Comparative genomics reveals the diversity of CRISPR-Cas systems among neonatal sepsis causing group B *Streptococcus agalactiae*

Sudeep D Ghate[1,4]*, R. Shyama Prasad Rao[1,4], Rajesh P Shastry[2]*, Deepak Pinto[1] and Praveenkumar Shetty[3,4]

[1]Center for Bioinformatics, NITTE deemed to be University, Mangaluru 575018, India

[2]Division of Microbiology and Biotechnology, Yenepoya Research Centre, Yenepoya (Deemed to be University), University Road, Deralakatte, Mangalore-575018, India

[3]Department of Biochemistry, K. S. Hegde Medical Academy, NITTE deemed to be University, Mangaluru, Karnataka, India

[4]Central Research Laboratory, K.S. Hegde Medical Academy, NITTE deemed to be University, Mangaluru 575018, India

*Correspondence

Dr. Sudeep D Ghate,

Center for Bioinformatics, NITTE deemed to be University, Mangaluru 575018, India, E-mail: sudeep.ghate@nitte.edu.in, Orcid ID: 0000-0001-9996-3605

AND

Dr. Rajesh P. Shastry, Division of Microbiology and Biotechnology, Yenepoya Research Centre, Yenepoya (Deemed to be University), University Road, Deralakatte, Mangalore-575018, India, E-mail: rpshastry@yenepoya.edu.in, Orcid ID: 0000-0001-8627-9759

Running title: CRISPR-Cas systems in group B *Streptococcus*.

ABSTRACT

The pathogen *Streptococcus agalactiae*, or Group B *Streptococcus* (GBS) infection is the leading cause of neonatal sepsis and meningitis in neonates.. In this study, we aimed to investigate the occurrence and diversity of the CRISPR-Cas system in *S. agalactiae* genomes using computational biology approaches. A total of 51 out of 52 complete genomes (98.07%) of *S. agalactiae* possess CRISPR arrays (75 CRISPR arrays) with 17 strains possessing multiple CRISPR arrays. There were only two CRISPR-Cas systems – type II-A system and type I-C system in *S. agalactiae* strains. RNA secondary structure analysis through direct repeat analysis showed that the analyzed strains could form stable secondary structures. The 16S rRNA phylogeny exhibited clustering of the strains into three major clades grouped on the type of CRISPR-Cas system. The anti-CRISPRs that contribute to CRISPR-Cas system diversity and prevent genome editing were also detected. These results provide valuable insights into elucidating the evolution, diversity, and function of CRISPR/Cas elements in this pathogen.

Key words: CRISPR-Cas system, Group B *Streptococcus*, Diversity, *Streptococcus agalactiae,* Genomics

INTRODUCTION

Streptococcus agalactiae (also known as Group B *Streptococcus*, GBS), a Gram-positive bacterium, is a common commensal of intestinal and reproductive tracts in healthy adults. It can be transmitted from mother to newborn during birth [1]. The GBS is a cause of stillbirth, chorioamnionitis, and neonatal infections including pneumonia, bacteremia, and meningitis. GBS-sepsis has a mortality rate of 10-18%, with a colonization rate of approximately 18–35% in pregnant women, and neonatal infection rates of 0.4 to 1.1 cases per 1000 live births [2,3]. The infection process is mediated by multifunctional GBS virulence factors that could be a challenge to the immune-deficient neonates [4]. GBS displays virulence factors including a potent hemolytic toxin, proteases, and multiple surface proteins to conquer host tissues [5].

Prokaryotes employ CRISPR-Cas systems (clustered regularly interspaced short palindromic repeat, with CRISPR-associated proteins), which provide sequence-based adaptive immunity against invasive transposable elements, conjugative plasmids, and phages [6]. About 40% of bacteria and 90% of archaea are equipped with CRISPR-Cas systems. Moreover, microbes may have more than one type of CRISPR–Cas system which function towards specific template based recognition, targeting, and degradation of exogenous nucleic acids [7]. These systems could differ in type of Cas proteins present and spacer sequences and also the length and number of CRISPR repeats. Although, initially known for its involvement in viral defense, recent findings suggest involvement of CRISPR-Cas systems in regulation of expression of virulence genes and escape host immunity [8]. CRISPR-Cas systems were earlier considered as adaptive immune system and widely studied in *Streptococcus thermophilus* [9].

Studies suggest that mainly three types of CRISPR-Cas9 systems are employed by

Streptococcus sp. type I, type II, and type III. In addition to these, they also harbor a single type V and unknown CRISPR loci.[10] . The recent reports on the emergence of hypervirulent *S. agalactiae* suggest the contribution of phages and other mobile genetic elements (MGE) in adaptation to different hosts and its virulence profile [11]. These phage-associated genes may play a major role in biological success of the strains by acting as delivery vehicles of resistance and virulence genes [12]. Recently, CRISPR analysis has been used as tool to follow maternal GBS colonization and also as a typing technique over traditional subtyping systems [13] .While extensive details are available on *S. thermophilus* and other animal pathogenic streptococci, detailed information on the CRISIPR-Cas systems in human pathogenic *S. agalactiae* are lacking. Therefore, in this regard, we sought to investigate the occurrence and diversity of CRISPR-Cas systems in *S. agalactiae* genomes. We used CRISPRminer2 server[14] and CRISPRCasFinder [15] - two most inclusive and widely used resources for the identification of CRISPR arrays and cas genes. We report here the diversity and provide insights into existing CRISPR-Cas systems in *S. agalactiae* based on 52 complete genomes of GBS of human origin.

METHODOLOGY

Sequence selection and retrieval

The data set included complete genomes of *Streptococcus agalactiae*. Only the complete genomes with human/*homo sapiens* as hosts were selected and retrieved from NCBI website (https://www.ncbi.nlm.nih.gov/, last accessed on 21-08-2022). Except for the reference strain NGBS128 none of the other genomes had any plasmid sequences. A total of 52 such sequences were selected and their fasta files downloaded for NCBI.

Detection of CRISPR-Cas features

The complete genomes of the 52 strains were screened for the presence of complete CRISPR–Cas loci using CRISPRminer2 server. CRISPRminer2 is a comprehensive tool that uses a comparative genomics approach to identify and annotate CRISPR–Cas loci. This tool also helps with multiple detection options, including anti-CRISPR detection and annotation, self-targeting spacer search, repeat type identification, bacteria–phage interaction detection, and prophage detection. Only "confirmed CRISPRs" identified by the CRISPRminer2 tool were selected for further analysis. The strains which then did not show CRISPR loci were eliminated and rest of the strains were retained for further analysis. The results were also corroborated by checking with CRISPRCasFinder.

Signature genes

The data from CRISPRminer2 was tabulated and a list of signature genes were determined. A tile map was generated to visualise the presence and distribution of these genes amongst the 43 strains. CRISPR map server was also used to obtain in depth information on each of the strains. The CRISPR repeats were analysed through multiple sequence alignment and the aligned direct repeats visualised using the WebLogo program [16].

RNAFold Webserver

Direct repeats (DR) obtained via CRISPRminer2 were then compiled and 11 unique repeats found were then used to generate free energy structures via the RNAFold server[17] . The RNAfold Webserver set to default parameters was used to predict the RNA secondary structure and minimum free energy (MFE) of each DR.

Phylogenetic Trees

To understand CRISPR-Cas distribution in the genomes from a phylogenetic perspective, complete 16S rDNA sequences from 52 genomes were retrieved from NCBI and aligned using MUSCLE in MEGAX [18]. ML statistical method with model selection was used to compute BIC score and AICc value of 24 different nucleotide substitution models. A maximum likelihood phylogenetic tree was constructed (Kimura-2 model of nucleotide substitution) and bootstrap analysis with 1000 random replicates. The cas1 and cas9 genes were aligned and a ML phylogenetic tree was constructed with 1000 bootstrap values. *Streptococcus pyogenes* was taken as the outgroup.

Spacer analysis

The spacer targets were identified using the CRIPSRminer2. The visual representation of the CRISPR spacers was performed using Excel macros, with each unique colour combination representing one unique spacer sequence

RESULTS

Sequence selection and retrieval

A search for *S. agalactiae* genomes in the NCBI database listed 1515 sequenced genomes amongst which 128 were complete genomes. Out of these only the ones affecting human/ *homo sapiens* hosts which resulted in a total of 52 genomes were considered for further analysis. Out of these 52 strains, 51 were determined to possess CRISPR arrays with a total of 75 CRISPR arrays detected and 21 strains possessing multiple of these CRISPR arrays. BJ01 and Sag27 were noted to have the highest number of individual arrays each with three CRISPR-Cas arrays.

Genomic context analysis of confirmed CRISPR-Cas loci

The selected strains were uploaded to CRISPRminer2 web server and results are tabulated in the Table 1. CRISPRminer2 provided details on the number and type of CRISPR locus found, number of spacers and direct repeats (DRs), including DR types, signature genes found within each locus. Other information such as number of prophages, anti-CRISPRs, mobile genetic elements and self-targeting spacers were all obtained from this server. All the strains were classified either into Type II, Type I or orphan CRISPR types (Supplementary table 1). Two types of DRs were found, II having 47 repeats and I having only 12, whilst the remaining 16 repeats were determined to be NA (Not applicable). GBS28 and NGB061 had the highest number of DRs with 31 whilst possessing only 1 CRISPR array. Meanwhile, Sag158 and BJ01 have 31 and 35 DRs respectively but with multiple CRISPR arrays. The individual CRISPR length was observed to have a wide range with FDAARGO_670 and B509 having 7693bp and 7363bp being on the higher end. Meanwhile, BGS-M002 has the shortest CRISPR with only 102 bp. Two types of Anti CRISPR (Acr) regions were also detected with AcrIIA21 being 108 aa long and being present in 26 strains whilst AcrIIA18 being 176 aa in length and present in just GBS1-NY. NGBS061 and BJ01 both possess the highest amount of self-targeting spacers with 11 each. Strain NGBS128 was found to have the greatest number of Mobile Genetic Elements (MGEs) with 37 in its single CRISPR array.

Cas genes

The tile map generated using the presence/absence matrix shows the distribution of signature genes amongst the 51 genomes (Figure 1). From the given cas genes only cas1, cas2, csn2 (Casein Beta which is a Protein Coding gene) and cas9 were seen to be present in all the 51 strains. The strains B507, CUGBS591, CU_GBS_08, CU_GBS_98, NGBS572, Sag153, Sag37, SG-M1, SG-M158, SG-M63, SG-M29 and SG-M50 possessed 2 CRISPR arrays and

hence have the maximum cas proteins. None of the genomes possessed any transposons in the CRISPR loci.

Direct repeats

The DRs from all the CRISPRs were collected and the duplicates were removed. The 11 unique repeats were then uploaded to the RNAFold Webserver from which the free energy structure was obtained as seen in Figure 2. Shorter or incomplete DR sequences were eliminated and the remaining 9 structures were taken into consideration. DR1 and DR2 are seen to have the highest minimum free energy (MFE) value whilst DR2 has the lowest making it the most stable out of the 9 DR structures. The MFE of ribonucleic acids (RNAs) increases at an apparent linear rate with sequence length and the lower the MFE, the more stable the structure [19]. In this case DR2 with -13.10 kcal/mol is seen to be the most stable out of the 9 predicted structures. Both DR1 and DR3 have the least stable structure with -0.3 kcal/mol and -0.4 kcal/mol respectively.

Spacer analysis

In total, 862 spacers were identified among GBS genomes positive for CRISPR loci. Of the identified spacers, 812 were unique (Supplementary figure 1). The spacers in each array ranged from 23 to 104. Among the genomes, the least number of spacers (1) was seen within the CRISPR locus of BJ01, while the highest number of spacers (31) was seen in GBS28 genome with an average of 11.4 spacers per array. An analysis of spacer sequences showed 212 spacers to match plasmids (24.79%) and 568 spacers (66.43%) to match phages. The CRISPRminer2 prediction indicated the absence of self-targeting spacers. Furthermore, 16 genomes had duplicate spacers within their genome with a total of 50 duplicate spacers across all the GBS genomes studied.

197 **Phylogenetic Trees**

198 Two separate phylogenetic trees were constructed for 16S sequences, and cas9 of the selected
199 genomes (Figure 3). The 16S phylogenetic analysis showed all sequences clustering into 2
200 major clades based on their CRISPR-Cas status. This close clustering of strains may be
201 indicative of close intra-genus relationship among them. Cas9 phylogenetic tree showed
202 clustering of the strains into 3 major clades grouped on the variations seen in their respective
203 genes (Supplementary figure 1).

204

205 **DISCUSSION**

206 In this study, we investigated the CRISPR–Cas systems in the GBS genomes isolated from
207 humans' origin to gain insights into the occurrence, diversity, and features of its adaptive
208 immune system. GBS had a high frequency of occurrence of the complete CRISPR–Cas system
209 (91.4%). This is comparable to the reported prevalence of complete CRISPR loci for
210 Streptococcus genera [9]. High CRISPR-Cas prevalence has been attributed to high viral
211 abundance coupled with lower viral diversity in the ecosystem [20]. Bacterial CRISPR-Cas
212 systems have been associated with interaction of pathogens with host cells, immune evasion
213 and other bacterial virulence [21]. Interestingly, contradictory functions have been reported on
214 the functioning of CRISPRs. Short or complete absence of CRISPR arrays have led to increased
215 pathogenicity as seen in gastroenteritis causing *Campylobacter jejuni* strains [8], while cas
216 genes has been shown to enhance virulence in *S. agalactiae* mutant studies [22]. On the one
217 hand, CRISPR-Cas system may lessen the potential virulence by preventing MGE from
218 introducing new virulence genes, while on the other hand, CRISPR-Cas may enhance virulence
219 by regulating gene expression and promoting host colonization. GBS expresses various surface
220 and secreted virulence factors to colonise and infect neonates, which also supports survival in
221 the bloodstream.

The CRISPR–Cas systems are classified into two classes, Class I and Class II, 6 types and 33 subtypes based on the crRNA–effector complex [23]. The genera *Streptococci* fundamentally harbor type I, type II and type III CRISPR-Cas systems in addition to the individual type V and unknown CRISPR loci [24]. The type II system is involved in pathogenesis, quorum sensing, invasion and stress response among others while type I systems drives DNA targeting and cleavage associated with antiviral defense. Type III systems provides transcription-dependent immunity against diverse nucleic acid invaders [25]. In our study, out of selected 52 genomes, 51 genomes contain CRISPR arrays with a total of 75 CRISPR arrays detected and 17 strains possessing multiple of these CRISPR arrays further classified into Type II, Type I or orphan CRISPR types.

A majority, 29 genomes (55.76%) of the CRISPR–Cas systems of the GBS genomes were of Type II-A, while 15 (28.84%) genomes contained both Type II-A and I-C type of the CRISPR–Cas system. This composition is similar to the that of CRISPR-Cas of other *Streptococcus* species like *S. canis* [26] and *S. pyogenes* [27]. The type I-C in GBS contains seven cas genes (cas3, cas5c, cas8c, cas7, cas4, cas1 and cas2) similar to the ones found in S. pyogenes [27]. Cas9 was found in all 51 genomes. The recent studies indicates that Type II CRISPR-associated protein 9 (cas9) influenced virulence in GBS strains [28,29]. The virulence factors of GBS have been implicated in vaginal colonization and invasive disease through Cas9 based regulators [29].None of the genomes contained any transposon or retrotransposon elements in the CRISPR loci.

Interestingly, we are also able to detect anti-CRISPRs from GBS, which contributes to CRISPR-Cas system diversity and which also prevents genome editing. Two types of Anti

CRISPR (Acr) regions were detected from selected strains as AcrIIA21 being present in 26 strains whilst AcrIIA18 in just GBS1-NY strain. AcrIIA21 exhibits broad spectrum action by inhibiting *Streptococcus pyogenes* Cas9 (SpyCas9), *Staphylococcus aureus* Cas9 (SauCas9), and *Streptococcus iniae* Cas9 (SinCas9), exhibiting high efficacy against SinCas9 [30]. An in depth understanding of its mechanism remains elusive. Furthermore, the modulation of Cas9 through sgRNA has also been reported from AcrIIA17 and AcrIIA18. The AcrIIA18 performs Cas9-dependent truncation of sgRNA which lead to generation of a shortened sgRNA which are incapable of triggering Cas9 activity [31].

CRISPR repeats are known to produce hairpin loops like secondary structure owing to its palindrome repeats. The stem-loop structure of DRs are known to facilitate the interaction between spacers and cas proteins. An investigation of the RNA secondary structures and their MFE values indicated that all but one DRs could form stable structures with $\Delta G < -10$ kcal mol^{-1}. DR1, DR3 and DR5 had lower MFE values in comparison to DR5, DR6, DR7, DR8 and DR9. Studies indicate that active CRISPR arrays tend to be long due to the continuous acquisition of spacers [32]. In this study, a maximum of 31 spacers were present in CRISPR loci indicating an active system. The average spacer length in the GBS genomes was 39 bp. In comparison, some genomes like that of *E. coli* contains an average length of 31 bp while it was found to be between 28 and 32 nucleotide bp length in *S. thermophilus* [33]. Studies indicate that CRISPR systems containing spacers of length >30 bp are more active than loci with shorter spacer lengths and more spacers allow bacteria to mount a better defense against viruses [34]. Many of the geographically close strains carried a CRISPR cassette with diverse spacers. Such observations have recorded earlier from *S. thermophilus* where spacer hypervariability has been directly linked to phage exposure [35]. Some of the spacers within the CRISPR loci were duplicated within the genome, the exact significance of this is not clear. Further experimental

evidences are needed to investigate the functioning of the CRISPR–Cas systems on gene expression and regulation especially during host-pathogen interaction in GBS genomes.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENT

R. P. Shastry was supported by DST-SYST, Government of India, New Delhi (SP/YO/2019/1046).

DATA AVAILABILITY STATEMENT

The sequence data used in this work were obtained from NCBI. The authors declare that all data supporting the findings of this study are available within the article and its supporting Information files.

AUTHOR CONTRIBUTIONS

SDG, RPS, and RSPR planned the work. RSPR, RPS and SDG performed the work and wrote the manuscript. DP helped in data curation. PKS gave critical comments and helped in the editing. All authors contributed intellectually, and edited/reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

ORCID ID

Rajesh P. Shastry https://orcid.org/0000-0001-8627-9759
R. Shyama Prasad Rao https://orcid.org/0000-0002-2285-6788
Sudeep D. Ghate https://orcid.org/0000-0001-9996-3605

SUPPLEMENTAL INFORMATION

Supplemental information for this article is available online.

REFERENCES

[1] P.J. Steer, A.B. Russell, S. Kochhar, P. Cox, J. Plumb, G. Gopal Rao, Group B streptococcal disease in the mother and newborn—A review, Eur J Obstet Gynecol Reprod Biol. 252 (2020) 526–533. https://doi.org/10.1016/j.ejogrb.2020.06.024.

[2] N. Medugu, K.C. Iregbu, R.E. Parker, J. Plemmons, P. Singh, L.I. Audu, E. Efetie, H.D. Davies, S.D. Manning, Group B streptococcal colonization and transmission dynamics in pregnant women and their newborns in Nigeria: implications for prevention strategies, Clin Microbiol Infect. 23 (2017) 673.e9-673.e16. https://doi.org/10.1016/j.cmi.2017.02.029.

[3] S. Kadambari, C.L. Trotter, P.T. Heath, M.J. Goldacre, A.J. Pollard, R. Goldacre, Group B Streptococcal Disease in England (1998 - 2017): A Population-based Observational Study, Clin Infect Dis. 72 (2021) e791–e798. https://doi.org/10.1093/cid/ciaa1485.

[4] K.A. Patras, V. Nizet, Group B Streptococcal Maternal Colonization and Neonatal Disease: Molecular Mechanisms and Preventative Approaches, Front Pediatr. 6 (2018) 27. https://doi.org/10.3389/fped.2018.00027.

[5] S. Shabayek, B. Spellerberg, Group B Streptococcal Colonization, Molecular Characteristics, and Epidemiology, Frontiers in Microbiology. 9 (2018). https://www.frontiersin.org/article/10.3389/fmicb.2018.00437 (accessed March 15, 2022).

[6] L.A. Marraffini, E.J. Sontheimer, CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea, Nat Rev Genet. 11 (2010) 181–190. https://doi.org/10.1038/nrg2749.

[7] A. Loureiro, G.J. da Silva, CRISPR-Cas: Converting A Bacterial Defence Mechanism into A State-of-the-Art Genetic Manipulation Tool, Antibiotics (Basel). 8 (2019) 18. https://doi.org/10.3390/antibiotics8010018.

[8] R. Louwen, R.H.J. Staals, H.P. Endtz, P. van Baarlen, J. van der Oost, The role of CRISPR-Cas systems in virulence of pathogenic bacteria, Microbiol Mol Biol Rev. 78 (2014) 74–88. https://doi.org/10.1128/MMBR.00039-13.

[9] C. Lemaire, B. Le Gallou, P. Lanotte, L. Mereghetti, A. Pastuszka, Distribution, Diversity and Roles of CRISPR-Cas Systems in Human and Animal Pathogenic Streptococci, Front Microbiol. 13 (2022) 828031. https://doi.org/10.3389/fmicb.2022.828031.

[10] T. Gong, M. Lu, X. Zhou, A. Zhang, B. Tang, J. Chen, M. Jing, Y. Li, CRISPR-Cas Systems in Streptococci, Curr Issues Mol Biol. 32 (2019) 1–38. https://doi.org/10.21775/cimb.032.001.

[11] C. Crestani, T.L. Forde, R.N. Zadoks, Development and Application of a Prophage Integrase Typing Scheme for Group B Streptococcus, Front Microbiol. 11 (2020) 1993. https://doi.org/10.3389/fmicb.2020.01993.

[12] A. Beceiro, M. Tomás, G. Bou, Antimicrobial Resistance and Virulence: a Successful or Deleterious Association in the Bacterial World?, Clin Microbiol Rev. 26 (2013) 185–230. https://doi.org/10.1128/CMR.00059-12.

[13] C. Beauruelle, A. Pastuszka, P. Horvath, F. Perrotin, L. Mereghetti, P. Lanotte, CRISPR: A Useful Genetic Feature to Follow Vaginal Carriage of Group B Streptococcus, Frontiers in Microbiology. 8 (2017). https://www.frontiersin.org/articles/10.3389/fmicb.2017.01981 (accessed February 9, 2023).

[14] F. Zhang, S. Zhao, C. Ren, Y. Zhu, H. Zhou, Y. Lai, F. Zhou, Y. Jia, K. Zheng, Z. Huang, CRISPRminer is a knowledge base for exploring CRISPR-Cas systems in microbe and phage interactions, Commun Biol. 1 (2018) 180. https://doi.org/10.1038/s42003-018-0184-6.

[15] D. Couvin, A. Bernheim, C. Toffano-Nioche, M. Touchon, J. Michalik, B. Néron, E.P.C. Rocha, G. Vergnaud, D. Gautheret, C. Pourcel, CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins, Nucleic Acids Research. 46 (2018) W246–W251. https://doi.org/10.1093/nar/gky425.

[16] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190. https://doi.org/10.1101/gr.849004.

[17] A.R. Gruber, R. Lorenz, S.H. Bernhart, R. Neuböck, I.L. Hofacker, The Vienna RNA Websuite, Nucleic Acids Res. 36 (2008) W70–W74. https://doi.org/10.1093/nar/gkn188.

[18] S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms, Mol Biol Evol. 35 (2018) 1547–1549. https://doi.org/10.1093/molbev/msy096.

[19] E. Trotta, On the Normalization of the Minimum Free Energy of RNAs by Sequence Length, PLOS ONE. 9 (2014) e113380. https://doi.org/10.1371/journal.pone.0113380.

[20] S. Meaden, A. Biswas, K. Arkhipova, S.E. Morales, B.E. Dutilh, E.R. Westra, P.C. Fineran, High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems, Current Biology. 32 (2022) 220-227.e5. https://doi.org/10.1016/j.cub.2021.10.038.

[21] D. Rath, L. Amlinger, A. Rath, M. Lundgren, The CRISPR-Cas immune system: Biology, mechanisms and applications, Biochimie. 117 (2015) 119–128. https://doi.org/10.1016/j.biochi.2015.03.025.

[22] Y. Dong, K. Ma, Q. Cao, H. Huang, M. Nie, G. Liu, M. Jiang, C. Lu, Y. Liu, CRISPR-dependent endogenous gene regulation is required for virulence in piscine Streptococcus agalactiae, Emerg Microbes Infect. 10 (n.d.) 2113–2124. https://doi.org/10.1080/22221751.2021.2002127.

[23] K.S. Makarova, Y.I. Wolf, J. Iranzo, S.A. Shmakov, O.S. Alkhnbashi, S.J.J. Brouns, E. Charpentier, D. Cheng, D.H. Haft, P. Horvath, S. Moineau, F.J.M. Mojica, D. Scott, S.A. Shah, V. Siksnys, M.P. Terns, Č. Venclovas, M.F. White, A.F. Yakunin, W. Yan, F. Zhang, R.A. Garrett, R. Backofen, J. van der Oost, R. Barrangou, E.V. Koonin, Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants, Nat Rev Microbiol. 18 (2020) 67–83. https://doi.org/10.1038/s41579-019-0299-x.

[24] J. Carte, R.T. Christopher, J.T. Smith, S. Olson, R. Barrangou, S. Moineau, C.V.C. Glover, B.R. Graveley, R.M. Terns, M.P. Terns, The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in Streptococcus thermophilus, Mol Microbiol. 93 (2014) 98–112. https://doi.org/10.1111/mmi.12644.

[25] J. Lin, M. Feng, H. Zhang, Q. She, Characterization of a novel type III CRISPR-Cas effector provides new insights into the allosteric activation and suppression of the Cas10 DNase, Cell Discov. 6 (2020) 1–16. https://doi.org/10.1038/s41421-020-0160-4.

[26] H. Yoshida, Y. Fukushima, M. Goto, Y. Tsuyuki, T. Takahashi, Analysis of the Type II-A CRISPR-Cas System in Streptococcus canis Isolated from Diseased Companion Animals and One Human Patient in Japan, Japanese Journal of Infectious Diseases. 72 (2019) 261–265. https://doi.org/10.7883/yoken.JJID.2018.492.

[27] A.L. Rhun, A. Escalera-Maurer, M. Bratovič, E. Charpentier, CRISPR-Cas in Streptococcus pyogenes, RNA Biology. 16 (2019) 380. https://doi.org/10.1080/15476286.2019.1582974.

[28] A.N. Dammann, A.B. Chamby, A.J. Catomeris, K.M. Davidson, H. Tettelin, J.-P. van Pijkeren, K.P. Gopalakrishna, M.F. Keith, J.L. Elder, A.J. Ratner, T.A. Hooven, Genome-Wide fitness analysis of group B Streptococcus in human amniotic fluid reveals a transcription factor that controls multiple virulence traits, PLOS Pathogens. 17 (2021) e1009116. https://doi.org/10.1371/journal.ppat.1009116.

[29] B.L. Spencer, L. Deng, K.A. Patras, Z.M. Burcham, G.F. Sanches, P.E. Nagao, K.S. Doran, Cas9 Contributes to Group B Streptococcal Colonization and Disease, Front Microbiol. 10 (2019) 1930. https://doi.org/10.3389/fmicb.2019.01930.

[30] S. Eitzinger, A. Asif, K.E. Watters, A.T. Iavarone, G.J. Knott, J.A. Doudna, F.U.A.A. Minhas, Machine learning predicts new anti-CRISPR proteins, Nucleic Acids Res. 48 (2020) 4698–4708. https://doi.org/10.1093/nar/gkaa219.

[31] X. Wang, X. Li, Y. Ma, J. He, X. Liu, G. Yu, H. Yin, H. Zhang, Inhibition mechanisms of CRISPR-Cas9 by AcrIIA17 and AcrIIA18, Nucleic Acids Res. 50 (2022) 512–521. https://doi.org/10.1093/nar/gkab1197.

[32] U. Gophna, D.M. Kristensen, Y.I. Wolf, O. Popa, C. Drevet, E.V. Koonin, No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales, ISME J. 9 (2015) 2021–2027. https://doi.org/10.1038/ismej.2015.20.

[33] T. Sinkunas, G. Gasiunas, S.P. Waghmare, M.J. Dickman, R. Barrangou, P. Horvath, V. Siksnys, In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus, EMBO J. 32 (2013) 385–394. https://doi.org/10.1038/emboj.2012.352.

[34] A. Martynov, K. Severinov, I. Ispolatov, Optimal number of spacers in CRISPR arrays, PLoS Comput Biol. 13 (2017) e1005891. https://doi.org/10.1371/journal.pcbi.1005891.

[35] P. Horvath, D.A. Romero, A.-C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, R. Barrangou, Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus, J Bacteriol. 190 (2008) 1401–1412. https://doi.org/10.1128/JB.01415-07.

433 **Figure legends**:

434 **Figure 1**: Heatmap of presence/absence of various signature cas genes amongst the 52 strains.
435 The tiles in dark blue denote the presence whilst the ones in light blue show absence.
436

437 **Figure 2:** The secondary structure for consensus 11 unique direct repeat sequences of CRISPR
438 arrays in GBS strains. The kcal/mol indicates the minimum free energy (MFE) which is known
439 to increase at a clear linear rate with sequence length. The colours represent the base-pair
440 probability range.
441

442 **Figure 3:** Phylogeny of GBS used in this study. The tree was based on 65 non-redundant
443 complete 16S rRNA sequences from 51 species of GBS. *S. pyogenes* was taken as the outgroup.
444 Numbers next to nodes indicate bootstrap values (%) based on 1000 iterations. Branch length
445 scale indicates the number of substitutions per site. The phylogeny tree was constructed in
446 MEGA10 using the maximum likelihood method.

447

448

449

450

451

452

453

454

455

456

457

458

459

460 **Table 1**: Genomic properties and CRISPR-Cas type of the 52 GBS strains used in this study.
461

#	Strain/Accession number	Source	Condition	CRISPR type	Length (Mb)	GC%	Protein Count
1	32790-3A / NZ_CP029561.1	Guangzhou, China	Hospital	II-A	2.15	35.7	2167
2	874391 / NZ_CP022537.1	Japan	Vagina	II-A	2.15	35.5	1991
3	B111 / NZ_CP021772.1	Shenzhen, China	Neonatal sepsis	II-A	2.15	35.4	2021
4	B507 / NZ_CP021771.1	Shenzhen, China	Vagina (mother)	II-A, I-C	2.08	35.4	1936
5	B508 / NZ_CP021770.1	Shenzhen, China	Vagina (mother)	II-A	2.20	35.6	2204
6	BJ01 / NZ_CP059383.1	Beijing, China	Neonate blood	Orphan	2.15	35.7	2037
7	CJB111 / NZ_CP063198.2	USA	Blood	II-A	2.09	35.5	1955
8	CNCTC 10_84/ NZ_CP006910.1	Atlanta, USA	Hospital	II-A	2.01	35.4	2046
9	COH1 / NZ_HG939456.1	Institute Pasteur	Sepsis (new-born)	II-A	2.07	35.4	1893
10	CU_GBS_08 / NZ_CP010874.1	Hong Kong	Hospital	II-A, I-C	2.08	35.4	1987
11	CU_GBS_98 / NZ_CP010875.1	Hong Kong	Meningitis (Hospital)	II-A, I-C	2.03	35.4	1916
12	CUGBS591 / NZ_CP021862.1	Hong Kong	Arthritis (Hospital)	II-A, I-C	2.23	35.8	2103
13	GBS11 / NZ_CP041999.1	Houston, USA	Blood	II-A	2.14	35.6	2180
14	GBS19 / NZ_CP042000.1	Houston, USA	Blood	II-A	2.10	35.5	2120
15	GBS1-NY / NZ_CP007570.1	USA	Blood	II-A	2.24	35.9	2059
16	GBS28 / NZ_CP042001.1	Tennessee, USA	Health Centre	II-A	2.14	35.7	2024
17	GBS2-NM / NZ_CP007571.1	USA	Hospital	II-A	2.21	35.9	2036
18	GBS30 / NZ_CP042002.1	Houston, USA	Blood	II-A	2.08	35.5	2096
19	GBS6 / NZ_CP007572.1	Houston, USA	Hospital	II-A	2.23	35.8	2054
20	GBS7 / NZ_CP041998.1	Houston, USA	Blood	II-A	2.09	35.5	1957
21	GBS85147 / NZ_CP010319.1	Rio de Janeiro, Brazil	New-born	II-A	2.00	35.5	1992
22	GBS-M002 / NZ_CP013908.1	Taiwan	Cervix	II-A	2.09	35.6	1955
23	H002 / NZ_CP011329.1	Guangxi, China	Vagina	II-A	2.15	35.7	1984
24	HU-GS5823 / NZ_AP018935.1	Hokkaido, Japan	Hospital	II-A	2.23	35.6	2233
25	NEM316 / NC_004368.1	Institute Pasteur	Septicemia	II-A	2.21	35.6	2227
26	NGBS061 / NZ_CP007631.2	Toronto, Canada	Health Centre	II-A	2.22	35.5	2275
27	NGBS572 / NZ_CP007632.1	Toronto, Canada	Health Centre	II-A, I-C	2.06	35.5	2079

28	PLGBS13 / NZ_CP029749.1	Alberta, Canada	Wound (Soft tissue)	II-A, I-C	2.10	35.5	2122
29	Sag153 / NZ_CP036376.1	Shanghai, China	Vagina	II-A, I-C	2.17	35.8	2223
30	Sag158 / NZ_CP019979.1	Shanghai, China	Hospital	II-A, I-C	2.10	35.7	1941
31	Sag27 / NZ_CP031556.1	Shanghai, China	Perianal region	II-A, Orphan (2)	2.21	35.7	2074
32	Sag37 / NZ_CP019978.1	Shanghai, China	Blood	II-A, I-C	2.20	35.8	2250
33	SG-M1 / NZ_CP012419.2	Singapore	Blood	II-A, I-C	2.12	35.5	2180
34	SG-M158 / NZ_CP021864.1	Singapore	Blood	II-A, I-C	2.11	35.5	2025
35	SG-M163 / NZ_CP021863.1	Singapore	Blood	II-A, I-C	2.12	35.5	2025
36	SG-M25 NZ_CP021867.1	Singapore	Blood	II-A, Orphan	2.21	35.7	2075
37	SG-M29 / NZ_CP021866.1	Singapore	Blood	II-A, I-C	2.12	35.5	2025
38	SG-M4 / NZ_CP021870.1	Singapore	Blood	II-A	2.07	35.5	2085
39	SG-M50 / NZ_CP021865.1	Singapore	Blood	II-A, I-C	2.12	35.5	2023
40	SG-M6 / NZ_CP021869.1	Singapore	Blood	II-A, I-C	2.11	35.6	1954
41	SG-M8 / NZ_CP021868.1	Singapore	Blood	II-A	2.17	35.6	2186
42	SS1 / NZ_CP010867.1	Houston, USA	Blood	II-A	2.09	35.5	2110
43	SS1168 / NZ_CP038809.1	Houston, USA	Hospital	II-A	2.04	35.4	1911
44	2012-845 / CP051842.1	Versailles, France	Blood	0 CRISPR	1.53	35.3	
45	B105 / NZ_CP021773.1	Shenzhen, China	Blood sample from a new-born	Orphan	2.27	35.7	2076
46	B509 / NZ_CP021769.1	Shenzhen, China	Vagina swab from a perinatal mother	II-A	2.06	35.5	1928
47	S9968 / NZ_CP058666.1	Seoul, South Korea	Urine	II-A, Orphan (2)	2.20	35.7	2088
48	NGBS128 / NZ_CP012480.1	Greater Toronto area/Peel, Canada	Infection sample	Orphan	2.08	35.7	1879
49	FDAARGOS_254 / NZ_CP020449.2	DC, USA	Blood	II-A	2.22	35.7	2060
50	FDAARGOS_512 / NZ_CP033822.1	DC, USA	Endotracheal aspirate	Orphan	2.13	35.6	2000
51	FDAARGOS_669 / NZ_CP044091.1	DC, USA	Clinical isolate	II-A	2.07	35.4	1937
52	FDAARGOS_670 / NZ_CP044090.1	DC, USA	Clinical isolate	II-A	2.21	35.8	2098

462

463

464

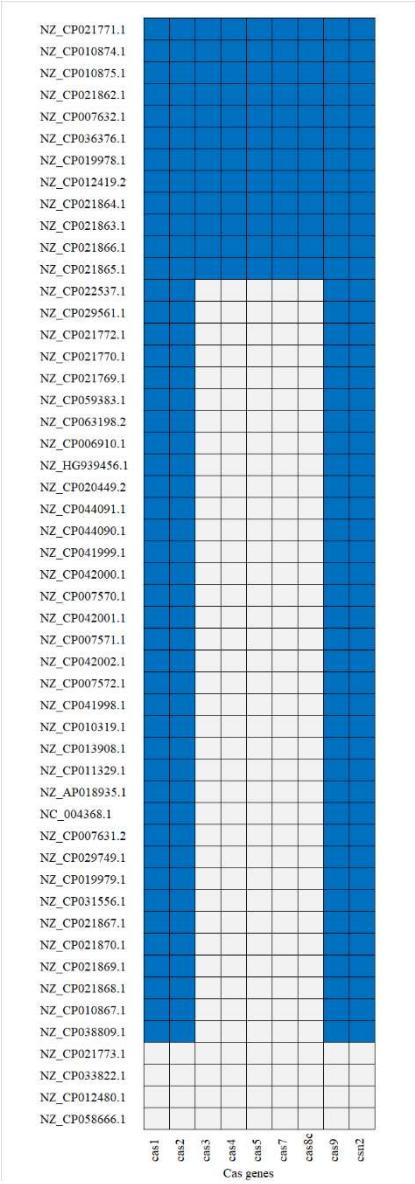

Fig. 1

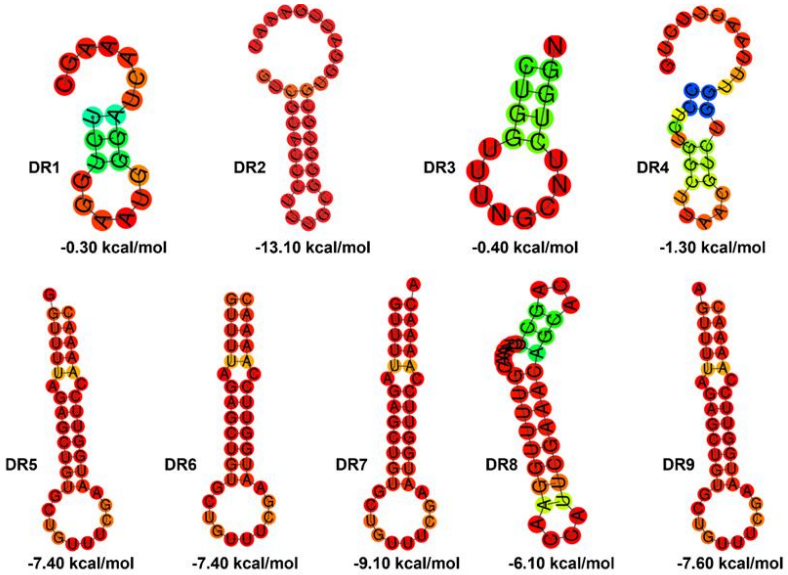

Fig. 2

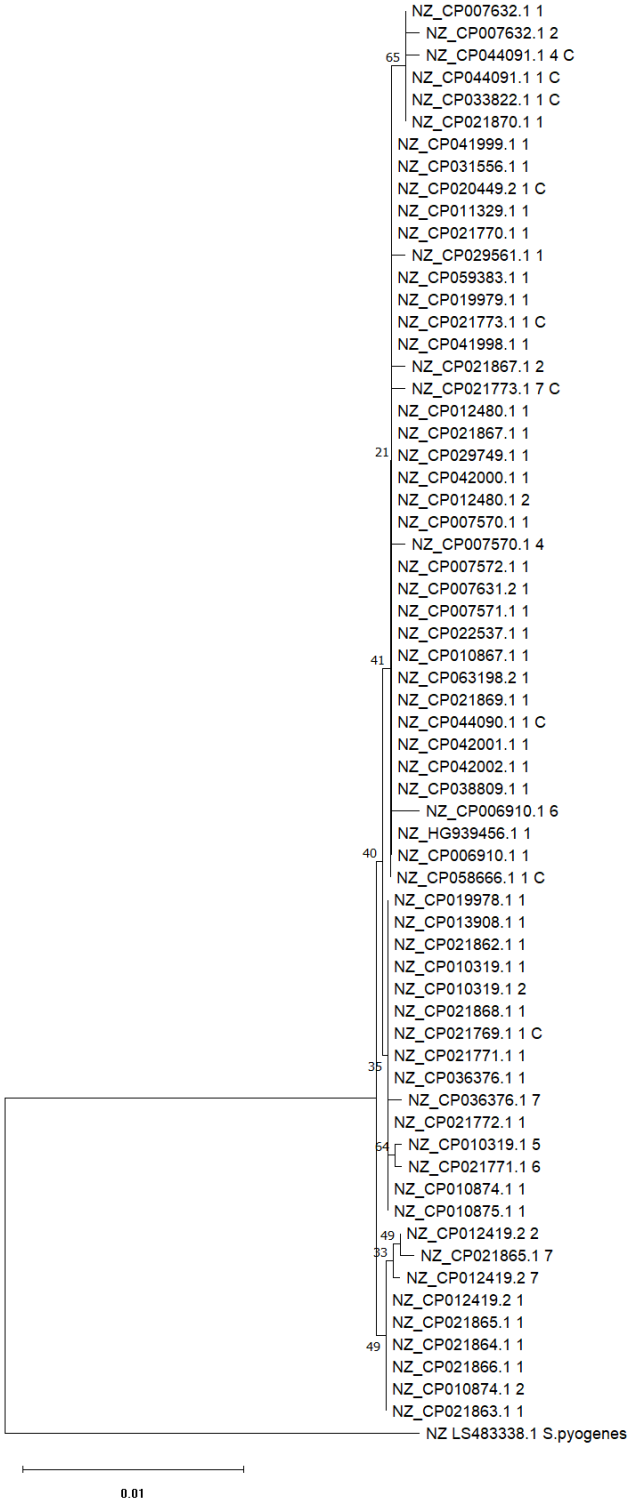

Fig. 3