nature portfolio

Corresponding author(s): NCOMMS-23-58269C

Last updated by author(s): Aug 5, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The Streptococcus mitis isolates were provided by BSAC and UKHSA together with a Microsoft Excel Metadata spreadsheet. Software and code was only used for data analysis after the isolates had undergone whole genome sequencing.
Data analysis	Post-sequencing quality control, genome assembly and speciation Illumina sequencing reads of presumed IE S. mitis, from our previous work, were checked for quality using FastQC (version 0.11.9) (https://github.com/s-andrews/FastQC), and trimmed using Trimmomatic (version 0.39) and a phred score of at least 33 per read was used as the minimum quality score threshold. De novo genome assembly was performed using default parameters in SPAdes (version 3.12) and genome quality was determined using the quality assessment tool for genome assemblies (QUAST version 5.0.2) with default parameters (Supplementary Data 7). Taxonomic classification of the sequenced S. mitis, that formed our curated dataset of 322 S. mitis genomes, was firstly done using KRAKEN (version 1.0) against the MiniKren DB_8GB database, and KRAKEN (version 2.0) against the minikraken2_v2_8GB_201904 database using default parameters (Supplementary Fig. 8). Genomes assigned as S. mitis were further screened by applying the online PathogenWatch Speciator tool (https://pathogen.watch/), where the in-house species identification tool applied MASH to search a curated NCBI RefSeq database. Genomes that were not assigned as S. mitis by both KRAKEN versions and PathogenWatch were excluded. In this current analysis, we reanalysed the curated 322 S. mitis genome dataset using the speciation methods described above, and as an additional screening step, average nucleotide identity (ANI) values were calculated using fastANI (version 1.32). All strain pairs were tested against each other and against a list of complete S. mitis genomes using the “many to many” method and by using the “–matrix” option. Previous studies have suggested that ANI values of 94–96% are generally accepted as a species boundary, however, S. mitis has been shown to have lower ANI values of up to 91% as the group consists of a continuum of lineages. Therefore, a relaxed approach using a 90% ANI threshold was used. Lastly, an S. mitis phylogeny was generated using the methods described below, and species assignment methods were assessed together to confirm the species.

Global S. mitis genomes were used to contextualise locally obtained isolates in a broader perspective. All publicly available S. mitis genome assemblies used in this project were downloaded from The National Center for Biotechnology Information (NCBI) genome database (https://www.ncbi.nlm.nih.gov/) and were from carriage, invasive disease, and unknown conditions (Supplementary Data 3).

Pairwise SNP distance, phylogeny, and population structure analysis
Snippy (version 4.6.0) (https://github.com/tseemann/snippy) was used to map confirmed UK IE S. mitis sequence reads to the S. mitis B6 reference genome (GenBank Accession: GCA_000027165.1) to obtain SNPs, determine genetic diversity, and the alignment was used to construct maximum-likelihood phylogenies using fasttree (version 2.1.10). We used the generalized time-reversible model of nucleotide evolution to generate the phylogenies, which were visualized and annotated using the online Interactive Tree of Life (iToL) software (version 3.0) and microreact (version 240). Isolates were clustered into Global Sequence Clusters (GSC) using PopPUNK (version 2.4.0), and STs were defined using a novel multi-locus sequence typing (MLST) scheme (https://pubmlst.org/organisms/streptococcus-mitis).

To obtain a core-genome alignment using global S. mitis, genome assemblies were first annotated using Prokka (version 1.13.4), and a core-genome analysis was conducted using Panaroo (version 1.2 .9) to obtain a core genome alignment. An alignment of SNPs was generated from the core-genome alignment using Snp-Sites (version 2.5.1), and phylogenies were constructed as described above. Acquired antimicrobial resistance (AMR) and virulence genes were identified among the streptococci using Abricate (version 0.9.8) (https://github.com/tseemann/abricate). The ResFinder and virulence finder databases were used as references for AMR genes and virulence genes, respectively. Since very few S. mitis genomes have been sequenced and studied, genotypic resistance was used to determine concordance with phenotypic data.

Bacterial genome-wide association study (GWAS)
We undertook a pilot bacterial genome-wide association study to identify specific genetic changes overrepresented in IE-associated S. mitis isolates when compared to those collected from nasopharyngeal carriage. Due to the high genetic diversity, and the modest dataset size, we only investigated the relative abundance of genes or gene clusters identified from the pangenome analysis using Panaroo. Because of the extremely high within-species genetic diversity of S. mitis and the challenges of collecting matched isolates from invasive disease and carriage from the same setting and time frame, we performed a two-stage bacterial genome-wide association study (GWAS) analysis. First, we generated a maximum likelihood phylogenetic tree of recently sequenced and publicly available confirmed S. mitis whole-genome sequences. We annotated the phylogenetic tree with the disease status of the isolates based on the body isolation site, i.e., blood as an 'IE-associate BSI' and oropharynx or nasopharynx as 'asymptomatic carriage'. Second, we selected pairs of genetically closest carriage and invasive disease isolates that shared the most recent ancestors regardless of their genetic divergence. We then pruned the initial phylogenetic tree of all the isolates to remain with a subtree with an equal number of invasive diseases and carriage S. mitis isolates, where each pair of carriage and invasive disease isolates formed monophyletic clades. This approach provided an approximate matching of the isolates, albeit with higher divergence than seen with similar analyses in other bacterial species, such as Staphylococcus aureus, Staphylococcus epidermidis, and Mycobacterium tuberculosis, to allow for a robust assessment of the genes potentially enriched in the carriage and invasive disease isolates. Due to the phylogenetic matching or pairing of the S. mitis isolates, we used the exact McNemar's test to identify genes or gene clusters overrepresented in IE or carriage-associated isolates. We used the function "mcnemar.test" in the stats (version 4.0.3) R package to perform the exact McNemar's test. Genes or gene clusters with P-value <0.05 were considered to be statistically significant. Overrepresented genes identified as hypothetical genes were further checked using the online NCBI BLAST tool, its databases, and default parameters to determine any known gene functions. NCBI BLAST matches with the highest total score, sequence coverage, and sequence identity were used to assign potential gene function.

Invasive Streptococcus pneumoniae genomes used for screening of overrepresented S. mitis genes were obtained from the Global Pneumococcal Sequencing Project (Supplementary Data 6). The Global Pneumococcal Sequencing Project (GPS) was screened to identify S. pneumoniae genomes obtained via blood from patients with bacteremia. The largest collection of S. pneumoniae genomes collected through bacteraemia surveillance was therefore used and is part of the Centers for Disease Control and Prevention's (CDC) active bacterial core surveillance. Abricate (version 1.0.1) was used with default settings and the overrepresented S. mitis genes as the database to screen invasive pneumococcal genomes for the presence of these genes.

Statistical analysis
Statistical tests and associated diagrams were generated in R (version 2.11.1) (R Core Team 2014; https://www.R-project.org/), GraphPad Prism (version 8.0) (GraphPad Software, San Diego, California, USA), and edited in Inkscape version 1.0.0. Parametric data collected included the age group of the IE cases and are presented as frequencies. Non-parametric data, which included antibiotic minimum inhibitory concentrations (MICs) and pairwise-SNP distances, are presented as individual data points and median values. The Kruskal-Wallis test was used to compare median MICs among isolates grouped by year, and the test was also used to compare population-level genetic diversity by pairwise SNPs and ANI values across the 16-year surveillance period. Statistical significance was defined as $p < 0.05$.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Genomes sequenced in this study have been deposited in the US National Center for Biotechnology Information (NCBI) database under BioProject accession code PRJEB55310 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB55310/]. Publicly available genomes used in this project are under BioProjects PRJNA480039 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA480039/], PRJEB42564 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB42564/], PRJEB42963 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB42963/], and PRJEB53188 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB53188/]. All genomes used in this study are also shared under genome assembly accessions listed in Supplementary Data 3, 4, and 6. Source data are provided with this paper

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Genomic findings of Streptococcus mitis do not apply to only one sex, therefore sex was not considered in the study design of this retrospective sequencing study. Data on sex of the infective endocarditis cases was assigned by the referring clinical laboratories that submitted the S. mitis isolates to the British Society of Antimicrobial Chemotherapy (BSAC) and the UK Health Security Agency (UKHSA) as part of the surveillance programme. UKHSA holds approvals to process patient-identifiable data for the purposes of infectious disease surveillance, in accordance with Section 60 of the Health and Social Care Act 2001. The S. mitis isolates from patients with infective endocarditis were from patients of all ages (0 to 99 years) with the age range 50-59 years having the highest frequency of S. mitis isolates (20.2%; 26/129). Of the 129 confirmed S. mitis isolates, 70 (54.3%) were collected from men and 58 (45.0%) from women. Sex-based analyses were not performed because the focus of the retrospective sequencing study was the genomic epidemiology of S. mitis invasive disease.
Reporting on race, ethnicity, or other socially relevant groupings	We did not have access to patient-identifiable data, including data on race or ethnicity from BSAC or UKHSA.
Population characteristics	We did not have access to patient-identifiable data or clinical information from BSAC or UKHSA. Only age in ranges and sex were provided.
Recruitment	All available isolates that were collected, archived, and identified as S. mitis from suspected IE cases by BSAC and UKHSA from 2001–2016 were included in the study. These isolates were collected as part of BSAC's Resistance Surveillance Project, and UKHSA's voluntary identification service. The diagnosis of IE assigned to the UKHSA and BSAC S. mitis isolates was made by the referring clinical teams. However, well-phenotyped isolates from more recent years were not available. The S. mitis isolates were not collected systematically from all regions of the UK and Ireland as bacterial surveillance isolate submission was voluntary, possibly introducing bias in the samples submitted by the hospital laboratories to BSAC and UKHSA.
Ethics oversight	This study complies with ethical regulations applied in public health surveillance. UKHSA holds approvals to process patient-identifiable data for the purposes of infectious disease surveillance, in accordance with Section 60 of the Health and Social Care Act 2001. All isolates were anonymised to the key researcher, and patient-identifiable information was not included in the study. Isolates submitted to BSAC were collected as part of routine clinical investigations and were processed as such by the original laboratory.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This was a retrospective whole genome sequencing study of Streptococcus mitis isolates collected from patients clinically diagnosed with infective endocarditis between 2001 and 2016. Isolates used in this study were originally submitted to the British Society of Antimicrobial Chemotherapy (BSAC) and the UK Health Security Agency (UKHSA) by referring clinical laboratories in the UK and Ireland as part of BSAC and UKHSA bacteraemia surveillance programmes. A total of 129 S. mitis isolates from BSAC and UKHSA were analysed. This dataset includes all available isolates that were collected, phenotyped, and archived. However, well-phenotyped isolates from more recent years were not available. We specify the organism taxa (S. mitis), source (Blood), sex of the IE cases, and their age ranges as described in Supplementary Table 2, Supplementary Data 1, and Supplementary Data 2. No manipulations of the organisms were performed. The S. mitis IE isolates may represent the strains present in the UK and Ireland. However, as one of our limitations in the manuscript we state that the S. mitis isolates were not collected systematically from all regions of the UK and Ireland as bacterial surveillance isolate submission was voluntary, possibly introducing bias in the samples submitted by the hospital laboratories to the British Society of Antimicrobial Chemotherapy (BSAC) and the UK Health Security Agency (UKHSA). The study involves other global S. mitis genomes from different sources described in Supplementary Data 3 and 4 where accession numbers are listed. We also utilise Streptococcus pneumoniae genomes to contextualise our findings, and accession numbers are listed in Supplementary Data 6.
Research sample	Streptococcus mitis isolates were obtained from patients diagnosed with infective endocarditis (IE) in the UK and Ireland. These isolates were obtained from BSAC and UKHSA from their bacteraemia surveillance programmes. A total of 129 S. mitis isolates from BSAC and UKHSA were analysed. This dataset includes all available isolates that were collected, phenotyped, and archived. The S.

mitis isolates from patients with IE were from patients of all ages (0 to 99 years). Of the 129 confirmed S. mitis isolates, 70 (54.3%) were collected from men and 58 (45.0%) from women.

Sampling strategy

No sample size calculation was done, however, the S. mitis isolates that were obtained spanned the years from 2001 to 2016. This included the entire collection of archived BSAC and UKHSA isolates. Well-phenotyped isolates from more recent years were not available. Due to the rarity of IE, achieving even a modest sample size is very challenging and our current sample size is a study limitation.

Data collection

The diagnosis of IE assigned to the UKHSA and BSAC S. mitis isolates was made and recorded by the referring clinical teams. The modified Duke/ESC 2023 diagnostic criteria for IE were not available. However, in view of the BSI, and the referral of the isolates for species confirmation and antibiotic sensitivity testing for the management of IE, the patients likely fulfilled the “definite” or “possible” modified Duke/ESC 2023 diagnostic categories. The isolates were cultured by the hospital clinical laboratories as part of clinical management, and the isolates were subsequently submitted to BSAC and UKHSA as part of the bacteraemia surveillance programmes. BSAC and UKHSA conducted further speciation and antimicrobial susceptibility testing on the isolates, where the data was recorded on Microsoft Excel Metadata spreadsheets. We obtained both the isolates and metadata from UKHSA and BSAC. Where phenotypic resistance data was not available, we derived penicillin MICs for presumed S. mitis isolates using the E-test method.

Timing and spatial scale

The S. mitis isolates that were part of the study covered the surveillance period from 2001 to 2016. This spanned the entire collection of archived BSAC and UKHSA S. mitis isolates. However, well-phenotyped isolates from more recent years were not available. Only the year of collection was available as the exact sampling dates were not provided by BSAC and UKHSA. The geographical locations of individual isolates were also not available, however, BSAC surveillance covered the UK and Ireland, while UKHSA surveillance covered England, Wales, and Northern Ireland. Laboratory source of the isolates and yearly frequency of Streptococcus mitis isolates is described in Supplementary Figure 1.

Data exclusions

Genomes that were not assigned as S. mitis by both KRAKEN versions and PathogenWatch were excluded. The exclusion criteria was pre-established as the aim of the study was to understand the genomic epidemiology of invasive S. mitis.

Reproducibility

The source of the S. mitis isolates, genomic extraction methods, and whole genome sequencing approach has been described in the Methods section. The sequencing data, genomic analysis methods, tools used, and their versions numbers have also been provided in the Methods and Data Availability sections for reproducibility.

Randomization

Randomisation was not relevant to our study as we conducted a retrospective whole genome sequencing study of invasive S. mitis isolates obtained from bacteraemia surveillance programmes.

Blinding

Blinding was not relevant to our study as we conducted a retrospective whole genome sequencing study of invasive S. mitis isolates obtained from bacteraemia surveillance programmes.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Plants

Seed stocks	This study involved the whole genome sequencing of Streptococcus mitis. No plants or plant material was used in this study.
Novel plant genotypes	This study involved the whole genome sequencing of Streptococcus mitis. No plants or plant material was used in this study.
Authentication	This study involved the whole genome sequencing of Streptococcus mitis. No plants or plant material was used in this study.