https://doi.org/10.1038/s41588-018-0343-1

nature genetics

Integrated analysis of population genomics, transcriptomics and virulence provides novel insights into *Streptococcus pyogenes* pathogenesis

Priyanka Kachroo[1,20], Jesus M. Eraso[1,20], Stephen B. Beres[1], Randall J. Olsen[1,2,3], 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[4,5], Kirsi Gröndahl-Yli-Hannuksela[4], Karl G. Kristinsson[6,7], Magnus Gottfredsson[7,8], Maiju Pesonen[9,10], Johan Pensar[9], Emily R. Davenport[11], Andrew G. Clark[11], Jukka Corander[9,12], Dominique A. Caugant[13], Shahin Gaini[14,15,16,17], Marita Debess Magnussen[7,18], Samantha L. Kubiak[1], Hoang A. T. Nguyen[1], S. Wesley Long[1], Adeline R. Porter[19], Frank R. DeLeo[19] and James M. Musser[1,2,3]*

***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 among 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 a novel understanding of the virulence mechanisms of this model organism. Genome-wide association study, expression quantitative trait loci 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 that we used is generally applicable to any microbe and may lead to new therapeutics for many human pathogens.**

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 among 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 the 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 among 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 strictly 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 postinfection 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 relationships among 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 United States and many European

[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, TX, USA. [2]Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA. [3]Department of Microbiology and Immunology, Weill Cornell Medical College, New York, NY, USA. [4]Institute of Biomedicine, Medical Microbiology and Immunology, University of Turku, Turku, Finland. [5]National Institute for Health and Welfare, Helsinki, Finland. [6]Department of Clinical Microbiology, Landspitali University Hospital, Reykjavik, Iceland. [7]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. [8]Department of Infectious Diseases, Landspitali University Hospital, Reykjavik, Iceland. [9]Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [10]Department of Computer Science, Aalto University, Espoo, Finland. [11]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [12]Department of Biostatistics, University of Oslo, Oslo, Norway. [13]Division for Infection Control and Environmental Health, Norwegian Institute of Public Health, Oslo, Norway. [14]Medical Department, Infectious Diseases Division, National Hospital of the Faroe Islands, Tórshavn, Denmark. [15]Department of Infectious Diseases, Odense University Hospital, Odense, Denmark. [16]Department of Clinical Research, University of Southern Denmark, Odense, Denmark. [17]Department of Science and Technology, Centre of Health Research, University of the Faroe Islands, Tórshavn, Denmark. [18]Thetis, Food and Environmental Laboratory, Torshavn, Denmark. [19]Laboratory of Bacteriology, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA. [20]These authors contributed equally: Priyanka Kachroo, Jesus M. Eraso. *e-mail: JMMusser@houstonmethodist.org

Table 1 | Summary of the 2,101 invasive *emm28* strains studied.

Country	State[a]/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 Active Bacterial Core surveillance study conducted by the Centers for Disease Control and Prevention. The strain from Texas is the genome reference strain MGAS6180 (MGAS, Musser GAS strain number)[63]. [a]For the US isolates, the states have been coded (A–J) at the request of the Centers for Disease Control.

countries[18–23]. For reasons that remain unexplained, strains of some *emm* types or M protein serotypes are nonrandomly 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 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 we 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. The application of statistical methods and machine learning facilitated the 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 and 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 Methods). Inferences in genetic relationships were made by using 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-kilobase (kb) horizontal gene transfer (HGT) block that contributes 520 core SNPs, and by 19 core SNPs located outside this HGT block (Supplementary Table 2). This 28.0-kb HGT block includes the *nga–ifs–slo* operon encoding the 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*[42] and was probably acquired by subclade 2A GAS via a recombination event.

SC1B comprised 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 United States, where ~55% of strains were SC2A. SC1B strains were predominantly isolated from patients in Finland, Norway, the 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 the strains (Fig. 1c).

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

Transcriptome signatures and population structure. To determine whether distinct patterns of gene expression were nonrandomly associated with the *emm28* population structure, we first conducted transcriptome RNA sequencing (RNA-seq) analysis on a subset of 50 strains genetically representative of the three numerically dominant subclades (that is, SC1A, SC1B, and SC2A) (Supplementary Table 7). Strains were selected for RNA-seq analysis from the main sample of 2,095 M28 strains according to the criteria described in the Methods. The 50 strains are from diverse years, countries, and regions within countries and have diverse MGE contents. RNA-seq analysis was conducted in triplicate (three biologic replicates) at midexponential and early-stationary growth phases (Methods, Supplementary Fig. 2b and Supplementary Table 7).

Although the 50 strains differed in genomic background, 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 alleles in all known major regulatory genes (Fig. 2c,d), the greatest number of differentially expressed genes occurred between strains assigned to different subclades (Supplementary Fig. 5). At the midexponential phase, the greatest number of differentially expressed genes was observed when we compared the transcriptomes of SC2A strains with SC1A and SC1B strains (32 (2%) and 15 (0.9%), respectively) (Supplementary Fig. 5a

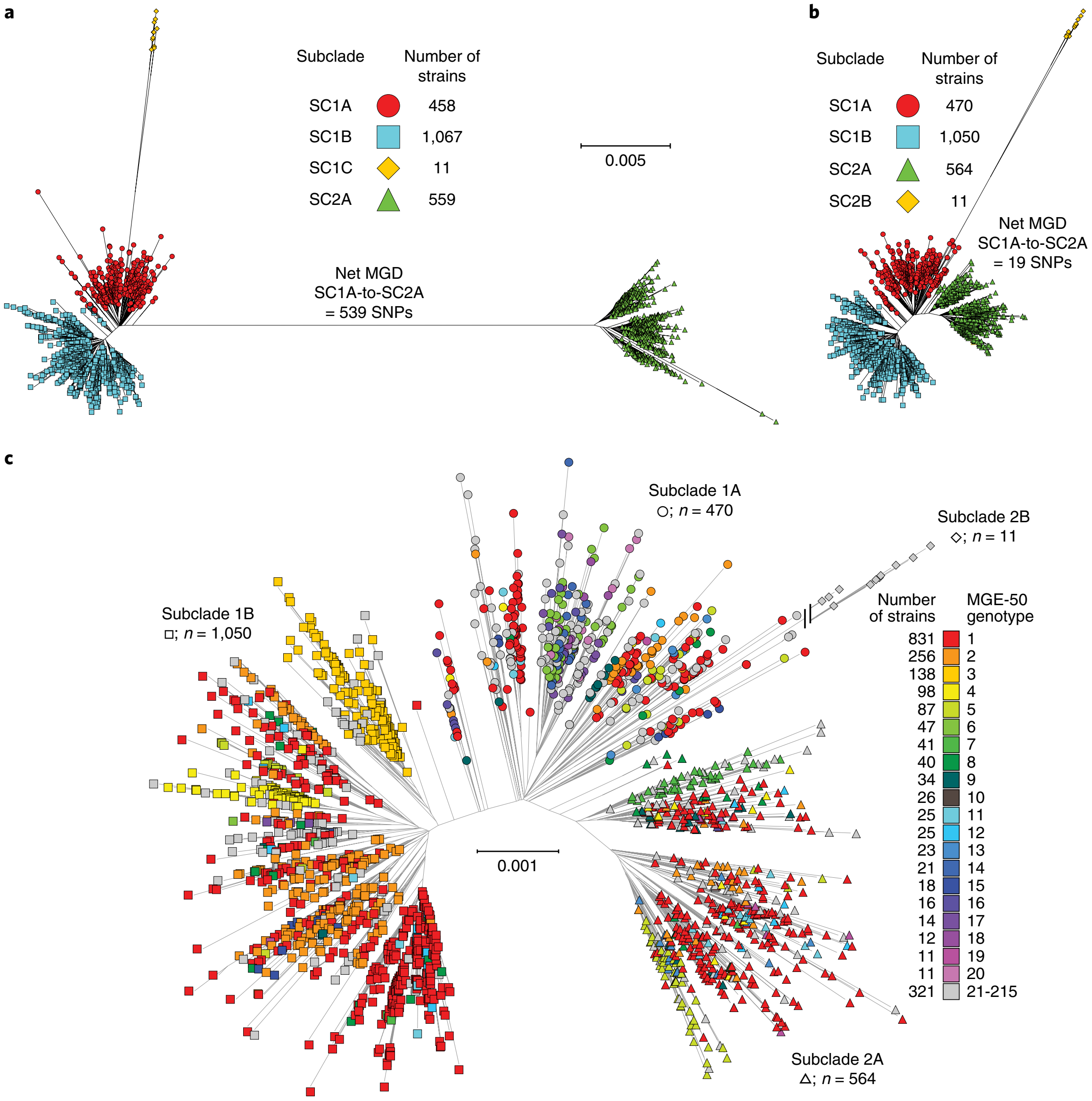

Fig. 1 | Population genetic structure for 2,095 *S. pyogenes emm28* invasive infection isolates. a, Genetic relationships inferred without correction for HGT and recombination events. Four genetic subclades (SC1A, SC1B, SC1C, and SC2A) inferred by BAPS are shown. **b**, Genetic relationships inferred with correction for HGT and recombination events by using Gubbins. hierBAPS was used to infer the genetic subclades (SC1A, SC1B, SC2A, and SC2B) within the population after exclusion of recombination events. After exclusion of HGT and recombination events, the SC1C strains in **a** ($n = 11$)—a distinct genetic lineage of *emm28* strains—were inferred as SC2B strains by hierBAPS. MGD, mean genetic distance. **c**, MGE-50 genotypes were defined according to the presence or absence of alleles for 50 MGE-encoded site-specific integrase ($n = 31$) and secreted virulence factor ($n = 19$) genes detected by using SRST2, as described in the Supplementary Note and presented in Supplementary Tables 1 and 3–6. Illustrated are the 20 most abundant MGE-50 genotypes, each present in ten or more strains and cumulatively accounting for 90% of the total strain sample. Phylogeny in all panels was inferred by neighbor joining (based on 20,135 core SNPs in **a** (recombined regions included) and 18,544 core SNPs in **b** and **c** (recombined regions excluded)). The trees in **a** and **b** are shown at the same scale. The strains are colored by subclade and MGE genotype (MGE-50), as indicated in the insets.

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, as compared with SC1A and SC1B strains (318 (19.9%) and 83 (5.2%), respectively) (Supplementary Fig. 5a and Supplementary Table 8).

A substantial proportion of the differentially expressed genes was located in the 28.0-kb region that was horizontally transferred (the HGT region) and included the *nga–ifs–slo* operon (Supplementary Table 8). At the midexponential phase, 35.7% (SC2A versus SC1A comparison: $P < 0.0001$) and 25% (SC2A versus SC1B comparison: $P < 0.001$) of genes within this HGT region (28 genes) were differentially expressed (P values assessed by Fisher's exact test), whereas 21.4% (SC2A versus SC1A comparison: $P = 0.81$) and 39.3% (SC2A versus SC1B comparison: $P < 0.001$) were differentially expressed at the early-stationary phase. The three most strongly upregulated genes in SC2A strains compared with SC1A and SC1B strains were *nga*, *ifs*, and *slo*, with an approximately fourfold increase in transcript levels at the midexponential phase and an approximately eightfold increase at the early-stationary phase (Supplementary Table 8 and Supplementary Fig. 5b).

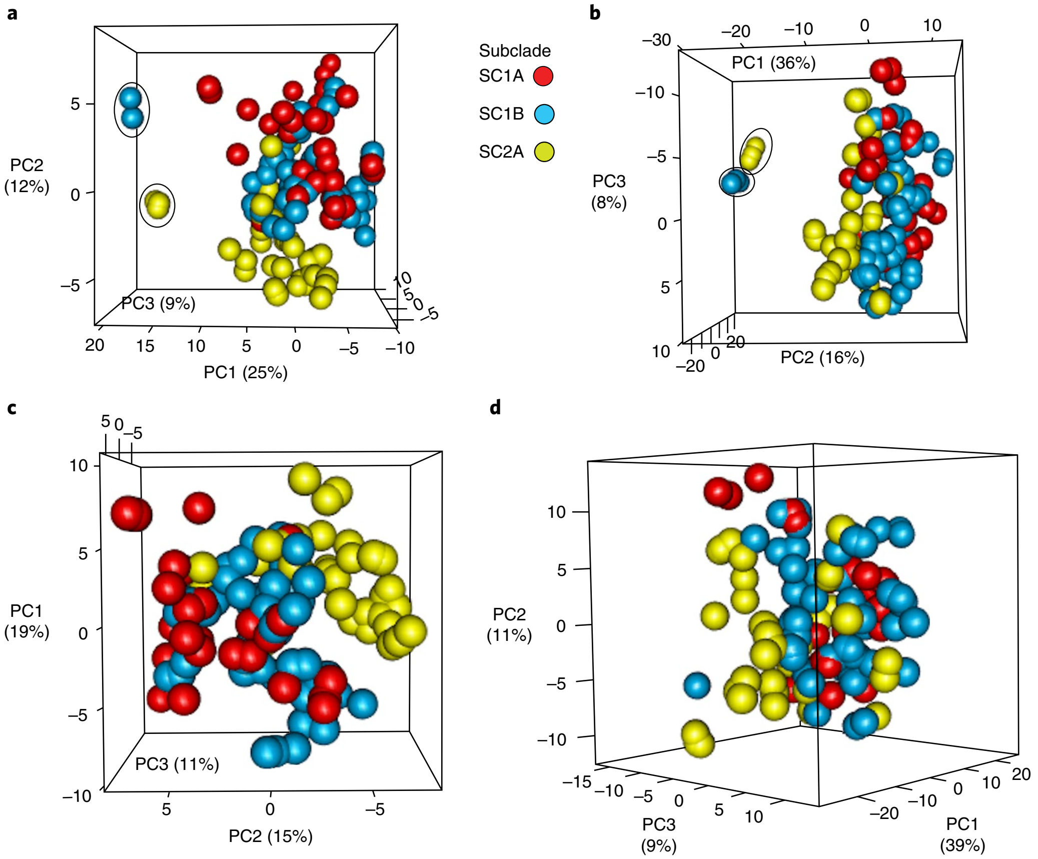

Fig. 2 | Transcriptome analysis of the subset of 50 strains. a–d, PCAs of transcriptome data for the subset of 50 strains (**a** and **b**) and global transcriptome data for wild-type-like strains (**c** and **d**) at midexponential (**a** and **c**) and early-stationary phases of growth (**b** and **d**). Highlighted within ovals in **a** and **b** are two strains with deletion frameshift mutations in *covS* that group distinctly apart from the other 48 strains at both phases. In **c** and **d**, SC2A isolates cluster together compared with SC1A and SC1B strains. Three biological replicates were analyzed at two time points. PC, principal component.

Infections caused by SC1B strains have increased in recent years in several countries, including the United States, Finland, Iceland, and Norway, whereas SC1A strains have decreased substantially (Supplementary Fig. 3a), thus raising the possibility that SC1B strains might 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 had two contiguous nonsynonymous mutations in RivR—a negative regulator of the *grab* (protein-G-related α_2-macroglobulin-binding) gene[46,47]. At the midexponential phase, *grab* was the only upregulated gene in SC1B strains compared with SC1A and SC2A strains (Supplementary Table 8 and Supplementary Fig. 5c). Similarly, higher *grab* transcript abundance was observed at the early-stationary phase in SC1B and SC2A strains compared with SC1A (Supplementary Fig. 5c). GRAB is a cell-surface-anchored protein that binds α_2-macroglobulin, allowing it to retain the proteolytically active form of the cysteine protease SpeB at the GAS surface, to protect GAS from killing by antimicrobial peptide LL-37 (refs. [48,49]), and to 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 has traditionally been conducted by using triplicate biological replicates of strains grown to two distinct growth phases. However, this approach is currently 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 (that is, lacking replicates), in concert with optimal sequencing depth (5–10 million sequencing reads) for a pathogen with a genome size of approximately 2 Mb (ref. [51]), would provide substantially 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 Fig. 6 and 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 the *k*-means clustering statistical strategy (Methods).

Population transcriptome analysis of diverse strains. To examine the relationships among 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 cluster B ($n=83$) (Fig. 3a). Density-Based Spatial Clustering of Applications with Noise validated the existence of these two clusters (Supplementary Fig. 8a). Wild-type-like strains (strains bioinformatically assessed to have wild-type alleles for all major regulatory genes) were predominantly associated with cluster A, except for ten outlier strains (Fig. 3a). Reexamination of the genome data for these outlier strains by using Pilon (Methods) identified undetected indels in the *covS* global transcriptional regulatory gene in eight of these ten strains. Hence, transcriptome-guided

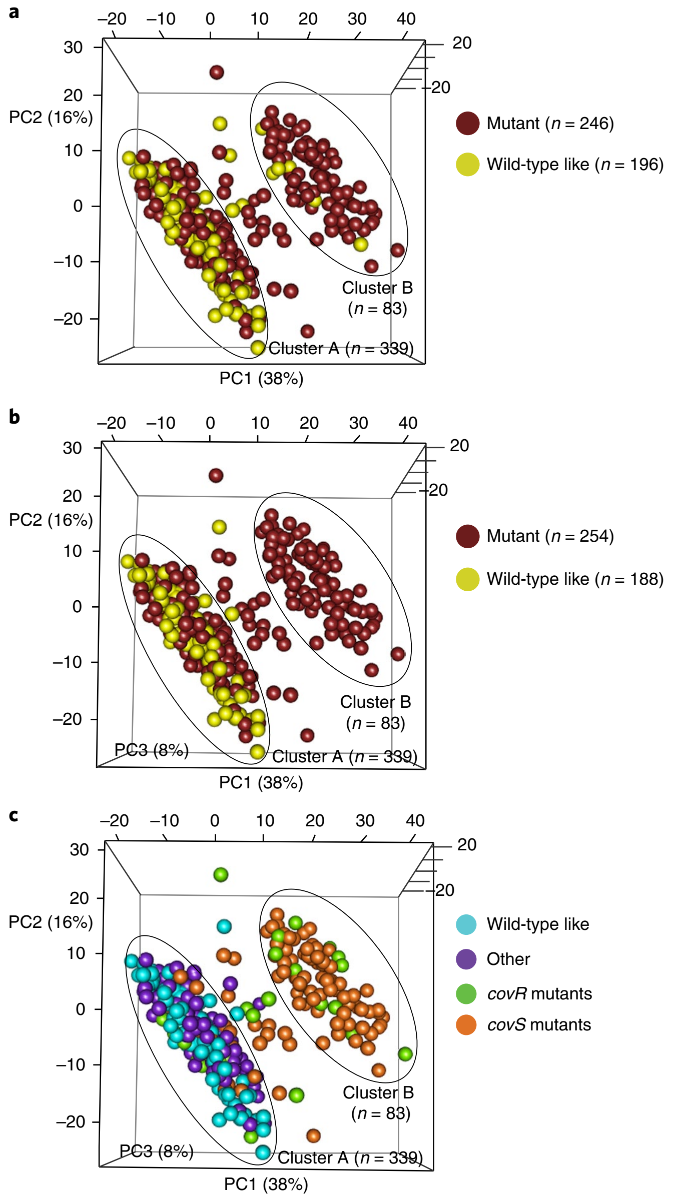

Fig. 3 | Singleton strains ($n = 442$) partition into two major transcriptome clusters according to their genome-wide expression profiles. a, Three-dimensional PCA plot displaying variation in the transcriptome data along the top three PCs. The greatest variance in transcriptome data along PC1 (38%) separated the strains into two major clusters, which are arbitrarily designated clusters A and B. Wild-type-like strains were predominantly associated with cluster A. **b**, The genome data were reexamined for ten outlier wild-type-like strains that did not group with the cluster A wild-type-like strains, eight of which were reassigned to cluster B after identification of previously missed polymorphisms. This resulted in cluster B containing mutant strains exclusively, whereas cluster A comprised both mutant and wild-type-like strains. **c**, Cluster B was composed exclusively of strains with mutations in *covR* or *covS*. Wild-type-like strains and strains with mutations in genes encoding regulators other than *covR* and *covS* (designated 'other') predominantly grouped into cluster A.

polymorphism discovery identified genetic causes underlying these aberrant transcriptomes.

Considering that the transcriptomes of the eight *covS* mutant strains differed markedly from the wild-type strains, and concordantly with previous results[53–58], we hypothesized that strains with mutations in specific major global regulators might 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 machine learning[59] to determine whether one of the four class labels (that is, wild type, *covR*, *covS*, or *ropB* mutant) could be assigned to the outlier strains with high probability. Briefly, on the basis of analysis of the transcriptome profiles of 283 singleton strains (see 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 wild type phenotypically (transcript profile) resembled wild-type 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 wild-type like) of strains with mutations in a major regulator gene.

Regulatory gene mutations and transcriptome changes. Reassignment of the outlier strains as *covS* mutants resulted in cluster A having both wild-type-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 had mutations in *covS* or *covR*, and second, most strains with either *covS* (68.5%) or *covR* (37.5%) mutations were assigned to cluster B (Fig. 3c). Third, most strains assigned to cluster A were wild-type like or had mutations in major regulatory genes other than *covRS* (described below). Fourth, cluster B strains had a significantly greater number of differentially expressed genes than 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 the expression of 15% of the transcriptome, including key virulence factors[44]. In agreement with this, inactivation of CovRS enhances virulence[53,56]. We compared the transcriptomes of 442 strains composed of 188 predicted wild-type 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 wild-type-like strains in cluster A (Fig. 3c). The finding that *covRS* mutant strains were 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 wild-type strains suggests that some polymorphisms may have fewer functional consequences than others. PCA of only *covRS* mutant strains in cluster A ($n = 33$) and cluster B ($n = 83$) recapitulated the grouping into two distinct clusters (Fig. 5a). Next, we used distance-based clustering to test the hypothesis that additional substructure not evident by PCA (Fig. 5a) might be present in the transcriptome data (Fig. 5b). The findings are reported in the Supplementary Note.

To test the hypothesis that cluster B *covRS* strains have distinctive transcriptomes compared with cluster A *covRS* strains, we examined the 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 as well as in the magnitude of the altered transcript changes (up or down) of the differentially expressed genes. Many (32%) differentially expressed genes had fivefold 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 with cluster A *covRS* strains, cluster B *covRS* strains had a significantly increased

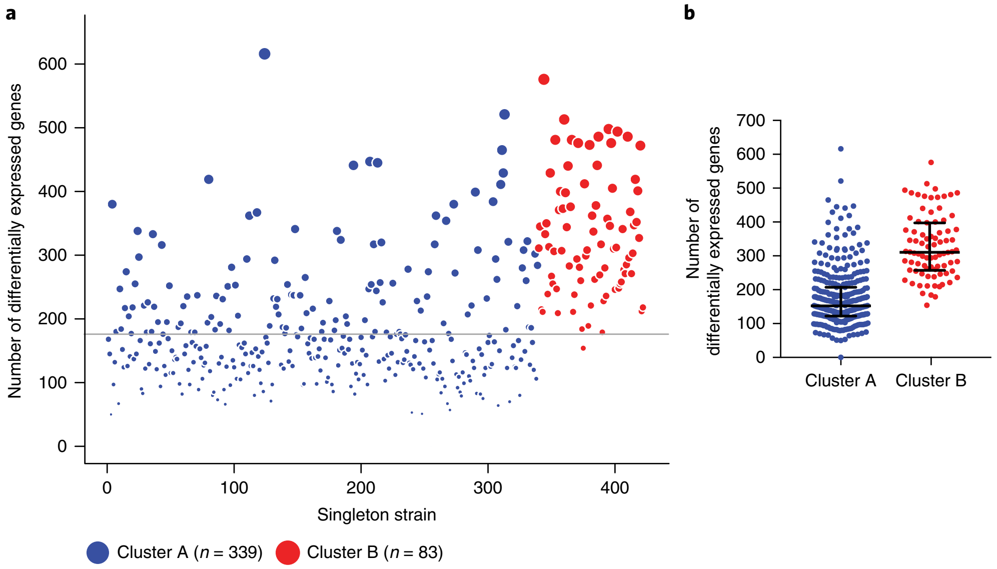

Fig. 4 | Variation in the numbers of differentially expressed genes between cluster A and B strains. a, Distribution of the numbers of differentially expressed genes for cluster A (blue circles) and cluster B strains (red circles). Differentially expressed genes were identified by using the strain MGAS28737 as a reference (Supplementary Note). The area of each circle is proportional to the number of differentially expressed genes. The horizontal gray line represents the median number of differentially expressed genes ($n = 176$) across 442 singleton strains. **b**, Cluster B strains ($n = 83$) had significantly more differentially expressed genes than cluster A strains ($n = 339$). Significance was determined by one-tailed Mann–Whitney *U* test ($P < 0.0001$). The median and interquartile range of genes for each cluster are depicted.

frequency of frameshift-inducing indels and nonsense mutations (two-tailed Fisher's exact test, $P < 0.0001$), probably to inactivate this regulatory system (loss-of-function mutations).

Given 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. In agreement with this hypothesis, analysis of virulence of four strains from each cluster by 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 and 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 (refs. [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 levels 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) using SEER[64,65] on de novo assemblies of all 442 strains for which we had associated transcriptome data. According to the transcript levels of the *Spy1336/R28* and *Spy1337* genes, we examined whether the strains with low- or high-transcript phenotypes were significantly associated with any genetic event (for example, SNP, indel, or 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 ten T residues, and those negatively associated with the high-transcript phenotype had nine T residues in this tract. The association of the 10T variants with the increased level of transcript of Spy1336/R28 and Spy1337 was also identified by an expression quantitative trait loci (eQTL) analysis[66,67] of the 50- and 442-strain datasets (Supplementary Fig. 10b).

Compared with a parental strain (nine T residues), an isogenic mutant strain (ten T residues) had significantly increased transcript levels of Spy1336/R28 and Spy1337 (Fig. 6d); caused significantly larger gross and microscopic lesions, and more near-mortality, as 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 the midexponential or early-stationary phase (Supplementary Table 15). Thus, insertion or deletion of a single T residue in this homopolymeric tract significantly altered 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 we reasoned that these genomic and transcriptomic changes might cause significant variation in virulence. To test this hypothesis, we assessed 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 the SC1A reference strain MGAS28426 in a

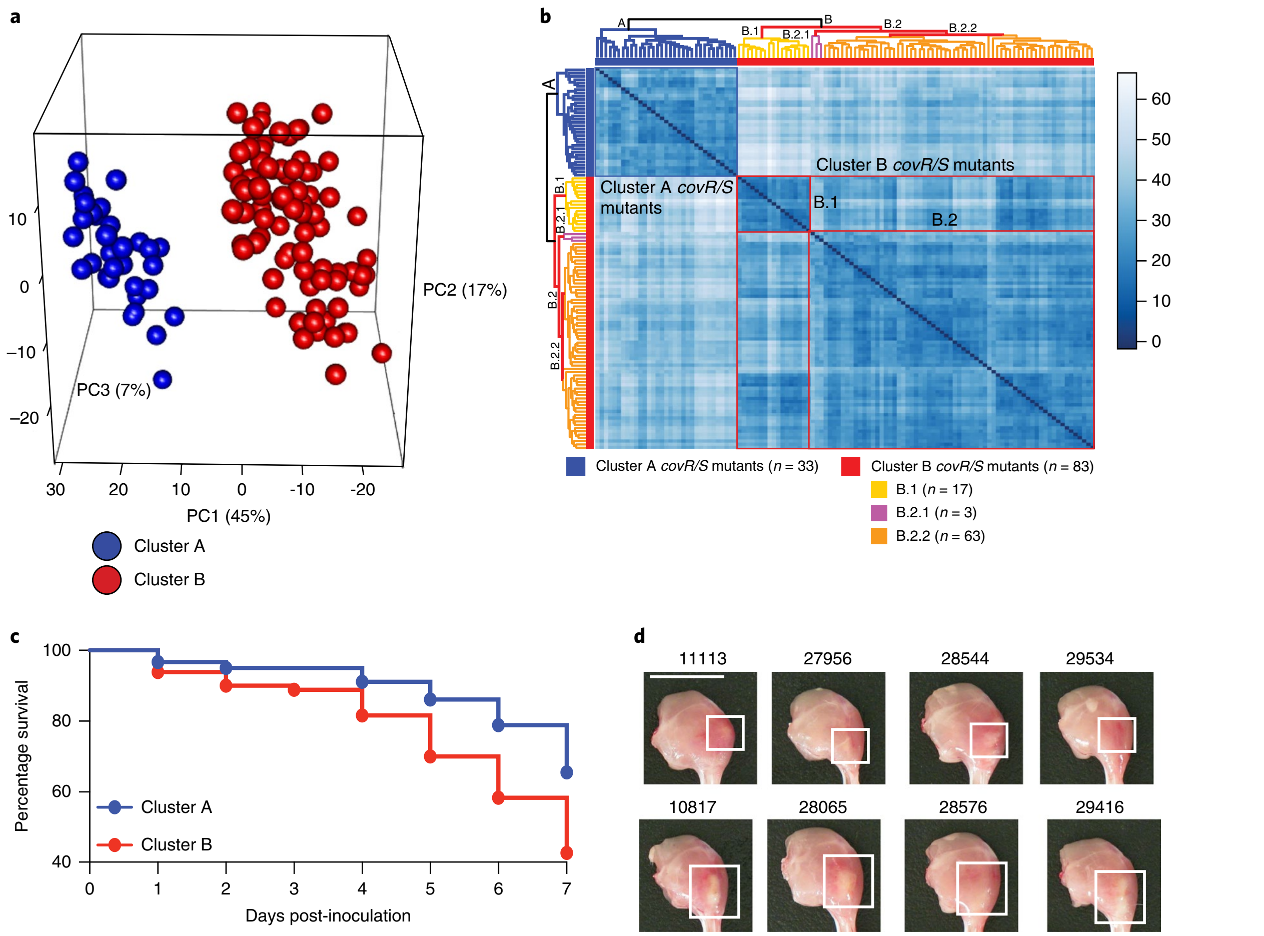

Fig. 5 | Clustering of *covR* and *covS* mutant strains, and associated virulence. **a**, PCA plot of *covR*/*covS* mutant strains from clusters A ($n=33$) and B ($n=83$), displaying distinct clustering. **b**, A distance-based hierarchical-clustering transcriptome profile validated the clustering evident 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 (Supplementary Note). **c**, Virulence of four cluster A and four cluster B strains in a mouse model of necrotizing myositis ($n=45$ mice per strain). A significantly increased ability to cause near-mortality was observed for cluster A strains compared with cluster B strains. Significance was determined using the log-rank test ($P<0.0001$). **d**, Representative gross pathology images of the hindlimb lesions from mice ($n=6$ mice per strain) infected with each of the four cluster A (top) and cluster B strains (bottom). Boxed areas demarcated in white illustrate major lesion areas. Scale bar, 1 cm.

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 (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 might be due, at least in part, to significantly increased expression of the *nga*, *ifs*, and *slo* genes, thus resulting in increased production of secreted SPN and SLO toxins by SC2A organisms, as shown for other GAS serotypes[1,6,15,16]. In agreeement with this hypothesis, SC2A strains had significantly higher *nga* transcript levels and SPN activity than SC1A or SC1B strains (Fig. 7c). The same was observed for *ifs* ($P<0.001$, Mann–Whitney *U*-test) and *slo* ($P<0.001$, Mann–Whitney *U*-test)—two other genes in the same operon.

To unambiguously demonstrate that the significantly greater virulence of SC2A strains than SC1A stains is due, in part, to greater *nga*–*ifs*–*slo* promoter activity, we replaced the *nga* promoter of the SC2A reference strain MGAS27961 with the SC1A *nga* promoter. The isogenic mutant strain with the SC1A promoter produced significantly less SPN 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 among population genomics, transcriptomics, and virulence in *emm28 S. pyogenes* strains. We discovered that there was no simple correlation between the magnitude of transcriptome changes (number of differentially expressed genes) and overall genome-to-genome genetic distance (Supplementary Fig. 9 and Supplementary Note). In aggregate, our analysis shows that a holistic approach involving multiple types of high-dimensional data applied to population-based strain samples can provide 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 used 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]. These findings add to an emerging theme in bacteria that

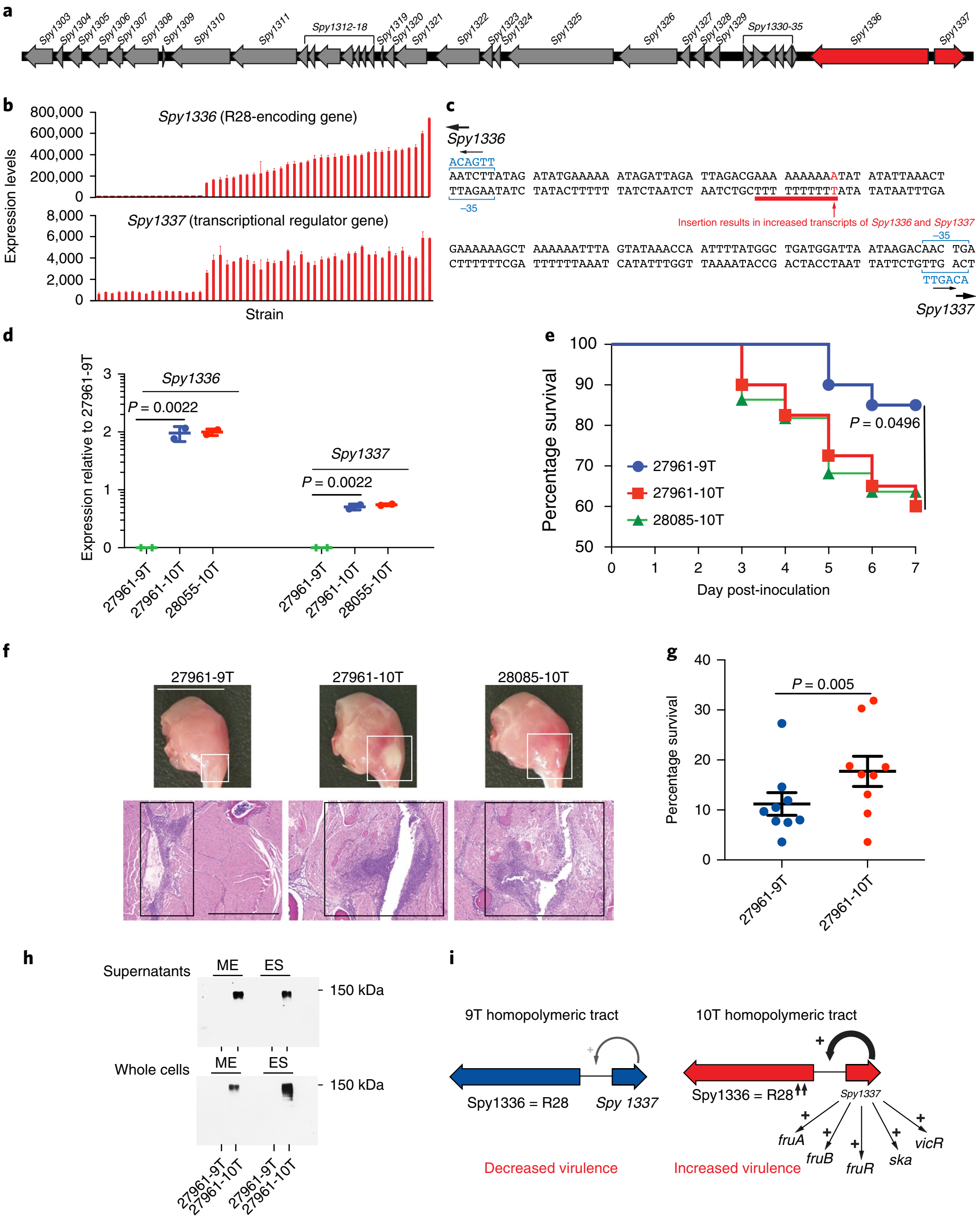

Fig. 6 | An intergenic single-nucleotide insertion increases *Spy1336/R28* expression and strain virulence. a, Schematic showing the divergently transcribed *Spy1336/R28* and *Spy1337* genes located in the 'region of difference 2' region of the M28 chromosome. The *Spy1336/R28* gene encodes the R28 protein (a virulence factor), and *Spy1337* encodes an inferred transcriptional regulator. **b**, A total of 17 (34%) of the 50 strains with low levels of *Spy1336/R28* transcript (top) also had reduced levels of *Spy1337* transcript (bottom), whereas 33 strains (66%) with high levels of *Spy1336/R28* transcript had high levels of *Spy1337* transcript. Error bars, s.d. **c**, Intergenic region between *Spy1336/R28* and *Spy1337*. The homopolymeric T tract is underlined in red. **d**, Reverse-transcription quantitative PCR results for low-expression wild-type parental (MGAS27961-9T) and isogenic mutant strains (MGAS27961-10T). Strain MGAS28055-10T is a naturally occurring high-expression strain with the 10T variant in the homopolymeric tract. Means ± s.d. are shown. The *y* axis is presented on a log scale. **e**, Virulence of MGAS27961-9T, MGAS27961-10T, and MGAS28085-10T in a mouse model of necrotizing myositis ($n = 20$ mice per strain). **f**, Representative gross (top) and microscopic images (bottom) of mouse hindlimbs ($n = 6$ mice strain^{-1}) infected with strains MGAS27961-9T, MGAS27961-10T, and MGAS28085-10T. Boxed areas demarcated in white illustrate major lesion areas. Scale bars, 1 cm (gross images) and 1 mm (microscopic images). **g**, Isogenic strains were exposed to purified polymorphonuclear neutrophils, and the percentage bacterial survival was assessed at 3 h. Means ± s.e.m. of the data from nine separate experiments are shown. **h**, Western blot analysis showing production of the Spy1336/R28 protein. Isogenic strains were collected at equivalent absorbance during the midexponential (ME) and early-stationary phase (ES). Cell-free supernatants and whole cells were assayed (two independent experiments) for the presence of the Spy1336/R28 protein with anti-R28. For analysis of R28, 24 µl (ME) and 8 µl (ES) were loaded in the cell-free supernatants, and 16 µl (ME) and 8 µl (ES) were loaded for whole cells. R28 from the reference strain MGAS6180 is predicted to have a molecular weight of 157 kDa. **i**, Schematic depicting how the single-nucleotide indel alters virulence. Weak binding of Spy1337 to the intergenic region containing the 9T homopolymeric tract led to lower expression of *Spy1336/R28*, *Spy1337*, and other virulence factors (left), whereas stronger binding of Spy1337 to the intergenic region containing the 10T homopolymeric tract led to higher expression of *Spy1336/R28*, *Spy1337*, and other virulence factors (right).

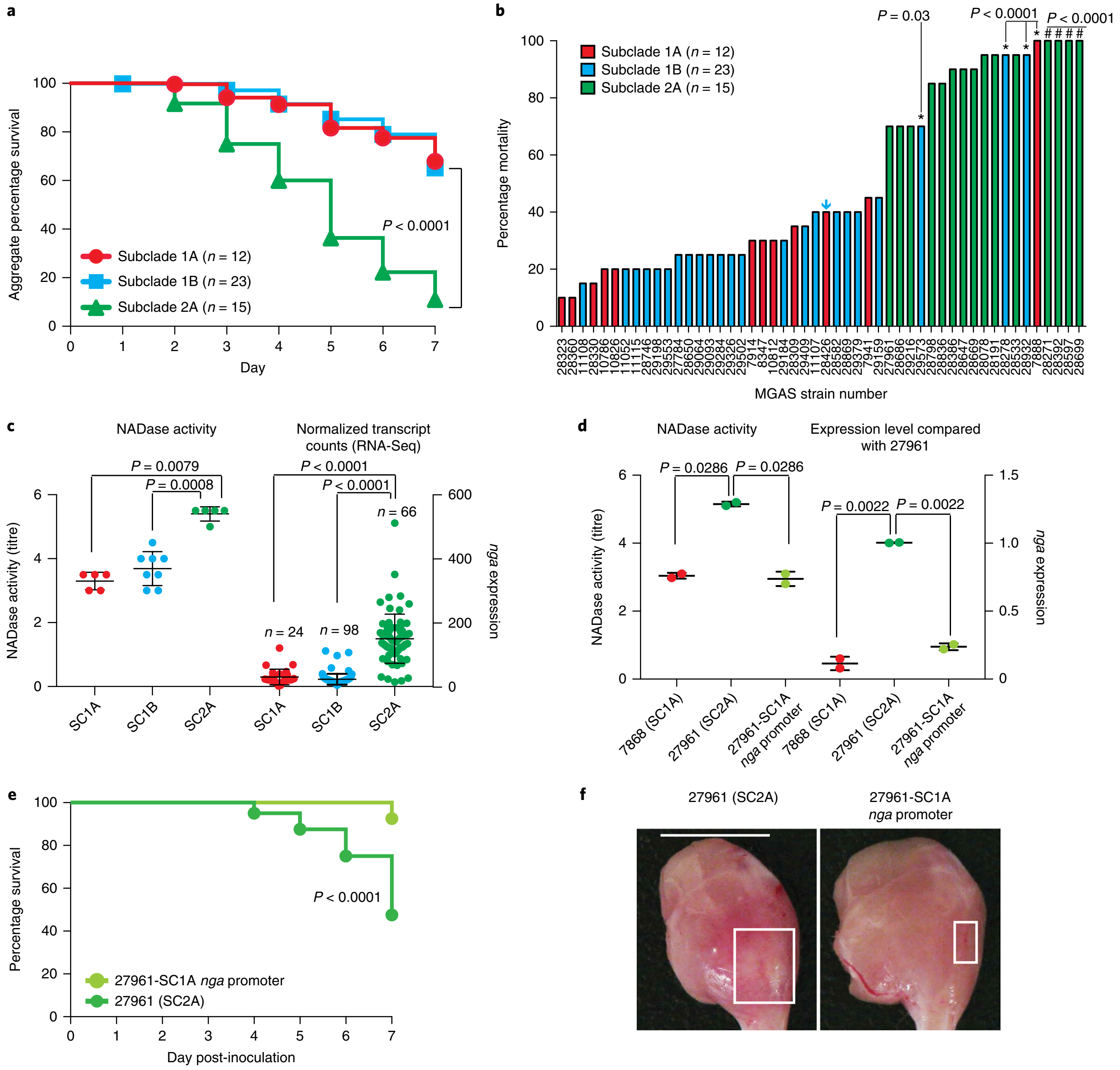

Fig. 7 | Mouse virulence data, NADase production, and *nga* transcript levels. **a**, Virulence of 50 *emm28* GAS strains in a mouse model of necrotizing myositis (*n* = 20 mice per strain). The 50 strains used for initial transcriptome analysis were studied. A significantly (log-rank test) increased ability to cause near-mortality was observed for strains of SC2A (green) compared with strains of SC1A (red) and SC1B (blue). A Kaplan–Meier curve for each strain tested is shown by subclade. **b**, Ability of each of the 50 strains assayed to cause near-mortality at 7 d post-inoculation. SC1A (red bars) and SC1B strains (blue bars) were compared with virulence reference strain MGAS28426 (arrow; **P* values relative to strain MGAS28426, log-rank test), and SC2A strains (green bars) were compared with the average of the SC2A strains overall (#*P* values relative to subclade 2A strains overall, log-rank test). **c**, NADase activity and *nga* transcript levels. NADase assays were performed by using two biological replicates on strains that are wild type for all known major transcriptional regulators. The numbers of strains (solid bars) analyzed per subclade were five (SC1A), eight (SC1B), and five (SC2A). NADase activity (*y* axis, left) is presented as the highest dilution with hydrolyzing activity against exogenously added NAD$^+$. Replicate data are expressed as means ± s.d. (two-tailed Mann–Whitney test). *nga* transcript levels (normalized transcript counts) are shown on the right-hand *y* axis. The numbers of strains analyzed per subclade were 24 (SC1A), 98 (SC1B), and 66 (SC2A). Strains that were wild type for all known major virulence regulators were assessed. **d**, NADase activity and *nga* transcript levels (reverse-transcriptase quantitative PCR) of the isogenic mutant strain (27961-SC1A *nga* promoter) were compared with its parental wild-type strain (MGAS27961; SC2A) and a representative SC1A strain (MGAS7868). Two biological replicates per strain are expressed as means ± s.d. (two-tailed Mann–Whitney 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**, Representative gross pathology images of infected mouse (*n* = 5) hindlimbs reflect the difference in virulence between the isogenic mutant and wild-type parental strains. Boxed areas demarcated in white illustrate major lesion areas. Scale bar, 1 cm.

seemingly modest changes in intergenic regions can alter gene expression and be adaptive[15,70–74]. Several possibilities exist regarding how the increased transcript levels of *Spy1336/R28* and *Spy1337* might 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]. Precisely how this protein contributes to virulence is not yet known, 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 RNA-seq analysis of the isogenic strains containing either nine or ten Ts and found that at the midexponential phase, only two genes (*Spy1336/R28* and *Spy1337*) were significantly upregulated, whereas at the 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 with 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]. In agreement 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 (Fig. 6g), and produced more secreted and cell-associated Spy1336/R28 protein (Fig. 6h; model in Fig. 6i).

The simplest hypothesis to explain how the 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 the 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, a homopolymeric tract with ten Ts (compared with nine Ts) would constitute a better consensus binding site, whereas in the second case, the presence of ten Ts in the spacer region (compared with nine Ts) 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 the use of machine learning to analyze and correctly classify regulator mutant strains that were misidentified on the basis of analysis of genome sequence data alone. Until recently, most studies on pathogenic microbes using machine learning have 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 applied 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* and *Spy1337*) and three in *trans* (*Spy1338*, *Spy1339*, and *Spy1340*)—by using transcript data from 50 strains in the midexponential phase. Similarly, *cis* (*Spy1336/R28* and *Spy1337*) and *trans* associations with 47 additional genes (false discovery rate <0.0005) were found by using transcript data from 442 strains in the early-stationary phase (Supplementary Fig. 10b and 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 enabled us to conclude that although HGT events can and do alter the transcriptome; most 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 the 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 United States. We found that, compared with non-SC2A strains from the United States, a significantly higher proportion of SC2A strains from the United States 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 the 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.

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

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0343-1.

Received: 12 April 2018; Accepted: 21 December 2018;
Published online: 18 February 2019

References

1. Beres, S. B. et al. Transcriptome remodeling contributes to epidemic disease caused by the human pathogen Streptococcus pyogenes. *mBio* **7**, e00403-16 (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**, 647–652 (2015).
4. Long, S. W. 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**, e00489-17 (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).
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. USA* **111**, E1768–E1776 (2014).
7. Bruchmann, S. et al. Deep transcriptome profiling of clinical *Klebsiella pneumoniae* isolates reveals strain and sequence type-specific adaptation. *Environ. Microbiol.* **17**, 4690–4710 (2015).
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, B. K. 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, T. M. & Krukonis, E. S. Three *Yersinia pestis* adhesins facilitate Yop delivery to eukaryotic cells and contribute to plague virulence. *Infect. Immun.* **78**, 4134–4150 (2010).
11. Swearingen, M. C., Porwollik, S., Desai, P. T., McClelland, M. & Ahmer, B. M. 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**, eaaf1283 (2017).
13. Carapetis, J. R., Steer, A. C., Mulholland, E. K. & Weber, M. The global burden of group A streptococcal diseases. *Lancet Infect. Dis.* **5**, 685–694 (2005).
14. Carapetis, J. R. 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**, 3545–3559 (2015).
16. Zhu, L., Olsen, R. J., Nasser, W., de la Riva Morales, I. & Musser, J. M. 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, E. T. The serotypes of *Streptococcus pyogenes* present in Britain during 1980–1990 and their association with disease. *J. Med. Microbiol.* **39**, 165–178 (1993).
18. Gherardi, G., Vitali, L. A. & Creti, R. Prevalent *emm* types among invasive GAS in Europe and North America since year 2000. *Front. Public Health* **6**, 59 (2018).
19. Smit, P. W. et al. Epidemiology and *emm* types of invasive group A streptococcal infections in Finland, 2008–2013. *Eur. J. Clin. Microbiol. Infect. Dis.* **34**, 2131–2136 (2015).
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**, 864–872 (2015).
21. Naseer, U., Steinbakk, M., Blystad, H. & Caugant, D. A. Epidemiology of invasive group A streptococcal infections in Norway 2010–2014: a retrospective cohort study.*Eur. J. Clin. Microbiol. Infect. Dis.* **35**, 1639–1648 (2016).
22. Nelson, G. E. et al. Epidemiology of invasive group A streptococcal infections in the United States, 2005–2012. *Clin. Infect. Dis.* **63**, 478–486 (2016).
23. Plainvert, C. et al. Invasive group A streptococcal infections in adults, France (2006–2010). *Clin. Microbiol. Infect.* **18**, 702–710 (2012).
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, M. R. et al. Emergence of scarlet fever *Streptococcus pyogenes emm12* clones in Hong Kong is associated with toxin acquisition and multidrug resistance. *Nat. Genet.* **47**, 84–87 (2015).
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**, 1522–1534 (2012).
27. Hamilton, S. M., Stevens, D. L. & Bryant, A. E. Pregnancy-related group a streptococcal infections: temporal relationships between bacterial acquisition, infection onset, clinical findings, and outcome. *Clin. Infect. Dis.* **57**, 870–876 (2013).
28. Johnson, D. R., Stevens, D. L. & Kaplan, E. L. Epidemiologic analysis of group A streptococcal serotypes associated with severe systemic infections, rheumatic fever, or uncomplicated pharyngitis. *J. Infect. Dis.* **166**, 374–382 (1992).
29. Shea, P. R. et al. Group A *Streptococcus emm* gene types in pharyngeal isolates, Ontario, Canada, 2002–2010. *Emerg. Infect. Dis.* **17**, 2010–2017 (2011).
30. Smoot, J. C. 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. USA* **99**, 4668–4673 (2002).
31. Ben Zakour, N. L., Venturini, C., Beatson, S. A. & Walker, M. J. Analysis of a *Streptococcus pyogenes* puerperal sepsis cluster by use of whole-genome sequencing. *J. Clin. Microbiol.* **50**, 2224–2228 (2012).
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**, 665–670 (2002).
33. Gaworzewska, E. & Colman, G. Changes in the pattern of infection caused by *Streptococcus pyogenes*. *Epidemiol. Infect.* **100**, 257–269 (1988).
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**, 455–461 (2005).
35. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
36. Bricker, A. L., Carey, V. J. & Wessels, M. R. Role of NADase in virulence in experimental invasive group A streptococcal infection. *Infect. Immun.* **73**, 6562–6566 (2005).
37. Bricker, A. L., Cywes, C., Ashbaugh, C. D. & Wessels, M. R. NAD$^+$-glycohydrolase acts as an intracellular toxin to enhance the extracellular survival of group A streptococci. *Mol. Microbiol.* **44**, 257–269 (2002).
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**, 771–782 (2005).
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).
40. Meehl, M. A., Pinkner, J. S., Anderson, P. J., Hultgren, S. J. & Caparon, M. G. 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**, 4253–4260 (2007).
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).
43. Carroll, R. K. et al. Naturally occurring single amino acid replacements in a regulatory protein alter streptococcal gene expression and virulence in mice. *J. Clin. Invest.* **121**, 1956–1968 (2011).
44. Graham, M. R. 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. USA* **99**, 13855–13860 (2002).
45. Ribardo, D. A. & McIver, K. S. 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).
46. Ramalinga, A., Danger, J. L., 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**, e00452-16 (2017).
47. Trevino, J., Liu, Z., Cao, T. N., Ramirez-Pena, E. & Sumby, P. RivR is a negative regulator of virulence factor expression in group A *Streptococcus*. *Infect. Immun.* **81**, 364–372 (2013).
48. Nyberg, P., Rasmussen, M. & Bjorck, L. α_2-Macroglobulin-proteinase complexes protect *Streptococcus pyogenes* from killing by the antimicrobial peptide LL-37. *J. Biol. Chem.* **279**, 52820–52823 (2004).
49. Rasmussen, M., Muller, H. P. & Bjorck, L. Protein GRAB of *Streptococcus pyogenes* regulates proteolysis at the bacterial surface by binding α_2-macroglobulin. *J. Biol. Chem.* **274**, 15336–15344 (1999).
50. Toppel, A. W., Rasmussen, M., Rohde, M., Medina, E. & Chhatwal, G. S. Contribution of protein G-related α_2-macroglobulin-binding protein to bacterial virulence in a mouse skin model of group A streptococcal infection. *J. Infect. Dis.* **187**, 1694–1703 (2003).
51. Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W. & Livny, J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* **13**, 734 (2012).
52. Shishkin, A. A. et al. Simultaneous generation of many RNA-Seq libraries in a single reaction. *Nat. Methods* **12**, 323–325 (2015).
53. Engleberg, N. C., Heath, A., Miller, A., Rivera, C. & DiRita, V. J. 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**, 1043–1054 (2001).
54. Li, J. et al. Neutrophils select hypervirulent CovRS mutants of M1T1 group A *Streptococcus* during subcutaneous infection of mice. *Infect. Immun.* **82**, 1579–1590 (2014).
55. Mayfield, J. A. 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, A. R., Graviss, E. A., DeLeo, F. R. & Musser, J. M. 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).
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**, 3141–3149 (2009).
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**, 208–219 (1999).
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**, 1593–1603 (1993).
62. Beres, S. B. & Musser, J. M. Contribution of exogenous genetic elements to the group A *Streptococcus* metagenome. *PLoS One* **2**, e800 (2007).
63. Green, N. M. 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**, 760–770 (2005).
64. Coll, F. et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **50**, 307–316 (2018).
65. Earle, S. G. 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, J. E. & Marigorta, U. M. Expression quantitative trait locus analysis for translational medicine. *Genome Med.* **7**, 60 (2015).
67. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
68. Olsen, R. J. & Musser, J. M. Molecular pathogenesis of necrotizing fasciitis. *Annu. Rev. Pathol.* **5**, 1–31 (2010).
69. Rodriguez-Ortega, M. J. et al. Characterization and identification of vaccine candidate proteins through analysis of the group A *Streptococcus* surface proteome. *Nat. Biotechnol.* **24**, 191–197 (2006).
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**, 2086–2093 (2016).
71. Hammarlof, D. L. et al. Role of a single noncoding nucleotide in the evolution of an epidemic African clade of *Salmonella.Proc. Natl Acad. Sci. USA* **115**, E2614–E2623 (2018).
72. Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* **489**, 513–518 (2012).
73. Zaunbrecher, M. A., Sikes, R. D. Jr, Metchock, B., Shinnick, T. M. & Posey, J. E. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase *eis* confers kanamycin resistance in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA* **106**, 20004–20009 (2009).
74. Puopolo, K. M. & Madoff, L. C. Upstream short sequence repeats regulate expression of the alpha C protein of group B *Streptococcus*. *Mol. Microbiol.* **50**, 977–991 (2003).
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**, 142–149 (2000).
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, K. M. 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**, 1016–1031 (2016).
78. Jeukens, J. et al. Genomics of antibiotic-resistance prediction in *Pseudomonas aeruginosa.Ann. NY Acad. Sci.* **1435**, 5–17 (2017).
79. Nguyen, M. et al. Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci. Rep.* **8**, 421 (2018).
80. Pesesky, M. W. 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).
81. Rishishwar, L., Petit, R. A. 3rd, Kraft, C. S. & Jordan, I. K. Genome sequence-based discriminator for vancomycin-intermediate *Staphylococcus aureus*. *J. Bacteriol.* **196**, 940–948 (2014).
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).
83. Li, Y. et al. Penicillin-binding protein transpeptidase signatures for tracking and predicting beta-lactam resistance levels in *Streptococcus pneumoniae*. *mBio* **7**, e00756-16 (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).
86. Ongen, H. et al. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676–1683 (2017).
87. Tung, J., Zhou, X., Alberts, S. C., Stephens, M. & Gilad, Y. The genetic architecture of gene expression levels in wild baboons. *eLife* https://doi.org/10.7554/eLife.04729.001 (2015).
88. Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S. & Kruglyak, L. Genetics of single-cell protein abundance variation in large yeast populations. *Nature* **506**, 494–497 (2014).
89. Parker, C. C. et al. Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. *Nat. Genet.* **48**, 919–926 (2016).
90. Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. *Nature* **505**, 208–211 (2014).
91. Beres, S. B. et al. Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc. Natl Acad. Sci. USA* **107**, 4371–4376 (2010).
92. Olsen, R. J. et al. The majority of 9,729 group A *Streptococcus* strains causing disease secrete SpeB cysteine protease: pathogenesis implications. *Infect. Immun.* **83**, 4750–4758 (2015).

Acknowledgements

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 a 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 N. Copeland, N. Jenkins, and D. Ginsburg for critical comments and suggestions to improve the manuscript; K. Stockbauer for critical comments and editorial assistance; E. Graviss, H. Erlendsdottir, W. Hong, and S. Linson for technical assistance; H.-L. Hyyryläinen, J. Jalava, and the Finnish clinical microbiology laboratories; A. A. Shishkin for helpful suggestions regarding the RNAtag-seq protocol; M. Todorovic and J. Jonsdottir Nielsen for banking strains from the Faroe Islands; A. McGeer for Ontario strains; C. Van Beneden, B. Beall, and the Active Bacterial Core Surveillance of the CDC's Emerging Infections Programs network; A. Ramstad Alme and A. Witsø for technical assistance; and M. Steinbakk (Norwegian Laboratory for Streptococci) for support.

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. P.K. and J.M.E. contributed equally to this work, as did S.B.B., R.J.O., and L.Z.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0343-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.M.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 were performed as described previously[1,6,93]. The pipeline used for bacterial genome analysis is shown in Supplementary Fig. 2a. The trimmed sequence reads were corrected by using Musket[94] and mapped to the genome of the reference serotype M28 strain MGAS6180 (GenBank accession number CP000056)[63] by using SMALT (see URLs). SNPs and insertions and deletions (indels) were identified by using FreeBayes (see URLs) as described[1] and Pilon[95], 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 versus noncoding, synonymous versus nonsynonymous, and so on). Alternatively, and in conjunction, the SPAdes algorithm was used for de novo genome assembly[96]. Short Read Sequence Typing version 2 (SRST2) was used to identify genes, alleles, and multilocus sequence types[97]. The algorithm Gubbins was used to detect HGT events[98]. hierBAPS[99] (hierarchical Bayesian analysis of population structure; see URLs) was used to determine the population structure, and SplitsTree was used to estimate phylogenetic trees and networks[100]. hierBAPS was run with five replicates of the estimation algorithm by using prior upper-bound values for the number of clusters ranging between 50 and 200, with each run converging to the same posterior mode estimate of the population structure.

Long-read sequencing of 24 strains (Supplementary Table 19) was performed by using an Oxford Nanopore MinION instrument with R9.5 flowcells and a Rapid Barcoding Kit (SQK-RBK004). These strains were selected from the 50 strains for which RNA-seq 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 differed in the R28 promoter region T-nucleotide homopolymeric tract. Hybrid assemblies of the nanopore long reads and Illumina short read data were generated with Unicycler[101], as described previously[102]. Quality metrics for read quality filtering and trimming were done with Trimmomatic, read error correction was done with Musket, de novo assemblies were generated with SPAdes/Unicycler, multilocus sequence types and *emm* determination were done with SRST2, read mapping was done with SMALT, and polymorphism discovery was done with FreeBayes (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 by using hierBAPS. The reference strain for transcriptome analysis was selected by using *k*-means (described 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[103]. First, a SNP-based pairwise distance matrix was calculated separately for all *emm28* isolates belonging to a given lineage (SC1A, SC1B, SC2A, or SC2B) by using core SNPs. We excluded MGEs and potential regions of recombination identified by Gubbins[98]. 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, as in SEER[103]. For the given total size *k* of the subset to be chosen from a lineage, the *k*-means algorithm with 200 random restarts[104] was used to identify an optimal set of *k* centroids to span the variation present in the three-dimensional multidimensional scaling 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. *emm28 strains grown in triplicate and harvested at two time points.* Fifty strains representative of the four major genetic subclades (Fig. 1a,b and Supplementary Table 7) were assayed in triplicate at the midexponential and early-stationary phases of growth. RNA was extracted with an RNeasy kit (Qiagen) according to the manufacturer's instructions. Ribosomal RNA (rRNA) was depleted with a Ribo-Zero rRNA removal kit for Gram-positive bacteria (Illumina), as described previously[1,105]. 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 complementary DNA (cDNA) libraries were prepared with indexed reverse primers from the ScriptSeq Index PCR Primers kit (Illumina) and purified with AMPure XP 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).

emm28 strains grown as singleton cultures and harvested at one time point. Transcriptome analysis of the 461 singletons (Supplementary Tables 1 and 10) was performed by RNAtag-seq as described[52], with the modifications described herein. Total RNA isolated from each strain was quantified fluorometrically by using a Qubit RNA BR Assay kit (Life Technologies). Some 400 ng from each sample was fragmented for 3 min at 94 °C in a volume of 16 µl in 1× FastAP buffer, dephosphorylated with FastAP alkaline phosphatase (Thermo Fisher Scientific) for 12 min at 37 °C in a final volume of 20 µl, and phosphorylated at the 5′ end with T4 polynucleotide kinase (New England Biolabs) for 30 min at 37 °C in a final volume of 82 µl. Fragmented, dephosphorylated total RNA was purified by using 2× the volume (164 µl) of Agencourt RNAClean XP paramagnetic beads, according to the manufacturer's instructions, in 1.5-ml microcentrifuge 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[52] (Supplementary Table 17). For each strain, 5 µl of fragmented, phosphorylated total RNAs was ligated to 1 µl of the respective oligoribonucleotide at a final concentration of 5 µM by 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 addition of 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 microcentrifuge tubes. Thus, six pools containing eight samples each were made for each set of 48 samples by successively passing the total RNAs corresponding to the eight 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 eight strains each and one additional pool with total RNA from five 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 by using RNA Pico chips (Agilent Technologies). First-strand cDNA synthesis was performed as described previously[52]. For each pool, 12 µl of ribodepleted RNAs was 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)) and denatured for 2 min at 70 °C. Then, first-strand cDNA synthesis was performed by 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 the addition of 14 µl of water, single-stranded cDNAs (sscDNAs) were purified by using 2.5× the 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[52] and ligated by using T4 RNA Ligase 1, in a volume of 20 µl. The reactions were incubated overnight at 22 °C, and this was followed by two consecutive cleanup reactions using a 2.5× volume of Agencourt RNAClean XP beads, and eluted with 25 µl of water.

Library amplifications were performed with the universal primer univP5 (ref. [52]) and one of four distinct P7 barcode adapters (Supplementary Table 17). The sscDNA pools were organized in sets of four and individually amplified in a final volume of 50 µl, after a PCR enrichment test to determine the correct amplification conditions; this was followed by two cleanup steps using Agencourt RNAClean XP beads and elution in 20 µl of low TE buffer (10 mM Tris and 0.1 mM EDTA). For each sample, the cDNA library average size was determined by 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 four pools, each corresponding to eight strains, were mixed together at equimolar amounts. This process was repeated 14 additional times; 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.

Analysis of the 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. The read quality of the sequencing data was evaluated by using FASTQC software (see URLs). Adapter contamination and read-quality filtering was performed by using Trimmomatic[106]. Reads were mapped to the genome of reference strain MGAS6180 by using EDGE-pro[107], and the reads mapping to rRNA and transfer RNA genes were excluded from subsequent analyses. Additionally, genes with low expression were excluded from downstream analysis according to the strategy described in the Supplementary Note. Differential expression analysis was performed by using DESeq2 (ref. [108]). This same pipeline was used for the comparative RNA-seq analysis of the isogenic strains containing either nine or ten Ts in the homopolymeric tract between *Spy1336/R28* and *Spy1337*.

Analysis of the 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 according to the inline barcodes by using FASTX-Toolkit (see URLs). The median number of reads per pool was 92.6 million (Supplementary Fig. 7b), and the median number of reads per sample per pool ranged between 8 and 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 by using the R package variancePartition[109], 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 by using NOISeq-sim implemented in the NOISeq package[110]. Differentially expressed genes were identified in each of the 441 strains compared with the reference strain MGAS28737, selected as described in the Supplementary Note. DESeq2 (ref. [108]) was used for differential expression analysis for instances in which two-group comparisons were being made, where each group comprised more than one strain.

Machine learning with random forest analysis. The random forest analysis was done with MATLAB by using the function TreeBagger (MATLAB R2016b, Statistics and Machine Learning Toolbox, MathWorks). The aim was to train a random forest[59] for classification of outlier strains into four categories (*covR*, *covS*, *ropB*, and wild type) according to the transcriptome profile of the strains. The 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 wild type 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 eight outlier strains for which the transcriptome profiles over the 1,614 genes were known, but the class labels were unknown. Before learning of 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 by using the built-in measure for feature importance. This process 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 twofold cross-validation, for which a random forest (100 trees) was built by using two-thirds of the training data and evaluated on one-third of the training data. The out-of-sample performance of the submodels 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, on this basis, the ten most informative genes were selected for the final model. The final model was used to predict the class probabilities of the eight outlier strains (Supplementary Table 11).

GWAS analysis. GWAS analysis using SEER[103] was performed with the de novo assemblies of 442 strains for which we also had RNA-seq data. GWAS was used to identify genetic variants significantly associated with a binary phenotypic grouping (high transcript expression = 1; low transcript expression = 0), defined according to the 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 (one-third of strains) and high expressers (two-thirds of strains)—analogous to our observations for the 50 strains (Fig. 6b). We identified the threshold and found that strains with normalized counts of less than 261.5 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 by using fsm-lite (see URLs). To account for the population structure, the distance matrix computed by Mash[111] 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. eQTL analysis was performed by using the R package MatrixEQTL[112]. 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 by using a mouse model of necrotizing myositis. Mouse necrotizing myositis studies with serotype M28 GAS strains were performed as described previously[39,68]. The 50 naturally occurring strains used in the initial RNA-seq experiment, including virulence reference strain MGAS28426, were used. The strain MGAS28426 was used as the virulence reference strain because it is genomically representative of SC1A strains, and it was used as a wild-type reference for the secreted SPN assays. Frozen stocks of each strain were prepared and quantified by counting colony-forming units (c.f.u.) recovered from thawed cultures after serial dilutions were made. Immunocompetent 4-week-old female outbred CD1 mice (Envigo) were randomly assigned to strain treatment groups and inoculated in the right lower hindlimb with 5×10^8 c.f.u. of each bacterial strain ($n=20$ mice per strain; 1,000 mice in total). This dose was selected on the basis of two pilot experiments that showed that the virulence reference strain MGAS28426 caused approximately 50% near-mortality at an inoculation dose of 5×10^8 c.f.u. This strategy facilitated the identification of comparator strains with significantly increased or decreased virulence. The mouse sample size was selected by using a power calculation with the following variables: $\alpha=0.05$; power $(1-\beta)=0.8$; difference in survival rates between groups = 0.4; ratio of group size = 1.

For the *ccovRS* mutant strain comparison, four cluster A strains and four cluster B strains ($n=45$ mice per strain) were used at a dose of 5×10^8 c.f.u. For the parental wild type (27961) and isogenic promoter mutant (27961-SC1A *nga* promoter) comparison ($n=40$ mice per strain), 5×10^8 c.f.u. was used. For the *ropB* mutant strain comparison, 3 group I and 4 group II strains ($n=40$ mice per strain) were used at a dose of 1×10^9 c.f.u. For the 9T and 10T isogenic strains (27961-9T and 27961-10T) and the control strain 28085-10T ($n=20$ mice per strain), we used a dose of 5×10^8 c.f.u. 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 (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 at the Methodist Hospital Research Institute. 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*[113]. Survival data were expressed as Kaplan–Meier curves, and statistically significant differences were determined with the log-rank test (Prism6; GraphPad).

Statistical analysis. Unless otherwise stated, error bars represent s.d., and P values were calculated with Fisher's exact, Mann–Whitney U, or log-rank tests. The false discovery rate 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 PCA and to find additional substructure in clusters, respectively. The random forest analysis was done in MATLAB by using the function TreeBagger. The R package variancePartition was used to confirm the absence of significant batch effects. Coefficient of determination (R^2) statistics were used to investigate whether a correlation existed between the 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 with RNAtag-seq data collected from strains grown as singletons.

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 under accession number GSE113058.

Ethics. All mouse studies were performed in accordance with a protocol (AUP-0318–0016) approved by the Institutional Animal Care and Use Committee at the 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 at the National Institute of Allergy and Infectious Diseases. All study volunteers gave written informed consent.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

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

93. Beres, S. B. et al. Genome sequence analysis of *emm89 Streptococcus pyogenes* strains causing infections in Scotland, 2010–2016. *J. Med. Microbiol.* **66**, 1765–1773 (2017).
94. Liu, Y., Schroder, J. & Schmidt, B. Musket: a multistage *k*-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**, 308–315 (2013).
95. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
96. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
97. Inouye, M. et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).
98. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
99. Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
100. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
101. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
102. Long, S. W., Kachroo, P., Musser, J. M. & Olsen, R. J. Whole-genome sequencing of a human clinical isolate of *emm28 Streptococcus pyogenes* causing necrotizing fasciitis acquired contemporaneously with Hurricane Harvey.*Genome Announc.* **5**, e01269-17 (2017).
103. Lees, J. A. et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).
104. Bishop, C. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
105. Eraso, J. M. 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).
106. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
107. Magoc, T., Wood, D. & Salzberg, S. L. EDGE-pro: estimated degree of gene expression in prokaryotic genomes. *Evol. Bioinform. Online* **9**, 127–136 (2013).
108. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome. Biol.* **15**, 550 (2014).
109. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
110. 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).
111. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome. Biol.* **17**, 132 (2016).
112. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
113. Committee for the Update of the Guide for the Care and Use of Laboratory Animals, Institute for Laboratory Animal Research & Division on Earth and Life Studies *Guide for the Care and Use of Laboratory Animals* 8th edn. (National Academies Press, Washington, DC, 2011).

Corresponding author(s): James M. Musser

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

- Trimmomatic: Read quality trimming. Version: 0.36
- Musket: Read error correction. Version: 1.1
- SRST2: detection of genes, alleles, and multi-locus sequence types (MLSTs). Version: 0.2.0
- SPAdes: de novo genome assembly. Version: 3.9.1
- SMALT: Read mapping to reference strain, (available at www.sanger.ac.uk/resources/software/smalt/). Version: 0.7.6
- FreeBayes: Identification of SNPs and InDels, (available at www.github.com/ekg/Freebayes/). Version: 1.0.2-16
- Pilon: Identification of InDels. Version: 1.22
- bcl2fastq: Demultiplexing of reads from superpools into separate pools, (https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html). Version: 2.20
- FASTX-toolkit: demultiplexing of reads from each pool into separate fastq files. Version: 0.0.14
- FASTQC: Read quality assessment, (available at https://www.bioinformatics. babraham.ac.uk/projects/fastqc/). Version: 0.11.6
- Unicycler: Hybrid assembly pipeline. Version: 0.4.1
- EDGE-Pro: Read mapping to the genome of reference strain MGAS6180. Version: 1.3.1

Data analysis

- Gubbins: Identification of potential regions of recombination.
- SNPfx: PERL script developed in house used to determine the nature of the SNPs (coding/noncoding, synonymous/nonsynonymous, etc). Custom script made in the laboratory.
- BAPS: Determination of population structure, (Bayesian Analysis of Population Structure, available at http://www.helsinki.fi/bsg/software/BAPS/)

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

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size

We performed whole genome sequencing on 2,101 strains, analyzed the transcriptome of 492 strains (23.4% of the entire population) and did virulence assays on 50 strains (2.4% of the entire population) for which we also had transcriptome data. We arbitrarily opted to use ~25% of strain cohort for the expression studies. For an unbiased selection of strains for the transcriptome analysis from the population of 2,101 strains, we used a projection-based approach. We sampled isolates proportionately to the fraction of the total population size represented by each genetic lineage. For the given total size size k of the subset to be chosen from a lineage, the k-means algorithm was used to identify an optimal set of k centroids to span the variation present with the lineage. Finally, the isolate with the minimum Euclidean distance to each centroid was chosen to determine the final subset of k representative isolates.

The 5x10^8 CFU inoculum used for the mouse infection studies was determined by Probit analysis (XLSTAT2011) of survival data generated from a dose-escalation study with virulence reference strain MGAS28426. A power calculation for the log-rank test (www.statstodo.com) predicted that a sample size of 20 mice per strain treatment group was needed to generate statistically significant results when using the pilot study survival rates and parameters alpha=0.05, beta=0.8, and r=1. Power calculations for other virulence experiments described in this study were performed in a manner as described above.
Transcriptome and virulence assays were done on strains numerically representative from the 3 major genetic subclades identified in our study. These values do not reflect the 19 strains that were excluded (see below).

Data exclusions

19 strains were excluded from the transcriptome analysis part of the study for technical reasons associated with contamination or low read coverage. The exclusion criteria were pre-established.

Replication

Transcriptome analyses for 50 strains were conducted in triplicate (3 biological replicates) and extremely high correlation coefficients were found for the 3 replicates representing each strain. For each strain, correlation between biological replicates, at each growth phase, was computed using Pearson correlation coefficient (r). At mid-exponential phase we found r>99% and at early-stationary phase r values >90% were obtained. 442 strains were analyzed as singleton (expression analyses were done without replicates). To assure the quality of the data we also set the criteria for sequencing depth at 5-10 million sequencing reads per strain or biological replicate. For the validation of the singleton strategy 7 of the 50 strains analyzed in triplicate were also analyzed as singletons, and yielded a very strong correlation in the global expression profile. Correlation was determined using Pearson correlation coefficient (r) and found to be between 86-92%
For the mouse experiments reproducibility was established in multiple ways. We used a minimum sample of 20 mice per bacterial strain per experiment. Prior to performing the 50 strain experiment, a pilot experiment was performed using reference strain MGAS28426 to select the appropriate dose. The pilot experiment established the baseline virulence phenotype of reference strain MGAS28426. The virulence phenotype of reference strain MGAS28426 was reproduced in the 50 strain experiment. Second, the virulence phenotype of 7 strains,

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

Randomization

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

Blinding

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

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

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

Antibodies

Antibodies used

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

Validation

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

Animals and other organisms

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

Laboratory animals

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

Wild animals

The study did involve the use of wild animals

Field-collected samples

The study did not involve samples collected from the field