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

Taryn B. T. Athey,^a Sarah Teatero,^a Aimin Li,^a Alex Marchand-Austin,^{a,b} Bernard W. Beall,^c Nahuel Fittipaldi^{a,b}

Public Health Ontario, Toronto, Ontario, Canada^a; Department of Laboratory Medicine and Pathobiology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada^b; Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, USA^c

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

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 antiphagocytic 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 Sequencing of a short, hypervariable region at the 5' end of *emm* has superseded serological methods and has become the *de facto* and most used GAS typing method (5–7).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 >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 U.S. Centers for Disease Control and Prevention (CDC).

It has long been noted 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 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 strains examined so far, *emm* and *emm*-like genes are found in a discrete GAS chromosomal region which is under direct transcriptional control of the stand-alone 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 (MLST) 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 (WGS) 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 of bacterial organisms using the WGS data may be considered the ultimate typing method, there is need to provide backward 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 with 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 on and discuss how WGS data analysis was instrumental in

Received 10 January 2014 Returned for modification 31 January 2014

Accepted 12 March 2014

Published ahead of print 19 March 2014

Editor: G. V. Doern

Address correspondence to Nahuel Fittipaldi, nahuel.fittipaldi@oahpp.ca.

Supplemental material for this article may be found at <http://dx.doi.org/10.1128/JCM.00029-14>.

Copyright © 2014, American Society for Microbiology. All Rights Reserved.

doi:10.1128/JCM.00029-14

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 was used (see Table S1 in the supplemental material). 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 during the time period. Strains were grown on Columbia blood agar plates containing 5% sheep blood 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 mini-kit (Qiagen, Toronto, Canada) following the manufacturer's 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). The multiplexed sequencing reads were parsed and barcode information was removed 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 CDC ftp sites (ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm and ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/emmsequ, respectively). The GAS MLST database was downloaded from the Imperial College (<http://spyogenes.mlst.net/>). MLST 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 (<https://code.google.com/p/mosaik-aligner/>) instead of Bowtie2 (29) to align reads to the reference database. We also used SRST2 to identify *emm* types. To resolve ambiguous results when particular GAS strains had both *emm* and *emm*-like genes (see Results and Discussion), we developed a custom script (named *emm_pipeline.pl*) that uses the scoring output of SRST2, *split_fasta.pl* (<http://code.google.com/p/nash-bioinformatics-codelets/downloads/list>), and BLAST (30) to map *emm* sequences present in the CDC database to *de novo* assemblies of the strain under investigation. This custom script uses the relative gene positions within the *mga* locus to assign strain *emm* types. The code and detailed instructions are available at https://github.com/streplab/emm_pipeline. We also used SRST2 to assess the presence or absence of genes *nra*, *rofa*, *sof*, *mrp*, and *enn* in the genome data of sequenced GAS strains.

Nucleotide sequence accession number. WGS data for the 191 isolates were submitted to the Sequence Read Archive (SRA) under accession no. SRP035244.

RESULTS AND DISCUSSION

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), which corresponds to an average coverage of 267× (maximum, 578×; minimum, 49×), considering a GAS genome size of 1.8 Mbp (see Table S1 in the supplemental material).

SRST and its successor SRST2 are software programs originally conceived as a means to derive MLST 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 data-

bases 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 *emm* type that had been previously determined using traditional Sanger sequencing-based *emm* typing. Notably, all strains belonging to *emm* types known or found not to contain *emm*-like genes were assigned an *emm* type which matched the one determined using Sanger sequencing-based *emm* typing (Fig. 1A). On the other hand, not all strains belonging to *emm* types known to possess *emm*-like genes had a match between Sanger sequencing-based and WGS-based *emm* typing. Indeed, we noted several instances of discrepancies between the two methods for strains possessing *emm*-like genes. To resolve these discrepancies, we manually inspected the SRST2 output and discovered that every time a mismatch between the Sanger sequencing-based and WGS-based typing was recorded, SRST2 had given very high scores to two or three *emm* types found on the CDC's *emm* database (data not shown). Remarkably, closer examination after *de novo* assembly of the Illumina short-reads into contigs and BLAST analysis revealed that all those highly scored “*emm*” sequences were indeed present in the genomes of the strains under investigation and they corresponded, one to the legitimate *emm* gene and the other(s) to *emm*-like genes (Fig. 1B). These findings strongly suggest that in a context of an ambiguous database such as what the current version of the CDC *emm* database appears to be, a mapping-based-only strategy such as the one offered by SRST2 is not sufficient to confidently assign *emm* types from WGS data.

The CDC *emm* typing database contains sequences found in *emm*-like genes that complicate WGS-based *emm* typing. Most *emm* designations and all recent *emm* sequence designations have relied solely on the sequencing results of amplicons generated by the use of oligonucleotide primers 1 and 2, specific for what has been thought to be the *emm* gene and not other *emm*-like genes (5–7, 31). Since this definition of *emm* type is broad, it may be difficult to establish whether an “*emm* type” found in the CDC database corresponds to sequences found in a legitimate *emm* gene. Instead, some “*emm* types” may correspond to *emm*-like genes that were somehow amplified using the *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 data 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 nontypeable by Sanger sequencing because an amplicon could not be generated using *emm* typing primers 1 and 2, was identified as an *emm*83 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 had a deletion of 1,390 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

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 of *mga* and upstream of *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 the 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 *emm*). 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.

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. First, it is not known how many more sequences matching *emm*-like genes are found in

the database, and second, and perhaps more important, 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

FIG 2 Schematics showing *emm* chromosomal arrangements in different strains. (A) Strain NGAS320, found to be nontypeable by Sanger sequencing (top) and strain NGAS300, representative of other *emm83* GAS strains in our collection (bottom). Although nontypeable by Sanger-based *emm* typing, NGAS320 was identified as an *emm83* by WGS-based *emm* typing. Further sequence analysis discovered a deletion of 1,390 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).

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, also had short-read sequences corresponding to allele *emm51*, 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 software to generate contigs for all GAS strains from the short-read WGS data and then 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 at https://github.com/streplab/emm_pipeline), the strain is flagged for further inspection by automatic *de novo* assembly and BLAST processing as described in the supplemental material. When this approach was attempted on the current data set, 77 of the 191 GAS strains were selected for *de novo* assembly and further inspection. For 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 (data not shown). In the remaining 66 GAS strains, at least two of the SRST2 top-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, the 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 100% agreement between Sanger sequencing-based and WGS-based *emm* typing, with the exception of the above-mentioned strain NGAS320, for which a PCR amplicon could not be obtained and, therefore, was nontypeable by Sanger sequencing-based *emm* typing. Thus, the use of our pipeline permitted the accurate determination of *emm* types 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 information directly from the short-read WGS data (28). Here, we were able to determine allele type and sequence type (ST) for most of the 191 GAS strains in our collection using SRST2 (see Table S1 in the supplemental material). Although some of the STs were not found in the MLST database and may represent novel variants, there were some occasions for which results obtained by SRST2 may need to be confirmed by traditional amplification and Sanger sequencing of the

TABLE 1 BLAST hits of sequences present in the CDC *emm* database against *emm* and *emm*-like genes present in the *de novo* assemblies of 77 GAS strains chosen by our pipeline for further inspection

Sanger sequencing-determined <i>emm</i> type	CDC database <i>emm</i> allele identified by BLAST for ^a :			No. of strains with these BLAST results
	<i>mrp</i> gene	<i>emm</i> gene	<i>enn</i> gene	
2		2	149	2
3		3 ^b		7
4	156	4	236	4
4	156	4		6
5		5 ^b		2
9		9	236	1
11		11	202	2
14		14	51	1
18		18	205	2
22	156	22		2
59		59	174	3
73		73	149	1
73		73 ^b		1
75		75	170	2
77		77	159	2
83		83	240	10
83		83 ^b		1
87		87	159	3
89		89	236	11
101		101	205	2
105	156	105		1
114		114	159	4
118		118	236	2
122		122	240	1
169		169	164	2
192		192	138	1
NT		83	240	1

^a Only the CDC database *emm* allele with the top BLAST score for each gene is shown.

^b All of the SRST2 top-scoring CDC database *emm* alleles had BLAST hits to the *emm* gene in strains of these *emm* types. Only the top-scored BLAST hit is presented.

alleles. SRST2 distinguishes those potentially ambiguous results with a question mark (see Table S1 in the supplemental material). Overall and despite the presence of these ambiguities, our data showed strong correlations between *emm* types and MLST STs for the vast majority of the isolates (see Table S1 in the supplemental material), 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* (see Table S1 in the supplemental material). 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) as containing 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 (see Table S1 in the supplemental material).

WGS-based typing of bacterial organisms is becoming an increasingly feasible option for diagnostic and public health labora-

tories (21–24). While we advance toward generalized WGS-based typing of GAS, it is important to rely on validated tools to rapidly and efficiently provide backward 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 the fact that high-quality, well-curated databases are crucial to fully take advantage of WGS data.

ACKNOWLEDGMENTS

We thank Jonas Winchell (U.S. 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

- Olsen RJ, Musser JM. 2010. Molecular pathogenesis of necrotizing fasciitis. *Annu. Rev. Pathol.* 5:1–31. <http://dx.doi.org/10.1146/annurev-pathol-121808-102135>.
- Lancefield RC. 1928. The antigenic complex of *Streptococcus haemolyticus*: I. Demonstration of a type-specific substance in extracts of *Streptococcus haemolyticus*. *J. Exp. Med.* 47:91–103.
- Scott JR, Pulliam WM, Hollingshead SK, Fischetti VA. 1985. Relationship of M protein genes in group A streptococci. *Proc. Natl. Acad. Sci. U. S. A.* 82:1822–1826. <http://dx.doi.org/10.1073/pnas.82.6.1822>.
- Manjula BN, Acharya AS, Fairwell T, Fischetti VA. 1986. Antigenic domains of the streptococcal M5 protein. Localization of epitopes crossreactive with type 6 M protein and identification of a hypervariable region of the M molecule. *J. Exp. Med.* 163:129–138.
- Facklam R, Beall B, Efstratiou A, Fischetti V, Johnson D, Kaplan E, Kriz P, Lovgren M, Martin D, Schwartz B, Totolian A, Bessen D, Hollingshead S, Rubin F, Scott J, Tyrrell G. 1999. *emm* typing and validation of provisional M types for group A streptococci. *Emerg. Infect. Dis.* 5:247–253. <http://dx.doi.org/10.3201/eid0502.990209>.
- Whatmore AM, Kapur V, Sullivan DJ, Musser JM, Kehoe MA. 1994. Non-congruent relationships between variation in *emm* gene sequences and the population genetic structure of group A streptococci. *Mol. Microbiol.* 14:619–631. <http://dx.doi.org/10.1111/j.1365-2958.1994.tb01301.x>.
- Li Z, Sakota V, Jackson D, Franklin AR, Beall B, Active Bacterial Core Surveillance/Emerging Infections Program Network. 2003. Array of M protein gene subtypes in 1064 recent invasive group A *Streptococcus* isolates recovered from the active bacterial core surveillance. *J. Infect. Dis.* 188:1587–1592. <http://dx.doi.org/10.1086/379050>.
- Kehoe MA, Kapur V, Whatmore AM, Musser JM. 1996. Horizontal gene transfer among group A streptococci: implications for pathogenesis and epidemiology. *Trends Microbiol.* 4:436–443. [http://dx.doi.org/10.1016/0966-842X\(96\)10058-5](http://dx.doi.org/10.1016/0966-842X(96)10058-5).
- Whatmore AM, Kapur V, Musser JM, Kehoe MA. 1995. Molecular population genetic analysis of the *emm* subdivision of group A streptococcal *emm*-like genes: horizontal gene transfer and restricted variation among *emm* genes. *Mol. Microbiol.* 15:1039–1048. <http://dx.doi.org/10.1111/j.1365-2958.1995.tb02279.x>.
- Podbielski A, Schnitzler N, Beyhs P, Boyle MD. 1996. M-related protein (Mrp) contributes to group A streptococcal resistance to phagocytosis by human granulocytes. *Mol. Microbiol.* 19:429–441. <http://dx.doi.org/10.1046/j.1365-2958.1996.377910.x>.
- Thern A, Stenberg L, Dahlback B, Lindahl G. 1995. Ig-binding surface proteins of *Streptococcus pyogenes* also bind human C4b-binding protein (C4BP), a regulatory component of the complement system. *J. Immunol.* 154:375–386.
- Horstmann RD, Sievertsen HJ, Knobloch J, Fischetti VA. 1988. Antiphagocytic activity of streptococcal M protein: selective binding of complement control protein factor H. *Proc. Natl. Acad. Sci. U. S. A.* 85:1657–1661. <http://dx.doi.org/10.1073/pnas.85.5.1657>.
- Hondorp ER, McIver KS. 2007. The Mga virulence regulon: infection where the grass is greener. *Mol. Microbiol.* 66:1056–1065. <http://dx.doi.org/10.1111/j.1365-2958.2007.06006.x>.
- Bessen DE, Lizano S. 2010. Tissue tropisms in group A streptococcal infections. *Future Microbiol.* 5:623–638. <http://dx.doi.org/10.2217/fmb.10.28>.
- Beall B, Gherardi G, Lovgren M, Facklam RR, Forwick BA, Tyrrell GJ. 2000. *emm* and *sof* gene sequence variation in relation to serological typing of opacity-factor-positive group A streptococci. *Microbiology* 146 (Part 5):1195–1209.
- Bessen DE, Manoharan A, Luo F, Wertz JE, Robinson DA. 2005. Evolution of transcription regulatory genes is linked to niche specialization in the bacterial pathogen *Streptococcus pyogenes*. *J. Bacteriol.* 187:4163–4172. <http://dx.doi.org/10.1128/JB.187.12.4163-4172.2005>.
- Johnson DR, Kaplan EL, VanGheem A, Facklam RR, Beall B. 2006. Characterization of group A streptococci (*Streptococcus pyogenes*): correlation of M-protein and *emm*-gene type with T-protein agglutination pattern and serum opacity factor. *J. Med. Microbiol.* 55:157–164. <http://dx.doi.org/10.1099/jmm.0.46224-0>.
- Enright MC, Spratt BG, Kalia A, Cross JH, Bessen DE. 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. *Infect. Immun.* 69:2416–2427. <http://dx.doi.org/10.1128/IAI.69.4.2416-2427.2001>.
- McGregor KF, Spratt BG, Kalia A, Bennett A, Bilek N, Beall B, Bessen DE. 2004. Multilocus sequence typing of *Streptococcus pyogenes* representing most known *emm* types and distinctions among subpopulation genetic structures. *J. Bacteriol.* 186:4285–4294. <http://dx.doi.org/10.1128/JB.186.13.4285-4294.2004>.
- Musser JM, Shelburne SA, 3rd. 2009. A decade of molecular pathogenomic analysis of group A *Streptococcus*. *J. Clin. Invest.* 119:2455–2463. <http://dx.doi.org/10.1172/JCI38095>.
- Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.* 10:599–606. <http://dx.doi.org/10.1038/nrmicro2850>.
- Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* 8:e1002824. <http://dx.doi.org/10.1371/journal.ppat.1002824>.
- Olsen RJ, Long SW, Musser JM. 2012. Bacterial genomics in infectious disease and the clinical pathology laboratory. *Arch. Pathol. Lab Med.* 136:1414–1422. <http://dx.doi.org/10.5858/arpa.2012-0025-RA>.
- Long SW, Williams D, Valson C, Cantu CC, Cernoch P, Musser JM, Olsen RJ. 2013. A genomic day in the life of a clinical microbiology laboratory. *J. Clin. Microbiol.* 51:1272–1277. <http://dx.doi.org/10.1128/JCM.03237-12>.
- Ben Zakour NL, Venturini C, Beatson SA, Walker MJ. 2012. Analysis of a *Streptococcus pyogenes* puerperal sepsis cluster by use of whole-genome sequencing. *J. Clin. Microbiol.* 50:2224–2228. <http://dx.doi.org/10.1128/JCM.00675-12>.
- Fittipaldi N, Beres SB, Olsen RJ, Kapur V, Shea PR, Watkins ME, Cantu CC, Laucirica DR, Jenkins L, Flores AR, Lovgren M, Ardanuy C, Linares J, Low DE, Tyrrell GJ, Musser JM. 2012. Full-genome dissection of an epidemic of severe invasive disease caused by a hypervirulent, recently emerged clone of group A *Streptococcus*. *Am. J. Pathol.* 180:1522–1534. <http://dx.doi.org/10.1016/j.ajpath.2011.12.037>.
- Tritt A, Eisen JA, Facciotti MT, Darling AE. 2012. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* 7:e42304. <http://dx.doi.org/10.1371/journal.pone.0042304>.
- Inouye M, Conway TC, Zobel J, Holt KE. 2012. Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics* 13:338. <http://dx.doi.org/10.1186/1471-2164-13-338>.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359. <http://dx.doi.org/10.1038/nmeth.1923>.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402. <http://dx.doi.org/10.1093/nar/25.17.3389>.
- Facklam RF, Martin DR, Lovgren M, Johnson DR, Efstratiou A, Thompson TA, Gowan S, Kriz P, Tyrrell GJ, Kaplan E, Beall B. 2002. Extension of the Lancefield classification for group A streptococci by addition of 22 new M protein gene sequence types from clinical isolates:

- emm103* to *emm124*. Clin. Infect. Dis. 34:28–38. <http://dx.doi.org/10.1086/324621>.
32. Whatmore AM, Kapur V, Musser JM, Sullivan DJ, Kehoe MA. 1995. Variation in *emm*-like gene sequences in the context of the population genetic structure of group A streptococci. Dev. Biol. Stand. 85:159–162.
33. Whatmore AM, Kehoe MA. 1994. Horizontal gene transfer in the evolution of group A streptococcal *emm*-like genes: gene mosaics and variation in Vir regulons. Mol. Microbiol. 11:363–374. <http://dx.doi.org/10.1111/j.1365-2958.1994.tb00316.x>.
34. Dowson CG, Barcus V., King S., Pickerill P., Whatmore A, Yeo M. 1997. Horizontal gene transfer and the evolution of resistance and virulence determinants in *Streptococcus*. Soc. Appl. Bacteriol. Symp. Ser. 26:42S–51S.