bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 # Microevolution of aquatic *Streptococcus agalactiae* ST-261
2 # from Australia indicates dissemination via imported tilapia
3 # and ongoing adaptation to marine hosts or environment
4
5

6 Minami Kawasaki[1], Jerome Delamare-Deboutteville[1], Rachel O Bowater[2], Mark J Walker[3], Scott
7 Beatson[3], Nouri L Ben Zakour[3a] & Andrew C Barnes[1*]
8 [1]*The University of Queensland, School of Biological Sciences and Centre for Marine Science, St*
9 *Lucia Campus, Brisbane, Queensland, Australia*
10 [2]*Biosecurity Queensland (North Region), Department of Agriculture and Fisheries Townsville,*
11 *Queensland, Australia*
12 [3]*The University of Queensland, School of Chemistry and Molecular Biosciences and Australian*
13 *Infectious Diseases Research Centre,, St Lucia Campus, Brisbane, Queensland, Australia*
14
15
16
17 **Keywords:**
18 Streptococcus agalactiae, evolution, genome analysis, epidemiology, fish, aquaculture
19
20
21
22
23
24
25
26 ***Correspondence**
27 **a.barnes@uq.edu.au**
28 [a]Present address: Westmead Institute for Medical Research, University of Sydney, NSW, Australia

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29 **Abstract**

30 *Streptococcus agalactiae* (GBS) causes disease in a wide range of animals. The serotype 1b lineage is
31 highly adapted to aquatic hosts, exhibiting substantial genome reduction compared with terrestrial
32 con-specifics. Here we sequence genomes from 40 GBS isolates including 25 from wild fish and
33 captive stingrays in Australia, six local veterinary or human clinical isolates, and nine isolates from
34 farmed tilapia in Honduras and compare with 42 genomes from public databases. Phylogenetic
35 analysis based on non-recombinant core genome SNPs indicated that aquatic serotype Ib isolates
36 from Queensland were distantly related to local veterinary and human clinical isolates. In contrast,
37 Australian aquatic isolates are most closely related to a tilapia isolate from Israel, differing by only 63
38 core-genome SNPs. A consensus minimum spanning tree based on core genome SNPs indicates
39 dissemination of ST-261 from an ancestral tilapia strain, which is congruent with several
40 introductions of tilapia into Australia from Israel during the 1970s and 1980s. Pan-genome analysis
41 identified 1,440 genes as core with the majority being dispensable or strain-specific with
42 non-protein-coding intergenic regions (IGRs) divided amongst core and strain-specific genes.
43 Aquatic serotype Ib strains have lost many virulence factors during adaptation, but six adhesins were
44 well conserved across the aquatic isolates and might be critical for virulence in fish and targets for
45 vaccine development. The close relationship amongst recent ST-261 isolates from Ghana, USA and
46 China with the Israeli tilapia isolate from 1988 implicates the global trade in tilapia seed for
47 aquaculture in the widespread dissemination of serotype Ib fish-adapted GBS.

48

49 **Importance**

50 *Streptococcus agalactiae* (GBS) is a significant pathogen of humans and animals. Some lineages
51 have become adapted to particular hosts and serotype Ib is highly specialized to fish. Here we show
52 that this lineage is likely to have been distributed widely by the global trade in tilapia for aquaculture,
53 with probable introduction into Australia in the 1970s and subsequent dissemination in wild fish
54 populations. We report variability in the polysaccharide capsule amongst this lineage, but identify a
55 cohort common surface proteins that may be a focus of future vaccine development to reduce the
56 biosecurity risk in international fish trade.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

57 **1. Introduction**

58 *Streptococcus agalactiae,* or Lancefield Group B *Streptococcus* (GBS), is a commensal and
59 occasionally pathogenic bacterium with a very diverse host range. A common commensal in the
60 urogenital tracts of humans, GBS is also a leading cause of morbidity in newborns causing
61 meningitis, septicaemia and pneumonia (1-4). *S. agalactiae* can cause septicaemic infections in
62 cattle, domestic dogs and cats, in camels, and reptiles and amphibians (5-8). In fish, disease outbreaks
63 caused by *S. agalactiae* have substantial impact on the aquaculture industry, particularly the
64 production of warm fresh water species such as tilapia (*Oreochromis* spp.)(9-12). Most outbreaks to
65 date in freshwater farmed fish have resulted from infection by highly adapted strains of GBS with
66 genomes that are 10-15% smaller than their terrestrial conspecifics (13). Unusually, *S. agalactiae*
67 also causes significant mortality in wild aquatic animals including grouper, stingrays and mullet (5)
68 suggesting further adaptation to marine as well as freshwater aquatic hosts.
69

70 Microevolution within a bacterial species can be driven by host or environmental adaptation (13, 14),
71 permitting inference of the epidemiology of disease outbreaks and how pathogens may have
72 transferred within and between geographic regions (14-16). This requires analysis of factors that
73 evolve at sufficiently rapid pace to be informative over relatively short timespans. In GBS, capsular
74 serotyping either with antibodies or by 'molecular serotyping' (sequencing of the capsular operon)
75 has become a widely used method of typing for population studies (17-20) and, currently, *S.*
76 *agalactiae* can be divided into ten capsular serotypes (Ia, Ib and II-IX) (18, 21). Determining capsular
77 serotypes is also critical for vaccine formulation since CPS is highly immunogenic and can confer
78 excellent protection against infections by the homologous CPS serotype (20, 22, 23). Further typing
79 resolution is provided by multilocus sequence typing (MLST), a method that has been employed to
80 great effect to conduct global population studies of isolates based on genetic variations amongst
81 relatively slowly evolving housekeeping genes (17). Combining molecular serotyping and MLST in

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

82 the analysis of *S. agalactiae* revealed that the majority of isolates associated with aquatic
83 environments and hosts fall within serotypes Ia and Ib, in which Ia isolates belong to ST-7 in clonal
84 complex (CC) 7 and ST-103 in CC103 (12, 18, 24-28). Serotype Ib strains isolated in Central and
85 South America are ST-260 and ST-552 in CC552 (27, 29) and strains isolated in Australia, Israel,
86 Belgium, China, Ghana, USA and Southeast Asia belong to ST-261 (5, 6, 9, 13, 29-31). Serotype III
87 is commonly causative of disease in humans, but has also been isolated from fish in Thailand, China
88 and recently in Brazil (26, 32-34).
89
90 Whilst capsular serotyping and MLST have been useful in inferring origin and dispersal of GBS
91 subtypes, they do not display sufficient resolution to explore spread and evolution within individual
92 sequence types nor can they reflect the complete genetic diversity of *S. agalactiae* (35). The rapid fall
93 in cost of whole genome sequencing coupled with multiplexing and rapid development of open
94 source bioinformatics tools has permitted much deeper analysis of evolution, host adaptation and
95 epidemiological modelling within single bacterial species (36) including those from aquatic hosts
96 (14, 15). Bacteria such as *S. agalactiae* that can colonise multiple host species often have greater
97 genomic intraspecies diversity (37). In GBS, two major evolutionary trends have been implicated in
98 rapid adaptation to new hosts, namely, acquisition of new genes by lateral gene transfer and genome
99 reduction via gene loss integral to host specialisation (13, 38). For example, *S. agalactiae* Ia strains,
100 GD201008-001 and ZQ0910 isolated from tilapia in China carry a 10 kb genomic island (GI), which
101 is absent from their closely related human isolate A909. Moreover, this 10 kb GI bears many
102 similarities with *Streptococcus anginosus* SK52/DSM 20563 genome sequence suggesting possible
103 transfer from *S. anginosus* to GBS, with implications for virulence in Tilapia (13, 39). During fish
104 host adaptation, serotype Ib strains have undergone reductive evolution resulting in 10-25% of their
105 genomes being lost compared to terrestrial *S. agalactiae* isolates and serotype Ia piscine strains (13).

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

106 Evolution of *S. agalactiae* by genome reduction is an ongoing process with a high number of
107 pseudogenes present in GBS genomes from aquatic sources (13).
108
109 Evolution of *S. agalactiae* and adaptation to aquatic hosts is an incomplete and ongoing process,
110 consequently sequencing the genomes from a few isolates is insufficient to understand the full
111 potential genetic diversity of *S. agalactiae* as a species (35). The pan-genome or supra-genome of a
112 bacterial species defines the full complement of genes, or the union of all the gene sets, within the
113 species (35). This pan-genome is subdivided into its core genome, which includes all the genes that
114 are present in all the strains of the same bacterial species and must therefore be responsible for
115 essential biological functions to allow the species to survive, and the accessory genome containing
116 species-specific genes that are unique to single strains or constrained to a cohort of strains within the
117 species; these genes contribute to the diversity makeup of the species. The pan-genome of a species
118 resolves the true genomic diversity of that species and permits the identification of gene cohorts that
119 are essential to the species as a whole, along with gene complements in the accessory genome that
120 permit host or habitat specialization (35). Moreover, by identifying potential antigens within the
121 pan-genome that are conserved across all strains that infect a particular host type, vaccine targets can
122 be specified that are likely to cross-protect (35, 40). Indeed, the first multicomponent
123 protein-containing universal vaccine against human *S. agalactiae* was developed using a pan-genome
124 reverse vaccinology approach by analysing eight human isolates to predict putative antigens that
125 were conserved amongst those strains (41). Some antigens in this vaccine are in the accessory
126 genome consequently it is important to analyse as large a dispensable genome as possible for vaccine
127 development (35, 41).
128

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

129 The *S. agalactiae* pan-genome is now well-advanced but still "open" (new genes continue to be added
130 with more sequenced genomes) and geographically constrained (35, 40). In the present study, we
131 sequenced genomes of new aquatic *S. agalactiae* strains isolated from tilapia in Honduras and from
132 wild and captive marine fish in Australia. We infer potential epidemiological distribution of ST-261
133 in aquatic hosts in Australia and show continuing adaptation to salt water fish. Moreover, we identify
134 conserved surface proteins across the ST-260 and ST-261 sequence types that may have potential for
135 incorporation into for aquaculture of important food fish species such as tilapia and grouper.
136

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

137 **2. Materials and Methods**

138 *2.1. Bacterial strains and culture conditions*

139 Forty *S. agalactiae* isolates comprising strains collected from fish in Honduras and Australia, along
140 with reptiles, humans and other terrestrial mammals from Australia were chosen for sequencing
141 (Table 1). Of these 40 isolates, 25 strains were collected from several species of fish in Queensland,
142 Australia, two human clinical strains were from Queensland, Australia, one strain was isolated from
143 saltwater crocodile (*Crocodylus porosus*) in the Northern Territory, Australia and three isolates were
144 collected from domestic animals including cats, dogs and cattle in Australia. Additionally, nine
145 isolates originating from disease in farmed tilapia in Honduras were sequenced during this study. All
146 isolates were maintained at -80ºC in Todd Hewitt broth (THB) (Oxoid) containing 25% glycerol as
147 frozen stock. The isolates were recovered from stock on Columbia agar containing 5% sheep blood
148 (Oxoid) for 24 h at 28ºC. For liquid culture, the isolates were grown in THB for 18 h with low
149 agitation at 28ºC.

150

151 *2.2. DNA extraction and sequencing*

152 Genomic DNA (gDNA) was extracted from 10 ml early-stationary phase THB cultures with the
153 Qiagen DNeasy mini kit (Qiagen) according to the manufacturer's instructions. The quantity of
154 extracted DNA was measured by Qubit fluorimetry (Invitrogen) and the quality was checked by
155 agarose gel electrophoresis. To confirm the purity of the gDNA, the 16S rRNA gene was amplified
156 by polymerase chain reaction (PCR) using universal primers 27F and 1492R (42) and the PCR
157 products were sent to Australian Genome Research Facility (AGRF, Brisbane) for Sanger
158 sequencing. The 16S amplicon sequences were assembled in Sequencher V5.2.2 and analysed by
159 BLAST. Once identity and purity were confirmed, Nextera XT paired-end libraries were generated
160 using gDNA from each isolate and sequenced on the Illumina HiSeq2000 platform system (AGRF,
161 Melbourne).

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

162

163 *2.3.* De novo *assembly and annotation*

164 Illumina sequencing yielded between 5,288,952 and 12,577,340 read pairs for each strain. Read
165 quality control and contaminant screening were performed using FastQC (43). Reads were trimmed
166 using the clip function in Nesoni (https://github.com/Victorian-Bioinformatics-Consortium/nesoni)
167 then assembled *de novo* with SPAdes Assembler version 3.11 (44). The assemblies of fish isolates in
168 Queensland comprised about 1.8 Mbp of assembled sequence while terrestrial strains comprised 2
169 Mbp. The assembled contigs for all Queensland strains were reordered against an internal curated
170 reference genome from *S. agalactiae* strain QMA0271, using Mauve contig ordering tool (45).
171 Automated annotation was performed using Prokka 1.12 (46).

172

173 *2.4. Molecular serotyping and multilocus sequence typing (MLST)*

174 Reference sequences for the nine CPS serotypes (Table 2) (21) were retrieved from GenBank to
175 generate a database for prediction of capsular serotype from the draft genomes with Kaptive using
176 default settings (47). To determine multilocus sequence types (MLST), all draft assemblies were
177 analysed using the Center for Genomic Epidemiology web-tools MLST ver. 1.8, using *S. agalactiae*
178 configuration (available at https://cge.cbs.dtu.dk//services/MLST/) (48).

179

180 *2.5. Phylogenetic analysis*

181 To estimate approximate phylogenetic relationship among our strains and other isolates with whole
182 genome sequence available in GenBank (Table 1), a core genome single nucleotide polymorphism
183 (SNP)-based phylogenetic tree was constructed. Whole genome sequences of Queensland and
184 Honduras tilapia isolates, terrestrial isolates and the genomes obtained from GenBank were aligned
185 with Parsnp in the Harvest Tools suite version 1.2 (49). The genome of *Streptococcus pyogenes* M1

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

186 GAS (Accession number: NC_002737.2) was also included as an outgroup for tree rooting.
187 Hypothetical recombination sites in the core genome alignment were detected and filtered out with
188 Gubbins (50). Maximum likelihood phylogenetic trees were inferred with RAxML version 8.2.9 (51)
189 based on non-recombinant core genome SNPs under general time-reversible nucleotides substitution
190 model (GTR) with 1000 bootstrap replicates. Ascertainment bias associated with analysis of only
191 variable sites was accounted for using Felsenstein's correction implemented in RAxML (51). The
192 resulting tree was exported, rooted and nodes with low bootstrap supported collapsed with
193 Dendroscope, and the resulting tree/cladogram annotated with Evolview V2 (52).
194
195 In order to infer possible phylogenetic relationships within the aquatic ST-261 clade of GBS,
196 minimum spanning trees were constructed in MSTGold 2.4 (53) from a distance matrix based on core
197 genome SNPs derived by alignment of 27 ST-261 genomes in Geneious V9.1 (Biomatters Inc). A
198 consensus tree was constructed based on inference of 2400 trees and only those edges occurring in
199 greater than 50% of trees were included in the consensus.
200
201 *2.6. Pan-genome analysis*
202 A reference pan-genome was built with GView server (54) using our curated genome of Queensland
203 ST-261 serotype Ib grouper isolate, QMA0271 as a seed and 38 complete genomes from public
204 databases added sequentially (Table 1). Only complete genomes were included into the reference
205 pan-genome to avoid incomplete genes associated with the high fragmentation of draft sequences.
206 For visualisation, draft and complete genomes were compared with the reference pangenome using
207 BRIG 0.95 (55). To investigate core and accessory protein-coding genes, sequences from all strains
208 were analysed with Roary (56) using default settings. Since Roary only considers protein-coding

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

209 sequences, we used Piggy (57) to examine non protein-coding intergenic regions (IGRs), which
210 comprise about 15% of the GBS genomes.
211
212 *2.7. Identification and comparison of virulence factors*
213 Virulence factor screening was performed using SeqFindr (58) by comparing the assembled genomes
214 of all strains in this study along with the genomes available through GenBank (Table 1) against a list
215 of 51 *S. agalactiae*-specific virulence factors sequences collected from the Virulence Factors
216 Database (VFDB) (59), complemented by six additional sequences identified in the sequenced
217 ST-261 *S. agalactiae* strain ND2-22 (Table 1).
218
219 *2.8. Analysis of effects of SNPs in ST-261 clade*
220 Putative effects of SNPs amongst and between the ST-261 clade were determined using SNPeff
221 version 4.3p (60). First, a new *S. agalactiae* database was constructed from the Genbank-formatted
222 curated reference genome for QMA0271 in accordance with the manual
223 (http://snpeff.sourceforge.net/SnpEff_manual.html#databases). Then a VCF file, generated from
224 curated SNPs generated by alignment of complete genomes of 27 ST-261 isolates in Geneious
225 version 9.1, was annotated for SNP effect with SNPeff using the database from QMA0271 as
226 reference (Supplementary information).
227

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

228 **3. Results and Discussion**

229 *3.1. GBS isolates from marine fish and rays in Queensland and tilapia in Honduras belong to*
230 *differing host-adapted lineages*

231 The average size of draft genomes of isolates from fish and rays in Queensland was 1,801,022 bp,
232 containing 1,881 genes on an average, while the mean genome size of strains from Honduran tilapia
233 was 1,801,133 bp with an average gene number of 1,869, consistent with the small genomes
234 associated with the host-adapted aquatic lineage (13). Molecular serotyping indicated that all
235 Queensland and Honduras aquatic strains belong to serotype Ib but Queensland isolates belong to
236 sequence type (ST)-261 while Honduras strains are ST-260. Both ST-260 and ST-261 have been
237 identified infecting aquatic animals previously with ST-260 isolates found in tilapia from Brazil and
238 Costa Rica and occupying clonal complex (CC) 552 along with ST-552 and ST-553 strains also
239 isolated from tilapia in Latin America (27), whilst ST-261 has been isolated from tilapia in USA,
240 China, Ghana and Israel (9, 13, 30).

241 The current isolates represent the first aquatic isolates sequenced from marine fish and stingrays
242 indicating a wider host range (rays and marine finfish), environmental (freshwater to marine) and
243 geographic distribution of ST-261 than previously observed (5, 6, 61). GBS strains isolated from
244 humans and terrestrial animals in Queensland and Northern Territory, Australia were also sequenced
245 and have larger genomes of 2,072,596 bp comprising 2,067 genes on average, suggesting that recent
246 possible local transfer from terrestrial origin to Australian fish is highly unlikely, although probable
247 transfer between humans and fish has been reported for ST-7 GBS elsewhere (8, 39, 62). Non-human
248 mammalian strains from Australia sequenced in this study QMA0300 and QMA0303 belong to ST-1
249 serotype V, and QMA0306 belong to ST-67 serotype III, whereas human isolates, QMA0355 and
250 QMA0357, and crocodile strain QMA0336 belong to ST-23 serotype Ia. Indeed, the high sequence
251 identity between the human ST-23 serotype Ia isolates and those from farm-raised crocodiles are
252 supportive of probable human transfer to these animals that has been implied previously (63).

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

253 To better inform possible origin of the ST-261 isolates from marine fish in Australia, a phylogenetic
254 tree was constructed by maximum likelihood from 29,689 non-recombinant core genome SNPs and
255 short indel positions derived by alignment of whole genome sequences of 25 Queensland fish
256 isolates, 9 Honduras isolates, 6 Queensland terrestrial isolates and 42 genomes from public databases
257 (Fig. 1A). Two distinct groups were resolved, the first comprising entirely of aquatic isolates
258 (including serotype Ib isolates from ST552, ST-260 and ST-261) while the second major group
259 comprised various isolates of terrestrial origin and some fish isolates from ST7 serotype Ia that may
260 have infected fish via transfer from terrestrial sources (Fig. 1A). The serotype Ib aquatic group
261 branched into three distinct lineages based on ST (Fig. 1A). One lineage comprised all Honduras
262 strains of ST-260, which were derived from a lineage comprising ST-552 isolates from Brazil (Fig.
263 1A). This is supportive of previous observations in which an extended typing system based on MLST,
264 virulence genes and serotype indicated geographic endemism within fish isolates from differing
265 regions of Brazil (64). The isolates belonging to ST-261 from USA, Israel, Ghana, China and
266 Queensland clustered together (Fig. 1A). The second major division, containing serotype Ia fish
267 strains along with terrestrial isolates, was more complex, but isolates largely clustered in line with
268 serotype and ST (Fig. 1A). The aquatic serotype Ia isolates clustered with human isolates of ST-7
269 (Fig 1A). These fish isolates have acquired a 10kb genomic island, putatively from *S. anginosus*, in
270 contrast to their human ST7 serotype Ia relatives (39). Lineage 10 comprised serotype III ST-17
271 human isolates (Fig. 1A). Serotype Ia strains were divided amongst several further ST groups: Three
272 Australian serotype Ia ST-23 strains (QMA0336, QMA0355 and QMA0357) isolated from humans
273 and a saltwater crocodile clustered together with strains of serotype III ST-23 (NEM316) and
274 serotype IV ST-452 (NGBS572) also of human origin (Fig. 1A). This lineage appears to be derived
275 from NEM316 which is a frequent cause of late-onset disease in human infants (65). A further
276 serotype Ia isolate was located in ST-103 and clustered together with two strains of serotype V

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

277 ST-609 and ST-617 isolated from camels (66) (Fig. 1A). Two newly sequenced Australian terrestrial
278 strains, QMA0300 isolated from a dog and QMA0303 isolated from a cat clustered with other
279 serotype V ST-1 strains isolated from humans, cattle and dog hosts (Fig. 1A). This lineage also
280 contained NGBS061 serotype IV ST-495 and GBS-M002 serotype VI ST-1 (Fig. 1A). ST-1 emerged
281 as a significant cause of infection and disease in humans during the 1990s, but was recently inferred
282 to have evolved from strains causing mastitis in cattle in the 1970s (67). Moreover, QMA0306 ST-1
283 serotype V from cattle in Queensland was closely related to SA111 ST-61 serotype II, which
284 represents a host-adapted lineage of *S. agalactiae* that is dominant in cattle in Europe (68)(Fig. 1A).
285
286 Our phylogeny based on whole genome SNPs does not support recent direct transfer of GBS from
287 Australian human clinical or terrestrial animal sources to marine fish and stingrays, in spite of close
288 proximity of many of the wild fish cases to human habitation (6). Consequently, we refined our
289 analyses to the ST-261 aquatic host-adapted lineage to attempt to infer possible route of introduction
290 and subsequent evolution in Australian marine fish. A consensus minimum spanning tree based on a
291 distance matrix comprised of all core genome SNPs derived from alignment of the ST-261 serotype
292 Ib strains revealed a likely original introduction via tilapia from Israel, with only 63 core genome
293 SNPs separating an Australian stingray isolate from the type strain of *S. difficile* (re-assigned as *S.*
294 *agalactiae* serotype Ib (69)), isolated from tilapia in Israel in 1988 (Fig 1B)(70). Tilapia were
295 imported on a number of occasions during the 1970s and 1980s from Israel into North Queensland
296 around Cairns and Townsville, and a number of strains and hybrids have since colonised rivers and
297 creeks throughout Queensland (71). Globally, the aquatic ST-261 lineage appears to have been
298 transferred through human movements of tilapia for aquaculture and other purposes over the last
299 several decades. The US and Chinese tilapia isolates also appear to derive from the early ND2-22
300 isolate, as do recent isolates from tilapia in Ghana (Fig 1B). Indeed, phylogenetic analysis by

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

301 maximum likelihood of draft whole genome alignments suggest that the Ghanaian and Chinese
302 isolates share a recent common ancestor that is derived from ND2-22, with only 60 SNPs separating
303 ND2-22 and the Ghanaian strains (72). We identified only 36 core genome SNPs separating the
304 Ghanaian isolates from ND2-22 but this reflects the smaller core genome in our study as a result of
305 the high number of GBS isolates analysed (40 in the present study compared with 9 in the previous
306 study (72).
307
308 The minimum spanning tree implicates continued adaptation of the ST-261 lineage post introduction
309 and suggests that grouper (marine Teleostei family) may have been infected via estuary stingrays
310 (Dasyatidae family) (Fig 1B). Stingrays are occasional prey for adult giant Queensland grouper and
311 stingray barbs have been found in the gut of grouper post-mortem (Bowater, unpublished
312 observation). ST-261 GBS has also caused mortality in captive stingrays in South East Queensland,
313 translocated from Cairns in North Queensland (61)
314
315 *3.2. The S. agalactiae pan-genome comprises a small core of protein-coding genes and is open*
316 To further elucidate adaptation amongst the fish pathogenic GBS types a pan-genome was built from
317 39 complete genomes retrieved from GenBank and using our curated ST-261 grouper isolate
318 QMA0271 as a high-quality reference seed genome. The resulting pan-genome was 4,074,275 bp
319 (Fig. 2A). All-vs-all BLAST analysis of the genomes implemented in BRIG clearly indicated the
320 substantial reduction in genome size amongst the fish pathogenic ST-261 cohort, as previously
321 reported for a limited number of ST-261 isolates (13). Here, we find that ST-260 and ST552
322 fish-pathogenic sequence types within serotype Ib are similarly reduced and that conservation
323 amongst the serotype Ib strains is high (Figure 2A). In total 4,603 protein-coding genes were
324 predicted in *S. agalactiae* pan-genome using Roary (Fig. 2B), which is consistent with previous
325 research (39). The number of core genes was 1,440 (representing 35% of the pan-genome) while

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

326 previous studies reported 1,202 to 1,267 genes in the pan-genome (39, 40). These differences may
327 result from the methods being used to examine the pan-genome, the difference in sequences number
328 being used to create the pan-genome and finally the use of draft sequences in the analysis (39). A
329 majority of protein-coding genes found in the pan-genome belonged to both the dispensable and
330 strain-specific genes. This could be a result of the inclusion of a high number of serotype Ib strain
331 sequences, which were all significantly smaller in size (approx. 1.8 Mbp) when compared with other
332 isolates. Liu et al. (2013) demonstrated that removing Ib piscine isolates from their analysis resulted
333 in an increase in the number of core genes.
334
335 Frequency analysis of IGRs in *S. agalactiae* pan-genome showed that the number of IGRs shared
336 across all strains was smaller than core protein-coding genes, whereas strain-specific IGRs was much
337 higher than protein-coding strain-specific genes (Fig. 2C). IGR analysis with Piggy excludes IGRs
338 that are less than 30 bp which may result in fewer IGRs than protein-coding genes in core regions
339 (57). Most IGRs identified in the pan-genome belonged to either core genes or strain-specific genes
340 (Fig. 2C) in line with previous findings in *Staphylococcus aureus* ST22 and *Escherichia coli* ST131
341 where similar distributions of IGRs were detected (57). The gradients of the accumulation curves for
342 both protein-coding genes and IGRs are still strongly positive therefore the *S. agalactiae* pan-genome
343 remains open.
344
345 *3.3. Aquatic serotype Ib have a reduced repertoire of virulence factors*
346 Almost all genes classed as adhesins, involved in immune evasion and host invasion, and most of the
347 toxin-related genes found in terrestrial isolates were absent from serotype Ib aquatic isolates (Fig. 3).
348 Rosinski-Chupin et al. (2013) reported ~60% of the virulence genes found in human strains were
349 present in a serotype Ib GBS strain from fish. We found that CAMP factor gene *cfa*/*cfb* was present in
350 all strains including serotype Ib ST-260, 261 and 552 isolates, but CAMP reaction was previously

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

351 reported to be negative for ST-260 and 261 strains (13, 28). These authors identified that CAMP
352 factor gene in ST-261 is disrupted while the gene in ST-260 is unaltered but the level of gene
353 expression may be too low to detected by the test (13). Most of the genes in *cyl* locus have been lost
354 from Ib piscine strains and only *cylB*, an ABC ATP binding cassette transporter was present in
355 ST-260 and 552 (Fig. 3). The *cyl* locus is responsible for hemolytic activity and production of
356 pigment via co-transcription of *cylF* and *cylL* (73). In ST-261, the *cyl* gene cluster is replaced by a
357 genomic island resulting in loss of hemolytic activity (13). Of particular relevance to virulence and
358 antigenic diversity, transmembrane immunoglobulin A-binding C protein beta-antigen *cba* was
359 absent in all aquatic isolates (Fig 3). This gene has been reported in type Ib and Ia GBS previously
360 (74), is implicated in virulence in neonates and is up regulated in GBS serotype Ia strain A909 in
361 response to human serum (74-76). In contrast, capsule-related genes were largely conserved and
362 associated with serotype (Fig. 3) supportive of the major role of capsule in virulence of fish
363 pathogenic streptococci (77, 78).
364
365 Although serotype Ib aquatic isolates have lost the majority of virulence factors found in terrestrial
366 isolates, most contained six sequences that were identified as probable adhesins by homology (Fig.
367 3). These ST-261 adhesins (named 0337, 0626, 0856, 1185, 1196 and 1648 based on position in the
368 annotated ST-261 genome from QMA0271) were fully conserved amongst aquatic serotype Ib
369 ST-261 (Fig. 3). Moreover, ST-261 adhesin 0856 was only present in ST-261 and two serotype II
370 isolates, GBS1-NY and SA111 strains (Fig. 3). ST-261 adhesins 1185 and 1648 were shown to be
371 unique to aquatic serotype Ib isolates, being absent from Ia fish isolates and all terrestrial strains (Fig.
372 3). The analysis indicated that ST-261 adhesins 0337, 0626 and 1196 were well conserved across
373 most of the isolates regardless of origin but 1196 was the only adhesin-like gene present in all strains
374 analysed (Fig. 3). ST-261 adhesin 0626 was also ubiquitous but contained deletions in Sag37 isolate

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

375 (Fig. 3). Some terrestrial strains, such as QMA0355, 0357, 0336 and NGBS572 lacked adhesin 0337
376 (Fig. 3).
377
378 As capsular polysaccharide is a requirement for full virulence in several fish pathogenic streptococci
379 (77, 78) and the major antigen in GBS (79-81), further analysis of the serotype Ib *cps* operon was
380 conducted. Within the serotype Ib lineage, the capsular polysaccharide operon was well conserved
381 (Fig. 4). However, QMA0281 from grouper had a deletion at position 341 in *cpsB,* encompassing
382 *cpsC, cpsD* and *cpsE* (Fig 4), resulting in a chimeric ORF. Moreover, an insertion in a TA repeat at
383 position 557 in *cpsH* resulted in an early stop codon marginally reducing gene size (Fig. 4).
384 QMA0368 had a TT insertion at nucleotide position 745 in *cpsE* resulting in insertion of an early stop
385 codon and truncation of the gene (Fig. 4). Deletion of this region that includes the priming glycosyl
386 transferase for capsular biosynthesis results in loss of capsule, attenuated virulence and modified
387 pathology in the fish pathogen *S. iniae* (82). Buoyant density analysis in Percoll indicated that
388 QMA0281 is also deficient in capsule (not shown). *cps* operon SNPs amongst the Australian serotype
389 Ib isolates were rare (Fig. 4). Indeed only 1 SNP *neuA* separated QMA0271 from the 1988 tilapia
390 isolate from Israel suggesting very little evolution of the *cps* operon since introduction of the lineage
391 (Fig. 4). This may reflect a well-adapted capsule for colonization naïve hosts as the Australian
392 isolates were from wild fish and captive stingrays recently after capture and transport thus placing
393 little selective pressure for novel capsular sequence types. Immunity in fish drives evolution of novel
394 capsular sequence types in *S. iniae* but these reported cases were all in high intensity aquaculture with
395 occasional use of autogenous vaccination and opportunity for development of cps-specific immunity
396 (82). This may explain the relatively high number of SNPs in genes encoding sugar and sialic acid
397 modifying enzymes amongst the isolates from tilapia farmed in Honduras relative to the other isolates
398 examined, including those from tilapia, as autogenous vaccination is occasionally used on farms

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

399 where isolates were sourced, and may apply selective pressure favouring modified polysaccharide
400 capsule.
401
402 **Conclusions**
403 The clade of aquatic *S. agalactiae* serotype Ib including ST-260, ST-261 and ST-552 is a highly
404 adapted fish pathogen with a substantially reduced genome compared to all other serotypes from
405 terrestrial mammalian, reptile and fish hosts. These variants were originally identified as *S. difficile*
406 due to their impoverished growth on laboratory media and hence difficulty in isolating from diseased
407 fish (70). The species was reassigned to serotype Ib GBS a few years later (69) but the recent
408 discovery of the major reduction in genomes size of ST-261 serotype Ib (13) explains the marked
409 phenotypic difference that merited early phenotypic assignation to its own species. Other serotypes
410 have been isolated from fish, notably serotype Ia and serotype III, but these seem to arise from
411 terrestrial transfer rather than being retained amongst the aquatic host population (28). In contrast,
412 serotype Ib has only been isolated from fish and stingrays, appears to be well-adapted and is likely to
413 have been transferred internationally via trade in domesticated tilapia, evidenced here by the very
414 close relationship (only a handful of SNPs) between a strain isolated from tilapia in 1988 in Israel and
415 those found in fish and stingrays in Australia since 2008, and in tilapia in the USA, China and Ghana.
416 The ST-261 lineage in Australia likely arrived with several introductions of tilapia in the 1970s and
417 1980s. Tilapia are classed as a noxious pest in Australia and import was banned, but not before
418 several lines became established throughout tropical and subtropical freshwater habitats in
419 Queensland (71). Although ST-261 serotype Ib GBS has not been isolated from farm fish in
420 Queensland in spite of proximity of freshwater farms to tilapia-infested creeks, this clade of GBS is a
421 substantial problem in farmed tilapia globally. The cohort of putative adhesins identified here to be
422 conserved through all fish pathogenic serotypes (including Ia and III in addition to Ib) may be

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

423 promising candidates for cost-effective cross-serotype protective vaccines for aquaculture and are
424 worthy of future research.

425 **Acknowledgements**

426 This work was funded by the Fisheries Research and Development Corporation Aquatic Animal
427 Health Subprogram Project 2010/34 awarded to ROB and ACB, and the Australian Research Council
428 Linkage Program LP130100242 awarded to ACB, MJW and SB.

429

430 **References**

431 1. Brochet M, Couvé E, Zouine M, Vallaeys T, Rusniok C, Lamy M-C, Buchrieser C,
432 Trieu-Cuot P, Kunst F, Poyart C. 2006. Genomic diversity and evolution within the species
433 Streptococcus agalactiae. Microbes and Infection 8:1227-1243.
434 2. Edwards MS, Gonik B. 2013. Preventing the broad spectrum of perinatal morbidity and
435 mortality through group B streptococcal vaccination. Vaccine 31:D66-D71.
436 3. Six A, Bellais S, Bouaboud A, Fouet A, Gabriel C, Tazi A, Dramsi S, Trieu-Cuot P, Poyart C.
437 2015. Srr2, a multifaceted adhesin expressed by ST-17 hypervirulent Group B Streptococcus
438 involved in binding to both fibrinogen and plasminogen. Molecular microbiology
439 97:1209-1222.
440 4. Doumith M, Mushtaq S, Martin V, Chaudhry A, Adkin R, Coelho J, Chalker V, MacGowan
441 A, Woodford N, Livermore DM. 2017. Genomic sequences of Streptococcus agalactiae with
442 high-level gentamicin resistance, collected in the BSAC bacteraemia surveillance. Journal of
443 Antimicrobial Chemotherapy 72:2704-2707.
444 5. Delamare-Deboutteville J, Bowater R, Condon K, Reynolds A, Fisk A, Aviles F, Barnes AC.
445 2015. Infection and pathology in Queensland grouper, Epinephelus lanceolatus, (Bloch),
446 caused by exposure to Streptococcus agalactiae via different routes. Journal of Fish Diseases
447 38:1021-1035.
448 6. Bowater RO, Forbes-Faulkner J, Anderson IG, Condon K, Robinson B, Kong F, Gilbert GL,
449 Reynolds A, Hyland S, McPherson G, Brien JO, Blyde D. 2012. Natural outbreak of
450 Streptococcus agalactiae (GBS) infection in wild giant Queensland grouper, Epinephelus
451 lanceolatus (Bloch), and other wild fish in northern Queensland, Australia. Journal of Fish
452 Diseases 35:173-186.
453 7. Elliott JA, Facklam RR, Richter CB. 1990. Whole-cell protein patterns of nonhemolytic
454 group B, type Ib, streptococci isolated from humans, mice, cattle, frogs, and fish. Journal of
455 Clinical Microbiology 28:628-630.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

456 8. Evans JJ, Klesius PH, Gilbert PM, Shoemaker CA, Al Sarawi MA, Landsberg J, Duremdez R,
457 Al Marzouk A, Al Zenki S. 2002. Characterization of beta-haemolytic Group B Streptococcus
458 agalactiae in cultured seabream, Sparus auratus L., and wild mullet, Liza klunzingeri (Day), in
459 Kuwait. Journal of Fish Diseases 25:505-513.
460 9. Verner-Jeffreys D, Wallis T, Cano Cejas I, Ryder D, Haydon D, Domazoro JF, Dontwi J,
461 Field T, Adjei-Boteng D, Wood G. 2017. Streptococcus agalactiae Multilocus sequence type
462 261 is associated with mortalities in the emerging Ghanaian tilapia industry. Journal of Fish
463 Diseases.
464 10. Li K, Liu L, Clausen JH, Lu M, Dalsgaard A. 2016. Management measures to control diseases
465 reported by tilapia (Oreochromis spp.) and whiteleg shrimp (Litopenaeus vannamei) farmers
466 in Guangdong, China. Aquaculture 457:91-99.
467 11. Chen M, Li L-P, Wang R, Liang W-W, Huang Y, Li J, Lei A-Y, Huang W-Y, Gan X. 2012.
468 PCR detection and PFGE genotype analyses of streptococcal clinical isolates from tilapia in
469 China. Veterinary Microbiology 159:526-30.
470 12. Zhang D, Li A, Guo Y, Zhang Q, Chen X, Gong X. 2013. Molecular characterization of
471 Streptococcus agalactiae in diseased farmed tilapia in China. Aquaculture 412:64-69.
472 13. Rosinski-Chupin I, Sauvage E, Mairey B, Mangenot S, Ma L, Da Cunha V, Rusniok C,
473 Bouchier C, Barbe V, Glaser P. 2013. Reductive evolution in Streptococcus agalactiae and the
474 emergence of a host adapted lineage. BMC genomics 14:252.
475 14. Brynildsrud O, Feil EJ, Bohlin J, Castillo-Ramirez S, Colquhoun D, McCarthy U, Matejusova
476 IM, Rhodes LD, Wiens GD, Verner-Jeffreys DW. 2014. Microevolution of *Renibacterium*
477 *salmoninarum*: evidence for intercontinental dissemination associated with fish movements.
478 ISME Journal 8:746-56.
479 15. Barnes AC, Delamare-Deboutteville J, Brosnahan C, Gudkovs N, Morrison RN, Carson J.
480 2016. Whole genome analysis of Yersinia ruckeri isolated over 27 years in Australia and New
481 Zealand reveals geographical endemism over multiple lineages and recent evolution under
482 host selection. Microbial Genomics 2.
483 16. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, Van Dijl JM, Laurent F, Grundmann H,
484 Friedrich AW, Markers ESGoE. 2013. Overview of molecular typing methods for outbreak
485 detection and epidemiological surveillance. Euro Surveill 18:20380.
486 17. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan M-S, Kunst F, Glaser P, Rusniok C,
487 Crook DWM, Harding RM. 2003. Multilocus sequence typing system for group B
488 streptococcus. Journal of Clinical Microbiology 41:2530-2536.
489 18. Carvalho-Castro GA, Silva JR, Paiva LV, Custódio DAC, Moreira RO, Mian GF, Prado IA,
490 Chalfun-Junior A, Costa GM. 2017. Molecular epidemiology of Streptococcus agalactiae
491 isolated from mastitis in Brazilian dairy herds. Brazilian Journal of Microbiology 48:551-559.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

492 19. Kong F, Gowan S, Martin D, James G, Gilbert GL. 2002. Serotype Identification of Group B
493 Streptococci by PCR and Sequencing. Journal of clinical microbiology 40:216-226.
494 20. Kapatai G, Patel D, Efstratiou A, Chalker VJ. 2017. Comparison of molecular serotyping
495 approaches of Streptococcus agalactiae from genomic sequences. BMC genomics 18:429.
496 21. Sheppard AE, Vaughan A, Jones N, Turner P, Turner C, Efstratiou A, Patel D, Walker AS,
497 Berkley JA, Crook DW. 2016. Capsular typing method for streptococcus agalactiae using
498 whole-genome sequence data. Journal of clinical microbiology 54:1388-1390.
499 22. Eldar A, Shapiro O, Bejerano Y, Bercovier H. 1995. Vaccination with whole-cell vaccine and
500 bacterial protein extract protects tilapia against< i> Streptococcus difficile</i>
501 meningoencephalitis. Vaccine 13:867-870.
502 23. Evans JJ, Klesius PH, Shoemaker CA. 2004. Efficacy of Streptococcus agalactiae (group B)
503 vaccine in tilapia (Oreochromis niloticus) by intraperitoneal and bath immersion
504 administration. Vaccine 22:3769-3773.
505 24. Chong S, Wong W, Lee W, Tan Z, Tay Y, Teo X, Chee L, Fernandez C. 2017. Streptococcus
506 agalactiae outbreaks in cultured golden pomfret, Trachinotus blochii (Lacépède), in
507 Singapore. Journal of fish diseases 40:971-974.
508 25. Sun J, Fang W, Ke B, He D, Liang Y, Ning D, Tan H, Peng H, Wang Y, Ma Y. 2016.
509 Inapparent Streptococcus agalactiae infection in adult/commercial tilapia. Scientific reports 6.
510 26. Kannika K, Pisuttharachai D, Srisapoome P, Wongtavatchai J, Kondo H, Hirono I, Unajak S,
511 Areechon N. 2017. Molecular serotyping, virulence gene profiling and pathogenicity of
512 Streptococcus agalactiae isolated from tilapia farms in Thailand by multiplex PCR. Journal of
513 Applied Microbiology 122:1497-1507.
514 27. Godoy D, Carvalho-Castro G, Leal C, Pereira U, Leite R, Figueiredo H. 2013. Genetic
515 diversity and new genotyping scheme for fish pathogenic Streptococcus agalactiae. Letters in
516 applied microbiology 57:476-483.
517 28. Evans JJ, Bohnsack JF, Klesius PH, Whiting AA, Garcia JC, Shoemaker CA, Takahashi S.
518 2008. Phylogenetic relationships among Streptococcus agalactiae isolated from piscine,
519 dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish
520 epidemic in Kuwait is also associated with human neonatal infections in Japan. Journal of
521 Medical Microbiology 57:1369-1376.
522 29. Delannoy CMJ, Crumlish M, Fontaine MC, Pollock J, Foster G, Dagleish MP, Turnbull JF,
523 Zadoks RN. 2013. Human Streptococcus agalactiae strains in aquatic mammals and fish.
524 BMC microbiology 13:41.
525 30. Pridgeon JW, Zhang D. 2014. Complete genome sequence of a virulent streptococcus
526 agalactiae strain, 138P, isolated from diseased Nile Tilapia. Genome announcements
527 2:e00295-14.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

528 31. Lusiastuti AM, Textor M, Seeger H, Akineden Ö, Zschöck M. 2014. The occurrence of
529 Streptococcus agalactiae sequence type 261 from fish disease outbreaks of tilapia
530 Oreochromis niloticus in Indonesia. Aquaculture Research 45:1260-1263.
531 32. Suanyuk N, Kong F, Ko D, Gilbert GL, Supamattaya K. 2008. Occurrence of rare genotypes
532 of Streptococcus agalactiae in cultured red tilapia Oreochromis sp. and Nile tilapia O.
533 niloticus in Thailand—Relationship to human isolates? Aquaculture 284:35-40.
534 33. Li L, Wang R, Liang W, Gan X, Huang T, Huang Y, Li J, Shi Y, Chen M, Luo H. 2013. Rare
535 serotype occurrence and PFGE genotypic diversity of Streptococcus agalactiae isolated from
536 tilapia in China. Veterinary microbiology 167:719-724.
537 34. Chideroli RT, Amoroso N, Mainardi RM, Suphoronski SA, de Padua SB, Alfieri AF, Alfieri
538 AA, Mosela M, Moralez ATP, de Oliveira AG, Zanolo R, Di Santis GW, Pereira UP. 2017.
539 Emergence of a new multidrug-resistant and highly virulent serotype of Streptococcus
540 agalactiae in fish farms from Brazil. Aquaculture 479:45-51.
541 35. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV,
542 Crabtree J, Jones AL, Durkin AS. 2005. Genome analysis of multiple pathogenic isolates of
543 Streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the
544 National Academy of Sciences of the United States of America 102:13950-13955.
545 36. Francis AR, Tanaka MM. 2012. Evolution of variation in presence and absence of genes in
546 bacterial pathways. BMC evolutionary biology 12:55.
547 37. Puymège A, Bertin S, Guédon G, Payot S. 2015. Analysis of Streptococcus agalactiae
548 pan-genome for prevalence, diversity and functionality of integrative and conjugative or
549 mobilizable elements integrated in the tRNALys CTT gene. Molecular genetics and genomics
550 290:1727-1740.
551 38. Richards VP, Palmer SR, Pavinski Bitar PD, Qin X, Weinstock GM, Highlander SK, Town
552 CD, Burne RA, Stanhope MJ. 2014. Phylogenomics and the dynamic genome evolution of the
553 genus Streptococcus. Genome biology and evolution 6:741-753.
554 39. Liu G, Zhang W, Lu C. 2013. Comparative genomics analysis of Streptococcus agalactiae
555 reveals that isolates from cultured tilapia in China are closely related to the human strain
556 A909. BMC genomics 14:775.
557 40. He E-M, Chen C-W, Guo Y, Hsu M-H, Zhang L, Chen H-L, Zhao G-P, Chiu C-H, Zhou Y.
558 2017. The genome of serotype VI Streptococcus agalactiae serotype VI and comparative
559 analysis. Gene 597:59-65.
560 41. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni
561 C, Iacobini ET, Rosini R. 2005. Identification of a universal Group B streptococcus vaccine
562 by multiple genome screen. Science 309:148-150.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

563 42. Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and in situ detection
564 of individual microbial cells without cultivation. Microbiol Rev 59:143-69.
565 43. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.
566 44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
567 Nikolenko SI, Pham S, Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and
568 its applications to single-cell sequencing. Journal of computational biology 19:455-477.
569 45. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. 2009. Reordering contigs
570 of draft genomes using the Mauve aligner. Bioinformatics 25:2071-3.
571 46. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068-9.
572 47. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, Holt KE. 2016.
573 Identification of Klebsiella capsule synthesis loci from whole genome data. Microb Genom
574 2:e000102.
575 48. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L,
576 Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus Sequence Typing
577 of Total-Genome-Sequenced Bacteria. Journal of Clinical Microbiology 50:1355-1361.
578 49. Treangen TJ, Ondov BD, Koren S, Phillippy AM. 2014. The Harvest suite for rapid
579 core-genome alignment and visualization of thousands of intraspecific microbial genomes.
580 Genome Biol 15:524.
581 50. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR.
582 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome
583 sequences using Gubbins. Nucleic Acids Res 43:e15.
584 51. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
585 large phylogenies. Bioinformatics 30:1312-3.
586 52. He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S. 2016. Evolview v2: an online
587 visualization and management tool for customized and annotated phylogenetic trees. Nucleic
588 acids research 44:W236-W241.
589 53. Salipante SJ, Hall BG. 2011. Inadequacies of minimum spanning trees in molecular
590 epidemiology. J Clin Microbiol 49:3568-75.
591 54. Petkau A, Stuart-Edwards M, Stothard P, Van Domselaar G. 2010. Interactive microbial
592 genome visualization with GView. Bioinformatics 26:3125-6.
593 55. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. 2011. BLAST Ring Image Generator
594 (BRIG): simple prokaryote genome comparisons. BMC Genomics 12:402.
595 56. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D,
596 Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis.
597 Bioinformatics 31:3691-3.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

598 57. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. 2017. Piggy: A Rapid, Large-Scale
599 Pan-Genome Analysis Tool for Intergenic Regions in Bacteria. bioRxiv doi:10.1101/179515.
600 58. Stanton-Cook M, Ben Zakour N, Alikhan N, Beatson S. 2013. SeqFindR.
601 59. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for
602 bacterial virulence factors. Nucleic Acids Res 33:D325-8.
603 60. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.
604 2012. A program for annotating and predicting the effects of single nucleotide
605 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;
606 iso-2; iso-3. Fly (Austin) 6:80-92.
607 61. Bowater RO, Dennis MM, Blyde D, Stone B, Barnes AC, Delamare-Deboutteville J, Horton
608 MA, White M, Condon K, Jones R. 2017. Epizootics of Streptococcus agalactiae infection in
609 captive rays from Queensland, Australia. J Fish Dis doi:10.1111/jfd.12701.
610 62. Evans JJ, Klesius PH, Pasnik DJ, Bohnsack JF. 2009. Human Streptococcus agalactiae Isolate
611 in Nile Tilapia (Oreochromis niloticus). Emerging Infectious Diseases 15:774-776.
612 63. Bishop EJ, Shilton C, Benedict S, Kong F, Gilbert GL, Gal D, Godoy D, Spratt BG, Currie BJ.
613 2007. Necrotizing fasciitis in captive juvenile Crocodylus porosus caused by Streptococcus
614 agalactiae: an outbreak and review of the animal and human literature. Epidemiol Infect
615 135:1248-55.
616 64. Godoy DT, Carvalho-Castro GA, Leal CAG, Pereira UP, Leite RC, Figueiredo HCP. 2013.
617 Genetic diversity and new genotyping scheme for fish pathogenic Streptococcus agalactiae.
618 Letters in Applied Microbiology 57:476-483.
619 65. Herbert MA, Beveridge CJ, McCormick D, Aten E, Jones N, Snyder LA, Saunders NJ. 2005.
620 Genetic islands of Streptococcus agalactiae strains NEM316 and 2603VR and their presence
621 in other Group B streptococcal strains. BMC Microbiol 5:31.
622 66. Fischer A, Liljander A, Kaspar H, Muriuki C, Fuxelius HH, Bongcam-Rudloff E, de Villiers
623 EP, Huber CA, Frey J, Daubenberger C, Bishop R, Younan M, Jores J. 2013. Camel
624 Streptococcus agalactiae populations are associated with specific disease complexes and
625 acquired the tetracycline resistance gene tetM via a Tn916-like element. Vet Res 44:86.
626 67. Flores AR, Galloway-Pena J, Sahasrabhojane P, Saldana M, Yao H, Su X, Ajami NJ, Holder
627 ME, Petrosino JF, Thompson E, Margarit YRI, Rosini R, Grandi G, Horstmann N, Teatero S,
628 McGeer A, Fittipaldi N, Rappuoli R, Baker CJ, Shelburne SA. 2015. Sequence type 1 group B
629 Streptococcus, an emerging cause of invasive disease in adults, evolves by small genetic
630 changes. Proc Natl Acad Sci U S A 112:6431-6.
631 68. Almeida A, Alves-Barroco C, Sauvage E, Bexiga R, Albuquerque P, Tavares F,
632 Santos-Sanches I, Glaser P. 2016. Persistence of a dominant bovine lineage of group B

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

633 Streptococcus reveals genomic signatures of host adaptation. Environ Microbiol
634 18:4216-4229.
635 69. Vandamme P, Devriese LA, Pot B, Kersters K, Melin P. 1997. Streptococcus difficile is a
636 nonhemolytic group B, type Ib Streptococcus. International Journal of Systematic
637 Bacteriology 47:81-85.
638 70. Eldar A, Bejerano Y, Bercovier H. 1994. Streptococcus shiloi andStreptococcus difficile:
639 Two new streptococcal species causing a meningoencephalitis in fish. Current Microbiology
640 28:139-143.
641 71. Mather PB, Arthington AH. 1991. An assessment of genetic differentiation among feral
642 Australian tilapia populations. Australian Journal of Marine and Freshwater Research
643 42:721-728.
644 72. Verner-Jeffreys DW, Wallis TJ, Cejas I, Ryder D, Haydon DJ, Domazoro JF, Dontwi J, Field
645 TR, Adjei-Boteng D, Wood G, Bean T, Feist SW. 2018. Streptococcus agalactiae Multilocus
646 sequence type 261 is associated with mortalities in the emerging Ghanaian tilapia industry.
647 Journal of Fish Diseases 41:175-179.
648 73. Spellerberg B, Martin S, Brandt C, Lutticken R. 2000. The cyl genes of Streptococcus
649 agalactiae are involved in the production of pigment. FEMS Microbiol Lett 188:125-8.
650 74. Nagano N, Nagano Y, Taguchi F. 2002. High expression of a C protein beta antigen gene
651 among invasive strains from certain clonally related groups of type Ia and Ib group B
652 streptococci. Infect Immun 70:4643-9.
653 75. Yang Q, Zhang M, Harrington DJ, Black GW, Sutcliffe IC. 2011. A proteomic investigation
654 of Streptococcus agalactiae reveals that human serum induces the C protein beta antigen and
655 arginine deiminase. Microbes Infect 13:757-60.
656 76. Yang Q, Zhang M, Harrington DJ, Black GW, Sutcliffe IC. 2010. A proteomic investigation
657 of Streptococcus agalactiae grown under conditions associated with neonatal exposure
658 reveals the upregulation of the putative virulence factor C protein beta antigen. Int J Med
659 Microbiol 300:331-7.
660 77. Locke JB, Colvin KM, Datta AK, Patel SK, Naidu NN, Neely MN, Nizet V, Buchanan JT.
661 2007. Streptococcus iniae capsule impairs phagocytic clearance and contributes to virulence
662 in fish. J Bacteriol 189:1279-87.
663 78. Delannoy CMJ, Zadoks RN, Crumlish M, Rodgers D, Lainson FA, Ferguson HW, Turnbull J,
664 Fontaine MC. 2016. Genomic comparison of virulent and non-virulent Streptococcus
665 agalactiae in fish. Journal of Fish Diseases 39:13-29.
666 79. Yao KH, Poulsen K, Maione D, Rinaudo CD, Baldassarri L, Telford JL, Sorensen UBS,
667 Kilian M, Grp DS. 2013. Capsular Gene Typing of Streptococcus agalactiae Compared to
668 Serotyping by Latex Agglutination. Journal of Clinical Microbiology 51:503-507.

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

669 80. Pasnik DJ, Evans JJ, Panangala VS, Klesius PH, Shelby RA, Shoemaker CA. 2005.
670 Antigenicity of Streptococcus agalactiae extracellular products and vaccine efficacy. Journal
671 of Fish Diseases 28:205-212.
672 81. Seib KL, Zhao X, Rappuoli R. 2012. Developing vaccines in the era of genomics: a decade of
673 reverse vaccinology. Clin Microbiol Infect 18:1-8.
674 82. Millard CM, Baiano JC, Chan C, Yuen B, Aviles F, Landos M, Chong RS, Benedict S, Barnes
675 AC. 2012. Evolution of the capsular operon of Streptococcus iniae in response to vaccination.
676 Appl Environ Microbiol 78:8219-26.
677 83. Yamamoto S, Miyake K, Koike Y, Watanabe M, Machida Y, Ohta M, Iijima S. 1999.
678 Molecular characterization of type-specific capsular polysaccharide biosynthesis genes of
679 Streptococcus agalactiae type Ia. J Bacteriol 181:5176-84.
680 84. Watanabe M, Miyake K, Yanae K, Kataoka Y, Koizumi S, Endo T, Ozaki A, Iijima S. 2002.
681 Molecular characterization of a novel beta1,3-galactosyltransferase for capsular
682 polysaccharide synthesis by Streptococcus agalactiae type Ib. J Biochem 131:183-91.
683 85. Martins ER, Melo-Cristino J, Ramirez M. 2007. Reevaluating the serotype II capsular locus
684 of Streptococcus agalactiae. J Clin Microbiol 45:3384-6.
685 86. Chaffin DO, Beres SB, Yim HH, Rubens CE. 2000. The serotype of type Ia and III group B
686 streptococci is determined by the polymerase gene within the polycistronic capsule operon. J
687 Bacteriol 182:4466-77.
688 87. Chaffin DO, McKinnon K, Rubens CE. 2002. CpsK of Streptococcus agalactiae exhibits
689 alpha2,3-sialyltransferase activity in Haemophilus ducreyi. Mol Microbiol 45:109-22.
690 88. Cieslewicz MJ, Chaffin D, Glusman G, Kasper D, Madan A, Rodrigues S, Fahey J, Wessels
691 MR, Rubens CE. 2005. Structural and genetic diversity of group B streptococcus capsular
692 polysaccharides. Infect Immun 73:3096-103.
693 89. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer.
694 Bioinformatics 27:1009-10.
695
696

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

697 **Tables**

698 **Table 1:** *S. agalactiae* isolates and sequences used in this study

Isolate	Origin	Year	Host	Serotype	ST	Size (Mbp)	Accession no.
QMA0264	QLD, AU	2008	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998566
QMA0266	QLD, AU	2008	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998567
QMA0267	QLD, AU	2008	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998568
QMA0268	QLD, AU	2009	*Pomadasys kaaken*	Ib	261	1.8	SAMN07998569
QMA0271	QLD, AU	2009	*Arius thalassinus*	Ib	261	1.8	SAMN07998570
QMA0273	QLD, AU	2009	*Arius thalassinus*	Ib	261	1.8	SAMN07998571
QMA0274	QLD, AU	2009	*Liza vaigensis*	Ib	261	1.8	SAMN07998572
QMA0275	QLD, AU	2009	*Aptychotrema rostrata*	Ib	261	1.8	SAMN07998573
QMA0276	QLD, AU	2009	*Himantura granulata*	Ib	261	1.8	SAMN07998574
QMA0277	QLD, AU	2009	*Dasyatis fluviorum*	Ib	261	1.8	SAMN07998575
QMA0280	QLD, AU	2010	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998576
QMA0281	QLD, AU	2010	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998577
QMA0284	QLD, AU	2010	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998578
QMA0285	QLD, AU	2010	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998579
QMA0287	QLD, AU	2010	*Pomadasys kaaken*	Ib	261	1.8	SAMN07998580
QMA0290	QLD, AU	2010	*Arius thalassinus*	Ib	261	1.8	SAMN07998581
QMA0292	QLD, AU	2010	*Aptychotrema rostrata*	Ib	261	1.8	SAMN07998582
QMA0294	QLD, AU	2010	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998583
QMA0320	QLD, AU	2010	*Dasyatis fluviorum*	Ib	261	1.8	SAMN07998584
QMA0321	QLD, AU	2010	*Dasyatis fluviorum*	Ib	261	1.8	SAMN07998585
QMA0323	QLD, AU	2010	*Dasyatis fluviorum*	Ib	261	1.8	SAMN07998586
QMA0326	QLD, AU	2010	*Dasyatis fluviorum*	Ib	261	1.8	SAMN07998587
QMA0347	QLD, AU	2010	*Dasyatis fluviorum*	Ib	261	1.8	SAMN07998588
QMA0368	QLD, AU	2010	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998589
QMA0369	QLD, AU	2011	*Epinephelus lanceolatus*	Ib	261	1.8	SAMN07998590
QMA0485	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998597
QMA0487	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998598
QMA0488	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998599
QMA0489	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998600
QMA0494	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998601
QMA0495	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998602

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

QMA0496	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998603
QMA0497	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998604
QMA0499	Honduras	2014	*Oreochromis niloticus*	Ib	260	1.8	SAMN07998605
QMA0355	QLD, AU	2011	*Homo sapiens*	Ia	23	2.0	SAMN07998591
QMA0357	QLD, AU	2011	*Homo sapiens*	Ia	23	2.0	SAMN07998592
QMA0336	NT, AU	2005	*Crocodylus porosus*	Ia	23	2.0	SAMN07998593
QMA0300	QLD, AU	2008	*Canis lapis familiaris*	V	1	2.1	SAMN07998594
QMA0303	QLD, AU	2009	*Felis catus*	V	1	2.1	SAMN07998595
QMA0306	QLD, AU	2005	*Bos taurus*	V	1	2.2	SAMN07998596
GS16-0008	Ghana	2016	*Oreochromis niloticus*	Ib	261	1.8	SRX2698682
GS16-0031	Ghana	2016	*Oreochromis niloticus*	Ib	261	1.8	SRX2698681
GS16-0035	Ghana	2016	*Oreochromis niloticus*	Ib	261	1.8	SRX2698680
GS16-0046	Ghana	2016	*Oreochromis niloticus*	Ib	261	1.8	SRX2698679
ND2-22	Israel	1988	*Oreochromis niloticus*	Ib	261	1.8	FO393392
138P	USA	2007	*Oreochromis niloticus*	Ib	261	1.8	CP007482
138spar	USA	2011	*Oreochromis niloticus*	Ib	261	1.8	CP007565.1
GX026	China	2011	*Oreochromis niloticus*	Ib	261	1.8	CP011328.1
S13	Brazil	2015	*Oreochromis niloticus*	Ib	552	1.8	CP018623.1
S25	Brazil	2015	*Oreochromis niloticus*	Ib	552	1.8	CP015976.1
SA20	Brazil		*Oreochromis niloticus*	Ib	552***	1.8	CP003919.2
GD201008-001	China	2010	*Oreochromis niloticus*	Ia	7	2.1	NC_018646.1
HN016	China	2011	*Oreochromis niloticus*	Ia	7	2.1	NZ_CP011325.1
WC1535	China	2015	*Oreochromis niloticus*	Ia	7	2.2	NZ_CP016501.1
A909	USA		*Homo sapiens*	Ia	7	2.1	NC_007432.1
GBS85147	Brazil		*Homo sapiens*	Ia	103	2.0	NZ_CP010319.1
Sag37	China	2014	*Homo sapiens*	Ib*	12	2.2	NZ_CP019978.1
GBS1-NY	USA	2012	*Homo sapiens*	II	22	2.2	NZ_CP007570.1
GBS2_NM	USA	2012	*Homo sapiens*	II	22	2.2	NZ_CP007571.1
GBS6	USA	2009	*Homo sapiens*	II	22	2.2	NZ_CP007572.1
FDAARGOS_254	USA	2014	*Homo sapiens*	II*	22	2.2	CP020449.1
COH1	USA		*Homo sapiens*	III	17	2.1	NZ_HG939456.1
NEM316	France		*Homo sapiens*	III	23	2.2	NC_004368.1
CU_GBS_08	Hong Kong	2008	*Homo sapiens*	III	283	2.1	NZ_CP010874.1
CU_GBS_98	Hong Kong	1998	*Homo sapiens*	III	283	2.0	NZ_CP010875.1
NGBS128	Canada	2010	*Homo sapiens*	III	17	2.1	NZ_CP012480.1
SG-M1	Singapore	2015	*Homo sapiens*	III	283	2.1	NZ_CP012419.2

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

H002	China	2011	*Homo sapiens*	III	736	2.1	NZ_CP011329.1
Sag158	China	2014	*Homo sapiens*	III*	19	2.1	NZ_CP019979.1
NGBS061	Canada	2010	*Homo sapiens*	IV	459	2.2	NZ_CP007631.1
NGBS572	Canada	2012	*Homo sapiens*	IV	452	2.1	NZ_CP007632.1
2603V/R	USA		*Homo sapiens*	V	110	2.2	NC_004116.1
CNCTC10/84	USA		*Homo sapiens*	V	26	2.0	NZ_CP006910.1
NGBS357	Canada	2011	*Homo sapiens*	V	1	2.2	NZ_CP012503.1
SS1	USA	1992	*Homo sapiens*	V	1	2.1	NZ_CP010867.1
GBS-M002	China	2014	*Homo sapiens*	VI*	1	2.1	NZ_CP013908.1
SA111	Portugal	2013	*Homo sapiens*	II*	61**	2.3	NZ_LT545678.1
FWL1402	China	2014	*Hoplobatrachus chinensis*	III*	739**	2.1	NZ_CP016391.1
09mas018883	Sweden		*Bos taurus*	V*	1	2.1	NC_021485.1
GBS ST-1	USA	2015	*Canis lapis familiaris*	V	1	2.2	NZ_CP013202.1
ILRI005	Kenya		*Camelus dromedarius*	V*	609	2.1	NC_021486.1
ILRI112	Kenya		*Camelus dromedarius*	VI*	617	2.0	HF952106.1

699 AU, Australia; QLD, Queensland; NT, Northern Territory; ST, Sequence Type.

700 *Serotype was detected with Kaptive. ** ST was determined via Center for Genomic Epidemiology.

701 *** gap in *glcK*

702

703 **Table 2:** Reference sequences used for molecular capsule serotyping.

Serotype	Accession Number	size (bp)	Reference
Ia	AB028896.2	25021	(83)
Ib	AB050723.1	9987	(84)
II	EF990365.1	12864	(85)
III	AF163833.1	17276	(86)
IV	AF355776.1	17596	(87)
V	AF349539.1	18239	(87)
VI	AF337958.1	16448	(87)
VII	AY376403.1	14202	(88)
VIII	AY375363.1	12637	(88)

704

705

bioRxiv preprint doi: https://doi.org/10.1101/301424; this version posted April 14, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

706 **Figures**

707 **Figure 1. A)** Maximum likelihood phylogeny of 82 *S. agalactiae* strains. The tree was inferred from
708 alignment of 6050 non-recombinant core genome SNPs. Branch length was adjusted for
709 ascertainment bias using Felsenstein's correction implemented in RAxML (51). Nodes are supported
710 by 1000 bootstrap replicates. The tree was rooted using *S. pyogenes* M1 GAS (Accession number:
711 NC_002737.2) as an outgroup. B) Minimum spanning tree showing relationship amongst ST-261
712 serotype 1b GBS isolates based on a distance matrix derived from non-recombinant core-genome
713 SNP alignment. The consensus network was computed in MST-Gold (53).

714 **Figure 2.** The pan-genome of *S. agalactiae*. A) BLASTN-based sequence comparison of 82 *S.*
715 *agalactiae* genomes against the *S. agalactiae* pan-genome as reference constructed with BRIG 0.95
716 (55). Rings from the innermost to the outermost shows GC content and GC skew of the pan-genome
717 reference, then sequence similarity of each of the 82 strains listed in the legend, from top to bottom
718 rings are coloured according to origin with fish isolates belonging to ST-261, ST-260 and ST-552
719 serotype Ib in blue, fish strains belonging to serotype Ia in purple and terrestrial strains in red. The
720 outermost ring (black) represents reference pan-genome. B) Proportion of protein-coding genes in the
721 core, soft core, shell and cloud of the pan-genome of 82 *S. agalactiae* isolates determined with Roary.
722 C) Histogram indicates the frequency of genes (protein-coding) in red and IGRs (non-protein-coding
723 intergenic regions) in blue-green from 82 *S. agalactiae* genomes analysed by Piggy (57).

724 **Figure 3.** Virulence gene presence and absence in *S. agalactiae*. Genes were identified from VFDB
725 to create a *S. agalactiae* database for comparison of 82 strains by BLAST using SeqFindr with default
726 settings (95% identity cut off).

727 **Figure 4.** The capsular polysaccharide (cps) operon of the ST-261 lineage. The operon was identified
728 in Genbank files manually and then compared by BLAST using EasyFig (89)

A

0.0001

BOOTSTRAP
< = 50
51~ 90
91~ 100

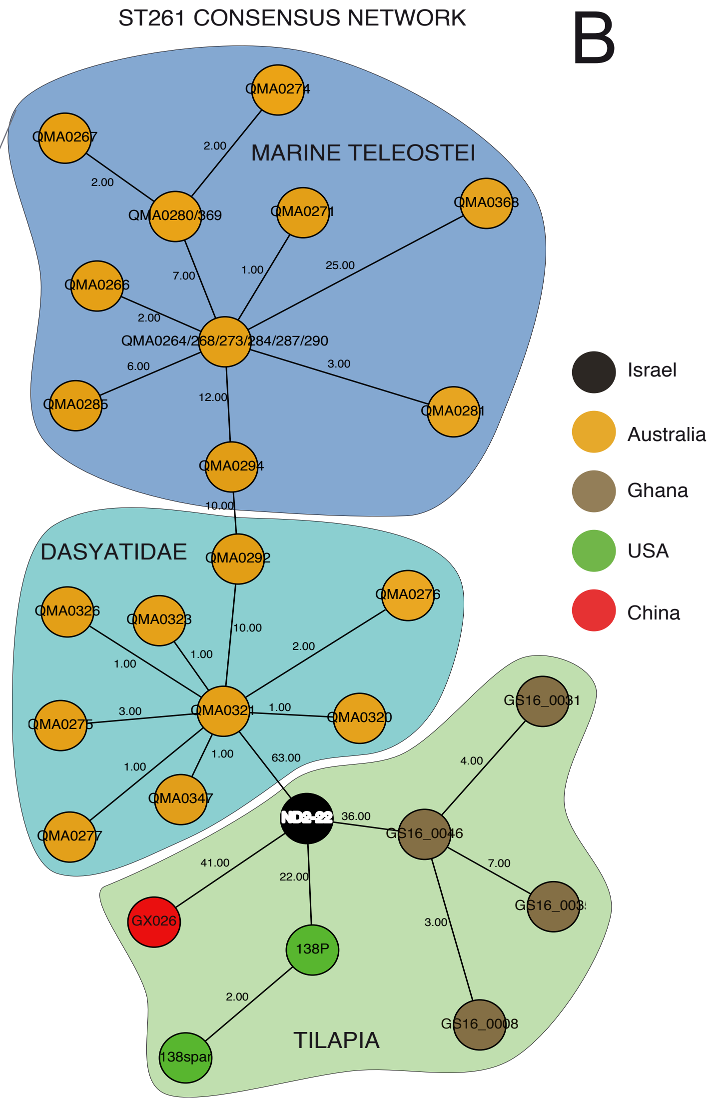
HOST
MAMMALS
Human
Canine
Feline
Bovine
Camel
REPTILES/AMPHIBIANS
Crocodile
Frog
TELEOST
Tilapia
Mullet
Catfish
Grouper
Javelin grunter
Trout
ELASMOBRANCH
Eastern shovel nose ray
Estuary ray
Mangrove whiptail ray

REGION
AMERICAS
Canada
USA
Honduras
Brazil
ASIA PACIFIC
Australia
China
Hong Kong
Singapore
EUROPE
Sweden
France
Portugal
Israel
AFRICA
Ghana
Kenya

YEAR
1988
1992
1998
2005
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

SEROTYPE ST HOST REGION YEAR

CNCTC1084 V 26
NGBS128 III 17
COH1
GD201008001 Ia 7
HN016
WC1535
A909
Sag37 Ib 12
CU GBS 98 III 283
FWL1402
CU GBS 08
SG M1 283
2603VR V 110
Sag158 III 19
H002 736
GBS M002 VI 1
NGBS061 IV 495
QMA0303 V 1
09mas018883
SS1
QMA0300
GBS ST1
NGBS357
NEM316 III 23
NGBS572 IV 452
QMA0336 Ia 23
QMA0357
QMA0355
SA111 II 61
QMA0306 V 1
GBS6 II 22
GBS2 NM
GBS1 NY
FDAARGOS 254
ILRI005 V 609
GBS85147 Ia 103
ILRI112 V 617
S25 552
SA20
S13
QMA0487 260
QMA0489
QMA0485
QMA0488
QMA0497
QMA0495
QMA0499
QMA0496
QMA0494
138P Ib 261
138spar
GX026
ND2-22
GS16 0008
GS16 0035
GS16 0031
GS16 0046
QMA0292
QMA0294
QMA0320
QMA0277
QMA0276
QMA0275
QMA0281
QMA0287
QMA0264
QMA0268
QMA0280
QMA0274
QMA0267
QMA0290
QMA0271
QMA0285
S.pyogenes

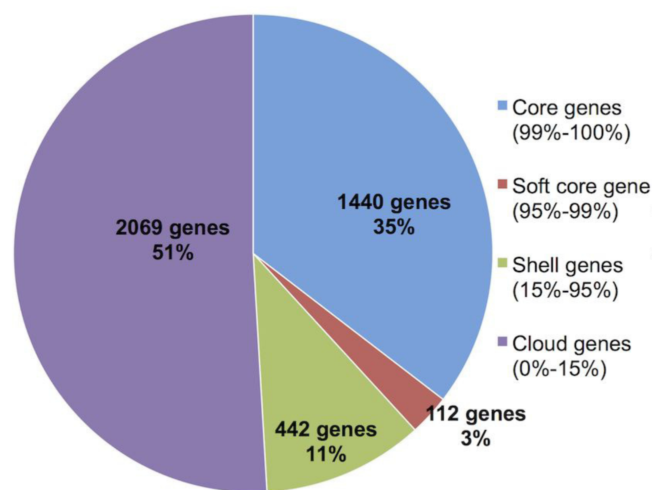


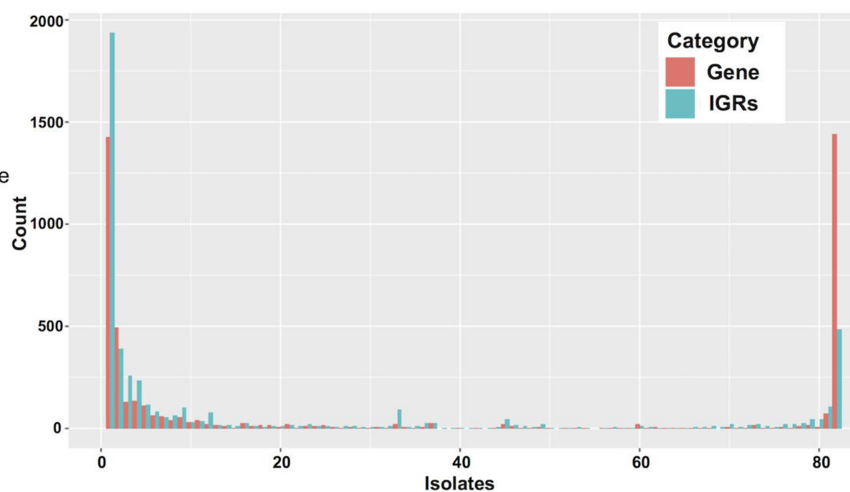
A

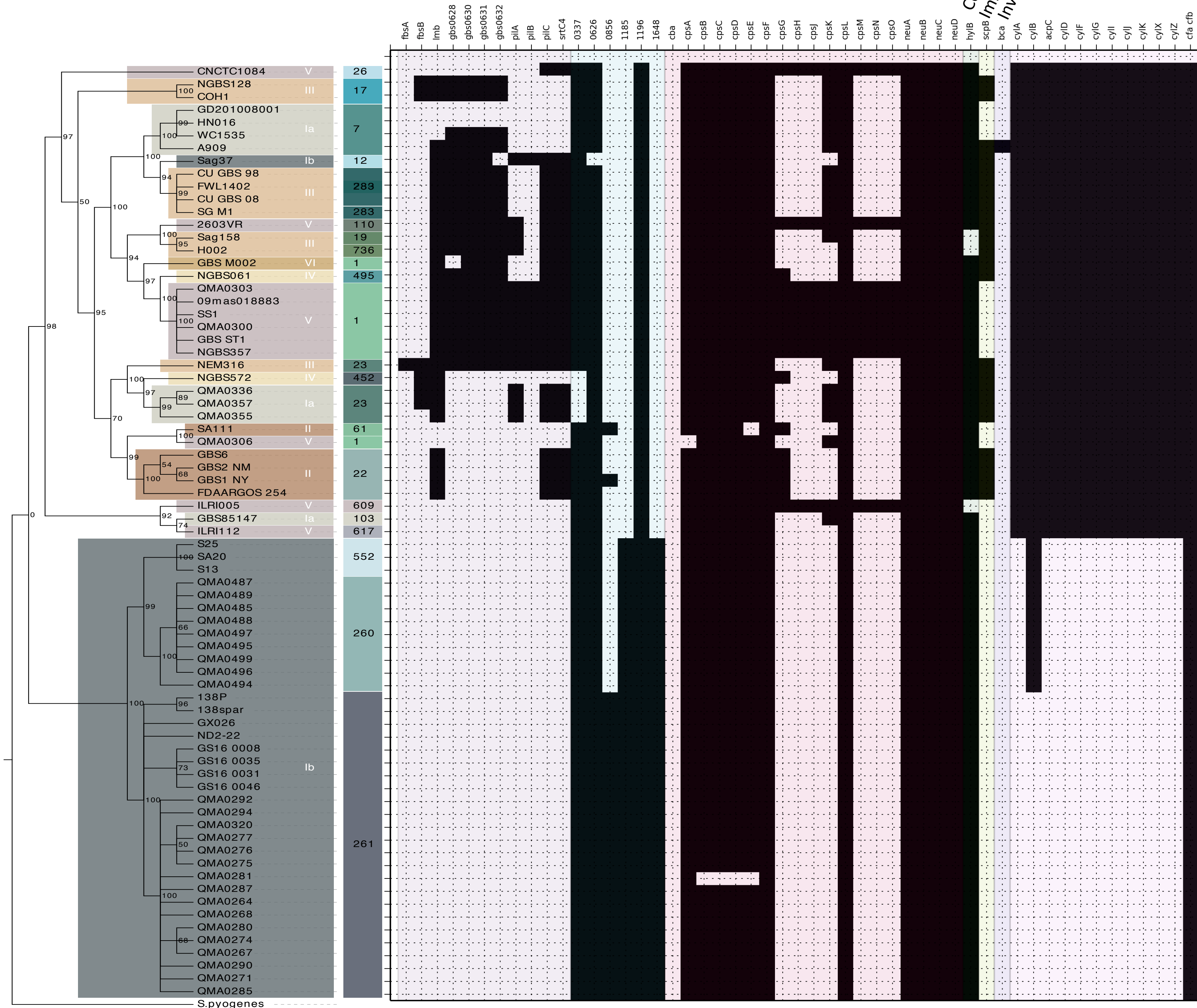
S. agalactiae
Pan-genome
4,074,275 bp

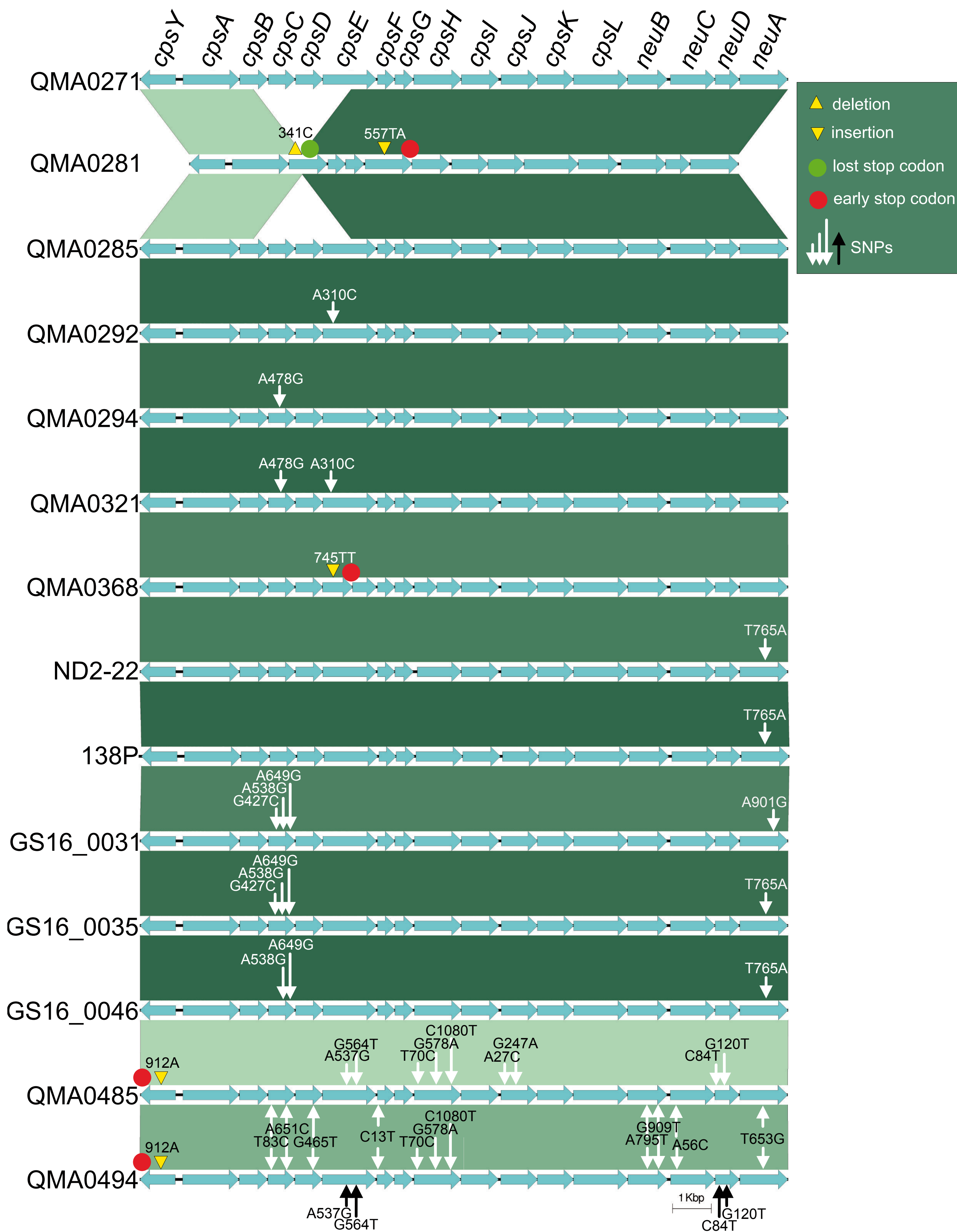
500 kbp
1000 kbp
1500 kbp
2000 kbp
2500 kbp
3000 kbp
3500 kbp
4000 kbp

GC Content
GC Skew
GC Skew(-)
GC Skew(+)
QMA0264
QMA0266
QMA0267
QMA0268
QMA0271
QMA0273
QMA0274
QMA0275
QMA0276
QMA0277
QMA0280
QMA0281
QMA0284
QMA0285
QMA0287
QMA0290
QMA0292
QMA0294
QMA0320
QMA0321
QMA0323
QMA0326
QMA0347
QMA0368
QMA0369
QMA0485
QMA0487
QMA0488
QMA0489
QMA0494
QMA0495
QMA0496
QMA0497
QMA0499
GS16-0008
GS16-0031
GS16-0035
GS16-0046
GX026
2-22
138P
138spar
S13
S25
SA20
GD201008-001
HN016
WC1535
A909
QMA0355
QMA0357
QMA0336
GBS85147
Sag37
GBS1-NY
GBS2-NM
GBS6
FDAARGOS_254
SA111
COH1
NGBS128
Sag158
NEM316
CU_GBS_08
CU_GBS_98
SG-M1
H002
FWL1402
NGBS061
NGBS572
NGBS357
SS1
09mas018883
GBS ST-1
QMA0300
QMA0303
QMA0306
CNCTC10/84
2603V/R
ILRI005
GBS-M002
ILRI112
Pan-genome






Adhesins
ST-261 Adhesins
Antiphagocytosis
Coenzyme
Immune evasion
Invasion
Toxin
fbsA
fbsB
lmb
gbs0628
gbs0630
gbs0631
gbs0632
pilA
pilB
pilC
srtC4
0337
0626
0856
1185
1196
1648
cba
cpsA
cpsB
cpsC
cpsD
cpsE
cpsF
cpsG
cpsH
cpsJ
cpsK
cpsL
cpsM
cpsN
cpsO
neuA
neuB
neuC
neuD
hylB
scpB
bca
cylA
cylB
acpC
cylD
cylF
cylG
cylI
cylJ
cylK
cylX
cylZ
cfa cfb
CNCTC1084
NGBS128
COH1
GD201008001
HN016
WC1535
A909
Sag37
CU GBS 98
FWL1402
CU GBS 08
SG M1
2603VR
Sag158
H002
GBS M002
NGBS061
QMA0303
09mas018883
SS1
QMA0300
GBS ST1
NGBS357
NEM316
NGBS572
QMA0336
QMA0357
QMA0355
SA111
QMA0306
GBS6
GBS2 NM
GBS1 NY
FDAARGOS 254
ILRI005
GBS85147
ILRI112
S25
SA20
S13
QMA0487
QMA0489
QMA0485
QMA0488
QMA0497
QMA0495
QMA0499
QMA0496
QMA0494
138P
138spar
GX026
ND2-22
GS16 0008
GS16 0035
GS16 0031
GS16 0046
QMA0292
QMA0294
QMA0320
QMA0277
QMA0276
QMA0275
QMA0281
QMA0287
QMA0264
QMA0268
QMA0280
QMA0274
QMA0267
QMA0290
QMA0271
QMA0285
S.pyogenes


cpsY
cpsA
cpsB
cpsC
cpsD
cpsE
cpsF
cpsG
cpsH
cpsI
cpsJ
cpsK
cpsL
neuB
neuC
neuD
neuA
QMA0271
QMA0281
QMA0285
QMA0292
QMA0294
QMA0321
QMA0368
ND2-22
138P
GS16_0031
GS16_0035
GS16_0046
QMA0485
QMA0494
deletion
insertion
lost stop codon
early stop codon
SNPs
341C
557TA
A310C
A478G
A478G A310C
745TT
T765A
T765A
A649G
A538G
G427C
A901G
A649G
A538G
G427C
T765A
A649G
A538G
T765A
912A
G564T
A537G
C1080T
G578A
T70C
G247A
A27C
G120T
C84T
912A
A651C
T83C G465T
C13T
C1080T
G578A
T70C
G909T
A795T
A56C
T653G
A537G
G564T
1 Kbp
G120T
C84T