Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A *Streptococcus*

Stephen B. Beres*, Ellen W. Richter*, Michal J. Nagiec*, Paul Sumby*, Stephen F. Porcella[†], Frank R. DeLeo[†], and James M. Musser*[‡]

*Center for Molecular and Translational Human Infectious Diseases Research, The Methodist Hospital Research Institute, Houston, TX 77030; and [†]Laboratory of Human Bacterial Pathogenesis, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 903 South Fourth Street, Hamilton, MT 59840

Edited by John J. Mekalanos, Harvard Medical School, Boston, MA, and approved February 8, 2006 (received for review November 29, 2005)

In recent years we have studied the relationship between strain genotypes and patient phenotypes in group A *Streptococcus* (GAS), a model human bacterial pathogen that causes extensive morbidity and mortality worldwide. We have concentrated our efforts on serotype M3 organisms because these strains are common causes of pharyngeal and invasive infections, produce unusually severe invasive infections, and can exhibit epidemic behavior. Our studies have been hindered by the lack of genome-scale phylogenies of multiple GAS strains and whole-genome sequences of multiple serotype M3 strains recovered from individuals with defined clinical phenotypes. To remove some of these impediments, we sequenced to closure the genome of four additional GAS strains and conducted comparative genomic resequencing of 12 contemporary serotype M3 strains representing distinct genotypes and phenotypes. Serotype M3 strains are a single phylogenetic lineage. Strains from asymptomatic throat carriers were significantly less virulent for mice than sterile-site isolates and evolved to a less virulent phenotype by multiple genetic pathways. Strain persistence or extinction between epidemics was strongly associated with presence or absence, respectively, of the prophage encoding streptococcal pyrogenic exotoxin A. A serotype M3 clone significantly underrepresented among necrotizing fasciitis cases has a unique frameshift mutation that truncates MtsR, a transcriptional regulator controlling expression of genes encoding iron-acquisition proteins. Expression microarray analysis of this clone confirmed significant alteration in expression of genes encoding iron metabolism proteins. Our analysis provided unprecedented detail about the molecular anatomy of bacterial strain genotype-patient phenotype relationships.

carrier | epidemic | genomic resequencing | SNP | phylogeny

Microbial epidemics have repeatedly altered the course of history by decimating human populations, killing domesticated animals, and blighting crops. Despite the impact of these destructive events, we know relatively little about the evolutionary genetic events contributing to bacterial strain emergence and diversification within and between epidemic waves. Also poorly understood is the molecular basis of intraspecies variation in disease phenotype. The recent confluence of genome sequencing and development of high-throughput techniques to interrogate genetic polymorphisms in large samples of strains now permits these topics to be studied at the individual nucleotide level.

Group A *Streptococcus* (GAS) infects humans worldwide and causes an array of diseases ranging from superficial infections such as pharyngitis to fulminant invasive infections characterized by high morbidity and mortality (1, 2). In recent years, we have used serotype M3 GAS strains as model pathogens for studying the molecular processes contributing to clone emergence, epidemic waves, and genotype–phenotype relationships (M protein is a highly polymorphic cell-surface molecule that is antiphagocytic and forms the basis of a scheme commonly used to classify GAS strains) (3, 4).

Our interest in strain genotype–patient phenotype relationships in serotype M3 strains stems from several important observations. Serotype M3 strains cause a disproportionate number of invasive disease cases, including necrotizing fasciitis, bacteremia, and streptococcal toxic shock syndrome (5–9). Importantly, serotype M3 strains cause a higher rate of lethal infections than strains of other M types (7–9). In addition, serotype M3 and other GAS strains can undergo rapid shifts in disease frequency and display epidemic behavior (5). The emergence of severe invasive infections caused by serotype M3 GAS induced us to sequence the genome of a strain that is genetically representative of the clone causing most contemporary episodes of serotype M3 disease (3). We then used the genome sequence data to study the relationship between bacterial strain genotype and patient disease phenotype on a genome-wide level by analysis of 255 serotype M3 invasive isolates collected in an 11-year population-based surveillance study conducted in Ontario, Canada (4). Genetic diversity in the strains was indexed by pulsed-field gel electrophoresis, DNA–DNA microarray, whole-genome PCR scanning, prophage genotyping, targeted gene sequencing, and SNP genotyping. All variation in gene content was attributable to gain or loss of prophages, a process that generates distinct combinations of virulence genes. Our analysis (4) suggested that distinct serotype M3 genotypes experienced rapid clonal expansion and caused infections that significantly differed in character.

Although the findings described above add to our understanding of the molecular events underlying clone emergence and epidemic behavior, many issues pertaining to these two epidemic waves remained unresolved. For example, although the M3 strains had been analyzed with state-of-the-art molecular methods, relatively little of the 1.9-Mb genome of each strain was studied, which means we have only a cursory understanding of genetic variation in these organisms. Furthermore, genetic relationship between strains causing invasive disease episodes to strains of the same serotype recovered from asymptomatic carriers was not examined. Importantly, it is not known whether strains cultured from asymptomatic carriers differ in virulence compared to invasive isolates. Finally, although we identified distinct subclone–patient phenotype associations, the molecular underpinnings of these relationships were not studied. These issues are universal to all microbial epidemics, and hence of widespread interest. Here, we report the results

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: GAS, group A *Streptococcus*; MLST, multilocus sequence typing; CGR, comparative genomic resequencing; CFU, colony-forming unit.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. CP000003, CP000017, CP000056, CP000259–CP000262, AE04092, AE009949, AE14074, and BA000034), and the expression microarray data have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE3966).

[‡]To whom correspondence should be addressed. E-mail: jmmusser@tmh.tmc.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0510279103

Table 1. Sequenced GAS strains

Strain	M type	Size, bp	CDS	Prophage	MLST	GenBank accession no.	ATCC no.
SF370	1	1,852,441	1,697	4	28	AE004092	700294
MGAS5005	1	1,838,554	1,865	3	28	CP000017	BAA-947
MGAS10270	2	1,928,252	1,987	5	55	CP000260	BAA-1063
MGAS315	3	1,900,521	1,865	6	15	AE14074	BAA-595
SSI-1	3	1,894,275	1,861	6	15	BA000034	—
MGAS10750	4	1,937,111	1,979	4	39	CP000262	BAA-1066
Manfredo	5	1,841,271	1,803	5	99	—	—
MGAS10394	6	1,899,877	1,886	8	382	CP000003	BAA-946
MGAS2096	12	1,860,355	1,898	2	36	CP000261	BAA-1065
MGAS9429	12	1,836,467	1,878	3	36	CP000259	BAA-1315
MGAS8232	18	1,895,017	1,845	5	42	AE009949	BAA-572
MGAS6180	28	1,897,573	1,894	4	52	CP000056	BAA-1064

ATCC, American Type Culture Collection.

of studies conducted to address these and other gaps in our knowledge.

Results and Discussion

Interserotype Comparative Genomics and Phylogenetic Relationships. To gain an enhanced understanding of the phylogenetic relationship of serotype M3 to other GAS strains commonly causing human infections (5–10), we sequenced the genome of a serotype M2, M4, and two M12 strains to closure and an average Q40 value throughout. Twelve whole-genome sequences (two strains each of serotype M1, M3, and M12, and one strain each of serotype M2, M4, M5, M6, M18, and M28) of GAS strains were available for analysis (Table 1) (4, 11–16). The genomes were aligned, and SNPs were identified relative to reference serotype M3 strain MGAS315 by using MUMMER (17). In the aggregate, 158,005 SNPs at 62,504 loci dispersed throughout the GAS metagenome were found (Table 3, which is published as supporting information on the PNAS web site). The core metagenome (the portion of the chromosome that lacks obvious foreign elements such as prophages and is largely conserved among all of the sequenced genomes) is ≈1,670 kbp. We identified 142,919 SNPs at 43,356 loci in this core metagenome, or one SNP every 38 bp. Between strains of different serotypes, the number of core SNPs ranged from 10,165 to 15,797, with an average of 14,475, or one SNP every 115 bp. In comparison, the estimated frequency for human autosomes is one SNP every 1,300 bp (18).

A SNP matrix (Fig. 6, which is published as supporting information on the PNAS web site) and corresponding phylogenetic network were derived by using the core SNP data (Fig. 1*A*). For comparison, genetic relationships also were inferred based on the seven gene segments used in a commonly used multilocus sequence typing (MLST) scheme (http://spyogenes.mlst.net) (Fig. 1*B*). These comparisons exclude exogenous gene content, largely encoded by putatively mobile genetic elements that might obscure evolutionary relationships based on vertical inheritance.

Estimates of genetic relationships based on MLST and core SNP data were largely concordant (Fig. 1). However, the MLST network failed to reveal that the M12 and M28 genomes were allied. Furthermore, it did not distinguish between genomes of strains of the same serotype, for example, M1 strains SF370 and MGAS5005 that differ genetically and phenotypically (16). Thus, the analysis shows that serotype M3 strains represent a distinct GAS phylogenetic lineage that has not evolved recently from a precursor strain represented by the organisms studied.

Virulence Assessment of Serotype M3 Strains. Relatively little is known about the molecular factors that contribute to development of an asymptomatic carrier state in microbial pathogens. This is important because, for GAS, evidence indicates that asymptomatic carriers are less likely than individuals with clinically significant infections to disseminate organisms and develop postinfection sequelae (19–21). In principle, asymptomatic carriage could occur because genetic changes in the pathogen mitigate virulence. Alternatively, host factors such as enhanced immune status may be a dominant contributor. To begin to address this issue, we next studied strain–genotype patient–phenotype associations in detail among 13 well characterized contemporary serotype M3 strains representing two groups with distinctly different clinical phenotypes (Table 2). Nine strains were cultured from patients with invasive infections (invasive strains), and four strains were throat isolates from asymptomatic young adult carriers with no recent history of GAS pharyngitis (carrier strains). The invasive strains studied are genetically representative of the six major serotype M3 subclones identified in the analysis of two epidemic waves in Ontario, Canada, centered in 1995 and 2000 (4).

To test the hypothesis that carrier strains are less virulent than invasive strains, we used a mouse model of systemic infection. As a group, invasive strains had a significantly lower mean 90% near-lethal dose (LD_{90}) compared to carrier strains ($P = 0.011$) after i.p. injection (Fig. 2 and Table 4, which is published as supporting information on the PNAS web site). This attenuated virulence was independent of differences in gene content. If the comparison is constrained to include only those invasive strains with the same

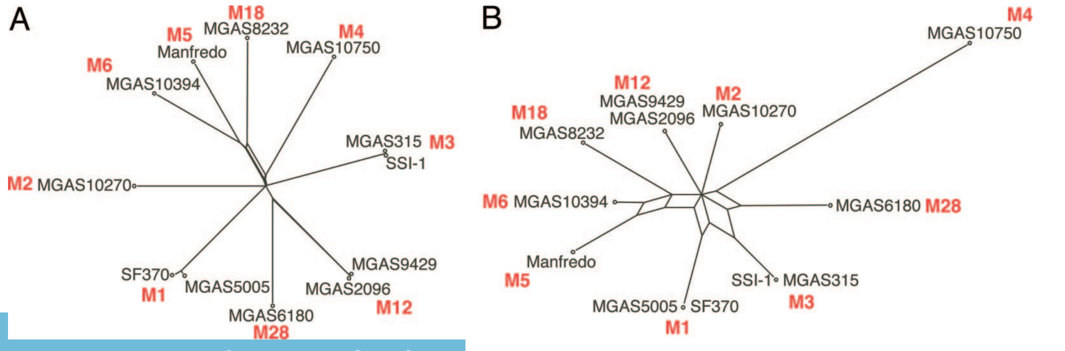


Fig. 1. Interserotype genetic relationships. (*A*) Relationships inferred from 43,356 core SNP loci concatenated nucleotides. (*B*) Relationships inferred from 3,134 MLST internal gene segment concatenated nucleotides.

Table 2. Serotype M3 strains

MGAS strain	Group	Date isolated	Infection	SC	SG	ΦG	*emm3* *allele*	MLST
315	Invasive	1985	STSS	Other	Other	ΦG3.01	3.1	15
3378	Invasive	1994	NF	SC4	SG3.02	ΦG3.03	3.1	15
3382	Invasive	1994	NF	SC3	SG3.02	ΦG3.01	3.1	15
3392	Invasive	1995	NF	SC1	SG3.01	ΦG3.01	3.1	15*
3394	Invasive	1995	NF	SC2	SG3.01	ΦG3.03	3.1	15
9887	Invasive	2000	BAC	SC5	SG3.01	ΦG3.01	3.2	15*
9919	Invasive	2000	NF	SC6	SG3.03	ΦG3.02	3.1	15
9967	Invasive	2000	NF	SC1	SG3.01	ΦG3.01	3.1	15
9982	Invasive	1999	NF	SC3	SG3.02	ΦG3.01	3.1	15
12501	Carrier	1993	ASM	SC3	SG3.02	ΦG3.01	3.1	15
12502	Carrier	1994	ASM	SC7	SG3.07	ΦG3.01	3.28	15
12503	Carrier	1994	ASM	SC1	SG3.01	ΦG3.01	3.1	15
12504	Carrier	1994	ASM	SC7	SG3.11	ΦG3.01	3.30	15

MGAS, Musser collection GAS strain; SC, subclone, SG, SNP genotype; ΦG, prophage genotype; STSS, streptococcal toxic shock syndrome; NF, necrotizing fasciitis; BAC, bacteremia; ASM, asymptomatic. Subclones are defined by their combination of SNP genotype, phage genotype, and *emm3* allele designations as described (4).

*Strain MGAS3392 differs by 1 nt and strain MGAS9887 by 2 nt from MLST *gtr* (glutamine transporter) allele 6. By allelic profile, these strains are closest to sequence type 15 in the GAS MLST database.

gene content as the carrier isolates (i.e., only the ΦG3.01 *speA*$^+$ strains), then the mean LD_{90} values are even more disparate (Fig. 2). The invasive strains caused near-mortality in a dose-dependent manner over a range spanning 3 to 4 orders of magnitude (5×10^5 to 5×10^8 colony-forming units, CFU). In striking contrast, the carrier strains had a much narrower response range that spanned <2 orders of magnitude, and were avirulent at doses $<10^7$ CFU. After 5 days, 78% (31 of 40) of the mice injected with the carrier isolates at a dose of $\approx 5 \times 10^7$ CFU survived, whereas only 31% (28 of 90) of the mice injected with the invasive isolates survived ($P = 0.050$). Taken together, the data argue that changes in the bacteria contribute to the carrier state.

Identification of Genetic Polymorphisms in Serotype M3 Strains. Previously we reported on several genetic characteristics of the invasive strains studied herein, but carrier strains were not studied. We analyzed four carrier strains and found that they had the same core gene content and the same prophage content as reference strain MGAS315 recovered from a patient with invasive disease (Table 2). Additionally, whole-genome PCR scanning indicated that all ΦG3.01 strains, invasive or carrier, have the same prophage-encoded exogenous gene content (ref. 4, and data not shown). These results imply that allelic variation is a key contributor to the attenuated virulence observed for the carrier strains. To identify allelic variation that may contribute to the observed virulence phenotypes, we used a two-stage DNA–DNA microarray hybridization-based comparative genomic resequencing (CGR) method (22). These arrays, in conjunction with polymorphism sequencing, defined genome-wide all strain-to-strain differences in gene content (Fig. 7, which is published as supporting information on the PNAS web site). Among the eight invasive and four carrier strains, an aggregate of 483 SNPs mapping to 249 loci were identified relative to the core genome sequence (≈1,670 kbp) of reference strain MGAS315 (Table 5, which is published as supporting information on the PNAS web site), translating to an average of one SNP every 6,707 bp. Thus, intra-M3 SNP frequency was 177-fold less than the interserotype average among the strains studied. A matrix of strain-to-strain SNP differences and a corresponding phylogenetic network were derived from the core SNPs (Figs. 3 and 4*A*). The number of SNPs between these 13 M3 strains ranged from 9 to 94 (average, 54 SNPs), or an average of one core SNP every 31 kbp, a frequency ≈270-fold less than the serotype-to-serotype average. This analysis confirmed that serotype M3 strains represent a closely related lineage of GAS.

Conventional sequencing was used to verify a subset of the SNPs identified by CGR. We targeted coding sequence (CDS) SNPs because they may be more likely to have functional consequences than SNPs found in intergenic regions. Sequence data were obtained for 88% (350 of 396) of the core CDS SNPs studied. Nearly all (98%; 343 of 350) CGR called SNPs were confirmed to be true, and no false-negative SNPs were identified. The few false-positive SNPs (2%) were present in genes with homologues located elsewhere in the genome (for example,

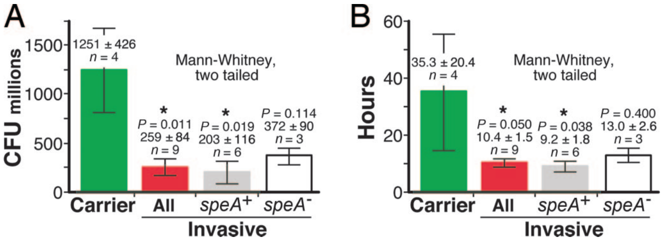
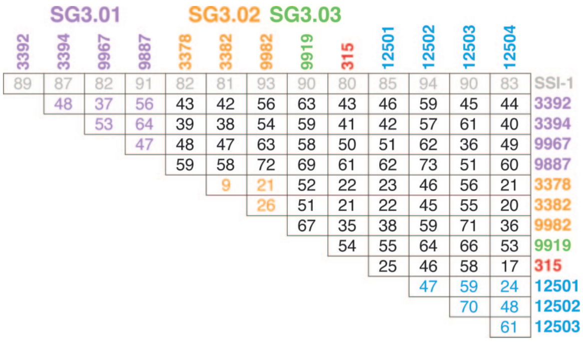


Fig. 2. Invasive and carrier isolate virulence comparison. (*A*) Comparison of mean LD_{90} values at 48 h after IP injection. (*B*) Comparison of mean survival time for $\approx 5 \times 10^8$ CFU IP dose. All *speA*$^+$ strains are ΦG3.01. Invasive groups that differ significantly from the carriers are marked with an asterisk. Bars are the SEM.



	SG3.01				SG3.02			SG3.03					
3392	3394	9967	9887	3378	3382	9982	9919	315	12501	12502	12503	12504	
89	87	82	91	82	81	93	90	80	85	94	90	83	SSI-1
	48	37	56	43	42	56	63	43	46	59	45	44	3392
		53	64	39	38	54	59	41	42	57	61	40	3394
			47	48	47	63	58	50	51	62	36	49	9967
				59	58	72	69	61	62	73	51	60	9887
					9	21	52	22	23	46	56	21	3378
						26	51	21	22	45	55	20	3382
							67	35	38	59	71	36	9982
								54	55	64	66	53	9919
									25	46	58	17	315
										47	59	24	12501
											70	48	12502
												61	12503

Fig. 3. Intraserotype M3 core SNP matrix. Invasive isolate strain designations are color-coded to indicate SNP genotypes as previously defined (4), and carrier strains are colored blue. SNP numbers are color-coded to indicate a comparison between strains of the same SNP genotype or between carriers. Strain SSI-1 is given in gray to indicate that its SNPs were determined by *in silico* comparison and not by CGR. The average number of core SNPs strain-to-strain is 54.

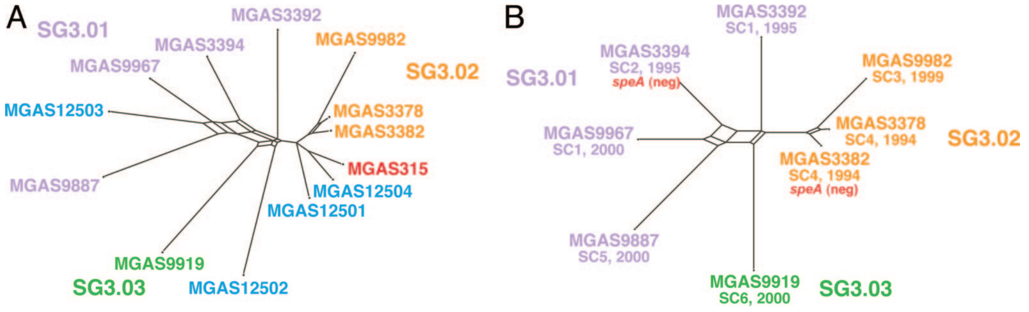


Fig. 4. Intraserotype M3 genetic relationships. (*A*) Relationships among contemporary M3 strains inferred from 248 core SNP loci concatenated nucleotides. (*B*) Relationships among invasive isolates inferred from 167 core SNP loci concatenated nucleotides. Invasive strain designations are color coded to indicate SNP genotypes as previously defined (4). Asymptomatic carrier isolates are colored aqua, and the high-virulence reference strain MGAS315 is colored red.

transposases and ABC transporters). Thus, the error rate in CGR SNP identification was very low and the errors likely were caused by cross-hybridization between similar DNA sequences.

CGR also identified 183 "other polymorphisms" (i.e., hybridization anomalies) that mapped to 148 loci. PCR amplification and sequencing determined that 53% of these anomalies (97 at 91 loci) were nonexistent, that is, were false-positive results. The remaining 86 anomalies coalesced to 51 polymorphisms that mapped to 27 loci (Table 6, which is published as supporting information on the PNAS web site). Indels that ranged in size from 1 to 195-bp accounted for 38 of these polymorphisms. Many (13 of 18) of the small indels (<5 bp) represented size variation in a homopolymeric tract (e.g., an additional T in a run of six Ts), whereas most (16 of 19) of the larger indels (>5 bp) were caused by variation in the number of tandem repeats. Ten of the indel and VNTR loci are predicted to produce reading frame shifts, nine resulting in the introduction of a premature stop codon (Table 6, which is published as supporting information on the PNAS web site). Eight more genes had SNPs that produce nonsense codons also resulting in truncated inferred proteins (Table 7, which is published as supporting information on the PNAS web site). Space constraints do not permit detailed treatment of all 17 truncated proteins, so we will focus on proteins known to have a role in host–pathogen interaction. This limitation should not be interpreted to mean that the other inferred truncated proteins are irrelevant.

Invasive and Carrier Relationships and Polymorphism Associations. Within the invasive and carrier groups, strains differed on average by 50 and 52 core SNPs, respectively, and between these two groups, strains differed on average by 49 SNPs. Hence, neither group was composed of strains that were excessively differentiated from one another, indicating relatively recent evolution from a common ancestor. Additionally, the carrier and invasive strains were interspersed throughout the M3 phylogenetic network (Fig. 4*A*), indicating that they do not represent two deeply differentiated genetic groups. Importantly, no SNP was uniformly present in one phenotypic group or the other. These findings strongly suggest that the carrier and invasive phenotypes can arise by multiple independent evolutionary paths.

Notably, only 25 and 17 core SNPs differentiate carrier strains MGAS12501 and MGAS12504, respectively, from highly virulent reference strain MGAS315. These two carrier strains had nearly identical LD_{90s} and were far less virulent (≈20-fold higher LD_{90s}) than strain MGAS315 (Table 4). This modest number of SNPs suggests that distinct phenotypes can be due to relatively few molecular events, perhaps a single mutation.

Strain MGAS12501 and MGAS12504 each has the same 12-bp deletion in the promoter region of the gene (*mga*) encoding a pleiotropic virulence regulator (Mga, multi-gene activator) (Fig. 8, which is published as supporting information on the PNAS web site). This deletion is located ≈20-bp downstream of the primary *mga* transcription start site (P2) and ≈60-bp upstream of the start of translation. It is possible that the 12-bp deletion alters *mga* expression by affecting regulatory factor binding and/or local secondary structure. Inasmuch as Mga directly regulates expression of M protein, a major GAS surface antigen and virulence factor, the 12-bp deletion could significantly alter host–pathogen interactions. Of interest, the *emm* gene in carrier strain MGAS12502 has a 195-bp deletion that completely removes the hypervariable N terminus of M protein (Fig. 8). This 65-aa in-frame deletion abuts the secretion signal sequence and therefore is not predicted to alter M protein expression. Analysis of the first 50 amino acids of the resultant M protein did not identify a homologue when searched against the CDC *emm* typing database (www.cdc.gov/ncidod/biotech/strep/strepblast.htm). Thus, three of four carrier strains have polymorphisms likely to influence M protein expression or function, suggesting a role in the carrier phenotype.

Genetic Relationships Between Serotype M3 Strains, Epidemic Behavior, and Disease Associations. Among the eight invasive isolates cultured from patients in Ontario, 327 core SNPs at 177 loci were identified and used to infer genetic relationships (Fig. 4*B*). The topology of this phylogenetic network is fully concordant with subclone designations and genetic relationships previously proposed for these strains (4). This finding is pivotal as these relationships were used to statistically link distinct M3 subclones with epidemic behavior and infection character (4).

We previously hypothesized (4) that M3 strains with the prophage that encodes SpeA (Φ315.5) were more fit than strains lacking this prophage. The hypothesis was based on the observation that strains lacking the prophage (subclones 2 and 4) failed to persist in the interepidemic period investigated. In addition, strains containing *speA* were significantly more likely to cause recurrent pharyngitis episodes than strains lacking *speA* (23). In the mouse model, as a group the *speA*$^+$ invasive isolates were more virulent than the *speA*$^-$, but the difference was not significant, possibly because of the small numbers of strains compared (Fig. 2). Importantly, no SNPs were identified by CGR that were uniquely present in subclone 2 and 4 strains (persistence failures) and absent in subclone 1 and 3 strains (successful persisters). Thus, the only genetic trait that consistently differentiated the subclones that persisted from those that did not was the SpeA-encoding prophage. This fact provides strong support to our hypothesis that SpeA enhances GAS persistence in natural populations.

We recently hypothesized that subclone 5 strains evolved from a subclone 1 precursor and rose to prominence in the second epidemic wave (centered in 2000) of infections as a consequence of host selective pressure. This hypothesis was based on the finding that a 4-aa duplication at the extreme N terminus of the mature M protein was the only genetic difference we found between subclone 5 (*emm3.2* allele) and subclone 1 strains (*emm3.1* allele). This duplication altered immune recognition of the extreme N terminus of the M protein based on ELISA and phagocytosis assays. Phylogenetic reconstruction using the whole-genome resequencing data strongly supports the proposed descent of subclone 5 strains from a subclone 1 ancestor (Fig. 4*B*). CGR identified only 28 core SNPs present in representative subclone 5 strain MGAS9887 and not in the other invasive isolates. Most of these SNPs (n = 25) were

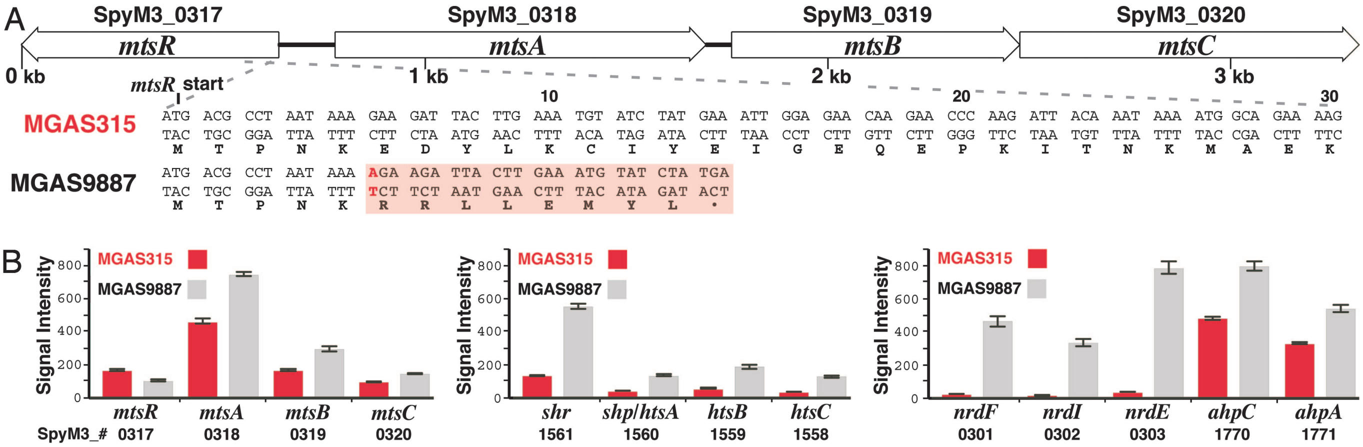


Fig. 5. Mts regulon and genes exhibiting altered expression. (*A*) Illustrated is the region of the M3 genome encoding the Mts regulon. Boxed in red is the frame-shifted translation product. (*B*) Iron acquisition and oxidative stress response genes exhibiting altered expression. Expression of each of these genes was significantly different ($P < 0.0001$) under conditions of exponential aerobic growth in metal replete rich media. Bars are the standard error of the mean ($n = 6$). In strain MGAS9887, expression of *mtsR* was decreased, whereas *mtsABC* was increased ≈1.5 fold. Expression of *shr* (streptococcal heme receptor) and *htsABC* was increased ≈3.5-fold. Expression of ribonucleotide-diphosphate reductase, *nrdFIE*, and alkyl hydroperoxide reductase, *ahpCA*, was increased ≈20- and ≈1.5-fold, respectively. Findings are consistent with the loss of MtsR repression of iron acquisition systems leading to intracellular iron accumulation, increased formation of reactive oxygen species, and induction of an oxidative stress response.

either intergenic or synonymous nucleotide changes, or nonsynonymous mutations that produced conservative amino acid replacements. In addition, no SNPs predicted to produce an amino acid replacement in an extracellular protein were identified. Together with previous findings, these results further support the hypothesis that reduced host immune recognition of the Emm3.2 variant was a fundamental contributor to the emergence and epidemic expansion of subclone 5 strains.

A critical finding from our previous work was that subclone 5 strains differed in infection character from the other invasive subclones in that they were significantly underrepresented in necrotizing fasciitis cases ($P = 0.0014$). The CGR data obtained for the invasive isolates permitted us to search for a genetic basis for the strain MGAS9887 difference in infection character. We discovered that relative to the other invasive isolates studied, strain MGAS9887 has a 1-bp insertion (T insertion in a run of three Ts) that results in a frameshift in *mtsR* (metal transporter of *Streptococcus*), a member of the metal-dependent DtxR family of transcription regulators (Table 6). The insertion is located 13 bp downstream of the *mtsR* translation start codon and introduces a stop codon after the first 13 aa (Fig. 5*A*).

The *mtsR* gene is located upstream of *mtsABC* and divergently transcribed. Moreover, an *mtsR* mutant derivative of serotype M49 strain NZ131 accumulated iron intracellularly, was more sensitive to oxidative stress, and was less virulent in zebrafish infection models (24). Importantly, we found that considerably more mice injected i.p. with MGAS9887 survived (31 of 50 at 120 h) than with reference strain MGAS315 (7 of 50). Interpretation of the contribution of MtsR in virulence is complicated by the redundancy of GAS iron acquisition systems and metal homeostasis regulators. Current understanding posits that this redundancy permits pathogenic bacteria to use different sources of iron in different host niches (25). It has been suggested that because of the different sources of iron (ferrichrome, free Fe ions, or heme) used by the GAS transporters, they may differ in importance to pathogenesis depending on the iron source available at the site of infection (26). In addition, MtsR might directly regulate the expression of virulence factors not involved in metal homeostasis, similar to DtxR regulation of diphtheria toxin production in *Clostridium diphtheriae*.

Expression Microarray Analysis of Subclone 5 Strain MGAS9887. Inasmuch as MtsR homologues act as repressors of metal uptake in other bacterial pathogens (27, 28), we hypothesized that expression of metal homeostasis genes regulated by MtsR would be altered in strain MGAS9887. To test this, we compared the transcriptomes of strains MGAS9887 and MGAS315 during exponential growth in THY broth. Consistent with the hypothesis, and the findings of Bates (24), we observed significant ($P < 0.001$) up-regulation of the *htsABC* operon (heme transporter of *Streptococcus*) in strain MGAS9887. Furthermore, transcript levels of *mtsABC* were significantly higher and *mtsR* lower ($P < 0.001$) in strain MGAS9887 than strain MGAS315 (Fig. 5*B*). This finding suggests that, in metal- and nutrient-rich media, MtsR enhances its own expression and represses that of *mtsABC*. Also significantly up-regulated were genes encoding an alkyl hydroperoxide reductase (*ahpCA*) and ribonucleotide diphosphate reductase (*nrdEIF*), both found to be up-regulated in response to oxidative stress in other bacterial pathogens (29, 30). Of note, the growth curves of strain MGAS9887 and MGAS315 were identical in THY broth (Fig. 9, which is published as supporting information on the PNAS web site), arguing that the transcriptome differences do not simply result from a difference in growth rate. Hence, the truncation mutation in *mtsR* was linked to significantly altered transcript levels of multiple operons involved with metal homeostasis and response to oxidative stress.

Conclusion. Epidemiologic investigations have repeatedly found nonrandom GAS serotype–disease type associations, but the molecular basis of these associations remains obscure. A clinically and genetically well defined cohort is critical for assessing strain genotype-disease type relationships because strains of the same serotype and MLST can differ extensively in exogenous virulence gene content. Here we have probed the molecular genetic anatomy of strain genotype–patient phenotype relationships at the nucleotide level in serotype M3 GAS, a model bacterial pathogen system. Our studies were facilitated by the availability of clinically very well characterized strain samples, the recent advent of a relatively inexpensive CGR technique, and expression microarray technology. Use of closely related strains of defined gene content that differ in clinical phenotype has greatly reduced the complexity of the analysis and has revealed specific naturally occurring genetic polymorphisms likely to contribute to the observed virulence differences. Taken together, we believe the data represent an important step toward personalized infectious disease research, a discipline that will become dominant in the next decade.

Materials and Methods

For further details, see *Supporting Text*, which is published as supporting information on the PNAS web site.

Bacterial Strains. The GAS strains used in this study are listed in Tables 1 and 2. The sequence of serotype M5 strain Manfredo can be obtained at www.sanger.ac.uk/Projects/S_pyogenes. Sequencing of the genome of one strain each of serotypes M2 and M4 and two strains of serotype M12 was accomplished by using methods described in refs. 13, 15, and 16. These four genomes were sequenced to closure, and an average base call error rate of <1 in 10,000 (Q40) throughout. Four strains, MGAS12501–MGAS12504, are asymptomatic carriage isolates recovered in a population-based carrier study of recruits at Lackland Air Force Base (San Antonio, TX) (31). An additional eight strains were recovered from January 1, 1992 to December 31, 2003 in an ongoing prospective population-based surveillance study of GAS invasive infections in Ontario, Canada (http://microbiology.mtsinai.on.ca/research/gas.shtml). These eight strains represent the major genetic subclones that caused two epidemic peaks of GAS serotype M3 invasive infections (4).

Bacterial Growth. Bacteria were grown at 37°C in atmosphere supplemented with 5% CO_2 and plated on tryptic soy agar supplemented with 5% sheep blood (TSAB). Broth cultures were grown statically in Todd Hewitt medium supplemented with 0.2% yeast extract (THY). Growth was monitored by measuring optical density at 600 nm.

Virulence Assessment. Virulence was assessed by using a mouse model of GAS systemic infection. Mean survival times and LD_{90} were determined by Kaplan–Meier and Probit analysis using PRISM (GraphPad, San Diego) and XLSTAT-DOSE (XLSTAT, New York), respectively.

Intraserotype M3 Polymorphism Mapping and CGR. Sequence polymorphisms were identified relative to reference strain MGAS315 by using a DNA–DNA hybridization microarray method (32). Polymorphism mapping microarrays were generated (NimbleGen Systems, Madison, WI) by using 29-mer oligonucleotide probes spaced every seven bases around the reference MGAS315 genome on both strands.

Identification of SNPs Among the Sequenced GAS Genomes. SNPs among the sequenced GAS strains were identified by using MUMMER 3.18 using the suggested alignment pipeline (www.tigr.org) (17). Potentially erroneous SNPs contained in repeated sequences in the genomes were excluded by using the "show-snps − C" option.

Verification of Polymorphisms Identified by CGR. PCR amplification and conventional DNA sequencing was used to assess the validity of polymorphisms identified by DNA–DNA microarray CGR. Nucleotide sequences obtained were aligned with the MGAS315 reference genome by using SEQUENCHER (GeneCodes, Ann Arbor, MI), and all polymorphisms were assessed by manual inspection.

Phylogenetic Reconstruction. Evolutionary relationships were reconstructed by using SPLITSTREE4 (www.splitstree.org) (33). Strain sequences analyzed were generated by concatenating nucleotides at all of the core chromosomal SNP loci identified in sequential order relative to the reference MGAS315 genome. Distances between strains were estimated by using the uncorrected P method that calculates the proportion of positions at which two sequences differ. Phylogenetic networks were computed by the split-decomposition method that does not force conflicting data into a bifurcating tree representative of only a subset of the total possible evolutionary paths.

Expression Microarray Analysis. A custom GeneChip (Affymetrix Santa Clara, CA) was used for expression microarray studies, as described (34). Image files were processed and expression data were analyzed by using DCHIP (www.dchip.org). RNA isolated from six replicates of strain MGAS9887 and MGAS315 was used. The concentration and quality of RNA were assessed with a 2100 Bioanalyzer (Agilent, Palo Alto, CA) and analysis of the A_{260}/A_{280} ratio.

1. Carapetis, J. R., Steer, A. C., Mulholland, E. K. & Weber, M. (2005) *Lancet Infect. Dis.* **5,** 685–694.
2. Musser, J. M. & Krause, R. M. (1998) in *Emerging Infections*, ed. Krause, R. M. (Academic, New York), Vol. 1, pp. 185–218.
3. Beres, S. B., Sylva, G. L., Barbian, K. D., Lei, B., Hoff, J. S., Mammarella, N. D., Liu, M. Y., Smoot, J. C., Porcella, S. F., Parkins, L. D., *et al*. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 10078–10083.
4. Beres, S. B., Sylva, G. L., Sturdevant, D. E., Granville, C. N., Liu, M., Ricklefs, S. M., Whitney, A. R., Parkins, L. D., Hoe, N. P., Adams, *et al*. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 11833–11838.
5. Davies, H. D., McGeer, A., Schwartz, B., Green, K., Cann, D., Simor, A. E. & Low, D. E. (1996) *N. Engl. J. Med.* **335,** 547–554.
6. Kaul, R., McGeer, A., Low, D. E., Green, K. & Schwartz, B. (1997) *Am. J. Med.* **103,** 18–24.
7. Li, Z., Sakota, V., Jackson, D., Franklin, A. R. & Beall, B. (2003) *J. Infect. Dis.* **188,** 1587–1592.
8. O'Brien, K. L., Beall, B., Barrett, N. L., Cieslak, P. R., Reingold, A., Farley, M. M., Danila, R., Zell, E. R., Facklam, R., Schwartz, B. & Schuchat, A. (2002) *Clin. Infect. Dis.* **35,** 268–276.
9. Sharkawy, A., Low, D. E., Saginur, R., Gregson, D., Schwartz, B., Jessamine, P., Green, K. & McGeer, A. (2002) *Clin. Infect. Dis.* **34,** 454–460.
10. Shulman, S. T., Tanz, R. R., Kabat, W., Kabat, K., Cederlund, E., Patel, D., Li, Z., Sakota, V., Dale, J. B. & Beall, B. (2004) *Clin. Infect. Dis.* **39,** 325–332.
11. Feretti, J. J., McShan, W. M., Ajdic, D., Savic, D. J., Savic, G., Lyon, K., Primeaux, C., Sezate, S., Suvorov, A. N., Kenton, S., *et al*. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 4658–4663.
12. Smoot, J. C., Barbian, K. D., Van Gompel, J. J., Smoot, L. M., Chaussee, M. S., Sylva, G. L., Sturdevant, D. E., Ricklefs, S. M., Porcella, S. F., Parkins, L. D., *et al*. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 4668–4673.
13. Banks, D. J., Porcella, S. F., Barbian, K. D., Beres, S. B., Philips, L. E., Voyich, J. M., DeLeo, F. R., Martin, J. M., Somerville, G. A. & Musser, J. M. (2004) *J. Infect. Dis.* **190,** 727–738.
14. Nakagawa, I., Kurokawa, K., Yamashita, A., Nakata, M., Tomiyasu, Y., Okahashi, N., Kawabata, S., Yamazaki, K., Shiba, T., Yasunaga, T., *et al*. (2003) *Genome Res.* **13,** 1042–1055.
15. Green, N. M., Zhang, S., Porcella, S. F., Nagiec, M. J., Barbian, K. D., Beres, S. B., LeFebvre, R. B. & Musser, J. M. (2005) *J. Infect. Dis.* **192,** 760–770.
16. Sumby, P., Porcella, S. F., Madrigal, A. G., Barbian, K. D., Virtaneva, K., Ricklefs, S. M., Sturdevant, D. E., Graham, M. R., Vuopio-Varkila, J., Hoe, N. P. & Musser, J. M. (2005) *J. Infect. Dis.* **192,** 771–782.
17. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004) *Genome Biol.* **5,** R12.
18. Miller, R. D., Phillips, M. S., Jo, I., Donaldson, M. A., Studebaker, J. F., Addleman, N., Alfisi, S. V., Ankener, W. M., Bhatti, H. A., Callahan, C. E., *et al*. (2005) *Genomics* **86,** 117–126.
19. Breese, B. B., Disney, F. A., Talpey, W. B. & Green, J. L. (1970) *Am. J. Dis. Child.* **119,** 18–26.
20. Hamburger, M., Green, M. J. & Hamburger, V. G. (1945) *J. Infect. Dis.* **77,** 96–108.
21. Wannamaker, L. W. (1954) in *Streptococcal Infections*, ed. McCarty, M. (Columbia Univ. Press, New York).
22. Albert, T. J., Dailidiene, D., Dailide, G., Norton, J. E., Kalia, A., Richmond, T. A., Molla, M., Singh, J., Green, R. D. & Berg, D. E. (2005) *Nat. Methods* **2,** 951–953.
23. Musser, J. M., Gray, B. M., Schlievert, P. M. & Pichichero, M. E. (1992) *J. Clin. Microbiol.* **30,** 600–603.
24. Bates, C. S., Toukoki, C., Neely, M. N. & Eichenbaum, Z. (2005) *Infect. Immun.* **73,** 5743–5753.
25. Garmory, H. S. & Titball, R. W. (2004) *Infect. Immun.* **72,** 6757–6763.
26. Hanks, T. S., Liu, M., McClure, M. J. & Lei, B. (2005) *BMC Microbiol.* **5,** 62.
27. Jakubovics, N. S., Smith, A. W. & Jenkinson, H. F. (2000) *Mol. Microbiol.* **38,** 140–153.
28. Ando, M., Manabe, Y. C., Converse, P. J., Miyazaki, E., Harrison, R., Murphy, J. R. & Bishai, W. R. (2003) *Infect. Immun.* **71,** 2584–2590.
29. Diaz, P. I., Zilm, P. S., Wasinger, V., Corthals, G. L. & Rogers, A. H. (2004) *Oral Microbiol. Immunol.* **19,** 137–143.
30. Monje-Casas, F., Jurado, J., Prieto-Alamo, M. J., Holmgren, A. & Pueyo, C. (2001) *J. Biol. Chem.* **276,** 18031–18037.
31. Hoe, N. P., Fullerton, K. E., Liu, M., Peters, J. E., Gackstetter, G. D., Adams, G. J. & Musser, J. M. (2003) *J. Infect. Dis.* **188,** 818–827.
32. Albert, T. J., Norton, J., Ott, M., Richmond, T., Nuwaysir, K., Nuwaysir, E. F., Stengele, K. P. & Green, R. D. (2003) *Nucleic Acids Res.* **31,** e35.
33. Huson, D. H. & Bryant, D. (2005) *Mol. Biol. Evol.* **23,** 254–267.
34. Graham, M. R., Virtaneva, K., Porcella, S. F., Barry, W. T., Gowen, B. B., Johnson, C. R., Wright, F. A. & Musser, J. M. (2005) *Am. J. Pathol.* **166,** 455–465.