1 Overlapping transmission of group A and C/G Streptococcus facilitates inter-species

2 mobile genetic element exchange

3

Ouli Xie^{1,2}, Cameron Zachreson³, Gerry Tonkin-Hill⁴, David J Price^{1,5}, Jake A Lacey^{1,6},
Jacqueline M Morris⁶, Malcolm I McDonald⁷, Asha C Bowen⁸, Philip M Giffard^{9,10}, Bart J
Currie^{9,11}, Jonathan R Carapetis⁸, Deborah C Holt⁹, Stephen D Bentley¹², <u>Mark R Davies^{6*}</u>,
Steven YC Tong^{1,13*#}

- 8
- ⁹ ¹ Department of Infectious Diseases, University of Melbourne, at the Peter Doherty Institute
- 10 for Infection and Immunity, Australia
- 11 ² Monash Infectious Diseases, Monash Health, Australia
- ³ School of Computing and Information Systems, University of Melbourne, Australia
- 13 ⁴ Department of Biostatistics, University of Oslo, Norway
- ¹⁴ ⁵ Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global
- 15 Health, University of Melbourne, Australia
- ⁶Department of Microbiology and Immunology, University of Melbourne, at the Peter Doherty
- 17 Institute for Infection and Immunity, Australia
- ⁷ Division of Tropical Health and Medicine, James Cook University, Australia
- 19⁸ Wesfarmers Centre for Vaccines and Infectious Diseases, Telethon Kids Institute, University
- 20 of Western Australia and Perth Children's Hospital, Australia
- ⁹ Global and Tropical Health Division, Menzies School of Health Research, Charles Darwin
- 22 University, Australia
- 23 ¹⁰ Faculty of Health, Charles Darwin University, Australia
- 24 ¹¹ Infectious Diseases Department, Royal Darwin Hospital, Australia
- 25 ¹² Wellcome Sanger Institute, Wellcome Genome Campus, UK NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

- 26 ¹³ Victorian Infectious Diseases Service, The Royal Melbourne Hospital, at the Peter Doherty
- 27 Institute for Infection and Immunity, Australia
- 28
- 29 * These authors contributed equally to the study
- 30
- 31 [#] Corresponding author:
- 32 Prof Steven Tong
- 33 The Peter Doherty Institute for Infection and Immunity
- 34 792 Elizabeth St, Melbourne, VIC, Australia, 3000
- 35 <u>steven.tong@mh.org.au</u>

36 Abstract

37 Streptococcus dysgalactiae subspecies equisimilis (SDSE) and Streptococcus pyogenes share 38 skin and throat niches with extensive genomic homology and horizontal gene transfer (HGT) 39 possibly underlying shared disease phenotypes. It is unknown if cross-species transmission 40 interaction occurs. We conducted a genomic analysis of a longitudinal household survey in 41 remote Australian First Nations communities for patterns of cross-species transmission 42 interaction and HGT. From 4,547 person-consultations, 294 SDSE and 315 S. pyogenes isolates 43 were sequenced. SDSE and S. pyogenes transmission intersected extensively among 44 households and the observed co-occurrence and transmission links were consistent with independent transmission without inter-species interference. At least one of three near-identical 45 46 cross-species mobile genetic elements (MGEs) carrying antimicrobial resistance or 47 streptodornase virulence genes was found in 55 (19%) SDSE and 23 (7%) S. pyogenes isolates. 48 These findings demonstrate extensive co-circulation of both pathogens and HGT and support a need to integrate SDSE and S. pyogenes surveillance and control efforts. 49

50 Introduction

51 dysgalactiae subspecies equisimilis (SDSE, commonly group C/G Streptococcus 52 Streptococcus) is closely related to the better-known human pathogen, Streptococcus pyogenes 53 (group A Streptococcus). SDSE shares much the same ecological niche on the human skin and 54 throat as S. pyogenes and the two pathogens exhibit overlapping disease manifestations such 55 as pharyngitis and invasive disease including necrotising fasciitis and streptococcal toxic shock 56 syndrome¹. In regions with a high burden of beta-haemolytic streptococcal disease and post-57 infectious sequelae, there has been evidence that superficial SDSE infection may trigger 58 immune responses which cross-react with cardiac myosin^{2, 3}. These findings raise the 59 possibility that SDSE may contribute to immune priming and the burden of rheumatic heart 60 disease in those regions². In high income regions, emerging evidence has also described crude rates of invasive SDSE disease comparable to, and in some jurisdictions, greater than S. 61 pvogenes⁴⁻⁶. 62

Whole genome comparisons of SDSE and *S. pyogenes* demonstrate extensive genomic homology including shared virulence factors such as the multi-functional surface M protein and evidence of horizontal gene transfer (HGT), frequently involving mobile genetic elements (MGEs)⁷⁻⁹. These similarities may contribute to shared disease phenotypes. Many *S. pyogenes* vaccine candidates are present in both species with evidence of cross-species homologous recombination⁹.

Despite extensive genomic homology, there is *in vitro* evidence of possible cross-species competition. Strains of the two pathogens possess shared quorum sensing genes such as the *sil* locus with evidence of cross-species signalling¹⁰. Furthermore, anti-microbial peptides or bacteriocins such as SpbN/SpbM and the SDSE-specific Dysgalacticin, are found in some strains of SDSE and *S. pyogenes* with cross-species activity^{11, 12}.

74 SDSE and S. pyogenes transmit by common pathways including respiratory droplets¹. Recently, we have shown that asymptomatic S. pyogenes throat carriage is an important 75 reservoir of transmission in high-endemic settings¹³. Transmission pathways of SDSE have not 76 77 previously been described. Further, it is uncertain if in real-world studies transmission of one 78 species competes with the other. In communities endemic for S. pyogenes infection with high 79 rates of skin infection, rheumatic heart disease and invasive disease, the current focus is largely 80 on S. pyogenes control through skin sore and scabies control programs, and vaccine 81 development. Understanding the transmission interactions of SDSE and S. pvogenes and 82 anticipating the potential impact of disease control measures on cross-species behaviour is important to inform the design of surveillance programs and infection control efforts. 83

In this study, we examine the transmission of SDSE at a whole genome sequence (WGS) resolution using isolates collected in a household-based surveillance study over two years in two remote communities in the Northern Territory of Australia¹⁴. These transmission networks were compared to that of co-collected *S. pyogenes* isolates to assess for inter-species transmission interactions, and in the setting of co-circulation, their genomes were systemically examined for evidence of cross-species HGT of MGEs carrying key virulence and antimicrobial resistance genes.

91

92 **Results**

93 Sampling and clinical epidemiology

Two remote Aboriginal communities in the Northern Territory of Australia were prospectively followed for a two-year period between 2003–2005^{14, 15}. Observations for one community (community 3) commenced in June 2004 as it replaced an initial community (community 2) with low recruitment. Communities 1 and 3 were included in this study.

98 Households (18 in community 1 and 20 in community 3) were visited approximately monthly 99 allowing for access affected by weather and cultural events (Supplementary Figure 1). At each 100 visit, throat swabs were taken regardless of symptoms and skin swabs were taken from 101 impetigo lesions. From a total of 4,547 person-consultations during 486 household-visits, 1,087 102 individuals (547 from community 1 and 540 from community 3) were sampled from which 330 103 SDSE isolates (252 community 1 and 78 community 3) were recovered. Of the 330 isolates, 8 104 were from skin and 322 from throat swabs of which only one case reported a sore throat. S. 105 progenes was recovered on 327 occasions (218 community 1 and 109 community 3) with 208 106 isolated from throat swabs and 119 from impetigo lesions. Detailed descriptions of the epidemiology of cases were described previously^{14, 15}. 107

There was a high rate of individual mobility in and out of households with a median of 28 people (range 6–57) enrolled per household over the study period. Each individual was observed at a median of 3 visits (range 1–19, intermittently sampled); and as such, duration of carriage in individuals could not be determined. Households were positive (i.e., at least one individual positive) for SDSE for a median of 56 days and then re-acquired SDSE a median of 37 days later.

114

115 Whole genome sequencing reveals detailed transmission clusters

From the 330 SDSE isolates, 294 (89%) were recovered for WGS. Using traditional epidemiological markers, *emm* type and multilocus sequence type (MLST), these isolates represented 19 *emm* types (23 *emm* subtypes), 21 MLSTs, and 26 *emm*-MLST combinations (Supplementary Table 1). Of these, 8/26 (31%) *emm*-MLST groups were found across both communities. Sequencing and analysis of 315/327 (96%) *S. pyogenes* isolates recovered from communities 1 and 3 for WGS were reported previously¹³.

122 To determine a WGS threshold for clustering of strains, we examined genomic variation of 123 isolates of the same strain found longitudinally on multiple occasions from the same individual. 124 Intra-host variation was used to predict longitudinal diversity of strains forming transmission 125 chain as well as technical variations in single nucleotide polymorphism (SNP) calling. SDSE 126 was found in 58 individuals on more than one occasion including three who were positive on 127 five occasions, four on four occasions, 15 on three occasions and 36 on two occasions. Using 128 emm and MLST as markers, 36 individuals had the same strain on more than one occasion 129 including six individuals with the same isolate on three occasions, one individual on four 130 occasions, and one on five occasions (Supplementary Figure 2). Pairwise SNP distances were calculated between these isolates and a threshold of <8 SNPs was determined for WGS 131 132 transmission clustering (Supplementary Figure 3).

133 Phylogenetic reconstruction supported 18 distinct SDSE lineages/global genomic sequence clusters⁹ present across both communities (Figure 1). High resolution genomic transmission 134 135 clusters based on single linkage clustering at a SNP threshold of <8 and >99% shared gene 136 content, revealed much finer detail than the traditional epidemiological markers (Figure 2). A 137 total of 37 SDSE transmission clusters representing 237 (81%) isolates were inferred with an 138 additional 57 singleton isolates (Supplementary Table 1). Transmission clusters were supported by core SNP phylogenies and presence-absence of virulence and/or antimicrobial resistance 139 140 genes (Supplementary Figure 4a-c) with significant diversity within emm types (Figure 2) and 141 evidence of mobile genetic element (MGE) gain/loss events carrying antimicrobial resistance 142 and/or virulence factor genes among closely related isolates.

The two largest transmission clusters consisted of 32 isolates each and clusters with four or more isolates made up a total of 204 isolates (69%). Transmission clusters were present across a mean of 3 households (range 1–16). Despite the finding of eight *emm*-MLST groups across both communities, the WGS analysis indicated that only a single transmission cluster spanned

both communities. The upper limit of the pairwise SNP distance between isolates of the same transmission cluster was 16 SNPs (median 4) compared to 791 SNPs (median 20) within the same MLST, 5491 SNPs (median 25) within the same *emm* type, 638 SNPs (median 19) within the same *emm*-MLST combination, and 1505 SNPs (median 21) within the same genomic sequence cluster (Supplementary Figure 5), highlighting the limitations of other markers in determining recent transmission clusters.

There was no clear pattern of *emm* type replacement of SDSE isolates over time in the two communities in contrast to sequential replacement of *S. pyogenes emm* types as reported previously^{13, 14}. Consistent with this finding, SDSE transmission clusters persisted for longer in the two communities (median of 349 days, 95% CI 189-440 days) compared to *S. pyogenes* (median of 241 days, 95% CI 181-259 days, log-rank p = 0.009) (Supplementary Figure 6).

158

159 Network analysis supports independent transmission dynamics for SDSE and S. pyogenes

160 SDSE transmission between households within each community was modelled by inferring 161 links between isolates of the same transmission cluster detected at successive community visits 162 (transmission window 12-44 days), including intra-household transmission events. Individuals 163 were grouped by household which formed the nodes of the transmission work. Analysis of the 164 transmission network revealed 123 SDSE putative transmission edges (events) in community 165 1 and 14 edges in community 3, which had a shorter duration of sampling and fewer isolates 166 detected (Table 1). All but one transmission edge was attributed to isolates from throat swabs 167 for SDSE in contrast to 50/173 (29%) edges attributed to a predicted skin source for S. pyogenes. 168

169 To test the hypothesis that transmission of SDSE or *S. pyogenes* may interfere with 170 transmission of the other species, the overlap between inferred transmission networks of the 171 two species was compared to a null model in which any cross-species interaction was removed.

"Transmission overlap" was defined as the proportion of inferred SDSE transmission edges 172 173 that corresponded to an inferred transmission of S. pvogenes. An overlapping edge 174 corresponded to transmission of both SDSE and S. pyogenes which occurred between the same 175 pair of households within the same transmission window without distinguishing which 176 household acted as source. To generate a null model of transmission overlap, household labels 177 in the inferred SDSE transmission network were randomised while preserving the S. pyogenes 178 network. This process preserves important structural features of the SDSE network including 179 degree distribution, and any clustering of SDSE transmission between households, while 180 removing any direct cross-species effects related to the transmission of S. pyogenes.

181 Overlaying the transmission networks of the two species found a highly interconnected 182 network with 11 shared transmission edges - nine in community 1 and two in community 3 183 (Figure 3). The number of shared edges in each community was consistent with the distribution 184 under the null model providing no evidence of inter-species transmission interference (one-185 sided p-value \leq observed value for community 1 = 0.75, community 3 = 0.94) (Supplementary 186 Figure 7a, c). Results were similar when restricting the analysis to isolates only from throat 187 swabs (Supplementary Figure 7b, d). These results indicate no evidence of an interaction between the two species in their household transmission patterns. 188

189 Although only 11/137 (8%) of total SDSE transmission edges were shared with S. pyogenes, 190 the combined transmission networks demonstrated extensive crossover of the two organisms at the household level — SDSE and *S. pyogenes* co-occurred in the same household on 100/486 191 (21%) of household-visits (Figure 4). To infer a null model of co-occurrence of SDSE and S. 192 193 pyogenes in households while removing cross-species transmission effects, SDSE and S. 194 pyogenes positive swabs were randomised across all swabs at each community visit. To account 195 for grouping of isolates within households, isolates from the same transmission cluster were 196 collapsed to a single positive result during the same household-visit. The observed co-

197 occurrence of SDSE and *S. pyogenes* within households was consistent with the model of 198 independent inter-species transmission without evidence of interference (one-sided p-value \leq 199 observed value across both communities = 0.62). Results from a sensitivity analysis limited to 200 isolates from throat swabs were consistent (Supplementary Figure 8a, b).

201

202 Co-occurrence of SDSE and S. pyogenes facilitates shared mobile genetic elements

We have previously demonstrated extensive genomic overlap between SDSE and *S. pyogenes* in the context of global genome databases⁹. In the setting of extensive household co-occurrence of the two species, we sought to find evidence of shared MGEs between the two species. Using a pangenome synteny-based approach, MGEs were systemically extracted from both SDSE and *S. pyogenes* isolates and examined for elements with >99% nucleotide identity across species^{9, 16}.

209 Three near identical MGEs were found to be present in SDSE and *S. pyogenes* with variable 210 presence across closely related isolates with as few as 0-11 core SNPs, suggestive of recent MGE gain/loss events within each of these strains (Figure 5a). A 53kbp prophage, ϕ 1207.3¹⁷, 211 carrying mef(A)/msr(D) macrolide efflux resistance genes was carried at a conserved cross-212 213 species genomic location (between SDEG RS07105 and SDEG RS07110 in reference 214 genome GGS 124 NC 012891.1) and was present in 5 S. pyogenes and 31 SDSE isolates 215 (Figure 5b). A second prophage, ϕ MGAS5005.3 carrying the streptodornase gene *sda1*, 216 previously described to be shared across species, was also found in a cross-species conserved 217 insertion region⁹. An 18kbp integrative conjugative element (ICE)-like segment carrying the 218 tetracycline resistance gene, tet(M), was present in four S. progenes and eight SDSE isolates at three distinct insertion regions (Figure 5c, Supplementary Figure 9). At least one of these 219 220 MGEs which carried antimicrobial resistance-associated genes or virulence-encoding genes, 221 was found in 55 (19%) of SDSE and 23 (7%) of S. pyogenes isolates. SDSE isolates carrying

these shared MGEs were found across both communities while *S. pyogenes* isolates carrying shared MGEs were restricted to single communities (community 1 for ϕ 1207.3, community 3 for ϕ MGAS5005.3 and the ICE-like element).

While directionality of MGE movement could not be inferred, including distinguishing between inter-species versus intra-species dissemination, the presence of near-identical elements at conserved insertion regions, suggests that overlapping transmission may facilitate shared MGEs from a common pool. The carriage of these MGEs across multiple distinct lineages suggests that these shared MGEs may lead to dissemination of antimicrobial resistance and virulence-associated genes.

231

232 Discussion

233 Using WGS-level resolution, we were able to reconstruct SDSE household transmission 234 networks and compare it to co-collected S. pyogenes isolates, demonstrating extensive co-235 circulation. Despite occupying similar niches on the skin and throat, we show that the two 236 organisms transmit independently without evidence of interference at the household level. In 237 the setting of extensive transmission cross-over in households, we find multiple MGEs present 238 across both populations carrying antimicrobial resistance or virulence factor genes with 239 evidence suggestive of recent gain/loss events. This analysis of a dataset of densely co-sampled 240 SDSE and S. pyogenes isolates provides a level of transmission detail and examination of realworld inter-species transmission dynamics and horizontal gene transfer which to our 241 242 knowledge, has not previously been described for beta-haemolytic streptococci.

SDSE is increasingly being recognised as an important cause of invasive human disease with recent studies suggesting incidence and mortality comparable to *S. pyogenes*⁴⁻⁶. While not traditionally considered as a cause of acute rheumatic fever/rheumatic heart disease (ARF/RHD), reports from northern Australia suggest that at least in high-incidence areas of ARF/RHD, SDSE throat carriage may have the potential to induce cardiac myosin cross-

reactive antibodies mimicking that seen with *S. pyogenes*^{2, 3}. Therefore, the finding of extensive
throat transmission of SDSE, including persistence of transmission clusters longer than that of *S. pyogenes*, underscores a need to further understand its contribution to immune priming for
ARF/RHD which in turn has important disease control implications.

252 Additionally, interactions between SDSE and *S. pyogenes* such as horizontal gene transfer and 253 homologous recombination are key drivers in bacterial population dynamics, and may influence S. pyogenes and SDSE biology⁹. Notably, genes encoding antigens currently under 254 255 investigation as S. pyogenes vaccine candidates are frequently also found in SDSE⁹. Our 256 findings of extensive household co-occurrence may provide an opportunity for HGT which we 257 demonstrate in the setting of shared MGEs. We show three near-identical MGEs were present 258 across different lineages in SDSE and S. pyogenes including presence and absence in closely 259 related isolates suggestive of recent gain/loss events. These MGEs carried antimicrobial 260 resistance and virulence genes such as the macrolide efflux genes mef(A)/msr(D), tetracycline 261 resistance tet(M), and the streptodornase gene *sda1*. While we cannot infer directionality of 262 HGT of MGEs across species compared to intra-species dissemination or acquisition from an intermediary species, at least one of these MGEs was present in 55 (19%) of SDSE and 23 263 264 (7%) of S. pyogenes isolates. This underscores the importance of integrating SDSE with S. 265 pyogenes surveillance as we seek to improve our understanding of transmission and disease 266 pathogenesis of the two organisms and as efforts move towards a possible S. pyogenes vaccine 267 which may introduce selection pressures across both organisms.

SDSE and *S. pyogenes* occupy similar ecological niches in the throat and on the skin with overlapping disease manifestations such as pharyngitis. Cross-species interaction and competition has been demonstrated such as the expression of bacteriocins which are able to inhibit the other species and cross-species quorum sensing involving the two-component regulator, *silAB* with its signalling peptide *silCR*¹⁰⁻¹². However, the *sil* locus and characterised

bacteriocins such as Dysgalacticin and SpbN/SpbM are variably present in SDSE and *S. pyogenes* and it is unclear if *in vitro* interactions translate to real-world transmission dynamics.
Our data demonstrate that despite evidence of possible *in vitro* interference, SDSE and *S. pyogenes* appear to transmit independently with highly interconnected household transmission
networks in a high burden setting.

SDSE was almost exclusively isolated from the throat in this study¹⁴. The mechanism behind 278 the predilection for the throat for SDSE in comparison to the wider presence of S. pyogenes 279 280 across throat and impetigo lesions is unclear. As described previously, the age of individuals 281 included in this study with SDSE was not different to those with S. pyogenes with the highest 282 rates in 5-14 year-olds and does not explain the throat predominance of SDSE^{14, 15}. Despite the 283 genomic similarities between SDSE and S. pyogenes, their virulence repertoires differ 284 including carriage of the cysteine proteinase SpeB which is exclusively present in S. pyogenes. 285 Experimental evidence suggests that SpeB activity may be important in establishing skin infection for S. pyogenes¹⁸. Cross-species genotype-phenotype associations could not be drawn 286 287 from this study due to the near perfect separation between skin and throat sites for SDSE. However, sensitivity analyses restricting cross-species transmission analyses to throat isolates 288 289 were concordant with the primary analysis without any evidence of cross-species interference. 290 Despite evidence of independent transmission at a household level, with household co-291 occurrence of SDSE and S. pyogenes on 100/486 (21%) of household-visits, the frequency of 292 presence of SDSE and S. pyogenes in the same swab is unclear. SDSE and S. pyogenes are both 293 large colony, beta-haemolytic streptococci and are generally indistinguishable by colony 294 morphology. Given only representative colonies were characterised in this study, the frequency 295 of co-colonisation of SDSE and S. pyogenes in the same individual could not be estimated. In 296 fact, this is a common limitation of carriage studies to date seeking to determine the prevalence of SDSE and S. pyogenes from throat swabs¹⁹⁻²¹. Given our findings of household-level 297

transmission dynamics, future studies should consider methods such as WGS from plate sweeps or deep sequencing of swabs to determine co-occurrence in individuals. These methods have also previously been shown to improve resolution of intra-host diversity and reconstructing transmission and may offer greater insight into cross-species transmission dynamics²².

303 Our study has some limitations. This study was carried out in a remote and tropical setting in northern Australia in Aboriginal communities with a high burden of S. pyogenes disease 304 305 including impetigo, ARF/RHD and invasive disease. Therefore, transmission dynamics and co-306 occurrence of the two organisms may differ in other settings. There was a high level of 307 population mobility in and out of households in these communities and thus individual level 308 transmission dynamics and duration of carriage could not be determined due to limited 309 longitudinal sampling of most individuals. Additionally, while SDSE was only found from 8 310 impetigo/skin sore swabs, intact skin was not sampled. Therefore, it is unclear if SDSE on 311 healthy skin may contribute to transmission.

312 In summary, this study demonstrates important transmission dynamics of SDSE and S. 313 pyogenes. The two closely related pathogens frequently co-occur within households with 314 interconnected transmission networks, but without evidence of inter-species interference across 315 households. Transmission overlap and shared niches, particularly in the human throat, may 316 facilitate interspecies gene flow including clinically important determinants such as 317 antimicrobial resistance genes. These findings emphasise a need to further understand the interactions between these pathogens including in the context of ARF/RHD in high burden 318 319 regions. The immunopathogenesis of ARF remains poorly understood despite many decades of 320 research and the specific events antecedent to each episode of ARF are elusive with respect to 321 the role of S. pyogenes in skin lesions and SDSE in the throat. As interventions targeting S. 322 pyogenes take place, it is possible that SDSE may also be affected. That impact could

potentially be a reduction in SDSE disease (e.g., by vaccines that may target common antigens) or conversely by SDSE filling an ecological niche if *S. pyogenes* infection or carriage is selectively targeted (e.g., in primary care interventions that expand the use of *S. pyogenes* rapid diagnostics for throat swabs). Incorporating research, surveillance and control efforts of SDSE with *S. pyogenes* will improve the understanding of both pathogens individually and crossspecies interactions in relation to clinical disease burden, disease phenotypes, and future response to vaccine interventions.

330

331 Methods

332 *Isolate collection and culture*

333 Isolates were collected from a previously reported prospective surveillance study in three 334 remote Aboriginal communities in remote Northern Territory, Australia, which were visited approximately monthly over a two-year period from August 2003 to June 2005¹⁴. Due to 335 336 waning community support and logistical difficulties in community 2, it was replaced with 337 another community in June 2004 (community 3). Only communities 1 and 3 were included in 338 this study. At each visit, researchers collected throat swabs regardless of symptoms from 339 participants and examined for skin sores both purulent and dry, which were also swabbed. Due 340 to high population mobility, individuals were identified as part of households for analyses, 341 including family groups residing in one or two adjacent houses.

Swabs were inoculated onto horse blood agar and selective media containing colistin and nalidixic acid and transported for culture at a central laboratory in Darwin, Australia. Plates were incubated at 37°C in 5% CO₂ and examined after 24 and 48 hours. A single representative colony was selected for typing (Streptococcal Grouping Kit, Oxoid Diagnostic Reagents) unless significant differences in colony morphology and/or haemolysis intensity was observed, in which case additional colonies were also selected.

The current study received ethics approval from the Human Research Ethics Committee of the
Northern Territory Department of Health and Menzies School of Health Research (approval
2015-2516).

351

352 Whole genome sequencing and typing

Lancefield group C/G streptococcal isolates were retrieved from stored glycerol stocks kept at
-70°C. Microbial DNA was extracted and 150bp paired-end libraries were prepared using the
Illumina TruSeq prep kit. Sequencing was performed using the Illumina HiSeq X Ten platform
(The Wellcome Trust Sanger Institute, United Kingdom). Fifty-four SDSE sequences were
previously published by Xie et al.⁹ *S. pyogenes* sequences were previously described by Lacey
et al.¹³ and available under Bioproject PRJNA879913.
Reads from Lancefield group C/G streptococcal isolates were checked for contamination using

360 Kraken2 v2.1.2.²³ Any sequences with >5% reads assigned to a species other than SDSE, with

361 the exception of *S. pyogenes*, was excluded. Genomes were assembled using a previously 362 described pipeline⁹.

In silico typing of the hypervariable N-terminal domain of the *emm* gene was performed using 363 364 emmtyper v0.2.0 (https://github.com/MDU-PHL/emmtyper) and MLST assigned using MLST v2.22.0 (https://github.com/tseemann/mlst)²⁴. Genomic sequence clusters, representative of 365 366 global SDSE populations, were assigned using PopPUNK v.2.60 with a scheme available at https://www.bacpop.org/poppunk/ (v1)^{9, 25}. Antimicrobial resistance and virulence genes were 367 inferred as previously described⁹. Genome metadata is available in Supplementary Table 1. S. 368 pyogenes genomic sequences clusters were assigned with a scheme available at 369 https://poppunk.net/pages/databases.html²⁶. 370

The pangenomes for SDSE and *S. pyogenes* were constructed using Panaroo v1.2.10²⁷ in 'strict'

mode with initial clustering at 98% length and sequence identity followed by a family threshold

of 70%. Core genes were defined as genes present in ≥99% of genomes. Pangenome gene
synteny was mapped using Corekaburra v0.0.5²⁸.

375 Maximum likelihood phylogenetic trees for SDSE and *S. pyogenes* isolates were inferred using IQ-tree v2.0.6 with a GTR+F+G4 model and 1000 UFBoot replicates^{29, 30}. Alignments for 376 377 SDSE were generated using Snippy v4.6.0 (https://github.com/tseemann/snippy) against reference genome GGS 124 (NC 012891.1) and S. pyogenes against reference genome 378 379 MGAS5005 (NC 007297.2) with MGE regions masked. Recombination was not masked. 380 Maximum parsimony trees were inferred within genomic sequence clusters to validate 381 predicted transmission clusters using phangorn v2.10.0 and SNP alignments generated by split kmer analysis (SKA v1.0) $^{31, 32}$. 382

383

384 MGE comparison

MGEs were systemically extracted from SDSE and *S. pyogenes* genomes using a pangenome synteny based approach based on proMGE and as previously described^{9, 16}. The pipeline was modified to extract nucleotide sequences of accessory genomic segments classified as MGEs. Sequences from SDSE and *S. pyogenes* were initially clustered using CD-HIT³³ v4.8.1 with a sequence identity threshold 0.8 and length difference cut-off 0.8. Clusters with sequences from both species were inspected and pairwise alignments generated using minimap2 v2.24³⁴.

391

392 *Transmission clustering*

Transmission clusters representing isolates predicted to have formed a recent transmission chain based on WGS data was determined using a SNP threshold of <8 and >99% shared gene content. A SNP threshold of <8 was determined using the maximal SNP distance between isolates of the same *emm* and MLST isolated within the same individual, a surrogate for SNP

diversity within a single infecting strain. A gene content threshold of >99% determined from
pangenome analysis was chosen to capture MGE gain/loss events.

399 Pairwise SNP distance between isolates within the same genomic sequence cluster was performed using SKA v 1.0^{31} from reads using the 'fastq' command with default coverage cut-400 401 off 4, minimum minor allele frequency 0.2, minimum base quality 20 and kmer size 15. Single 402 linkage clustering with a SNP distance of <8 was calculated using the 'distance' command. 403 SKA requires exact kmer matches in order to detect SNPs between the flanking/split kmers and 404 therefore may miss SNPs between more divergent sequences. The 150bp hypervariable N-405 terminal region of the *emm* gene poses such a challenge and has previously been shown to be 406 able to undergo recombination. As such, isolates clustered by SKA were checked for matching 407 emm types. In the case of different emm subtypes, alignments of the hypervariable region emm 408 region were manually inspected to determine the number of SNPs. Finally, a pangenome 409 comparison and single linkage clustering within each SKA cluster at 20 genes (approximates 99% gene similarity) was performed to generate the final transmission clusters. 410

Transmission cluster persistence was calculated by Kaplan-Meier estimation from time of first detection of a transmission cluster in a community to the first visit where the transmission cluster was not detected without subsequence re-appearance. The difference between SDSE and *S. pyogenes* was calculated by Cox Proportional Hazards as implemented in survival v3.4.0 and survival v0.4.9.

416

417 Household transmission network

Transmission networks between households in each community were inferred using a modified version of the model described by Lacey *et al.*¹³ and the R packages igraph v1.3.5 for network analysis, and ggraph v2.1.0, and scatterpie v0.1.8 for visualisation. Networks were initially inferred for SDSE and *S. pyogenes* separately. Transmission clusters predicted using WGS

422 were mapped against epidemiological metadata to generate adjacency matrices from which 423 transmission networks were inferred. Each household was represented by a node within the transmission network and unweighted edges (transmission events or links) were drawn 424 425 between households, or within the same household, when unique individuals carried isolates 426 from the same transmission cluster across successive community visits (transmission window 427 between 12 to 44 days). Isolates could be linked to multiple other isolates within the same 428 transmission cluster and respective transmission window. The isolate detected at the earlier 429 community visit/time point was denoted as putative source for the purposes of predicting the 430 contribution of throat and skin carriage/infection to transmission. Edges were drawn for each transmission window and assigned to the latter, 'recipient', community visit. As households 431 432 were sampled in order over a short window (range 1-4 days) within each community visit, 433 transmission edges were not inferred within the same community visit given the uncertainty in 434 predicting a source.

To determine transmission overlap between SDSE and *S. pyogenes*, the intersection between the transmission networks for SDSE (G_{sdse}) and *S. pyogenes* (G_{pyo}) at each transmission window (w) was taken. Here, households were nodes and with undirected edges (E) drawn if they were linked by a transmission event. The percentage of shared edges (f) was calculated as:

440
$$f = \frac{\sum_{w=1}^{n} |(E_{sdse,w} \cap E_{pyo,w})|}{\sum_{w=1}^{n} |(E_{sdse,w} \cup E_{pyo,w})|} \times 100$$

441

442 Models of independent transmission

A null model of independent transmission was generated for the SDSE and *S. pyogenes*transmission networks by node-label permutation. The household labels of the transmission
adjacency matrix generated from the observed SDSE data were permuted over 10,000 iterations

and compared to the *S. pyogenes* transmission network inferred from observed data as described above. The number of overlapping edges at each transmission window was calculated, summed for each iteration, and compared to the observed number of shared transmission edges. A one-sided p-value testing the hypothesis of SDSE and *S. pyogenes* transmission interference was calculated by the proportion of permutations with shared edges \leq observed shared edges.

452 A model of independent inter- and intra-species transmission for household co-occurrence was 453 generated by permutation of positive SDSE and S. pyogenes swabs across individuals and households within each community visit while accounting for grouping of isolates from the 454 455 same transmission cluster within households. Sequenced isolates from the same transmission 456 cluster and household-visit (159/609 SDSE and S. pyogenes combined, 26%) were collapsed 457 into a single positive result (159 collapsed to 69 positive swabs). Positive SDSE and S. 458 pyogenes swabs at each community visit were then permuted across all individuals sampled at 459 that respective community visit. A co-occurrence within a household was counted when SDSE 460 and S. pyogenes were present simultaneously in individuals in a household regardless if they 461 were from the same individual or across multiple individuals. After 10,000 iterations, a one-462 sided p-value testing the hypothesis of SDSE and S. pyogenes transmission interference at a 463 household level was calculated by the proportion of permutations with co-occurrences \leq 464 observed co-occurrences.

465

466 Data availability

467 Accessions for newly sequenced SDSE isolates are listed in Supplementary Table 1 under the468 BioProject identifier PRJEB35476.

469

470 *Code availability*

471 Scripts used to generate the transmission networks and null models are available at 472 <u>https://github.com/OuliXie/SDSE_transmission</u>. Scripts for MGE extraction and classification 473 from the pangenome are updated from that described previously⁹ and are available at 474 <u>https://github.com/OuliXie/Strep_MGE_pipeline</u>.

475

476 Acknowledgements

We thank the participants, communities, councils, Aboriginal research officers and health 477 478 centres for their involvement in the original surveillance study. We thank Ross M. Andrews for 479 his role in the original surveillance study. We acknowledge the assistance of the sequencing and pathogen informatics core teams at the Wellcome Sanger Institute, UK where this work 480 481 was supported by the Wellcome Trust core grants 206194 and 108413/A/15/D. OX was 482 supported by the Australian Health and Medical Research Council (NHMRC) postgraduate 483 scholarship (GNT2013831) and Avant Foundation Doctors in Training Research Scholarship 484 (2021/000017). MRD was supported by a University of Melbourne CR Roper Fellowship.

485

486 **Contributions**

487 OX worked on study design, analysis, data interpretation and manuscript preparation. MRD 488 and SYCT contributed to conception of the project and data interpretation. CZ, GTH, DJP and 489 JAL contributed to transmission model design and data interpretation. JMM, MIM, ACB, 490 PMG, BJC, JRC, and DCH contributed to data collection and curation. SDB contributed to 491 genomic sequencing. All authors contributed to manuscript preparation and review.

492 **Table 1.** Number of inferred unweighted transmission edges between households at successive

493 community visits in two remote communities.

Streptococcus dysgalactiae subspecies equisimilis					
	Household		Community		
	All	Throat	All	Throat	Total
	isolates	source	isolates	source	
Community 1	22 (18%)	22 (18%)	101 (82%)	100 (81%)	123
Community 3	3 (21%)	3 (21%)	11 (79%)	11 (79%)	14
Streptococcus pyo	genes				
	Household		Community		
	All	Throat	All	Throat	Total
	isolates	source	isolates	source	
Community 1	15 (11%)	9 (7%)	116 (89%)	75 (57%)	131
Community 3	4 (10%)	3 (7%)	38 (90%)	36 (86%)	42

494





496 Figure 1. Maximum-likelihood phylogeny of 294 Streptococcus dysgalactiae subsp. 497 equisimilis (SDSE) isolates. Sequences were aligned against SDSE reference genome 498 GGS 124 (NC 012891.1) with mobile genetic element regions masked. Distinct genomic sequence clusters determined by PopPUNK²⁵ as previously defined by a global SDSE dataset⁹, 499 500 are denoted by alternating blue and grey highlights from internal nodes. Site of isolation is 501 coloured by blue (throat) and red (skin) tips. The inner ring denotes the Lancefield group 502 carbohydrate and the outer ring the community of isolation. Bootstrap supports are shown as 503 branch colour gradients and were calculated using the ultrafast bootstrap approximation 504 demonstrating some uncertainty in deep branches of the phylogeny³⁰.

Transmission cluster - 10_11 (2) 9_11 (1) - 8_11 (1) - 7_11 (1) - 6_11 (1) 5_11 (6) - 4_11 (2) — 3_11 (1)́ 2_11 (32) 1_11 (8) $-7_2(1)$ $-5_2(1)$ $-10_2(1)$ Genomic **9_2** (3) emm subtype $-8_{2}(1)$ $-6_{2}(1)$ sequence emm type cluster MLST **-** 4_2 (2) (PopPUNK) 3_2 (6) stC1400.0 (53) 2_2 (4) 1_2 (4) stC1400 (55) 66 (55) 11 (55) stC1400.3 (1) stC1400.4 (1) 1_10 (32) stC6979 (3) stC6979.0 (3) 137 (2) 6_5 (1) stC74A (17) stC74A.0 (17) - 12_5 (1) - 11_5 (1) - 10_5 (1) 2 (24) 17 (22) stG2078 (4) stG2078.0 (4) $-9_{5}(2)$ $-8_{5}(1)$ $-7_{5}(1)$ $-5_{5}(1)$ 3 (32) 10 (32) stC839.0 (28) stC839 (32) · 717 (1) stC839.2 (4) 4_5 (2) stG480.3 (1) 5 (29) — 3_5 (1) stG480 (29) 8 (28) 2_5 (6) stG480.8 (28) 1_5 (11) 13 (35) - 15_13 (1) - 14_13 (1) - 13_13 (1) - 12_13 (1) - 12_13 (1) stG4831 (35) 75 (35) stG4831.0 (35) 26 (25) — 11_13 (1) — 10_13 (1) — 9_13 (1) stG643 (25) 171 (25) stG643.0 (25) stC36 (2) 8_13 (3) 9 (37) stC36.0 (2) **—** 7_13 (2) 4 (37) stG6792 (35) **—** 6_13 (2) — 5_13 (1) stG6792.0 (35) 4_13 (7) - 3_13 (1) 2_13 (2) 1_13 (10) - 4_26 (1) 3_26 (14) 2_26 (2) 1_26 (8) 8_9 (4) 7_9 (5) - 6_9 (1) - 5_9 (1) - 4_9 (1) 3 9 (8) 2_9(1) 1_9 (16)

506 Figure 2. Alluvial plot of the relationship between the largest 7/18 Streptococcus dysgalactiae 507 subspecies *equisimilis* (SDSE) genomic sequence clusters (representing 237/294 isolates) as determined by PopPUNK against *emm* type, *emm* subtype, multilocus sequence type (MLST) 508 509 and transmission clusters determined using single linkage clustering at a SNP threshold of < 8510 and >99% shared gene content. The number of isolates in each category is denoted within 511 brackets. From 10 emm types (14 emm subtypes) and 9 MLSTs shown, 60 transmission clusters 512 were determined. Traditional markers such as emm subtype in some cases over-split SDSE 513 clusters as demonstrated by stC839.0 and stC839.2 which differ by only one SNP within their 514 hypervariable emm region and otherwise fall within the same transmission cluster.





516 **Figure 3.** Transmission links between households across consecutive community visits in 517 community 1 **a**), and community 3 **b**). Households are represented by nodes proportional in

size to the number of participants enrolled at each household and coloured by the proportion of *S. pyogenes* (red) and *S. dysgalactiae* subsp. *dysgalactiae* (SDSE, blue) isolates detected in the household across the entire study period. Transmission links are represented by undirected and unweighted edges between households and coloured by species with shared edges highlighted in green. Loops correspond to predicted transmission edges between unique individuals within the same household. Only community visits where transmission edges were predicted are shown.





525

526 Figure 4. Co-occurrence of Streptococcus dysgalactiae subsp. equisimilis (SDSE) and S. 527 *pyogenes* in households in communities 1 and 3 at each community visit (light blue highlights) 528 during the study period. Detection of SDSE (blue) and S. pyogenes (red) are denoted by points with the size of each point proportional to the number of isolates. Community visits where a 529

- 530 household was not sampled are denoted by crosses. SDSE and S. pyogenes co-occurred on
- 531 100/486 (21%) of household-visits.



533 Figure 5. Shared mobile genetic elements (MGE) across *Streptococcus dysgalactiae* subsp. equisimilis (SDSE) and Streptococcus pyogenes isolates. a) Maximum likelihood trees of 534 SDSE and *S. pyogenes* with isolates carrying three near-identical (>99% nucleotide identity) 535 536 MGEs highlighted by tree tip points. Genomic sequence clusters without the three MGEs of 537 interest are collapsed and denoted by blue (SDSE) and red (S. pyogenes) triangles at tree tips. 538 Flows link corresponding shared MGEs across the species but do not imply directionality of 539 transfer. Bootstrap supports are shown as branch colour gradients and were calculated using ultrafast bootstrap approximation³⁰. **b**) A 54kbp prophage, ϕ 1207.3, carrying macrolide efflux 540 541 resistance genes mef(A) and msr(D) was present with >99.9% nucleotide identity across SDSE and S. pvogenes. A representative SDSE element from isolate NS4595 was aligned against a 542 543 representative S. progenes sequence (NS3871) with percentage nucleotide identity calculated 544 using Hamming distance and plotted in 100bp sliding windows. The element was present in a 545 cross-species conserved insertion region with flanking core genes highlighted in green. 546 \$\$\phi1207.3 was present in 5/8 emm58.8, MLST 549 S. pyogenes isolates with as few as 11 single 547 nucleotide polymorphisms (SNPs) between isolates with and without the prophage. The same 548 element was present in 31 SDSE isolates across subsets of five different lineages suggestive of 549 recent gain/loss events of ϕ 1207.3 in both the SDSE and S. pyogenes populations. c) An 18kbp 550 integrative conjugative element (ICE)-like MGE carrying the tetracycline resistance gene, 551 tet(M), was present with >99.9% nucleotide identity in 4 S. pyogenes and 8 SDSE isolates, 552 including two distinct SDSE populations. The element was present at three different genomic 553 insertion regions and thus flanking core genes are not shown. In the example shown, a 12bp 554 in-frame deletion was present at the 5' end of tet(M) in the SDSE element which was distant 555 from the active ribosomal binding domain. d) A 41kbp prophage ϕ MGAS5005.3 carrying the streptodornase gene sda1, was shared across species with >99.9% nucleotide identity at a cross-556 species conserved insertion region as has been described previously⁹. ϕ MGAS5005.3 was 557

- present in 14/16 S. pyogenes emm1.0, MLST 28 isolates with a maximum of 3 SNPs between
- 559 isolates and 16 SDSE isolates across subsets of two different lineages.

560 **References**

561 1. Brandt CM, Spellerberg B. Human infections due to Streptococcus dysgalactiae
562 subspecies equisimilis. *Clin Infect Dis* 2009; **49**: 766-72.

563 2. Haidan A, Talay SR, Rohde M et al. Pharyngeal carriage of group C and group G 564 streptococci and acute rheumatic fever in an Aboriginal population. *Lancet* 2000; **356**: 1167-9.

565 3. Sikder S, Williams NL, Sorenson AE et al. Group G Streptococcus Induces an 566 Autoimmune Carditis Mediated by Interleukin 17A and Interferon γ in the Lewis Rat Model of 567 Rheumatic Heart Disease. *The Journal of Infectious Diseases* 2017; **218**: 324-35.

568 4. Oppegaard O, Glambek M, Skutlaberg DH et al. Streptococcus dysgalactiae
569 Bloodstream Infections, Norway, 1999-2021. *Emerg Infect Dis* 2023; 29: 260-7.

570 5. Wajima T, Morozumi M, Hanada S et al. Molecular Characterization of Invasive 571 Streptococcus dysgalactiae subsp. equisimilis, Japan. *Emerg Infect Dis* 2016; **22**: 247-54.

572 6. Wright CM, Moorin R, Pearson G et al. Invasive Infections Caused by Lancefield
573 Groups C/G and A Streptococcus, Western Australia, Australia, 2000-2018. *Emerg Infect Dis*574 2022; 28: 2190-7.

575 7. McMillan DJ, Bessen DE, Pinho M et al. Population genetics of Streptococcus
576 dysgalactiae subspecies equisimilis reveals widely dispersed clones and extensive
577 recombination. *PLoS One* 2010; **5**: e11741.

578 8. McNeilly CL, McMillan DJ. Horizontal gene transfer and recombination in
579 Streptococcus dysgalactiae subsp. equisimilis. *Front Microbiol* 2014; **5**: 676.

580 9. Xie O, Morris JM, Hayes AJ et al. Inter-species gene flow drives ongoing evolution of
581 Streptococcus pyogenes and Streptococcus dysgalactiae subsp. equisimilis. *bioRxiv* 2023:
582 2023.08.10.552873.

583 10. Belotserkovsky I, Baruch M, Peer A et al. Functional analysis of the quorum-sensing
584 streptococcal invasion locus (sil). *PLoS Pathog* 2009; **5**: e1000651.

585 11. Armstrong BD, Herfst CA, Tonial NC et al. Identification of a two-component Class
586 IIb bacteriocin in Streptococcus pyogenes by recombinase-based in vivo expression
587 technology. *Sci Rep* 2016; 6: 36233.

Heng NCK, Ragland NL, Swe PM et al. Dysgalacticin: a novel, plasmid-encoded
antimicrobial protein (bacteriocin) produced by Streptococcus dysgalactiae subsp. equisimilis. *Microbiology (Reading)* 2006; 152: 1991-2001.

13. Lacey JA, Marcato AJ, Chisholm RH et al. Evaluating the role of asymptomatic throat
carriage of Streptococcus pyogenes in impetigo transmission in remote Aboriginal
communities in Northern Territory, Australia: a retrospective genomic analysis. *Lancet Microbe* 2023.

McDonald M, Towers RJ, Andrews RM et al. Epidemiology of Streptococcus
dysgalactiae subsp. equisimilis in tropical communities, Northern Australia. *Emerg Infect Dis*2007; 13: 1694-700.

McDonald MI, Towers RJ, Andrews RM et al. Low rates of streptococcal pharyngitis
and high rates of pyoderma in Australian aboriginal communities where acute rheumatic fever
is hyperendemic. *Clin Infect Dis* 2006; **43**: 683-9.

601 16. Khedkar S, Smyshlyaev G, Letunic I et al. Landscape of mobile genetic elements and
602 their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Research* 2022; 50:
603 3155-68.

604 17. Iannelli F, Santagati M, Santoro F et al. Nucleotide sequence of conjugative prophage 605 Φ 1207.3 (formerly Tn1207.3) carrying the mef(A)/msr(D) genes for efflux resistance to 606 macrolides in Streptococcus pyogenes. *Front Microbiol* 2014; **5**: 687.

Sumitomo T, Mori Y, Nakamura Y et al. Streptococcal Cysteine Protease-Mediated
Cleavage of Desmogleins Is Involved in the Pathogenesis of Cutaneous Infection. *Front Cell Infect Microbiol* 2018; 8: 10.

610 19. Steer AC, Jenney AW, Kado J et al. Prospective surveillance of streptococcal sore throat
611 in a tropical country. *Pediatr Infect Dis J* 2009; 28: 477-82.

612 20. Jose JJM, Brahmadathan KN, Abraham VJ et al. Streptococcal group A, C and G
613 pharyngitis in school children: a prospective cohort study in Southern India. *Epidemiol Infect*614 2018; 146: 848-53.

615 21. Turner JC, Hayden FG, Lobo MC et al. Epidemiologic evidence for Lancefield group
616 C beta-hemolytic streptococci as a cause of exudative pharyngitis in college students. *J Clin*617 *Microbiol* 1997; **35**: 1-4.

618 22. Tonkin-Hill G, Ling C, Chaguza C et al. Pneumococcal within-host diversity during
619 colonization, transmission and treatment. *Nat Microbiol* 2022; 7: 1791-804.

620 23. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome*621 *Biol* 2019; **20**: 257.

622 24. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb

623 software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018; **3**: 124.

624 25. Lees JA, Harris SR, Tonkin-Hill G et al. Fast and flexible bacterial genomic
625 epidemiology with PopPUNK. *Genome Res* 2019; 29: 304-16.

626 26. Davies MR, McIntyre L, Mutreja A et al. Atlas of group A streptococcal vaccine
627 candidates compiled using large-scale comparative genomics. *Nature Genetics* 2019; **51**: 1035628 43.

629 27. Tonkin-Hill G, MacAlasdair N, Ruis C et al. Producing polished prokaryotic
630 pangenomes with the Panaroo pipeline. *Genome Biol* 2020; 21: 180.

631 28. Jespersen MG, Hayes A, Davies MR. Corekaburra: pan-genome post-processing using
632 core gene synteny. *Journal of Open Source Software* 2022; 7: 4910.

633 29. Minh BQ, Schmidt HA, Chernomor O et al. IQ-TREE 2: New Models and Efficient

634 Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530-4.

- 635 30. Hoang DT, Chernomor O, von Haeseler A et al. UFBoot2: Improving the Ultrafast
- 636 Bootstrap Approximation. *Mol Biol Evol* 2018; **35**: 518-22.
- 637 31. Harris SR. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology.
- 638 *bioRxiv* 2018: 453142.
- 639 32. Higgs C, Sherry NL, Seemann T et al. Optimising genomic approaches for identifying
- 640 vancomycin-resistant Enterococcus faecium transmission in healthcare settings. *Nat Commun*
- 641 2022; **13**: 509.
- 642 33. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation
- 643 sequencing data. *Bioinformatics* 2012; **28**: 3150-2.
- 644 34. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 2021;
- 645 **37**: 4572-4.

646