nature research

Corresponding author(s): J. Ross Fitzgerald

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software used

Data analysis	Completed genomes downloaded from the NCBI database were converted into pseudo-fastq files using Wgsim (https://github.com/lh3/wgsim). Multiple assemblies using VelvetOptimiser v2.2.5 and Velvet v1.2. The assemblies were improved by scaffolding the best N50 and contigs using SSPACE and sequence gaps filled using GapFiller. Sequence reads were mapped to a relevant reference genome using SMALT. Consensus sequences were obtained using samtools. A maximum likelihood tree was constructed using RAxML. Time scaled trees were generated using BEAST 1.8.2. and Beagle. Pseudogenes were predicted during the PROKKA. Proteins in a genome was searched against UniProtKB (Swiss-Prot) using BLASTp 54 or UniProtKB (TrEMBL). The UniProt ID Mapping tool was used to assign Gene ontology (GO) terms to all pseudogenes. GO was assigned to all non-pseudogenes (CDS features) using InterProScan. The R package topGO with Fisher's exact test was used to identify enriched GO terms. Pan-genome association analysis were performed using Nucmer, Get_homologues, the distmat function in EMBOSS and the graph was processed in BioLayout. Identification of gene acquisitions or losses associated with host-switching events were performed using the R package APE. Codon usage bias analysis was performed using the EMBOSS tools cusp and cai. Genome-wide positive selection analysis were performed using get_homologues, MUSCLE 3.8.31, pal2nal v14, PhiPack and PAML. Functional categories were annotated using the COGs and GO databases as reference. Identification of overrepresented GO categories of positively selected genes in different hosts, we used BiNGO and REVIGO.

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The sequence datasets generated during the current study are available in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) with the accession number PRJEB20741.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

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

Study description	Population genomic study of bacterial evolution in the context of the host-species
Research sample	For selection of isolates, the literature was reviewed (date: November 2013) and all available S. aureus strains associated with animals and humans for which genomes had been determined were identified.
Sampling strategy	We aimed to include isolates to represent the breadth of clonal complexes, host-species diversity, geographical locations and as wide a temporal scale as possible. Publicly available sequences were selected as follows; 74 reference genomes, 302 from the EARSS project29, and 252 from other published studies of the authors. Furthermore, to be as representative of the known S. aureus host, clonal, and geographic diversity as possible we selected an additional 172 isolates for whole genome sequencing (Supplementary Table 1).
Data collection	sequencing of existing bacterial isolates
Timing and spatial scale	n/a
Data exclusions	Isolates were excluded from the analysis for the following reasons that are indicative of contamination or poor quality sequence data; a large number of contigs and a large number of 'N's in the assemblies or genome size larger than expected for S. aureus (>2.9 Mb).
Reproducibility	All findings are reproducible
Randomization	Isolates were allocated into groups according to their respective host-species. Random sub-sampling was carried out to limit the effects of sample bias.
Blinding	n/a

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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