B	99580	101051	GTTGAACTTTAACATTGGATGT ATTTAAAT	F	yes	23
	P38_NZ_JSVI01000107.1_111	I-B	446	1784	GTTGAACTTTAACATTGGATGT ATTTAAAT	R	orphan	21
XZL9	XZL9_NZ_ANAP01000038.1_000038	I-B	122	953	GTTTGTATTTTACCTATGAGGA ATTGAAAC	R	yes	13
	XZL9_NZ_ANAP01000157.1_000160	I-B	97	1197	GTTTGTATTTTACCTATGAGGA ATTGAAAC	F	orphan	17
	XZL9_NZ_ANAP01000220.1_000226	I-B	68	3018	GTTTGTATTTTACCTATGAGGA ATTGAAAC	R	orphan	45

The *B. coagulans* strains contain 1 to 5 CRISPR arrays each, for a total of 30 arrays of which 13 are annotated as orphan CRISPRs (i. e. not located in the vicinity of a *cas* operon). Additionally, while strains MA-13, P38, and XZL9 have one consensus direct repeat sequence, strains 36D1, DSM1, CSIL1 and H1 have two to three direct repeat sequences despite all strains harboring one single adaptation cassette (Table 3).

At present, it is not clear whether orphan CRISPR arrays are indeed functional, although several lines of evidence indicate that at least some isolated CRISPR arrays are active [61,62]. Strains 2-6, 36D1, and CSIL1 have orphan arrays with a repeat different to the one of the CRISPR array associated with their *cas* operon and adaptation cassette. Moreover, in strains 36D1 and CSIL1 this repeat is predicted to belong to another CRISPR subtype. A similar picture is seen for strain DSM1, which has an adaptation cassette associated with an I-C CRISPR system, but an orphan array with a different direct repeat. These arrays could represent a remnant of a *cas* operon previously present in the genome that was subsequently lost. Additionally, strain H-1 bears interference cassettes for both I-B and I-C subtypes, but an adaptation cassette only for the I-B operon (Table 3). The repeat of the associated CRISPR arrays is different for each subtype, suggesting that both *cas* operons are capable of interference, but only the I-B cassette may be able to acquire new spacers. The low number of spacers in the I-C CRISPR arrays (3 to 5) could reflect its inability to acquire new spacers. Taken together, the distribution of CRISPR-Cas modules in *B. coagulans* strains suggests that loss and gain of CRISPR-Cas cassettes is a common trait in these organisms. From an applicative point of view, the endogenous CRISPR-Cas adaptive immune systems can be repurposed for genome editing of *B. coagulans,* potentially constituting a valuable tool in the genetic toolbox for these strains. Analysis of CRISPR spacer matches allowed the identification of viruses infecting the *B. coagulans* strains. From a pool of 651 spacers extracted from the strains, 113 spacers (17.3%) matched to 77 *Firmicutes*-infecting viruses in the IMG-VR database [63]. These viruses are classified as members of the class *Caudoviricetes,* with the majority of them (66%) belonging to the *Siphoviridae* family. Several viruses match to CRISPR spacers of strains from both clades of *B. coagulans* (group A and group B), suggesting that these strains share a common pool of viruses. Additionally, most viruses in the network originate from an engineered environment (e.g. lab enrichment), including all spacers from strain MA-13 (Figure 10b), while only 4 viruses originated from environmental ecosystems. Interestingly, one single spacer from strain H-1 matched 38 viruses, of which 36 originate from host-associated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

environments, particularly the human digestive system (Figure 10). The source of the *B. coagulans* viruses is in agreement with the prevalent agri-food origin of the isolated *B. coagulans* strains.

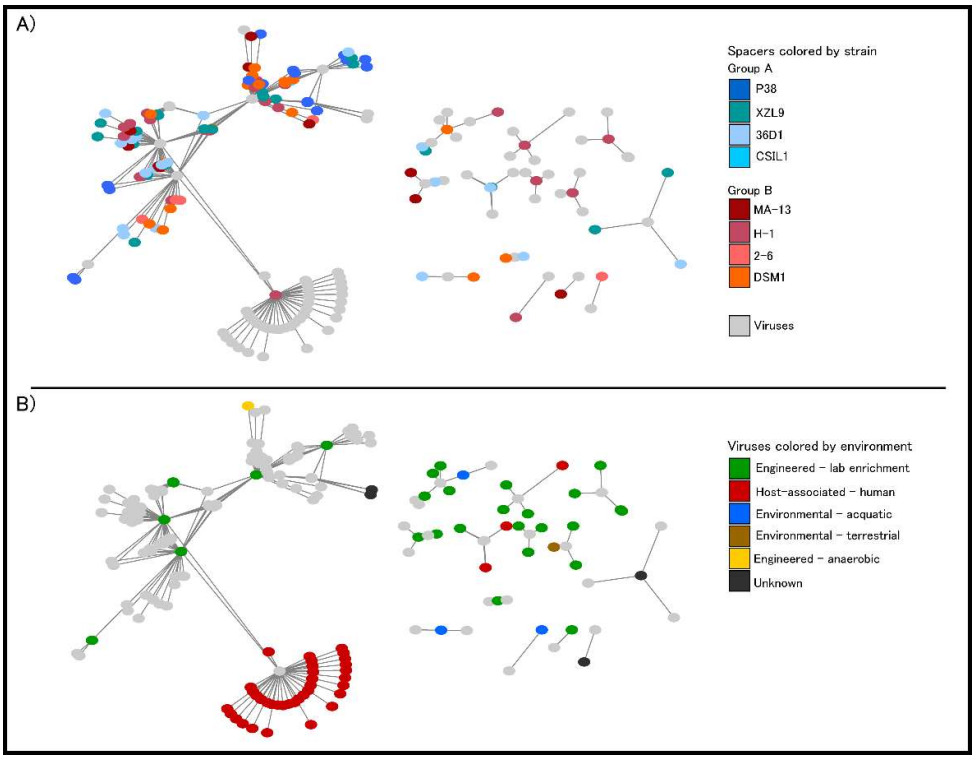

Figure 10. Graphical representation of interaction networks between spacers and viruses. Nodes represent spacers and virus sequences, with lines representing a match between them. A) Spacer nodes colored by their strain of origin. B) Virus nodes are colored by their ecosystem of isolation.

The analysis of the adaptive immunity in MA-13 was extended also to the identification of genomic islands (GIs, regions of probable horizontal origin) and prophage regions [64,65]. GIs are non-self-mobilizing integrative regions encoding factors that support the adaptability and competitiveness of the microbes within a niche, including virulence factors and other medically or environmentally important adaptations [61,63]. Using IslandViewer, the prediction of GIs is carried out with a precise definition of boundaries since integration splits off a gene fragment that marks the distal ends of the island. 297 genes were assigned as belonging to GIs (Table S8 and Fig 11), among which the following *cas* genes have been identified: *cas3* (MBF8417476.1), *cas5* (MBF8417475.1), *cas8c* (MBF8417474.1), *cas7* (MBF8417473.1, MBF8417526.1), *cas4* (MBF8417527.1), *cas1* (MBF8417528.1) and *cas2* (MBF8417529.1). A similar analysis conducted on strains 2-6 and 36D1 indicates that the CRISPRs-Cas systems are located in genomic islands that are flanked by transposases [4]. Several CRISPR-Cas systems have been discovered within GIs and for *V. cholerae* all CRISPR-Cas arrays are located in mobile genetic elements (MGEs) [64]. Among the other genetic traits found in GIs that enhance the fitness of MA-13, worth noting is the presence of i) two genes encoding for putative thiazole-containing heterocyclic bacteriocins (MBF8418016.1 bacteriocin biosynthesis cyclodehydratase) and uberolysin/carnocyclin family circular bacteriocin (MBF8417079.1); ii) a full set of genes belonging to type I restriction systems that are large pentameric proteins with separate restriction (R, MBF8419113.1), methylation (M, MBF8419111.1) and DNA sequence-recognition (S, MBF8419112.1) subunits; iii) a complete cluster of genes encoding for arsenic resistance made up of *arsABCD* (MBF8417195.1, MBF8417192.1, MBF8417193.1, MBF8417194.1). ArsB and D are arsenite

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

efflux transporters, ArsC is an arsenate reductase and ArsA is an arsenical pump-driving ATPase that overall contributes to confer resistance to arsenic (Figure 11, Table S8) [65].

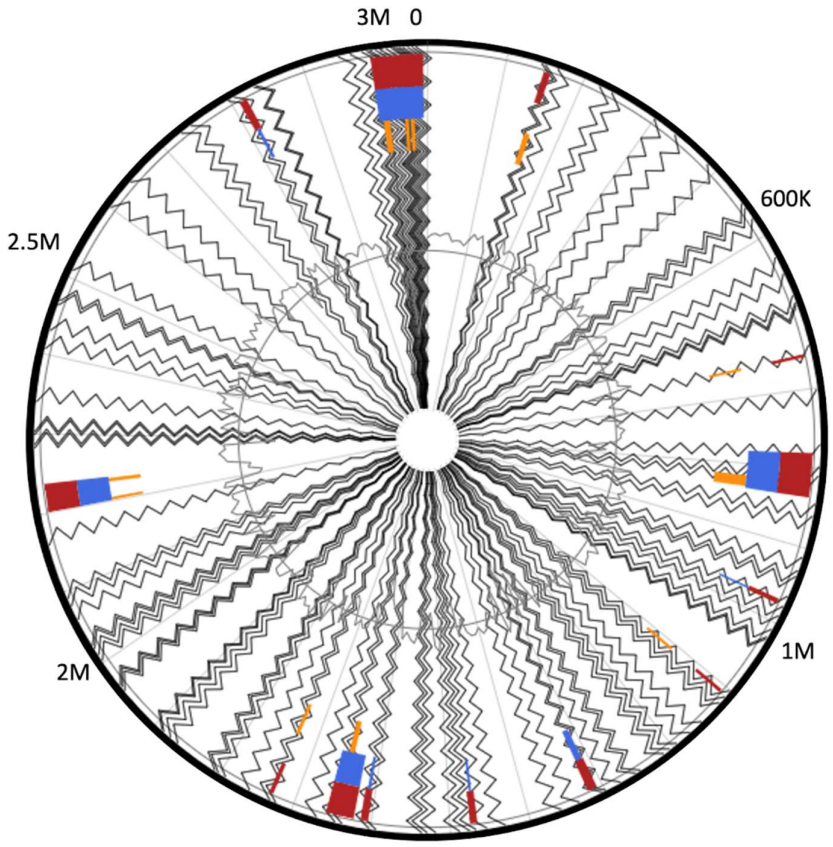

Figure 11. Circular visualization of predicted genomic islands. Blocks are colored according to the prediction method: IslandPath-DIMOB (blue), SIGI-HMM (orange), as well as the integrated results (dark red).

Furthermore, the prophage finding software PHASTER was used to identify intact prophages and their integrase genes in MA-13, utilizing a BLASTP comparison of the query genome with a frequently updated prophage sequence database [66]. Only one prophage is present in the MA-13 genome whose integrity is indicated by the presence of the attachment junctions *attL* and *attR* (also named hybrid attachment sites) at the extremities of the prophage (Figure 11). A total of 58 genes were identified in this genomic region that overall represent a blend of bacterial and bacteriophage features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

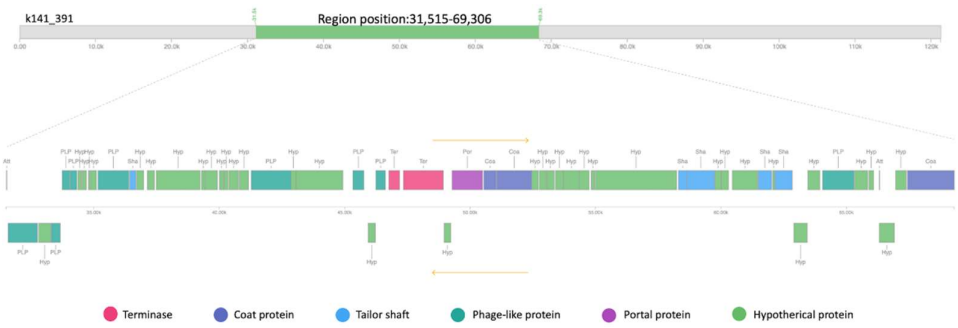

Figure 12. Linear genomic map of prophage phBC6A52 identified with PHASTER from MA-13 genome.

Specifically, 20 hypothetical proteins were annotated using Bacterial Database and the remaining 38 belong to diverse phages (Figure 12 and Table S9). Among these genes, 13 are linked to *Bacillus cereus* prophage phBC6A52 (Genbank NC_004821.1) assigned to the *Podoviridae* family [67]. Besides MA-13, also P38 bears a complete sequence related to a *B. cereus* prophage. PHASTER analysis revealed that it shares 100% identity with *Bacillus* Phage vB_BtS_B83 that belongs to the *Bembunaquatrovirus* bacteriophages genus. Although bacteriophages typically exhibit a narrow host range, yet the presence of the same phage in two different *Bacillus* species suggests that the horizontal gene transfer events can overcome the interspecies boundaries to allow the bacteria-phage coevolution in a given ecotype. Indeed, the exchange of phage attachment molecules could even occur in an interspecies context, enabling phage adsorption to non-host species, providing an alternative route for HGT. In *in vitro* models, species of *Bacillus*, including *B. coagulans*, *B. subtilis,* and *B. cereus* showed a similar life cycle in artificial gastrointestinal tract models [68].

Since *B. coagulans* and *B. cereus* strains have been both isolated mainly from food sources, it is reasonable that they share the same lifestyle and/or ecological environment, making it possible for genetic material exchanges.

3. Materials and Methods

3.1. Genome assembly and annotation

MA-13 genomic sequence was ri-assembled using MEGAHIT v 1.2.9 [69]. The result was analyzed using QUAST v 4.4 and the assembly quality was evaluated using CheckM v 1.0.18 [70]. PATRIC (i.e. NIAID/PathoSystems Resource Integration Center) was used for resistome analysis as well as for comparative genomic analysis [71]. AntiSmash v2.0 (https://antismash.secondarymetabolites.org) was employed to analyze secondary metabolite biosynthesis gene clusters [72].

3.2. Construction of whole-genomes sequence-based phylogenetic tree and comparative analysis of orthologous genes

Several genomes of *B. coagulans* strains available at the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genome) were imported into KBase system to create a genome set, that subsequently was employed to build the tree, using SpeciesTreeBuilder v.0.1. Afterward, the genomic comparison was performed using Build Pangenome with OrthoMCL (Pangenome Orthomcl v0.0.7) and visualized with Pangenome Circle Plot app (v1.2.0). The whole-genome sequences of the nine strains, namely *B. coagulans* MA-13 (SMSP00000000), *B. coagulans* 2-6 (NC_015634), *B. coagulans* 36D1 (NC_016023), *B. coagulans* CSIL1 (NZ_AXVW01000001), *B. coagulans* DSM1 (NZ_CP009709), *B. coagulans* H-1 (NZ_ANAQ01000001), *B. coagulans* P38 (NZ_JSVI01000001), *B. coagulans* XZL4 (NZ_AFWM01000001), and *B. coagulans* XZL9 (NZ_ANAP01000008) were included to evaluate the clusters of orthologous groups (COGs) and to identify the core, non-core, and singletons. These last ones were also

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

compared by annotating the functional domains in the whole genome set with DomainAnnotation (version 1.03) using TIGRFAM and viewing the results using the View Function Profile for Genomes app (version 1.0.1). Furthermore, the CAZome was annotated using "Search with dbCAN HMMs of CAZy families" v.6 in KBase [24].

Two strains i.e MA-13 and DSM1 were selected based on their genetic relatedness and by pairwise comparative analysis by calculating the average nucleotide identity (ANI) using FastANI. The percentage threshold for species boundary is 95% ANI [73]. *In silico* DNA–DNA hybridization (DDH) values were estimated using the Genome-to-Genome Distance Calculator (GGDC).

3.3. Analysis of resistance mechanisms to environmental challenges

Identification of prokaryotic innate defense systems in the *B. coagulans* genomes was done using the webtool PADLOC (https://padloc.otago.ac.nz/padloc/). CRISPR/Cas systems were identified with the web tool CRISPRCasTyper (https://crisprcastyper.crispr.dk/) [74] and the orientation of the CRISPR arrays was determined using CRISPRDetect (http://crispr.otago.ac.nz/CRISPRDetect/predict_crispr_array.html). CRISPR spacers predicted by CRISPRCasTyper were compared to the 2,332,702 sequences in the IMG-VR database (downloaded on the 22nd August 2021) [63] using BLAST (blastn-short task, evalue of 1E-5). The resulting network of virus-host interactions was visualized using Cytoscape (https://cytoscape.org/v.3.7.0). The IslandViewer4 web server (http://www.pathogenomics.sfu.ca/islandviewer/) was utilized for the prediction and interactive visualization of genomic islands (GIs) whose validation has to be performed with at least one of the software (IslandPick, SIGI-HMM or IslandPath-DIMOB) [75]. The web server PHASTER (http://phaster.ca) uses BLAST against a phage-specific sequence database to predict prophage(s) integrated into the bacterial genomes [66]. Only prophages identified as "complete" or "intact" are considered.

4. Conclusions

Thermophilic microorganisms are a reservoir for biodiversity, molecular phylogeny, and the production of unique industrially valuable enzymes and other compounds [76–84]. The purpose of this work is to expand our understanding of the intra-strains genomic diversity of *B. coagulans* and to provide new insights into its genetic potential in biotechnological applications. Since MA-13 has been successfully used for LA production from lignocellulose biomasses and as prebiotics producer, we attempted to explain its favorable fermentation features through this comparative genomic analysis. Several interesting characteristics emerged from this study, pointing to MA-13 as a platform for the efficient production of bio-chemicals. Indeed, the high operating temperature used during fermentation and its innate and adaptive immunity could protect the MA-13 strain from phage infection and contamination. Moreover, the presence of diverse secretion systems can be efficiently employed for homologous/heterologous protein production.

Finally, our analysis contributes also to explore genomic diversity and sheds light on genetic traits that are essential to the basic lifestyle of *B. coagulans*. Furthermore, we have expanded the knowledge of the selective advantages that shape the organization and dynamics of *B. coagulans* genomes, including niche adaptation, antibiotic resistance, and genetic tractability.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: title, Table S1: title, Video S1: title.

Author Contributions: MA and LMA performed the analyisis. MA, GF, DL, PC and XP supervised the project. MA and PC drafted the manuscript. All authors read and approved the final manuscript.

Funding: This work was supported by PRIN 2017-PANACEA CUP:E69E19000530001 and by GoodbyWaste: ObtainGOOD products–exploit BY-products–reduce WASTE, MIUR 2017-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

JTNK78.006, Italy. Novo Nordisk Foundation/Hallas-Moeller Ascending Investigator Grant to X.P (NNF17OC0031154).

Institutional Review Board Statement: "Not applicable."

Informed Consent Statement: "Not applicable."

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Acknowledgments: None

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Jönsson: L.J.; Martín, C. Pretreatment of Lignocellulose: Formation of Inhibitory by-Products and Strategies for Minimizing Their Effects. *Bioresource Technology* **2016**, *199*, 103–112, doi:10.1016/j.biortech.2015.10.009.
2. Aulitto, M.; Tom, L.M.; Ceja-navarro, J.A.; Simmons, B.A. Whole-Genome Sequence of *Brevibacillus borstelensis* SDM , Isolated from a Sorghum-Adapted Microbial Community. *Microbiology Resource Announcements* **2020**, 8–9.
3. Hofvendahl, K.; Hahn-Hägerdal, B. Factors Affecting the Fermentative Lactic Acid Production from Renewable Resources. *Enzyme and Microbial Technology* **2000**, *26*, 87–107, doi:10.1016/S0141-0229(99)00155-6.
4. Su, F.; Xu, P. Genomic Analysis of Thermophilic *Bacillus coagulans* Strains: Efficient Producers for Platform Bio-Chemicals. *Scientific Reports* **2014**, *4*, doi:10.1038/srep03926.
5. Aulitto, M.; Fusco, S.; Bartolucci, S.; Franzén, C.J.; Contursi, P. *Bacillus coagulans* MA-13: A Promising Thermophilic and Cellulolytic Strain for the Production of Lactic Acid from Lignocellulosic Hydrolysate. *Biotechnology for Biofuels* **2017**, *10*, 210, doi:10.1186/s13068-017-0896-8.
6. Aulitto, M.; Fusco, S.; Nickel, D.B.; Bartolucci, S.; Contursi, P.; Franzén, C.J. Seed Culture Pre-Adaptation of *Bacillus coagulans* MA-13 Improves Lactic Acid Production in Simultaneous Saccharification and Fermentation. *Biotechnology for Biofuels* **2019**, *12*, 45, doi:10.1186/s13068-019-1382-2.
7. Ouyang, J.; Cai, C.; Chen, H.; Jiang, T.; Zheng, Z. Efficient Non-Sterilized Fermentation of Biomass-Derived Xylose to Lactic Acid by a Thermotolerant *Bacillus coagulans* NL01. *Applied Biochemistry and Biotechnology* **2012**, *168*, 2387–2397, doi:10.1007/s12010-012-9944-9.
8. Ou, M.S.; Ingram, L.O.; Shanmugam, K.T. L(+)-Lactic Acid Production from Non-Food Carbohydrates by Thermotolerant *Bacillus coagulans*. *Journal of Industrial Microbiology and Biotechnology* **2011**, *38*, 599–605, doi:10.1007/s10295-010-0796-4.
9. Ye, L.; Zhou, X.; Hudari, M.S. Bin; Li, Z.; Wu, J.C. Highly Efficient Production of L-Lactic Acid from Xylose by Newly Isolated *Bacillus coagulans* C106. *Bioresource Technology* **2013**, *132*, 38–44, doi:10.1016/j.biortech.2013.01.011.
10. Walton, S.L.; Bischoff, K.M.; Van Heiningen, A.R.P.; Van Walsum, G.P. Production of Lactic Acid from Hemicellulose Extracts by *Bacillus coagulans* MXL-9. *Journal of Industrial Microbiology and Biotechnology* **2010**, *37*, 823–830, doi:10.1007/s10295-010-0727-4.
11. Konuray, G.; Erginkaya, Z. Potential Use of *Bacillus coagulans* in the Food Industry. *Foods* **2018**, *7*, doi:10.3390/foods7060092.
12. Kapse, N.G.; Engineer, A.S.; Gowdaman, V.; Wagh, S.; Dhakephalkar, P.K. Functional Annotation of the Genome Unravels Probiotic Potential of *Bacillus coagulans* HS243. *Genomics* **2019**, *111*, 921–929, doi:10.1016/j.ygeno.2018.05.022.
13. Cao, J.; Yu, Z.; Liu, W.; Zhao, J.; Zhang, H.; Zhai, Q.; Chen, W. Probiotic Characteristics of *Bacillus coagulans* and Associated Implications for Human Health and Diseases. *Journal of Functional Foods* **2020**, *64*, 103643, doi:10.1016/j.jff.2019.103643.
14. Aulitto, M.; Strazzulli, A.; Sansone, F.; Cozzolino, F.; Monti, M.; Moracci, M.; Fiorentino, G.; Limauro, D.; Bartolucci, S.; Contursi, P. Prebiotic Properties of *Bacillus coagulans* MA-13: Production of Galactoside Hydrolyzing Enzymes and Characterization of the Transglycosylation Properties of a GH42 β-Galactosidase. *Microbial Cell Factories* **2021**, *20*, 1–18, doi:10.1186/s12934-021-01553-y.
15. le Marrec, C.; Hyronimus, B.; Bressollier, P.; Verneuil, B.; Urdaci, M.C. Biochemical and Genetic Characterization of Coagulin, a New Antilisterial Bacteriocin in the Pediocin Family of Bacteriocins, Produced by *Bacillus coagulans* I4. *Applied and Environmental Microbiology* **2000**, *66*, 5213–5220, doi:10.1128/AEM.66.12.5213-5220.2000.
16. Yu, Y.; Yu, X.; Ouyang, J.; Ma, X. Complete Genome Sequence of *Bacillus coagulans* BC01, a Promising Human Probiotic Strain Isolated from Thick Broad Bean Sauce. *Microbiology Resource Announcements* **2021**, *10*, doi:10.1128/mra.00392-21.
17. Rhee, M.S.; Moritz, B.E.; Xie, G.; Glavina del Rio, T.; Dalin, E.; Tice, H.; Bruce, D.; Goodwin, L.; Chertkov, O.; Brettin, T.; et al. Complete Genome Sequence of a Thermotolerant Sporogenic Lactic Acid Bacterium, *Bacillus coagulans* Strain 36D1. *Standards in Genomic Sciences* **2011**, *5*, 331–340, doi:10.4056/sigs.2365342.
18. Su, F.; Tao, F.; Tang, H.; Xu, P. Genome Sequence of the Thermophile *Bacillus coagulans* Hammer, the Type Strain of the Species. *Journal of Bacteriology* **2012**, *194*, 6294–6295, doi:10.1128/JB.01380-12.
19. Nema, V. The Role and Future Possibilities of Next-Generation Sequencing in Studying Microbial Diversity. *Microbial Diversity in the Genomic Era* **2019**, 611–630, doi:10.1016/B978-0-12-814849-5.00034-4.
20. Rodríguez-Beltrán, J.; DelaFuente, J.; León-Sampedro, R.; MacLean, R.C.; San Millán, Á. Beyond Horizontal Gene Transfer: The Role of Plasmids in Bacterial Evolution. *Nature Reviews Microbiology* **2021**, *19*, 347–359, doi:10.1038/s41579-020-00497-1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

21. Aulitto, M.; Fusco, S.; Franzén, C.J.; Strazzulli, A.; Moracci, M.; Bartolucci, S.; Contursi, P. Draft Genome Sequence of *Bacillus coagulans* MA-13, a Thermophilic Lactic Acid Producer from Lignocellulose. *Microbiology Resource Announcements* **2019**, *8*, 341–360, doi:10.1128/mra.00341-19.
22. Tettelin, H.; Medini, D. The Pangenome: Diversity, Dynamics and Evolution of Genomes; 2020; ISBN 9783030382810.
23. Li, L.; Stoeckert, C.J.; Roos, D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome research* **2003**, *13*, 2178–2189, doi:10.1101/GR.1224503.
24. Arkin, A.; Stevens, R.; Cottingham, R.; Maslov, S.; Perez, F. KBase: The Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology* **2015**, *36*, 566–569.
25. Abdul Manas, N.H.; Md. Illias, R.; Mahadi, N.M. Strategy in Manipulating Transglycosylation Activity of Glycosyl Hydrolase for Oligosaccharide Production. *Critical Reviews in Biotechnology* **2018**, *38*, 272–293, doi:10.1080/07388551.2017.1339664.
26. Bang, W.Y.; Ban, O.H.; Lee, B.S.; Oh, S.; Park, C.; Park, M.K.; Jung, S.K.; Yang, J.; Jung, Y.H. Genomic-, Phenotypic-, and Toxicity-Based Safety Assessment and Probiotic Potency of *Bacillus coagulans* IDCC 1201 Isolated from Green Malt. *Journal of industrial microbiology & biotechnology* **2021**, *48*, doi:10.1093/jimb/kuab026.
27. Aulitto, M.; Fusco, F.A.; Fiorentino, G.; Bartolucci, S.; Contursi, P.; Limauro, D. A Thermophilic Enzymatic Cocktail for Galactomannans Degradation. *Enzyme and Microbial Technology* **2018**, *111*, 7–11, doi:10.1016/j.enzmictec.2017.12.008.
28. Aulitto, M.; Fusco, S.; Fiorentino, G.; Limauro, D.; Pedone, E.; Bartolucci, S.; Contursi, P. *Thermus thermophilus* as Source of Thermozymes for Biotechnological Applications: Homologous Expression and Biochemical Characterization of an α-Galactosidase. *Microbial Cell Factories* **2017**, *16*, 1–10, doi:10.1186/s12934-017-0638-4.
29. Aulitto, M.; Fusco, S.; Limauro, D.; Fiorentino, G.; Bartolucci, S.; Contursi, P. Galactomannan Degradation by Thermophilic Enzymes: A Hot Topic for Biotechnological Applications. *World Journal of Microbiology and Biotechnology* **2019**, *35*, 1–13, doi:10.1007/s11274-019-2591-3.
30. Bourne, Y.; Henrissat, B. Glycoside Hydrolases and Glycosyltransferases: Families and Functional Modules. *Current Opinion in Structural Biology* **2001**, *11*, 593–600, doi:10.1016/S0959-440X(00)00253-0.
31. Xavier, J.R.; Ramana, K.V.; Sharma, R.K. β-Galactosidase: Biotechnological Applications in Food Processing. *Journal of Food Biochemistry* 2018, *42*.
32. Ing, N.; Deng, K.; Chen, Y.; Aulitto, M.; Gin, J.W.; Pham, T.L.M.; Petzold, C.J.; Singer, S.W.; Bowen, B.; Sale, K.L.; et al. A Multiplexed Nanostructure-Initiator Mass Spectrometry (NIMS) Assay for Simultaneously Detecting Glycosyl Hydrolase and Lignin Modifying Enzyme Activities. *Scientific Reports* **2021**, *11*, 1–9, doi:10.1038/s41598-021-91181-8.
33. Yang, Y.J.; Lin, W.; Singh, R.P.; Xu, Q.; Chen, Z.; Yuan, Y.; Zou, P.; Li, Y.; Zhang, C. Genomic, Transcriptomic and Enzymatic Insight into Lignocellulolytic System of a Plant Pathogen Dickeya Sp. WS52 to Digest Sweet Pepper and Tomato Stalk. *Biomolecules* **2019**, *9*, 753, doi:10.3390/biom9120753.
34. Coutinho, P.M.; Deleury, E.; Davies, G.J.; Henrissat, B. An Evolving Hierarchical Family Classification for Glycosyltransferases. *Journal of Molecular Biology* **2003**, *328*, 307–317, doi:10.1016/S0022-2836(03)00307-3.
35. Ulvskov, P.; Paiva, D.S.; Domozych, D.; Harholt, J. Classification, Naming and Evolutionary History of Glycosyltransferases from Sequenced Green and Red Algal Genomes. *PLOS ONE* **2013**, *8*, e76511, doi:10.1371/JOURNAL.PONE.0076511.
36. Lombard, V.; Golaconda Ramulu, H.; Drula, E.; Coutinho, P.M.; Henrissat, B. The Carbohydrate-Active Enzymes Database (CAZy) in 2013. *Nucleic Acids Research* **2014**, *42*, D490–D495, doi:10.1093/nar/gkt1178.
37. Levasseur, A.; Drula, E.; Lombard, V.; Coutinho, P.M.; Henrissat, B. Expansion of the Enzymatic Repertoire of the CAZy Database to Integrate Auxiliary Redox Enzymes. *Biotechnology for Biofuels* **2013**, *6*, doi:10.1186/1754-6834-6-41.
38. Nikolaivits, E.; Siaperas, R.; Agrafiotis, A.; Ouazzani, J.; Magoulas, A.; Gioti, A.; Topakas, E. Functional and Transcriptomic Investigation of Laccase Activity in the Presence of PCB29 Identifies Two Novel Enzymes and the Multicopper Oxidase Repertoire of a Marine-Derived Fungus. *Science of the Total Environment* **2021**, *775*, 145818, doi:10.1016/j.scitotenv.2021.145818.
39. Gygli, G.; de Vries, R.P.; van Berkel, W.J.H. On the Origin of Vanillyl Alcohol Oxidases. *Fungal Genetics and Biology* **2018**, *116*, 24–32, doi:10.1016/j.fgb.2018.04.003.
40. Green, E.R.; Mecsas, J. Bacterial Secretion Systems: An Overview. *Microbiology Spectrum* **2016**, *4*, doi:10.1128/microbiolspec.vmbf-0012-2015.
41. Gentschev, I.; Dietrich, G.; Goebel, W. The *E. coli* α-Hemolysin Secretion System and Its Use in Vaccine Development. *Trends in Microbiology* **2002**, *10*, 39–45, doi:10.1016/S0966-842X(01)02259-4.
42. Rivera-Calzada, A.; Famelis, N.; Llorca, O.; Geibel, S. Type VII Secretion Systems: Structure, Functions and Transport Models. *Nature Reviews Microbiology* **2021**, *19*, 567–584, doi:10.1038/s41579-021-00560-5.
43. Song, S.; Wood, T.K. A Primary Physiological Role of Toxin/Antitoxin Systems Is Phage Inhibition. *Frontiers in Microbiology* **2020**, *0*, 1895, doi:10.3389/FMICB.2020.01895.
44. Van Melderen, L. Toxin–Antitoxin Systems: Why so Many, What For? *Current Opinion in Microbiology* **2010**, *13*, 781–785, doi:10.1016/J.MIB.2010.10.006.
45. Pellegrini, O.; Mathy, N.; Gogos, A.; Shapiro, L.; Condon, C. The *Bacillus subtilis* YdcDE Operon Encodes an Endoribonuclease of the MazF/PemK Family and Its Inhibitor. *Molecular Microbiology* **2005**, *56*, 1139–1148, doi:10.1111/J.1365-2958.2005.04606.X.
46. Davis, J.J.; Boisvert, S.; Brettin, T.; Kenyon, R.W.; Mao, C.; Olson, R.; Overbeek, R.; Santerre, J.; Shukla, M.; Wattam, A.R.; et al. Antimicrobial Resistance Prediction in PATRIC and RAST. *Scientific Reports* **2016**, *6*, 1–12, doi:10.1038/srep27930.
47. Tran, T.T.; Panesso, D.; Mishra, N.N.; Mileykovskaya, E.; Guan, Z.; Munita, J.M.; Reyes, J.; Diaz, L.; Weinstock, G.M.; Murray, B.E.; et al. Daptomycin-Resistant Enterococcus Faecalis Diverts the Antibiotic Molecule from the Division Septum and Remodels Cell Membrane Phospholipids. *mBio* **2013**, *4*, doi:10.1128/MBIO.00281-13.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

48. Ohki, R.; Giyanto; Tateno, K.; Masuyama, W.; Moriya, S.; Kobayashi, K.; Ogasawara, N. The BceRS Two-Component Regulatory System Induces Expression of the Bacitracin Transporter, BceAB, in *Bacillus subtilis*. *Molecular Microbiology* **2003**, *49*, 1135–1144, doi:10.1046/j.1365-2958.2003.03653.x.
49. Harel, Y.M.; Bailone, A.; Bibi, E. Resistance to Bacitracin as Modulated by an *Escherichia coli* Homologue of the Bacitracin ABC Transporter BcrC Subunit from *Bacillus licheniformis*. *Journal of Bacteriology* **1999**, *181*, 6176–6178, doi:10.1128/jb.181.19.6176-6178.1999.
50. Heath, R.J.; Su, N.; Murphy, C.K.; Rock, C.O. The Enoyl-[Acyl-Carrier-Protein] Reductases FabI and FabL From *Bacillus subtilis*. *Journal of Biological Chemistry* **2000**, *275*, 40128–40133, doi:10.1074/JBC.M005611200.
51. Demain, A.L.; Fang, A. The Natural Functions of Secondary Metabolites. *Advances in biochemical engineering/biotechnology* **2000**, *69*, 1–39, doi:10.1007/3-540-44964-7_1.
52. Perry, E.K.; Meirelles, L.A.; Newman, D.K. From the Soil to the Clinic: The Impact of Microbial Secondary Metabolites on Antibiotic Tolerance and Resistance. *Nature Reviews Microbiology 2021* **2021**, 1–14, doi:10.1038/s41579-021-00620-w.
53. Kemperman, R.; Jonker, M.; Nauta, A.; Kuipers, O.P.; Kok, J. Functional Analysis of the Gene Cluster Involved in Production of the Bacteriocin Circularin A by *Clostridium beijerinckii* ATCC 25752. *Applied and Environmental Microbiology* **2003**, *69*, 5839–5848, doi:10.1128/AEM.69.10.5839-5848.2003.
54. Doron, S.; Melamed, S.; Ofir, G.; Leavitt, A.; Lopatina, A.; Keren, M.; Amitai, G.; Sorek, R. Systematic Discovery of Anti-Phage Defense Systems in the Microbial-Genome. *Science (New York, N.Y.)* **2018**, *359*, doi:10.1126/SCIENCE.AAR4120.
55. Gao, L.; Altae-Tran, H.; Böhning, F.; Makarova, K.S.; Segel, M.; Schmid-Burgk, J.L.; Koob, J.; Wolf, Y.I.; Koonin, E. V.; Zhang, F. Diverse Enzymatic Activities Mediate Antiviral Immunity in Prokaryotes. *Science* **2020**, *369*, 1077–1084, doi:10.1126/science.aba0372.
56. Panas, M.W.; Jain, P.; Yang, H.; Mitra, S.; Biswas, D.; Wattam, A.R.; Letvin, N.L.; Jacobs, W.R. Noncanonical SMC Protein in Mycobacterium Smegmatis Restricts Maintenance of *Mycobacterium fortuitum* Plasmids. *Proceedings of the National Academy of Sciences* **2014**, *111*, 13264–13271, doi:10.1073/PNAS.1414207111.
57. Millman, A.; Melamed, S.; Amitai, G.; Sorek, R. Diversity and Classification of Cyclic-Oligonucleotide-Based Anti-Phage Signalling Systems. *Nature Microbiology 2020 5:12* **2020**, *5*, 1608–1615, doi:10.1038/s41564-020-0777-y.
58. Cheng, R.; Huang, F.; Wu, H.; Lu, X.; Yan, Y.; Yu, B.; Wang, X.; Zhu, B. A Nucleotide-Sensing Endonuclease from the Gabija Bacterial Defense System. *Nucleic Acids Research* **2021**, *49*, 5216–5229, doi:10.1093/NAR/GKAB277.
59. Weinberger, A.D.; Gilmore, M.S. CRISPR-Cas: To Take up DNA or Not - That Is the Question. *Cell Host and Microbe* **2012**, *12*, 125–126, doi:10.1016/j.chom.2012.07.007.
60. Newsom, S.; Parameshwaran, H.P.; Martin, L.; Rajan, R. The CRISPR-Cas Mechanism for Adaptive Immunity and Alternate Bacterial Functions Fuels Diverse Biotechnologies. *Frontiers in Cellular and Infection Microbiology* 2021, *10*, 1.
61. Almendros, C.; Guzmán, N.M.; García-Martínez, J.; Mojica, F.J.M. Anti-Cas Spacers in Orphan CRISPR4 Arrays Prevent Uptake of Active CRISPR–Cas I-F Systems. *Nature Microbiology 2016 1:8* **2016**, *1*, 1–8, doi:10.1038/nmicrobiol.2016.81.
62. Shmakov, S.A.; Utkina, I.; Wolf, Y.I.; Makarova, K.S.; Severinov, K. V.; Koonin, E. V. CRISPR Arrays Away from Cas Genes. *CRISPR Journal* **2020**, *3*, 535–549, doi:10.1089/crispr.2020.0062.
63. Roux, S.; Páez-Espino, D.; Chen, I.-M.A.; Palaniappan, K.; Ratner, A.; Chu, K.; Reddy, T.B.K.; Nayfach, S.; Schulz, F.; Call, L.; et al. IMG/VR v3: An Integrated Ecological and Evolutionary Framework for Interrogating Genomes of Uncultivated Viruses. *Nucleic Acids Research* **2021**, *49*, D764–D775, doi:10.1093/NAR/GKAA946.
64. Li, W.; Wang, A. Genomic Islands Mediate Environmental Adaptation and the Spread of Antibiotic Resistance in Multiresistant Enterococci - Evidence from Genomic Sequences. *BMC Microbiology* **2021**, *21*, doi:10.1186/S12866-021-02114-4.
65. Owen, S. v; Wenner, N.; Dulberger, C.L.; Rodwell, E. v; Bowers-Barnard, A.; Quinones-Olvera, N.; Rigden, D.J.; Rubin, E.J.; Garner, E.C.; Baym, M.; et al. Prophage-Encoded Phage Defence Proteins with Cognate Self-Immunity., doi:10.1101/2020.07.13.199331.
66. Arndt, D.; Grant, J.R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D.S. PHASTER: A Better, Faster Version of the PHAST Phage Search Tool. *Nucleic Acids Research* **2016**, *44*, W16–W21, doi:10.1093/NAR/GKW387.
67. Gillis, A.; Mahillon, J. Phages Preying on *Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis*: Past, Present and Future. *Viruses* **2014**, *6*, 2623–2672, doi:10.3390/v6072623.
68. Bernardeau, M.; Lehtinen, M.J.; Forssten, S.D.; Nurminen, P. Importance of the Gastrointestinal Life Cycle of *Bacillus* for Probiotic Functionality. *Journal of Food Science and Technology* **2017**, *54*, 2570, doi:10.1007/S13197-017-2688-3.
69. Li, D.; Liu, C.M.; Luo, R.; Sadakane, K.; Lam, T.W. MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph. *Bioinformatics* **2015**, *31*, 1674–1676, doi:10.1093/bioinformatics/btv033.
70. Parks, D.H.; Imelfort, M.; Skennerton, C.T.; Hugenholtz, P.; Tyson, G.W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Research* **2015**, *25*, 1043–1055, doi:10.1101/gr.186072.114.
71. Davis, J.J.; Boisvert, S.; Brettin, T.; Kenyon, R.W.; Mao, C.; Olson, R.; Overbeek, R.; Santerre, J.; Shukla, M.; Wattam, A.R.; et al. Antimicrobial Resistance Prediction in PATRIC and RAST. *Scientific Reports* **2016**, *6*, 1–12, doi:10.1038/srep27930.
72. Blin, K.; Shaw, S.; Kloosterman, A.M.; Charlop-Powers, Z.; van Wezel, G.P.; Medema, M.H.; Weber, T. AntiSMASH 6.0: Improving Cluster Detection and Comparison Capabilities. *Nucleic Acids Research. Nucleic Acids Research* **2021**, 1, doi:10.1093/nar/gkab335.
73. Jain, C.; Rodriguez-R, L.M.; Phillippy, A.M.; Konstantinidis, K.T.; Aluru, S. High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nature Communications* **2018**, *9*, 1–8, doi:10.1038/s41467-018-07641-9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2021 doi:10.20944/preprints202112.0372.v1

74. Russel, J.; Pinilla-Redondo, R.; Mayo-Muñoz, D.; Shah, S.A.; Sørensen, S.J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *https://home.liebertpub.com/crispr* **2020**, *3*, 462–469, doi:10.1089/CRISPR.2020.0059.
75. Bertelli, C.; Laird, M.R.; Williams, K.P.; Lau, B.Y.; Hoad, G.; Winsor, G.L.; Brinkman, F.S.L. IslandViewer 4: Expanded Prediction of Genomic Islands for Larger-Scale Datasets. *Nucleic Acids Research* **2017**, *45*, W30–W35, doi:10.1093/nar/gkx343.
76. Fusco, S.; Aulitto, M.; Iacobucci, I.; Crocamo, G.; Pucci, P.; Bartolucci, S.; Monti, M.; Contursi, P. The Interaction between the F55 Virus-Encoded Transcription Regulator and the RadA Host Recombinase Reveals a Common Strategy in Archaea and Bacteria to Sense the UV-Induced Damage to the Host DNA. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **2020**, *1863*, doi:10.1016/j.bbagrm.2020.194493.
77. Beeby, M.; O'Connor, B.D.; Ryttersgaard, C.; Boutz, D.R.; Perry, L.J.; Yeates, T.O. The Genomics of Disulfide Bonding and Protein Stabilization in Thermophiles. *PLoS Biology* **2005**, *3*, 1549–1558, doi:10.1371/journal.pbio.0030309.
78. Turner, P.; Mamo, G.; Karlsson, E.N. Potential and Utilization of Thermophiles and Thermostable Enzymes in Biorefining. *Microbial Cell Factories* **2007**, *6*, 1–23, doi:10.1186/1475-2859-6-9.
79. Atalah, J.; Cáceres-Moreno, P.; Espina, G.; Blamey, J.M. Thermophiles and the Applications of Their Enzymes as New Biocatalysts. *Bioresource technology* **2019**, *280*, 478–488.
80. Notomista, E.; Falanga, A.; Fusco, S.; Pirone, L.; Zanfardino, A.; Galdiero, S.; Varcamonti, M.; Pedone, E.; Contursi, P. The Identification of a Novel *Sulfolobus islandicus* CAMP-like Peptide Points to Archaeal Microorganisms as Cell Factories for the Production of Antimicrobial Molecules. *Microb Cell Fact* **2015**, *14*, 126, doi:10.1186/s12934-015-0302-9.
81. Contursi, P.; Pisani, F.M.; Grigoriev, A.; Cannio, R.; Bartolucci, S.; Rossi, M. Identification and Autonomous Replication Capability of a Chromosomal Replication Origin from the Archaeon *Sulfolobus solfataricus*. *Extremophiles* **2004**, *8*, 385–391, doi:10.1007/s00792-004-0399-y.
82. Prato, S.; Vitale, R.M.; Contursi, P.; Lipps, G.; Saviano, M.; Rossi, M.; Bartolucci, S. Molecular Modeling and Functional Characterization of the Monomeric Primase-Polymerase Domain from the *Sulfolobus solfataricus* Plasmid PIT3. *FEBS Journal* **2008**, *275*, 4389–4402, doi:10.1111/j.1742-4658.2008.06585.x.
83. Contursi, P.; Cannio, R.; Prato, S.; She, Q.; Rossi, M.; Bartolucci, S. Transcriptional Analysis of the Genetic Element PSSVx: Differential and Temporal Regulation of Gene Expression Reveals Correlation between Transcription and Replication. *Journal of Bacteriology* **2007**, *189*, 6339–6350, doi:10.1128/JB.00638-07.
84. Contursi, P.; Fusco, S.; Limauro, D.; Fiorentino, G. Host and Viral Transcriptional Regulators in *Sulfolobus*: An Overview. *Extremophiles* **2013**, *17*, 881–895, doi:10.1007/s00792-013-0586-9.