Natural variation of the streptococcal Group A carbohydrate biosynthesis genes impacts host-pathogen interaction

Sara M. Tamminga[1], Kim Schipper[1], Nicholas Murner[2], Matthew Davies[1,5], Paul Berkhout[1], Debra E. Bessen[4], Astrid Hendriks[1], Natalia Korotkova[2,3], Yvonne Pannekoek[1] and Nina M. van Sorge[1,5,#]

[1]Department of Medical Microbiology and Infection Prevention, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands

[2]Department of Microbiology, Immunology and Molecular Genetics, University of Kentucky, Lexington, KY, USA

[3]Department of Molecular and Cellular Biochemistry, University of Kentucky, Lexington, KY, USA

[4]Department of Pathology, Microbiology, and Immunology, New York Medical College, Valhalla, New York, USA

[5]Netherlands Reference Laboratory for Bacterial Meningitis (NRLBM), Amsterdam UMC location AMC, Amsterdam, The Netherlands

#Corresponding author: Prof. N.M. van Sorge, PhD, Amsterdam UMC location University of Amsterdam, Department of Medical Microbiology and Infection Prevention, Meibergdreef 9, IWO building IA3.211, 1105 AZ Amsterdam, the Netherlands. E: n.m.vansorge@amsterdamumc.nl

Abstract

Streptococcus pyogenes (*S. pyogenes*) is a leading cause of infection-related mortality in humans globally. The characteristic cell wall-anchored Group A Carbohydrate (GAC) is expressed by all *S. pyogenes* strains and consists of a polyrhamnose backbone with alternating *N*-acetylglucosamine (GlcNAc) side chains, of which 25% are decorated with glycerol phosphate (GroP). The genes in the *gacA-L* cluster are critical for GAC biosynthesis with *gacI-L* being responsible for the characteristic GlcNAc-GroP decoration, which confers the agglutination in rapid test diagnostic assays and contributes to *S. pyogenes* pathogenicity. Seminal research papers described *S. pyogenes* isolates, so-called A-variant strains, that lost the characteristic GlcNAc side chain following serial animal passage. We performed genomic analysis of a single viable historic parent/A-variant strain pair to reveal a premature inactivating stop codon in *gacI*, explaining the described loss of the GlcNAc side chain. Subsequently, we analyzed the genetic variation of the 12 *gacA-L* genes in a collection of 2,044 *S. pyogenes* genome sequences. Although all *gac* genes (*gacA-L*) displayed genetic variation, we only identified 31 isolates (1.5%) with a premature stop codon in one of the *gac* genes. Nearly 40% of these isolates contained a premature stop codon in *gacH*. To study the functional consequences of the different premature stop codons for GacH function, we plasmid-expressed three *gacH* variants in a *S. pyogenes gacH*-deficient strain. Cell wall analysis confirmed GacH loss-of-function through the significant reduction of GroP. Complementary, we showed that strains expressing *gacH* loss-of-function variants were completely resistant to the human bactericidal enzyme group IIA-secreted phospholipase. Overall, our data provide a comprehensive overview of the genetic variation of the *gacA-L* gene cluster in a global population of *S. pyogenes* strains and the functional consequences of *gacH* variation for immune recognition and clearance.

Keywords

Streptococcus pyogenes, Group A *Streptococcus*, group A carbohydrate, group IIA-secreted phospholipase, PubMLST, *N*-acetylglucosamine, glycerol phosphate

Abbreviations

S. pyogenes, *Streptococcus pyogenes;* GAC, group A carbohydrate; GlcNAc, *N*-acetylglucosamine; GroP, glycerol phosphate; hGIIa, human group IIA-secreted phospholipase A2.

Data summary

All *S. pyogenes* genome sequences used for this analysis are available within the publication by Davies *et al.* (2019), 'Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics' Nature Genetics, 51(6):1035-43

Introduction

Streptococcus pyogenes (*S. pyogenes*) is a human-restricted pathogen, responsible for hundreds of millions of infections globally each year (1, 2). *S. pyogenes* causes a wide spectrum of clinical manifestations ranging from pharyngitis and impetigo to invasive infections *e.g.* puerperal sepsis, necrotizing fasciitis, and streptococcal toxic shock syndrome. Moreover, repeated infections can result in the development of post-infectious sequelae such as glomerulonephritis and rheumatic heart disease. Together, infections by this single pathogen result in an estimated 500,000 deaths annually worldwide, with resource-limited areas and Indigenous populations being affected disproportionally (1-3). Developing new strategies for effective treatment and prevention of *S. pyogenes* infection and their complications remains a critical public health priority, as safe and effective vaccines are not yet available.

The serological classification of *S. pyogenes* depends on the expression of the cell wall–anchored Group A Carbohydrate (GAC) (4). The GAC is universally conserved by all *S. pyogenes* strains, which has resulted in application of GAC-reactive rapid diagnostic test kits for species identification. These GAC glycopolymers comprise up to half of the cell wall mass and are composed of a linear polyrhamnose chain decorated with *N*-acetylglucosamine (GlcNAc) side chains (5, 6). Recently researchers showed that 25% of GlcNAc side chains are further modified with negatively-charged glycerol phosphate (GroP) moieties (7). GAC is important for the structural integrity of the bacterial cell wall, and biosynthesis of the polyrhamnose backbone is essential for the viability of *S. pyogenes* (5, 8-10). Removal of the GlcNAc-GroP side chain resulted in increased *in vitro* killing by human whole blood, neutrophils, and platelet releasate and attenuated virulence in murine and rabbit infection models (11), whereas removal of just the GroP moiety resulted in resistance against human cationic antimicrobial proteins, human bactericidal enzyme group IIA-secreted phospholipase (hGIIa), lysozyme and histones, while increasing susceptibility to zinc (7, 12).

The 12-gene *gac* cluster is crucial for GAC biosynthesis and exhibits limited genetic variation across the *S. pyogenes* population (11, 13, 14). The first seven genes of the cluster (*gacABCDEFG*) encode proteins catalyzing the biosynthesis and transport of the polyrhamnose backbone (8, 11, 15). GacIJKL,

encoded by *gacIJKL*, are involved in the decoration of the polyrhamnose backbone with the GlcNAc side chain. Finally, GacH is the GroP transferase enzyme that cleaves membrane phosphatidylglycerol to release and attach GroP to C-6 of the GlcNAc side chains (7). Interestingly, historic research reported isolates, referred to "A-variants", in which GAC lost its characteristic GlcNAc side chain after serial passage in mice and rabbits (16, 17). However, the underlying mechanisms responsible for the loss of the GAC GlcNAc side chain were never reported. Additionally, this A-variant phenotype was not detected among a stock collection of human isolates (17).

With the current availability of whole genome sequences, we aimed to re-examine the concept that strains with a non-canonical GAC may arise in humans. To this end, we first analyzed the genomic sequences of a historic *S. pyogenes* A-variant/parent strain pair to pinpoint the genetic alteration that could explain the loss of the GlcNAc side chain in the evolved A-variant strain. Additionally, we analyzed the genetic variation of *gacA-L* in a collection of 2,044 *S. pyogenes* genomes and identified potential inactivating mutations in *gacH*. We investigated the impact of these mutations on GacH function by complementing a *gacH*-deficient strain with plasmid-expressed truncated *gacH* variants and subsequent detection of GroP by biochemical, phenotypical, and functional analysis.

Methods

Antibody ab9191 binding assay

S. pyogenes strains were grown to midlog phase in THY (OD_{600} 0.4), centrifuged, resuspended in PBS containing 0.1% Bovine serum albumin (BSA; Sigma) and incubated with 1 µg/ml ab9191, a goat polyclonal Group A carbohydrate antibody (Abcam), for 20 min at 4 °C. After washing, bacteria were resuspended in PBS 0.1% BSA containing Protein G Alexa fluor 488 conjugate (1:1,000) for ab9191 conditions. Fluorescence was analyzed by FACSCanto II Flow Cytometer (BD Bioscience). Per sample, 10,000 gated events were collected and fluorescence was expressed as the geometric mean.

Group A *streptococcus* PubMLST analysis of *gac* gene cluster

Whole genome sequences (Illumina short-read) of 2,109 *S. pyogenes* isolates were uploaded to the

open-access PubMLST database (www.pubmlst.org) (13, 18). *gacA-gacL* were annotated and screened for genetic variation, where allele 1 was assigned to reference strain MGAS5005 (19). Strains of which one of the *gac* genes was truncated because it was located at the end of a contig (n = 65) were excluded from the analysis. The sequence of allele 1 was blasted against the genomes of strains without an allele number using the PubMLST database to identify sequences with a premature stop codon. Default BLAST settings were used: BLASTN word size:11; BLASTN scoring: reward: 2; penalty: -3; gap open: 5; gap extend: 2; Hits per isolate: 1; Flanking length (bp): 0.

Plasmids, bacterial strains and culture conditions

All plasmids and *S. pyogenes* strains used in this study for experimental assays are listed in Table S1. *Escherichia coli* (*E. coli*) was grown in Lysogeny Broth (LB, Oxoid) medium or LB agar plates, supplemented with 500 µg/ml erythromycin at 37 °C. *S. pyogenes* strains were grown on Todd Hewitt supplemented with 0.5 % yeast (THY, Oxoid) agar plates or in THY broth supplemented with 5 µg/ml erythromycin for the plasmid-complemented strains. No antibiotics were added for wild type strains and the *gacH* knockout. Overnight cultures were grown at 37 °C, subcultured the next day in fresh THY and grown to mid-exponential growth phase, corresponding to an optical density at 600 nm (OD_{600}) of 0.4.

Generation of complemented *S. pyogenes* Δ*gacH* strains with *gacH* variant

All primers used in this study are listed in Table S2. Expression vector pDCerm_*gacH* (*gacH* allele 1) was isolated from *E. coli* MC1061. The plasmid was digested with EcoRI and BglII. For complementation of the *S. pyogenes* Δ*gacH* knockout strain with the different *gacH* variants, different strategies were used. To create *gacH* complementation plasmids that contained a premature stop codon at nucleotide position 310 and 313, sequences were codon optimized and gblocks were ordered. The gblocks were amplified with primer STOP310AAF/ STOP310AAR or STOP313AAF/STOP313AAR, digested with EcoRI and BglII and ligated into EcoRI/BglII-digested pDCerm. Correct insertion was confirmed by Sanger sequencing with primers pDCermF, pDCermR, *gacH*310_checkF1, *gacH*310_checkR1 or pDCermF, pDCermR, STOP313checkF, STOP313checkR. For complementation with *gacH* containing a premature stop codon at nucleotide position 774, a gblock

could not be generated, even after codon optimization. Therefore, wild type *S. pyogenes* strain 20162146, which was identified to contain this *gacH* stop codon, was obtained from the Centers for Disease Control (CDC, Atlanta, GO, USA). The *gacH* gene was amplified with primers *gacH*EcoRIF and *gacH*BglIIR, followed by a similar procedure as described for the gblocks. The correct insert was confirmed with primers *gacH*check1, *gacH*check2, *gacH*check3, pDCermF and pDCermR. Plasmids with the correct inserts were transformed to *S. pyogenes* 5448Δ*gacH*.

Measurement of relative phosphate concentrations on GAC

S. pyogenes cell wall was isolated from late exponential phase cultures (OD_{600} ~0.8) by the SDS-boiling procedure as previously described (12, 20). Purified cell wall samples were lyophilized and stored at −20°C before the analysis. GAC was released from cell wall preparations by mild acid hydrolysis as previously described (12). After hydrolysis, samples were purified by running over a PD-10 desalting column (VWR, 17-0851-01), with de-ionized water as the exchange buffer. Phosphate was released from GAC as outlined (7). Briefly, soluble GAC was incubated with 2 N HCl at 100 °C for 2 h to cleave GroP. Samples were neutralized with NaOH in the presence of 62.5 mM HEPES pH 7.5. To release phosphate from GroP, samples (100 μL) were incubated with 2 μL of 1 U/μL alkaline phosphatase (ThermoFisher) in alkaline phosphatase buffer at 37°C overnight. Phosphate concentrations were measured using the malachite green method. The reactions were diluted to 160 μL with water and 100 μL was transferred to a flat-bottom 96-well culture plate. Malachite Green reagent (0.2 mL) was added and the absorbance at 620 nm was read after 10 min at room temperature. Malachite Green reagent contained one volume 4.2% ammonium molybdate tetrahydrate (by weight) in 4 M HCl, 3 volumes 0.045% malachite green (by weight) in water and 0.01% Tween 20. Phosphate concentrations were determined using a phosphate standard curve. Concentrations of rhamnose (Rha) in GAC were measured by an anthrone assay as previously described (12). The concentrations of phosphate were normalized to total Rha content and presented as a percentage of the ratios in the WT strain.

Human group IIA-secreted phospholipase A2 killing assay

hGIIa killing was determined as described before (21). Briefly, *S. pyogenes* strains were grown to midlog phase (OD_{600} 0.4), diluted 1,000-fold in HEPES solution (20 mM HEPES, 2 mM Ca^{2+}, 1% BSA (pH 7.4)) with or without 0.5 µg/ml hGIIa. Samples were incubated for 2 hours at 37˚C and viability was assessed using serial dilutions plating on blood agar plates. The next day, colony forming units (CFU) were counted. Percentage survival was calculated as follows: Survival (%) = (counted CFU * dilution)/CFU count at 0 µg/ml hGIIa * dilution)*100.

Statistical analysis

Flow cytometry data were analyzed using FlowJo 10 (FlowJo). Data were analyzed using GraphPad Prism 10.2.0 (GraphPad software). Statistical significance was tested using an unpaired t test or a one-way ANOVA followed by a Dunnett's test for multiple comparison. The *P* values are depicted in the figures or mentioned in the caption, and $P<0.05$ was considered significant.

Results

We had a single viable A-variant/wild type strain pair available from the historical Lancefield streptococcal collection (https://www2.rockefeller.edu/vaf/lanceindex.php). The A-variant phenotype of strain D315/87/3 was confirmed by absence of StrepTex latex agglutination and loss of binding of a GAC-GlcNAc reactive polyclonal antibody (Figure 1). Next, we compared whole genome sequences of the strain pair and identified a nucleotide substitution T382C in the *gacI* sequence of D315/87/3. This mutation resulted in a premature stop codon at amino acid position 127 (out of 232 amino acids) of GacI, thereby reducing GacI size by 45%. The identified *gacI* mutation likely explains the absence of GlcNAc in the historic mouse-passaged A variant D315/87/3.

To comprehensively analyze the genetic variation in *gacA-L*, we analyzed allelic variation and presence of specific mutations in these genes in 2,044 *S. pyogenes* genomes (13). The number of alleles varied between 35 (*gacJ*) and 287 (*gacH*) (Table 1). Estimation of the numbers of synonymous and nonsynonymous nucleotide substitutions revealed that none of the *gac* genes are subjected to positive

or purifying selection (22, 23). Overall and as expected, the *gac* gene cluster was genetically highly conserved in the *S. pyogenes* population.

Both the number of alleles and unique proteins of each *gac* gene correlated strongly with gene size (Figure 2A, B). *gacH* displayed the highest number of unique protein sequences (Figure 2B). The genetic diversity in *gacH* was also reflected by the fact that the most abundant GacH protein sequence was present in <10% of the 2,044 *S. pyogenes* isolates (Figure 2C). This indicates that *gacH*, responsible for the recently discovered GroP modification on GlcNAc, is the most variable gene in the *gac* cluster based on our dataset. Interestingly, for all *gac* genes except *gacJ*, we identified variants that contained a premature stop codon (Table 1). Overall, we identified 31 isolates (1.5% of all analyzed strains) that contained a premature stop codon in one of the *gac* genes. For 12 strains (38%), this premature stop codon was located in *gacH* (Table 1). These isolates did not cluster based on *emm* type, multi locus sequence type (MLST), or year of isolation (ranged from 1990-2015). Furthermore, these strains were isolated from different disease manifestations (pharyngitis n=3, skin/soft tissue infection n=2, invasive infection n= 4, meningitis n=1, unknown n=2) and originated from different continents (Oceania n=5, Africa n=2, Asia n=1, North America n=3 and Europe n=1).

GacH consists of a transmembrane domain (Figure 3A; amino acids 1 – 395; green) and a catalytic domain on the extracellularly site of the membrane (Figure 3A; amino acids 444 – 822, red). The premature stop codons were located at amino acid position 15 (n=1), 310 (n=3), 313 (n=1), 330 (n=1), 754 (n=1) and 774 (n=5) (Figure 3A). We aimed to analyze the impact of the naturally-occurring premature stop codons on GacH protein function. We therefore visualized the locations of the stop codons in a 2D-scaled model, based on previously published crystal structure and enzymatic analysis (7). Figure 3B shows the full-length WT GacH (allele 1; MGAS5005). Premature stop codons at positions 310 and 313 are predicted to result in only a partial transmembrane domain, and a complete loss of the extracellular catalytic domain of GacH (Figure 3C). The premature stop codon at amino acid position 774 is predicted to result in partial loss of the catalytic domain (Figure 3D), which may result in a destabilized or misfolded GacH protein.

To study the functional consequences of the premature stop codons in *gacH*, we expressed three *gacH* variants (stop codon at position 310, 313, and 774) on a plasmid in a *S. pyogenes* mutant lacking *gacH* (5448Δ*gacH*). Furthermore, we included a wild-type strain (20162146) that contained the naturally-occurring premature stop codon in *gacH* at amino acid position 774. To determine the presence of GroP on the GAC GlcNAc side chain, we measured the phosphate content in the isolated GAC in the different *S. pyogenes* strains (7). As expected, GAC isolated from 5448Δ*gacH* contained a significantly reduced amount of phosphate compared to WT 5448 GAC (Figure 4A). This phenotype could be restored by complementation with a plasmid containing WT *gacH*, but not with *gacH* variants encoding truncated GacH variants from amino acid position positions 310, 313 or 774 (Figure 4A). Similarly, the WT strain 20162146, which naturally acquired the premature stop codon at position 774 in *gacH*, showed strongly reduced levels of phosphate in the isolated GAC (Figure 4A).

The presence of GroP confers susceptibility of *S. pyogenes* to the bactericidal enzyme hGIIa (7). To confirm the functional implications of GroP loss in the complemented strains expressing *gacH* variants, we determined bacterial survival of these *gacH* variants in hGIIa-killing assay. Similar to the 5448Δ*gacH* mutant, all premature stop codon variants displayed resistance towards the bactericidal enzyme hGIIa compared to the isolate that plasmid-expressed wild-type *gacH* (Figure 4B). Overall, both the biochemical and functional assay confirmed that *gacH* variants with an early stop codon (amino acid positions 310 and 313) and even a premature stop codon close to the C-terminus (amino acid position 774 out of 825) were defective in modifying of GAC with GroP (Figure 4).

Discussion

GAC is a major and characteristic cell wall component of *S. pyogenes* and plays important roles in bacterial physiology and pathogenesis. Functional and structural analysis of the GAC has been performed for a few well-known *S. pyogenes* laboratory strains, where deletion of *gacI* results in reduced survival in human blood and animal models of systemic infection, and removal of *gacH* renders *S. pyogenes* susceptible to zinc and resistant to the host cationic antimicrobial peptides including hGIIA

(7, 11, 12, 21). Here, we identified that a mutation in *gacI*, resulting in a premature stop codon, likely underlies the historic A-variant phenotype that has lost the characteristic GAC GlcNAc side chain after frequent animal passage. Furthermore, by analyzing 2,044 *S. pyogenes* genomes, we observed *gacH* to be the most variable gene in the well-conserved *gac* gene cluster. Although rare, we identified a small number of clinical *S. pyogenes* isolates that contain a *gacH* allele with a premature stop codon. By expressing these allelic variants in a *gacH*-deficient strain, we demonstrated that these genetic variants are severely attenuated in their enzymatic activity, resulting in loss of the GroP side chain and acquiring resistance to hGIIA.

In this study, we analyzed a geographically and clinically diverse collection of *S. pyogenes* genome sequences, comprising 150 different *emm* types and 484 MLST types (13), to obtain a more comprehensive overview of the variability of the *gac* genes across the *S. pyogenes* population. We uploaded the genome sequences to PubMLST, which comprises an important freely-accessible tool for *S. pyogenes* research on the presence and genetic variation of specific genes at the population level (18). Our results thereby expand observations from previous work, where variation in *gac* genes was analyzed in 520 of these 2,044 *S. pyogenes* strains (14). Similar to our study, this study also reported the existence of strains with premature stop codons although these variants were not functionally confirmed. In the expanded data set, we identified 10 isolates with premature stop codons in genes *gacA-gacG*, which are critical for biosynthesis of the polyrhamnose backbone. In addition, 21 isolates were identified that contained a premature stop codon in genes that are critical for decoration of polyrhamnose with GlcNAc-GroP (*gacH-gacL*). Twelve (38%) of these 21 isolates contained a premature stop codon in *gacH* and had a unique *emm* type and MLST profile.

We further analyzed the *gacH* premature stop codons for functional consequences to GAC biosynthesis. In five of the 12 isolates, the premature stop codon was located at amino acid position 774 in GacH. Since these five strains were isolated from different pathologies and different continents, the acquisition of this premature stop codon evolved independently. Despite preserving 93% of the protein, the stop codon at position 774 resulted in the loss of GroP, suggesting that the last 51 amino acids of the C-terminus of GacH are crucial for its enzymatic activity.

gacH was the most variable gene compared to the other genes in the *gac* gene cluster. Although this may suggest greater selective pressure on *gacH* and the GroP modification, dN/dS analysis did not provide evidence for this. The function of the recently identified GroP modification is not completely clear yet. Researchers reported that the presence of the negatively-charged GroP group protects *S. pyogenes* from zinc toxicity (7), which is used by the host to reduce intracellular bacterial survival (24). In contrast, GroP renders *S. pyogenes* more susceptible to hGIIA-mediated killing, which is likely important in defense against *S. pyogenes* in human blood or tissues (21). However, since *S. pyogenes* already shows high intrinsic resistance towards hGIIA, we do not expect high selective pressure of hGIIA on *S. pyogenes* to lose the GroP group. Indeed, we only found 12 (0.6%) isolates that contained a *gacH* variant with a premature stop codon.

In addition to gene sequence variation, environmental conditions may also affect *gacI* and *gacH* expression or enzymatic activity. Historically, it was reported that GlcNAc is present in a 1:2 ratio to the rhamnose backbone (5, 25, 26). Furthermore, approximately 25% of GlcNAc residues contains a GroP group (7). Possibly, the enzymes implicated in biosynthesis of the GlcNAc-GroP epitope exhibit different expression levels or activities under different environmental conditions. Whether and how the amount of GlcNAc and GroP present on GAC is regulated remains to be determined.

In conclusion, the *gacA-L* gene cluster is highly conserved in presence and genetic sequence. We showed that there are few exceptions in which *gac* genes are present but contained a premature stop codon. For three of these *gacH* stop codon variants, we confirmed that the GAC-GroP modification is lost, which resulted in increased resistance to hGIIA. Studying the genetic diversity of the GAC gene cluster in a representative global collection of *S. pyogenes* strains provides more insight into the structural and functional diversity of GAC, which can help in the development of anti-*S. pyogenes* preventive and therapeutic interventions, such as monoclonal antibodies and vaccines.

Funding information

This work was supported by the Vici (09150181910001) research programme to N.M.v.S. which is financed by the Dutch Research Council (NWO) and NIH grant (R01 AI143690) from the NIAID to N.K. This study also made use of the *S. pyogenes* PubMLST database (https://pubmlst.org/spyogenes), which is funded by the Wellcome Trust.

Author contributions

Conceptualization: NvS, YP; Data curation: KS, YP; Formal analysis: ST, KS; Funding acquisition: NvS, NK; Investigation: ST, KS, NM; Resources: DB, NK, NvS; Software: MD, PB; Supervision: NvS, YP, AH; Validation: ST, KS, NM; Visualization: ST, KS, NM, NK; Writing – original draft: ST, KS; Writing – review & editing: NvS, AH, NK, YP. All authors read and contributed to the final manuscript.

Conflicts of interest

NMvS declares royalties related to a patent (WO 2013/020090 A3) on vaccine development against *Streptococcus pyogenes* (Vaxcyte; Licensee: University of California San Diego with NMvS as co-inventor). NMvS is a member of the Science advisory Board for the Rapua project (paid to institution; Project facilitating Strep A vaccine development for Aotearoa New Zealand). The other authors declare no conflict of interest for the submitted work.

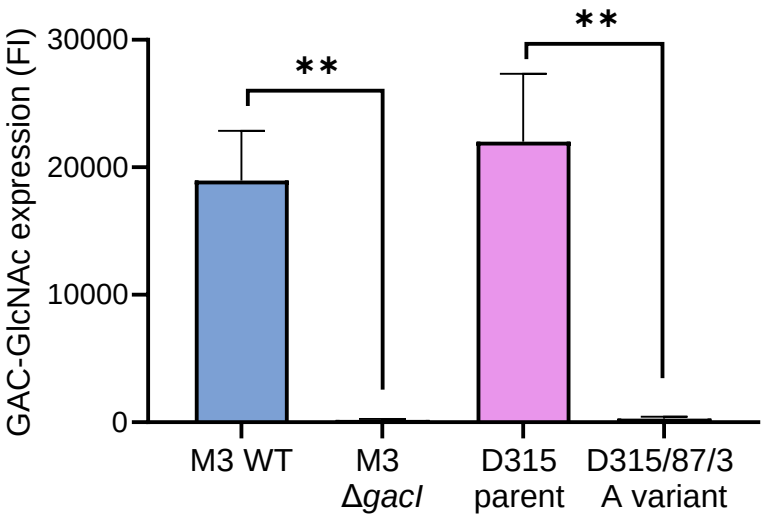

Figure 1 Historical A-variant lacks GlcNAc side chain

Binding of goat polyclonal *Streptococcus pyogenes* Group A carbohydrate antibody ab9191 (1 µg/ml) to wild type (WT) *S. pyogenes* M3, an isogenic *gacI* mutant, the historic A variant (D315/87/3) and its parent strain (D315). Data are depicted as geometric mean fluorescence intensity (FI) of three individually displayed biological replicates (mean + standard deviation). *P* values were calculated by two unpaired t tests; ** $P<0.01$.

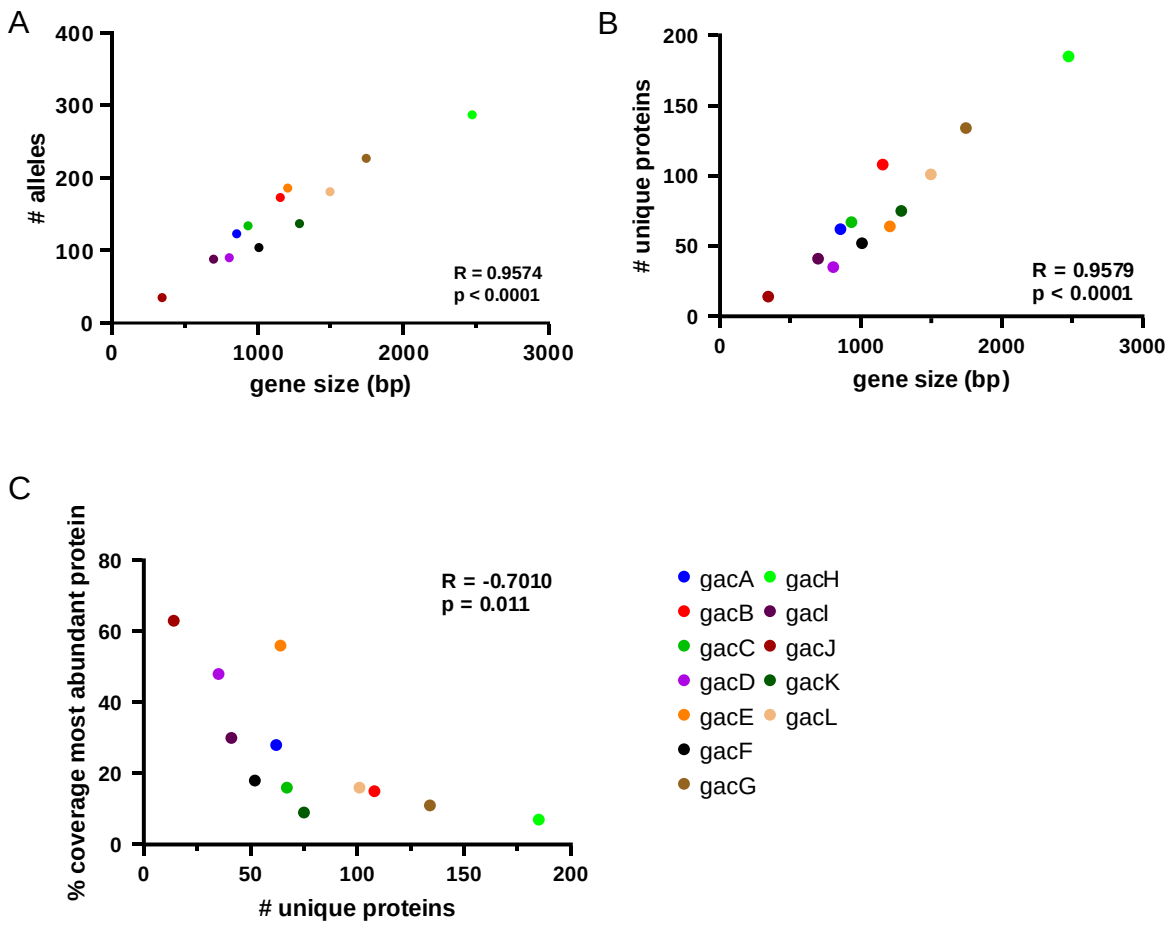

Figure 2 Unique allelic and protein variants of *gacA-L* in 2,044 *S. pyogenes* isolates

The number of (A) unique alleles in and (B) unique protein sequences encoded by each of the *gacA-L* genes and correlation with gene size. (C) The percentage of isolates that contains the most abundant protein sequence for each *gac* gene, versus the amount of unique protein sequences per gene in the *gac* gene cluster.

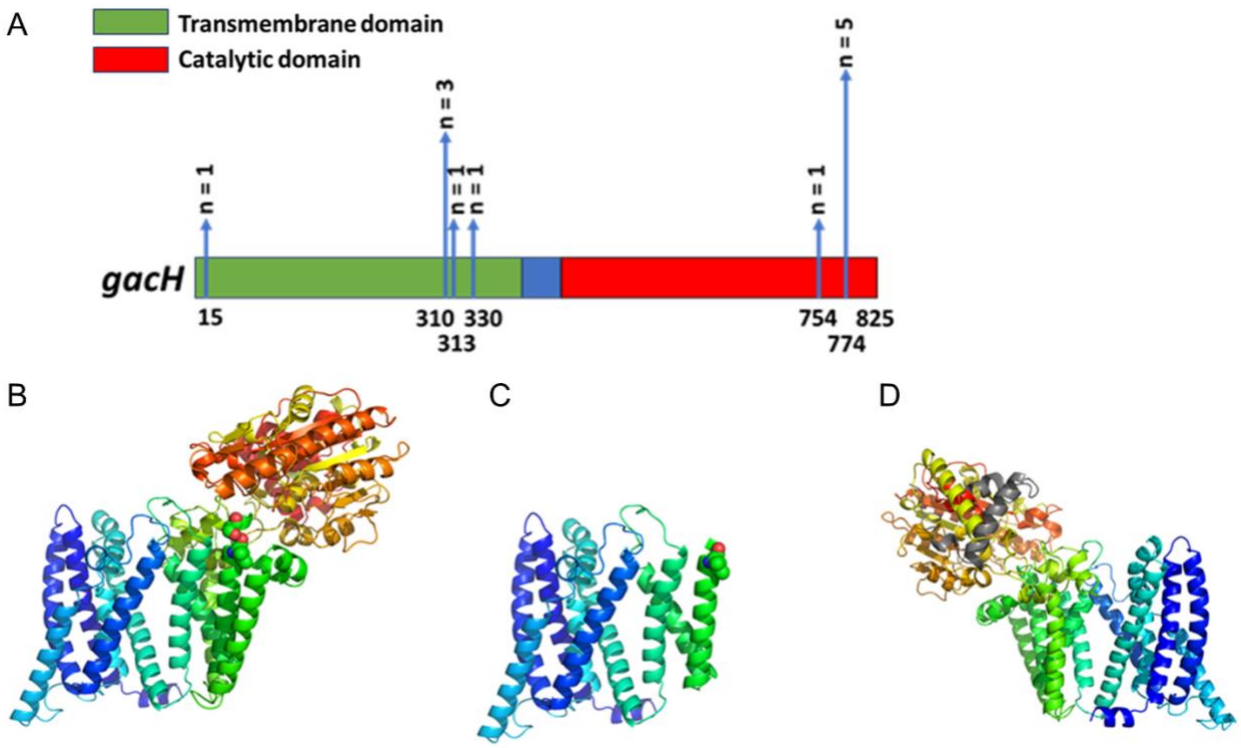

Figure 3 *In silico* analysis of premature stop codons in *gacH*

(A) Scaled 2D representation of *gacH* (2,475 bp, 825 amino acids), containing a transmembrane domain (amino acids 1-395; green) and a catalytic domain (amino acids 444-822; red). Premature stop codons were identified in 12 isolates and are indicated by the vertical blue arrows that show the amino acid position and frequency. (B) Full-length GacH structure, amino acids on position 310 and 313 visualized as spheres. Transmembrane domain (amino acids 1-395) depicted as blue/green, catalytic domain (amino acids 444-822) depicted in yellow/red. (C) Structure of the partial transmembrane domain that remains with premature stop codons at amino acid positions 310 or 313. (D) Disruption of the catalytic domain with premature stop codon at amino acid position 774 depicted in grey. Rest of the remaining structure colored as in (A).

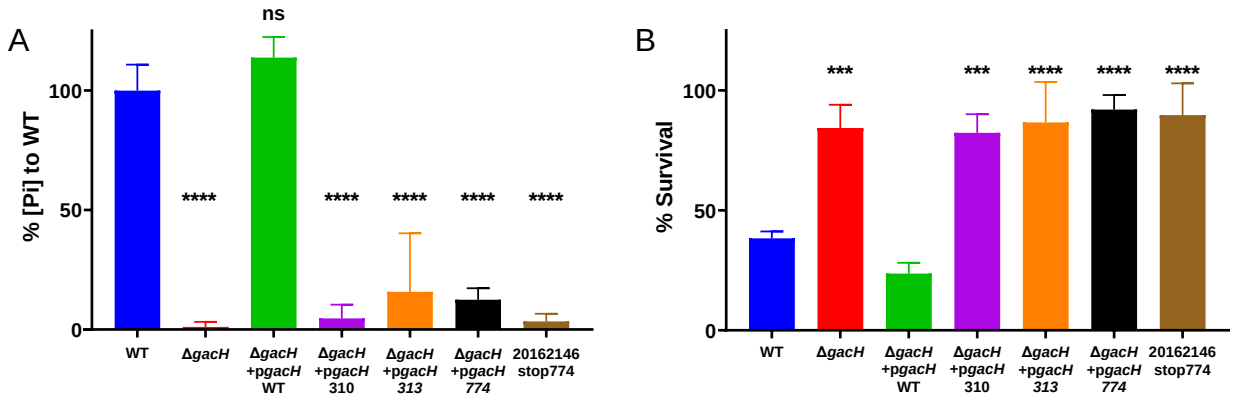

Figure 4 Biochemical and functional analysis of *gacH* variants with premature stop codon in *S. pyogenes*.

(A) Analysis of phosphate content in GAC isolated from WT 5448 *S. pyogenes*, 5448Δ*gacH*, and 5448Δ*gacH* complemented with plasmid-expressed WT *gacH*, premature stop codon *gacH* (on positions 310, 313 or 774) or WT *S. pyogenes* that naturally acquired a premature stop codon in *gacH* (strain 20162146). The concentrations of phosphate are relative to *S. pyogenes* WT 5448. Bars and error bars represent the mean relative phosphate concentrations measured in three biological replicates and the standard deviation, respectively. *P* values were calculated by one-way analysis of variance (ANOVA) comparing all strains to *S. pyogenes* WT 5448; * $P<0.05$, ** $P<0.01$. (B) Survival of all strains mentioned in (A) after exposure to 0.5 µg/mL recombinant hGIIA. Bars and error bars represent the mean percentage survival and standard deviation, respectively (n=3 biological replicates). *P* values were calculated by one-way analysis of variance (ANOVA) comparing all strains to WT 5448 *S. pyogenes*; ns non-significant, **** $P<0.0001$.

Table 1. Overview of genetic variation of *gacA-L* in the 2,044 *S. pyogenes* collection. Shaded rows indicate genes implicated in polyrhamnose biosynthesis, white rows indicate genes critical for formation of the GlcNAc-GroP side chain.

Gene	Gene size (bp)	Alleles (n)	Unique proteins (n)	Unique variants with premature stopcodon (n)	Isolates with truncated proteins (n) (% of total)
gacA	855	123	62	1	1 (0.05)
gacB	1155	173	108	1	1 (0.05)
gacC	933	134	67	1	1 (0.05)
gacD	804	90	35	1	1 (0.05)
gacE	1206	186	64	3	3 (0.15)
gacF	1008	104	52	1	1 (0.05)
gacG	1746	227	134	2	2 (0.1)
gacH	2475	287	185	6	12 (0.6)
gacI	696	88	41	3	5 (0.2)
gacJ	342	35	14	0	0 (0)
gacK	1287	137	75	1	1 (0.05)
gacL	1497	181	101	3	3 (0.2)

References

1. Sims Sanyahumbi A, Colquhoun S, Wyber R, Carapetis JR. Global Disease Burden of Group A Streptococcus. In: Ferretti JJ, Stevens DL, Fischetti VA, editors. *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations. Oklahoma City (OK), 2016.

2. Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis. 2005;5(11):685-94.

3. Walker MJ, Barnett TC, McArthur JD, Cole JN, Gillen CM, Henningham A, et al. Disease manifestations and pathogenic mechanisms of Group A Streptococcus. Clin Microbiol Rev. 2014;27(2):264-301.

4. Lancefield RC. The Antigenic Complex of *Streptococcus Haemolyticus*: II. Chemical and Immunological Properties of the Protein. J Exp Med. 1928;47(3):469-80.

5. McCarty M. The lysis of group A hemolytic streptococci by extracellular enzymes of *Streptomyces albus*. II. Nature of the cellular substrate attacked by the lytic enzymes. J Exp Med. 1952;96(5):569-80.

6. Kreis UC, Varma V, Pinto BM. Application of two-dimensional NMR spectroscopy and molecular dynamics simulations to the conformational analysis of oligosaccharides corresponding to the cell-wall polysaccharide of Streptococcus group A. International Journal of Biological Macromolecules. 1995:117-30.

7. Edgar RJ, van Hensbergen VP, Ruda A, Turner AG, Deng P, Le Breton Y, et al. Discovery of glycerol phosphate modification on streptococcal rhamnose polysaccharides. Nat Chem Biol. 2019;15(5):463-71.

8. van der Beek SL, Le Breton Y, Ferenbach AT, Chapman RN, van Aalten DM, Navratilova I, et al. GacA is essential for Group A Streptococcus and defines a new class of monomeric dTDP-4-dehydrorhamnose reductases (RmlD). Mol Microbiol. 2015;98(5):946-62.

9. van der Beek SL, Zorzoli A, Canak E, Chapman RN, Lucas K, Meyer BH, et al. Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification. Mol Microbiol. 2019;111(4):951-64.

10. Gao NJ, Rodas Lima E, Nizet V. Immunobiology of the Classical Lancefield Group A Streptococcal Carbohydrate Antigen. 2021;89(12):e0029221.

11. van Sorge NM, Cole JN, Kuipers K, Henningham A, Aziz RK, Kasirer-Friede A, et al. The classical lancefield antigen of Group A Streptococcus is a virulence determinant with implications for vaccine design. Cell Host Microbe. 2014;15(6):729-40.

12. Rush JS, Parajuli P, Ruda A, Li J, Pohane AA, Zamakhaeva S, et al. PplD is a de-N-acetylase of the cell wall linkage unit of streptococcal rhamnopolysaccharides. Nat Commun. 2022;13(1):590.

13. Davies MR, McIntyre L, Mutreja A, Lacey JA, Lees JA, Towers RJ, et al. Atlas of Group A streptococcal vaccine candidates compiled using large-scale comparative genomics. Nat Genet. 2019;51(6):1035-43.

14. Henningham A, Davies MR, Uchiyama S, van Sorge NM, Lund S, Chen KT, et al. Virulence Role of the GlcNAc Side Chain of the Lancefield Cell Wall Carbohydrate Antigen in Non-M1-Serotype Group A Streptococcus. mBio. 2018;9(1):e02294-17

15. Zorzoli A, Meyer BH, Adair E, Torgov VI, Veselovsky VV, Danilov LL, et al. Group A, B, C, and G Streptococcus Lancefield antigen biosynthesis is initiated by a conserved alpha-d-GlcNAc-beta-1,4-l-rhamnosyltransferase. J Biol Chem. 2019;294(42):15237-56.

16. Wilson AT. Loss of Group Carbohydrate During Mouse Passages of a Group A Hemolytic Streptococcus. J Exp Med. 1945;81(6):593-6.

17. McCarty M, Lancefield RC. Variation in the group-specific carbohydrate of group A streptococci. I. Immunochemical studies on the carbohydrates of variant strains. J Exp Med. 1955;102(1):11-28.

18. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.

19. Sumby P. Porcella SF, Madrigal AG, Barbian KD, Virtaneva K, Ricklefs SM, 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. 2005;192(5):771-82.

20. Bui NK, Eberhardt A, Vollmer D, Kern T, Bougault C, Tomasz A, et al. Isolation and analysis of cell wall components from *Streptococcus pneumoniae*. Anal Biochem. 2012;421(2):657-66.

21. van Hensbergen VP, Movert E, de Maat V, Luchtenborg C, Le Breton Y, Lambeau G, et al. Streptococcal Lancefield polysaccharides are critical cell wall determinants for human Group IIA secreted phospholipase A2 to exert its bactericidal effects. PLoS Pathog. 2018;14(10):e1007348.

22. Nei M, Gojobori T. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. Mol Biol Evol. 1986;3(5):418-426.

23. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021;38(7):3022-7.

24. Djoko KY, Ong CL, Walker MJ, McEwan AG. The Role of Copper and Zinc Toxicity in Innate Immune Defense against Bacterial Pathogens. J Biol Chem. 2015;290(31):18954-61.

25. McCarty M. Variation in the group-specific carbohydrate of group A streptococci. II. Studies on the chemical basis for serological specificity of the carbohydrates. J Exp Med. 1956;104(5):629-43.

26. Rush JS, Edgar RJ, Deng P, Chen J, Zhu H, van Sorge NM, et al. The molecular mechanism of N-acetylglucosamine side-chain attachment to the Lancefield group A carbohydrate in *Streptococcus pyogenes*. J Biol Chem. 2017;292(47):19441-57.

27. Parker MT, Bassett DC, Maxted WR, Arneaud JD. Acute glomerulonephritis in Trinidad: serological typing of group A streptococci. J Hyg (London). 1968;66(4):657–675.

28. Fox EN. M Proteins of Group A Streptococci. Bacteriol Rev. 1974;38(1):57-86.

29. Chochua S, Metcalf BJ, Li Z, Rivers J, Mathis S, Jackson D, et al. Population and Whole Genome Sequence Based Characterization of Invasive Group A Streptococci Recovered in the United States during 2015. mBio. 2017;8(5).

Supplemental information

Table S1. Bacterial strains used in this study

Strain	Source	Information
S. pyogenes (M3)	(14)	
S. pyogenes (M3) Δ*gacI*	(14)	
S. pyogenes D315 (M58) (NCTC10876; R67/3884)	(27, 28). Obtained from UK Health Security Agent.	Wild type
S. pyogenes D315/87/3	Rockefeller.edu/vaf/strainsDB/index.php. Provided by Dr. D. Bessen, New York Medical College	'A variant', passaged in mice
E. coli MC1061 pDCerm_*gacH*	(7)	
S. pyogenes 5448	(7)	
S. pyogenes 5448 Δ*gacH*	(7)	
S. pyogenes 5448 Δ*gacH* + pDCerm_*gacH*	(7)	*gacH* from 5448
S. pyogenes 5448 Δ*gacH* + pDCerm_*gacH*_STOP310	This study	
S. pyogenes 5448 Δ*gacH* + pDCerm_*gacH*_STOP313	This study	
S. pyogenes 5448 Δ*gacH* + pDCerm_*gacH*_STOP774	This study	*gacH* from 20162146
S. pyogenes 20162146	CDC, (29)	Wildtype but *gacH* stop codon at position 774

Table S2. Primers used in this study

Primer	Sequence (5' to 3')
STOP310AAF	CCGGAATTCATGATTAAAGATACTTTCCTTAAGACTAAC
STOP310AAR	GAAGATCTTCAGCGTGAGATTTTAAAGAAGGTCTTGTG
gacH310_checkF1	CTTAGTTGCCTATTTCGCAATCG
gacH310_checkR1	CTCTCAGACGGATTGATCAATG
STOP313AAF	CCGGAATTCATGATTAAGGACACATTTTTAAAAACCAATTG
STOP313AAR	GAAGATCTTCAGCGTGAGATTTTAAAGAAGGTCTTGTG
STOP313checkF	GTGGCCTACTTTGCTATTG
STOP313checkR	GATCACGCTAAAGAATTGAG
gacHEcoRIF	CCGGAATTCATGATTAAAGACACATTTTTAAAAACCAAT
gacHBglIIR	GAAGATCTTTAACGTGATATCTTAAAAAAAGTTTTGTGT
gacHcheck1	CTGAATTAGTAACAGCAGGTAATAG
gacHcheck2	GCTAGGTACAAATCTTTAGC
gacHcheck3	GGTAGCTTGGTATCTACATG
pDCermF	GTACCGTTACTTATGAGCAAG
pDCermR	CACGGCATAAATCGCTCAG