



Population genomics of Group B *Streptococcus* reveals the genetics of neonatal disease onset and meningeal invasion

Chrispin Chaguza ^{1,2,8✉}, Dorota Jamrozy ^{1,8}, Merijn W. Bijlsma^{3,8}, Taco W. Kuijpers^{4,5}, Diederik van de Beek³, Arie van der Ende ^{6,7,9✉} & Stephen D. Bentley ^{1,9✉}

Group B *Streptococcus* (GBS), or *Streptococcus agalactiae*, is a pathogen that causes preterm births, stillbirths, and acute invasive neonatal disease burden and mortality. Here, we investigate bacterial genetic signatures associated with disease onset time and meningeal tissue infection in acute invasive neonatal GBS disease. We carry out a genome-wide association study (GWAS) of 1,338 GBS isolates from newborns with acute invasive disease; the isolates had been collected annually, for 30 years, through a national bacterial surveillance program in the Netherlands. After controlling for the population structure, we identify genetic variation within noncoding and coding regions, particularly the capsule biosynthesis locus, statistically associated with neonatal GBS disease onset time and meningeal invasion. Our findings highlight the impact of integrating microbial population genomics and clinical pathogen surveillance, and demonstrate the effect of GBS genetics on disease pathogenesis in neonates and infants.

¹Parasites and Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. ²Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, CT, USA. ³Department of Neurology, Amsterdam Neuroscience, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ⁴Department of Immunopathology, Sanquin Research and Landsteiner Laboratory of the Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ⁵Department of Paediatric Haematology, Immunology and Infectious Diseases, Emma Children's Hospital, Amsterdam University Medical Center, Amsterdam, The Netherlands. ⁶Department of Medical Microbiology, Amsterdam Infection and Immunity, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands. ⁷Netherlands Reference Laboratory for Bacterial Meningitis, Center of Infection and Immunity Amsterdam, Amsterdam University Medical Center, Amsterdam, The Netherlands. ⁸These authors contributed equally: Chrispin Chaguza, Dorota Jamrozy, Merijn W. Bijlsma. ⁹These authors jointly supervised this work: Arie van der Ende, Stephen D. Bentley. ✉email: cc19@sanger.ac.uk; a.vanderende@amsterdamumc.nl; sdb@sanger.ac.uk

Group B *Streptococcus* (GBS), or *Streptococcus agalactiae*, is an emerging β -haemolytic pathogen, which causes substantial neonatal disease burden and mortality worldwide¹. Global estimates showed that GBS colonises approximately 21 million pregnant women annually, leading to ascending infections associated with approximately 3.5 million preterm births and more than 57,000 fetal infections and stillbirths^{2–5}. In neonates and infants, GBS is a cause of approximately 319,000 invasive disease episodes globally on a yearly basis; however, this underestimates the true global disease burden, especially in low-income countries, where little or no data has been reported³. These acute invasive neonatal GBS diseases include pneumonia, bacteraemia, and meningitis; broadly classified, based on the time of occurrence, as early-onset disease (EOD) and late-onset disease (LOD), occurring within 0 to 6 and 7 to 89 days after birth, respectively^{6–8}. To reduce the risk for vertical transmission of GBS at birth^{9–11}, risk-based or universal screening and intrapartum antibiotic prophylaxis for pregnant women with risk factors are implemented in the third trimester of pregnancy, particularly in high-income countries^{12,13}. Despite this, intrapartum antibiotics are ineffective against GBS-associated stillbirths¹⁴ and LOD, as seen by its increasing incidence globally^{7,15,16}, and there is conflicting evidence regarding its impact in preventing EOD^{7,17,18}. Furthermore, these interventions are less likely to be feasible in low-income settings¹⁹. Therefore, the World Health Organisation (WHO) has called for developing maternal GBS vaccines⁵, widely regarded as the most effective strategy for reducing invasive neonatal GBS diseases^{4,14}. However, no vaccine has been licensed to date, although a few candidates are undergoing preclinical development and early-phase clinical trials^{20–24}.

The sialic acid capsular polysaccharide (*cps*) is the primary virulence determinant for GBS, which promotes immune evasion by inhibiting phagocytosis²⁵, complement deposition and activation²⁶, and platelet-mediated killing^{26–28}. GBS also contains an arsenal of other virulence factors involved in immune evasion^{29–31}, toxin-mediated virulence^{32,33}, transcription regulation³⁴, and adhesion to the epithelial tissues and host cell entry^{35–37}. Except for the *cps* genes, most of the virulence genes are core genes, ubiquitously found across the GBS species. Therefore, the mere presence and absence patterns of these genes are unlikely to explain the inter-strain variability in GBS phenotypes and disease outcomes. However, accessory genes that are variable present, and allelic variation within the core genome, may contribute to inter-strain phenotypic differences and clinical manifestations, such as the onset time of disease and the tissues that are invaded. Although previous studies have reported mutations and lineage-specific genes in GBS^{38,39}, that potentially affect virulence and niche adaptation, genetic variation in GBS influencing the onset time of acute invasive neonatal disease and meningeal invasion remains poorly understood. Revealing such pathogenicity loci could accelerate the development of diagnostics, therapies, and especially vaccines¹, which are universally considered the most effective strategy to reduce GBS-associated stillbirths, preterm births, and invasive burden and sequelae in neonates globally. The application of agnostic and unbiased comparative genomic analysis approaches, particularly genome-wide association studies (GWAS), has shown remarkable potential for uncovering the genetic basis of bacterial phenotypes^{40,41}, such as disease susceptibility^{42–48}, tissue invasion^{42,49} and virulence⁵⁰, antimicrobial resistance^{51–55}, and niche adaptation^{56–58}.

In this study, we performed well-controlled GWAS of an extensive collection of GBS clinical isolates to investigate the genetic basis of the disease onset time and central nervous system (CNS) tissue invasion of GBS isolates in neonates with acute

invasive disease. We leveraged a catalogue of 1,338 whole-genome sequences of GBS isolates sampled over thirty years, 1987 to 2016, through a long-term nationwide bacterial surveillance programme in the Netherlands⁷. We show that genomic variation within and outside the capsule biosynthesis locus region influences disease onset time and CNS invasion of GBS in neonates. These findings highlight the critical role of the capsule and other genomic loci in the pathogenicity of GBS, implicating them as potential candidates for the development of capsule- and protein-based vaccines, diagnostics, and treatments to reduce the neonatal GBS burden and mortality globally.

Results

Thirty years of invasive neonatal GBS sampling through a national surveillance programme. We analysed a collection of 1,338 whole-genome sequences of GBS from clinical isolates, sampled from neonates and infants with acute invasive diseases in the Netherlands, to understand the genetic basis for the disease onset time and invasion of the central nervous system (CNS) tissue (Figs. 1a and 2; and Supplementary Data 1 and 2). The isolates were collected over 30 years (1987 to 2016) by the Netherlands Reference Laboratory for Bacterial Meningitis (NRLBM) through well-established national surveillance of meningitis and bacteraemia⁵⁹ (Fig. 1b; and Supplementary Fig. 1 and Supplementary Data 1). Of these isolates, 494 were sampled from cerebrospinal fluid (CSF), representing CNS invasion, while 844 were collected from blood or non-CNS site. When the isolates were stratified by disease onset time, 826 isolates were sampled from neonates with EOD, and 515 isolates were associated with LOD (Fig. 1b and Supplementary Fig. 2). We found the ten GBS capsular serotypes known to date⁶⁰, 134 sequence types (ST), and six clonal complexes (CC) based on the multilocus sequence typing (MLST) scheme⁶¹, and six previously defined clades or lineages^{59,62}, based on the genomic sequence clustering algorithms using Bayesian approaches⁶² (Fig. 1c and Supplementary Fig. 2). The most common serotype was III, both in EOD (51.76%) and LOD (75.53%), and CNS (78.74%), and non-CNS disease (50.47%). Besides serotype III, serotype Ia was the second most prevalent serotype found in 19.43% of the isolates. The incidence of GBS serotypes and lineages or clades, especially serotype III associated with the clonal complex [CC] 17, increased over time (Fig. 1b–d)⁶³, consistent with studies regionally^{64,65} and globally^{66,67}. The increasing trend of this CC17 lineage correlated with the emergence of multidrug resistance⁶⁴.

GBS serotypes and lineages influence disease onset time and meningeal infection.

We first compared the frequency of serotypes, clonal complexes, and lineages among the isolates stratified by the acute neonatal invasive disease onset time and CNS infection. We found five capsular serotypes relatively less common among the isolates associated with LOD than EOD, namely serotype Ia (odds ratio: 0.59, $P = 0.0004$), Ib (odds ratio: 0.45, $P = 0.002$), II (odds ratio: 0.21, $P = 4.47 \times 10^{-07}$), and V (odds ratio: 0.36, $P = 0.0009$) (Fig. 3a, Supplementary Fig. 2, and Supplementary Data 3). In contrast, serotype III showed a higher relative frequency among the LOD than EOD isolates (odds ratio: 2.88, $P = 2.07 \times 10^{-18}$), consistent with the findings elsewhere showing that serotype III is the predominant serotype associated with LOD^{65,68,69}. Overall, more LOD isolates were associated with CNS disease compared to EOD (odds ratio: 3.93, $P = 3.39 \times 10^{-31}$), a pattern supported by epidemiological studies^{68,70}. Correspondingly, the associations between serotypes and MLST clonal complexes, lineages and clades showed similar patterns (Fig. 3b, c and Supplementary Fig. 2). In terms of the type of the infected tissues, serotype III showed a higher relative

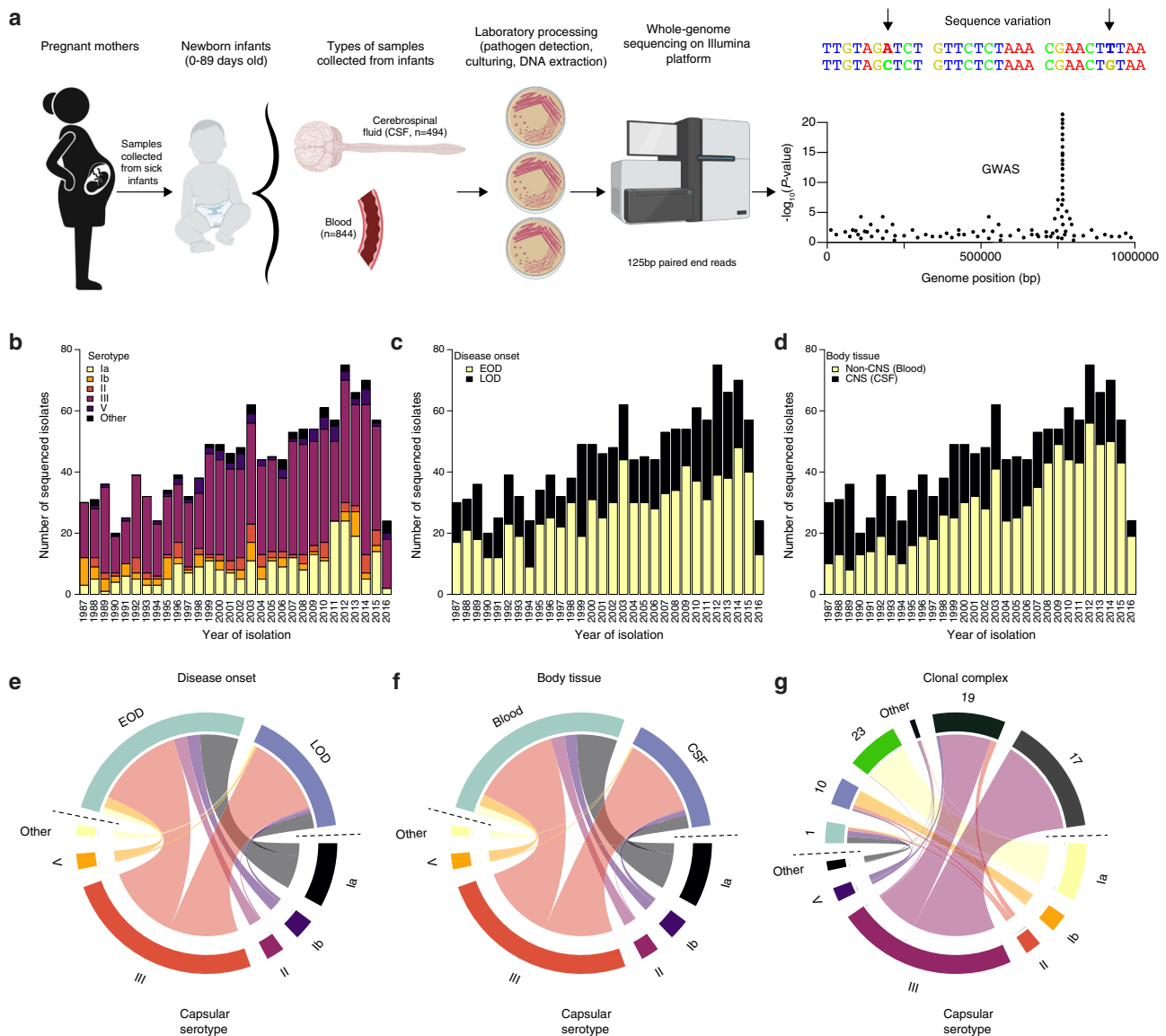


Fig. 1 Study design and characteristics of the 1,338 GBS isolates sampled from newborns over 30 years. **a** Schematic diagram showing sample collection and study design. GBS samples were collected annually for 30 years, 1987–2016, from blood and cerebrospinal fluid (CSF) of newborns aged 0–89 days through a national bacterial surveillance programme in the Netherlands. The samples were processed as described in the methods, and whole genomes were sequenced on the Illumina HiSeq platform with 125 bp reads for downstream analyses, particularly GWAS. The number of sequenced GBS isolates annually is stratified by the **b** proportion of serotypes, **c** proportion of isolates classified as EOD and LOD. **d** proportion of isolates by isolation tissue. The total number of sequenced GBS isolates stratified by **e** disease onset time and capsular serotype, **f** isolation tissue and capsular serotype, and **g** clonal complex and capsular serotype. The icons in panel **a** were created with permission in BioRender.com (<https://biorender.com/>). Source data used to generate figures in panel **b–g** is available in Supplementary Data 2.

frequency among isolates sampled from the CNS than from non-CNS sites (odds ratio: 3.62, $P = 1.75 \times 10^{-25}$). Conversely, the frequency of the other serotypes, except serotype Ib, was lower among CNS isolates than among non-CNS isolates (Fig. 3d–f). These findings showed that the capsular serotype and genetic background of GBS influence the disease onset time and invasion of the CNS.

GWAS implicates genomic loci influencing GBS disease onset time. We next investigated whether genomic variation in the GBS genome influences the onset time for acute invasive diseases in neonates and infants (Supplementary Fig. 3). Firstly, we specified the disease onset time as a categorical binary target variable, whereby EOD and LOD were defined as the affected and unaffected status, respectively. To account for the population structure,

which typically confounds bacterial GWAS analyses, if not accounted for⁴¹, we included a pairwise genetic relatedness matrix of the isolates as a random covariate in the linear mixed model. We sequenced the genomes of the isolates using the same protocol and identical read lengths to control for potential batch effects, as unequal read lengths can significantly confound bacterial GWAS⁷¹ (Supplementary Figs. 4 and 5). After correcting for multiple statistical tests, we found no single-nucleotide polymorphism (SNP), in the GBS reference genome (GenBank accession: AP018935.1), statistically associated with the disease onset time (Fig. 4a and Supplementary Data 4). To address the potential impact of frequent genetic exchanges in bacteria via recombination and horizontal gene transfer⁷², we next performed the GWAS using the presence and absence patterns of unitigs—unique high-confidence contiguous sequences⁷³. The unitigs are advantageous since they

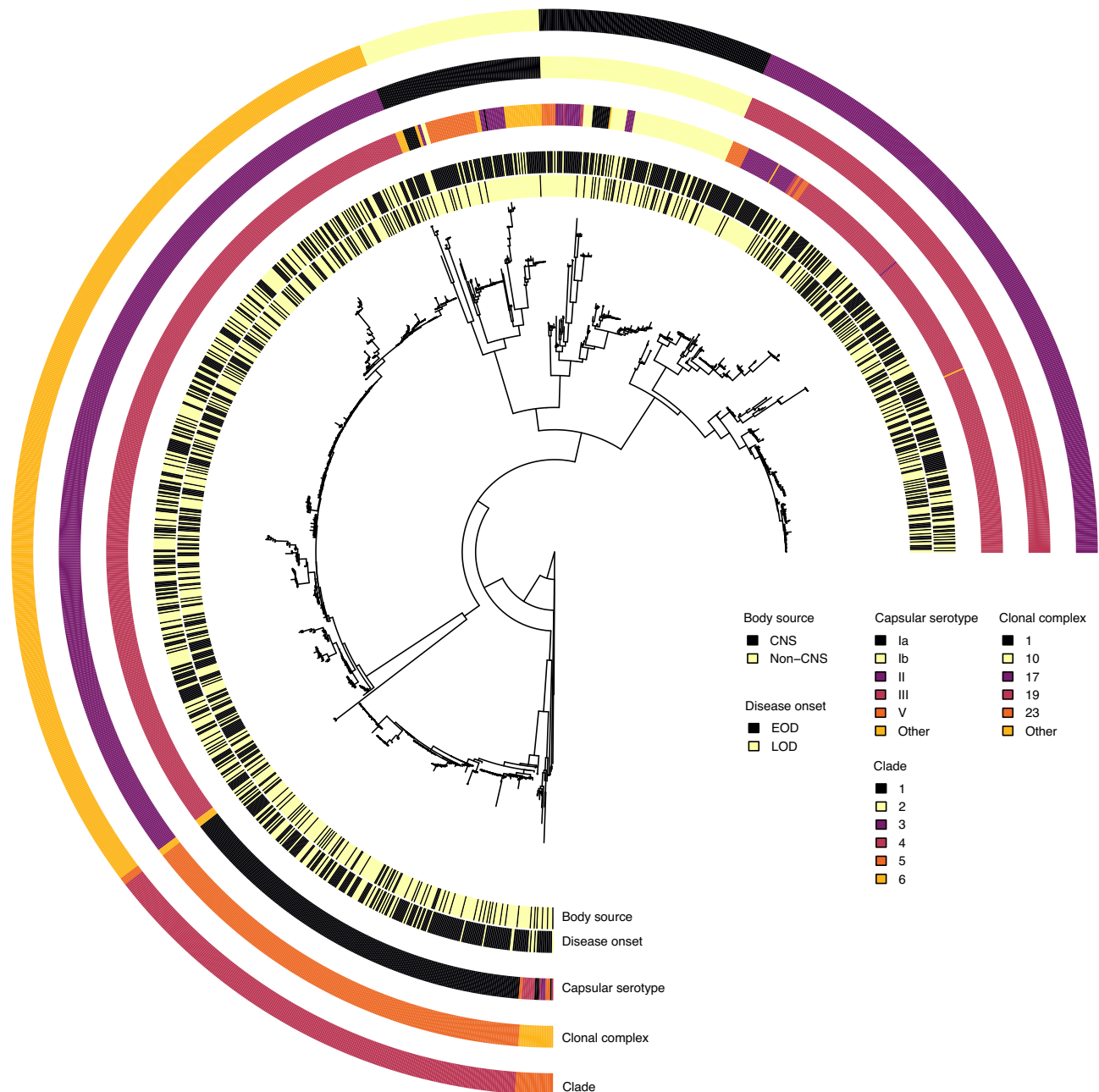


Fig. 2 Maximum-likelihood phylogenetic tree of 1,338 GBS isolates from the Netherlands. Each circular ring at the tips of the phylogenetic tree represents the body isolation site or source, disease onset time, capsular serotype, clonal complex based on the MLST approach, and the clade defined based on the Bayesian clustering approach, which yielded concordant groupings with the clonal complex. The phylogeny was rooted using a genome from a related but different species, *Streptococcus pneumoniae*, as an outgroup (not shown in the tree).

efficiently capture inter-strain genomic variation—SNPs, insertions and deletions, and genomic rearrangements—in intergenic and coding regions of the core and accessory sequences⁷⁴. We found a single unitig, located in an intergenic region, statistically associated with the categorical disease onset time of the isolates (odds ratio: 0.76, adjusted $P = 1.38 \times 10^{-07}$) (Fig. 4b and Supplementary Data 4). However, the AP018935.1 reference used for visualisation of the genomic context of the variants did not contain this unitig, therefore, it is not shown in Fig. 4b.

As the categorical classification of the GBS disease onset time is mainly for convenience, clinically, we posited that the GWAS based on the continuous values for the disease onset time would improve the statistical power to identify statistically significant

genotype–phenotype associations (Supplementary Figs. 7 and 8). Therefore, we next repeated the GWAS using the continuous target variable defined as the number of days from birth to disease onset, while similarly controlling the clonal population structure. Since the number of days from birth to disease onset is right skewed, we applied a rank-based inverse normal transformation to generate a normally distributed values to improve the power to uncover associations. We found no unitig sequences with statistically significant association with the transformed disease onset time (Fig. 4c, d). Similarly, we found no association of the accessory genes with disease onset time (Supplementary Fig. 3). The resulting Q–Q plots showed no issues for GWAS analyses due to the population structure (Fig. 4e–h). Therefore, we

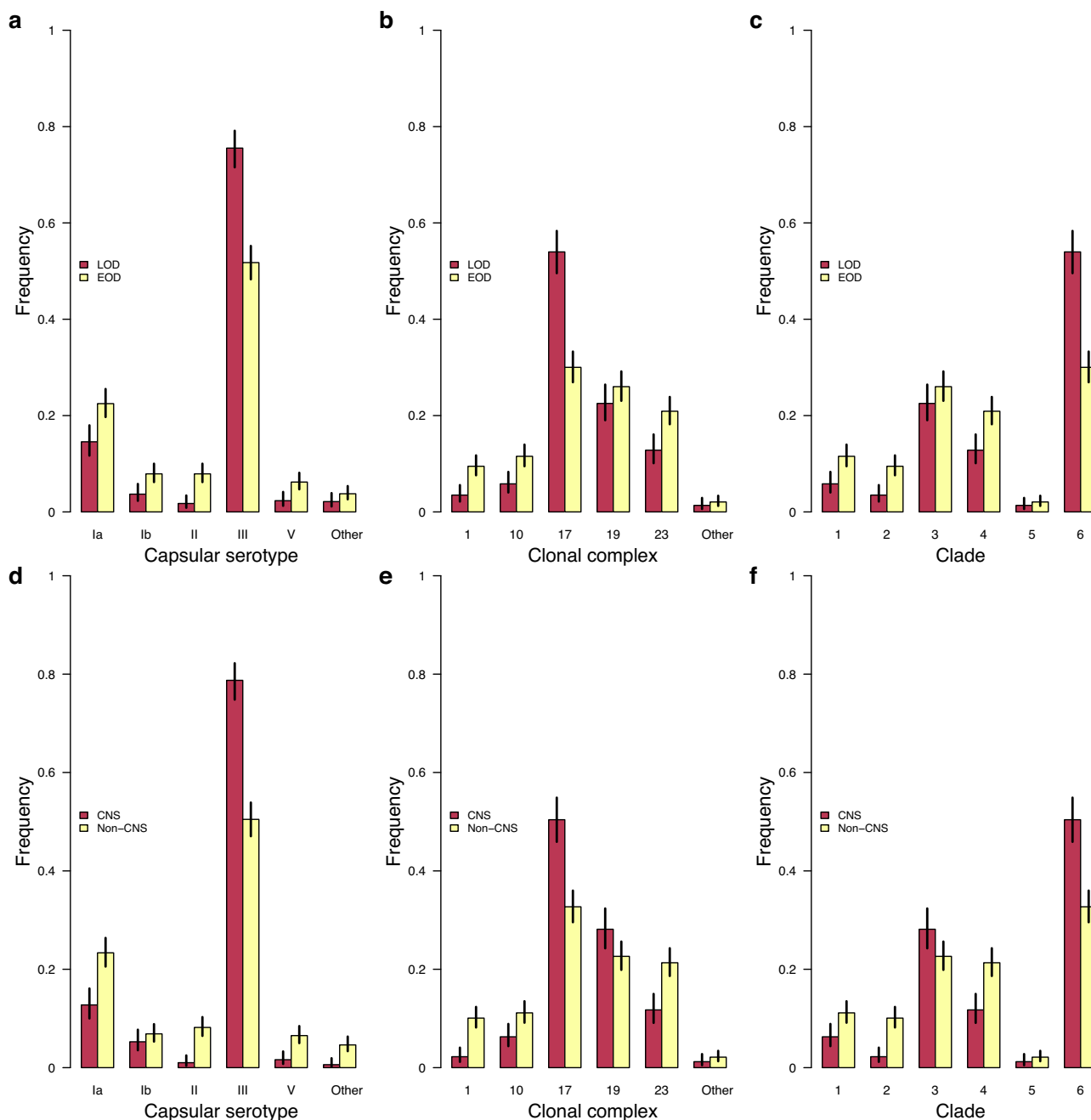


Fig. 3 Relative frequency of GBS strains stratified by disease onset time and body isolation tissue. **a** Relative frequency of GBS capsular serotypes in LOD ($n = 515$) and EOD ($n = 823$) isolates. The statistical significance for each serotype based on the test of given proportions were as follows: Ia ($P = 0.0004$), Ib ($P = 0.0017$), II ($P = 4.47 \times 10^{-07}$), III ($P = 2.07 \times 10^{-18}$), V ($P = 0.0009$), and Other ($P = 0.1082$). **b** Relative frequency of GBS clonal complexes in LOD ($n = 515$) and EOD ($n = 823$) isolates. The statistical significance for each clonal complex based on the test of given proportions were as follows: 1 ($P = 2.75 \times 10^{-05}$), 10 ($P = 0.0005$), 17 ($P = 3.75 \times 10^{-18}$), 19 ($P = 0.1710$), 23 ($P = 0.0002$), and Other ($P = 0.4024$). **c** Relative frequency of GBS clades in LOD ($n = 515$) and EOD ($n = 823$) isolates. The statistical significance for each clade based on the test of given proportions were as follows: 1 ($P = 0.0005$), 2 ($P = 2.75 \times 10^{-05}$), 3 ($P = 0.1710$), 4 ($P = 0.0002$), 5 ($P = 0.4024$), and 6 ($P = 3.75 \times 10^{-18}$). **d** Relative frequency of GBS capsular serotypes in isolates sampled from the CNS ($n = 494$) and non-CNS ($n = 844$) tissues. The statistical significance for each serotype based on the test of given proportions were as follows: Ia ($P = 1.50 \times 10^{-06}$), Ib ($P = 0.2932$), II ($P = 1.32 \times 10^{-09}$), III ($P = 1.75 \times 10^{-25}$), V ($P = 2.23 \times 10^{-05}$), and Other ($P = 1.22 \times 10^{-05}$). **e** Relative frequency of GBS capsular serotypes in isolates sampled from the CNS ($n = 494$) and non-CNS ($n = 844$) tissues. The statistical significance for each serotype based on the test of given proportions were as follows: 1 ($P = 9.87 \times 10^{-09}$), 10 ($P = 0.0034$), 17 ($P = 2.23 \times 10^{-10}$), 19 ($P = 0.0256$), 23 ($P = 7.96 \times 10^{-06}$), and Other ($P = 0.2874$). **f** Relative frequency of GBS lineages in isolates sampled from the CNS ($n = 494$) and non-CNS ($n = 844$) tissues. The statistical significance for each serotype based on the test of given proportions were as follows: 1 ($P = 0.0034$), 2 ($P = 9.87 \times 10^{-09}$), 3 ($P = 0.0256$), 4 ($P = 7.96 \times 10^{-06}$), 5 ($P = 0.2874$), and 6 ($P = 2.23 \times 10^{-10}$). All the error bars in each plot represents 95% confidence intervals. Source data used to generate figures in panel **a-f** is available in Supplementary Data 3.

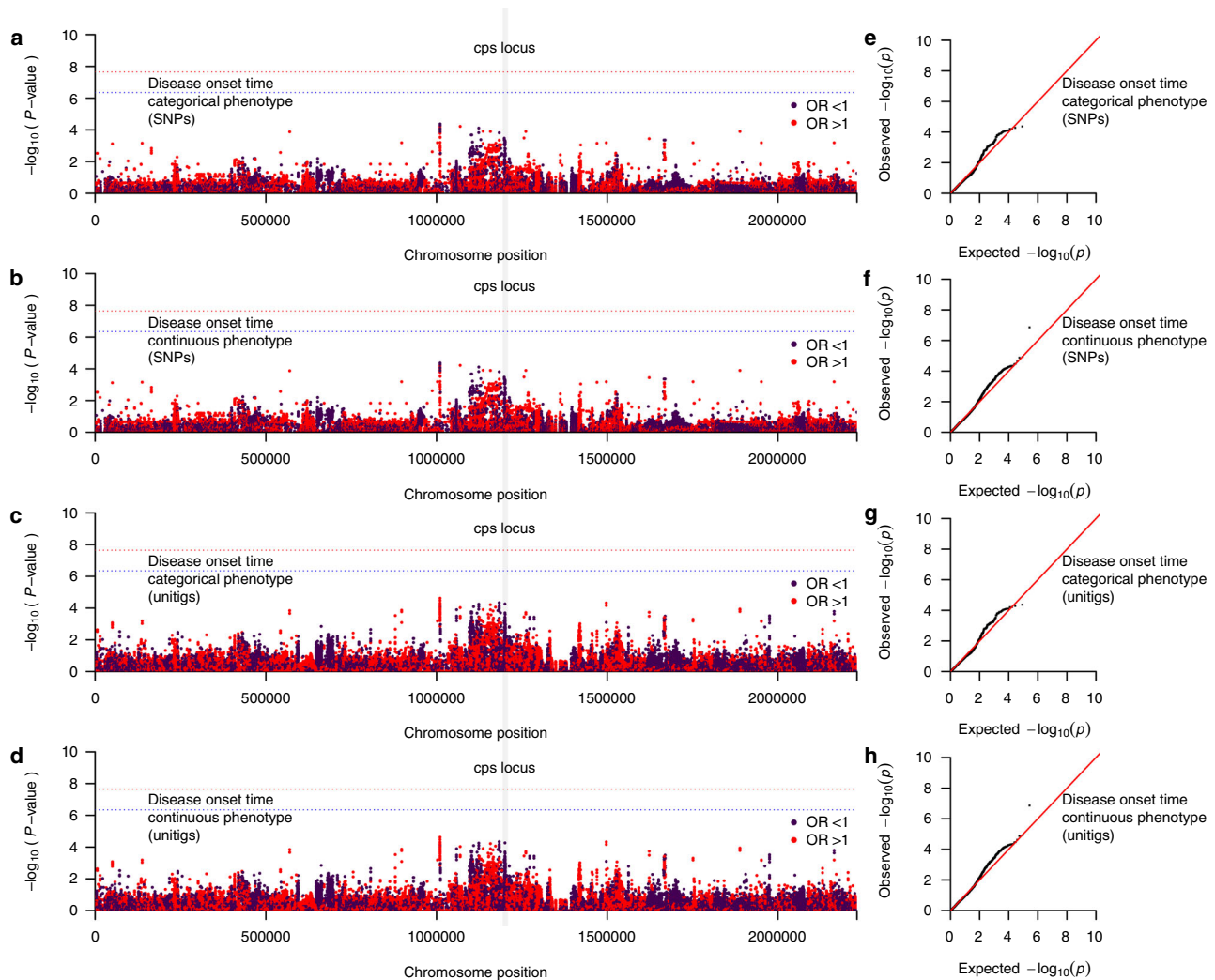


Fig. 4 Manhattan plots showing the association between GBS genomic variation and GBS disease onset time. Statistical significance ($-\log_{10}[P\text{-value}]$) of the GBS genomic variants based on the likelihood ratio test are coloured by the exponentiated fixed effect coefficients or odds ratios of the minor allele as the effect alleles in the GWAS. **a** SNP-based GWAS using disease onset time as a categorical target variable, defined as EOD and LOD. **b** Unitig-based GWAS using disease onset time as a categorical target variable, defined as EOD and LOD. **c** SNP-based GWAS using disease onset time as a continuous target variable defined as the rank-based inverse normal transformation number of days from birth to disease onset. **d** Unitig-based GWAS using disease onset time as a categorical target variable. **e** SNP-based GWAS using the transformed disease onset as a continuous target variable. Q-Q plots showing the relationship between the observed statistical significance and the expected statistical significance, **f** Unitig-based GWAS using disease onset time as a categorical target variable, **g** SNP-based GWAS using the transformed disease onset time as a continuous target variable, and **h** Unitig-based GWAS using disease onset time as a continuous target variable. The red and blue dotted lines represent the genome-wide significance and suggestive threshold, respectively. The variants with odds ratios (OR) >1 is coloured in red while those with odds ratio <1 is coloured in dark purple. Source data for panel **a-f** is available in Supplementary Data 4 and on GitHub (https://github.com/ChrispinChaguza/GBS_Study_NL).

concluded that specific genomic loci in the GBS genomes had minimal influence on the disease onset time of acute invasive neonatal diseases.

Genetic variation within the capsule locus influences meningeal invasion of GBS. Considering the differences in the relative frequency of capsular serotypes in the isolates sampled from the CNS and non-CNS sites (Fig. 3), we next investigated differences in the abundance of genetic variation in the isolates sampled from the CNS and non-CNS tissues, which could influence meningeal tissue invasion of GBS. Similarly, we performed a GWAS with sampled tissue as the target variable, coding samples isolated from the CNS as the affected status and from non-CNS sites as the unaffected status. As done in the GWAS for the disease onset time (Fig. 4), we controlled the population structure to account

for the clonality of the isolates. We identified four SNPs from the GWAS based on SNP genetic variation (Fig. 5a and Supplementary Data 4). These SNPs were in genes within the *cps* locus, with locus tags SAGS_1212 (*epsJ_2*) (odds ratio: 0.86, adjusted $P = 5.14 \times 10^{-20}$) and three in SAGS_1213 (odds ratio: 0.85, adjusted $P = 0.045$) in the GBS genome. Interestingly, we implicated nine *cps* genes in the GWAS using the genetic variation captured by the unitig sequences, including genes identified in the SNP-based GWAS (Fig. 5b, Table 1, and Supplementary Data 4). These unitigs were in several genes, all within the *cps* locus region (odds ratio: 0.85 to 0.86, adjusted $P = 2.30 \times 10^{-20}$ to 0.04). These genes included locus tags SAGS_1212 (*epsJ_2*) and SAGS_1213. We also identified 77 suggestive hits in several genes within and outside the *cps* locus, including SAGS_1212 (*epsJ_2*), SAGS_1213, SAGS_1226 (*arsC*), SAGS_1214, SAGS_1209, SAGS_1200 (*parC*), SAGS_1201 (*parE*), SAGS_1207 (*neuC*) and SAGS_1228 (*rpiA*).

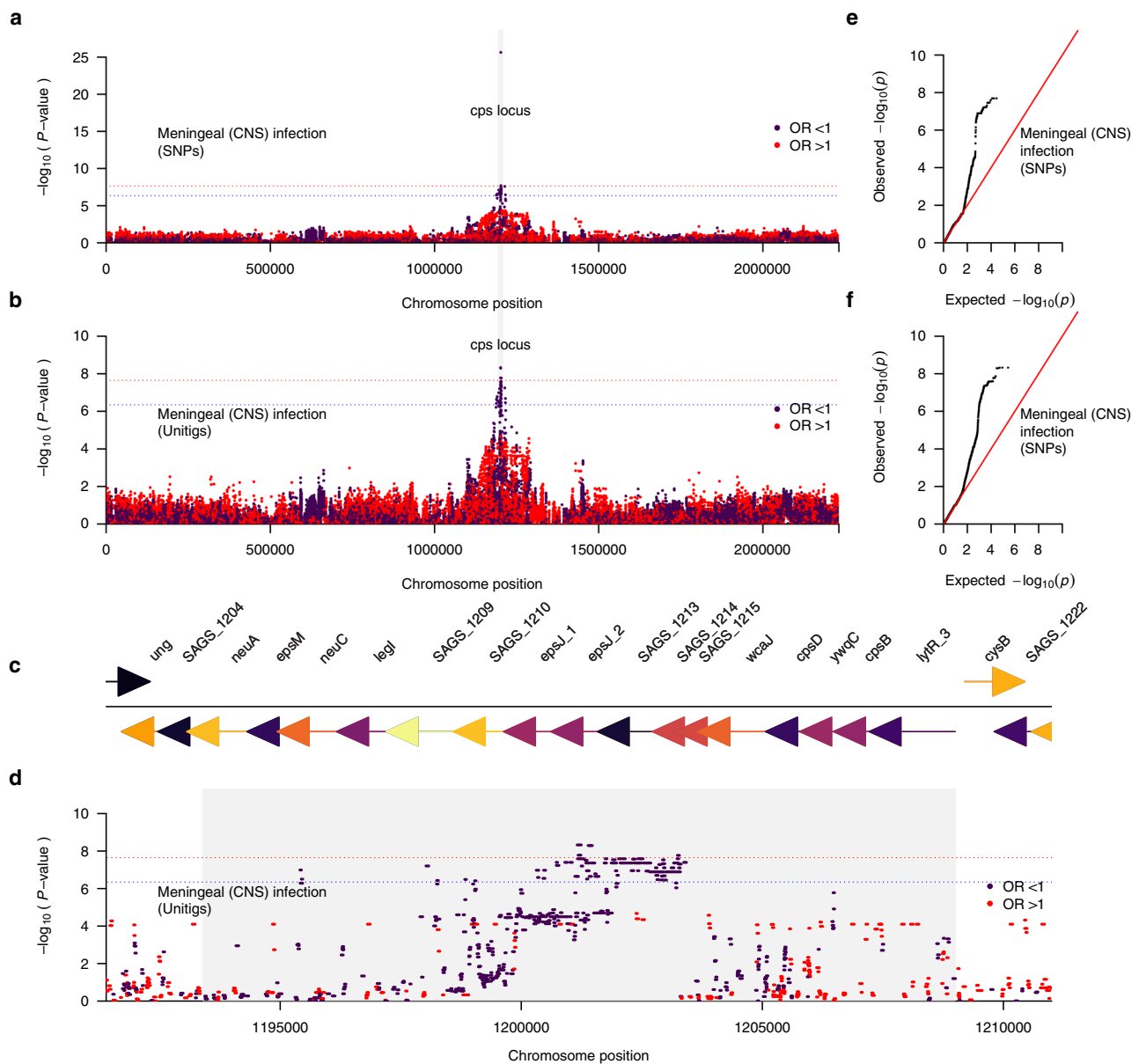


Fig. 5 Manhattan plots showing the association between GBS genomic variation and the CNS infection status. Statistical significance ($-\log_{10}[P\text{-value}]$) of the GBS genomic variants based on the likelihood ratio test are coloured by the exponentiated linear mixed model coefficients or odds ratio as the target variable in the GWAS. **a** SNP-based GWAS and **b** unitig-based GWAS using CNS infection status of the GBS isolates as the target variable specified as isolation from the CNS or non-CNS tissue. **c** Genomic features in the GBS capsular polysaccharide (*cps*) locus, and **d** corresponding Zoom plot showing the statistical significance and location of the genetic variants (unitigs), which mapped to the *cps* region of the complete GBS reference genome (GenBank accession: AP018935.1). Q-Q plots showing the relationship between the observed statistical significance and the expected statistical significance for **e** SNP-based GWAS, and **f** Unitig-based GWAS using body isolation tissue as the target variable. The red and blue dotted lines represent the genome-wide significance and suggestive threshold, respectively. The variants with odds ratios (OR) >1 is coloured in red while those with odds ratio <1 is coloured in dark purple. Source data for panel **a-f** is available in Supplementary Data 4 and on GitHub (https://github.com/ChrispinChaguza/GBS_Study_NL).

Additionally, we identified a total of 123 suggestive unitigs mostly mapping to the *cps* locus and other genomic region (Supplementary Data 4). The *cps* genes tagged by the unitigs included those encoding a multidrug major facilitator superfamily (MFS) transporter (*cpsG*), glycosyltransferase CpsJ (*cpsJ*), a capsular polysaccharide biosynthesis protein CpsA (*cpsA*), and a polysialic acid biosynthesis protein P7 (*neuC*) while the other genes flanking the *cps* locus, included DNA topoisomerase 4 subunit A (*parC*), DNA topoisomerase IV subunit B (*gyrB*) and an arsenate reductase (*arsC*) (Fig. 5c). Consistent with the GWAS of the disease onset time outcome, the Q-Q plots for the analysis based on the CNS infection phenotype suggested no apparent issues

when controlling the population structure (Fig. 4e, f). Complementary GWAS based on the presence and absence of accessory genes revealed a single statistically significant association for a phosphopentomutase encoding gene, and two suggestive hits for capsular biosynthesis genes (Supplementary Fig. 6c). Altogether, these findings suggested that genetic variation within the capsule biosynthesis locus influences bloodstream-to-meningeal tissue invasion and potentially survival in the CNS.

Genetic variation for CNS invasion varies by GBS serotype. To understand the distribution of identified genetic variants within

Table 1 Summary of the unitig sequences statistically associated with infection of the CNS in the GWAS using FaST-LMM.

Locus tag	Gene name	Reference genome	Number of unitigs	P-value range	Adjusted P-value (Q) range	Odds ratio range	Gene product
CHF17_01256	<i>cpsG</i>	CP022537.1	1	1.696×10^{-08}	0.038	0.845	Glycosyl transferase CpsG(V)
No match	No match	No match	1	1.361×10^{-08}	0.030	0.850	No match
AV644_06110		CP013908.1	1	4.745×10^{-09}	0.011	0.857	Arsenate reductase
BB165_05995		CP021870.1	1	5.086×10^{-09}	0.011	0.866	Capsular biosynthesis protein
CCZ24_04050		CP021773.1	1	1.656×10^{-08}	0.037	0.847	Capsular biosynthesis protein
CWQ20_06175		CP025029.1	3	4.706×10^{-09} – 5.086×10^{-09}	0.011–0.011	0.864–0.866	Capsular biosynthesis protein
GBS222_1012		FO393392.1	1	1.617×10^{-08}	0.036	0.843	Hypothetical protein

the *cps* locus associated with CNS invasion, we compared the relative abundance of the *cps*-associated unitigs in the GBS isolates sampled from the CNS and non-CNS tissues stratified by serotype. We found certain unitigs were differentially abundant among the isolates sampled from the CNS and non-CNS tissues, especially for serotype IV and non-typeable isolates^{75,76} the latter account for ~10% of the GBS isolates in Europe⁷⁷ (Supplementary Fig. 9). These two serotypes were relatively less abundant in the GBS isolates in the Netherlands. However, the rest of the unitigs showed similar abundance among the CNS and non-CNS isolates for the other serotypes. Therefore, these findings suggested that the genetic variants had a small effect on invading the CNS tissue.

Heritability highlights a moderate effect of genetics on disease onset time and CNS invasion. We then formally quantified the variability in the neonatal disease onset time and CNS infection phenotypes explained by the genetic variability in the GBS genomes. To achieve this, we estimated the narrow-sense heritability (h^2) for each phenotype using several methods, namely GEMMA⁷⁸, FaST-LMM⁷⁹, and GCTA⁸⁰ (see methods). Previous GWAS of bacteria suggested a negligible contribution of pathogen genetics to the variability in the phenotypes associated with disease outcomes, for example, severity^{44,81}. Contrary to colonisation, genetic variation correlated with invasive disease is unlikely to be positively selected by natural selection, because invasive disease is an evolutionary dead-end. Essentially, either the host immune system clears the pathogen, or the host dies without impacting onward transmission of the pathogen and frequency of the variants in the population⁸², which obscures the genetic–phenotype association signal in the GWAS. Therefore, we hypothesised that the disease onset time and CNS infection phenotypes have low heritability. Consistent with our hypothesis, we found low estimates for the narrow-sense heritability using GEMMA for the disease onset time as categorical ($h^2 = 0.07$ to 0.21) and continuous ($h^2 = 0.06$ to 0.21) variables, and CNS invasion ($h^2 = 0.06$ to 0.14) based on different types of genetic variation (Fig. 6a and Supplementary Data 5). In support of these findings, we found similar heritability estimates using FaST-LMM⁷⁹ (Fig. 6b and Supplementary Data 5), although slightly lower values were inferred with GCTA⁸⁰ (Fig. 6c and Supplementary Data 5). Overall, these findings suggested a modest but non-negligible impact of GBS genetics on the inter-strain variability in the disease onset time and infection of the CNS tissue.

Discussion

This study leveraged an extensive collection of acute invasive neonatal GBS clinical isolates routinely collected over thirty years through a robust and well-established national bacterial surveillance programme in the Netherlands⁸³. By applying well-controlled linear mixed model GWAS approaches, we have systematically identified genomic variation in GBS associated with the disease

onset time and invasion of the CNS. These findings suggest that pathogen genetics modulates the timing of GBS disease in neonates and bloodstream-to-meningeal invasion, which increases the risk for meningitis—a severe clinical manifestation of GBS disease typically associated with long-term neurologic sequelae⁸⁴ and mortality⁶³. Previous studies have not implicated the loci associated with disease onset time identified in this study with GBS pathogenicity and virulence, highlighting the utility of agnostic and unbiased GWAS approaches to unravel novel genotype–phenotype associations. Furthermore, although the GBS capsule is a well-known virulence determinant critical for immune evasion and virulence^{25–28}, our results provide the evidence that genomic variation within the capsule biosynthesis locus influences the pathogenesis of meningitis by modulating bloodstream-to-meningeal invasion of the CNS compartment.

Bloodstream infection is the predominant transient invasive disease state preceding infection of the CNS, especially meningitis. Therefore, such an aetiology of meningitis implies that GBS isolated from the CNS are also capable of causing bloodstream infection. However, the converse may not necessarily hold if the pathogen genetics influenced CNS invasion. Our findings show that certain GBS isolates infecting the CNS harbour genetic variation within the capsule biosynthetic locus, which may influence meningeal invasion by modulating translocation across the blood-brain-barrier into CNS, possibly through interactions with the host endothelial cells, as similarly seen with other adhesins and host transmembrane receptors⁸⁵. Similar to other encapsulated bacteria, such as *Streptococcus pneumoniae*⁸⁶, the polysaccharide capsule of GBS is a critical virulence determinant⁸⁷, which promotes immune evasion by inhibiting phagocytosis²⁵, complement deposition, and activation²⁶, and platelet-mediated killing^{26–28}. The *cps* genes containing genetic variation associated with CNS infection included those encoding for a transcriptional regulator for the *cps* operon (*cpsA*), synthesis and transport of oligosaccharides to the outside of the cell membrane (*cpsJ* and *cpsG*), and synthesis transport of sialic acids (*neuC*)⁸⁸. Therefore, the identified genomic variation associated with the CNS invasion may also promote GBS survival in the CNS, ultimately modulating the risk of meningitis. However, we found no statistically significant genetic variation tagging other known virulence factors important for meningeal tropism, such as HvgA⁸⁹, which suggest although such genes are generally essential for GBS meningeal tropism, their allelic variability does not influence the ability of GBS to invade the CNS. Our findings highlight the importance of the *cps* and other genomic loci in the pathogenicity and virulence of GBS, implicating them as potential targets for the development of capsule- and protein-based vaccines, treatments, and diagnostics to reduce the short- and long-term neonatal GBS disease burden and death toll globally. Such GBS vaccines are currently undergoing preclinical²⁰ and early phase I and II clinical trials^{20–23}. Reassuringly, a capsule-based vaccine, for example, a hexavalent polysaccharide-protein conjugate vaccine

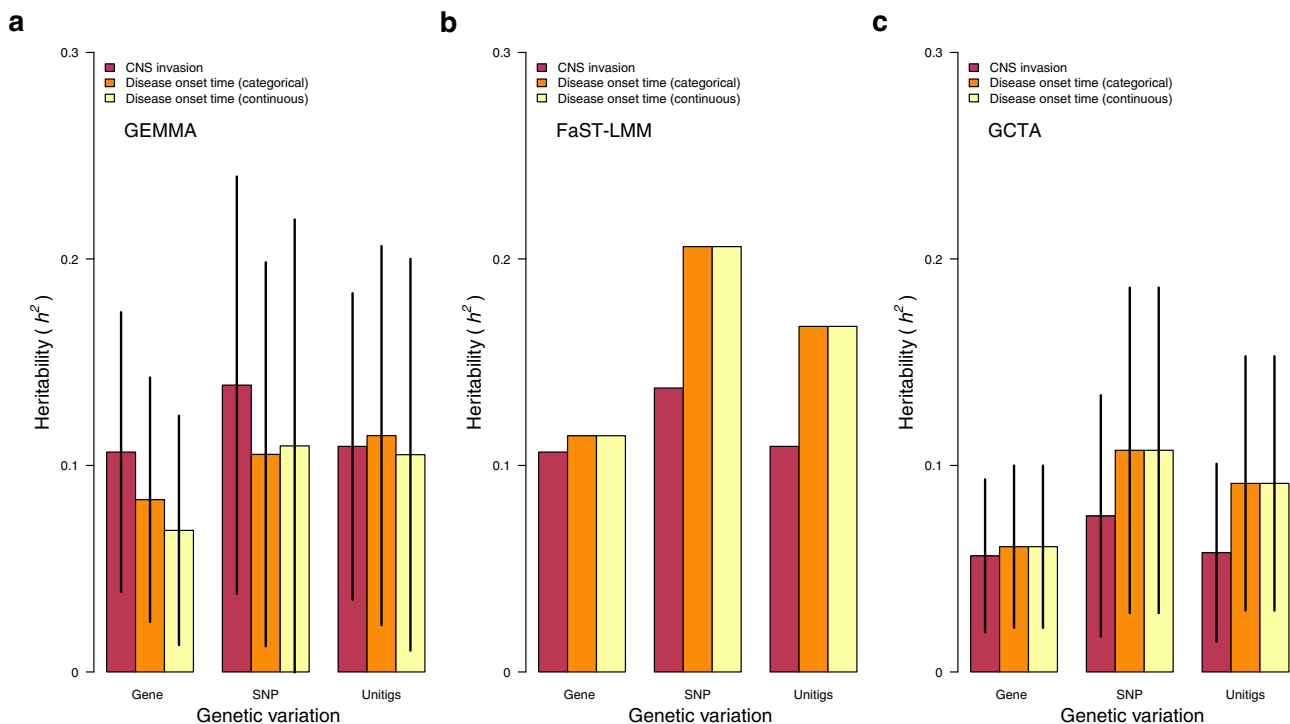


Fig. 6 Estimates of the narrow-sense heritability (h^2) using different genetic variants and methods. **a** Heritability estimated for accessory genes, SNPs, and unitigs using GEMMA. **b** Heritability estimated for accessory genes, SNPs, and unitigs using FaST-LMM. **c** Heritability estimated for accessory genes, SNPs, and unitigs using GCTA. FaST-LMM reported no standard errors for the heritability estimates, therefore, we do not show the confidence intervals in the panel **b**. Heritability is expressed as a proportion with values ranging from 0 to 1. The error bars in each plot represents 95% confidence intervals. The estimates in panels **a–c** are based on $n = 1338$ isolates. Source data used to generate figures in panel **a–c** is available in Supplementary Data 5.

formulation targeting GBS serotypes Ia, Ib, II, III, IV, and V, could target between 93 to 99% of isolates from maternal colonisation, maternal invasive disease, and neonatal and infant invasive disease;⁶⁰ potentially reducing preterm and stillbirths, neurologic sequelae, mortality, and economic burden globally⁸⁴.

Although GBS genetics appears to substantially affect the disease onset time and CNS invasion, the amount of variability in these phenotypes accounted for by the genomic variation, i.e., the narrow-sense heritability, appears to be moderate but not negligible. Such heritability estimates reflect a polygenic nature of the phenotypes, as complex traits are modulated by multiple variants with potentially small effect sizes, requiring larger datasets to implicate them in the GWAS and account for the missing heritability. These findings contrast with the genomic variation associated with other bacterial phenotypes, including antimicrobial resistance⁹⁰, host niche adaptation⁵⁷, and invasiveness^{44,46}, which typically exhibit high heritability reflecting substantial natural selection pressures on these genomic loci. However, a current challenge remain that an unknown proportion of bloodstream infection cases result in an unidentified meningitis, i.e., some of the cases with a positive blood culture would also have had a positive CSF culture if a lumbar puncture would have been performed, which is not done in all neonates with signs of infection. Therefore, the GBS isolates sampled from the CNS and non-CNS sites may not be completely genetically distinguishable likely dampening the genotype–phenotype association signal and heritability. Altogether, these findings suggest that although pathogen genetics partly influence the variability in the onset of acute invasive GBS disease and tissue invasion in the neonates and infants, the maternal and neonatal host factors, host–pathogen interactions, and the environmental factors contribute to a more considerable extent to the GBS disease pathogenesis.

This study shows that genomic variation of GBS, within and outside the capsule biosynthesis locus, influences the disease onset

time and bloodstream-to-meningeal translocation and invasion of the CNS in neonates with acute invasive disease. These findings highlight the crucial role of the sialic acid polysaccharide capsule in the virulence of GBS, emphasising the need for evaluating capsule-based vaccines to prevent and control invasive diseases in neonates and infants. Furthermore, our study highlights the utility of microbial population genomics combined with well-established clinical bacterial surveillance programmes to generate novel and unbiased insights into the contribution of bacterial genetics to the population-level pathogen traits, potentially challenging to study experimentally, to inform disease prevention and control strategies. As more GBS whole-genome sequences, as well as strain and patient-level metadata, become available, we are optimistic that exploiting these big data by applying robust GWAS and possibly machine learning approaches, as highlighted in this study, will not only validate but also unravel additional novel cryptic pathogenicity loci influencing several GBS phenotypes and clinical outcomes, including mother-to-child transmission, human intestinal and vaginal niche, animal adaptation, and disease severity.

Though this work provides robust evidence of the contribution of GBS genetics to disease onset and CNS invasion, there are some limitations worth noting. Meningitis diagnosis could have been missed in some sepsis cases because a lumbar puncture was postponed, not done, or false negative, which could result in an underestimation of meningitis incidence. Therefore, although the CSF-positive cases were from meningitis cases, some blood culture isolates are likely from patients with unidentified meningitis, which may have dampened the associations in the GWAS of CNS and non-CNS isolates. A nationwide guideline recommending intrapartum antibiotic prophylaxis to women with risk factors of a newborn with early-onset disease was implemented around 1998; however, this is unlikely to introduce bias in the dataset. We have previously used the concurrent *Escherichia coli* collection to

make the case that there had been no apparent shift in laboratory surveillance practice⁷. We have also compared meningococcal submissions to the reference laboratory to another mandatory notification system and found a similar pattern over time⁸³. Furthermore, detailed clinical background information was not available, which restricted the adjustments in the GWAS analyses.

In conclusion, we have shown using population genomics of an extensive and well-sampled collection of neonatal GBS clinical isolates that variation in the GBS genomes, within and outside the capsule biosynthesis region, influence the onset time for acute invasive disease and invasion of the meningeal tissue, highlighting a genetic basis for the inter-strain variability of the GBS disease outcomes in neonates and infants.

Methods

Samples, microbiological processing, and ethical approvals. All GBS isolates cultured from cerebrospinal fluid or blood from patients were submitted to the National Reference Laboratory of Bacterial Meningitis (NRLBM) at the Amsterdam UMC, University of Amsterdam, for further typing and storage, as part of the continuous surveillance of bacterial meningitis in the Netherlands. One thousand and three hundred and thirty-eight GBS isolates were selected from a dataset of isolates collected from nationwide surveillance of infants with bacterial meningitis and bacteraemia conducted by the NRLBM⁵⁹ (Supplementary Data 1). The isolates were collected over 30 years from January 1987 to January 2016. Of these isolates, 823 and 515 were from neonates with EOD (0 to 6 days, post-birth) and LOD (6 to 89 days, post-birth). The age of the infant was estimated as the time from birth to sample collection as previously described⁵⁹. By isolation source, 494 isolates were sampled from the cerebrospinal fluid (CSF), while 844 isolates were from non-CSF sites, namely blood ($n = 844$). For the present study patient data were anonymized. Additional institutional review board approval is not required for studying submitted strains with anonymised patient data.

Whole-genome sequencing and molecular typing of GBS isolates. Genomic DNA was extracted using the Wizard[®] Genomic DNA Purification Kit from Promega following the manufacturer's instructions⁵⁹. The genomic libraries were created using the Illumina protocol, and whole genomes were sequenced with 125 bp reads on the HiSeq 2000 platform (Illumina, CA, USA). Genome assembly was done using SPAdes genome assembler (version 3.14.0)⁹¹. The serotype for each isolate was determined in silico using whole-genome sequence data^{92,93}. Sequence typing using the multilocus sequence typing (MLST) scheme for GBS⁶¹ was done based on the sequencing data using SRST2 (version 0.2.0)⁹⁴.

Phylogenetic and population structure analysis. A multi-sequence whole-genome alignment was generated based on consensus sequences of each isolate inferred after mapping reads against a complete GBS reference genome for an invasive human strain HU-GS5823 (GenBank accession: AP018935.1) belonging to sequence type (ST335) and serotype III using Snippy (version 4.6.0) (<https://github.com/tseemann/snippy>). The genomic positions in the consensus sequences containing variable nucleotide sites or SNPs were extracted from the alignment as multi-FASTA, and variant call format (VCF) files using SNP-sites (version 2.3.2)⁹⁵. The identified SNPs were then used for population structure analysis to identify sequence clusters or lineages using the hierarchical clustering approach implemented in BAPS (version 6)⁹⁶. A maximum-likelihood phylogenetic tree of the entire GBS isolates was generated based on the whole-genome SNP alignment using the general time-reversible (GTR) and Gamma model in FastTree (version 2.1.10)^{97,98}. We used a *Streptococcus pneumoniae* strain (ENA accession: ERS812015) as an outgroup to root the inferred phylogenetic tree of the GBS isolates. Visual exploration and analysis of the phylogenetic trees was done using the APE package (version 4.3)⁹⁹. Annotation of the phylogenetic tree with the isolate metadata was done using the "gridplot" and "phylo4d" functions in phyloSignal (version 1.3) and phyloBase (version 0.8.6) (<https://cran.r-project.org/package=phyloBase>) packages, respectively¹⁰⁰.

Generating variant data for bacterial GWAS. The input data for the GWAS were generated using only bi-allelic SNPs in the VCF file of each GBS isolate using VCFtools (version 0.1.16)¹⁰¹. The SNPs with minor allele frequency <5% or missingness >5% were filtered out from the final dataset to exclude rare variants using PLINK (version 1.90b4)¹⁰². To generate the input dataset for the GWAS using the presence and absence patterns of the accessory genes, we first clustered the predicted gene sequences predicted using Prokka (version 1.11)¹⁰³ into clusters of orthologous genes (COGs) with Panaroo (version 1.2.2)¹⁰⁴. We specified the moderate stringency mode when running Panaroo. The COGs are referred to as genes for simplicity. The presence and absence patterns of the predicted genes were merged with the isolate metadata and converted to the pedigree format for the GWAS. Similar to the SNP variant data, the genes with minor allele frequency <5% were filtered out using PLINK (version 1.90b4)¹⁰². To identify the maximal unitig sequences, i.e., non-branching paths in a compacted De Bruijn graph, we first build

the graph for the entire dataset based on 31 bp k-mer sequences using Bifrost (version 1.0.1)⁷³. The unitig sequences generated based on the entire isolate collection were queried against a De Bruijn graph of each genome using Bifrost to determine the presence and absence patterns of each unitig sequence in the genomes. A unitig was considered present when exact matches for all the k-mers in the query unitig sequence were found in each isolate's genome graph. The presence and absence patterns of the unitigs were merged with the affection status (disease onset time and CNS infection) to generate the pedigree data files required for the GWAS. Similarly, the unitigs with minor allele frequency <5% were filtered out using PLINK before the GWAS. The genes and unitigs were not filtered based on missingness as missingness for any reason was regarded as the absence of the gene since it was not possible to distinguish missingness due to either sequencing or assembly errors from true absence due to the variability in the accessory genome.

Genome-wide association analysis. We first compared the relative frequency of capsular serotypes, clonal complexes, and lineages in the GBS isolates associated with EOD and LOD and CNS and non-CNS tissue, using Fisher's exact test. We used the clonal complexes and lineages previously defined by Jamrozny et al.⁵⁹. To identify genomic variation, defined in terms of the presence and absence patterns of SNPs, unitig and accessory gene sequence, associated with the GBS disease phenotypes, namely disease onset time, either as a categorical (EOD and LOD) or continuous (days from birth to disease onset) values, and infection of the CNS (CNS and non-CNS) where GBS was isolated; we performed univariate GWAS using robust linear mixed models implemented in FaST-LMM (FastLmmC, version 2.07.20140723)⁷⁹. We applied a rank-based inverse normal transformation to the disease onset time as a continuous phenotype to generate normally distributed values as required by the GWAS methods (https://github.com/ChrispinChaguza/GBS_Study_NL). We specified a genetic relatedness matrix of the GBS isolates generated using the unitigs as a random covariate to account for the clonal population structure during the GWAS analysis. Since bacterial chromosomes are haploid, we coded the genotypes as originating from the human mitochondrial genome, designated as chromosome 26, as previously described^{45,52}. We adjusted the raw *P*-values for each variant, inferred using the likelihood ratio test using the Bonferroni correction method to control the false discovery rate due to multiple testing. Since the frequency of genomic variants tested, i.e., accessory genes, SNPs, and unitigs, varied greatly, we used a fixed value for the GBS genome size to represent the possible maximum number of realised genomic variants. This approach is more conservative than adjusting based on the observed variants, minimising false positives but potentially increasing false negatives slightly. However, crucially, our approach ensures the use of a consistent *P*-value threshold when assessing the statistical significance of different types of genomic variation.

Genetic variants with *P*-values < 2.24×10^{-08} , i.e., α/G where the statistical significance threshold $\alpha = 0.05$ and the genome size $G = 2,231,314$ bp for the GBS reference genome of the strain HU-GS5823, were deemed statistically significant. Similarly, variants with *P*-values < 8.12×10^{-07} , i.e., α/G where the statistical significance threshold $\alpha = 1$, were considered suggestive. The statistical significance was assessed to compare the expected and observed *P*-values to visually check potential issues with the population structure, using the quantile-quantile plots generated with qqman (version 0.1.7)¹⁰⁵. The overall proportion of phenotypic variability explained by variation in the genome (narrow-sense heritability) was estimated using FaST-LMM, GEMMA (version 0.98.1)⁷⁸, and GCTA (version 1.93.2)⁸⁰. The genomic features associated with each SNP, accessory gene, and unitigs were identified by comparing them with a panel of GBS reference genomes using BioPython (version 1.78)¹⁰⁶. In addition, we used BLASTN (version 2.5.0+)¹⁰⁷ to identify genomic regions containing the gene and unitig sequences. The GWAS results were summarised to visually identify genomic regions containing statistically significant genotype-phenotype associations using Manhattan plots generated in R (version 4.0.3) (<https://www.R-project.org/>).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequence reads for the isolates used in this study are available in the European Nucleotide Archive under study accession code PRJEB14124. The accession numbers and information for individual isolates are provided in Supplementary Data 1. The authors declare that all other data supporting the findings of this study are available within the paper and its supplementary information files. Additional data for the SNPs, accessory genes, and unitig sequences used in this study are available at https://github.com/ChrispinChaguza/GBS_Study_NL.

Code availability

All tools and methods used for the analysis are publicly available and fully described in the Methods section. The scripts used in this analysis are available at https://github.com/ChrispinChaguza/GBS_Study_NL.

Received: 19 January 2022; Accepted: 6 July 2022;

Published online: 21 July 2022

References

- Johri, A. K. et al. Group B Streptococcus: global incidence and vaccine development. *Nat. Rev. Microbiol.* <https://doi.org/10.1038/nrmicro1552> (2006).
- Nizet, V. I., Ferrieri, P. A. & Rubens, C. E. Molecular pathogenesis of group B streptococcal disease in newborns. *Streptococcal Infect. Clin. Asp. Microbiol. Mol. Pathog.* 180221 (Oxford Univ. Press, New York, NY, 2000).
- Davies, H. G., Carreras-Abad, C., Le Doare, K. & Heath, P. T. Group B Streptococcus: trials and tribulations. *Pediatr. Infect. Dis. J.* <https://doi.org/10.1097/INF.0000000000002328> (2019).
- Seale, A. C. et al. Estimates of the burden of Group B Streptococcal disease worldwide for pregnant women, stillbirths, and children. *Clin. Infect. Dis.* <https://doi.org/10.1093/cid/cix664> (2017).
- World Health Organisation. *Group B Streptococcus Vaccine: Full Value of Vaccine Assessment. Financial Analysis.* (World Health Organisation, 2021).
- Schuchat, A. Epidemiology of group B streptococcal disease in the United States: Shifting paradigms. *Clin. Microbiol. Rev.* <https://doi.org/10.1128/cmr.11.3.497> (1998).
- Bekker, V., Bijlsma, M. W., van de Beek, D., Kuijpers, T. W. & Van der Ende, A. Incidence of invasive group B streptococcal disease and pathogen genotype distribution in newborn babies in the Netherlands over 25 years: a nationwide surveillance study. *Lancet Infect. Dis.* [https://doi.org/10.1016/S1473-3099\(14\)70919-3](https://doi.org/10.1016/S1473-3099(14)70919-3) (2014).
- Bevan, D., White, A., Marshall, J. & Peckham, C. Modelling the effect of the introduction of antenatal screening for group B Streptococcus (GBS) carriage in the UK. *BMJ Open* <https://doi.org/10.1136/bmjopen-2018-024324> (2019).
- Ancona, R. J., Ferrieri, P. & Williams, P. P. Maternal factors that enhance the acquisition of group-B streptococci by newborn infants. *J. Med. Microbiol.* **13**, 273–280 (1980).
- Tazi, A. et al. Risk factors for infant colonization by hypervirulent CC17 group B Streptococcus: toward the understanding of late-onset disease. *Clin. Infect. Dis.* **69**, 1740–1748 (2019).
- Gizachew, M. et al. Proportion of Streptococcus agalactiae vertical transmission and associated risk factors among Ethiopian mother-newborn dyads, Northwest Ethiopia. *Sci. Rep.* **10**, 3477 (2020).
- Hung, L.-C. et al. Risk factors for neonatal early-onset group B streptococcus-related diseases after the implementation of a universal screening program in Taiwan. *BMC Public Health* **18**, 438 (2018).
- Schrag, S. J. et al. A population-based comparison of strategies to prevent early-onset group B Streptococcal disease in neonates. *N. Engl. J. Med.* **347**, 233–239 (2002).
- Lawn, J. E. et al. Every country, every family: time to act for group B Streptococcal disease worldwide. *Clin. Infect. Dis.* ciab859, <https://doi.org/10.1093/cid/ciab859> (2021).
- Romain, A.-S. et al. Clinical and laboratory features of group B Streptococcus meningitis in infants and newborns: study of 848 cases in France, 2001–2014. *Clin. Infect. Dis.* **66**, 857–864 (2018).
- Allhazmi, A., Hurteau, D. & Tyrrell, G. J. Epidemiology of invasive group B Streptococcal disease in Alberta, Canada, from 2003 to 2013. *J. Clin. Microbiol.* **54**, 1774–1781 (2016).
- Moore, M. R., Schrag, S. J. & Schuchat, A. Effects of intrapartum antimicrobial prophylaxis for prevention of group-B-streptococcal disease on the incidence and ecology of early-onset neonatal sepsis. *Lancet Infect. Dis.* **3**, 201–213 (2003).
- Ohlsson, A. & Shah, V. S. Intrapartum antibiotics for known maternal Group B streptococcal colonization. *Cochrane Database Syst. Rev.* <https://doi.org/10.1002/14651858.CD007467.pub4> (2014).
- Nishihara, Y., Dangor, Z., French, N., Madhi, S. & Heyderman, R. Challenges in reducing group B Streptococcus disease in African settings. *Arch. Dis. Child.* **102**, 72 LP–72 77 (2017).
- Buurman, E. T. et al. A novel hexavalent capsular polysaccharide conjugate vaccine (GBS6) for the prevention of neonatal group b streptococcal infections by maternal immunization. *J. Infect. Dis.* <https://doi.org/10.1093/infdis/jiz062> (2019).
- Madhi, S. A. et al. Safety and immunogenicity of an investigational maternal trivalent group B streptococcus vaccine in healthy women and their infants: a randomised phase 1b/2 trial. *Lancet Infect. Dis.* [https://doi.org/10.1016/S1473-3099\(16\)00152-3](https://doi.org/10.1016/S1473-3099(16)00152-3) (2016).
- Heyderman, R. S. et al. Group B streptococcus vaccination in pregnant women with or without HIV in Africa: a non-randomised phase 2, open-label, multicentre trial. *Lancet Infect. Dis.* [https://doi.org/10.1016/S1473-3099\(15\)00484-3](https://doi.org/10.1016/S1473-3099(15)00484-3) (2016).
- Nilo, A. et al. Anti-group B Streptococcus glycan-conjugate vaccines using pilus protein GBS80 as carrier and antigen: comparing lysine and tyrosine-directed conjugation. *ACS Chem. Biol.* <https://doi.org/10.1021/acscchembio.5b00247> (2015).
- Absalon, J. et al. Safety and immunogenicity of a novel hexavalent group B streptococcus conjugate vaccine in healthy, non-pregnant adults: a phase 1/2, randomised, placebo-controlled, observer-blinded, dose-escalation trial. *Lancet Infect. Dis.* **21**, 263–274 (2021).
- Martin, T. R., Ruzinski, J. T., Rubens, C. E., Chi, E. Y. & Wilson, C. B. The effect of type-specific polysaccharide capsule on the clearance of group B Streptococci from the lungs of infant and adult rats. *J. Infect. Dis.* **165**, 306–314 (1992).
- Marques, M. B., Kasper, D. L., Pangburn, M. K. & Wessels, M. R. Prevention of C3 deposition by capsular polysaccharide is a virulence mechanism of type III group B streptococci. *Infect. Immun.* <https://doi.org/10.1128/iai.60.10.3986-3993.1992> (1992).
- Uchiyama, S. et al. Dual actions of group B Streptococcus capsular sialic acid provide resistance to platelet-mediated antimicrobial killing. *Proc. Natl. Acad. Sci. USA.* <https://doi.org/10.1073/pnas.1815572116> (2019).
- Herbert, M. A., Beveridge, C. J. E. & Saunders, N. J. Bacterial virulence factors in neonatal sepsis: group B streptococcus. *Curr. Opin. Infect. Dis.* <https://doi.org/10.1097/00001432-200406000-00009> (2004).
- Lynskey, N. N. et al. Multi-functional mechanisms of immune evasion by the streptococcal complement inhibitor C5a peptidase. *PLOS Pathog.* **13**, e1006493 (2017).
- Bryan, J. D. & Shelver, D. W. Streptococcus agalactiae CspA is a serine protease that inactivates chemokines. *J. Bacteriol.* **191**, 1847–1854 (2009).
- Poyart, C. et al. Contribution of Mn-cofactored superoxide dismutase (SodA) to the virulence of Streptococcus agalactiae. *Infect. Immun.* **69**, 5098–5106 (2001).
- Gibson, R. L., Nizet, V. & Rubens, C. E. Group B Streptococcal β -hemolysin promotes injury of lung microvascular endothelial cells. *Pediatr. Res.* **45**, 626–634 (1999).
- Zhu, L. et al. Genetic basis underlying the hyperhemolytic phenotype of Streptococcus agalactiae strain CNCTC10/84. *J. Bacteriol.* **202**, e00504–e00520 (2020).
- Deng, L. et al. Characterization of a two-component system transcriptional regulator, LtdR, that impacts group B Streptococcal colonization and disease. *Infect. Immun.* **86**, e00822–17 (2018).
- Wang, N.-Y. et al. Group B streptococcal serine-rich repeat proteins promote interaction with fibrinogen and vaginal colonization. *J. Infect. Dis.* **210**, 982–991 (2014).
- Buscetta, M. et al. FbsC, a novel fibrinogen-binding protein, promotes Streptococcus agalactiae-host cell interactions. *J. Biol. Chem.* **289**, 21003–21015 (2014).
- Doran, K. S. et al. Blood-brain barrier invasion by group B Streptococcus depends upon proper cell-surface anchoring of lipoteichoic acid. *J. Clin. Invest.* **115**, 2499–2507 (2005).
- Almeida, A. et al. Whole-genome comparison uncovers genomic mutations between group B Streptococci sampled from infected newborns and their mothers. *J. Bacteriol.* **197**, 3354–3366 (2015).
- Andrea, G. et al. Pan-GWAS of Streptococcus agalactiae highlights lineage-specific genes associated with virulence and niche adaptation. *MBio* **11**, e00728–20 (2021).
- Read, T. D. & Massey, R. C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med.* **6**, 109 (2014).
- Power, R. A., Parkhill, J. & De Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* <https://doi.org/10.1038/nrg.2016.132> (2016).
- Lees, J. A. et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb. Genomics* **3**, e000103–e000103 (2017).
- Lilje, B. et al. Whole-genome sequencing of bloodstream Staphylococcus aureus isolates does not distinguish bacteraemia from endocarditis. *Microb. Genomics* **3**, e000138 (2017).
- Lees, J. A. et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat. Commun.* <https://doi.org/10.1038/s41467-019-09976-3> (2019).
- Li, Y. et al. Genome-wide association analyses of invasive pneumococcal isolates identify a missense bacterial mutation associated with meningitis. *Nat. Commun.* **10**, 178 (2019).
- Young, B. C. et al. Pantone-valentine leucocidin is the key determinant of staphylococcus aureus pyomyositis in a bacterial GWAS. *Elife* <https://doi.org/10.7554/eLife.42486> (2019).
- Kulohoma, B. W. et al. Comparative genomic analysis of meningitis- and bacteremia-causing pneumococci identifies a common core genome. *Infect. Immun.* **83**, 4165–4173 (2015).

48. Davies, M. R. et al. Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nat. Genet.* **51**, 1035–1043 (2019).
49. Chaguzza, C. et al. Bacterial genome-wide association study of hyper-virulent pneumococcal serotype 1 identifies genetic variation associated with neurotropism. *Commun. Biol.* **3**, 559 (2020).
50. Laabei, M. et al. Predicting the virulence of MRSA from its genome sequence. *Genome Res.* <https://doi.org/10.1101/gr.165415.113> (2014).
51. Coll, F. et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* <https://doi.org/10.1038/s41588-017-0029-0> (2018).
52. Chewapreecha, C. et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.* **10**, e1004547–e1004547 (2014).
53. Farhat, M. R. et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **45**, 1183–1189 (2013).
54. Suzuki, M., Shibayama, K. & Yahara, K. A genome-wide association study identifies a horizontally transferred bacterial surface adhesin gene associated with antimicrobial resistant strains. *Sci. Rep.* **6**, 37811 (2016).
55. Hicks, N. D., Carey, A. F., Yang, J., Zhao, Y. & Fortune, S. M. Bacterial genome-wide association identifies novel factors that contribute to ethionamide and prothionamide susceptibility in *Mycobacterium tuberculosis*. *MBio* **10**, e00616–e00619 (2019).
56. Sieber, R. N. et al. Genome investigations show host adaptation and transmission of LA-MRSA CC398 from pigs into Danish healthcare institutions. *Sci. Rep.* **9**, 18655 (2019).
57. Ma, K. C. et al. Adaptation to the cervical environment is associated with increased antibiotic susceptibility in *Neisseria gonorrhoeae*. *Nat. Commun.* **11**, 4126 (2020).
58. Chewapreecha, C. et al. Genetic variation associated with infection and the environment in the accidental pathogen *Burkholderia pseudomallei*. *Commun. Biol.* **2**, 428 (2019).
59. Jamrozny, D. et al. Increasing incidence of group B streptococcus neonatal infections in the Netherlands is associated with clonal expansion of CC17 and CC23. *Sci. Rep.* **10**, 9539 (2020).
60. Bianchi-Jassir, F. et al. Systematic review of Group B Streptococcal capsular types, sequence types and surface proteins as potential vaccine candidates. *Vaccine* <https://doi.org/10.1016/j.vaccine.2020.08.052> (2020).
61. Nicola, J. et al. Multilocus sequence typing system for group B Streptococcus. *J. Clin. Microbiol.* **41**, 2530–2536 (2003).
62. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
63. van Kassel, M. N. et al. Molecular epidemiology and mortality of group B streptococcal meningitis and infant sepsis in the Netherlands: a 30-year nationwide surveillance study. *Lancet Microbe* **2**, e32–e40 (2021).
64. Plainvert, C. et al. Multidrug-resistant hypervirulent group B streptococcus in neonatal invasive infections, France, 2007–2019. *Emerg. Infect. Dis.* <https://doi.org/10.3201/eid2611.201669> (2020).
65. O’Sullivan, C. P. et al. Group B streptococcal disease in UK and Irish infants younger than 90 days, 2014–15: a prospective surveillance study. *Lancet Infect. Dis.* **19**, 83–90 (2019).
66. Francois Watkins, L. K. et al. Epidemiology of invasive group B Streptococcal infections among nonpregnant adults in the United States, 2008–2016. *JAMA Intern. Med.* **179**, 479–488 (2019).
67. Angela, M. et al. Epidemiological characterization of group B Streptococcus infections in Alberta, Canada: an update from 2014 to 2020. *Microbiol. Spectr.* **9**, e01283–21 (2021).
68. Joubrel, C. et al. Group B streptococcus neonatal invasive infections, France 2007–2012. *Clin. Microbiol. Infect.* <https://doi.org/10.1016/j.cmi.2015.05.039> (2015).
69. Nanduri, S. A. 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.* **173**, 224–233 (2019).
70. Seale, A. C. et al. Maternal colonization with *Streptococcus agalactiae* and associated stillbirth and neonatal disease in coastal Kenya. *Nat. Microbiol.* <https://doi.org/10.1038/nmicrobiol.2016.67> (2016).
71. Young, B. C. et al. Antimicrobial resistance determinants are associated with *Staphylococcus aureus* bacteraemia and adaptation to the healthcare environment: a bacterial genome-wide association study. *Microb. Genomics* **7**, 000700 (2021).
72. Arnold, B. J., Huang, I.-T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* <https://doi.org/10.1038/s41579-021-00650-4> (2021).
73. Holley, G. & Melsted, P. Bifrost: Highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biol.* <https://doi.org/10.1186/s13059-020-02135-8> (2020).
74. Jaillard, M. et al. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. *PLoS Genet.* **14**, e1007758 (2018).
75. Rosini, R. et al. Genomic analysis reveals the molecular basis for capsule loss in the group B Streptococcus population. *PLoS ONE* **10**, e0125985 (2015).
76. Ramaswamy, S. V., Ferrieri, P., Flores, A. E. & Paoletti, L. C. Molecular characterization of nontypeable group B streptococcus. *J. Clin. Microbiol.* **44**, 2398–2403 (2006).
77. Slotved, H.-C., Fuursted, K., Kavalari, I. D. & Hoffmann, S. Molecular identification of invasive non-typeable group B Streptococcus isolates from Denmark (2015 to 2017). *Front. Cell. Infect. Microbiol.* **11**, 239 (2021).
78. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* <https://doi.org/10.1038/ng.2310> (2012).
79. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* <https://doi.org/10.1038/nmeth.1681> (2011).
80. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* <https://doi.org/10.1016/j.ajhg.2010.11.011> (2011).
81. Kremer, P. H. C. et al. Genetic variation in *Neisseria meningitidis* does not influence disease severity in meningococcal meningitis. *Front. Med.* **7**, 826 (2020).
82. Culyba, M. J. & Van Tyne, D. Bacterial evolution during human infection: Adapt and live or adapt and die. *PLoS Pathog.* **17**, e1009872 (2021).
83. Bijlsma, M. W. et al. Epidemiology of invasive meningococcal disease in the Netherlands, 1960–2012: an analysis of national surveillance data. *Lancet Infect. Dis.* **14**, 805–812 (2014).
84. Horváth-Puhó, E. et al. Mortality, neurodevelopmental impairments, and economic outcomes after invasive group B streptococcal disease in early infancy in Denmark and the Netherlands: a national matched cohort study. *Lancet Child Adolesc. Heal.* [https://doi.org/10.1016/S2352-4642\(21\)00022-5](https://doi.org/10.1016/S2352-4642(21)00022-5) (2021).
85. Deshayes de Cambronne, R. et al. CC17 Group B Streptococcus exploits integrins for neonatal meningitis development. *J. Clin. Invest.* <https://doi.org/10.1172/jci136737> (2021).
86. Hyams, C., Camberlein, E., Cohen, J. M., Bax, K. & Brown, J. S. The Streptococcus pneumoniae capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. *Infect. Immun.* **78**, 704–715 (2010).
87. Rubens, C. E., Wessels, M. R., Heggen, L. M. & Kasper, D. L. Transposon mutagenesis of type III group B Streptococcus: correlation of capsule expression with virulence. *Proc. Natl Acad. Sci. USA* **84**, 7208–7212 (1987).
88. Berti, F. et al. Structure of the type IX Group B Streptococcus capsular polysaccharide and its evolutionary relationship with types V and VII. *J. Biol. Chem.* **289**, 23437–23448 (2014).
89. Tazi, A. et al. The surface protein HvgA mediates group B streptococcus hypervirulence and meningeal tropism in neonates. *J. Exp. Med.* **207**, 2313–2322 (2010).
90. Farhat, M. R. et al. GWAS for quantitative resistance phenotypes in *Mycobacterium tuberculosis* reveals resistance genes and regulatory regions. *Nat. Commun.* **10**, 2128 (2019).
91. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* <https://doi.org/10.1089/cmb.2012.0021> (2012).
92. Claire, P. et al. Multiplex PCR assay for rapid and accurate capsular typing of group B Streptococci. *J. Clin. Microbiol.* **45**, 1985–1988 (2007).
93. Fanrong, K. et al. Use of phenotypic and molecular serotype identification methods to characterize previously nonserotypeable Group B streptococci. *J. Clin. Microbiol.* **46**, 2745–2750 (2008).
94. Inouye, M. et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).
95. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2**, e000056–e000056 (2016).
96. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkz361> (2019).
97. Tavaré, S. *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences* (National Agricultural Library, 1986).
98. Yang, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* <https://doi.org/10.1093/oxfordjournals.molbev.a040082> (1993).
99. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
100. Keck, F., Rimet, F., Bouchez, A. & Franc, A. phylsignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* **6**, 2774–2780 (2016).
101. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

102. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* <https://doi.org/10.1086/519795> (2007).
103. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
104. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
105. Turner, S. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. <https://doi.org/10.1101/005165> (2014).
106. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
107. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Acknowledgements

The authors would like to thank the study participants and guardians, the clinical and laboratory staff who collected and processed the samples at various laboratories in the Netherlands, and the sequencing, core, and pathogen teams, and the Bentley lab at the Wellcome Sanger Institute for their support and feedback on the genomic analysis. We would also like to thank Dr. John Lees at European Bioinformatics Institute for providing advice on the GWAS. The study was funded by the Meningitis Research Foundation Project grant 1502.0 (A.v.d.E.), Wellcome Trust grant 098051 (D.J., S.D.B.), Netherlands Organization for Health Research and Development (ZonMw; NWO-Vici 918.19.627 (D.v.d.B.), Amsterdam Medical Centre Innovation grant (D.v.d.B.), and the Bill and Melinda Gates Foundation grant for the Juno project (SDB) [<https://www.gbgsen.net/>]. C.C. and S.D.B. were supported by funding from the Joint Initiative for Antimicrobial Resistance (JPIAMR) grant no. MR/R003076/1 (S.D.B.), the Bill and Melinda Gates Foundation grant number OPP1034556 (S.D.B.), and Wellcome Trust (2016–2021 core award grant no. 206194).

Author contributions

C.C., D.J., M.W.B., and S.D.B. conceived the study. A.v.d.E., T.W.K., M.W.B., and D.v.d.B. collected and processed the samples for whole-genome sequencing. D.J. and S.D.B. oversaw whole-genome sequencing. C.C. and D.J. conducted the analysis. D.J., M.W.B., and S.D.B. contributed to the data interpretation and discussions. D.J., A.v.d.E., D.v.d.B., T.W.K., and S.D.B. administered the project. D.J., S.D.B., D.v.d.B., A.v.d.E., and S.D.B. acquired funding for the study. C.C., D.J., and S.D.B. drafted the first version of

the manuscript. The manuscript was reviewed by C.C., D.J., A.v.d.E., T.W.K., M.W.B., D.v.d.B., and S.D.B.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-31858-4>.

Correspondence and requests for materials should be addressed to Chrispin Chaguz, Arievander Ende or Stephen D. Bentley.

Peer review information *Nature Communications* thanks Thomas Hooven and the other, anonymous, reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022