Comparative Genomics and Reverse Vaccinology of *Streptococcus iniae*: Blueprints for Affordable Aquaculture Vaccines

Quentin Andres*[1,2], Anurak Bunnoy[1,2], Prapansak Srisapoome[1,2]

1. Laboratory of Aquatic Animal Health Management, Department of Aquaculture, Faculty of Fisheries, Kasetsart University, Chatuchak, Bangkok 10900, Thailand

2. Center of Excellence in Aquatic Animal Health Management, Faculty of Fisheries, Kasetsart University, Chatuchak, Bangkok 10900, Thailand

Correspondence:*
Corresponding Author
quentinludovicstephane.a@ku.th

Keywords: Comparative genomics, *Streptococcus iniae*, aquatic animal pathogens, pan-genome, reverse vaccinology, quality by design (QbD), aquaculture, vaccines.

Abbreviations: QbD quality by design, SNP single nucleotide polymorphism, SV structural variant, PAM presence absence matrix, IS insertion sequence, AMR antimicrobial resistance.

Author notes: Genome sequences are available under the NCBI bioproject number **PRJNA933632** .

Full Text:

Abstract

The high cost of vaccines relative to antimicrobials in aquaculture limits their adoption, particularly in low-value freshwater fish that nonetheless play a crucial role in global food security. Current vaccines, killed or live attenuated are highly effective but are no longer cost-efficient. Here, we present a new framework that leverages epidemiology, pangenomics, and reverse vaccinology with Quality By Design (QbD) principles to design more affordable vaccines. Targeting a range of aquatic streptococcal species, we emphasize the fish pathogen *Streptococcus iniae*. We identify a subset of protein antigens from the proteome that are well-adapted for low-cost mass production using QbD manufacturing strategies. This integrated approach is a crucial step towards making vaccines more accessible and cost-effective for global aquaculture, ultimately aiding in combating antimicrobial resistance and ensuring food security and consumer health.

Introduction

General context and problematic *[Your text here]*

The initial stage in developing recombinant protein, DNA and mRNA vaccines involves identifying a protective antigen for the target animal derived from a panel of proteins of the pathogen \cite{Rawal2021, Liu2008,Moriel2010}. Although vaccine research can be empirical, the abundancy of public knowledge on vaccine research , a lot of knowledge and research on GBS GAS s iniae.. then modern pan-genomics and epidemiology can provide elements that support a list of potential candidates. Ultimately each vaccine must meet several criteria to be adopted (safety price cost efficiency, demand, efficiency).

Firstly, while unknown proteins classified as hypothetical proteins may be of relevance as protective antigens \cite{Dey2022}, proteins incompatible with vaccine development, such as transfer RNA (tRNA), DNA polymerization proteins, and those known to be poor immunogens, must be excluded from the selection.

Re-explain the objectives of the work and introduce the datasets *[Your text here]*

and explain the significance of the work *[Your text here]*

Results

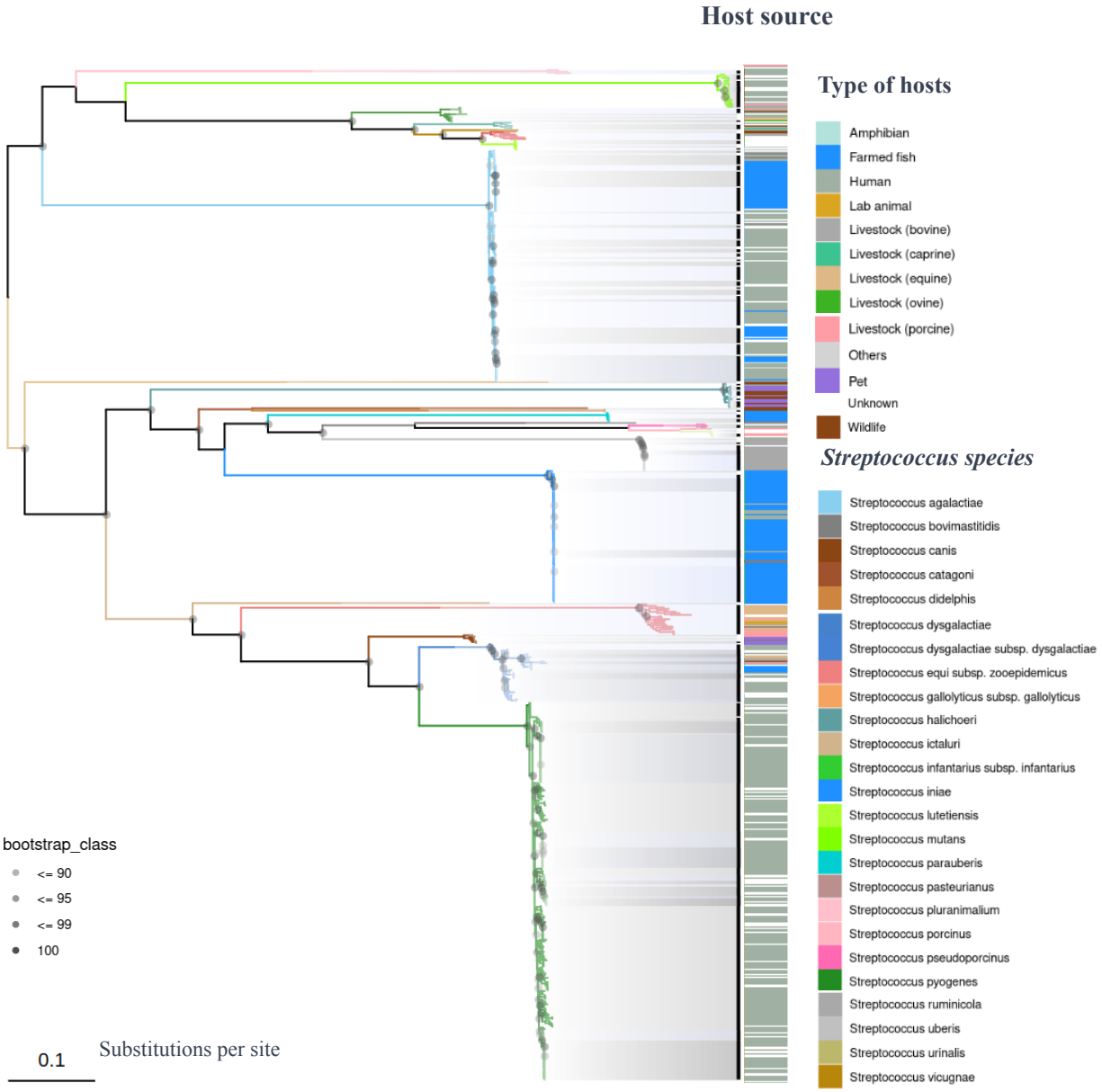
Streptococci species of aquatic animals are mostly found within the pyogenic phylogroup and are clustered separately into lineages

Introduction to Streptococcus Genus, present the group of species including bovis and pyogenic, then present Streptococcus species specific to aquatic animals *[Your text here]*

Explain the datasets, where the data comes from, which collection, how many samples, when it was sequenced and what was the isolation host/tissue, and the reason for the sequencing *[Your text here]*

Explain the methods used, why, and for what and what do we learn, i.e., what information do we obtain after the clustering of all Streptococcus of the genera but also after the Maximum likelihood phylogenetic tree constructed from the core-gene alignment analysis of the closest relatives of S iniae (=Bovis and Pyogenic group of species), and interpret it, put in perspective with the context/literature Describe the phylogenetic structure of the Genus at different "scales/levels" *[Your text here]*

Fig. 1: Phylogeny of Streptococcus Mutans, Bovis and Pyogenic group of species.



a) Maximum likelihood phylogenetic tree with lineages constructed from a core-gene alignment of a representative set of Streptococcus Pyogenic and Bovis groups of species using panaroo with a threshold of 95% for sequence identity, 70% for family identity and clean-mode="moderate" and built considering only non-monomorphic nucleotide sites with IQ-tree2 using GTR+F+ASC+R6 substitution model chosen according to BIC, rooted mid-point based on preliminary trees rooted with L. *lactis* as outgroup. The tree is colored by species, the background is split in rows based on fastBAPS clustering in gray shades for contrast. Branches are colored for the closest whole genomes fastANI clades at 95%. **b)** The legend represents labeled species (e.g. S. *iniae*), sources of isolations of the pathogens (i.e. hosts) and datasets. On the upper left a tree with a star topology of the genera Streptococcus with a background coloring for each group of species (e.g. Bovis, Mutans, Pyogenic) and on the lower left, a table of species and common hosts based on the frequent source of isolation.

Explain the figure and comment on it *[Your text here]*

Genome synteny and phylogenetic analysis highlight shared evolutionary history and important mosaicism between S. *iniae*, aquatic animal streptococci, and cattle streptococci, revealing conserved and divergent genomic regions across lineages

https://github.com/SamiLhll/macrosyntR

Explain the methods used, why, and for what and what do we learn, i.e., what information do we obtain after the analysis and interpret it, put in perspective *[Your text here]*

Describe the difference in terms of synteny and put it in perspective using phylogenetic data and clustering data *[Your text here]*

Fig. 2: The phylogroup pan-genome and phylogenetic context of aquatic animal pathogens from the Streptococcus Mutans, Bovis and Pyogenic group of species including *Streptococcus iniae.*

a) A presence/absence matrix of N genes in the *Streptococcus iniae* pan-genome, gen-erated using Panaroo, is displayed with lineage-specific shading alongside the maximum-likelihood tree in Fig. 1. The matrix is sorted by gene class as determined by Twilight. Gene classes are initially established within each lineage based on whether genes are core (≥ 95% of strains in each lineage), intermediate (>15% and ≤ 95% of strains), or rare (≤ 15% of strains). Subsequent inter-lineage comparisons categorize gene groups as either collection core, multi-lineage core, or lineage-specific core. Genes found at intermediate or rare occurrence across multiple or single lineages are classified as such. Coloring is used to indicate these classes: core in blue shades, intermediate in pink shades, and rare in orange shades. Genes that belong to one classification in a particular lineage but another classification in a different lineage are referred to as hybrid classes (in green shades). **b)** UpSetR Plot and Lineage Membership: An UpSetR plot reveals the N largest intersections of lineage-specific core genomes (genes present in ≥ 95% of strains in each lineage). Black dots in the matrix below the stacked bar plot indicate lineages contributing to each intersection. Stacked bar plots show the number of genes in each intersection, color-coded according to Twilight-assigned gene classes. Singleton lineages are included for completeness. Rows in the matrix correspond to each lineage and are color-coded based on *Streptococcus iniae* species, as defined by fastANI. Red boxes indicate intersections represented in Supplementary Figs. X–Y. **c)** Pan-Genome Accumulation Curves: The estimated pan-genome accumulation curves for each Streptococcus group within Asian seabass are depicted. The shaded region represents the standard deviation. Colors correspond to different species according to the key in part c. Source data are provided as a Source Data file.

Explain the figure and comment on it *[Your text here]*

Pangenomes and metabolic pathways across lineages of the pyogenic group help to differentiate between streptococcus of aquatic animals and reveal inter-specific preferences in ecotypes

Explain the methods used, why, and for what and what do we learn, i.e., what information do we obtain after the analysis and interpret it, put in perspective *[Your text here]*

Fig. 3: Predicted metabolic pathways in *Streptococcus iniae* and across other Streptococci and correspondence with aquatic health diagnosis testing historical data.

a) Predicted Metabolic Pathways Across *Streptococcus iniae* Metabolic pathways across the *Streptococcus iniae* lineages are predicted using Pathway Tools, following re-annotation of assemblies with functional annotation of representative protein sequences, which is carried out through Interproscan/EggNOG. These pathways are displayed alongside the maximum-likelihood tree from Fig. 1. **b)** Presence/Absence of Selected Metabolic Pathways: The presence/absence of selected metabolic pathways critical for *Streptococcus iniae* in Asian seabass is shown. For example, presence of X degradative pathways and that of Y pathway are colored in blue. The presence of a specific gene cluster required for the biosynthesis of Z molecule is colored in red to reflect the corresponding bio-indicator. Pathways are selected based on a subset of biochemical tests or health indicators used for aquatic animal health management. **c)** Variability in Metabolic Pathways: Tables representing variability (v, variable) in the pathways across different strains or isolates are presented. **d)** Habitat Source and Health Indicators: Habitat source and associated health indicators for different *Streptococcus iniae* strains are detailed, adapted from (your reference here).

Explain the figure and comment on it *[Your text here]*

There are infra-specific variations in S. *iniae* in terms of gene absence-presence, structural variants, genetic diversity, and distribution of mobile genetic elements resulting in potentially different observable phenotype in accordance with historical data

Description of Streptococcus iniae *[Your text here]*

Description of knowledge about Streptococcus iniae for pathogenesis and for immunology (host-pathogen) [Your text here]

Explain the methods used, why, and for what and what do we learn i.e. what information do we obtain after the analysis and interpret it, put in perspective [Your text here]

Fig. 4: Comprehensive pan-genomic and phylogenetic analysis of *Streptococcus iniae* the principal bacterial pathogen of Asian Seabass (Lates *calcarifer)*.

This figure serves as a multi-layered representation of the pangenome, phylogenetic relationships, and genomic plasticity within the *Streptococcus iniae* species, particularly in Asian seabass. Using Panaroo, the pangenome network highlights the core, intermediate, and rare genes across different lineages, giving an in-depth view of the functional genome landscape. Adjacent to the network is a maximum-likelihood phylogenetic tree, showcasing the evolutionary relationships among the analyzed strains. Also featured are distinct 'regions of plasticity,' which are demarcated areas within the genomes that display a high rate of insertions, deletions, or horizontal gene transfers, and are of particular interest for their potential roles in virulence, antibiotic resistance, or host adaptation. A linear representation of the pangenome, created using pggb and odgi, is displayed in both one-dimensional and two-dimensional formats. The 1D representation linearizes the complex pangenome structure, whereas the 2D representation allows for the visualization of higher-order relationships and structural rearrangements. Together, these components provide a comprehensive view of the genetic architecture, evolutionary dynamics, and functional capabilities of *Streptococcus iniae* in Asian seabass.

Explain the figure and comment it [Your text here]

Explain the methods used, why, and for what and what do we learn i.e. what information do we obtain after the analysis and interpret it, put in perspective [Your text here]

Fig. 5: Mobile genetic elements (MGEs) and corresponding phylogeny in *Streptococcus iniae* and closely related species.

This figure provides a focused examination of mobile genetic elements, specifically Insertion Sequences (IS) and their transposases, across *Streptococcus iniae* and its closest phylogenetic relatives. Using ISfinder and ISMapper for in-depth annotation and mapping, the distribution of various IS types is depicted across different strains or isolates. Aligned with this is a maximum-likelihood phylogenetic tree, which enables a simultaneous view of evolutionary relationships and the dynamics of mobile genetic elements within this lineage. The presence, absence, or variation in IS types across lineages could be indicative of mechanisms driving genome plasticity, virulence, or antibiotic resistance. By integrating the phylogenetic tree with the mapping of IS transposases, this figure offers a holistic understanding of how mobile genetic elements may influence the evolutionary trajectory and functional capabilities of *Streptococcus iniae* and its closest relatives.

Explain the figure and comment it [Your text here]

Homologous recombination is minimal in S. iniae and is not an active driver of evolution, and does not yet contribute to vaccine escape or antibiotic resistance but it has played a role in the acquisition of a Type VII secretion system at one point of evolution shared by most lineages

Explain the methods used, why, and for what and what do we learn i.e. what information do we obtain after the analysis and interpret it, put in perspective [Your text here]

In Streptococcus iniae, evidence suggests that homologous recombination plays a minimal role in the organism's evolutionary dynamics. Unlike in some bacterial species where recombination actively contributes to genetic diversity, vaccine escape, or antibiotic resistance, this mechanism appears to be subdued in S. iniae. Nevertheless, genomic analyses have revealed that homologous recombination has had a tangible impact at specific points in the species' evolutionary history.

One noteworthy event is the acquisition of a Type VII secretion system (T7SS), which is shared by most S. iniae lineages. The presence of T7SS, generally involved in host-pathogen interactions and virulence, is indicative of the role recombination has played in shaping some functional aspects of the S. iniae genome. The conservation of this system across multiple lineages further underscores its potential importance, perhaps in niche adaptation or interbacterial competition.

Fig. 6: Homologous recombination and foreign DNA imports in S. *iniae*.

This figure, generated using Gubbins, offers a comprehensive analysis of homologous recombination events and the acquisition of foreign DNA in various S. *iniae* strains. Notably, the data shows an almost complete absence of foreign DNA import across most lineages, except for a conspicuous presence of the TSS7 element, indicated as a red band on the genomic maps. This singular instance of foreign DNA incorporation is plotted alongside its genomic locations in multiple lineages. The integrative display aims to illuminate the role of genetic exchange—or the lack thereof—in the evolutionary history and pathogenicity of S. *iniae*.

While homologous recombination might not currently contribute to vaccine escape or antibiotic resistance in S. iniae, its past influence on the acquisition of systems like the T7SS should not be overlooked. Ongoing comparative genomics and functional studies could shed further light on how mechanisms like recombination intersect with the ecology and pathogenicity of S. iniae, providing crucial insights for future therapeutic strategies, including vaccine development.

Leveraging pangenomics and epidemiology for robust vaccine strategy and monitoring: a case study on capsular operon variations in S. iniae

Explain the methods used, why, and for what and what do we learn i.e. what information do we obtain after the analysis and interpret it, put in perspective [Your text here]

Fig. 7: S. *iniae* capsular operon, its single nucleotide polymorphisms and structural variations.

https://europepmc.org/articles/PMC8006571/figure/F2/

This figure delves into the structural and functional aspects of the capsular operon in S. *iniae*. The distribution of SNPs and SVs within this operon is highlighted, offering insights into possible variations that could influence the bacterial capsule's role in virulence and host interactions.

Explain the figure and comment on it, Give directions [Your text here]

Leveraging pangenomics and epidemiology for robust vaccine strategy and monitoring: finding conserved proteins with limited structural variations

Pan-genome analysis is useful as there may be strain specific variation in gene content, such as some genes present in some strains but absent in other therefore a pan-genome at the species level would allow to understand strain specific variation and to find and target cross-strain protective antigens \cite{Monk2022}, whilst a pan-genome at the genus level would allow to understand species specific variation and to find and target cross-species protective antigens. ***Williams2022,Bobay2018***

Core proteins can be identified from the pan-genome analysis and in general they have lower recombination rates than non-core genes but it is not always the case \cite{Preska_Steinberg_2022}, one way to study the mutations of a certain gene by comparing all individuals to each other is to create a pan-genome of SNPs and structural variations with Pandora \cite{Colquhoun2021}. A list of orthologues \cite{Zielke2016} in other pathogenic streptococci of the Asian seabass can also be studied using bidirectional BLASTn for genes \cite{HernndezSalmern2020,MorenoHagelsieb2007} or bidirectional Diamond \cite{Buchfink2014} for proteins.

Explain the methods used, why, and for what and what do we learn i.e. what information do we obtain after the analysis and interpret it, put in perspective [Your text here]

Fig. 8: Genome-wide single nucleotide polymorphisms (SNPs) and structural variants (SVs) in *Streptococcus iniae* and antimicrobial resistance (AMR) genes.

This figure presents an integrative view of the genome-wide distribution of SNPs and SVs in various *Streptococcus iniae* strains, plotted alongside the phylogenetic tree to enable comparison of genotypic diversity. Utilizing high-throughput sequencing data, SNPs and SVs were identified and characterized to provide insights into genomic variability and its potential functional implications. To emphasize the clinical relevance, this figure also includes a parallel annotation of AMR genes, highlighting their location and frequency across the different strains. The analysis of AMR genes is crucial for understanding the potential resistance mechanisms that could be present in these bacterial strains. Together, the SNPs, SVs, and AMR genes provide a comprehensive snapshot of the genomic landscape of *Streptococcus iniae*, aiding in both epidemiological understanding and targeted therapeutic interventions.

Explain the figure and comment it [Your text here]

By applying quality by design (QbD) methodologies to reverse vaccinology, we can more efficiently identify the most suitable antigens for S. iniae vaccine development, tailored to specific cost-effective manufacturing strategies and use-cases

An ideal antigen for a protein based vaccine should have certain characteristics that make it suitable for efficient production and purification. One such characteristic is its isoelectric point (pI), which is the pH at which a protein carries no net charge. An ideal antigen should have a pI between 7 and 9, making it negatively charged under physiological conditions (pH ~7.4). This is because proteins with pI values between 7 and 9 will be negatively charged at neutral pH, which facilitates the purification process. \cite{Widmann2010,Freitas2022} Purification can be achieved through chromatography using a positively charged histidine tag in a nickel column. The negatively charged protein will bind to the positively charged histidine tag, allowing it to be separated from other proteins in the mixture. Alternatively, the protein can be purified using a column containing a silicon dioxide-based filter, which also utilizes the charge interactions between the protein and the filter material. \cite{Freitas2022,Spriestersbach2015} If the protein is too positively charged, it becomes more difficult to separate it from other proteins, as it will not bind as effectively to the histidine tag or the silicon dioxide-based filter. In such cases, higher concentrations of eluting competitors like imidazole or L-lysine are needed to help separate the protein. However, this complicates the process and increases production costs, making it less desirable for vaccine development. \cite{Freitas2022,Spriestersbach2015}

Table 1: Reverse vaccinology scoring and proposed S. *iniae* antigens per type of vaccine manufacture strategy.

This table compiles candidate antigens identified through reverse vaccinology for S. *iniae*. Each antigen is scored based on various criteria, including predicted immunogenicity, conservation across strains, and functional relevance. The proposed antigens serve as a roadmap for potential vaccine development.

In addition to the isoelectric point, other protein attributes like stability and folding \cite{Scheiblhofer2017,Thai2004}, solubility \cite{Zayas1997}, and codon usage \cite{Zhou2016} should be examined to determine if the protein can be effectively expressed and for a better immunogenicity. Protein conservation is also relevant, a comparative genomics approach can be employed to identify conserved proteins among various *Streptococcus iniae* species, as well as proteins conserved in other fish-pathogenic streptococcal species. This approach can provide valuable information on the protein's suitability for use as an antigen. An ideal antigen should possess favorable production characteristics, such as ease of production and scalability. The protein size should not be too large, as this reduces the surface area in contact with the immune system. Conversely, proteins that are too small tend to be unstable and exhibit improper folding. Typically, a protein size between 20kDa and 80kDa is preferred. \cite{Bachmann2010} Membrane proteins, however, differ as they are anchored in the membrane, which may affect their conformation when expressed. For instance, the M-like membrane protein Sim-A, which was unsuccessfully used as a vaccine in the past, likely adopted an inadequate conformation when expressed alone. \cite{Aviles2013,ErrastiMurugarren2021}

To simplify the selection process, it is beneficial to choose soluble proteins, such as secreted proteins or those known to be soluble. Some cytoplasmic proteins, like enolase, are also soluble and have been expressed in other streptococcal species. Moreover, the selected antigen should be non-toxic, as activated toxins can harm the fish and impose excessive stress on their health. Moreover, it is important to consider proteins with known epitopes that stimulate T and B lymphocytes \cite{SanchezTrincado2017}. This is because there is a 30\% similarity between fish and human immunoglobulins, which play a crucial role in the immune response to pathogens. By selecting proteins with known immunogenic properties, a more effective vaccine can be developed, targeting the immune system's ability to recognize and respond to the pathogen. \cite{Matz2021} \textcolor{red}{30\% to be verified.. ref}

The most effective vaccine strategy may involve protecting a lower percentage of individuals against the disease, provided the protein is produced more cost-effectively and in greater quantities. The initial design should be simple, and genes with higher GC content, particularly GC3, may be more highly expressed in vivo and suitable for a DNA vaccine. \cite{Beaudoin2022,Smon2004,Vinogradov2005} Gene length is not as critical, but the absence of restriction sites for Golden Gate cloning, such as BsaI or Eco31I, simplifies cloning by avoiding the need for mutagenesis through Polymerase chain reaction. \cite{Bird2022}

It is possible to create a custom scoring system to rank the potential candidate proteins to be included in a vaccine (see below.), based on all of the features and characteristics that were explained above. By giving positive and negative points to the proteins in S. *iniae* proteome and setting custom filters, we can reduce the list of potential candidates from around 300 manually reviewed and curated proteins believed to be immunogenic based on their protein signatures (from interproscan) and their function, to approximately 30 proteins which are believed to be adequate candidates for the characteristics listed above.

%In silico prediction of operons and regulons using OperonFinder \cite{Tomar2022,Assaf2021}, GOST \cite{Li2011,Reimand2011,Liu2016} and BoBro 2.0 motifs \cite{Ma2013} helps identify simultaneously expressed prokaryotic genes and proteins, offering insights into potential vaccine candidates most likely to be produced during infection. This information can be cross-referenced with existing literature on previous vaccines for S. *iniae* and other streptococcal species. Transcriptome

data from the pathogen during infection, such as protein expression levels over time, can also be considered. \cite{Lee2021} An ideal antigen should be soluble and exhibit virulence factor patterns or immune recognition molecular patterns (PAMPs) \cite{Mogensen2009}, enabling the immune system to detect bacterial presence even when the bacteria attempt to modulate the immune response. Targeting virulence factors in a vaccine can provide appropriate epitopes to resident immune cells in the skin and muscle, supporting a memory-based type B immune response that detects proteins expressed during infection. \cite{Merrikh2018} However, virulence factors may not always be the best candidate antigens, necessitating a compromise between maximizing purification/production and achieving a stable, protective protein.

Explain the methods used, why, and for what and what do we learn i.e. what information do we obtain after the analysis and interpret it, put in perspective [Your text here]

Fig. 9: Global amino acid variation in S. *iniae* Enolase and GAPDH.

a) Frequency of amino acid variations within the genomic dataset for enolase and GAPDH. **b)** Schematic representation of the enolase and GAPDH open reading frames, detailing the locations of amino acids within the mature enzymes (blue blocks). **c)** and **d)** Structural models of the consensus sequences of the mature enzymes of enolase (c) and GAPDH (d). The frequency of amino acid variation is plotted against each structure, as represented in the color gradient from 1\% (blue) to \textcolor{red}{value}\% (red); invariant sites are colored light gray ***Davies2019***

Explain the figure and comment it [Your text here]

Discussion

Streptococcus iniae, akin to numerous other aquatic animal pathogens, poses a substantial threat to farmed fish populations. The intricate challenges unique to aquaculture, marked by high fish densities and aquatic environments, amplify the intricacies of pathogen control. This hurdle is further exacerbated by the disproportionally high costs associated with producing vaccines relative to the lower economic value of the fish species. In fact, the aquaculture sector has bifurcated into two distinct segments: one focused on high-value vertebrate aquatic animal species (such as salmonids and koi carps), which has advanced considerably in vaccine research and modernization, and another concerning low-value species, where advancements have been limited due to cost-benefit considerations.

This disparity can be attributed to multifaceted factors, primarily hinging on the cost-effectiveness of vaccination for these species. Amidst the progress in modern genomics, vaccines are becoming increasingly sophisticated. While ethical considerations and regional regulations play a role, the inclination in low-value aquatic animal aquaculture tends to lean toward inactivated vaccines and live attenuated vaccines rather than recombinant DNA technology (protein / DNA / mRNA), driven by cost-efficiency. However, the landscape is evolving with the rise of synthetic biology, the broader dissemination of manufacturing process knowledge, and the potential to render DNA vaccines and protein vaccines up to 1000-fold more affordable than inactivated counterparts.

Simultaneously, the global aquaculture industry is expanding, with concerns about antibiotic resistance becoming more pronounced. Notably, aquaculture itself isn't a direct contributor to antimicrobial resistance (AMR); however, there's a growing need for alternative strategies to antibiotics in aquatic animal health management. Promising avenues of research include developing disease-resistant breeds, resilient brood-stock, immuno-metabolomics for improved feed, potent probiotics, efficient water management systems, novel antibiotics responsibly used and novel vaccines. In addition manufacturing vaccines that are durable in time is essential, in contrast to polysaccharide vaccines, protein based vaccines have the potential to be more resilient to the adaption by mutations of the pathogen population to vaccines. Therefore genomic data is a cornerstone to vaccine success and sampling should be increased in a directed manner for both sampling of non-clinical isolates (environmental isolates) and clinical isolates (diseased aquatic animals) in under-represented areas of the world were aquaculture is of importance such as in the African continent and in India

In this context, we presented a comparative genomic blueprint, seeking to foster the more cost-effective development of vaccines to address the specific challenges in aquatic animal health. Moreover, it's essential to underscore that pivotal patents related to DNA vaccines, and soon protein vaccines, are either expiring or belong to the public domain. Green vaccines are gaining traction as environmentally friendly alternatives. As developing countries enhance their aquaculture management practices, frameworks for vaccine development and deployment are being established, particularly for DNA and protein vaccines.

Simultaneously, Quality by Design frameworks are taking shape, bolstered by a deeper understanding of the effects of vaccines on hosts. This confluence of factors creates a conducive environment for the advancement of recombinant vaccines in aquatic health management.

Fig. 11: Summary of S. *iniae* pathogenesis and common vaccination strategies.

This figure serves as an integrative overview of the molecular mechanisms underlying the pathogenesis of S. *iniae* and the various strategies employed for vaccination. It juxtaposes key pathogenic factors, virulence genes, and the host immune response mechanisms against the innovative vaccine approaches currently under investigation. Icons and flow diagrams are used to illustrate the complex interplay between bacterial virulence and host defenses, while side panels provide a snapshot of both traditional and recombinant vaccine methodologies. This summary aims to offer a comprehensive view of the current state-of-the-art in combating S. *iniae* infections, guiding future research and application.

Explain the figure and comment it [Your text here]

Fig. 12: Summary of S. *iniae* pathogenesis and vaccination strategies.

a) This figure outlines the pathogenesis of S. *iniae* and associated vaccination strategies. Comparative genomics is used for antigen discovery and vaccine optimization. Epidemiology aids in tracking disease spread and vaccine effectiveness. Quality by Design is employed for vaccine formulation and production. **b)** Previous Vaccines for S. *iniae* Made Using Recombinant Technology and Associated Survival Rates: This figure provides a review of previously developed recombinant vaccines against S. *iniae*, showing their efficacy in terms of survival rates across different aquatic animals. The data aims to provide context for the current state of S. *iniae* vaccine development.

Explain the figure and comment it, conclude give directions [Your text here]

Methods

Data-sets

Table S1: Datasets

Dataset ID	Source_db	Type of Data	Format	Species	Sequences	Authors
DS1	This Study	Read Sequences	.fastq	S. *iniae*	5	Kasetsart U.
		Genomic assemblies	.fna	S. *iniae*	-	
DS2	NCBI (RefSeq)	Genomic assemblies	.fna	S. *iniae*		Multiple
DS3	NCBI (RefSeq)	Reference genomes	.fna	*Streptococcus spp.*	115	Multiple
				S. Bovis, Mutans, Pyogenic spp.	28	Multiple
DS4	NCBI (RefSeq)	Genomic assemblies	.fna	*S. Bovis, Mutans, Pyogenic spp.*	690	Multiple
DS5	NCBI (Genbank)	Genomic assemblies	.fna	*S. agalactiae.*		Multiple
DS6	NCBI (SRA)	Read Sequences	.fastq	S. *iniae*		A.C. Barnes
DS7	NCBI (RefSeq)	Genomic assemblies	.fna	S. *iniae*		Multiple
DS8	ISfinder	IS Transposases	.fasta DNA	S. *iniae*		Multiple
DS9	RegPrecise	Regulons	.fasta DNA	*Streptococcus spp.*		Multiple
DS10	IEDB	T/B Cell Epitopes	.fasta PROT	*Streptococcus spp.*		Multiple

Bacterial strains, DNA sequencing, genome assemblies, annotations and NCBI submissions

Sample isolation and DNA extraction: *Streptococcus iniae* isolated from farmed Asian seabass in (5 diseased animals, same farm and same location, in the region of Chachoengsao, Thailand) and sub-cultured, gram-stained and incubated overnight in lysozyme + proteinase K, 0.3-1ug/uL of purified gDNA for each pure culture was extracted using a nucleo-spin kit and sequenced on Illumina HiSeq2000 platform.
Quality control and read preparation: FastQC assessed the quality of five 3M Paired-end reads (151bp). Poly-G stretches not present in bacteria (affecting 6000 reads) were removed with Awk for R1 and R2 .fastq files. SeqPurge and Trimmomatic removed low-quality reads using a sliding window of 10 bases evaluated base quality, confirmed by a second FastQC run. Kraken2's k-mer search against an 8 Gb minimal bacterial database confirmed no sample contamination. For simplicity we used the same files for reference genome mapping and for de-novo assembly. . **Genome assembly and mapping:** Trimmed PE reads (Dataset **DS1**, 5 .fastqR1 + 5 .fastqR2) were de-novo assembled (=without prior genome information) using Unicycler (SPAdes), visualized in BANDAGE and quality metrics computed using QUAST. Complete genomes of *Streptococcus iniae* were downloaded from the NCBI Genomes Refseq database (Dataset **DS2**) aligned in Progressive mauve. Trimmed PE reads from **DS1** were aligned to each **DS2** reference using Bowtie2 fast-local option (less conservative), the read pileup was used to create a consensus to yield mapped contigs visualized in IGV. **DS2** reference genomes, **DS1** de-novo assemblies and mapped contigs were compared in Progressive mauve.

Draft corrections and variant calling: An “intermediate assembly” was created by merging and deleting DNA fragments in an approach known as "reference guided de-novo assembly". DNA fragments that are present in the **DS1** samples but absent in each of the **DS2** reference genomes were merged in a stepwise manner using QMA0248 as a starting point. Hence, gaps are defined by sequences that are absent in the reference but present in the sample and were resolved by merging some fragments of the de novo assembly from Unicycler onto the Bowtie2 mapped read consensus or deleting DNA regions of the intermediate assembly when evidenced by read mate information. Similarly the loci for CRISPR was also merged from de-novo contigs of each respective isolate resulting in 5 intermediate assemblies for which **DS1** trimmed PE reads were mapped using Bowtie2 end-to-end --no-unal --no-mixed --no-discordant (more conservative) with a quality filter of MAPQ > 10. More than 99.7% of the reads were mapped while preserving the synteny. The read pileup was used to create a consensus but without using the reference sequences when no reads were mapped onto it (Bcftools mpileup options --no-reference --ploidy 1). The variant calling and was done with PILON. The reads were finally remapped with SNIPPY(BWA-MEM). Each assembly was manually inspected using the IGV viewer.
Genome annotations and submission: S. *iniae* assemblies (**DS1** N=5 genomes) were annotated with NCBI's prokaryotic genome annotation pipeline (PGAP) standalone version. CheckM analysis (v1.2.2) was performed before submissions to NCBI Genbank. Illumina reads were uploaded in the SRA database under the bioproject number **PRJNA933632** .

Phylogenomic analysis

wgANI values were obtained using ANIclustermap 1. all *Streptococcus spp.* reference genomes (Dataset **DS3** N= 115 *species*) 2. a subset of DS3 including only Streptococcus from Mutans, Bovis and Pyogenic groups (N = 28 species).

Genome assemblies for *Streptococcus spp.* of Mutans, Bovis and Pyogenic groups and were downloaded from Refseq (Dataset **DS4**) or Genbank (Dataset **DS5**) and annotated with Prokka. Their metadata was downloaded from NCBI BioSample database using efetch and xtract (see Supplementary_Table_S1). *L. lactis* genome assemblies were also downloaded and annotated (N= 25 .fna) to serve as an outgroup for phylogenetic analyses.

Pan-genomes with and without *L. lactis* were built using Panaroo with options (clean_mode="moderate", family_threshold=0.7, core_threshold=0.95, aligner="mafft") resulting in a filtered-core-gene-alignment.aln, non-monomorphic sites were subsequently extracted using SNP-sites (option -c) to obtain core gene SNP alignments. IQ-TREE2 options () produced trees with and without *L. lactis*. *L. lactis* was excluded from the analysis for subsequent steps.

FastBAPS was run with options () on the previous core gene SNP alignments to obtain clusters at 4 levels. POPPUNK was run on each species separately using DbScan and linear models 1 to 16 to get clusters. fastANI was run on whole genome assemblies using ANIclustermap and a custom script extracted clusters for various thresholds (95, 96, 97, 98, 99% wgANI). ANI phylogroups were based on clusters at 95%. Lineages were based on level 3 fastBAPS clusters.

Pan-genomic studies

Pan-genomes were constructed using panaroo and the same options, core gene alignments were created for each phylogroups but also for each pair of lineages and for each pair of phylogroups (all-versus-all). Upset plots were made using custom scripts.

Comparative genomics

Insertion sequences were mapped using ISmapper, and intergenic regions were analyzed using Piggy. Draft genome was improved using a "reference guided de-novo assembly" approach.

Reverse vaccinology and quality by design

Antigen candidates were ranked based on a custom scoring matrix implemented in R. Epitope mapping was done using the DIAMOND algorithm against the IEDB database.

Data availability

The MiniKraken reference database used for species identification is available at https://ccb.jhu.edu/software/kraken/dl/minikraken_20171019_8GB.tgz. Source data are provided with this paper.

Code availability

All custom scripts for which github repositories are not specified above can be found at https://github.com/djw533/Serratia_genus_paper/analysis_scripts. Other packages used can be found at https://github.com/djw533/hamburger93, https://github.com/djw533/micro.gen.extra94, and https://github.com/djw533/pathwaytools_gff2gbk95. Rscripts used to plot figures can also be found in this repository at https://github.com/djw533/Serratia_genus_paper/figure_scripts. Rscripts make use of the tidyverse96 collection of packages. R version 4.0.3 was used for all analysis and generation of plots.

References

1. **Rawal, Kamal, et al. (2021)**. "Identification of vaccine targets in pathogens and design of a vaccine using computational approaches." Scientific Reports 11.1. DOI. https://doi.org/10.1038/s41598-021-96863-x

2. **Liu, Lina, et al. (2008)**. "Identification and experimental verification of protective antigens against streptococcus suis serotype 2 based on genome sequence analysis." Current Microbiology 58.1: 11-17. DOI. https://doi.org/10.1007/s00284-008-9258-x
3. **Moriel, Danilo Gomes, et al. (2010)**. "Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic escherichia coli." Proceedings of the National Academy of Sciences 107.20: 9072-9077. DOI. https://doi.org/10.1073/pnas.0915077107
4. **Dey, Supantha, et al. (2022)**. "Functional annotation of hypothetical proteins from the enterobacter cloacae b13 strain and its association with pathogenicity." Bioinformatics and Biology Insights 16. DOI. https://doi.org/10.1177/11779322221115535
5. **Monk, Jonathan M. (2022)**. "Genome-scale metabolic network reconstructions of diverse escherichia strains reveal strain-specific adaptations." Philosophical Transactions of the Royal Society B: Biological Sciences 377.1861. DOI. https://doi.org/10.1098/rstb.2021.0236
6. **Williams, David J., et al. (2022)**. "The genus serratia revisited by genomics." Nature Communications 13.1. DOI. https://doi.org/10.1038/s41467-022-32929-2
7. **Bobay, Louis-Marie, and Howard Ochman. (2018)**. "Factors driving effective population size and pan-genome evolution in bacteria." BMC Evolutionary Biology 18.1. DOI. https://doi.org/10.1186/s12862-018-1272-4
8. **Steinberg, Asher Preska, et al. (2022)**. "Core genes can have higher recombination rates than accessory genes within global microbial populations." eLife 11. DOI. https://doi.org/10.7554/elife.78533
9. **Colquhoun, Rachel M., et al. (2021)**. "Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs." Genome Biology 22.1. DOI. https://doi.org/10.1186/s13059-021-02473-1
10. **Zielke, Ryszard A., et al. (2016)**. "Proteomics-driven antigen discovery for development of vaccines against gonorrhea." Molecular and Cellular Proteomics 15.7: 2338-2355. DOI. https://doi.org/10.1074/mcp.m116.058800
11. **Hernández-Salmerón, Julie E., and Gabriel Moreno-Hagelsieb. (2020)**. "Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2." BMC Genomics 21.1. DOI. https://doi.org/10.1186/s12864-020-07132-6
12. **Moreno-Hagelsieb, G., and K. Latimer. (2007)**. "Choosing BLAST options for better detection of orthologs as reciprocal best hits." Bioinformatics 24.3: 319-324. DOI. https://doi.org/10.1093/bioinformatics/btm585
13. **Buchfink, Benjamin, et al. (2014)**. "Fast and sensitive protein alignment using DIAMOND." Nature Methods 12.1: 59-60. DOI. https://doi.org/10.1038/nmeth.3176
14. **Widmann, Michael, et al. (2010)**. "The isoelectric region of proteins: A systematic analysis." PLoS ONE 5.5: e10546. DOI. https://doi.org/10.1371/journal.pone.0010546
15. **Freitas, Ana I., et al. (2022)**. "Bare silica as an alternative matrix for affinity purification/immobilization of his-tagged proteins." Separation and Purification Technology 286: 120448. DOI. https://doi.org/10.1016/j.seppur.2021.120448

16. Anne Spriestersbach, Jan Kubicek, Frank Schäfer, Helena Block, and Barbara Maertens. "Purification of his-tagged proteins." In Laboratory Methods in Enzymology: Protein Part D, pp. 1-15. Elsevier, 2015. DOI. https://doi.org/10.1016/bs.mie.2014.11.003
17. Sandra Scheiblhofer, Josef Laimer, Yoan Machado, Richard Weiss, and Josef Thalhamer. "Influence of protein fold stability on immunogenicity and its implications for vaccine design." Expert Review of Vaccines, Vol. 16, No. 5, pp. 479-489, March 2017. DOI. https://doi.org/10.1080/14760584.2017.1306441
18. Robert Thai, Gervaise Moine, Michel Desmadril, Denis Servent, Jean-Luc Tarride, André Menez, and Michel Léonetti. "Antigen stability controls antigen presentation." Journal of Biological Chemistry, Vol. 279, No. 48, pp. 50257-50266, November 2004. DOI. https://doi.org/10.1074/jbc.m405738200
19. Joseph F. Zayas. "Solubility of proteins." In Functionality of Proteins in Food, pp. 6-75. Springer Berlin Heidelberg, 1997. DOI. https://doi.org/10.1007/978-3-642-59116-7_2
20. Zhipeng Zhou, Yunkun Dang, Mian Zhou, Lin Li, Chien hung Yu, Jingjing Fu, She Chen, and Yi Liu. "Codon usage is an important determinant of gene expression levels largely through its effects on transcription." Proceedings of the National Academy of Sciences, Vol. 113, No. 41, September 2016. DOI. https://doi.org/10.1073/pnas.1606724113
21. Martin F. Bachmann and Gary T. Jennings. "Vaccine delivery: a matter of size, geometry, kinetics and molecular patterns." Nature Reviews Immunology, Vol. 10, No. 11, pp. 787-796, October 2010. DOI.https://doi.org/10.1038/nri2868
22. Fabian Aviles, Meiman May Zhang, Janlin Chan, Jerome Delamare-Deboutteville, Timothy J. Green, Cecile Dang, and Andrew C. Barnes. "The conserved surface m-protein SiMA of streptococcus iniae is not effective as a cross-protective vaccine against differing capsular serotypes in farmed fish." Veterinary Microbiology, Vol. 162, No. 1, pp. 151-159, February 2013. DOI https://doi.org/10.1016/j.vetmic.2012.08.018
23. Ekaitz Errasti-Murugarren, Paola Bartoccioni, and Manuel Palacín. "Membrane protein stabilization strategies for structural and functional studies." Membranes, Vol. 11, No. 2, Article 155, February 2021. DOI https://doi.org/10.3390/membranes11020155
24. Jose L. Sanchez-Trincado, Marta Gomez-Perosanz, and Pedro A. Reche. "Fundamentals and methods for t- and b-cell epitope prediction." Journal of Immunology Research, 2017, pp. 1-14. DOI https://doi.org/10.1155/2017/2680160
25. Hanover Matz, Danish Munir, James Logue, and Helen Dooley. "The immunoglobulins of cartilaginous fishes." Developmental and Comparative Immunology, Vol. 115, Article 103873, February 2021. DOI. https://doi.org/10.1016/j.dci.2020.103873
26. Christopher A. Beaudoin, Martin Bartas, Adriana Volná, Petr Pečinka, and Tom L. Blundell. "Are there hidden genes in DNA/RNA vaccines?" Frontiers in Immunology, Vol. 13, February 2022. DOI. https://doi.org/10.3389/fimmu.2022.801915
27. Marie Sémon, Dominique Mouchiroud, and Laurent Duret. "Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance." Human Molecular Genetics, Vol. 14, No. 3, pp. 421-427, December 2004. DOI. https://doi.org/10.1093/hmg/ddi038

28. Alexander E. Vinogradov. "Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth." Trends in Genetics, Vol. 21, No. 12, pp. 639-643, December 2005. DOI. https://doi.org/10.1016/j.tig.2005.09.002 .
29. Jasmine E. Bird, Jon Marles-Wright, and Andrea Giachino. "A user's guide to golden gate cloning methods and standards." ACS Synthetic Biology, Vol. 11, No. 11, pp. 3551-3563, November 2022. DOI. https://doi.org/10.1021/acssynbio.2c00355
30. Mark R. Davies et al. "Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics." Nature Genetics, Vol. 51, No. 6, pp. 1035-1043, May 2019. DOI. https://doi.org/10.1038/s41588-019-0417-8

Acknowledgements

Author information

Ethics declarations

Peer review

Additional information

Supplementary information

Source data

Rights and permissions