Optimizing nanopore adaptive sampling for pneumococcal serotype surveillance in complex samples using the graph-based GNASTy algorithm

Samuel T. Horsfield, Basil C.T. Fok, Yuhan Fu, et al.

Genome Res. published online March 4, 2025
Access the most recent version at doi:10.1101/gr.279435.124

Optimising Nanopore Adaptive Sampling for pneumococcal serotype surveillance in complex samples using the graph-based GNASTy algorithm.

Samuel T. Horsfield[1,2,**], Basil Fok[1], Yuhan Fu[1], Paul Turner[3], John A. Lees[1,2*], and Nicholas J. Croucher[1*]

[1]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, W12 0BZ, UK
[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK
[3]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, OX3 7LG, UK
[*]These authors contributed equally to this work.
[**]Corresponding author - shorsfield@ebi.ac.uk

ABSTRACT

Serotype surveillance of *Streptococcus pneumoniae* (the pneumococcus) is critical for understanding the effectiveness of current vaccination strategies. However, existing methods for serotyping are limited in their ability to identify co-carriage of multiple pneumococci and detect novel serotypes. To develop a scalable and portable serotyping method that overcomes these challenges, we employed Nanopore Adaptive Sampling (NAS), an on-sequencer enrichment method which selects for target DNA in real-time, for direct detection of *S. pneumoniae* in complex samples. Whereas NAS targeting the whole *S. pneumoniae* genome was ineffective in the presence of non-pathogenic streptococci, the method was both specific and sensitive when targeting the capsular biosynthetic locus (CBL), the operon that determines *S. pneumoniae* serotype. NAS significantly improved coverage and yield of the CBL relative to sequencing without NAS, and accurately quantified the relative prevalence of serotypes in samples representing co-carriage. To maximise the sensitivity of NAS to detect novel serotypes, we developed and benchmarked a new pangenome-graph algorithm, named GNASTy. We show that GNASTy outperforms the current NAS implementation, which is based on linear genome alignment, when a sample contains a serotype absent from the database of targeted sequences. The methods developed in this work provide an improved approach for novel serotype discovery and routine *S. pneumoniae* surveillance that is fast, accurate and feasible in low resource settings. Although NAS facilitates whole genome enrichment under ideal circumstances, GNASTy enables targeted enrichment to optimise serotype surveillance in complex samples.

Keywords Pneumococcus · Nanopore · Adaptive Sampling · Graphs

Running title

Graph-based Nanopore Adaptive Sampling Typing.

Introduction

Streptococcus pneumoniae (the 'pneumococcus') is a human nasopharyngeal commensal which can cause severe diseases, such as pneumonia, bacteremia and meningitis, disproportionately affecting young children and the elderly (Weiser *et al.* 2018). *S. pneumoniae* infections cause a significant global health burden, being associated with more than 800,000 deaths annually (Ikuta *et al.* 2022), and are the leading cause of death in children under 5 years of age (Wahl *et al.* 2018; Wang *et al.* 2016). The species can be divided into >100 serotypes (Ganaie *et al.* 2020), each of which expresses an immunologically-distinct polysaccharide capsule that enables the bacterium to evade the host's immune response (Hyams *et al.* 2010).

Polysaccharide conjugate vaccines (PCVs) target a subset of *S. pneumoniae* serotypes that cause a substantial proportion of invasive pneumococcal disease (IPD) (Croucher *et al.* 2018), driving a reduction in the global IPD burden (Wahl *et al.* 2018). This is achieved through a significant perturbation of the pneumococcal population carried in the nasopharynx. Consequently, vaccine-targeted serotypes have been replaced through the expansion of already common serotypes not included in current formulations, and the emergence of previously rare or unknown serotypes, changing the frequency of antimicrobial resistance (AMR) and incidence of disease in *S. pneumoniae* (Lo *et al.* 2019; van Tonder *et al.* 2019; Ladhani *et al.* 2018). Ongoing serotype surveillance is critical to identify significant increases in non-vaccine serotype prevalence, particularly if a serotype is associated with AMR or high invasiveness (Lo *et al.* 2022). Such dynamics can be monitored through analysis of nasopharyngeal samples, although the frequent carriage of multiple serotypes within a single individual, known as 'co-carriage' or 'co-colonisation', makes identification of all circulating serotypes challenging (Huebner *et al.* 2000). This problem is exacerbated by the recent discovery that minority serotypes are often present at low frequency ($< 25\%$ of pneumococcal cells within an individual), but are still responsible for a notable proportion of transmission events (Tonkin-Hill *et al.* 2022). Therefore, scalable high sensitivity serotype assays that can deconvolute mixed samples, and identify novel serotypes, are necessary to update vaccine formulations and public health strategies in response to pneumococcal epidemiological dynamics (Colijn *et al.* 2020).

The original methods for serotyping pneumococci assay the ability of an unknown isolate to agglutinate in the presence of different antisera that recognise known serotypes (Turner *et al.* 2011; Habib *et al.* 2014). Agglutination assays have high specificity when applied by experts, but have extensive training requirements, as precise typing requires a succession of tests with different antisera (Satzke *et al.* 2015). When applied to individual colonies, such methods have a low sensitivity for detecting co-carriage, although this can be improved using latex agglutination of plate sweeps (Turner *et al.* 2011). Nevertheless, these methods cannot identify novel serotypes; these may be discovered through whole genome sequencing (WGS) approaches, which detect specific sequence variants of the Capsular Biosynthetic Locus (CBL) (Sheppard *et al.* 2022; Epping *et al.* 2018; Kapatai *et al.* 2016), the operon which defines pneumococcal serotype (Bentley *et al.* 2006). Yet WGS of individual colonies is difficult to deploy at scale in resource-limited settings as it is expensive and time-consuming, requiring specific expertise and access to specialist laboratory equipment

(Jauneikaite *et al.* 2015). The limited number of colonies from a single patient that can be feasibly sequenced limits the ability of WGS to detect co-carriage, unless a sample is subjected to deep sequencing (Tonkin-Hill *et al.* 2022). However, this reduces the number of samples that can be analysed, and therefore lowers overall throughput. Additionally, both agglutination assays and WGS rely on prior selective culture of *S. pneumoniae* as means of enrichment to improve sensitivity. Selective culture adds additional time, resource and expertise requirements to already complex laboratory workflows, limiting throughput, and potentially resulting in false negatives if cells fail to grow (Ricketson *et al.* 2021). Purely genotypic approaches, such as PCR and DNA microarrays, target CBL DNA sequences directly present in the sample and therefore do not require selective culture. These methods can identify co-carriage, and are less laborious and expensive than agglutination assays or WGS, and can therefore be used in high-throughput settings (Jauneikaite *et al.* 2015). However, these methods require target CBL sequences to be specified *a priori*, and so cannot detect novel serotypes. Overall, no current serotyping method can scalably and sensitively identify both known and novel serotypes, as well as co-carriage.

Novel nucleotide sequencing approaches have the potential to allow accurate, simple and relatively inexpensive culture-free *S. pneumoniae* surveillance. Nanopore sequencing, developed by Oxford Nanopore Technologies (ONT), is a portable long-read nucleotide sequencing technology in which DNA or RNA molecules are sequenced as they move across an impermeable membrane through protein nanopores (Ip *et al.* 2015; Quick *et al.* 2016). Reads are generated in real-time, enabling on-flowcell enrichment of sequences of interest, referred to as 'target' DNA, via rejection of all other sequences, referred to as 'nontarget' DNA. These methods, known collectively as Nanopore Adaptive Sampling or 'NAS', align the first segment of DNA fragments as they pass through a nanopore to a reference database, before sending a signal back to the sequencer to either 'accept', where the fragment is sequenced to completion, or 'reject', where voltage across the nanopore is reversed, ejecting the fragment (Payne *et al.* 2021). NAS increases target sequence yield by rejecting nontarget DNA, increasing the sensitivity for detecting of sequences of interest (Payne *et al.* 2021; Weilguny *et al.* 2023). This makes NAS well-suited for metagenomics, the culture-free DNA sequencing-based analysis of mixed samples (Ye *et al.* 2019), such as nasopharyngeal communities. NAS has been shown to increase target yield approximately four-fold (Marquet *et al.* 2022; Su *et al.* 2023), and by extension enabling multiplexing of samples on ONT devices to increase throughput. Furthermore, increased target yield has been shown to improve accuracy of downstream analyses such as variant calling and assembly when analysing metagenomes (Weilguny *et al.* 2023; Martin *et al.* 2022; Wrenn and Drown 2023).

NAS is available as part of the standard ONT sequencing software platform. However, there has been limited quantification of its accuracy, particularly in metagenomics. It has been previously shown that NAS sensitivity is highest when a target present in a metagenome is closely related to a sequence in the reference database (Martin *et al.* 2022; Viehweger *et al.* 2023). However, high genetic relatedness between nontarget and target taxa in the same sample has the potential to negatively impact NAS specificity, as target and nontarget reads will be more difficult to distinguish between during the rejection process. The sequence similarity between *S. pneumoniae* and other members of the *Streptococcus* genus (Marttinen *et al.* 2015; D'Aeth *et al.* 2021), which are also present as part of the upper respiratory

tract microbiome (Bek-Thomsen *et al.* 2008), is comparable with the error rate of individual ONT reads (Delahaye and Nicolas 2021). Hence attempts to enrich for a whole *S. pneumoniae* genome may be limited by the challenge of resolving pneumococcal DNA from that of non-pathogenic streptococci. Alternatively, targeting loci which are specific to *S. pneumoniae* will improve target enrichment, as such sequence are typically absent from benign commensals. Hence CBL sequences are a promising candidate for targeted metagenomics enrichment (Bentley *et al.* 2006; Croucher *et al.* 2018; Løchen *et al.* 2022).

Here, we benchmark the sensitivity and specificity of NAS for detection of *S. pneumoniae* in mixed samples, and assess the ability of NAS to quantify serotype prevalence in co-carriage samples using different target databases. To improve performance when detecting novel serotypes, we develop a graph-based bioinformatic method for NAS, named GNASTy (Graph-based Nanopore Adaptive Sampling Typing, pronounced 'nasty'), and benchmark it against the current NAS implementation, which uses linear alignment. Overall, we demonstrate the advantages and caveats of NAS for application in metagenome-based *S. pneumoniae* surveillance, and introduce a new method for detection and discovery of novel serotypes in metagenomes.

Results

NAS performance depends on microbiome composition

We first set out to determine the taxonomic range across which NAS can effectively enrich for a target sequence, whilst still correctly rejecting nontarget sequences. We hypothesised that NAS would fail to enrich for target loci when the sequence similarity between target and nontarget genomes was comparable to the single-strand error rate of ONT reads ($\sim 6\%$ (Delahaye and Nicolas 2021)), resulting in incorrect selection of nontarget DNA that ultimately reduces target enrichment.

To test this hypothesis, we generated mock communities containing mixtures of genomic DNA from *S. pneumoniae*, the serotype 23F pneumococcal isolate ATCC 700669 (Croucher *et al.* 2009), referred to as 'Spn23F', with that of closely and distantly related nontarget species. Spn23F DNA was mixed with DNA from species from a different phylum, represented by *Escherichia coli* DH5-α; the same genus but different species, represented by *Streptococcus mitis* SK142; and the same species but different strain, represented by *S. pneumoniae* R6 (Figure 1a). To test NAS sensitivity at low target DNA concentrations, Spn23F DNA was titrated from a proportion of 0.5 to 0.001 ($50\% - 0.1\%$) (Figure 1b) in nontarget DNA. These proportions describe the ratio of total DNA bases within a sample which belong to target DNA. The choices of alignment parameters used for NAS performance comparisons are detailed in the Supplemental Material (Section C.1, Supplementary Table 3, Supplementary Figures 15 and 16). All libraries were size-selected to remove DNA fragments < 10 kb in length, as this was shown to improve enrichment (Supplemental Material, Section C.2, Supplementary Tables 4-6, Supplementary Figures 17-21). NAS was carried out using Readfish (Payne *et al.* 2021), targeting either the whole Spn23F genome or 23F CBL (Figure 1c). All samples were multiplexed into a combined sequencing library and run on a single flow cell to control for batch effects. Half of the 'channels' (a

group of four pores, of which only one is sequencing at one time) sequenced the library using NAS, while the other half sequenced the same library normally without NAS (termed 'control'). Splitting the flow cell in this way provides an internal control which is used for calculation of enrichment by composition (referred to further as 'enrichment', see Methods) (Martin *et al.* 2022). Using enrichment allows direct comparison of NAS performance across sequencing runs which may otherwise be confounded by between-run variability. Enrichment > 1 indicates that a target was successfully enriched, with a greater proportion of target bases generated using NAS relative to the control.

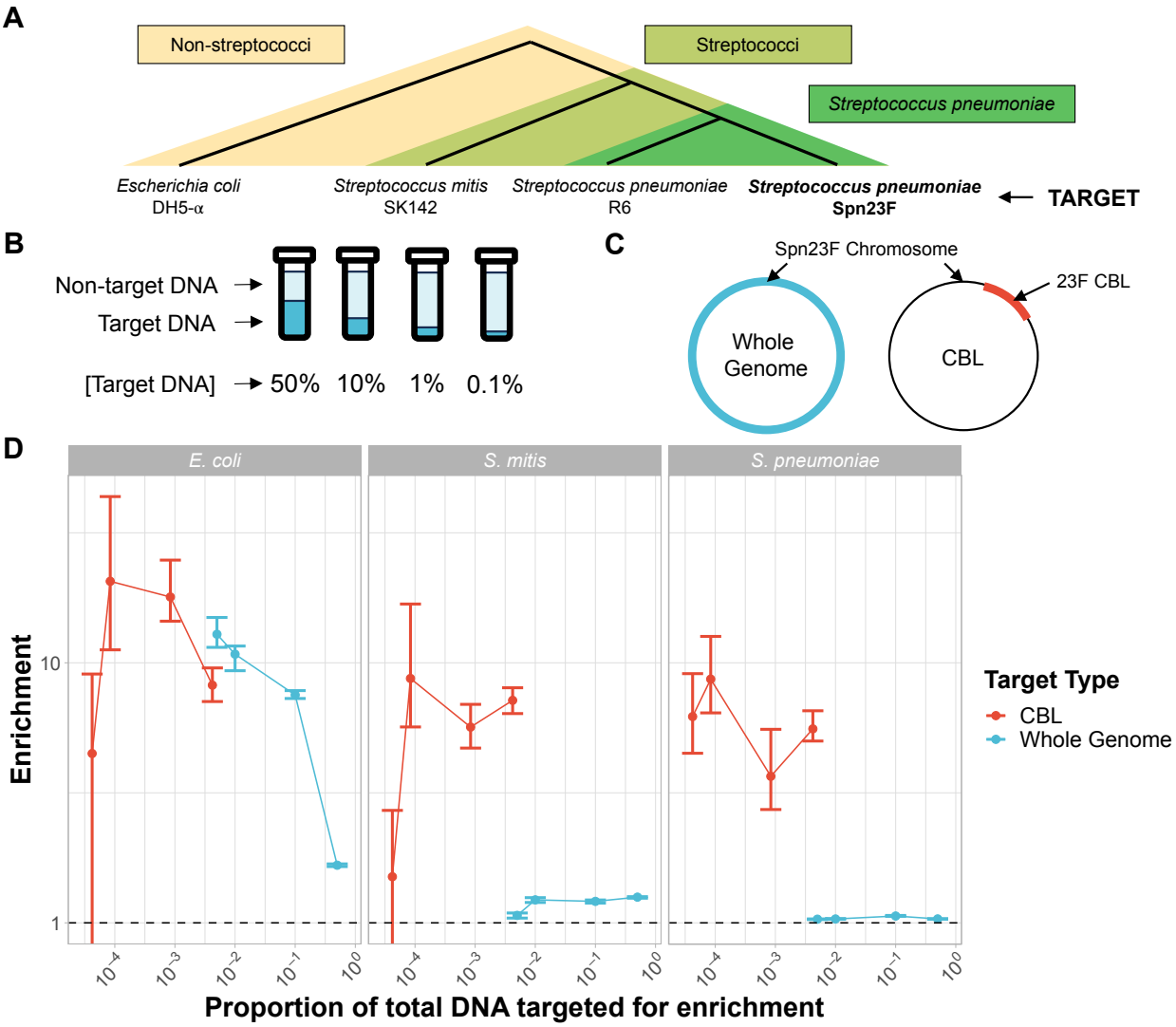

Figure 1: Enrichment of *S. pneumoniae* Spn23F in samples containing closely and distantly related nontarget species. (**a**) Representation of evolutionary relatedness of nontarget species and the target, *S. pneumoniae* Spn23F. (**b**) Experimental setup of target DNA dilution series with nontarget DNA. (**c**) Representation of two enrichment experiments, either targeting the whole Spn23F genome (blue) or the 23F CBL sequence present on the Spn23F Chromosome. (**d**) Enrichment results of Spn23F whole genome or 23F CBL at different concentrations of target DNA. Bar ranges are inter-quartile range of enrichment from 100 bootstrap samples of reads. Data points connected by lines are observed enrichment values for each library, with solid lines connecting target DNA diluted at different concentrations with nontarget DNA. Columns describe the nontarget species within each mixture. To plot on a log scale, all enrichment values had 0.01 added to them. Horizontal dashed line describes enrichment = 1 i.e. no enrichment has occurred.

Comparison of NAS performance based on enrichment is shown in Figure 1d (blue). Spn23F whole genome enrichment was highest in mixtures containing *E. coli* for all target proportions. Conversely, mixture with *S. mitis* and *S. pneumoniae* resulted in notably lower enrichment, although enrichment was slightly higher in mixtures with *S. mitis*. For example, for the 0.1 target dilution, enrichment of the Spn23F genome was 7.51, 1.20 and 1.05 for the *E. coli*, *S. mitis* and *S. pneumoniae* mixtures respectively. Additionally, enrichment increased monotonically as Spn23F DNA proportions decreased in *E. coli* mixtures, as observed previously (Martin *et al.* 2022), whilst for mixtures with *S. mitis* and *S. pneumoniae* enrichment remained relatively constant between dilutions. These results indicate that the NAS alignment process is not able to effectively reject sequences from nontarget species when their divergence is similar to the read error rate. This result has particular significance for use of NAS in *S. pneumoniae* surveillance, as the presence of commonly co-occurring streptococci in the nasopharynx greatly impacts NAS performance.

To determine the effect of non-specific enrichment on downstream analyses, we then assembled reads using MetaFlye (Kolmogorov *et al.* 2020) and analysed assembly quality using Inspector (Chen *et al.* 2021), overlaying results on the Spn23F reference genome for the 0.1 target DNA dilutions. We compared the relative read coverage and aligned contig coverage of the reference genome, as well as the presence of small (< 50 bp) and large (≥ 50 bp) assembly errors (Figure 2). Greater coverage by aligned contigs indicates that read coverage, and therefore target yield, was sufficiently high to generate a contiguous assembly, whilst presence of small or large errors suggests problems with the assembly process, such as insufficient read coverage or integration of nontarget reads into assemblies.

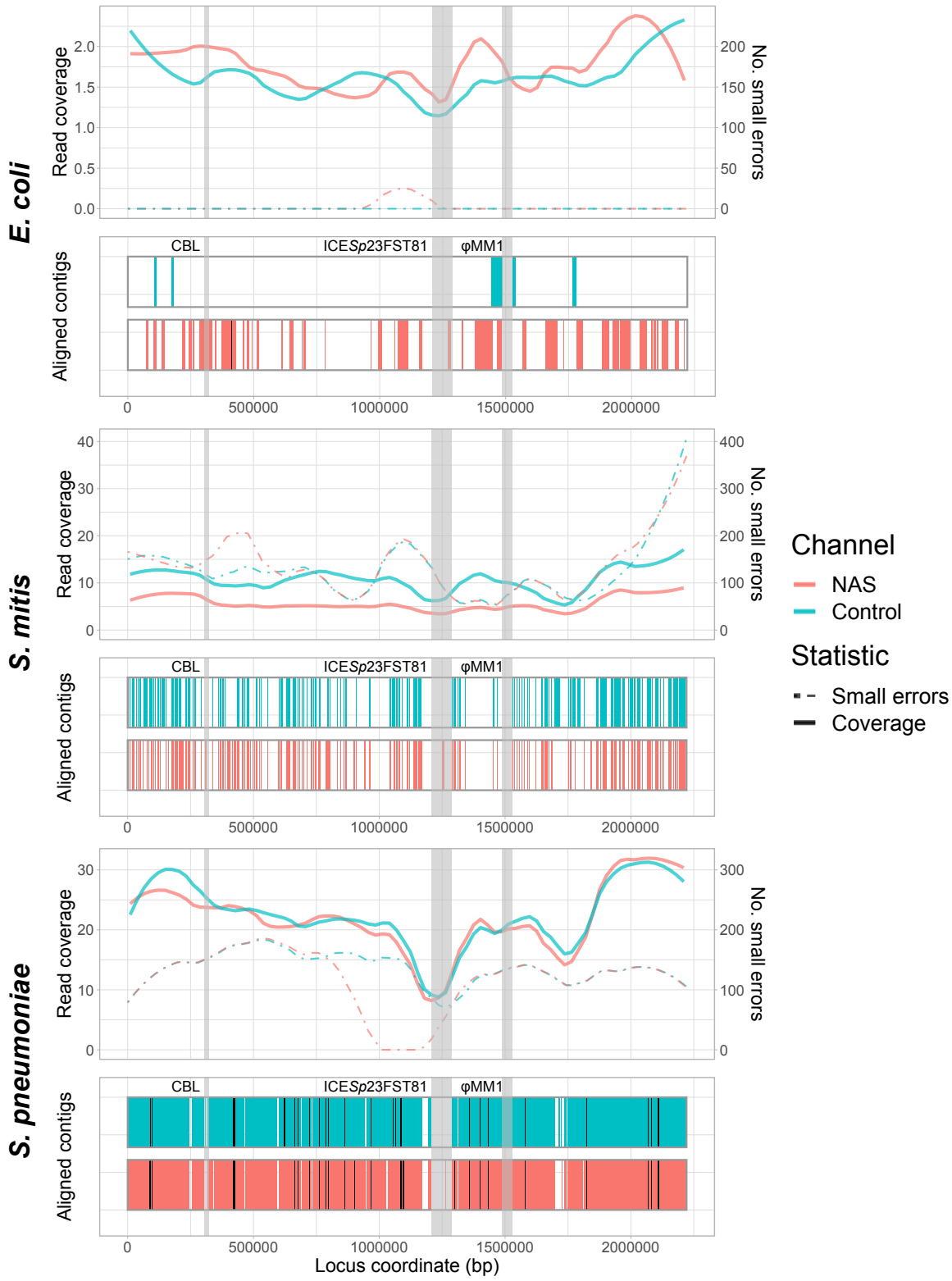

Figure 2: Spn23F whole genome enrichment assembly comparison. Each panel describes a Spn23F assembly generated from 0.1 Spn23F dilutions with each nontarget organism. For each panel, the top plot shows the read coverage (solid), defined as the absolute number of bases aligning to a locus, and number of small errors ($\leq$ 50 bp, dashed), whilst the bottom plot shows aligned contigs (colours) and large errors (black bars, $>$ 50 bp) in each assembly. Loci of interest are annotated by grey bars; CBL, as well as ICE*Sp*23FST81 and ψMM1 prophage which are missing in this isolate of Spn23F (Croucher *et al.* 2012).

For the *E. coli* mixture, the Spn23F whole genome assembly contained very few errors, although read coverage was low and the assembly covered only a small portion of the Spn23F genome (Figure 2, top). The resulting assembly from NAS channels had a greater coverage of aligned contigs than that from control channels, coupled with higher read coverage across the Spn23F genome. The assemblies from the dilution with *S. mitis* had greater overall genome coverage than the equivalent *E. coli* mixture, although the respective aligned contigs were short and contained larger numbers of small errors (Figure 2, middle). Based on read coverage, which was higher in the *S. mitis* mixture over *E. coli* despite Spn23F being at equivalent concentrations, these errors are likely due to incorporation of nontarget *S. mitis* reads into Spn23F assemblies, ultimately resulting in mismatches with the reference sequence. The dilution of Spn23F with *S. pneumoniae* R6 produced assemblies with the greatest coverage of aligned contigs, although the assemblies also had large numbers of both small and large errors (Figure 2, bottom). There was also a gap in assemblies at the 23F CBL; *S. pneumoniae* R6 is unencapsulated and so does not possess a CBL, meaning that these assemblies likely contained a large number of nontarget *S. pneumoniae* R6 reads. Comparing NAS and control assemblies across all mixtures, both read and assembly coverage were similar for the *S. mitis* and *S. pneumoniae* mixtures between control and NAS channels, whilst for *E. coli* the NAS channels outperformed the control channels. These results highlight the inability of NAS to distinguish between closely-related target and nontarget sequences, resulting in lowered enrichment and chimeric assemblies.

NAS can effectively enrich for pneumococcal CBL

To improve enrichment by NAS, we specifically enriched for the pneumococcal CBL, which is generally absent from streptococci other than *S. pneumoniae* (Bentley *et al.* 2006). We sequenced the same library as described in Figure 1, targeting 106 distinct pneumococcal CBL sequences using NAS (see Methods for details of sequences), measuring the enrichment of the 23F CBL present in the Spn23F genome. Targeting all known CBL sequences using NAS would be the best approach when serotyping a novel isolate in the field, as this practice would provide the highest probability of detecting and enriching for a previously observed serotype. Most CBL are approximately 20 kb, with a 2.2 Mb genome, and therefore the enrichment values were scaled by 8×10^{-3} to account for the smaller target sequence size.

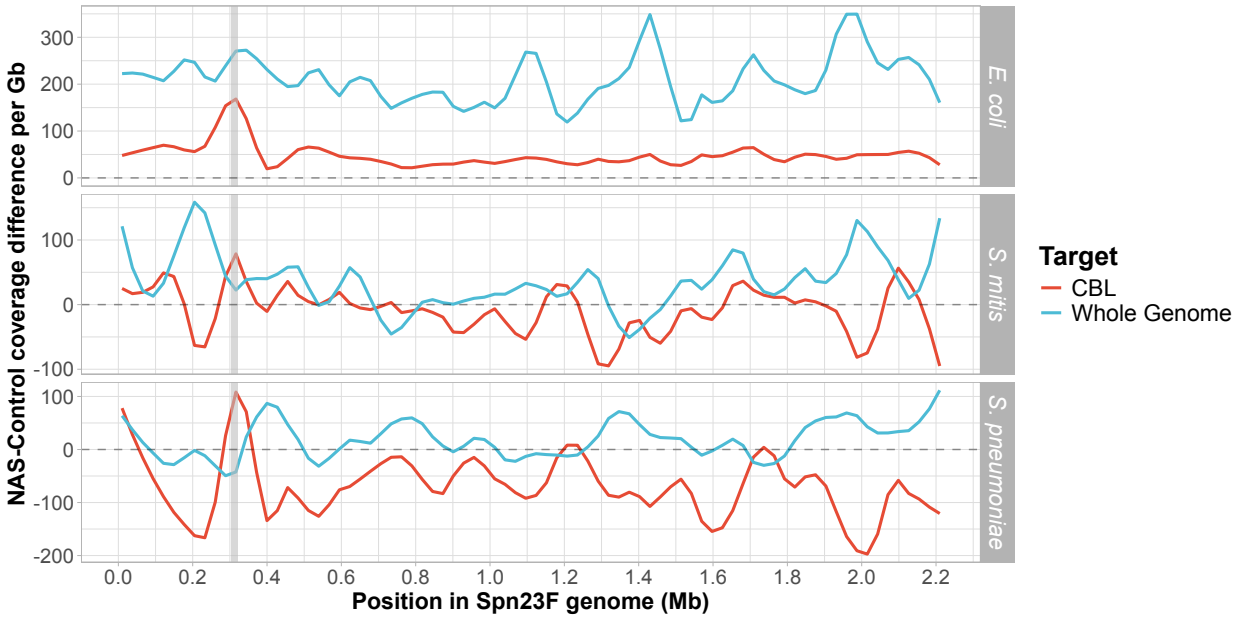

Figure 3: Difference in normalised coverage per locus between NAS and control channels across the Spn23F Chromosome when targeting whole genome (blue) or CBL (red) from 0.1 Spn23F dilutions with each nontarget organism. NAS-control coverage difference per gigabase (Gb) was calculated by normalising the read coverage for each locus by the amount of data generated (in Gb) for each respective sample and channel, and then negating the normalised coverage for control channels from NAS channels for each locus. The grey dashed line at 0 indicates equivalent coverage at a given locus between NAS and control channels; > 0 indicates NAS channels generated greater coverage, < 0 indicates control channels generated greater coverage. Data shown for 0.1 dilutions of Spn23F only. Grey column in each plot highlights 23F CBL locus. Rows show different species for the nontarget which was mixed with Spn23F in each sample.

We observed a notable improvement in enrichment when only targeting the 23F CBL, particularly in mixtures containing *S. mitis* and *S. pneumoniae* (Figure 1d, red). For example, for the 4×10^{-3} target dilution, CBL enrichment was 7.16 and 5.46, whilst for the 5×10^{-3} target dilution, whole genome enrichment was 1.06 and 1.02 for *S. mitis* and *S. pneumoniae* mixtures respectively. For the *E. coli* mixture with Spn23F at 0.1 proportion, the coverage difference between NAS and control channels at the 23F CBL locus was greater when enriching for the whole Spn23F genome than for the 23F CBL, whilst the reverse was true for the *S. mitis* and *S. pneumoniae* mixtures (Figure 3). These results indicate that when a nontarget species is sufficiently divergent from target species, both whole genome and CBL enrichment are viable means of serotyping, exemplified by the *E. coli* mixture. However, directly targeting the CBL boosts NAS performance when nontarget species are closely related to the target, exemplified by the *S. mitis* and *S. pneumoniae* mixtures. Therefore, CBL sequences are sufficiently divergent from the rest of the *S. pneumoniae* genome, as well as other closely related genomes, to be differentiated and enriched for. We did not observe the same monotonic increase in CBL enrichment with decreasing target concentration, as observed with whole genome enrichment. Additionally, bootstrap inter-quartile ranges were wider for CBL samples compared to whole genome samples. This is consistent with Martin *et al.* (2022), in which a predictive model of target enrichment was less accurate at lower target concentrations, indicating that low target concentrations produce more noisy enrichment measures. Overall, CBL enrichment works consistently, independently of the population composition, whilst whole genome en-

richment is dependent on concomitant nontarget species. Furthermore, 23F DNA was still detectable at the lowest concentration tested, meaning that NAS can enrich for target DNA at concentrations as low as 1 in 10,000 bases (targeting 20 kb of 2.2 Mb pneumococcal genome ($\sim 1\%$) in a 0.01 dilution) in a mixed sample. Taken together, these results indicate that targeting CBL for identification and serotyping of *S. pneumoniae* is a viable alternative to whole genome enrichment in complex microbial samples.

We then generated and compared 23F CBL assemblies as before, focusing on 0.1 Spn23F dilutions, equating to 8×10^{-4} 23F CBL DNA (Figure 4). For mixtures containing *E. coli* and *S. mitis*, NAS channels generated more read coverage than control channels, resulting in more complete 23F CBL assemblies containing very few errors. For both whole genome and CBL assemblies for mixtures containing *E. coli*, we observed slightly higher numbers of small errors for NAS over control channels (Figures 2, top and 4, top). However, the numbers of small errors for *E. coli* mixtures are relatively low in comparison to the other mixtures, where we additionally observed similar or lower numbers of small errors for NAS compared to control channels, meaning these small errors are likely random noise. For the mixture containing *S. pneumoniae* R6, the 23F CBL assembly was conversely more complete for control channels, likely due to low read counts making the assembly process noisy and leading to patchy coverage for both NAS and control channels (Supplementary Table 2). Overall, NAS improved assembly quality at lower target DNA concentrations over normal sequencing, although low read count made assembly accuracy more variable between samples.

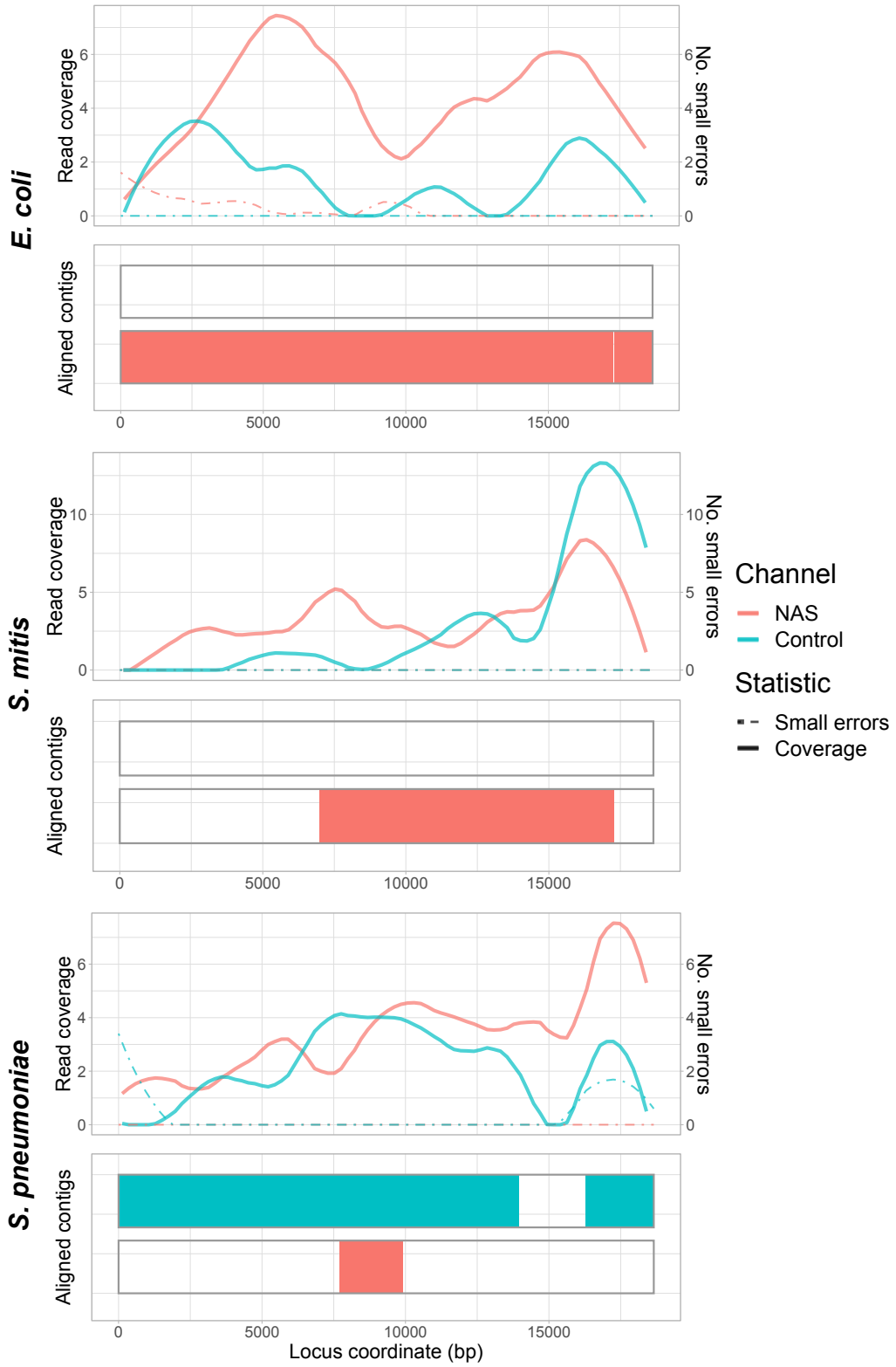

Figure 4: Spn23F CBL enrichment assembly comparison. Each panel describes a 23F CBL assembly generated from 0.1 Spn23F dilutions (8×10^{-4} 23F CBL proportion) with each nontarget organism. For each panel, the top plot shows the read coverage (solid), defined as the absolute number of bases aligning to a locus, and number of small errors (≤ 50 bp, dashed), whilst the bottom plot shows the aligned contigs (colours) and large errors (black bars > 50 bp) in each assembly.

Previous studies have shown that although NAS increases the proportion of target bases within the read dataset, it may reduce the absolute yield for an equivalent sequencing time (Payne *et al.* 2021; Martin *et al.* 2022). In these instances, normal sequencing would give increased coverage of the target genome, and therefore NAS should be avoided. To determine whether this was the case with enrichment of the whole Spn23F genome and 23F CBL, we compared the total number of bases aligning to target sequences across control and NAS channels (Supplementary Figure 1). For whole genome enrichment, the absolute yield was lower for NAS channels on average; however, this difference was not significant. For CBL enrichment, there was a significant increase in absolute yield using NAS (2.17-fold on average, p=0.0049). Furthermore, mean read lengths for aligned reads were 4.2-fold higher ($p=7\times10^{-6}$) on average in NAS channels than unaligned reads for CBL enrichment, whilst there was no difference for whole genome enrichment (p=0.37) (Supplementary Figure 2, results for unselected libraries in Supplementary Figure 3). Greater difference in read length indicates better performance of NAS, as short unaligned reads and long aligned reads suggest correct rejection of nontarget sequences and acceptance of target reads respectively (Payne *et al.* 2021). Therefore, when targeting sequences that are divergent from nontarget DNA, NAS increases both proportional and absolute yield due to better distinction between target and nontarget reads.

NAS can simultaneously enrich for multiple pneumococcal CBL in the same mixture

NAS is therefore capable of distinguishing encapsulated *S. pneumoniae* from other streptococci, but effective serotype surveillance requires the identification of multiple serotypes in cases of co-carriage. CBL are highly structurally diverse (Bentley *et al.* 2006), potentially allowing differentiation of multiple CBL in co-carriage by phasing contiguous structural variants using long reads (Cretu Stancu *et al.* 2017). To determine whether NAS can differentiate and enrich for multiple CBL sequences, we generated a set of mock communities where Spn23F was mixed in 50:50 proportions with other *S. pneumoniae* strains with different genotypes and serotypes (Figure 5a). We then targeted CBL sequences using NAS; however, we increased the number of times a read can be realigned to the reference sequence before it is rejected ('maxchunks' = 4, rather than 0) to determine whether this would improve enrichment of poorly aligned short reads.

All CBL sequences were enriched in across all mixtures, independent of serotype or genotype (Figure 5b), with NAS significantly increasing the yield of reads aligning to the CBL locus relative to control channels by 1.9-fold on average ($p=9.5\times10^{-7}$) (Supplementary Figure 4). Therefore, NAS can be used for targeted sequencing in cases of co-carriage, regardless of respective *S. pneumoniae* serotypes or genotypes. However, CBL enrichment was slightly lower than that observed in mixtures containing a single encapsulated isolate at equivalent concentrations. Comparing the enrichment of the 23F CBL in the 50:50 mixture with the unencapsulated strain, R6, (Figure 5b, GPSC622), with that observed previously (Figure 1d, 4×10^{-3} target dilution with *S. pneumoniae*), enrichment was reduced (4.9 vs. 5.6). Therefore, increasing the 'maxchunks' had a detrimental impact on enrichment and should kept at zero.

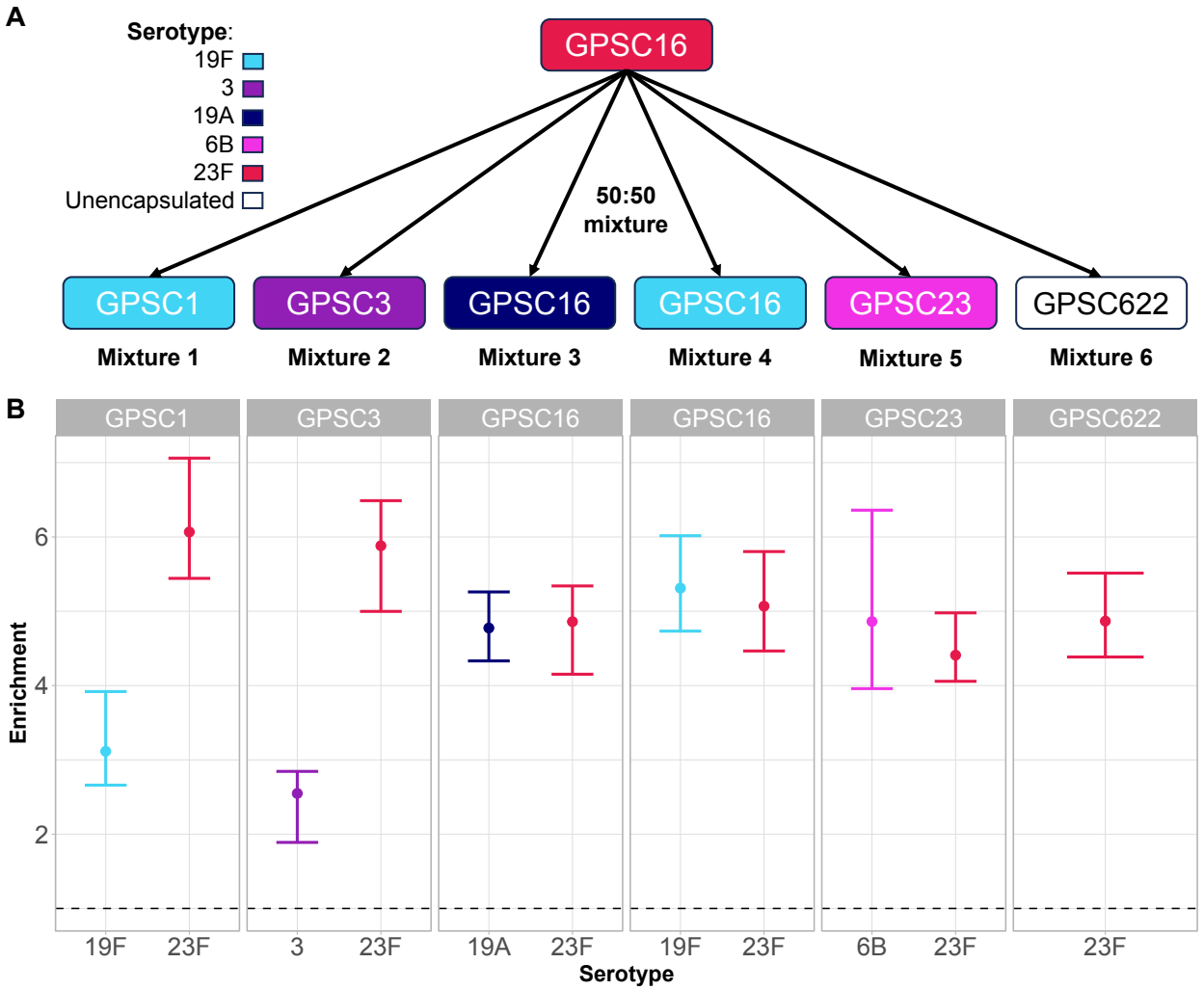

Figure 5: CBL enrichment in mixtures of multiple pneumococci. (**a**) Experimental setup. Spn23F DNA (GPSC16, serotype 23F, red) was mixed in 50:50 proportions with other *S. pneumoniae* isolates with different serotypes (given by colour) and genotypes (given by Global Pneumococcal Sequence Cluster, GPSC). (**b**) Enrichment of multiple CBL in mixtures. Bar ranges are inter-quartile range of enrichment from 100 bootstrap samples of reads. Data points are observed enrichment values for each CBL per library. X-axis and colour describes the serotype combination of the *S. pneumoniae* isolate mixed with Spn23F, columns describe the GPSC. Dashed line describes enrichment = 1 i.e. no enrichment has occurred. GPSC = Global Pneumococcal Sequence Cluster.

To determine whether NAS improves serotype prediction accuracy in mixed samples, we then analysed reads from mixed samples using PneumoKITy (Sheppard *et al.* 2022), a tool for pneumococcal serotype prediction from read data. Serotype predictions were correct for all but mixture 1 using reads from NAS channels, which missed a prediction of 19F (Table 1); however, this serotype was successfully identified in mixture 3. Although we did observe enrichment of the 19F CBL in this mixture (Figure 5b) the number of bases generated by NAS (53.4 kb) and control (32.4 kb) channels was not sufficient for PneumoKITY to confidently assign 19F as being present in the mixture. For control channels, three mixtures had incorrect serotype predictions, where either one or both serotypes were missed. For samples where two serotypes were correctly predicted, estimated proportions did not deviate substantially from 50%,

the expected values for these mixtures, and were similar between NAS and control channels. Therefore, NAS improved the accuracy of co-carriage detection over normal sequencing.

Table 1: Serotype predictions from mixed samples. Each row describes a mixture from Figure 5, with the expected serotypes in the 50:50 mixtures and relative proportions estimated by PneumoKITy (Sheppard *et al.* 2022) for reads generated from NAS and control channels. Prediction errors are highlighted in red. 'None' represents the unencapsulated isolate.

	Expected		NAS		Control	
Mixture	**Serotype A**	**Serotype B**	**A %**	**B %**	**A %**	**B %**
1	23F	19F	100	0	0	0
2	23F	3	36.4	63.6	0	0
3	23F	19A	39.1	60.9	36.4	63.6
4	23F	19F	41.2	58.8	0	100
5	23F	6B	37.5	62.5	46.2	53.8
6	23F	None	100	0	100	0

Optimising serotyping sensitivity from metagenomes with graph-based alignments using GNASTy

We have shown that pneumococcal CBL sequences are a more suitable target for metagenome-based serotype surveillance than whole genomes. However, high sequence divergence between CBL may limit NAS application in the discovery of previously unobserved serotypes. Although SNPs and short variants can usually be aligned to a divergent reference, larger structural variation, present between CBL of different pneumococcal serotypes, can hinder read alignment when variants are not captured in a reference database (Garrison *et al.* 2018). Such variation can be captured using a pangenome graph, which is a compact representation of multiple linear DNA sequences. Pangenome graphs are constructed by merging similar sequences into nodes, with variation between genomes represented by edges. Pangenome graphs provide a means of recapitulating unobserved structural variation, enabling greater flexibility in alignment to capture novel recombinants (Dilthey *et al.* 2015) and alignment across assembly gaps (Horsfield *et al.* 2023). We therefore explored the application of graph alignment in NAS to enrich for novel *S. pneumoniae* CBL.

We developed and implemented a read-to-graph alignment method to replace the linear alignment method currently used in NAS methods (Figure 6). Our method employs 'pseudoalignment', whereby short overlapping nucleotide sequences, known as '*k*-mers', are matched between a read and a de Bruijn graph (DBG), a type of pangenome graph built from matching shared *k*-mers between reference sequences (Iqbal *et al.* 2012; Bray *et al.* 2016). Pseudoalignment is faster than conventional graph alignment, which uses a seed-and-extend approach between a query and reference sequence, and has been used previously in metagenomic read classification (Mäklin *et al.* 2021; Alanko *et al.* 2023). We implemented graph psuedoalignment using Bifrost (Holley and Melsted 2020), which builds coloured compacted DBGs, whereby *k*-mers are 'coloured' by their source genomes, with non-branching paths of *k*-mers 'compacted' into sequences known as 'unitigs', reducing graph size. We named this method 'GNASTy' (Graph-based Nanopore

Adaptive Sampling Typing, pronounced ‘nasty’). A detailed description of the GNASTy method is available in the Supplemental Material (Section C.3, Supplementary Figure 22).

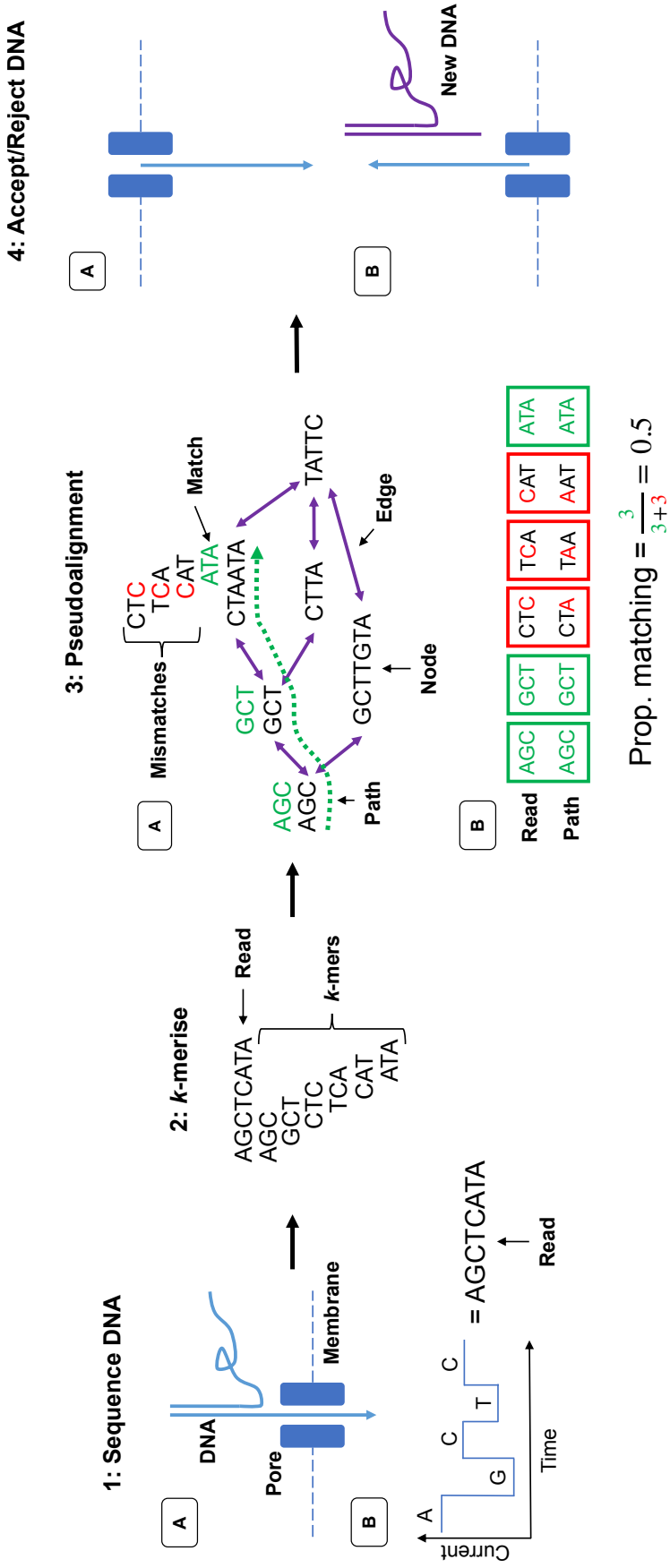

Figure 6: NAS using graph pseudoalignment in GNASTy. (**1a**) The start of a DNA fragment passes through a nanopore, disrupting movement of ions and causing a change in current determined by the base passing through the pore. (**1b**) This current change is used to basecall the read. (**2**) The read is *k*-merised, depending on the *k*-mer size used to build the DBG. (**3a**) The *k*-mers are matched to those in the graph via pseudoalignment, analogous to traversing a hypothetical path (dotted green line). (**3b**) The number of matches (green) and mismatches (red) are used to calculate the proportional number of *k*-mer matches between the read and the hypothetical path in the graph. (**4a**) If the read surpasses the pre-defined identity threshold, the remainder of the DNA is sequenced. (**4b**) If not, the voltage is reversed across the membrane, pushing the read in the revere direction and freeing the pore to sequence a new DNA fragment.

To benchmark the accuracy of GNASTy against the current linear alignment used during NAS, we generated a simulated dataset of nanopore reads from the Spn23F and *E. coli* DH5-α reference genomes. Reads originating from the 23F CBL were classified as target reads, and all others were classified as nontarget. We compared classification accuracy of graph pseudoalignment, implemented in GNASTy, against minimap2 (Li 2018), the linear aligner used in Readfish (Payne *et al.* 2021). The alignment index for both methods was generated from the 106 CBL sequences used previously. Graph pseudoalignment using GNASTy was carried out with a *k*-mer size of 19 bp based on simulation sequencing run performance (Supplemental Material, Sections C.4 and C.5, Supplementary Figures 23-34), with a percentage identity, defined by minimum read-graph identity or 'S', of 75% between a read and a path within the DBG for a read to be classified as a target. Additionally, minimum read length was set to 50 bp, with unaligned reads below this length being rejected.

Table 2: Alignment accuracy comparison between graph pseudoalignment in GNASTy (referred to as 'graph') and minimap2. Sensitivity is defined as $TP/(TP+FN)$, specificity is defined as $TN/(TN+FP)$.

Tool	**No. TP**	**No. FN**	**No. TN**	**No. FP**	**Sensitivity**	**Specificity**
graph	1149	495	494790	3566	0.699	0.993
minimap2	713	931	497514	842	0.434	0.998

We found that alignment sensitivity was higher for graph pseudoalignment than minimap2, whilst specificity was similar between the two methods (Table 2). Therefore, minimap2 had a greater tendency to incorrectly reject target reads than graph pseudoalignment, whilst correct rejection of nontarget reads by graph pseudoalignment was only slightly lower than minimap2. Higher graph pseudoalignment sensitivity is likely due to two factors; graph pseudoalignment does not rely on mapping contiguous blocks of sequence to identify read matches unlike minimap2, allowing more sensitive alignment of reads with structural variants introduced by the read simulator (Yang *et al.* 2017; Břinda *et al.* 2018). Furthermore, graph pseudoalignment enables alteration of alignment identity parameters, which is not possible in the implementation of minimap2 in ReadFish, where these are locked to default values. Both of these factors will contribute to the increased sensitivity of graph pseudoalignment over minimap2.

When comparing computation speed between the two methods, minimap2 outperformed Bifrost/GNASTy during index generation and read alignment. Minimap2 was 30-fold faster at index generation than Bifrost and used 4.5-fold less memory, although Bifrost generated an index 2-fold smaller than minimap2 (Supplementary Table 1). This is an upfront cost not relevant during sequencing. Per-read alignment times for graph pseudoalignment were notably higher than those for minimap2 (Supplementary Figure 5). For graph pseudoalignment, all reads were individually aligned in less than $1/8^{th}$ of a millisecond, equivalent to sequencing 0.056 bases assuming a rate of 450 bases sequenced per second (Payne *et al.* 2021). If 512 reads were aligned in a single chunk (the maximum number of reads that could be generated at once on a MinION), this would be equivalent to an additional 29 bases being sequenced per pore before a decision is made to accept or reject each read. Therefore, we tested whether GNASTy's greater sensitivity for detecting target reads, at the cost of slower rejection of nontarget reads, would increase the enrichment of CBL sequences.

Graph-based alignment facilitates the discovery of novel CBL

We investigated whether GNASTy's graph representation of CBL variation would enable it to discover and enrich novel CBL variants more accurately than conventional NAS. To evaluate this, we tested whether graph pseudoalignment in GNASTy could outperform linear alignment when the target sequence was not present in a reference database. We used the 106 CBL sequences used previously as a reference database, removing the 23F CBL, along with all closely related CBL (cluster two from Mavroidi *et al.* (2007)). We then sequenced the same samples used previously (Figure 1a and b), this time using V14 rather than V12 Nanopore chemistry, and calculated enrichment of the 23F CBL. These experiments used V14 sequencing chemistry, which generates reads faster than now-discontinued V12 chemistry, but has similar read quality (Sinclair Dokos 2022). We conducted sequencing runs using three different alignment methods. Minimap2 was compared with graph psuedoalignment in GNASTy with $k = 19$ and minimum read length of 50 bp as before. We tested two percentage identity thresholds for graph pseudoalignment, $S = 75\%$ and $S = 90\%$, to understand the effect of increasing graph pseudoalignment stringency on enrichment.

Results showed that enrichment could be achieved by all NAS methods and parameter values, although graph pseudoalignment ($S = 75\%$) performed best, with equivalent or higher enrichment than minimap2 across all samples (Figure 7). The highest observation of 23F CBL enrichment exceeded $10,000$ for graph pseudoalignment ($S = 75\%$) in the *E. coli* mixture (identified by a red asterisk), which was due to no 23F CBL control reads being generated for this sample, whilst NAS enabled detection of target DNA. As observed in Martin *et al.* (2022), targets at low concentration produce more variable enrichment values due to the low numbers of reads detected by both NAS and control channels. Therefore, the slower read alignment speed of graph pseudoalignment compared to minimap2 did not have a large enough effect to negatively impact enrichment. In addition to enrichment, absolute yield of 23F bases was significantly increased using graph pseudoalignment ($S = 75\%$) relative to control channels (Supplementary Figure 6). Graph pseudoalignment ($S = 75\%$) achieved a mean yield increase of 2.75-fold (p=9.8×10^{-4}), which was greater than for minimap2, which achieved a mean yield increase of 2.0-fold (p=2.4×10^{-3}). Furthermore, graph pseudoalignment ($S = 75\%$) performed similarly to minimap2 when the 23F CBL was included in the reference database, meaning that graph pseudoalignment in GNASTy can be used as a direct replacement for minimap2 for NAS (Supplemental Material, Section C.6, Supplementary Figures 35-38). Graph pseudoalignment ($S = 90\%$) performed worst of the three methods, resulting in lower enrichment and reduced absolute yield which was not significantly different from control channels. Enrichment fell below 1 at the lowest target concentrations in *E. coli* and *S. mitis* mixtures, indicating target depletion. This result highlights that $S = 90\%$ is too stringent for graph pseudoalignment, resulting in incorrect rejection of target reads.

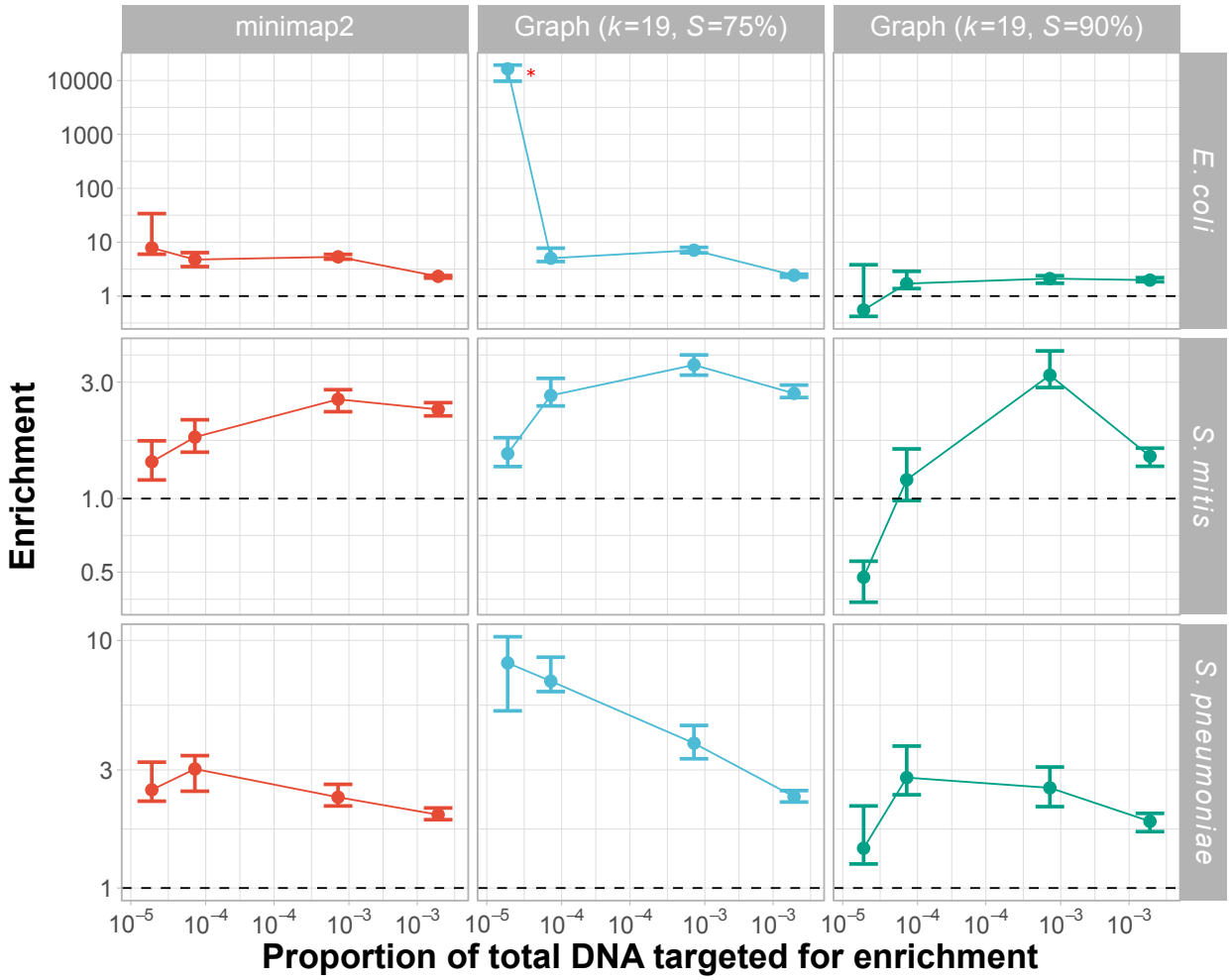

Figure 7: Enrichment comparison of 23F CBL at different concentrations of target between minimap2 and graph pseudoalignment in GNASTy when aligning to a partial CBL reference database. Bar ranges are inter-quartile range of enrichment from 100 bootstrap samples of reads. Data points connected by lines are observed enrichment values for each library, with solid lines connecting the same genome diluted at different concentrations. Rows describe the nontarget species in the mixture, columns describe the alignment method used. Each column represents data from a single flow cell. To plot on a log scale, all enrichment values had 0.01 added to them. Horizontal dashed line describes enrichment = 1 i.e. no enrichment has occurred. Red asterisk marks high enrichment observed using graph pseudoalignment (k= 19, $S = 75\%$) in the *E. coli* mixture at 4×10^{-5} target proportion.

The increased enrichment we observed for graph pseudoalignment in GNASTy over minimap2 may be due to biased over-sequencing of a specific position in the 23F CBL, rather than even coverage of the entire 23F CBL. To enable accurate assembly of the full 23F CBL, coverage should ideally be increased evenly across the target sequence, rather than over-represented in specific regions. Comparison of normalised coverage across the full 23F CBL for minimap2 and graph pseudoalignment showed similar read coverage variability across the 23F CBL all three methods (Supplementary Figure 7). For example, at the highest target proportion (4×10^{-3}), there were coverage spikes at both ends of the CBL locus for all three methods. Therefore, enrichment achieved by linear alignment can be explained by mapping of the start of read to shared regions at the ends of CBL (Supplemental Material, Section C.5, Supplementary Figure 32). Coverage for both minimap2 and graph pseudoalignment fell in the center of the CBL, which is particularly not-

able at 4×10^{-3} target dilutions. At the lowest target dilutions (4×10^{-5}), a spike in coverage can be observed at the 18 kb position in the 23F CBL for mixtures containing *S. mitis* and *S. pneumoniae*. As the nontarget isolates *S. mitis* SK142 and *S. pneumoniae* R6 both contain *aliA* homologues (Sørensen *et al.* 2016; Hoskins *et al.* 2001), which is also found at the end boundary of all *S. pneumoniae* CBL (Bentley *et al.* 2006), this peak can be attributed to non-specific enrichment of a gene common to streptococci. However, coverage was equivalent or higher for graph pseudoalignment ($S = 75\%$) over minimap2 across all target concentrations and nontarget species. In summary, although NAS can enrich for novel serotypes using linear alignment, using GNASTy increases NAS sensitivity.

Next, we compared the ability for minimap2 and GNASTy to correctly identify the 23F serotype in the mixtures using PneumoKITy (Supplementary Figure 8). We compared the proportion of the 23F CBL reference sequence covered by the reads, which is used by PneumoKITy as a proxy for serotype prediction confidence (Sheppard *et al.* 2022). Minimap2 and graph pseudoalignment ($S = 75\%$) performed similarly, with reads from NAS channels providing more support for the 23F CBL call than for controls in all cases. Even at low target concentrations ($\leq 8 \times 10^{-5}$ target proportion), these alignment methods were still able to identify 23F as the most likely serotype, with the exception of the mixture with *S. pneumoniae* R6, where serotype 2 was predicted to be the most likely serotype. *S. pneumoniae* R6 is derived from a serotype 2 strain via deletion of its respective CBL (Iannelli *et al.* 1999); however, presence of CBL flanking sequences in *S. pneumoniae* R6, as described above, likely lead to false detection of serotype 2 CBL.

We then compared assemblies of the 23F CBL across the three alignment methods. We chose samples containing 0.1 Spn23F dilution with *S. mitis* to mimic carriage of a single isolate (Supplementary Figure 9). For all alignment methods, read coverage was higher for NAS channels than for control channels, although graph pseudoalignment ($S = 90\%$) had the lowest absolute coverage for both channel types. Despite variation in coverage, all assemblies covered a majority of the CBL with minimal errors of any kind. Assembly completeness was similar between control and NAS channels, except at the right end of the CBL, where minimap2 and graph pseudoalignment ($S = 90\%$) were unable to generate an aligning contig. This effect was also observed when using a full CBL database for enrichment (Supplemental Material, Section C.6), and may be due to uneven local read coverage affecting assembly contiguity, as Metaflye expects uniform coverage for individual strain genomes (Kolmogorov *et al.* 2020). Additionally, two central regions ($\sim$ 7.5 kb and $\sim$ 12 kb), and a small region in the 18 kb end of the CBL, were missing in the control assembly for graph pseudoalignment ($S = 75\%$). However, these were correctly identified when reads were enriched with graph pseudoalignment. When graph pseudoalignment was run with the suboptimally high alignment specificity parameter ($S = 90\%$), the NAS assembly was missing a single region ($\sim$ 7.5 kb) present in the control assembly. Therefore, whilst assemblies were largely similar between NAS and control channels, these small differences indicate higher graph pseudoalignment stringency slightly lowered assembly quality compared to the control, whilst greater sensitivity for CBL reads improved assembly quality.

Graph-based alignment enriches CBL in complex samples mimicking the nasopharynx microbiome

Previous experiments demonstrated graph pseudoalignment in GNASTy was capable of enriching for CBL from simple mixtures. Therefore, we tested whether the method was also effective with more realistic microbial compositions that would be observed in the nasopharynx or oral cavity. We used samples containing a mixed culture generated from nasopharyngeal swabs, spiking in Spn23F as before. As there was no ground truth for these samples, it was unknown whether *S. pneumoniae* strains were already present prior to spiking. Spn23F DNA was added to give a final proportion of 0.1 of total DNA in each sample, reflecting typically observed *S. pneumoniae* prevalences in the nasopharynx (Salter *et al.* 2017), resulting in a final 23F CBL DNA proportion of 8×10^{-4}. Libraries were run without size selection, as we observed a detrimental effect on extracted DNA yield with mixed culture samples which did not affect single isolate samples (Supplementary Figure 10). NAS was conducted using graph pseudoalignment ($k = 19$, $S = 75\%$) in GNASTy using a database containing all 106 CBL sequences, including the 23F CBL. As a control, a sample containing Spn23F mixed with *S. pneumoniae* R6 without size selection at 0.1 and 0.5 proportions was also run, and compared with equivalent samples with size selection.

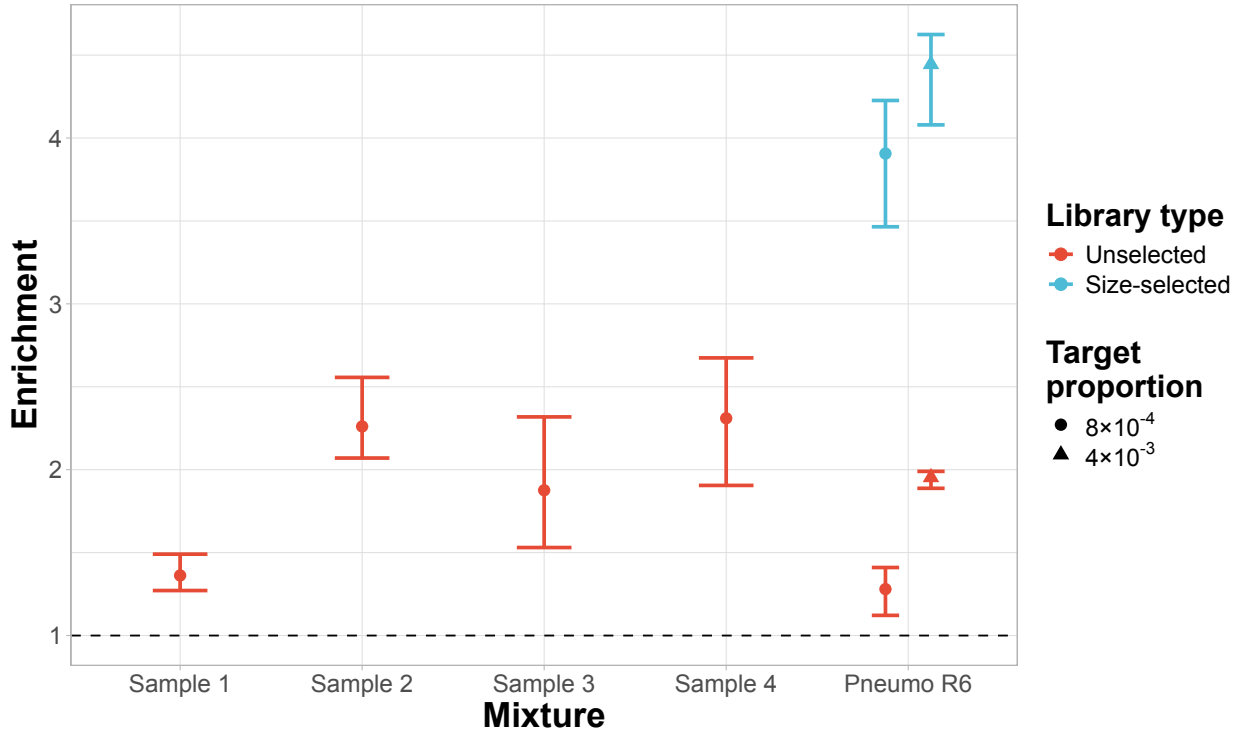

Figure 8: Enrichment of 23F CBL across samples containing mixed cultures from nasopharyngeal swabs. All nasopharyngeal samples (denoted 'Sample X') were run without size selection, with control samples containing Spn23F mixed with *S. pneumoniae* R6 (denoted 'Pneumo R6') without size selection at 0.1 and 0.5 proportions (8×10^{-4} and 4×10^{-3} 23F CBL DNA proportions respectively) run alongside. Equivalent control samples from a run with size selection are plotted for comparison. Bar ranges are inter-quartile range of enrichment from 100 bootstrap samples of reads. Data points are observed enrichment values for each library. Dashed line describes enrichment = 1 i.e. no enrichment has occurred.

Enrichment of the 23F CBL was achieved for all mixed culture samples (Figure 8), with all samples performing equivalently to or better than the unselected *S. pneumoniae* R6 sample at the equivalent concentration (8×10^{-4}). Size selection had a notable positive impact on enrichment in the *S. pneumoniae* R6 mixtures, increasing enrichment from 1.3-1.9 fold to 3.9-4.4 fold for 8×10^{-4} and 4×10^{-3} target proportions, respectively. This was consistent with the method's performance with the simpler DNA mixtures (Supplemental Material C.2). Therefore the lowered performance of graph pseudoalignment in GNASTy with these complex samples relative to the simpler mixtures was a consequence of the lack of size selection during DNA sample preparation. This factor also explains the similar target yield between NAS and control channels (Supplementary Figure 11). Therefore, we advise size selection be used where possible to boost NAS efficiency, although it may not be suitable in all cases due to high DNA yield loss.

The 23F CBL was identified as the most likely CBL in samples 1 and 2, as well as the *S. pneumoniae* R6 mixtures by PneumoKITy, although samples 1, 3 and 4 had evidence of co-carriage (Supplementary Figure 12), with graph pseudoalignment able to enrich for multiple CBL identified as present by PneumoKITY (Supplementary Figure 13). In samples 3 and 4, the 23F CBL was not identified as the most likely serotype, although the proportion of the reference 23F CBL sequence matched was above PneumoKITy's confidence cutoff (70%) for reads originating from NAS channels, meaning that these samples were identified as containing a mixture of serotypes. Notably, the prediction for the 23F CBL did not meet the confidence cutoff for reads from control channels in sample 4, meaning that graph pseudoalignment in GNASTy enabled the detection of a low level secondary serotype that would have otherwise been missed. Assembly quality was similar between NAS and control channels, although NAS channel read coverage was equivalent or higher for most samples (Supplementary Figure 14). Nevertheless, the read coverage was sufficiently high (>20) to enable assembly in principle, which likely could be achieved by algorithms capable of correcting for the uneven read coverage across the 23F CBL in these datasets. Overall, we have shown that NAS, employing graph pseudoalignment with our tool GNASTy, can be used to enrich for *S. pneumoniae* CBL in mock communities resembling real nasopharyngeal samples.

Discussion

The complex population dynamics of *S. pneumoniae* reflect the antigenic and genetic diversity of the species, and the serotype replacement that has been driven by the widespread use of serotype-specific vaccines (Lo *et al.* 2019; Ladhani *et al.* 2018; Tonkin-Hill *et al.* 2022). Current methods for *S. pneumoniae* serotype surveillance are limited either by their requirement for application to individual colonies (e.g. WGS or antiserum agglutination), reducing their sensitivity for detecting co-carriage, or by being restricted to only the serotypes known at the time the assay was designed (e.g. PCR genotyping, microarrays). Deep sequencing has revealed the need for sensitive sequencing-based assays that can detect novel serotypes (van Tonder *et al.* 2019; Tonkin-Hill *et al.* 2022). However, such datasets are resource-intensive to produce, and necessitate substantial pre-existing infrastructure for library generation, sequencing, and data processing. Therefore, simpler, quicker and cheaper surveillance methods are required to provide a comprehensive view of pneumococcal serotype diversity and prevalence to inform public health strategies.

In this work, we explored the application of NAS to pneumococcal serotype surveillance, which has the potential to fulfill all above criteria. We found that when targeting a whole *S. pneumoniae* genome in the presence of closely-related species, enrichment was reduced, as alignment lacks the specificity to distinguish between target and nontarget reads. However, we showed that targeting the *S. pneumoniae* CBL increases both enrichment and yield of NAS due to the strong association of the sequence with *S. pneumoniae*. We showed this enables the detection, enrichment and serotyping of multiple CBL simultaneously, and can detect a serotype that only comprises 1% of the sample.

Direct detection of pneumococcal CBL using NAS promises to be a simple, scalable surveillance method. NAS does not rely on culture, reducing the time required to generate a result compared to WGS or agglutination assays. Library construction takes a few hours, and NAS required one day of sequencing on a portable MinION device, which can be shortened if sufficient read coverage is reached during sequencing. Although NAS is slower than PCR, which takes a few hours, this time is comparable to microarrays (Jauneikaite *et al.* 2015). Unlike NAS and microarrays, however, PCR of multiple serotypes cannot be conducted in a single reaction (Pai *et al.* 2006), increasing workflow complexity despite its shorter runtime. Additionally, NAS provides increased resolution over these genotypic methods, enabling distinction of serotypes separated by few single or multi-nucleotide polymorphisms (Mauffrey *et al.* 2017). NAS is also fully portable and simple enough to be run with limited equipment, requiring only bench-top apparatus such as a centrifuge, thermocycler (Quick *et al.* 2016) and laptop with a suitable GPU, as used here. NAS is more expensive than PCR and microarrays due to the cost of sequencing reagents (Jauneikaite *et al.* 2015). However, NAS has lower entry costs compared to other sequencing technologies such as Illumina or PacBio, with higher target yield enabling sample multiplexing to reduce per-sample costs. We observed $>$ 2-fold increases in target yield compared to standard sequencing using NAS, enabling twice the number of samples to be run on the same flow cell to achieve the same target coverage, therefore halving per-sample costs. Lowered costs, coupled with the portability of Nanopore sequencing, make NAS attractive for applications in low resource settings where pneumoccocal disease burdens are highest (Troeger *et al.* 2018).

Despite the potential for NAS to be used in serotype surveillance, the extensive structural variation distinguishing different CBL caused us to hypothesise that the standard linear alignment employed by NAS would have limited sensitivity when applied to novel or variant loci. To address this issue, we developed a pangenome graph-based alignment method for NAS, GNASTy. We showed that GNASTy enables greater enrichment of novel CBL over linear alignment, and is therefore capable of discovering rare or previously-unknown serotypes. Therefore, GNASTy combines the advantages of NAS described above with the added benefit of increased sensitivity to enrich for novel serotypes. Unlike PCR and microarrays, GNASTy is capable of identifying novel serotypes and easily adding new targets, with any updates to the serotyping panel achieved through simply extending the reference database, without alterations to the laboratory protocol. Therefore, GNASTy is well suited for surveillance of diverse pathogen biomarkers, such as CBL, where novel variants are discovered frequently (Ganaie *et al.* 2020; Ganaie *et al.* 2023), necessitating repeated panel updates which would otherwise be time consuming and expensive using PCR or microarrays. Overall, GNASTy

provides a balance of accuracy, simplicity and cost-effectiveness, making it well-suited for routine pneumococcal surveillance in both high and low resource settings.

A key improvement of targeted sequencing over shotgun metagenomic sequencing is the improved limit of detection, meaning more rare sequences can be identified. We showed that NAS can increase the proportion of target DNA more than 10-fold over that of the control channels, based on the normalised measure; enrichment by composition, when applied to CBL sequences at concentrations $< 0.01\%$, in line with previous evaluations of NAS efficiency (Martin *et al.* 2022; Weilguny *et al.* 2023). We also showed that NAS significantly increased absolute yield of target reads, which improved assembly coverage and accuracy, and increased sensitivity of DNA-based serotyping in samples mimicking co-carriage. Finally, we showed that GNASTy can enrich for CBL DNA in samples mimicking the complexity of the nasopharyngeal microbiome, and improves serotyping accuracy over normal sequencing. We note that these conclusions are based on single sequencing runs, which is common practice when analysing NAS performance due to the cost of Nanopore sequencing (Payne *et al.* 2021; Martin *et al.* 2022). Although we used the normalised measure of enrichment to negate effects of variability between experiments, and bootstrapping to account for noise in data generation, this lack of replicate datasets should motivate future additional validation of this method with different sample types. Furthermore, in the mock communities used throughout this work, our measures of *S. pneumoniae* abundance were relative proportions based on measures from observed nasopharyngeal microbiomes (Salter *et al.* 2017). We did not convert target concentrations into absolute concentration values (e.g. in ng/µL of target DNA), as sequencing sensitivity will be dependent on the number of bases generated per sample, which itself is contingent on multiplexing and variability in DNA loading onto the flowcell.

Multiplexing is key to enabling the batch processing necessary for NAS-based methods to be viable for use in routine surveillance. Based on our experience, we recommend sequencing between 12-24 samples on a single flowcell to provide sufficient coverage to detect *S. pneumoniae* DNA, whilst reducing the cost per sample through multiplexing. Such relatively small batches are practical in routine local surveillance applications where small clusters of samples are available, contrasting with the hundreds of samples that need be multiplexed for higher-throughput sequencing methods to be maximally cost effective.

The current limitations of NAS and GNASTy primarily represent the challenges of optimising DNA sample preparation. The mock communities tested did not contain human reads; however, oral and nasopharyngeal samples often contain substantial host DNA, which will ultimately impact target yield. Therefore, GNASTy will require further optimisation to include host DNA depletion, for which suitable laboratory methods are available (Nelson *et al.* 2019; Charalampous *et al.* 2019). One potential solution is to use NAS to deplete human DNA as off-target reads, which is possible if the human genome were supplied as a database of unwanted sequence. Such depletion sequencing is more effective than targeted sequencing when sampling the full bacterial diversity of a sample (Marquet *et al.* 2022). Hence GNASTy may have additional utility when applied to culture-free nasopharyngeal samples using depletion sequencing. Additionally, when targeting pneumococcal CBL sequences, GNASTy will not be able to distinguish similar or identical sequences found in non-pneumococcal species which co-inhabit the nasopharynx, such as *Strep-*

tococcus mitis or *Streptococcus oralis* (Gertz *et al.* 2021). This issue can be addressed by identifying species-specific flanking regions present in reads that start within the CBL but end outside of it (D'Aeth *et al.* 2021). Here, GNASTy generated up to 41 kb of flanking sequence for reads aligning to target CBL, which can be used to assign a detected CBL to a given species.

NAS has the potential to enable accurate, direct and relatively inexpensive *S. pneumoniae* surveillance. However, this work highlights the current limitations of enriching for low-abundance species with NAS in mixtures containing closely-related taxa, and the suboptimal sensitivity for identifying loci that are not present in the target database. We have developed and tested NAS for the detection and serotyping of *S. pneumoniae* in complex samples, providing methodological recommendations and a novel pangenome graph-based method, GNASTy, for use by public health researchers, which we hope will improve access to accurate *S. pneumoniae* surveillance in low resource settings. GNASTy promises to be a powerful method both for routine epidemiology, and novel serotype discovery.

Methods

Isolate and sample acquisition

All isolate bacterial strains used in this work included: *E. coli* DH5-α, *Moraxella catarrhalis* 0193-3, *Haemophilus influenzae* 0456-2, *S. mitis* SK142, *Streptococcus oralis* SK23, *S. pneumoniae* ATCC 706669 (referred to as 'Spn23F', GPSC16, serotype 23F), *S. pneumoniae* R6 (GPSC622, unencapsulated), *S. pneumoniae* 110.58 (GPSC81, unencapsulated), *S. pneumoniae* MalM6 (GPSC16, serotype 19F), *S. pneumoniae* 8140 (GPSC16, serotype 19A), *S. pneumoniae* Tw01-0057 (GPSC1, serotype 19F), *S. pneumoniae* K13-0810 (GPSC23, serotype 6B), and *S. pneumoniae* 99-4038 (GPSC3, serotype 3).

Nasopharyngeal swab samples were chosen from a collection originating from a study of mother-infant pairs in the Maela camp for refugees in Thailand (Turner *et al.* 2012; Turner *et al.* 2013). This research complied with all relevant ethical regulations, and was approved by the Ethics Committee of The Faculty of Tropical Medicine, Mahidol University, Thailand (MUTM-2009-306), and by the Oxford Tropical Research Ethics Committee, Oxford University, UK (OXTREC-031-06). All women gave written informed consent to participate in the study. Individuals did not receive monetary compensation for their participation.

Bacterial culture and DNA extraction

For culture, glycerol stocks containing bacterial isolates and nasopharyngeal swab (referred to as 'mixed culture') samples were inoculated in 10 mL of Todd-Hewitt broth (Oxoid, UK) and 2% yeast extract (Sigma-Aldrich, UK) and cultured overnight overnight at 35°C in 5% CO_2 atmosphere. For culture of *M. catarrhalis* and *H. influenzae*, 3 mM hemin (X factor) and 22.5 mM nicotinamide-adenine-dinucleotide (NAD, V factor) were also added to respective inocula. Liquid cultures of *M. catarrhalis*, *H. influenzae* and *E. coli* and mixed cultures were incubated with shaking at

150 rpm. Following incubation, the inocula were centrifuged at 16000 *g* for 10 min, with supernatant being discarded to obtain cell pellets.

DNA was extracted from cell pellets using the Wizard Genomic DNA Extraction Kit (Promega UK, Catalogue number: A1120). For *S. pneumoniae* isolates and mixed culture samples, cell pellets were re-suspended in 480 µL 50 mM EDTA, before addition of 120 µL freshly prepared lysozyme (30 mg/mL). The solution was incubated at 37°C for 60 min, before centrifugation at 16,000 *g* for 2 min, with the supernatant being discarded. For all isolates, 600 µL nuclei lysis solution was added to pellets and incubated at 80°C for 5 min. Three µL RNase solution was then added and incubated at 37 °C for 15 min, before cooling to room temperature. 50 µL of 20 mg/mL recombinant Proteinase K solution (Life Technologies, Catalogue number: AM2548) was then added, with the sample being incubated at 55°C for one hour. Two hundred µL protein precipitation solution was then added and incubated on ice for 5 min, before solutions were centrifuged at 16000 *g*, and the supernatant transferred to a clean tube. Six hundred µL of room temperature 100% isopropanol was then added to the supernatant and centrifuged at 16000 *g*, with the supernatant being discarded. Six hundred µL of room temperature 70% ethanol was then added to the pellet and mixed to resuspend the pellet. The solutions were centrifuged at 16000 *g*, with the supernatant being discarded, and the pellets were allowed to air-dry for 15 min. DNA pellets were then resuspended in 150 µL DNA rehydration solution.

Extracted DNA was size selected using the SRE XS kit (PacBio US, Catalogue number: SKU 102-208-200) following manufacturer's instructions to remove fragments < 10 kb in length.

DNA Quality control

Extracted DNA was quantified using a dsDNA broad-range assay kit (Catalogue number: Q32850) on the Qubit 3 fluorimeter (ThermoFisher Scientific UK) following manufacturer's instructions. DNA was also sized using a Genomic DNA ScreenTape Assay (Catalogue numbers: 5067-5366 [Reagents], 5067-5365 [Screentape]) on the TapeStation 2200 system (Agilent UK) following manufacturer's instructions. DNA samples with modal peaks > 45 kb were carried forward for library construction and sequencing.

Library construction

Library construction was conducted using the native barcoding kits (ONT UK, Catalogue numbers: SQK-NBD112.24 [V12 chemistry], SQK-NBD114.24 [V14 chemistry]) following manufacturer's instructions. Briefly, 400 ng DNA was aliquoted per barcoded sample for end and single-strand nick repair using NEBNext Ultra II End repair/dA-tailing Module and NEBNext FFPE Repair Mix (New England Biolabs UK, Catalogue numbers: M6630S, E7546S), with samples then being cleaned using AMPure XP Beads (Beckman Coulter UK) and 70% or 80% ethanol for V12 and V14 chemistry respectively. Barcode ligation followed, using the barcodes provided and the NEB Blunt/TA Ligase Master Mix (New England Biolabs UK, Catalogue number: M0367L), with samples then being pooled together and cleaned as before. Finally, adapter ligation was conducted using the NEBNext Quick Ligation Module (New England Biolabs UK, Catalogue number: E6056S), with the library cleaned using AMPure XP Beads and the long-fragment

buffer provided with the ONT library construction kit. Libraries were loaded onto MIN112 or MIN114 flowcells for V12 and V14 chemistries respectively.

Sequencing and adaptive sampling

All analysis scripts and CBL reference sequences used in this work are available on Zenodo (Horsfield 2024b). GenBank reference sequence accession numbers for whole genome assemblies include: *E. coli* DH5-α (NZ_JRYM01000009.1), *M. catarrhalis* (NZ_CP018059.1), *H. influenzae* (NZ_CP007470.1), *S. mitis* SK142 (NZ_JYGP01000001.1), *S. oralis* SK23 (NZ_LR134336.1), *S. pneumoniae* Spn23F (FM211187.1), *S. pneumoniae* R6 (NC_003098.1) and *S. pneumoniae* 110.58 (CP007593.1).

Sequencing was conducted using a MinION Mk1B instrument and a Dell Mobile Precision 7560 with an Intel Xeon processor and 128 GB RAM, and a NVIDIA RTX A5000 GPU with 16 GB GPU RAM running MinKNOW v22.12.7 (ONT UK) and MinKNOW core v5.4.3 (ONT UK). Local GPU base-calling was conducted using Guppy v6.4.6 (ONT UK) with the fast base-calling model and reads were rejected immediately if they did not align to the reference genome by setting 'maxchunks' to 0 in the Readfish '.toml' file. For each new library, a control sequencing run was conducted for 1 hour with no adaptive sampling with bulk capture, providing a 'recording' for simulation playback.

Adaptive sampling was carried out using Readfish v0.0.10dev2 (Payne *et al.* 2021). Graph pseudoalignment was carried out using a custom fork from the Readfish GitHub repository (Horsfield 2024a). Readfish was installed using the 'readfish.yml' file present in the GitHub repository by running the command 'conda create -f readfish.yml'. During sequencing, Readfish was run using the command 'sudo runuser minknow -c '/path/to/readfish targets --device [device] --experiment-name [name] --channels 1-256 --toml /path/to/toml --logfile [logfile] --port 9502 --graph [True/False] --align_threshold [threshold] --len_cutoff [cutoff]".

Adaptive sampling was used on channels 1-256 of the flowcell, with the remaining 256 channels run as controls without adaptive sampling. Linear alignment for adaptive sampling was carried out using Mappy v2.24 (https://pypi.org/project/mappy/). Sequencing was carried out for 24 hours for each experiment, based on the sequencing time used in Payne *et al.* (2021), after which the run was terminated. No flowcell flushing or library reloading was conducted. Each sequencing experiment was run once. Metadata for all sequencing runs and samples, including the number of bases generated and aligned, the number of reads generated and aligned, and calculated enrichment, are available in Supplemental Data S1. This file also links each sequencing run archived on the European Nucleotide Archive (Horsfield *et al.* 2024) to individual barcoded samples.

Enrichment analysis

Enrichment analysis was based on analysis performed by Martin *et al.* (2022). Enrichment by composition of the target x, denoted by E_x, was calculated as described in Equation 1. Each flowcell was bioinformatically split into two halves; one half contained channels (a segment of a flowcell containing a nanopore) which were 'adaptive' (using

NAS), the other half contained channels which were 'controls' (not using NAS). E_x was calculated as the fold increase in the proportion of read bases aligning to target sequence x in NAS channels, a, versus control channels, c:

$$E_x = \frac{\left(\frac{N_{x,a}}{N_{\text{total},a}}\right)}{\left(\frac{N_{x,c}}{N_{\text{total},c}}\right)} \quad (1)$$

where N_x is the number of bases aligning to target sequence x, and N_{total} is the total bases sequenced in either adaptive (a) or control (c) channels. Using enrichment by composition enables results to be compared across sequencing runs, which may vary in the amount of data generated. If no aligning control reads were generated for a given library, $N_{x,c}$ was set to 1 to avoid division by 0. A merged table of values used to calculated enrichment is present in Supplementary Table S1 (sheet 'enrichment_calculation'), where enrichment is calculated as ('bases_mapped_adaptive' / 'bases_total_adaptive') / ('bases_mapped_control' / 'bases_total_control').

To calculate enrichment post-sequencing, all reads, including those passing and failing the Phred-score filter (Q-score ≥ 8), were aligned to a reference sequences using Mappy v2.24 using the custom script 'analyse_RU.py'. Reads were aligned to specific reference sequences based on known isolates present within each sample ('-t <target>'). All reads were used to avoid any potential biases introduced by read filtering, such as flow cell spatial effects, in the calculation of enrichment. Reads were split by channel ('-c 1-256') to identify which reads were sequenced under NAS (channels 1-256) or control (channels 257-512) conditions. Reads aligning above a specified minimum identity threshold (84% identity within the aligned block, '-p 0.84') were assigned as target reads, with the highest-identity alignment for multi-mapping reads being taken as the only alignment. Only regions of reads aligning to a reference sequence were included in enrichment calculations. Quartiles were generated from 100 bootstrapped samples of aligned reads ('-bs 100').

Serotype prediction

Serotype prediction was conducted using a customised version of PneumoKITy which can be run using a single FASTQ file, as opposed to paired FASTQ files as in the original version, available on Zenodo (https://doi.org/10.5281/zenodo.10590659) (Horsfield 2024c). Reads were split using a custom script (split_by_channel.py) to generate files for reads sequenced under adaptive (channels 1-256) and control (channels 257-512) conditions ('--channels 1-256'). PneumoKITy was run in 'mix' mode using a minimum median-multiplicity value of 4 ('-n 4') and a minimum kmer percentage of 85% ('-p 85') for reference CBL sequence matching.

Assembly and quality control

All reads were first re-basecalled using Guppy v6.4.6 with the super-high accuracy model using the following command: 'guppy_basecaller --compress_fastq --input_path [input_path] --save_path [output_path] --config dna_r10.4.1_e8.2_400bps_sup.cfg --device cuda:0 --recursive --barcode_kits SQK-NBD114-24 --enable_trim_barcodes --trim_adapters --trim_primers'. Reads were then assembled using MetaFlye v2.9.2

(Kolmogorov *et al.* 2020) in '--nano-raw' mode. We did not use the high accuracy '--nano-hq' mode, as testing showed this was too stringent and resulted in no assembly being generated for some samples. Assembly quality was then analysed using Inspector v1.2 (Chen *et al.* 2021), with reads mapped to respective assemblies, and assembly contigs mapped to respective reference sequences to identify errors. Errors were identified in contigs $\geq$ 50 bp in length ('--min_contig_length_assemblyerror 50', '--min_contig_length 50'). BED files generated by Inspector, containing contig alignment and error positions on respective reference sequences, were visualised using a custom script ('plotting_scripts /generate_linear_assembly_plot.R'). Read alignment for coverage analysis was conducted using the custom script, 'analyse_coverage.py', using the original reads basecalled using Guppy's fast basecalling model. Alignment and read parsing settings were the same as 'analyse_RU.py' described above. All alignment was carried out using Mappy v2.24. Assembly statistics are available in Supplemental Data S2.

Nanopore sequencing simulation and analysis

Simulations of nanopore sequencing runs were conducted using bulk capture recordings from previous sequencing runs, as described on the Readfish GitHub repository (https://github.com/LooseLab/readfish). Results were analysed using a custom script (analyse_unblocks.py). This script aligns reads to a specified target sequence using Mappy v2.24 and classifies them as either accepted or rejected by the adaptive sampling process. Reads that align to a target sequence and were accepted or rejected are classified as true positives and false negatives respectively. Reads that did not align to a target sequence and were accepted or rejected are classified as false positives and true negatives respectively. For all experiments described here, the reference sequence was the Spn23F Chromosome ('--ref data/cps/sequences/SP_ATCC700669.fasta') and the target sequence was the 23F CBL sequence ('--loci data/cps/split_cps/23F.fa').

For benchmarking of alignment speed, a bespoke simulation model was generated using Nanosim-H v1.1.0.2 (Yang *et al.* 2017; Břinda *et al.* 2018). Model training used FASTQ files from a V14 chemistry nanopore sequencing run containing 50%-50% dilutions of *S. pneumoniae* Spn23F and *E. coli* DH5-α, and their respective reference sequences ('nanosim-h-train -i training_reads.fasta reference/genome.fasta output'). Using this model, 500,000 simulated nanopore reads were generated ('nanosim-h -n 500000 -p output reference_genome.fasta'). Simulated reads were then split into true positive and true negative reads based on whether they originated from the 23F CBL using the custom script split_simulated.py, which parses reads simulated by Nanosim-H based on their original locus. Reads overlapping by at least 50 bp with the 23F CBL (position 303558-322212 bp within the Spn23F Chromosome) were classified as true positives ('--pos 303558-322212 --min-overlap 50'). CBL sequences (updated_cps.fasta, N=106) were indexed using minimap2 v2.26 (Li 2018) and Bifrost v1.2.0 (Holley and Melsted 2020) with $k = 19$. The time taken to align all 500,000 simulated reads for Mappy v2.24 and graph pseudoalignment was measured using a custom script (simulate_readuntil.py) which parses the start of each read, with length defined by Poisson sampling ('--avg-poi 180', based on Payne *et al.* (2021)). This fragment is then aligned using both Mappy and graph pseudoalignment, with alignment timed using the Python 'timeit' module. Graph pseudoalignment was run using minimum read identity 75% (--id 0.75)

and minimum read length 50 bp ('--min-len 50'). Mappy was run with default parameters. Alignment accuracy was measured based on whether a read was accepted or rejected, depending on whether it originated from the 23F CBL or not. Comparisons were carried out on a server cluster with dual processor x86-64 nodes, running CentOS v8.2.

Pseudoalignment simulation

Pseudoalignment simulations proceeds as follows. A specified number of target and nontarget sequences of given lengths are generated by random sampling of DNA bases. Constituent *k*-mers of these sequences are then generated, and reads with specified mutation rates are simulated from target sequences. Read *k*-mers are then matched back to the respective target and nontarget *k*-mer sets, enabling calculation of recall and precision respectively. The code for this process can be found in the 'kmer_simulation.R' script.

Software availability

Code for GNASTy is available on Zenodo (https://zenodo.org/records/13358697) (Horsfield 2024a) and on GitHub (https://github.com/bacpop/readfish/tree/graph_alignment_bifrost) under the GPL-3.0 license. All analysis scripts used in this work are available on Zenodo (https://zenodo.org/records/12636613) (Horsfield 2024b) and on GitHub (https://github.com/bacpop/adaptive_sampling_scripts) under the GPL-3.0 license. This repository also contains 106 *S. pneumoniae* CBL sequences and associated sources (updated_cps.fasta, updated December 19^{th} 2022) used as reference sequences for NAS. The updated version of PneumoKITY used in this manuscript is also available on Zenodo (https://doi.org/10.5281/zenodo.10590659) (Horsfield 2024c) under the GPL-3.0 license. All code is also available in the Supplemental Code file.

Data access

All raw and processed sequencing data generated in this study have been submitted to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under accession number PRJEB72455.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

S.T.H. was funded by the MRC Centre for Global Infectious Disease Analysis (studentship grant ref.: MR/S502388/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth and Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 program supported by the European Union. S.T.H. acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. N.J.C. and J.A.L. were funded by the UK Medical Research Council and Department for International Development (grants MR/R015600/1 and MR/T016434/1). N.J.C. was also supported by a Sir Henry Dale fellowship jointly funded by Wellcome and the Royal Society (grant 104169/Z/14/A). J.A.L. and S.T.H. were also supported by the European Molecular Biology Laboratory. P.T. was funded by the Wellcome Trust (grants 083735 and 220211). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Author contributions

Conceptualization: S.T.H., N.J.C., and J.A.L. Methodology: S.T.H., N.J.C., and J.A.L. Software: S.T.H. Validation: S.T.H., B.F., Y.F. Formal analysis: S.T.H. Investigation: S.T.H. Resources: S.T.H., P.T, N.J.C. and J.A.L. Data curation: S.T.H. Writing - original draft: S.T.H. Writing - review and editing: All authors. Visualization: S.T.H. Supervision: N.J.C. and J.A.L. Project Administration: S.T.H., N.J.C. and J.A.L. Funding acquisition: S.T.H., P.T., N.J.C., and J.A.L.

References

Alanko JN, Vuohtoniemi J, Mä Klin T and Puglisi SJ. 2023. Themisto: a scalable colored k-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. *Bioinformatics*. **39**: i260–i269.

Bek-Thomsen M, Tettelin H, Hance I, Nelson KE and Kilian M. 2008. Population diversity and dynamics of Streptococcus mitis, Streptococcus oralis, and Streptococcus infantis in the upper respiratory tracts of adults, determined by a nonculture strategy. *Infection and Immunity*. **76**: 1889–1896.

Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA *et al.* 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genetics*. **2**: 0262–0269.

Bray NL, Pimentel H, Melsted P and Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*. **34**: 525–527.

Břinda K, Yang C, Chu J, Linthorst J and Franus W 2018. NanoSim-H; a simulator of Oxford Nanopore reads; a fork of NanoSim. URL: https://zenodo.org/record/1341250#.XqlLHahKiUk.

Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, Rae D, Grundy S, Turner DJ, Wain J *et al.* 2019. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature Biotechnology*. **37**: 783–792.

Chen Y, Zhang Y, Wang AY, Gao M and Chong Z. 2021. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology*. **22**: 1–21.

Colijn C, Corander J and Croucher NJ. 2020. Designing ecologically optimized pneumococcal vaccines using population genomics. *Nature Microbiology*. **5**: 473–485.

Cretu Stancu M, Van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, De Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J *et al.* 2017. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications*. **8**: 1–13.

Croucher NJ, Harris SR, Barquist L, Parkhill J and Bentley SD. 2012. A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathogens*. **8**:

Croucher NJ, Løchen A and Bentley SD. 2018. Pneumococcal Vaccines: Host Interactions, Population Dynamics, and Design Principles. *Annual Review of Microbiology*. **72**: 521–549.

Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, Mitchell AM, Quail MA, Andrew PW, Parkhill J *et al.* 2009. Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone Streptococcus pneumoniae Spain23F ST81. *Journal of Bacteriology*. **191**: 1480.

D'Aeth JC, van der Linden MP, McGee L, de Lencastre H, Turner P, Song JH, Lo SW, Gladstone RA, Sá-Leão R, Ko KS *et al.* 2021. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. *eLife*. **10**:

Delahaye C and Nicolas J. 2021. Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*. **16**:

Dilthey A, Cox C, Iqbal Z, Nelson MR and McVean G. 2015. Improved genome inference in the MHC using a population reference graph. *Nature Genetics*. **47**: 682–688.

Epping L, van Tonder AJ, Gladstone RA, Bentley SD, Page AJ and Keane JA. 2018. SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. *Microbial Genomics*. **4**:

Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW, Gladstone RA, Turner P, Keenan JD, Breiman RF *et al.* 2020. A new pneumococcal capsule type, 10D, is the 100th serotype and has a large cps fragment from an oral streptococcus. *mBio*. **11**:

Ganaie FA, Saad JS, Lo SW, McGee L, van Tonder AJ, Hawkins PA, Calix JJ, Bentley SD and Nahm MH. 2023. Novel pneumococcal capsule type 33E results from the inactivation of glycosyltransferase WciE in vaccine type 33F. *The Journal of Biological Chemistry*. **299**:

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF *et al.* 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*. **36**: 875–881.

Gertz RE, Pimenta FC, Chochua S, Larson S, Venero AK, Bigogo G, Milucky J, Carvalho MdG and Beall B. 2021. Nonpneumococcal Strains Recently Recovered from Carriage Specimens and Expressing Capsular Serotypes Highly Related or Identical to Pneumococcal Serotypes 2, 4, 9A, 13, and 23A. *mBio*. **12**:

Habib M, Porter BD and Satzke C. 2014. Capsular serotyping of Streptococcus pneumoniae using the Quellung reaction. *Journal of Visualized Experiments*.

Holley G and Melsted P. 2020. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*. **21**: 249.

Horsfield ST 2024a. Graph-based Nanopore Adaptive Sampling Typing (GNASTy). URL: https://zenodo.org/records/13358697.

Horsfield ST 2024b. Nanopore Adaptive Sampling Analysis Scripts. URL: https://zenodo.org/records/12636613.

Horsfield ST 2024c. PnuemoKITy-Nanopore_v1.0.1. URL: https://zenodo.org/records/10590659.
Horsfield ST, Fok B, Fu Y, Turner P, Lees JA and Croucher NJ 2024. Nanopore Adaptive Sampling for pneumococcal surveillance using serotyping. URL: https://www.ebi.ac.uk/ena/browser/view/PRJEB72455.
Horsfield ST, Tonkin-Hill G, Croucher NJ and Lees JA. 2023. Accurate and fast graph-based pangenome annotation and clustering with ggCaller. *Genome Research*. **33**: gr.277733.123.
Hoskins J, Alborn J, Arnold J, Blaszczak LC, Burgett S, Dehoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W *et al.* 2001. Genome of the Bacterium Streptococcus pneumoniae Strain R6. *Journal of Bacteriology*. **183**: 5709.
Huebner RE, Dagan R, Porath N, Wasas AD, T M and Klugman KP. 2000. Lack of utility of serotyping multiple colonies for detection of simultaneous nasopharyngeal carriage of different pneumococcal serotypes. *The Pediatric Infectious Disease Journal*. **19**: 1017–1020.
Hyams C, Camberlein E, Cohen JM, Bax K and Brown JS. 2010. The Streptococcus pneumoniae Capsule Inhibits Complement Activity and Neutrophil Phagocytosis by Multiple Mechanisms. *Infection and Immunity*. **78**: 704.
Iannelli F, Pearce BJ and Pozzi G. 1999. The Type 2 Capsule Locus of Streptococcus pneumoniae. *Journal of Bacteriology*. **181**: 2652.
Ikuta KS, Swetschinski LR, Aguilar GR, Sharara F, Mestrovic T, Gray AP, Weaver ND, Wool EE, Han C, Hayoon AG *et al.* 2022. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*. **400**: 2221–2248.
Ip CL, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, Leggett RM, Eccles DA, Zalunin V, Urban JM *et al.* 2015. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*. **4**: 1075.
Iqbal Z, Caccamo M, Turner I, Flicek P and McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*. **44**: 226–232.
Jauneikaite E, Tocheva AS, Jefferies JM, Gladstone RA, Faust SN, Christodoulides M, Hibberd ML and Clarke SC. 2015. Current methods for capsular typing of Streptococcus pneumoniae. *Journal of Microbiological Methods*. **113**: 41–49.
Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG and Fry NK. 2016. Whole genome sequencing of Streptococcus pneumoniae: Development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ*.
Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TP *et al.* 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*. **17**: 1103–1110.
Ladhani SN, Collins S, Djennad A, Sheppard CL, Borrow R, Fry NK, Andrews NJ, Miller E and Ramsay ME. 2018. Rapid increase in non-vaccine serotypes causing invasive pneumococcal disease in England and Wales, 2000-17: a prospective national observational cohort study. *The Lancet Infectious Diseases*. **18**: 441–451.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**: 3094–3100.
Lo SW, Gladstone RA, van Tonder AJ, Lees JA, du Plessis M, Benisty R, Givon-Lavi N, Hawkins PA, Cornick JE, Kwambana-Adams B *et al.* 2019. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *The Lancet Infectious Diseases*. **19**: 759–769.
Lo SW, Mellor K, Cohen R, Alonso AR, Belman S, Kumar N, Hawkins PA, Gladstone RA, von Gottberg A, Veeraraghavan B *et al.* 2022. Emergence of a multidrug-resistant and virulent Streptococcus pneumoniae lineage mediates serotype replacement after PCV13: an international whole-genome sequencing study. *The Lancet Microbe*. **3**: e735–e743.
Løchen A, Truscott JE and Croucher NJ. 2022. Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates. *PLoS Computational Biology*. **18**: e1009389.
Mäklin T, Kallonen T, Alanko J, Samuelsen Ø, Hegstad K, Mäkinen V, Corander J, Heinz E and Honkela A. 2021. Bacterial genomic epidemiology with mixed samples. *Microbial Genomics*. **7**: 691.
Marquet M, Zöllkau J, Pastuschek J, Viehweger A, Schleußner E, Makarewicz O, Pletz MW, Ehricht R and Brandt C. 2022. Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. *Scientific Reports*. **12**:
Martin S, Heavens D, Lan Y, Horsfield S, Clark MD and Leggett RM. 2022. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biology*. **23**: 1–27.
Marttinen P, Croucher NJ, Gutmann MU, Corander J and Hanage WP. 2015. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*. **1**:
Mauffrey F, Fournier É, Demczuk W, Martin I, Mulvey M, Martineau C, Lévesque S, Bekal S, Domingo MC, Doualla-Bell F *et al.* 2017. Comparison of sequential multiplex PCR, sequetyping and whole genome sequencing for serotyping of Streptococcus pneumoniae. *PLoS ONE*. **12**:
Mavroidi A, Aanensen DM, Godoy D, Skovsted IC, Kaltoft MS, Reeves PR, Bentley SD and Spratt BG. 2007. Genetic Relatedness of the Streptococcus pneumoniae Capsular Biosynthetic Loci. *Journal of Bacteriology*. **189**: 7841.

Nelson MT, Pope CE, Marsh RL, Wolter DJ, Weiss EJ, Hager KR, Vo AT, Brittnacher MJ, Radey MC, Hayden HS *et al.* 2019. Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles. *Cell Reports*. **26**: 2227.

Pai R, Gertz RE and Beall B. 2006. Sequential Multiplex PCR Approach for Determining Capsular Serotypes of Streptococcus pneumoniae Isolates. *Journal of Clinical Microbiology*. **44**: 124.

Payne A, Holmes N, Clarke T, Munro R, Debebe BJ and Loose M. 2021. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*. **39**: 442.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A *et al.* 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. **530**: 228–232.

Ricketson LJ, Lidder R, Thorington R, Martin I, Vanderkooi OG, Sadarangani M and Kellner JD. 2021. PCR and culture analysis of streptococcus pneumoniae nasopharyngeal carriage in healthy children. *Microorganisms*. **9**:

Salter SJ, Turner C, Watthanaworawit W, de Goffau MC, Wagner J, Parkhill J, Bentley SD, Goldblatt D, Nosten F and Turner P. 2017. A longitudinal study of the infant nasopharyngeal microbiota: The effects of age, illness and antibiotic use in a cohort of South East Asian children. *PLoS Neglected Tropical Diseases*. **11**: e0005975.

Satzke C, Dunne EM, Porter BD, Klugman KP, Mulholland EK, Vidal JE, Sakai F, Strachan JE, Hay Burgess DC, Holtzman D *et al.* 2015. The PneuCarriage Project: A Multi-Centre Comparative Study to Identify the Best Serotyping Methods for Examining Pneumococcal Carriage in Vaccine Evaluation Studies. *PLoS Medicine*. **12**:

Sheppard CL, Manna S, Groves N, Litt DJ, Amin-Chowdhury Z, Bertran M, Ladhani S, Satzke C and Fry NK. 2022. PneumoKITy: A fast, flexible, specific, and sensitive tool for Streptococcus pneumoniae serotype screening and mixed serotype detection from genome sequence data. *Microbial Genomics*. **8**:

Sinclair Dokos R. 2022. Update from Oxford Nanopore Technologies. *London Calling 2022*.

Sørensen UB, Yao K, Yang Y, Tettelin H and Kilian M. 2016. Capsular Polysaccharide Expression in Commensal Streptococcus Species: Genetic and Antigenic Similarities to Streptococcus pneumoniae. *mBio*. **7**:

Su J, Lui WW, Lee Y, Zheng Z, Siu GKH, Ng TTL, Zhang T, Lam TTY, Lao HY, Yam WC *et al.* 2023. Evaluation of Mycobacterium tuberculosis enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for identification and variant calling. *Scientific Reports*. **13**: 5237.

Tonkin-Hill G, Ling C, Chaguza C, Salter SJ, Hinfonthong P, Nikolaou E, Tate N, Pastusiak A, Turner C, Chewapreecha C *et al.* 2022. Pneumococcal within-host diversity during colonization, transmission and treatment. *Nature Microbiology*. **7**: 1791–1804.

Troeger C, Blacker B, Khalil IA, Rao PC, Cao J, Zimsen SR, Albertson SB, Deshpande A, Farag T, Abebe Z *et al.* 2018. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Infectious Diseases*. **18**: 1191.

Turner P, Turner C, Green N, Ashton L, Lwe E, Jankhot A, Day NP, White NJ, Nosten F and Goldblatt D. 2013. Serum antibody responses to pneumococcal colonization in the first 2 years of life: results from an SE Asian longitudinal cohort study. *Clinical Microbiology and Infection*. **19**: E551.

Turner P, Hinds J, Turner C, Jankhot A, Gould K, Bentley SD, Nosten F and Goldblatt D. 2011. Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray. *Journal of Clinical Microbiology*. **49**: 1784–1789.

Turner P, Turner C, Jankhot A, Helen N, Lee SJ, Day NP, White NJ, Nosten F and Goldblatt D. 2012. A Longitudinal Study of Streptococcus pneumoniae Carriage in a Cohort of Infants and Their Mothers on the Thailand-Myanmar Border. *PLoS ONE*. **7**:

Van Tonder AJ, Gladstone RA, Lo SW, Nahm MH, du Plessis M, Cornick J, Kwambana-Adams B, Madhi SA, Hawkins PA, Benisty R *et al.* 2019. Putative novel cps loci in a large global collection of pneumococci. *Microbial Genomics*. **5**: e000274.

Viehweger A, Marquet M, Hölzer M, Dietze N, Pletz MW and Brandt C. 2023. Nanopore based enrichment of antimicrobial resistance genes - a case-based study. *GigaByte*. 1–15.

Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y, Lukšić I, Nair H, McAllister DA, Campbell H *et al.* 2018. Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000-15. *The Lancet Global Health*. **6**: e744–e757.

Wang H, Naghavi M, Allen C, Barber RM, Carter A, Casey DC, Charlson FJ, Chen AZ, Coates MM, Coggeshall M *et al.* 2016. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*. **388**: 1459–1544.

Weilguny L, De Maio N, Munro R, Manser C, Birney E, Loose M and Goldman N. 2023. Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design. *Nature Biotechnology*. **41**: 1018–1025.

Weiser JN, Ferreira DM and Paton JC. 2018. Streptococcus pneumoniae: transmission, colonization and invasion. *Nature Reviews Microbiology*.

Wrenn DC and Drown DM. 2023. Nanopore adaptive sampling enriches for antimicrobial resistance genes in microbial communities. *Gigabyte*. 1–14.

Yang C, Chu J, Warren RL and Birol I. 2017. NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience*. **6**: 1–6.

Ye SH, Siddle KJ, Park DJ and Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*. **178**: 779–794.