JCM Accepted Manuscript Posted Online 9 March 2016
J. Clin. Microbiol. doi:10.1128/JCM.03142-15

Capsular typing method for *Streptococcus agalactiae* using whole genome sequence data

Running title: GBS capsular typing using whole genome sequence data

Anna E Sheppard,[1#] Alison Vaughan,[1] Nicola Jones,[1] Paul Turner,[2,3] Claudia Turner,[2,3] Androulla Efstratiou,[4,5] Darshana Patel,[4] the Modernising Medical Microbiology (MMM) Informatics Group,[1] A Sarah Walker,[1] James A Berkley,[2,6] Derrick W Crook,[1] Anna C Seale.[2,6#]

1. Modernising Medical Microbiology Consortium, Nuffield Department of Clinical Medicine, University of Oxford, UK.
2. Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, UK.
3. Shoklo Malaria Research Unit, Thailand
4. Microbiology Reference Division, Public Health England, London, UK
5. Imperial College, London, UK
6. KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya.

[#]Corresponding authors: Anna E Sheppard (anna.sheppard@ndm.ox.ac.uk) or Anna C Seale (aseale@nhs.net).

Word count:

Abstract 47/50

Main text 933/1000

Abstract

Group B streptococcus (GBS) capsular serotype is a major determinant of virulence, and affects potential vaccine coverage. Here we report a whole genome sequencing-based method for GBS serotype assignment. This shows high agreement (kappa=0.92) with conventional methods, and increased serotype assignment (100%) to all ten capsular types.

Main text

Streptococcus agalactiae, or Group B Streptococcus (GBS), is an important pathogen in neonates (1-3), with early infection acquired from the maternal genito-urinary tract (4). In addition, GBS is now recognised as an increasingly important pathogen in high-income regions in immunosuppressed and elderly individuals (5, 6).

GBS expresses a capsular polysaccharide, which is involved in virulence and immune evasion. Ten different variants, or serotypes, have been described (Ia, Ib, II, III, IV, V VI, VII, VIII and IX), which differ in their disease-causing ability. Conjugate vaccines targeting the most common disease-causing serotypes are currently in development (7). Establishing vaccine serotype coverage is important, as is surveillance post-introduction to monitor for potential serotype replacement, as has been seen following the introduction of other conjugate vaccines (8).

Current methods for GBS serotype allocation rely on latex agglutination assays or PCR (9). Recent advances in whole genome sequencing (WGS) have enabled the development of approaches that can be used in place of traditional microbiological methods, such as strain typing and antibiotic susceptibility profiling (10-12). A major advantage of this approach is that the cost of sequencing can be mitigated by the ability to use the same data to generate multiple outputs. Given the decreasing cost of WGS (13), and the ongoing increase in WGS data generation, we sought to establish and validate a WGS-based method for GBS capsular typing.

We developed an algorithm for serotype assignment on the basis of sequence similarity between a given *de novo* assembly and capsular gene sequences of the ten GBS serotypes. For nine serotypes, published sequences were used as references (Table 1), while for serotype IX, only a partial capsular locus sequence has been published (14). A suitable reference for the full capsular locus region was therefore determined by WGS of a serotype IX isolate obtained from the Statens Serum Institute, Denmark.

To assign serotype for a given isolate, a BLAST database was generated from the *de novo* assembly and queried with the variable region of the capsular locus sequence for each serotype (*cpsG-cpsK* for serotypes Ia-VII and IX, *cpsR-cpsK* for serotype VIII) using BLASTn with an evalue threshold of 1e-100 and otherwise default parameters. A serotype was considered as correct if it showed ≥95% sequence identity over ≥90% of the sequence length. These thresholds were chosen on the basis of being stringent enough to provide differentiation between the various reference sequences, while maximising serotype allocation on an initial test set of publicly available GBS WGS data, where serotype information was not available (so we had no way of knowing whether the assigned serotypes were in fact correct).

This sequence-based method for serotype allocation was validated using WGS on a set of 223 colonising or invasive human isolates from Canada, Latin America, Singapore, UK, USA, and Thailand, for which serotype had previously been determined using conventional latex agglutination assays, with PCR used to confirm weak positives or negatives in a subset (15-17). For two rare serotypes (Serotype VIII and IX), one isolate of each was obtained from the Statens Serum Institute, Denmark. GBS isolates stored at -80°C were sub-cultured on Columbia blood agar for 24-48 hours, followed by DNA extraction from a single colony using a commercial kit (QuickGene, Fujifilm, Tokyo, Japan). High throughput sequencing was undertaken at the Wellcome Trust Centre for Human Genetics (Oxford University, UK) using the Illumina HiSeq2500 platform, generating 150 base paired-end reads. *De novo* assembly was performed using Velvet and VelvetOptimiser (18, 19). Agreement between serotype allocations was tested with the Kappa statistic.

High quality sequence data were obtained for all 223 GBS isolates (median read number: 2,975,508, range: 1,798,744-13,073,718; median contig number: 46, range 16-106; median assembly length: 2.05 Mb, range: 1.94-2.22 Mb). Each isolate was allocated to a single serotype using the WGS data (Table 2). Three isolates that did not have a capsular type assigned by latex agglutination methods had serotypes Ib, VI and VIII assigned. For all previously serotyped GBS isolates with a known capsule type, the kappa statistic (0.92) indicated very high agreement between WGS-predicted and conventional serotype. There were nine discordant isolates. In each case there was strong support for the sequence-allocated serotype, with >98% sequence identity over 100% of the reference length in all nine cases (Figure 1). Across all isolates, differences in relatedness between the capsular locus sequences of the different serotypes led to characteristic relationships between the allocated (best match) serotype and the second-best match. For example, all isolates assigned as serotype Ia had serotype III as the second-best match. In all cases, the second-best match was substantially poorer than the best match, demonstrating that there was no ambiguity in predicted serotype (Figure 1, Table 3).

The nine discordant and three non-typeable isolates were retested by latex agglutination (Table 4) and resequenced using the Illumina MiSeq platform, with sequence processing and WGS-based serotype prediction performed as above. In all cases, resequencing was consistent with the initial WGS classification. For 6/9 discordant isolates, the new latex agglutination results matched the WGS-based prediction, suggesting that the initial discordance may have resulted from incorrect latex agglutination typing or sample mislabelling. The other three initially discordant isolates, and the three non-typeable isolates, were all classified as non-typeable on retesting.

This WGS-based method for GBS serotyping, validated using 223 isolates typed using conventional methods, was therefore highly accurate. Although WGS may not currently be cost-effective for directly replacing traditional serotyping, costs are likely to further decrease. Furthermore, WGS may already be the cheapest option for combined studies, with possibilities to utilise the resulting data for additional analyses such as multi-locus sequence typing, relatedness to other sequenced isolates, and detailed phylogenetic analysis.

Funding

This publication presents independent research commissioned by The Wellcome Trust ([www.wellcome.ac.uk 093804, 098532] ACS, JAB), the Health Innovation Challenge Fund (grant HICF-T5-358 and WT098615/Z/12/Z), a parallel funding partnership between the Department of Health and the Wellcome Trust (DWC, AES and ASW); the UK Clinical Research Collaboration (a parallel funding partnership between the Medical Research Council [G0800778], Biotechnology and Biological Sciences Research Council and the Wellcome Trust [087646/Z/08/Z]); and the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. DWC is a NIHR (UK) Senior Investigator. The views expressed in this publication are those of the author(s) and not necessarily those of the funders.

Acknowledgements

We would like to thank The Wellcome Trust Centre for Human Genetics, University of Oxford where the whole genome sequencing was done, and we thank the library and sequencing teams. We thank the Shoklo Malaria Research Unit, Thailand for providing GBS isolates.

Modernising Medical Microbiology (MMM) informatics group

Jim Davies, Charles Crichton, Milind Acharya, Carlos del Ojo Elias

References

1. **Baker CJ, Barrett FF, Gordon RC, Yow MD.** 1973. Suppurative meningitis due to streptococci of Lancefield group B: a study of 33 infants. J Pediatr **82:**724-729.

2. **Barton LL, Feigin RD, Lins R.** 1973. Group B beta hemolytic streptococcal meningitis in infants. J Pediatr **82:**719-723.

3. **Communicable Disease Surveillance Centre London.** 1985. Neonatal meningitis: a review of routine national data 1975-83. Br Med J (Clin Res Ed) **290:**778-779.

4. **Dillon HC, Jr., Gray E, Pass MA, Gray BM.** 1982. Anorectal and vaginal carriage of group B streptococci during pregnancy. J Infect Dis **145:**794-799.

5. **Schuchat A.** 1998. Epidemiology of group B streptococcal disease in the United States: shifting paradigms. Clin Microbiol Rev **11:**497-513.

6. **Phares CR, Lynfield R, Farley MM, Mohle-Boetani J, Harrison LH, Petit S, Craig AS, Schaffner W, Zansky SM, Gershman K, Stefonek KR, Albanese BA, Zell ER, Schuchat A, Schrag SJ, Active Bacterial Core surveillance/Emerging Infections Program N.** 2008. Epidemiology of invasive group B streptococcal disease in the United States, 1999-2005. JAMA **299:**2056-2065.

7. **Madhi SA, Dangor Z, Heath PT, Schrag S, Izu A, Sobanjo-Ter Meulen A, Dull PM.** 2013. Considerations for a phase-III trial to evaluate a group B Streptococcus polysaccharide-protein conjugate vaccine in pregnant women for the prevention of early- and late-onset invasive disease in young-infants. Vaccine **31 Suppl 4:**D52-57.

8. **Mulholland K, Satzke C.** 2012. Serotype replacement after pneumococcal vaccination. Lancet **379:**1387; author reply 1388-1389.

9. **Imperi M PM, Alfarone G, Baldassarri L, Orefici G, Creti R.** 2010. A multiplex PCR assay for the direct identification of the capsular type (Ia to IX) of Streptococcus agalactiae. J Microbiol Methods **80:**212-214.

10. **Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TE, Crook DW, Walker AS, Golubchik T.** 2014. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. J Clin Microbiol **52:**1182-1191.

11. **Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TE, Crook DW.** 2013. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother **68:**2234-2244.

12. **Walker TM, Kohl TA, Omar SV, Hedge J, Del Ojo Elias C, Bradley P, Iqbal Z, Feuerriegel S, Niehaus KE, Wilson DJ, Clifton DA, Kapatai G, Ip CL, Bowden R, Drobniewski FA, Allix-Beguec C, Gaudin C, Parkhill J, Diel R, Supply P, Crook DW, Smith EG, Walker AS, Ismail N, Niemann S, Peto TE, Modernizing Medical Microbiology Informatics G.** 2015. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis doi:10.1016/S1473-3099(15)00062-6.

13. **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.

14. **Slotved HC, Kong F, Lambertsen L, Sauer S, Gilbert GL.** 2007. Serotype IX, a Proposed New Streptococcus agalactiae Serotype. J Clin Microbiol **45:**2929-2936.

15. **Bisharat N, Crook DW, Leigh J, Harding RM, Ward PN, Coffey TJ, Maiden MC, Peto T, Jones N.** 2004. Hyperinvasive neonatal group B streptococcus has arisen from a bovine ancestor. J Clin Microbiol **42:**2161-2167.

16. **Jones N, Oliver K, Jones Y, Haines A, Crook D.** 2006. Carriage of group B streptococcus in pregnant women from Oxford, UK. J Clin Pathol **59:**363-366.

17. **Davies HD, Jones N, Whittam TS, Elsayed S, Bisharat N, Baker CJ.** 2004. Multilocus sequence typing of serotype III group B streptococcus and correlation with pathogenic potential. J Infect Dis **189:**1097-1102.

18. **Gladman S, Seeman T.** 2012. VelvetOptimiser. http://bioinformatics.net.au/software.velvetoptimiser.shtml

19. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821-829.

20. **Yamamoto S, Miyake K, Koike Y, Watanabe M, Machida Y, Ohta M, Iijima S.** 1999. Molecular characterization of type-specific capsular polysaccharide biosynthesis genes of Streptococcus agalactiae type Ia. J Bacteriol **181:**5176-5184.

21. **Watanabe M, Miyake K, Yanae K, Kataoka Y, Koizumi S, Endo T, Ozaki A, Iijima S.** 2002. Molecular characterization of a novel beta1,3-galactosyltransferase for capsular polysaccharide synthesis by Streptococcus agalactiae type Ib. J Biochem **131:**183-191.

22. **Martins ER, Melo-Cristino J, Ramirez M.** 2007. Reevaluating the serotype II capsular locus of Streptococcus agalactiae. J Clin Microbiol **45:**3384-3386.

23. **Chaffin DO, Beres SB, Yim HH, Rubens CE.** 2000. The serotype of type Ia and III group B streptococci is determined by the polymerase gene within the polycistronic capsule operon. J Bacteriol **182:**4466-4477.

24. **Cieslewicz MJ, Chaffin D, Glusman G, Kasper D, Madan A, Rodrigues S, Fahey J, Wessels MR, Rubens CE.** 2005. Structural and genetic diversity of group B streptococcus capsular polysaccharides. Infect Immun **73:**3096-3103.

Table 1. Reference sequences used for sequence-based serotype allocation

Serotype	Accession	Region	Reference
Ia	AB028896.2	6982-11695	Yamamoto et al.(20)
Ib	AB050723.1	2264-6880	Watanabe et al.(21)
II	EF990365.1	1915-8221	Martins et al.(22)
III	AF163833.1	6592-11193	Chaffin et al.(23)
IV	AF355776.1	6417-11656	Cieslewicz et al.(24)
V	AF349539.1	6400-12547	Cieslewicz et al.(24)
VI	AF337958.1	6437-10913	Cieslewicz et al.(24)
VII	AY376403.1	3403-8666	Cieslewicz et al.(24)
VIII	AY375363.1	2971-7340	Cieslewicz et al.(24)
IX	NA	NA	This study

Table 2. Serotype allocation by WGS to serotype allocation by latex agglutination

		Serotype allocated by WGS										
		Ia	**Ib**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**Total**
Serotype by latex agglutination	**Ia**	**34**	0	0	1	0	0	0	0	0	0	**35**
	Ib	0	**9**	1	0	0	0	0	0	0	0	**10**
	II	0	0	**25**	0	0	0	0	0	0	0	**25**
	III	3	0	0	**111**	0	0	0	0	0	1	**115**
	IV	0	0	0	0	**1**	0	1	0	0	0	**2**
	V	0	0	0	0	0	**16**	0	0	0	0	**16**
	VI	0	0	0	0	0	1	**8**	0	0	0	**9**
	VII	0	0	0	0	0	0	0	**5**	0	0	**5**
	VIII	0	0	0	0	0	0	0	0	**1***	0	**1**
	IX	0	1	0	0	0	0	0	0	0	**1***	**2**
	Non-typeable	0	1	0	0	0	0	1	1	0	0	**3**
	Total	37	11	26	112	1	17	10	6	1	2	**223**

*Reference GBS isolates from Statens Serum Institute serotypes VIII and IX

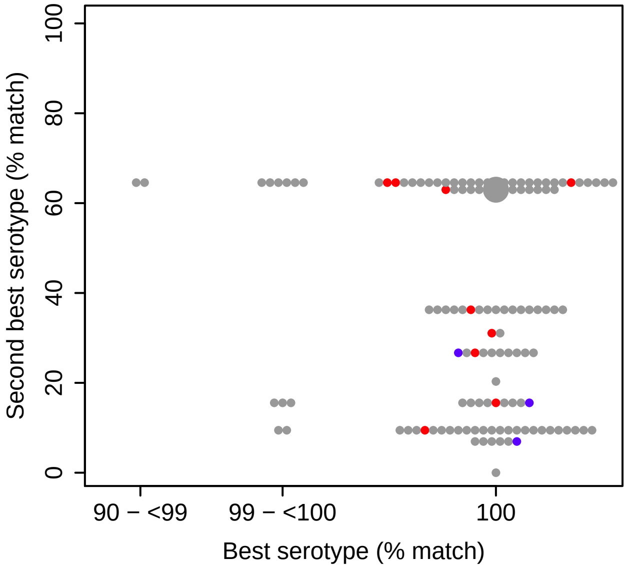
Table 3. Relationship between allocated serotype and second-best match (see also Figure 1)

Allocated serotype	**% match**	**Second-best serotype**	**% match**
Ia	93.91-100	III	64.56
III	100	Ia	62.98
V	100	IX	36.26
IX	100	V	31.05
VI	100	III	26.68
IV	100	Ia	20.3
Ib	99.61-100	VI	15.55
II	99.86-100	IV	9.45
VII	100	Ib	6.95
VIII	100	none	0

Table 4. Retyping of discordant and non-typable isolates

		Latex agglutination		WGS	
Isolate	**Reason for retyping**	**Initial**	**Repeat**	**Initial**	**Repeat**
CB466	Discordant	III	Ia	Ia	Ia
IW8194	Discordant	III	IX	IX	IX
IW8466	Discordant	Ia	III	III	III
IW8471	Discordant	III	Ia	Ia	Ia
IW7157	Discordant	Ib	II	II	II
SMRU1	Discordant	VI	V	V	V
SMRU25	Discordant	IV	NT	VI	VI
SMRU4	Discordant	IX	NT	Ib	Ib
SMRU59	Discordant	III	NT	Ia	Ia
Z41	Non-typeable	NT	NT	Ib	Ib
UK22	Non-typeable	NT	NT	VII	VII
IW2723	Non-typeable	NT	NT	VI	VI
CB454	Control	III	III	III	III
IW4445	Control	Ia	Ia	Ia	Ia
IW4077	Control	II	II	II	II

Figure 1 Discordant isolates show high support for sequence-based serotype allocation. For each isolate, the percentage of the capsular locus region present (≥95% sequence identity) for the assigned serotype is shown on the X axis, and that for the serotype showing the next best match on the Y axis. Isolates showing agreement between sequence-based and conventional serotyping are shown in grey, those classified as non-typeable by conventional methods in blue, and discordant isolates in red. Small circles represent single isolates, the large circle represents 100 isolates. For each serotype, the second-best match is identical in all cases, leading to the observed horizontal banding (details in Table 3).


Second best serotype (% match)
0
20
40
60
80
100
90 − <99
99 − <100
100
Best serotype (% match)