

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



LSHTM Research Online

Hackman, JN; (2023) Application of pathogen genomics to infer the transmission direction of respiratory infection. PhD thesis, London School of Hygiene & Tropical Medicine. DOI: <https://doi.org/10.17037/PUBS.04671561>

Downloaded from: <https://researchonline.lshtm.ac.uk/id/eprint/4671561/>

DOI: <https://doi.org/10.17037/PUBS.04671561>

Usage Guidelines:

Please refer to usage guidelines at <https://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license. To note, 3rd party material is not necessarily covered under this license: <http://creativecommons.org/licenses/by-nc-nd/4.0/>

<https://researchonline.lshtm.ac.uk>

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



長崎大学
NAGASAKI UNIVERSITY

APPLICATION OF PATHOGEN GENOMICS TO INFER THE TRANSMISSION DIRECTION OF RESPIRATORY INFECTION

Jada N. Hackman

Thesis submitted in accordance with the requirements for the degree of
Doctor of Philosophy of the University of London

June 2023

Centre for the Mathematical Modelling of Infectious Disease
Department of Infectious Disease Epidemiology
Faculty of Epidemiology and Population Health
LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Department of Paediatric Infectious Diseases
School of Tropical Medicine and Global Health
NAGASAKI UNIVERSITY

Funded by Nagasaki University "Doctoral Program for World-leading Innovative and Smart Education" for Global Health, "Global Health Elite Programme for Building a Healthier World"

DECLARATION OF AUTHORSHIP

I, Jada Hackman, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed,



31 May 2023

ABSTRACT

Understanding transmission direction is important for the epidemiological assessment of infectious diseases to identify sources and risks for infection and thus implement preventative measurements against disease acquisition. Thus, phylogenetic reconstruction methods, which infer transmission direction using genomic data of sampled individuals, are well suited for these investigations. Recent developments in sequencing technologies and bioinformatic tools have streamlined phylogenetic inference in the transmission direction of human pathogens with reduced cost and required resources.

Previous approaches to infer transmission direction have mostly relied on epidemiological data which are time and cost intensive to follow and can lead to unreliable self-reported data from the study participants. In outbreak settings where the transmission involves multiple individuals such as hospital, household, and school settings, determining the source of the infection can be difficult to disentangle in the absence of genomic data.

This PhD focuses on the capacity at which we can infer the transmission direction of *Streptococcus pneumoniae* and SARS-CoV-2 using whole-genome next-generation sequencing data from household settings with “known” transmission direction according to the epidemiological records. In addition to highlighting the potential role of within-host genetic diversity, in the context of *Streptococcus pneumoniae* co-carriage, in transmission events.

In summary, the context of *Streptococcus pneumoniae*, increased sequencing read lengths and intra-host diversity, in the form of single nucleotide polymorphisms, increased our ability to infer the correct direction of transmission. Moreover, the presence of a transmission bottleneck can aid in identifying the source of infection. While for the SARS-CoV-2 study, the transmission direction inferred from the genomic data suggests reclassification of the household index case. Findings from this PhD show promising results that we can infer the linkage and transmission direction of respiratory pathogens, *Streptococcus pneumoniae* and SARS-CoV-2.

ACKNOWLEDGEMENT

Firstly, I would like to thank my supervisory team, Stéphane Hué, Stefan Flasche, Michiko Toizumi, and Lay-Myint Yoshida, for their unconditional support, consistent guidance, and breadth and depth of expertise throughout this joint LSHTM and Nagasaki University PhD. It was not an easy feat to maintain an international collaboration, particularly not amid a pandemic. There were times of uncertainty that resulted in a massive shift in the aims and objectives of the PhD. However, they did their best to maintain continuous conversations and frequent updates to manoeuvre around those difficult times. Likewise, I would like to thank Martin Hibberd, who has been instrumental to my PhD both as a collaborator and mentor. I would also like to thank Jody Phelan and Sam Clifford who, patiently, helped me with bioinformatics and coding more efficiently in R and allowed me to ask the “dumb” questions.

Secondly, I would like to thank the Japanese Ministry of Education, Culture, Sports, Science and Technology WISE Program for funding and making this PhD possible.

Thirdly, I would like to thank my colleagues at both LSHTM and Nagasaki University. I am grateful to have shared office space, knowledge, pints, laughs, and struggles with so many of you. It made the journey less lonely despite lockdown a few months after starting my PhD, especially the Early Career Research group within CMMID for being a welcoming and safe space for newcomers.

Fourthly, on a more personal side, during the third year of my PhD, I was in a life-threatening cycling accident, and I would like to extend additional gratitude and appreciation to my supervisors for their patience, understanding, and support as I got back on my feet and continued recovery. Similarly, I would also like to extend the utmost gratitude to my family (my mom, Judi, and my sisters, Jess, Jamie, Julianne), my partner (Pablo), my friends (near and far), the girls Camden Eagles Football Club, Adidas Runners, my physios, therapist, and of course, my cat (Eli) for helping me get through the most mentally, emotionally, and physically difficult part of my life. Without them, I would not have been able to cross the PhD finish line.

Lastly, I would like to thank my dad, Robert, who is no longer with us, for always being my quietest but biggest cheerleader. From an early age, he instilled in me a profound sense of curiosity, and this was the initial spark that kindled my desire to pursue science which led to the completion of this PhD.

ABBREVIATIONS

BNT	BioNTech, Pfizer vaccine COVID-19
CI	Confidence interval
COG-UK	COVID-19 Genomics UK
COVID	Coronavirus disease
cps	Capsule polysaccharide
DNA	Deoxyribonucleic acid
GAVI	Global Vaccine Alliance
HIV	Human immunodeficiency viruses
IPD	Invasive pneumococcal diseases
ML	Maximum-likelihood
MLST	Multilocus sequence typing
MM	Mixture modelling
MRSA	Methicillin-resistant <i>Staphylococcus aureus</i>
NGS	Next-generation sequencing
NT	Non-typeable
PCR	Polymerase chain reaction
PCV	Pneumococcal conjugate vaccine
PPV	Pneumococcal polysaccharide-based vaccine
RDP	Recombination detection programs
RNA	Ribonucleic acid
RT-PCR	Reverse transcription polymerase chain reaction
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SD	Standard deviation
SNP	Single nucleotide polymorphism
SNV	Intrahost single nucleotide variants
Sp	<i>Streptococcus pneumoniae</i>
st	Serotype
TB	Tuberculosis
UK	United Kingdom
WGS	Whole-genome sequencing
WHO	World Health Organisation

Table of Contents

DECLARATION OF AUTHORSHIP	1
ABSTRACT	2
ACKNOWLEDGEMENT	3
ABBREVIATIONS	4
CHAPTER 1: INTRODUCTION	6
1.1 <i>A brief history of genomics</i>	6
1.2 <i>Role of pathogen genomics in public health</i>	7
1.3 <i>Detection of linked infection or transmission clusters</i>	8
1.4 <i>Transmission directionality in epidemiological studies</i>	10
1.5 <i>Limitations when inferring the directionality of bacterial infections</i>	15
1.6 <i>Introducing Streptococcus pneumoniae</i>	17
1.7 <i>Introducing SARS-CoV-2</i>	22
1.8 <i>Aim</i>	24
1.9 <i>Objectives</i>	24
<i>References</i>	25
CHAPTER 2: PHYLOGENETIC INFERENCE OF PNEUMOCOCCAL TRANSMISSION FROM CROSS-SECTIONAL DATA, A PILOT STUDY	43
CHAPTER 3: EVALUATING METHODS IN IDENTIFYING AND QUANTIFYING STREPTOCOCCUS PNEUMONIAE SUBPOPULATION USING NEXT-GENERATION SEQUENCING DATA	77
CHAPTER 4: EFFECTIVENESS OF BNT162B2 AND CHADOX1 AGAINST SARS-COV-2 HOUSEHOLD TRANSMISSION: A PROSPECTIVE COHORT STUDY IN ENGLAND	102
CHAPTER 5: CHALLENGES IN PHYLOGENETIC INFERENCE OF WHO INFECTED WHOM WITH SARS-COV-2, A PROSPECTIVE HOUSEHOLD STUDY	116
CHAPTER 6: DISCUSSION AND CONCLUSION	136
6.1 <i>Summary on findings</i>	136
6.2 <i>Context on transmission directionality inference using phylogenetic approaches</i> ..	139
6.3 <i>Study limitations</i>	141
6.4 <i>The future of inferring transmission directionality using pathogen genomics</i>	143
6.5 <i>Conclusion</i>	145
<i>References</i>	146

CHAPTER 1: INTRODUCTION

1.1 A brief history of genomics

Genomics is the study of deoxyribonucleic acid (DNA) and genes and the interaction between those genes and their environment and a DNA sequence refers to a specific order of nucleotides in a DNA molecule. Polymerase chain reaction (PCR) was developed in 1985 to amplify or make multiple copies of targeted DNA sequences by combining the sample DNA with a mixture that contains nucleotides which then undergoes a process of denaturation, annealing, and extension of the sequences via temperature-mediated reactions¹. The final PCR product is multiple copies of the targeted DNA which is ready to undergo sequencing. Massive improvements have been made since the major breakthrough of Sanger's first-generation sequencing method in 1977² which has led to the automated sequencing of more complex species in 1991³. In the mid-2000s, decreased costs, increased efficiency, and throughput data via parallelisation led to second-generation or next-generation sequencing methods⁴ including the 454, Ion Torrent, and various Illumina platforms (Table 1). The high throughput of short read fragments generated from second-generation sequencing methods such as the MiSeq was ideal for studying gene expression and genotyping, however, it fell short on full genome assemblies⁵ due to the short sequence fragments. This led to third-generation sequencing methods including PacBio single-molecule real-time technology and Oxford Nanopore Technologies. Third-generation, compared to previous ones, does not require an amplification step and thus can produce much longer reads at the cost of read depth e.g. fewer number of sequences for specific regions on the genome (Table 2).

Table 1. Comparison of various second-generation sequencing platforms (from Kang *et al.*⁵)

Instrument	454 series	Ion Torrent	MiSeq series	NextSeq series	HiSeq X series	NovaSeq series
Amplification method	Emulsion PCR on based	Emulsion PCR on based	Solid-phase bridge amplification	Solid-phase bridge amplification	Solid-phase bridge amplification	Solid-phase bridge amplification
Sequencing mechanism	Pyrosequencing	Semiconductor sequencing	Sequencing by synthesis	Sequencing by synthesis	Sequencing by synthesis	Sequencing by synthesis
Run time	10 - 23 h	2 - 8 h	4 - 55 h	12 - 30 h	< 3 d	16 - 44 h
Maximum Output	700 Mb	15 Gb	15 Gb	120 Gb	1,800 Gb	6,000 Gb
Maximum Reads Per Run	1 Mb	80 Mb	25 million	400 million	6 billion	20 billion
Maximum Read length	Up to 1000 bp	200 - 400 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Error profile	1%, Indel	1%, Indel	0.1%, substitution	0.1%, substitution	0.1%, substitution	0.1%, substitution

Table 2. Comparison of various third-generation sequencing platforms (from Kang *et al.*⁵)

Instrument	PacBio RS II	PacBio Sequel	MinION	GridION
Sequencing mechanism	Single molecule real time (SMRT)	Single molecule real time (SMRT)	Nanopore sequencing	Nanopore sequencing
Run time	2 h	Up to 20 h	12 - 48 h	12 - 48 h
Maximum output	500 Mb - 1 Gb	3.5 - 7 Gb	10 - 20 Gb	50 - 100 Gb
Maximum reads per run	- 55,000	- 350,000	> 100,000	> 500,000
Maximum read length	- 20 kb	8 - 12 kb	Up to 100 kb	Up to 100 kb
Error profile	10 - 15% single pass, ≤1% circular consensus read, Indel	10 - 15% single pass, ≤1% circular consensus read, Indel	- 12%, Indel	- 12%, Indel

First-generation sequencing methods resulted in the first organism to be fully sequenced, *Haemophilus influenzae*⁶ in 1995 with 1.8 million base pairs (bp) followed by *Saccharomyces cerevisiae*⁷ in the subsequent year with 12 million base pairs. The Human Genome Project was initiated in 1990 and the first draft of the entire human genome was published using next-generation sequencing methods in 2000⁸ which has revolutionised our understanding of biology, human genetics, and infectious diseases. Since then, genome sequencing has had an important role across various studies including comparative and functional genomics, environmental studies, and public health. Further, the improvements to sequencing technologies have been fundamental to researchers and have made genome research more accessible globally with the caveat that high-resource countries having the most access⁹ while the equity gap is currently being addressed for low-resource areas that are underrepresented¹⁰.

1.2 Role of pathogen genomics in public health

Pathogen genomics has become more prevalent with the advent of affordable, rapid whole-genome sequencing (WGS) data that provides an additional level of detail and thus impacts the development of diagnostics, vaccines, therapies, and strategies for disease control. Currently, there are over 93 million bacterial sequences that are publicly available on the National Center for Biotechnology Information GenBank Database and over 15 million SARS-CoV-2 sequences on the Global Initiative on Sharing All Influenza Data (GISAID database)¹¹. Some of its vast utilities include molecular epidemiology techniques that have allowed us to rapidly and effectively detect and characterize pathogens using genotyping tools which in turn, can improve outbreak response times and surveillance of circulating strains¹². Additionally, sequencing data can provide information on phenotypic traits including antigenic type¹³ and antimicrobial resistance profiles¹⁴.

With the decreased cost of high-throughput methods, deep sequencing provides a new level of insight into infectious diseases that were not previously possible with consensus sequencing. Consensus sequences focus on the representative base at a particular position taking into account the quality of the base call while deep sequences (e.g. MiSeq) produce a large volume of short sequence reads in fragments which allows the identification of minor mutations and detection of multiple strains¹⁵ and quasispecies of the same pathogen during infection¹⁶. Detection of accumulated mutations over time via sequencing data in combination with epidemiological data can provide insight into communicable diseases including respiratory pathogens' origins, transmission routes, and evolutionary changes in the absence and presence of vaccines. Mutations accumulate in an organism's genome which is then passed on to the offspring allowing identification of related genomes. Additionally, in conjunction with temporal and spatial metadata, the clonality of sampled pathogen genomes can also help determine if cases are linked to an outbreak¹⁷ or not¹⁸. Integrating this information into epidemiological models can inform epidemic growth rates and reproductive numbers¹⁹.

1.3 Detection of linked infection or transmission clusters

Linkage or cluster of related infections can shed light on recent outbreaks or transmission events. To elucidate linked infections and then who infected whom, samples from an infected population are sequenced, then the sequences undergo phylogenetic reconstruction²⁰, then clusters of closely related sequences are identified, and then inference on transmission direction is validated against the epidemiological data. Phylogenetic reconstruction is well suited to infer linkage amongst isolates of related sequences by assessing the accumulated mutations between the transmission source and recipient e.g. nucleotide substitutions and can be measured using a distance-based or subtree-based method (Figure 1). The distance-based method measures the genetic distance or divergence that two sequences have relative to their common ancestor. Divergence can be calculated from the pairwise distance of the two sequences which is calculated as a percentage of matching nucleotides over the total number of aligned nucleotides. Alternatively, linkage can be detected using a phylogeny, which represents the evolutionary relationship among pathogens of sampled individuals (tips of the tree) and their unsampled common ancestors (internal nodes of the tree) using the patristic distance which is a measurement of branch lengths between two tips on the tree. The threshold defining linkage or cluster is variable and can be set for both pairwise and patristic distance calculations²¹.

The subtree-based method can be a portion of the phylogeny or clusters of tips that is defined by the demographics of the individuals represented by that particular portion e.g. geographical or risk groups. Alternatively, subtree-based clustering can be derived from a portion of the tree

that is within an assigned bootstrap support value which is the confidence value assigned to a branch on the tree²². The distance-based clustering methods, pairwise and patristic, can be computed rapidly, while the subtree-based clustering method is more computationally intensive due to the bootstrapping step where phylogenies are generated from sampling columns of the multiple sequence alignment with replacement meaning the same columns can be sampled more than once or not at all.

More information on limitations with linkage detection is in “**Identifying transmission pairs**”.

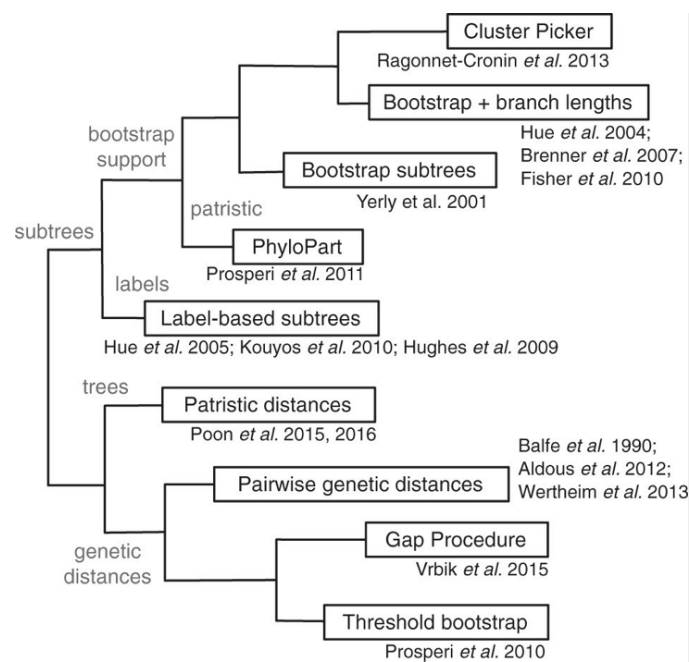


Figure 1. “A hierarchical clustering dendrogram of nonparametric genetic clustering methods. This dendrogram was generated from a binary character state matrix that encodes ten different features for nine categories of nonparametric methods. Internal nodes of the dendrogram are labelled with features that distinguish the categories below the node. Each category is annotated with a small number of citations to publications that either describe the method or provide examples of its usage; these are not meant to be exhaustive lists.” (Figure adapted from Poon (2016) (CC BY 4.0).²³

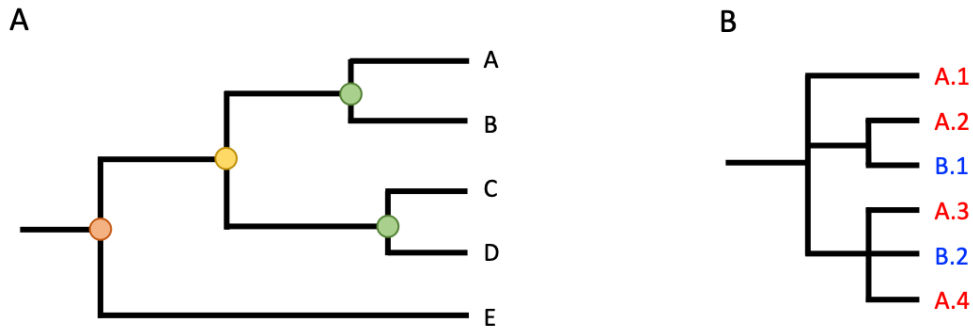


Figure 2. A) A basic phylogenetic tree with consensus sequences from individuals (A, B, C, D, E) located on the terminal branches of the tree and horizontal black lines represents the branch length or differences between samples. The internal nodes represent the common ancestry e.g. the two green internal nodes represent the common ancestors for A and B, and C and D, while the yellow internal node represents the common ancestors for A, B, C, and D which forms a cluster of four individuals. Lastly, the orange internal node represents the common ancestors for all the samples. The common ancestors can be inferred from tree reconstruction using distance-based methods, maximum likelihood, or Bayesian methods. B) A phylogenetic tree representing within-host sequences for individuals A and B only where individual A has four reads and individual B has two reads.

1.4 Transmission directionality in epidemiological studies

Methods to describe temporal or spatial clusters of closely related infection can only describe individuals at risk of disease acquisition without any additional information on directionality and thus factors associated with individuals contributing to ongoing spread and ongoing acquisition. Directionality plays an important role in the epidemiological assessment of infectious diseases to identify the sources and risks for infection and thus implement preventative measurements against disease acquisition which has been successful for investigating HIV in public health settings^{24–26}.

1.4.1 Current approaches to investigate directionality

Current approaches to detect transmission direction between individuals include using i) contact tracing data, ii) spatial or temporal inference, iii) phylogenetic reconstruction using pathogen genomes, or iv) a combination of either three²⁷. Contact tracing data including the time of symptom onset can aid in identifying the source and recipient of the infection, however, these inferred directions can be undermined by pre-symptomatic and asymptomatic cases²⁸ in addition to being costly and time-consuming. Alternatively, transmission chains can be elucidated from temporal data such as their symptom onsets to infer the probability that the recipient has been infected by the source given their time interval and duration of the infectious period of the pathogen of interest while accounting for missing and erroneous data²⁹. A

phylogenetic approach, using pathogen sequences, can aid in identifying more closely related pathogens than what would be expected by chance. However, this method can overlook the complexities of transmission chains including a cluster of tips that can represent multiple transmission scenarios and branching events that do not necessarily represent a transmission event³⁰ (Figure 2). Despite these limitations, the use of deep sequencing data from next-generation sequencing (NGS) methods to detect intra-host genetic diversity^{31–33} can be valuable in inferring transmission direction due to the ease of cross-sectional sample collection and widely adopted NGS data in research settings.

Previous approaches to infer transmission direction have mostly relied on epidemiological data which are time and cost intensive to follow and can lead to unreliable self-reported data from the study participants. Self-reported data can be affected by social desirability bias when the questionnaire asks about sensitive topics such as sexual partners. For example, in settings where there is a stigma associated with homosexuality, participants with HIV would report being heterosexual despite the phylogenetic data showing evidence of potential non-disclosed same-sex behaviour^{34,35} which would have public health implications on interventions. Self-reported data can also be affected by recall bias where the participant's ability to recall past events can be unreliable such as the date of symptom onset and thus leading to unreliable classification of the source and recipient in the transmission pair³⁶. In outbreak settings where the transmission involves multiple individuals such as within hospital, household, and school settings, determining the source of the infection can be difficult to disentangle in the absence of genomic data³⁷ due to a lack of evidence that could link patients to the outbreak. Recent developments in sequencing technologies and bioinformatic tools have streamlined the phylogenetic inference in the transmission direction of human pathogens with reduced cost and required resources.

The majority of phylogenetic reconstruction tools have been developed to study fast-evolving viral pathogens including human immunodeficiency viruses (HIV), hepatitis C virus, and Dengue^{38–40} where even a small portion of their genomes can accumulate sufficient mutations over a few months to inform phylogenetic inferences. Both HIV and hepatitis C viruses are ~10 kbp long and accumulate mutations between 10^{-3} and 10^{-5} substitutions/site/replication cycle⁴¹. Our ability to improve inference accuracy is observable from two studies using the same cohort, Rose *et al.* and Zhang *et al.* who inferred directionality among HIV-infected partners with epidemiological support for directionality^{42,43}. The increased accuracy by Zhang *et al.* compared to Rose *et al.* can be attributed to higher sequencing depth in addition to longer reads. The increased sequencing depth would increase the phylogenetic genetic signal by detecting minor variations from the between- and within-host genomes. While the longer

read fragments would aid in the genome assembly and thus provide more robust genomes additionally capturing more mutations on a single read resulting in more robust tree reconstruction. This suggests increasing the depth and/or read lengths could potentially increase the phylogenetic signals in inferring the transmission direction.

1.4.2 Ancestral state reconstruction for inferring directionality

Until the development of Phyloscanner, most of the available tools for phylogenetic inference lacked sufficient sensitivity to infer the direction of transmission. There are insufficient variations at the consensus genome level between transmission pairs to detect within-host diversity. Phyloscanner automates the phylogenetic analysis of next-generation sequencing short reads to detect the transmission direction, presence of multiple infections, recombination, and contamination for both viruses and bacterial sequencing data³³.

In Phyloscanner, directionality is determined from phylogenetic tree reconstruction for each of a set of sliding windows across an alignment of the generated reads (Figure 3) and the ancestral state at the root of the phylogeny (either the source or the recipient of the infection) is determined for each of the trees (subtrees) using the concepts proposed by Romero-Severson *et al.*⁴⁴. For each of the phylogenies reconstructed, the source of the infection is determined through a modified maximum-parsimony framework, where the most likely identity of the paired member is inferred at each node.

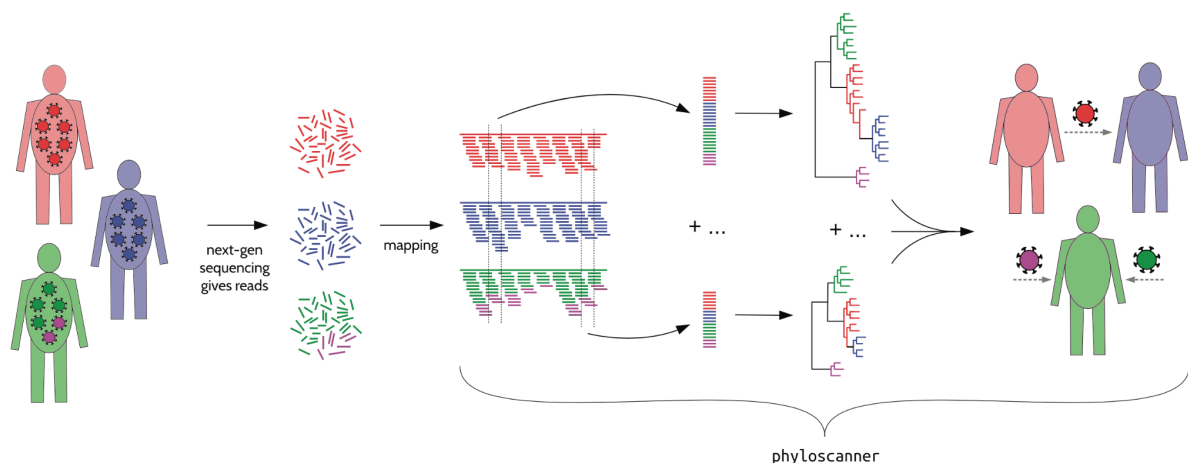


Figure 3. “Phyloscanner schematic for whole-genome deep sequence data. In this schematic, pathogens are sampled from the population infecting three hosts. NGS deep sequencing produces reads, which are fragments of the genome sequence of one pathogen particle (after amplification if necessary). Mapping to a reference means aligning each read to the appropriate location in the genome; this must be done beforehand, as mapped reads are the inputs to phyloscanner. Phyloscanner produces alignments of reads in sliding windows along

the genome, automatically adjusting for the fact that the reference may be different for each sample. Phylogenies are inferred for each alignment. These phylogenies are analysed separately using ancestral host-state reconstruction (e.g., assigning hosts to internal nodes), and their information is combined to give biologically and epidemiologically meaningful summaries. For example, here, we infer that the red individual infected the blue individual directly or indirectly, and the green individual has two distinct pathogen strains.”(Figure adapted from Wymant (2018) (CC BY 4.0).³³

Based on the topology and ancestral reconstruction, each subtree is classified as one of the following three relationships based on the subgraphs which are defined as the connected regions of the tree with the same host state: (i) single ancestry, where the subgraphs, form a paraphyletic (source) - monophyletic (recipient) relationship, (ii) equivocal, where the source and recipient subgraphs form dual monophyletic groups and thus the direction of infection is unclear, (iii) complex is where the subgraphs form paraphyletic - paraphyletic groups where the ancestral state is assigned to both the source and recipient depending on the subgraph (Figure 4). The subtrees or relationships are then aggregated and the one that occurs the most often is the most likely overall scenario for the individuals analysed and the amount of support for alternative relationships indicates the robustness of the inference. This approach minimizes the effects of random reconstruction errors that could have been introduced from contamination or sequencing methods.

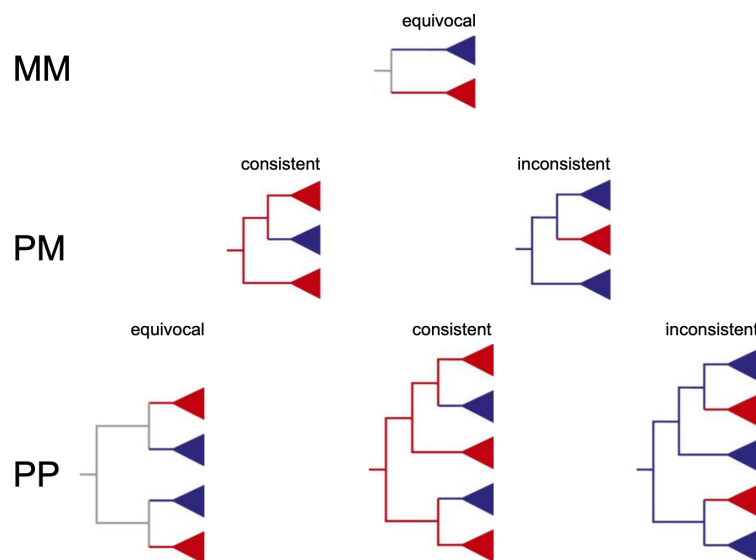


Figure 4. Classes of topological signal. When one host (red) is epidemiologically linked to another host (blue), the resulting virus populations upon sampling may relate to each other such that both populations are monophyletic (MM), or one is paraphyletic and the other

monophyletic (PM), or one is paraphyletic and the other polyphyletic relative to the other (PP). If the red host was infected first, the deduced root label of the phylogeny may be equivocal (the root node could be assigned to either host), consistent (correct root assignment in direct or indirect transmission cases), or inconsistent (incorrect root assignment in direct or indirect transmission cases).” (Figure adapted from Romero-Severson (2016) (CC BY 4.0).⁴⁴

The capacity to answer who infected whom using phylogenetics is limited by various factors in the data acquisition and parameters in the data inference. Factors include imperfect sampling such as not sampling the direct transmission pair; sequencing of the pathogen including not capturing sufficient intra-host diversity⁴⁵.

1.4.3 Implications of transmission bottleneck on inferring directionality

The transmission bottleneck, which refers to the amount of pathogen genetic diversity that is passed from the source to the recipient of the infection, has been proven to be difficult to estimate particularly among respiratory pathogens (Figure 5). This difficulty is due to variability in transmissibility and modes of transmission e.g. through direct physical contact, indirect physical contact, large droplets, or aerosols. An estimated 1-3 distinct viral particles are transmitted amongst both influenza and SARS-CoV-2 infections⁴⁶ and a single particle is sufficient to initiate a new infection⁴⁷. A narrow bottleneck reduces the amount of genetic information that is present in the recipient⁴⁸, there needs to be a wide bottleneck to be able to detect an adequate amount of within-host genetic diversity to assess both transmission linkage and direction of infection⁴⁹. If the bottleneck is too narrow, the bacterial population between linked individuals will be homogenous and if it is too large, the source and recipient will share the same distribution of mutations. For example, the transmission of influenza A virus, a respiratory pathogen, is affected by the mode of transmission; a narrow bottleneck would result in a reduction in the recipient within-host diversity but a wide bottleneck would result in increased diversity and potential acquisition of drug-resistant viruses^{50,51}.

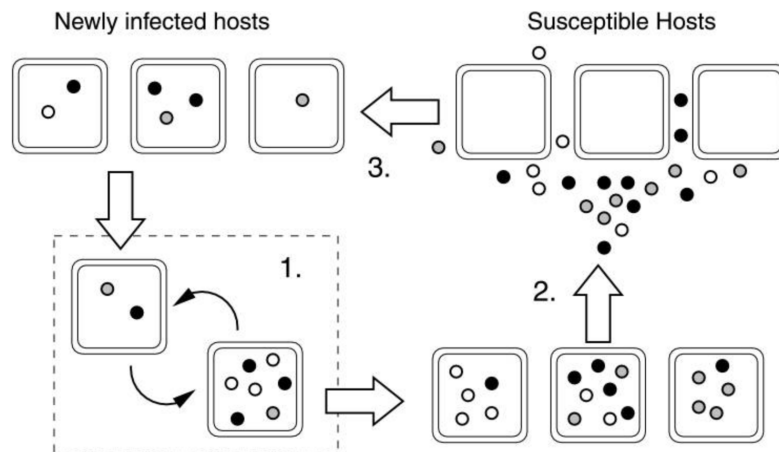


Figure 5. “The horizontal transmission model. Starting with a population of n newly infected hosts, the transmission cycle proceeds as follows. (1) Within each host, the virus undergoes t periods of intrahost process of replication and mutation. (2) Hosts carrying mature virus populations release viral particles into the environment. In this model, all hosts contribute viral particles, in proportion to their viral titer, to a common pool (e.g., a host carrying twice as many virions as a second host contributes twice as many viral particles to the common pool). (3) Susceptible hosts are infected by b viral particles drawn at random from this common pool, thereby generating a new set of newly infected hosts.” (Figure adapted from Bergstrom (1999) (CC BY 4.0).⁴⁷

1.5 Limitations when inferring the directionality of bacterial infections

1.5.1 Identifying transmission pairs

Determining linked infections is the first step in phylogenetic inference and understanding the evolutionary rates of the bacterial genome plays an important role in identifying putative transmission pairs. The evolutionary rate of bacteria is relatively slow compared to fast-evolving RNA viruses, between 10^{-4} to 10^{-3} substitutions/site/year⁴⁹, with bacteria evolving between 10^{-7} to 10^{-5} nucleotide substitutions/site/year. Further, the estimated rate negatively correlates with the sampling time frame e.g. estimates over a short period such as outbreak data will contain more deleterious mutations and would inflate the overall long-term evolutionary rate⁵².

The slow mutation rate of bacteria has implications for determining transmission linkage and how much genetic differences are allowed within the transmission model to determine if two individuals are directly linked. No or low genetic difference between two pathogens sampled from an infected population is indicative of potential recent transmission, however, bacterial genomes tend to have low genetic diversity at the population level due to slow evolutionary changes and large genomes compared to viruses. More specifically, the mutations that are observed are generally dispersed along the large genome thus there are only a few regions

that are conserved and hypervariable. The 16S rRNA region is one of the few which contains approximately 1,550 base pairs across the nine regions⁵³ and while this region is suitable for identifying strains and taxa, it lacks the resolution and accuracy in comparison to WGS for phylogenetic inference. While investigating direct transmission linkage, the number of mutations from suspected linked individuals should be distinguishable e.g. from independent clusters from randomly paired individuals⁵⁴, however, the definition of linkage is subjective and there are currently no formal guidelines on clustering criteria²³. More specifically, pairwise and patristic distance thresholds and bootstrap support values are subjective and dependent on the pathogen of interest.

1.5.2 Limited within-host diversity

Bacterial genome structures can be broadly categorised as i) closed or specialist or ii) open or generalist which is reflective of the ecological niches and selective pressures in which the bacteria has evolved⁵⁵. Additionally, bacterial genomes are comprised of three components i) the core genome which refers to the portion of the genome that is under purifying selection, the removal of harmful mutations over time, and contains the conserved functions of the bacteria, ii) the accessory genome which refers to the complement genes that is not a part of the core genome, and iii) pangenome is the full complement genes found within a bacterial population⁵⁵. The proportion of the three genome components will vary depending on if the bacteria is a specialist, which will generally have a proportionally larger core genome, or a generalist, which will generally have a proportionally smaller core genome⁵⁶.

The current standard approach for bacterial phylogenetics is to analyse the core genome using single-nucleotide polymorphisms (SNP). However, the bacterial core genome tends to be between 10^6 and 10^7 base pairs long while most viruses rarely exceed 10^6 base pairs⁵⁷. In conjunction with slow rates of evolution, this results in mutations that are sparsely distributed across the genome reducing population diversity compared to most viruses. Current standard sequencing methods primarily focus on targeted amplicon sequencing such as 16S-23S rRNA region for bacterial genomic data analysis which lacks sufficient diversity signal for phylogenetic analysis. The whole-genome sequencing approach is better able to characterise bacteria genotypes and phenotypes and provide insight into antimicrobial resistance⁵⁸ because it provides more information than targeted gene sequencing. Subsequently, it is better equipped to identify linked infections in a population of homogenous genotype infections compared to sequencing only a portion of the genome. While the lack of standardisation has been previously flagged in genomic analysis⁵⁹, we are still far from streamlining the analysis of bacterial whole-genome sequencing⁶⁰.

The inability to account for within-host diversity at the sampling sites such as the upper respiratory tract can lead to inaccurate inference on transmission directionality⁴⁸. There is a large overlap at the consensus genome level of infected individuals making source identification difficult. The pathogen population infecting an individual can be genetically diverse at the time of infection, if containing multiple strains or variants, and can further diversify due to within-host selective pressures⁴⁸. The within-host population dynamic is then maintained through natural selection, genetic drift, and changing population size⁶¹. Ancestral state reconstruction using next-generation sequencing data is suited to capture the within-host genetic population improving the transmission inference³³.

1.5.3 Recombination

Exchanging genetic material via recombination is fundamental to bacterial evolution and adaptation which results in mosaic genomes where the components exhibit different evolutionary histories. Bacteria can acquire antibiotic resistance genes and virulence traits through horizontal gene transfer and homologous recombination including antiphagocytic properties, adherence factors, invasion genes, and host-defence evasion⁶²⁻⁶⁴, resulting in new pathogenic strains and serotypes^{65,66}. Recombination events are important to better understand pathogen evolution and there is a considerable amount of interest in detecting recombinant events⁶⁷.

Recombined pathogen genomes violate the assumptions of tree reconstruction that isolates share only one ancestor, therefore, recombinant genomes are often discarded in phylogenetic studies, despite their potential role in a better understanding of the pathogen's molecular evolution⁶⁸. Tree reconstruction is impacted by strong selective pressures resulting in biased exchanges of DNA⁶⁹, therefore highly conserved genes that evolve under high selective pressure might be good candidates for tree reconstruction⁷⁰. Recombination Detection Programs (RDP), including RDP4 and RDP5^{71,72}, are able to detect and characterise recombination events by identifying potential recombination breakpoints, regions with different phylogenetic relationships from a multiple sequence alignment then statistically test the significance of the changes⁷³. One approach to handle recombinant events is to partition the data into genomic regions of recombination or non-recombination and apply different evolutionary models to each partition or implement a step-by-step approach as proposed by *Didelot et al.* by constructing dated phylogenies that account for recombination⁷⁴.

1.6 Introducing *Streptococcus pneumoniae*

Streptococcus pneumoniae (pneumococcus) is an opportunistic Gram-positive bacterium that asymptotically colonises the human upper respiratory tract. Pneumococcus is considered

invasive when it proliferates to other areas of the respiratory tract or aspirates to sterile body fluids, which can lead to invasive pneumococcal diseases (IPD) including pneumonia or meningitis⁷⁵. Pneumococcal disease contributes substantially to worldwide morbidity and mortality amongst young children (1 to 5 years old), older adults, and immunocompromised individuals⁷⁶. Children are more likely to acquire pneumococci compared to adults⁷⁷. Moreover, pneumococcal disease is one of the largest contributors to mortality amongst children less than five years old (Figure 6)^{78,79}.

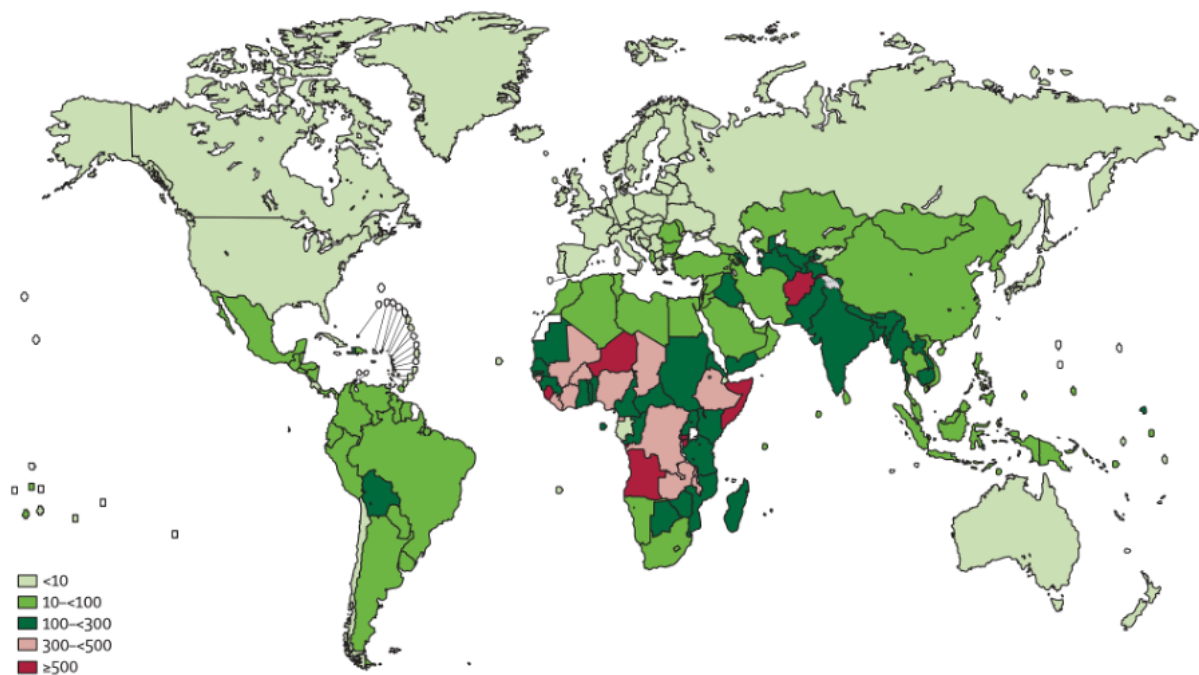


Figure 6. “Pneumococcal deaths in children aged 1–59 months per 100 000 children younger than 5 years (HIV-negative pneumococcal deaths only). The boundaries shown and the designations used on this map do not imply the expression of any opinion by WHO concerning the legal status of any country, territory, city, or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.”(Figure adapted from O’Brien (2009) (CC BY 4.0)).⁷⁹

1.6.1 *Pneumococcal colonisation*

Pneumococcal colonisation is a prerequisite for IPD. Pneumococci can colonise children within the first days of birth and carriage is typically higher amongst children in lower-resource settings^{79,80}. Child carriage rates in high-income settings are between 25% to 50% while in low-to middle-income settings rates are between 20% to 90%⁸¹. Colonisation is usually dominated by a single serotype; however, multiple serotype colonisation has been previously observed⁸².

Moreover, carriage rates can vary by demographics such as age, household settings, and presence/absence of upper respiratory infection^{79,83,84}. Carriage duration also differs for children (weeks) compared to adults (months)⁸⁵. Further, the duration of carriage amongst children can vary by serotype such as 28 days for serotype 20 and up to 124 days for serotype 6A^{86–88}.

1.6.2 Transmission

The pneumococcus is a commensal inhabitant of healthy individuals' nasopharynx and upper respiratory tract, however, there have been previous reports of outbreaks in certain settings^{89,90}. There is also evidence children could be possible pneumococcal reservoirs during outbreaks⁹¹. The main route of transmission is predominantly through direct contact with contaminated respiratory secretion between members within the same house, infants, and children⁸⁵. The primary source of transmission is from the nasopharynx mostly via aerosol⁹² and less frequently through fomites⁹³ to another individual's nasopharynx. While pneumococcus can be detected in other parts of the body including blood, urine, and cerebrospinal fluid,⁹⁴ they are an unlikely transmission source that would seed an infection in the recipient's nasopharynx. A study revealed an association between increased carriage prevalence and ethnicity which was a result of heightened transmission due to a higher frequency of physical contact⁹⁵. Other studies have determined young children were the main source of transmission^{96,97}. Young children contribute substantially to ongoing transmission highlighting the importance of high pneumococcal conjugate vaccine (PCV) coverage to induce herd protection against vaccine type *Streptococcus pneumoniae*.

1.6.3 Invasive pneumococcal disease

IPD is defined by the isolation of pneumococcus from a sterile site including the inner ear, blood, and central nervous system which could lead to pneumonia, bacteraemia, and meningitis⁹⁸. Similarly to carriage rates, disease outcome is also dependent on age, genetic background, socio-demographics, and immune status^{85,99}. For children in lower- to middle-income settings, the higher risk of childhood pneumonia is increased amongst those who are malnourished, immunocompromised, and those who have been exposed to tobacco smoke or other air pollutants¹⁰⁰.

There is a heightened risk of IPD for children, the elderly, and people with underlying comorbidities⁹¹. Other studies have suggested infection with other respiratory viruses can increase pneumococcal infection^{101–103}. More specifically, Wolter *et al.* observed respiratory viruses were associated with increased colonisation density, and thus IPD¹⁰² while Launes *et*

al. reported that nearly half of the children with IPD were also co-infected with respiratory viruses¹⁰¹.

1.6.4 *Pneumococcal serotypes and their global distribution*

Most pneumococci are encapsulated with a complex polysaccharide that contributes to their virulence and pathogenicity¹⁰⁴. Previous findings demonstrate serotypes with thicker capsules are associated with increased mortality^{105,106}. The capsule polysaccharide locus (*cps*) encodes for the capsule biosynthesis gene cluster. All typeable *Streptococcus pneumoniae* are characterised by the *cps* locus which is flanked by the *dexB* and *aliA* genes^{107,108} and there are currently more than 100 different pneumococcal serotypes identified¹⁰⁹. The Global Pneumococcal Sequencing project has provided surveillance and insight into circulating pneumococcal strains and a better understanding of the impact of vaccines¹¹⁰ with the potential to forecast emerging strains enabling interventions. A systematic review of countries that have already introduced PCV reported, overall, non-PCV13 serotypes contribute to 42% of childhood IPD cases in descending order 22F, 12F, 33F, 24F, 15C, 15B, 23B, 10A, and 38¹¹¹.

Serotyping relies on the capsule region and a complete or partial deletion of the *cps* results in defective capsules, thus do not have evidence of capsule expression or possess a capsule for which there are no current typing antisera and are categorised as non-typeable (NT)^{112,113}. A previous study demonstrated a single-point mutation in the *wchA* gene resulted in serotype change from 7F to NT¹¹⁴. Isolate classified as NT could be due to current tools only being limited to serotyping based on the capsular protein or the limitation of the currently available antisera. Following the introduction of PCV in the early 2000s, a study revealed an increase in NT pneumococcal isolates from 1.5% in 2001 to 5.6% in 2006¹¹⁵. Moreover, carriage of non-typeable pneumococcal isolates has predisposed those who develop infectious conjunctivitis and IPD^{116,117}.

1.6.5 *Pneumococcal conjugate vaccines and potential impact on transmission*

The diversity of pneumococcus is based on the capsular polysaccharides and is the basis of pneumococcal vaccines¹⁰⁷. The current vaccines against *Streptococcus pneumoniae* on the market include the 23-valent pneumococcal polysaccharide-based vaccine (PPV23) and three polysaccharide-protein conjugate vaccines (PCV13, PCV15, PCV20) where PCV13 is recommended for children under 5 years old and PCV15 or PCV20 for adults over 65 years old¹¹⁸. Indirect protection can be established by reducing vaccine-type carriage by vaccinating thus mitigating onward spread¹¹⁹⁻¹²². Routine infant PCV vaccination in high-income countries has greatly reduced carriage and IPD of vaccine serotypes¹²³⁻¹²⁵. PCVs are effective in reducing pneumococcal disease but they are costly and not effective against emerging non-

vaccine types^{126,127}. There has been serotype replacement as a result of types that are not included in the current vaccines within vaccinated communities^{128–130}. Moreover, capsular switching is a recombination event of large DNA fragments usually including the capsular gene resulting in a different serotype¹³⁰. Both events do limit overall vaccine effectiveness.

Despite the financial aid from the Global Vaccine Alliance (GAVI) to support the uptake of PCV in many low-income countries, the cost associated with either of the World Health Organisation (WHO) recommended 3-dose schedule is still a barrier to PCV uptake in middle-income countries who are not able to receive support from GAVI despite the countries' limited resources^{126,131,132}. One potential approach in ameliorating these costs is to reduce the number of PCV from 3 to 2 doses^{133,134} and a better understanding of pneumococcal transmission dynamics can aid vaccination strategies without compromising herd immunity.

1.6.6 Introducing co-carriage and its potential role in transmission dynamics

While most bacterial carriage is dominated by a single strain, multi-strain carriage is frequently observed among carriers of *Streptococcus pneumoniae*¹⁵. Multiple carriage can promote recombination such as acquiring genetic elements from other microbes through transformation. Pneumococci is highly transformable and can recombine at the *cps* locus resulting in a different serotype also known as serotyping switching and this could aid in vaccine escape¹⁵ and contribute to its pathogenicity^{135,136}. Carriage with multiple sequence types can increase the within-host bacteria diversity, however, inferring the transmission direction in the presence of a strong transmission bottleneck where a single strain is transmitted from the source to the recipient can be difficult due to potential unsampled lineages from either the source or recipient³¹. To mitigate this limitation, adequate sampling at the collection and sequencing steps would be needed for both the source and recipient of the transmission. Carriage of multiple unique pneumococcal strains is common in settings with high carriage prevalence¹³⁷ with an estimated 40% of children having co-carriage within the first year of life in The Gambia^{15,138}. Carriage of multiple serotypes provides an opportunity for horizontal gene transfer could lead to serotype switching and thus result in vaccine escape^{139,140}. This highlights the importance of mixed-serotype carriage surveillance when monitoring vaccine impact. Due to the complexities of co-carriage, studies have mostly relied on single serotype carriage from a representative genome or from purified single colony picks which limits the sensitivity of carriage surveillance underestimating the carriage rates¹⁴¹.

The Quellung reaction is the current gold standard for serotyping pneumococcus using serotype-specific monoclonal antibodies where the pneumococcal isolates are sequentially tested first with a pooled antisera and then against each individual antisera¹⁴². The antibody

binds to the pneumococcal capsule and apparent swelling, under a microscope, indicates the presence of the specific polyclonal antibody. This method is labour-intensive requires training and expertise and is not scalable to large studies¹⁴² thus is now primarily used by reference laboratories¹⁴³. Other methods include the dot blot assay and latex agglutination, however, these methods are time-consuming¹⁴⁴ and are only limited to a few serogroups/serotypes¹⁴⁵, respectively. DNA microarray is another serotyping technique with high sensitivity and specificity that uses serotype and serogroup-specific probes to target the highly variable glycosyltransferase genes^{146–148}. DNA microarray analysis is rapid and can be used to detect and quantify serotypes and co-carriage of multiple serotypes¹⁴⁹. However, this method requires trained personnel and is not as accessible due to logistical and cost constraints.

WGS for pneumococcal serotyping has been widely used in IPD surveillance and can effectively identify serotypes from a single carriage and co-carriage^{150–153}. Tools such as PneumoCaT and SeroBA use a k-mer-based method to detect concordance of the *cps* locus with the pneumococcus Capsular Type Variant database with high sensitivity, 99% and 98%, respectively but offer little to no co-carriage detection^{150,153}. PneumoKITy has been developed using the same approaches as SeroBA but includes flexibility to allow multiple serotypes to be detected with 92% sensitivity to identify co-carriage¹⁵¹. Further, SeroCall uses a mapping approach, and while this method is 100% sensitive for major serotypes and 86% for minor, it is also more computationally intensive compared to PneumoKITy¹⁵². Additionally, SeroCall was validated from co-carriage of up to five distinct serotypes and implemented more costly sequencing methods to generate greater read depths. Co-carriage is important for pneumococcal surveillance and there are free-access genomic serotyping tools available to detect co-carriage.

1.7 Introducing SARS-CoV-2

According to the WHO (<https://covid19.who.int/>) the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has caused over seven million deaths since it was first identified in 2020 with many more cases being mild or leading up to severe pneumonia. The efficiency of SARS-CoV-2 human-to-human transmission during the rapid rise of cases prior to national lockdowns was evident. Transmission arises from the respiratory tract of an infectious individual and can be transmitted via inhalation of fine aerosols and small droplets or contamination of the eyes, nose, or mouth from large droplets¹⁵⁴. Transmission efficiency is affected by the viral load of the infector in addition to external factors including ventilation, temperature, humidity, and behaviour factors including mask-wearing, cleaning of the exposed surface, vaccination, and immunity level¹⁵⁴. While most individuals were infectious around the time of symptom onset, contact-tracing information revealed frequent pre-symptomatic^{155,156}

and asymptomatic infection¹⁵⁷. Additionally, even in high-income countries, most cases were not ascertained due to limited testing capacity, asymptomatic cases, and fear of loss of livelihood from infection^{158,159}. The high transmissibility is largely due to virological factors of SARS-CoV-2 including high viral load and thus high virus shedding during the first week of symptom onset¹⁶⁰.

An estimated 10% of infectious individuals have contributed to 80% of secondary SARS-CoV-2 transmission¹⁶¹ with indoor settings such as hospitals, care homes, schools, and households acting as potential hubs for linked cases¹⁶². A review and meta-analysis from Madewell *et al.* estimated a 16.6% household secondary attack with higher attack rates from symptomatic cases compared to asymptomatic and transmission to adult contacts compared to transmission to child contacts¹⁶³. A more recent review and meta-analysis estimated a secondary attack rate from household child index cases to be lower at 7.6%¹⁶⁴. Inconsistent study definitions of index cases contribute to heterogeneous secondary attack rate (SAR) across different studies¹⁶³. Inference of household transmission pairs can be unreliable thus genomic data has the potential to reduce the risk of coincidental infections being attributed to household settings by identifying dissimilar infections using a phylogenetic approach. The inclusion of sequencing data has been previously analysed in SAR which excluded potential outside-of-the-household infections and thus added confidence in the household case-contact designation¹⁶⁵.

Routine sequencing of coronavirus disease-positive (COVID-19) samples during the pandemic has helped track and monitor the spread of variants of concerns. The COVID-19 Genomics UK Consortium (COG-UK) had sequenced over half a million viral samples by the summer of 2021 which allowed the epidemic in England to be characterised^{166,167}. Dominant circulating variant B.1.1.7¹⁶⁸⁻¹⁷⁰ (alpha) was rapidly replaced by B.1.617 (delta) over the summer of 2021 in the UK¹⁷¹, simultaneously, vaccine efficacy against delta infection diminished depending on the vaccine type¹⁷² and was even lower in household compared to community settings¹⁷³.

Whole-genome analysis has been routinely implemented and applied to track and identify mutations and subsequently variants of concern in the UK and these data help guide public health interventions during the COVID-19 pandemic¹⁷⁴. Large-scale genomic epidemiology coupled with phylogenetic analysis has provided insight into the epidemic in the UK with an estimate of more than 1,000 independent introductions of SARS-CoV-2 into the UK during the early stages of the epidemic¹⁶⁶.

The SARS-CoV-2 genome is about ~30 kb and most infections have limited within-host

diversity and with a large number of mutations inevitably lost during the transmission process from the source to the recipient¹⁷⁵. In cross-sectional sampling near the time of the transmission event, the presence of a lower genetic diversity in the recipient compared to the source due to the small founding population^{176,177} can aid in determining the transmission direction.

1.8 Aim

This PhD thesis aimed to evaluate the potential of phylogenetic inference methods for detecting putative transmission pairs and assessing the direction of transmission for both *Streptococcus pneumoniae* and SARS-CoV-2.

1.9 Objectives

The objectives of this PhD thesis are stratified into the following:

1. *Streptococcus pneumoniae*
 - a. Evaluate the potential of phylogenetic inference to identify the direction of pneumococcal transmission (**Chapter 2**)
 - b. Assess the potential to discriminate multiple pneumococcal serotypes from co-carriage NGS data (**Chapter 3**)
2. SARS-CoV-2
 - a. The uses of genomic data to identify unlinked infections in household studies (**Chapter 4**)
 - b. Evaluate the potential of phylogenetic inference to identify the transmission direction of SARS-CoV-2 (**Chapter 5**)

References

1. Mullis, K. *et al.* Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**, 5463–5467 (1977).
3. Hunkapiller, T., Kaiser, R. J., Koop, B. F. & Hood, L. Large-Scale and Automated DNA Sequence Determination. *Science* **254**, 59–67 (1991).
4. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16–18 (2008).
5. Kang, Y., Kang, C.-S. & Kim, C. History of Nucleotide Sequencing Technologies: Advances in Exploring Nucleotide Sequences from Mendel to the 21st Century. *Hortic. Sci. Technol.* **37**, 549–558 (2019).
6. Fleischmann, R. D. *et al.* Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
7. Goffeau, A. *et al.* Life with 6000 Genes. *Science* **274**, 546–567 (1996).
8. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
9. Phillips, K. A., Douglas, M. P., Wordsworth, S., Buchanan, J. & Marshall, D. A. Availability and funding of clinical genomic sequencing globally. *BMJ Glob. Health* **6**, e004415 (2021).
10. Jooma, S., Hahn, M. J., Hindorff, L. A. & Bonham, V. L. Defining and Achieving Health Equity in Genomic Medicine. *Ethn. Dis.* **29**, 173–178 (2019).
11. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly.* **3**, 1049–1051 (2021).
12. Armstrong, G. L. *et al.* Pathogen Genomics in Public Health. *N. Engl. J. Med.* **381**, 2569–2580 (2019).
13. Croucher, N. J. *et al.* Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proc. Natl. Acad. Sci.* **114**, (2017).

14. Metcalf, B. J. *et al.* Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clin. Microbiol. Infect.* **22**, 1002.e1-1002.e8 (2016).
15. Kamng'ona, A. W. *et al.* High multiple carriage and emergence of *Streptococcus pneumoniae* vaccine serotype variants in Malawian children. *BMC Infect. Dis.* **15**, 234 (2015).
16. Dudouet, P. *et al.* SARS-CoV-2 quasi-species analysis from patients with persistent nasopharyngeal shedding. *Sci. Rep.* **12**, 18721 (2022).
17. Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of methicillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* **13**, 130–136 (2013).
18. Walker, T. M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
19. Stockdale, J. E., Liu, P. & Colijn, C. The potential of genomics for infectious disease forecasting. *Nat. Microbiol.* **7**, 1736–1743 (2022).
20. Volz, E. M. & Frost, S. D. W. Inferring the source of transmission with phylogenetic data. *PLoS Comput. Biol.* **9**, e1003397 (2013).
21. Aldous, J. L. *et al.* Characterizing HIV Transmission Networks Across the United States. *Clin. Infect. Dis.* **55**, 1135–1143 (2012).
22. Felsenstein, J. Confidence Limits on Phylogenies: an Approach Using the Bootstrap. *Evolution* **39**, 783–791 (1985).
23. Poon, A. F. Y. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evol.* **2**, vew031 (2016).
24. Paraskevis, D., Nikolopoulos, G. K., Magiorkinis, G., Hodges-Mameletzis, I. & Hatzakis, A. The application of HIV molecular epidemiology to public health. *Infect. Genet. Evol.* **46**, 159–168 (2016).
25. Wertheim, J. O., Chato, C. & Poon, A. F. Y. Comparative analysis of HIV sequences in

- real time for public health. *Curr. Opin. HIV AIDS* **14**, 213–220 (2019).
26. Capoferri, A. A., Bale, M. J., Simonetti, F. R. & Kearney, M. F. Phylogenetic inference for the study of within-host HIV-1 dynamics and persistence on antiretroviral therapy. *Lancet HIV* **6**, e325–e333 (2019).
 27. Robert, A. *et al.* Quantifying the value of viral genomics when inferring who infected whom in the 2014–16 Ebola virus outbreak in Guinea. *Virus Evol.* **9**, vead007 (2023).
 28. Kucharski, A. J. *et al.* Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. *Lancet Infect. Dis.* **20**, 1151–1160 (2020).
 29. Hens, N., Calatayud, L., Kurkela, S., Tamme, T. & Wallinga, J. Robust Reconstruction and Analysis of Outbreak Data: Influenza A(H1N1)v Transmission in a School-based Population. *Am. J. Epidemiol.* **176**, 196–203 (2012).
 30. Kenah, E., Britton, T., Halloran, M. E. & Longini, I. M. Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. *PLoS Comput. Biol.* **12**, e1004869 (2016).
 31. Sashittal, P. & El-Kebir, M. Sampling and summarizing transmission trees with multi-strain infections. *Bioinformatics* **36**, i362–i370 (2020).
 32. De Maio, N., Worby, C. J., Wilson, D. J. & Stoesser, N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput. Biol.* **14**, e1006117 (2018).
 33. Wymant, C. *et al.* PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Mol. Biol. Evol.* **35**, 719–733 (2018).
 34. Ragonnet-Cronin, M. *et al.* Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data. *Lancet HIV* **5**, e309–e316 (2018).
 35. Chen, Y. *et al.* Inferring potential non-disclosed men who have sex with men among self-reported heterosexual men with HIV in Southwest China: A genetic network study. *PLoS ONE* **18**, e0283031 (2023).

36. Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. *Int. J. Infect. Dis.* **93**, 284–286 (2020).
37. Snitkin, E. S. *et al.* Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Sci. Transl. Med.* **4**, (2012).
38. Hackman, J. *et al.* Correlates of hepatitis C viral clustering among people who inject drugs in Baltimore. *Infect. Genet. Evol.* **77**, 104078 (2020).
39. Bbosa, N. *et al.* Phylogeography of HIV-1 suggests that Ugandan fishing communities are a sink for, not a source of, virus from general populations. *Sci. Rep.* **9**, 1051 (2019).
40. Faria, N. R. *et al.* Genomic and epidemiological characterisation of a dengue virus outbreak among blood donors in Brazil. *Sci. Rep.* **7**, 15216 (2017).
41. Martinez, M. A. Diversity and Evolution of HIV and HCV. *Viruses* **13**, 642 (2021).
42. Zhang, Y. *et al.* Evaluation of Phylogenetic Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV) Transmission: HIV Prevention Trials Network (HPTN) 052. *Clin. Infect. Dis.* **72**(1), 30-37. <https://doi.org/10.1093/cid/ciz1247>
43. Rose, R. *et al.* Phylogenetic Methods Inconsistently Predict the Direction of HIV Transmission Among Heterosexual Pairs in the HPTN 052 Cohort. *J. Infect. Dis.* **220**, 1406–1413 (2019).
44. Romero-Severson, E. O., Bulla, I. & Leitner, T. Phylogenetically resolving epidemiologic linkage. *Proc. Natl. Acad. Sci.* **113**, 2690–2695 (2016).
45. Hall, M. D., Woolhouse, M. E. J. & Rambaut, A. Using genomics data to reconstruct transmission trees during disease outbreaks: -EN- -FR- L'utilisation des données sur le génome pour reconstituer l'arborescence de la transmission lors d'un foyer de maladie - ES- Utilización de datos genómicos para reconstruir árboles de transmisión durante brotes infecciosos. *Rev. Sci. Tech. OIE* **35**, 287–296 (2016).
46. Bendall, E. E. *et al.* Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. *Nat. Commun.* **14**, 272 (2023).
47. Bergstrom, C. T., McElhany, P. & Real, L. A. Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proc. Natl. Acad. Sci.* **96**, 5095–5100 (1999).

48. Worby, C. J., Lipsitch, M. & Hanage, W. P. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS Comput. Biol.* **10**, e1003549 (2014).
49. Didelot, X., Walker, A. S., Peto, T. E., Crook, D. W. & Wilson, D. J. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150–162 (2016).
50. Frise, R. *et al.* Contact transmission of influenza virus between ferrets imposes a looser bottleneck than respiratory droplet transmission allowing propagation of antiviral resistance. *Sci. Rep.* **6**, 29793 (2016).
51. Varble, A. *et al.* Influenza A Virus Transmission Bottlenecks Are Defined by Infection Route and Recipient Host. *Cell Host Microbe* **16**, 691–700 (2014).
52. Duchêne, S. *et al.* Genome-scale rates of evolutionary change in bacteria. *Microb. Genomics* **2**, 11 (2016).
53. Clarridge, J. E. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clin. Microbiol. Rev.* **17**, 840–862 (2004).
54. Rose, R. *et al.* Complex patterns of Hepatitis-C virus longitudinal clustering in a high-risk population. *Infect. Genet. Evol.* **58**, 77–82 (2018).
55. Touchon, M. *et al.* Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet.* **16**, e1008866 (2020).
56. Ingle, D. J., Howden, B. P. & Duchene, S. Development of Phylodynamic Methods for Bacterial Pathogens. *Trends Microbiol.* **29**, 788–797 (2021).
57. Chao, E. *et al.* Molecular source attribution. *PLoS Comput. Biol.* **18**, e1010649 (2022).
58. Balloux, F. *et al.* From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. *Trends Microbiol.* **26**, 1035–1048 (2018).
59. Lubin, I. M. *et al.* Principles and Recommendations for Standardizing the Use of the Next-Generation Sequencing Variant File in Clinical Settings. *J. Mol. Diagn.* **19**, 417–426 (2017).
60. Croucher, N. J., Harris, S. R., Grad, Y. H. & Hanage, W. P. Bacterial genomes in

- epidemiology—present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120202 (2013).
61. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* **13**, 601–612 (2012).
 62. Moran, G. J. *et al.* Methicillin-resistant *S. aureus* infections among patients in the emergency department. *N. Engl. J. Med.* **355**, 666–674 (2006).
 63. Kelly, C. P. & LaMont, J. T. *Clostridium difficile*—more difficult than ever. *N. Engl. J. Med.* **359**, 1932–1940 (2008).
 64. Martín-Galiano, A. J., Wells, J. M. & de la Campa, A. G. Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. *Microbiol. Read. Engl.* **150**, 2313–2325 (2004).
 65. Levin, B. R. & Cornejo, O. E. The Population and Evolutionary Dynamics of Homologous Gene Recombination in Bacteria. *PLoS Genet.* **5**, e1000601 (2009).
 66. Gürtler, V. & Mayall, B. C. Genomic approaches to typing, taxonomy and evolution of bacterial isolates. *Int. J. Syst. Evol. Microbiol.* **51**, 3–16 (2001).
 67. Shikov, A. E., Malovichko, Y. V., Nizhnikov, A. A. & Antonets, K. S. Current Methods for Recombination Detection in Bacteria. *Int. J. Mol. Sci.* **23**, 6257 (2022).
 68. Posada, D. How does recombination affect phylogeny estimation? *Trends Ecol. Evol.* **15**, 489–490 (2000).
 69. Shapiro, B. J. *et al.* Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* **336**, 48–51 (2012).
 70. Stott, C. M. & Bobay, L.-M. Impact of homologous recombination on core genome phylogenies. *BMC Genomics* **21**, 829 (2020).
 71. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, 1 (2015).
 72. Martin, D. P. *et al.* RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **7**, veaa087 (2021).

73. González-Torres, P., Rodríguez-Mateos, F., Antón, J. & Gabaldón, T. Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. *mBio* **10**, e02494-18 (2019).
74. Didelot, X. & Parkhill, J. A scalable analytical approach from bacterial genomes to epidemiology. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210246 (2022).
75. Henriques-Normark, B. & Tuomanen, E. I. The Pneumococcus: Epidemiology, Microbiology, and Pathogenesis. *Cold Spring Harb. Perspect. Med.* **3**, a010215–a010215 (2013).
76. (CDC), C. for D. C. *Global Pneumococcal Disease and Vaccine*. <https://www.cdc.gov/pneumococcal/global.html#disease>.
77. Mosser, J. F. *et al.* Nasopharyngeal carriage and transmission of *Streptococcus pneumoniae* in American Indian households after a decade of pneumococcal conjugate vaccine use. *PLoS One* **9**, e79578 (2014).
78. Wahl, B. *et al.* 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. *Lancet Glob. Health* **6**, e744–e757 (2018).
79. O'Brien, K. L. *et al.* Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet* **374**, 893–902 (2009).
80. Granat, S. M. *et al.* Longitudinal study on pneumococcal carriage during the first year of life in Bangladesh. *Pediatr. Infect. Dis. J.* **26**, 319–324 (2007).
81. Croucher, N. J., Løchen, A. & Bentley, S. D. Pneumococcal Vaccines: Host Interactions, Population Dynamics, and Design Principles. *Annu. Rev. Microbiol.* **72**, 521–549 (2018).
82. Hare, K. M., Morris, P., Smith-Vaughan, H. & Leach, A. J. Random colony selection versus colony morphology for detection of multiple pneumococcal serotypes in nasopharyngeal swabs. *Pediatr. Infect. Dis. J.* **27**, 178–180 (2008).
83. Regev-Yochay, G. *et al.* Nasopharyngeal carriage of *Streptococcus pneumoniae* by adults and children in community and family settings. *Clin. Infect. Dis.* **38**, 632–639 (2004).

84. Le Polain de Waroux, O., Flasche, S., Prieto-Merino, D. & Edmunds, W. J. Age-dependent prevalence of nasopharyngeal carriage of *Streptococcus pneumoniae* before conjugate vaccine introduction: a prediction model based on a meta-analysis. *PLoS One* **9**, e86136 (2014).
85. van der Poll, T. & Opal, S. M. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *The Lancet* **374**, 1543–1556 (2009).
86. Smith, T. *et al.* Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children. *Epidemiol. Infect.* **111**, 27–39 (1993).
87. Sleeman, K. L. *et al.* Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. *J. Infect. Dis.* **194**, 682–688 (2006).
88. Lipsitch, M. *et al.* Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model. *Epidemiol. Camb. Mass* **23**, 510–519 (2012).
89. Romney, M. G. *et al.* Large community outbreak of *Streptococcus pneumoniae* serotype 5 invasive infection in an impoverished, urban population. *Clin. Infect. Dis.* **47**, 768–774 (2008).
90. Crum, N. F. *et al.* Halting a pneumococcal pneumonia outbreak among United States Marine Corps trainees. *Am. J. Prev. Med.* **25**, 107–111 (2003).
91. Zivich, P. N., Grabenstein, J. D., Becker-Dreps, S. I. & Weber, D. J. *Streptococcus pneumoniae* outbreaks and implications for transmission and control: a systematic review. *Pneumonia* **10**, 11 (2018).
92. Weiser, J. N., Ferreira, D. M. & Paton, J. C. *Streptococcus pneumoniae*: transmission, colonization and invasion. *Nat. Rev. Microbiol.* **16**, 355–367 (2018).
93. Morimura, A., Hamaguchi, S., Akeda, Y. & Tomono, K. Mechanisms Underlying Pneumococcal Transmission and Factors Influencing Host-Pneumococcus Interaction: A Review. *Front. Cell. Infect. Microbiol.* **11**, 639450 (2021).
94. Samra, Z., Shmueli, H., Nahum, E., Paghis, D. & Ben-Ari, J. Use of the NOW

- Streptococcus pneumoniae* urinary antigen test in cerebrospinal fluid for rapid diagnosis of pneumococcal meningitis. *Diagn. Microbiol. Infect. Dis.* **45**, 237–240 (2003).
95. Neal, E. F. G. *et al.* A Comparison of Pneumococcal Nasopharyngeal Carriage in Very Young Fijian Infants Born by Vaginal or Cesarean Delivery. *JAMA Netw. Open* **2**, e1913650 (2019).
 96. Flasche, S., Lipsitch, M., Ojal, J. & Pinsent, A. Estimating the contribution of different age strata to vaccine serotype pneumococcal transmission in the pre vaccine era: a modelling study. *BMC Med.* **18**, 129 (2020).
 97. Weinberger, D. M., Pitzer, V. E., Regev-Yochay, G., Givon-Lavi, N. & Dagan, R. Association Between the Decline in Pneumococcal Disease in Unimmunized Adults and Vaccine-Derived Protection Against Colonization in Toddlers and Preschool-Aged Children. *Am. J. Epidemiol.* **188**, 160–168 (2019).
 98. Gierke, R., Wodi, P. & Kobayashi, M. *Pneumococcal Disease*. <https://www.cdc.gov/vaccines/pubs/pinkbook/pneumo.html> (2021).
 99. Simell, B. *et al.* The fundamental link between pneumococcal carriage and disease. *Expert Rev. Vaccines* **11**, 841–855 (2012).
 100. Marangu, D. & Zar, H. J. Childhood pneumonia in low-and-middle-income countries: An update. *Paediatr. Respir. Rev.* **32**, 3–9 (2019).
 101. Launes, C. *et al.* Viral coinfection in children less than five years old with invasive pneumococcal disease. *Pediatr. Infect. Dis. J.* **31**, 650–653 (2012).
 102. Wolter, N. *et al.* High nasopharyngeal pneumococcal density, increased by viral coinfection, is associated with invasive pneumococcal pneumonia. *J. Infect. Dis.* **210**, 1649–1657 (2014).
 103. Vu, H. T. T. *et al.* Association Between Nasopharyngeal Load of *Streptococcus pneumoniae*, Viral Coinfection, and Radiologically Confirmed Pneumonia in Vietnamese Children. *Pediatr. Infect. Dis. J.* **30**, 11–18 (2011).
 104. Watson, D. A., Musher, D. M. & Verhoef, J. Pneumococcal virulence factors and host immune responses to them. *Eur. J. Clin. Microbiol. Infect. Dis.* **14**, 479–490 (1995).

105. Weinberger, D. M. *et al.* Association of Serotype with Risk of Death Due to Pneumococcal Pneumonia: A Meta-Analysis. *Clin. Infect. Dis.* **51**, 692–699 (2010).
106. Weinberger, D. M. *et al.* Pneumococcal Capsular Polysaccharide Structure Predicts Serotype Prevalence. *PLoS Pathog.* **5**, e1000476 (2009).
107. Bentley, S. D. *et al.* Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. *PLoS Genet.* **2**, e31 (2006).
108. Mavroidi, A. *et al.* Genetic Relatedness of the *Streptococcus pneumoniae* Capsular Biosynthetic Loci. *J. Bacteriol.* **189**, 7841–7855 (2007).
109. Ganaie, F. *et al.* A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large *cps* Fragment from an Oral Streptococcus. *mBio* **11**, e00937-20, /mbio/11/3/mBio.00937-20.atom (2020).
110. Bentley, S. D. & Lo, S. W. Global genomic pathogen surveillance to inform vaccine strategies: a decade-long expedition in pneumococcal genomics. *Genome Med.* **13**, 84 (2021).
111. Balsells, E., Guillot, L., Nair, H. & Kyaw, M. H. Serotype distribution of *Streptococcus pneumoniae* causing invasive disease in children in the post-PCV era: A systematic review and meta-analysis. *PLoS ONE* **12**, e0177113 (2017).
112. Salter, S. J. *et al.* Variation at the capsule locus, *cps*, of mistyped and non-typable *Streptococcus pneumoniae* isolates. *Microbiol. Read. Engl.* **158**, 1560–1569 (2012).
113. Geno, K. A. *et al.* Pneumococcal Capsules and Their Types: Past, Present, and Future. *Clin. Microbiol. Rev.* **28**, 871–899 (2015).
114. Melchiorre, S. *et al.* Point mutations in *wchA* are responsible for the non-typability of two invasive *Streptococcus pneumoniae* isolates. *Microbiol. Read. Engl.* **158**, 338–344 (2012).
115. Sá-Leão, R. *et al.* Changes in pneumococcal serotypes and antibiotypes carried by vaccinated and unvaccinated day-care centre attendees in Portugal, a country with widespread use of the seven-valent pneumococcal conjugate vaccine. *Clin. Microbiol. Infect.* **15**, 1002–1007 (2009).

116. Hathaway, L. J., Meier, P. S., Bättig, P., Aebi, S. & Mühlemann, K. A Homologue of aliB Is Found in the Capsule Region of Nonencapsulated *Streptococcus pneumoniae*. *J. Bacteriol.* **186**, 3721–3729 (2004).
117. Martin, M. *et al.* An outbreak of conjunctivitis due to atypical *Streptococcus pneumoniae*. *N. Engl. J. Med.* **348**, 1112–1121 (2003).
118. Centers for Disease Control and Prevention. *Pneumococcal Vaccine Recommendations*. [https://www.cdc.gov/vaccines/vpd/pneumo/hcp/recommendations.html#:~:text=CDC%20recommends%20routine%20administration%20of%20pneumococcal%20conjugate%20vaccine%20\(PCV15%20or,of%20PPSV23%20one%20year%20later.](https://www.cdc.gov/vaccines/vpd/pneumo/hcp/recommendations.html#:~:text=CDC%20recommends%20routine%20administration%20of%20pneumococcal%20conjugate%20vaccine%20(PCV15%20or,of%20PPSV23%20one%20year%20later.) (2023).
119. Principi, N. & Esposito, S. Prevention of Community-Acquired Pneumonia with Available Pneumococcal Vaccines. *Int. J. Mol. Sci.* **18**, 1 (2016).
120. Poolman, J. T., Peeters, C. C. A. M. & van den Dobbelen, G. P. J. M. The history of pneumococcal conjugate vaccine development: dose selection. *Expert Rev. Vaccines* **12**, 1379–1394 (2013).
121. Grijalva, C. G. *et al.* Decline in pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. *Lancet Lond. Engl.* **369**, 1179–1186 (2007).
122. O'Brien, K. L. & Dagan, R. The potential indirect effect of conjugate pneumococcal vaccines. *Vaccine* **21**, 1815–1825 (2003).
123. Flasche, S. *et al.* Effect of pneumococcal conjugate vaccination on serotype-specific carriage and invasive disease in England: a cross-sectional study. *PLoS Med.* **8**, e1001017 (2011).
124. Jayasinghe, S. *et al.* Long-term Impact of a '3 + 0' Schedule for 7- and 13-Valent Pneumococcal Conjugate Vaccines on Invasive Pneumococcal Disease in Australia, 2002-2014. *Clin. Infect. Dis.* **64**, 175–183 (2017).
125. Steens, A., Caugant, D. A., Aaberge, I. S. & Vestheim, D. F. Decreased Carriage and Genetic Shifts in the *Streptococcus pneumoniae* Population After Changing the Seven-valent to the Thirteen-valent Pneumococcal Vaccine in Norway. *Pediatr. Infect. Dis. J.* **34**,

- 875–883 (2015).
126. van Zandvoort, K. *et al.* Pneumococcal conjugate vaccine use during humanitarian crises. *Vaccine* **37**, 6787–6792 (2019).
 127. Lo, S. W. *et al.* 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. *Lancet Infect. Dis.* **19**, 759–769 (2019).
 128. Golden, A. R. *et al.* Molecular characterization of predominant *Streptococcus pneumoniae* serotypes causing invasive infections in Canada: the SAVE study, 2011-15. *J. Antimicrob. Chemother.* **73**, vii20–vii31 (2018).
 129. Chochua, S. *et al.* Invasive Serotype 35B Pneumococci Including an Expanding Serotype Switch Lineage, United States, 2015-2016. *Emerg. Infect. Dis.* **23**, 922–930 (2017).
 130. Wyres, K. L. *et al.* Pneumococcal capsular switching: a historical perspective. *J. Infect. Dis.* **207**, 439–449 (2013).
 131. World Health Organization. *SAGE Working Group on Pneumococcal Conjugate Vaccines (December 2016 - December 2019)*. [https://www.who.int/groups/strategic-advisory-group-of-experts-on-immunization/working-groups\(2020\)](https://www.who.int/groups/strategic-advisory-group-of-experts-on-immunization/working-groups(2020)).
 132. Temple, B. *et al.* Evaluation of different infant vaccination schedules incorporating pneumococcal vaccination (The Vietnam Pneumococcal Project): protocol of a randomised controlled trial. *BMJ Open* **8**, e019795 (2018).
 133. Goldblatt, D. *et al.* Immunogenicity and boosting after a reduced number of doses of a pneumococcal conjugate vaccine in infants and toddlers. *Pediatr. Infect. Dis. J.* **25**, 312–319 (2006).
 134. Choi, Y. H., Andrews, N. & Miller, E. Estimated impact of revising the 13-valent pneumococcal conjugate vaccine schedule from 2+1 to 1+1 in England and Wales: A modelling study. *PLoS Med.* **16**, e1002845 (2019).
 135. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2009).
 136. Kim, J. O. & Weiser, J. N. Association of intrastrain phase variation in quantity of capsular

- polysaccharide and teichoic acid with the virulence of *Streptococcus pneumoniae*. *J. Infect. Dis.* **177**, 368–377 (1998).
137. Murad, C. *et al.* Pneumococcal carriage, density, and co-colonization dynamics: A longitudinal study in Indonesian infants. *Int. J. Infect. Dis.* **86**, 73–81 (2019).
138. Chaguza, C. *et al.* Carriage Dynamics of Pneumococcal Serotypes in Naturally Colonized Infants in a Rural African Setting During the First Year of Life. *Front. Pediatr.* **8**, 587730 (2021).
139. Everett, D. B. *et al.* Genetic Characterisation of Malawian Pneumococci Prior to the Roll-Out of the PCV13 Vaccine Using a High-Throughput Whole Genome Sequencing Approach. *PLoS ONE* **7**, e44250 (2012).
140. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).
141. Turner, P. *et al.* Improved Detection of Nasopharyngeal Cocolonization by Multiple Pneumococcal Serotypes by Use of Latex Agglutination or Molecular Serotyping by Microarray. *J. Clin. Microbiol.* **49**, 1784–1789 (2011).
142. Selva, L., del Amo, E., Brotons, P. & Muñoz-Almagro, C. Rapid and Easy Identification of Capsular Serotypes of *Streptococcus pneumoniae* by Use of Fragment Analysis by Automated Fluorescence-Based Capillary Electrophoresis. *J. Clin. Microbiol.* **50**, 3451–3457 (2012).
143. O'Brien, K. L. & Nohynek, H. Report from a WHO Working Group: standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*. *Pediatr. Infect. Dis. J.* **22**, e1-11 (2003).
144. Bronsdon, M. A. *et al.* Immunoblot Method To Detect *Streptococcus pneumoniae* and Identify Multiple Serotypes from Nasopharyngeal Secretions. *J. Clin. Microbiol.* **42**, 1596–1600 (2004).
145. Slotved, H.-C., Kalsoft, M., Skovsted, I. C., Kern, M. B. & Espersen, F. Simple, Rapid Latex Agglutination Test for Serotyping of Pneumococci (Pneumotest-Latex). *J. Clin. Microbiol.* **42**, 2518–2522 (2004).

146. Satzke, C. *et al.* The PneuCarriage Project: A Multi-Centre Comparative Study to Identify the Best Serotyping Methods for Examining Pneumococcal Carriage in Vaccine Evaluation Studies. *PLoS Med.* **12**, e1001903 (2015).
147. Tomita, Y. *et al.* A new microarray system to detect *Streptococcus pneumoniae* serotypes. *J. Biomed. Biotechnol.* **2011**, 352736 (2011).
148. Wang, Q. *et al.* Development of a DNA microarray to identify the *Streptococcus pneumoniae* serotypes contained in the 23-valent pneumococcal polysaccharide vaccine and closely related serotypes. *J. Microbiol. Methods* **68**, 128–136 (2007).
149. Newton, R., Hinds, J. & Wernisch, L. Empirical Bayesian models for analysing molecular serotyping microarrays. *BMC Bioinformatics* **12**, 88 (2011).
150. Epping, L. *et al.* SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data. *Microb. Genomics* **4**, 8 (2018).
151. Sheppard, C. L. *et al.* PneumoKITy: A fast, flexible, specific, and sensitive tool for *Streptococcus pneumoniae* serotype screening and mixed serotype detection from genome sequence data. *Microb. Genomics* **8**, 12 (2022).
152. Knight, J. R. *et al.* Determining the serotype composition of mixed samples of pneumococcus using whole genome sequencing. <http://biorxiv.org/lookup/doi/10.1101/741603> (2019) doi:10.1101/741603.
153. Kapatai, G. *et al.* Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ* **4**, e2477 (2016).
154. Rutter, H. *et al.* Visualising SARS-CoV-2 transmission routes and mitigations. *BMJ* **375**:e065312. <https://doi.org/10.1136/bmj-2021-065312>
155. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
156. Johansson, M. A. *et al.* SARS-CoV-2 Transmission From People Without COVID-19 Symptoms. *JAMA Netw. Open* **4**, e2035057 (2021).
157. Qiu, X. *et al.* The role of asymptomatic and pre-symptomatic infection in SARS-CoV-2

- transmission—a living systematic review. *Clin. Microbiol. Infect.* **27**, 511–519 (2021).
158. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020).
159. Seemann, A., Becker, U., He, L., Maria Hohnerlein, E. & Wilman, N. Protecting livelihoods in the COVID-19 crisis: A comparative analysis of European labour market and social policies. *Glob. Soc. Policy* **21**, 550–568 (2021).
160. Zou, L. *et al.* SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients. *N. Engl. J. Med.* **382**, 1177–1179 (2020).
161. Endo, A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]. *Wellcome Open Res.* **5**, 67 (2020).
162. Leclerc, Q. *et al.* What settings have been linked to SARS-CoV-2 transmission clusters? [version 2; peer review: 2 approved]. *Wellcome Open Res.* **5**, 83 (2020)
163. Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E. & Dean, N. E. Household Transmission of SARS-CoV-2: A Systematic Review and Meta-analysis. *JAMA Netw. Open* **3**, e2031756 (2020).
164. Viner, R. *et al.* Transmission of SARS-CoV-2 by children and young people in households and schools: a meta-analysis of population-based and contact-tracing studies. *J. Infect.* **83**, 3(2021) doi:10.1016/j.jinf.2021.12.026.
165. Peng, J. *et al.* Estimation of Secondary Household Attack Rates for Emergent Spike L452R Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants Detected by Genomic Surveillance at a Community-Based Testing Site in San Francisco. *Clin. Infect. Dis.* **74**, 1 (2021) doi:10.1093/cid/ciab283.
166. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
167. Vöhringer, H. S. *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).

168. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
169. San, J. E. *et al.* Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol.* **7**, veab041 (2021).
170. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10**, e66857 (2021).
171. Torjesen, I. Covid-19: Delta variant is now UK's most dominant strain and spreading through schools. *BMJ* 2021;373:n1445
172. Lopez Bernal, J. *et al.* Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *N. Engl. J. Med.* **385**, 585–594 (2021).
173. Clifford, S. *et al.* Effectiveness of BNT162b2 and ChAdOx1 against SARS-CoV-2 household transmission: a prospective cohort study in England. <http://medrxiv.org/lookup/doi/10.1101/2021.11.24.21266401> (2021)
doi:10.1101/2021.11.24.21266401.
174. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100 (2020).
175. Braun, K. M. *et al.* Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLoS Pathog.* **17**, e1009849 (2021).
176. Zwart, M. P. & Elena, S. F. Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annu. Rev. Virol.* **2**, 161–179 (2015).
177. Gutiérrez, S., Michalakis, Y. & Blanc, S. Virus population bottlenecks during within-host progression and host-to-host transmission. *Curr. Opin. Virol.* **2**, 546–555 (2012).

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1902896	Title	Miss
First Name(s)	Jada Nicole		
Surname/Family Name	Hackman		
Thesis Title	APPLICATION OF PATHOGEN GENOMICS TO INFER THE TRANSMISSION DIRECTION OF RESPIRATORY INFECTION		
Primary Supervisor	Stéphane Hué		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Wellcome Open Research
Please list the paper's authors in the intended authorship order:	Jada Hackman, Carmen Sheppard, Jody Phelan, William Jones-Warner, Ben Sobkowiak, Sonal Shah, David Litt, Norman K. Fry, Michiko Toizumi, Lay-Myint Yoshida, Martin Hibberd, Elizabeth Miller, Stefan Flasche,

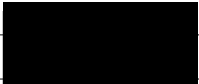
	Stéphane Hué,
Stage of publication	Submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed all of the bioinformatic analysis for this study, interpreted the results, and wrote/edited the manuscript for submission.
--	--

SECTION E

Student Signature	
Date	29/ May 2023

Supervisor Signature	
Date	31/05/23

CHAPTER 2: PHYLOGENETIC INFERENCE OF PNEUMOCOCCAL TRANSMISSION FROM CROSS-SECTIONAL DATA, A PILOT STUDY

Author List

Jada Hackman,^{1,6} Carmen Sheppard,² Jody Phelan,³ William Jones-Warner,³ Ben Sobkowiak,³ Sonal Shah,³ David Litt,² Norman K. Fry,^{2,4} Michiko Toizumi,^{5,6} Lay-Myint Yoshida,^{5,6} Martin Hibberd,³ Elizabeth Miller,¹ Stefan Flasche,^{1*} Stéphane Hué,^{1*}

Affiliations

¹ Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

² Vaccine Preventable Bacteria Section, UK Health Security Agency, London, United Kingdom

³ Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

⁴ Immunisation & Countermeasures Division, UK Health Security Agency, London, United Kingdom

⁵ Department of Paediatric Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan

⁶ School of Tropical Medicine and Global Health, Nagasaki, Japan

Corresponding Author

Jada Hackman, jada.hackman@lshtm.ac.uk

* Contributed equally

Keywords:

within-host diversity, phylogenetic, transmission direction, *Streptococcus pneumoniae*, pneumococcus

0. Abstract

Inference on pneumococcal transmission has mostly relied on longitudinal studies which are costly and resource intensive. Therefore, we conducted a pilot study to test the ability to infer who infected whom from cross-sectional pneumococcal sequences using phylogenetic inference.

Five suspected transmission pairs, for which there was epidemiological evidence of who infected whom were selected from a household study. For each pair, *Streptococcus pneumoniae* full genomes were sequenced from nasopharyngeal swabs collected on the same day. The within-host genetic diversity of the pneumococcal population was used to infer the transmission direction and then cross-validated with the direction suggested by the epidemiological records.

The pneumococcal genomes clustered into the five households from which the samples were taken. The proportion of concordantly inferred transmission direction generally increased with increasing minimum genome fragment size and single nucleotide polymorphisms. We observed a larger proportion of unique polymorphic sites in the source bacterial population compared to that of the recipient in four of the five pairs, as expected in the case of a transmission bottleneck. The only pair that did not exhibit this effect was also the pair that had consistent discordant transmission direction compared to the epidemiological records suggesting potential misdirection as a result of false-negative sampling.

This pilot provided support for further studies to test if the direction of pneumococcal transmission can be reliably inferred from cross-sectional samples if sequenced with sufficient depth and fragment length.

1. Introduction

Pneumococcal disease is a major contributor to global mortality amongst children less than five years old (O'Brien *et al.* 2009; Wahl *et al.* 2018). The main route of *Streptococcus pneumoniae* (*Sp*) transmission is through close physical interpersonal contact and exposure to contaminated respiratory secretion (van der Poll and Opal 2009; le Polain de Waroux *et al.* 2018; Neal *et al.* 2019). Children are the main reservoir for infection and transmission (Zivich *et al.* 2018; Weinberger *et al.* 2019; Flasche *et al.* 2020; Qian *et al.* 2022). Reduction of vaccine-type carriage via Pneumococcal Conjugate Vaccines enhances direct vaccine impact beyond the vaccinated children by mitigating onward spread (O'Brien and Dagan 2003; Grijalva *et al.* 2007; Poolman *et al.* 2013; Principi and Esposito 2016). With a more in-depth understanding of pneumococcal transmission, vaccination strategies may be further improved, but classical epidemiological approaches to understanding transmission rely on time and resource-intensive longitudinal studies.

Phylogenetic inference is particularly well suited for the exploration of infectious disease dynamics at the between-host and within-host level and may allow inference of transmission even from more easily collected cross-sectional infection surveys, including those for pneumococcal carriage (PANGEA Consortium and Rakai Health Sciences Program *et al.* 2019; Xu *et al.* 2020; Gouliouris *et al.* 2021). The phylogenetic analysis of pathogen genomes sampled from an infected population in principle not only allows the identification of transmission partners or clusters but also, the direction of transmission (who infected whom) (Rose *et al.* 2020; Zhang *et al.* 2020). These approaches have so far been mainly developed for and applied to study viral pathogens, particularly human immunodeficiency virus (HIV) and hepatitis C virus (Jacka *et al.* 2014; Hall *et al.* 2019; Leitner 2019; Rose *et al.* 2020; Street *et al.* 2020).

PhyloScanner is a phylogenetic algorithm that infers the direction of transmission from similarities in within-host pathogen diversity by reconstructing phylogenetic trees from deep sequencing data using a sliding window approach across the alignment thus each of the sliding window results in a tree. Until the development of PhyloScanner, most of the available tools lacked sufficient sensitivity to infer the direction of transmission due to limited use of the within-host genetic signal (Wymant *et al.* 2018). Moreover, PhyloScanner has been validated in the context of HIV direction of transmission with high concordance with the epidemiological records (Zhang *et al.* 2020).

Bacteria's large genome size, slow rates of evolution, and frequent horizontal gene transfer characteristics make the application of phylogenetic approaches more difficult for these organisms than for most viruses. The decrease in genetic diversity that accompanies the

transmission bottleneck limits the amount of genetic information that is detectable even further (Worby *et al.* 2014). A weak transmission bottleneck is needed to detect an adequate amount of within-host genetic diversity in both source and recipient to assess transmission linkage and its direction (Didelot *et al.* 2016). Despite these inherent limitations, the methodology applied to viral infectious diseases could still be applicable to bacterial infectious diseases.

This pilot study explored within-host pneumococcal bacterial diversity from whole-genome next-generation sequencing (NGS) data. We test and adapt currently available phylogenetic approaches to infer linked pneumococcal infections and their transmission direction from cross-sectional pneumococcal carriage data.

2. Results

2.1 Streptococcus pneumoniae study samples

The bacterial populations analysed in this study are from a prospective longitudinal household pneumococcal colonisation study (Hussain *et al.* 2005). The previous study enrolled and followed 121 families in monthly intervals for 10 consecutive visits. The carriage prevalence was 52% for children 0-2 years old, 45% for 3-4 years, 21% for 5-17 years, and 8% for ≥ 18 -year-old adults. A total of 10 transmission events across nine households met this study's inclusion criteria where there is epidemiological evidence to support a transmission event and its direction.

Across the nine households, 37 samples were connected to suspected transmission events and were thus sequenced. Among those, 5 pairs containing the same serotypes were available and thus included in the main direction of transmission analysis (same-visit samples). Moreover, there were 10 pairs containing the same serotypes that were collected one month apart that were included in the sensitivity analysis (subsequent-visit samples). There were 10 individuals that had swabs with the same serotype across consecutive visits and thus included in the within-host evolutionary rate estimation. Of these 10 individuals, five had up to three consecutive swabs while the others had up to two (Table 1).

2.2 Whole-genome sequencing and sequence quality control

The isolates were cultured and whole plate scrapes were processed for whole-genome sequencing using Illumina MiSeq. The mean sequencing coverage of the genomes was 112 reads per position (standard deviation (SD), 31 reads), with the lowest mean coverage of 26 reads per position (samples H3IAM3 and H9IBM3) and the highest of 337 reads per position (sample H1IBM2). Overall, 85.6% (± 9.7) of the raw reads matched with *S. pneumoniae* genomic positions (range, 33.1% (H9IBM3) - 93.0% (H9IAM2)), and unmatched reads were filtered out for the downstream analysis (Supplemental Table 1).

2.3 Serotyping

Of the 37 samples, 29 were previously serotyped using DNA microarray, while the remaining 8 were serotyped using the Quellung reaction. For quality assurance, the isolates were then serotyped from the raw NGS reads using SeroBA genomic serotyping tool. The sequence-based serotype assignments were concordant with microarray serotyping, except for three of the 37 samples. Samples H9IAM2 and H9IBM3 were both originally identified as serotype 6A using the Quellung reaction but as 6C in the genomic serotyping. This was due to the reclassification of sub-lineages of serotype 6A to 6C subsequent to the original serotyping (In *et al.* 2007). Furthermore, all three consecutive-visit samples from individual H2IB were classified as serotype 23F according to the microarray typing, however, sequence-based methods determined swab H2IBM6 as serotype 6B while H2IBM5 and H2IBM7 were concordant with the microarray data. Since we could not exclude the possibility that this discrepancy was the result of a sample mix-up, isolate H2IBM6 was excluded from the analysis but the subsequent-visit samples from H2IB were still included in the sensitivity analysis (Figure 1A).

2.4 Multiple-carriage's role in phylogenetic tree reconstruction

Samples were tested for the presence of multiple distinct pneumococcal populations. Clusters of single nucleotide polymorphism (SNP) frequencies below 100% were indicative of the presence of multiple pneumococcal haplotypes. Sample H4IBM7 demonstrated two SNP clusters, one at 20% and the other at 80% which were designated as the minor and major strain, respectively. The reads from both strains were separated using a SNP frequency cut-off of 50%. The major strain from H4IBM7 was genetically more similar to the linked isolate H4IAM8 (distance 0.11 nuc sub/site) compared to the minor strain to H4IBM7 (distance 0.44 nuc sub/site) (Supplemental Figure 1).

2.5 Putative transmission pairs identified with consensus SNP phylogenetic reconstruction

A maximum-likelihood (ML) phylogeny of the five putative transmission pairs was reconstructed from the consensus SNP sequences of the respective cross-sectional samples. The tree confirmed the clustering of isolate pairs that belonged to the same serotypes and were collected from the same households (Figure 1A). The average genetic distance between the putative source-recipients pairs was 0.045 nuc sub/site (range, 0.038-0.057 nuc sub/site).

Consecutive-visit swabs from the same individuals were also included in the consensus SNP tree reconstruction for households 3 and 8. The phylogeny revealed there was an insufficient phylogenetic signal to distinguish samples collected from the same individual a month apart

compared to samples collected cross-sectionally from transmission pairs within a month after the transmission event (Figure 1B).

In the sensitivity analysis, we reconstructed a tree using consensus SNP sequences from likely transmission pairs but taken at subsequent visits e.g. one month apart (N=10 pairs). Of the 10 putative pairs, 9 pairs (90%) clustered concordantly with the epidemiological data with $\geq 90\%$ bootstrap support. Amongst those clustered pairs, eight demonstrated short genetic distances (≤ 0.10 nuc sub/site) except for pair H4IBM7_major and H4IAM8 which could be due to the imperfect haplotype reconstruction and for H6IAM6 and H6IBM5 we found >0.10 nuc sub/site difference between the two isolates suggesting a potential indirect transmission event (Supplemental Figure 3).

2.6 Direction of transmission using within-host genomic variation

The direction of transmission was inferred from the five pairs of same-visit samples using Phyloscanner, a tool that implements a sliding window approach across the genomes and reconstructs sub-trees using the reads present in a given window. For each sub-trees reconstructed, the source of the infection is determined through a modified maximum-parsimony ancestral state reconstruction inference, where the most likely identity of the pair member is inferred at each node.

We conducted a total of 200 inferences of the direction of transmission conducted across the five transmission pairs. The inferences were generated from a combination of varying sliding window sizes (50, 75, 100, 125, 150 bp) and varying minimum number of SNP (1,3,5,7,9,11,13,15 SNP) in the sub-tree reconstruction as these parameters would most likely affect the phylogenetic signal. Sub-trees were filtered for a minimum of two reads per individual and a clear ancestral state assignment to one of the individuals. This resulted in 102 inferences (51.5%) being viable to infer the direction of transmission. As expected, increasing either the minimum number of SNP threshold or the window size decreased the number of sub-trees included in the inference (Figure 2 and Supplemental Figure 4).

For small window size and a low SNP threshold, concordance with the epidemiologically inferred direction of transmission was two to three out of the five pairs, with 50% being the expected concordance if inference was no better than random chance. The proportion of pairs in which the direction of transmission was inferred in concordance with the epidemiological records generally increased with larger window sizes and/or more SNP. At least four out of the five inferred directions of transmission were concordant if using sliding window sizes of

125 bp, however, no analyses with further increased window size were possible due to the lack of samples with sufficient read lengths in the present sequencing approach (Figure 2).

Increasing the sliding window size and/or minimum number of SNP resulted in a higher level of concordance with the epidemiological evidence in the directionality inferred for pairs H3IBM7 and H3ICM7; and H7IAM10 and H7IBM10. Pairs H3IBM3 and H3IAM3; and H1IBM3 and H1IAM3 demonstrated consistently concordant directionality independent of window size and/or minimum number of SNP. Conversely, the pair H8IAM3 and H8IBM3 demonstrated consistent discordant directionality (Figure 3A).

In a sensitivity analysis using a different reference genome, serotype 23F, the findings were qualitatively similar in the direction of transmission analysis with subsequent-visit sample pairs, albeit the association was less apparent (Supplemental Figure 4).

2.7 Within-host diversities of source-recipient pairs

The proportion of unique SNP in the source-recipient pairs when sampled during the same visit was used as a proxy for the presence of a transmission bottleneck effect; expecting the source to have had more time to evolve before transmitting a subset of the acquired within-host heterogeneity and thus presenting more unique SNP than the recipient.

The average number of polymorphic sites between the source and recipient of a pair was 11,975 per transmission pair (SD, \pm 1067). The source and recipient of all five same-visit pairs shared a large proportion of SNP (mean 91.6%; SD, \pm 8.6%). The source of infection as determined by the epidemiological records had a higher proportion of unique polymorphic sites compared to the recipient for 4 of the 5 pairs; 7.3% vs 1.1% (range of unique SNP source vs recipient, 0.7%-22.6% vs 0.3%-2.7%). The only pair where the putative sources had a smaller proportion of unique polymorphic sites was H8IAM3 (source) and H8IBM3 (recipient); the pair was found to consistently suggest a direction of transmission discordant to the epidemiological records (Figure 3B, 3C).

The direction of transmission inferred by the larger number of unique SNP was compared to that inferred by Phyloscanner. Pair H3IBM3 and H3IAM3 had the largest difference in the proportion of unique SNP as previously mentioned, while pair H1IBM3 (source) and H1IAM3 (recipient) had a relatively moderate difference with 4.7% and 0.03% unique SNP, respectively. Both of these pairs had a consistent concordant transmission direction across all permutations of window sizes and a minimum number of SNP. Conversely, pair H3IBM7 (source) and H3ICM7 (recipient) had the smallest differences in the proportion of unique SNP

and mixed inferences. Further, pair H7IAM10 (source) and H7IBM10 (recipient) had relatively large differences in the proportion of unique SNP and also had mixed inferences (Figure 3B). Interestingly, the only pair that exhibited a larger proportion of unique SNP in the recipient compared to the source, H8IAM3 (source) and H8IBM3 (recipient), had a consistent discordant directionality despite an increase in window sizes or minimum number of SNP (Figure 3).

2.8 Estimation of the within-host rate of nucleotide substitution

The within-host rate of nucleotide substitution for *S. pneumoniae* was 65 SNP/month (range, 15-1539 SNP) and the within-host evolutionary rate 1.8E-5 nucleotide substitutions/site/year (range, 6.0E-5, 1.7E-6) (Figure 4).

3. Discussion

In this study, a genomic approach was used to infer the direction of *S. pneumoniae* transmission and cross-validated with the direction of transmission inferred from epidemiological evidence. We found that linkage was concordantly identified from reconstructed phylogenies in all five of the same-visit pairs and nine of the ten subsequent-visit pairs. Albeit, the phylogenetic linkage of the same-visit pairs may in part be attributable to the serotype heterogeneity. To address this, paired isolates from subsequent months were assessed where there is more serotype homogeneity and more transmission pairs and the phylogenetic reconstruction revealed distinguishable linkage in addition to the indistinguishable linkage of pairs within their respective serotypes. The indistinguishable linked pairs within a serotype cluster could be due to the difference in sampling time between the two consecutive months which could contribute to genetic drift and the accumulation of variation in the recipient of the infection. These results imply that linked pneumococcal infection is identifiable from genomic data alone, however, more stringent phylogenetic criteria e.g. more conservative bootstrap cut-off or larger intra-cluster genetic distance thresholds might have to be placed in settings where there is more serotype homogeneity and less population diversity.

The two parameters that were likely to affect the probability of identifying the concordant source-recipient relationship within a transmission pair using a sliding-window phylogenetic approach were the sliding window sizes and the minimum number of SNP present within those windows. These two parameters indeed impact the phylogenetic signal of the read alignment used to reconstruct the sub-trees, e.g. on the capacity to reconstruct a robust phylogeny from which conclusions can be drawn with sufficient statistical certainty. Under optimal conditions, the direction of transmission was concordant between the epidemiological records and

phylogenetic inference for all five same-visit transmission pairs with a window size of 125 bp and a minimum number of 3 SNP. Moreover, based solely on the genomic data, the phylogenetic inference and the transmission bottleneck analysis were concordant in all five same-visit pairs.

In general, these results suggest an increased concordant direction inferred from combinations of longer window sizes and a larger minimum number of SNP, however, the sample size and the maximum window size were too low to allow a definitive conclusion. Hence further studies are needed to determine whether higher coverage and/or read lengths can increase the phylogenetic signal for inferring the direction of transmission. More sequencing coverage would increase the phylogenetic genetic signal by detecting minor variations between source-recipient pairs while the longer reads would aid in the genome assembly and thus provide more robust genomes.

To our knowledge, the only studies that have attempted to validate genomic approaches against epidemiological data on the direction of transmission were using HIV transmission pairs (Rose *et al.* 2020; Zhang *et al.* 2020; Villabona-Arenas *et al.* 2022). Villabona-Arenas *et al.* investigated the phylogenetic inference of known transmission direction of HIV-1 transmission partners. They observed an increase in correct transmission direction up to 93% when inferring from paraphyletic-monophyletic tree topology highlighting the importance of sufficient intra-host diversity to distinguish HIV-1 populations amongst partners (Villabona-Arenas *et al.* 2022). Rose *et al.* looked at HIV transmission partners where the accuracy of transmission direction was inferred concordantly for 55%-74% of the pairs and the range was dependent on the sequencing and inference methods used (Rose *et al.* 2020). While a more recent study from Zhang *et al.*, using the same cohort as Rose *et al.*, increased the accuracy up to 93.3% (Zhang *et al.* 2020). Zhang *et al.* speculated the higher accuracy for inferring transmission direction compared could be attributable to higher sequencing coverage in addition to the longer sequencing reads up to 400 bp. Zhang *et al.* also used Phyloscanner for their analysis and similarly explored the impact of varying window sizes across the entire HIV genome. They reconstructed sub-trees between 280 - 400 bp in 20 bp increments and observed higher accuracy using larger window sizes. This prompts further investigation to assess if increased coverage and/or sequencing reads would also increase phylogenetic signal in bacterial pathogen transmission.

The evolutionary rate of bacteria is relatively slow compared to fast-evolving RNA viruses such as HIV where bacteria evolve between 10^{-7} to 10^{-5} substitutions/site/year and amongst the fastest evolving pathogens, between 10^{-4} to 10^{-3} substitutions/site/year (Didelot *et al.* 2016).

The relatively slower evolutionary rate of bacteria to viruses substantially affects the number of accumulated mutations, therefore, the number of genetic fingerprints to link transmission pairs and their direction.

The comparison of within-host bacterial diversity within the transmission pairs showed evidence of a transmission bottleneck of varying strengths, with a higher percentage of unique SNP in the source's bacterial population compared to the recipient's in 4 of 5 of the studied pairs implying the direction of transmission according to the epidemiological records could be incorrect which could be explained by false negative sampling (Thindwa *et al.* 2021). This directed reduction of diversity could aid in determining the direction of transmission when the latter is not known.

Hall *et al.* used a similar approach to investigate the transmission direction of Methicillin-resistant *Staphylococcus aureus* (MRSA), in a high-transmission setting (Hall *et al.* 2019). They observed varying transmission bottleneck strengths among their source-recipient pairs. The bottleneck strength ranged from strong where a single lineage was transmitted from the source to the recipient to weak where the transmission pairs shared multiple lineages, however, the direction was ambiguous. In conjunction with our study, this suggests the presence of a transmission bottleneck for bacteria, however, the strength of the bottlenecks is not associated with a higher probability of inferring the concordant direction of transmission. In other words, while we observed more unique SNP in the source of the infection compared to the recipient, a larger proportion of unique SNP in the source compared to the recipient is not associated with higher chances of inferring the concordant direction. These results imply that the observed bottleneck effect is not random and a comparison of the number of unique SNP in the members of a suspected transmission pair can aid in supporting the direction of transmission inferences, under the assumption that the recipient will be the individuals with the bacterial population exhibiting the least number of unique SNP.

The inclusion of additional longitudinal samples from the same individual, sampled over a couple of months, confounded the ability to detect true transmission pairs. This suggests that there is relatively little within-host diversity within that time frame to distinguish transmission pairs from within-host samples. The evolutionary rate that was extrapolated from the SNP accumulated over time is relatively small and there would be less diversity accumulated especially when looking at a 1-month or even 2-month sampling time difference. The within-host evolutionary rate for *S. pneumoniae* that we estimated is similar to the estimates by Chaguzza *et al.* who looked at the natural colonisation of longitudinal samples with estimates around 10^{-5} substitutions/site/year for most serotypes and as low as 10^{-6} substitutions/site/year

for serotype 19A (Chaguza *et al.* 2020). Moreover, the rates are dependent upon the carrier, serotype, and colonisation episodes, suggesting the importance of the host-microbe interaction during the evolution of pneumococcus.

Rather than longitudinal within-host diversity, Hall *et al.* looked at within-host MRSA diversity between samples from different body sites and similarly saw no evidence for decreased or increased genetic diversity between the within-host samples. Other studies, in the context of *Clostridioides difficile* and slow-evolving bacteria such as *Mycobacterium tuberculosis*, observed difficulty capturing within-host level diversity from whole-genome sequences (Martin *et al.* 2018; Balaji *et al.* 2019). As expected, the within-host diversity of bacteria is difficult to capture, especially in the absence of relatively high-coverage sequencing data. While most pneumococcal infections are dominated by a major serotype, there are settings of mixed high carriage rates, and being able to capture the within-host diversity is crucial for understanding transmission dynamics (Kamng'ona *et al.* 2015).

The transmission directions that were phylogenetically inferred and discordant with the epidemiological records could be attributable to multiple factors and inherent limitations of the studies. The first is the imperfect sensitivity of the swab collection in combination with the imperfect sensitivity of the culturing technique to detect pneumococci and identify the dominant serotype. Pneumococcal testing has been previously reported with 85% sensitivity (95% CI, 73%-94%) which would result in up to 15% false-negative tests (Abdullahi *et al.* 2007; Thindwa *et al.* 2021). With false-negative testing, a carriage episode could have been missed and thus led to a different interpretation of transmission direction based on the epidemiological data on the sequence of pneumococcal positivity within the households.

The second includes potential unsampled intermediary transmission partners that were not included in the study. Since the transmission is predominantly through close contact and within households, it is unlikely an individual outside of the household is introduced to the transmission chain. However, the possibility of an unsampled person within the link cannot be discarded. If there was an intermediary individual within the chain between the time of sampling of the source and recipient pairs, then the directionality would be more difficult to determine due to the decreased mutation similarities between the source and recipient.

The third factor includes the phylogenetic uncertainty that is limited by the short-read fragments. An increase in read lengths would result in improved genome assembly and therefore increased genomic signal (Mantere *et al.* 2019). Other sequencing methods such as PacBio can yield longer read lengths, up to 10 kbp, and should be further investigated and

assessed if improved genome assemblies improve phylogenetic inference in assessing the directionality of transmission.

In summary, in this pilot study we find evidence that conventional NGS may offer too little phylogenetic signal to allow robust inference for the direction of transmission for cross-sectionally sampled pairs of pneumococcal carriage, but that with increased sequencing depth and particular fragment size, such inference may be possible. This motivates further studies to explore the feasibility and limits of inference of who infected whom with pneumococci from genomic data.

4. Materials and Methods

4.1 Study design and study samples

This study cohort was from a prospective, longitudinal household study of pneumococcal colonisation conducted in the county of Hertfordshire, United Kingdom in 2001-2002. The original study is described in detail elsewhere (Hussain *et al.* 2005). In summary, preschool children and their household contacts were enrolled and followed up monthly for 10 consecutive months. At each visit, nasopharyngeal swabs were collected and any *S. pneumoniae* bacteria isolated by culture were serotyped using DNA microarray or the Quellung reaction to identify carriage type (Southern *et al.* 2018).

A total of 10 within-household putative source-recipient transmission events were included based on the following inclusion criteria which were also the epidemiological evidence supporting a transmission event and its direction: (i) the recipient is tested positive for carrying a single pneumococcal serotype, (ii) the potential source of infection is an individual within the same household who was carrying the same serotype in the month before the recipient was tested positive, and (iii) in the two visits prior to the carriage episode of the recipient, the remainder of the household were found to not carry pneumococci of the same serotype (Table 1).

The epidemiological inclusion criteria aimed to maximise the probability of correctly identifying a transmission pair. In five instances, the source also carried pneumococci of the same serotype on the following visit resembling cross-sectional sampling of source and recipient. These five same-visit paired samples were used for the main direction of transmission analysis. We defined the sample ID in the following format: household (H), individual ID (I), and the month the swab was collected from (M); e.g. sample H1IAM1 was collected from household 1, individual A, from month 1 of the study.

The selected study samples were included in two different analyses:

(i) The main analysis tested the direction of transmission and included same-visit swabs of putative transmission pairs (N=10 individuals), simulating a cross-sectional carriage survey. Of the five pairs where same-visit samples were available, two pairs had a second same-serotype same-visit instance to assess within-host diversity (Table 1). The same-visit samples were additionally used to estimate the proportion of unique SNP in the source-recipient pairs. Alongside this, the 10 pairs (N=20 individuals) where samples of source and recipient were taken from subsequent visits (one month apart) were used to also test the direction of transmission to assess the sensitivity of the method on more temporally distant samples. (ii) The second analysis was to estimate the within-host evolutionary rate from 10 individuals who had at least 2 consecutive swabs of the same serotype (N=25 sequences) (Table 1).

4.2 Isolate culturing and whole-genome sequencing

Isolates were grown overnight on horse blood agar with 5% CO₂. The isolates used were from stock cultures stored at -80°C in glycerol blood broth medium since 2001/02. The stocks used were pneumococcal isolates obtained from the culture plates directly inoculated with the swab in the original study. Samples from the glycerol blood broths were partially thawed when plated and DNA was extracted until a minimum concentration of 20 ng/uL (Kapatai *et al.* 2016).

Whole-genome sequencing was carried out on the Illumina MiSeq platform on the DNA extracts. Library preparation was done using QIAseq FX DNA Library Kit (96 – Cat no:180475) as per the manufacturer's protocol yielding a DNA fragment size of 300 bp, including adaptors. Sequencing was completed using the Illumina MiSeq in conjunction with the MiSeq Reagent Kit v2 (300-cycles – Cat no: MS-102-2002). The sequencing was run in duplicates and were later merged. Adaptors were removed from the raw sequencing data using Trimmomatic v0.39, along with low-quality reads based on an average quality and sliding window approach (Bolger *et al.* 2014). Additional quality control of the reads was carried out with Kraken2 v2.0.9 and unmatched *S. pneumoniae* reads were filtered out from the downstream analysis (Wood *et al.* 2019).

4.3 Genomic serotyping

Genomic serotyping of the isolates was carried out on the *S. pneumoniae* sequencing reads using SeroBA v1.0.1, a tool that predicts pneumococcal serotypes using a k-mer-based approach from raw fastq data (Epping *et al.* 2018). Then the reads were aligned to a reference genome strain KK0981 (serotype 3, GenBank accession number AP01797) with the Burrow-Wheeler Alignment (BWA-MEM) and SAMtools mpileup software (Li *et al.* 2009; Li 2013; Chiba *et al.* 2017). Variant calling format files (VCF) containing information on SNP were

generated using FreeBayes v1.3.2 (Garrison and Marth 2012) A consensus sequence of all polymorphic positions was generated for each of the isolates which were then included in the phylogenetic reconstruction to identify linkage.

4.4 Multi-carriage detection

We tested samples for multiple pneumococcal populations by assessing the distribution of SNP frequencies in all of the samples using LoFreq, a sensitive-variant calling tool (Wilm *et al.* 2012). The presence of more than one cluster or peak of SNP was considered as evidence for carriage of multiple haplotypes, under the assumption that clusters of SNPs are associated with common polymorphic sites within the reads (Supplemental Figure 1).

4.5 Phylogenetic reconstruction of putative transmission pairs

The phylogenies of the sequenced bacterial genomes were reconstructed by maximum-likelihood inference using RAxML v2.0.2, under the General Time Reversible model of nucleotide substitutions and with 1,000 bootstrap replicates from the alignment of consensus single nucleotide polymorphisms (Stamatakis 2014). Transmission pairs were identified from the consensus SNP tree topology as clusters of sequences (≤ 0.10 nuc sub/site) with branch support $\geq 90\%$.

4.6 Inference of transmission direction

The most likely direction of transmission within a transmission pair was inferred using PhyloScanner v1.4.7 (Wymant *et al.* 2018). Each phylogeny inferred from PhyloScanner is classified as one of the following three relationships (i) single ancestry, where the subgraphs from the two populations form a paraphyletic (source) - monophyletic (recipient) relationship, (ii) equivocal, where the source and recipient subgraphs form dual monophyletic groups and thus the direction of infection is unclear, and (iii) complex ancestry, where the subgraphs form paraphyletic - paraphyletic groups and where the ancestral state is assigned to both the source and recipient depending on the subgraph (Chiba *et al.* 2017). The sub-trees, relationships identified with reads within a restricted sliding window, are then aggregated and the one that occurs the most often was considered to be the most likely scenario for the pair of individuals analysed. See Wymant *et al.* for more details on the methods implemented (Wymant *et al.* 2018).

Given the size of the pneumococcal genome analysed, approximately 2.1 million bp, and its low mutation rate, we restricted PhyloScanner to only process those windows that contained a predefined minimum number of SNP across the reads, to increase phylogenetic signal, and tested a range of window sizes. In addition, sub-trees (i) that had less than 2 tips from each

host and (ii) where sequences from both hosts were equidistant from the reference sequence used as an outgroup were excluded to further enhance the accuracy of the inference (Supplemental Figure 2). This approach was used for the inference of transmission direction from both the same-visit and the subsequent-visit pairs.

As a sensitivity analysis to test the presence of bias in the inference in direction, the Phyloscanner analysis was carried out using reference strain ATCC700669 (serotype 23F, GenBank accession: NC_011900) as the mapping genome (Croucher *et al.* 2009).

4.7 Identifying unique SNP among source-recipient pairs

The count and proportion of unique SNP detected in both members of a suspected transmission pair were estimated from the VCF files containing polymorphic sites mapped to the reference genome. The average percent of unique SNP in each individual was reported with standard deviation. The percent of unique SNP was compared between the putative source and recipient of each pair, using 95% confidence intervals (95% CI) and a two-tailed t-test.

4.8 Comparison of Within-host Diversities

S. pneumoniae within-host rate of nucleotide substitution, expressed as the number of nucleotide substitution/site/year, was estimated from the number of unique polymorphic sites accumulated between consecutive pneumococcal isolates from the same individual using the same methods as the proportion of unique SNP in recipient-source pairs.

Supplemental Materials

Supplemental materials will be made available online.

Acknowledgements

We thank the participants who made this study possible. We thank the nurses who recruited and followed up on the participants and the staff who isolated, characterised, and archived the swabs. We also thank Oliver Ratmann for his invaluable feedback for the data analysis.

Funding

This study was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the WISE scheme and EU grant QL4-CT-2000- 00640. SF is funded by a Sir Henry Dale Fellowship through the Wellcome Trust and the Royal Society (208812/Z/17/Z). EM receives support from the National Institute for Health Research (NIHR) Health Protection Research Unit in Immunisation at the London School of Hygiene and Tropical Medicine in partnership with the UKHSA (Grant Reference NIHR200929). The funders had no input in the study design, data analysis, or manuscript draft.

Author Contributions

S.H, S.F, E.M, M.H, N.K.F, and D.L designed the project.

S.S, C.S, and W.J organised, cultured, and sequenced the samples, respectively.

J.P and B.S provided bioinformatic support throughout the data analysis.

L.Y and M.T provided feedback and direction on the data analysis.

J.H carried out the data analysis and drafted the manuscript.

All authors read, edited, and approved the article.

Data Availability

The whole-genome sequencing data has been made available for download on the European Nucleotide Archive under study accession “PRJEB60532”. The specific data for this study are available from the authors upon request and subject to a data-sharing agreement.

Conflict of Interest

C.L.S., D.J.L. and N.K.F. received grant funding from Pfizer and GSK for investigator-led research projects into carriage and disease caused by *S. pneumoniae* in England.

Ethics

Ethical approval for the study was granted by LSHTM's institutional ethics board (reference number: 17642).

TABLES AND FIGURES

Table 1. Samples were selected for inference of the direction of transmission of *S. pneumoniae* and within-host diversity.

Household (H)	Individual (I)	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10
1	A			6B*	6B	6B	+	+	+		
1	B	6B	6B*	6B	+	+	+	+	+		
2	A						23F*				
2	B					23F*	23F**	23F**			
3	A			23F*					+		NA
3	B		23F*	23F	+	+	6A*	6A	6A	+	+
3	C	+						6A*	6A		
4	A								22F*	NA	
4	B		+		+		+	22F*		NA	
5	A			23F*	+				+		
5	B		23F*	+		NA	NA	NA	NA	NA	NA
6	A						6B*	6B	+	+	+
6	B				6B	6B*			NA	NA	
7	A	+		+						14*	14
7	B			+			+				14*
8	A		19F*	19F	19F	+		+	+	+	+
8	B			19F*	19F	19F	+			+	+
9	A	+	6A*	NA		NA	NA		NA	NA	NA
9	B			6A*						NA	NA

Green highlights paired same-visit samples used to infer the direction of transmission

Box line highlights consecutive-visit samples used to estimate the within-host evolutionary rate

* Paired subsequent-visit samples used for the sensitivity analysis

** Discordant serotyping between epidemiological data and genomic serotyping data for individual H2IB at month 6

+ A positive nasal swab for pneumococci, but samples were not included in the analysis because they did not satisfy the epidemiological inclusion criteria

Empty cell, a negative test for pneumococcal carriage

“NA”, samples that were not obtained in the respective month

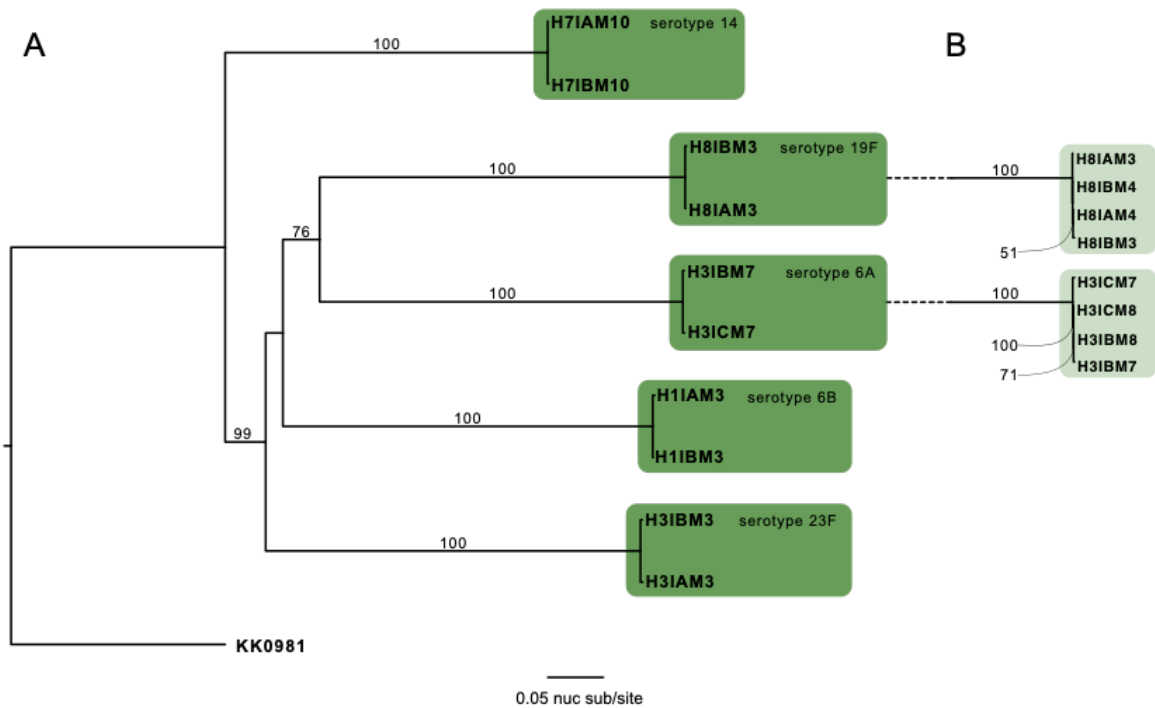


Figure 1. Maximum-likelihood phylogeny of the 10 *S. pneumoniae* genomes from five same-visit putative transmission pairs rooted to the reference genome, KK0981. (A) The consensus SNP tree was reconstructed from an alignment of polymorphic sites along the genomes (42,499 base pairs). Branch supports $\geq 50\%$, as determined by 1,000 bootstrap replicates, are denoted on the relevant branches. Branch length represents nucleotide substitutions per site (nuc sub/site), as denoted by the scaled bar. Clusters of two sequences supported by a bootstrap score $\geq 90\%$ were considered as putative transmission pairs and are highlighted by the dark green boxes. (B) An additional same-visit transmission pair was included from household 3 (H3IBM8 and H3ICM8) and household 8 (H8IAM4 & H8IBM4). The light green boxes highlight the intermingling of transmission pairs with their respective within-host longitudinal swabs.

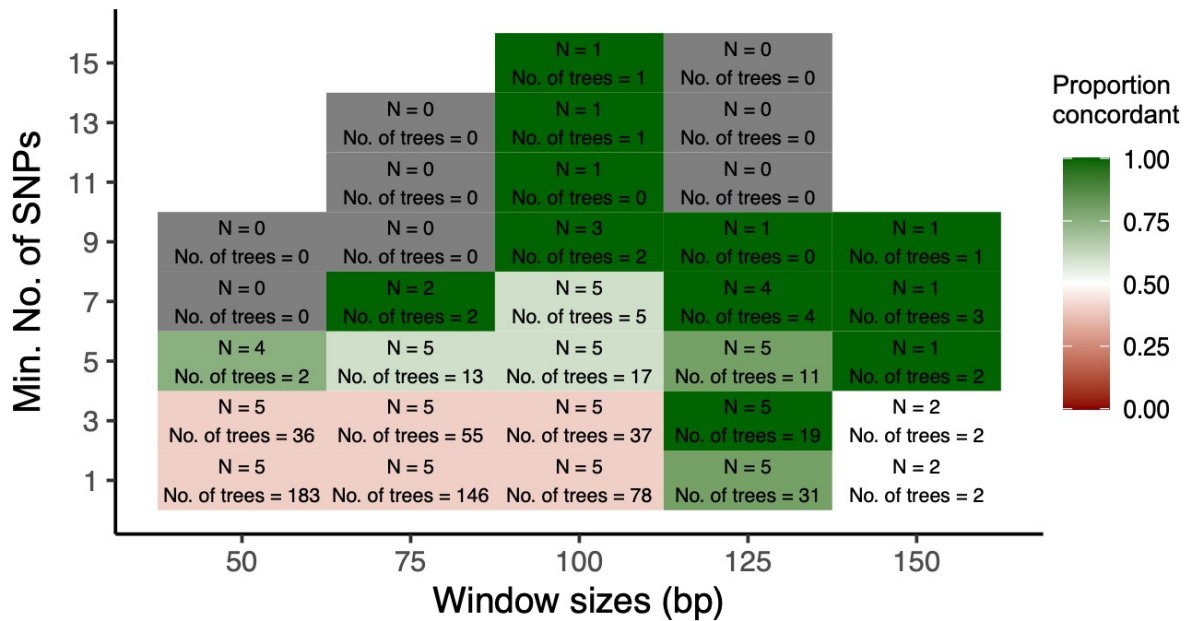


Figure 2. The proportion of concordant directionality with the epidemiological data inferred per minimum number of SNP per read (1,3,5,7,9,11,13, or 15 SNP) and read window sizes (50, 75, 100, 125, or 150 bp). (A) Inference from samples collected during the same visit. Green and red-coloured boxes denote the proportion of pairs for which the inferred transmission direction was concordant with the epidemiological data, green is equivalent to 100% and red is equivalent to 0%. White boxes denote equal distributions of concordant and discordant inferred directions (proportion = 0.50). While grey boxes denote that phylogenies were generated, however, they were classified as “unlinked” or “ambiguous directions” and empty boxes denote that no sub-trees were generated for this combination of window size and SNP. The “N” represents the number of pairs analysed for the respective window size and SNP combination and the “N of Trees” is the average number of sub-trees used for the direction of transmission for those pairs analysed.

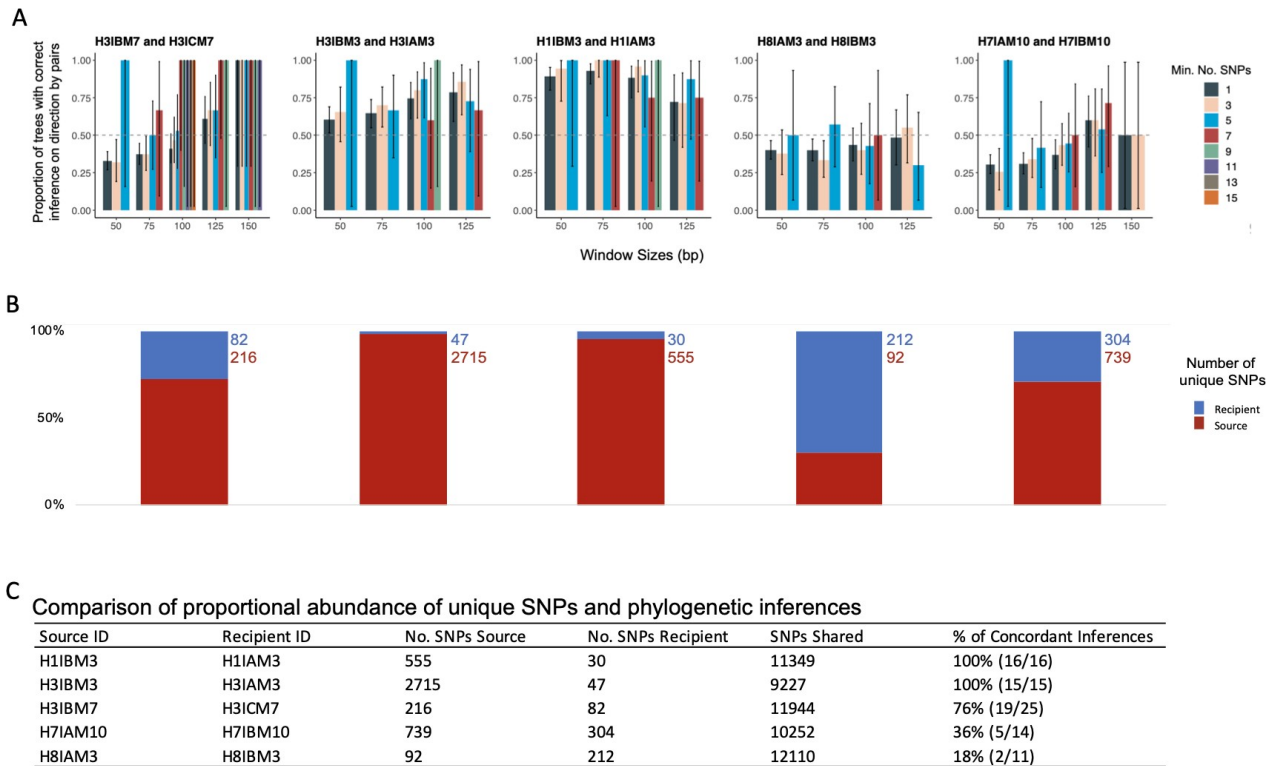


Figure 3. (A) The proportion of sub-trees concordant with the epidemiological data, for each pair, with the different combinations of window sizes and minimum number of SNP represented by the coloured bars. (B) Proportional abundances of unique SNP in source-recipient pairs. The proportional abundances are observed in source and recipients with the red bar denoting the percentage of unique SNP from the suspected source of infection, while the blue bar is the recipient (C) The raw number of unique SNP detected for the source, recipient, and variants that are shared. The % of concordant inferences represents the number of inferences (combinations of a minimum number of SNP and window sizes) that were analysed and concordant with the epidemiological data.

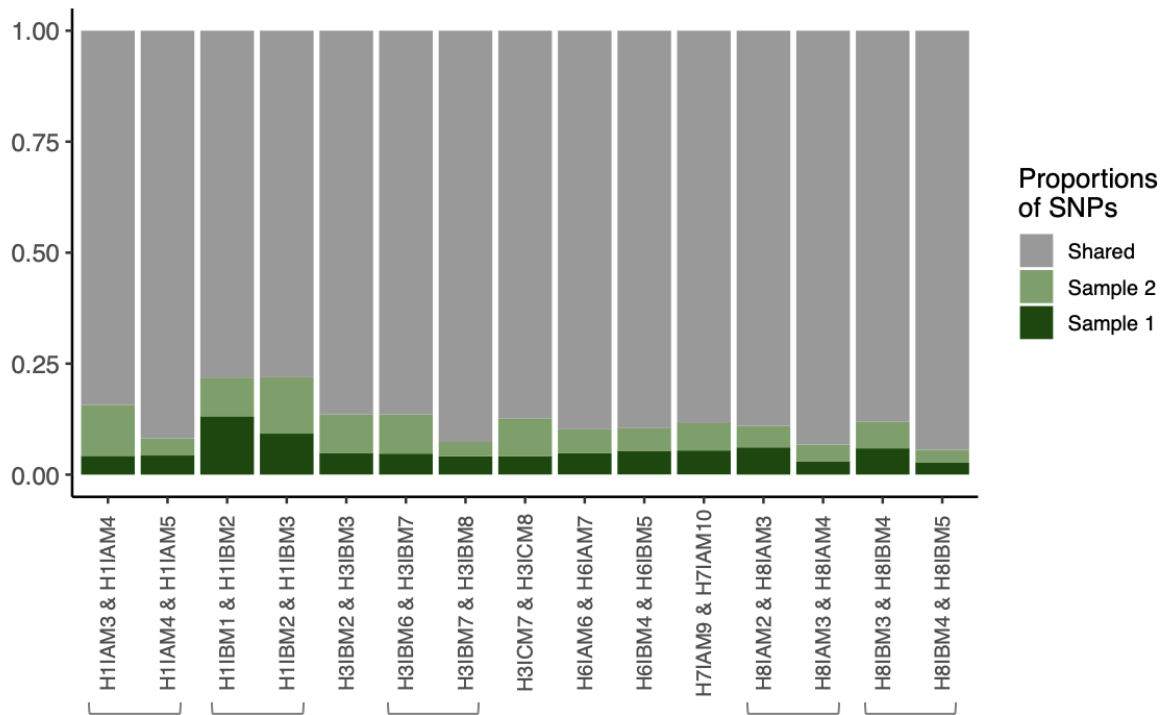
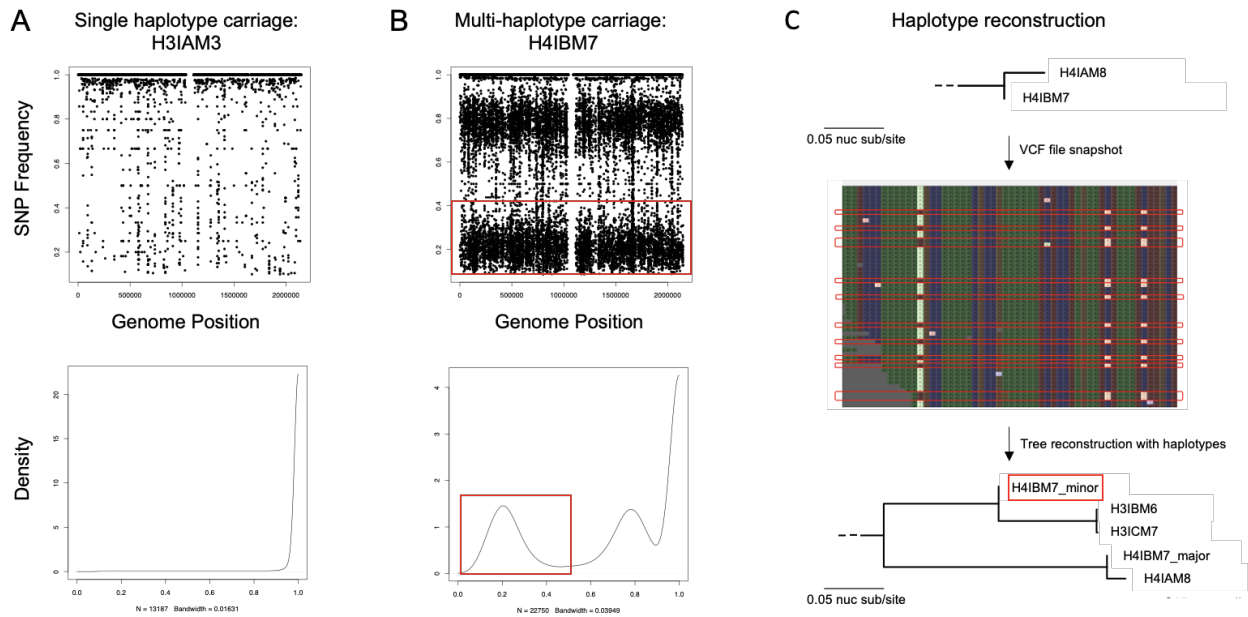


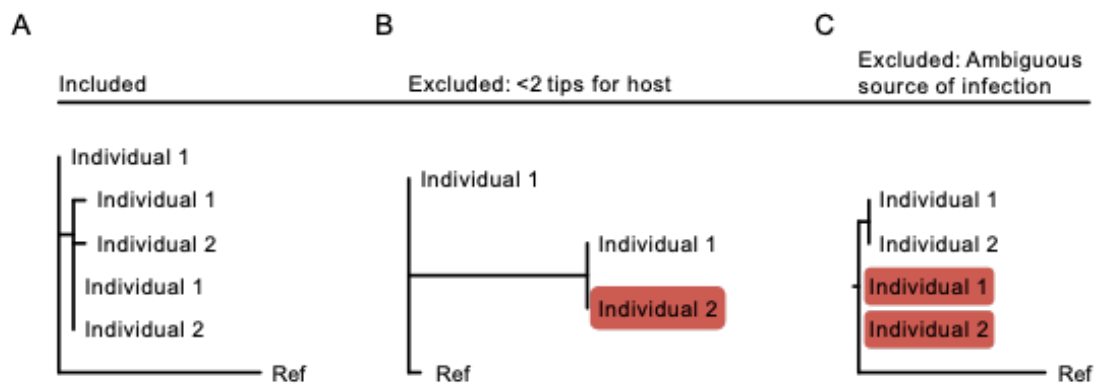
Figure 4. Proportional abundances of unique SNP count from 1-month intervals from within-host longitudinal samples. Where individuals had at least two consecutive swabs, the first time point was compared to the second time point and subsequently, the second time point was compared to the third time point. Instances of individuals having more than two consecutive swabs are denoted by the grey brackets. The light green represents the proportion of SNP from the first time point and the dark green represents the count from the second time point of the consecutive sets. The grey represents the shared SNP counts present in both time points. The proportions of the unique number of SNP are explicitly written within each of the corresponding coloured bars.

Supplemental Table 1. Sequencing quality of the whole-genome NGS reads for all 37 isolates included in the study.

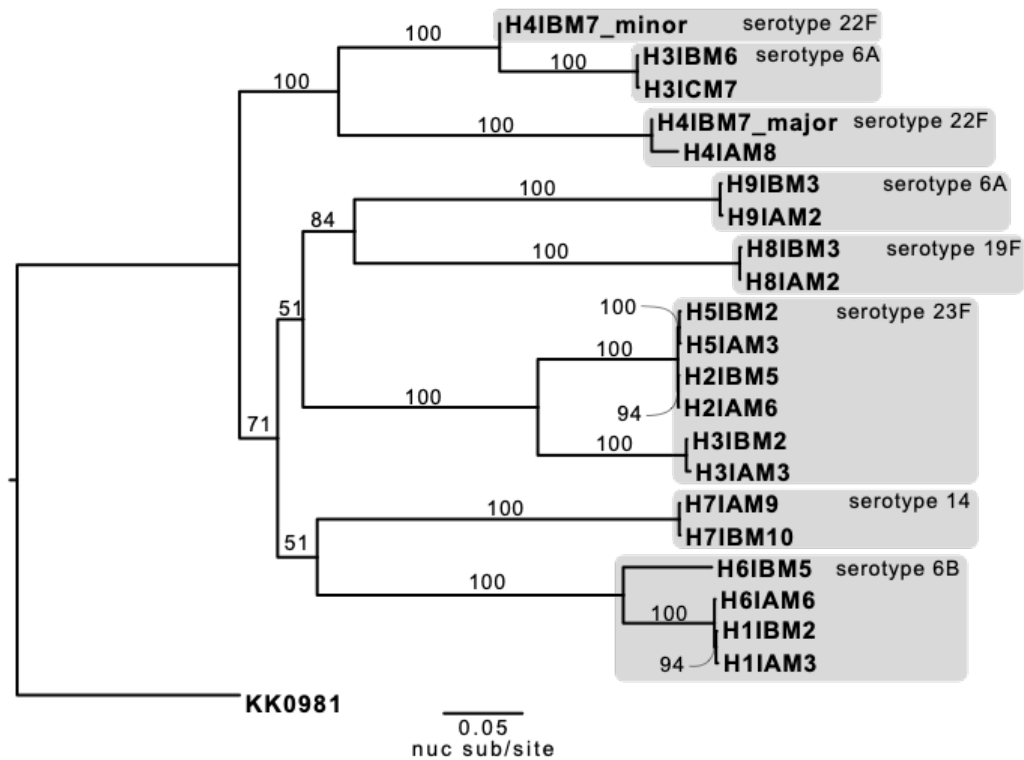
ID	Mean Coverage	SD	<i>Streptococcus pneumoniae</i> read match (%)
H1IAM3	53	24	81.1
H1IAM4	130	47	84.2
H1IAM5	200	70	84.9
H1IBM1	186	66	86.7
H1IBM2	337	135	90.7
H1IBM3	189	66	85.0
H2IAM6	95	39	90.6
H2IBM5	71	30	87.6
H2IBM6	151	52	84.6
H2IBM7	141	56	80.0
H3IAM3	26	19	77.1
H3IBM2	91	37	92.9
H3IBM3	169	66	86.9
H3IBM6	76	29	90.7
H3IBM7	182	68	85.5
H3IBM8	166	61	83.9
H3ICM7	80	32	92.4
H3ICM8	168	62	85.8
H4IAM8	133	55	92.6
H4IBM7	67	25	91.0
H5IAM3	128	53	89.7
H5IBM2	90	37	90.4
H6IAM6	115	48	90.3
H6IAM7	50	19	88.6
H6IBM4	42	18	85.4
H6IBM5	110	41	85.4
H7IAM10	116	48	80.2
H7IAM9	66	28	89.2
H7IBM10	86	49	90.2
H8IAM2	59	27	91.8
H8IAM3	106	44	83.4
H8IAM4	103	43	83.3
H8IBM3	75	32	90.9
H8IBM4	103	43	82.6
H8IBM5	125	51	84.0
H9IAM2	39	17	92.9
H9IBM3	26	26	33.1



Supplemental Figure 1. Haplotype reconstruction. (A) This is an example where there is no evidence to support that the individual is infected with multiple haplotypes. A single point on the SNP frequency plot represents a single polymorphic site to the reference genome. SNP that occur at a frequency of 1.0 indicate the SNP is present in all of the sample's reads while the density plot shows the density of the SNP frequencies. (B) This is an example where there is evidence to support that the individual is infected with multiple haplotypes. The points on the SNP frequency plot reveal there are two populations with distinct clusters of polymorphic sites at 20% and 80% likewise in the density plot. The distribution occurring at 20% is designated as the minor strain and is highlighted in a red box throughout. (C) Shows a snapshot of the phylogenetic consensus SNP tree with H4IBM7 (no haplotype isolation) and the linked isolate, H4IAM8. The snapshot of the variant calling format files highlights reads that correspond to the minor strain while the remainder corresponds to the major strain. The phylogenetic consensus SNP tree reconstruction after haplotype isolation reveals clustering of H4IBM7_major and H4IAM8 while H4IAM8_minor is more distantly related to H4IAM8.



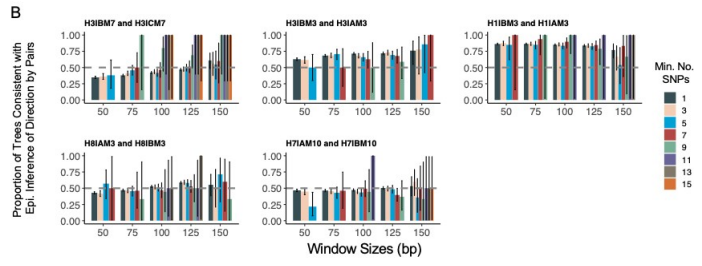
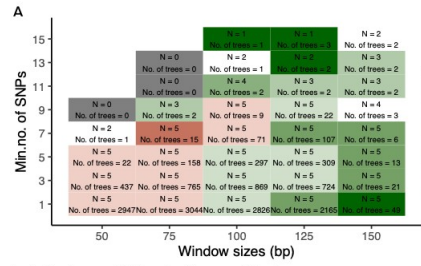
Supplemental Figure 2. Two additional quality control steps were included in the direction of transmission analysis. (A) Shows a simplified sub-tree that would pass the quality control steps and would be included in the call for directionality where individual 1 is the source of the infection. (B) Highlights the first step of the quality control which was to exclude sub-trees that were revealed to have only one tip from either individual (highlighted in red). (C) Highlights the second step which is the excluded sub-trees that demonstrate both individuals being equally the source of the infection (highlighted in red).



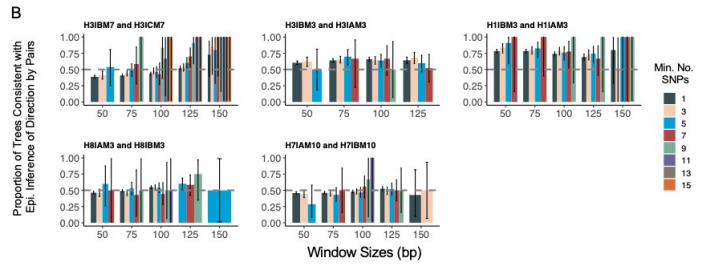
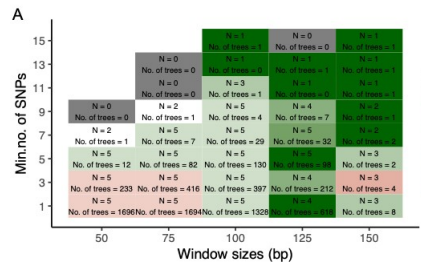
Supplemental Figure 3. Maximum-likelihood phylogeny of the 20 *S. pneumoniae* genomes from the 10 pairs of isolates from subsequent visits rooted to the reference genome, KK0981. The consensus SNP tree was reconstructed from an alignment of all polymorphic sites along the genomes (51,682 bp). Branch supports $\geq 50\%$, as determined by 1,000 bootstrap replicates, are denoted on the relevant branches. Branch length represents nucleotide substitutions per site (nuc sub/site), as denoted by the scaled bar. Within-serotype clustering is highlighted in grey boxes.

Direction of transmission analysis using same-visit samples

No tree filtering

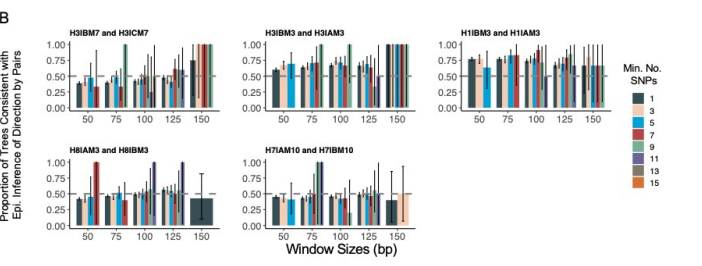
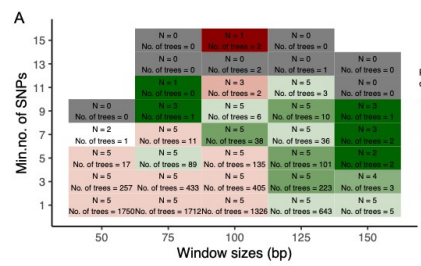


Including trees with 2 nodes from each host

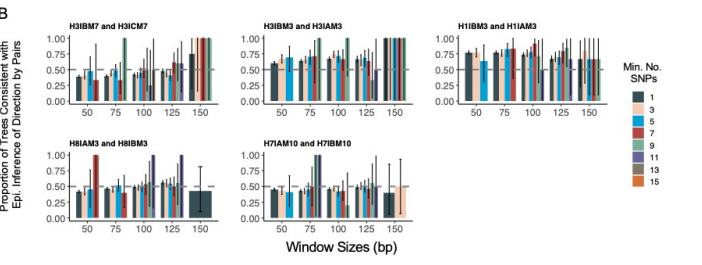
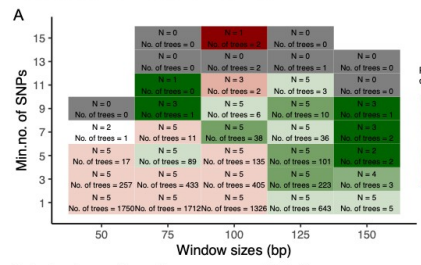


Direction of transmission analysis using same-visit samples & reference serotype 23F

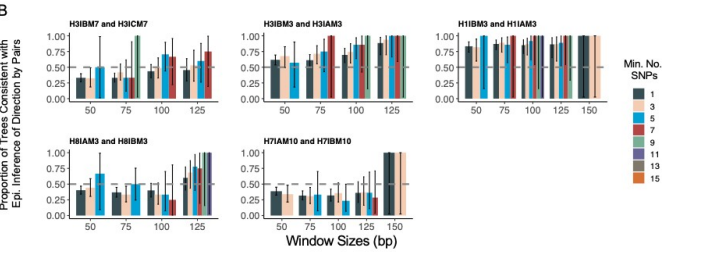
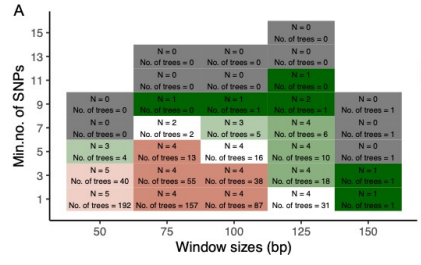
No tree filtering



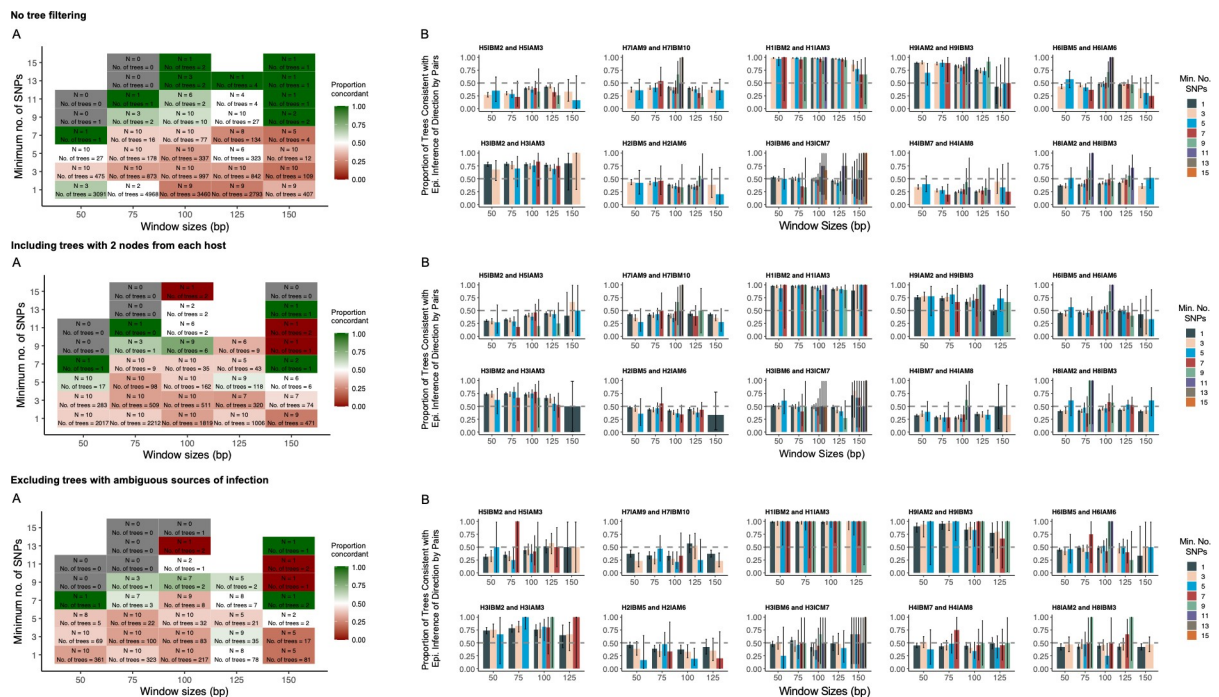
Including trees with 2 nodes from each host



Excluding trees with ambiguous sources of infection



Direction of transmission analysis using subsequent-visit samples



Supplemental Figure 4. Sensitivity analysis inferring the direction of transmission. The proportion of concordant inferred directionality per minimum number of SNP per read (1,3,5,7,9,11,13, or 15) and read window sizes (50, 75, 100 125, or 150 base pairs). (A) Inference from samples collected during the same visit. Green and red-coloured boxes denote the proportion of pairs for which the inferred direction of transmission was concordant with the epidemiological data, green is equivalent to 100% and red is equivalent to 0%. White boxes denote equal distributions of concordant and discordant inferred directions (proportion = 0.50). While grey boxes denote that phylogenies were generated, however, they were classified as “unlinked” or “ambiguous directions” and empty boxes denote that no sub-trees were generated for this combination of window size and SNP. The “N” represents the number of pairs analysed for the respective window size and SNP combination and the “N of Trees” is the average number of sub-trees used for the direction of transmission for those pairs analysed. (B) The proportion of sub-trees concordant with the epidemiological data, for each pair, with the different combinations of window sizes and minimum number of SNP represented by the coloured bars.

References

- Abdullahi O, Wanjiru E, Musyimi R, Glass N, Scott JAG. 2007. Validation of nasopharyngeal sampling and culture techniques for detection of *Streptococcus pneumoniae* in children in Kenya. *J. Clin. Microbiol.* 45:3408–3410.
- Balaji A, Ozer EA, Kociolek LK. 2019. Clostridioides difficile whole-genome sequencing reveals limited within-host genetic diversity in a paediatric cohort. *J. Clin. Microbiol.* 57:1–6.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
- Chaguza C, Senghore M, Bojang E, Gladstone RA, Lo SW, Tientcheu PE, Bancroft RE, Worwui A, Foster-Nyarko E, Ceesay F, et al. 2020. Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation. *Nat. Commun.* 11:1–14.
- Chiba N, Murayama SY, Morozumi M, Iwata S, Ubukata K. 2017. Genome Evolution to Penicillin Resistance in Serotype 3 *Streptococcus pneumoniae* by Capsular Switching. *Antimicrob. Agents Chemother.* 61:e00478-17, e00478-17.
- 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. *J. Bacteriol.* 191:1480–1489.
- Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* 14:150–162.
- Epping L, van Tonder AJ, Gladstone RA, The Global Pneumococcal Sequencing Consortium, Bentley SD, Page AJ, Keane JA. 2018. SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data. *Microb. Genomics* [Internet] 4. Available from: <https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000186>
- Flasche S, Lipsitch M, Ojal J, Pinsent A. 2020. Estimating the contribution of different age strata to vaccine serotype pneumococcal transmission in the pre vaccine era: a modelling study. *BMC Med.* 18:129.
- Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* [Internet]. Available from: <http://arxiv.org/abs/1207.3907>
- Gouliouris T, Coll F, Ludden C, Blane B, Raven KE, Naydenova P, Crawley C, Török ME, Enoch DA, Brown NM, et al. 2021. Quantifying acquisition and transmission of *Enterococcus faecium* using genomic surveillance. *Nat. Microbiol.* 6:103–111.
- Grijalva CG, Nuorti JP, Arbogast PG, Martin SW, Edwards KM, Griffin MR. 2007. Decline in

- pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. *Lancet Lond. Engl.* 369:1179–1186.
- Hall MD, Holden MTG, Srisomang P, Mahavanakul W, Wuthiekanun V, Limmathurotsakul D, Fountain K, Parkhill J, Nickerson EK, Peacock SJ, *et al.* 2019. Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife* 8:1–22.
- Hussain M, Melegaro A, Pebody RG, George R, Edmunds WJ, Talukdar R, Martin SA, Efstratiou A, Miller E. 2005. A longitudinal household study of *Streptococcus pneumoniae* nasopharyngeal carriage in a UK setting. *Epidemiol. Infect.* 133:891–898.
- In HP, Pritchard DG, Cartee R, Brandao A, Brandileone MCC, Nahm MH. 2007. Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. *J. Clin. Microbiol.* 45:1225–1233.
- Jacka B, Applegate T, Kraiden M, Olmstead A, Harrigan PR, Marshall BDL, DeBeck K, Milloy M-J, Lamoury F, Pybus OG, *et al.* 2014. Phylogenetic clustering of hepatitis C virus among people who inject drugs in Vancouver, Canada: HEPATOLOGY, Vol. 00, No. X, 2014. *Hepatology* 60:1571–1580.
- Kamng'ona AW, Hinds J, Bar-Zeev N, Gould KA, Chaguza C, Msefula C, Cornick JE, Kulohoma BW, Gray K, Bentley SD, *et al.* 2015. High multiple carriage and emergence of *Streptococcus pneumoniae* vaccine serotype variants in Malawian children. *BMC Infect. Dis.* 15:234.
- Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, 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* 4:e2477.
- Leitner T. 2019. Phylogenetics in HIV transmission. *Curr. Opin. HIV AIDS* 14:181–187.
- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* [Internet]. Available from: <http://arxiv.org/abs/1303.3997>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* 25:2078–2079.
- Mantere T, Kersten S, Hoischen A. 2019. Long-read sequencing emerging in medical genetics. *Front. Genet.* 10:1–14.
- Martin MA, Lee RS, Cowley LA, Gardy JL, Hanage WP. 2018. Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. *Microb. Genomics* 4.
- Neal EFG, Nguyen C, Ratu FT, Matanitobua S, Dunne EM, Reyburn R, Kama M, Devi R,

- Jenkins KM, Tikoduadua L, *et al.* 2019. A Comparison of Pneumococcal Nasopharyngeal Carriage in Very Young Fijian Infants Born by Vaginal or Cesarean Delivery. *JAMA Netw. Open* 2:e1913650.
- O'Brien KL, Dagan R. 2003. The potential indirect effect of conjugate pneumococcal vaccines. *Vaccine* 21:1815–1825.
- O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T. 2009. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *The Lancet* 374:893–902.
- PANGEA Consortium and Rakai Health Sciences Program, Ratmann O, Grabowski MK, Hall M, Golubchik T, Wymant C, Abeler-Dörner L, Bonsall D, Hoppe A, Brown AL, *et al.* 2019. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nat. Commun.* 10:1411.
- le Polain de Waroux O, Flasche S, Kucharski AJ, Langendorf C, Ndazima D, Mwanga-Amumpaire J, Grais RF, Cohuet S, Edmunds WJ. 2018. Identifying human encounters that shape the transmission of *Streptococcus pneumoniae* and other acute respiratory infections. *Epidemics* 25:72–79.
- van der Poll T, Opal SM. 2009. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *The Lancet* 374:1543–1556.
- Poolman JT, Peeters CCAM, van den Dobbelsteen GPJM. 2013. The history of pneumococcal conjugate vaccine development: dose selection. *Expert Rev. Vaccines* 12:1379–1394.
- Principi N, Esposito S. 2016. Prevention of Community-Acquired Pneumonia with Available Pneumococcal Vaccines. *Int. J. Mol. Sci.* 18.
- Qian G, Toizumi M, Clifford S, Le LT, Papastylianou T, Satzke C, Quilty B, Iwasaki C, Kitamura N, Takegata M, *et al.* 2022. Association of pneumococcal carriage in infants with the risk of carriage among their contacts in Nha Trang, Vietnam: A nested cross-sectional survey. Kretzschmar MEE, editor. *PLOS Med.* 19:e1004016.
- Rose R, Hall M, Redd AD, Lamers S, Barbier AE, Porcella SF, Hudelson SE, Piwowar-Manning E, McCauley M, Gamble T, *et al.* 2020. Phylogenetic methods inconsistently predict the direction of HIV transmission among heterosexual pairs in the HPTN 052 cohort. *J. Infect. Dis.* 221:1406–1413.
- Southern J, Andrews N, Sandu P, Sheppard CL, Waight PA, Fry NK, Van Hoek AJ, Miller E. 2018. Pneumococcal carriage in children and their household contacts six years after introduction of the 13-valent pneumococcal conjugate vaccine in England. Miyaji EN, editor. *PLOS ONE* 13:e0195799.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
- Street NW, Street NW, Suite C. 2020. Evaluation of phylogenetic methods for inferring the

direction of HIV transmission : HPTN © The Author (s) 2020 . Published by Oxford University Press for the Infectious Diseases Society of America . All rights reserved .
For permissions , e-mail : jour.

- Thindwa D, Wolter N, Pinsent A, Carrim M, Ojal J, Tempia S, Moyes J, McMorro M, Kleynhans J, Gottberg A von, *et al.* 2021. Estimating the contribution of HIV-infected adults to household pneumococcal transmission in South Africa, 2016–2018: A hidden Markov modelling study. Althouse B, editor. *PLoS Comput. Biol.* 17:e1009680.
- Villabona-Arenas CJ, Hué S, Baxter JAC, Hall M, Lythgoe KA, Bradley J, Atkins KE. 2022. Using phylogenetics to infer HIV-1 transmission direction between known transmission pairs. *Proc. Natl. Acad. Sci.* 119:e2210604119.
- 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. *Lancet Glob. Health* 6:e744–e757.
- Weinberger DM, Pitzer VE, Regev-Yochay G, Givon-Lavi N, Dagan R. 2019. Association Between the Decline in Pneumococcal Disease in Unimmunized Adults and Vaccine-Derived Protection Against Colonization in Toddlers and Preschool-Aged Children. *Am. J. Epidemiol.* 188:160–168.
- Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40:11189–11201.
- Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257.
- Worby CJ, Lipsitch M, Hanage WP. 2014. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. Koelle K, editor. *PLoS Comput. Biol.* 10:e1003549.
- Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, Gall A, Cornelissen M, Fraser C, STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration. 2018. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Mol. Biol. Evol.* 35:719–733.
- Xu Y, Stockdale JE, Naidu V, Hatherell H, Stimson J, Stagg HR, Abubakar I, Colijn C. 2020. Transmission analysis of a large tuberculosis outbreak in London: a mathematical modelling study using genomic data. *Microb. Genomics* [Internet] 6. Available from: <https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000450>
- Zhang Y, Wymant C, Laeyendecker O, Grabowski MK, Hall M, Hudelson S, Piwowar-Manning E, McCauley M, Gamble T, Hosseinipour MC, *et al.* 2020. Evaluation of phylogenetic

methods for inferring the direction of HIV transmission: HPTN 052. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.*

Zivich PN, Grabenstein JD, Becker-Dreps SI, Weber DJ. 2018. *Streptococcus pneumoniae* outbreaks and implications for transmission and control: a systematic review. *Pneumonia* 10:11.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1902896	Title	Miss
First Name(s)	Jada Nicole		
Surname/Family Name	Hackman		
Thesis Title	APPLICATION OF PATHOGEN GENOMICS TO INFER THE TRANSMISSION DIRECTION OF RESPIRATORY INFECTION		
Primary Supervisor	Stéphane Hué		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

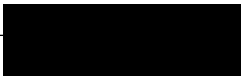
Where is the work intended to be published?	To be determined
Please list the paper's authors in the intended authorship order:	Jada Hackman, Martin L. Hibberd, Todd D. Swarthout, Jason Hinds, James Ashall, Carmen Sheppard, Gerry Tonkin-Hill, Kate Gould, Comfort Brown, Jacqueline Msefula, Andrew A Mataya, Michiko Toizumi, Lay-Myint

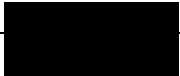
	Yoshida, Neil French, Robert S. Heyderman, Stefan Flasche, Brenda Kwambana, Stéphane Hué
Stage of publication	Not yet submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed all of the bioinformatic analysis for this study, interpreted the results, and wrote/edited the manuscript for submission.
--	--

SECTION E

Student Signature	
Date	29 May 2023

Supervisor Signature	
Date	31/05/23

CHAPTER 3: EVALUATING METHODS IN IDENTIFYING AND QUANTIFYING *STREPTOCOCCUS PNEUMONIAE* SUBPOPULATION USING NEXT-GENERATION SEQUENCING DATA

Authors

Jada Hackman (1), Martin L. Hibberd (2), Todd D. Swarthout (3,4,5), Jason Hinds (8,9), James Ashall (2), Carmen Sheppard (4,8), Gerry Tonkin-Hill (6), Kate Gould (7), Comfort Brown (5), Jacqueline Msefula (5), Andrew A Mataya (5), Michiko Toizumi (11,12), Lay-Myint Yoshida (11,12), Neil French (9,10), Robert S. Heyderman (3), Stefan Flasche (1)*, Brenda Kwambana (13)*, Stéphane Hué (1)*

Affiliation

(1) Faculty of Epidemiology and Population Health, Department of Infectious Disease Epidemiology, The London School of Hygiene and Tropical Medicine, London, UK

(2) Faculty of Infectious and Tropical Diseases, The London School of Hygiene and Tropical Medicine, London, UK

(3) NIHR Global Health Research Unit on Mucosal Pathogens, Division of Infection and Immunity, University College London, London, UK

(4) Julius Center for Health Sciences and Primary Care, Dept Epidemiology, University Medical Centre Utrecht, Utrecht, Netherlands

(5) Malawi Liverpool Wellcome Research Programme, Blantyre, Malawi

(6) Department of Biostatistics, University of Oslo, Blindern, Norway

(7) BUGS Bioscience, London Bioscience Innovation Centre, London, UK

(8) Vaccine Preventable Bacteria Section, UKHSA, London, UK

(9) Institute for Infection and Immunity, St George's University of London, London, UK

(10) University of Liverpool, Institute of Infection Veterinary & Ecological Science, Liverpool, UK

(11) Department of Pediatric Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, Nagasaki, Japan

(12) School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan

(13) Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

*Contributed equally

Corresponding author

Jada Hackman, Jada.hackman@lshtm.ac.uk, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, United Kingdom

Keywords

Co-carriage, pneumococcus, Africa, *Streptococcus pneumoniae*, sequencing, microarray, serotyping

ABSTRACT

Background

Detection of multiple pneumococcal serotype carriage is important for monitoring vaccine impact, particularly among populations in which pneumococcal co-carriage is common. We compared pipelines for identifying pneumococcal subpopulations using whole-genome sequencing data.

Methods

We selected 24 paediatric nasopharyngeal samples from Blantyre, Malawi, previously assessed by DNA microarray serotyping with confirmed pneumococcal co-carriage including up to six different serotypes. Pneumococcal DNA from culture plate sweeps were sequenced using Illumina MiSeq and genomic serotyping was carried out using SeroCall and PneumoKITy. We also used a mixture modelling on mutation frequency distributions to identify respective serotype subpopulations. Six samples were re-sequenced at higher depths to improve the detection of low-abundance serotypes.

Results

DNA microarray detected a total of 79 non-unique serotypes, of which 41 occur at high abundance (>10%) while the remaining 37 occur at low abundance (<10%). The average sequencing depth for the 24 samples was 57X. In comparison with DNA microarray, SeroCall had 100% sensitivity in determining the dominant serotype while PneumoKITy had 92% (22/24) concordance. SeroCall's sensitivity for identifying high abundance serotypes was 98% (95% CI, 0.68-1.00); low abundance was 54% (95% CI, 0.22-0.86), any abundance was 66% (95% CI: 0.44-1.00) for any abundance. While PneumoKITy's sensitivity for identifying high abundance serotypes was 86% (95% CI, 0.56-1.00); low abundance was 19% (95% CI, 0.00-0.51), any abundance was 54% (95% CI: 0.32-0.76) for any abundance.

An average 3-fold increase in sequencing depth slightly increased sensitivity for low-abundance serotype identification. Mixture modelling showed some potential for identifying serotypes in the sample through their associated SNP frequency.

Conclusion

Genomic serotyping pipelines have high sensitivity for identifying serotypes unless carried at low abundance.

INTRODUCTION

Carriage of *Streptococcus pneumoniae* is typically asymptomatic but can lead to the development of invasive pneumococcal disease (IPD) and it is one of the most common causes of childhood pneumonia. Most pneumococci are encapsulated with a complex capsular polysaccharide (cps) that contributes to its virulence and pathogenicity¹. All typeable pneumococci are typed by the cps locus flanked by the dexB and aliA genes^{2,3} and there are currently more than 100 distinct serotypes identified⁴.

Carriage of multiple unique pneumococcal strains (co-carriage) is common in settings with high carriage prevalence⁵, with an estimated 40% of children found to carry multiple serotypes within their first year of life in The Gambia and Malawi^{6,7}. Monitoring of co-carriage is an important part of surveillance activities, including in the characterisation of the ecological response of pneumococci to vaccine pressures. Additionally, there is a limited understanding of co-carriage within transmission dynamics which is likely due to limited sensitivity to detect minor variants from genomic data.

DNA microarray that uses serotype and serogroup-specific probes to target the highly variable glycosyltransferase genes has been established as the gold standard for reliable detection of co-carriage⁸⁻¹⁰. Recently, whole-genome sequencing (WGS) has become a cost-effective alternative for serotyping IPD samples in routine surveillance to identify the serotypes that are present¹¹. Bioinformatic tools such as PneumoCaT and SeroBA use a k-mer-based method to identify concordance between query cps locus next-generation sequencing reads and the pneumococcus Capsular Type Variant database¹¹, with high sensitivity (99% and 98%, respectively). However, these tools have limited capacity to identify serotypes in specimens with pneumococcal co-carriage. However, the recently developed pipelines PneumoKITy and SeroCall can identify co-carriage with high sensitivity (<85%), making them an attractive alternative to previous approaches.

In this study, we compared pneumococcal serotyping methods for identifying pneumococcal co-carriage using whole-genome sequencing and assessed the potential to differentiate these variants for further analyses.

MATERIALS AND METHODS

Sample collection

A total of 24 *S. pneumoniae*-positive nasopharyngeal swab samples were included in this study. These were part of a larger study, the study design and sample collection of which were detailed elsewhere¹³. In summary, the nasopharyngeal swabs were collected from asymptomatic children as a part of a prospective observational study using random sampling to monitor pneumococcal carriage in Blantyre, Malawi, following the introduction of Pneumococcal Conjugate Vaccine (PCV) over 3.5 years. Two samples, S13 and S16 (Table 1), were also a mixture with other *non-Streptococcus pneumoniae* including *Bifidobacterium infantis*, and *Streptococcus mitis*, *oralis*, and *parasanguinis* for the purpose of this analysis.

Sample processing

Nasopharyngeal swabs were stored in milk–tryptone–glucose–glycerol (STGG) medium of which 30 µL was plated on gentamicin-sheep blood agar and incubated overnight at 37°C in 5% CO₂. The presence of *Streptococcus pneumoniae* was identified by optochin sensitivity and colony morphology.

DNA microarray for co-carriage detection

The nasopharyngeal swab specimens were prepared for *S. pneumoniae* microarray serotyping at BUGS Bioscience Ltd. (London, United Kingdom) as previously described³. The DNA purified from the pneumococcal plate sweeps prepared for microarray analysis was stored at -20°C. The 24 samples were selected to represent a mix of colonisation with a range of 1 to 6 pneumococcal serotypes present at varying frequencies, as determined by microarray (Table 1). The subsequent genomic serotyping analysis was initially carried out blinded to the results of the microarray data.

Sequencing and sequence processing

Aliquots of the samples' DNA were transported to the London School of Hygiene and Tropical Medicine (London, United Kingdom) for whole-genome sequencing on the Illumina MiSeq platform, using Qiagen FX library kit (Qiagen), with enzymatic fragmentation for 12 minutes targeting 300-400 bps fragments. Of the original 24 samples, six were selected to be resequenced at a higher sequencing depth.

Adaptors from the raw data were trimmed using Trimmomatic v0.39¹⁴. The forward and reverse FASTQ files containing the reads were aligned using the reference genome KK0981, with Burrow-Wheeler Alignment v.0.7.17 (BWA-MEM) and SAMtools mpileup v1.9.114¹⁵. The

quality of the sequencing data was assessed using Kraken2 and non-*S. pneumoniae* reads were excluded for subsequent analysis except for the genomic serotyping. Sequencing coverage and depth were calculated from mapped reads in the bam files containing only *S. pneumoniae* reads, using SAMtools. For the 6 samples resequenced at greater depth, original and resequencing reads were pooled, resulting in higher sequencing depth.

Genomic serotyping

Two genomic serotyping approaches were implemented, SeroCall¹⁶ and PneumoKITy¹⁷, to identify the occurrence of co-carriage. These were carried out using the sequencing raw reads (e.g. not filtered for *S. pneumoniae* reads). There were no options to modify the SeroCall algorithm, however, PneumoKITy was initially run with the default parameters, including the requirement that 90% of k-mers were found in the reference. This was later lowered to 80% and 70% to investigate the corresponding trade-offs in sensitivity and specificity for serotyping (Supplemental Table 1).

Concordance of genomic serotyping with DNA microarray

The serotyping results from SeroCall and PneumoKITy were compared to DNA microarray serotyping. For both SeroCall and PneumoKITy, a sample was categorised as (i) *completely concordant* if the serotypes detected by genomic serotyping matched those detected by microarray exactly, (ii) *semi-concordant* when only serotypes present at high abundance (>10%) were observed by genomic serotyping, (iii) *semi-discordant* when some serotypes present at high abundance (>10%) were not observed, and (iv) *complete discordance* when the dominant serotype (>50%) was not observed as the most abundant by genomic serotyping. The genomic serotyping methods were compared to DNA microarray and sensitivity is defined as being able to detect the serotype levels and not just the serogroup level, additionally, non-typeables are included in the sensitivity calculation. Sensitivity was reported as a percentage with a 95% confidence interval (95% CI).

Identifying subpopulations based on SNP frequency

Variant calling format files containing the distribution and frequency of single nucleotide polymorphisms (SNP) found in the samples were generated using Freebayes v1.3.2¹⁸. They were then visualised using LoFreq v2¹⁹, which plots the relative frequency of observed SNPs. The number of frequency peaks and relative SNPs abundance were first estimated by visual inspection of the plots. A mixture modelling approach was then implemented to estimate the number of subpopulations based on the SNPs frequency distributions. This was carried out in R version 4.2.2, using the package *gamlss.mx* version 4.3-5. The modelling approach fitted

one-to-size normal distributions to the SNPs frequency data for subpopulation number estimates and the best estimates were assessed using Akaike Information Criterion values.

SNPs were filtered in an attempt to increase the genomic signal prior to serotyping: SNPs that occurred at 100% frequency were removed, as these were not informative for intra-host diversity. Additionally, SNPs at very low frequencies were removed due to potential sequencing artefacts by setting a density threshold which was determined by visual inspection as these areas, densities <0.3, under the curve were commonly observed between two defined peaks (Figure 3).

RESULTS

Sample description

The number and frequency of serotypes detected in the samples by microarrays are shown in Table 1. Of the 24 nasopharyngeal swab samples, DNA microarray detected six serotypes in two samples, five serotypes in four, four serotypes in four, three serotypes in four, two serotypes in four, and a single serotype in four samples (Table 1). The most abundant dominant serotype was 35B (4/24) followed by 23F (3/24) and the most prevalent serotype that was co-carried was serotype 14 (5/20). These results were used as a point of comparison to assess the genomic methods.

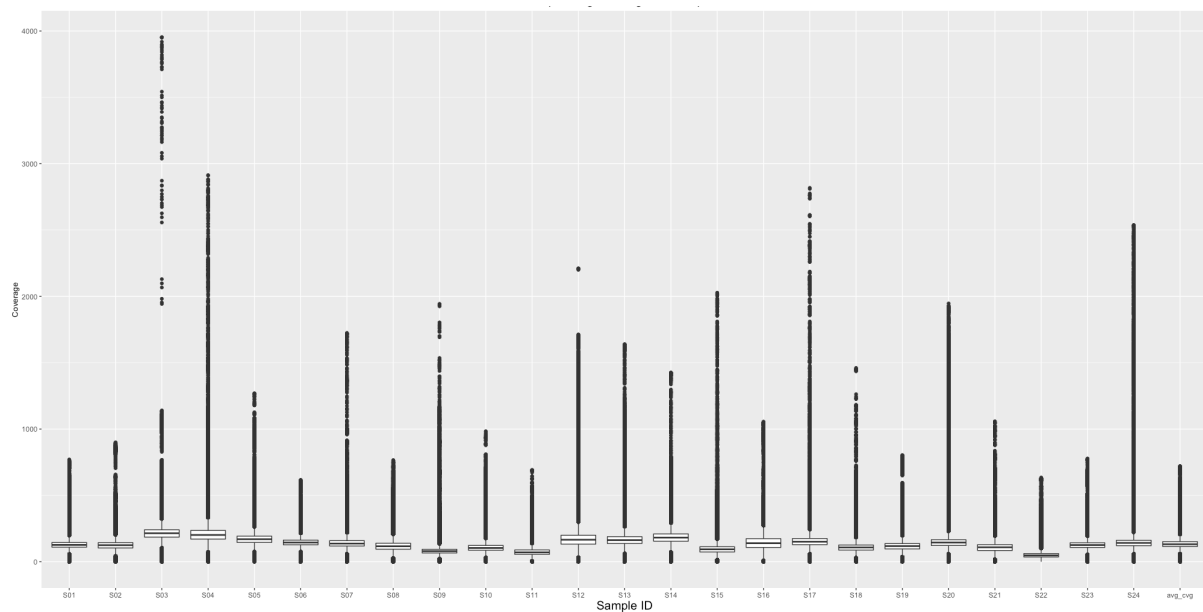
Table 1. Overview of the 24 sequence data quality and the DNA microarray results on the serotypes detected.

Sample ID	Number of mapped reads	Coverage	Mean depth	% Sp reads	DNA Microarray						
					No. of st	st no. 1 (%)	st no. 2 (%)	st no. 3 (%)	st no. 4 (%)	st no. 5 (%)	st no. 6 (%)
1	675683	92	51	81	3	35B (92)	6A/B[6A] (6)	14 (2)			
2	639079	91	53	87	1	35B (100)					
3	1053737	92	86	86	2	3 (82)	13 (18)				
4	1031636	94	88	86	4	23A (92)	20A/B[20B] (4)	23F (3)	15F (1)		
5	892799	96	74	83	6	35B (44)	14 (22)	19B (12)	6A/B[6A] (11)	21 (6)	NT2 (5)
6	785997	98	67	84	5	23F (64)	38 (24)	NT3b (6)	NT2 (5)	9L-like (1)	
7	776809	97	65	82	2	23F (96)	11A/D/E[11A] (4)				
8	586890	93	48	83	4	14 (79)	23F (9)	23A (8)	19A (4)		
9	450136	93	34	80	6	19F (72)	14 (22)	22F (3)	35B (1)	19B (1)	NT2 (1)
10	561528	95	45	82	5	19A (63)	15B/C[15B] (20)	1 (15)	11A/D/E[11A] (1)	23B-like (1)	
11	409984	93	32	81	3	15B/C[15B] (67)	13 (30)	21 (3)			
12	914487	95	71	82	5	12F/44[12F] (80)	28A/F[28F] (8)	20A/B[20B] (5)	14 (4)	6A/B[6B] (3)	
13*	819610	90	65	79	3*	22A (96)	36-like* (3)	18B/C[18C] (1)			
14	977902	90	79	86	2	35B (99)	38 (1)				
15	506404	98	41	82	4	23F (57)	19F (38)	11A/D/E[11A] (3)	6A/B[6B] (2)		
16*	701352	92	56	62	4*	40/7B/7C[7C] (54)	17F (28)	19F (14)	7F-like* (4)		
17	792903	93	66	81	1	6A/B[6B] (100)					
18	556650	94	44	84	5	13 (54)	14 (36)	3 (7)	34 (2)	38 (1)	
19	623422	95	51	85	4	3 (65)	40/7B/7C[7C] (25)	38 (7)	11A/D/E[11A] (3)		
20	748438	90	62	86	1	12F/44[12F] (100)					
21	573636	95	47	85	3	16F (45)	34 (32)	3 (23)			
22	273984	90	21	40	2	19B (62)	NT2 (38)				
23	655286	94	55	84	3	34 (83)	10B (13)	19F (4)			
24	745529	88	60	83	1	23B-like (100)					

* Artificial mixture of *Streptococcus Pneumoniae* with *Bifidobacterium infantis*, and *Streptococcus mitis, oralis*, and *parasanguinis*
Abbreviations: *Streptococcus pneumoniae* (Sp), number (no.), serotype (st)
Green text, highlight st detected >10% with DNA microarray

Sequencing results

The average sequencing coverage for the original 24 samples was 93% (range, 88%-98%), with a corresponding average sequencing depth of 57X (standard deviation, $\pm 17X$). Sample S22 had the lowest average depth at 21X (Table 1, Supplemental Figure 1). The average percentage of reads that matched the pneumococcal genome was 81% (standard deviation, $\pm 10\%$). Three samples had more than 20% of reads that did not match *Streptococcus pneumoniae*, the two mixture samples, S13 and S16, and S22 (Table 1). S13 had 86% of reads that matched with *Streptococcus pneumoniae*, while the remainder (14%) of the reads were non-pneumoniae *Streptococcus*, while S16 had 62% of reads matched with *Streptococcus pneumoniae* and the remainder were mostly non-pneumoniae *Streptococcus* and *Actinobacteria* (7%), *Eukaryota* (1%), and unclassified (3%). S22, a non-mixture sample, had 40% of reads match with Sp, while the remainder were mostly non-pneumoniae *Streptococcus* (7%), *Lactococcus* (24%), *Enterococcaceae* (21%), and unclassified (2%).



Supplemental Figure 1. Sequencing depth for all 24 original samples. Boxes and whiskers display the sequencing depth's lower, median, and upper bounds.

Sensitivity of genomic serotyping methods compared to DNA microarray

Compared to DNA microarray, SeroCall correctly identified the dominant serotypes in all 24 samples (100%, [95% CI, 0.85 - 1.00]) and correctly identified up to five extra serotypes in cases of co-carriage. There was complete concordance for 7/24 (29%) of the samples, of which, four were single serotype carriage, one dual co-carriage, and two co-carriage of three serotypes. There was semi-concordance for 16/24 (67%) and semi-discordance for 1/24 (4%) which was due to an unobserved non-typeable that was detected by DNA microarray at 38%

(Table 2). DNA microarray detected a total of 79 non-unique serotypes, of which 41 occur at high abundance while the remaining 37 occur at low abundance. Of the total detected by DNA microarray, SeroCall was able to identify 40 (98% [95% CI, 0.68-1.00]) at high abundance, 20 (54% [95% CI, 0.22-0.86]) at low abundance, and 52 (66% [95% CI, 0.44-1.00]) at any abundances (Figure 1).

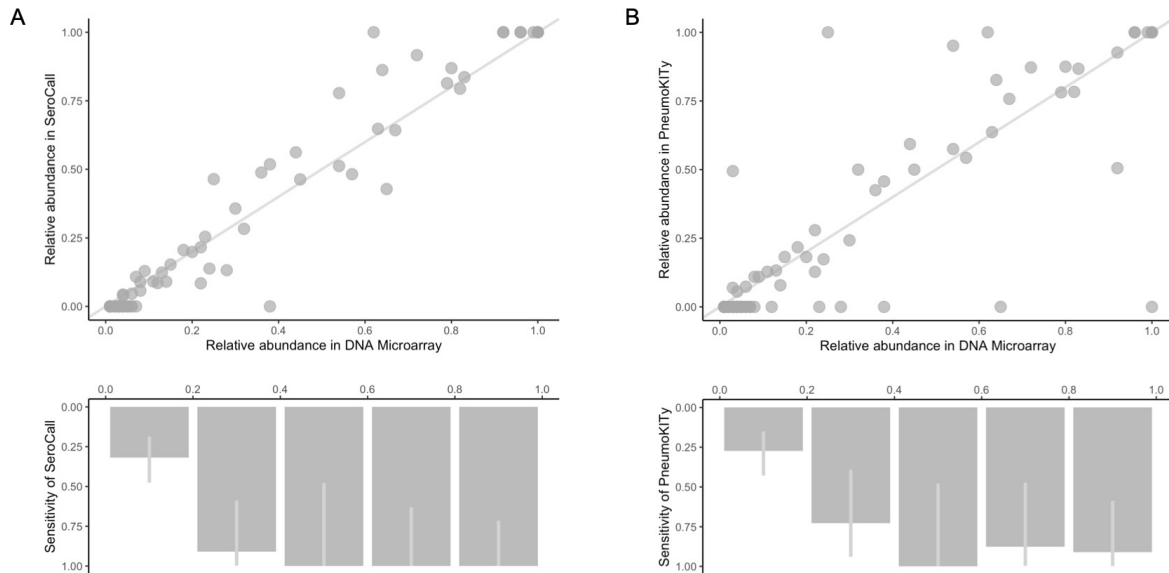


Figure 1. (A) Top, relative abundance of serotypes detected by DNA Microarray and their relative abundances observed by SeroCall; bottom, SeroCall sensitivity (%) to identify serotypes detected by DNA microarray, regardless of their relative abundance, with the light grey line representing 95% confidence interval. (B) Same as (A) but using PneumoKITy as the genomic serotyping method.

Table 2. Sensitivity of SeroCall and PneumoKITy compared to DNA Microarray

Sample ID	Visual no. of peaks	SNP frequency				SeroCall					PneumoKITy, $\rho = 80\%$					
		% of peaks	MM no. of st	no. of st.	no. cap reads	stno. 1 (%)	stno. 2 (%)	stno. 3 (%)	stno. 4 (%)	stno. 5 (%)	discordance / notes	no. of st.	stno. 1 (%)	stno. 2 (%)	stno. 3 (%)	discordance / notes
S01	2	100%	2	1	580	358 (100)						2	358/350 (92.65)	6A_6B_6C_6D (7.35)		
S02	1	100%	1	1	689	358 (100)	13 (20.6)					1	378 (3)	13 (21.7)		
S03	2	80%, 20%	3	2	638	03 (79.4)						2	23A (50.54)	23F (49.46)	PK detect higher 23F	
S04	1	100%	2	1	1931	23A (100)						2	35B/35D (99.3)	14 (27.91)	19B	
S05	3	50%, 30%, 20%	3	5	1227	35B/56.2)	14 (21.6)	08E/6A (9.1)	199 (8.5)	21 (4.6)		3	23F (82.67)	25F_55A_38 (17.33)		
S06	2	80%, 20%	3	2	1322	23F (86.2)	38 (13.8)					2	23F (100)	23A (10.94)		
S07	1	100%	3	1	1332	23F (100)						1	14 (78.12)	14 (12.77)		
S08	2	90%, 10%	5	3	670	14 (81.4)	23F (12.9)	23A (5.7)				3	19F (87.23)	23A (10.94)		
S09	2	90%, 10%	3	2	584	19F (91.6)						2	19A_19AF (63.84)	14 (12.77)		
S10	2	80%, 20%	3	3	764	19A (64.8)	15B/15C (19.9)	01 (1.53)				3	15B/15C (75.76)	15B/15C (18.18)		
S11	2	80%, 20%	6	2	402	15B/15C (64.3)	13 (35.7)					2	13 (24.24)	1 (18.18)		
S12	2	90%, 10%	2	3	1194	12F (86.9)	28F (8.9)	14 (4.3)				3	12F_12A_12B_44_46 (87.5)	14 (5.56)		
S13*	1	100%	1	1	1039	22A (100)						1	22A (100)			
S14	1	100%	1	1	990	35B (100)						1	35B/35D (100)			
S15	2	50%, 15%	4	2	536	19F (51.8)	23F (48.2)					2	23F (54.29)	19F (45.71)	st17F	
S16*	2	80%, 20%	3	3	1148	07C (77.8)	17F (13.2)	19F (9.0)				3	7B/0 (45.67)	7C (46.46)	st8B	
S17	1	100%	1	1	1435	08E/6B1 (100)						1	Snogroup_6_(6E) (100)		st17F	
S18	3	50%, 40%, 10%	3	2	587	13 (61.2)	14 (48.8)					2	13 (67.5)	14 (42.5)	st10	
S19	3	70%, 20%, 10%	3	3	541	07C (46.4)	03 (42.8)	38 (10.8)				1	7C (100)		st10	
S20	1	100%	1	1	1608	12F (100)						1	12F (100)		st13	
S21	2	70%, 30%	4	3	503	16F (48.3)	34 (28.3)	03 (25.4)				2	16F (50.0)	34 (50.0)	st13	
S22	1	100%	2	2	411	19B (100.0)						1	19B (100)		only NT	
S23	2	90%, 10%	2	3	864	34 (83.6)	10B (12.4)	19F (4.0)				2	34 (86.76)	10B (13.24)	only NT	
S24	1	100%	1	1	1092	23B (100.0)						1	23B (100)			

* Artificial mixture of *Streptococcus Pneumoniae* with *Bifidobacterium infantis* and *Streptococcus mitis*, *oralis*, and *parvaeruginis*
 Abbreviations: *Streptococcus pneumoniae* (Sp), number (no.), serotype (st), mixture modelling (MM)
 Red text, additional false-positive serotypes detected with decrease specificity

PneumoKITy, with an 80% k-mer percentage cut-off, was able to identify the dominant serotypes in 23/24 samples (96%) and identified co-carriage with up to three serotypes. There was complete concordance for 4 (17%) of the samples, of which, one sample had co-carriage of up to two serotypes. There was semi-concordance for 14 (58%) of the samples, semi-discordance for 4 (17%), and complete discordance for 2 (8%). Semi-discordant samples were due to unidentified high abundance serotypes including serotypes 19B, 17F, 3, and a non-typeable population while the two complete discordant cases were due to dominant serotype 3 that was not observed for one sample and the other only identified to the serogroup level, serogroup 6 (Table 2). In comparison to the 79 unique serotypes detected by DNA microarray, PneumoKITy was able to identify 36 (86% [95% CI, 0.56-1.00]) at high abundance, 7 (19% [95% CI, 0.00-0.51]) at low abundance, and 43 (54% [95% CI, 0.32-0.76]) at any abundances (Figure 1). Of the six serotypes that were unobserved by PneumoKITy at high abundances, two were serotype 3, one was 17F, one was 19B, one was 6B, and one was non-typeable-2.

Overall, increasing the sequencing depth of the 6 samples had no impact on the sensitivity of the genomic serotyping methods. The specificities of the resequenced samples were reiterated through genomic serotyping using both SeroCall and PneumoKITy where samples maintained the same pneumococcal serotype mixtures and abundance. Pooling the original and resequenced samples to further increase the sequencing depth by 3-fold increased the sensitivity of the genomic serotyping methods to further identify low abundance serotypes. Of the 37 low abundance serotypes identified by DNA microarray, SeroCall was able to find an additional two serotypes, serotype 21 at 2.7% for sample S11, and serotype 11A at 5.4% for sample S19. Similarly, PneumoKITy was also able to identify one additional serotype at high abundance, serotype 38 at 15.56%, and two low abundance serotypes, 11A/11D at 8.89% for sample S19 and serotype 17F at 8.92% for sample S16 (Table 3).

Table 3. Effect of increased sequencing depth on the sensitivity of genomic serotyping methods

Sample ID	Fold increase no. of reads	Fold increase mean seq depth	no. of st	SeroCall				PneumKITy p = 80%			
				st.no. 1 (%)	st.no. 2 (%)	st.no. 3 (%)	st.no. 4 (%)	st.no. 1 (%)	st.no. 2 (%)	st.no. 3 (%)	st.no. 4 (%)
S03	Ref	Ref	2	03 (79.4)	13 (20.6)			3 (78.3)	13 (21.7)		
S03.reseq	3.8	3.7	2	03 (78.1)	13 (21.9)			3 (78.82)	13 (21.18)		
S03.pooled	4.7	4.8	2	03 (78.0)	13 (22.0)			3 (78.86)	13 (21.14)		
S09	Ref	Ref	2	19F (91.6)	14 (8.4)			19F (87.23)	14 (12.77)		
S09.reseq	2.0	2.0	2	19F (92.4)	14 (7.6)			19F (86.96)	14 (13.04)		
S09.pooled	3.0	3.1	2	19F (93.3)	14 (6.7)			19F (87.67)	14 (12.33)		
S11	Ref	Ref	2	15B/15C (64.3)	13 (35.7)			15B/15C (75.76)	13 (24.24)		
S11.reseq	4.3	4.2	2	15B/15C (70.5)	13 (29.5)			15B/15C (75.35)	13 (24.65)		
S11.pooled	5.3	5.4	3	15B/15C (88.8)	13 (28.6)			15B/15C (75.98)	13 (24.02)		
S16*	Ref	Ref	3	07C (77.8)	17F (13.2)	19F (9.0)		7B/40 (45.67)	7C (46.46)	19F (7.87)	
S16.reseq*	1.9	1.8	3	07C (79.2)	17F (13.4)	19F (7.3)		7B/40 (46.32)	7C (46.75)	19F (6.93)	
S16.pooled*	2.9	3.1	3	07C (80.4)	17F (12.4)	19F (7.2)		7B/40 (42.49)	7C (42.96)	19F (5.83)	17F (8.92)
S19	Ref	Ref	3	07C (46.4)	03 (42.8)	38 (10.8)		7C (100)			
S19.reseq	1.1	1.1	3	07C (28.2)	03 (64.9)	38 (6.8)		7C (100)			
S19.pooled	2.2	2.4	4	07C (37.2)	03 (48.7)	38 (8.8)		7C (75.56)			
S22	Ref	Ref	1	19B (100.0)				19B (100)			
S22.reseq	3.0	3.0	1	19B (100.0)				19B (100)			
S22.pooled	4.0	4.3	1	19B (100.0)				19B (100)			

* Artificial mixture of *Streptococcus Pneumoniae* with *Bifidobacterium infantis*, and *Streptococcus mitis, oralis*, and *parasanginis*

Abbreviations: *Streptococcus pneumoniae* (Sp), number (no.), serotype (st)

Blue text: additional serotypes detected from increased sequencing depth

Subpopulation identification from SNP frequency

The frequency distribution of polymorphic sites revealed some samples having clear and distinct peaks of frequencies of SNPs (e.g. S03) that likely indicate distinct subpopulations and potentially different capsular serotypes, while other samples had more ambiguous distributions (e.g. S05), and samples with minor genomic signals but no defined distributions of SNPs (e.g. S04) (Figure 3). For all 24 samples, the range of frequency peaks that were visually observable ranged from one to three peaks with two peaks being observed the most (Supplemental Figure 2). Compared to the number of serotypes detected by DNA microarray, 5/24 samples had a concordant number of peaks, four of which were single serotype carriage, while the remaining one had co-carriage of two serotypes. Additionally, the sample with concordant frequencies between visual estimates and microarray was also similar, 80%/20% and 82%/18%, respectively. The number of SNP frequency peaks identified for the remainder of 19/24 samples were on average two fewer serotypes detected by DNA microarray (Table 2). Of the original 24 samples, mixture modelling identified between one to six subpopulations from the SNP frequencies and estimated the same number of subpopulations as DNA microarray in 6 (25%), more in 5 (21%), and fewer in 13 (54%); SeroCall and PneumoKITy only identified fewer in 17 (71%) and 19 (80%), respectively.

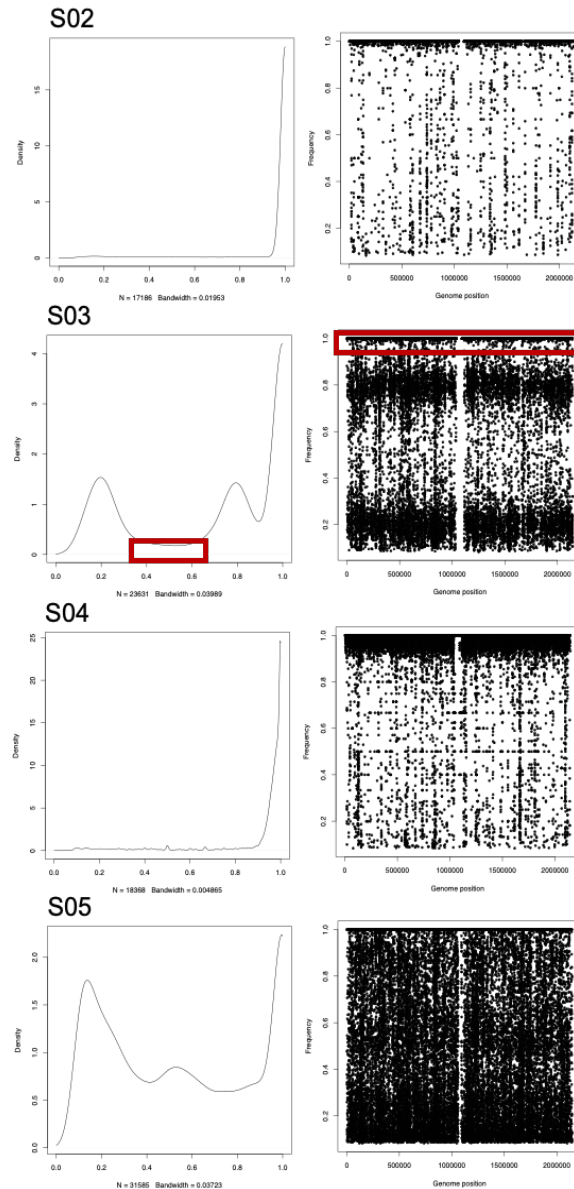
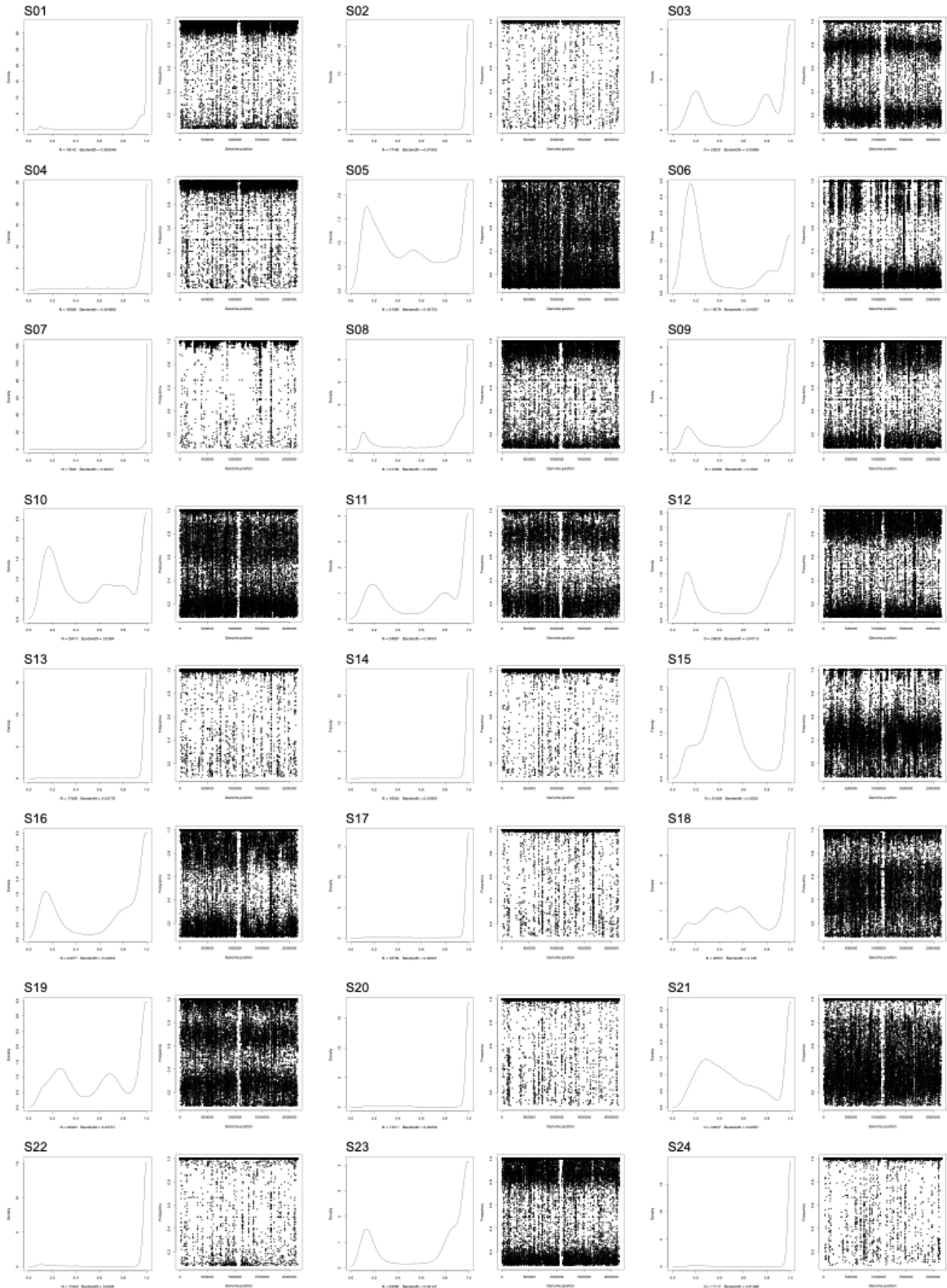


Figure 2. Density plot of SNP frequencies (left) and genome plot of SNPs in reference to KK0981 whole genome (right) where a single point on the SNP frequency plot represents a single polymorphic site to the reference genome. S02 is an example where there is no evidence to support that the individual is infected with multiple haplotypes. S03 is an example where there is evidence to support that the individual is infected with multiple haplotypes. S04 is an example where there is evidence there is probably a single population, however, there is some signal represented by the small peaks indicating potential unobserved minor variants. S05 is an example of clear co-carriage, however, it is difficult to distinguish. The red box, in the density plot, highlights the threshold (<0.3) that was set to exclude potential artefacts due to sequencing error, and in the genome plot, highlights the SNPs that occur at a frequency of 1.00 which are SNPs that are present in both samples.



Supplemental Figure 2. Density (left) and genome (right) plots for the original 24 sequences based on SNP frequencies.

The resequenced samples with increased sequencing depth had a qualitative impact on the SNPs frequency distribution for three of the six samples. Samples S03, S11, and S22 demonstrate darker frequency bands in the resequenced runs, highlighting a more mutation specificity, however, the same number of frequency bands remain, indicating the sensitivity has not been impacted. The remaining three samples maintained quantitatively similar frequency distributions. The mixture modelling revealed that S03 and S19 maintained the same number of subpopulations, and S09, S11, and S22 reduced the number of subpopulations by one, while S16 increased the subpopulation from three to seven (Table 1).

Sensitivity analyses

For PneumoKITy, configuring the alternative filter cut-off value for k-mer percentage parameter from the default (90%) to 80% resulted in higher sensitivity for identifying serotypes without compromising specificity. The adjustment increased sensitivity to identify an additional 10 serotypes across 9 samples. However, lowering the threshold to 70%, lowered the specificity and thus false positive serotypes were observed (Supplemental Table 1).

Supplemental Table 1. Sensitivity of PneumoKITy with varying levels of specificity.

Sample ID	no. of st. at no. 1 (%)	PneumoKITy p = 90%			discrepance / notes	no. of st. at no. 1 (%)	PneumoKITy p = 80%			discrepance / notes	no. of st. at no. 1 (%)	PneumoKITy p = 70%			discrepance / notes
		st no. 2 (%)	st no. 3 (%)	st no. 4 (%)			st no. 2 (%)	st no. 3 (%)	st no. 4 (%)			st no. 2 (%)	st no. 3 (%)	st no. 4 (%)	
S01	1	356/950 (100)				356/950 (92.65)	6A_08_6C_0D (7.35)				356/950 (92.65)	6A_08_6C_0D (7.35)			
S02	1	356/950 (100)				356 (100)					356 (100)				
S03	1	3 (100)				3 (78.3)	13 (21.7)				3 (78.3)				
S03.mseq	1	3 (100)				3 (78.3)	13 (21.7)				3 (78.3)				
S04	1	23A (100)				23A (90.54)	23F (49.46)				23A (90.54)	23F (49.46)			
S05	3	356/950 (93)	14 (27.9)	6A_08_6C_0D (12.79)		356/950 (92.65)	14 (27.9)	6A_08_6C_0D (12.79)		356/950 (92.65)	14 (27.9)	6A_08_6C_0D (12.79)			
S06	2	23F (100)	23F_23A_38 (16.22)			23F (100)	23F_23A_38 (17.33)				23F (100)	23F_23A_38 (17.33)			
S07	1	23F (100)				23F (100)					23F (100)				
S08	2	14 (87.72)	23F (12.28)			14 (78.12)	23A (10.94)	23F (10.94)		14 (78.12)	23A (10.94)	23F (10.94)			
S09	1	19F (100)				19F (86.96)	14 (13.04)	1 (1.818)		19F (86.96)	14 (13.04)	1 (1.818)			
S09.mseq	3	19A_19AF (63.64)	15B/15C (18.18)			19A_19AF (63.64)	15B/15C (18.18)	1 (1.818)		19A_19AF (63.64)	15B/15C (18.18)	1 (1.818)			
S10	1	15B/15C (100)				15B/15C (73.59)	13 (24.69)			15B/15C (73.59)	13 (24.69)				
S11	1	15B/15C (100)				15B/15C (73.59)	13 (24.69)			15B/15C (73.59)	13 (24.69)				
S11.mseq	1	15B/15C (100)				15B/15C (73.59)	13 (24.69)			15B/15C (73.59)	13 (24.69)				
S12	1	22A (100)				22A (100)	14 (3.96)	6A_08_6C_0D (6.94)		22A (100)	14 (3.96)	6A_08_6C_0D (6.94)			
S13*	1	356/950 (100)				356/950 (100)				356/950 (100)					
S14	1	23F (54.29)	19F (45.71)			23F (54.29)	19F (45.71)			23F (54.29)	19F (45.71)				
S15	2	7C (65.51)	19F (14.49)			7C (46.46)	19F (7.87)			7C (46.46)	19F (7.87)				
S16*	2	7C (67.1)	19F (12.9)			7C (46.46)	19F (6.93)			7C (46.46)	19F (6.93)				
S16.mseq*	1	14 (100)				14 (42.5)				14 (42.5)					
S17	1	7C (100)				7C (100)				7C (100)					
S18	1	7C (100)				7C (100)				7C (100)					
S19	1	7C (100)				7C (100)				7C (100)					
S19.mseq	1	7C (100)				7C (100)				7C (100)					
S20	1	16F (50.0)	34 (50.0)			16F (50.0)	34 (50.0)			16F (50.0)	34 (50.0)				
S21	2	198 (100)				198 (100)				198 (100)					
S22	1	198 (100)				198 (100)				198 (100)					
S22.mseq	2	34 (86.76)	10B (13.24)			34 (86.76)	10B (13.24)			34 (86.76)	10B (13.24)				
S23	1	23B (100)				23B (100)				23B (100)					
S24	1	23B (100)				23B (100)				23B (100)					

* Artificial mixture of *Streptococcus Pneumoniae* with *Haemophilus influenzae*, and *Streptococcus mitis*, *oralis*, and *parvus*

Abbreviations: *Streptococcus pneumoniae* (Sp), number (no.), serotype (st)

Blue box: additional serotypes detected with decrease specificity

Red box: additional false-positive serotypes detected with decrease specificity

Co-carriage detection sensitivity was cross-validated using the pipeline implemented by Tonkin-Hill *et al.*²⁰ which combines SeroCall with a deconvolution-based strategy and the results were comparable.

A subset of samples was resequenced at a higher depth in an attempt to improve sensitivity. Samples S03, S09, S16, and S22 increased sequencing depth from an average of 49X to 138X. While S19 and S11 had a slightly decreased coverage from the original run to the rerun 51X to 56X and 45X to 32X, respectively (Table 1).

DISCUSSION

Detecting co-carriage of *S. pneumoniae* serotypes through genomic sequencing could be advantageous over other serotyping methods by providing additional information on phylogenetic relationships and antimicrobial resistance. We demonstrate that both genomic pneumococcal serotyping methods tested in this study, SeroCall and PneumoKITy, can reliably identify the dominant serotypes of a mixed population, with a steep drop of sensitivity for serotypes carried at low abundance (<10%). However, increasing sequencing depth can increase the sensitivity of these methods and identify low-abundance serotypes.

SeroCall identified all the dominant serotypes in all 24 samples. However, PneumoKITy did not identify the majority populations in two samples, one of which was only identified at the serogroup level and the other an unobserved serotype 3. The developers of PneumoKITy, Sheppard *et al.*, noted that there is a limitation in identifying serotype 3, particularly in co-carriage at low abundances and that this could be potentially mitigated by lowering the specificity parameter. In our study, serotype 3 was co-carried at a high abundance and was only observable when the k-mer percentage threshold was lowered from 90% to 80%. The 80% threshold resulted in 100% specificity across the study samples, however, when the threshold was lowered from 80% to 70%, false-positive serotypes were observed.

Most of the discordance between microarray and genomic methods was due to the genomic serotyping methods lacking sufficient sensitivity to identify serotypes at relatively low abundance, highlighting the importance of read depth in the genomic detection of multi-carriage. However, increasing sequencing depth did not necessarily result in better detection rates, and there were instances where PneumoKITy was also not able to identify non-dominant high-abundance serotypes. This later observation could largely be explained by the reference database used by the program lacking a sufficient number of reference sequences that are reflective of current circulating diverse strains in Africa¹⁷.

Partial or complete deletion of the *cps* gene cluster can result in a serotype being non-typeable meaning they do not react with current antisera. Sample genomically serotyped as non-typeable could be due to current tools only being limited to serotyping based on the capsular protein or the limitation of the currently available antisera. Currently, PneumoKITy is not programmed to identify non-typeables, however, SeroCall was able to observe non-typeables. A previous study demonstrated a single-point mutation in the *wchA* gene resulted in serotype change from 7F to non-typeable²¹, highlighting the difficulties in distinguishing non-typeables.

Previous studies have evaluated serotyping methods but they have been limited to single carriage or have compared genomic methods between them, without comparison to microarray detection. Sheppard *et al.* did include a small comparison between PneumoKITy and SeroCall in their study, however, it was limited to a combination of a small number of unique serotypes (n=10), while our study had 34 unique serotypes¹⁷. Similarly, a study from Swarthout *et al.*, using the same dataset as our study, observed high concordance of serotype identification of single carriage between latex agglutination, genomic serotyping (PneumoCaT), and DNA microarray using 1,347 samples from community carriage surveillance in Blantyre, Malawi²². Manna *et al.* observed discordant results between PneumoCaT, SeroBA, and SeroCall, as well as discordant results within SeroCall in the identification of single carriage of serotype 14-like identifying them as serotype 14 and/or non-typeable²³, highlighting the importance of additional phenotypic testing to validate serotyping data.

Knight *et al.* highlighted that read depth would affect the sensitivity of SeroCall and recommended that samples should have between 2-3 million reads per sample. Only 5 of the 30 sequenced samples had >1 million *S. pneumoniae* reads, of which 3 samples were resequenced at a higher frequency. Despite the increased sequencing depth, SeroCall was unable to identify additional serotypes in these cases. However, when the original and resequenced samples were pooled, increasing the overall sequencing coverage, additional serotypes were picked up with SeroCall and PneumoKITy, most of which were present at low abundances. These results concur with the notion that higher sequencing depth could improve sensitivity for identifying co-carriage of low abundant serotypes from genomic data.

Limitations of the study

The first limitation of this study is the small sample size of 24 which limited the variation in combination and the quantity of co-carriage we were able to study. The second limitation is the use of a single next-generation sequencing method (Illumina MiSeq). Other sequencing methods such as Illumina HiSeq or Oxford Nanopore Technologies could result in a higher

depth of coverage or longer sequencing reads, which could impact the sensitivity and specificity of the genomic serotyping. On the other hand, some alternative methods have a higher sequencing error rate (e.g. Oxford Nanopore), which could also potentially affect detection. Increasing sequencing depth would increase the sensitivity of identifying low abundance variants as we observed when we pooled the duplicate sequencing runs together which improved our identification of serotypes <10%. Additionally, we suspect that increasing the read lengths would improve the alignment thus increasing specificity using the genomic serotyping methods. The third limitation is the potential degradation of the DNA between the sample preparation for DNA microarray and whole-genome sequencing resulting in potentially less optimal sequencing data. The fourth limitation of genomic serotyping methods is the potential bias induced by the reference database used in the pipeline, which could lead to misclassification or misquantification. For example, genetically similar serotypes could then be misclassified due to phenotypic differences or closely related serotypes would be difficult to quantify compared to distantly related serotypes.

Future work

Future work will focus on parsing the SNPs frequencies in an attempt to reconstruct the serotypes that are present in co-carriage that were identified by the mixture modelling. The reconstruction of the serotypes present in co-carriage will be added benefit.

Conclusion

One of the major limitations of the genomic-based serotyping approach is its lack of sensitivity compared to DNA microarray in its detection of minority serotypes. Despite that, genomic serotyping can identify high-abundant populations, which are likely to be the ones with the highest public biological relevance. NGS has become a cost-effective method compared to DNA microarray and can be easily implemented as a part of routine monitoring, particularly in resource-poor settings with higher rates of co-carriage. Additionally, there is an added benefit to including sequencing as a part of routine surveillance including additional information for phylogenetic inference to investigate transmission dynamics. While Quellung/latex agglutination might be more cost-effective at detecting dominant serotypes, this method is not optimal for identifying co-carriage²⁴. It has been reported that 120 colonies must be sampled by latex agglutination to identify serotypes that are co-carried at 5%⁸ and 299 colonies must be sampled to identify serotypes that are co-carried at 1%²⁵. SeroCall and PneumoKITy are free-access options with sufficient sensitivity for routine carriage surveillance to characterise dominant serotypes, additionally, SeroCall is able to identify co-carriage of serotypes >10% relative abundance.

Declaration of interests

Pending

Contributions

Project design - JH, MLH, SF, BK, SH

Sample collection, storage, culturing - Pending

Sample sequencing - MLH, JA

Sample processing for DNA microarray - JH

Data analysis - JH, CS, GTH, BK, SF, SH

Writing of first draft - JH

Writing and editing of the manuscript - JH, SH, BK

All authors read, edited, and approved the article.

Ethics

The study protocol was approved by the College of Medicine Research and Ethics Committee, University of Malawi (P.02/15/1677), and the Liverpool School of Tropical Medicine Research Ethics Committee (14.056). Adult participants and parents/guardians of child participants provided written informed consent; children 8 years old and older provided informed assent. This included consent for publication.

Data availability

Sequencing data will be made available upon acceptance of manuscripts. The specific data for this study are available from the authors upon request and subject to a data-sharing agreement.

Funding

This study was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the WISE scheme and EU grant QL4-CT-2000- 00640. SF is funded by a Sir Henry Dale Fellowship through the Wellcome Trust and the Royal Society (208812/Z/17/Z).

Acknowledgements

We thank the participants who made this study possible and the local authorities for their support.

References

1. Watson, D. A., Musher, D. M. & Verhoef, J. Pneumococcal virulence factors and host immune responses to them. *Eur. J. Clin. Microbiol. Infect. Dis.* **14**, 479–490 (1995).
2. Bentley, S. D. *et al.* Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. *PLoS Genet.* **2**, e31 (2006).
3. Mavroidi, A. *et al.* Genetic Relatedness of the *Streptococcus pneumoniae* Capsular Biosynthetic Loci. *J. Bacteriol.* **189**, 7841–7855 (2007).
4. Ganaie, F. *et al.* A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large *cps* Fragment from an Oral Streptococcus. *mBio* **11**, e00937-20, /mbio/11/3/mBio.00937-20.atom (2020).
5. Murad, C. *et al.* Pneumococcal carriage, density, and co-colonization dynamics: A longitudinal study in Indonesian infants. *Int. J. Infect. Dis.* **86**, 73–81 (2019).
6. Chaguza, C. *et al.* Carriage Dynamics of Pneumococcal Serotypes in Naturally Colonized Infants in a Rural African Setting During the First Year of Life. *Front. Pediatr.* **8**, 587730 (2021).
7. Kamng'ona, A. W. *et al.* High multiple carriage and emergence of *Streptococcus pneumoniae* vaccine serotype variants in Malawian children. *BMC Infect. Dis.* **15**, 234 (2015).
8. Satzke, C. *et al.* The PneuCarriage Project: A Multi-Centre Comparative Study to Identify the Best Serotyping Methods for Examining Pneumococcal Carriage in Vaccine Evaluation Studies. *PLOS Med.* **12**, e1001903 (2015).
9. Tomita, Y. *et al.* A new microarray system to detect *Streptococcus pneumoniae* serotypes. *J. Biomed. Biotechnol.* **2011**, 352736 (2011).
10. Wang, Q. *et al.* Development of a DNA microarray to identify the *Streptococcus pneumoniae* serotypes contained in the 23-valent pneumococcal polysaccharide vaccine and closely related serotypes. *J. Microbiol. Methods* **68**, 128–136 (2007).
11. Kapatai, G. *et al.* Whole genome sequencing of *Streptococcus pneumoniae* : development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ* **4**, e2477 (2016).
12. Epping, L. *et al.* SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data. *Microb. Genomics* **4**, (2018).
13. Swarthout, T. D. *et al.* High residual carriage of vaccine-serotype *Streptococcus pneumoniae* after introduction of pneumococcal conjugate vaccine in Malawi. *Nat. Commun.* **11**, 2222 (2020).
14. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
15. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*

- 25**, 2078–2079 (2009).
16. Knight, J. R. *et al.* Determining the serotype composition of mixed samples of pneumococcus using whole genome sequencing. <http://biorxiv.org/lookup/doi/10.1101/741603> (2019) doi:10.1101/741603.
 17. Sheppard, C. L. *et al.* PneumoKITy: A fast, flexible, specific, and sensitive tool for *Streptococcus pneumoniae* serotype screening and mixed serotype detection from genome sequence data. *Microb. Genomics* **8**, (2022).
 18. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* (2012).
 19. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
 20. Tonkin-Hill, G. *et al.* Pneumococcal within-host diversity during colonization, transmission and treatment. *Nat. Microbiol.* **7**, 1791–1804 (2022).
 21. Melchiorre, S. *et al.* Point mutations in wchA are responsible for the non-typability of two invasive *Streptococcus pneumoniae* isolates. *Microbiol. Read. Engl.* **158**, 338–344 (2012).
 22. Swarthout, T. D. *et al.* Evaluation of pneumococcal serotyping in nasopharyngeal carriage isolates by latex agglutination, whole genome sequencing (PneumoCaT) and DNA microarray in a high pneumococcal carriage prevalence population in Malawi. <http://biorxiv.org/lookup/doi/10.1101/2020.08.17.255224> (2020) doi:10.1101/2020.08.17.255224.
 23. Manna, S. *et al.* Variants of *Streptococcus pneumoniae* Serotype 14 from Papua New Guinea with the Potential to Be Mistyped and Escape Vaccine-Induced Protection. *Microbiol. Spectr.* **10**, e01524-22 (2022).
 24. O'Brien, K. L. & Nohynek, H. Report from a WHO Working Group: standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*. *Pediatr. Infect. Dis. J.* **22**, e1-11 (2003).
 25. Huebner, R. E. *et al.* Lack Of Utility Of Serotyping Multiple Colonies For Detection Of Simultaneous Nasopharyngeal Carriage Of Different Pneumococcal Serotypes: *Pediatr. Infect. Dis. J.* **19**, 1017–1020 (2000).

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1902896	Title	Miss
First Name(s)	Jada Nicole		
Surname/Family Name	Hackman		
Thesis Title	APPLICATION OF PATHOGEN GENOMICS TO INFER THE TRANSMISSION DIRECTION OF RESPIRATORY INFECTION		
Primary Supervisor	Stéphane Hué		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Wellcome Open Research		
When was the work published?	22 February 2023		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	

Stage of publication	Choose an item.
----------------------	-----------------

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed the bioinformatic analysis on the DNA sequences and the phylogenetic analysis for this study to detect within household transmission and edited the manuscript for submission. I had no involvement in the statistical analyses or interpretation other than the genomic identification of likely transmission links.
--	---

SECTION E

Student Signature	[Redacted]
Date	29 May 2023

Supervisor Signature	[Redacted]
Date	31/05/23

CHAPTER 4: EFFECTIVENESS OF BNT162B2 AND CHADOX1 AGAINST SARS-COV-2 HOUSEHOLD TRANSMISSION: A PROSPECTIVE COHORT STUDY IN ENGLAND



RESEARCH ARTICLE

Effectiveness of BNT162b2 and ChAdOx1 against SARS-CoV-2 household transmission: a prospective cohort study in England [version 1; peer review: 2 approved with reservations]

Samuel Clifford^{1,2}, Pauline Waight³, Jada Hackman^{1,2}, Stephane Hué^{1,2}, Charlotte M. Gower³, Freja CM Kirsebom³, Catriona Skarnes³, Louise Letley³, Jamie Lopez Bernal³, Nick Andrews³, Stefan Flasche^{1,2*}, Elizabeth Miller^{2,3*}

¹Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

²Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, UK

³National Infection Service, UK Health Security Agency, London, NW9 5EQ, UK

* Equal contributors

V1 First published: 22 Feb 2023, 8:96
<https://doi.org/10.12688/wellcomeopenres.17995.1>
Latest published: 22 Feb 2023, 8:96
<https://doi.org/10.12688/wellcomeopenres.17995.1>

Abstract

Background: The ability of SARS-CoV-2 vaccines to protect against infection and onward transmission determines whether immunisation can control global circulation. We estimated the effectiveness of Pfizer-BioNTech mRNA vaccine (BNT162b2) and Oxford AstraZeneca adenovirus vector vaccine (ChAdOx1) vaccines against acquisition and transmission of the Alpha and Delta variants in a prospective household study in England.

Methods: Households were recruited based on adult purported index cases testing positive after reverse transcription-quantitative (RT-q)PCR testing of oral-nasal swabs. Purported index cases and their household contacts took oral-nasal swabs on days 1, 3 and 7 after enrolment and a subset of the PCR-positive swabs underwent genomic sequencing conducted on a subset. We used Bayesian logistic regression to infer vaccine effectiveness against acquisition and transmission, adjusted for age, vaccination history and variant.

Results: Between 2 February 2021 and 10 September 2021, 213 index cases and 312 contacts were followed up. After excluding households lacking genomic proximity (N=2) or with unlikely serial intervals (N=16), 195 households with 278 contacts remained, of whom 113 (41%) became PCR positive. Delta lineages had 1.53 times the risk (95% Credible Interval: 1.04 – 2.20) of transmission than Alpha; contacts older than 18 years old were 1.48 (1.20 – 1.91) and 1.02 (0.93 – 1.16) times more likely to acquire an Alpha or Delta infection than children. Effectiveness of two doses of BNT162b2 against transmission of Delta was 36% (-1%, 66%) and 49% (18%, 73%) for ChAdOx1, similar to their effectiveness for Alpha. Protection against infection with Alpha

Open Peer Review

Approval Status ? ?

	1	2
version 1 22 Feb 2023	 view	 view
1. John Kubale , University of Michigan, Ann Arbor, USA		
2. Tim K Tsang , The University of Hong Kong, Hong Kong Special Administrative Region, China		

Any reports and responses or comments on the article can be found at the end of the article.

was higher than for Delta, 69% (9%, 95%) vs. 18% (-11%, 59%), respectively, for BNT162b2 and 24% (-41%, 72%) vs. 9% (-15%, 42%), respectively, for ChAdOx1.

Conclusions: BNT162b2 and ChAdOx1 reduce transmission of the Delta variant from breakthrough infections in the household setting, although their protection against infection within this setting is low.

Keywords

covid, vaccination, secondary attack rate, SARS-CoV-2, household transmission

Corresponding author: Elizabeth Miller (liz.miller@shrm.ac.uk)

Author roles: **Clifford S:** Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Waight P:** Data Curation, Investigation, Software, Writing – Review & Editing; **Hackman J:** Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hués S:** Investigation, Methodology, Supervision, Writing – Review & Editing; **Gower CM:** Data Curation, Writing – Review & Editing; **Kirsebom FC:** Data Curation, Writing – Review & Editing; **Skarnes C:** Data Curation, Writing – Review & Editing; **Letley L:** Data Curation, Investigation, Project Administration, Supervision, Writing – Review & Editing; **Lopez Bernal J:** Conceptualization, Funding Acquisition, Writing – Review & Editing; **Andrews N:** Conceptualization, Data Curation, Methodology, Writing – Review & Editing; **Flasche S:** Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Miller E:** Conceptualization, Data Curation, Funding Acquisition, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [208812, <https://doi.org/10.35802/208812>]; a Sir Henry Dale Fellowship award to Stefan Flasche and supporting Samuel Clifford. This study was funded by the UK Health Security Agency (formerly Public Health England) (an executive agency of the Department of Health) as part of the COVID-19 response. Samuel Clifford is funded by the UK Medical Research Council (MC_PC_19065 - Covid 19: Understanding the dynamics and drivers of the COVID-19 epidemic using real-time outbreak analytics). Jada Hackman is funded by the Nagasaki University-London School of Hygiene and Tropical Medicine Doctoral Programme under the WISE scheme. EM receives support from the National Institute for Health Research (NIHR) Health Protection Research Unit in Immunisation at the London School of Hygiene and Tropical Medicine in partnership with the UKHSA (Grant Reference NIHR200929). The non-UK HSA funders had no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. The authors have conducted their research independent from the funders. All authors, external and internal, had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis is also required.

Copyright: © 2023 Clifford S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Clifford S, Waight P, Hackman J *et al.* Effectiveness of BNT162b2 and ChAdOx1 against SARS-CoV-2 household transmission: a prospective cohort study in England [version 1; peer review: 2 approved with reservations] Wellcome Open Research 2023, 8:96 <https://doi.org/10.12688/wellcomeopenres.17995.1>

First published: 22 Feb 2023, 8:96 <https://doi.org/10.12688/wellcomeopenres.17995.1>

Introduction

The rapid development of safe and effective coronavirus disease 2019 (COVID-19) vaccines using both novel and traditional platforms, is an unprecedented scientific achievement. The United Kingdom was the first country to launch a national COVID-19 vaccination programme with the rollout of the Pfizer-BioNTech mRNA vaccine (BNT162b2) on 8th December 2020, followed shortly after by the Oxford AstraZeneca adenovirus vector vaccine (ChAdOx1). By September 2021, over 40% of the world's population had received at least one dose of a COVID-19 vaccine, whether an mRNA, adenovirus vector, or inactivated whole virion vaccine¹. In most countries, vaccine deployment has been focussed on direct protection of those individuals at the greatest risk of a severe outcome of SARS-CoV-2 infection, including the elderly and those with co-morbidities. Health care workers and others who, if infected, pose a transmission risk to vulnerable individuals, have also been identified as a priority group for vaccination.

The primary outcome of the efficacy trials of the currently authorised COVID-19 vaccines was symptomatic laboratory confirmed SARS-CoV-2 infection, with little information generated on protection against severe COVID-19 infection nor on the ability of the vaccines to prevent onward transmission in those infected. There is now a growing body of evidence from observational studies showing high protection against severe COVID-19 from inactivated whole virion, mRNA, and adenovirus vector vaccines²⁻⁴ but information on protection against transmission is still limited⁵. Attempts have been made to infer protection against transmission by comparing the viral load in the nasopharynx of vaccinated individuals with breakthrough infections with that in unvaccinated cases, using cycle threshold (Ct) values as a proxy⁶. Other approaches have used routine diagnostic PCR testing data, constructing households based on individuals' addresses or identifying them with contact tracing, and to estimate secondary attack rates by vaccination status of the index case. However, these studies are potentially subject to ascertainment bias as they are reliant on the testing behaviour of household contacts⁷⁻⁹.

Here we report the results of a prospective household transmission study set up by Public Health England (PHE) (now the UK Health Security Agency) in January 2021 to assess the effect of the vaccination history of index cases with COVID-19 on transmission of SARS-CoV-2 to household contacts, and the protection afforded to vaccinated contacts under conditions of household exposure.

Methods

Data

Households. The procedures for household recruitment and laboratory testing are the same as those used in the household transmission study conducted prior to vaccine availability and are detailed elsewhere¹⁰. In brief, infected index cases, identified *via* community testing in England (known as Pillar 2 testing), and their consenting household contacts are recruited by study nurses, on average, three days after their initial PCR test.

No additional measures were taken in the study to prevent household transmission. The vaccination status of index cases and their household contacts is obtained by data linkage with the National Immunisation Management System (NIMS) for England and checked with participants by the study nurse at the time of recruitment. Self-testing kits for the index case and household contacts to take combined nose and throat swabs on Day 1 (day of recruitment), Day 3 and Day 7 are couriered to households and subsequently tested by dual target PCR at PHE Colindale (ORF and E genes). PCR positive swabs are sequenced as part of the COG-UK initiative¹¹. Household contacts were defined as infected if one or more swabs was PCR positive.

The household transmission study is ongoing and inclusion in this analysis is based on participants having returned at least one swab, being either unvaccinated or vaccinated with one or two doses of either BNT162b2 or ChAdOx1 with the vaccination dates recorded in the national vaccination register, and the age at time of recruitment and the date of onset of symptoms (fever, cough, runny nose, sore throat, shortness of breath, loss of taste or smell, nausea, diarrhoea, muscle/body pain, headache or other) recorded.

The analysis code used can be found as *Extended data*¹².

Statistical analysis

All analysis was conducted in **R Project for Statistical Computing** (RRID:SCR_001905) 4.1.1¹³ with Bayesian models fit using the **rjags package** (RRID:SCR_017573)¹⁴. The secondary attack rate (SAR) for each combination of case and contact is estimated here by predicting the probability an unseen contact acquires an infection from an infected case given the vaccination history and age of each and the index case's variant. As the observed SARs in this study were high, model-estimated odd ratios poorly approximate relative risks. Thus, effect estimates are calculated as risk ratios (RRs) of SARs. Unless mentioned otherwise, the baseline age groups for such comparisons were adult index cases younger than 50 years old and contacts at least 18 years old. The predicted SARs and RRs are summarised with medians and 95% credible intervals.

Household secondary attack rate. We fit a Bayesian hierarchical linear model with Bernoulli likelihood for the probability that a household contact of an index case acquires a SARS-CoV-2 infection within a week of recruitment. The model estimates both a protective effect for vaccinated contacts against infection and a reduction in transmission for vaccinated cases, which are assumed to be independent. The effect of the first dose is assumed to only occur 21 days after the vaccination is received, and an additional effect of the second dose requires at least seven days have passed since the second vaccination as in the SIREN study, which considers the effectiveness of BNT162b2 in healthcare workers in England¹⁵. These effects are assumed to depend on the vaccine product, and number of doses thereof, received by both the index case and the contact (Table 1). The probability of acquiring infection

Table 1. Number of contacts with listed vaccine status for each case vaccine status (N). Numbers in brackets show the additional individuals included in the sensitivity analysis (n). BNT162b2, Pfizer-BioNTech mRNA vaccine; ChAdOx1, Oxford AstraZeneca adenovirus vector vaccine.

Case	1 ChAdOx1 N(n)	2 ChAdOx1 N(n)	1 BNT162b2 N(n)	2 BNT162b2 N(n)	None N(n)
1 ChAdOx1	17 (4)	1	3	4	23 (5)
2 ChAdOx1	2 (1)	26 (5)	7 (1)	12 (1)	21 (9)
1 BNT162b2	6	1	15 (2)	2	33 (2)
2 BNT162b2	9	8	4	10	9
None	6	2	4	5	48 (2)

is also assumed to depend on the age of both the case and contact, and the circulating lineage. Vaccine effectiveness is calculated for both protection against infection and reduction of transmission as 1 -RR for RRs of household SARs with and without the vaccine. For such, the SARs were sampled during the Markov Chain Monte Carlo (MCMC) sampling, for each combination of variant and case and contact vaccine status (1 or 2 doses for each product) and age group, against a baseline of that case-contact pair and variant in the absence of any vaccination.

For the model of secondary attack rate, the likelihood is

$$\begin{aligned}
 y_i &\sim \text{Bern}(p_i) \\
 \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \delta_{V_i} + \\
 &\beta_{1,V_i} \mathbf{I}(d_{1,i,\text{contact}} \geq k_1) + \beta_{2,V_i} \mathbf{I}(d_{2,i,\text{contact}} \geq k_2) + \\
 &\gamma_{1,V_i} \mathbf{I}(d_{1,i,\text{case}} \geq k_1) + \gamma_{2,V_i} \mathbf{I}(d_{2,i,\text{case}} \geq k_2) + \\
 &\epsilon_{\text{contact}} \mathbf{I}(A_{i,\text{contact}} < a_{\text{contact}}) + \epsilon_{\text{case}} \mathbf{I}(A_{i,\text{case}} \geq a_{\text{case}})
 \end{aligned}$$

where \mathbf{I} is an indicator function, which is 1 when its input is true and 0 otherwise, and $d_{j,i,c}$ is the number of days since the j th dose of vaccine product V was given to either contact i or their household index case (indexed by c) who is infected with variant V . A fixed effect, δ_i accounts for the increased infectivity of Delta beyond that of Alpha. Protection afforded by dose j is assumed to begin after $k = \{21,7\}$ days. These are currently fixed, but a distribution may be used instead if there is some observed variability we wish to include. Age effects, ϵ_i are assumed to be non-zero when the contact is younger than $A_{1,\text{contact}} = 18$ and when the case is at least as old as $A_{1,\text{case}} = 50$. Where V_i was missing due to that household's swabs not being sequenced, it was sampled at each step of the MCMC from a Bernoulli distribution with its single parameter representing the modelled proportion of sequenced Pillar 2 swabs with Delta lineage at time of that household's Day 1 swabs.

The priors for the model parameters associated with transmission reduction are parameterised as weakly informative normal distributions (with means and precision ($\tau = \sigma^{-2}$))

$$\begin{aligned}
 \gamma_{j,v,V} &\sim \mathcal{N}(\gamma_{j,0,v}, \tau_\gamma) \\
 \gamma_{j,0,v} &\sim \mathcal{N}(0, 10^6) \\
 \tau_\gamma &= \sigma_\gamma^{-2} \\
 \sigma_\gamma &\sim \text{Exp}(0.3)
 \end{aligned}$$

The penalised complexity prior on the standard deviation implies a prior probability of it exceeding 10 of 0.05.

For the infection protection parameters, informative priors are derived from reported vaccine efficacy of the two vaccine products against the Alpha (B.1.1.7) and Delta (B.1.617.2) variants of SARS-CoV-2¹⁶. The priors are normally distributed for the log-odds ratios, with mean and precision parameters,

$$\begin{aligned}
 \beta_{1,1,A} &\sim \mathcal{N}(\log 0.5, 2) & \beta_{1,1,D} &\sim \mathcal{N}(\log 0.65, 2) \\
 \beta_{1,2,A} &\sim \mathcal{N}(\log 0.43, 0.25) & \beta_{1,2,D} &\sim \mathcal{N}(\log 0.67, 0.25) \\
 \beta_{2,1,A} &\sim \mathcal{N}\left(\log \frac{0.25}{0.5}, 2\right) & \beta_{2,1,D} &\sim \mathcal{N}\left(\log \frac{0.33}{0.65}, 2\right) \\
 \beta_{2,2,A} &\sim \mathcal{N}\left(\log \frac{0.06}{0.43}, 0.25\right) & \beta_{2,2,D} &\sim \mathcal{N}\left(\log \frac{0.12}{0.67}, 0.25\right)
 \end{aligned}$$

Here we have taken the point estimates of the log odds ratios and scaled the standard errors up by a factor of two, rounding to the nearest 0.5, in order to provide informative priors with additional variance that ensure that the posteriors are still sensitive to the data. As the $\beta_{2,v,V}$ represent the marginal effect of the second dose, we derive $\beta_{2,v,V} = \beta_{1,v,V} + \beta_{2,v,V}$, the log-odds of the effect of double vaccination against variant V with vaccine product v .

The effects of age have informative priors derived from Davies *et al.*¹⁷, for under-18s acquiring infection, $\epsilon_{\text{contact}} \sim \mathcal{N}(\log 0.50, 24)$,

and from Yousaf *et al.*,¹⁸ cited in Goldstein *et al.*,¹⁹ for case transmission, $\epsilon_{\text{case}} \sim N(\log 1.86, 4.67)$.

To determine how informative the priors above are, we replace the informative priors above for all $\beta_{j,u,v}$ with a weakly informative $N(0, \sigma_\beta^{-2})$ prior with $\sigma_\beta \sim \text{Exp}(0.3)$ and the effects of age each having a weakly informative $N(0, 10^{-6})$ prior.

Lineage. At the start of data collection, the B.1.1.7 (Alpha) SARS-CoV-2 variant was most prevalent in the United Kingdom, and an increasing proportion of swabs sequenced by Pillar 2 testing were identified as B.1.617.2 (Delta) variant over time²⁰. Where sequencing was not available to determine the variant for a positive swab, the probability that it was the Delta variant was estimated from the date of sampling and a logistic regression model fit to the number of weekly cases identified through Pillar 2 that were either Alpha or Delta variant.

Participants' age. Vaccine eligibility and type is correlated with age and date of vaccination. This is because from 7th April 2021 the BNT162b2 vaccine was recommended for individuals under 30 years old in preference to ChAdOx1, which was then extended to those between 30 and 40 years old from 7th May 2021²¹ and also because, apart from those in high risk groups, vaccination was not offered to the general 16–17 year old population until August 2021²² and the general 12–15 year old population until September 2021²³. We account for age in the model by considering that children under 18 years old will have decreased susceptibility to infection, compared to adults¹⁷, and that older adults are more likely to transmit¹⁹. While the study did not specifically recruit only adult index

cases, the minimum age of index cases was 21 years old. The median age of index cases was 48 years old and so we split adults into younger (between 18 and 49 years old) and older (at least 50 years old) age groups. Few participants were older than 65 years old, so we did not distinguish between groups aged 50–64 and 65+ years old. We did not adjust for prior infection status as information on this was incomplete at the time of data lock, nor for gender as this was previously shown not to be a factor in determining household transmission¹⁰. Table 2 shows the age and vaccine status breakdown of index cases and their household contacts.

Infection history dynamics. PCR positivity relative to the onset of symptoms was estimated using data from all symptomatic cases and contacts, with pseudo-absences generated to simulate the time of infecting exposure. Comparison is made for each combination of vaccine product, number of doses, and variant against the corresponding unvaccinated group.

Identification of non-household transmission. As per the study design, the index case for each household was by default considered to be the individual who presented for Pillar 2 testing. To reduce the risk of misclassification bias we excluded from the analyses all households where both the index case and an infected household contact were symptomatic and the index case's symptoms appeared more than two days after the contact's symptoms.

To further reduce the potential for misclassification bias, a phylogenetic approach was used to identify apparent secondary cases in the household who were in fact infected elsewhere. If none of the sequences from a contact clustered with at least

Table 2. Number of index cases and their household contacts with listed vaccine status for each age group (N). Numbers in brackets show the additional individuals included in the sensitivity analysis (n). There are no index cases younger than 18 years old. BNT162b2, Pfizer-BioNTech mRNA vaccine; ChAdOx1, Oxford AstraZeneca adenovirus vector vaccine.

Status	Vaccine	Age group			
		<18 N(n)	18-49 N(n)	50-64 N(n)	65+ N(n)
Case	1 ChAdOx1	0	15 (2)	17 (2)	3 (1)
	2 ChAdOx1	0	20 (3)	26 (4)	1
	1 BNT162b2	0	22 (1)	13 (2)	2
	2 BNT162b2	0	13	17	1
	None	0	33	12 (1)	0
Contact	1 ChAdOx1	0 (1)	13 (2)	22 (2)	5
	2 ChAdOx1	0	13 (2)	22 (2)	3 (1)
	1 BNT162b2	0	22 (2)	9	2 (1)
	2 BNT162b2	0	14 (1)	16	3
	None	67 (7)	55 (8)	11 (2)	1 (1)

one of the sequences from the household's index case, then this was considered as evidence for an infection acquired outside of the household; therefore, the contact was excluded from the downstream analysis.

Whole-genome Illumina reads were retrieved from the European Nucleotide Archive (ENA) (RRID:SCR_006515) under the accession PRJEB37886. Consensus genomes were generated using the Snippy pipeline mapping to the reference genome NC_045512.2²⁴. Highly ambiguous and/or homoplastic sites were masked in the consensus alignment as described by de Maio *et al.*²⁵. A maximum-likelihood phylogeny was reconstructed from the consensus genomes under the Hasegawa-Kishino-Yano (HKY) model of nucleotide substitution with 1,000 ultrafast bootstrap replicates to assess branch supports and visualized in iTOL (RRID:SCR_018174)^{26,27}.

ClusterPicker was used to identify clusters of transmission in the phylogeny²⁸. These were defined as clusters of sequences with patristic distances of no more than 2 SNP (6.6×10^{-5} substitutions/site)²⁹ and bootstrap support of at least 70%.

Results

By September 10th, 2021, a total of 213 index cases and 312 contacts had been recruited and met the criteria for inclusion at that time. Two contacts were removed due to lack of genomic proximity (outlined below), which resulted in the removal of each of their households as there were no further contacts. The serial interval was two (95% range: -6, 10) days. A total of 16 households with their respective index cases and a total of 32 contacts were excluded from the main analysis because at least one infected household contact presented symptoms more than two days before the index case. Thus, the main analysis was performed on 195 index cases and their 278 contacts. Households had between one and seven contacts, with a mean of 2.2, median of 2, and standard deviation of 1.2. The mode number of household contacts was 1.

Of the included individuals, 175 index cases (90%) and 113 (41%) contacts tested positive for SARS-CoV-2 at least once in the week since recruitment. Sequencing information was available for 122 (69%) and 81 (71%) of those, respectively.

A total of 24% of contacts were less than 18 years old, and therefore not eligible for vaccination at the time. The proportion of at least partially vaccinated (adult) household index cases and contacts was 77% and 69%, respectively (Table 2). Only 10 index cases (5%) were asymptomatic, reflecting the bias of Pillar 2 testing in the UK towards detecting mostly symptomatic infections. Fully vaccinated cases had received their second dose on average 70 days before enrolment and fully vaccinated contacts 71 days.

Prevalence of lineages

Of the 195 index cases analysed here, 99 were identified as infected with B.1.1.7 (Alpha), 24 with B.1.617.2 (Delta), 20 did not test positive again after recruitment, and 52 were of unknown lineage as their PCR-positive swabs had not yet

been sequenced. Of the 72 individuals without information on the infecting lineage, we estimated that 18 were likely of Alpha and 54 were likely of Delta lineage based on the date of sampling and the national prevalence of lineages at the time. That is, 60% of index cases had an Alpha variant infection and the remainder were Delta.

Identification of non-household transmission

Sequencing information for both index case and contact were available for 92 PCR positive case-contact pairs across 79 households. In total, 345 whole-genome sequences (including longitudinal samples) were available for analyses, a majority of which were of Alpha variant (82.6%) and the remainder were Delta (17.4%).

The phylogeny provided evidence that in two households the contact of the recruited index case had acquired infection elsewhere (Figure 1, households HH002 and HH007). Five households that did not form unique clusters in the phylogeny did not meet the exclusion criteria: in two a sequence from an index case did not cluster with the remaining household sequences but another sequence from the same index case did (HH004 and HH006), while the other three households did not have sufficient bootstrap support to be a part of a cluster (HH001, HH003, and HH005). Of the remaining households, 72 (91%), formed unique, household-specific clusters that included all and only sequences of members of the household, indicating likely direct transmission within the household.

Age and lineage effects

We estimate that in the absence of vaccination of either case or contacts, Delta lineage infections were much more transmissible within the household than Alpha lineage infections (RR: 1.53, 95% Credible Interval: 1.04, 2.20 for adult cases <50 years old). Children younger than 18 years old were less likely than adults to acquire an Alpha infection (RR: 0.67, 95%: 0.52, 0.83) and just as likely to acquire a Delta infection (RR: 0.98, 95%: 0.86, 1.07). Compared to a baseline of index cases aged between 18 and 49 years old, those 50 years old and over did not transmit either an Alpha (RR: 1.15, 95%: 0.89, 1.47) or Delta (1.08, 95%: 0.96, 1.30) infection to any greater degree at a 95% level of credibility.

Effectiveness of vaccination

Either one or two doses of BNT162b2 provide contacts with a protective effect against infection from a symptomatic index case with Alpha variant SARS-CoV-2 with a vaccine effectiveness of 51%, (95% credible interval: 4%, 83%) and 69% (95% credible interval: 9%, 95%), respectively (Table 3). At 0% (-33%, 39%) and 18% (-11%, 59%) the effectiveness of one and two doses of BNT162b2 against infection with the Delta variant was lower than against Alpha. The protection offered by ChAdOx1 to either variant after two doses was also low, with effectiveness against Alpha of 24% (-41%, 72%) and against Delta of 9% (-15%, 42%).

We estimate that the effectiveness of one and two doses of BNT162b2 against onward transmission from cases infected

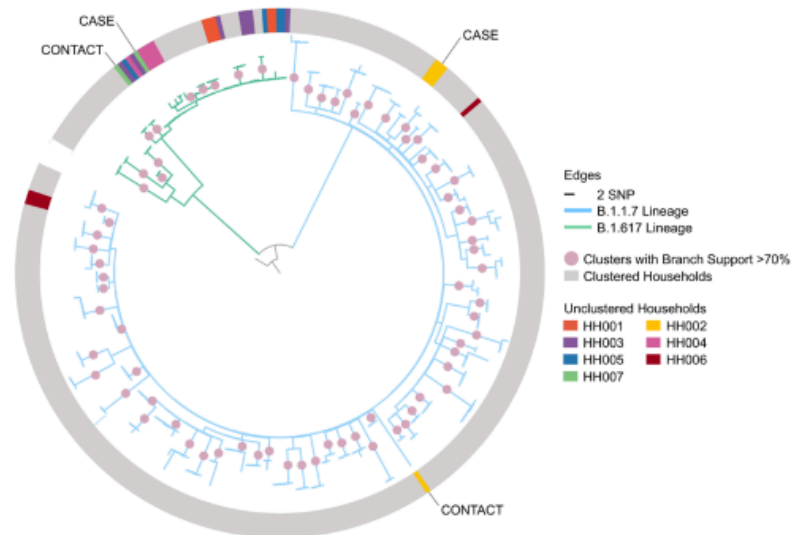


Figure 1. Maximum-likelihood phylogeny of household index cases and contacts' sequences with 1,000 ultrafast bootstrap replicates rooted to the reference sequence with a scaled bar of 2 SNP (6.6×10^5 substitutions/site). The dotted line at bottom left indicates where a single long branch was collapsed for visualisation. The non-grey shading on the outer ring represents non-clustered households where sequences are coloured by their households. HH002 and HH007 were the only households where none of the contacts' sequences clustered with that of their household's index case and this is evidence that the contact could have acquired the infection elsewhere and is thus excluded from the analysis. SNP, Single Nucleotide Polymorphism.

Table 3. Median VE and 95% credible intervals for infection protection in contacts and transmission reduction in cases, by variant, vaccine product, and number of doses. VE, vaccine effectiveness; BNT162b2, Pfizer-BioNTech mRNA vaccine; ChAdOx1, Oxford AstraZeneca adenovirus vector vaccine.

Variant	Vaccine	Doses	VE infection	VE transmission
Alpha	ChAdOx1	1	-1% (-42%, 36%)	-9% (-63%, 28%)
		2	24% (-41%, 72%)	36% (-29%, 74%)
	BNT162b2	1	51% (4%, 83%)	23% (-18%, 54%)
		2	69% (9%, 95%)	57% (2%, 85%)
Delta	ChAdOx1	1	-1% (-28%, 28%)	15% (-17%, 58%)
		2	9% (-15%, 42%)	49% (18%, 73%)
	BNT162b2	1	0% (-33%, 39%)	10% (-20%, 54%)
		2	18% (-11%, 59%)	36% (-1%, 66%)

with the Alpha variant was 23% (-18%, 54%) and 57% (2%, 85%), respectively, and for Delta variant one and two doses reduce transmission by 10% (-20%, 54%) and 36% (-1%, 66%), respectively. RRs for the protective effect of BNT162b2 over ChAdOx1 for one and two doses of against both Alpha

and Delta variants indicate that at 95% credibility there is no difference between the effectiveness of the two vaccine products. Specifically, the RRs for acquisition of Alpha and Delta after two doses of BNT162b2 vs. ChAdOx1 for an adult contact are 0.41 (0.06, 1.70) and 0.92 (0.52, 1.31), respectively.

Secondary attack rates

The estimated secondary household attack rate among adults in an unvaccinated household was 50% (37%, 64%) for the Alpha variant and 78% (54%, 95%) for the Delta variant (Figure 2).

BNT162b2 is very effective against Alpha variant infection when either the case or contact are vaccinated, and especially when both have received two doses (Figure 2). SARs for Delta variant infection in unvaccinated case-contact pairs are substantially higher. Full (two dose) vaccination with either vaccine is still effective against Delta infection when both the case and contact are vaccinated, at least halving the SAR; e.g., case and contact both fully vaccinated with BNT162b2 has an SAR of 31% (13%, 59%). Notably, the reduced susceptibility to infection of (unvaccinated) individuals under 18 years old results in Alpha SARs that are no greater than those seen in adult contacts who have received two doses of ChAdOx1. Conversely, for Delta infections, there is no reduced susceptibility for those aged under 18 years and so unvaccinated under- and over-18s have similar probability of becoming infected, with a single dose of either vaccine providing no discernible protection.

Sensitivity analysis

Sensitivity analysis was conducted by including the 16-index case-contact pairs with serial intervals less than two days. This did not qualitatively change our results. The absence of informative priors on the protective vaccine effects against infection led some of the vaccine effectiveness against infection in our study to be re-attributed to effectiveness against onward transmission or to age effects.

Infection history dynamics

We estimate that within a week of symptom onset, the relative risk of symptomatic cases testing PCR positive is near identical for vaccinated and unvaccinated participants. For cases infected with the Alpha variant, there was little difference in PCR positivity generally between vaccinated and unvaccinated cases, while in cases infected with the Delta variant the proportion of participants with PCR detectable infection in participants fully vaccinated with BNT162b2 declined about four days before that in unvaccinated participants. At two to three days the effect in participants fully vaccinated with ChAdOx1 was slightly less pronounced.

Discussion

In this prospective household-based study of SARS-CoV-2 infection, we showed that both the ChAdOx1 and BNT162b2 vaccines are effective in reducing transmission of the Alpha and Delta variants from those who develop breakthrough infections despite having received two doses. The estimated vaccine effectiveness against acquisition of a Delta infection in the household setting was however low; 9% (-15%, 42%) and 18% (-11%, 59%) after two doses of ChAdOx1 and BNT162b2, respectively. This is much lower than that estimated from cases presenting for Pillar 2 testing in the community for which the effectiveness of two doses of ChAdOx1 against symptomatic infection is estimated as 67.0% (61.3%, 71.8%) and 88.0% (85.3%, 90.1%) for BNT162b2¹⁵. Effectiveness against acquisition of an Alpha infection in the household was substantially higher in our study than that against Delta but still lower than that estimated from Pillar 2 community testing. The lower protection against acquisition in the household likely reflects

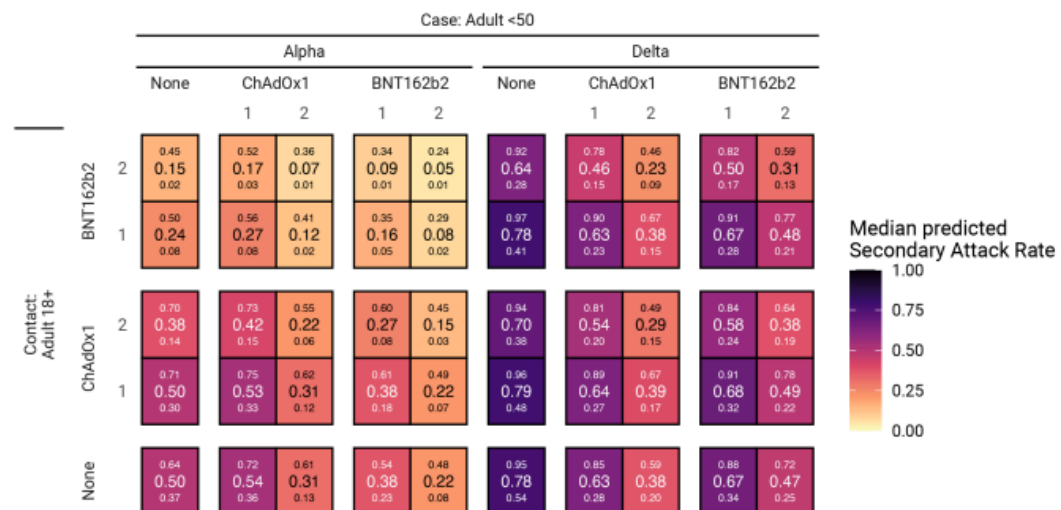


Figure 2. Predicted SARs for each combination of vaccine status of case and contact. Large numbers inside cells are the median SAR, with the small numbers below and above corresponding to the 95% credible interval. SAR, secondary attack rate; BNT162b2, Pfizer-BioNTech mRNA vaccine; ChAdOx1, Oxford AstraZeneca adenovirus vector vaccine.

the prolonged and intense exposure that occurs in this setting. Similarly, although the effectiveness estimates against Delta transmission within the household were moderate at 49% (18%, 73%) and 36% (-1%, 66%) after two doses of ChAdOx1 and BNT162b2, respectively, the protective effect in those with breakthrough infections may be higher in the community where exposure is less intense and of shorter duration. The reduction in duration of PCR positivity in breakthrough infections (average of four days shorter for the Delta variant for those infected after two doses of BNT162b2 and around two to three days for ChAdOx1) will also have more of an impact in the community than in the household setting where generation times between infections are short – around 3.5 days for the Delta variant³⁰. Our household contacts were actively followed up with repeated swabbing and showed the high secondary attack rates that occur in this setting; 81% for Delta infections in unvaccinated households but that reduced to 25–40% in households where both index case and contacts were fully vaccinated.

Comparison with other studies

Our finding of a moderate level of protection against onward transmission from fully vaccinated individuals, with either vaccine and against either variant, is in apparent contrast to a study that similarly followed up contacts reported by the UK test and trace system prospectively, about 90% of whom were in the same household as the index case³¹. The study estimated a moderate effect of vaccination against infection but no difference in secondary attack rates with the delta variant between fully vaccinated and unvaccinated index cases (24% and 23%, respectively). However, such estimates were neither controlled for age nor vaccination status of the contact. Notably, only four out of 17 (24%) unvaccinated contacts were infected by fully vaccinated index cases, whereas eight out of 20 (40%) unvaccinated contacts were infected by unvaccinated

index cases; a reduction in transmission of 41% albeit based on very small numbers. In a similar study from Singapore Delta-exposed fully vaccinated household, contacts were 56.4% less likely to test positive on quarantine exit screening³². Also, this study had insufficient power to detect a significant vaccine effect in onward transmission; the odds of a positive exit screening test for Delta-exposed household contacts was 27% (95% CI: -40, 62) lower in contacts of fully vaccinated index cases. Secondary attack rates in our study were much higher, potentially owing to the regular testing during quarantine and the absence of measures to prevent transmission in the household. This has helped through providing statistical power to the point estimates of the other two studies and show a protective effect against onward transmission. Vaccine effectiveness against onward transmission of 40–80% has been suggested by several retrospective observational studies using either information on the household structure⁷ or contact tracing^{8,9} in combination with routine national COVID-19 notification systems to estimate reductions in secondary attack rates from breakthrough infections. While observational studies are prone to biases introduced by testing behaviour particularly for mild disease manifestations, our study combines prospectively collected longitudinal data from recruited households with a robust analytical framework to confirm that both vaccines reduce transmissibility of breakthrough infections in fully vaccinated individuals.

Among symptomatic index cases and contacts, we found a lower rate of PCR positivity within two weeks of symptom onset in all vaccinated groups (Figure 3). PCR positivity for Delta declined fastest (four days ahead of unvaccinated) in individuals fully vaccinated with BNT162b2. These results largely mirror those in other studies that found enhanced clearance following vaccination³¹, but raise the question whether enhanced clearance can be the driving mechanism for reduced transmission

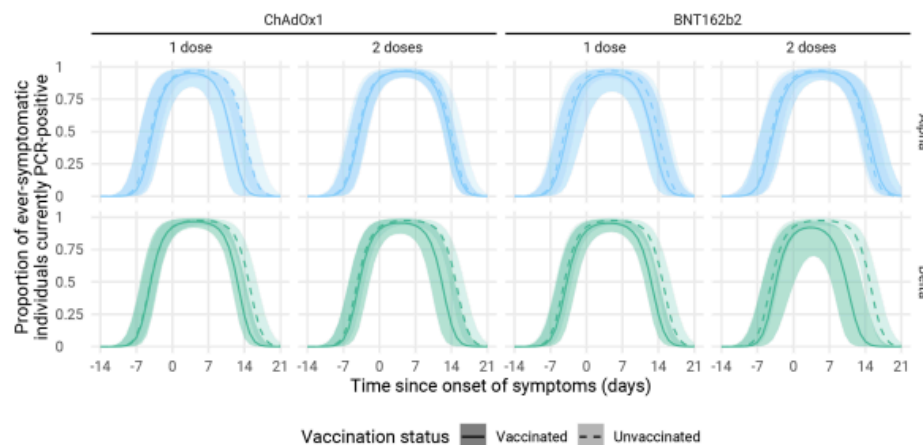


Figure 3. PCR positivity by variant and vaccination status for symptomatic infections (index cases recruited from Pillar 2 testing and the symptomatic household contacts they infected). Lines represent median trajectories, and the ribbon is the 95% credible interval. BNT162b2, Pfizer-BioNTech mRNA vaccine; ChAdOx1, Oxford AstraZeneca adenovirus vector vaccine.

in a frequent contact household setting. Another mechanism may be that while positivity with the highly sensitive PCR test is similar to that in the unvaccinated, vaccination can reduce³³ both peak viral load^{6,34} and viral shedding³⁵ although such effects have not been reported in all studies and may be masked by age effects.

Strengths and limitations of this study

Our study comes with limitations. To minimise the potential for misclassification we restricted the main analyses to only those putative transmission pairs where there was no evidence against direct transmission based on phylogenetic distance (which was available for 63% of all putative transmission pairs) and where symptom onset in the contact did not pre-date that of the index case by more than two days. Our model does not account for the introduction of SARS-CoV-2 into the household from two index cases who acquired the infection separately, opting instead to exclude the most recently PCR-positive index case. Importantly, if there is residual misclassification between infector and infected this would attribute infection protection to transmission protection and *vice versa*.

Only households with adult index cases were recruited into the study. During the time of data collection, children were ineligible for vaccination and so there would be no vaccine effectiveness for reducing transmission to estimate even if such data were included. The inclusion of households with child index cases would, however, provide useful further information on the protection against acquisition for both adult and child contacts of unvaccinated cases, particularly in understanding the risk of children who acquire infection in their school environment and may transmit to family members.

Prior infection status is not included in the model as the data were not available at the time of data lock. This can be incorporated in the model under the same structure as a vaccine product, though this may be difficult when considering protection against a specific variant provided by infection with previously seen variants when the prior infection's variant is unidentifiable.

The ability to detect an infection in contacts relies on the sensitivity of PCR and the timing of swabs. A vaccinated contact who acquires infection may be less detectable due to a reduction in viral load and/or shorter shedding period³⁵ and may have been detectable between the swabs on days three and seven.

We did not include waning of vaccine protection in our analyses. In the analysed dataset the longest reported time since vaccine receipt was 169 days. While some individuals in the analysis have since become eligible for a booster vaccination over concerns of waning protection, some of this potential effect will have been absorbed in our model in the age structuring because of the strong correlation between age and timing of vaccine eligibility as per the vaccine roll-out strategy in the UK. Lastly, data collection spanned a period of multiple months during which Delta became the dominant strain in circulation in the UK and included participants vaccinated with two different vaccine products; thus, requiring sub-strata analyses and reducing the effective sample size for each strata. We

used a Bayesian model that allowed the borrowing of strength through the model hierarchy, and priors allowing us to make use of the heterogeneity in risk factors and not only estimate vaccine effectiveness against transmission in these strata but simultaneously estimate the difference in transmissibility in Alpha and Delta variants and the effectiveness of partially completed dosing schedules. The use of informative priors was integral to disentangling the confounded age and vaccine history effects, which arose due to vaccine product prioritisation and were exacerbated by low counts for case-contact vaccine history combinations. Additionally, we assume that carriage of multiple variants does not occur, with genomic sequencing only showing a single variant.

Conclusions

Our findings provide robust evidence that vaccination with either BNT162b2 or ChAdOx1 can help to substantially reduce, but not completely prevent, household transmission with SARS-CoV-2. This highlights the importance of vaccines to limit circulation of SARS-CoV-2 particularly in close and prolonged contact indoor settings. The effectiveness of booster doses to further enhance protection against transmission will need to be evaluated to better understand the extent to which we can rely on vaccination for the control of SARS-CoV-2 infection, particularly during winter seasons when most contacts occur in households or household-like settings.

Ethics approval

UKHSA Research Ethics and Governance Group Statement: Surveillance of COVID-19 testing and vaccination is undertaken under Regulation 3 of The Health Service (Control of Patient Information) Regulations 2002 to collect confidential patient information (<http://www.legislation.gov.uk/uksi/2002/1438/regulation/3/made>) under Sections 3(1) (a) to (c), 3(1)(d) (i) and (ii) and 3(3). The study protocol was subject to an internal review by the PHE Research Ethics and Governance Group and was found to be fully compliant with all regulatory requirements. As no regulatory issues were identified, and ethical review is not a requirement for this type of work, it was decided that a full ethical review would not be necessary. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived. Oral informed consent for sampling and follow up was obtained by the nurses from household members who were free to decline to participate in the surveillance at any time. Consent for children was obtained by a parent or legal guardian. Only anonymised data were provided to non-UKHSA authors.

Data availability

Underlying data

The data necessary to replicate results are available from the authors on request, subject to a data sharing agreement. Requests for the underlying data should be made via the UKHSA office for data release: <https://www.gov.uk/government/publications/accessing-ukhsa-protected-data>.

Extended data

Analysis code available from: <https://doi.org/10.5281/zenodo.7618847>¹²

License: [MIT](#)

Author contributions

EM developed the household transmission protocol; NA and JLB contributed to the study design; SF advised on the overall analytic approach; PW was responsible for developing and curating the database; CG, FK and CS assisted in data management; LL managed the team of study nurses; SC developed and conducted the Bayesian analysis; JH, SH and SC conducted the genomic analysis; all authors contributed to the interpretation of the data. SC, SF, EM and JH drafted the paper; SC and JH generated the tables and figures. All authors revised the manuscript approved for final submission.

Acknowledgements

We thank the nurses in the Immunisation and Vaccine Preventable Diseases Division of the UK Health Security Agency who recruited and followed up the households and the administrative

staff who sent out the swabbing kits to households and arranged for their collection. We also thank the staff of the Virus Reference Department, Central Sequencing Laboratory and the Core Bioinformatics Team of PHE Colindale who performed the molecular testing and sequencing. Sequencing was financially supported in part by the COG-UK Consortium. COG-UK is supported by funding from the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute. The authors also wish to thank Prof. Neil Ferguson (Imperial College London) for his comments and questions about earlier versions of this analysis.

An earlier version of this article can be found on medRxiv (<https://www.medrxiv.org/content/10.1101/2021.11.24.21266401v2>), which was evaluated on [society](https://society.org/articles/activity/10.1101/2021.11.24.21266401) (<https://society.org/articles/activity/10.1101/2021.11.24.21266401>).

References

- Ritchie H, Mathieu E, Rodes-Guirao L, et al.: **Coronavirus pandemic (COVID-19)**. Our World Data. 2020. [Reference Source](#)
- Chung H, He S, Nasreen S, et al.: **Effectiveness of BNT162b2 and mRNA-1273 covid-19 vaccines against symptomatic SARS-CoV-2 infection and severe covid-19 outcomes in Ontario, Canada: test negative design study**. *BMJ*. 2021; **374**: n1943. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bernal JL, Andrews N, Gower C, et al.: **Effectiveness of the Pfizer-BioNTech and Oxford-AstraZeneca vaccines on covid-19 related symptoms, hospital admissions, and mortality in older adults in England: test negative case-control study**. *BMJ*. 2021; **373**: n1088. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ranzani OT, Hitchings MDT, Dorion M, et al.: **Effectiveness of the CoronaVac vaccine in older adults during a gamma variant associated epidemic of covid-19 in Brazil: test negative case-control study**. *BMJ*. 2021; **374**: n2015. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- World Health Organization: **Landscape of observational study designs on the effectiveness of COVID-19 vaccination**. World Health Organization; 2021; [cited 2021 Oct 11]. [Reference Source](#)
- Chia PY, Ong SWX, Chiew CJ, et al.: **Virological and serological kinetics of SARS-CoV-2 Delta variant vaccine breakthrough infections: a multicentre cohort study**. *Clin Microbiol Infect*. 2022; **28**(4): e12.e1–e12.e7. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Prunas O, Warren JL, Crawford FW, et al.: **Vaccination with BNT162b2 reduces transmission of SARS-CoV-2 to household contacts in Israel**. *medRxiv*. 2021; 2021.07.13.21260393. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Eyre DW, Taylor D, Purver M, et al.: **The impact of SARS-CoV-2 vaccination on Alpha & Delta variant transmission**. 2021; 2021.09.28.21264260. [Publisher Full Text](#)
- de Gier B, Andeweg S, Joosten R, et al.: **Vaccine effectiveness against SARS-CoV-2 transmission and infections among household and other close contacts of confirmed cases, the Netherlands, February to May 2021**. *Euro Surveill*. 2021; **26**(31): 2100640. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Miller E, Waight PA, Andrews NJ, et al.: **Transmission of SARS-CoV-2 in the household setting: A prospective cohort study in children and adults in England**. *J Infect*. 2021; **83**(4): 483–489. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- The COVID-19 Genomics UK (COG-UK) consortium: **An integrated national scale SARS-CoV-2 genomic surveillance network**. *Lancet Microbe*. 2020; **1**(3): E99–E100. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Clifford S, Hackman J: **Effectiveness of BNT162b2 and ChAdOx1 against SARS-CoV-2 household transmission**. *Zenodo*. [Code]. 2023. <http://www.doi.org/10.5281/zenodo.7618848>
- R Core Team: **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing; 2021. [Reference Source](#)
- Plummer M: **rjags: Bayesian Graphical Models using MCMC**. 2019. [Reference Source](#)
- Hall VJ, Foulkes S, Saei A, et al.: **COVID-19 vaccine coverage in health-care workers in England and effectiveness of BNT162b2 mRNA vaccine against infection (SIREN): a prospective, multicentre, cohort study**. *Lancet*. 2021; **397**(10286): P1725–1735. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bernal JL, Andrews N, Gower C, et al.: **Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant**. *N Engl J Med*. 2021; **385**(7): 585–594. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Davies NG, Abbott S, Barnard RC, et al.: **Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England**. *Science*. 2021; **372**(6538): eabg3055. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yousaf AR, Duca LM, Chu V, et al.: **A Prospective Cohort Study in Nonhospitalized Household Contacts With Severe Acute Respiratory Syndrome Coronavirus 2 Infection: Symptom Profiles and Symptom Change Over Time**. *Clin Infect Dis*. 2021; **73**(7): e1841–e1849. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goldstein E, Lipsitch M, Cevik M: **On the Effect of Age on the Transmission of SARS-CoV-2 in Households, Schools, and the Community**. *J Infect Dis*. 2021; **223**(3): 362–369. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wellcome Sanger Institute: **COVID-19 Genomic Surveillance**. 2021; [cited 2021 Oct 18]. [Reference Source](#)
- Andrews NJ, Stowe J, Ramsay ME, et al.: **Risk of venous thrombotic events and thrombocytopenia in sequential time periods after ChAdOx1 and BNT162b2 COVID-19 vaccines: a national cohort study in England**. *Lancet Reg Health Eur*. In press; 2022; **13**: 100260. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- UK Government: **All young people aged 16 and 17 in England to be offered vaccine by next week**. GOV.UK. 2021; [cited 2021 Oct 21]. [Reference Source](#)
- UK Department of Health and Social Care: **Universal vaccination of children and young people aged 12 to 15 years against COVID-19**. GOV.UK. 2021; [cited 2021 Oct 21]. [Reference Source](#)

24. Seemann T: **Rapid haploid variant calling and core genome alignment.** 2020.
[Reference Source](#)
25. De Maio N, Walker C, Borges R, et al.: **Masking strategies for SARS-CoV-2 alignments.** *Virological.org*; 2020; [cited 2021 Oct 13].
[Reference Source](#)
26. Nguyen LT, Schmidt HA, von Haeseler A, et al.: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.** *Mol Biol Evol.* 2015; **32**(1): 268-74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Letunic I, Bork P: **Interactive Tree Of Life (ITOL) v5: an online tool for phylogenetic tree display and annotation.** *Nucleic Acids Res.* 2021; **49**(W1): W293-W296.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
28. Ragonnet-Cronin M, Hodcroft E, Hué S, et al.: **Automated analysis of phylogenetic clusters.** *BMC Bioinformatics.* 2013; **14**: 317.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Lythgoe KA, Hall M, Ferretti L, et al.: **SARS-CoV-2 within-host diversity and transmission.** *Science.* 2021; **372**(6539): eabg0821.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Hart W, Miller E, Andrews N, et al.: **Generation time of the Alpha and Delta SARS-CoV-2 variants.** *Epidemiology.* 2021; [cited 2021 Nov 12].
[Publisher Full Text](#)
31. Singanayagam A, Hakki S, Dunning J, et al.: **Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study.** *Lancet Infect Dis.* 2021; **22**(2): 183-195.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Ng OT, Koh V, Chiew CJ, et al.: **Impact of Delta Variant and Vaccination on SARS-CoV-2 Secondary Attack Rate Among Household Close Contacts.** *Lancet Reg Health West Pac.* 2021; **17**: 100299.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Regev-Yochay G, Amit S, Bergwerk M, et al.: **Decreased infectivity following BNT162b2 vaccination: A prospective cohort study in Israel.** *Lancet Reg Health Eur.* 2021; **7**: 100150.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Levine-Tiefenbrun M, Yelin I, Katz R, et al.: **Initial report of decreased SARS-CoV-2 viral load after inoculation with the BNT162b2 vaccine.** *Nat Med.* 2021; **27**(5): 790-792.
[PubMed Abstract](#) | [Publisher Full Text](#)
35. Shamier MC, Tostmann A, Bogers S, et al.: **Virological characteristics of SARS-CoV-2 vaccine breakthrough infections in health care workers.** 2021.
[Publisher Full Text](#)

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1902896	Title	Miss
First Name(s)	Jada Nicole		
Surname/Family Name	Hackman		
Thesis Title	APPLICATION OF PATHOGEN GENOMICS TO INFER THE TRANSMISSION DIRECTION OF RESPIRATORY INFECTION		
Primary Supervisor	Stéphane Hué		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

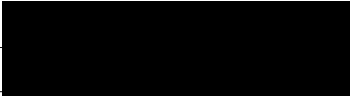
Where is the work intended to be published?	Microbial Genomics
Please list the paper's authors in the intended authorship order:	Jada Hackman, Samuel Clifford, Pauline Waight, Elizabeth Miller, Michiko Toizumi, Lay-Myint Yoshida, Stefan Flasche, Stéphane Hue

Stage of publication	Not yet submitted
----------------------	--------------------------

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed all of the bioinformatic analysis for this study, interpreted the results, and wrote/edited the manuscript for submission.
--	--

SECTION E

Student Signature	
Date	29 May 2023

Supervisor Signature	
Date	31/05/23

CHAPTER 5: CHALLENGES IN PHYLOGENETIC INFERENCE OF WHO INFECTED WHOM WITH SARS-COV-2, A PROSPECTIVE HOUSEHOLD STUDY

Authors

Jada Hackman,^{1,2} Samuel Clifford,^{1,2} Pauline Waight,³ Elizabeth Miller,^{1,2} Michiko Toizumi,^{4,5} Lay-Myint Yoshida,^{4,5} Stefan Flasche,^{1,2} Stephane Hue^{1,2}

Affiliations

1. Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, United Kingdom
2. Department for Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, United Kingdom
3. National Infection Service, UK Health Security Agency, London NW9 5EQ, United Kingdom
4. Department of Pediatric Infectious Diseases, Institute of Tropical Medicine, Nagasaki University, Nagasaki, 852-8521, Japan
5. School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, 852-8521, Japan

Corresponding author

Jada Hackman, Jada.hackman@lshtm.ac.uk, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, United Kingdom

Keywords: within-host diversity, phylogenetic, transmission direction, COVID-19

ABSTRACT

Households have been linked to the rapid transmission of SARS-CoV-2. Accurately identifying household index cases and secondary transmission ensures rigorous estimation of attack rates and vaccine effectiveness against transmission. We evaluated the potential of phylogenetic inference to identify the direction of SARS-CoV-2 transmission in household infection pairs.

Positive SARS-CoV-2 nasopharyngeal swabs were collected between February and September 2021 from a prospective longitudinal household study in the UK. We inferred the direction of transmission among household infection pairs using Phyloscanner on whole-genome sequences and cross-validated findings on a study population level through the distribution of resulting serial intervals.

Of the 146 putative within-household transmission events, sequencing information was available for 92, of which, 58 had sufficient intra-host phylogenetic diversity to infer a direction of transmission and had a date of symptom onset for cross-validation. Longer sequence read length increased the phylogenetic signal to infer transmission direction. The inferred direction of transmission was consistent across phylogenies constructed from different sequence read lengths. In the cross-validation, we found that the phylogenetically inferred index case was no more likely to be the first of the pair to report symptoms.

Phylogenetic detection of who infected whom with SARS-CoV-2 in a household context was possible, but it lacked robustness and conflicted with epidemiological information in some instances.

IMPACT STATEMENT

Household settings are a major source of rapid SARS-CoV-2 transmission, however, there is limited research exploring the transmission direction in households using phylogenetic inference methods. A better understanding of the transmission dynamics in these settings is critical for assessing secondary attack rates, risk factors associated with infection, the role of asymptomatic infections, and the effectiveness of the current vaccines against transmission. To address this gap, this study uses routine whole-genome sequencing data and epidemiological data to investigate the potential of phylogenetic inference in identifying the direction of SARS-CoV-2 transmission in households. This study revealed that longer sequencing reads improved the robustness of the phylogenetic inferences, thereby reducing ambiguity among putative transmission pairs. Additionally, the inferred transmission direction was consistent across various sequencing read lengths, highlighting the reliability of this method. However, this study also noted a high level of discordance between phylogenetically inferred index cases and epidemiologically inferred index cases. This suggests that while phylogenetic methods can be useful in detecting the transmission direction of SARS-CoV-2 in household settings, the routine sequencing data used may lack the necessary robustness and should be used with caution in future studies. Overall, this study contributes to the understanding of the transmission dynamics of SARS-CoV-2 in households and can inform public health interventions aimed at controlling its spread.

INTRODUCTION

Crowded and poorly ventilated indoor settings such as hospitals, care homes, schools, and households have been hubs of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission.¹² The analysis of viral transmission in these settings is critical in characterising secondary attack rates (SAR), the risk factors associated with infection, the contribution of asymptomatic infections, and the effectiveness of vaccines against transmission.³

Studies have reported a wide range of SARs for different settings and SARS-CoV-2 variants. A review and meta-analysis from Madewell *et al.* estimated a 16.6% household secondary attack with higher rates from symptomatic cases than asymptomatic ones and to adult contacts compared to children contacts.⁴ Another review and meta-analysis estimated the SAR from household child index cases to be as low as 7.6%.⁵ A household study reported a 21% SAR for Delta variant infection and an increased transmission for unvaccinated individuals.⁶ Inconsistent definitions of household index cases and contacts across studies and lack of accuracy in self-reported symptom onsets are likely to result in discordant serial intervals and SAR estimates.⁴

Phylogenetic inference has been widely used to characterise and investigate localised SARS-CoV-2 transmission events in various settings such as hospitals, urban districts, care homes, and universities.⁷⁻¹¹ A few studies have used viral sequencing of SARS-CoV-2 to better identify direct transmission within household settings based on genomic similarities.^{12,13} This can reveal apparent transmission events where in fact the infection was acquired not directly from the index case or even as part of a different transmission chain; thus reducing misclassification in SAR estimates.^{11,13}

Furthermore, while most SARS-CoV-2 transmission studies use consensus viral genomes to infer the direction of transmission, intra-host viral diversity has been vastly ignored. Approaches to estimate directionality have been refined to include some level of within-host diversity, opening new perspectives for the study of transmission through genomic analyses.¹⁴⁻¹⁶ Inferring the transmission direction from viral sequences can improve accuracy when identifying household index cases and mitigate potential misclassification biases for SAR or vaccine effectiveness.

We used a large household study conducted during the Alpha and Delta waves of the SARS-CoV-2 pandemic in England to phylogenetically infer the likely direction of transmission among household transmission pairs and cross-validated these against the order and time of reported dates of illness onset. We found differences in the inferred direction of transmission when comparing the phylogenetic inference to the order of symptom onset.

METHODS

Study design

The SARS-CoV-2 household transmission study has been described in detail elsewhere.¹⁷ In summary, between 2 February 2021 and 10 September 2021, adult PCR-positive index cases were identified and enrolled via Pillar 2 community testing in England, together with their consenting households' contacts. Thus, as per the study design, the index case is defined as the first individual recruited to the study. Self-taken nasal-throat swabs were obtained from both index cases and contacts on days 1, 3, and 7 and tested by RT-qPCR. Contacts were classified as infected if they had at least one PCR-positive sample. Participants were asked to self-report the first day of symptoms including fever, cough, runny nose, sore throat, shortness of breath, loss of taste or smell, nausea, diarrhoea, muscle pain, and/or headache.

Sequence data

The COVID-19 Genomics UK (COG-UK) consortium has facilitated large-scale routine sequencing of COVID-19-positive samples.^{18,19} All PCR-positive samples were sequenced on the Illumina NextSeq 550 or HiSeq 2500 platform and sequencing reads were deposited to the COG-UK consortium.²⁰ The sequenced samples were then labelled according to World Health Organisation variant nomenclature, "Alpha" (Pango lineage B.1.1.7), or "Delta" (Pango lineage B.1.617.2).²⁰ Whole-genome sequencing reads from households with suspected transmission (e.g. at least one positive swab from at least one contact) were retrieved from the European Nucleotide Database under the accession PRJEB37886 accessed September 2021 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB37886>).²¹

Consensus genomes were generated using the Snippy pipeline mapping to the reference genome NC_045512.2.21 (Wuhan strain). Highly ambiguous and/or homoplasic sites were masked in the consensus alignment as described by de Maio *et al.* to reduce artefact mutations that arise due to contamination during the sample preparation, sequencing, and or consensus calling methods.^{22,23}

Genetic diversity calculations

Mean read lengths were calculated using BAMPEFragmentSize from deepTools.²⁴ Pairwise genetic distances were calculated from all of the available consensus genomes stratified by variant. Longitudinal samples from the same individuals were included in the pairwise genetic distance calculation. Pango lineages were assigned using the pangolin web application <https://pangolin.cog-uk.io/> version 4.3²⁵. SNP distance comparisons were carried out for both Alpha variant sequences and Delta variant sequences from consensus genomes that were generated, as previously described.²¹ Following Lythgoe *et al.*, three or more pairwise differences between paired case-contact sequences were considered as evidence for indirect

transmission and sequences with three or more consensus SNPs were excluded from the direction of transmission analysis.¹⁴ More specifically, infections are usually cleared within two weeks and SARS-CoV-2 evolutionary rate is relatively slow thus an individual would most likely not accumulate more than 2 consensus SNP mutations during the short infection period.

Phylogenetic inference of SARS-CoV-2 household transmission

Within household transmission pairs were previously identified from the reconstructed phylogeny of the studied genomes that were collected from the UK.²¹ In summary, a maximum-likelihood phylogeny was reconstructed from the consensus genomes under the Hasegawa-Kishino-Yano model of nucleotide substitution, with 1,000 ultrafast bootstrap replicates, and rooted against the SARS-CoV-2 reference genome NC_045512.2.21. The clustering of genomes within the phylogeny was deemed indicative of a household transmission event if at least one sequence from an index case and one sequence from their household contacts formed a monophyletic cluster of support greater than 70%. If none of the contact sequences clustered with their household index case sequence, then the whole household was excluded from the analysis.

The software PhyloScanner was used to infer the direction of SARS-CoV-2 transmission between putative case-contact pairs from viral intra-host genetic similarities.²⁶ PhyloScanner reconstructs phylogenies within sliding windows of deep sequencing reads alignments of the case-contact pairs. For each tree, the direction of transmission is inferred from the tree topology and the relative positioning of the individuals' sequences in the trees, with 4 possible outcomes:

- I. Individual A infected individual B
- II. Individual B infected individual A
- III. A and B are linked but the direction of transmission is ambiguous
- IV. A and B are unlinked

The primary phylogenetic inference was carried out for each household independently using the largest available sliding window lengths for inferring directionality. All genomes from individuals with multiple samples were included in the phylogenetic inference e.g. if an individual had two sequenced genomes they would be labelled as A1, A2 etc. Moreover, households can have multiple transmission events e.g. households can only have one index case but multiple contacts. Potential transmission pairs were excluded from the analysis if either of the paired individuals had mean read lengths less than 70 bp. The within-host diversity penalty (k parameter) was set to allow up to 3.4% within-host genetic diversity ($s = 29.903$). The direction of transmission was supported by the largest proportion of phylogenies, e.g. the relationship with the highest number of supporting sub-trees.

Secondary analysis was carried out to explore the trade-off between the number of sliding windows within a genome (shorter windows) and the robustness of phylogeny (longer windows, spanning multiple mutations) by using multiple sliding window sizes for the clusters of sequencing read lengths observed in the data, e.g. 70, 95, or 120 base pairs (bp). The consistency of inferred transmission direction in dependence of window size choice was assessed for pairs with genome mean read lengths greater than or equal to 120 bp each (thus allowing inference for each of the chosen window sizes). This analysis was carried out for all permutations of case-contact pairs within their respective households, regardless of multiple longitudinal samples from the same individual. This resulted in more pairs available for testing the sensitivity to sequence read length from which the sub-trees are reconstructed.

Serial intervals calculation

Serial intervals were calculated for symptomatic case-contact pairs as the number of days after the recruited Pillar 2 positive household member symptoms appeared that the contact's symptoms appeared. Serial intervals were calculated (i) based on enrolment, the recruited cases being the index case, and (ii) based on the direction of transmission inferred from the genomic analysis.

RESULTS

Study samples

Of the 146 putative within-household transmission events (based on the results and samples from Chapter 4) based on positive swabs, sequencing information for both index case and contact was available in 92 (63%) unique pairs across 79 households. All of the analysis assumes the first recruited case is the index case. All of the index cases that were recruited were at least 21 years of age. The median age of the index case was 48 years old. For every household index case, there was a mean of one within-household contact (range 1-3 individuals) with an average of two longitudinal sequences per individual (range 1-4 sequences/individual). This resulted in a total of 345 whole-genome sequences (including longitudinal samples) available for the study. Mean sequencing read lengths of all 345 whole-genome distributions showed three distinct clusters and these clusters were used to determine the sliding window sizes, 70, 95, or 120 bp, in the PhyloScanner analysis for inferring the direction of transmission (Figure 1).

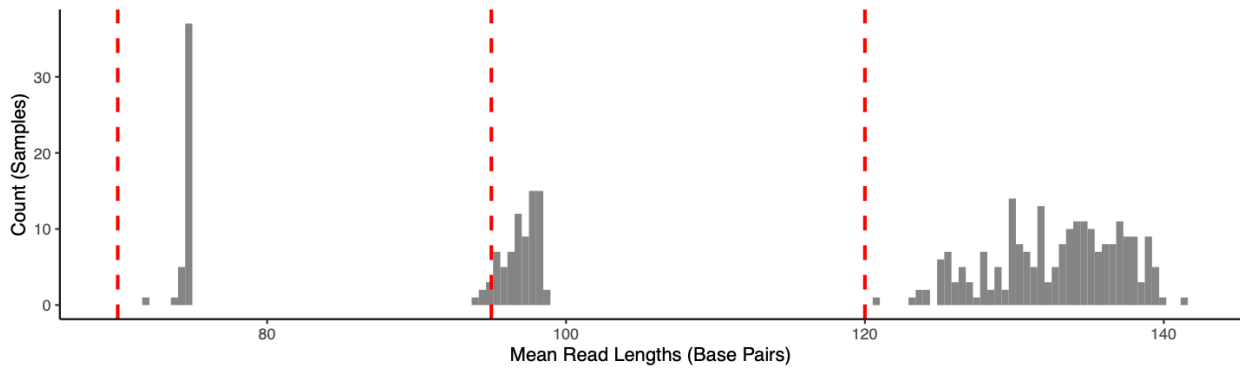


Figure 1. Mean read length distribution of the 345 whole-genome sequences (including longitudinal samples) available on the European Nucleotide Archive. Red dashed lines mark the three window sizes used in the analysis (70, 95, and 120 base pairs).

Genetic diversity

The majority of the sequenced viruses were Alpha variants (82.6%; World Health Organisation lineage designation), and the remainder were Delta variants (17.4%). The distribution of pairwise SNP distance for Alpha variant sequences was unimodal with a mean of 12 SNPs per genome [Range 0 and 29] except for a small number of zero-distance SNP pairs (Figure 2). Delta variant genetic distance distribution was multimodal, suggesting the presence of distinct sub-lineages in the dataset. Pango lineages were assigned to the consensus genomes and revealed that 4/60 Delta infections were B.1.617.2, and the remainder 56/60 were AY.* variants with most (40/56) belonging to AY.4. The mean pairwise SNP distance was 11 SNPs per genome [Range 0-31], even though the distribution is near-zero around 11. As expected, given the more recent introduction and spread of variant Delta in the study area, a higher proportion of the Delta variant genomes differed by less than 2 SNPs per genome (30.7%) compared to Alpha genomes (1.5%).

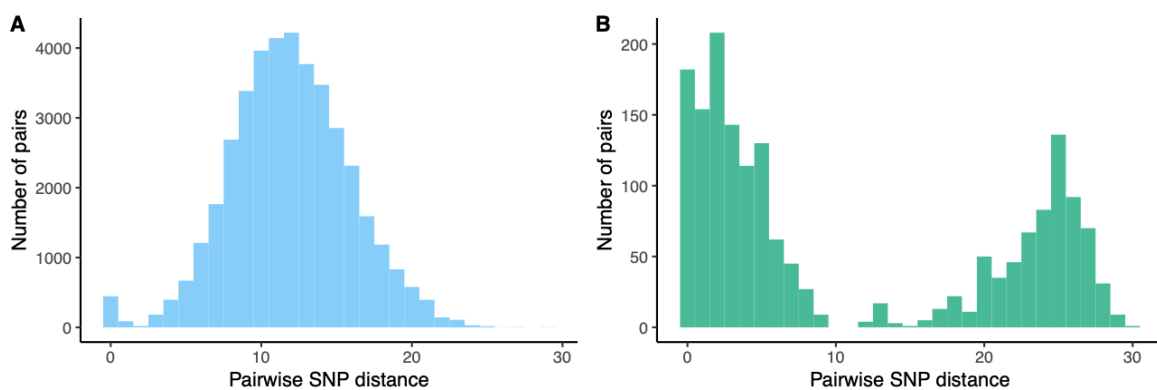
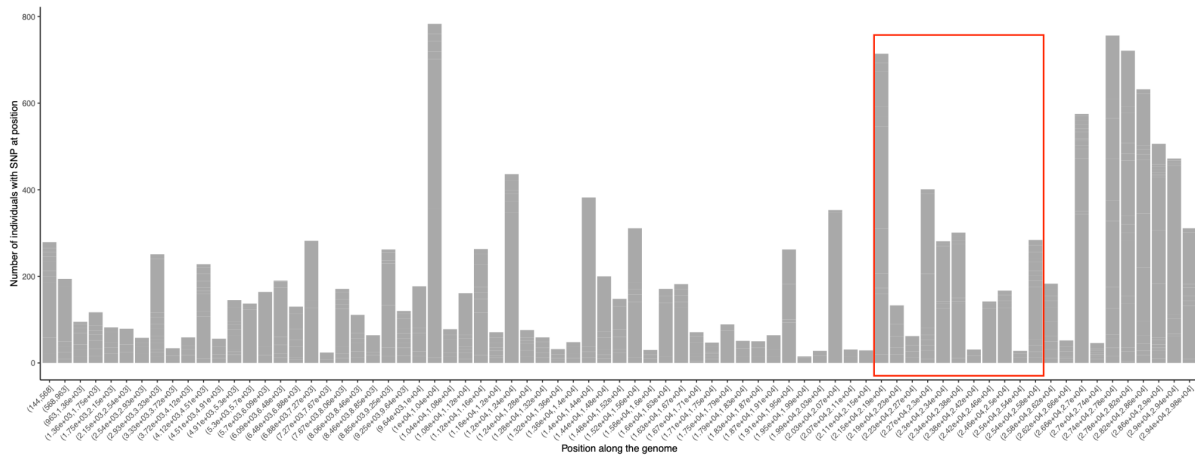


Figure 2. Pairwise genetic differences among available SARS-CoV-2 consensus genomes, expressed as the number of SNPs per genome, for Alpha variant (A) and Delta variant (B) sequences (B.1.617.2 and AY.* variants).

The average distance between the SNPs along the whole genome for all 345 samples included in the study was 62 base pairs (range, 174 - 29975 base positions, standard deviation, 73 bases).



Supplemental Figure 1. Number of samples with a detected SNP along the whole genome in reference to NC_045512.2.21. The distribution of SNPs is split into 75 breaks with the red box highlighting the spike regions of the genome (21,563 - 25,384 base positions).

Phylogenetic inference of SARS-CoV-2 household transmission

As previously reported, phylogenetic evidence to support multiple introductions is defined as the purported household case-contact paired individuals' viral sequences being too distantly related (≥ 3 SNPs) to represent direct transmission. One contact belonged to Alpha and the other contact belonged to Delta variant resulting in a total of 339 of 345 (including longitudinal samples) sequences across 77 of 79 households included in the direction of transmission analysis.²¹

Inference of direction of transmission

Of the 77 included households, all had sufficient minimum read length for analyses with a sliding window size of 70 bp, 69 (89%) for a window size of 95 bp, and 52 (68%) for a window size of 120 bp. For the primary phylogenetic inference, the largest window size available was used for the inference on household transmission and each household was analysed independently. The index case was identified, using phylogenetic inference, for 60 of the 77 households when the direction of transmission was analysed with the largest possible window size. While the index case was not identifiable for the remaining 17 households due to high phylogenetic tree support for either “unlinked”, or “ambiguous” direction of transmission or conflicting directions where there was support for A1 infected B1 and B1 infected A2. The 60 putative transmission pairs were included in the serial interval calculation comparison.

Only 52 households had sufficient read lengths (≥ 120 bp) to be included in the sensitivity analysis which looks at the effect of the sliding window sizes and the ability to infer transmission direction. A total of 139 transmission events generated from permutations of index cases and contacts within the study were included in this secondary analysis.

Across all three window sizes analysed, there was insufficient evidence to support a direction of transmission (e.g. the relationship with the highest tree support was “ambiguous”) in most instances, while for 10 (7.2%), 15 (10.8%), and 26 (18.7%) transmission pairs a direction of transmission was inferred as the most likely outcome for window sizes of 70, 95, and 120 bp, respectively (Table 1, *Max support*). The expected little within-host diversity of SARS-CoV-2 would result in little phylogenetic signal thus making inference on transmission direction difficult. Intuitively, we observed a high number of “ambiguous” relationships. When we excluded ambiguous as a potential relationship due, we found that 39 (28.0%), 32 (23.0%), and 35 (25.2%) pairs were more likely unlinked than samples from a direct transmission event for window sizes 70, 95, and 120 bp, respectively (Table 1, *Ambiguous excluded*). There were only 10 pairs that never had an “ambiguous” relationship for any of the window sizes (70, 95, or 120 bp) (Table 1, *Ambiguous ever*). There was an increased relationship of “a direction of transmission” in either direction and also an increase in the proportion of supporting trees with increasing sliding window sizes.

Of the 139 pairs analysed, 123 pairs did not yield information on the underlying transmission dynamics when the possible outcomes were i) either direction of transmission ii) unlinked or iii) ambiguous (Figure 3A). However, 120 of the 139 pairs were consistent in the directionality inferred when the possible outcomes were restricted to i) either direction of transmission or ii) unlinked (Figure 3B). Lastly, of the 16 pairs that did yield information on the underlying transmission (Figure 3A), 10 of the 16 pairs had sufficient sequence read length for a comparison analysis resulting in a consistent direction of transmission across all three window sizes for all 10 pairs (Figure 3C).

Table 1. Overview of the sliding window sizes tested in the sensitivity analysis and the phylogenetic inference for the relationship with the highest tree support

	Sliding window sizes					
	70 bp	%	95 bp	%	120 bp	%
Max support N=139 pairs						
Average no. of trees among pairs	393		279		210	
No. of pairs with ambiguous relationship	129		124		113	
Average supporting trees	271	69	182	65	120	57
No. of pairs with a DoT	10		15		26	
Average supporting trees	248	63	178	64	137	65
Ambiguous excluded N=139 pairs						
Average no. of trees among pairs	392		279		210	
No. of pairs with unlinked relationship	39		32		35	
Average supporting trees	66	17	52	19	40	19
No. of pairs with a DoT	100		107		104	
Average supporting trees	100	26	86	31	83	40
Ambiguous ever removed N=10 pairs						
Average no. of trees among pairs	404		294		221	
No. of pairs with a DoT	10		10		10	
Average supporting trees	248	61	193	66	162	73

Max support: possible relationships were “ambiguous”, “A infected B or B infected A”, or “unlinked”

Ambiguous excluded, possible relationships were “A infected B or B infected A” or “unlinked”

Ambiguous ever removed, if a putative pair resulted in an “ambiguous” relationship in *Max support* analysis using sliding window sizes of 70, 95, or 120 bp. The pairs were removed from the analysis. The possible relationship was “A infected B or B infected A” or “unlinked”

% is calculated from overall mean trees for a given analysis

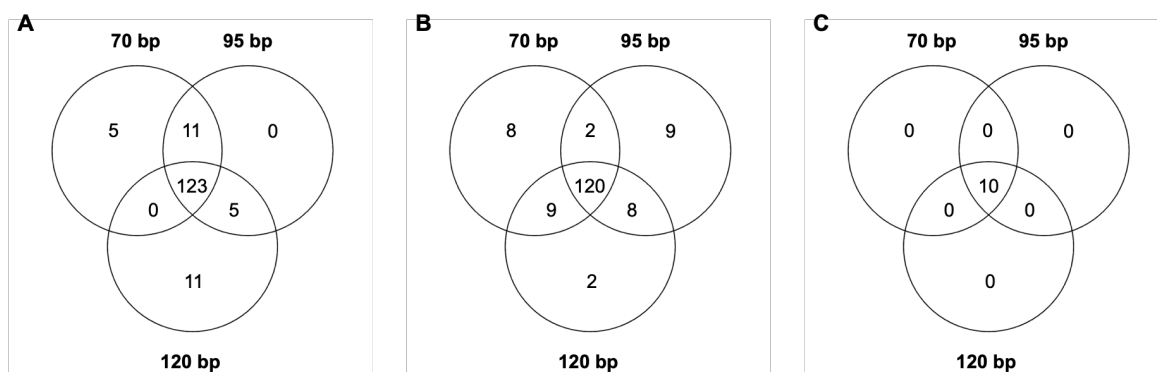


Figure 3. Comparing the most frequently inferred relationship either “ambiguous”, “unlinked”, or “a direction of transmission” to test the sensitivity of sliding window sizes in Phyloscanner.

Window sizes analysed were selected based on the mean read length distribution and the outcome is the relationship with the highest tree support. A) Possible relationships include "unlinked", "ambiguous", A infected B, or B infected A. B) Possible relationships include "unlinked", A infected B or B infected A. C) Based on pairs where "ambiguous" never had the highest tree support in A).

Inferred serial interval

Of the 60 case-contact pairs with phylogenetic support for a given direction of transmission in the primary analysis, 4 (7%) did not have data on symptom onset for either case or contact and thus were excluded resulting in 58 case-contact pairs for the comparison. Of those, 13 (22%) pairs had a serial interval of 0 days meaning same-day case-contact symptom onset. While 19 (32%) pairs retained the epidemiological index case classification and the remaining 24 (40%) pairs had phylogenetic support for index case reclassification. Of the 24 households with index case reclassifications, 20 were reclassified with negative serial intervals while 4 were reclassified with positive serial intervals (Figure 4B).

Assuming that the recruited index case was the first case in the household, the average serial interval was 2 days with a median of 2 days and a standard deviation of 4 days. In comparison, based on the self-reported date of symptom onset and the phylogenetic inference of the direction of transmission the average serial interval was -1 day with a median of -1 day and a standard deviation of 4 days, suggesting a large amount of misclassification in the inferred direction of transmission (Figure 4A).

Of the 43 pairs with non-zero serial intervals, the phylogenetically inferred index case was recruited first in 4 (9%) instances while the contacts were recruited first in 3 (7%) instances and both index case and contacts were recruited on the same days for the remainder 36 (84%) instances. Additionally, the phylogenetically inferred index case experienced symptoms first in 15 (35%) instances.

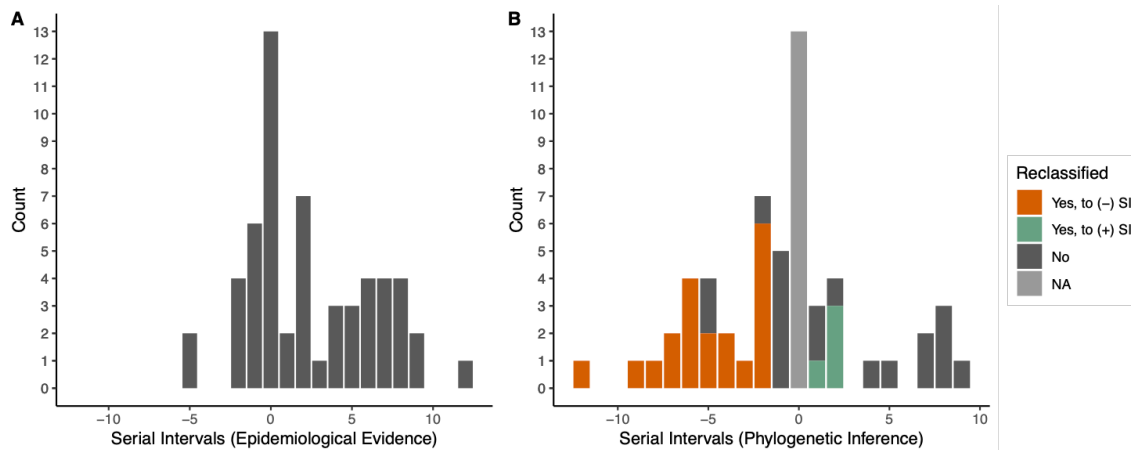


Figure 4. A) Serial intervals calculated from the index case-contact statuses classified according to the epidemiological evidence. B) Serial intervals calculated from the index case-contacts status reclassified, if there was support from the phylogenetic data. Dark grey fill represents serial interval concordance between the phylogenetic inference and the epidemiological evidence while light grey fill represents those with a serial interval of zero days. Orange fill represents households that now have a negative serial interval and green fill represents households that now have a positive serial interval due to phylogenetic evidence to support the reclassification of the household index case.

DISCUSSION

We inferred the direction of putative SARS-CoV-2 household transmission events from viral genome sequences and compared our estimates to patient-derived epidemiological evidence. We then investigated the impact these estimates have on the calculation of serial intervals, a key epidemiological parameter. We observed that increased sequence read length increased phylogenetic signal to infer the direction of transmission and directionality was consistent among transmission pairs across sequence read lengths. Moreover, we found a high level of discordance between the serial intervals calculated from epidemiological evidence compared to phylogenetic inference.

One parameter that affects the phylogenetic signal is the length from which the phylogenetic trees are reconstructed. As we increased the window sizes from which the phylogenies are generated, we noticed a decreased proportion of ambiguous phylogenies and thus an increased proportion of phylogenies that support a direction of transmission. This observation is intuitive because longer sequencing read lengths result in more phylogenetic signals, although this will also decrease the number of trees we can draw inferences from.

The three window sizes analysed resulted in a high proportion of "ambiguous" relationships

between the index and contact cases. However, when "ambiguous" relationships were excluded, the direction of transmission was concordant between window sizes. We observed high concordance (>85%) similarities between the three window sizes that were analysed therefore we selected the largest window sizes allowed for each household's sequences.

Some studies have highlighted the challenges of sequence-based transmission inference,²⁷ and similarly observe highly similar consensus SARS-CoV-2 genomes among household members.

While other studies have looked at the direction of SARS-CoV-2 transmission using within-host diversity which was calculated from the intrahost single nucleotide variants (iSNVs) and the direction of transmission was tested on a range of minor allele frequency thresholds from as low as 1% up to 50%.^{14–16} However, these methods were not tested on samples from routine sequences. An analysis with simulated genomic data using Phyloscanner demonstrated that as sequencing length increased, the accuracy of phylogenetic reconstruction also increased. However, accuracy was not affected by the number of viruses sampled per host, highlighting the inherent lack of genomic diversity in SARS-CoV-2 genomes due to short infection periods and slow mutation rates.²⁸ Our data revealed an increase in consensus SNPs around the spike region. Additionally, we observed high frequencies of SNPs which align with previously identified regions of elevated mutations²⁹. This includes regions downstream of the spike (25,800 base position and higher) which corresponds to the nucleocapsid, membrane, and non-structural protein 3 (*nsp3*) and upstream of the spike region (~10,000 base position) which corresponds to the non-structural protein 5 (*nsp5*).²⁹

Viral RNA can be detected in the respiratory tract up to 2-3 days before patients' symptom onset³⁰ and individuals can be infectious asymptotically or presymptomatically^{31–33}, with most patients with mild cases of COVID-19 infectious for up to 10 days. Of the 62 putative transmission pairs analysed, 14 of them had same-day contact-case symptom onsets, indicating presymptomatic transmission. There could be self-reported bias, however, when a household individual's symptom onset date influences the reported date of another household individual's symptom onset date.³⁴ Lastly, we cannot exclude the idea that both individuals could have been infected by a common infector that was not captured in the study resulting in the index and contact same-day symptoms and an inferred direction of transmission.

A previous study calculated shorter mean household serial intervals for Delta infections than Alpha (1.8 days, 95% CI 1.0–2.4 vs 3.5 days, CI 2.7–4.1). Other studies from household data estimated Alpha infection serial interval of 2.38 days (95% CI, 2.30–2.47)³⁵, while Delta infection serial interval was 3 days (95% 95% CI, 2 - 3 days).³⁶

There are limitations to this study. First, a major limitation of this study is the high proportion of individuals with relatively short sequencing read lengths (<100 bp) which massively limits our ability to reconstruct trees using Phyloscanner. Second, inherent limitations using genomes from SARS-CoV-2 in inferring the direction of transmission include the slow evolutionary rate of 2 mutations per month and short infection times,³⁷ and the difficulty in adequately sequence low-viral-load samples.³⁸ Third, other limitations include batch effects, sampling biases, and study-specific definitions of linked cases and the direction of transmission.³⁹ Fourth, the study designs limit the ability to test for a common infector of infection for index and contact cases with zero-day serial intervals.

In summary, we found little correlation between the phylogenetically inferred direction of transmission and that suggested by either the order of symptom onset or the order of recruitment. While phylogenetic inference holds great promise for such inference, our results show a potentially high rate of misclassification in either inference or reporting of symptom onset. Thus, phylogenetic inference to detect who infected whom should be applied with great caution in the absence of supporting additional data or further validation analyses.

Declaration of interests:

The authors have no conflicts of interest to declare.

Author contributions:

EM developed the household transmission protocol; PW developed and curated the database, SF, SH, LY, and MT advised on the data analysis; JH and SC conducted the data analysis; JH, SF, and SH drafted the paper; all authors reviewed and contributed to the manuscript prior to submission.

Data Sharing:

The whole-genome sequencing data is publicly available for download on the European Nucleotide Archive under study accession "PRJEB37886". The specific data for this study are available from the authors upon request and subject to a data-sharing agreement.

Funding statement:

This study was funded by the UK Health Security Agency as part of the COVID-19 response effort. Jada Hackman is funded by the Nagasaki University-London School of Hygiene and Tropical Medicine Doctoral Programme under the WISE scheme. Samuel Clifford is funded by the UK Medical Research Council (MC_PC_19065 - Covid 19: Understanding the dynamics and drivers of the COVID-19 epidemic using real-time outbreak analytics). Samuel Clifford and Stefan Flasche are funded by a Sir Henry Dale Fellowship through the Wellcome Trust and the Royal Society (208812/Z/17/Z). Elizabeth Miller receives support from the National Institute for Health Research (NIHR) Health Protection Research Unit in Immunisation at the London School of Hygiene and Tropical Medicine in partnership with the UKHSA (Grant Reference NIHR200929).

Ethics:

This study was approved by the UK Health Security Agency Research Ethics and Governance Group. Participant consent was obtained by the nurses and consent for children was obtained via a parent or legal guardian. The data was anonymised to non-UK Health Security Agency authors.

Acknowledgements:

We thank the participants who made this study possible. We thank the nurses of the UK Health Security Agency who recruited, followed up, and organized the distribution and collection of the nasal swabbing kits. We also thank the teams at UK Health Security Agency Colindale who carried out the molecular testing and genomic sequencing. We thank COG-UK Consortium for financially supporting the genomic sequencing.

REFERENCES

1. Endo, A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]. *Wellcome Open Res* **5**, (2020).
2. Leclerc, Q. *et al.* What settings have been linked to SARS-CoV-2 transmission clusters? [version 2; peer review: 2 approved]. *Wellcome Open Res* **5**,.
3. WHO. *Household transmission investigation protocol for 2019-novel coronavirus (COVID-19) infection*. [https://www.who.int/publications-detail/household-transmission-investigation-protocol-for-2019-novel-coronavirus-\(2019-ncov\)-infection](https://www.who.int/publications-detail/household-transmission-investigation-protocol-for-2019-novel-coronavirus-(2019-ncov)-infection) (2020).
4. Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E. & Dean, N. E. Household Transmission of SARS-CoV-2: A Systematic Review and Meta-analysis. *JAMA Netw. Open* **3**, e2031756 (2020).
5. Viner, R. *et al.* Transmission of SARS-CoV-2 by children and young people in households and schools: a meta-analysis of population-based and contact-tracing studies. *J. Infect.* S0163445321006332 (2021) doi:10.1016/j.jinf.2021.12.026.
6. Lyngse, F. P. *et al.* SARS-CoV-2 Omicron VOC Transmission in Danish Households. <http://medrxiv.org/lookup/doi/10.1101/2021.12.27.21268278> (2021) doi:10.1101/2021.12.27.21268278.
7. Løvestad, A. H., Jørgensen, S. B., Handal, N., Ambur, O. H. & Aamot, H. V. Investigation of intra-hospital SARS-CoV-2 transmission using nanopore whole-genome sequencing. *J. Hosp. Infect.* **111**, 107–116 (2021).
8. Popa, A. *et al.* Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci. Transl. Med.* **12**, eabe2555 (2020).
9. Meredith, L. W. *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* **20**, 1263–1271 (2020).
10. Aggarwal, D. *et al.* Genomic epidemiology of SARS-CoV-2 in a UK university identifies

- dynamics of transmission. *Nat. Commun.* **13**, 751 (2022).
11. Lindsey, B. B. *et al.* Characterising within-hospital SARS-CoV-2 transmission events using epidemiological and viral genomic data across two pandemic waves. *Nat. Commun.* **13**, 671 (2022).
 12. Soto, J. C. *et al.* Outbreak investigation of SARS-CoV-2 transmission in an emergency childcare centre. *Can. J. Public Health.* **112**, 566–575 (2021).
 13. Peng, J. *et al.* Estimation of Secondary Household Attack Rates for Emergent Spike L452R Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Variants Detected by Genomic Surveillance at a Community-Based Testing Site in San Francisco. *Clin. Infect. Dis.* ciab283 (2021) doi:10.1093/cid/ciab283.
 14. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
 15. San, J. E. *et al.* Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evol.* **7**, veab041 (2021).
 16. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10**, e66857 (2021).
 17. Hart, W. *et al.* *Generation time of the Alpha and Delta SARS-CoV-2 variants.* <http://medrxiv.org/lookup/doi/10.1101/2021.10.21.21265216> (2021)
doi:10.1101/2021.10.21.21265216.
 18. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
 19. Vöhringer, H. S. *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).
 20. Nicholls, S. M. *et al.* CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biol.* **22**, 196, s13059-021-02395-y (2021).
 21. Clifford, S. *et al.* *Effectiveness of BNT162b2 and ChAdOx1 against SARS-CoV-2 household transmission: a prospective cohort study in England.*

<http://medrxiv.org/lookup/doi/10.1101/2021.11.24.21266401> (2021)

doi:10.1101/2021.11.24.21266401.

22. De Maio, N. *et al.* *Masking strategies for SARS-CoV-2 alignments*. <https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480> (2020).
23. De Maio, N. *et al.* *Issues with SARS-CoV-2 sequencing data*. <https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473> (2020).
24. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
25. O’Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
26. Wymant, C. *et al.* PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. 17.
27. Bendall, E. E. *et al.* SARS-CoV-2 Genomic Diversity in Households Highlights the Challenges of Sequence-Based Transmission Inference. *mSphere* **7**, e00400-22 (2022).
28. Dhar, S., Zhang, C., Mandoiu, I. I. & Bansal, M. S. TNet: Transmission Network Inference Using Within-Host Strain Diversity and its Application to Geographical Tracking of COVID-19 Spread. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 230–242 (2022).
29. Yuan, F., Wang, L., Fang, Y. & Wang, L. Global SNP analysis of 11,183 SARS-CoV-2 strains reveals high genetic diversity. *Transbound. Emerg. Dis.* **68**, 3288–3304 (2021).
30. Rhee, C., Kanjilal, S., Baker, M. & Klompas, M. Duration of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infectivity: When Is It Safe to Discontinue Isolation? *Clin. Infect. Dis.* **72**, 1467–1474 (2021).
31. Pan, X. *et al.* Asymptomatic cases in a family cluster with SARS-CoV-2 infection. *Lancet Infect. Dis.* **20**, 410–411 (2020).
32. Bai, Y. *et al.* Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA* **323**, 1406 (2020).
33. Kimball, A. *et al.* Asymptomatic and Presymptomatic SARS-CoV-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March

2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 377–381 (2020).
34. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* 211 (2016) doi:10.2147/JMDH.S104807.
35. Manica, M. *et al.* Intrinsic generation time of the SARS-CoV-2 Omicron variant: An observational study of household transmission. *Lancet Reg. Health - Eur.* **19**, 100446 (2022).
36. Kang, M. *et al.* Transmission dynamics and epidemiological characteristics of SARS-CoV-2 Delta variant infections in Guangdong, China, May to June 2021. *Eurosurveillance* **27**, (2022).
37. Callaway, E. The coronavirus is mutating — does it matter? *Nature* **585**, 174–177 (2020).
38. Lam, C. *et al.* SARS-CoV-2 Genome Sequencing Methods Differ in Their Abilities To Detect Variants from Low-Viral-Load Samples. *J. Clin. Microbiol.* **59**, e01046-21 (2021).
39. Wang, Y., Zhao, Y. & Pan, Q. Advances, challenges and opportunities of phylogenetic and social network analysis using COVID-19 data. *Brief. Bioinform.* bbab406 (2021) doi:10.1093/bib/bbab406.

CHAPTER 6: DISCUSSION AND CONCLUSION

6.1 Summary on findings

This thesis aimed to investigate to which capacity we can use genomics to infer the transmission direction of respiratory pathogens. Pathogen genomics plays an important role in epidemiological studies and its utilisation has been routine in investigating SARS-CoV-2 and is widely used but not yet routine for *Streptococcus pneumoniae*. The use of pathogen whole-genome sequences for identifying linkage or determining transmission directionally can aid in the identification and prevention of infection of respiratory pathogens.

6.1.1 Summary on transmission direction of *Streptococcus pneumoniae* (Chapter 2)

The initial objective of the thesis was to investigate the capacity at which we can use genomic data to link *Streptococcus pneumoniae* infections and infer the direction of transmission where there is suspicion of infection between two individuals. I found that linked *Streptococcus pneumoniae* infections were identifiable from phylogenetic tree reconstruction using consensus SNP sequences and that household transmission pairs were distinguishable from transmission occurring outside of their households. However, the introduction of longitudinal samples from the same individual into the tree reconstruction highlighted the little within-host diversity accumulated over the weeks following transmission, making longitudinal genomic sequences of an individual indistinguishable from those collected from the other member of the transmission event.

The NGS data from the phylogenetically linked individuals were then used for transmission direction inference, using ancestral state reconstruction within PhyloScanner, and the inferred direction was compared to the epidemiological records on who infected whom. Various parameters, including a minimum number of SNPs and the sliding window size lengths were tested and the sensitivity analysis revealed that both parameters did affect our abilities on the transmission direction inference. Increasing the minimum number of SNPs in the subtree reconstruction resulted in higher concordance between the transmission direction inferred from phylogenetic inference and the epidemiological records. Interestingly, four of the five phylogenetically identified sources largely agreed with the epidemiological data and demonstrated a higher number of unique SNPs compared to the recipient except for one pair where the phylogenetic data suggest the source-recipient status could have potentially been flipped. This study demonstrates promising results in bacterial transmission direction inference and how we can improve our inference abilities by increasing the robustness of the phylogenies which we infer from.

6.1.2 Summary of detecting multiple serotypes during co-carriage (Chapter 3)

While most bacterial genomes have low within-host diversity, thus hindering our abilities to

infer transmission direction, co-carriage of multiple *Streptococcus pneumoniae* serotypes can result in high within-host diversity and thus aid in transmission direction inference. However, there is little understanding of the role of minor serotypes within transmission dynamics which is largely due to our current lack of sensitivity to detect and disentangle minor variants from the genomic data. This led to us testing our ability to detect the occurrence and quantify co-carriage of multiple pneumococcal serotypes.

I used whole-genome NGS data to detect the occurrence and relative abundance of pneumococcal serotypes from individuals with co-carriage and compared current genomic serotyping methods to identify co-carriage. I investigated two popular genomic-based methods for detecting co-carriage, SeroCall and PneumoKITy, and then compared these results to DNA microarray, a highly sensitive and specific method for serotyping multiple pneumococcal populations. Both SeroCall and PneumoKITy had high sensitivity for detecting the dominant serotypes and low sensitivity for detecting serotypes at low abundances (<10%). Increasing the sequencing depth did have an impact and improved the detection of low-abundance serotypes for both genomic serotyping methods.

The current methods to capture the within-host diversity at the sequencing level and minor serotypes at the serotyping level have implications for transmission direction inference. More specifically, the inability to detect serotypes at low abundances would result in underestimating the within-host genetic diversity and thus impact the ancestral state reconstruction resulting in more ambiguous subtrees reconstructed. The importance of these findings suggests that increased sequencing depth can identify more within-host genetic diversity and thus lead to improved detection of minor variants.

In addition to comparing genomic methods to detect co-carriage, mixture modelling was implemented to estimate the sub-population based on the SNP frequency distribution alone. This method aimed to isolate the sub-populations in hopes of reconstructing the haplotypes in future studies. Initial attempts have been made for haplotype reconstruction for this thesis with little success. Further work is needed to improve the pipeline and to validate the reconstructed haplotypes, ideally against genomic data based on the extraction of DNA from morphologically different bacterial growth at culture.

6.1.3 Summary on transmission direction of SARS-CoV-2 (Chapter 4)

Transmission direction of SARS-CoV-2 alpha and delta variants was investigated from samples collected as a part of a UK prospective household transmission study. I was able to confirm putative household transmission pairs from the whole-genome sequencing data using patristic distance and bootstrap thresholds. I identified non-direct transmission using

phylogenetics thus excluded these individuals from estimates on the effectiveness of BNT162b2 and ChAdOx1 vaccines against household transmission of SARS-CoV-2.

The direction of transmission was inferred for the phylogenetically confirmed putative transmission pairs, using the method implemented in Chapter 2. I explored the effects of sequencing read length on directionality and compared the distribution of serial intervals obtained between serial intervals inferred phylogenetically and those inferred from the epidemiological data. I observed that the most likely direction inferred was not sensitive to the read lengths, however, the subtree inferred from longer reads provided better resolution e.g. fewer ambiguous relationships and stronger support for the most likely scenario. Further, the direction inferred from Phyloscanner was used to recalculate the serial interval between the suspected source and recipient and this was compared to the serial intervals calculated from the epidemiological data between the index case and the contact. The phylogenetic evidence supported 19/43 pairs to retain their epidemiological index case classification while 24/43 had evidence for reclassification. From this, I observed that phylogenetic inference was able to identify who infected whom, however, it should be interpreted with caution due to lack of robustness and conflict with the epidemiological data.

6.2 Context on transmission directionality inference using phylogenetic approaches

There are a limited number of studies that have validated transmission direction using phylogenetics approaches against epidemiological records¹⁻³, all of which were from known source-recipient of HIV-infected pairs. Those studies, like ours, had varying degrees of concordance between the phylogenetically and epidemiologically inferred source-recipient which were influenced by the sequencing depth and read lengths. Rose *et al.* observed between 55%-74% concordance while Zhang *et al.* had up to 93%. Zhang *et al.* highlight the impact of sequencing depth and read lengths from which the subtrees are reconstructed and discovered increased accuracy, compared to the previous study led by Rose *et al.* on the transmission direction inference^{2,3}. Both studies were from the same cohort but the sequencing method, Rose *et al.* used amplicon sequencing targeting different HIV genomic regions while Zhang *et al.* used an ultra-deep whole-genome NGS approach. The latter approach can generate depths over 10,000-fold which allows the detection of low-frequency variants, however, the cost and resources required for analysis of such large data associated with this method might not be feasible for routine surveillance studies².

Additionally, there are a limited number of studies looking at the transmission direction of bacterial infections using whole genome sequencing and phylogenetic approaches. One of which is a study from Hall *et al.* who used a similar approach to our study, PhyloScanner, to investigate the transmission direction of Methicillin-resistant *Staphylococcus aureus* from WGS NGS data⁴. An interesting finding from this study was that despite the presence of a large transmission bottleneck, e.g. a large number of shared lineages between source-recipient, the transmission direction was still ambiguous. This indicates that there is a presence of a bottleneck effect, however, its strength might not be associated with a higher probability of inferring the correct transmission direction. While in our study we observed a higher number of unique SNPs in the source compared to the recipient which implies that the bottleneck effect is not random. This assumes the pathogen population from the source has had more time to accumulate within-host genetic diversity compared to the recipient and thus the individual with the higher number of unique SNPs is also likely to be the source within the transmission pair.

Laboratory experiments have postulated that pneumococcal transmission usually involves a single cell from the source to the recipient resulting in a very narrow bottleneck that likely occurs following the exit from the source but prior to the establishment in the recipient⁵. However, a different study by Tonkin-Hill *et al.* revealed that human-to-human transmission bottleneck probably results in more than one transmitted bacterial cell⁶. Capturing within-host diversity can improve inference on transmission links⁷ and only considering the dominant variant can substantially underestimate the number of transmission links⁶. Additionally, if we

can reconstruct the haplotypes detected in co-carriage, this would help us better understand the role of minor variants in pneumococcus transmission dynamics.

Despite carriage of multiple unique serotypes being common in settings with high carriage prevalence^{8,9} the role of co-carriage of multiple pneumococcal serotypes is poorly understood in the transmission dynamics. There are tools available to detect co-carriage of multiple *Streptococcus pneumoniae* serotypes and previous studies have compared different pneumococcal serotyping methods¹⁰ but there is a limited head-to-head comparison that includes the use of genomic data for determining mixed serotypes in co-carriage. The genomic serotyping tools that are available to detect multiple serotypes lack sensitivity for detecting low abundant serotypes when compared to DNA microarray. Despite the lack of sensitivity, there is an added benefit to including sequencing as a part of routine surveillance including additional information for phylogenetic inference to investigate transmission dynamics. Tonkin-Hill *et al.* demonstrated the high sensitivity of genomic serotyping of pneumococcal co-carriage and highlighted the added insights on drug resistance and within-host evolution⁶.

Other studies have investigated the transmission of SARS-CoV-2 using phylogenetic approaches and also observed highly similar consensus genomes among household transmission pairs¹¹, with less than two consensus mutations. Most SARS-CoV-2 infections have limited within-host diversity and most of the mutations are inevitably lost during the transmission from the infection source to the recipient¹². In cross-sectional sampling near the time of the transmission event, the presence of a lower genetic diversity in the recipient compared to the source due to the small founding population^{13,14} can aid in determining the transmission direction. Despite the inherent lack of within-host genomic diversity¹⁵, directionality was able to be inferred from the whole-genome NGS data using ancestral state reconstruction. Most of the subtrees resulted in ambiguous relationships amongst the putative pairs, however, increasing the sliding window size which the subtrees are reconstructed from, resulted in more non-ambiguous relationships and the directionality inferred was consistent across the tested sliding window sizes. This implies less robust subtrees will result in more ambiguous relationships rather than incorrect transmission direction, while more robust trees will result in a transmission direction. However, these results should be interpreted with caution as there are inherent limitations specifically due to the sequencing read lengths (<200 bp) which limited the length of the sliding windows I was able to test.

The key takeaway from previous and current studies on phylogenetic inference on transmission direction is the necessity to consider intra-host diversity. The inability to account for within-host diversity can lead to inaccurate inference on transmission directionality¹⁶. The question remains, how much sequencing depth is sufficient for detecting linkage and

subsequently transmission direction inference? Linkage is observable using standard Illumina sequencing methods and consensus genomes for viruses and bacterial genomes. However, transmission direction has only been validated on HIV partners which used sequences up to 10,000-fold, but this method might not be feasible for most settings studying bacterial transmission.

Additionally, disease persistence should be considered when using within-host dynamics to infer transmission direction e.g. how much diversity should be expected at a certain stage and type of the infection. For example, HIV infections go through an acute stage resulting in rapid replication and high within-host diversity which then later usually transitions into a chronic stage^{17,18}. While the within-host diversity slows down in the chronic stage, the persistence of HIV infections allows a longer timeframe for additional within-host evolution to occur. Conversely, both pneumococcal carriage and COVID-19 are considered acute where carriage of pneumococcus usually lasts up to a month¹⁹ and COVID-19 infections are usually cleared within two weeks²⁰. The shorter timeframe compared to HIV results in a limited time frame in which within-host evolution can occur.

While pneumococcus and SARS-CoV-2 are both usually transmitted by respiratory routes in settings with close contact, there are key differences to take into account when assessing transmission dynamics. Individuals can asymptotically carry pneumococcus while SARS-CoV-2 infection can be asymptotically or presymptomatically²¹⁻²³ thus making it difficult to determine transmission direction from epidemiological data. Even in the presence of genomic data, there is still uncertainty with the transmission direction inference due to slow mutation rates and small regions of mutational hotspots for SARS-CoV-2 and large genomes with mutations spread along the genomes for pneumococcus²⁴.

6.3 Study limitations

The below summaries highlight important limitations that can be observed across multiple chapters of this thesis and trends that can often be observed when using pathogen genomics to detect within-host diversity thus affecting abilities to infer directionality.

6.3.1 Short and low-depth sequencing reads

The sequencing reads for the PhD were generated by various Illumina methods resulting in short-read fragments (<200 base pairs) for the subtree reconstruction in PhyloScanner. Short reads limited our abilities during the genome assembly and subsequently affected the phylogenetic signal for Chapters 2, 4, and 5, and potentially excluded unmapped mutations for Chapter 3 which could impact our ability to detect co-carriage. Increasing the sequencing read lengths would improve the mapping step which would improve the genomic signal for better

detection of within-host diversity and thus the phylogenetic signal for inferring directionality. Additionally, longer sequencing would most probably help aid in the haplotype reconstruction from *Streptococcus pneumoniae* population with co-carriage. Similar to the limitations imposed by short sequencing reads, low depth of sequencing also affects our abilities to detect within-host diversity and thus infer directionality for both pneumococcal infection and SARS-CoV-2.

6.3.2 Directionality inferred from cross-sectional household pairs excludes potential intermediary infections

Transmission direction inference from cross-sectional data can capture genetically similar pathogen populations from the source and recipient because the pathogen has less time to evolve. The approach to infer directionality attempts to maximise the probability of linking putative transmission pairs within their respective households. While I did test for linkage amongst putative household transmission pairs, the phylogenetic inference on transmission direction was only tested on those household pairs for both pneumococcal and SARS-CoV-2 infections based on our assumption that it is unlikely that individuals for either study were infected by a non-household individual. This assumption excludes potential unsampled intermediary transmission links or a common source of infection making directionality difficult to determine due to decreased mutation similarities between the assumed household transmission pairs. To mitigate this limitation and increase inference certainty, future studies should sequence community samples or close contacts of the household members to potentially identify non-household links with the exception of mother-child transmission where there is an unlikely unknown transmission link.

6.3.3 Reference genome and potential biases with mapping

Short sequencing reads are usually processed two different ways, either by mapping to a reference genome or through *de novo* assembly, where you do not need a reference genome. However, the small number of representative and reliable pneumococcal whole-genome sequences limited our selection pool. The reference genomes I used to map the *Streptococcus pneumoniae* Illumina sequencing reads could have introduced biases in the polymorphic sites that were detected or undetected. More explicitly, not all the reads necessarily map to the reference genome which could have potentially excluded particular reads or polymorphic sites that go undetected because too few reads were mapped to that region. Thus we would not be able to fully exploit the within-host diversity and this would affect the transmission direction inference, most likely resulting in more ambiguous relationships.

Particularly for Chapter 2, where I was investigating *Streptococcus pneumoniae* transmission directionality, I tried to capture as much within-host diversity as possible to get the maximal possible phylogenetic signal, however, I was analysing a heterogenous population of

serotypes and the reference genome selected for mapping all the samples was serotype 3. I performed a sensitivity analysis using another representative genome, serotype 23F, which resulted in similar inference outcomes, however, reference genome selection could still be improved for more accurate SNP detection. If available, a local representative genome should be used to map the study samples, similar to what Lee *et al.* did for inferring the transmission direction of tuberculosis (TB) where they mapped their samples, generated by Illumina, to a novel local reference genome, generated by PacBio²⁵. They were able to identify a previously undetected TB super-reader and demonstrated that the previous reference genome resulted in false positive SNPs in the study population. The false-positive SNPs that were detected were due to an alignment error and a local reference genome should be considered for identifying accurate variants.

6.3.4 False-negative diagnoses impacting linked infections

In addition to excluding potential intermediary infections and missing potential transmission links, false-negative nasal swab results for both pneumococcus and SARS-CoV-2 could also miss a potential transmission link. Pneumococcal detection sensitivity using culture-based methods is about 85%, however false negative rates for detecting pneumococcal carriage are up to 15%^{26,27}.

While RT-PCR is a sensitive method and the current gold standard for testing SARS-CoV-2, false negative rates for the initial detection can be as high as 54%²⁸ which can be partially explained by the incorrect administration and sample collection using the home nasal swabs²⁹ and thus administered by a medical professional. An undetected infected household individual due to false-negative testing would impact our assumptions on who the source and recipient are in the putative transmission pairs. Additionally, it would impact our ability to infer directionality similar to if a potential intermediary was not sampled.

6.4 The future of inferring transmission directionality using pathogen genomics

While pathogen genomics is useful and sensitive to detecting linked infections and able to infer transmission directions, there are clear limitations that need to be addressed. The decreased amount of cost and time associated with pathogen whole-genome sequencing has fallen over the past decade with applications to small local outbreaks to large pandemic scale. Genomic data has provided insights into pathogen evolution which has led to better disease control measurements.

6.4.1 Improving sequencing methods to better detect within-host diversity

One of the biggest limitations for inferring linkage and directionality of respiratory pathogens is the limited ability to capture sufficient within-host variation amongst linked infections which limits our ability to reconstruct the ancestral state and thus infer who infected whom. However,

despite this limitation, our capacity to sequence pathogens at increasing lengths and depths is one promising way to minimise the impacts of the limited within-host diversity⁶. The findings of this PhD show promising results which prompt further investigations into the impacts of longer and greater depth sequencing reads on phylogenetic inferences, particularly for bacterial infections where there is limited within-host diversity.

Results from Chapter 2 revealed that sequencing read length has an impact on inference abilities thus, I hypothesize that an increased read length would improve the transmission direction inferred e.g. less ambiguous trees. However, this was not tested during the PhD due to time constraints, therefore, future work on inferring directionality should aim to use sequencing methods that produce longer reads. This approach would improve the mapping and result in more robust genome alignments. Additionally, longer reads would allow the flexibility to reconstruct more robust subtrees that would be able to capture more within-host diversity with the longer sliding window lengths as previously observed by Rose *et al.* and Zhang *et al.*^{2,3}.

6.4.2 Validating transmission linkage from homogeneous *Streptococcus pneumoniae* population

Transmission direction was inferred from cross-sectional samples of putative household pairs of heterogeneous serotypes. While this approach maximised our chances of selecting true transmission pairs, it limited our abilities to explore the capacity at which we can infer linked *Streptococcus pneumoniae* infections. Thus, future work should rely on putative transmission pairs from a homogeneous population of serotypes to determine a genetic threshold for determining link and non-linked infections. This has been previously established for HIV³⁰ where the distribution of the patristic distance was bimodal revealing genetic distances of closely (<0.05% substitution per site) and distantly (>0.05% substitution per site) related infections.

6.4.3 Haplotype reconstruction from *Streptococcus pneumoniae* co-carriage

Co-carriage is detectable and quantifiable from genomic data and the mixture modelling shows promising results in being able to distinguish subpopulations. Haplotype reconstruction is largely based on the assumption that polymorphic sites that occur at the same frequencies belong to the same haplotype and thus the reads containing those polymorphic sites can be parsed based on those frequencies. The main limitation of this approach is haplotypes that occur at the same or similar frequencies. However, more work is needed to improve the reconstruction pipeline, particularly in validating the haplotypes that are reconstructed. More specifically, due to the already low read depths for the samples with co-carriage, once the serotype populations are parsed, this reduces the number of reads belonging to a specific serotype, making it difficult to serotype the parsed samples. Future work on haplotype

reconstruction will include using the merged samples with higher read depths with the hypothesis that increasing the number of reads will also increase the read count for the respective haplotypes that are present in the mixed population and thus I will be better able to validate the subpopulations. After, validate the haplotype reconstructed by reconstructing a phylogenetic tree using closely related circulating isolates to see if the reconstructed haplotype clusters with their respective serotypes.

6.4.4 Scalability for bacterial genomics

Future endeavours in pathogen epidemiology should consider investment in infrastructure for scaling NGS whole-genome sequencing and processing on the computational and analytical front. Most of the current tools have been developed for fast-evolving viruses and bacterial genomes are often magnitudes larger than viral genomes and undergo more frequent recombination. Currently, there are integrated pipelines that have streamlined the data processing and analysis of viral phylogenetics⁴, however, to our knowledge, such tools currently do not exist for bacterial phylogenetics due to scalability. *Didelot et al.* proposed a step-by-step approach for scaling bacterial phylogenetics in epidemiology studies by integrating bacterial genomic tools into pipelines that were previously developed for viral phylogenetics³³. While the proposed step-by-step approach can reduce computational time due to the parallelisation of large datasets, the limitation is the practicality of input and output compatibility from one tool to another which could result in inconsistency and is at risk for errors.

6.5 Conclusion

In summary, this thesis revealed the capacity at which we can determine linked infection and transmission of respiratory pathogens, *Streptococcus pneumoniae* and SARS-CoV-2, using next-generation whole-genome sequencing data. While there are some limitations in our studies particularly using sequencing methods that can be routinely generated resulting in limited detection of within-host variation thus impacting our ability to detect linkage and infer transmission direction. Despite these limitations, I was able to observe key findings including the influence of sequencing coverage and sequencing depth which has implications for future studies on respiratory pathogens particularly for bacterial species.

References

1. Villabona-Arenas, C. J. *et al.* Number of HIV-1 founder variants is determined by the recency of the source partner infection. *Science* **369**, 103–108 (2020).
2. Zhang, Y. *et al.* Evaluation of Phylogenetic Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV) Transmission: HIV Prevention Trials Network (HPTN) 052. *Clin. Infect. Dis.* **72**, 1 (2020) doi:10.1093/cid/ciz1247.
3. Rose, R. *et al.* Inconsistent temporal patterns of genetic variation of HCV among high-risk subjects may impact inference of transmission networks. *Infect. Genet. Evol.* **71**, 1–6 (2019).
4. Hall, M. D. *et al.* Improved characterisation of MRSA transmission using within-host bacterial sequence diversity. *eLife* **8**, e46402 (2019).
5. Kono, M. *et al.* Single Cell Bottlenecks in the Pathogenesis of *Streptococcus pneumoniae*. *PLoS Pathog.* **12**, e1005887 (2016).
6. Tonkin-Hill, G. *et al.* Pneumococcal within-host diversity during colonization, transmission and treatment. *Nat. Microbiol.* **7**, 1791–1804 (2022).
7. De Maio, N., Worby, C. J., Wilson, D. J. & Stoesser, N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput. Biol.* **14**, e1006117 (2018).
8. Murad, C. *et al.* Pneumococcal carriage, density, and co-colonization dynamics: A longitudinal study in Indonesian infants. *Int. J. Infect. Dis.* **86**, 73–81 (2019).
9. Swarthout, T. D. *et al.* High residual carriage of vaccine-serotype *Streptococcus pneumoniae* after introduction of pneumococcal conjugate vaccine in Malawi. *Nat. Commun.* **11**, 2222 (2020).
10. Swarthout, T. D. *et al.* Evaluation of pneumococcal serotyping in nasopharyngeal carriage isolates by latex agglutination, whole genome sequencing (PneumoCaT) and DNA microarray in a high pneumococcal carriage prevalence population in Malawi. <http://biorxiv.org/lookup/doi/10.1101/2020.08.17.255224> (2020) doi:10.1101/2020.08.17.255224.
11. Bendall, E. E. *et al.* SARS-CoV-2 Genomic Diversity in Households Highlights the

- Challenges of Sequence-Based Transmission Inference. *mSphere* **7**, e00400-22 (2022).
12. Braun, K. M. *et al.* Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLoS Pathog.* **17**, e1009849 (2021).
 13. Zwart, M. P. & Elena, S. F. Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annu. Rev. Virol.* **2**, 161–179 (2015).
 14. Gutiérrez, S., Michalakis, Y. & Blanc, S. Virus population bottlenecks during within-host progression and host-to-host transmission. *Curr. Opin. Virol.* **2**, 546–555 (2012).
 15. Dhar, S., Zhang, C., Mandoiu, I. I. & Bansal, M. S. TNet: Transmission Network Inference Using Within-Host Strain Diversity and its Application to Geographical Tracking of COVID-19 Spread. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 230–242 (2022).
 16. Worby, C. J., Lipsitch, M. & Hanage, W. P. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS Comput. Biol.* **10**, e1003549 (2014).
 17. Ragonnet-Cronin, M. *et al.* Genetic Diversity as a Marker for Timing Infection in HIV-Infected Patients: Evaluation of a 6-Month Window and Comparison With BED. *J. Infect. Dis.* **206**, 756–764 (2012).
 18. Kouyos, R. D. *et al.* Ambiguous Nucleotide Calls From Population-based Sequencing of HIV-1 are a Marker for Viral Diversity and the Age of Infection. *Clin. Infect. Dis.* **52**, 532–539 (2011).
 19. van der Poll, T. & Opal, S. M. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *The Lancet* **374**, 1543–1556 (2009).
 20. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
 21. Pan, X. *et al.* Asymptomatic cases in a family cluster with SARS-CoV-2 infection. *Lancet Infect. Dis.* **20**, 410–411 (2020).
 22. Bai, Y. *et al.* Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA* **323**, 1406 (2020).
 23. Kimball, A. *et al.* Asymptomatic and Presymptomatic SARS-CoV-2 Infections in

- Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 377–381 (2020).
24. Callaway, E. The coronavirus is mutating — does it matter? *Nature* **585**, 174–177 (2020).
 25. Lee, R. S., Proulx, J.-F., McIntosh, F., Behr, M. A. & Hanage, W. P. Previously undetected super-spreading of *Mycobacterium tuberculosis* revealed by deep sequencing. *eLife* **9**, e53245 (2020).
 26. Abdullahi, O., Wanjiru, E., Musyimi, R., Glass, N. & Scott, J. A. G. Validation of Nasopharyngeal Sampling and Culture Techniques for Detection of *Streptococcus pneumoniae* in Children in Kenya. *J. Clin. Microbiol.* **45**, 3408–3410 (2007).
 27. Thindwa, D. *et al.* Estimating the contribution of HIV-infected adults to household pneumococcal transmission in South Africa, 2016–2018: A hidden Markov modelling study. *PLoS Comput. Biol.* **17**, e1009680 (2021).
 28. Arevalo-Rodriguez, I. *et al.* False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS ONE* **15**, e0242958 (2020).
 29. Higgins, T. S., Wu, A. W. & Ting, J. Y. SARS-CoV-2 Nasopharyngeal Swab Testing—False-Negative Results From a Pervasive Anatomical Misconception. *JAMA Otolaryngol. Neck Surg.* **146**, 993 (2020).
 30. PANGAEA Consortium and Rakai Health Sciences Program *et al.* Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nat. Commun.* **10**, 1411 (2019).
 31. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, 1 (2018).
 32. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
 33. Didelot, X. & Parkhill, J. A scalable analytical approach from bacterial genomes to epidemiology. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210246 (2022).

