Unique Genomic Arrangements in an Invasive Serotype M23 strain of Streptococcus pyogenes

Identify Genes that Induce Hypervirulence

Yunjuan Bao, Zhong Liang, Claire Booyjzsen, Jeffrey A. Mayfield, Yang Li, Shaun W. Lee, Victoria A. Ploplis, Hui Song, Francis J. Castellino

* Co-first authors, contributed equally to the work.

#To whom correspondence should be addressed: Francis J. Castellino, W.M. Keck Center for Transgene Research, University of Notre Dame, Notre Dame, IN 46556, USA. Telephone: 574.631.9152; telefax: 574.631.8017; email: fcastell@nd.edu.

Running title: emm23 GAS Genome

Key words: emm 23 genome; bacterial pathogenicity; host-pathogen interactions; bacterial evolution; bacteriophages; psuedogenes; genomic arrangement

JB Accepts, published online ahead of print on 15 September 2014
Copyright © 2014, American Society for Microbiology. All Rights Reserved.
The first genome sequence of a Group A *Streptococcus pyogenes* M23 (*emm23*) serotype 24 (M23ND), isolated from an invasive human infection, has been completed. This opacity factor-negative (SOF-) strain is composed of a circular chromosome of 1,846,477 bp. Gene profiling showed that this strain contained six phage-encoded and 24 chromosomally-inherited well-known virulence factors, as well as 11 pseudogenes. The bacterium has acquired four large prophage elements, ΦM23ND.1-4, harboring genes encoding streptococcal superantigen (*ssa*), streptococcal pyrogenic exotoxins (*spe*) C, H, and I, and DNases *spd1* and *spd3*, with phage integrase genes present at one flank of each phage insertion, suggesting that the phages were integrated by horizontal gene transfer. Comparative analyses revealed unique large-scale genomic rearrangements that differ from previously sequenced GAS strains. These rearrangements resulted in an imbalanced genomic architecture and translocations of chromosome-encoded virulence genes. The *covS* sensor in M23ND was identified as a pseudogene, resulting in attenuation of *speB* function and increased expression of the chromosomal virulence factors, *mga*, *emm23*, *scpA*, fibronectin-binding proteins (*sfb1* and *fbp54*), streptolysin O (*slo*), hyaluronic acid capsule (*hasA*), streptokinase (*ska*), and a 2 DNases (*spd* and *spd3*), which were verified by PCR. These genes are responsible for facilitating host epithelial cell binding and/or immune evasion, thus further contributing to the virulence of M23ND. In conclusion, strain M23ND has become highly pathogenic as the result of a combination of multiple genetic factors, particularly gene composition and mutations, prophage integrations, unique genomic rearrangements, and regulated expression of critical virulence factors.
INTRODUCTION

Streptococcus pyogenes (Group A streptococcus; GAS) is a pathogenic low G+C content β-hemolytic Gram+ bacterium (1). GAS is responsible for ~700 million infections worldwide per year (2), ranging from simple pharyngitis and impetigo to more invasive life-threatening infections that include necrotizing fasciitis and streptococcal toxic shock syndrome (3). Acute rheumatic fever (ARF) and acute glomerulonephritis are among the more serious non-suppurative sequelae that can result from GAS infection (4).

Currently >200 strains of S. pyogenes have been identified from emm genotyping (5). The genomes of 20 strains that span a range of serotypes have been completely sequenced and assembled. Of the many known virulence factors of this bacterial strain, e.g., C5a peptidasw (speB), the hyaluronic acid (hya) capsule, streptolysin S (sls), and streptolysin O (slo), the M-protein that is encoded by the emm gene is one of the most important features of this group of bacteria (6). The M-protein is composed of multiple N-terminal A and B modules, which are highly variable among M-types, along with well conserved C-terminal C and D modules. The N-terminal A region is the most variable of these domains (7), thus rendering it suitable for distinct serotyping. In this regard, GAS strains are accordingly serologically classified as different M-types based on the first ~50 amino acid residues of this hypervariable N-terminus (8).

In addition to the emm gene, up to two additional subfamily emm-like genes, fcR and enn, that encode Fc binding regions of IgG and IgA, respectively, have been found in a cluster in different GAS strains. This emm subfamily of genes can be identified through the nucleotide sequences of their 3'-peptidoglycan-spanning domains (9, 10). Not all of these genes are present in every GAS strain and their presence and chromosomal arrangement have been used
to further map GAS as chromosomal pattern A-E, with the intent of correlating these genomic patterns with tissue tropicity and virulence (11). Patterns A-C are associated with pharyngeal disease, pattern D with skin disease, and pattern E with both (12). In addition, two additional classes of GAS have been designated. Class I is a serum opacity factor (sof/sfbII-)negative (SOF-) strain that ordinarily contains the novel fbpB genotype of a fibronectin binding protein (FBP), prfF2 (13-15). Additionally, some of these strains encode the FBP sfbl gene, which is an important determinant for epithelial cell (EpC) binding and invasion (16). These strains are also linked to a surface-exposed antigen in the C repeat region of the M-protein, which interacts with ARF antibodies (17). Class II sof/sfbII-positive (SOF+) GAS strains normally contain the ρfb genotype of prfF2 (6, 14). Further, many of these strains also encode the FBP genes, sfbl/prfF1 (18) and sfbX, the latter of which is located immediately downstream of the sof/shfII bicistronic tandem (19). SOF+ strains generally do not immunoreact with ARF antibodies to the class I C-repeat surface of the M-protein (17). Many Class I and Class 2 strains contain a gene for another FBP, fbpA, which is regulated by the one component multiple gene activator (Mga) and exists as a member of its cis regulon (14). Most SOF- and SOF+ strains display FBP54 (fbp54), which is also a fibronectin/fibrinogen binding protein (20, 21).

The M-protein is most prominently regulated by the first component (Mga) of the mga regulon, formerly known as vir or mry (22, 23). Mga is maximally expressed during the logarithmic growth phase (LP) in response to changing environmental conditions, e.g., temperature, pH, CO2 levels, and/or the iron concentration (24). The largest mga cis-regulon is present in pattern D strains (15) and consists of the following sequential order of genes: mga-feR-emm-enn-scpA (C5a peptidase)-fbpA. Expression of all of these genes is controlled by Mga (15, 25, 26). Mga also regulates in-trans a number of other GAS proteins, e.g., streptococcal
inhibitor of complement (sic), and the bicistronic FBPs, sof-sbfX (19, 27). Mga expression is 
in-turn regulated by itself (Mga), as well as by the transcriptional regulator genes, rgg (ropB) 
and rofA (nra) (23).

Thus, GAS produces numerous proteins and regulatory systems that are strain-specific 
and help the bacteria circumvent the host innate immune system. Of the known virulence-
enhancing extracellular proteins produced by GAS, many are capable of triggering a severe 
non-specific host immune response (28). Because of the large number of strains, and the broad 
variation in their properties and characteristics, a large variance in virulence is displayed 
phenotypically as differing levels of severity and invasiveness. Therefore, sequencing full 
genomes of GAS, particularly those harboring different M-types, is integral to understanding 
the nature of this organism. This in-turn, is critical to developing effective strategies for the 
host against this bacterium. Most importantly, such knowledge will allow us to understand 
GAS evolution and development, which will allow the assessment and prediction of important 
patterns in GAS phylogeny. Toward this end, we have sequenced the full genome of S. 
pyogenes strain M23ND, the first M23 strain to be reported in this manner. This circular 
genome possesses ~1.85 Mbp and appears to have a high rate of genetic recombination. As a 
unique isolate with many unusual properties, M23ND offers an excellent opportunity to 
examine some of the more variable and subtle characteristics of GAS that are associated with 
severe GAS infections.

Materials and Methods

GAS strain. S. pyogenes strain, ATCC 21059, is a serotype M23 GAS. This bacterium was 
isolated as strain Sv in 1965 from a patient with a case of severe streptococcal disease (29). We
will refer to this fully sequenced strain as M23ND. The genomic sequence of M23ND has been deposited at NCBI GenBank under accession number CP008695. The sequence and strain information can be obtained at the Castellino-Keck Center GAS Database (http://www3.nd.edu/~transgen/GASdatabase).

Strain handling. The GAS isolate was cultivated from glycerol stock cultures on blood agar at 37°C and 5% CO₂ for 24 hr. Genomic DNA (gDNA) was extracted using a Mini-DNA kit (Qiagen, Valencia, CA).

Genome sequencing and gene annotation. The entire genome of M23ND was sequenced using Illumia Miseq (Illumina, CA), with read lengths of 150 bp x2 and 454-pyrosequencing (Roche 454 Life Science, Basel, CH). A high quality draft genome assembly was obtained which comprised eight scaffolds. The gaps were closed using PCR primer walking and the complete circular genome was derived. The protein coding sequences were predicted using Glimmer3.02b (30), the rRNA sequences were predicted using RNAmmer (31), and tRNA sequences were detected by tRNAscan-SE (31). Genome annotation was performed using the automated RAST annotation server (32) and manual curation.

Comparative analysis of GAS strains. Genome sequences for the 20 fully-sequenced GAS strains currently available were downloaded from the NCBI Genome Database (33). The genome sequence map of *S. pyogenes* M23ND, in comparison to the other 20 available whole genome GAS sequences, was generated using BRIG (34). The comparative genome architecture of the GAS genomes was determined by BLASTn and graphically represented using the Artemis Comparison Tool (ACT) (35). Visualization of phage locations was implemented using Geneious 7.0.6 (Biomatters; Auckland, NZ).
Phylogenetic analyses. Phylogenetic analyses were conducted based on multi-locus sequence typing (MLST) (36) of seven housekeeping marker genes, viz., glucose kinase (gki); glutamine transporter protein (gtr); glutamate racemase (murI); DNA mismatch repair protein (mutS); transketolase (recP); xanthine phosphoribosyltransferase (xpt); and acetyl CoA acetyltransferase (yqiL), as well as SNP detection of virulence genes. The construction and visualization of phylogenetic trees or networks were implemented in SplitsTree (37).

Quantitative reverse transcription polymerase chain reaction (qRT-PCR). Total gDNA was isolated from a selected single colony (from 10) of M23ND, after checking growth rates and the presence of some known genes. The sample was cultured overnight at 37° C/5% CO₂ in Todd-Hewitt broth (BD Biosciences, San Jose, CA) supplemented with 10% yeast extract (THY). The gDNA was isolated after treatment of the cells with lysozyme/proteinase K and cell lysis buffer (100 mM Tris/5 mM EDTA/0.2% SDS/200 mM NaCl, pH 8.5) and extracted with phenol/chloroform/isoamyl alcohol (25/24/1, v/v/v). The gDNA was precipitated with isopropanol and washed with 70% ethanol. Approximately 100 ng of gDNA was used in a 30 μl PCR for gene validation.

Total RNA was isolated from single colonies grown to logarithmic phase (LP; A₆₀₀nm = 0.5-0.6) or stationary phase (SP; A₆₀₀nm > ~1.0), at 37° C/5% CO₂ in THY media as described earlier. Two treatments with DNase I were performed to eliminate any DNA contamination (15). Approximately 50 ng total RNA was used in 30 μl RT-PCR for the gene transcription validation. A sample without reverse transcriptase added as the negative control. Primers (Table S1) were specifically designed in the CDs for a PCR amplicon of ~200-500 bp for both gDNA detection and mRNA transcription. Relative gene expression levels were analyzed by the 2^{ΔΔCT} method (15). Triplicate measurements of the threshold cycle number, C_T, were
collected for each chosen gene relative to the housekeeping gene, *gapdh*, for both *covS* and *covS*† strains grown to LP and SP. The expression ratio was calculated relative to the *covS* strain grown to LP.

### RESULTS

#### General background of GAS strain M23ND.
GAS strain M23ND was originally isolated from a case of severe streptococcal infection for its cytotoxic anti-cancer properties by virtue of cytolytic SLS production (U.S.P 3,477,914 and 4,328,218). At that time it was taxonomically classified as *Streptococcus hemolyticus* and designated as strain Sv (ATCC 21059) (29, 38, 39).

Subsequently, strain Sv was serologically identified as having a type 23 M-protein, which was confirmed by comparison of its amino acid sequence with that of the GenBank deposited isolate, M23-Memphis (accession number U11953). Additionally, the mga regulon of the Sv strain was classified as the small variety. The 4791 bp incomplete sequence nonetheless allowed identification of the gene sequence: 5'-*mga-emm-scpA*-3', thus representing a Pattern A strain (12).

#### Overall features of the genome sequence of GAS strain M23ND.
The genome of *S. pyogenes* M23ND consists of a single circular chromosome of 1,846,477 bp, with an average G+C content of 38.6% (**Fig. 1**). The genome encodes 1851 predicted open reading frames (ORFs) and 11 pseudogenes which in-total cover 86.3% of the entire genome sequence. This M23 strain contains 44 unique and unidentified ORFs based on a minimum 85% homology by BLASTn comparison. Approximately 12% of the ORFs are contained in four prophage segments which account for 8.5% of the genome. The ORFs within the prophage regions
encode multiple known or putative virulence factors that include toxins, superantigens,mitogenic factors, and fibronectin-binding proteins.

Comparative study of GAS strain M23ND and characterization of phage elements contributing to its genetic diversity. Comparative genome mapping (Fig. 2) illustrates the genomic profile of M23ND in comparison with the 20 other previously fully-sequenced GAS strains, the properties of which are summarized in Table 1. This comparison showed that the sequences are conserved throughout the genome, except at the points of insertion of short mobile genetic elements and at four large prophage regions where the genome sequences exhibit mosaic profiles across the compared GAS strains. The short mobile elements of M23ND have orthologs in only some of the other strains. However, the average homologies for aligned prophage sequences (91-94%) between different strains are lower than that for non-prophage sequences (>98%). It generally appears that mobile genetic elements, including prophages, are major sources of genetic diversity among different GAS strains.

In order to more fully assess the contributions of M23ND prophage elements to genetic diversity, we identified the locations and lengths of the prophage segments based on gene annotations and BLAST comparisons (Table 2 and Fig. 3A). There are four large prophage-related elements, designated ΦM23ND.1–ΦM23ND.4, with individual lengths ranging from 34-42 Kbp. Three of the four prophages are distributed within the clockwise half-circle proximal to the origin of replication site (ori) of the genome. Comparisons with other S. pyogenes sequenced strains show that prophages are generally inserted in the second replicihore. However, for M23ND, three of the four prophages are inserted in the first replicihore, thus highlighting the imbalanced architecture of the M23ND strain. Four phage integrase genes are situated on the terminal ends of each of these prophages and are located at 168,633 - 170,048
bp; 409,285 - 410,331 bp; 572,756 - 573,884 bp; and 875,643 - 876,785 bp. Only the latter prophage encodes most genes in the forward direction (Fig. 1).

In addition, the mismatch repair (MRR) genes coding for mutS (1,789,917-1,792,472 bp)-mutL 1,792,601-1,794,583), present on a polycistronic gene cluster and controlled by a single promoter, are uninterrupted in M23ND, such that both genes are expressed throughout its growth cycle. This assists in maintenance of the genetic material of this strain and limits mutability of the already virulent and established M23ND. This is unlike the less virulent SF370 strain, which contains a prophage (ΦSF370.4) insertion between mutS and mutL that inactivates mutL in a growth-dependent manner, and allows for early growth phase mutations to occur, perhaps randomly generating more virulent strains (40).

The prophages in M23ND have different genomic integration locations compared to their orthologs in other strains (Fig. 3A). For example, prophage ΦM23ND.1 is very similar to ΦManfredo.4 (M5), ΦMGAS10394.3 (M6), ΦMGAS1882.2 (M18), and those in several other strains, but this prophage is inserted at different genomic locations in these strains. Prophage ΦM23ND.3 shares high homology with ΦM23ND.1, but is integrated at two distinct sites. This implies that these two prophages arose from a common ancestor and likely underwent horizontal gene transfer within or between strains. These intra- and inter-strain recombination events suggest the important role of horizontal gene transfer in shaping the genetic diversity of S. pyogenes evolution (41, 42).

An immune defense system, clustered regulatory interspaced short palindromic repeats (CRISPR), has recently been identified in many bacterial strains, including S. pyogenes, as a protective mechanism against exogenous prophages or plasmid DNAs (43-45) via inhibition of the acquisition of phage elements (46). Considering the prevalence of phages in the genome of
M23ND, it was of interest to determine whether CRISPR was involved in the protection system of this strain. Remarkably, we did not identify CRISPR sequences and/or the CRISPR-associated genes (cas) in M23ND. Further examination of genome sequences of the other 20 sequenced GAS strains showed that several other strains from different serotypes do not contain the CRISPR sequences, including M5 Manfredo, M6 10394, M18 8232, M53 Alab49, M3 SSI-1, M3 315, and M14 HSC5. These strains also do not contain the cas genes, except the two serotype M3 strains, M3 SSI-1 and M3 315. The absence of the CRISPR system in GAS strains represented the clustering of the phylogenetically-related strains (see additional details below), i.e., M23ND was closest to M5 Manfredo and M6 10394 and shared with them a common ancestor, M18 8232; M53 was closest to the two M3 strains (M3 SSI-1 and M3 315) and they are located in the same evolutionary branch; M14 HSC5 was itself in a separate branch. This indicated that CRISPR may represent a type of evolutionary product inherited via selective pressure. The absence of CRISPR in some GAS strains, and their specific amounts of phage elements, could be a balanced result of the advantages of enhanced virulence and disadvantages of excess foreign toxins from phage acquisition.

Sources of genomic diversity in *S. pyogenes* M23ND: extensive unique genomic arrangements and imbalanced genome architecture. Further inspection of the comparative genomic architectures between the fully-sequenced GAS strains (Fig. S1) showed that strain M23ND exhibited unique genomic rearrangements, compared with M5 Manfredo, M12 HKU16, and M3 SSI-1, which contain similar central inversions across the replication origin (ori) and *dif*-like replication terminus (ter) axis (47). M23ND differs from these three strains in that the *dif*-like replication terminus site (ter) was found to be located within the half-circle of the genome (at 702,433 bp) and the inversions and translocations are asymmetrical across the
ori/ter replication axes. Additionally, M23ND exhibits gene translocations over and above those found in M3 SSI-1, M5 Manfredo, and M12 HKU16. For example, the critical gene, sagA that encodes streptolysin S (SLS) that is responsible for the β-hemolytic characteristic of GAS, is found in flanking locations of the ter site on the various fully sequenced genomes, due to inversion around ori/ter (Fig. 3B). It is symmetrical for all genomes, except for M23ND, due to the nonsymmetrical ter with respect to ori. A similar situation exists for sen, a gene essential for GAS survival. For srv, a transcriptional regulator of virulence, it would appear that inversions around a symmetrical ori-ter placed this gene on the opposite replichore for M3 SSI-1, M5 Manfredo, and M12 HKU16, and then another inversion in the latter three genomes around ter of M23ND uniquely placed srv in M23ND. These examples highlight the unique extent of gene translocations in strain M23ND. Of particular note, this genome also contains an extra short inversion within the last 100 kb of the chromosome (Fig. S1). This feature has not been seen previously in any other sequenced GAS genome. These data suggest that M23ND may share a common ancestor with M3 SSI-1, M5 Manfredo, and M12 HKU16, but experienced distinct rearrangement events and, thus, underwent a disparate evolutionary path.

By comparing the genome sequences of M23ND with M5 Manfredo and the representative strain, M18 8232, that does not exhibit significant rearrangements, we found that the prophage segments are located at the breakpoints, or within the rearranged genomic regions (Fig. 4). This characteristic has also been reported previously for another GAS strain, M3 SSI-1 (48). It was proposed that the genomic rearrangements were triggered by prophage recombinations to balance global genome architecture (48). However, in M23ND, the clustering of all four prophages and the ter site within a single half circle of the M23ND genome has disrupted the balance of global genomic architecture. The replichores in the
clockwise and counterclockwise directions are unequal in length. This may point to a positive correlation between an unbalanced replichore architecture and the severity of invasive infection induced by M23ND. Previous research using *E. coli* found that such an imbalance can affect bacterial growth and fitness (49, 50). However, it is difficult to establish a firm relationship between this genotype and clinical phenotypes displayed in GAS, due to limited examples of imbalanced replichores in the currently available fully sequenced genomes.

### Profiling of the prophage encoded virulence genes of *S. pyogenes*.

Virulence factors in GAS genomes are major contributors to the pathogenesis of *S. pyogenes*. A variety of these factors have been identified in previous GAS studies as phage-encoded or chromosomally inherited. We investigated a total of 38 virulence factors of interest to our studies, of which 30 were found to be present in M23ND (*Table S2*). We profiled all of these virulence genes in M23ND and compared them with those in other GAS strains to assess their patterns in genomic distributions.

M23ND contains six well-established virulence genes encoded by prophage elements. Specifically, genes for *ssa* and exotoxin type I/H (*speI/speH*) were encoded by ΦM23ND.2 and ΦM23ND.4, respectively, while exotoxin C (*speC*) was encoded by ΦM23ND.1. Two mitogenic factors, *spd1* and *spd3*, were carried by two related prophages, ΦM23ND.1 and ΦM23ND.3, respectively. The acquisition of such toxins and pyogenic genes has been reported to correlate with severe invasive infections and epidemic outbreaks of *S. pyogenes* (51). Very likely, the carriage of the combination of several virulence factors (*ssa*, *speC*, *speI*, and *speH*) and endonucleases (*spd1* and *spd3*) in strain M23ND is a major factor responsible for the virulence of this strain (52). M23ND is one of few fully-sequenced strains to carry *ssa*, in addition to strains of the M3, M4, and M12 lineages.
Despite the properties of the phage-encoded virulence factors, their spatial distribution across species is discordant with that of the carrier prophages (Table S2 and Fig. 3B). Generally, each virulence factor could be encoded by divergent prophage elements integrated at diverse sites across different GAS strains. In order to extend this finding into other phage-encoded virulence factors, viz., sda, speA, speK, slaA, speL, and speM, we mapped the locations of all of the established phage-encoded virulence genes across the 21 fully-sequenced GAS strains (Fig. 5). It was found that the virulence genes are scattered nearly randomly throughout the chromosomes, although strains of the same M-type tend to cluster these genes in similar areas of the chromosome. However, there are currently too few fully-sequenced and assembled S. pyogenes genomes to make any definitive conclusions regarding this point. These phenomena reflect the complex recombinative evolution of prophage and phage-encoded virulence genes in GAS strains and underscore their dominant contributions to genetic diversity.

**Profiling the bacterial chromosomally encoded virulence genes of S. pyogenes.** The genome of strain M23ND also encodes 24 chromosomally-inherited established virulence genes (Table S2). In contrast to the phage-encoded virulence genes, the chromosome-encoded genes are present in nearly all of the sequenced GAS strains, with the exception of the absence of speG in M4 10750; smeZ in M2 10270, M49 NZ131; hasA in M4 10750; slaA in M18 8232; and endoS in M49 NZ131. The major deviations in this regard are sfb1, sic, and speJ, which are encoded by 10/21, 4/21 and 7/21 sequenced GAS strains, respectively. Sfb1 is a novel virulence gene product in M23ND with, at most, 66% nucleotide homology across orthologs in other strains. Specifically, while this gene is present in almost half of the fully-sequenced GAS strains, 34%-85% of the sequence of sfb1 in M23ND is unique to this strain. This gene is critically involved in streptococcal adherence to host epithelial cells (EpC) via fibronectin. A
BLASTp analysis of the protein sequence of Sfb1-M23ND shows that it contains a fibronectin
binding protein signal sequence and five fibronectin binding repeat domains positioned between
amino acids 410-583 near the COOH-terminus. Sfb1 in M23ND is accordingly a major
candidate for promoting bacterial internalization into host EpC, as well as invasion and
infection specificity (53).

We discovered that genomic locations of chromosomally-inherited virulence factors
(Fig. 6) and regulatory genes (Fig. 7) are more conserved than those of phage-encoded genes
across the GAS strains, with the exception of M23ND, M5 Manfredo, M12 HKU16, and M3
SSI-1. The streptococcal pyrogenic exotoxin genes, viz., speG, speJ, and smeZ, along with
other key genes, viz., sagA, slo, sfb1, ska, and sen, and extracellular toxins, viz., spyA, prtS,
ideS, lmp, sclA, hylA, nga, graB, dltA, dltC, cfa, and plr (Fig. 6A,B), are translocated away from
these conserved sites. A similar situation exists with regulatory genes, viz., covRS, mga, rgg,
and srv (Fig. 7), along with the M23ND M-protein gene, emm23, which coexists within the mga
cis-regulon. A close examination of the distribution of these genes revealed that the
translocations correspond to the large genomic rearrangements outlined above (Fig. 4). Large
segmental inversions and translocations carried the genes to their present locations.

Further examination of the sequences of the chromosomally-inherited virulence genes
via multiple sequence alignment showed that the orthologous genes share high similarity
(>95%) between different GAS strains, except for genes such as sfb1 and sic, which are not
present in all GAS strains and the binding-related genes, viz., graB, ska, endoS, and ideS, which
contain long divergent regions between GAS strains. The exceptional genes containing
divergent regions mainly encode binding-related proteins and the divergent regions fall into the
binding domains within the parent protein, e.g., ska. Furthermore, the divergent sequences can
be grouped into a few clusters, independent of the serotypes of the GAS strains. It is likely that the divergence of the virulence genes was induced from horizontal gene exchange and was functionally related with binding specificity in adaptation to particular host challenges. In order to assess the adaptation roles played by the highly conserved virulence genes, we performed single nucleotide polymorphism (SNP) detection for those genes. We determined a frequency of 39 SNPs/kb, which is compatible with the 41 SNPs/kb observed for the seven commonly used housekeeping genes (36). This indicates that the evolutionary drive for enhanced adaptation of GAS strains may not originate from point mutations, but from horizontal gene transfer, a feature that is strikingly similar to phage-encoded genes. This conclusion is also supported by the phylogenetic network constructed from SNP detection of the conserved virulence genes (Fig. S2-A). The network was topologically similar to the background evolutionary structure inferred from the pair-wise whole-genome comparisons (Fig. S2-B) and based on multi-locus sequence typing (MLST) of seven housekeeping genes (Fig. S2-C). M23ND is most closely related to serotype M5 Manfredo and M6 10394, which may also share a common ancestor with M18 8232, having diverged more recently.

**Pseudogenes candidates for hypervirulence of M23ND.** Gene inactivations can play a major role in the growth and invasive properties of GAS and some occur during the course of infection to maintain GAS virulence during different challenges by the host (54). Two pseudogenes of particular relevance have been identified in the MD23 genome. One gene, the superantigen, speH, encoded within prophage ΦM23ND.4, and regulated by bacterial rgg (55), resulted from a one bp frame-shift deletion in the ORF and formed a hypothetically dysfunctional gene. Another critically important pseudogene, the CovS component of the CovRS two component regulatory system, was found to contain a 5-nucleotide deletion in strain
M23ND, resulting in attenuated speB expression and increased virulence and invasiveness in mice (56). Another pseudogene, trx, is essential for inhibiting O2-mediated killing of mycobacterium. Inactivation of this gene can be harmful to bacteria but is unlikely to be essential for GAS elimination, since GAS is killed by O2-independent mechanisms (57).

Other M23ND pseudogenes of lesser-studied relevance are mainly involved with catabolism, biosynthesis, or signaling. These genes are an ammonium transporter (amt), chloride channel protein, glutamine 5-kinase, 2-(5'-triphosphoribosyl)-3'-dephosphocoenzyme A-synthetase, asparaginyl-tRNA-synthetase-related protein, lanthionine biosynthesis protein, a mobile element protein, and a protein of unknown function. While we cannot know whether these genes were inactivated during evolution and/or during infection to allow GAS survival at different points of invasion, they clearly are compatible with survival of GAS at the final invasion stage.

Identification of the genetic properties that contribute to pyogenic invasion and virulence. A previous study reported the association between mutations in the two-component regulatory system, covRS, of GAS with increased virulence (58). These mutations resulted in decreased expression of the cysteine protease, speB, and upregulation of multiple virulence genes (59). To explore the mechanism of severe invasion and virulence of the currently studied strain M23ND, we examined the genomic mutations in covRS and in vitro expression of the related virulence genes. We detected that covS is indeed a pseudogene in M23ND and presents a highly attenuated expression of speB, determined by Western immunoblots (data not shown). Similarly, inactivating mutations in covS, and corresponding attenuated expression of speB, were also reported in other serotype GAS strains, such as the serotype M1 invasive isolate (54, 60), a highly virulent M3 isolate (61), M53 (15, 54), and M81 (62) invasive strains, and in the
current M23ND strain. SpeB plays a complex role in virulence, initially being a factor that enhances invasion of the bacteria in EpC (63) but during later stages this protease can have detrimental effects since it can catalyze degradation of other important virulence factors, such as M-protein and FBP (64). Thus, under optimal conditions, this protease would be upregulated during initial infection phases and downregulated after initial invasion. The CovRS system is capable of accomplishing this task by rapidly generating inactivating mutations in CovRS (54).

In addition to the point mutations, the invasiveness of GAS requires expression of a combination of virulence genes which function at different infectious stages of infection. Invasion is initiated by adherence to EpC mediated by surface-binding proteins, such as FBP, LBP, collagen-binding protein, and fibrinogen-binding protein. After invasion, the GAS strains developed mechanisms to escape host innate immune systems by expressing the DNases, hyaluronic acid capsule synthesis genes (hasABC), streptolysin O (slo), NAD glycohydrolase (nga), and pyrogenic exotoxin B (speB), among others. Invasive GAS strains also contain multicistronic regulatory systems, which, via mutations, can upregulate or downregulate important virulence determinants. The regulation of these genes during the course of infection allows invasive GAS strains to resist killing and persist at the infection sites (65).

Expression of critical virulence genes in M23ND. While the gene composition of M23ND was identified through genome sequencing, we verified the presence of these genes of interest to our work by RT-PCR (data not shown). Next, q-RT-PCR was employed to assess the expression properties of these genes at LP and SP growth in both the clinical isolate covS strain and the isogenic covS+ strain. Expression of the cysteine protease, SpeB, is well known to be regulated by the growth phase and by the CovRS system. As a verification that M23ND functioned similarly to many other GAS cell lines, we find in agreement with previous work on
other strains that the speB mRNA is produced at the highest level in SP growth and with the CovRS system intact (Fig. 8A).

A examples of the data obtained with other genes, we find that genes facilitating host surface binding or initial immune invasion show qualitatively higher mRNA levels at LP, including that of the M-protein, emm23, and the C5a peptidase, scpA (Fig. 8B), as well as the mRNA of a fibronectin-binding protein (sfb1), the pore forming protein, streptolysin O (slo), and the capsule encoding gene, hasA (Fig. 8C). An opposite result is found for the chromosomal gene encoding the DNase, spd (Fig. 8C), which is slightly upregulated at SP (Fig. 8C). This result is logical since chromosomal spd and phage-encoded spd1 and spd3 are the only extracellular DNases present in M23ND, and may be required at a later stage of infection when DNA nets encapsulate the bacteria. While some of these genes are universally present in different serotypes of GAS strains, e.g., mga, emm, scpA, ska, fbp54, plr, and spd, due to their essential role in GAS pathogenesis, the existence of several genes are serotype-specific, e.g., spd1, spd3, sfb1, enn, fbpA (the latter two are absent in M23ND), or linked to tissue tropicity, e.g., ska, sfb1. Some critical genes are regulated by the two component regulator, CovRS, and the ability of CovRS to become inactivated during the course of infection is a process that is particularly relevant with speB, hasA, and slo expression in M23ND to enhance its virulence (Fig. 8A,C).

Of special interest to our work, the ska gene, which is typically under strong CovRS regulation, is nearly equally expressed in both CovR⁻S⁻ and CovR⁺S⁺ M23ND strains (Fig. 8D). In addition, the mRNA of both emm23 and scpA appear to deviate from strict Mga-mediated expression, as both transcripts are attenuated at SP while mga remains nearly constant (Fig. 8B). This finding indicates that factors other than Mga influence the expression of emm23 and...
Lastly, the mRNA of the fibronectin binding protein, *sfbI* is regulated in a growth and CovRS dependent manner, since the CovR<sup>+</sup>S<sup>+</sup> strain produces significantly more *sfbI* transcript at LP, supporting its role in initial adherence and colonization. All of these observations suggest unique gene regulation in M23ND consistent with the variation observed in genetic organization. While greater differences are seen in expression of several of these genes at LP and SP and in CovR<sup>+</sup>S<sup>+</sup> and CovR<sup>+</sup>S<sup>-</sup> in AP53 cells (15) than in M23ND cells, any strain-dependent variations are likely to contain substantial contributions from the particular genomic architectural features of the strains.

**DISCUSSION**

The first complete genome of a M23 GAS strain, M23ND, was sequenced and compared to the 20 other fully-sequenced *S. pyogenes* strains currently available at NCBI. Our nucleotide sequence analysis showed that the genome contained four externally integrated prophage elements that encoded six virulence genes, including four *Streptococcal* pyrogenic exotoxins (*ssa, speC, speI* and *speH*) and two endonucleases (*spd1* and *spd3*). The acquisition of virulence factors via prophage integration plays an important role in the pathogenesis of GAS isolates. In the present study, we propose that the combined recombination of the six phage-encoded virulence genes is one of the major contributing factors responsible for the severity of *S. pyogenes* M23ND infection. A comparative study revealed large-scale genomic rearrangements, unique to M23ND, that are different from previously sequenced GAS strains. However, the rearranged genomic architecture is imbalanced, yielding two unequal replichores. It is possible that this resultant imbalance may contribute to the invasive nature of M23ND.
We also identified several known chromosomally-inherited virulence factors, mainly related to host cell-surface adherence and host immune system interactions. This indicates that a multitude of adaptive virulence factors has evolved to allow *S. pyogenes*, in general, and M23ND, in particular, to become a versatile human pathogen. The genomic locations of these bacterial chromosomal genes are generally conserved across various GAS strains, but are translocated in M23ND, resulting in global genomic rearrangement patterns unique to this genome. Translocations in chromosomally-encoded virulence genes may provide an alternative hypothesis for the enhanced adaptation of *Streptococci* to particular environmental pressures via altered gene regulation patterns. However, whether the translocation of virulence genes is related to pathogenicity in *S. pyogenes* is still unclear. The small number of fully-assembled *S.* pyogenes genomes prevents extensive and detailed assessments of the effects of gene translocations on virulence. Bacterial virulence is a complex process, with many end-points. Some of these genes may be needed for certain stages of an infection, while a combination of genes is likely essential for severe virulence and death of the host. Therefore, we believe that point mutations in virulence genes should be the dominant factor for altering gene expression. Point mutations in regulatory virulence genes or extracellular toxins provoke survival advantages and enhanced fitness to environments via modulating gene expression.

Phylogenetic analyses of housekeeping genes, based on MLST, showed that M23ND most likely shared a common ancestor (M18 8232) with strains M5 Manfredo and M6 10394. However, the genetic development, based on phage-encoded virulence factors, reflects a complex evolutionary path for M23ND, mediated by extensive horizontal gene transfer. Ultimately, the pathogenesis and invasiveness of *S. pyogenes* isolate M23ND is unlikely to have been induced by a single factor. It is reasonable to conclude that the demonstrated
lethality is a combined consequence of multiple factors, including the acquisition of virulence genes from prophages, global genomic rearrangements, and mutations in critical virulence factors. In addition, adaptive gene expression is also an important factor for the invasive infection and environmental fitness.

It should necessarily be considered that all strains of GAS that have been analyzed normally originate from a host and are oftentimes further genetically manipulated and further passaged in the laboratory setting. These practices result in gene interactions between host and bacteria that may alter the genetic characteristics of the bacteria, and the nature of the originating infecting agent may genetically differ from the isolate. The most obvious example of this is the inactivation of covR and covS during the course of infection (54), and the profound consequences of covRS inactivations toward virulence, especially in regard to speB production. SpeB, produced in CovR+ cells, is beneficial to the dissemination of the initial infection since speB catalyzes degradation of extracellular matrix components, e.g., fibronectin and vitronectin (66); activates proinflammatory proteins, e.g., IL-1β (67); degrades IgG and IgA (68), and inactivates complement factors (69), thereby circumventing the host immune response. This protease can also cleave EpC junction proteins, thereby facilitating GAS translocation across the epithelium (70). However, in late infection, the presence of speB may not be beneficial to bacterial dissemination since this protease inactivates GAS virulence factors. Thus, it is an advantage for GAS to downregulate speB expression at late infection stages. Since inactivating mutants of speB have not been found to date during the course of infection, mutagenic inactivation of covR, covS, and/or another speB regulator, rgg (71), serve as mutagenic loci during infection to downregulate expression of speB. This highly tuned phase switch mechanism with covRS would initiate hyperinvasive disease (72), to a major extent via SpeB
regulation, by maximizing conditions for initial GAS localized protease-base invasion into tissue, and at later stages, by rapidly attenuating SpeB production and preserving GAS virulence factors that are needed for dissemination.

Another manner of M23ND assembling a virulent proteolytic surface is via binding of host plasminogen (hPg) and plasmin (hPm). Upon examining the amino acid sequence of the major predicted hPg/hPm binding protein, M23, we propose that M23 should not bind hPg/hPm directly, but should employ its ability to bind host fibrinogen (hFg) which then will allow hPg/hPm binding. We have previously established that a coinheritance of isoforms of SK and the mode of binding hPg occurs (73-75). The amino acid sequence of SK secreted by M23ND is the SK2a form, which maximally activates hPg bound to GAS via hFg/M protein. Thus, the principles of coinheritance of the forms of SK and M-protein are verified with M23ND.

In conclusion, on a gene content level, the hypervirulence of M23ND is consistent with the covS- mutation found, the presence of the prophage superantigen gene, ssa, and the expression of the critical fibronectin binding gene, sfb1. These genes, plus the gene architectural features described throughout this manuscript, explain the presence of this form of the bacterium in the hyperinfectious human isolate from which it was discovered.

Acknowledgments

This study was supported by NIH Grant HL013423.
REFERENCES


FIG 1 Circular representation of the 1,846,477 bp genome of *S. pyogenes* strain M23ND.

Data are shown from the outermost to innermost circles. Circles 1 and 2 display annotated coding sequences for the reverse (pink) and forward (blue) DNA strands, respectively. Circle 3 shows the locations of four phage elements ΦM23ND.1-ΦM23ND.4 (red boxes) and short mobile elements (black lines). Circle 4 illustrates virulence genes identified in M23ND (orange). Circles 5 and 6 represent the locations of tRNA (turquoise) and rRNA (brown) genes, respectively. The genome contains 57 tRNA genes and 5 rRNA operons. Circles 6 and 7 display the GC content and GC skew ([G-C]/[G+C]). The triangles on the outermost circumference indicate the positions of the replication origin (*ori*; green triangle) at zero and replication terminus (*ter*; red triangle) at 702,443 bp. The genome contains 1851 ORFs, 231 of which are present in phage-encoded regions. Approximately 1389 genes have assigned functions.

FIG 2 Global genome sequence comparisons of fully sequenced *S. pyogenes* genomes. The sequences of the 20 previously sequenced GAS genomes were obtained from the NCBI database and were correlated with a basic local alignment with *S. pyogenes* strain M23ND (innermost circle). The areas of similarity and divergence within the sequences are contrasted, with white gapped areas indicating regions of highest variance. The profile of M1 5005 and M1 A20 are similar to M1 476, and are therefore not included in the figure. Phage proteins and short mobile genetic elements are indicated by the black and red arrows, respectively, in the outer circle. The strains, compared to M23ND, which is positioned on the central ring, are: M1 SF370, M1 476, M2 10270, M3 315, M3 SSI-1, M4 10750, M5 Manfredo, M6 10394, M12...
FIG 3 DNA characterizations of genomic sequences of *S. pyogenes*. (A) A circular visualization of the comparative locations of phage elements across M23ND (*red*) with the 20 previously fully sequenced GAS genomes available through NCBI (*blue*). The integration sites of phages are generally clustered at several regions, but the locations of orthologous phages are non-conserved. For example, prophage \( \Phi \)M23ND.1 is closely similar to \( \Phi \)Manfredo.4 (M5), \( \Phi \)MGAS10394.3 (M6), \( \Phi \)MGAS1882.2 (M18), and several others, but these prophages are inserted at distinct sites, suggesting that phage recombination via horizontal gene transfer plays an important role in genetic diversity. The black triangles on the circumference indicate the positions of the replication origin (ori) at zero and terminus (ter) at 702,433 bp, respectively. (B) A circular visualization of the comparative locations of the virulence factors of interest within the genomes of the 21 fully-sequenced GAS genomes. The elongated bars represent the regions in which each gene can be found across all of the 21 fully sequenced and assembled GAS genomes, unless the gene appears elsewhere e.g. *covRS* for SSI-1, Manfredo, HKU16 and M23ND; or *sen*, *srv* and *sagA* for M23ND. Inversions around ori (white triangle) are clearly observed, especially for *srv*, and inversions around ter (white triangle) are readily seen in the cases of *sen*, *sagA*, and *fhp* (*fhp54*).

FIG 4 Whole genome comparisons of *S. pyogenes* M23ND with two phylogenetic neighbors, M5 Manfredo and M18 8232. The red and blue lines represent forward and
reverse alignments respectively; prophage elements are indicated by colored boxes, and black arrows indicate the replication terminus in each genome. Large segmental inversions and translocations are observed in M23ND relative to its neighbors. These rearrangements result in an imbalanced global genomic architecture, where the replication terminus (ter) and three prophage insertions were located within the same replicore of the chromosome. M23ND also contains an additional short inversion in the final 100 kb of the sequence. Prophage elements are located on the breakpoints of, or within, the rearrangements themselves.

**FIG 5 Profiling of phage encoded virulence factors across S. pyogenes genomes.** Specific genes are represented by colored triangles and the length of each gene is scaled by the triangle size. M23ND contains six known phage-encoded virulence factors incorporated in four prophage regions, including speC, spd1, ssa, spd3, speI, and speH. Profiling of these six virulence factors, together with others across 21 GAS sequences, showed that they are randomly distributed throughout the chromosome.

**FIG 6 Profiling of chromosomally inherited virulence factors across S. pyogenes genomes.** Genes are represented by colored triangles and the length of each gene is scaled by the triangle width. These genes are present in almost all of the 21 fully-sequenced GAS genomes, except for fibronectin-binding protein (sfb1), streptococcal inhibitor of complement (sic), and pyrogenic exotoxin J (speJ), which are encoded by 10, 4, and 7 GAS strains, respectively. The overall genomic locations of chromosomally-encoded virulence factors are conserved across different GAS strains, with the exception of M23ND, M5 Manfredo, M12 HKU16, and M3 SSI-1, where gene locations are obscured by large-scale genomic arrangements including...
translocations and inversions. A total of 25 known virulence factors present in M23ND include:

(A) endoS, hasA, hylA, ideS, scpA, sk, smeZ, spd, speB, speG, and spyA, and (B) cfa, dltA, dltC, sen, graB, lmb, nga, ptr, sagA, sxlA, sfb1, sic, and slo.

**FIG 7 Profiling of six regulatory genes across S. pyogenes genomes.** Genes are represented by colored triangles and the length of each gene is scaled by the triangle width. All 21 currently available genome sequences were analyzed. The genes examined include emm, mga, covR, covS, rgg, and srv. The M-protein gene, emm, has been artificially shifted in order to avoid overlap with mga. The locations of regulatory genes are highly conserved across all GAS strains, except for srv in M23ND. This gene is displaced from a position of ~1.5 Mbp to ~1.2 Mbp, induced by translocation of a large fragment within the genome. Similar translocations are evident for other genes in M23ND, e.g., sagA and sfb1.

**FIG 8 Gene expression in full-length and truncated CovS M23ND strains.** mRNAs isolated from the two strains (CovS+ and CovS−) at LP (A_{600nm} = 0.6) and SP (A_{600nm} > 1.0) were analyzed for virulence factor gene expression using q-RT-PCR. Primers specific for each gene (Table S1) were used to measure relative transcription levels compared to the housekeeping gene, gapdh. (A) The relative gene expression of the extracellular cysteine protease, speB. (B) Gene expression in the multigene regulon (Mga) of GAS, which, in the case of M23ND contains the minimal gene content, viz., M protein (emm23) and C5α peptidase (scpA). (C) q-RT-PCR analyses of cell surface and secreted virulence factors involved in fibronectin binding proteins (sfb1 and fbp54), hyaluronic acid capsule biosynthesis (hasA), DNase activity (spd and spd3),
and host cell lysis (slo). (D) q-RT-PCR analysis of host-plasminogen activator, ska, showing
very similar expression in both strains and throughout both LP and SP growth.
### TABLE 1 Characteristics of fully sequenced *S. pyogenes* genomes

<table>
<thead>
<tr>
<th>Strain</th>
<th>M-type</th>
<th>RefSeq Accession</th>
<th>Clinical Source</th>
<th>Genome Size</th>
<th>Prophages</th>
<th>Genes</th>
<th>Proteins</th>
<th>%GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>476</td>
<td>1</td>
<td>NC_020540</td>
<td>STSS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1,831,128</td>
<td>3</td>
<td>1,644</td>
<td>1,571</td>
<td>38.5</td>
</tr>
<tr>
<td>5005</td>
<td>1</td>
<td>NC_007297</td>
<td>CSF&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1,838,554</td>
<td>3</td>
<td>1,950</td>
<td>1,865</td>
<td>38.5</td>
</tr>
<tr>
<td>A20</td>
<td>1</td>
<td>NC_18395</td>
<td>Blood, NF&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1,837,281</td>
<td>3</td>
<td>1,915</td>
<td>1,828</td>
<td>38.5</td>
</tr>
<tr>
<td>SF370</td>
<td>1</td>
<td>NC_002737</td>
<td>Wound Infection</td>
<td>1,852,441</td>
<td>4</td>
<td>1,810</td>
<td>1,696</td>
<td>38.5</td>
</tr>
<tr>
<td>10270</td>
<td>2</td>
<td>NC_008022</td>
<td>Pharyngeal</td>
<td>1,928,252</td>
<td>5</td>
<td>2,068</td>
<td>1,987</td>
<td>38.4</td>
</tr>
<tr>
<td>315</td>
<td>3</td>
<td>NC_04070</td>
<td>STTS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1,900,521</td>
<td>6</td>
<td>1,951</td>
<td>1,865</td>
<td>38.6</td>
</tr>
<tr>
<td>SSI-1</td>
<td>3</td>
<td>NC_004606</td>
<td>STTS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1,894,275</td>
<td>6</td>
<td>1,931</td>
<td>1,859</td>
<td>38.6</td>
</tr>
<tr>
<td>10750</td>
<td>4</td>
<td>NC_008024</td>
<td>Pharyngeal</td>
<td>1,937,111</td>
<td>4</td>
<td>2,059</td>
<td>1,978</td>
<td>38.3</td>
</tr>
<tr>
<td>Manfredo</td>
<td>5</td>
<td>NC_009332</td>
<td>ARF&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1,841,271</td>
<td>5</td>
<td>1,906</td>
<td>1,745</td>
<td>38.6</td>
</tr>
<tr>
<td>10394</td>
<td>6</td>
<td>NC_006086</td>
<td>Pharyngeal</td>
<td>1,899,877</td>
<td>8</td>
<td>1,971</td>
<td>1,886</td>
<td>38.7</td>
</tr>
<tr>
<td>2096</td>
<td>12</td>
<td>NC_008023</td>
<td>Glomerulonephritis</td>
<td>1,860,355</td>
<td>2</td>
<td>1,979</td>
<td>1,898</td>
<td>38.7</td>
</tr>
<tr>
<td>9429</td>
<td>12</td>
<td>NC_008021</td>
<td>Infection</td>
<td>1,836,467</td>
<td>3</td>
<td>1,962</td>
<td>1,877</td>
<td>38.5</td>
</tr>
<tr>
<td>HKU16</td>
<td>12</td>
<td>NZ_AFROY01000001</td>
<td>Blood, Scarlet Fever</td>
<td>1,908,100</td>
<td>3</td>
<td>1,950</td>
<td>1,865</td>
<td>38.4</td>
</tr>
<tr>
<td>HSC5</td>
<td>14</td>
<td>NC_021807</td>
<td>Infection</td>
<td>1,818,351</td>
<td>3</td>
<td>1,829</td>
<td>1,744</td>
<td>38.5</td>
</tr>
<tr>
<td>8232</td>
<td>18</td>
<td>NC_003485</td>
<td>ARF&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1,895,017</td>
<td>5</td>
<td>1,924</td>
<td>1,839</td>
<td>38.5</td>
</tr>
<tr>
<td>M23ND&lt;sup&gt;d&lt;/sup&gt;</td>
<td>23</td>
<td>Invasive Infection</td>
<td>1,846,477</td>
<td>4</td>
<td>1,851</td>
<td>1,620</td>
<td>38.6</td>
<td></td>
</tr>
<tr>
<td>6180</td>
<td>28</td>
<td>NC_007296</td>
<td>Invasive Infection</td>
<td>1,897,573</td>
<td>4</td>
<td>1,977</td>
<td>1,894</td>
<td>38.4</td>
</tr>
<tr>
<td>NZ131</td>
<td>49</td>
<td>NC_001375</td>
<td>Glomerulonephritis</td>
<td>1,815,785</td>
<td>3</td>
<td>1,788</td>
<td>1,700</td>
<td>38.6</td>
</tr>
<tr>
<td>Alab49</td>
<td>53</td>
<td>NC_017596</td>
<td>Impetigo Lesion</td>
<td>1,827,308</td>
<td>4</td>
<td>1,866</td>
<td>1,773</td>
<td>38.6</td>
</tr>
<tr>
<td>1882</td>
<td>59</td>
<td>NC_017053</td>
<td>Infection</td>
<td>1,781,029</td>
<td>2</td>
<td>1,792</td>
<td>1,691</td>
<td>38.5</td>
</tr>
<tr>
<td>15252</td>
<td>59</td>
<td>NC_017040</td>
<td>Invasive Infection</td>
<td>1,750,832</td>
<td>2</td>
<td>1,757</td>
<td>1,662</td>
<td>38.5</td>
</tr>
</tbody>
</table>

---

<sup>a</sup> Current study.

<sup>b</sup> CSF = cerebrospinal fluid.

<sup>c</sup> NF = necrotizing faciitis.

<sup>d</sup> STTS = streptococcal toxic shock.

<sup>e</sup> ARF = acute rheumatic fever.
### TABLE 2 Phage elements in M23ND and their orthologs in other GAS strains

<table>
<thead>
<tr>
<th>GAS Strain</th>
<th>ΦM23ND.1(^a) ((168,433-209,722))^b</th>
<th>ΦM23ND.2(^a) ((409,085-450,619))^b</th>
<th>ΦM23ND.3(^a) ((572,556-612,004))^b</th>
<th>ΦM23ND.4(^a) ((842,696-876,985))^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1 SF370</td>
<td>Φ10270.1</td>
<td>Phi10270.1</td>
<td>Phi10270.1</td>
<td>PhiSF370.3</td>
</tr>
<tr>
<td>M2 10270</td>
<td>Φ315.2</td>
<td>Phi315.6</td>
<td>Phi315.2</td>
<td>Phi315.3</td>
</tr>
<tr>
<td>M3 315</td>
<td>ΦSSI.5</td>
<td>PhiSSI.1</td>
<td>PhiSSI.5</td>
<td>PhiSSI.4</td>
</tr>
<tr>
<td>M4 10750</td>
<td>Φ315.2</td>
<td>Phi315.6</td>
<td>Phi315.2</td>
<td>Phi315.3</td>
</tr>
<tr>
<td>M5 Manfredo</td>
<td>ΦManfredo.4</td>
<td>PhiManfredo.3</td>
<td>PhiManfredo.4</td>
<td>PhiManfredo.2</td>
</tr>
<tr>
<td>M6 10394</td>
<td>Φ10394.3</td>
<td>Phi10394.1</td>
<td>Phi10394.3</td>
<td>Phi10394.5</td>
</tr>
<tr>
<td>M12 2096</td>
<td>Φ2096.3</td>
<td>Phi2096.3</td>
<td>Phi2096.3</td>
<td></td>
</tr>
<tr>
<td>M12 9429</td>
<td>Φ9429.1</td>
<td>Phi9429.3</td>
<td>Phi9429.1</td>
<td></td>
</tr>
<tr>
<td>M12 HKU16</td>
<td>ΦHKU16.2</td>
<td>PhiHKU16.2</td>
<td>PhiHKU16.2</td>
<td></td>
</tr>
<tr>
<td>M14 HSC5</td>
<td>ΦMANFREDO.4</td>
<td>PhiMANFREDO.3</td>
<td>PhiMANFREDO.4</td>
<td>PhiHSC5.2</td>
</tr>
<tr>
<td>M18 8232</td>
<td>Φ8232.2</td>
<td>Phi8232.2</td>
<td>Phi8232.2</td>
<td>Phi8232.5</td>
</tr>
<tr>
<td>M49 NZ131</td>
<td>ΦNZ131.2</td>
<td>PhiNZ131.2</td>
<td>PhiNZ131.2</td>
<td>PhiNZ131.2</td>
</tr>
<tr>
<td>M59 1882</td>
<td>Φ1882.1</td>
<td>Phi1882.1</td>
<td>Phi1882.1</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) The phage element ΦM23ND.1 encodes the mitogenic factors, *spd1*, and exotoxin C, *speC*; ΦM23ND.2 encodes superantigen A (*ssa*); ΦM23ND.3 encodes the mitogenic factor, *spd3*; and ΦM23ND.4 encodes exotoxin type I/H (*speI/speH*).

\(^b\) The genomic locations of phage elements in M23ND denoted in the parenthesis after the phage.
S. pyogenes M23ND
1,846,477 bp
A

M23ND prophages
GAS prophages
ICE

1.8M
0.2M
0.4M
0.6M
0.8M
1.0M
1.6M
1.4M
1.2M

S. pyogenes prophage distribution

B

1. M1 476
2. M1 MGAS5005
3. M1 SF370
4. M1 A20
5. M2 MGAS10270
6. M3 MGAS315
7. M4 MGAS10750
8. M6 MGAS10394
9. M12 MGAS9429
10. M12 MGAS2096
11. M14 HSC5
12. M18 MGAS8232
13. M28 MGAS6180
14. M49 NZ131
15. M53 Alab49
16. M59 MGAS15252
17. M59 MGAS19882
18. M3 SSI-1
19. M5 Manfredo
20. M12 HKU16
21. M23ND