

# Comparative genomics of *Nocardia seriolae* reveals recent importation and subsequent widespread dissemination in mariculture farms in the South Central Coast region, Vietnam

Cuong T. Le<sup>1,2</sup>, Erin P. Price<sup>1,3</sup>, Derek S. Sarovich<sup>1,3</sup>, Thu T. A. Nguyen<sup>4</sup>, Daniel Powell<sup>1</sup>, Hung Vu-Khac<sup>5</sup>, D. İpek Kurtböke<sup>1</sup>, Wayne Knibb<sup>1</sup>, Shih-Chu Chen<sup>6</sup> and Mohammad Katouli<sup>1,7,\*</sup>

## Abstract

Between 2010 and 2015, nocardiosis outbreaks caused by *Nocardia seriolae* affected many permit farms throughout Vietnam, causing mass fish mortalities. To understand the biology, origin and epidemiology of these outbreaks, 20 *N. seriolae* strains collected from farms in four provinces in the South Central Coast region of Vietnam, along with two Taiwanese strains, were analysed using genetics and genomics. PFGE identified a single cluster amongst all Vietnamese strains that was distinct from the Taiwanese strains. Like the PFGE findings, phylogenomic and SNP genotyping analyses revealed that all Vietnamese *N. seriolae* strains belonged to a single, unique clade. Strains fell into two subclades that differed by 103 SNPs, with almost no diversity within clades (0–5 SNPs). There was no association between geographical origin and subclade placement, suggesting frequent *N. seriolae* transmission between Vietnamese mariculture facilities during the outbreaks. The Vietnamese strains shared a common ancestor with strains from Japan and China, with the closest strain, UTF1 from Japan, differing by just 220 SNPs from the Vietnamese ancestral node. Draft Vietnamese genomes range from 7.55 to 7.96 Mbp in size, have an average G+C content of 68.2% and encode 7602–7958 predicted genes. Several putative virulence factors were identified, including genes associated with host cell adhesion, invasion, intracellular survival, antibiotic and toxic compound resistance, and haemolysin biosynthesis. Our findings provide important new insights into the epidemiology and pathogenicity of *N. seriolae* and will aid future vaccine development and disease management strategies, with the ultimate goal of nocardiosis-free aquaculture.

## DATA SUMMARY

Sequence read files (SRX10462095, SRX10462096, SRX10462097, SRX10462093, SRX10462094, SRX10462092, SRX10462098) and the draft genome assemblies of all seven Vietnamese *Nocardia seriolae* strains are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject PRJNA551736 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA551736> and <https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/14550/>).

## INTRODUCTION

The genus *Trachinotus*, of the family Carangidae, comprises a group of marine, medium-sized, migratory, pelagic finfish that are widely distributed in subtropical and tropical waters worldwide [1, 2]. Many members of the genus, such as *T. carolinus*, *T. blochii*,

Received 01 December 2021; Accepted 16 May 2022; Published 04 July 2022

**Author affiliations:** <sup>1</sup>Centre for Bioinnovation, University of the Sunshine Coast, Sippy Downs, Queensland, Australia; <sup>2</sup>Institute for Aquaculture, Nha Trang University, Nha Trang, Vietnam; <sup>3</sup>Sunshine Coast Health Institute, Birtinya, Queensland, Australia; <sup>4</sup>Institute for Biotechnology and Environment, Nha Trang University, Nha Trang, Vietnam; <sup>5</sup>Central Vietnam Veterinary Institute, Nha Trang, Vietnam; <sup>6</sup>Department of Veterinary Medicine, College of Veterinary Medicine, National Pingtung University of Science and Technology, Pingtung, Taiwan, ROC; <sup>7</sup>School of Science, Technology and Engineering, University of the Sunshine Coast, Sippy Downs, Queensland, Australia.

\*Correspondence: Mohammad Katouli, [mkatouli@usc.edu.au](mailto:mkatouli@usc.edu.au)

**Keywords:** *Nocardia seriolae*; trachinotus; nocardiosis; genomics; aquaculture; permit fish; fish infection; infectious disease; fish mortality.

**Abbreviations:** BHI, Brain–Heart Infusion; HPD, highest posterior density; MRCA, most recent common ancestor; PFGE, pulsed-field gel electrophoresis; SYBR-MAMA, SYBR green-based mismatch amplification mutation assay; UPGMA, unweighted pair group method with arithmetic mean; WGS, whole genome sequencing.

Repositories: The Illumina data and genome assemblies for the seven Vietnamese *N. seriolae* genomes generated in this study are available at NCBI BioProject PRJNA551736.

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary Figures are available with the online version of this article.

000845 © 2022 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

## Impact Statement

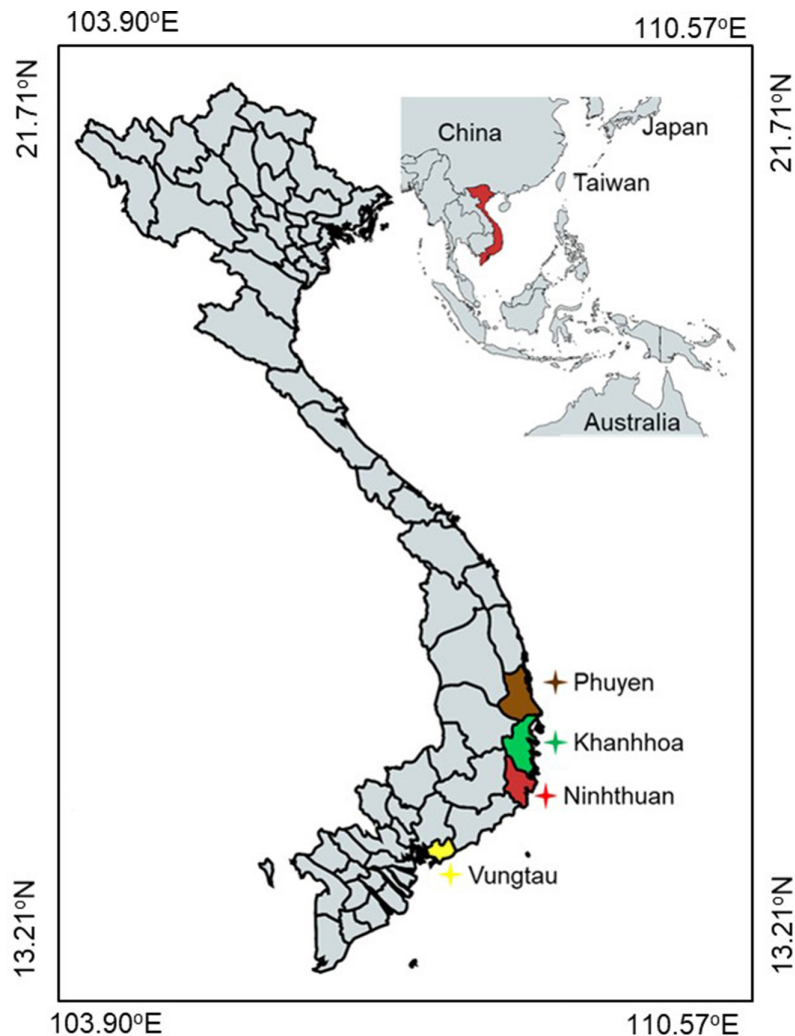
*Nocardia seriolae*, the aetiological agent of a lethal granulomatous disease known as fish nocardiosis, has caused high fish mortalities to global aquaculture sectors in recent decades, particularly in Asia and the Americas. This pathogen possesses a highly conserved genome and minimal genetic diversity, which limits the discriminatory power of existing genotyping techniques such as PFGE, leading to insufficient resolution among genetically related strains. To overcome resolution issues using genotyping methods such as PFGE, we employed whole-genome sequencing (WGS) to create highly resolved time-calibrated phylogenies from all available *N. seriolae* genomes ( $n=20$ ), including seven newly sequenced strains we retrieved from Vietnamese fish farms, where nocardiosis outbreaks are increasingly imposing a significant commercial burden. This comprehensive comparative genomic analysis provides the first global phylogenetic analysis of *N. seriolae* strains, allowing the elucidation of the temporal and spatial dynamics of this pathogen, particularly in Vietnam. Using the comparative genomic data, we developed two SNP-based genotyping assays for differentiating Vietnamese from non-Vietnamese strains, and for distinguishing between the two Vietnamese clades, offering an inexpensive tool for rapidly discriminating and tracing the origin of new nocardiosis outbreaks. Our WGS and SNP assays identified the rapid and undetected spread of *N. seriolae* throughout South Central Coast aquaculture facilities, reflecting the need for better surveillance measures for this emerging pathogen. Finally, our genomic analysis also identified multiple virulence factors and antimicrobial resistance genes, which provide valuable information for better understanding the pathogenicity and persistence of this important aquaculture pathogen.

*T. ovatus*, and *T. falcatus* are of great economic importance for fisheries and aquaculture sectors in America and Asia due to their high-quality meat, fast growth, high market price, and strong adaptability to a variety of captive environments [3–7]. In Asia, the farming of permit fish, particularly the snub nose permit, *T. falcatus*, has commercially taken place in ponds, raceways, and floating sea cages in both brackish and sea waters. Since 2010, Asian mariculture farms have produced over 2 million tonnes of fish meat, significantly contributing to the food security, poverty alleviation, and economic growth of the region [8]. However, the shortage of quality seed stock and the risk of fish disease outbreaks in several countries are key obstacles and challenges for the sector's sustainable development.

*T. falcatus* fingerlings were first imported into Vietnam from Taiwan and China in the 2000s and have quickly gained popularity, with permit fish now the third largest group of commercially cultured marine fish after seabass and grouper. However, high mortality rates of *T. falcatus* weighing between 5 and 350 g (6–45 cm in length) emerged in 2010 during an epizootic event that affected sea cage farms in Khánh Hòa province, in the South Central Coast region of Vietnam. Since this initial outbreak, large-scale outbreaks have occurred at several other farming sites in southern and central parts of the country [9, 10]. Infected fish showed clinical signs of nocardiosis such as lethargy, skin blisters, ulcers, and multiple yellowish to whitish nodules affecting both internal and external organs. Based on analyses of 16S rRNA gene sequences and biochemical characteristics, the bacterial pathogen *Nocardia seriolae* was confirmed as the causative agent [10]; however, the origin of *N. seriolae* affecting Vietnamese permit fish farms has not yet been identified.

*N. seriolae* is a Gram-positive, branching, filamentous intracellular bacterium of the family *Nocardiaceae* that was initially described as *N. kampachi* in farmed yellowtail, *Seriola quinqueradiata* [11], following large outbreaks in Mie Prefecture, Japan. An estimated loss of approximately 260 tonnes of cultured yellowtails due to the disease was recorded in 1989 [12]. Nocardiosis has also impacted several other important fish species within the Japanese aquaculture industry such as amberjack (*Seriola dumerili*), Japanese flounder (*Paralichthys olivaceus*), and chub mackerel (*Scomber japonicas*). *N. seriolae* has subsequently been documented in Taiwan, China, Korea, USA, and Mexico, where high mortalities and associated economic losses due to nocardiosis have been reported in freshwater and marine fish species in both cultured and wild populations [13–23]. Despite causing significant economic losses in fish aquaculture worldwide, there are currently no effective measures against nocardiosis.

Five complete and eight draft *N. seriolae* genome sequences were publicly available prior to our study, representing isolates retrieved from Japan, South Korea, and China [24–28]. These genomes have provided important insights into *N. seriolae* epidemiology, transmission, pathogenesis, and infection control strategies; however, isolates from other nocardiosis-prevalent regions such as Taiwan, USA, Mexico, and Vietnam have not yet been examined, leaving major gaps in our understanding of this devastating infectious disease. In the current study, we sequenced the genomes of seven *N. seriolae* isolates isolated from different permit fish farm locations across Vietnam and compared them with the 13 previously genome-sequenced *N. seriolae* isolates, allowing a comparison of isolates spanning a decade in time and from a variety of sources and geographical locations. Using this information, we developed two novel SNP-based PCR assays to rapidly differentiate Vietnam and non-Vietnam strains, and strains representing the two Vietnamese clades. We also characterized potential virulence factors and antimicrobial/toxin resistance determinants to gain insights into pathogenicity and survival mechanisms. Finally, we functionally annotated our *N. seriolae* genomes to determine whether differences in gene content might contribute to physiological variability among isolates.



**Fig. 1.** Four Vietnamese provinces where *Nocardia seriolae* isolates were collected from infected permit fish (*Trachinotus falcatus*).

## METHODS

### Bacterial strains

Due to a ban on *N. seriolae* culture importation into Australia, all live culture work was carried out in laboratories at the Institute for Aquaculture, Nha Trang University, Vietnam (for Vietnamese strains) and the Department of Veterinary Medicine, College of Veterinary Medicine, National Pingtung University of Science and Technology, Pingtung, Taiwan (for Taiwanese strains).

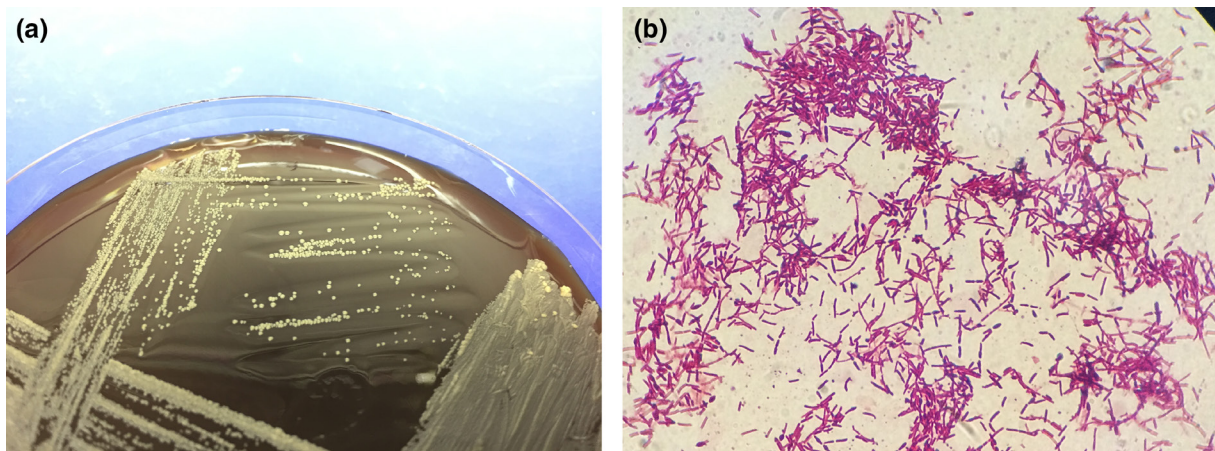
Twenty-two *N. seriolae* strains isolated from fish were examined in this study, comprising 20 from Vietnam and two from Taiwan. Vietnamese strains were isolated from cultured permit fish (*T. falcatus*) (31.0–85.8 g) during nocardiosis outbreaks occurring between 2014 and 2015 in four provinces (Phú Yên, Khánh Hòa, Ninh Thuận, and Vũng Tàu) in the South Central Coast region, and the Taiwanese strains were isolated from largemouth bass (*Micropterus salmoides*) and mullet (*Mugil cephalus*) in 2007 (Fig. 1 and Table 1). Isolates were confirmed as *N. seriolae* based on morphological observations, Ziehl-Neelsen staining (Fig. 2), 16S rRNA gene sequencing, and biochemical characteristics. The 20 Vietnamese strains were subject to PFGE analyses, of which seven isolates were selected for whole-genome sequencing (WGS) to enable more detailed genetic analyses. All 22 isolates were tested using our SNP genotyping assays.

Isolates were preserved in Brain Heart Infusion (BHI; Difco) broth mixed with 25% (v/v) glycerol and stored at  $-80^{\circ}\text{C}$ . For culturing, strains were grown in BHI broth at  $28^{\circ}\text{C}$  for 5 days, with orbital shaking at 150 r.p.m. For DNA extraction, 0.3 ml of bacterial cells were pelleted at  $6000g$  at  $4^{\circ}\text{C}$  for 5 min and washed twice with  $1\times$  sterile PBS. To test for a haemolytic reaction, *N. seriolae* colonies grown in BHI broth were streaked onto 5% (v/v) sheep blood agar and incubated at  $28^{\circ}\text{C}$  for 3 weeks (Fig. 2a).

**Table 1.** *Nocardia seriolae* strains collected in this study, their *AseI* and *XbaI* PFGE profiles, and their SNP genotypes

Country	Strain	Fish species	Host tissue	Origin	Collection date	<i>AseI</i>	<i>XbaI</i>	SNP genotype*
Taiwan	96127	<i>Micropterus salmoides</i>	Unknown	Taiwan	2007	A1	X1	S1
Taiwan	96994	<i>Mugil cephalus</i>	Unknown	Taiwan	2007	A4	X5	S1
Vietnam	KH_11	<i>Trachinotus falcate</i>	Muscle	Khánh Hòa, Vietnam	March 2014	NsA2	NsX3	S2C1
Vietnam	KH_14	<i>Trachinotus falcatus</i>	Spleen	Khánh Hòa, Vietnam	April 2014	NsA1	NsX1	S2C2
Vietnam	KH_15	<i>Trachinotus falcatus</i>	Kidney	Khánh Hòa, Vietnam	May 2014	NsA1	NsX5	S2C1
Vietnam	KH_17	<i>Trachinotus falcatus</i>	Spleen	Khánh Hòa, Vietnam	March 2014	NsA1	NsX3	S2C1
Vietnam	KH_21	<i>Trachinotus falcatus</i>	Kidney	Khánh Hòa, Vietnam	April 2014	NsA2	NsX3	S2C2
Vietnam	NT_01	<i>Trachinotus falcatus</i>	Muscle	Ninh Thuận, Vietnam	April 2014	NsA3	NsX5	S2C2
Vietnam	NT_02	<i>Trachinotus falcatus</i>	Spleen	Ninh Thuận, Vietnam	April 2014	NsA3	NsX2	S2C1
Vietnam	NT_03	<i>Trachinotus falcatus</i>	Liver	Ninh Thuận, Vietnam	April 2014	NsA5	NsX1	S2C2
Vietnam	NT_50	<i>Trachinotus falcatus</i>	Spleen	Ninh Thuận, Vietnam	April 2014	NsA2	NsX3	S2C2
Vietnam	PY_22	<i>Trachinotus falcatus</i>	Spleen	Phú Yên, Vietnam	April 2014	NsA4	NsX1	S2C1
Vietnam	PY_23	<i>Trachinotus falcatus</i>	Muscle	Phú Yên, Vietnam	April 2014	NsA9	NsX1	S2C1
Vietnam	PY_30	<i>Trachinotus falcatus</i>	Liver	Phú Yên, Vietnam	April 2014	NsA8	NsX1	S2C2
Vietnam	PY_31	<i>Trachinotus falcatus</i>	Bone	Phú Yên, Vietnam	April 2014	NsA10	NsX4	S2C1
Vietnam	PY_35	<i>Trachinotus falcatus</i>	Spleen	Phú Yên, Vietnam	April 2014	NsA7	NsX1	S2C2
Vietnam	PY_37	<i>Trachinotus falcatus</i>	Spleen	Phú Yên, Vietnam	April 2014	NsA6	NsX1	S2C2
Vietnam	PY_39	<i>Trachinotus falcatus</i>	Spleen	Phú Yên, Vietnam	April 2014	NsA7	NsX1	S2C2
Vietnam	PY_40	<i>Trachinotus falcatus</i>	Kidney	Phú Yên, Vietnam	April 2014	NsA6	NsX1	S2C1
Vietnam	VT_45	<i>Trachinotus falcatus</i>	Spleen	Vũng Tàu, Vietnam	June 2015	NsA10	NsX3	S2C1
Vietnam	VT_61	<i>Trachinotus falcatus</i>	Spleen	Vũng Tàu, Vietnam	June 2015	NsA11	NsX1	S2C1
Vietnam	VT_62	<i>Trachinotus falcatus</i>	Liver	Vũng Tàu, Vietnam	June 2015	NsA12	NsX1	S2C2

\*S1, non-Vietnamese SNP genotype; S2, Vietnamese SNP genotype; C1, Vietnam Clade 1; C2, Vietnam Clade 2.



**Fig. 2.** Morphology of *Nocardia seriolae* isolated from Vietnam mariculture farms. (a) Chalky white non-haemolytic colonies of *N. seriolae* on sheep blood agar (3-week-old culture); and (b) Ziehl–Neelsen-stained *N. seriolae*, showing purple red, filamentous branching bacteria.



## PFGE typing

PFGE was performed using 50 U *Xba*I or *Ase*I (New England BioLabs) as previously described [20]. The type strain, *N. seriolae* BCRC 13745 (JCM 3360; isolated from the spleen of farmed yellowtail in Nagasaki Prefecture, Japan, ca. 1974), was included for comparative purposes. Gels of DNA fragments were analysed using GelCompar II software version 6.5 (Applied Maths). Gel bands were automatically assigned by the software and were checked and corrected manually. Only clearly resolved bands were considered for further analysis. A dendrogram was constructed using an unweighted pair group method with arithmetic mean (UPGMA) approach and the Dice similarity coefficient, with band optimization and band position tolerances of 1.0%. Isolates that showed similarity between the banding profiles of  $\geq 80\%$  (fewer than six bands of difference) were defined as indistinguishable or clonally related, whereas patterns with  $< 80\%$  similarity (six or more bands of difference) represented different clusters of unrelated strains [29, 30].

## DNA extraction

Total genomic DNA of bacterial isolates was extracted using the Wizard Genomic DNA Purification Kit (Promega) as per the manufacturer's instructions. DNA was checked for sterility and shipped to the University of the Sunshine Coast, Queensland, Australia. The quantity and purity of extracted DNA were assessed using a NanoDrop 2000 (Thermo Scientific) and 1% gel electrophoresis. DNA for Illumina WGS was submitted on dry ice to the Australian Genome Research Facility (AGRF; North Melbourne, VIC, Australia).

## WGS and comparative genomic analyses

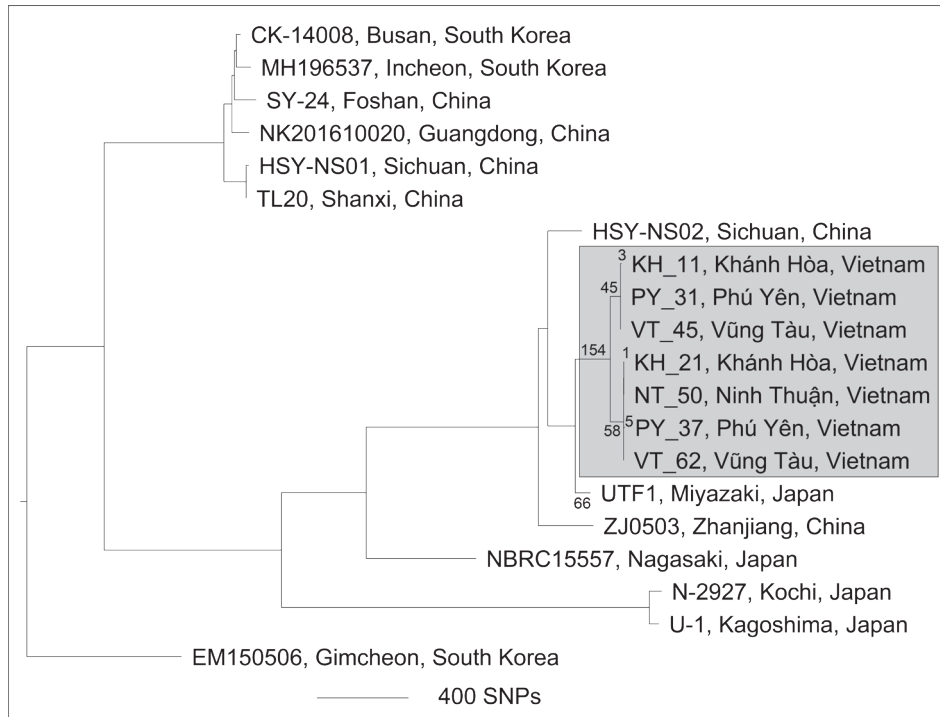
NextEra DNA Flex Illumina libraries for seven Vietnamese *N. seriolae* isolates were sequenced in four lanes of a single flowcell on the NextSeq 500 platform (Illumina), to produce 150 bp paired reads at an average depth of  $\sim 390\times$  (range: 326–433 $\times$ ). Raw read quality was assessed with FastQC v0.11.5 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). These seven genomes are available on the Sequence Read Archive database under BioProject PRJNA551736. Thirteen publicly available genome assemblies (strains EM150506, CK-14008, HSY-NS01, HSY-NS02, MH196537, N-2927, NBRC 15557, NK201610020, SY-24, U-1, UTF1, ZJ0503 and TL20, corresponding to GenBank assembly references ASM186585v1, ASM188553v1, ASM301359v1, ASM366707v1, ASM1411730v1, ASM58371v2, ASM799071v1, ASM1520982v1, ASM209393v1, ASM119293v1, ASM235603v1, ASM76316v1 and ASM1822368v1, respectively) were converted to simulated Illumina reads using ART v2016.06.05 [31] prior to analysis. EM150506, the largest complete *N. seriolae* genome (GenBank accession number CP017839.1) [28], was used as the reference sequence for read mapping and gene annotation. Biallelic, orthologous SNPs from the 20 *N. seriolae* genomes were identified using the default settings of SPANdX v4.0.1 [32], which integrates the Burrows–Wheeler Aligner [33], Sequence Alignment/Map (SAM) tools [34], BEDTools [35], VCFtools [36], Picard Tools (<http://broadinstitute.github.io/picard>) and Genome Analysis Toolkit [37] into a single pipeline.

We performed a hierarchical rooted phylogenetic approach to identify the appropriate root for our *N. seriolae*-only phylogeny (Fig. 3). First, we identified the nearest genetic neighbour to *N. seriolae* via a SPANdX phylogenomic comparison of 134 *Nocardia* species genomes belonging to 78 assigned species and 10 unassigned species (Fig. S1, available in the online version of this article). Next, we reconstructed a rooted phylogeny using the closest relative, *N. concava* NBRC 100430 (RefSeq accession: GCF\_000308815.1) (Fig. S2), to determine the most ancestral *N. seriolae* strain for phylogenetic rooting.

Using the SPANdX SNP matrices, maximum-parsimony phylogenomic trees were reconstructed by Phylogenetic Analysis Using Parsimony (PAUP\*) v4.0a168 software [38], with trees visualized using FigTree v1.4.0 (<http://tree.bio.ed.ac.uk/software/figtree/>). For the *N. seriolae*-only phylogeny, variant annotation was also carried out using SnpEff [39] (Data S1). To determine similarity among *N. seriolae* genomes, and to check for potential rearrangements, contigs in all genome assemblies were oriented and arranged against the reference genome using MAUVE v2.3.1 [40]. BLAST Ring Image Generator (BRIG) [41] was subsequently used to visualize genome relatedness and structural variation. Finally, temporal analysis was performed with BEAST v1.10.4 [42] using the approach outlined by Holt and colleagues [43, 44].

## SNP genotyping

The SPANdX SNP matrix was used to identify SNPs that: (i) distinguished Vietnamese from non-Vietnamese *N. seriolae* strains (220 SNPs; SNP1 assay), and (ii) differentiated the two Vietnamese clades (103 SNPs; SNP2 assay). We selected SNPs at positions 60409 and 587171 in EM150506 for SNP1 and SNP2 assay design, respectively (Data S1). SYBR Green-based mismatch amplification mutation assay (SYBR-MAMA) real-time PCRs were developed to permit rapid genotyping of all strains from this study against these two SNPs. SYBR-MAMA, also known as allele-specific PCR or amplification-refractory mutation system, exploits the differential 3' amplification efficiency of *Taq* polymerase in real time via allele-specific primers targeting each SNP allele at their ultimate 3'-end [45]. SYBR-MAMA has been used for SNP genotyping in many bacteria [46, 47] due to its low cost and simplicity. Each SNP assay consisted of one common primer and two allele-specific primers, matching either the non-Viet allele or the Viet allele for the SNP1 assay, and the Viet Clade 1 allele or Viet Clade 2 allele for the SNP2 assay (Table 2). The same destabilizing mismatch (A for SNP1 and G for SNP2) was incorporated at the penultimate (−2) 3' base of both allele-specific



**Fig. 3.** Rooted maximum-parsimony phylogenomic analysis of seven Vietnamese (KH\_11, KH\_21, NT\_50, PY\_31, PY\_37, VT\_62 and VT\_45; grey box) and 13 non-Vietnamese *Nocardia seriolae* genomes using EM150506 (Fig. S2) as an outgroup. A total of 7343 high-confidence biallelic, orthologous, core-genome SNPs were used to reconstruct the phylogeny. Branch lengths within the Vietnamese clade are labelled and refer to the number of SNPs along each branch. Consistency index=0.998.

primers to increase allele specificity [48]. Cycles-to-threshold ( $C_T$ ) values for each allele-specific reaction were used to determine the SNP genotype for each strain via a change in  $C_T$  value ( $\Delta C_T$ ).

To validate SNP genotypes for our newly developed assays, we first established the reference  $\Delta C_T$  values for each assay by running against the two Taiwanese and seven genome-sequenced Vietnamese strains. Assays were then tested against the 13 remaining Vietnamese isolates to determine their genotypes. For each PCR run, control DNA samples representing the matching and non-matching allele genotypes were used as positive controls, and at least two no-template controls were included.

SYBR-MAMAs contained 1  $\mu$ l of target DNA template at  $\sim 1$  ng  $\mu$ l<sup>-1</sup>, 0.2  $\mu$ M allele-specific primer, 0.2  $\mu$ M common primer (Macrogen), 1 $\times$  Platinum SYBR Green qPCR SuperMix-UDG (cat. no. 11733038; Thermo Fisher Scientific) and RNase/DNase-free PCR-grade water (Cat No. 10977015; Thermo Fisher Scientific), to a total reaction volume of 5  $\mu$ l. Thermocycling conditions

**Table 2.** Primer sequences of SYBR-MAMAs designed in this study for the differentiation of Vietnamese *Nocardia seriolae* strains

SNP assay and target	SNP position*	Variation (allele base)	Primer name	Primer sequence†
SNP1 (Vietnam vs. non-Vietnam strains)	60409	C/T	CtS1_nonViet_For	CAAACCGGCTGGATATCGaC
			CtS1_Viet_For	CAAACCGGCTGGATATCGa <b>T</b>
			SNP1_Rev	CACGCCGACGCTAGTACCTG
SNP2 (Vietnam subclades 1 vs. 2)	587171	A/C	CtS2_Clade1_Rev	CATACCGACTTCCAGGTGTG <b>T</b>
			CtS2_Clade2_Rev	ACCGACTTCCAGGTGTG <b>G</b>
			SNP2_For	AGCCCATTAGCAGTCGTGTA

\*SNP position as per *N. seriolae* EM150506 [28] (GenBank reference CP017839.1).

†Single 3' penultimate mismatch bases are shown in lowercase; SNP-specific nucleotides are indicated in bold. SYBR-MAMA, SYBR Green-based mismatch amplification mutation assay.;

comprised an initial 2 min denaturation at 95 °C, followed by 45 cycles of 95 °C for 15 s and 60 °C for 15 s. All samples were run in duplicate.

### Genome assembly and annotation

Assemblies of the seven Vietnamese *N. seriolae* genomes were constructed from the quality-filtered sequence data using the Microbial Genome Assembly Pipeline (MGAP) v1.1 (<https://github.com/dsarov/MGAP---Microbial-Genome-Assembler-Pipeline>) and EM150506 (GenBank reference CP017839.1) as the scaffolding reference. MGAP wraps Trimmomatic [49], Velvet [50], VelvetOptimiser (<https://github.com/tseemann/VelvetOptimiser>), ABACAS [51], IMAGE [52], SSPACE [53, 54], GapFiller [55, 56] and Pilon [57] into a single tool. Assemblies were primarily annotated using the Rapid Annotations using Subsystems Technology (RAST) server v2.0 with SEED data with default features (RAST annotation scheme: RASTtk, automatically fix errors, fix frameshifts, build metabolic model, backfill gaps, turn on debug, verbose level: 0, and disable replication: yes). RAST was also used to group genes into functional subsystems (akin to Clusters of Orthologous Groups). Annotated genomes were then compared with results provided by Prokka v1.8 [58]. In cases where aberrant results arose between the two tools, the functional prediction of RAST was checked and manually corrected by using BLASTP to search for similar proteins in the UniProtKB database (<http://www.uniprot.org/blast/>). The clustered regularly interspaced short palindromic repeat (CRISPR)-Cas region finder program (<https://crisprcas.i2bc.paris-saclay.fr>) was used to identify regular repeats and the intervening spacer sequences [59]. The assembled genomes for all Vietnamese strains are available from NCBI under BioProject PRJNA551736 (Table 3).

### Virulence and antimicrobial resistance profile determination

The identification of antimicrobial resistance- and virulence-related genes among the Vietnamese *N. seriolae* genomes were performed using RAST and the Virulence Factor Database (VFDB), Victors and PATRIC Virulence Factor (VF) databases available on the Pathosystems Resource Integration Center (PATRIC) [60, 61]. In addition, homologues of experimentally verified pathogenicity determinants within other members of the genus *Nocardia* were searched for in the *N. seriolae* genomes.

## RESULTS

### PFGE genotypes

Twenty *N. seriolae* isolates from four Vietnamese coastal provinces (Fig. 1) were subjected to *Xba*I and *Ase*I digestion to determine isolate relatedness across provinces. Restriction fragment sizes ranged from 40 kb to 1.1 Mbp. PFGE with *Xba*I alone resulted in between 19 and 21 restriction fragments among the Vietnamese strains; similarly, between 16 and 20 fragments were identified using *Ase*I. Seven distinct patterns (labelled as pulsotypes NsX1–NsX7) were present using *Xba*I-digested DNA fragments, and ten patterns (labelled as pulsotypes NsA1–NsA10) for *Ase*I. Using the ≥80% similarity cut-off and ‘fewer than six bands of difference’ Tenover criteria, only one cluster was identified for each enzyme [29, 30]. Even when combining data from both enzymes, the 20 Vietnamese isolates were still closely related, irrespective of their geographical origin, as shown by their categorization into a single cluster that was distinct from the Japanese type strain (Fig. 4).

### Phylogenomic analysis

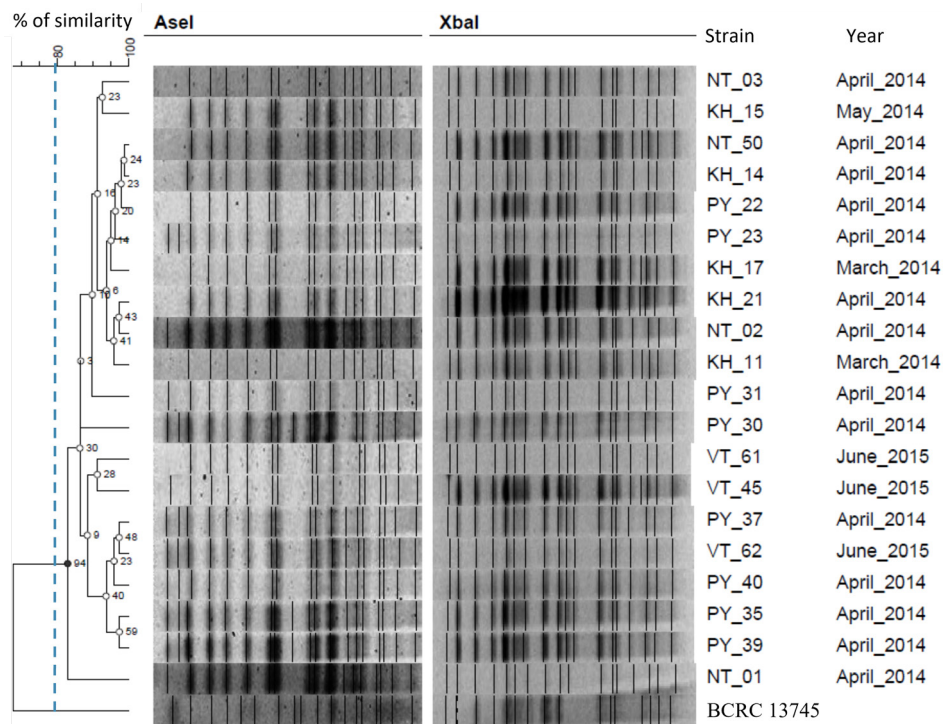
Based on the PFGE results, seven geographically diverse Vietnamese isolates were Illumina-sequenced, resulting in high-coverage draft genomes (Table 3). These genomic data were generated to address two questions: (i) whether comparative genomics, as with PFGE, would reveal minimal genetic diversity among the Vietnamese *N. seriolae* strains, and (ii) whether phylogenomic analysis could identify a potential origin for nocardiosis in Vietnamese aquaculture facilities. The seven Vietnamese genomes generated in this study, plus the sequences of 13 publicly available *N. seriolae* strains (all from other Asian countries), were compared to identify phylogenetically informative SNPs. A total of 8206 SNPs were identified; 7517 (91.6%) were located in coding regions and comprised 126 nonsense, 5163 missense and 1531 silent variants. Of the 8206 SNPs, 7275 high-confidence, orthologous, core genome, biallelic SNPs were identified among the 20 *N. seriolae* strains; these SNPs were used for phylogenomic reconstruction.

The phylogenomic dendrogram revealed five distinct strain clusters (Fig. 3). As with PFGE, the seven Vietnamese isolates were highly clonal, with all strains clustering into a single unique ‘Vietnamese’ clade. Within this clade were two subclades that differed by 103 SNPs. These subclade SNPs were well distributed across the genome, with no evidence of SNP clusters due to recombination. The phylogenomic analysis also suggested that *N. seriolae* undergoes very little, if any, recombination, as demonstrated by a very high consistency index of 0.997; in other words, homoplastic SNP characters, which are more common following recombination events [62], were essentially absent. Within the two Vietnamese subclades, isolates were virtually identical (0–5 SNPs), indicating limited genomic alterations among these lineages (Fig. 3). Notably, there was no link between geographical region and subclade placement, with strains from Phú Yên, Khánh Hòa and Vũng Tàu falling into both Vietnamese subclades, indicating frequent *N. seriolae* transmission events between regions. The most recent common ancestor of the Vietnamese strains

**Table 3.** Genetic and genomic features of the Vietnamese *Nocardia seriolae* strains compared with the South Korean EM150506 strain according to RAST

Strains/ feature	Country	Genome size (Mbp)	Level of completion	Sequencing platform	Sequencing depth	GC%	L50 (bp)	Total no. of proteins	No. of RNAs	No. of hypothetical proteins	No. of proteins with function prediction	No. of proteins assigned to subsystem	NCBI accession no.
KH_11	Vietnam	7.66	Draft	NextSeq 500	340x	68.3	90	7655	58	3560	4465	2055	WMIKE000000000.1
KH_21	Vietnam	7.72	Draft	NextSeq 500	424x	68.2	58	7657	66	3597	4428	2033	WMKFO0000000.1
NT_50	Vietnam	7.96	Draft	NextSeq 500	395x	68.2	86	7640	66	3571	4437	2063	WMKKG0000000.1
PY_31	Vietnam	7.68	Draft	NextSeq 500	408x	68.3	62	7602	62	3212	4818	2220	WMKC000000000.1
PY_37	Vietnam	7.55	Draft	NextSeq 500	326x	68.3	126	7707	51	3549	4525	2087	WMKD000000000.1
VT_45	Vietnam	7.94	Draft	NextSeq 500	404x	68.2	70	7958	67	3609	4718	2054	WMKKB0000000.1
VT_62	Vietnam	7.7	Draft	NextSeq 500	433x	68.3	62	7643	63	3580	4428	2052	WMKH000000000.1
UTF1	Japan	8.12	Complete	PacBio	133x	68.1	1	7890	75	3572	4683	2219	AP017900.1
U-1	Japan	7.77	Draft	Roche 454; MiSeq	179x	68.3	56	7757	69	3645	4497	2291	BBYQ000000000.1
N-2927	Japan	7.76	Draft	Roche 454	160x	68.3	54	7627	66	3225	4841	2245	BAWD000000000.2
NBRC15557	Japan	7.61	Draft	Roche 454; HiSeq 1000	112x	68.3	51	7527	64	3190	4768	2211	NZ_BJWY010000001.1
SY-24	China	7.89	Draft	MiSeq	100x	68.2	52	7632	66	3227	4845	2230	MVAC000000000.1
NK201610020	China	8.31	Complete	HiSeq; PacBio	100x	68.1	1	8133	78	3398	5185	2306	NZ_CP063662.1
HSY-NS01	China	7.91	Draft	HiSeq	126x	68.2	50	7947	70	3727	4605	2133	PXZE000000000.1
HSY-NS02	China	7.76	Draft	HiSeq	110x	68.2	51	7801	69	3301	4932	2225	RCNK000000000.1
ZJ0503	China	7.71	Draft	MiSeq	100x	68.3	50	7579	66	3212	4798	2204	JNCT000000000.1
TL20	China	8.3	Complete	PacBio	200x	68.1	1	7710	66	3212	4798	2204	GCA_018223685.1
CK-14008	Korea	8.37	Draft	PacBio	139x	68.1	1	8212	78	3422	5244	2347	MOYO000000000.1
MHI196537	Korea	8.26	Complete	PacBio	118x	68.1	1	8074	78	3368	5155	2296	CP059737.1
EM150506	Korea	8.3	Complete	PacBio	156x	68.1	1	8068	77	3338	5175	2277	CP017839.1





**Fig. 4.** PFGE dendrogram of *AseI*- and *XbaI*-digested genomic DNA from 20 representative *Nocardia seriolae* strains collected in four Vietnamese provinces. A type strain, BCRC 13745 (Japan), was included for comparison. Cluster analysis of genetic distances was performed using the Dice coefficient and UPGMA method (tolerance and optimization 1%). Two pulsotypes were identified based on an 80% similarity cut-off. Numbers at tree nodes indicate the percentage of replicate trees in which the same clusters were found after 1000 bootstrap replicates.

differed by 220 SNPs from the next closest known strain, UTF1, which was isolated from cultured yellowtail that succumbed to nocardiosis in 2008 in Miyazaki Prefecture, Japan [27].

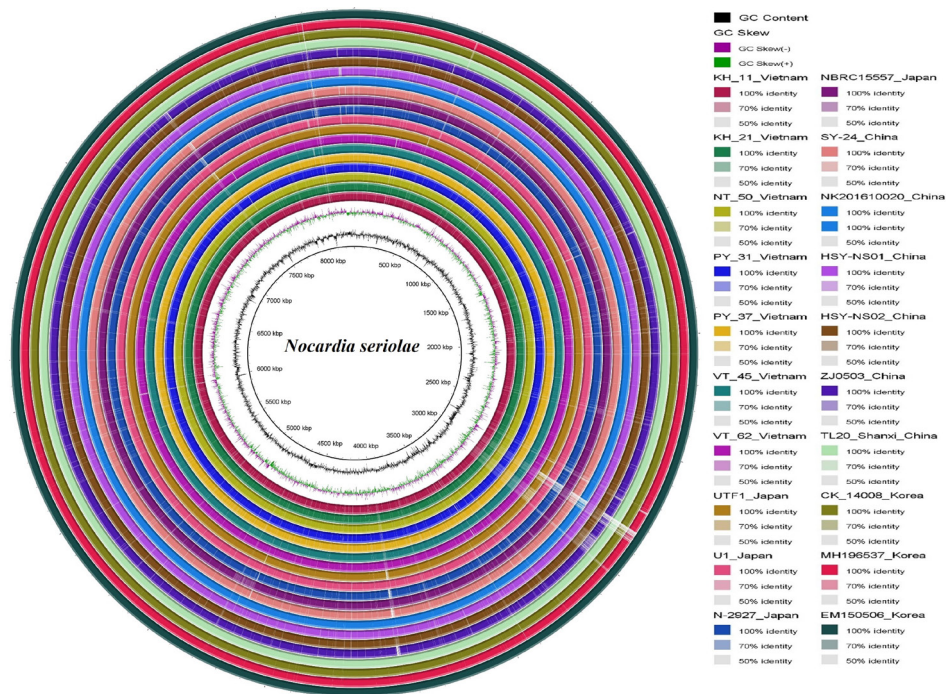
BEAST analysis (Fig. S3) showed that the most recent common ancestor (MRCA) for the Vietnamese and Japanese *N. seriolae* strains occurred in ~1998 [95% highest posterior density (HPD): 1997–1999], and all Vietnam *N. seriolae* strains shared an MRCA in 2001 (95% HPD: 1999–2003).

### SNP genotyping

SYBR-MAMAs demonstrated clear distinction of SNP genotypes. For the SNP1 assay, the two Taiwanese strains amplified the non-Viet allele earlier than the Viet allele ( $\Delta C_T$  range: 2.8–5.5); in contrast, all Vietnamese strains amplified the Viet allele earlier than the non-Viet allele ( $\Delta C_T$  range: 6.0–9.3). For the SNP2 assay, 10 Vietnamese strains belonging to Clade 1 amplified the Clade 1 allele earlier than the Clade 2 allele ( $\Delta C_T$  range: 9.9–13.4), whereas 10 Clade 2 strains amplified the Clade 2 allele earlier ( $\Delta C_T$  range: 4.5–8.1) (Table 1). No amplification was observed for the no-template controls.

### Genome assembly and functional annotation

To gain deeper insights into the seven Vietnamese *N. seriolae* genomes, we conducted a comparative analysis of genome assembly metrics and gene function. The Vietnamese genomes possess 6937 core genes and encode 1–6 rRNA genes and 49–63 transfer RNA genes. Total assembly length ranged from 7.55 to 7.96 Mbp, smaller than the closed genomes EM150506 (8.30 Mbp), MH196537 (8.26 Mbp), UTF1 (8.12 Mbp), and draft genomes reported for CK-14008 (8.37 Mbp) and NK201610020 (8.31 Mbp), but similar to other draft genomes of this species (range: 7.61 to 7.91 Mbp). GC content (68.2–68.3%) was comparable to previously sequenced *N. seriolae* genomes (Table 3). Multiple genome alignment of all strains against EM150506 using BRIG showed a high degree of homology (Fig. 5), demonstrating high conservation among *N. seriolae* genomes. There were four main non-homologous regions (positions 2700000–3100000, 3900000–4100000, 7500000–7600000 and 8000000–8200000 bp) that were present in the reference genome but absent in all other genomes; these regions may indeed be absent or may simply reflect differences in assembly quality [5]. Most genes at these loci were classified as hypothetical proteins, mobile element proteins and repeat regions; the remaining loci are mainly genes involved in membrane transport, biosynthesis, metabolism and transcription (Data S2).



**Fig. 5.** Whole-genome comparison of *Nocardia seriolae* strains from Vietnam and other Asian countries against the EM150506 (South Korean) reference genome using the circular BLASTn alignment in BLAST Ring Image Generator [39]. The innermost circle shows genome scale (bp), the black irregular ring represents %GC content, and the irregular purple/green ring represents %GC skew. Outer colour rings (innermost first) represent Vietnamese strains (KH\_11, KH\_21, NT\_50, PY\_31, PY\_37, VT\_45, VT\_62) and 13 strains from Japan, China, and South Korea. The outermost circle (dark green) represents the EM150506 reference genome.

RAST predicted between 7602 and 7958 coding DNA sequences in the Vietnamese *N. seriolae* genomes, of which 45.8% (range: 42.2–47.0%) are of unknown function ('hypothetical proteins'). Of the 59.1% (range: 57.8–63.4%) coding DNA sequences with RAST function predictions, 45.8% (range: 43.5–50.9%) grouped into 308–330 functional subsystems belonging to 24 protein family categories. These predictions are similar to the previously reported *N. seriolae* genomes (Table 4). Little difference was found in the number of genes in family categories among Vietnamese vs. non-Vietnamese strains (Table 4). No plasmids were identified in any of the Vietnamese genomes, consistent with most *N. seriolae* genomes lacking plasmids; the only exception is CK-14008 from South Korea, which potentially harbours two plasmids [28].

A typical CRISPR-Cas system contains both a CRISPR array of repeat and spacer units, and associated *cas* genes; however, many systems are devoid of one of these components. These atypical CRISPR configurations are known as 'orphan' or 'isolated' CRISPR arrays and *cas* loci depending on which component is lacking. Between three and six CRISPR arrays were found in the Vietnamese strains, with lengths varying from 73 to 114 bp. Each array is made up of two direct repeats and one spacer without nearby Cas (CRISPR-associated) genes (Data S3). Notably, the same CRISPR array structure was found in all 20 *N. seriolae* genomes.

### Virulence and antimicrobial/toxin resistance profiles

To explore the pathogenic potential of the Vietnamese *N. seriolae* strains, we assessed their virulence and antimicrobial/toxin resistance gene content in comparison to non-Vietnamese genomes. The RAST, VFDB, Victors and VF databases found between 182 and 202 genes that encode virulence and resistance factors, including gene products associated with Adherence ( $n=50-54$ ), Cellular metabolism and nutrient uptake ( $n=10$ ), Damage ( $n=6-7$ ), Invasion and intracellular survival ( $n=33-36$ ), Resistance to antibiotics and toxic compounds ( $n=65-81$ ), and Other ( $n=16-18$ ) (Data S3). In general, virulence factors and antimicrobial/toxin resistance factors were almost identical in number among the Vietnamese strains and were comparable to non-Vietnamese strains. However, some genes were absent in most Vietnamese strains but present in most non-Vietnamese strains, such as 'MCE-family protein Mce1D', 'MCE-family protein Mce1F', 'Chromate transport protein ChrA', 'NAD(P)H oxidoreductase YRKL (EC 1.6.99.-) Putative NADPH-quinone reductase (modulator of drug activity B) Flavodoxin 2' and 'Tellurite resistance protein TerB'. In contrast, 'Hemolysins and related proteins containing cystathionine- $\beta$ -synthase domains' was found only in EM150506. Several experimentally verified virulence factors identified in *N. seriolae* and other *Nocardia* species, including catalase, superoxide

**Table 4.** Number of genes for each *Nocardia seriolae* strain associated with the 24 general Clusters of Orthologous Groups functional categories predicted by RAST

Functional category	KH_11	KHL_21	NTL_50	PY_31	PY_37	VT_45	VT_62	UTF1	U-1	N-2927	NBRC 15557	SY-24	NK 201610020	HSY- NS01	HSY- NS02	ZM0503	TL20	CK- 14008	MHI96537	EMI50506
Cofactors, Vitamins, Prosthetic Groups, Pigments	198	195	196	207	206	195	194	204	211	208	209	204	210	199	205	202	198	212	209	208
Cell Wall and Capsule	32	31	31	36	31	31	31	36	36	36	36	34	36	31	36	36	31	38	36	36
Virulence, Disease and Defence	50	47	48	56	50	53	47	55	58	59	55	57	58	49	55	55	49	60	59	62
Potassium metabolism	10	10	10	11	10	11	10	11	10	11	10	11	10	10	10	10	10	11	12	10
Miscellaneous	30	30	30	33	33	30	30	33	32	32	32	32	32	29	33	33	29	32	32	31
Phages, Prophages, Transposable elements, Plasmids	7	5	5	13	6	5	7	10	16	12	8	15	16	11	12	11	9	17	16	10
Membrane Transport	31	31	31	35	31	31	31	35	37	37	37	37	37	32	35	35	32	37	37	36
Iron acquisition and metabolism	14	14	14	15	14	14	14	15	14	15	15	15	15	14	15	15	14	15	15	15
RNA metabolism	56	58	58	59	56	60	58	61	58	59	57	58	62	58	59	56	60	63	62	62
Nucleosides and Nucleotides	96	96	96	107	98	95	97	101	100	100	106	99	101	95	106	101	95	103	101	100
Protein Metabolism	219	224	225	228	212	229	221	242	238	234	233	233	246	229	236	230	237	248	246	248
Regulation and Cell signalling	23	23	23	26	23	23	23	26	26	26	26	26	26	23	27	26	24	26	26	26
Secondary metabolism	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
DNA metabolism	100	99	100	100	105	101	99	102	101	101	100	102	101	99	101	102	101	105	101	100
Fatty Acids, Lipids and Isoprenoids	226	219	243	274	229	223	239	272	310	275	273	273	311	280	273	270	281	319	308	304
Nitrogen Metabolism	32	32	32	35	32	32	32	35	36	36	28	36	35	33	35	35	33	35	36	36
Dormancy and Sporulation	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Respiration	101	100	100	104	107	103	99	103	104	103	77	102	103	99	104	104	98	104	104	104
Stress Response	56	54	55	59	55	56	54	58	58	61	58	61	58	54	60	60	52	59	57	57
Metabolism of Aromatic Compounds	26	26	26	32	27	27	27	32	33	32	32	33	33	26	33	33	27	32	33	34
Amino Acids and Derivatives	365	369	369	391	371	365	367	394	411	406	414	404	415	387	392	392	385	417	412	399
Sulphur Metabolism	14	13	14	13	16	13	14	12	12	14	12	14	13	14	13	14	13	13	13	13
Phosphorus Metabolism	27	27	26	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
Carbohydrates	337	325	326	354	343	325	326	350	358	356	361	352	356	329	353	352	329	369	349	354

dismutase, phospholipase C and protease [63], were present in all Vietnamese and non-Vietnamese strains, indicating that they are highly conserved genes within this genus.

## DISCUSSION

PFGE has conventionally been considered the 'gold standard' for studying the genetic diversity of many different pathogenic bacteria species, including *N. seriolae* [19, 20, 30, 64]. PFGE has previously identified multiple pulsotypes among isolates retrieved from fish in Japan and Taiwan [19, 20]. Notably, one study identified identical pulsotypes between certain Taiwanese 1997–2007 outbreak strains and Japanese *N. seriolae* isolated from yellowtail in 2002 (pulsotypes X1 and A1) and 2005 (pulsotype X11) [20], suggesting at least two transmission events between Taiwan and Japan. Unlike *N. seriolae* from Japan and Taiwan, all 20 Vietnamese isolates fell into a single cluster, even when using a combination of *Xba*I and *Ase*I. However, PFGE lacked the resolution to differentiate Vietnamese isolates into the two clades identified using phylogenomic analysis. This limited resolution has also been documented for other bacteria such as *Salmonella enterica* [65], *Listeria monocytogenes* [66] and *Escherichia coli* [67]. It was unfortunately not practical to compare the Vietnamese pulsotypes with published studies due to known challenges with interlaboratory standardization using PFGE [68]; therefore, it is not known whether the Vietnamese PFGE cluster has been previously reported.

Next-generation sequencing provides excellent resolution, accuracy and data portability, and as such, has begun replacing PFGE as the new gold standard for nocardiosis outbreak analyses [69]. To illustrate the value of WGS for nocardiosis epidemiological investigations, we sequenced seven representative Vietnamese *N. seriolae* strains and compared them with all publicly available genomes ( $n=13$ ). Like PFGE, the limited genomic variation (0–5 SNPs; Fig. 3) observed among Vietnamese strains confirms a recent, single introduction into Vietnam, with subsequent dissemination across multiple mariculture facilities within the South Central Coast region. Phylogenomic analysis showed that Vietnamese strains were most closely related to UTF1, which was isolated from farmed yellowtail in Japan in 2008 [27]; this strain differed from the Vietnamese common ancestor by just 220 SNPs (MRCA: ~1998). Shimahara and colleagues [20] have previously postulated that transboundary translocation of live fish stocks asymptotically infected with *N. seriolae* from China and Hong Kong may have introduced new strains into Japan. Wild-caught amberjack juveniles, one of the most susceptible host species for *N. seriolae* infection, was also reportedly imported into Japan from Vietnam in 2000 [70]. However, there has not yet been a case of nocardiosis reported in Vietnam in other aquatic species besides *Trachinotus* species, and the first of these cases were only recorded in 2012 [9]; therefore, it is unlikely that the Japanese *N. seriolae* was introduced from amberjack imported from Vietnam. Based on our genomic analysis, it is plausible that *N. seriolae* from Japan has been introduced into other countries such as Vietnam given that international export of valuable aquaculture fish species is relatively common; however, there is a paucity of information about import–export of live fish stocks from Japan or Vietnam, and, as such, this hypothesis cannot be confirmed.

Our BEAST results (Fig. S3) add further to our hypothesis of a recent introduction of *N. seriolae* into Vietnam from infected *Trachinotus* species. Our analysis showed that *N. seriolae* introduction into Vietnam occurred in ~2001 (95% HPD: 1999–2003), which fits with the Taiwanese/Japanese outbreaks occurring in the late 1990s and early 2000s. We unfortunately lack isolate data from Taiwan that could suggest the directionality of transfer, or that could provide more accurate source attribution; nevertheless, we have been able to make some interesting and useful insights into the evolutionary history of *N. seriolae* in Vietnam based on this dated phylogeny.

Whilst our results suggest a probable Asian origin for the Vietnamese outbreaks, there are few publicly available *N. seriolae* genomes (only 20 as of 11 February 2022, including seven from our study), and none from other Asian regions such as Taiwan [20], Singapore, Malaysia, or Indonesia [71], or non-Asian regions such as Mexico [23] and USA [21] where *N. seriolae* outbreaks have been documented; therefore, the precise origin of the Vietnamese outbreaks and mode of *N. seriolae* introduction currently remain unresolved. Concerningly, our results, and those of others, demonstrate that, unchecked, *N. seriolae* transmission may represent a substantial unmitigated risk to fish aquaculture. It is thus an utmost imperative to establish domestic and international monitoring processes for *N. seriolae* for both farmed and wild species, including the implementation of molecular methods to characterize new outbreaks, to prevent the spread of this devastating pathogen into new environments, and associated heavy economic losses and food security concerns.

To facilitate the rapid identification of *N. seriolae* genotypes among our Vietnamese strains, we designed inexpensive SYBR-MAMAs targeting two phylogenetically informative SNPs. The first SNP assay robustly differentiates Vietnam from non-Vietnamese strains, thereby permitting prospective identification of newly transmitted strains into Vietnam, an essential facet in future fish importation biocontrol efforts. This assay can also be used to monitor for the emergence of Vietnamese strains in new regions, such as new aquaculture facilities in Vietnam, or prior to export of fingerlings to other countries. The second SNP assay rapidly differentiates strains belonging to the two Vietnamese clades. By applying this second assay to the 20 Vietnamese strains, we observed that both clades were well disseminated across all four provinces: Khánh Hòa, Ninh Thuận, Phú Yên and Vũng Tàu. Phylogenomic analysis of seven representative Vietnamese strains also showed dispersal of these two clades among three of the four provinces. Although unconfirmed, it is probable



that the widespread trade of eggs, fingerlings and live permit fish for aquaculture in Vietnam since industry inception in the early 2000s, including local unmonitored trade among fish farmers, has driven the successful dissemination of *N. seriolae* among Vietnamese permit farms. Taken together, our findings highlight the large risk of undetected *N. seriolae* dispersal among mariculture facilities and the need for establishing strict monitoring practices to prevent further pathogen transmission.

WGS is currently laborious, expensive and inaccessible to most laboratories in Vietnam and many other Asian countries. Using comparative genomics, we established a catalogue of SNPs specific to each clade and subclade. This SNP database may be useful for both targeted resequencing efforts and the design of phylogenetically robust genotyping methods to permit source tracing of future *N. seriolae* outbreaks without the requirement for further WGS or bioinformatic analyses. The SYBR-MAMAs developed in this study successfully detected two phylogenetically informative SNPs, with genotyping results fully concordant with WGS, confirming that SYBR-MAMA is a valuable and inexpensive diagnostic method for SNP characterization.

Very little is known about the pathogenesis of *Nocardia* species, which are capable of invading host macrophages and preventing the fusion of phagosomes with lysosomes, leading to long-term survival and proliferation in host cells [72]. Due to the paucity of available genomic data for this pathogen, a final aspect of this study was to better understand virulence and antimicrobial resistance factors encoded by the *N. seriolae* genome. Our analysis of 2020 *N. seriolae* genomes is the largest genomic assessment of this pathogen to date, and largely corroborates the conclusions drawn from a previous analysis of seven *N. seriolae* genomes, which showed that *N. seriolae* have >99.9% Orthologous Average Nucleotide Identity values [28]. Analysis of the genome content of seven Vietnamese *N. seriolae* strains revealed that, like non-Vietnamese strains, they encode a high proportion of 'hypothetical protein' genes (i.e. 45.8%), a finding that highlights the need for more studies to investigate the functions of these genes. More than 180 core genes (present in all strains) were found to code for antimicrobial resistance and virulence factors in the Vietnamese strains, including genes associated with Adherence ( $n=49$ ), Cellular metabolism and nutrient uptake ( $n=10$ ), Damage ( $n=6$ ), Invasion and intracellular survival ( $n=33$ ), Resistance to antibiotics and toxic compounds ( $n=26$ ), and Others ( $n=11$ ) that may possibly account for the main virulence traits of this fish pathogen. The presence of conserved genes encoding  $\beta$ -lactamase class C-like and penicillin-binding proteins ( $n=11$ ), multidrug resistance protein ErmB ( $n=1$ ), probable multidrug resistance protein NorM ( $n=1$ ) and a small multidrug resistance family protein ( $n=1$ ) in all *N. seriolae* genomes may explain observed antimicrobial resistance towards penicillin and cephalexin, two  $\beta$ -lactam antibiotics that are commonly used to treat nocardiosis in Vietnamese permit fish farms (data not shown).

CRISPRs, which are encoded by many bacterial and archaeal species, defend against invasive mobile genetic elements such as viral or plasmid DNA [73], and also play a role in bacterial pathogenesis, biofilm formation, adherence, programmed cell death and quorum sensing [74]. Acquisition and maintenance of CRISPR-Cas systems are greatly influenced by environmental conditions and microbial communities [75]. Recent research has shown that 40% of CRISPR-Cas loci are away from any associated *cas* genes or are not associated with *cas* genes, which are known as orphan CRISPR arrays [76]. Like many other bacterial species such as *Listeria monocytogenes*, *Aggregatibacter actinomycetemcomitans*, *Enterococcus faecalis*, *Staphylococcus* spp., *Pseudomonas aeruginosa* and *Salmonella enterica* [77–81], orphan CRISPR arrays were found in *N. seriolae* genomes. These incomplete CRISPR-Cas systems may be a remnant of decaying loci that are recruited and/or selectively maintained to perform important, but as yet unknown, biological functions [73]. Alternatively, our results may be an artefact of current CRISPR-Cas prediction tools, which predict the CRISPRs primarily based on the typical CRISPR structure [77]. As the role of these CRISPR loci in *N. seriolae* is not yet known, further work is needed to uncover their precise role in this pathogen.

In conclusion, our study provides novel insights into the epidemiology of *N. seriolae* outbreaks in farmed permit fish in Vietnam. Our detailed molecular and genomic analyses revealed minimal genomic diversity among Vietnamese *N. seriolae* isolates. Unlike PFGE, WGS detected strain variation at single-base resolution, and identified two distinct Vietnamese clades that share recent ancestry. Our results indicate recent importation of a single *N. seriolae* clone into Vietnam, which has then led to a nationwide outbreak of nocardiosis in permit fish farms. The analysis of additional genomes, particularly from other geographical regions, will be important for better understanding *N. seriolae* evolution, and will enable more precise investigations into the origin and transmission of this devastating pathogen. Finally, our SNP assays provide a rapid and inexpensive method for genotyping of ongoing and future nocardiosis outbreaks in Vietnam.

---

#### Funding information

This research was supported by an Australia Awards PhD scholarship to C.L., which is funded by the Australian Department of Foreign Affairs and Trade. D.S.S. and E.P.P. were supported by Advance Queensland fellowships (AQR13016-17RD2 and AQIRF0362018, respectively)

#### Acknowledgements

We gratefully acknowledge the financial support and laboratory facilities provided by the Centre for Bioinnovation, the University of the Sunshine Coast, and Nha Trang University.



**Author contributions**

C.L.: Project design, sample collection, sample and data analysis, results interpretation, drafting paper. D.S.S.: Data analyses and interpretation, drafting and revising paper. E.P.P.: Supervision, data analyses and interpretation, drafting and revising paper. T.T.A.N.: Assistance in the sample preparation and drafting paper. D.P.: Sample collection guidance, drafting and revising paper. H.V.-K.: Sample collection guidance, drafting and revising paper. I.D.K.: Assisting with the project design, revising paper. W.K.: Supervision, advising on project design, drafting paper. S.-C.C.: Assistance in PFGE analyses, drafting paper. M.K.: Supervision, project design, revising paper. All authors read and approved the final manuscript.

**Conflicts of interest**

The authors have no competing interests to declare.

**References**

- Berry F, Iversen ES. Pompano: biology, fisheries, and farming potential. *Proc Annu Gulf Caribb Fish Inst* 1967;19:116–128.
- Finucane JH. Ecology of the pompano (*Trachinotus carolinus*) and the permit (*T. falcatus*) in Florida. *Trans Am Fish Soc* 1969;98:478–486.
- Muller RG, Tisdell K, Murphy MD. *The 2002 update of the stock assessment of Florida pompano (Trachinotus carolinus)*. St Petersburg, FL: Florida Fish and Wildlife Conservation Commission, Florida Marine Research Institute; 2002.
- McMaster M, Kloth T, Coburn J. Prospects for commercial mariculture. In: *Aquaculture America 2003 Exhibition and Conference (18th-21st Feb 2003, Louisville, KY, USA)*. 2003. <https://mariculturerechnology.com/wp-content/uploads/2019/07/AquacultureAmerica03.pdf>
- Tutman P, Glavić N, Kožul V, Skaramuca B, Glamuzina B. Preliminary information on feeding and growth of pompano, *Trachinotus ovatus* (Linnaeus, 1758) (Pisces; Carangidae) in captivity. *Aquaculture International* 2004;12:387–393.
- Klinkhardt M, Myrseth B, editors. *New aquaculture candidates*. Global Trade Conference on Aquaculture; 2007.
- Juniyanto MN, Akbar S. Breeding and seed production of silver pompano (*Trachinotus blochii*, Lacepede) at the mariculture development center of batam. *Providing Claims Services to the Aquaculture Industry* 2008;8:46–48.
- FAO. *Global Aquaculture Production*. online query; 2021. <http://www.fao.org/fishery/statistics/global-aquaculture-production/query/en> [accessed 10 June 2021].
- Giang N, Binh D, Hoa D. Preliminary study of white spot disease in internal organs in snubnose pompano (*Trachinotus blochii*). *J Fish Sci Technol* 2012;4:26–33.
- Vu-Khac H, Duong VQ, Chen S-C, Pham TH, Nguyen TT, et al. Isolation and genetic characterization of *Nocardia seriolae* from snubnose pompano *Trachinotus blochii* in Vietnam. *Dis Aquat Organ* 2016;120:173–177.
- Kariya T, Kubota S, Nakamura Y, Kira K. Nocardial infection in cultured yellowtails (*Seriola quinqueradiata* and *S. purpurascens*)—I. *Fish Pathol* 1968;3:16–23.
- Kusuda R, Salati F. Major bacterial diseases affecting mariculture in Japan. *Annu Rev Fish Dis* 1993;3:69–85.
- Kudo T, Hatai K, Seino A. *Nocardia seriolae* sp. nov. causing nocardiosis of cultured fish. *Int J Syst Bacteriol* 1988;38:173–178.
- Chen S, Tung M, Tsai W. An epizootic in Formosa snake-head fish, *Channa maculata* Lacepede, caused by *Nocardia asteroides* in fresh water pond in Southern Taiwan. *COA Fisheries Series* 1989;15:42–48.
- Chen S, Tung M. An epizootic in large mouth bass, *Micropterus salmoides*, lacepede caused by *Nocardia asteroides* in freshwater pond in Southern Taiwan. *J Chin Soc Vet Sci* 1991;17:15–22.
- Chen S-C, Lee J-L, Lai C-C, Gu Y-W, Wang C-T, et al. Nocardiosis in sea bass, *Lateolabrax japonicus*, in Taiwan. *J Fish Dis* 2000;23:299–307.
- Huang S. Isolation and characterization of the pathogenic bacterium, *Nocardia seriolae*, from female broodstock of striped mullet (*Mugil cephalus*). *J Fish Res* 2004;12:61–69.
- Park M, Lee D-C, Cho M-Y, Choi H-J, Kim J-W. Mass mortality caused by nocardial infection in cultured snakehead, *Channa argus* in Korea. *J Fish Pathol* 2005;18:157–165.
- Shimahara Y, Nakamura A, Nomoto R, Itami T, Chen S-C, et al. Genetic and phenotypic comparison of *Nocardia seriolae* isolated from fish in Japan. *J Fish Dis* 2008;31:481–488.
- Shimahara Y, Huang Y-F, Tsai M-A, Wang P-C, Yoshida T, et al. Genotypic and phenotypic analysis of fish pathogen, *Nocardia seriolae*, isolated in Taiwan. *Aquaculture* 2009;294:165–171.
- Cornwell ER, Cinelli MJ, McIntosh DM, Blank GS, Wooster GA, et al. Epizootic *Nocardia* infection in cultured weakfish, *Cynoscion regalis* (Bloch and Schneider). *J Fish Dis* 2011;34:567–571.
- Kim JD, Lee N-S, Do JW, Kim MS, Seo HG, et al. *Nocardia seriolae* infection in the cultured eel *Anguilla japonica* in Korea. *J Fish Dis* 2018;41:1745–1750.
- Del Rio-Rodriguez RE, Ramirez-Paredes JG, Soto-Rodriguez SA, Shapira Y, Huchin-Cortes MDJ, et al. First evidence of fish nocardiosis in Mexico caused by *Nocardia seriolae* in farmed red drum (*Sciaenops ocellatus*, Linnaeus). *J Fish Dis* 2021;44:1117–1130.
- Imajoh M, Fukumoto Y, Yamane J, Sukeda M, Shimizu M, et al. Draft genome sequence of *Nocardia seriolae* Strain N-2927 (NBRC 110360), isolated as the causal agent of nocardiosis of yellowtail (*Seriola quinqueradiata*) in Kochi Prefecture, Japan. *Genome Announc* 2015;3:e00082-15.
- Xia L, Cai J, Wang B, Huang Y, Jian J, et al. Draft genome sequence of *Nocardia seriolae* ZJ0503, a fish pathogen isolated from *Trachinotus ovatus* in China. *Genome Announc* 2015;3:e01223-01214.
- Imajoh M, Sukeda M, Shimizu M, Yamane J, Ohnishi K, et al. Draft genome sequence of erythromycin- and oxytetracycline-sensitive *Nocardia seriolae* strain U-1 (NBRC 110359). *Genome Announc* 2016;4:e01606-01615.
- Yasuike M, Nishiki I, Iwasaki Y, Nakamura Y, Fujiwara A, et al. Analysis of the complete genome sequence of *Nocardia seriolae* UTF1, the causative agent of fish nocardiosis: the first reference genome sequence of the fish pathogenic *Nocardia* species. *PLoS One* 2017;12:e0173198.
- Han H-J, Kwak M-J, Ha S-M, Yang S-J, Kim JD, et al. Genomic characterization of *Nocardia seriolae* strains isolated from diseased fish. *Microbiologyopen* 2019;8:e00656.
- Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, et al. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* 1995;33:2233–2239.
- Calvez S, Fournel C, Douet D-G, Daniel P. Pulsed-field gel electrophoresis and multi locus sequence typing for characterizing genotype variability of *Yersinia ruckeri* isolated from farmed fish in France. *Vet Res* 2015;46:73.
- Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;28:593–594.
- Sarovich DS, Price EP. SPANdx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. *BMC Res Notes* 2014;7:618.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* 2010;20:623–635.

36. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156–2158.
37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
38. Swofford DL. PAUP\*. Phylogenetic analysis using parsimony (\*and other methods), V4.0a168. Sunderland, MA: Sinauer Associates, 2003.
39. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 2012;6:80–92.
40. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;14:1394–1403.
41. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011;12:402.
42. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 2018;4:vey016.
43. Holt K, Kenyon JJ, Hamidian M, Schultz MB, Pickard DJ, et al. Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant *Acinetobacter baumannii* global clone 1. *Microb Genom* 2016;2:e000052.
44. Holt K, Kenyon JJ, Hamidian M, Schultz MB, Pickard DJ, et al. Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant *Acinetobacter baumannii* global clone 1. *Microb Genom* 2016;2:e000052.
45. Germer S, Holland MJ, Higuchi R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Res* 2000;10:258–266.
46. Price EP, Matthews MA, Beaudry JA, Allred JL, Schupp JM, et al. Cost-effective interrogation of single nucleotide polymorphisms using the mismatch amplification mutation assay and capillary electrophoresis. *Electrophoresis* 2010;31:3881–3888.
47. Birdsell DN, Pearson T, Price EP, Hornstra HM, Nera RD, et al. Melt analysis of mismatch amplification mutation assays (Melt-MAMA): a functional study of a cost-effective SNP genotyping assay in bacterial models. *PLoS One* 2012;7:e32866.
48. Hézard N, Cornillet P, Droullé C, Gillot L, Potron G, et al. Factor V Leiden: detection in whole blood by ASA PCR using an additional mismatch in antepenultimate position. *Thromb Res* 1997;88:59–66.
49. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
50. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;18:821–829.
51. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 2009;25:1968–1969.
52. Tsai IJ, Otto TD, Berriman M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol* 2010;11:R41.
53. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011;27:578–579.
54. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano WJB. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011;27:578–579.
55. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol* 2012;13:R56.
56. Nadalin F, Vezzi F, Policriti A. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 2012;13:14.
57. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 2014;9:e112963.
58. Seemann TJB. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
59. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, et al. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 2018;46:W246–W251.
60. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;9:75.
61. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 2014;42:D581–91.
62. Crispell J, Balaz D, Gordon SV. HomoplasmyFinder: a simple tool to identify homoplasies on a phylogeny. *Microb Genom* 2019;5.
63. Vera-Cabrera L, Ortiz-Lopez R, Elizondo-Gonzalez R, Ocampo-Candiani J. Complete genome sequence analysis of *Nocardia brasiliensis* HJEG-1 reveals a saprobic lifestyle and the genes needed for human pathogenesis. *PLoS One* 2013;8:e65425.
64. Sun J, Fang W, Ke B, He D, Liang Y, et al. Inapparent *Streptococcus agalactiae* infection in adult/commercial tilapia. *Sci Rep* 2016;6:26319.
65. Bakker HC den, Switt AIM, Cummings CA, Hoelzer K, Degoricija L, et al. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to A common *Salmonella enterica* subsp. *enterica* serovar Montevideo pulsed-field gel electrophoresis type. *Appl Environ Microbiol* 2011;77:8648–8655.
66. Kwong JC, Mercouliou K, Tomita T, Easton M, Li HY, et al. Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. *J Clin Microbiol* 2016;54:333–342.
67. Lee K-I, Morita-Ishihara T, Iyoda S, Ogura Y, Hayashi T, et al. A geographically widespread outbreak investigation and development of a rapid screening method using whole genome sequences of enterohemorrhagic *Escherichia coli* O121. *Front Microbiol* 2017;8:701.
68. Seifert H, Dolzani L, Bressan R, van der Reijden T, van Strijen B, et al. Standardization and interlaboratory reproducibility assessment of pulsed-field gel electrophoresis-generated fingerprints of *Acinetobacter baumannii*. *J Clin Microbiol* 2005;43:4328–4335.
69. Uelze L, Grütze J, Borowiak M, Hammerl JA, Juraschek K, et al. Typing methods based on whole genome sequencing data. *One Health Outlook* 2020;2:3.
70. Nakada M. Capture-based aquaculture of yellowtail. In: Lovatelli A and Holthuis PF (eds). *Capture-Based Aquaculture Global Overview*. Rome: FAO Fisheries Technical Paper 2008. No. 508 FAO; . pp. 199–215.
71. Labrie L, Ng J, Tan Z, Komar C, Ho E, et al. Nocardial infections in fish: an emerging problem in both freshwater and marine aquaculture systems in Asia. *Diseases in Asian aquaculture VI Fish Health Section, Asian Fisheries Society, Manila* 2008:297–312.
72. Davis-Scibienski C, Beaman BL. Interaction of *Nocardia asteroides* with rabbit alveolar macrophages: association of virulence, viability, ultrastructural damage, and phagosome-lysosome fusion. *Infect Immun* 1980;28:610–619.
73. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 2015;13:722–736.
74. Newsom S, Parameshwaran HP, Martin L, Rajan R. The CRISPR-cas mechanism for adaptive immunity and alternate bacterial functions fuels diverse biotechnologies. *Front Cell Infect Microbiol* 2020;10:619763.
75. Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, et al. High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. *Curr Biol* 2022;32:220–227.

76. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. CRISPRcastyper: automated identification, annotation, and classification of CRISPR-cas loci. *CRISPR J* 2020;3:462–469.
77. Zhang Q, Ye Y. Not all predicted CRISPR-Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics* 2017;18:1–12.
78. Zhang M, Bi C, Wang M, Fu H, Mu Z, et al. Analysis of the structures of confirmed and questionable CRISPR loci in 325 *Staphylococcus* genomes. *J Basic Microbiol* 2019;59:901–913.
79. Shmakov SA, Utkina I, Wolf YI, Makarova KS, Severinov KV, et al. CRISPR arrays away from cas genes. *CRISPR J* 2020;3:535–549.
80. Tanmoy AM, Saha C, Sajib MSI, Saha S, Komurian-Pradel F, et al. CRISPR-cas diversity in clinical *Salmonella enterica* serovar Typhi isolates from South Asian countries. *Genes* 2020;11:E1365.
81. de Oliveira Luz AC, da Silva Junior WJ, do Nascimento Junior JB, da Silva JMA, de Queiroz Balbino V, et al. Genetic characteristics and phylogenetic analysis of Brazilian clinical strains of *Pseudomonas aeruginosa* harboring CRISPR/Cas systems. *Curr Genet* 2021;67:663–672.

#### **Five reasons to publish your next article with a Microbiology Society journal**

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

**Find out more and submit your article at [microbiologyresearch.org](https://microbiologyresearch.org).**