**Supplementary Appendix**

Table of Contents

**Supplementary methods** 2

**Supplementary Figure 1 -** Proportion of *emm89.0* among invasive group A *Streptococcus* isolates, ABCs, 1995-2015. 5

**Supplementary Table S1 -** Outbreak status of cases with date of collection of specimens 7

**Supplementary Table S2 -** Beast prior distribution selection using log marginal likelihoods for five different phylogenetic models 14

**References** 15

## SUPPLEMENTARY METHODS:

**Core SNP phylogenetic comparisons of 102 *emm89.0* genomes**

Core genome SNP identification and alignment was carried out using kSNP v3.0 [1]. The optimal kmer length was calculated using the Kchooser script and the kSNP3 program was run with the ‘-core’ option. A maximum-likelihood phylogenetic tree was generated from the core SNP alignment using RaxML v7.3.0 [2]. RaxML was run with an ASC\_GTRGAMMA DNA substitution model and employed the Lewis method for ascertainment bias correction. Node support was assessed using 500 bootstrap replicates. The tree was visualized and annotated using the iTOL v3.2.4 [3] online visualization tool.

**Time-scaled Facility A Outbreak Phylogeny**

Reads were mapped using the Bowtie 2 v2.1.0 [4] read aligner. The Bowtie 2 software was run with the ‘--very-sensitive-local’ and ‘-a’ options. PCR duplicates were removed using the Samtools v0.1.18 [5] rmdup command and genome coverage was calculated using the bedtools v2.17.0 [6] genomecov software. The genome coordinates pertaining to regions below 10X depth were saved to a bed file. Variant calling was performed using Freebayes v0.8.21 [7] and run with the ‘-j -q 20 -m 20’ options. The vcffilter tool from the vcflib C++ library was used to filter potential variants with the following thresholds: ‘QUAL > 20 -g "DP > 10 & GQ > 30 & AO > 10 & GT = 1’. The filtered vcf file was then passed to the VCFtools v0.1.11 [8] command vcf-consensus to generate a whole genome consensus sequence for each outbreak isolate. However, one obstacle in creating this consensus sequence was masking out low depth regions. The coordinates in the low depth bed file were relative to the reference H293 genome and since the isolate genome may contain indels there is no guarantee that the sample consensus sequence will share the same coordinates. Because vcf-consensus builds a sequence by applying variants to the reference genome, a solution was found by masking the H293 reference before it was passed to vcf-consensus to create the consensus sequence for each sample. A small modification was made to the vcf-consensus script to allow reference files with the ambiguous “N” characters to be processed by the software. A whole genome multiple sequence alignment (MSA) was carried out using Mugsy v1r2.2 [9] and low quality alignment regions were filtered with the trimAl v1.2 [10] program using the ‘-automated1’ option. Alignment and post-processing yielded a final shared genome size of 1750034 bases. Finally, manual curation masked four additional regions [158973-159036, 359866-359925, 1486640-1486751, 1704941-1705338] where suspected read misalignments created false variants.

To track the evolution of the tip-dated samples over the course of the outbreak, a time-measured phylogeny was generated using the BEAST v1.8.2 [11] software. A default phylogeny was reconstructed with a HKY substitution model, a Gamma site heterogeneity model, strict molecular clock (with a Gamma [α=0.001, β=1000] ‘clock.rate’ prior) and a constant coalescent tree prior. One of the earlier outbreak isolates, 0569\_14, did not have a precise sample time and was instead estimated by adding a precision attribute to the date element. Additional models were generated using priors shown in Table S2. Using the Tracer v1.6 program it was determined that each Beast model should be run with MCMC chain lengths of 35 million in order to achieve an effective sample size (ESS) > 200 for all traces. Model selection was performed using Bayes factor comparisons with maximum likelihood estimates (MLE) generated by path sampling/stepping-stone sampling. Model comparison results indicated that no other model performed significantly better (2lnBF > 6) than the default. The final model was replicated 7 times to ensure reproducibility of the results. The TreeAnnotator v1.8.2 program was used to generate a maximum clade credibility tree with a 10% burnin and median node heights. The time-scaled tree was visualized and annotated using the FigTree v1.4.2 visualization tool. Finally, a quantitative measurement of the strength of ancestral signal in the post-chemoprophylaxis strains was calculated with the BaTS v1.0 [12] trait association software using 500 replicates. This trait analysis was replicated with 5 of the 7 BEAST runs to ensure reproducibility of the results. Seven patients were sampled at multiple times within 2 days of each other. This could include sampling at multiple sites or within at the same site multiple times. If these samples shared a MRCA then it was likely that these samples were essentially replicates of the same sequence. There was risk that this could skew the BaTS analysis since BaTS assumes it is comparing independently sampled sequences. To address this issue a single replicate was chosen for any sample that had a MRCA with another sample isolated from the same patient within 2 days of each other. Using this threshold, an additional BEAST phylogeny was generated with the following isolates removed: R38\_2, R18B\_2, R18B\_3, and CR10\_1. The new BaTS analysis validated the original results with an association index (AI) score of 1.49 and a p-value < 0.005. All of the in-house scripts and xml files used in this analysis are available at https://github.com/BenJamesMetcalf

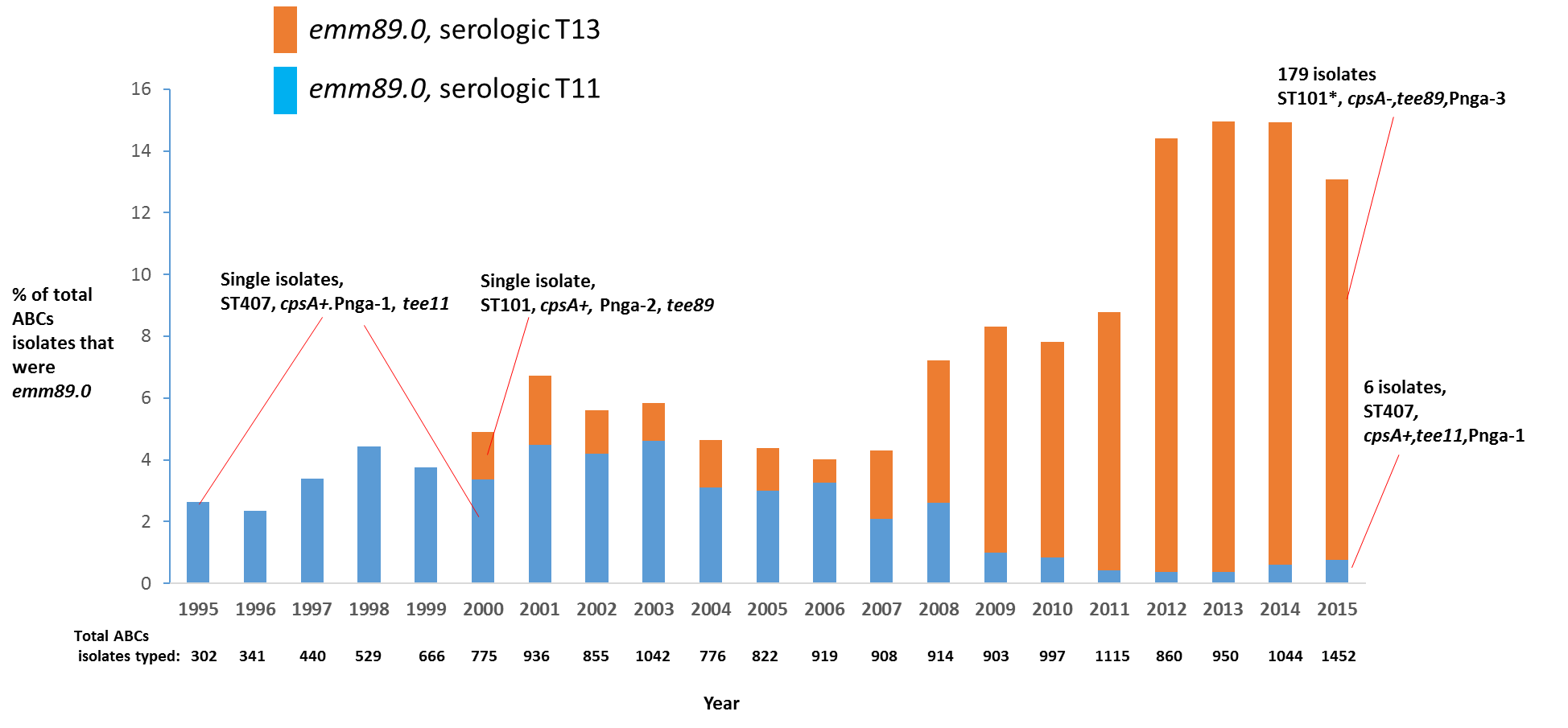


Figure S1. Proportion of *emm89.0* among invasive group A *Streptococcus* isolates, ABCs, 1995-2015. Thered lines indicate the isolates sequenced at CDC for this study. Note that 866 of *emm89.0* isolates recovered during 1995-2013 through ABCs were sequenced previously [13]. Our phenotypic data correlating the emergence of a serologic T13 variant associated with the Pnga-3 promoter and lack of *cpsA* capsular biosynthetic gene, along with the limited sequencing data from 1995 and 2001, and the complete year 2015 sequencing data, is consistent with the previously published 1995-2013 genomic data that noted the international emergence of the clade 3 *emm89* lineage featuring the Pnga-3 promoter and lack of *hasA* capsular biosynthetic locus. ST407 is a single locus variant of ST101.

Table S1:Outbreak status of cases with date of collection of specimens

|  |  |  |
| --- | --- | --- |
| **Outbreak ID\*** | **Case status** | **Date of collection of specimen** |
| R1 | Non-invasive | 5/2014# |
| R2 | Invasive | 5/2014# |
| R3 | Invasive | 5/2014# |
| CS1 | Colonized | 5/2014# |
| CS2 | Colonized | 5/2014# |
| CS3 | Colonized | 6/2014# |
| CS4 | Colonized | 6/2014# |
| CS5 | Colonized | 6/2014# |
| R4 | Invasive | 6/2014# |
| R5 | Non-invasive | 6/2014# |
| R6 | Non-invasive | 6/2014# |
| R7 | Non-invasive | 6/2014# |
| R8 | Non-invasive | 7/2014# |
| R9 | Non-invasive | 7/2014# |
| R10 | Non-invasive | 7/2014# |
| R11 | Non-invasive | 7/2014# |
| R12 | Invasive | 7/2014# |
| R13A | Invasive | 11/16/2014 |
| S1 | Non-invasive | 1/31/2015 |
| S2 | Non-invasive | 2/1/2015 |
| R14 | Non-invasive | 2/5/2015 |
| R15 | Non-invasive | 2/5/2015 |
| R16 | Non-invasive | 2/8/2015 |
| R17A | Non-invasive | 2/13/2015 |
| R18A | Non-invasive | 2/14/2015 |
| R19 | Invasive | 2/15/2015 |
| R20 | Non-invasive | 2/25/2015 |
| S3 | Non-invasive | 3/6/2015 |
| S4 | Non-invasive | 3/7/2015 |
| R21 | Invasive | 3/8/2015 |
| R22 | Non-invasive | 3/8/2015 |
| R23A | Non-invasive | 3/11/2015 |
| R24 | Invasive | 3/11/2015 |
| CS6 | Colonized | 3/12/2015 |
| CS7 | Colonized | 3/12/2015 |
| CS8 | Colonized | 3/12/2015 |
| CS9 | Colonized | 3/12/2015 |
| CS10 | Colonized | 3/12/2015 |
| CS11 | Colonized | 3/13/2015 |
| CS12 | Colonized | 3/13/2015 |
| CS13 | Colonized | 3/13/2015 |
| CS14 | Colonized | 3/13/2015 |
| R25A | Non-invasive | 3/13/2015 |
| R26 | Non-invasive | 3/13/2015 |
| R27 | Non-invasive | 3/15/2015 |
| CS15 | Colonized | 3/15/2015 |
| CS16 | Colonized | 3/16/2015 |
| R28 | Invasive | 3/18/2015 |
| R29 | Non-invasive | 3/18/2015 |
| CR1 | Colonized | 3/22/2015 |
| CR2 | Colonized | 3/22/2015 |
| CR3 | Colonized | 3/22/2015 |
| CR4 | Colonized | 3/23/2015 |
| CR5 | Colonized | 3/23/2015 |
| CR6 | Colonized | 3/23/2015 |
| S5 | Non-invasive | 3/24/2015 |
| R30 | Non-invasive | 3/27/2015 |
| R31 | Non-invasive | 3/29/2015 |
| R32 | Invasive | 3/29/2015 |
| S6 | Non-invasive | 3/29/2015 |
| CS17 | Colonized | 3/30/2015 |
| R17B | Non-invasive | 4/1/2015 |
| S7 | Non-invasive | 4/3/2015 |
| R33 | Non-invasive | 4/7/2015 |
| S8 | Non-invasive | 4/7/2015 |
| R34 | Non-invasive | 4/8/2015 |
| S9 | Non-invasive | 4/9/2015 |
| S10 | Non-invasive | 4/9/2015 |
| R35 | Non-invasive | 4/10/2015 |
| R36 | Non-invasive | 4/14/2015 |
| R37 | Invasive | 4/15/2015 |
| R38 | Non-invasive | 4/15/2015 |
| R39 | Non-invasive | 4/16/2015 |
| R13B | Invasive | 4/16/2015 |
| S11 | Non-invasive | 4/17/2015 |
| S12 | Non-invasive | 4/19/2015 |
| R40 | Non-invasive | 4/22/2015 |
| R41 | Non-invasive | 4/26/2015 |
| R42 | Invasive | 6/30/2015 |
| R43 | Invasive | 8/28/2015 |
| R44 | Invasive | 9/16/2015 |
| R25B | Invasive | 9/21/2015 |
| R45 | Non-invasive | 9/22/2015 |
| S13 | Non-invasive | 9/24/2015 |
| CR7 | Colonized | 9/29/2015 |
| CR8 | Colonized | 9/30/2015 |
| S14A | Non-invasive | 9/30/2015 |
| R23B | Invasive | 10/13/2015 |
| R46 | Invasive | 10/26/2015 |
| S15 | Non-invasive | 11/1/2015 |
| S16 | Non-invasive | 11/1/2015 |
| R47 | Non-invasive | 11/10/2015 |
| S17 | Non-invasive | 11/13/2015 |
| S18 | Non-invasive | 11/14/2015 |
| R18B | Non-invasive | 11/14/2015 |
| CR9 | Colonized | 11/15/2015 |
| CR10 | Colonized | 11/16/2015 |
| CS18 | Colonized | 11/17/2015 |
| CR11 | Colonized | 11/17/2015 |
| R48 | Non-invasive | 11/17/2015 |
| S14B | Non-invasive | 11/22/2015 |
| S19 | Non-invasive | 12/1/2015 |
| S20 | Non-invasive | 12/3/2015 |
| S21 | Non-invasive | 12/27/2015 |
| S22 | Non-invasive | 1/9/2016 |
| R49 | Non-invasive | 1/30/2016 |
| S23 | Non-invasive | 2/2/2016 |
| R50 | Invasive | 8/20/2016 |

**\*** Labels that begin with ‘C’ indicate a GAS colonized sample while the ‘S’ and ‘R’ designation represent staff and resident respectively. The number following ‘S’ or ‘R’ is the staff or resident index number. Recurrent cases are coded by a letter after the index number.

# Record of exact date was unavailable**.**

## Table S2

Beast prior distribution selection using log marginal likelihoods for five different phylogenetic models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Beast Run | DNA substitution Model | Molecular Clock | Tree Prior | Log Marginal Likelihood |
| Default | HKY Gamma | Strict | Constant Coalescent | -2341155.89 |
| 1 | HKY Gamma | Random Local | Constant Coalescent | -2341165.23 |
| 2 | GTR Gamma | Random Local | Constant Coalescent | -2341165.23 |
| 3 | GTR Gamma | Strict | Constant Coalescent | -2341166.30 |
| 4 | HKY Gamma | Strict | Constant Exponential Growth | No Convergence |

## References:

1. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics.* 2015;31(17):2877-2878.
2. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014.
3. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 2016;44(W1):W242-245.
4. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359.
5. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079.
6. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842.
7. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv e-prints.* 2012;1207:3907.
8. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27(15):2156-2158.
9. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics.* 2011;27(3):334-342.
10. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972-1973.
11. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 2012;29(8):1969-1973.
12. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol.* 2008;8(3):239-246.
13. Zhu L, Olsen RJ, Nasser W, et al. A molecular trigger for intercontinental epidemics of group A *Streptococcus*. J Clin Invest 2015;125:3545-59.