GBS-SBG - GBS Serotyping by Genome Sequencing

Suma Tiruvayipati¹, Tan Wen Ying², Timothy Barkham², Swaine L. Chen^{1,3}

¹Infectious Diseases Translational Research Programme, Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

²Department of Laboratory Medicine, Tan Tock Seng Hospital, Singapore

³Laboratory of Bacterial Genomics, Genome Institute of Singapore, 60 Biopolis Street, Genome #02-01, Singapore 138672

Corresponding author: Swaine L. Chen slchen@gis.a-star.edu.sg

Keywords: Group B Streptococcus serotypes, Genome sequencing, Short read typing, Molecular epidemiology

ABSTRACT

Group B *Streptococcus agalactiae* (GBS; *Streptococcus agalactiae*) is the most common cause of neonatal meningitis and a rising cause of sepsis in adults. Recently, it has also been shown to cause foodborne disease. As with many other bacteria, the polysaccharide capsule of GBS is antigenic, enabling its use for strain serotyping. Recent advances in DNA sequencing have made sequence-based typing attractive (as has been implemented for several other bacteria, including *Escherichia coli*, *Klebsiella pneumoniae* species complex, *Streptococcus pyogenes*, and others). For GBS, existing WGS-based serotyping systems do not provide complete coverage of all known GBS serotypes (specifically including subtypes of serotype III), and none are simultaneously compatible with the two most common data types, raw short reads and assembled sequences. Here, we create a serotyping database (GBS-SBG, GBS Serotyping by Genome Sequencing), with associated scripts and running instructions, that can be used to call all currently described GBS serotypes, including subtypes of serotype III, using both direct short-read- and assembly-based typing. We achieved higher concordance using GBS-SBG on a previously reported data set of 790 strains. We further validated GBS-SBG on a new set of 572 strains, achieving 99.8% concordance with PCR-based molecular serotyping using either short-read- or assembly-based typing. The GBS-SBG package is publicly available and will accelerate and simplify serotyping by sequencing for GBS.

DATA SUMMARY

1. The GBS-SBG package is open source and available for at Github under the MIT license (URL - <https://github.com/swainechen/GBS-SBG>)
2. Accession numbers of the sequencing reads and reference sequences used in the study from earlier reports have been provided within the article and the supplementary data
3. The WGS data for the 572 isolates used in the study is available at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA293392>

INTRODUCTION

Group B *Streptococcus* (GBS, also known as *Streptococcus agalactiae*) was named for its common association with mastitis in cows (agalactiae = “no milk”). It is a common colonizer of the human gastrointestinal and urinary tract, being present in

up to one third of apparently healthy individuals (1–3). Since the late 19th century, it has become increasingly associated with neonatal meningitis, today representing the most common cause (4–7). Neonatal GBS infections are classified clinically into early-onset (< 7 days of age, EOD) and late-onset (> 7 days of age, LOD) disease. In addition, GBS is an increasingly common pathogen in immunocompromised and elderly adults (8–11). Recently, GBS has been shown to cause foodborne infection associated with the consumption of raw fish in otherwise healthy adults in Singapore (and likely throughout Southeast Asia) (12–14). Accordingly, GBS is also well known to colonize and infect (often resulting in severe invasive disease) multiple other species, including fish (where it has a large impact on aquaculture) as well as other mammals, amphibians, and reptiles (15–20). As an important pathogen of humans, cows, and fish, GBS is therefore of concern for public health, economic, and zoonotic reasons.

Several decades ago, a serotyping system was established (21) based on differences in antigenicity of the polysaccharide capsule. Currently, there are 10 main serotypes (Ia, Ib, II–IX); furthermore, serotype III has been subtyped into 4 subtypes (III-1 through III-4) (22). These serotypes are encoded by the *cps* locus, which consists of both conserved (*cpsD* - *cpsG*) and variable (*cpsG* - *cpsK*) regions; the sequence differences, primarily in the variable region, form the basis of a commonly-used PCR-based molecular serotyping scheme (23). Epidemiological studies have made clear that different serotype (and subtype) distributions are associated with different host species, disease states, and geographical distributions (24). For example, mastitis in cows is mostly caused by serotype Ia GBS in China (25), while serotype III (subtype III-3) is most common among cows in Canada (26). Furthermore, outbreaks in infected fish are largely caused by serotype Ia, Ib, and III (the latter specifically being subtype III-4 and occurring predominantly in Southeast Asia, with recent introductions into Brazil) (14,27). Additional resolution afforded by other typing systems (multilocus sequence typing, virulence gene typing, mobile genetic elements, antibiotic resistance profiles, and whole genome sequencing) have overall confirmed these initial epidemiological observations based on serotypes (19,24,28–33). Therefore, serotyping for GBS is still a useful and important method, and continues to be particularly relevant for vaccine development (24,34).

Various experimental methods have been used for serotyping of bacteria, such as enzyme immunoassay, immunoprecipitation, co-agglutination, inhibition enzyme-linked immunosorbent assay, latex agglutination, and fluorescence microscopy (21,35–39). These traditional methods have mostly been replaced by genotyping methods, often based on polymerase chain reaction (PCR)-based assays (40–43). Given recent advances in the throughput, availability, and affordability of sequencing technologies, whole genome sequencing (WGS) has now become a practical method (with advantages for automation and scale) to call serotypes in multiple bacteria, such as *Salmonella* (44,45), *Streptococcus pneumoniae* (46), *Escherichia coli* (47–49), and *Klebsiella pneumoniae* (50).

To date, three reports have explored using WGS to serotype GBS. The first used a database of 9 serotypes, focusing on the variable region of the *cps* locus (*cpsG* - *cpsK*); this database did not include serotype IX (51) and was designed for use with assembled genomes. The second database included all of the 10 main serotypes, including the full *cps* locus with both conserved and variable regions (52). This latter database was tested on raw short reads (using a mapping strategy) and assembled genomes (using a blast-based strategy); the mapping strategy was found to have higher concordance with latex agglutination-based serotyping (52). The final mapping strategy from this report does not seem to have been implemented in any publicly available software. A third WGS serotyping strategy employed partial gene sequences of the *cps* locus as a reference database for a short-read mapping strategy (53); the use of short (100-300 bp) partial gene sequences makes it ill-suited for analysis using an assembly-based strategy. Importantly, none of these databases included subtypes of serotype III, which were defined based on SNPs in a portion of the conserved region of the *cps* locus (22).

We thus aimed to devise a serotyping solution using WGS for GBS that would (i) include all existing serotypes, including subtypes of serotype III and (ii) enable accurate serotyping by both short-read-mapping and assembly-based strategies. We validated our method on the 790 strains reported in (52), achieving equivalent or higher concordance using either mapping- or assembly-based strategies. We further validated our method using an additional 572 strains, achieving 99.8% concordance with traditional PCR-based serotyping, WGS short read-based calling, and WGS

assembly-based calling. Our software is called GBS-SBG and is available at <https://github.com/swainechen/GBS-SBG>.

MATERIALS & METHODS

Previously published GBS capsular locus reference sequences

Reference sequences representing the 10 serotypes (Ia – IX) were taken from (52); we refer to these as the “Kapatai database” (Figure 1, Table 1). The sequences in the Kapatai database range in length from 15,090bp to 17,514bp. Reference sequences for 9 serotypes (Ia – VIII) with lengths ranging from 4,477bp to 6,307bp were obtained from (51). We refer to these 9 sequences as the “Sheppard database”. Sequences representing two subtypes of serotype III (III-2 (AF332896) and III-3 (AF332897)) were obtained from (22).

GBS strains

572 GBS isolates were collected between November 2000 and July 2018. Of these, 547 were isolated from humans and 25 from fish. 487 Human isolates were collected at Tan Tock Seng Hospital (TTSH), Singapore, and stored at -70°C. All isolates were from nonpregnant adults, as TTSH does not offer paediatric or obstetric care. The study was approved by the Institutional Review Board of TTSH, Singapore, DSRB2016/00202. 45 (human), 6 (fish), 20 (1 human, 19 fish), 14 (human) isolates were obtained/ collected from Laos, Malaysia, Thailand, and Vietnam, respectively.

PCR-based molecular serotyping

GBS isolates were subcultured onto blood agar and re-identified using a MALDI-TOF system (Bruker). GBS DNA was extracted with the EasyMag system (Biomerieux) according to the manufacturer's instructions. Molecular serotyping of the GBS isolates was performed using 2 PCRs: "Multiplex PCR 1" and "Multiplex PCR 2" as described by Poyart *et al.* (54), with segregation into capsular polysaccharide serotypes Ia, Ib, II, III, IV, V, VI, VII, and VIII achieved using agarose gel electrophoresis based on amplicon size. If an isolate remained untypeable after these two PCRs, a third PCR step with another multiplex PCR as described by Imperi *et al.* (23) was then performed.

Whole genome sequencing and analysis

Genomic DNA was extracted from overnight cultures on blood agar using the QIAamp DNA mini kit (cat. no. 51306). Whole-genome sequencing was performed by the Next Generation Sequencing Platform at the Genome Institute of Singapore, as previously described (13). Briefly, sequencing library preparation was done with the use of the TruSeq Nano DNA LT Library Prep Kit (Illumina) or the Nextera XT Library Prep Kit (Illumina) according to the manufacturer's instructions. The sequencing libraries were sequenced using a NextSeq 500 or HiSeq 4000 sequencer with 2x151-bp reads (Illumina, San Diego, CA, USA). All sequencing data is uploaded in GenBank under BioProject PRJNA293392.

SRST2 version 0.2.0 (55) was used to call multilocus sequence types (MLST) using reference sequences downloaded from PubMLST (<https://www.pubmlst.org>). SRST2 was also used to call serotypes using a mapping strategy using the GBS-SBG database as a custom reference database, according to the SRST2 documentation. The default minimum percentage coverage cutoff (--min-coverage option; default 90%) for gene reporting was used.

Raw short read (FASTQ) sequences were also assembled with velvet (version-1.2.10) (56) using the VelvetOptimizer helper script (version 2.2.4) and a minimum contig cutoff of 500bp, scaffolded with OPERA-LG (version 2.0.6) (57), and finished with FinIS (version 0.3) (58). Alignment-based calling of serotypes was done using a single script that performed alignment against the GBS-SBG database using BLASTN (59) and processed the results (available at <https://github.com/swainechen/GBS-SBG>). For consistency with SRST2, a minimum nucleotide identity of 90% across 90% of the reference serotype length (i.e. 90% coverage) was required for making a serotype call.

Table 1. GBS serotype reference databases used in the current study.

Boxes in white - Start, End, Id%, Region relative to Kapatai et al. 2017 sequences.

Boxes in grey - Start, End, Id%, Region relative to serotype III Kapatai et al. 2017 sequence.

(*Isolate names from the current study, Id% – Identity in percentage, L(bp) – Length in base pairs)

Serotype	Katapai et al. 2017			Metcalf et al. 2017						Sheppard et al. 2016						Current Study					
	Accession	L(bp)	Region	Accession	L(bp)	Start	End	Id%	Region	Accession	L(bp)	Start	End	Id%	Region	Accession	L(bp)	Start	End	Id%	Region
Ia	LT671983	15799	cps locus	AB028896.2	143	6284	6426	94.41	<i>cpsH</i>	AB028896.2	4714	5538	10254	99.15		-	6712	3539	10250	100	
Ib	LT671984	15807		AB050723.1	143	6396	6538	100	<i>cpsH</i>	AB050723.1	4617	5650	10266	100		-	6616	3651	10266	100	
II	LT671985	17514		AY375362.1	139	11345	11483	100	<i>cpsK</i>	EF990365.1	6307	5575	11881	99.84		-	8315	3567	11881	100	
III	LT671986	15760		AF163833.1	172	6631	6802	99.42	<i>cpsH</i>	AF163833.1	4602	5592	10193	99.33		N/A	-	-	-	-	
III-1	-	-		-	-	-	-	-	-	-	-	-	-	-		SG-M340*	6610	3593	10202	99.33	<i>cpsD - cpsK</i>
III-2	-	-		-	-	-	-	-	-	-	-	-	-	-		SG-M40*	6602	3593	10194	99.35	
III-3	-	-		-	-	-	-	-	-	-	-	-	-	-		SG-M61*	6610	3593	10202	99.67	
III-4	-	-		-	-	-	-	-	-	-	-	-	-	-		SG-M918*	6610	3593	10202	99.45	
IV	LT671987	16549		AF355776.1	222	6670	6891	100	<i>cpsH</i>	AF355776.1	5240	5731	10971	99.98		-	7249	3723	10971	100	
V	LT671988	17319		AF349539.1	171	9107	9277	100	<i>cpsO</i>	AF349539.1	6148	5578	11725	100		-	8156	3570	11725	100	
VI	LT671989	15847		AF337958.1	99	7385	7483	100	<i>cpsH</i>	AF337958.1	4477	5777	10254	99.89		-	6477	3778	10254	100	
VII	LT671990	16473		AY376403.1	160	7535	7694	100	<i>cpsM</i>	AY376403.1	5264	5578	10842	99.45		-	7272	3570	10841	100	
VIII	LT671991	15090		AY375363.1	113	8167	8279	100	<i>cpsJ</i>	AY375363.1	4370	5117	9487	99.98		-	5915	3541	9455	100	
IX	LT671992	16440		GQ499301.1	130	8665	8794	100	<i>cpsO</i>	N/A	-	-	-	-		-	7273	3581	10853	100	

Creation of the GBS-SBG database

For all serotypes except for serotype III, the sequences reported by (52) (i.e. the Kapatai database) were used (Table 1). Multiple sequence alignments were made using MAFFT version 7.45 (60). Sequence alignments were visualized using Jalview 2.11.1.3 (61). Sequences were trimmed to a uniform 5' and 3' end (see results) (Figure 1). For the subtypes of serotype III, reference sequences for III-2 and III-3 were taken from (22). SNPs specific for III-1 and III-4 (22) were introduced into the reference serotype III sequence from (52) (of note, this serotype III sequence (LT671986) appeared to be a hybrid between III-2 and III-4 (based on characteristic SNPs for these subtypes) and was therefore not clearly subtypeable itself (Supplementary Table 1)). To generate reference sequences for III-1 and III-4, all of the new strains that were typed as serotype III by PCR (173/572 strains) were considered. All of these were unambiguously typable as one of the four subtypes of serotype III based on the SNPs reported by (22) (Supplementary Table 1, Supplementary Figure 1). For each subtype, one corresponding assembled sequence (Table 1) was then included as a reference sequence in the final database.

Pairwise distance and phylogenetic analysis

Pairwise distances and maximum likelihood phylogenetic analysis using a Tamura-Nei model (62) was performed using MEGA X (63). The initial tree(s) for the heuristic search were obtained automatically by applying the Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL).

RESULTS

Existing GBS serotyping strategies do not include all known serotypes and only work with a single data type

A valuable dataset, comprising 790 isolates with WGS data and experimental serotyping by latex agglutination, was made publicly available in the report of one of the short-read mapping GBS serotypers (52). This study achieved a concordance of 725/790 between WGS (using short read mapping) and latex agglutination. The authors also tested an assembly-based method using their same sequence database, but found lower concordance (664/790) (Supplementary Data 2). The lower concordance using assemblies, of course, meant that the concordance between a mapping- and assembly-based strategy was also not perfect (721/790).

We performed serotyping of these 790 isolates using a short read mapping strategy (with the popular SRST2 program) against the Kapatai database; this resulted in 431/790 correct serotypes. The 358 discrepant calls were all miscalls between serotypes III and Ia (355 were miscalled as Ia while 3 were miscalled as III). Of note, high similarity between the *cps* loci for serotypes Ia and III has previously been reported (52,64).

We further leveraged this data set of 790 strains to evaluate the suitability of the reference database reported in (51) (designed for an assembly-based strategy) using a mapping-based strategy. In agreement with (65), we found that this database performed very well, resulting in 780/790 correct serotype calls. Of the 10 discordant calls, 9 were serotype IX (which was not present in the Sheppard database) while 1 was miscalled as serotype III (instead of Ia).

The report by (53) was also originally designed for mapping-based typing. We did not assess its performance for assembled sequences because the reference sequences were very short (100-300 bp) (Table 1).

Therefore, none of the three existing databases was definitively usable as-is for both mapping- and assembly-based typing. Furthermore, none included the possibility of calling serotype III subtypes.

Construction of a complete reference serotype database usable for both assembly- and short read-based analyses

We therefore sought to construct a single database that could provide high accuracy using both short reads and assemblies as well as provide serotype III subtyping. Previous studies have noted that the variable region of the *cps* locus is important for assembly-based typing; this was the main region used for the assembly-based method (51). Short read mapping-based typing, such as that performed by SRST2 (which was used by (53)), is typically designed to differentiate between closely related alleles (i.e. conserved regions), making it well suited for multilocus sequence typing (MLST), for which the 5' and 3' ends of the reference sequences are strictly trimmed to provide uniform-length alignments of the typing alleles. Interestingly, we found that using SRST2 also worked well with the Sheppard database but less well with the Kapatai database, despite the latter also including the variable region of the *cps* locus.

We therefore hypothesized that a new database should include both conserved and variable portions of the *cps* locus, in order to accommodate both short read- and assembly-based typing. We also sought to align the start and end of each sequence. We built upon the database created by (52), as it already achieved good performance with short read mapping (though not with SRST2).

We constructed a new database as follows (see Methods for more details):

1. Take all 10 main serotypes from the Kapatai database and align them.
2. Trim all sequences on the 5' end to the region that includes the SNPs reported to differentiate serotype III subtypes (22).

3. Trim all sequences on the 3' end to the edge of the variable sequence, which corresponds to the right edge of the sequences in the Sheppard database; this resulted in a ~8kbp reference sequence for each serotype.
4. Replace the single serotype III sequence with representatives from the four subtypes (see below).

The serotype III subtypes have been defined by SNPs that are found in a conserved region of the *cps* locus (22). Full length sequences for these have not been explicitly identified, however. We therefore used the WGS assemblies for our newly sequenced strains (see below) that were typed as serotype III by PCR (173 strains in total). These were unambiguously classifiable into the four subtypes based on the SNPs reported by (22), resulting in 26 III-1, 22 III-2, 8 III-3, and 117 III-4 isolates. For each subtype, we then took the consensus sequence over the ~8kbp typing region (for each subtype, this consensus sequence exactly matched the sequence of multiple assemblies in our data set) and included this in our new reference serotyping database. We refer to this as the GBS-SBG database.

The GBS-SBG database enables serotyping by both mapping and assembly-based strategies

We tested the GBS-SBG database on the 790 isolates from (52), using both short read- (with SRST2) and assembly-based (using BLASTN, with 90% identity and 90% coverage cutoffs; see Methods) strategies. We used the mapping-based calls from Supplementary Data 2 in the original report (52) as the gold standard (as noted above, these mapping-based calls were only concordant with assembly-based calls for 721/790 strains in the original report). We found 100% concordance (790/790) and 94.6% (748/790) with the reported serotypes when using read mapping (SRST2) and assembled sequences (BLASTN), respectively, with the GBS-SBG database. The differences in the 42 strains from the assembly-based analysis were all due to calls as non-typeable. Further examination showed that these 42 strains had lower sequencing depth ($p < 1.044e-12$, 2-tailed Mann-Whitney U-test; Supplementary Figure 4), which led to an inability to assemble portions of the *cps* locus. The regions that were not assembled correlated with areas of low coverage, which still largely remained above the minimum read depth (5x) required by SRST2 to make a call (Supplementary Figure 5). Use of the original assemblies by (52), which were

performed using SPAdes instead of velvet, led to a different set of nontypeable strains, which also had generally lower coverage (Supplementary Data 2).

Validation of GBS-SBG on an a previously unanalyzed set of strains

A total of 572 previously unanalyzed GBS isolates were serotyped by PCR and whole genome sequenced. By PCR, none of these were typed as serotype VIII, and one isolate (SG-M122) was non-typeable (Supplementary Data 3). We again used both short read- (with SRST2) and assembly-based (using BLASTN) strategies with the Sheppard and Kapatai databases as well as with our new database. The concordance with the PCR-based serotyping is summarized in Supplementary Data 3. Using the GBS-SBG database achieved the best concordance by both mapping- and assembly-based strategies (571/572, 99.8% concordance). Furthermore, we achieved 100% accuracy in calling serotype III subtypes. For assembly-based analysis only, the Sheppard database was equivalently good, with the exception of not including a serotype IX reference sequence.

Interestingly, using a mapping-based strategy (although with SRST2), the Kapatai database led to concordant calls with PCR serotyping in only 386/572 isolates. The majority of the discrepancies (168/186) were due to a miscall of serotype Ia (by WGS) for serotype III (by PCR). The remaining 18 isolates had the opposite problem (called serotype III by WGS using the Kapatai database and serotype Ia by PCR) (Supplementary Data 3). As noted above, this is consistent with the close similarity between serotype Ia and III (52,64), which we saw with the reference sequences we used as well (Figure 2, Supplementary Figure 2). Furthermore, the serotype III sequence included in the Kapatai database itself was not clearly one of the previously described subtypes, but instead appeared to be a hybrid of III-2 and III-4 (Supplementary Data 1). Regardless, these miscalls between serotypes Ia and III were resolved by using our new database, possibly in part by the inclusion of all four subtypes and by the alignment of the 5'- and 3'-ends of the reference sequence, facilitating discrimination by SRST2.

We found only one strain (SG-M666) that had a discordant call with PCR (Ia) and WGS (V) typing (both short reads and assemblies gave a serotype V call). On repeat PCR typing, this strain was again called a serotype Ia. Examination of the WGS data

showed that the *cps* locus was 100% identical to the serotype V reference sequence, with a 614 bp deletion affecting the *cpsN* and *cpsO* genes. As this deletion is <10% of the length of the serotype V reference sequence, GBS-SBG called this as serotype V (Supplementary Figure 3). In contrast, the SG-M666 *cps* locus was only 98.8% identical to the serotype Ia reference sequence in regions where it was aligned by BLASTN. Furthermore, it had a 3,423 bp deletion (>10% of the total length) relative to the serotype Ia reference sequence (such that the *cpsG-cpsJ* genes were missing). Closer examination of the target regions for the typing PCRs showed that the 614 bp deletion eliminated one of the priming sites for the serotype V PCR; in this multiplex PCR, serotype Ia and serotype V have two PCR products of identical size, but serotype V should have a third PCR product that overlapped this deletion. This led to a miscall of this strain as serotype Ia by PCR (Supplementary Figure 3). Therefore, while discordant between PCR and WGS, we believe that in this case the WGS typing (serotype V) is the correct call, though the 614 bp deletion may mean this strain is actually nontypeable by latex agglutination.

DISCUSSION

The dramatic increase in availability of whole genome sequencing has increased the importance of *in silico* approaches for analysis of multiple bacteria. Serotyping remains an important epidemiological adjunct for many bacteria of public health importance. For GBS, three previous studies have reported WGS-based serotyping, two using a mapping approach (one using SRST2 and another using a custom SNP-based method) (52,53) and one using an assembly approach (51). Having a database that is compatible with a general tool like SRST2 (as opposed to a custom mapping method) provides advantages for enabling standard analysis of multiple bacteria, such as for MLST, antibiotic resistance, and virulence genes. SRST2 also includes a reference database enabling serotyping of *E. coli*. However, many WGS-based bacterial serotyping methods (such as for *E. coli*, *K. pneumoniae*, *Salmonella*, Group A *Streptococcus*, etc.) still use an assembly-based analysis (50,65–67). Furthermore, long-read sequencing technologies are also becoming more popular, which likely will be better suited to assembly-based methods. We therefore speculate that both short read- and assembly-based methods will be important for some time; however, finding a common reference database which works equivalently well for mapping- and assembly-based serotype calling for GBS has been challenging (52).

Here, we have developed a single database (GBS-SBG) that provides the highest accuracy (as assessed by concordance with PCR-based molecular serotyping) for serotyping GBS regardless of whether a mapping- or assembly-based strategy is used. This database can be used directly by the popular SRST2 program. Previous studies had already noted that use of the full *cps* locus, the conserved *cpsD-G* genes (51), or the variable *cpsG-K* genes (52,53) were not simultaneously suitable for use with both mapping- and assembly-based analyses. Therefore, we included portions of both the variable and conserved regions. Inclusion of a part of the conserved region enabled us to further incorporate the discriminating SNPs for subtypes of serotype III (22). Finally, to match other databases provided with SRST2, we ensured that the reference sequences had aligned 5'- and 3'-edges.

We validated the accuracy of serotyping with this database using a previously published dataset of 790 sequenced strains (typed with latex agglutination) (52), achieving equal or higher concordance than the original report, regardless of whether a mapping- or assembly-based approach was used. Calls based on analysis of assemblies appear more sensitive to sequencing depth, with lower sequencing depth correlating with an increase in non-typeable calls. To handle these cases, GBS-SBG can be set to report the next best call (i.e. with a coverage < 90%), which were all concordant with the mapping-based analysis and the gold standard serotypes reported by (52). We further validated the accuracy with a new dataset of 572 sequenced strains (typed by PCR), achieving 99.8% (571/572) concordance using either analysis method. On careful examination, the single discordance appeared to be caused by an error in the PCR-based serotyping, due to a 614 bp deletion in an otherwise canonical serotype V locus that affected the priming sites for one of the PCRs. It remains possible that the WGS call (serotype V) is also incorrect, as the 614 bp deletion affects genes important for capsule assembly, and this strain may actually be nontypable by latex agglutination. However, this example highlights the advantage of WGS in providing more data for serotyping and in being less susceptible to sequence variants or mutations that affect PCR priming sites.

We attempted to use data sets reported by the other two reported WGS serotyping methods (51,53). Unfortunately, serotype data for individual strains was not reported

in (53); however, the number of strains for each serotype was reported in aggregate, and WGS-based serotyping using our new database, with either analysis method, reproduced a similar serotype distribution (data not shown). We were unable to locate sequencing data for the strains analyzed by (51) in public repositories.

In conclusion, we found that use of our new database enables accurate WGS-based serotyping of GBS by both mapping- and assembly-based strategies. The mapping-based strategy relies on the commonly used SRST2 tool. The ability to also use an assembly-based strategy with the same reference database increases the generality and anticipates a continued transition to long-read sequencing. Finally, this new database provides a further advantage over all previously reported databases in allowing subtyping of serotype III.

Author contributions

Conceptualization, Methodology: S.L.C.; Software: S.T., S.L.C.; Validation: S.T., T.W.Y.; Formal Analysis: S.T., S.L.C., Investigation: S.T., T.W.Y.; Resources: T.B., S.L.C.; Data Curation: S.T., T.W.Y., T.B., S.L.C.; Writing – Original Draft Preparation: S.T., S.L.C., Writing – Review and Editing: S.T., T.B., S.L.C., Visualisation: S.T., S.L.C., Supervision: T.B., S.L.C., Project Administration: T.B., S.L.C., Funding: T.B., S.L.C.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

This work was supported by the National Medical Research Council, Ministry of Health, Singapore (grants NMRC/CIRG/1467/2017 and CIRG19NOV-0024); the Temasek Foundation Innovates through its Singapore Millennium Foundation Research Grant Programme; and the Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research (A*STAR).

Acknowledgements

434 We would like to thank the members of the Chen lab for helpful comments and
435 inputs. We thank Kurosh Mehershahi and the Genome Institute of Singapore's Next
436 Generation Sequencing Platform for assisting with library preparation and
437 sequencing.

438

439

Figures and figure legends

Figure 1. Genetic organization of the GBS *cps* locus and serotyping database strategies. (a) Colored arrows at the top depict the genetic organization of the serotype II sequence. Arrows are drawn to scale, as indicated by the axis at the top. Alignments for the other 9 serotype sequences (indicated by Roman numerals on the left) are shown in gray bars below. Vertical red lines indicate single nucleotide polymorphisms relative to the serotype II sequence. Black bars indicate genes that are absent from a given serotype. Orange arrows indicate genes that are unique to a given serotype, and dotted lines indicate where they fall relative to the serotype II reference. All sequences used for this figure panel are taken from (52). The black lines demarcated by diamonds at each end at the bottom of the panel indicate the region of the *cps* locus spanned by the serotype sequences in the Sheppard database, Kapatai database, and GBS-SBG databases, as indicated to the right of each line. (b) Colored arrows depict the genetic organization of the serotype III sequence from (52), using the same scale bar as for (a). Dotted lines between the panels show relative alignment of the serotype II and serotype III reference sequences. Gray bars and vertical red lines indicate alignment and SNPs for the serotype III subtypes (indicated on the left). The yellow box indicates the region in which SNPs reported by (22) are located.

Figure 2. Phylogenetic relationship of serotype reference sequences in the GBS-SBG database. A maximum likelihood tree is shown; a total of 13 serotype sequences were used to plot an unrooted binary tree. The scale bar for branch length at the bottom indicates the number of substitutions per site. There were a total of 11911 positions in the alignment. The labels for serotypes Ia and subtypes of serotype III are highlighted in blue for ease of visualization.

References

1. Manning SD, Neighbors K, Tallman PA, Gillespie B, Marrs CF, Borchardt SM, et al. Prevalence of group B streptococcus colonization and potential for transmission by casual contact in healthy young men and women. *Clin Infect Dis* [Internet]. 2004 Aug 1;39(3):380–8. Available from: <http://dx.doi.org/10.1086/422321>
2. Foxman B, Gillespie B, Manning SD, Howard LJ, Tallman P, Zhang L, et al. Incidence and duration of group B Streptococcus by serotype among male and female college students living in a single dormitory. *Am J Epidemiol* [Internet]. 2006 Mar 15;163(6):544–51. Available from: <http://dx.doi.org/10.1093/aje/kwj075>
3. Bliss SJ, Manning SD, Tallman P, Baker CJ, Pearlman MD, Marrs CF, et al. Group B Streptococcus colonization in male and nonpregnant female university students: a cross-sectional prevalence study. *Clin Infect Dis* [Internet]. 2002 Jan 15;34(2):184–90. Available from: <http://dx.doi.org/10.1086/338258>
4. Baker CJ, Barrett FF, Gordon RC, Yow MD. Suppurative meningitis due to streptococci of Lancefield group B: a study of 33 infants. *J Pediatr* [Internet]. 1973 Apr;82(4):724–9. Available from: [http://dx.doi.org/10.1016/s0022-3476\(73\)80606-7](http://dx.doi.org/10.1016/s0022-3476(73)80606-7)
5. Barton LL, Feigin RD, Lins R. Group B beta hemolytic streptococcal meningitis in infants. *J Pediatr* [Internet]. 1973 Apr;82(4):719–23. Available from: [http://dx.doi.org/10.1016/s0022-3476\(73\)80605-5](http://dx.doi.org/10.1016/s0022-3476(73)80605-5)
6. Berardi A, Tzialla C, Riva M, Cerbo RM, Creti R. Group B streptococcus: early- and late-onset infections. *J Chemother* [Internet]. 2007 Oct;19 Suppl 2:24–7. Available from: <http://dx.doi.org/10.1080/1120009x.2007.11782439>
7. Nanduri SA, Petit S, Smelser C, Apostol M, Alden NB, Harrison LH, et al. Epidemiology of Invasive Early-Onset and Late-Onset Group B Streptococcal Disease in the United States, 2006 to 2015: Multistate Laboratory and Population-Based Surveillance. *JAMA Pediatr* [Internet]. 2019 Mar 1;173(3):224–33. Available from: <http://dx.doi.org/10.1001/jamapediatrics.2018.4826>
8. Farley MM, Harvey RC, Stull T, Smith JD, Schuchat A, Wenger JD, et al. A population-based assessment of invasive disease due to group B Streptococcus in nonpregnant adults. *N Engl J Med* [Internet]. 1993 Jun 24;328(25):1807–11. Available from: <http://dx.doi.org/10.1056/NEJM199306243282503>
9. Francois Watkins LK, McGee L, Schrag SJ, Beall B, Jain JH, Pondo T, et al. Epidemiology of Invasive Group B Streptococcal Infections Among Nonpregnant Adults in the United States, 2008-2016. *JAMA Intern Med* [Internet]. 2019 Apr 1;179(4):479–88. Available from: <http://dx.doi.org/10.1001/jamainternmed.2018.7269>
10. Phares CR, Lynfield R, Farley MM, Mohle-Boetani J, Harrison LH, Petit S, et al.

- Epidemiology of invasive group B streptococcal disease in the United States, 1999-2005. JAMA [Internet]. 2008 May 7;299(17):2056–65. Available from: <http://dx.doi.org/10.1001/jama.299.17.2056>
11. Schuchat A. Epidemiology of group B streptococcal disease in the United States: shifting paradigms. Clin Microbiol Rev [Internet]. 1998 Jul;11(3):497–513. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/9665980>
12. Tan S, Lin Y, Foo K, Koh HF, Tow C, Zhang Y, et al. Group B Streptococcus Serotype III Sequence Type 283 Bacteremia Associated with Consumption of Raw Fish, Singapore. Emerg Infect Dis [Internet]. 2016 Nov;22(11):1970–3. Available from: <http://dx.doi.org/10.3201/eid2211.160210>
13. Kalimuddin S, Chen SL, Lim CTK. 2015 Epidemic of Severe Streptococcus agalactiae Sequence Type 283 Infections in Singapore Associated With the Consumption of Raw Freshwater Fish: A Clin Infect Dis [Internet]. 2017; Available from: https://academic.oup.com/cid/article-abstract/64/suppl_2/S145/3782664
14. Barkham T, Zadoks RN, Azmai MNA, Baker S, Bich VTN, Chalker V, et al. One hypervirulent clone, sequence type 283, accounts for a large proportion of invasive Streptococcus agalactiae isolated from humans and diseased tilapia in Southeast Asia. PLoS Negl Trop Dis [Internet]. 2019 Jun;13(6):e0007421. Available from: <http://dx.doi.org/10.1371/journal.pntd.0007421>
15. Bishop EJ, Shilton C, Benedict S, Kong F, Gilbert GL, Gal D, et al. Necrotizing fasciitis in captive juvenile Crocodylus porosus caused by Streptococcus agalactiae: an outbreak and review of the animal and human literature. Epidemiol Infect [Internet]. 2007 Nov;135(8):1248–55. Available from: <http://dx.doi.org/10.1017/S0950268807008515>
16. Edelstein RM, Pegram RG. Contagious skin necrosis of Somali camels associated with Streptococcus agalactiae. Trop Anim Health Prod [Internet]. 1974 Nov;6(4):255–6. Available from: <http://dx.doi.org/10.1007/BF02383286>
17. Robinson JA, Meyer FP. Streptococcal fish pathogen. J Bacteriol [Internet]. 1966;92(2):512. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/pmc276271/>
18. Tavella A, Bettini A, Cocchi M, Idrizi I, Colorio S, Viel L, et al. Isolation of Streptococcus agalactiae in a female llama (Lama glama) in South Tyrol (Italy). BMC Vet Res [Internet]. 2018 Nov 13;14(1):343. Available from: <http://dx.doi.org/10.1186/s12917-018-1676-9>
19. Evans JJ, Bohnsack JF, Klesius PH, Whiting AA, Garcia JC, Shoemaker CA, et al. Phylogenetic relationships among Streptococcus agalactiae isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan. J Med Microbiol [Internet]. 2008 Nov;57(Pt 11):1369–76. Available from: <http://dx.doi.org/10.1099/jmm.0.47815-0>

20. Sørensen UBS, Poulsen K, Ghezzi C, Margarit I, Kilian M. Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones. MBio [Internet]. 2010 Aug 24;1(3). Available from: <http://dx.doi.org/10.1128/mBio.00178-10>
21. Wilkinson HW, Moody MD. Serological relationships of type I antigens of group B streptococci. J Bacteriol [Internet]. 1969 Feb;97(2):629–34. Available from: <http://dx.doi.org/10.1128/JB.97.2.629-634.1969>
22. Kong F, Gowan S, Martin D, James G, Gilbert GL. Serotype Identification of Group B Streptococci by PCR and Sequencing [Internet]. Vol. 40, Journal of Clinical Microbiology. 2002. p. 216–26. Available from: <http://dx.doi.org/10.1128/jcm.40.1.216-226.2002>
23. Imperi M, Pataracchia M, Alfarone G, Baldassarri L, Orefici G, Creti R. A multiplex PCR assay for the direct identification of the capsular type (Ia to IX) of *Streptococcus agalactiae*. J Microbiol Methods [Internet]. 2010 Feb;80(2):212–4. Available from: <http://dx.doi.org/10.1016/j.mimet.2009.11.010>
24. Chen SL. Genomic Insights Into the Distribution and Evolution of Group B Streptococcus. Front Microbiol [Internet]. 2019 Jun 28;10:1447. Available from: <http://dx.doi.org/10.3389/fmicb.2019.01447>
25. Yang Y, Liu Y, Ding Y, Yi L, Ma Z, Fan H, et al. Molecular characterization of *Streptococcus agalactiae* isolated from bovine mastitis in Eastern China. PLoS One [Internet]. 2013 Jul 10;8(7):e67755. Available from: <http://dx.doi.org/10.1371/journal.pone.0067755>
26. Zhao Z, Kong F, Martinez G, Zeng X, Gottschalk M, Gilbert GL. Molecular serotype identification of *Streptococcus agalactiae* of bovine origin by multiplex PCR-based reverse line blot (mPCR/RLB) hybridization assay. FEMS Microbiol Lett [Internet]. 2006 Oct;263(2):236–9. Available from: <http://dx.doi.org/10.1111/j.1574-6968.2006.00428.x>
27. Leal CAG, Queiroz GA, Pereira FL, Tavares GC, Figueiredo HCP. *Streptococcus agalactiae* Sequence Type 283 in Farmed Fish, Brazil. Emerg Infect Dis [Internet]. 2019 Apr;25(4):776–9. Available from: <http://dx.doi.org/10.3201/eid2504.180543>
28. Finch LA, Martin DR. Human and bovine group B streptococci: two distinct populations. J Appl Bacteriol [Internet]. 1984 Oct;57(2):273–8. Available from: <http://dx.doi.org/10.1111/j.1365-2672.1984.tb01391.x>
29. Oliveira ICM, De Mattos MC, Areal MFT, Ferreira-Carvalho BT, Figueiredo AMS, Benchetrit LC. Pulsed-field gel electrophoresis of human group B streptococci isolated in Brazil. J Chemother [Internet]. 2005 Jun;17(3):258–63. Available from: <http://dx.doi.org/10.1179/joc.2005.17.3.258>
30. Ip M, Cheuk ESC, Tsui MHY, Kong F, Leung TN, Gilbert GL. Identification of a *Streptococcus agalactiae* serotype III subtype 4 clone in association with adult invasive disease in Hong Kong. J Clin Microbiol [Internet]. 2006

- Nov;44(11):4252–4. Available from: <http://dx.doi.org/10.1128/JCM.01533-06>
31. Pereira UP, Mian GF, Oliveira ICM, Benchetrit LC, Costa GM, Figueiredo HCP. Genotyping of Streptococcus agalactiae strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. Vet Microbiol [Internet]. 2010 Jan 6;140(1-2):186–92. Available from: <http://dx.doi.org/10.1016/j.vetmic.2009.07.025>
32. Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia [Internet]. 2011 Dec;16(4):357–72. Available from: <http://dx.doi.org/10.1007/s10911-011-9236-y>
33. Kalimuddin S, Chen SL, Lim CTK, Koh TH, Tan TY, Kam M, et al. 2015 Epidemic of Severe Streptococcus agalactiae Sequence Type 283 Infections in Singapore Associated With the Consumption of Raw Freshwater Fish: A Detailed Analysis of Clinical, Epidemiological, and Bacterial Sequencing Data. Clin Infect Dis [Internet]. 2017 May 15;64(suppl_2):S145–52. Available from: <http://dx.doi.org/10.1093/cid/cix021>
34. Liu G, Zhu J, Chen K, Gao T, Yao H, Liu Y, et al. Development of Streptococcus agalactiae vaccines for tilapia. Dis Aquat Organ [Internet]. 2016 Dec 21;122(2):163–70. Available from: <http://dx.doi.org/10.3354/dao03084>
35. Arakere G, Flores AE, Ferrieri P, Frasch CE. Inhibition enzyme-linked immunosorbent assay for serotyping of group B streptococcal isolates. J Clin Microbiol [Internet]. 1999 Aug;37(8):2564–7. Available from: <http://dx.doi.org/10.1128/JCM.37.8.2564-2567.1999>
36. Cropp CB, Zimmerman RA, Jelinkova J, Auernheimer AH, Bolin RA, Wyrick BC. Serotyping of group B streptococci by slide agglutination fluorescence microscopy, and microimmunodiffusion. J Lab Clin Med [Internet]. 1974 Oct;84(4):594–603. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/4213377>
37. Håkansson S, Burman LG, Henrichsen J, Holm SE. Novel coagglutination method for serotyping group B streptococci. J Clin Microbiol [Internet]. 1992 Dec;30(12):3268–9. Available from: <http://dx.doi.org/10.1128/JCM.30.12.3268-3269.1992>
38. Holm SE, Håkansson S. A simple and sensitive enzyme immunoassay for determination of soluble type-specific polysaccharide from group B streptococci. J Immunol Methods [Internet]. 1988 Jan 21;106(1):89–94. Available from: [http://dx.doi.org/10.1016/0022-1759\(88\)90275-x](http://dx.doi.org/10.1016/0022-1759(88)90275-x)
39. Zuerlein TJ, Christensen B, Hall RT. Latex agglutination detection of group-B streptococcal inoculum in urine. Diagn Microbiol Infect Dis [Internet]. 1991 May;14(3):191–4. Available from: [http://dx.doi.org/10.1016/0732-8893\(91\)90031-a](http://dx.doi.org/10.1016/0732-8893(91)90031-a)
40. Ke D, Ménard C, Picard FJ, Boissinot M, Ouellette M, Roy PH, et al.

Development of conventional and real-time PCR assays for the rapid detection of group B streptococci. Clin Chem [Internet]. 2000 Mar;46(3):324–31. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/10702518>

41. Nagano Y, Nagano N, Takahashi S, Murono K, Fujita K, Taguchi F, et al. Restriction endonuclease digest patterns of chromosomal DNA from group B beta-haemolytic streptococci. J Med Microbiol [Internet]. 1991 Nov;35(5):297–303. Available from: <https://www.microbiologyresearch.org/content/journal/jmm/10.1099/00222615-35-5-297>
42. Rolland K, Marois C, Siquier V, Cattier B, Quentin R. Genetic features of Streptococcus agalactiae strains causing severe neonatal infections, as revealed by pulsed-field gel electrophoresis and hylB gene analysis. J Clin Microbiol [Internet]. 1999 Jun;37(6):1892–8. Available from: <http://dx.doi.org/10.1128/JCM.37.6.1892-1898.1999>
43. Sellin M, Olofsson C, Håkansson S, Norgren M. Genotyping of the Capsule Gene Cluster (cps) in Nontypeable Group B Streptococci Reveals Two Majorcps Allelic Variants of Serotypes III and VII [Internet]. Vol. 38, Journal of Clinical Microbiology. 2000. p. 3420–8. Available from: <http://dx.doi.org/10.1128/jcm.38.9.3420-3428.2000>
44. Ibrahim GM, Morin PM. Salmonella Serotyping Using Whole Genome Sequencing [Internet]. Vol. 9, Frontiers in Microbiology. 2018. Available from: <http://dx.doi.org/10.3389/fmicb.2018.02993>
45. Diep B, Barretto C, Portmann A-C, Fournier C, Karczmarek A, Voets G, et al. Salmonella Serotyping; Comparison of the Traditional Method to a Microarray-Based Method and an in silico Platform Using Whole Genome Sequencing Data [Internet]. Vol. 10, Frontiers in Microbiology. 2019. Available from: <http://dx.doi.org/10.3389/fmicb.2019.02554>
46. Epping L, van Tonder AJ, Gladstone RA, The Global Pneumococcal Sequencing Consortium, Bentley SD, Page AJ, et al. SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. Microb Genom [Internet]. 2018 Jul;4(7). Available from: <http://dx.doi.org/10.1099/mgen.0.000186>
47. Jenkins C. Whole-Genome Sequencing Data for Serotyping Escherichia coli—It's Time for a Change! [Internet]. Vol. 53, Journal of Clinical Microbiology. 2015. p. 2402–3. Available from: <http://dx.doi.org/10.1128/jcm.01448-15>
48. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, et al. In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. Microb Genom [Internet]. 2016 Jul;2(7):e000064. Available from: <http://dx.doi.org/10.1099/mgen.0.000064>
49. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak

- 679 detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* [Internet]. 2014
680 May;52(5):1501–10. Available from: <http://dx.doi.org/10.1128/JCM.03617-13>
- 681 50. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive Web: User-Friendly Capsule and
682 Lipopolysaccharide Serotype Prediction for *Klebsiella* Genomes. *J Clin Microbiol*
683 [Internet]. 2018 Jun;56(6). Available from: [http://dx.doi.org/10.1128/JCM.00197-](http://dx.doi.org/10.1128/JCM.00197-18)
684 18
- 685 51. Sheppard AE, Vaughan A, Jones N, Turner P, Turner C, Efstratiou A, et al.
686 Capsular Typing Method for *Streptococcus agalactiae* Using Whole-Genome
687 Sequence Data. *J Clin Microbiol* [Internet]. 2016 May;54(5):1388–90. Available
688 from: <http://dx.doi.org/10.1128/JCM.03142-15>
- 689 52. Kapatai G, Patel D, Efstratiou A, Chalker VJ. Comparison of molecular
690 serotyping approaches of *Streptococcus agalactiae* from genomic sequences.
691 *BMC Genomics* [Internet]. 2017 Jun 1;18(1):429. Available from:
692 <http://dx.doi.org/10.1186/s12864-017-3820-5>
- 693 53. Metcalf BJ, Chochua S, Gertz RE Jr, Hawkins PA, Ricaldi J, Li Z, et al. Short-
694 read whole genome sequencing for determination of antimicrobial resistance
695 mechanisms and capsular serotypes of current invasive *Streptococcus*
696 *agalactiae* recovered in the USA. *Clin Microbiol Infect* [Internet]. 2017
697 Aug;23(8):574.e7–574.e14. Available from:
698 <http://dx.doi.org/10.1016/j.cmi.2017.02.021>
- 699 54. Poyart C, Tazi A, Réglier-Poupet H, Billoët A, Tavares N, Raymond J, et al.
700 Multiplex PCR assay for rapid and accurate capsular typing of group B
701 streptococci. *J Clin Microbiol* [Internet]. 2007 Jun;45(6):1985–8. Available from:
702 <http://dx.doi.org/10.1128/JCM.00159-07>
- 703 55. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al.
704 SRST2: Rapid genomic surveillance for public health and hospital microbiology
705 labs. *Genome Med* [Internet]. 2014 Nov 20;6(11):90. Available from:
706 <http://dx.doi.org/10.1186/s13073-014-0090-6>
- 707 56. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using
708 de Bruijn graphs. *Genome Res* [Internet]. 2008 May;18(5):821–9. Available
709 from: <http://dx.doi.org/10.1101/gr.074492.107>
- 710 57. Gao S, Sung W-K, Nagarajan N. Opera: reconstructing optimal genomic
711 scaffolds with high-throughput paired-end sequences. *J Comput Biol* [Internet].
712 2011 Nov;18(11):1681–91. Available from:
713 <http://dx.doi.org/10.1089/cmb.2011.0170>
- 714 58. Gao S, Bertrand D, Nagarajan N. FinIS: Improved in silico Finishing Using an
715 Exact Quadratic Programming Formulation. In: *Algorithms in Bioinformatics*
716 [Internet]. Springer Berlin Heidelberg; 2012. p. 314–25. Available from:
717 http://dx.doi.org/10.1007/978-3-642-33122-0_25
- 718 59. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
719 BLAST+: architecture and applications. *BMC Bioinformatics* [Internet]. 2009 Dec

- 15;10:421. Available from: <http://dx.doi.org/10.1186/1471-2105-10-421>
60. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* [Internet]. 2013 Apr;30(4):772–80. Available from: <http://dx.doi.org/10.1093/molbev/mst010>
61. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* [Internet]. 2009 May 1;25(9):1189–91. Available from: <http://dx.doi.org/10.1093/bioinformatics/btp033>
62. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* [Internet]. 1993 May;10(3):512–26. Available from: <http://dx.doi.org/10.1093/oxfordjournals.molbev.a040023>
63. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* [Internet]. 2018 Jun 1;35(6):1547–9. Available from: <http://dx.doi.org/10.1093/molbev/msy096>
64. Chaffin DO, Beres SB, Yim HH, Rubens CE. The serotype of type Ia and III group B streptococci is determined by the polymerase gene within the polycistronic capsule operon. *J Bacteriol* [Internet]. 2000 Aug;182(16):4466–77. Available from: <http://dx.doi.org/10.1128/jb.182.16.4466-4477.2000>
65. Kapatai G, Coelho J, Platt S, Chalker VJ. Whole genome sequencing of group A Streptococcus: development and evaluation of an automated pipeline for emmgene typing. *PeerJ* [Internet]. 2017 Apr 27;5:e3226. Available from: <http://dx.doi.org/10.7717/peerj.3226>
66. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, et al. Salmonella serotype determination utilizing high-throughput genome sequencing data. *J Clin Microbiol* [Internet]. 2015 May;53(5):1685–92. Available from: <http://dx.doi.org/10.1128/JCM.00323-15>
67. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol* [Internet]. 2015 Aug;53(8):2410–26. Available from: <http://dx.doi.org/10.1128/JCM.00008-15>

Supplementary Figures

Supplementary Figure 1. Phylogenetic tree of serotype III sequences. A maximum likelihood tree is shown; a serotype Ia sequence was used as an outgroup to root the tree. This analysis involved 175 nucleotide sequences. The scale bar for branch length at the bottom indicates the number of substitutions per site. There were a total of 6789 positions in the final dataset. Branch colors indicate the subtype of serotype III, as indicated by the legend at the top left. The sequence labelled in black and highlighted with an arrow is the reference serotype III sequence from the Kapatai database.

Supplementary Figure 2. Typeability based on best versus second-best serotype match. Each graph plots the total Bitscore (reported by BLASTN) for the second best-scoring serotype (y-axis) against the total Bitscore for the best-scoring serotype matched (x-axis). A solid black line is drawn at $y=x$; thus, points further to the right of the black line indicate better typeability (i.e. a larger difference in total Bitscore between the best and second-best match). (a) Each data point represents one strain ($n = 790$ total). Points are colored based on the best serotype matched, as indicated by the legend on the right. (b) Only strains with a best match to one of the serotype III subtypes are plotted ($n = 356$ total). Each strain is represented by two data points: one point is as plotted in panel (a), while a second data point is plotted in which the second-best match was taken to be the best non-serotype III match. Thus, the two data points for each strain are aligned vertically. Points are colored based on the serotype of the second-best match as indicated by the legend on the right.

Supplementary Figure 3. Detailed analysis of PCR versus WGS serotyping of SG-M666. Colored arrows in the middle track depict the genetic organization of the reference sequences for serotype Ia and V. Colors for each gene are the same as shown in Figure 1. Genes conserved between serotype Ia and V are shown in the middle. Genes unique to serotype Ia and serotype V are shown offset above and below, respectively, the conserved genes. Dotted lines join contiguous sequences, which are offset due to different lengths of unique genes. A nucleotide scale bar for each serotype is shown (Ia above, V below). Solid colored bars adjacent to the nucleotide scale indicate the percent nucleotide identity (as reported by BLASTN) between SG-M666 and the serotype Ia and serotype V reference sequences, according to the color legend at the bottom left. The location of the PCR amplicons used for PCR-based serotyping (Poyart C, et al. 2007, Imperi M, et al. 2010) are shown as colored lines above (for serotype Ia) and below (for serotype V) the nucleotide identity tracks; those with coordinates overlapping black bars on the “% nucleotide identity” track are predicted to yield no amplification. Of the three PCRs from Imperi, et al. 2010, the PCR that distinguishes serotype V from serotype I (at ~4.6-5.2kb on the serotype V scale; highlighted with an asterisk) is predicted to be negative, thus mimicking a serotype I result.

Supplementary Figure 4. Box plot showing the distribution of the average read depths in the Kapatai datasets. The median is indicated by the horizontal line in the middle of each box and the 75th and 25th percentiles indicated by the top and bottom borders of the box, respectively. Whiskers indicate 1.5x the inter-quartile range. Outliers are represented by individual data points. The Mann–Whitney U test was used to compare the median read depths of the non-typeable ($n=43$) and typeable ($n=747$) datasets ($p < 1.044e-16$).

Supplementary Figure 5. Detailed analysis of a GBS-SBG non-typeable and typeable dataset. Each graph plots the read depth per reference base (y-axis) against the serotype II reference coordinate (x-axis). A solid red line is drawn at y=5, which is the minimum read depth required by SRST2 to consider a base “covered”. (a) PHEGBS0096 (ERR1741774), which was typed as serotype II by SRST2 using short reads (20.967x average read depth). This was called as nontypeable by GBS-SBG using the assembly; the next best hit was serotype II, with 99.9% identity over 68.5% coverage. (b) PHEGBS0074 (ERR1742026) was typed as serotype II by both SRST2 (for short reads) (23.678x average read depth) and GBS-SBG (for assemblies). Note that the coverage does not fall as low as it does for PHEGBS0096, and the assembly of the cps locus is more complete. Solid colored bars at the bottom of each graph indicate the percent identity (as reported by BLASTN) between the serotype II reference sequence (8315bp) and PHEGBS0096 and PHEGBS0074 samples, according to the legend between the panels.