HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2021 October 26.

Published in final edited form as:

Nat Genet. 2019 March ; 51(3): 548–559. doi:10.1038/s41588-018-0343-1.

Integrated Analysis of Population Genomics, Transcriptomics and Virulence Provides Novel Insights into *Streptococcus pyogenes* Pathogenesis

Priyanka Kachroo[1,‡], **Jesus M. Eraso**[1,‡], **Stephen B. Beres**[1,¶], **Randall J. Olsen**[1,2,¶], **Luchang Zhu**[1,¶], **Waleed Nasser**[1], **Paul E. Bernard**[1], **Concepcion C. Cantu**[1], **Matthew Ojeda Saavedra**[1], **María José Arredondo**[1], **Benjamin Strope**[1], **Hackwon Do**[1], **Muthiah Kumaraswami**[1], **Jaana Vuopio**[3,4], **Kirsi Gröndahl-Yli-Hannuksela**[3], **Karl G. Kristinsson**[5,7], **Magnus Gottfredsson**[6,7], **Maiju Pesonen**[8,9], **Johan Pensar**[8], **Emily R. Davenport**[11], **Andrew G. Clark**[11], **Jukka Corander**[8,10], **Dominique A. Caugant**[12], **Shahin Gaini**[13,14,15], **Marita Debess Magnussen**[16,17], **Samantha L. Kubiak**[1], **Hoang A. T. Nguyen**[1], **S. Wesley Long**[1], **Adeline R. Porter**[18], **Frank R. DeLeo**[18], **James M. Musser**[1,2,*]

[1]Center for Molecular and Translational Human Infectious Diseases Research, Department of Pathology and Genomic Medicine, Houston Methodist Research Institute and Houston Methodist Hospital, Houston, Texas, USA

[2]Departments of Pathology and Laboratory Medicine, and Microbiology and Immunology, Weill Cornell Medical College, New York, New York, USA

[3]Institute of Biomedicine/Medical Microbiology and Immunology, University of Turku, Turku, Finland

[4]National Institute for Health and Welfare, Helsinki, Finland

[5]Department of Clinical Microbiology, Landspitali University Hospital, 101 Reykjavik, Iceland

[6]Department of Infectious Diseases, Landspitali University Hospital, 101 Reykjavik, Iceland

[7]Faculty of Medicine, School of Health Sciences, University of Iceland, 101 Reykjavik, Iceland

[*]**Corresponding author** James M. Musser; jmmusser@houstonmethodist.

[‡]Priyanka Kachroo and Jesus M. Eraso contributed equally to this work.

[¶]Stephen B. Beres, Randall J. Olsen, and Luchang Zhu contributed equally to this work.

AUTHOR CONTRIBUTIONS

J.M.M. conceptualized the study. P.K., J.M.E, and J.M.M. designed the study. P.K., J.M.E., S.B.B., R.J.O., L.Z., W.N., P.E.B., C.C.C., M.O.S., M.J.A., B.S., M.P., J.P., J.C., S.L.K., H.A.T.N., S.W.L., and A.R.P. produced the data. P.K., J.M.E., S.B.B., R.J.O., L.Z., H.D., M.K., M.P., J.P., J.C., S.W.L., and F.R.D. analyzed the data. P.K. led the analyses of the transcriptome data. M.P., J.P., E.R.D., A.G.C. and J.C. provided scholarly input on the statistical analysis and presentation strategies. J.V., K.G.-Y.-H., K.G.K., M.G., D.A.C., S.G., and M.D.M. provided strains and metadata. All authors contributed to writing the manuscript. All authors reviewed and approved the final draft.

ACCESSION CODES. The sequences reported in this paper have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (see URLs) with the Bioproject accession number **PRJNA434389** and the NCBI Gene Expression Omnibus (GEO) under accession number **GSE113058**.

ETHICS STATEMENT

All mouse studies were performed in accordance with a protocol (AUP-0318–0016) approved by the Institutional Animal Care and Use Committee at Houston Methodist Research Institute. All studies with human blood and blood components were performed in accordance with a protocol (01-I-N055) approved by the Institutional Review Board for human subjects, National Institute of Allergy and Infectious Diseases. All study volunteers gave written informed consent.

COMPETING INTERESTS

The authors declare no competing interests, either financial or non-financial.

[8]Helsinki Institute of Information Technology (HIIT), Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland

[9]Department of Computer Science, Aalto University, 00076 Espoo, Finland

[10]Department of Biostatistics, University of Oslo, 0317 Oslo, Norway

[11]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA

[12]Division for Infection Control and Environmental Health, Norwegian Institute of Public Health, P.O. Box 222 Skøyen, 0213 Oslo, Norway

[13]Medical Department, Infectious Diseases Division, National Hospital Faroe Islands, Tórshavn, Faroe Islands, Denmark

[14]Infectious Diseases Research Unit, Odense University Hospital and University of Southern Denmark, Odense, Denmark

[15]Centre of Health Research, Department of Science and Technology, University of the Faroe Islands, Tórshavn, Faroe Islands, Denmark

[16]Faculty of Medicine, School of Health Sciences, University of Iceland, 101 Reykjavik, Iceland

[17]Thetis, Food and Environmental Laboratory, Torshavn, Faroe Islands, Denmark

[18]Laboratory of Bacteriology, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, Montana, USA

Abstract

Streptococcus pyogenes causes 700 million human infections annually worldwide, yet, despite a century of intensive effort, there is no licensed vaccine against this bacterium. Although a number of large-scale genomic studies of bacterial pathogens have been published, the relationships between the genome, transcriptome, and virulence in large bacterial populations remain poorly understood. We sequenced the genomes of 2,101 *emm28 S. pyogenes* invasive strains, from which we selected 492 phylogenetically diverse strains for transcriptome analysis and 50 strains for virulence assessment. Data integration provided novel understanding of the virulence mechanisms of this model organism. Genome-wide association study (GWAS), expression quantitative trait loci (eQTL) analysis, machine learning, and isogenic mutant strains identified and confirmed a one-nucleotide indel in an intergenic region that significantly alters global transcript profiles and ultimately virulence. The integrative strategy we used is generally applicable to any microbe and may lead to new therapeutics for many human pathogens.

Editorial Summary

This study presents the genomes of 2,101 emm28 Streptococcus pyogenes invasive strains, of which 492 were transcriptionally profiled and 50 were assessed for virulence. GWAS, eQTL analysis, and study of isogenic mutant strains identified an intergenic region that alters global transcript profiles and bacterial virulence.

INTRODUCTION

Regardless of ecological niche or host range, all bacterial species comprise genetically diverse strains. One poorly understood area of the molecular genetics of microbes is the complex interplay between the genome, transcriptome, and virulence in large populations of infectious strains. Genetic variation may affect gene transcript levels, but the extent to which this is true and what consequences it has for pathogenesis remain unclear. Although large genomics studies have been published[1–6], far less has been done in the areas of comparative transcriptome[7–9] and virulence analyses involving natural populations[10,11]. Moreover, with the exception of one study involving a relatively small sample of strains of *Escherichia* coli[12], the relationships between the genome, transcriptome, and virulence have not been studied. Integrative analysis of diverse data from population-based strain samples may have implications for translational research efforts in areas such as vaccine formulation, new therapeutics and diagnostics, and public health.

Streptococcus pyogenes (group A streptococcus, GAS) is a strict human pathogen that causes more than 700 million infections annually in children and adults worldwide[13]. Human infections range in severity from relatively mild pharyngitis (“strep throat”) to extremely severe and life-threatening infections such as septicemia and necrotizing fasciitis/myositis, commonly known as “flesh-eating” disease. The organism also causes skin and soft-tissue infections and is responsible for post-infection sequelae such as rheumatic fever and rheumatic heart disease, important causes of morbidity globally[13,14].

GAS has been used as a model organism for studying the relationship between strain type and disease phenotype, and epidemics[1,6,15–17]. *emm28* strains are among the top five *emm*-types associated with invasive GAS infections in the USA, and many European countries[18–23]. For reasons that remain unexplained, strains of some *emm* types or M protein serotypes are non-randomly associated with particular types of human infections[17,24–30]. As an example, *emm28* GAS strains are repeatedly overrepresented among cases of puerperal sepsis (childbed fever) and neonatal infections[17,31–34].

Despite important advances in the genomics of selected organisms, little is known about the nature and extent of transcriptome diversity among clonally-related progeny of bacterial strains that have shared a recent common ancestor. Data bearing on this issue are critical for enhanced understanding of bacterial evolution in natural populations, phenotypic diversification, and microbial epidemics. To address these knowledge gaps, we sequenced the genomes of 2,101 strains of type *emm28* GAS recovered in comprehensive, population-based studies and used the resulting phylogenetic information to select representative strains for analyses of transcriptomes (n = 492 strains) and virulence (n = 50 strains). Data integration provided new understanding about the biology of this model organism, including a striking magnitude of transcriptome variation in a relatively closely related clade of organisms. Application of statistical methods and machine learning facilitated discovery of a new molecular genetic process that underpins enhanced virulence in some GAS strains.

RESULTS

Population structure and temporal distribution

We sequenced the genomes of 2,101 *emm28* GAS strains isolated from invasive infections in six countries in North America and Europe during a 26-year period, 1991 through 2016 (Table 1, Supplementary Table 1, Supplementary Fig. 1). All strains were recovered as part of comprehensive, population-based studies. The genomes were sequenced to 202-fold mean coverage[35] (Supplementary Fig. 2a, Supplementary Table 1, and Online Methods). Inference of genetic relationships were made using single nucleotide polymorphisms (SNPs) present in the core genome, that is, the genome devoid of mobile genetic elements (MGEs) such as prophages and integrative-conjugative elements (ICEs) (Supplementary Table 1). The major *emm28* GAS population was distributed into two primary clades (Clades 1 and 2) and four subclades (designated SC1A, SC1B, SC2A, and SC2B) by Bayesian clustering (Fig. 1a,b). Clade 2 organisms are differentiated from Clade 1 in part by a 28.0-Kb horizontal gene transfer (HGT) bloc that contributes 520 core SNPs and by 19 core SNPs located outside of this HGT bloc (Supplementary Table 2). This 28.0-Kb HGT bloc includes the *nga-ifs-slo* operon encoding secreted toxins NAD^+-glycohydrolase (SPN) and streptolysin O (SLO), known key contributors to GAS virulence[6,16,36–39]. *ifs* encodes an endogenous inhibitor of SPN[40]. Importantly, recombinogenic acquisition of high-expression variants of the *nga-ifs-slo* operon can increase survival in the primate upper respiratory tract, enhance virulence and trigger intercontinental epidemics[1,6,16,41]. Clade 2 organisms have an *nga-ifs-slo* region that has 99% sequence identity to the analogous three-gene operon present in *Streptococcus dysgalactiae* subspecies *equisimilis* (SDSE)[42] and likely was acquired by subclade 2A GAS via a recombination event.

SC1B comprises the most strains, accounting for 49.7% of the isolates, followed by SC2A (26.8%), SC1A (22.3%), and SC2B (0.53%). Strains belonging to subclades SC1A, SC1B, and SC2A varied by year, geographic location and MGE content (Fig. 1c and Supplementary Fig. 3). Marked temporal displacement of SC1A strains occurred concomitantly with a surge of SC2A strains in the USA, where ~55% of strains were SC2A. SC1B strains were predominantly isolated from patients in Finland, Norway, Faroe Islands, and Iceland, whereas SC1A was the prevalent subclade in Canada. MGE diversity in the cohort was assessed (Supplementary note and Supplementary Tables 3-6) and the 20 most abundant MGE-50 genotypes accounted for 90% of strains (Fig. 1c).

We next used an integrative strategy to investigate the complex interplay between genome variation, transcriptome changes and virulence differences in an animal infection model from a population perspective.

Transcriptome signatures and population structure

To determine if distinct patterns of gene expression are nonrandomly associated with the *emm28* population structure, we first conducted transcriptome (RNA-seq) analysis on a subset of 50 strains genetically representative of the three numerically dominant subclades (i.e., SC1A, SC1B, and SC2A) (Supplementary Table 7). Strains were selected for RNA-seq analysis from the main sample of 2,095 M28 strains based on the criteria described

in the Online Methods. The 50 strains are from diverse years, countries, regions within countries, and MGE content. RNA-seq analysis was conducted in triplicate (three biologic replicates) at mid-exponential (ME) and early-stationary (ES) growth phases (Online Methods, Supplementary Fig. 2b and Supplementary Table 7).

Although the 50 strains differ in genomic backgrounds, the number of strains analyzed, coupled with the extremely high correlation coefficients among the transcript levels in the triplicate samples (Supplementary Fig. 4), permitted identification of distinct transcriptome alterations with respect to the population structure (Fig. 1). We identified two strains that unexpectedly had “outlier” transcriptomes (Fig. 2a,b). Manual inspection of the genome sequence data for these two outlier strains identified two separate large deletion events in the *covS* global regulatory gene (Fig. 2a,b). Mutations in genes encoding global transcriptional regulators such as CovR/CovS, RopB, and Mga can alter substantial proportions (5–25%) of the transcriptome[43–45].

For the 46 strains with wild-type (WT) alleles in all known major regulatory genes (Fig. 2c,d), the greatest number of differentially expressed (DE) genes occurs between strains assigned to different subclades (Supplementary Fig. 5). At the mid-exponential phase, the greatest number of differentially expressed genes was observed when comparing the transcriptomes of SC2A strains to SC1A and SC1B strains (32; 2% and 15; 0.9% differentially expressed genes, respectively) (Supplementary Fig. 5a and Supplementary Table 8). A similar pattern was evident when the early-stationary transcriptomes of the three genetic subclades were compared, but the number of differentially expressed genes was considerably greater (5–9 fold) (Supplementary Fig. 5a). SC2A strains had the greatest number of differentially expressed genes compared to SC1A and SC1B strains (318; 19.9% and 83; 5.2% DE genes, respectively) (Supplementary Fig. 5a and Supplementary Table 8).

A significant proportion of the differentially expressed genes was located in the 28.0-Kb region that was horizontally transferred (HGT region) and includes the *nga-ifs-slo* operon (Supplementary Table 8). At mid-exponential phase, 35.7% (SC2A/SC1A comparison, $P < 0.0001$) and 25% (SC2A/SC1B comparison, $P < 0.001$) of genes within this HGT region (28 genes) were differentially expressed (P value assessed by Fisher exact test) and 21.4% (SC2A/SC1A comparison) and 39.3% (SC2A/SC1B comparison, $P < 0.001$) were differentially expressed at early stationary phase. The three most strongly up-regulated genes in SC2A strains compared to SC1A and SC1B strains were *nga*, *ifs*, and *slo*, with ~4-fold increase in transcript levels at mid-exponential phase and an ~8-fold increase at the early stationary phase (Supplementary Table 8 and Supplementary Fig. 5b).

Infections caused by SC1B strains have increased in recent years in several countries, including the US, Finland, Iceland, and Norway, whereas SC1A strains have decreased substantially (Supplementary Fig. 3a) raising the possibility that SC1B strains have evolved to be more fit than SC1A strains. Therefore, we inspected the genetic differences in the core chromosome (Supplementary Table 9) that differentiate all SC1A from SC1B strains and found that all SC1B strains have two contiguous nonsynonymous mutations in RivR, a negative regulator of *grab* (protein-G-related α_2-macroglobulin-binding)[46,47]. At mid-exponential phase, *grab* was the only up-regulated gene in SC1B strains compared to

SC1A and SC2A strains (Supplementary Table 8 and Supplementary Fig. 5c). Similarly, higher *grab* transcript abundance was observed at early stationary phase in SC1B and SC2A strains compared to SC1A (Supplementary Fig. 5c). GRAB is a cell-surface anchored protein that binds to α_2-macroglobulin, allowing it to retain the proteolytically active form of cysteine protease SpeB at the GAS surface, protect GAS from killing by antimicrobial peptide LL-37[48,49] and contribute to invasive infection in a mouse model[50]. We cannot rule out stochastic processes contributing to subclade displacement.

Validation of singleton (RNAtag-seq) analysis

Transcriptome analysis of bacteria traditionally has been conducted using triplicate biologic replicates of strains grown to two distinct growth phases. However, this approach currently is not economically feasible for studying many hundreds of strains. Given the improved accuracy, sensitivity, and reproducibility of RNA-seq, we hypothesized that large-scale transcriptome analysis using singleton strains (i.e., lacking replicates), in concert with optimal sequencing depth (5–10 million sequencing reads) for a pathogen with a genome size of approximately 2 Mb[51], would provide significantly enhanced understanding of the transcriptome landscape of a group of relatively closely related strains. To test our hypothesis, we used RNAtag-seq[52] to increase strain throughput for transcriptome analysis. Expression data from strains without (singletons) and with biological replicates were found to be highly correlated. (Supplementary Figure 6, Supplementary note). Thus, we proceeded with population transcriptome analysis of 442 genetically representative and diverse singleton strains (Supplementary Table 10 and Supplementary Fig. 7) chosen by *k*-means clustering statistical strategy (Online Methods).

Population transcriptome analysis of diverse strains

To examine the relationship between the transcriptomes of 442 singleton strains, we first used principal component analysis (PCA) on the normalized expression data and identified two major clusters, referred to as Cluster A ($n = 339$) and B ($n = 83$) (Fig. 3a). DBSCAN clustering validated the existence of these two clusters (Supplementary Fig. 8a). WT-like strains (strains bioinformatically assessed to have WT alleles for all major regulatory genes) were predominantly associated with Cluster A except for 10 outlier strains (Fig. 3a). Reexamination of the genome data for these outlier strains using Pilon (Online Methods) identified undetected indels in the *covS* global transcriptional regulatory gene in eight of these 10 strains. Hence, transcriptome-guided polymorphism discovery identified genetic causes underlying these aberrant transcriptomes.

Inasmuch as the transcriptomes of the eight *covS* mutant strains differed markedly from the WT strains and consistent with previous results[53–58] we hypothesized that strains with mutations in specific major global regulators have distinct underlying patterns of gene expression that could be exploited to distinguish specific classes of regulator gene mutants. To test this hypothesis, we used Random Forest (RF) machine learning[59] to determine if one of the four class labels (i.e. WT, *covR*-, *covS*- or *ropB*-mutant) could be assigned to the outlier strains with high probability. Briefly, based on analysis of the transcriptome profile of 283 singleton strains (see Online Methods), Random Forest classification was used to predict class labels for the eight outlier strains. Transcriptome-based classification correctly

identified all eight organisms as *covS* mutant strains (Supplementary Table 11). Among the 81 *covRS* and 21 *ropB* strains, 85.2 and 61.9% of the strains, respectively, were accurately classified (Supplementary Table 11). *covRS* strains misclassified as WT phenotypically (transcript profile) resemble WT strains, grouping with Cluster A strains (Fig. 3c). Thus, machine learning classification of the transcript profiles accurately predicted the genotype (regulatory gene mutation status) and predicted the transcript phenotype (mutant-like or WT-like) of strains with mutations in a major regulator gene.

Regulatory gene mutations and transcriptome changes

Re-assignment of the outlier strains as *covS* mutants resulted in Cluster A having both WT-like and mutant strains whereas Cluster B was composed exclusively of mutant strains (Fig. 3b). Inspection of transcriptomic and genomic data for Clusters A and B produced five findings. First, all Cluster B strains have mutations in *covS* or *covR*, and second, the majority of strains with either *covS* (68.5%) or *covR* (37.5%) mutations are assigned to Cluster B (Fig. 3c). Third, the great majority of strains assigned to Cluster A were WT-like or had mutations in major regulatory genes other than *covRS* (see below). Fourth, Cluster B strains had a significantly increased number of differentially expressed genes compared to Cluster A strains (Fig. 4). Fifth, no simple genomic subclade-specific association was evident with respect to the two major transcriptome clusters (Supplementary Fig. 8b).

The CovRS two-component system negatively regulates expression of 15% of the transcriptome, including key virulence factors[44]. Consistent with this, inactivation of CovRS enhances virulence[53,56]. We compared the transcriptomes of 442 strains composed of 188 predicted WT strains, 132 strains with diverse types of mutations in *covRS*, and 122 strains with varied mutations in other major regulator genes. Although Cluster B contained only *covRS* mutant strains (Fig. 3c), a sizeable proportion of *covS* (20.4%) and *covR* (45.8%) mutant strains grouped with nearly all WT-like strains in Cluster A (Fig. 3c). The finding that *covRS* mutant strains are predominantly of two distinct transcriptome clusters suggests that polymorphisms in *covRS* are not equivalent, as reported previously[57,58], and the grouping of Cluster A *covRS* mutants with WT strains suggest some polymorphisms may have fewer functional consequences than others. PCA of only *covRS* mutant strains in Cluster A ($n = 33$) and B ($n = 83$) recapitulated the grouping into two distinct clusters (Fig. 5a). We next used distance-based clustering to test the hypothesis that additional substructure not evident by PCA (Fig. 5a) was present in the transcriptome data (Fig. 5b) and the findings are reported in the Supplementary note.

To test the hypothesis that Cluster B *covRS* strains have distinctive transcriptomes compared to Cluster A *covRS* strains, we examined transcription of genes in both groups and found 142 differentially expressed genes (Supplementary Table 12). The two clusters differed in the complement of differentially expressed genes and also in magnitude of the altered transcript changes (up or down) of the differentially expressed genes. Many (32%) differentially expressed genes had 5-fold or greater altered transcript levels, including critical CovRS-regulated genes encoding virulence factors such as SPN and SLO, the Mga-regulon, HasABC and SpeB (Supplementary Table 12). Moreover, compared to Cluster A *covRS* strains, Cluster B *covRS* strains have a significantly increased frequency of frame-

shift inducing indels and nonsense mutations (Fisher exact test, two tailed $P < 0.0001$), likely to inactivate this regulatory system (loss-of-function mutations).

Based on the increase in transcripts of genes encoding multiple key virulence factors in Cluster B strains we hypothesized that Cluster B strains would be more virulent than Cluster A *covRS* mutant strains. Consistent with this hypothesis, analysis of virulence of four strains from each cluster using a mouse infection model showed that cluster B strains caused significantly higher mortality (Fig. 5c) and larger lesions with more tissue destruction (Fig. 5d). An analogous study comparing mutations in *ropB* variably affecting the expression and activity of the SpeB cysteine protease virulence factor is presented in the Supplementary note, Supplementary Fig. 8d-h, and Supplementary Tables 13-14.

A single nucleotide indel significantly alters virulence

The secreted R28 protein is a GAS virulence factor that has been studied as a potential vaccine candidate[60,61]. This protein is encoded by the Spy1336/R28 gene located on an ICE-like element annotated as the region of difference 2 (RD2)[62,63] (Fig. 6a). This 37.4 Kb segment of DNA is >99% identical to a region present in the chromosome of group B streptococci[63]. In our transcriptome study of the initial 50 strains, we observed that approximately one-third of the strains expressed low levels of Spy1336/R28 transcript, whereas two-thirds of the strains expressed high levels of Spy1336/R28 transcript. The adjacent gene (Spy1337) had the same pattern of expression (Fig. 6b). There was no correlation between Spy1336/R28 and Spy1337 transcript level and genetic structure, geographic location, year of isolation or MGE-50 genotype. This perplexing finding prompted us to conduct a genome-wide association study (GWAS) analysis using SEER[64,65] on *de novo* assemblies of all 442 strains for which we had associated transcriptome data. Based on the transcript levels of the Spy1336/R28 and Spy1337 genes, we examined if the strains with low or high-transcript phenotype were significantly associated with any genetic event (e.g., SNP, indel, recombination). For both phenotypes (high and low transcript levels), 100% of the significant *k*-mers mapped to the intergenic region between Spy1336/R28 and Spy1337 (Supplementary Fig. 10a), and this led to identification of a variant in a poly(T) homopolymeric tract located in the intergenic region between the Spy1336/R28 and Spy1337 genes (Fig. 6c). Significant *k*-mers positively associated with the high-transcript phenotype had 10T residues and negatively associated with high-transcript phenotype had 9T residues in this tract. The association of the 10T variant with increased level of transcript of Spy1336/R28 and Spy1337 also was identified by an expression quantitative trait loci (eQTL) analysis[66,67] of the 50- and 442-strain data sets (Supplementary Fig. 10b).

Compared to a parental strain (9T residues), an isogenic mutant strain (10T residues) had significantly increased transcript levels of Spy1336/R28 and Spy1337 (Fig. 6d), caused significantly larger gross and microscopic lesions, and more near-mortality, assessed in a mouse necrotizing myositis infection model (Fig. 6e,f), was significantly more resistant to killing by human polymorphonuclear leukocytes *ex vivo* ($P < 0.05$, Fig. 6g), and produced more secreted and cell-associated Spy1336/R28 protein (Fig. 6h). Altered levels of Spy1336/R28 and Spy1337 caused by variation in the number of T residues in this homopolymeric nucleotide tract was further confirmed by RNA-seq analysis of the isogenic

strains grown to mid-exponential or early stationary phase (Supplementary Table 15). Thus, insertion or deletion of a single T residue in this homopolymeric tract significantly alters the transcript levels of Spy1336/R28 and Spy1337, the transcriptome, and strain virulence.

SC2A subclade strains are more virulent in mice

The whole genome sequence and transcriptome data showed considerable differences among the *emm28* strains, and these genomic and transcriptomic changes may cause significant variation in virulence. To test this hypothesis, the virulence of 50 *emm28* strains (the same phylogenetically diverse strain set used in the initial transcriptome studies described above; Supplementary Table 7) relative to SC1A reference strain MGAS28426 was assessed in a mouse model of necrotizing myositis[16,39,68]. Virtually all of these strains (96%) were wild-type for all known major regulatory genes. As a population, the virulence of SC1A and SC1B strains did not differ significantly from one another (Fig. 7a,b). In striking contrast, SC2A strains were significantly more virulent than SC1A and SC1B strains (Fig. 7a,b). We hypothesized that the increased virulence of SC2A strains is due, at least in part, to significantly increased expression of the *nga-ifs-slo* genes, resulting in increased production of secreted NAD^+-glycohydrolase and SLO toxins by SC2A organisms, as shown for other GAS serotypes[1,6,15,16]. Consistent with this hypothesis, SC2A strains had significantly higher *nga* transcript levels and NAD^+-glycohydrolase activity compared to either SC1A or SC1B strains (Fig. 7c). The same was observed for *ifs* ($P < 0.001$, Mann-Whitney test) and *slo* ($P < 0.001$, Mann-Whitney test), two other genes in this operon.

To unambiguously demonstrate that the significantly increased virulence of SC2A strains compared to SC1A stains is due, in part, to greater *nga-ifs-slo* promoter activity, we replaced the *nga* promoter of SC2A reference strain MGAS27961 with the SC1A *nga* promoter. The isogenic mutant strain with the SC1A promoter produced significantly less NAD^+-glycohydrolase activity *in vitro* (Fig. 7d) and caused significantly less mortality and tissue destruction in a mouse necrotizing myositis infection model (Fig. 7e,f).

DISCUSSION

We used GAS as a model pathogen to investigate the complex interplay between population genomics, transcriptomics, and virulence in *emm28 S. pyogenes* strains. We discovered there was no simple correlation between the magnitude of transcriptome changes (number of differentially expressed genes) and the overall genome-to-genome genetic distance (Supplementary Fig. 9 and Supplementary note). In the aggregate, our analysis shows that a holistic approach involving multiple types of high-dimensional data applied to population-based strain samples can reveal new understanding of pathogen-host interactions not readily discovered by less integrative and comprehensive approaches. By exploiting transcriptome signature analysis, we found that ~ 5% of strains had mutations in global regulators missed by commonly employed bioinformatics methods. This finding serves as a cautionary note for studies investigating genome-transcriptome relationships.

We exploited the population-based multidimensional genome and transcriptome datasets, and statistical methods including GWAS and eQTL to identify the molecular event responsible for altering the transcript level of Spy1336/R28, a gene encoding the R28

protein virulence factor and vaccine candidate[60,63,69]. This adds to an emerging theme in bacteria that seemingly modest changes in intergenic regions can alter gene expression and be adaptive[15,70–74]. Several possibilities exist for how the increased transcript levels of Spy1336/R28 and Spy1337 enhance virulence. One possibility is that the altered-virulence phenotype is solely or predominantly caused by increased production of Spy1336/R28, a known virulence factor[60,75]. It is not yet known precisely how this protein contributes to virulence, although it has been reported to promote adhesion to human epithelial cells[60,76]. A second possibility is that the Spy1337 regulatory protein directly or indirectly alters transcription of itself and other genes that may influence virulence. To test this hypothesis, we conducted RNAseq analysis of the isogenic strains containing either 9Ts or 10Ts and found that at mid-exponential phase, only two genes (Spy1336/R28 and Spy1337) were significantly upregulated whereas at early stationary phase, Spy1336/R28 and Spy1337 and 33 other genes were upregulated, and 165 were downregulated (Supplementary Table 15). The three-gene *fruRBA* operon (Spy0641, Spy0642, and Spy0643 encoding proteins involved in fructose utilization) was highly upregulated in the 10T isogenic mutant strain compared to the 9T parental organism. Inactivation of the *fruR* or *fruB* gene significantly decreases survival of the mutant strains in human whole blood or in the presence of polymorphonuclear leukocytes[77]. Consistent with this finding, the 10T isogenic mutant strain was significantly more virulent in a mouse model of necrotizing myositis (Fig. 6e,f), had significantly enhanced resistance to killing by human polymorphonuclear neutrophils (PMNs, Fig. 6g), and produced more secreted and cell-associated Spy1336/R28 protein (Fig. 6h); see model (Fig. 6i).

The simplest hypothesis to explain how insertion or deletion of one T residue in this intergenic region alters the transcript levels of Spy1336/R28 and Spy1337 is that the transcriptional regulator encoded by Spy1337 binds directly to this intergenic region and increases transcription of both genes simultaneously. Under this hypothesis, the homopolymeric nucleotide tract might either (i) be part of, or constitute the entire Spy1337 consensus binding site, or (ii) be located in a spacer region flanked by two consensus binding sites. In the first case above a homopolymeric tract with 10Ts (compared to 9Ts) would constitute a better consensus binding site, whereas in the second case above the presence of 10Ts in the spacer region (compared to 9Ts) would place flanking putative consensus binding sites in a more favorable spatial orientation in the DNA helix for binding of the Spy1337 transcriptional regulator (Fig. 6c). Additional studies are underway to resolve this matter.

Machine learning and eQTL analysis were used in novel ways in this study. Our transcriptome dataset facilitated use of machine learning to analyze and correctly classify regulator mutant strains that were misidentified based on analysis of genome sequence data alone. Until recently the majority of studies on pathogenic microbes using machine learning had used DNA sequence-based data, commonly focused on predicting resistance to antimicrobial agents[4,78–83]. eQTL analysis has been used with expression data in humans and other eukaryotic organisms[84–90]; this study applies eQTL analysis to a bacterial dataset, made possible by the generated extensive transcriptome data. We discovered that an indel in an intergenic region was significantly associated with altered expression of five genes; two in *cis* (Spy1336/R28, Spy1337) and three in *trans* (Spy1338, Spy1339, Spy1340),

using transcript data from 50 strains in mid-exponential phase. Similarly, *cis* (Spy1336/R28, Spy1337) and *trans*-associations with 47 additional genes (FDR < 0.0005), were found using transcript data from 442 strains in early stationary phase (Supplementary Fig. 10b, Supplementary Table 16). Importantly, 60% of the 49 genes identified by eQTL analysis were also differentially expressed by RNA-seq analysis of the isogenic (10T) mutant and (9T) parental strains.

The transcriptome data also permitted us to conclude that although HGT events can and do alter the transcriptome, the vast majority of transcriptome changes are caused by SNPs (missense or nonsense mutations) and short indels that affect major regulatory genes such as *covR*/*covS*, *ropB*, and *mga* and result in truncation of the cognate encoded protein[53–55,57]. The findings are consistent with the observation that these regulatory genes are among the genes with the highest densities of polymorphisms in population genomic analyses of GAS strains[6,91,92].

For unknown reasons, *emm28* GAS strains are overrepresented among cases of puerperal sepsis (childbed fever), female genital tract infections, and neonatal infections[17,31–34]. Although our study was not designed to address the very complicated relationships between bacterial population structure and detailed clinical phenotype of the infecting strains, one observation warrants comment. Reasonably detailed infection-type information was available for the 951 isolates from patients in the U.S. We found that compared to non-SC2A strains from the U.S., a significantly higher proportion of SC2A strains from the U.S. was associated with puerperal sepsis, neonatal infections and female genital tract infections (χ^2 (1) = 5.854, P = 0.015; Supplementary Table 1). In this regard, we note that as a group, SC2A strains were also significantly more virulent in the mouse necrotizing myositis experiments (Fig. 7a, b).

In summary, our study serves as an exemplar for how multidimensional datasets generated from population-based samples can be effectively integrated to yield new knowledge about microbial genetics and pathogen-host interactions. Integration of the three different types of data resulted in a more enhanced understanding of the molecular genetics of a pathogen than study of any one or two of the three types of data. The strategy is generally applicable to any microbe, pathogenic or otherwise, and may lead to new therapeutics.

ONLINE METHODS

Whole genome sequencing and polymorphism analysis.

Strain growth, isolation of chromosomal DNA, generation of paired-end libraries, and multiplexed sequencing using an Illumina NextSeq 550 instrument (San Diego, CA) were performed as described previously [93–95]. The pipeline used for bacterial genome analysis is shown in Supplementary Fig. 2a. The trimmed sequence reads were corrected using Musket[96], and mapped to the genome of reference serotype M28 strain MGAS6180 (GenBank accession number CP000056)[97] using SMALT (see URLs). Single nucleotide

URLs
FaBox, http://users-birc.au.dk/biopv/php/fabox/;
SMALT, www.sanger.ac.uk/resources/software/smalt/;

polymorphisms (SNPs) and insertions and deletions (indels) were identified using FreeBayes (see URLs) as described[93], and Pilon[98], which was used to detect indels. SNPfx.pl (a PERL script developed in-house, see URLs) was used to determine the nature of the SNPs (coding/noncoding, synonymous/nonsynonymous, etc). Alternatively, and in conjunction with, the SPAdes algorithm was used for *de novo* genome assembly[99]. SRST2 was used to identify genes, alleles, and multi-locus sequence types (MLSTs)[100]. The algorithm Gubbins was used to detect horizontal gene transfer events[101]. hierBAPS [102] (hierarchical Bayesian Analysis of Population Structure, see URLs) was used to determine population structure and SplitsTree was used to estimate phylogenetic trees and networks[103]. hierBAPS was run with five replicates of the estimation algorithm using prior upper bound values for the number of clusters ranging between 50–200, each run converging to the same posterior mode estimate of the population structure.

Long-read sequencing of 24 strains (Supplementary Table 19) was performed using an Oxford Nanopore MinION instrument with R9.5 flowcells and the Rapid Barcoding Kit (SQK-RBK004). These strains were selected from the 50 strains for which RNAseq expression analysis was done in triplicate to be numerically representative of the major genetic clades, encompass a diversity of prophage and ICE content (MGE genotypes), and include strains that differ in the R28 promoter region T nucleotide homopolymeric tract. Hybrid assemblies of the nanopore long reads and Illumina short read data were performed with Unicycler[104], as described previously[105]. Quality metrics for read quality filtering and trimming were done with Trimmomatic, read error correction with Musket, *de novo* assemblies with SPAdes/Unicycler, MLST and *emm* determination with SRST2, read mapping with SMALT, and polymorphism discovery with FreeBayes, are listed in Supplementary Table 20.

Phylogeny among the strains was inferred by Neighbor-Joining based on concatenated sequential core chromosomal SNPs, and clades of related strains were defined by Bayesian analysis based on entire core genome sequences using hierBAPS. The reference strain for transcriptome analysis was selected using *k*-means (see below and in the Supplementary note).

Strain selection for transcriptome analysis.

To choose a subset of isolates from the sequenced population of 2,101 *emm28* strains that would approximately span as much genetic variation as possible for a given size of the subset, we used a projection-based approach, similar to the population structure correction used in the bacterial GWAS method SEER[106]. First, a single-nucleotide polymorphism (SNP)-based pairwise distance matrix was calculated separately for all *emm28* isolates belonging to a given lineage (SC1A, SC1B, SC2A, SC2B) using core SNPs. We excluded

FreeBayes, www.github.com/ekg/Freebayes/;
SNPfx.pl, https://github.com/codinghedgehog/SNPfx/;
hierBAPS, http://www.helsinki.fi/bsg/software/BAPS/;
Illumina bcl2fastq, https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2–20.html;
FASTQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc;
FASTX-toolkit, http://hannonlab.cshl.edu/fastx_toolkit/;
(NCBI) Sequence Read Archive, http://www.ncbi.nlm.nih.gov/sra;
fsm-lite, https://github.com/nvalimak/fsm-lite.

MGEs and potential regions of recombination identified by Gubbins[101]. Subsequently we sampled isolates proportionately to the fraction of the total population size represented by each lineage. The distance matrix for each lineage was then used to project all lineage isolates into a three-dimensional Euclidean space with multidimensional scaling (MDS), as in SEER[106]. For the given total size *k* of the subset to be chosen from a lineage, the *k*-means algorithm with 200 random restarts[107] was used to identify an optimal set of *k* centroids to span the variation present in the 3-dimensional MDS projection of the genetic variation present within the lineage. The isolate with the minimum Euclidean distance to each centroid was chosen to determine the final subset of *k* representative isolates.

RNA-seq library preparation and sequencing.

a) 50 *emm28* strains grown in triplicate and harvested at two time points.— 50 strains representative of the four major genetic subclades (Fig. 1a,b; Supplementary Table 7) were assayed in triplicate at mid-exponential and early-stationary phases of growth. RNA was extracted with the RNeasy kit (Qiagen) following the manufacturer's instructions. rRNA was depleted using the Ribo-Zero rRNA removal kit for Gram-positive bacteria (Illumina), as described previously[93,108]. The quality of the total RNA and rRNA-depleted RNA was evaluated with RNA Nano, and Pico chips, respectively (Agilent Technologies), and an Agilent 2100 Bioanalyzer. The cDNA libraries were prepared with indexed reverse primers from the ScriptSeq Index PCR primers kit (Illumina), and purified with AMPureXP beads (Beckman Coulter). The quality of the cDNA libraries was evaluated with High-Sensitivity DNA chips (Agilent Technologies). For each sample, the cDNA library concentration was measured fluorometrically with Qubit™ dsDNA HS assay kits (Invitrogen). The cDNA libraries were diluted, pooled, and sequenced with an Illumina NextSeq instrument. This same protocol was used for the comparative RNA-seq analysis of the 9T and 10T isogenic strains (Supplementary Table 15).

b) 461 *emm28* strains grown as singleton cultures and harvested at one time point.—Transcriptome analysis of the 461 singletons (Supplementary Table 1, 10) was performed by RNAtag-seq as described[109], with the modifications described herein. Total RNA isolated from each strain was quantified fluorometrically using the Qubit RNA BR assay kit (Life Technologies). 400 ng from each sample were fragmented for 3 min at 94°C in a volume of 16 µl in 1X FastAP buffer, dephosphorylated using FastAP alkaline phosphatase (ThermoFisher Scientific) for 12 min at 37°C in a final volume of 20 µl, and phosphorylated at the 5'-end using T4 polynucleotide kinase (T4 PNK) (New England Biolabs), for 30 min at 37°C in a final volume of 82 µl. Fragmented, dephosphorylated total RNA was purified using 2X volume (164 µl) of Agencourt RNAClean XP paramagnetic beads, according to the manufacturer's instructions, in 1.5-ml Eppendorf tubes and Dynamag-2 magnets (Invitrogen). The final elution volume was 12 µl. Pooling of the total RNAs during the RNAtag-seq procedure was enabled via ligation of barcodes such that all RNA fragments from the same strain were distinctly labeled with an individual barcode. We used 16 uniquely barcoded oligoribonucleotides described in an earlier study[109] and shown in Supplementary Table 17. For each strain 5 µl of fragmented, phosphorylated total RNAs were ligated to 1 µl of the respective oligoribonucleotide at 5 µM final concentration using T4 RNA Ligase 1 (ssRNA ligase) (New England Biolabs) in a volume of 20.1 µl. The

reaction was carried out at 22°C for 90 min. After the ligation the volume of each sample was increased to 80 μl by adding 59.9 μl RLT buffer (RNeasy mini kit, Qiagen), and mixed with a 1:1 mixture containing 80 μl of RNA binding buffer (RNA Clean & concentrator-5, Zymo Research), and 80 μl 100% ethanol in 1.5 ml Eppendorf tubes. Thus, six pools containing 8 samples each were made for each set of 48 samples by successively passing the total RNAs corresponding to the 8 strains constituting one particular pool sequentially through one Zymo column, and concentrating them together, as shown in Supplementary Fig. 7a. The final eluted volume per pool was 32 μl.

The quality of the total RNA pools was evaluated with RNA Pico chips (Agilent Technologies). We made 57 pools containing total RNA from 8 strains each, and 1 additional pool with total RNA from 5 strains, for a total of 461 strains. The Ribo-Zero rRNA removal kit (Gram-Positive Bacteria) was used to eliminate unwanted rRNAs from the pools. The quality of the ribodepleted RNA was analyzed using RNA Pico chips (Agilent Technologies). First strand cDNA synthesis was performed as described previously[109]. For each pool 12 μl of ribodepleted RNAs were mixed with 2 μl of AR2 oligonucleotide, which is complementary to a region present in all 16 barcoded oligoribonucleotides used in this study (Supplementary Table 17), denatured for 2 min at 70°C, and first-strand cDNA synthesis was performed using AffinityScript reverse transcriptase (Agilent) in a volume of 20 μl, at 55°C for 55 min. RNA was subsequently degraded in 0.09 N NaOH at 70°C for 12 min, and neutralized with acetic acid at a final concentration of 76.9 mM, in a final volume of 26 μl. After addition of 14 μl of water, single-stranded cDNAs (sscDNAs) were purified using 2.5X volume (100 μl) of Agencourt RNAClean XP paramagnetic beads, and the sscDNAs along with the beads were resuspended in 5 μl of water. While in the beads, the sscDNAs were mixed with 2 μl of 3Tr3 adapter[109], and ligated, using T4 RNA Ligase 1 in a volume of 20 μl. The reactions were incubated overnight at 22°C, followed by two consecutive cleanup reactions using 2.5X volume of Agencourt RNAClean XP beads, and eluted with 25 μl of water.

Library amplifications were performed with the universal primer univP5[109] and 1 of 4 distinct P7 barcode adapters (Supplementary Table 17). Namely, sscDNA pools were organized in sets of 4, and individually amplified in a final volume of 50 μl, after a PCR enrichment test to determine the correct amplification conditions, followed by two cleanup steps using Agencourt RNAClean XP beads, and elution in 20 μl of low TE (10 mM Tris, 0.1 mM EDTA). For each sample, the cDNA library average size was determined using High-Sensitivity DNA chips (Agilent Technologies), and the cDNA library concentration was measured fluorometrically with Qubit™ dsDNA HS assay kits (Invitrogen).

Samples were pooled one additional time at this point in the protocol. The cDNA libraries corresponding to 4 pools, each corresponding to 8 strains, were mixed together at equimolar amounts. This process was repeated 14 additional times, and thus we ended up with 15 superpools, amounting to 58 pools, and representing 461 strains (Supplementary Fig. 7a). The libraries corresponding to each superpool were individually spiked with a 10% PhiX library to improve cluster diversity and sequenced with an Illumina NextSeq instrument.

Analysis of RNA-seq data

The bioinformatics pipeline used to process RNA-seq data is presented in Supplementary Fig. 2b.

a) Analysis of 50 *emm28* strains grown in triplicate and harvested at two time points.—For each sequencing run, Illumina bcl2fastq software (see URLs) was used to convert Illumina generated BCL base call files to FASTQ files. Read quality of sequencing data was evaluated using FASTQC software (see URLs). Adapter contamination and read quality filtering was performed using Trimmomatic[110]. Reads were mapped to the genome of reference strain MGAS6180 using EDGE-pro[111] and the reads mapping to rRNA and tRNA genes were excluded from subsequent analyses. Additionally, genes with low expression were excluded from downstream analysis based on the strategy described in the Supplementary note. Differential expression analysis was performed using DESeq2[112]. This same pipeline was used for the comparative RNA-seq analysis of the isogenic strains containing either 9T or 10T in the homopolymeric tract between Spy1336 and Spy1337.

b) Analysis of 442 *emm28* singleton strains at one time point.—Demultiplexing of reads from superpools into separate pools was performed with bcl2fastq. Read quality assessment and read quality trimming were done with FASTQC and Trimmomatic, respectively. Reads from each pool were demultiplexed into separate fastq files corresponding to individual samples based on the inline barcodes using FASTX-toolkit (see URLs). Median reads per pool was 92.6 million (Supplementary Fig. 7b) and median number of reads per sample per pool ranged between 8 to 24 million (Supplementary Fig. 7c). Exclusion of low-expression genes is described in the Supplementary note. No significant batch effects were found to be associated with the expression data. As estimated using the R package variancePartition[113], the percentage of variance explained by batch effects was found to be <2%.

Differential expression analysis of the final set of 442 singleton strains was performed using NOISeq-sim implemented in the NOISeq package[114]. Differentially expressed genes were identified in each of the 441 strains compared to the reference strain MGAS28737, selected as described in the Supplementary note. DESeq2[112] was used for differential expression analysis for instances where two-group comparisons were being made, where each group comprised of more than one strain.

Machine learning using random forest analysis.

The random forest analysis was done with MATLAB using the function TreeBagger (MATLAB R2016b - Statistics and Machine Learning Toolbox, The Mathworks, Natick, MA, USA). The aim was to train a Random Forest[115] for classification of outlier strains into four categories: *covR*, *covS*, r*opB* and WT based on the transcriptome profile of the strains. Random forest was trained with transcriptome data generated for strains with mutations in only a single major global regulatory gene (*covR*, *covS* or *ropB*) and strains known to be WT-for all known major regulatory genes ($n = 283$). Hence, the training data consisted of 283 strains, for which the transcriptome profiles over the 1,614 genes and the class labels were known. The test data consisted of 8 outlier strains, for which the transcriptome profiles

over the 1,614 genes were known but the class labels were unknown. Prior to learning the final model, the following feature-selection procedure was applied. An initial random forest (1,000 trees) over all genes was built and the predictive importance of the genes was estimated using the built-in measure for feature importance. This was repeated ten times and the final feature importance values were taken as the average from the individual runs. Starting with the most important feature, we successively included more features according to the given order of importance. For each subset of features, we performed a two-fold cross-validation, for which a random forest (100 trees) was built using two-thirds of the training data and evaluated on one-third of the training data. The out-of-sample performance of the sub-models was measured by the average out-of-sample classification accuracy over 100 cross-validation iterations. The increase in classification accuracy quickly plateaued as more features were added and, based on this, the 10 most informative genes were selected for the final model. The final model was used to predict the class probabilities of the 8-outlier strains (Supplementary Table 11).

Genome-wide association analysis.

Genome-wide association analysis (GWAS) using SEER[106] was performed with the *de novo* assemblies of 442 strains for which we also had RNA-seq data. GWAS was used to identify genetic variant(s) significantly associated with a binary phenotypic grouping (high transcript expression = 1 or low transcript expression = 0), defined based on transcript levels of the Spy1336/R28 gene. Plotting of the normalized transcript level (counts) of the Spy1336/R28 gene for the 442 strains, resulted in two visually very distinct groups - low (1/3rd strains) and high (2/3rd strains) expressers, analogous to our observations for the 50 strains (Fig. 6b). We identified the threshold and found strains with less than 261.5 normalized counts were considered low-expressers (coded as 0) and strains with equal to or greater than 261.5 counts were considered high expressers (coded as 1). The binary phenotype file and *de novo* assembled fasta files were supplied to SEER and the *k*-mers were counted from assembled reads using fsm-lite (see URLs). To account for the population structure, the distance matrix computed by Mash[116] was used. Running SEER yielded 17 and 13 significant k-mers (adjusted *P*value < 10^{-8}) that were positively or negatively associated with high or low transcript expression, respectively.

eQTL analysis.

Expression quantitative trait loci (eQTL) analysis was performed using the R package Matrix eQTL[117]. For eQTL analysis, population structure was accounted for in the model by using the top ten principal components as covariates. Associations were considered in *cis* if the polymorphism was within 1 Kb of the gene under consideration.

Virulence studies of 50 naturally occurring and isogenic serotype M28 GAS mutant strains using a mouse model of necrotizing myositis.

Mouse necrotizing myositis studies with serotype M28 GAS strains were performed as described previously[118,119]. The 50 naturally occurring strains used in the initial RNA-seq experiment, including virulence reference strain MGAS28426, were used. Strain MGAS28426 was used as the virulence reference strain because it is genomically representative of SC1A strains and was used as a wild-type reference for the secreted

NAD^+-glycohydrolase assays. Frozen stocks of each strain were prepared and quantified by counting colony forming units (CFUs) recovered from thawed cultures after serial dilutions were made. Immunocompetent 4-week old female outbred CD1 mice (Envigo Laboratories) were randomly assigned to strain treatment groups and inoculated in the right lower hind limb with 5 X 10^8 CFUs of each bacterial strain (n = 20 mice/strain; 1,000 mice total). This dose was selected based on two pilot experiments that showed the virulence reference strain MGAS28426 caused approximately 50% near-mortality at an inoculation dose of 5 X 10^8 CFUs. This strategy facilitated identification of comparator strains with significantly increased or decreased virulence. The mouse sample size was selected using a power calculation with the following variables: $\alpha = 0.05$, power (1-ß) = 0.8, difference in survival rates between groups = 0.4, and ratio of group size = 1.

For the CovRS mutant strain comparison, four cluster A strains and four cluster B strains (n = 45 mice per strain) were used at a dose of 5 X 10^8 CFU. For the parental wildtype (27961) and isogenic promoter mutant (27961-SC1A-*nga*-promoter) comparison (n = 40 mice per strain), 5 X 10^8 CFU were used. For the RopB mutant strain comparison, three Group I and four Group II strains (n = 40 mice per strain) were used at a dose of 1 X 10^9 CFU. For the 9T and the 10T isogenic strains, 27961–9T and 27961–10T, and the control strain 28085–10T (n = 20 mice per strain) were used at a dose of 5 X 10^8 CFU. Representative gross and microscopic images of limbs taken from mice assigned to histopathology analysis were obtained. Oligonucleotides used to create the isogenic mutants are shown in Supplementary Table 18 (see the Supplementary note).

All animal studies were performed in accordance with a protocol (AUP-0615–0041) reviewed and approved by the Institutional Care and Use Committee, The Methodist Hospital Research Institute, Houston, TX. Mice were monitored at least once daily, and near-mortality was determined with internationally recognized criteria guidelines provided by the National Research Council (US) Committee for the Update of the Guide for the Care and Use of Laboratory Animals 2011, and the Guide for the Care and Use of Laboratory Animals, 8th ed. National Academies Press, Washington, DC. Survival data were expressed as Kaplan-Meier curves, and statistically significant differences were determined with the log-rank test (Prism6, GraphPad Software).

Statistical analysis.

Unless otherwise stated, error bars represent SD, and P values were calculated using Fisher exact, Mann-Whitney, or log-rank tests. False discovery rate (FDR) was used as reported by the MatrixEQTL package. Bayesian clustering was used to define clades and subclades in the *emm28* population. k-means and distance-based clustering were used to identify centroids in a two-dimensional space clustering of strains generated by principal component analysis (PCA) and find additional substructure in clusters, respectively. The random forest analysis was done with MATLAB using the function TreeBagger. The R package variancePartition was used to confirm the absence of significant batch effects. R-squared (R^2) statistics was used to investigate if a correlation existed between genetic distance and extent of transcriptome remodeling. A one-tailed test of proportions was used establish that Group II RopB strains contained a significant proportion of mutations affecting functional

domains. The Pearson correlation coefficient (r) was used to compare RNA-seq data from strains analyzed in triplicate to RNAtag-seq data collected from strains grown as singletons.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This study was supported in part by the Fondren Foundation, Houston Methodist Hospital and Research Institute (to J.M.M.), the Academy of Finland (grant 255636) (to J.V.), a European Research Council grant number 742158 (to J.C.), and National Institutes of Health grant 1R01AI109096-01A1 (to M.K). This research was also supported in part by the Intramural Research Program of the National Institute of Allergy and Infectious Disease, National Institutes of Health (to F.R.D.). We thank Neal Copeland, Nancy Jenkins and David Ginsburg for critical comments and suggestions to improve the manuscript; Kathryn Stockbauer for critical comments and editorial assistance; Edward Graviss, Helga Erlendsdottir, Willa Hong, and Sarah Linson for technical assistance; Hanne-Leena Hyyryläinen, Jari Jalava, and the Finnish clinical microbiology laboratories; Alexander A. Shishkin for helpful suggestions regarding RNAtag-seq protocol; Marija Todorovic and Jenny Jonsdottir Nielsen for banking strains from the Faroe Islands; Allison McGeer for Ontario strains; Chris Van Beneden, Bernard Beall, and the Active Bacterial Core Surveillance of the CDC's Emerging Infections Programs network; Anne Ramstad Alme and Anne Witsø for technical assistance; and Martin Steinbakk, Norwegian Laboratory for Streptococci, for support.

DATA AVAILABILITY.

Whole-genome sequencing data for the 2,101 isolates studied were deposited in the NCBI Sequence Read Archive with 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.** The data that support the findings of this study are available from the corresponding author upon request.

REFERENCES

1. Beres SB et al. Transcriptome Remodeling Contributes to Epidemic Disease Caused by the Human Pathogen Streptococcus pyogenes. MBio 7(2016).
2. Chewapreecha C. et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet 10, e1004547 (2014).
3. Fernandez-Romero N. et al. Uncoupling between core genome and virulome in extraintestinal pathogenic Escherichia coli. Can J Microbiol 61, 647–52 (2015). [PubMed: 26063294]
4. Long SW et al. Population Genomic Analysis of 1,777 Extended-Spectrum Beta-Lactamase-Producing Klebsiella pneumoniae Isolates, Houston, Texas: Unexpected Abundance of Clonal Group 307. MBio 8(2017).
5. Mukherjee S. et al. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol 35, 676–683 (2017). [PubMed: 28604660]
6. Nasser W. et al. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. Proc Natl Acad Sci U S A 111, E1768–76 (2014).
7. Bruchmann S. et al. Deep transcriptome profiling of clinical Klebsiella pneumoniae isolates reveals strain and sequence type-specific adaptation. Environ Microbiol 17, 4690–710 (2015). [PubMed: 26261087]
8. Dotsch A. et al. The Pseudomonas aeruginosa Transcriptional Landscape Is Shaped by Environmental Heterogeneity and Genetic Variation. MBio 6, e00749 (2015).
9. Sharma-Kuinkel BK et al. Potential Influence of Staphylococcus aureus Clonal Complex 30 Genotype and Transcriptome on Hematogenous Infections. Open Forum Infect Dis 2, ofv093 (2015).
10. Felek S, Tsang TM & Krukonis ES Three Yersinia pestis adhesins facilitate Yop delivery to eukaryotic cells and contribute to plague virulence. Infect Immun 78, 4134–50 (2010). [PubMed: 20679446]

11. Swearingen MC, Porwollik S, Desai PT, McClelland M. & Ahmer BM Virulence of 32 Salmonella strains in mice. PLoS One 7, e36043 (2012).
12. Schreiber H.L.t. et al. Bacterial virulence phenotypes of Escherichia coli and host susceptibility determine risk for urinary tract infections. Sci Transl Med 9(2017).
13. Carapetis JR, Steer AC, Mulholland EK & Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis 5, 685–94 (2005). [PubMed: 16253886]
14. Carapetis JR et al. Acute rheumatic fever and rheumatic heart disease. Nat Rev Dis Primers 2, 15084 (2016).
15. Zhu L. et al. A molecular trigger for intercontinental epidemics of group A Streptococcus. J Clin Invest 125, 3545–59 (2015). [PubMed: 26258415]
16. Zhu L, Olsen RJ, Nasser W, de la Riva Morales I. & Musser JM Trading Capsule for Increased Cytotoxin Production: Contribution to Virulence of a Newly Emerged Clade of emm89 Streptococcus pyogenes. MBio 6, e01378–15 (2015).
17. Colman G, Tanna A, Efstratiou A. & Gaworzewska ET The serotypes of Streptococcus pyogenes present in Britain during 1980–1990 and their association with disease. J Med Microbiol 39, 165–78 (1993). [PubMed: 8366514]
18. Gherardi G, Vitali LA & Creti R. Prevalent emm Types among Invasive GAS in Europe and North America since Year 2000. Front Public Health 6, 59 (2018). [PubMed: 29662874]
19. Smit PW et al. Epidemiology and emm types of invasive group A streptococcal infections in Finland, 2008–2013. Eur J Clin Microbiol Infect Dis 34, 2131–6 (2015). [PubMed: 26292935]
20. Ikebe T. et al. Increased prevalence of group A streptococcus isolates in streptococcal toxic shock syndrome cases in Japan from 2010 to 2012. Epidemiol Infect 143, 864–72 (2015). [PubMed: 25703404]
21. Naseer U, Steinbakk M, Blystad H. & Caugant DA Epidemiology of invasive group A streptococcal infections in Norway 2010–2014: A retrospective cohort study. Eur J Clin Microbiol Infect Dis 35, 1639–48 (2016). [PubMed: 27311458]
22. Nelson GE et al. Epidemiology of Invasive Group A Streptococcal Infections in the United States, 2005–2012. Clin Infect Dis 63, 478–86 (2016). [PubMed: 27105747]
23. Plainvert C. et al. Invasive group A streptococcal infections in adults, France (2006–2010). Clin Microbiol Infect 18, 702–10 (2012). [PubMed: 21883669]
24. Al-Shahib A. et al. Emergence of a novel lineage containing a prophage in emm/M3 group A Streptococcus associated with upsurge in invasive disease in the UK. Microb Genom 2, e000059 (2016).
25. Davies MR et al. Emergence of scarlet fever Streptococcus pyogenes emm12 clones in Hong Kong is associated with toxin acquisition and multidrug resistance. Nat Genet 47, 84–7 (2015). [PubMed: 25401300]
26. Fittipaldi N. et al. Full-genome dissection of an epidemic of severe invasive disease caused by a hypervirulent, recently emerged clone of group A Streptococcus. Am J Pathol 180, 1522–34 (2012). [PubMed: 22330677]
27. Hamilton SM, Stevens DL & Bryant AE Pregnancy-related group a streptococcal infections: temporal relationships between bacterial acquisition, infection onset, clinical findings, and outcome. Clin Infect Dis 57, 870–6 (2013). [PubMed: 23645851]
28. Johnson DR, Stevens DL & Kaplan EL Epidemiologic analysis of group A streptococcal serotypes associated with severe systemic infections, rheumatic fever, or uncomplicated pharyngitis. J Infect Dis 166, 374–82 (1992). [PubMed: 1634809]
29. Shea PR et al. Group A Streptococcus emm gene types in pharyngeal isolates, Ontario, Canada, 2002–2010. Emerg Infect Dis 17, 2010–7 (2011). [PubMed: 22099088]
30. Smoot JC et al. Genome sequence and comparative microarray analysis of serotype M18 group A Streptococcus strains associated with acute rheumatic fever outbreaks. Proc Natl Acad Sci U S A 99, 4668–73 (2002). [PubMed: 11917108]
31. Ben Zakour NL, Venturini C, Beatson SA & Walker MJ Analysis of a Streptococcus pyogenes puerperal sepsis cluster by use of whole-genome sequencing. J Clin Microbiol 50, 2224–8 (2012). [PubMed: 22518858]

32. Chuang I, Van Beneden C, Beall B. & Schuchat A. Population-based surveillance for postpartum invasive group a streptococcus infections, 1995–2000. Clin Infect Dis 35, 665–70 (2002). [PubMed: 12203162]
33. Gaworzewska E. & Colman G. Changes in the pattern of infection caused by Streptococcus pyogenes. Epidemiol Infect 100, 257–69 (1988). [PubMed: 3128449]
34. Raymond J, Schlegel L, Garnier F. & Bouvet A. Molecular characterization of Streptococcus pyogenes isolates to investigate an outbreak of puerperal sepsis. Infect Control Hosp Epidemiol 26, 455–61 (2005). [PubMed: 15954483]
35. Sims D, Sudbery I, Ilott NE, Heger A. & Ponting CP Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121–32 (2014). [PubMed: 24434847]
36. Bricker AL, Carey VJ & Wessels MR Role of NADase in virulence in experimental invasive group A streptococcal infection. Infect Immun 73, 6562–6 (2005). [PubMed: 16177331]
37. Bricker AL, Cywes C, Ashbaugh CD & Wessels MR NAD+-glycohydrolase acts as an intracellular toxin to enhance the extracellular survival of group A streptococci. Mol Microbiol 44, 257–69 (2002). [PubMed: 11967084]
38. Sumby P. et al. Evolutionary origin and emergence of a highly successful clone of serotype M1 group a Streptococcus involved multiple horizontal gene transfer events. J Infect Dis 192, 771–82 (2005). [PubMed: 16088826]
39. Zhu L. et al. Contribution of Secreted NADase and Streptolysin O to the Pathogenesis of Epidemic Serotype M1 Streptococcus pyogenes Infections. Am J Pathol 187, 605–613 (2017). [PubMed: 28034602]
40. Meehl MA, Pinkner JS, Anderson PJ, Hultgren SJ & Caparon MG A novel endogenous inhibitor of the secreted streptococcal NAD-glycohydrolase. PLoS Pathog 1, e35 (2005).
41. Tatsuno I. et al. Characterization of the NAD-glycohydrolase in streptococcal strains. Microbiology 153, 4253–60 (2007). [PubMed: 18048938]
42. Shimomura Y. et al. Complete genome sequencing and analysis of a Lancefield group G Streptococcus dysgalactiae subsp. equisimilis strain causing streptococcal toxic shock syndrome (STSS). BMC Genomics 12, 17 (2011). [PubMed: 21223537]
43. Carroll RK et al. Naturally occurring single amino acid replacements in a regulatory protein alter streptococcal gene expression and virulence in mice. J Clin Invest 121, 1956–68 (2011). [PubMed: 21490401]
44. Graham MR et al. Virulence control in group A Streptococcus by a two-component gene regulatory system: global expression profiling and in vivo infection modeling. Proc Natl Acad Sci U S A 99, 13855–60 (2002).
45. Ribardo DA & McIver KS Defining the Mga regulon: Comparative transcriptome analysis reveals both direct and indirect regulation by Mga in the group A streptococcus. Mol Microbiol 62, 491–508 (2006). [PubMed: 16965517]
46. Ramalinga A, Danger JL, Makthal N, Kumaraswami M. & Sumby P. Multimerization of the Virulence-Enhancing Group A Streptococcus Transcription Factor RivR Is Required for Regulatory Activity. J Bacteriol 199(2017).
47. Trevino J, Liu Z, Cao TN, Ramirez-Pena E. & Sumby P. RivR is a negative regulator of virulence factor expression in group A Streptococcus. Infect Immun 81, 364–72 (2013). [PubMed: 23147037]
48. Nyberg P, Rasmussen M. & Bjorck L. alpha2-Macroglobulin-proteinase complexes protect Streptococcus pyogenes from killing by the antimicrobial peptide LL-37. J Biol Chem 279, 52820–3 (2004).
49. Rasmussen M, Muller HP & Bjorck L. Protein GRAB of streptococcus pyogenes regulates proteolysis at the bacterial surface by binding alpha2-macroglobulin. J Biol Chem 274, 15336–44 (1999).
50. Toppel AW, Rasmussen M, Rohde M, Medina E. & Chhatwal GS Contribution of protein G-related alpha2-macroglobulin-binding protein to bacterial virulence in a mouse skin model of group A streptococcal infection. J Infect Dis 187, 1694–703 (2003). [PubMed: 12751026]
51. Haas BJ, Chin M, Nusbaum C, Birren BW & Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC Genomics 13, 734 (2012). [PubMed: 23270466]

52. Shishkin AA et al. Simultaneous generation of many RNA-seq libraries in a single reaction. Nat Methods 12, 323–5 (2015). [PubMed: 25730492]
53. Engleberg NC, Heath A, Miller A, Rivera C. & DiRita VJ Spontaneous mutations in the CsrRS two-component regulatory system of Streptococcus pyogenes result in enhanced virulence in a murine model of skin and soft tissue infection. J Infect Dis 183, 1043–54 (2001). [PubMed: 11237829]
54. Li J. et al. Neutrophils select hypervirulent CovRS mutants of M1T1 group A Streptococcus during subcutaneous infection of mice. Infect Immun 82, 1579–90 (2014). [PubMed: 24452689]
55. Mayfield JA et al. Mutations in the control of virulence sensor gene from Streptococcus pyogenes after infection in mice lead to clonal bacterial variants with altered gene regulatory activity and virulence. PLoS One 9, e100698 (2014).
56. Sumby P, Whitney AR, Graviss EA, DeLeo FR & Musser JM Genome-wide analysis of group a streptococci reveals a mutation that modulates global phenotype and disease specificity. PLoS Pathog 2, e5 (2006).
57. Tatsuno I, Okada R, Zhang Y, Isaka M. & Hasegawa T. Partial loss of CovS function in Streptococcus pyogenes causes severe invasive disease. BMC Res Notes 6, 126 (2013). [PubMed: 23537349]
58. Trevino J. et al. CovS simultaneously activates and inhibits the CovR-mediated repression of distinct subsets of group A Streptococcus virulence factor-encoding genes. Infect Immun 77, 3141–9 (2009). [PubMed: 19451242]
59. Breiman L. Random Forests. Mach. Learn 45, 5–32 (2001).
60. Stalhammar-Carlemalm M, Areschoug T, Larsson C. & Lindahl G. The R28 protein of Streptococcus pyogenes is related to several group B streptococcal surface proteins, confers protective immunity and promotes binding to human epithelial cells. Mol Microbiol 33, 208–19 (1999). [PubMed: 10411737]
61. Stalhammar-Carlemalm M, Stenberg L. & Lindahl G. Protein rib: a novel group B streptococcal cell surface protein that confers protective immunity and is expressed by most strains causing invasive infections. J Exp Med 177, 1593–603 (1993). [PubMed: 8496678]
62. Beres SB & Musser JM Contribution of exogenous genetic elements to the group A Streptococcus metagenome. PLoS One 2, e800 (2007).
63. Green NM et al. Genome sequence of a serotype M28 strain of group a streptococcus: potential new insights into puerperal sepsis and bacterial disease specificity. J Infect Dis 192, 760–70 (2005). [PubMed: 16088825]
64. Coll F. et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet 50, 307–316 (2018). [PubMed: 29358649]
65. Earle SG et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol 1, 16041 (2016).
66. Gibson G, Powell JE & Marigorta UM Expression quantitative trait locus analysis for translational medicine. Genome Med 7, 60 (2015). [PubMed: 26110023]
67. Nicolae DL et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6, e1000888 (2010).
68. Olsen RJ & Musser JM Molecular pathogenesis of necrotizing fasciitis. Annu Rev Pathol 5, 1–31 (2010). [PubMed: 19737105]
69. Rodriguez-Ortega MJ et al. Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. Nat Biotechnol 24, 191–7 (2006). [PubMed: 16415855]
70. Zhu L. et al. Intergenic Variable-Number Tandem-Repeat Polymorphism Upstream of rocA Alters Toxin Production and Enhances Virulence in Streptococcus pyogenes. Infect Immun 84, 2086–93 (2016). [PubMed: 27141081]
71. Hammarlof DL et al. Role of a single noncoding nucleotide in the evolution of an epidemic African clade of Salmonella. Proc Natl Acad Sci U S A (2018).
72. Blount ZD, Barrick JE, Davidson CJ & Lenski RE Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature 489, 513–8 (2012). [PubMed: 22992527]

73. Zaunbrecher MA, Sikes RD Jr., Metchock B, Shinnick TM & Posey JE Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 106, 20004–9 (2009).

74. Puopolo KM & Madoff LC Upstream short sequence repeats regulate expression of the alpha C protein of group B Streptococcus. Mol Microbiol 50, 977–91 (2003). [PubMed: 14617155]

75. Stalhammar-Carlemalm M, Areschoug T, Larsson C. & Lindahl G. Cross-protection between group A and group B streptococci due to cross-reacting surface proteins. J Infect Dis 182, 142–9 (2000). [PubMed: 10882591]

76. Weckel A. et al. The N-terminal domain of the R28 protein promotes emm28 group A Streptococcus adhesion to host cells via direct binding to three integrins. J Biol Chem 293, 16006–16018 (2018).

77. Valdes KM et al. The fruRBA Operon Is Necessary for Group A Streptococcal Growth in Fructose and for Resistance to Neutrophil Killing during Growth in Whole Human Blood. Infect Immun 84, 1016–31 (2016). [PubMed: 26787724]

78. Jeukens J. et al. Genomics of antibiotic-resistance prediction in Pseudomonas aeruginosa. Ann N Y Acad Sci (2017).

79. Nguyen M. et al. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci Rep 8, 421 (2018). [PubMed: 29323230]

80. Pesesky MW et al. Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. Front Microbiol 7, 1887 (2016). [PubMed: 27965630]

81. Rishishwar L, Petit RA 3rd, Kraft CS & Jordan IK Genome sequence-based discriminator for vancomycin-intermediate Staphylococcus aureus. J Bacteriol 196, 940–8 (2014). [PubMed: 24363339]

82. Li Y. et al. Validation of beta-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. BMC Genomics 18, 621 (2017). [PubMed: 28810827]

83. Li Y. et al. Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting beta-Lactam Resistance Levels in Streptococcus pneumoniae. MBio 7(2016).

84. Hao K. et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet 8, e1003029 (2012).

85. Naranbhai V. et al. Genomic modulators of gene expression in human neutrophils. Nat Commun 6, 7545 (2015). [PubMed: 26151758]

86. Ongen H. et al. Estimating the causal tissues for complex traits and diseases. Nat Genet 49, 1676–1683 (2017). [PubMed: 29058715]

87. Tung J, Zhou X, Alberts SC, Stephens M. & Gilad Y. The genetic architecture of gene expression levels in wild baboons. Elife 4(2015).

88. Albert FW, Treusch S, Shockley AH, Bloom JS & Kruglyak L. Genetics of single-cell protein abundance variation in large yeast populations. Nature 506, 494–7 (2014). [PubMed: 24402228]

89. Parker CC et al. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat Genet 48, 919–26 (2016). [PubMed: 27376237]

90. Francesconi M. & Lehner B. The effects of genetic variation on gene expression dynamics during development. Nature 505, 208–11 (2014). [PubMed: 24270809]

91. Beres SB et al. Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. Proc Natl Acad Sci U S A 107, 4371–6 (2010). [PubMed: 20142485]

92. Olsen RJ et al. The majority of 9,729 group A streptococcus strains causing disease secrete SpeB cysteine protease: pathogenesis implications. Infect Immun 83, 4750–8 (2015). [PubMed: 26416912]

METHODS REFERENCES

93. Beres SB et al. Transcriptome Remodeling Contributes to Epidemic Disease Caused by the Human Pathogen Streptococcus pyogenes. MBio 7(2016).

94. Nasser W. et al. Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. Proc Natl Acad Sci U S A 111, E1768–76 (2014).
95. Beres SB et al. Genome sequence analysis of emm89 Streptococcus pyogenes strains causing infections in Scotland, 2010–2016. J Med Microbiol 66, 1765–1773 (2017). [PubMed: 29099690]
96. Liu Y, Schroder J. & Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics 29, 308–15 (2013). [PubMed: 23202746]
97. Green NM et al. Genome sequence of a serotype M28 strain of group a streptococcus: potential new insights into puerperal sepsis and bacterial disease specificity. J Infect Dis 192, 760–70 (2005). [PubMed: 16088825]
98. Walker BJ et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963 (2014).
99. Bankevich A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19, 455–77 (2012). [PubMed: 22506599]
100. Inouye M. et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 6, 90 (2014). [PubMed: 25422674]
101. Croucher NJ et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43, e15 (2015).
102. Cheng L, Connor TR, Siren J, Aanensen DM & Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30, 1224–8 (2013). [PubMed: 23408797]
103. Huson DH SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68–73 (1998). [PubMed: 9520503]
104. Wick RR, Judd LM, Gorrie CL & Holt KE Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13, e1005595 (2017).
105. Long SW, Kachroo P, Musser JM & Olsen RJ Whole-Genome Sequencing of a Human Clinical Isolate of emm28 Streptococcus pyogenes Causing Necrotizing Fasciitis Acquired Contemporaneously with Hurricane Harvey. Genome Announc 5(2017).
106. Lees JA et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun 7, 12797 (2016).
107. Bishop C. Pattern Recognition and Machine Learning (Springer, New York, 2006).
108. Eraso JM et al. Genomic Landscape of Intrahost Variation in Group A Streptococcus: Repeated and Abundant Mutational Inactivation of the fabT Gene Encoding a Regulator of Fatty Acid Synthesis. Infect Immun 84, 3268–3281 (2016). [PubMed: 27600505]
109. Shishkin AA et al. Simultaneous generation of many RNA-seq libraries in a single reaction. Nat Methods 12, 323–5 (2015). [PubMed: 25730492]
110. Bolger AM, Lohse M. & Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–20 (2014). [PubMed: 24695404]
111. Magoc T, Wood D. & Salzberg SL EDGE-pro: Estimated Degree of Gene Expression in Prokaryotic Genomes. Evol Bioinform Online 9, 127–36 (2013). [PubMed: 23531787]
112. Love MI, Huber W. & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014). [PubMed: 25516281]
113. Hoffman GE & Schadt EE variancePartition: interpreting drivers of variation in complex gene expression studies. BMC Bioinformatics 17, 483 (2016). [PubMed: 27884101]
114. Tarazona S. et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 43, e140 (2015).
115. Breiman L. Random Forests. Mach. Learn 45, 5–32 (2001).
116. Ondov BD et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132 (2016). [PubMed: 27323842]
117. Shabalin AA Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–8 (2012). [PubMed: 22492648]
118. Olsen RJ & Musser JM Molecular pathogenesis of necrotizing fasciitis. Annu Rev Pathol 5, 1–31 (2010). [PubMed: 19737105]

119. Zhu L. et al. Contribution of Secreted NADase and Streptolysin O to the Pathogenesis of Epidemic Serotype M1 Streptococcus pyogenes Infections. Am J Pathol 187, 605–613 (2017). [PubMed: 28034602]

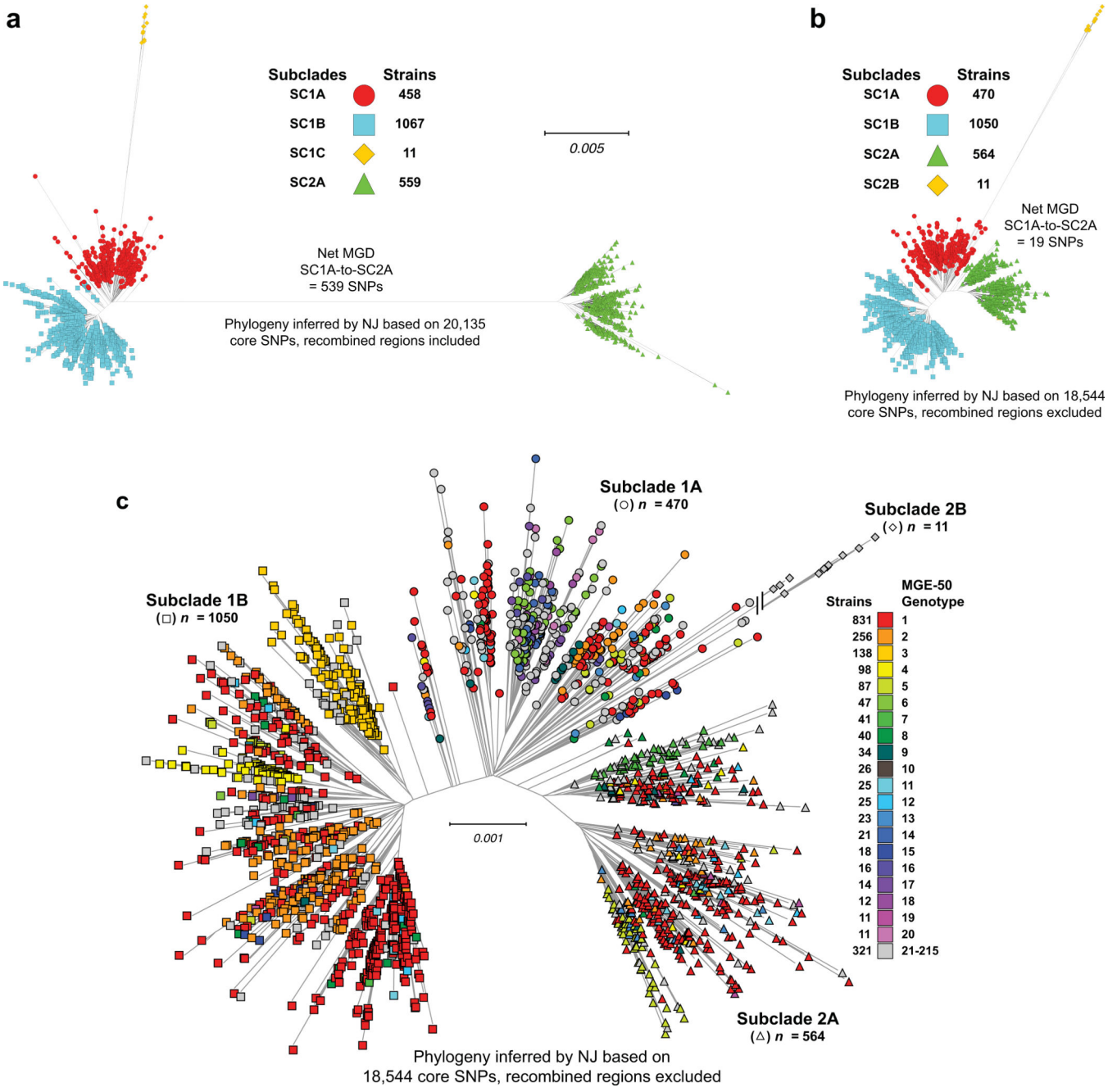

Figure 1. Population genetic structure for 2,095 *S. pyogenes emm28* invasive infection isolates.

(**a**) Genetic relationships inferred without correction for horizontal gene transfer (HGT) and recombination events. Four genetic subclades (SC1A, SC1B, SC1C, and SC2A) inferred by BAPS are shown. (**b**) Genetic relationships inferred with correction for HGT and recombination events using Gubbins. hierBAPS was used to infer genetic subclades (SC1A, SC1B, SC2A, and SC2B) within the population after exclusion of recombination events. Post exclusion of HGT and recombination events, SC1C strains (n = 11; panel a), a distinct genetic lineage of *emm*28 strains, were inferred as SC2B strains by hierBAPS. (**c**) Mobile genetic element genotypes (MGE-50) were defined based on the presence/absence of alleles

for 50 MGE-encoded site-specific integrase (n = 31) and secreted virulence factor (n = 19) genes detected using SRST2 as described in the Supplementary note, and presented in Supplementary Tables 1, 3-6. Illustrated are the 20 most abundant MGE-50 genotypes, each present in 10 or more strains and cumulatively accounting for 90% of the total strain sample. Trees in panels **a** and **b** are shown at the same scale. Strains are colored by clades and mobile genetic element genotype (MGE-50) as indicated in the panel insets. NJ, Neighbor Joining.

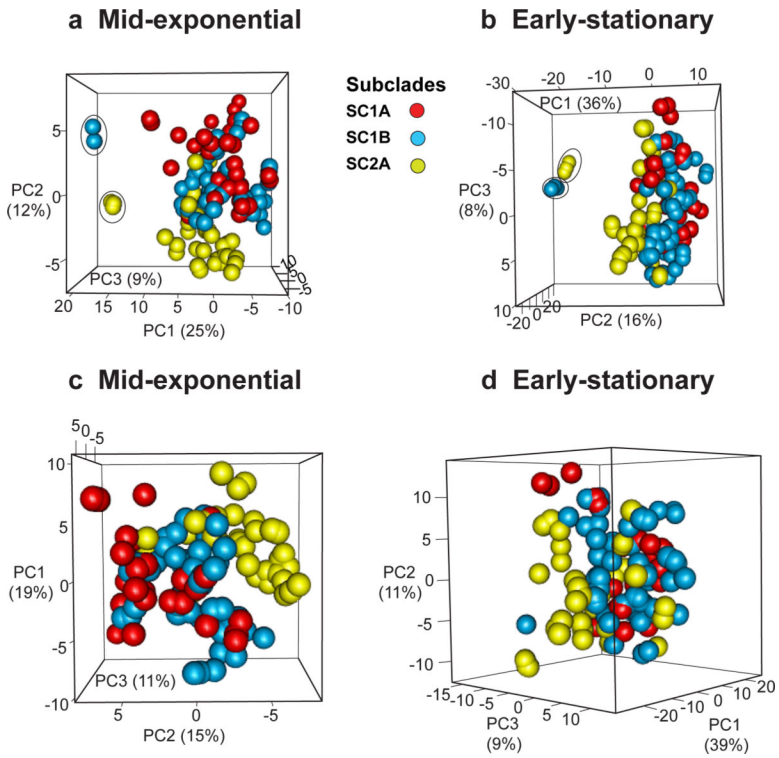

Figure 2. Transcriptome analysis of 50 strains with three biological replicates at two time points.

Principal component analysis of transcriptome data for 50 strains at mid-exponential (**a**) and early-stationary (**b**) phases of growth. Highlighted within ovals are two strains with deletion frameshift mutations in *covS* that group distinctly away from the other 48 strains at both mid-exponential and early-stationary phase (**a & b**). Based on the global transcriptome profile of WT-like strains, SC2A isolates clustered together compared to SC1A or SC1B strains (**c & d**). Red, SC1A; blue, SC1B; and yellow, SC2A.

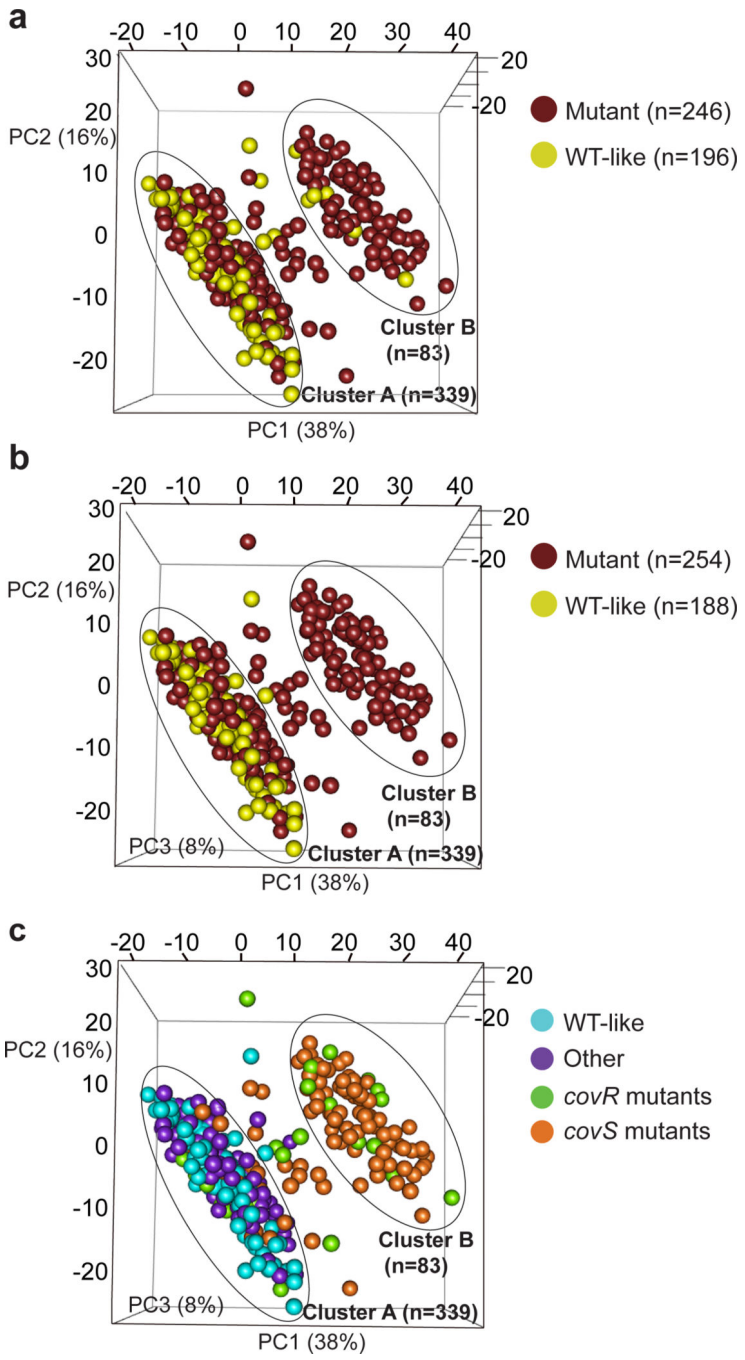

Figure 3. 442 singleton strains partition into two major transcriptome clusters based on their genome-wide expression profiles.

(**a**) Figure depicts three-dimensional PCA plots displaying variation in the transcriptome data along the top three principal components (PCs). The greatest variance in transcriptome data along PC1 (38%) separates strains into two major clusters, arbitrarily designated Cluster A and B. Wild-type (WT)-like strains are predominantly associated with Cluster A. The genome data were reexamined for 10 outlier WT-like strains that did not group with Cluster A WT-like strains. (**b**) 8 of the 10 outlier WT-like strains were re-assigned to

Cluster B after identification of previously missed polymorphisms, resulting in Cluster B containing mutant strains exclusively whereas Cluster A is comprised of both mutant and WT-like strains. (**c**) Cluster B is composed exclusively of strains with mutations in covR or covS. WT-like strains and strains with mutations in genes encoding regulators other than covR/covS (designated Other) predominantly grouped into Cluster A.

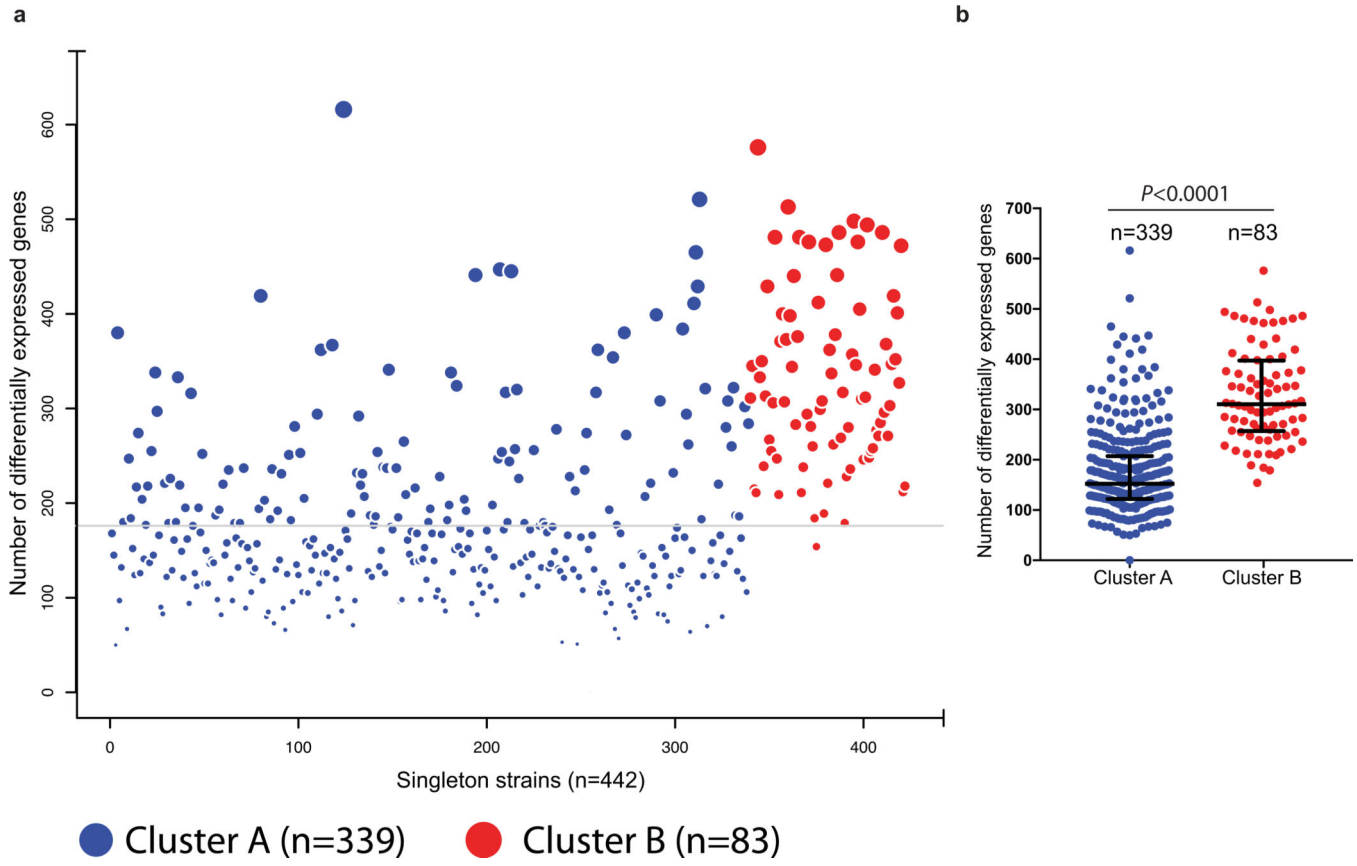

Figure 4. Variation in the number of differentially expressed (DE) genes between Cluster A and B strains.

(**a**) Plot displays the distribution of the number of DE genes for Cluster A (blue circles) and Cluster B (red circles) strains. DE genes were identified using strain MGAS28737 as a reference (see Supplementary note). The area of each circle is proportional to the number of DE genes. The horizontal grey line represents the median number of DE genes (n = 176) across 442 singleton strains. (**b**) Cluster B (n = 83) strains have significantly more DE genes compared to Cluster A (n = 339) strains (Mann-Whitney test, one-tailed, $P < 0.0001$). Median and interquartile range of DE genes for each cluster are depicted.

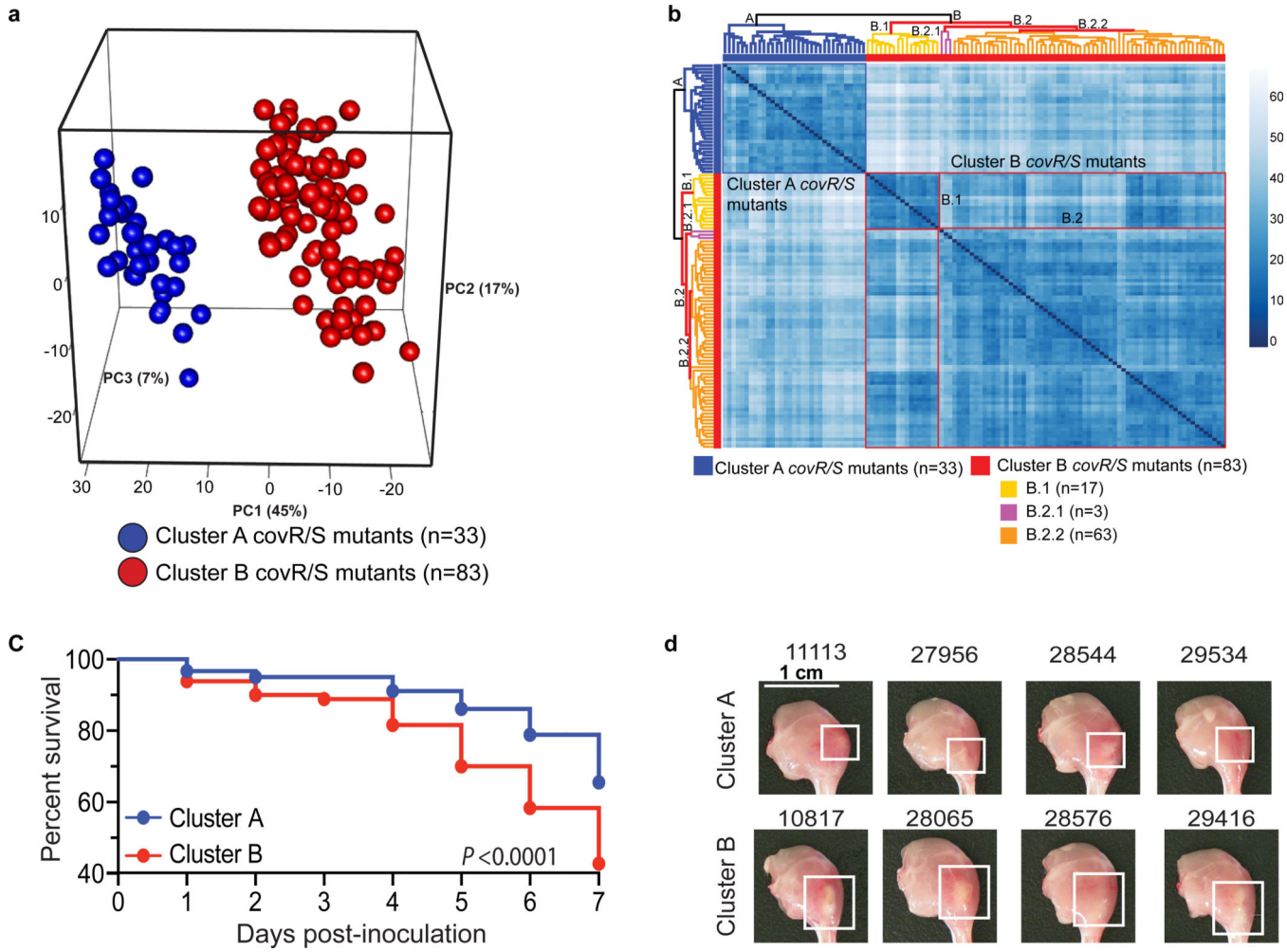

Figure 5. Clustering of *covR* and *covS* mutant strains and associated virulence.

(**a**) PCA plot of *covR*/*covS* mutant strains from Cluster A and Cluster B display distinct clustering. (**b**) Distance based hierarchical clustering transcriptome profile validated the clustering evident by PCA and also showed additional clustering. Strains in cluster B partitioned into additional subgroups arbitrarily designated B.1 and B.2. The analysis further refined subgroup B.2 strains into subgroups B.2.1 and B.2.2 (see Supplementary note). (**c**) Virulence of four cluster A and four cluster B strains in a mouse model of necrotizing myositis (n = 45 mice/strain). A significantly increased ability to cause near-mortality was observed for cluster A strains compared to cluster B strains. P values were determined using the log-rank test. (**d**) Representative gross pathology images of the hindlimb lesions from the mice (n = 6 mice per strain) infected with each of the four Cluster A (top panels) and Cluster B (bottom panels) strains are shown. Scale bar, 1cm.

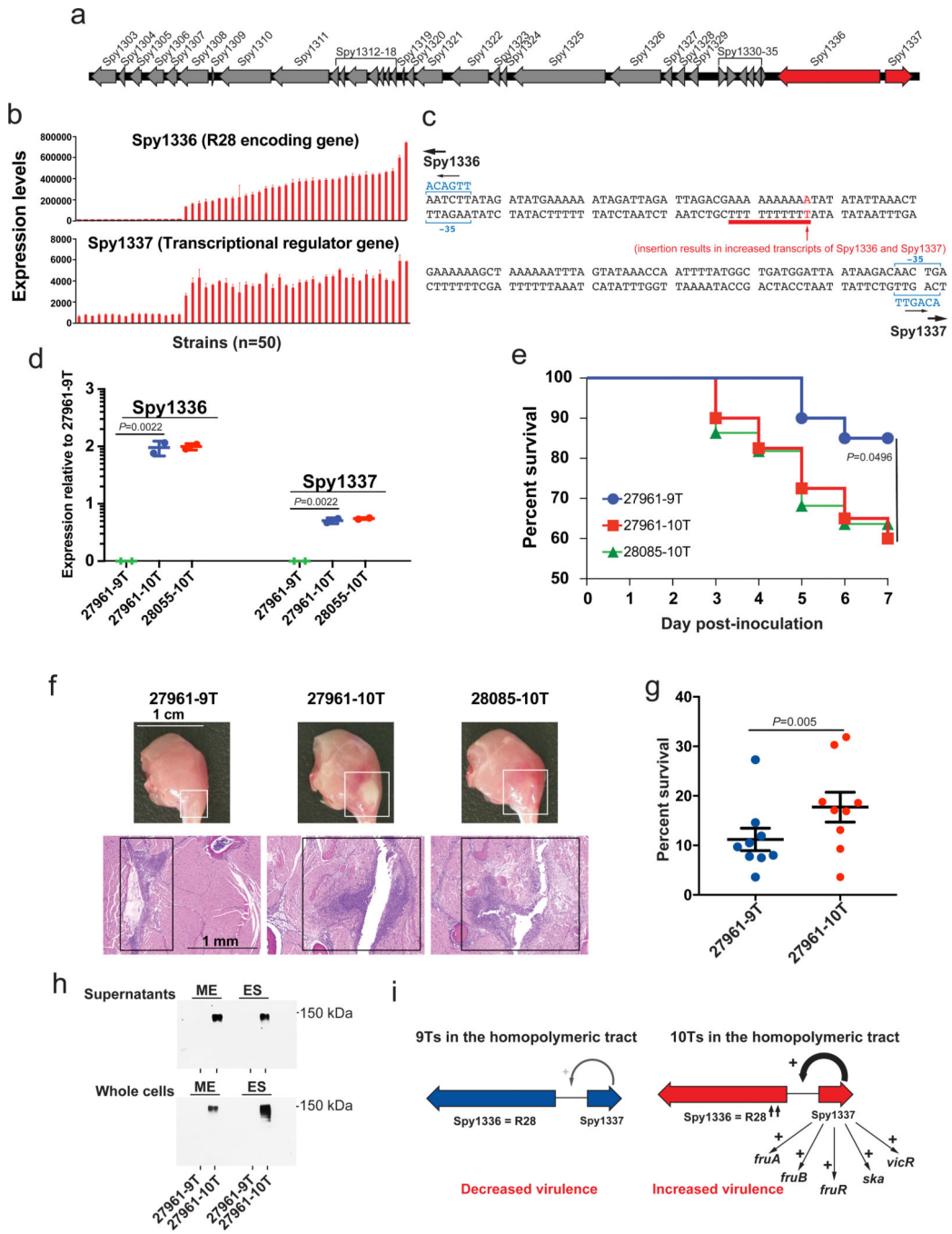

Figure 6. An intergenic single nucleotide insertion increases Spy1336/R28 expression and strain virulence.

(**a**) Schematic showing the divergently transcribed Spy1336/R28 and Spy1337 genes located in the RD2 region of the M28 chromosome. The Spy1336/R28 gene encodes the R28 protein, a virulence factor, and Spy1337 encodes an inferred transcriptional regulator. (**b**) Seventeen of 50 (34%) strains with low levels of Spy1336/R28 transcript (top) have reduced levels of Spy1337 transcript (bottom), whereas 33 strains (66%) with high levels of Spy1336/R28 transcript have high levels of Spy1337 transcript. Whiskers represent

minimum and maximum values. (**c**) Intergenic region betweenSpy1336/R28 and Spy1337. Homopolymeric T tract is underlined in red. (**d**) qRT-PCR results for low expresser wild-type parental (MGAS27961–9T) and isogenic mutant strain (MGAS27961–10T). Strain MGAS28055–10T is a naturally occurring high expresser strain with the 10T variant in the homopolymeric tract. Mean and SD are shown; y-axis is presented in log scale. (**e**) Virulence of MGAS27961–9T, MGAS27961–10T, and MGAS28085–10T in a mouse model of necrotizing myositis (n = 20 mice/strain). (**f**) Shown are representative gross and microscopic images of mice (n = 6 mice per strain) hindlimbs infected with strains MGAS27961–9T, MGAS27961–10T, or MGAS28085–10T. Scale bars, 1 cm (gross), and 1 mm (microscopic images). (**g**) Isogenic strains were exposed to purified PMNs, and percent bacterial survival was assessed at 3 h. Results are means and SEM of data from 9 separate experiments. (**h**) Western immunoblot analysis showing production of Spy1336/R28 protein. Isogenic strains were collected at equivalent ODs during mid-exponential (ME) and early stationary phase (ES). Cell-free supernatants and whole cells were assayed (2 independent experiments) for presence of Spy1336/R28 protein with an anti-R28 antibody. 24 μl (ME) and 8 μl (ES) were loaded for analysis of R28 in the cell-free supernatants, and 16 μl (ME) and 8 μl (ES) for whole cells. R28 from the reference strain MGAS6180 is predicted to have a molecular weight of 157 kDa. (**i**) Model depicting how the single nucleotide indel alters virulence. Weak binding of Spy1337 to the intergenic region containing the 9T homopolymeric tract leads to lower expression of Spy1336/R28, Spy1337 and other virulence factors (left panel), whereas stronger binding of Spy1337 to the intergenic region containing the 10T homopolymeric tract leads to higher expression of Spy1336/R28, Spy1337 and other virulence factors (right panel).

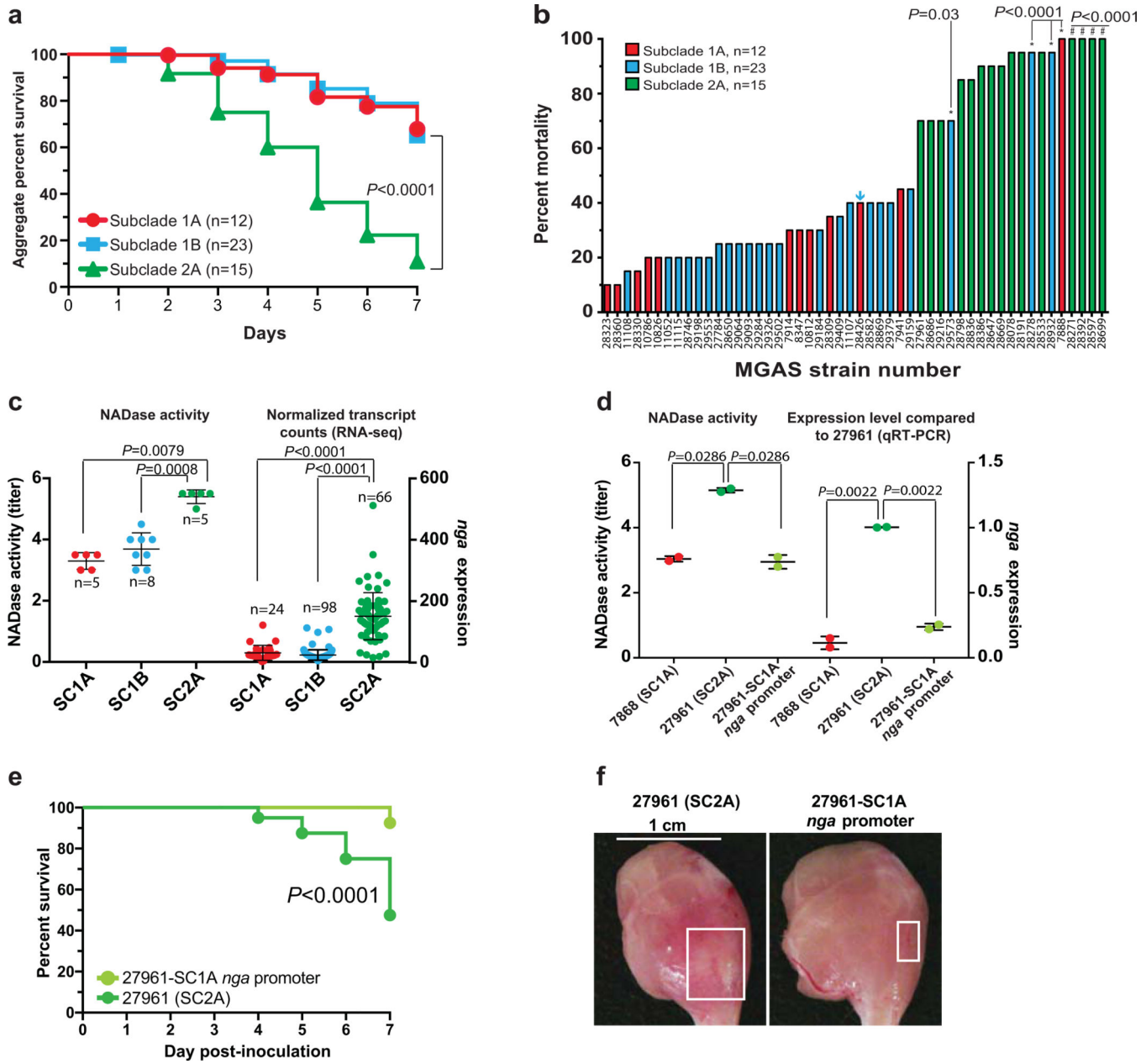

Figure 7. Mouse virulence data, NADase production and *nga* transcript level.

(**a**) Virulence of 50 *emm28* GAS strains in a mouse model of necrotizing myositis (*n* = 20 mice/strain). The 50 strains used for initial transcriptome analysis were studied. A significantly (log-rank test) increased ability to cause near-mortality was observed for strains of SC2A (green) compared to strains of SC1A (red) and SC1B (blue). Kaplan-Meier curve for all strains tested is shown by subclade. (**b**) Ability of each of the 50 strains assayed to cause near-mortality at 7 days post-inoculation is shown. SC1A (red bars) and 1B (blue bars) strains were compared to virulence reference strain MGAS28426 (arrow; * indicates *P* values relative to strain MGAS28426, log-rank test), and SC2A (green bars) strains were compared to the average of SC2A strains overall (# indicates *P* values relative to subclade 2A strains overall, log-rank test). (**c**) NADase activity and *nga* transcript levels. NADase

assays were performed using two biological replicates on strains that are wild-type for all known major transcriptional regulators. Number of strains (solid bars) analyzed per subclade was 5 (SC1A), 8 (SC1B), and 5 (SC2A). NADase activity (Y-axis, left) is presented as the highest dilution with hydrolyzing activity against exogenously added NAD^+. Replicate data are expressed as mean ± SD, Mann-Whitney two-tailed test. *nga* transcript levels (normalized transcript counts) are shown (Y-axis, right). The number of strains (hatched bars) analyzed per subclade was 24 (SC1A), 98 (SC1B), and 66 (SC2A). Strains wild-type for all known major virulence regulators were assessed. (**d**) NADase activity and *nga* transcript levels (qRT-PCR) of the isogenic mutant strain (27961-SC1A-*nga*-promoter) were compared to its parental wild-type strain (MGAS27961; SC2A) and a representative SC1A strain (MGAS7868). Two biological replicates per strain are expressed as mean ± SD, Mann-Whitney two-tailed test. (**e**) Kaplan-Meier curve showing that the isogenic mutant and wild-type parental strains differ significantly (log-rank test) in virulence in a mouse necrotizing myositis infection model. (**f**) Gross pathology images of infected mouse (n = 5) hindlimbs reflect the difference in virulence between the isogenic mutant and wild-type parental strains. Scale bar, 1 cm.

Table 1.

Summary of the 2,101 invasive *emm28* strains studied

Country	State[1]/Region	Years	Number of strains
Canada	Ontario	1991–2002	247
Denmark	Faroe Islands	2002–2014	7
Finland	Countrywide	1995–2015	704
Iceland	Countrywide	1992–2012	27
Norway	Countrywide	2006–2016	164
USA	A	1995–2012	105
USA	B	2000–2012	99
USA	C	1995–2011	61
USA	D	1995–2012	103
USA	E	1997–2012	103
USA	F	1995–2012	239
USA	G	2004–2012	34
USA	H	1998–2012	89
USA	I	1996–2012	53
USA	J	2000–2012	65
USA	Texas	1990s	1

The complete list of 2,101 strains analyzed in this study is presented in Supplementary Table 1. The total number of strains isolated in the USA was 952, of which 951 strains were collected as part of the ABC surveillance study conducted by the Centers for Disease Control and Prevention[22,93–95]. The strain from Texas is the genome reference strain MGAS6180[63].

[1]For the U.S. isolates, the states have been coded (A-J) at the request of the Centers for Disease Control.