nature research

Corresponding author(s): James M. Musser

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	- Trimmomatic: Read quality trimming. Version: 0.36 - Musket: Read error correction. Version: 1.1 - SRST2: detection of genes, alleles, and multi-locus sequence types (MLSTs). Version: 0.2.0 - SPAdes: de novo genome assembly. Version: 3.9.1 - SMALT: Read mapping to reference strain, (available at www.sanger.ac.uk/resources/software/smalt/). Version: 0.7.6 - FreeBayes: Identification of SNPs and InDels, (available at www.github.com/ekg/Freebayes/). Version: 1.0.2-16 - Pilon: Identification of InDels. Version: 1.22 - bcl2fastq: Demultiplexing of reads from superpools into separate pools, (https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html). Version: 2.20 - FASTX-toolkit: demultiplexing of reads from each pool into separate fastq files. Version: 0.0.14 - FASTQC: Read quality assessment, (available at https://www.bioinformatics. babraham.ac.uk/projects/fastqc/). Version: 0.11.6 - Unicycler: Hybrid assembly pipeline. Version: 0.4.1 - EDGE-Pro: Read mapping to the genome of reference strain MGAS6180. Version: 1.3.1
Data analysis	- Gubbins: Identification of potential regions of recombination. - SNPfx: PERL script developed in house used to determine the nature of the SNPs (coding/noncoding, synonymous/nonsynonymous, etc). Custom script made in the laboratory. - BAPS: Determination of population structure, (Bayesian Analysis of Population Structure, available at http://www.helsinki.fi/bsg/software/BAPS/)

- Splits Tree: Estimation of phylogenetic trees and networks. Version: 4.14.4
- k-means: Selection of genetically representative singleton strains for transcriptome analysis
- MEGA: Determination of mean genetic distances
- DESeq2:Differential expression analysis of strains with biological replicates. Versions 1.14.1 and 1.16.1
- NOISeq: Differential expression analysis of the 442 singleton strains. Version: 2.20.0
- MATLAB with the function TreeBagger: Random Forest Analysis
- GraphPad Prism software v6, 7.0a, and 7.03
- SEER: Sequence element enrichment analysis. Used for genome-wide determination of genetic variants associated with a phenotypic trait. Version: 1.1.4
- Matrix eQTL: Identification of genomic loci that contribute to variation in expression levels of mRNAs. Version: 2.2

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

Whole-genome sequencing data for the 2,101 isolates studied were deposited in the NCBI Sequence Read Archive under the Bioproject accession number PRJNA434389. The slightly updated complete genome sequence of emm28 reference strain MGAS6180 (GenBank accession number CP000056) has been deposited in the NCBI GenBank database under the same accession number. Transcriptome data has been deposited in the Gene Expression Omnibus under accession GSE113058.

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

Life sciences study design

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

Sample size	We performed whole genome sequencing on 2,101 strains, analyzed the transcriptome of 492 strains (23.4% of the entire population) and did virulence assays on 50 strains (2.4% of the entire population) for which we also had transcriptome data. We arbitrarily opted to use ~25% of strain cohort for the expression studies. For an unbiased selection of strains for the transcriptome analysis from the population of 2,101 strains, we used a projection-based approach. We sampled isolates proportionately to the fraction of the total population size represented by each genetic lineage. For the given total size size k of the subset to be chosen from a lineage, the k-means algorithm was used to identify an optimal set of k centroids to span the variation present with the lineage. Finally, the isolate with the minimum Euclidean distance to each centroid was chosen to determine the final subset of k representative isolates. The 5x10^8 CFU inoculum used for the mouse infection studies was determined by Probit analysis (XLSTAT2011) of survival data generated from a dose-escalation study with virulence reference strain MGAS28426. A power calculation for the log-rank test (www.statstodo.com) predicted that a sample size of 20 mice per strain treatment group was needed to generate statistically significant results when using the pilot study survival rates and parameters alpha=0.05, beta=0.8, and r=1. Power calculations for other virulence experiments described in this study were performed in a manner as described above. Transcriptome and virulence assays were done on strains numerically representative from the 3 major genetic subclades identified in our study. These values do not reflect the 19 strains that were excluded (see below).
Data exclusions	19 strains were excluded from the transcriptome analysis part of the study for technical reasons associated with contamination or low read coverage. The exclusion criteria were pre-established.
Replication	Transcriptome analyses for 50 strains were conducted in triplicate (3 biological replicates) and extremely high correlation coefficients were found for the 3 replicates representing each strain. For each strain, correlation between biological replicates, at each growth phase, was computed using Pearson correlation coefficient (r). At mid-exponential phase we found r>99% and at early-stationary phase r values >90% were obtained. 442 strains were analyzed as singleton (expression analyses were done without replicates). To assure the quality of the data we also set the criteria for sequencing depth at 5-10 million sequencing reads per strain or biological replicate. For the validation of the singleton strategy 7 of the 50 strains analyzed in triplicate were also analyzed as singletons, and yielded a very strong correlation in the global expression profile. Correlation was determined using Pearson correlation coefficient (r) and found to be between 86-92% For the mouse experiments reproducibility was established in multiple ways. We used a minimum sample of 20 mice per bacterial strain per experiment. Prior to performing the 50 strain experiment, a pilot experiment was performed using reference strain MGAS28426 to select the appropriate dose. The pilot experiment established the baseline virulence phenotype of reference strain MGAS28426. The virulence phenotype of reference strain MGAS28426 was reproduced in the 50 strain experiment. Second, the virulence phenotype of 7 strains,

including reference strain MGAS28426, was reproduced in a subsequent experiment conducted to confirm reproducibility of the data (data not intended for publication). Third, the virulence phenotype of reference strain MGAS28426 was reproduced again in another experiment that is not part of the submitted research. Overall, all expression data and mouse virulence experiments could be successfully reproduced.

Randomization

Random allocation of strains is not relevant to our study. Genetic relationships were inferred and strains were assigned to genetic clades and subclades using the software hierBAPS (hierarchical Bayesian analysis of population structure). This process identified 2 major clades, divided into 4 subclades, and the number of strains selected for transcriptome analysis from each subclade was selected as being proportional to the total number of strains in each respective subclade, as described in Sample size.

Blinding

Strains were assigned numbers from 1-461 (singletons), and 1-50 (analyzed in triplicate). The exact identity of these strains was unknown until the experimental part and the analyses were completed.

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Every country and agency has their own set of rules regarding the export of biological materials. Thus, availability of strains will be dependent on the particular country's regulations.

Antibodies

Antibodies used

A custom-made rabbit affinity-purified polyclonal antibody (GenScript) was used to detect the R28 protein. The anti-R28 antibody was made against an R28 internal peptide with the sequence VVDPRTDADKNDPA. A 1:1000 dilution of this antibody was used.

Validation

The custom-made anti-R28 antibody did not cross-react against any other protein made by emm28 Group A streptococci.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

- species: Mus musculus
- strain: CD1 (Envigo Laboratories)
- sex: female
- age: 4 weeks

Wild animals

The study did involve the use of wild animals

Field-collected samples

The study did not involve samples collected from the field