Deriving Group A *Streptococcus* Typing Information from Short-Read Whole Genome Sequencing Data

Taryn B.T. Athey\(^1\), Sarah Teatero\(^1\), Aimin Li\(^1\), Alex Marchand-Austin\(^{1,2}\), Bernard W. Beall\(^3\) and Nahuel Fittipaldi\(^{1,2}\)#

Public Health Ontario, Toronto, Ontario, Canada\(^1\), Department of Laboratory Medicine and Pathobiology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada\(^2\), Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA\(^3\).

**Running title:** Deriving GAS typing information from WGS data.

**Keywords:** Group A *Streptococcus*, epidemiology, bacterial typing, *emm* typing, MLST, whole-genome sequencing.

**Author for correspondence:** Nahuel Fittipaldi, PhD. Public Health Ontario, Toronto Laboratories, 81 Resources Rd, Toronto, Ontario M9P 3T1, Canada. Phone: (416) 235-5714. Fax: (416) 235-6063. E-mail: nahuel.fittipaldi@oahpp.ca.
Abstract

Typing of Group A Streptococcus is crucial for infection control and epidemiology. While whole-genome sequencing (WGS) is revolutionizing the way bacterial organisms are typed, it is necessary to provide back-compatibility with currently used typing schemas to facilitate comparison and understanding of epidemiological trends. Here, we used bioinformatics tools to derive commonly used GAS typing information directly from short-read WGS data of 191 GAS isolates representing 42 different emm types. We show that acquisition of emm typing and multilocus sequence typing, as well as presence or absence of genes associated with GAS tissue tropism can be achieved rapidly and efficiently using this approach. We also report on how WGS data analysis was instrumental in identifying ambiguities present in the commonly used emm-type database hosted by the United States Centers for Disease Control and Prevention.
Introduction

Group A Streptococcus (GAS, also known as Streptococcus pyogenes) is a human-specific pathogen that causes diseases ranging in severity from uncomplicated pharyngitis to life-threatening necrotizing fasciitis (1). Until relatively recently, strains of GAS were typed based on a serological reaction against M protein, a polymorphic cell-surface adhesin and anti-phagocytic factor encoded by gene emm (2-4). Sequencing of a short, hypervariable region at the 5' end of emm has superseded serological methods and has become the de facto and most used GAS typing method (5-7). Currently, the sequences of more than 200 distinct emm types, defined simply as the sequence type derived from an amplicon generated from a specific primer pair (5-7), are listed in a database curated by the United States Centers for Disease Control and Prevention (CDC).

It has long been described that, in addition to gene emm, some GAS strains may possess up to two additional emm-like genes, also known as mrp and enn, which encode M-like proteins designated Mrp and Enn, respectively (8, 9). M and M-like proteins are belong to a family of cell-wall associated, structurally related proteins that have affinity for several plasma proteins, including fibrinogen, immunoglobulins G and A, and the complement regulatory proteins. It has been suggested that the antiphagocytic properties conferred by M and M-like proteins result from binding of these plasma proteins (10-12). In all GAS examined so far, emm and emm-like genes are found in a discrete GAS chromosomal region which is under direct transcriptional control of the standalone virulence regulator Mga (13). The first gene of this discrete chromosomal region is mga, which encodes and is itself regulated by Mga; the last gene is scpA, encoding a cell wall-associated C5a peptidase (13).

In GAS types with only emm, this gene is found immediately downstream of mga. In strains with both emm and emm-like genes, mrp is found downstream of mga, followed by emm and then by enn (13, 14). Other more complex chromosomal arrangements have also been described (14). It has been suspected that at least one of...
the sequences present in the CDC emm database may correspond to mrp or enn in a minority of strains (15).

Together with the presence and/or absence of other markers such as nra and rofA, which encode regulators of pili expression, pili genes themselves, and sof, encoding a serum opacity factor, emm chromosomal arrangements have been used as putative predictors of host-tissue tropism for GAS strains (14, 16, 17).

Multilocus sequence typing schemes have also been developed and used to type GAS isolates (18, 19).

Recent advances in molecular microbial characterization by whole genome analysis are opening up tremendous new opportunities for a better understanding of the pathogenicity, evolution, and spread of human pathogens and the epidemiology of the diseases they cause (20, 21). Whole-genome sequencing holds the promise of improving the resolution and predictive value of microbial typing as applied to public health objectives such as disease surveillance and epidemic investigation, as well as in hospital laboratories (21-26).

Despite these exciting opportunities, a number of key hurdles for routine implementation of WGS for public health purposes and identification and typing of infectious agents remain unresolved (22, 24). In addition, while typing bacterial organisms using the WGS data may be considered the ultimate typing method, there is need to provide back-compatibility with currently used typing schemas to facilitate comparison and enhanced understanding of epidemiological trends.

Following the sequencing of 191 genomes of GAS strains representative of all 42 emm-types isolated in Ontario during the last 3 years, we show here that using appropriate bioinformatics tools it is possible to accurately and rapidly derive currently used typing information directly from the short-read WGS data. We also report and discuss on how WGS data was instrumental in identifying ambiguities present in the commonly used CDC emm type database.
Materials and Methods

Strains, culture conditions and DNA preparation. A convenience sample of 191 strains representing all 42 GAS emm types isolated province-wide in Ontario from 2010-2013 were used (Table S1). The emm types of these strains were determined through traditional Sanger sequencing using previously described primers and conditions (15). Samples were selected so that all 42 emm types found in Ontario during the specified time period had at least one representative strain included in the collection. In most cases, the number of replicates represented the proportion of isolates of each of these 42 emm types. Strains were cultured on Columbia blood agar plates containing 5% sheep blood, and grown at 37°C with 5% CO₂. Liquid cultures were grown in Todd-Hewitt broth supplemented with 0.2% yeast extract. DNA was prepared from overnight GAS cultures using the QIAamp DNA minikit (Qiagen, Toronto, ON, Canada) following the manufacturers’ protocol for Gram positive organisms.

Whole-genome sequencing and data analysis. Genomic libraries were prepared using Nextera XT kits (Illumina, San Diego CA) and sequenced as paired-end (101 bp) in a HiSeq 2500 instrument (Illumina). Parsing of the multiplexed sequencing reads and removal of barcode information was done using onboard software. The A5 pipeline was used for de novo assembly of newly sequenced GAS strains (27). The trimmed and untrimmed emm type reference databases were downloaded directly from the CDC ftp site at ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsem and ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/emmsequ, respectively. The GAS MLST database was downloaded from the Imperial College at http://spyogenes.mlst.net/. MLST types were determined directly from the short read WGS data using SRST2 (https://github.com/katholt/srst2) (28), with some modifications. Briefly, we modified the SRST2 original code to use Mosaik Aligner
Deriving emm typing information from short-read WGS. We sequenced the genomes of 191 GAS strains representing 42 different emm types using a HiSeq 2500 instrument. The average number of 101 bp reads per strain was 2,379,140 (maximum 5,147,286, minimum 435,775) and the coverage considering an average GAS genome size of 1.8 Mbp was 267X, on average (maximum 578X, minimum 49X) (Table S1). WGS data for the 191 isolates has been submitted to the SRA archive under accession number SRP035244.

SRST and its successor SRST2 are software originally conceived as a means to derive MLST typing information from bacterial short-read WGS data (28). However, since SRST and SRST2 are database driven, they can be used for other WGS-based typing tasks beyond MLST, such as finding drug resistance genes and determining virulence gene alleles (provided that appropriate databases are used as input) (28). Here, we used a modified SRST2 and the CDC emm database to derive emm type information from the short-read WGS data. Using this...
approach, we were able to match 177 of the 191 GAS strains (92.7 %) to the \textit{emm} type that had been previously determined using traditional, Sanger sequencing-based \textit{emm} typing. Notably, all strains belonging to \textit{emm} types known or found not to contain \textit{emm}-like genes were assigned an \textit{emm} type which matched the one determined using Sanger sequencing-based \textit{emm} typing (Fig. 1A). On the other hand, many, but not all, strains belonging to \textit{emm} types known to possess \textit{emm}-like genes had a match between Sanger sequencing-based and WGS-based \textit{emm} typing. Indeed, we noted several instances of discrepancies between the two methods for strains possessing \textit{emm}-like genes. To resolve these discrepancies, we manually inspected the SRST2 output and discovered that every time a mismatch between Sanger sequencing-based and WGS-based typing was recorded, SRST2 had given very high scores to two or three \textit{emm} types found on the CDC’s \textit{emm} database (data not shown). Remarkably, closer examination after \textit{de novo} assembly of the Illumina short-reads into contigs and BLAST analysis revealed that all those highly scored “\textit{emm}” sequences were indeed present in the genomes of the strains under investigation and they corresponded, one to the legitimate \textit{emm} gene and the other(s) to \textit{emm}-like genes (Fig. 1B). These findings strongly suggest that in a context of an ambiguous database such appears to be the current version of the CDC \textit{emm} database, a mapping-based-only strategy such as the one offered by SRST2 is not sufficient to confidently assign \textit{emm} types from WGS data.

The CDC \textit{emm} typing database contains sequences found in \textit{emm}-like genes that complicate WGS-based \textit{emm}-typing. Most \textit{emm} designations and all recent \textit{emm} sequence designations have relied solely on sequencing results of amplicons generated by the use of oligonucleotide primers 1 and 2, specific for what has been thought to be the \textit{emm} gene and not other \textit{emm}-like genes (5-7, 31). Since this definition of \textit{emm} type is broad, it may be difficult to establish whether an “\textit{emm} type” found in the CDC database corresponds to sequences found in a legitimate \textit{emm} gene. Instead, some “\textit{emm} types” may correspond to \textit{emm}-like genes that were somehow amplified using the \textit{emm} typing primers. Indeed, we identified here several examples of
these emm-like sequences in the CDC emm database (Fig. 1B). The presence of these emm-like sequences complicates automated mapped-based emm-typing from short-read WGS because SRST2 actually finds reads aligning to these sequences in the WGS data of the strains under investigation and ranks them highly. When we did not consider these emm types matching emm-like sequences (we removed them from our working database), SRST2 was able to assign all strains to the emm type previously determined using Sanger sequencing. Moreover, one strain, NGAS320, which was originally recorded as non-typable by Sanger sequencing because an amplicon could not be generated using emm typing primers 1 and 2, was identified as an emm83 by WGS-based emm typing. Upon further investigation of the emm chromosomal arrangement of strain NGAS320 following de novo assembly of the WGS data, and also by PCR using primers annealing to mga and scpA, we discovered that this strain has a deletion of 1390 bp resulting in a hybrid emm gene and loss of the annealing site for primer 1 used in Sanger-based emm typing (Fig. 2A).

The workaround we describe above (i.e., removing those emm types with sequences matching emm-like genes from the input database that SRST2 uses) is very difficult to implement, because it is not trivial to determine a priori what emm-like sequences may need to be removed from the CDC database. Firstly, it is unknown how many more sequences matching emm-like genes are found in the database and, secondly and perhaps more importantly, even if these sequences were easier to identify, it is possible that they are novel, legitimate emm alleles containing sequences from emm-like genes which arose following genetic rearrangements in a region of the genome which has been described to be prone to recombination (9, 32, 33). As an example, we discovered that one of our strains, NGAS128, originally typed as emm14 using Sanger sequencing, had also short-read sequences corresponding to allele emmS1 (GenBank accession number U11977) (6), a well-documented serotype, that mapped to the enn gene (Fig. 2B). These findings replicate previous reports by Dowson et al. who
found that the 5'-end of emm51 was identical to sequences found in enn14 and in enn46 (34). Interestingly, emm51 strains are rarely described in surveillance studies.

A pipeline to confidently call emm type from short-read WGS in GAS strains with both emm and one or more emm-like genes. As shown above, it is possible to use de novo assembly to generate contigs for all GAS strains from the short-read WGS, and use BLAST analysis to identify the correct emm type of a particular strain.

However, this approach is time consuming and sometimes superfluous, as for many GAS emm types, a mapping-based approach using SRST2 is sufficient to accurately derive emm typing information. We thus developed a pipeline which uses SRST2 to initially map the short reads to the CDC emm database, and then automatically assess the SRST2-assigned scores from the initial mapping. Then, if a GAS strain has more than one SRST2 alignment to emm alleles present in the CDC emm database which show score values below a predetermined threshold (set to 10 in this study; the threshold can be changed manually within the pipeline script available in the supplementary materials), the strain is flagged for further inspection by automatic de novo assembly and BLAST processing as described in the supplementary methods. When this approach was attempted on the current dataset, 77 of the 191 GAS strains were selected for de novo assembly and further inspection. In 11 of these strains, all of the top scoring alleles had BLAST hits to the legitimate emm gene. BLAST scores were used to assign the most probable emm type which in all cases coincided with the one determined by Sanger sequencing, with the exception of the non-typable strain mentioned above (data not shown). In the remaining 66 GAS strains, at least two of the top SRST2-scored alleles had BLAST hits to two different regions of the emm region. In 4 of those 66 strains we discovered that in addition to a highly ranked BLAST hit to the emm gene, other two BLAST hits were to two different emm-like genes (Table 1). The remaining 62 GAS strains had BLAST hits to the emm gene and only one emm-like gene (Table 1). Using this pipeline we observed a 100 % agreement between Sanger sequencing-based and WGS-based emm typing, with the exception of the
The aforementioned strain NGAS320, for which a PCR amplicon could not be obtained and therefore, was non-typable by Sanger sequencing-based emm typing. Thus, the use of our pipeline permitted the accurate determination of emm type in all GAS strains, including those whose genomes contained both emm and emm-like genes.

**Deriving other commonly used GAS typing information from the short-read WGS.** As mentioned above, SRST2 was originally described to derive MLST typing information directly from the short-read WGS (28). Here, we were able to determine allele type and ST for most of the 191 GAS strains in our collection using SRST2 (Table S1). Although some of the STs were not found in the MLST database and may represent novel variants, there were some occasions in which results obtained by SRST2 may need to be confirmed by traditional amplification and Sanger sequencing of the alleles. SRST2 distinguishes those potentially ambiguous results with a question mark (?; Table S1). Overall, and despite the presence of these ambiguities, our data showed a strong correlation between emm types and MLST STs for the vast majority of the isolates (Table S1), which probably is a reflection of the clonal nature of the GAS strains circulating in Ontario. We also used SRST2 to determine the presence or absence in our strains of markers associated with host-tissue tropism in GAS (14, 16, 17), including genes sof, nra and rofA (Table S1). We identified a few discrepancies between our results and previously found patterns of genes in GAS strains (16). Namely, some of the GAS emm types (emm73 and emm105) described in previous reports (16) to contain gene nra were found in this study to instead contain gene rofA, while type emm29 strains were found here to contain gene nra and not gene rofA as had been reported previously (16). Finally we used SRST2 to reveal whether the GAS strains possessed the mga1 or mga2 alleles of the mga gene directly from the short-read WGS data (Table S1).
Typing of bacterial organisms based on the full-extent genome sequence is becoming an increasingly feasible option for diagnostic and public health laboratories (21-24). While we advance towards generalized WGS-based typing of GAS, it is important to rely on validated tools to rapidly and efficiently provide back-compatibility with currently used typing schemas. Our results demonstrate that deriving this typing information from the short-read WGS data is achievable, but they also highlight that high-quality, well-curated databases are crucial to fully take advantage of WGS data.

Acknowledgements

We thank Jonas Winchell (Centers for Disease Control and Prevention) for critical reading of an earlier version of the manuscript. We also thank Irene Martin (National Microbiology Laboratory, Winnipeg) for the original Sanger-based emm typing of the GAS strains used in this study. We are grateful to Prasad Rawte (Public Health Ontario, Toronto) for his help with strain identification. We used the GAS MLST database which is hosted at the Imperial College and curated by D. Bessen and the CDC emm database curated by Velusamy Srinivasan. This work was funded by Public Health Ontario.

References


Table 1: BLAST hits of sequences present in the CDC \textit{emm} database against \textit{emm} and \textit{emm}-like genes present in the \textit{de novo} assemblies of 77 GAS strains chosen by our pipeline for further inspection.

<table>
<thead>
<tr>
<th>Sanger sequencing-determined \textit{emm} type</th>
<th>CDC database allele blasting to MRP gene\textsuperscript{a}</th>
<th>CDC database allele blasting to \textit{emm} gene\textsuperscript{a}</th>
<th>CDC database allele blasting to \textit{enn} gene\textsuperscript{a}</th>
<th>Number of strains with these BLAST results</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-</td>
<td>2</td>
<td>149</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>3\textsuperscript{b}</td>
<td>-</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>156</td>
<td>4</td>
<td>236</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>156</td>
<td>4</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>5\textsuperscript{b}</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>9</td>
<td>236</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>11</td>
<td>202</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>14</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>18</td>
<td>205</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>156</td>
<td>22</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>59</td>
<td>-</td>
<td>59</td>
<td>174</td>
<td>3</td>
</tr>
<tr>
<td>73</td>
<td>-</td>
<td>73</td>
<td>149</td>
<td>1</td>
</tr>
<tr>
<td>73\textsuperscript{b}</td>
<td>-</td>
<td>73</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>-</td>
<td>75</td>
<td>170</td>
<td>2</td>
</tr>
<tr>
<td>77</td>
<td>-</td>
<td>77</td>
<td>159</td>
<td>2</td>
</tr>
<tr>
<td>83</td>
<td>-</td>
<td>83</td>
<td>240</td>
<td>10</td>
</tr>
<tr>
<td>83\textsuperscript{b}</td>
<td>-</td>
<td>83</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>87</td>
<td>-</td>
<td>87</td>
<td>159</td>
<td>3</td>
</tr>
<tr>
<td>89</td>
<td>-</td>
<td>89</td>
<td>236</td>
<td>11</td>
</tr>
<tr>
<td>101</td>
<td>-</td>
<td>101</td>
<td>205</td>
<td>2</td>
</tr>
<tr>
<td>105</td>
<td>156</td>
<td>105</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>114</td>
<td>-</td>
<td>114</td>
<td>159</td>
<td>4</td>
</tr>
<tr>
<td>118</td>
<td>-</td>
<td>118</td>
<td>236</td>
<td>2</td>
</tr>
<tr>
<td>122</td>
<td>-</td>
<td>122</td>
<td>240</td>
<td>1</td>
</tr>
<tr>
<td>169</td>
<td>-</td>
<td>169</td>
<td>164</td>
<td>2</td>
</tr>
<tr>
<td>192</td>
<td>-</td>
<td>192</td>
<td>138</td>
<td>1</td>
</tr>
<tr>
<td>NT</td>
<td>-</td>
<td>83</td>
<td>240</td>
<td>1</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Only the CDC database \textit{emm} allele with the best BLAST score for each gene is shown.

\textsuperscript{b}All of the SRST2 top-scoring CDC database \textit{emm} alleles had BLAST hits to the \textit{emm} gene in strains of these \textit{emm} types. Only the highest-scored BLAST hit is presented.
Fig. 1. Schematics showing emm chromosomal arrangements in different emm GAS types. A) Many emm types have an emm chromosomal arrangement in which gene emm is found downstream mga and upstream scpA. These emm types do not possess emm-like genes. However, some emm types may have other genes between emm and scpA (generically represented by the triangle). All GAS strains in our collection belonging to emm types with this type of emm chromosomal arrangement (listed to the right of scpA) were successfully assigned an emm type using SRST2 WGS data mapping-based approach. B) Several GAS emm types contain emm-like genes (mrp and/or enn) in addition to gene emm. We were able to correctly derive emm type information using the same WGS mapping-based approach for several GAS emm types (listed to the right of scpA). However, the mapping-based strategy failed in cases where the CDC emm database contained sequences found also in emm-like genes. These emm types (listed below the red box in the gene schematics) were confounded by SRST2 with “emm” types whose sequences actually match either mrp or enn (listed below the green and blue boxes, respectively, in the schematics). The correct emm type could however be determined after de novo assembly of WGS data and BLAST and positional analysis. Dotted lines link “emm” sequences found in the CDC database that were identified in the same strain. Red arrows indicate annealing positions of primers 1 and 2 used in Sanger-based emm typing.

Fig. 2. Schematics showing emm chromosomal arrangements in A) Strain NGAS320, found to be non-typable by Sanger sequencing (top) and NGAS300, representative of other emm83 GAS strains in our collection (bottom). Although non-typable by Sanger-based emm typing, NGAS320 was identified as an emm83 by WGS-based emm typing. Further sequence analysis discovered a deletion of 1390 bp in this strain that knocked out the annealing site of primer 1 used in traditional Sanger-based emm typing. The deletion was confirmed by PCR amplification.
using primers annealing to the 3’ end of mga and the 5’ end of scpA (indicated by the blue arrows). Red arrows indicate annealing positions of primers 1 and 2 used in Sanger-based emm typing. B) Strain NGAS128. This GAS has an emm14 gene and was correctly typed by both Sanger-based and WGS-based emm typing. However, the enn gene of this strain possesses sequences identical to those found in the CDC emm database for emm51 (indicated by the blue box) and the originally designated emm51 sequence (6).