

1 Complete genome sequence and comparative genomics of the fish 2 pathogen *Streptococcus iniae* QMA0248

3
4 Areej S. Alsheikh-Hussain^{1,2,3}, Nouri L. Ben Zakour^{1,2,3,5}, Brian M. Forde^{1,2,3},
5 Oleksandra Silayeva⁴, Andrew C. Barnes⁴, and Scott A. Beatson^{*1,2,3}.

6
7 ¹School of Chemistry & Molecular Biosciences, ²Australian Infectious Diseases
8 Research Centre, ³Australian Centre for Ecogenomics, and ⁴School of Biological
9 Science, The University of Queensland, Brisbane, Queensland, Australia;
10 Present address: ⁵The Westmead Institute for Medical Research and the University
11 of Sydney, Sydney, New South Wales, Australia

12
13
14 Key words: SMRT sequencing, Insertion Sequence, reference guided assembly,
15 misassembly, mobile genetic elements

16
17
18 * Corresponding author.

19 Mailing address: School of Chemistry and Molecular Biosciences, Building 76, The
20 University of Queensland, Brisbane QLD 4072, Australia.

21 Phone: +617 33654863; E-mail: s.beatson@uq.edu.au

22

23

Abstract

Fish mortality caused by *Streptococcus iniae* is a major economic problem in fish aquaculture in warm and temperate regions globally. There is also risk of zoonotic infection by *S. iniae* through handling of contaminated fish. In this study, we present the complete genome sequence of *S. iniae* strain QMA0248, isolated from farmed barramundi in South Australia. The 2.12 Mb genome of *S. iniae* QMA0248 carries a 32 Kb prophage, a 12 Kb genomic island, and 92 discrete insertion sequence (IS) elements. These include 9 novel IS types that belong mostly to the IS3 family. Comparative and phylogenetic analysis between *S. iniae* QMA0248 and publicly available complete *S. iniae* genomes revealed discrepancies that are likely due to misassembly in the genomes of isolates ISET0901 and ISNO. We also determined by long-range PCR that a tandem duplication of an rRNA region in the PacBio assembly of QMA0248 was an assembly error. We suggest that a similar rRNA duplication in the PacBio genome of *S. iniae* 89353 may also be a misassembly. Our study not only highlights assembly problems in existing genomes, but provides a high quality reference genome for *S. iniae* QMA0248, including manually curated mobile genetic elements, that will assist future *S. iniae* comparative genomic and evolutionary studies.

Introduction

Streptococcus iniae is a fish pathogen that causes mortality in a wide range of fish species in wild and farmed, marine and freshwater environments, resulting in large economic losses to aquaculture (1, 2). *S. iniae* is also considered an opportunistic human pathogen, causing sporadic infections mostly in the elderly who have more than one underlying health condition such as diabetes mellitus, or

chronic rheumatic heart disease (3, 4). *S. iniae* pathogenesis is imparted through a repertoire of virulence factors (VF) including surface proteins, secreted toxins, and capsular polysaccharide (CPS) (4). VFs can be acquired through lateral gene transfer (LGT) of mobile genetic elements (MGE) such as composite transposons, genomic islands (GI) or prophages.

MGEs are a means by which bacterial pathogens acquire traits that help adapt to changing conditions including vaccination, antibiotics, a new host or environment (5, 6). Indeed, they are considered the main drivers of gene flux in bacteria, contributing to diversity within species (7). Insertion sequence (IS) elements, for instance, are small MGEs (0.7-3.5 Kb) that have an important role in evolution and genome plasticity. IS insertion within bacterial chromosomes or plasmids can result in genetic modifications through insertional inactivation of genes or up-regulation of adjacent intact genes through outward-facing promoter sequences carried by some IS (8, 9). In some cases pairs of IS can mobilize intervening sequence as a composite transposon (8). The mobility of IS elements leads to their expansion or loss within bacterial lineages. Expansion is associated with accumulation of pseudogenes, which is considered an early stage in genome reduction as a mechanism for adaptation (9). Accordingly, to obtain the complete evolutionary picture within bacterial species it is important to study the distribution and abundance of IS elements.

As yet there is no study that focuses on the diversity and distribution of MGEs in *S. iniae* genomes. In fact, only four complete *S. iniae* genomes were deposited in Genbank at the commencement of the present study, none of which had

comprehensive annotations for MGEs: SF1, YSFST01-82, ISET0901, and ISNO. *S. iniae* SF1 (accession: CP005941) was cultured from moribund flounder from a fish farm that experienced an epidemic in North China (10). YSFST01-82 (accession: CP010783) is an isolate from a diseased Olive flounder from South Korea (2012) (11). ISET0901 (accession: CP007586) is reported to be a highly virulent strain, which was isolated from diseased Nile tilapia during an outbreak in Israel (2005) (12). The strain ISNO (accession: CP007587) was obtained through selection for resistance of *S. iniae* ISET0901 to novobiocin (13).

There are currently 25 different IS families in the ISFinder database, but only one IS described for *S. iniae* (IS*Stin1* of the family IS256). As is typical for complete bacterial genomes, the annotation of the four complete *S. iniae* genomes available at Genbank are limited to some of the transposase genes associated with IS elements without definition of the IS boundaries. Indeed, the difficulty in annotating IS elements and the lack of reliable automated annotation methods means that only a small subset of complete genomes have accurate IS annotations. Small partial IS elements are often disregarded although they reflect valuable insights on ancestral recombination events. With the increasing availability of long-read sequencing there is a need for a high quality, well-characterized *S. iniae* reference genome that better enables the impact of IS elements on evolution and diversity to be determined.

In this study, we have completely characterized the genome of *S. iniae* QMA0248. Manual curation of annotations for IS, genomic islands, prophages, and CRISPR was carried out along with a comparison with the four publicly available complete

genomes from NCBI (SF1, YSFST01-82, ISET0901, and ISNO). Comparative and phylogenetic analyses revealed discrepancies between the MGE content indicating likely misassembly in the genomes ISNO, and ISET0901. The complete genome of QMA0248 will provide an important scaffold for future phylogenomic studies of the *S. iniae* species.

Results and Discussion

Genomic features of *S. iniae* QMA0248

The genome of *S. iniae* QMA0248 consists of a single circular chromosome of 2,116,570 bp with no plasmids (Figure 1 and Table 1). The QMA0248 chromosome has an average GC content of 36.8 %, consistent with the other four *S. iniae* genomes (SF1, YSFST01-82, ISET0901, and ISNO) and in common with several other *Streptococcus* spp. such as *S. agalactiae* (35.6%) (14), *S. pneumoniae* (39.7%) (15), and *S. pyogenes* (38.5%) (16). Of note, there is a high degree of strand bias in the genome of *S. iniae* QMA0248 where genes are preferentially oriented in the leading strand, which is typical for Firmicutes (Figure 1). Only seven percent of QMA0248 protein-coding genes are annotated as hypothetical proteins compared with a quarter in SF1 and 11-14% in the other three published genomes: YSFST01-82, ISET0901, and ISNO. There are 68 pseudogenes identified in QMA0248 (GenBank assembly accession: CP022392.1), most due to interruption by IS (Figure 1, Supplementary Table S1). This is approximately five times greater than the number of pseudogenes predicted in SF1. Collectively, these differences between the compared genomes QMA0248, SF1, YSFST01-82, ISET0901, and ISNO are likely to reflect different approaches to annotation. The

other 25 pseudogenes are caused by in-frame stop codons or frame-shifts, all supported by additional mapping of Illumina reads of the strain QMA0248 against its PacBio assembly.

QMA0248 has a single CRISPR locus (Clustered regularly interspaced short palindromic repeats) (Figure 1), which harbors a tandem array of 15 identical 36 bp repeats, separated by 14 distinct 30 bp spacers, which is about double the size of the CRISPR array in SF1, ISET0901, and ISNO (Supplementary Figure S1). Upstream of QMA0248 CRISPR are four Cas genes, *csn2*, *cas2*, *cas1*, and *cas9*, which alongside the CRISPR locus provide adaptive immunity against foreign DNA (e.g. phage and plasmids) (17).

QMA0248 harbors 58 tRNA genes and 15 rRNA loci, consisting of 5S, 16S, and 23S genes, arranged in five loci. In contrast, there is one rRNA operon fewer in SF1, ISET0901, and ISNO than in QMA0248 (Table 1). Furthermore, during the preparation of this manuscript, a PacBio complete genome was published for the strain *S. iniae* 89353 (accession: [CP017952](https://www.ncbi.nlm.nih.gov/nuccore/CP017952)) which has an identical rRNA arrangement to QMA0248 except that one rRNA locus encodes a ~7 Kb tandem duplication (i.e. six loci in total) (18). Such intra-species variation in the number of rRNA genes (and tRNA genes) is not uncommon in bacteria (including streptococci) (19, 20), and prompted us to investigate further.

Confirmation of rRNA assembly in *S. iniae* QMA0248 genome

PacBio reads for QMA0248 were initially assembled into a large ~2 Mb contig representing most of the chromosome of *S. iniae* QMA0248 and three contigs less than 10 Kb in length. The short contigs appeared to be single read chimeras that were discarded from the final assembly. However, the identification of the tandem rRNA region in *S. iniae* 89353 prompted us to review the assembled short contigs. One of these ~7 Kb discarded contigs encoded an rRNA region (5S, 16S, and 23S genes in tandem with an intervening cluster of tRNA genes). Subsequent reassembly and visualization of mapped raw reads indicated that the additional rRNA contig could be placed in three of the five rRNA operon locations to form a ~13 Kb putative tandem duplication of 5S, 16S and 23 rRNA genes as seen in the 89353 genome (18). Closer examination of the read pileup for the ~7 Kb rRNA contig revealed that the tandem duplication was not well supported by overlapping reads (Supplementary Figure S2) suggesting that it may be a chimeric assembly of reads from more than one rRNA locus. To investigate the location of the putative tandem rRNA duplication in *S. iniae* QMA0248 we carried out long-range PCR across each of the five potential rRNA loci in the chromosome. PCR revealed no tandem rRNA duplication in any locus (Supplementary Figure S3). Never-the-less, we cannot rule out that during the culturing step, prior to DNA extraction and genome sequencing, there existed a sub-population of QMA0248 cells with the tandem rRNA duplication. This would be consistent with finding only a small number of reads spanning the tandem repeat. Similar doubt exists over the tandem rRNA repeat in 89353, which should also be investigated at the read-level.

Characterization of large mobile genetic elements in *S. iniae* QMA0248

Our investigation of the rRNA discrepancies in *S. iniae* SF1, ISET0901, and ISNO also revealed differences in the mobile genetic element (MGE) content within the available complete genomes. The chromosome of QMA0248 has a single ~12 Kb genomic island (GI-Leu) inserted within the tRNA-Leu downstream of a large number of consecutive ribosomal genes (Table 2). GI-Leu encodes an integrase at its 5' end (QMA0248_0125), which is predicted to be responsible for the GI insertion. A ~2.8 Kb region at the 5' end of the GI (87881–90708) is homologous to the fish pathogen *S. parauberis* KCTC 11537, with 90% nucleotide sequence identity, including part of the integrase, plasmid replication gene, and two hypothetical proteins. The island encodes a collagen-binding surface protein (Cna, B-type domain), which is a virulence factor with an LPxTG cell wall anchor motif, a conserved Gram-positive *cocci* pentapeptide (21). An IS30 family transposon insertion truncates the collagen-binding domain encoded by *cna*, likely leading to loss of functionality in QMA0248. This adhesin has been shown to play an important pathogenic role in *Staphylococcus aureus* by facilitating bacterial cells adherence to host collagen (22). Most of this GI appears to have been deleted in the genomes of SF1, ISET0901, and ISNO only, along with the rRNA operon upstream of it (Figure 3), which explains the difference with QMA0248 in total number of rRNA and tRNA genes (Table 1).

The genome of *S. iniae* QMA0248 harbors a single ~32 Kb incomplete phage (Phi1) (1991661–2023508), including a 5' integrase gene (QMA0248_1936), inserted upstream of a tRNA-Cys gene (Table 2). A total of 44 genes encoding phage proteins were identified, including genes involved in DNA replication such as DNA polymerase III, tail morphogenesis, as well as host lysis such as holin and lysin, in

addition to 24 phage hypothetical proteins. More than half of the genes encoding phage proteins carried by QMA0248 are homologous to proteins encoded by temperate bacteriophage *Streptococcus* PH10 (56.8% according to PHAST) (23). Furthermore, Phi1 in QMA0248 exhibits a remarkable nucleotide sequence identity (99%) to a prophage encoded within the SF1 genome in the same locus, whereas it is entirely absent in YSFST01-82, ISET0901, and ISNO (Figure 3).

Characterization of *S. iniae* QMA0248 insertion sequences

Insertion sequences (IS) were analyzed in the *S. iniae* QMA0248 genome using the ISFinder database coupled with manual curation. The analysis revealed 92 IS (Table 3), which is higher than the average number per bacterial genome (n=38) but consistent with the lifestyle of *S. iniae* as a facultative pathogen (24, 25). Furthermore, the number of IS found in *S. iniae* QMA0248 is substantially higher than other Streptococci such as *S. mitis* strain B6 (n=63) but comparable to that of the Gram positive fish pathogen *Lactococcus garvieae* (26, 27). The 92 IS elements belong to 7 different IS families and 20 IS types. These include 9 novel types belonging to IS3, IS30, IS1182, and IS200/IS605 families, which we have submitted to the ISFinder database (ISStin2–ISStin10) (Table 3 and Supplementary Tables S2–4). Around half of all IS copies in QMA0248 belong to these 9 novel types consistent with expansion of *S. iniae* specific IS since speciation (Table 3). Amongst those genes disrupted by IS in QMA0248 is the restriction enzyme component of a type II Restriction Methylation system that probably recognizes “GCNGC” (28). This insertion renders the cognate methyltransferase (MTase) (QMA0248_0516) an orphan and, given the high number of GCNGC sites (3814/Mb in QMA0248) across the genome, suggests a potential role for this

MTase in global gene regulation. By comparison, there are 7762/Mb available GATC sites in the *E. coli* K-12 MG1655 genome available for methylation by Dam, the archetypal orphan type II MTase known to play a major role in gene regulation (29). Further discussion of the methylome data generated by PacBio sequencing of QMA0248 is provided as Supplementary Information. . Taken altogether, the *S. iniae* genome harbors a large repertoire of IS elements, which may be associated with adaptation to host or environment. Indeed, it is well accepted that IS expansion is an early sign of genome reduction as a mechanism of adaptation to host (25, 30, 31).

Phylogenetic and comparative analysis of *S. iniae* QMA0248, SF1, YSFST01-82, ISET0901, and ISNO

The core genome of *S. iniae* QMA0248, SF1, YSFST01-82, ISET0901, and ISNO accounts for ~75% of the chromosome. IS were compared between the reference chromosome of QMA0248 and each of the *S. iniae* chromosomes (SF1, YSFST01-82, ISET0901, and ISNO) using Artemis Comparison Tool (ACT) (32). Although IS elements typically result in genomic rearrangements and loss of synteny, this is not seen in *S. iniae*. This lack of rearrangement is reflected by the consistent pattern of GC skew in the genome of *S. iniae* QMA0248 (Figure 1). Eight IS copies out of the 92 detected in QMA0248 are absent in the genomes of SF1, ISET0901 and ISNO only (Supplementary Table S2). Other IS elements are unique to the genomes SF1, ISET0901 and ISNO in syntenic positions. An interesting example is an IS981 (SF1 locus_tag: K710_0799 and K710_0800), inserted in *cas9* gene in the CRISPR/Cas region only in those three genomes (Figure 2). Another example is

IS*Stin1*, inserted only in SF1, ISET0901, and ISNO (SF1 locus_tag: K710_0761). Additionally, three IS copies are present in syntenic positions only in QMA0248, ISET0901, and ISNO. Eight insertions are absent in YSFST01-82 only and another seven IS copies are found exclusive to QMA0248 (Supplementary Table S2).

In addition to the IS differences, the five *S. iniae* genomes have major differences in the content of other mobile genetic elements (MGE) that reflect variations in the length of their respective chromosomes (Table 1). These MGEs include two prophages (Phi1 and Phi2), a genomic island (GI), and two other regions of difference (ROD) (Table 2, Figure 3). This includes a ~28 Kb region that is only found in YSFST01-82 (ROD1), and a ~20 Kb region that is present in four genomes but almost entirely absent from YSFST01-92 (ROD2) (Table 2, Figure 3).

Most variations in MGEs (including IS) were found to be incongruent with the core SNP phylogeny. For instance, the deleted ~12 Kb GI-Leu and the *cas9* gene disrupted by IS981 exist only in SF1, ISET0901, and ISNO (Figures 2 and 3), but these three isolates appear on divergent branches, indicating potential independent events (Supplementary Figure S4). To investigate the discrepancies between MGEs and phylogeny we compared multiple phylogenetic trees that were constructed using different methods, including the core genome, core SNP and using different software (Supplementary Figure S4). All phylogenies consistently revealed that *S. iniae* isolates QMA0248 and SF1 cluster together in one clade, whereas ISET0901 and ISNO cluster in another, and all four isolates cluster

separately to YSFST01-82, the latter diverging earliest from the root (Supplementary Figure S4).

Discrepancies between the *S. iniae* genomes are likely due to misassembly

Taken together, our comparative analyses suggest that one or more of the *S. iniae* genomes under comparison have been misassembled. The genomes ISET0901 and ISNO were both assembled from Illumina data using BioNumerics (Applied Math) and the genome *S. iniae* SF1 as a reference (12, 13). *S. iniae* SF1 was assembled *de novo* from a combination of 454 GS FLX+, Illumina MiSeq and Sanger sequencing (12). During the preparation of the present manuscript SF1 was removed from the NCBI RefSeq database (<https://www.ncbi.nlm.nih.gov/refseq/>). YSFST01-92 was also a hybrid assembly (454 GS FLX Titanium, Opgen optical mapping and Sanger sequencing) but it remains in the RefSeq database. Reference-guided assembly was introduced to enable comparisons between two very closely related isolates. However, this practice can result in the erroneous inclusion of MGEs that exist in the template genome but are absent from the comparison strain. Even with careful curation it is impossible to avoid misplacing repetitive sequences such as IS, as observed here in the case of *cas9* insertion and the other 8 IS copies that are absent in SF1, ISET0901, and ISNO only. Moreover, reference-guided assemblies may result in the loss of novel regions that are only present in the newly sequenced strain, in which case a *de novo* approach is always required (33). Although reference-guided assembly is no longer generally accepted for prokaryote genomes, a number of examples remain available in public repositories such as GenBank. For both ISET0901 and ISNO the assembly strategy is clearly outlined in the comment field of the GenBank file, and the primary publications (12, 13).

Never-the-less, the consequences of using such genomes in downstream analyses may not be apparent to all (for example, all three genomes are available in widely used genome databases such as PATRIC (www.patricbrc.org version 3.5.36)(34)). Removal of some early hybrid 454 complete genomes from public repositories like RefSeq should help maintain the quality of available complete genomes. Long-read sequencing data from Pacific Biosciences and Oxford Nanopore bring complete bacterial genomes within reach of most laboratories, but here also significant care is often required to avoid misassembly. Furthermore, as illustrated here and in other studies (35, 36), what appear to be misassemblies may in fact be biologically relevant. Ultimately the onus is on the user of public data to exercise caution when validating the source, assembly strategy, and quality of available complete genomes.

Conclusions

We assembled and annotated a high quality complete genome sequence for *S. iniae* QMA0248, including manual curation of 92 insertion sequences. Comparative analysis with publicly available complete genomes *S. iniae* SF1, YSFST01-82, ISET0901, and ISNO revealed discrepancies in the MGE content consistent with errors introduced by reference guided assembly of ISNO and ISET0901. Such problems are not new, but many bacterial genomes assembled in this way remain in public repositories of complete genomes. Our results emphasize the need to critically appraise complete genome assemblies prior to comparative analysis. Despite long-read sequencing becoming the gold standard for complete genome assembly of bacterial isolates, caution is needed to avoid misassembly. To better understand how insertion sequences, genomic islands, and other mobile elements

contribute to *S. iniae* diversity, there is a need for larger genomic studies using global collections of *S. iniae* isolates from dissimilar origins. The genome of *S. iniae* QMA0248 represents an important resource for future *S. iniae* comparative genomic and evolutionary studies.

Materials and Methods

Bacterial strain and sequencing

Streptococcus iniae strain QMA0248 was isolated from diseased barramundi (*Lates calcarifer*) from a farm in South Australia in 2009 (37). Genomic DNA was prepared from several well-isolated colonies of *S. iniae* QMA0248 grown for 18 h on Todd-Hewitt agar from a master seed stock (non-subcultured) with the Genomic Tip 20 kit (Qiagen). Pre-incubation for 2 h at 37°C of cells suspended in 500 µL 50 mM EDTA containing 200 units of mutanolysin and 2 mg lysozyme was found to improve cell lysis prior to following the manufacturers protocol for purification of high molecular weight DNA. The genome of the strain QMA0248 was sequenced using 3 SMRT cells on the Pacific Biosciences (PacBio) RS II platform and P4C2 sequencing chemistry, which generated a total of 57,083 reads with an average length of 6,178 bp. Reads were deposited at BioProject [PRJNA385746](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA385746) under accession [SRP109617](https://www.ncbi.nlm.nih.gov/sra/SRP109617). The genome of the strain QMA0248 was also sequenced using Illumina Nextera libraries on HiSeq2000.

Genome assembly and modified bases detection

PacBio sequencing reads derived from *S. iniae* QMA0248 genomic DNA were assembled using HGAP (hierarchical genome assembly process) using the PacBio Single Molecule Real Time (SMRT) Portal (V2.3.0) (38), with default settings and

minimum seed read length of 4,509 bp. Contiguity was used to visualize the assembly and the overlap between contigs using BLASTn (39, 40). The resulting assembly was used as a reference in the RS Resequencing module of PacBio's SMRT Analysis v2.3.0 to map the raw reads onto the reference genome producing a highly accurate genome consensus. Illumina-sequenced reads of the strain QMA0248 were mapped to the PacBio assembly using Snippy v3.0 (<https://github.com/tseemann/snippy>). To analyze read pileup in the potential rRNA tandem operon, raw reads of QMA0248 were mapped onto the ~7 Kb rRNA contig using BLASR v2.2, as part of the PacBio's SMRT Analysis Suite, and visualized using Artemis (41). Methylated DNA bases were identified in the resulting genome assembly using the RS Modification and Motif Analysis protocol and Motif Finder v1 within the SMRT Analysis suite v2.3.0 using Quality Value (QV) cutoff of 30. DNA Methyltransferases (MTases) were identified using nucleotide comparisons against REBASE (42).

PCR to investigate the rRNA tandem duplication

To investigate the presence of a rRNA tandem repeat, long range PCR from unique adjacent region to unique adjacent region was performed using specific primers for each of 5 rRNA regions (Supplementary Table S1), which were designed using Primer3 (43). PCR was done using LongAmp® Taq PCR Kit (NEB) from 40 ng of QMA0248 *S. iniae* genomic DNA as follows: 5 min at 94°C; 30 cycles of (30 sec at 94°C, 30 sec at 56°C, and 15 min at 65°C); and 10 min final extension at 65°C. The gel was loaded with 5 µL QUICK-LOAD® 1 Kb Extend DNA Ladder and 1µL of PCR products and run for 90 min at 70V using 0.7% TAE buffer solution and stained with HydraGreen.

Genome and mobile genetic elements annotation

Automated genome annotation was done using Prokka v1.11 (Prokaryotic Genome Annotation System) (44) and then manually curated using Artemis (41) and GenePrimp (45). The start codons of known coding sequences (CDS), such as in the capsule and streptolysin S operons, were further adjusted where appropriate using UniProtKB (<http://www.uniprot.org/>) and Pfam (46). CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) were predicted using CRISPR finder tool (<http://crispr.u-psud.fr/crispr/>) (47). Prophage annotation was done using PHAST (Phage Search Tool) (<http://phast.wishartlab.com/>) (48). Island Viewer (<http://www.pathogenomics.sfu.ca/islandviewer/>) was used to predict genomic islands (GI) (49). Boundaries of phage and GI regions were manually adjusted to their respective attachment sites. IS Saga (<http://www-genome.biotoul.fr/>) was used for the initial identification of insertion sequences (IS). Additional manual curation was carried out to confirm the boundaries of complete and partial IS elements using the IS Finder database (<http://www-is.biotoul.fr/>). IS element matches against the database that were at $\geq 95\%$ nucleotide identity were assigned the top matching IS name. IS elements of less percent identity are novel and were assigned the names IS*Stin2*–IS*Stin10* by ISFinder. The impact of IS elements on flanking coding sequences was analyzed using Artemis and by searching the amino acid sequence in UniProt KB (<http://www.uniprot.org/>), and Pfam databases (41, 46). The complete annotated genome sequence was deposited at GenBank under the accession number CP022392.

Comparative genomics and phylogenetic analysis

Alignments of the whole-genome or genomic sub-regions, such as the CRISPR and genomic islands, were done using BLASTn implemented in EasyFig v2.1 (40, 50). Detailed analysis of regions of difference and comparison of insertion sequences (IS) were done using Artemis Comparison Tool (ACT) (32). The core genome of the five genomes QMA0248, SF1, YSFST01-82, ISET0901, and ISNO was defined using Roary (51). Phylogenetic trees were constructed using the core genome and the core SNP methods, using multiple programs (see below), using the strain QMA0140 as the out-group (52). The strain *S. iniae* QMA0140 is a dolphin isolate from USA in 1976, and was sequenced using Illumina HiSeq 2000 (See Bioproject PRJNA417543). For quality control, the first 20 bp of each read derived from QMA0140 genomic DNA were hard trimmed using Nsoni v0.132 (<https://github.com/Victorian-Bioinformatics-Consortium/nsoni>) with a minimum length and quality of 70, and 20, respectively. Hard trimmed filtered reads of QMA0140 were assembled using SPAdes v3.9.0 where contigs < 10X coverage and smaller than 100 bp were removed (53). For core genome phylogenies, whole-genome alignment of the five complete genomes and QMA0140 draft assembly was done using Mauve v2.4.0 and Parsnp v1.2 with default parameter settings (54, 55). For the alignment using Mauve, conserved blocks in all 6 genomes longer than 500 bp were selected and concatenated using the stripSubsetLCBs script producing the core genome alignment. For core SNP phylogenies, error-free simulated reads were created using wgsim v0.3.2 (<https://github.com/lh3/wgsim>) and mapped to the reference genome QMA0248 along with QMA0140 hard trimmed and filtered raw reads using bowtie v1.0.0 (56), where variants were called using Nsoni v0.132

419 (<https://github.com/Victorian-Bioinformatics-Consortium/nesoni>) using default
 420 parameters. Core SNPs were also identified by mapping the reads of the 6
 421 genomes to QMA0248 using BWA-MEM v0.7.15 (r1140), implemented in Snippy
 422 v3.0 (<https://github.com/tseemann/snippy>) (57). All trees were produced by
 423 RAxML v8.2.9 (58) using the general time-reversible (GTR) and GAMMA
 424 distribution model of among-site rate variation with bootstrapping from 1000
 425 replicates, and viewed using FigTree v1.4.0
 426 (<http://tree.bio.ed.ac.uk/software/figtree>).
 427

- 428 1. Agnew W, Barnes AC. Streptococcus iniae: an aquatic pathogen of global
429 veterinary significance and a challenging candidate for reliable vaccination.
430 Veterinary microbiology. 2007;122(1-2):1-15.
- 431 2. Shoemaker CA, Klesius PH, Evans JJ. Prevalence of Streptococcus iniae in
432 tilapia, hybrid striped bass, and channel catfish on commercial fish farms in the
433 United States. American journal of veterinary research. 2001;62(2):174-7.
- 434 3. Lau SK, Woo PC, Luk W-k, Fung AM, Hui W-t, Fong AH, et al. Clinical
435 isolates of Streptococcus iniae from Asia are more mucoid and β -hemolytic than
436 those from North America. Diagnostic microbiology and infectious disease.
437 2006;54(3):177-81.
- 438 4. Baiano J, Barnes AC. Towards control of Streptococcus iniae. Emerg Infect
439 Dis. 2009;15(12):1891.
- 440 5. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements:
441 the agents of open source evolution. Nature reviews Microbiology.
442 2005;3(9):722.
- 443 6. van Elsas JD, Bailey MJ. The ecology of transfer of mobile genetic
444 elements. FEMS microbiology ecology. 2002;42(2):187-97.
- 445 7. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the
446 nature of bacterial innovation. Nature. 2000;405(6784):299-304.
- 447 8. Partridge SR. Analysis of antibiotic resistance regions in Gram-negative
448 bacteria. FEMS microbiology reviews. 2011;35(5):820-55.
- 449 9. Siguier P, Gournayre E, Chandler M. Bacterial insertion sequences: their
450 genomic impact and diversity. FEMS microbiology reviews. 2014;38(5):865-91.

- 451 10. Cheng S, Hu YH, Jiao XD, Sun L. Identification and immunoprotective
452 analysis of a *Streptococcus iniae* subunit vaccine candidate. *Vaccine*.
453 2010;28(14):2636-41.
- 454 11. Rajoo S, Jeon W, Park K, Yoo S, Yoon I, Lee H, et al. Complete genome
455 sequence of *Streptococcus iniae* YSFST01-82, isolated from olive flounder in Jeju,
456 South Korea. *Genome announcements*. 2015;3(2):e00319-15.
- 457 12. Pridgeon JW, Zhang D, Zhang L. Complete genome sequence of a virulent
458 strain, *Streptococcus iniae* ISET0901, isolated from diseased tilapia. *Genome*
459 *announcements*. 2014;2(3):e00553-14.
- 460 13. Pridgeon JW, Zhang D, Zhang L. Complete genome sequence of the
461 attenuated novobiocin-resistant *Streptococcus iniae* vaccine strain ISNO.
462 *Genome announcements*. 2014;2(3):e00510-14.
- 463 14. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, et al.
464 Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive
465 neonatal disease. *Molecular microbiology*. 2002;45(6):1499-513.
- 466 15. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, et al.
467 Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*.
468 *Science*. 2001;293(5529):498-506.
- 469 16. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, et al. Complete
470 genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the*
471 *National Academy of Sciences*. 2001;98(8):4658-63.
- 472 17. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et
473 al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*.
474 2007;315(5819):1709-12.

- 475 18. Gong H-Y, Wu S-H, Chen C-Y, Huang C-W, Lu J-K, Chou H-Y. Complete
476 Genome Sequence of Streptococcus iniae 89353, a Virulent Strain Isolated from
477 Diseased Tilapia in Taiwan. Genome Announcements. 2017;5(4):e01524-16.
- 478 19. Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and
479 redundancy of 16S rRNA sequences in genomes with multiple rrn operons.
480 Journal of bacteriology. 2004;186(9):2629-35.
- 481 20. Lim K, Furuta Y, Kobayashi I. Large variations in bacterial ribosomal RNA
482 genes. Molecular biology and evolution. 2012;29(10):2937-48.
- 483 21. Fischetti V, Pancholi V, Schneewind O. Conservation of a hexapeptide
484 sequence in the anchor region of surface proteins from gram - positive cocci.
485 Molecular microbiology. 1990;4(9):1603-5.
- 486 22. Switalski LM, Patti JM, Butcher W, Gristina AG, Speziale P, Höök M. A
487 collagen receptor on Staphylococcus aureus strains isolated from patients with
488 septic arthritis mediates adhesion to cartilage. Molecular microbiology.
489 1993;7(1):99-107.
- 490 23. van der Ploeg JR. Genome sequence of the temperate bacteriophage PH10
491 from Streptococcus oralis. Virus genes. 2010;41(3):450-8.
- 492 24. Aziz RK, Breitbart M, Edwards RA. Transposases are the most abundant,
493 most ubiquitous genes in nature. Nucleic Acids Res. 2010;38(13):4207-17.
- 494 25. Ochman H, Davalos LM. The nature and dynamics of bacterial genomes.
495 Science. 2006;311(5768):1730-3.
- 496 26. Denapate D, Brückner R, Nuhn M, Reichmann P, Henrich B, Maurer P, et
497 al. The genome of Streptococcus mitis B6-what is a commensal? PLoS One.
498 2010;5(2):e9426.

499 27. Eraclio G, Ricci G, Fortina MG. Insertion sequence elements in *Lactococcus*
500 *garvieae*. *Gene*. 2015;555(2):291-6.

501 28. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE—a database for DNA
502 restriction and modification: enzymes, genes and genomes. *Nucleic acids*
503 *research*. 2009;38(suppl_1):D234-D6.

504 29. Lobner-Olesen A, Skovgaard O, Marinus MG. Dam methylation:
505 coordinating cellular processes. *Curr Opin Microbiol*. 2005;8(2):154-60.

506 30. Moran NA, Plague GR. Genomic changes following host restriction in
507 bacteria. *Current opinion in genetics & development*. 2004;14(6):627-33.

508 31. Richards VP, Lang P, Bitar PDP, Lefébure T, Schukken YH, Zadoks RN, et al.
509 Comparative genomics and the role of lateral gene transfer in the evolution of
510 bovine adapted *Streptococcus agalactiae*. *Infection, Genetics and Evolution*.
511 2011;11(6):1263-75.

512 32. Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG,
513 Parkhill J. ACT: the Artemis comparison tool. *Bioinformatics*. 2005;21(16):3422-
514 3.

515 33. Nijkamp J, Winterbach W, Van den Broek M, Daran J-M, Reinders M, De
516 Ridder D. Integrating genome assemblies with MAIA. *Bioinformatics*.
517 2010;26(18):i433-i9.

518 34. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al.
519 Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis
520 Resource Center. *Nucleic Acids Res*. 2017;45(D1):D535-D42.

521 35. Forde BM, Ben Zakour NL, Stanton-Cook M, Phan MD, Totsika M, Peters
522 KM, et al. The complete genome sequence of *Escherichia coli* EC958: a high

523 quality reference sequence for the globally disseminated multidrug resistant E.
524 coli O25b:H4-ST131 clone. PLoS One. 2014;9(8):e104400.

525 36. Draper JL, Hansen LM, Bernick DL, Abedrabbo S, Underwood JG, Kong N,
526 et al. Fallacy of the Unique Genome: Sequence Diversity within Single
527 *Helicobacter pylori* Strains. MBio. 2017;8(1).

528 37. Millard CM, Baiano JC, Chan C, Yuen B, Aviles F, Landos M, et al. Evolution
529 of the capsular operon of *Streptococcus iniae* in response to vaccination. Applied
530 and environmental microbiology. 2012;78(23):8219-26.

531 38. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al.
532 Nonhybrid, finished microbial genome assemblies from long-read SMRT
533 sequencing data. Nature methods. 2013;10(6):563-9.

534 39. Sullivan MJ, Zakour NLB, Forde BM, Stanton-Cook M, Beatson SA.
535 Contiguity: contig adjacency graph construction and visualisation. PeerJ
536 PrePrints; 2015. Report No.: 2167-9843.

537 40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment
538 search tool. Journal of molecular biology. 1990;215(3):403-10.

539 41. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, et al.
540 Artemis: sequence visualization and annotation. Bioinformatics.
541 2000;16(10):944-5.

542 42. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE—a database for DNA
543 restriction and modification: enzymes, genes and genomes. Nucleic acids
544 research. 2009:gkp874.

545 43. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et
546 al. Primer3—new capabilities and interfaces. Nucleic acids research.
547 2012;40(15):e115-e.

548 44. Seemann T. Prokka: rapid prokaryotic genome annotation.
549 Bioinformatics. 2014;btu153.

550 45. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A,
551 et al. GenePRIMP: a gene prediction improvement pipeline for prokaryotic
552 genomes. Nature methods. 2010;7(6):455-7.

553 46. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths - Jones S, et al.
554 The Pfam protein families database. Nucleic acids research. 2004;32(suppl
555 1):D138-D41.

556 47. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify
557 clustered regularly interspaced short palindromic repeats. Nucleic acids
558 research. 2007;35(suppl 2):W52-W7.

559 48. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage
560 search tool. Nucleic acids research. 2011:gkr485.

561 49. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, et al.
562 IslandViewer 3: more flexible, interactive genomic island discovery, visualization
563 and analysis. Nucleic acids research. 2015:gkv401.

564 50. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison
565 visualizer. Bioinformatics. 2011;27(7):1009-10.

566 51. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al.
567 Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics.
568 2015;31(22):3691-3.

569 52. Baiano JC, Tumbol RA, Umapathy A, Barnes AC. Identification and
570 molecular characterisation of a fibrinogen binding protein from *Streptococcus*
571 *iniae*. BMC microbiology. 2008;8:67.

572 53. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al.
573 SPAdes: a new genome assembly algorithm and its applications to single-cell
574 sequencing. *Journal of computational biology*. 2012;19(5):455-77.

575 54. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of
576 conserved genomic sequence with rearrangements. *Genome research*.
577 2004;14(7):1394-403.

578 55. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid
579 core-genome alignment and visualization of thousands of intraspecific microbial
580 genomes. *Genome biology*. 2014;15(11):524.

581 56. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-
582 efficient alignment of short DNA sequences to the human genome. *Genome*
583 *biology*. 2009;10(3):R25.

584 57. Li H. Aligning sequence reads, clone sequences and assembly contigs with
585 BWA-MEM. *arXiv preprint arXiv:13033997*. 2013.

586 58. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-
587 analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-3.

588 59. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter:
589 circular and linear interactive genome visualization. *Bioinformatics*.
590 2009;25(1):119-20.

591

592 **Table 1: General features of five *S. iniae* complete genomes.**

Strain	QMA0248		SF1	YSFST01-82	ISSET0901	ISNO
Accession Number	CP022392		CP005941	CP010783	CP007586	CP007587
Genome Size (bp)	2,116,570		2,149,844	2,086,959	2,070,822	2,070,182
GC Content (%)	36.8		36.7	36.8	36.8	36.8
Total CDS number	1,946		2,125	1,897	1,872	1,865
Total gene number	2,196		2,196	2,029	1,997	1,996
rRNA (5S, 16S, 23S)	15		12	15	12	12
tRNA	58		45	58	45	45
Reference	This study		(12)	(11)	(12)	(13)
Assembly type	PacBio P2C4	RSII	454 FLX+/ Illumina MiSeq/ Sanger	454 FLX Titanium/ Opgen/ Sanger	Illumina 1500 HiSeq/ Reference guided assembly	Illumina 1500 HiSeq/ Reference guided assembly

593

Table 2: Large mobile genetic elements (MGE) and regions of difference (ROD) identified in the 5 *S. iniae* genomes analyzed (QMA0248, SF1, YSFST01-82, ISET0901, and ISNO).

Region	Genomic (GI-leu)	Island	ROD1	Prophage 2	ROD2	Prophage 1 (Phi1)
Coordinates¹	87374-100014		177819-206359 (YSFST01-82)	848479-890501 (SF1)	1767227-1787619	1991661-2023508
Length (Kb)	12.6		28.5	42.0	20.4	31.8
GC content (%)	35.9		37.2	35.3	34.3	37.6
Features	MGE; 13 tRNA and 1 rRNA operon (5S, 16S, and 23S) upstream; integrase; IS30 and IS256 family IS elements		Integrase; IS3 and IS256 family IS elements	MGE; integrase	IS3 family IS element	MGE; integrase; 1 tRNA-Cys
No. of CDSs	11		33	63	18	53
Major CDSs	Cro/CI family transcriptional regulator; ECF subfamily RNA polymerase sigma factor; plasmid replication protein; membrane protein		ESAT-6-like protein; two-component sensor histidine kinase; galactose mutarotase; 3 PTS galactitol transporter subunits (IIA, IIC, and IIB)	Phage DNA replication protein; prophage antirepressor, phage capsid and scaffold protein; putative tail protein; holin; endolysin; antigen C; several phage hypothetical proteins	ESAT-6-like protein; O-glycosyl hydrolase; phage infection protein; 3 lipoproteins; protein kinase	DNA helicase; Cro/CI family transcriptional regulator; tail and capsid proteins; holin; lysin; DNA N-4 cytosine methyltransferase; site specific recombinase; several phage hypothetical proteins
Best hit (% identity, % coverage)	<i>S. parauberis</i> KCTC 11537 (90, 41)		<i>S. dygalactiae</i> subsp. <i>equisimilis</i> ATCC 12394 (70, 34)	<i>S. parauberis</i> KCTC 11537 (91, 45)	<i>S. thermophilus</i> JIM 8232 (81, 13)	Bacteriophage PH10 of <i>Streptococcus</i> (71, 34)

¹Coordinates are in QMA0248 GenGenBank annotation (GCA_002220115.1) unless otherwise indicated.

598 **Table 3: Summary of all insertion sequences (IS) identified in QMA0248. Partial IS are suffixed by -p.**

IS Family in QMA0248	Number of IS copies	IS types (copy number, mean % amino acid identity)
IS3	32	ISSag2 (10, 98.6), ISSag2-p (2, 98.7) IS981 (8, 99.8) ISSpy1 (1, 86.7), ISSpy1-p (1, 90.5) *ISStin6 (3, 94.5), *ISStin7 (4, 90.0), *ISStin5 (2, 69.3), unclassified most similar to *ISStin5-p (1, 73.8)
IS30	22	ISSag9 (3, 99.6), ISSag9-p (5, 99.6) *ISStin4 (2, 93.3), *ISStin4-p (2, 89.6), *ISStin9 (9, 81.4), *ISStin2 (1, 84.7)
IS256	17	ISStin1 (16, 91.0), unclassified most similar to ISStin1-p (1, 90.2)
IS1182	13	*ISStin8 (7, 86.3), *ISStin8-p (2, 89.8), *ISStin3 (1, 87.7), *ISStin3-p (2, 87.1), unclassified most similar to *ISStin8-p (1, 70.5)
IS200/IS605	5	*ISStin10 (3, 99), *ISStin10-p (1, 98.6), unclassified most similar to *ISStin10 (1, 86.5)
ISL3	1	Unclassified most similar to ISStin1-p (1, 77.8)
IS110	2	Unclassified most similar to ISL4 (2, 65.8)

* Novel IS element.

599

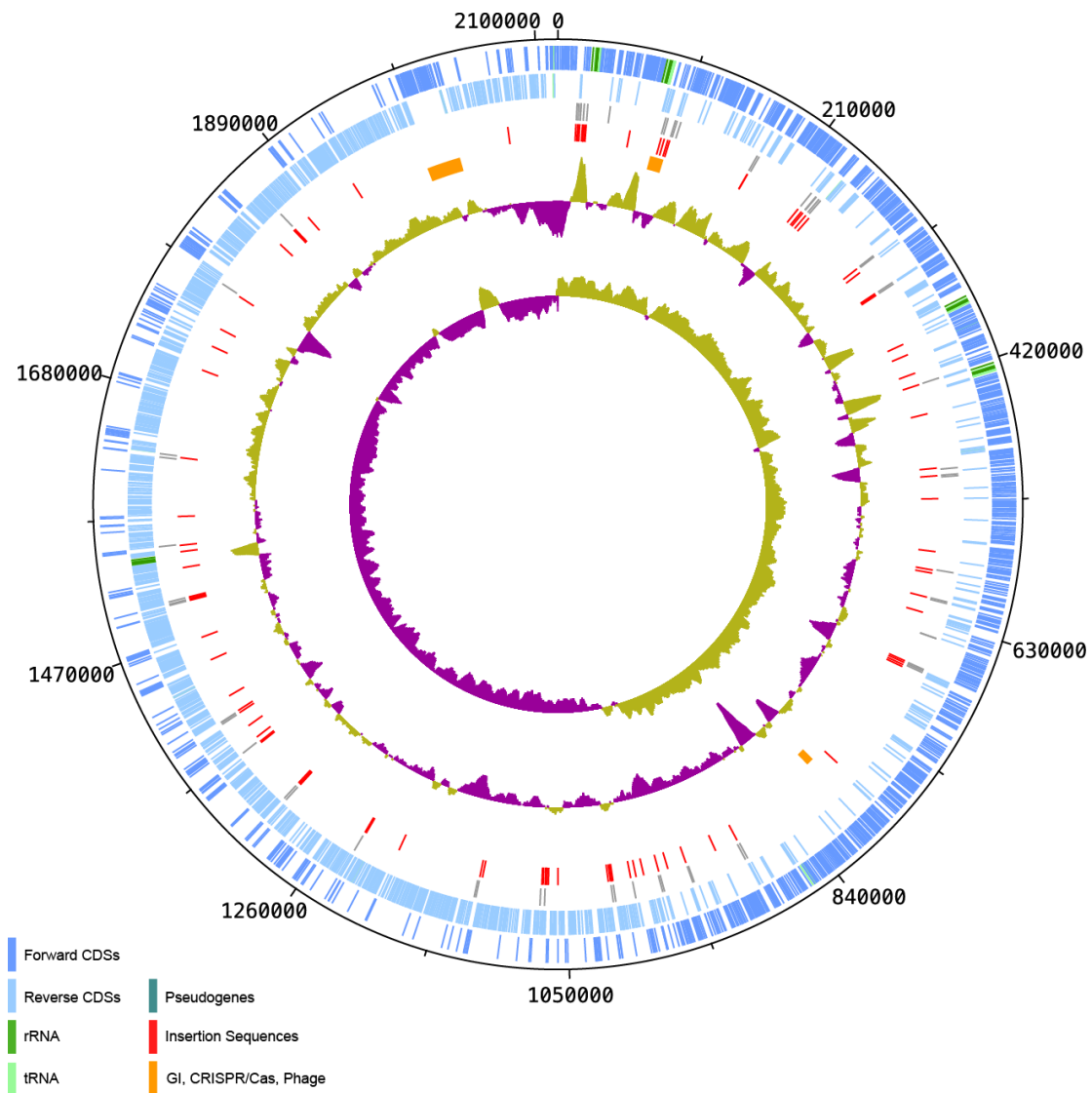


Figure 1: Circular map of the genome *Streptococcus iniae* QMA0248. Genomic features from outer ring to the inner ring are described in the key to the left, where the innermost two rings correspond to the GC skew (inner), and GC plot (outer). CDS: coding sequence. GI: genomic island. The circular map was generated using DNAPlotter (59).

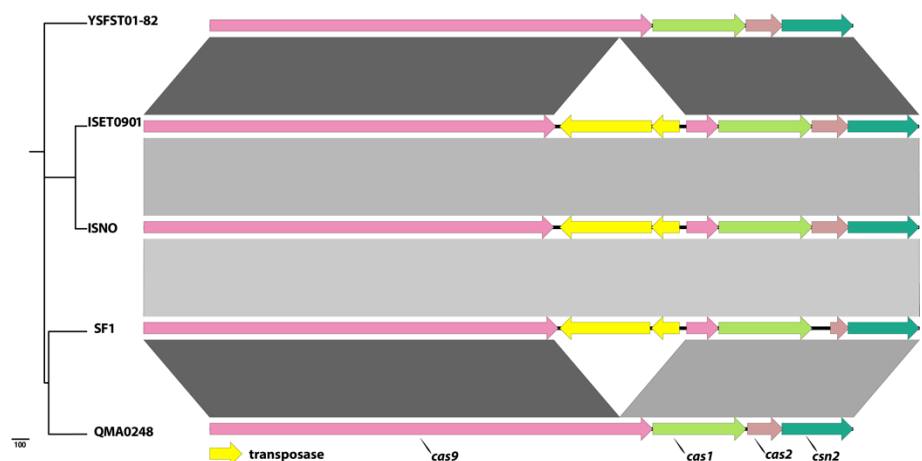


Figure 2: Comparison of the CRISPR/Cas region between QMA0248, SF1, YSFST01-82, ISET0901, and ISNO. Alignment of Cas genes where the genomes are ordered according to their position in the phylogenetic tree (left). The maximum Likelihood (ML) phylogeny was rooted to QMA0140 (not shown) and built using 1,111 SNPs. Arrows correspond to Cas genes, which are labeled at the bottom. Figure was produced using EasyFig (50) using 500 bp as minimum length, 90% as minimum identity value, and 0.001 as maximum e-value.

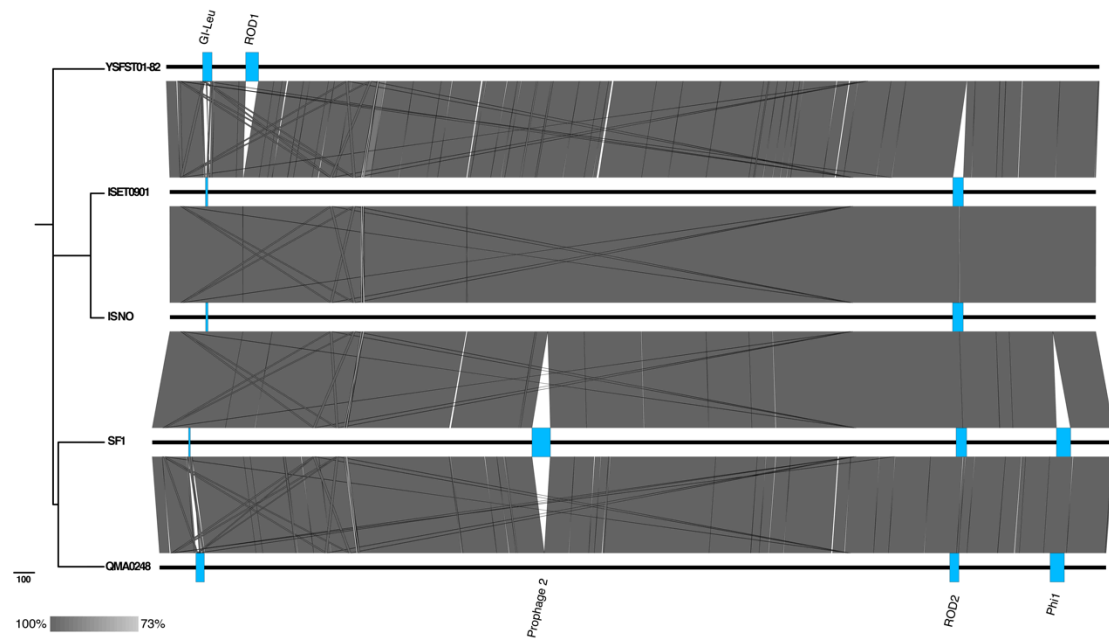


Figure 3: Whole-genome alignment of the 5 genomes QMA0248, SF1, YSFST01-82, ISET0901, and ISNO. The genomes are ordered according to their position in the core SNP-based phylogenetic tree. The maximum Likelihood (ML) phylogeny was rooted to QMA0140 (not shown) and built using 1,111 SNPs. The scale bar indicates the number of substitutions represented by branch lengths. BLASTn comparison was produced using EasyFig (50) using 2000 bp as minimum length, 50% as minimum identity value, and 1×10^{-17} as maximum e-value.

1 Importance

2 The present manuscript provides *Streptococcus iniae* PacBio sequenced genome (QMA0248) with
3 complete characterization of its mobile genetic elements (MGE) content including genomic islands,
4 prophages, and insertion sequences (IS). The genome of *S. iniae* QMA0248 harbours 92 IS elements
5 including 9 novel ones belonging mostly to IS3, which were submitted to ISFinder database. Here we
6 also present the first analysis for *S. iniae* methylome. We also performed comparative and phylogenetic
7 analysis between *S. iniae* QMA0248 and publicly available complete *S. iniae* genomes, which to our
8 knowledge has not been done for *S. iniae* before. This analysis revealed discrepancies that are likely due
9 to misassembly in the genomes of strains ISET0901 and ISNO. Our study not only highlights assembly
10 problems in existing *S. iniae* genomes, but provides a high quality reference genome for *S.*
11 *iniae* QMA0248, including manually curated MGEs, that will assist future *S. iniae* comparative genomic
12 and evolutionary studies.