Inter-species gene flow drives ongoing evolution of *Streptococcus pyogenes* and *Streptococcus dysgalactiae* subsp. *equisimilis*

Ouli Xie[1,2], Jacqueline M. Morris[3], Andrew J. Hayes[3], Rebecca J. Towers[4], Magnus G. Jespersen[3], John A. Lees[5], Nouri L. Ben Zakour[6], Olga Berking[6], Sarah L. Baines[7], Glen P. Carter[7], Gerry Tonkin-Hill[8], Layla Schrieber[9], Liam McIntyre[3], Jake A. Lacey[1], Taylah B. James[3], Kadaba S. Sriprakash[10,11], Scott A. Beatson[6], Tadao Hasegawa[12], Phil Giffard[4], Andrew C. Steer[13], Michael R. Batzloff[10,14], Bernie W. Beall[15], Marcos D. Pinho[16], Mario Ramirez[16] , Debra E. Bessen[17], Gordon Dougan[18], Stephen D. Bentley[18], Mark J. Walker[6,19], Bart J. Currie[4], Steven Y. C. Tong[1,20], David J. McMillan[11,21#], Mark R. Davies[3 #, *]

[#] Contributed equally to this work

* For Correspondence: Dr Mark Davies, mark.davies1@unimelb.edu.au

1. Department of Infectious Diseases, the University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

2. Monash Infectious Diseases, Monash Health, Melbourne, Australia

3. Department of Microbiology and Immunology, the University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

4. Menzies School of Health Research, Charles Darwin University, Darwin, Australia

5. European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton, UK

6. Australian Infectious Diseases Research Centre and School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia

25 7. Doherty Applied Microbial Genomics, Department of Microbiology and Immunology, The
26 University of Melbourne at The Doherty Institute for Infection and Immunity, Melbourne,
27 Australia
28 8. Department of Biostatistics, University of Oslo, Blindern, Norway
29 9. Faculty of Veterinary Science, The University of Sydney, Australia
30 10. Infection and Inflammation Laboratory, Queensland Institute of Medical Research
31 Berghofer (QIMR) Berghofer Medical Research Institute, Brisbane, Australia
32 11. School of Science & Technology, University of New England, Armidale, NSW, Australia
33 12. Department of Bacteriology, Nagoya City University Graduate School of Medical
34 Sciences, Nagoya, Japan
35 13. Tropical Diseases, Murdoch Children's Research Institute, Parkville, Australia
36 14. Institute for Glycomics, Griffith University, Southport, Australia.
37 15. Respiratory Disease Branch, National Center for Immunizations and Respiratory
38 Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
39 16. Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina,
40 Universidade de Lisboa, Lisboa, Portugal
41 17. Department of Pathology, Microbiology and Immunology, New York Medical College,
42 Valhalla, NY, USA
43 18. Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
44 19. Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia.
45 20. Victorian Infectious Disease Service, The Royal Melbourne Hospital, Peter Doherty
46 Institute for Infection and Immunity, Melbourne, Victoria, Australia
47 21. School of Science and Technology, Engineering and Genecology Research Centre,
48 University of the Sunshine Coast, Maroochydore, QLD, Australia
49

50 **Key words:**

51 *Streptococcus dysgalactiae*; *Streptococcus pyogenes*; gene flow; population genomics

52 **Abstract**

53 *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) is an emerging cause of human infection
54 with invasive disease incidence and clinical manifestations comparable to the closely related
55 species, *Streptococcus pyogenes*. Through systematic genomic analyses of 501 disseminated
56 SDSE strains, we demonstrate extensive overlap between the genomes of SDSE and *S.*
57 *pyogenes.* More than 75% of core genes are shared between the two species with one third
58 demonstrating evidence of cross-species recombination. Twenty-five percent of mobile genetic
59 element (MGE) clusters and 16 of 55 SDSE MGE insertion regions were found across species.
60 Assessing potential cross-protection from leading *S. pyogenes* vaccine candidates on SDSE,
61 12/34 preclinical vaccine antigen genes were shown to be present in >99% of isolates of both
62 species. Relevant to possible vaccine evasion, six vaccine candidate genes demonstrated
63 evidence of inter-species recombination. These findings demonstrate previously unappreciated
64 levels of genomic overlap between these closely related pathogens with implications for
65 streptococcal pathobiology, disease surveillance and prevention.

66 **Introduction**

67 *Streptococcus dysgalactiae* subspecies *equisimilis* (SDSE), a beta-hemolytic *Streptococcus*
68 normally possessing the Lancefield group C/G antigen, is an emerging cause of human disease
69 with recently reported incidences of invasive disease comparable to or surpassing that of the
70 closely related and historically important pathogen, *Streptococcus pyogenes* (group A
71 *Streptococcus*) [1-8]. SDSE and *S. pyogenes* occupy the same ecological niche and possess
72 overlapping disease manifestations including pharyngitis, skin and soft tissue infections,
73 necrotising fasciitis, streptococcal toxic shock syndrome and osteoarticular infections[9, 10].
74 Gene transfer between SDSE and *S. pyogenes,* including housekeeping multi-locus sequence
75 typing (MLST) loci, major virulence factors including the *emm* gene, and antimicrobial
76 resistance (AMR) determinants has been reported[11-17]. Inter-species transfer of genes encoding
77 the serogroup carbohydrate have led to SDSE isolates which express the group A
78 carbohydrate[15, 18]. Transfer of accessory virulence or AMR genes is thought to occur by cross-
79 species exchange of mobile genetic elements (MGEs) such as prophage or integrative
80 conjugative elements (ICE) [19-21]. However, the mechanism that enables the exchange of genes
81 present in all strains of a species, termed 'core' genes, is not well understood. The extent of
82 genetic exchange between these two pathogens has not been defined within a global population
83 genomic framework.
84 An analysis of the population structure of a globally diverse collection of *S. pyogenes* genomes
85 provided insights into the drivers of population diversity and global utility of candidate
86 vaccines. In contrast, studies of SDSE whole genome diversity have generally been[7, 22-25]. With
87 increasing efforts to develop a *S. pyogenes* vaccine[26, 27], an improved understanding of the
88 overlap and extent of genetic similarity between human isolates of SDSE and *S. pyogenes* in a
89 global context is required. Here, we have compiled a globally diverse collection of 501 SDSE
90 genomes isolated from human hosts. Through a gene synteny-based approach, we conducted a

91 systematic analysis of the population genetics and pangenome of SDSE. Using the same
92 framework, we compare SDSE to a previously published global *S. pyogenes* dataset[22] to reveal
93 extensive genomic overlap between the two closely related pathogens including genes
94 encoding candidate *S. pyogenes* vaccine antigens.
95
96 **Results**
97 *A globally diverse SDSE genomic database.*
98 To assess SDSE population diversity (Figure 1), we compiled a genomic database of 501
99 geographically distributed SDSE isolates from 17 countries including 228 newly reported
100 genomes with one new complete reference quality genome, NS3396[28] (Supplementary Table
101 1a). The database includes 53 *emm* sequence types, 88 *emm* sub-types, and 129 MLSTs.
102
103 *Global SDSE population structure.*
104 A phylogeny of the global SDSE population was constructed using a recombination aware
105 pipeline implemented in Verticall. The minimum evolution phylogeny demonstrated a deep
106 radially branching structure forming multiple distinct lineages similar to that observed for the
107 global population structure of *S. pyogenes*[22] (Figure 1a and Supplementary Figure 1).
108 Whole genome clustering of the global SDSE population using PopPUNK[29] detected 59
109 distinct population clusters (akin to 'lineages' or 'genome clusters') (Figure 2a and
110 Supplementary Figure 2). These genome clusters were geographically dispersed and were
111 highly concordant with the inferred phylogeny.
112 Our data show a limited concordance between the inferred whole genome population structure
113 and the classical SDSE molecular epidemiological markers, *emm* type and MLST
114 (Supplementary Figure 3). Of the 24 *emm* types represented by three or more isolates, 11 (46%)

115 were present in multiple genome clusters. Of the 27 genome clusters with three or more
116 isolates, 18 (67%) contained two or more *emm* types.
117 MLST was in stronger agreement with the inferred genome clusters than *emm* type.
118 Nevertheless, the largest genetic distance between isolates in the same MLST was frequently
119 greater than that separating isolates from different MLST types (Supplementary Figure 3d).
120 Supporting this observation, 67% (18/27) of genome clusters with three or more members
121 contained more than one MLST. While many of these differed at only a single MLST locus,
122 single locus variant MLST clonal complexes delineated distant lineages poorly and were
123 present across multiple distinct genetic backgrounds (Supplementary Figure 3e). These
124 findings indicate that while *emm* and MLST typing for SDSE have been useful epidemiological
125 markers in jurisdictional or regional contexts, they have limitations when defining global SDSE
126 evolution and diversity. The presented whole genome clustering scheme implemented in
127 PopPUNK provides a solution to the limitations of *emm* and MLST at the global population
128 level[29].
129 The Lancefield group carbohydrate was predicted by the presence of a 14 gene carbohydrate
130 synthesis gene cluster in group C SDSE[18], 15 gene cluster in group G[15] and 12 gene cluster in
131 group A *Streptococcus*[15] (Supplementary Figure 4a) in a conserved genomic location
132 immediately upstream of *pepT*. Of the 27 genome clusters which contained three or more
133 isolates, 16 consisted of only group G isolates and 9 consisted of only group C (Supplementary
134 Figure 4b). Two genome clusters consisted of isolates with more than one group carbohydrate.
135 The Lancefield group A carbohydrate, normally associated with *S. pyogenes*, was found in
136 isolates from 4 different genome clusters (Figure 2a), including one genome cluster that also
137 contained isolates with the group C and group G antigen. These findings suggest that while
138 closely related isolates generally express the same Lancefield group carbohydrate antigen,

139 horizontal transfer of the group carbohydrate locus can occur, including the group A
140 carbohydrate locus normally associated with *S. pyogenes*.
141 Bayesian dating of the most recent common ancestor (MRCA) of the two most well sampled
142 genome clusters using BactDating[30] with recombination masked by Gubbins[31], revealed
143 emergence of these lineages within the last 100 years. A genome cluster containing members
144 present in 7 countries and inclusive of *emm* type *stG*6792/ST 17 isolates, until recently the
145 most frequent cause of invasive disease in Japan[7, 10, 32], had an estimated root date in the mid-
146 1970s (credible interval CrI 1960 – 1991, Supplementary Figure 5a) suggesting recent
147 emergence and global dispersion. A genome cluster consisting of isolates with *emm* type
148 *stG*62647, an *emm* type described to be increasing in frequency in multiple jurisdictions[24, 33],
149 had an estimated root date in the early 1930s (CrI 1855 – 1969, Supplementary Figure 5b).
150 These findings suggest that major genome clusters in this global database represent modern,
151 globally disseminated lineages.
152 Virulence genes or regulators known to influence expression of virulence genes were enriched
153 in a subset of genome clusters. Isolates were screened for the presence of accessory toxins
154 including DNAses, superantigens, adhesins, immune escape factors and regulators described
155 in SDSE or *S. pyogenes* (Supplementary Table 1b). We found that certain genome clusters
156 were enriched for accessory factors including: chromosomally encoded exotoxin *speG*,
157 immune escape factor *drsG*, adhesin *gfbA* and an accessory/secondary FCT locus, presence of
158 the negative regulatory *sil* locus, and three accessory prophage streptodornase genes *spd3*, *sda1*
159 and *sdn* (all $p < 0.001$, χ^2 test of independence). In particular, the accessory FCT locus[34], *sil*
160 locus, *gfbA*, *speG* and *drsG* were frequently either present or entirely absent from different
161 genome clusters. These findings suggest that the SDSE genetic population structure is
162 characterised by distinct virulence repertoires. Non-random sampling prevented associations
163 with clinical manifestations from being inferred.

164

165 *Signatures of recombination in the SDSE core genome.*

166 The SDSE pangenome consisted of 6,228 genes of which 1,561 were core genes (present in
167 ≥99% of genomes) and 4,667 were accessory genes, of which 3,849 genes were present in
168 <15% of genomes (Figure 2b). While several SDSE genes have previously been reported to be
169 recombinogenic, we next aimed to infer the number of SDSE core genes with signatures of
170 recombination in their evolutionary history using fastGEAR[35]. Of 1,543 core genes (excluding
171 18 genes that had over 25% gaps in their alignment), 837 (54%) genes had a recombinatorial
172 signature (Supplementary Table 2a). These notably included all seven MLST genes. These
173 findings correlate with uncertainties when classifying isolates using MLST compared to whole
174 genome clusters.

175 To further quantify the contribution of recombination to SDSE population diversity, we
176 measured the ratio of recombination-derived mutation vs vertically inherited mutation (r/m)
177 using Gubbins[31] for the 12 largest genome clusters (385 isolates). The median r/m per genome
178 cluster was 4.46 (range 0.38 – 7.05) which was comparable to a r/m of 4.95 estimated
179 previously for *S. pyogenes*[22] (Supplementary Table 3). The median recombination segment
180 length was 4,647 bp (range 6 – 97,789 bp) and the median number of events per genome cluster
181 was 42 (range 3 – 64), again similar to that previously described for *S. pyogenes*[22], supporting
182 multiple small fragments of homologous recombination as a major source of genetic diversity
183 in SDSE.

184

185 *Contribution of the accessory genome to SDSE diversity.*

186 To investigate the contribution of MGE and non-MGE genes to accessory genome diversity,
187 an identification and characterisation workflow was developed using genome synteny and a
188 classification algorithm adapted from proMGE[36]. Segments of accessory genes (hereafter

189 referred to as 'accessory segments') were classified as prophage, phage-like, ICE, insertion
190 sequence/transposon (IS) or non-MGE based on the presence of MGE-specific
191 integrase/recombinase genes and prophage or ICE structural genes (Supplementary Figure 6a).
192 Accessory segments were classed as mobility elements (ME) when insufficient information
193 was present to classify a segment – such as with degraded elements, complex nested elements
194 or elements fragmented by assembly breaks with insufficient contextual information. Genes
195 were classified into MGE categories based on the frequency of their presence in different
196 elements (Figure 2b). Prophage elements contributed the largest number of genes to the
197 accessory genome (36%, 1,697/4,667 genes) followed closely by ME (25%, 1,164/4,667
198 genes) and ICE (23%, 1,084/4,667 genes). Non-MGE accessory genes constituted 8%
199 (393/4,667 genes) of the accessory genome. A mean of 0.8 prophage (range 0 – 5), 1.1 phage-
200 like (range 0 – 4), 1.1 ICE (range 0 – 4) and 4.7 ME (range 0 – 10) were found per genome.
201 MGE counts were similar when restricted to four complete SDSE genomes (mean 1.3
202 prophage, 1 phage-like, 1 ICE, 2.8 ME). A mean of 10.5 IS/transposon elements (range 3 – 30)
203 were found per genome but was likely to be an underestimation as parameters used to construct
204 the pangenome resulted in the removal of infrequent genes at genome assembly breaks, which
205 are commonly associated with IS/transposons.
206 Using genome location defined between two syntenic core genes, MGE chromosomal
207 'insertion regions' were mapped in the SDSE dataset. A total of 55 insertion regions were found
208 in SDSE (Figure 3a, Supplementary Table 4a). Prophage or phage-like elements were found at
209 32 insertion regions (58%), ICE at 21 regions (38%), and 10 regions (18%) were occupied by
210 either prophage or ICE. Twelve regions (22%) were occupied by elements classified as ME
211 only, reflective of the limitations imposed by sequence breaks around MGEs in draft genomes.
212 The MGE insertion regions were broadly occupied across SDSE genome clusters
213 (Supplementary Figure 6b).

214

215 *Conservation of MGE insertion regions across SDSE and S. pyogenes are associated with*
216 *shared elements.*

217 Applying the same accessory identification and categorisation workflow to the 2,083 *S.*
218 *pyogenes* genomes published previously[22], enabled a systematic comparison of MGE and
219 associated insertion sites both within and between the two species. An average of 1.9 prophage
220 (range 0 – 6), 0.4 phage-like (range 0 – 3), 0.2 ICE (range 0 – 2), and 1.9 ME (range 0 – 7)
221 were found per *S. pyogenes* genome. Compared to SDSE, *S. pyogenes* isolates had more
222 prophage elements but less ICE ($p < 0.001$, Wilcoxon rank sum).

223 Overlaying the SDSE and *S. pyogenes* pangenomes while accounting for genome synteny, 31
224 previously published[37, 38] and 13 new *S. pyogenes* MGE insertion regions (31 prophage, 9 ICE,
225 and 2 mixed prophage and ICE) were mapped and compared to 55 insertion regions in SDSE
226 (Supplementary Table 4c). Of these, 16 (29%) insertion regions were shared across the species
227 (Figure 3a). At the 16 cross-species insertion regions, 1,443 accessory genes (54% of SDSE
228 and 59% of *S. pyogenes* accessory genes at these regions) were shared across species,
229 suggesting likely shared MGEs at these regions. At insertion regions which were not conserved
230 across the two species, 816 accessory genes (34% of SDSE and 42% of *S. pyogenes* accessory
231 genes) were shared, significantly less than the proportion at conserved insertion regions ($p <$
232 0.001, χ^2 test).

233 Hypothesising that shared MGE insertion regions may provide a basis for shared prophage and
234 ICE between species, all completely assembled and fragmented putative MGEs >10-15kb at
235 the 16 cross-species insertion regions were examined. A total of 3,335 fully assembled and 295
236 high-quality fragmented putative MGEs, 710 from SDSE and 2,920 from *S. pyogenes*, were
237 extracted. To identify similar elements within this database, we used mge-cluster[39] to cluster
238 ICE, ME and prophage elements which overcomes limitations of sequence homology-based

239 approaches which are restricted by the modular and highly recombinogenic nature of MGEs.
240 Using this approach, 40 clusters containing 2,897 MGEs were identified of which 10 (25%)
241 ICE and prophage clusters across 13 insertion regions were found in both species (Figure 3b).
242 The clusters were generally well defined and separated by MGE type (Supplementary Figure
243 7). Of the 40 clusters, 14 were present at more than one insertion region (median 1, range 1 –
244 7) indicating broad sharing of MGE clusters across these common insertion regions and
245 between species.
246 We further examined MGE clusters for examples of shared near-identical elements (at least
247 >89% nucleotide identity and coverage). Four MGEs (prophage, ICE and nested IS/transposon
248 elements) carrying streptodornase genes *sda1* and *spd1*, the exotoxin gene *speC*, and AMR
249 genes *ermB*, *ant(6)-Ia* and *aph(3')-IIIa* were detected across both species (Table 1,
250 Supplementary Figure 8). A completely assembled prophage element that was near-identical
251 (>99% nucleotide identity) to the *S. pyogenes* M1T1 prophage ϕ5005.3 was shared across 12
252 genomes in SDSE and 104 genomes in *S. pyogenes* (Figure 3c). The M1 prophage ϕ5005.3
253 carries the streptodornase gene *sda1*, an extracellular virulence factor thought to have been
254 acquired during the emergence of the global M1T1 *S. pyogenes* lineage[40, 41]. To investigate the
255 distribution of this prophage further in SDSE, including MGEs fragmented by contig breaks,
256 co-presence of the same integrase and *sda1* allele was examined across the SDSE database. An
257 additional 8 genomes with co-carriage of these genes at the same insertion region were found.
258 The ϕ5005.3 prophage was inferred to be present in SDSE isolates spanning 10 different
259 genome clusters across seven countries (Australia, Canada, China, Denmark, Norway, Portugal
260 and USA), over a 20-year period (1999 to 2018). Within these genome clusters, a mean of
261 28.8% of isolates carried the ϕ5005.3 phage (range 1.5 – 100%). These findings indicate that
262 MGEs carrying virulence and/or AMR determinants can cross species boundaries including
263 evidence of dispersal into more than one global SDSE lineage.

264

265 *Extensive overlap and exchange in the core genome between SDSE and S. pyogenes.*

266 To examine core genome overlap across SDSE and *S. pyogenes*, we found that 1,166 core genes
267 comprising 75% (1,166/1,547) of the SDSE core genome and 88% (1,166/1,320) of the *S.*
268 *pyogenes* core genome were shared in the merged pangenome (Figure 4a, Supplementary Table
269 2b). A small number of genes/coding sequences were combined when merging pangenomes,
270 resulting in a slightly smaller merged core pangenome compared to SDSE or *S. pyogenes* alone.
271 To investigate cross species recombination in these core genes, 1,166 genes that were core in
272 both species and with <25% length variation were aligned and assessed using fastGEAR[35]. A
273 total of 526 core genes (45%) had clusters with members from both species consistent with
274 either shared ancestry or whole gene recombination. Putative cross-species recombination was
275 identified in 393 (34%) unique genes, including the MLST genes *gki*, *murI*, and *recP* and the
276 penicillin-binding protein-encoding gene *pbp1b*, which have previously been documented to
277 be affected by inter-species recombination[11, 16] (Supplementary Table 2b). Of these, 216 genes
278 had a signature of multiple unique events between species. Recombination affected genes
279 across all functional categories with no significant difference between classes ($p = 0.26$, χ^2 test
280 of independence). While net directionality and absolute frequency of events cannot be inferred
281 using this data, predicted cross-species recombination events affected genes from across the
282 genome with few hot-spot regions of higher density or restriction (Figure 4b).

283

284 *Predicted conserved metabolic pathways between species identified by pangenome analysis.*

285 Examination of non-MGE genes unique to the pangenomes of each species found well-defined
286 KEGG modules predicting metabolic differences between the species. Core to SDSE but absent
287 from the *S. pyogenes* pangenome were modules encoding glycogen biosynthesis (M00854) and
288 threonine biosynthesis (M00018). *S. pyogenes* is known to be auxotrophic for threonine and

289 the absence of threonine and glycogen biosynthesis genes may reflect greater host dependence
290 and/or adaptation. Unique to *S. pyogenes* were multiple genes encoding V/A-type ATPases
291 (M00159). While additional differences are likely to exist beyond described KEGG modules,
292 69% of SDSE and 86% of *S. pyogenes* metabolic genes were shared indicating extensive
293 overlap between the species (Supplementary Figure 9).
294
295 *Conservation of leading S. pyogenes vaccine candidates in the global SDSE population.*
296 Given the extensive overlap in gene content between SDSE and *S. pyogenes*, we next assessed
297 the carriage of 34 leading *S. pyogenes* candidate vaccine antigens (Supplementary Table 5) [22]
298 in the global SDSE population. Of the 26 candidate vaccine antigens using a full or near-full
299 length gene product, 12 were highly represented in both species (present in >99% of isolates at
300 70% identity, Figure 5a). Mean amino acid sequence divergence in SDSE for these 12
301 candidates from the *S. pyogenes*-derived reference sequence varied from 80.9% to 99.9%
302 (Figure 5a, Supplementary Figure 10a). Of the four small peptide vaccine candidates, the
303 multivalent N-terminal M protein and multivalent Tee antigen candidates, which were searched
304 at 100% identity, only J8.0 (C-terminal fragment of M protein) was present in all SDSE
305 isolates. However, none were present in >99% of isolates in both species (Figure 5b). Potential
306 coverage by five of 11 leading preclinical multicomponent vaccines was >99% in both
307 populations (Figure 5c, Supplementary Table 6).
308 Shared antigens may provide cross-species vaccine coverage but conversely, interspecies
309 recombination could yield increased antigenic diversity. Interspecies recombination analysis
310 using fastGEAR was carried out on nine full length antigens which had <25% length variation
311 and were highly present in both species. Five candidates, SpyAD, OppA, Shr, PulA and Fbp54,
312 demonstrated signatures of recombination between a single-species cluster and a genome of
313 the other species (Figure 5d and Supplementary Figure 10b). Three candidates, SLO, ADI and

314 TF, had sequence clusters containing both species. Alleles of *srtA* were separated by species
315 between SDSE and *S. pyogenes*. Examination of shared clusters found evidence of SDSE
316 isolates which had acquired a TF allele from *S. pyogenes* (Supplementary Figure 10b).
317 However, in the setting of limited sequence diversity, recombination could not be inferred for
318 SLO and ADI.
319
320 **Discussion**
321 With increasing disease control efforts for *S. pyogenes* including vaccine development[26, 27, 42]
322 and reports of high disease burden associated with invasive SDSE infection[1-8], an improved
323 understanding of the SDSE genomic population structure and overlap between these closely
324 related pathogens provides a new framework for understanding the evolution and prevention
325 of disease associated with infection by these human pathogens.
326 The population structure of SDSE was found to have many similarities to that of *S. pyogenes*
327 with multiple evolutionary distinct lineages. Recombination of small genomic fragments was
328 found to be a major driver of diversity of the core genome in SDSE with variation in the
329 accessory genome due to MGE-related genes. These features mirror findings described
330 previously for *S. pyogenes*, suggesting similar evolutionary dynamics at a global population
331 level across these pathogens[22].
332 At the level of the core genome, more than 75% of core genes were shared between each
333 species. Despite the extensive overlap, metabolic differences were found in well characterised
334 KEGG modules such as the presence of threonine and glycogen biosynthesis modules which
335 are core in SDSE but absent in *S. pyogenes*. SDSE has on average, a larger genome than *S.*
336 *pyogenes* (2.1Mbp vs 1.8Mbp). The threonine and glycogen biosynthesis modules were also
337 present in 8 complete *Streptococcus dysgalactiae* subsp. *dysgalactiae* genomes available on
338 RefSeq (accessed 6 March 2023). While theoretical, it is plausible that these metabolic

339 differences may reflect greater levels of genome reduction in *S. pyogenes*, under the assumption
340 that human isolates of SDSE may have more recently diverged from a multi-host reservoir[43].
341 We found extensive horizontal gene transfer through homologous recombination across the
342 SDSE and *S. pyogenes* core genome, with over a third of SDSE genes demonstrating evidence
343 of cross-species recombination. These findings extend previous recognition of such
344 interspecies recombination of MLST genes[11, 17] and is likely a combination of both ancestral
345 and ongoing genetic transfers. While we do not quantify the frequency of cross-species
346 recombination, the extent of recombination is comparable to that reported by Diop *et al.* who
347 used measures of homoplasy and sliding window-based sequence identity[44]. This dataset may
348 also provide a foundation for development of future methods controlling for population
349 structure to infer net directionality and frequency of interspecies recombination.
350 Over 25% of MGE insertion regions and MGE clusters were found to be conserved across
351 SDSE and *S. pyogenes*, suggesting much greater sharing of MGEs across the species than
352 previously appreciated. Within MGE clusters, near-identical complete MGEs were detected
353 including ϕ5005.3, the prophage carrying streptodornase Sda1 found in the globally successful
354 M1T1 *S. pyogenes* lineage[41, 45]. The prophage was present across multiple distinct lineages in
355 SDSE indicating sharing and dissemination of the element in the SDSE population.
356 Extending these findings to *S. pyogenes* vaccine candidates, 12 of 34 antigens and five of 11
357 multicomponent vaccine candidates were predicted to contain antigens present in >99% of
358 isolates from both species. Of the 12 antigens, 6 had evidence of interspecies recombination
359 including components of multivalent vaccines, SpyAD, OppA, TF and PulA. This suggests that
360 SDSE may represent an additional reservoir for antigenic diversity particularly if vaccine
361 candidates enhance selection in an antigenically diverse region. Even greater diversity likely
362 exists considering current limited whole genome sampling of SDSE in the continents of Africa

363 and South America. Thus, surveillance of SDSE should be considered in the context of *S.*
364 *pyogenes* vaccine development and monitoring.
365 Although published data on preclinical efficacy of vaccine candidates in SDSE is limited
366 beyond preliminary data on the J8 peptide[46], the prevalence of conserved antigens suggest there
367 may be cross-species effects of vaccines intended to target *S. pyogenes*. However, the
368 sequence-based approach used in this study does not consider potential conserved structural
369 epitopes which may confer antibody cross-reactivity or cross-opsonisation to divergent alleles.
370 It should be noted though that an immunological correlate of protection has yet to be
371 determined for *S. pyogenes* or SDSE.
372 This systematic and detailed analysis of the overlap between SDSE and *S. pyogenes* at the level
373 of the core genome, recombination, and MGEs reveals the extensive shared genomic content
374 between these closely related pathogens and provides a platform for further investigations into
375 their shared biology. Genomic exchange however is not limited to movement between these
376 two organisms. Indeed, horizontal gene transfer between SDSE and another beta-haemolytic
377 *Streptococcus*, *Streptococcus agalactiae*, has been described and elements of SDSE biology
378 such as a second FCT locus, are more closely associated with *S. agalactiae* than *S. pyogenes*[34].
379 These methods could therefore be applied to other closely related species to provide insight
380 into the shared biology between disease-causing streptococci in humans.

381 **Methods**

382 *Bacterial isolates*

383 The collection of 501 global SDSE sequences included publicly available short-read sequence
384 data from NCBI sequence read archive (SRA) and complete genome assemblies from NCBI
385 RefSeq until 4 May 2022. These included studies of invasive and non-invasive SDSE from
386 Japan[47-49], Germany[50], India[51], China[52], Canada[23], Norway[24, 34, 43], USA[18, 19], Denmark[53],
387 Switzerland[54] and The Gambia[55]. A further 228 invasive and non-invasive isolates were
388 collected from Australia, France, USA, India, USA, Argentina, Trinidad, Japan, Fiji, and
389 Portugal to maximise geographical diversity. Metadata for the isolates are available in
390 Supplementary Table 1.

391

392 *Genome sequencing, assembly and quality control*

393 For the newly sequenced genomes, microbial DNA was extracted and 75-100bp paired-end
394 libraries were sequenced using the Illumina HiSeq 2500 platform (The Wellcome Sanger
395 Institute, United Kingdom). Reads were trimmed using Trim Galore v0.6.6
396 (https://github.com/FelixKrueger/TrimGalore) with a Phred score threshold of 20-25 and
397 filtered to remove reads <36bp. Reads were examined for contamination using Kraken2
398 v2.1.2[56]. Draft assemblies were generated using Shovill v1.1.0
399 (https://github.com/tseemann/shovill) with SPAdes assembler v3.14.0[57] and a minimum contig
400 length of 200bp. Only assemblies with <150 contigs (mean 94, range 57-149), total assembly
401 size between 1.9-2.4Mb (mean, 2.11Mb, range 1.91 – 2.30Mb), and GC% between 38-40%
402 (mean 39.3%, range 38.7 – 39.6%) were included. The mean N50 was 72,091bp (range 42,987
403 – 146,734bp). Annotations were generated using Prokka v1.14.6[58] using the '–proteins' flag
404 with four Refseq *S. pyogenes* and SDSE genomes to supplement annotations for consistency.
405 *emm* sequence typing was performed using emmtyper v0.2.0 (https://github.com/MDU-

406 PHL/emmtyper) and MLST assigned using MLST v2.22.0
407 (https://github.com/tseemann/mlst)[59]. The Lancefield group carbohydrate was inferred by the
408 presence genes in the 14 gene group C carbohydrate synthesis locus[18], 15 gene group G[15] locus
409 and 12 gene group A[15] locus in the SDSE pangenome. The carbohydrate synthesis locus was
410 present in a conserved location between core genes *mscF* (SDEG_RS03715) and *pepT*
411 (SDEG_RS03795). Reads for all newly sequenced genomes are available on SRA with
412 accession numbers provided in Supplementary Table 1.
413 To make a complete genome sequence of the *S. dysgalactiae* subsp. *equisimilis* strain NS3396,
414 high quality genomic DNA was extracted using the GenElute Bacterial Genomic DNA Kit
415 (Sigma). The complete genome assembly of NS3396 (GenBank accession CP128987) was
416 performed using SMRT analysis system v2.3.0.140936 (Pacific Biosciences). Raw sequence
417 data was *de novo* assembled with the HGAP3 protocol. Polished contigs were error corrected
418 with Quiver v1, and the final assembly structure check by mapping raw reads against the
419 alignment with BridgeMapper v1 as previously described[60].
420 Complete genomes were checked for genome arrangement and duplications around RNA
421 operons using socru v2.2.4[61]. Genomes with unverified large-scale rearrangements and
422 duplications around RNA operons compared to published complete genomes NC_012891.1
423 GGS_124, NC_018712.1 RE378 and NC_019042.1 AC-2713 were excluded. Four complete
424 genomes were included in the final database.
425
426 *Phylogenetic analysis*
427 Recombination masked distances between SDSE genomes were calculated using Verticall
428 v0.4.0 (https://github.com/rrwick/Verticall) which conducts pairwise comparisons between
429 genomes and masks regions with increased and/or decreased sequence divergence as putatively
430 recombinogenic. Using recombination masked distances, a minimum evolution phylogenetic

431 tree was generated using fastME v2.1.6.1[62] with nearest neighbour interchange using BME
432 criterion for optimisation and subtree pruning and regrafting.
433 A comparative maximum-likelihood tree, without masking of recombination, was generated
434 by alignment of all genomes against reference genome NC_012891.1 GGS_124 using Snippy
435 v4.6.0 (https://github.com/tseemann/snippy). MGE regions were masked and tree inference
436 conducted with IQ-tree v2.0.6[63] using a GTR+F+G4 model. Tree comparisons were performed
437 using TreeDist v2.5.0[64] to calculate generalised Robinson-Foulds distances with mutual
438 clustering information and matched splits were visualised with the 'VisualizeMatching'
439 function.
440 Bayesian dating of the MRCA of the two most sampled genome clusters (groups '1' and '2')
441 was performed using BactDating v1.1[30] with an additive relaxed clock model and 10^8 iterations
442 to ensure Markov Chain Monte Carlo convergence and parameter effective sample size >190.
443 Recombination-masked alignments were used as input. Alignments were generated using
444 Snippy v4.6.0 against a reference genome within the same genome cluster when available,
445 NC_018712.1 for genome cluster '2', or a high-quality draft genome (SRR3676046) for
446 genome cluster '1'. Regions affected by recombination were masked using Gubbins v3.1.2[31]
447 and IQ-TREE with a GTR+G4 model for tree building, and set with maximum 10 iterations,
448 minimum of 5 SNPs to identify a recombination block, minimum window size of 100bp and
449 maximum window size of 10,000bp. Contigs were padded with 10,000 Ns, the maximum
450 window size, for the genome cluster '1' alignment to prevent Gubbins calling recombination
451 across contigs. Four non-*stG*62647 isolates were excluded from genome cluster '1' for dating
452 analysis as they were >1500 SNPs distant and formed a distinct sub-lineage within the cluster.
453
454 *SDSE population genomics*

455 Evolutionarily related clusters (genome clusters) were defined using PopPUNK v2.4.0[29] which
456 has previously been used to describe the population genomics of *S. pyogenes*. PopPUNK
457 assigns clusters based on core and accessory distances calculated using sliding k-mers.
458 Distances were calculated using k-mer sizes between 13 to 29 at steps of four (Supplementary
459 Figure 3a). A three-distribution Bayesian Gaussian Mixture Model (BGMM) was fit with 2D
460 cluster boundary refinement to obtain a network score of 0.93 and was chosen after
461 benchmarking against a model built using HDBSCAN and BGMM models with different
462 numbers of distributions (Supplementary Figure 3b and 3c). The PopPUNK model (v1) and
463 genome cluster designations from this study are available at https://www.bacpop.org/poppunk/
464 and can be iteratively expanded.
465 Genomic distances used to compare PopPUNK genome clusters with traditional
466 epidemiological markers, *emm* and MLST, were recombination masked distances calculated
467 by Verticall as above. Distances are analogous to 100% minus average nucleotide identity.
468 Single locus variant MLST clonal complexes were calculated using the goeBURST algorithm
469 as implemented in PHYLOViZ v2.0[65].
470 Known SDSE and *S. pyogenes* virulence factors were screened in the SDSE database using
471 screen_assembly v1.2.7[22] with sequence accessions as listed in Supplementary Table 1b at 70%
472 nucleotide identity and 70% coverage. Virulence factors were supplemented using Abricate
473 v1.0.1 (https://github.com/tseemann/abricate) with VFDB[66] at 70% nucleotide identity and
474 70% coverage. Genes with length variation and assembly breaks due to large repeat regions,
475 were screened using Magphi v1.0.1[67] for conserved 5' and 3' sequences at a distance equal to
476 the maximum known length of the gene. Antimicrobial resistance genes were screened using
477 Abritamr v1.0.6 (https://github.com/MDU-PHL/abritamr), a wrapper for AMRfinder plus
478 v3.10.18[68], at default 90% identity and 50% coverage.
479

480 *SDSE pangenome and MGE analysis*

481 The SDSE pangenome was defined using Panaroo v1.2.10[69] which utilises a pangenome graph-
482 based approach for pangenome clustering. Panaroo was run in 'strict' mode with initial
483 clustering at 98% length and sequence identity followed by a family threshold of 70% to
484 collapse syntenic gene families. Core genes were defined as genes present in ≥99% of genomes,
485 shell accessory genes between 15%-99% of genomes, and cloud accessory genes in <15% of
486 genomes. COG functional categories were assigned to core genes using eggNOG-mapper
487 v2.1.7[70] with default Diamond mode. COG categories J, K, L, A, B and Y were summarised as
488 genes involved in 'information storage and processing', categories T, D, V, U, M, N, O, W and
489 Z were summarised as genes involved in 'cellular processing and signalling', categories C, G,
490 E, F, H, I, P and Q were summarised as genes involved in 'metabolism'.

491 MGE detection and classification was performed by adapting an automated classification
492 algorithm by Khedkar *et al.* [36] and enhanced by Corekaburra[71]. Corekaburra maps core gene
493 consensus synteny from pangenomes and was used to find stretches of accessory genes or
494 'accessory segments' between core genes which are used as anchor points. Full details of the
495 logic for classification of MGEs is as described by Khedkar *et al.* [36] and Hidden Markov Models
496 (HMMs) are available at http://promge.embl.de/. Briefly, accessory segments were
497 investigated for integrase/recombinase subfamilies using HMMER v3.3.2[72] 'hmmsearch' with
498 model-specific gathering thresholds. Representative translated protein sequences from the
499 pangenome were used and resulted in almost identical results compared to searches using
500 sequences from individual genomes. Integrase/recombinase subfamilies were mapped to
501 specific MGE classes: prophage, ICE, IS/transposon and ME (or mobility islands). For
502 subfamilies associated with more than one MGE class, additional information including
503 presence of prophage or ICE structural genes was required. Prophage structural genes were
504 annotated using eggNOG-mapper v2.1.7[70]. At least two prophage genes with an appropriate

505 integrase/recombinase were required to classify an element as a prophage. ICE structural genes
506 were classified using HMMs from TXSScan[73] and 'hmmsearch' with an E-value threshold of
507 0.001. Unlike the original method by Khedkar *et al.*, we required the presence of at least one
508 coupling protein and one T4SS ATPase with an appropriate integrase/recombinase for
509 classification of an ICE to improve specificity. Elements with a prophage specific
510 integrase/recombinase, but no prophage structural genes were classified as phage-like.
511 Accessory segments without enough contextual information or which contained both prophage
512 and ICE structural genes were classified as ME. ME may also represent degraded elements,
513 integrative mobilizable elements (IME), or accessory segments fragmented by sequence breaks
514 with a disconnect between the integrase/recombinase and structural genes. The boundaries of
515 nested segments with more than one recombinase were not resolved and we did not define the
516 exact *attR* and *attL* sites.
517 Classification of accessory genes into non-MGE or MGE classes was based on the frequency
518 each gene was detected in each class. Accessory segments with prophage or ICE structural
519 genes but no integrase or unclassified segments with a sequence break did not contribute to the
520 count as they may represent part of a fragmented MGE. Genes were classified as MGE-related
521 if it appeared on any MGE more frequently than it was found on a non-MGE element. A small
522 number of genes only present adjacent to contig breaks without an accompanying integrase
523 were considered unclassified as they could not be confidently grouped into an MGE or non-
524 MGE category. For the purposes of gene classification, phage-like elements were grouped with
525 prophage.
526
527 *SDSE recombination detection*
528 To examine evidence of recombination within the core SDSE genome, fastGEAR[35] was run on
529 1,543 core gene alignments from the 501 SDSE strains included in the study. Gene alignments

530 were performed using MAFFT v7.505[74]. Alignments with greater than 25% gap characters
531 were excluded. fastGEAR infers population structure for each alignment, allowing for the
532 detection of lineages or clusters that have 'ancestral' and 'recent' recombination events
533 between them. Default parameters were used with a minimum threshold of 4 bp applied for the
534 recombination length.
535 The relative ratio of mutation due to recombination to vertically inherited mutation (r/m) was
536 determined for the 12 most frequently sampled genome clusters (385 isolates) using Gubbins
537 v3.1.2[31] as described above for phylogenetic and BactDating analysis. Complete genomes
538 within the same genome cluster were available and used for alignment for four genome clusters.
539 The remaining genome clusters were aligned against a high-quality draft genome within the
540 same genome cluster. MGE regions were masked in the alignment. The r/m, number of
541 recombination events and length of recombination segments was calculated for each genome
542 within a genome cluster by summing along branches from the Gubbins output. The median r/m
543 was calculated for each genome cluster, and the median of these values was given as the species
544 r/m.
545
546 *MGE insertion site mapping across SDSE and S. pyogenes*
547 SDSE MGE insertion regions were defined between two core genes using Corekaburra[71] to call
548 pangenome synteny. Only prophage, ICE and ME insertion sites were mapped. MGE counts
549 presented in Supplementary Table 4a were estimated by summarising fragmented MGEs on
550 the same contig as a core gene, to the corresponding core gene pair insertion region. Phage-
551 like elements were grouped with prophages. Alternative insertion sites were defined as the less
552 commonly found connection between two core genes and represent genome rearrangements
553 around MGEs.

554 Insertion regions found in *S. pyogenes* were collated from McShan *et al.* [37] and Berbel *et al.* [38]
555 in addition to 13 newly reported insertion sites using Corekaburra. Insertion regions were
556 mapped across species by merging the SDSE and *S. pyogenes* pangenome. The *S. pyogenes*
557 pangenome was defined using Panaroo v1.2.10[69] with the same parameters as that used for
558 SDSE. The pangenomes of the two species were then merged using 'panaroo-merge' which
559 overlays pangenome graphs. Merging was performed with an initial clustering threshold of
560 90% identity and 90% length followed by a threshold of 70% for collapsing syntenic genes.
561 Insertion sites in each species were then matched using the merged pangenome. Where no
562 match was obtained, including where genes were core in only one species, core genes within
563 two genes either side of the insertion region were manually inspected for a match.
564
565 *Shared genes and MGEs at insertion sites*
566 MGE genes at shared insertion sites were determined using Corekaburra[71] by listing all
567 accessory genes present between core gene insertion sites which were conserved across SDSE
568 and *S. pyogenes*. Genes were considered shared across SDSE and *S. pyogenes* if they
569 overlapped in the merged pangenome.
570 MGEs were clustered using mge-cluster[39] which maps distances between MGEs using presence
571 and absence of shared unitigs, projects the distances in two dimensions using t-SNE and
572 clusters similar elements using HDBSCAN. mge-cluster has been used with plasmid sequences
573 and we now extend it to use with prophage and ICE. mge-cluster was run with t-SNE perplexity
574 of 75 and HDBSCAN minimum cluster size of 30. MGE sequences were extracted using
575 Magphi v1.0.1[67] with insertion site core genes used as seed sequences, maximum seed distance
576 determined by the largest distance between the respective core pair calculated by Corekaburra,
577 and '--print_breaks' selected which attempts to extract sequences across contig breaks. When
578 cross-species seed hits could not be obtained using nucleotide sequences, input of the '--

579 protein_seed' flag with translated protein sequences were used. To ensure only high quality
580 MGE sequences were obtained, only sequences longer than 10-15kb were included.
581 Fragmented MGEs were concatenated for input into mge-cluster. Some segments were unable
582 to be extracted by Magphi as seed core genes were present on small contigs resulting in
583 difficulty determining whether to extract sequences in the 5' or 3' direction.
584 Highly similar MGEs present in SDSE and *S. pyogenes* were found by sequence-based
585 clustering using CD-HIT v4.8.1[75] 'cd-hit-est' with word size 5, sequence identity threshold 0.8
586 and length difference cut-off 0.8. MGE alignment figures were generated using Easyfig
587 v2.2.3[76].
588
589 *Comparing SDSE and S. pyogenes pangenome*
590 The accessory gene classification scheme described for SDSE and COG annotations were
591 mapped to the *S. pyogenes* pangenome. The merged pangenome was then used to map
592 functional classes of core genes and accessory gene MGE classes across species. Panaroo[69]
593 combines a small number of genes/coding sequences when merging pangenomes resulting in
594 slightly fewer pangenome genes in the merged pangenome compared to that for the individual
595 species.
596 Metabolic differences between SDSE and *S. pyogenes* were inferred by searching for well-
597 defined complete KEGG modules present in one species but not the other. Non-MGE and core
598 genes unique to each species were extracted from the pangenomes of each respective species
599 and KO identifiers were assigned using eggNOG-mapper v2.1.7[70]. KO numbers of unique
600 genes from both species were then mapped simultaneously to KEGG modules using KEGG
601 mapper Reconstruct[77].
602
603 *Recombination detection across species*

604 Gene alignments of the 1,116 shared core genes of SDSE and *S. pyogenes* with <25% gap
605 characters were created and analysed with fastGEAR[35]. Interspecies recombination was
606 inferred by two criteria from the fastGEAR results. Lineages or sequence clusters inferred by
607 fastGEAR could contain sequences from one species or both. A predicted recombination event
608 from a cluster containing only one species to a genome from the other species, an event
609 classified as 'recent' by fastGEAR, was classified as a putatively cross-species. The presence
610 of both species within the same lineage or cluster when more than one lineage was predicted,
611 could occur in the setting of cross-species recombination of the whole gene or with shared
612 ancestry. As these could not be easily separated, cross-species clusters were recorded
613 separately, and locus tags are provided in Supplementary Table 2b individual interrogation.
614
615 *Vaccine antigen screen*
616 *S. pyogenes* vaccine candidates were screened for presence and sequence diversity in SDSE.
617 Vaccine candidates and screening methods were adapted from a previous report of vaccine
618 antigenic diversity in *S. pyogenes* and updated with new multivalent antigens and
619 formulations[22, 42] (Supplementary Table 5). The presence of vaccine antigen genes was
620 determined using screen_assembly v1.2.7[22] with a cut-off of 70% nucleotide identity and 70%
621 coverage. Sequence diversity was presented as nucleotide divergence calculated by BLASTn
622 or predicted amino acid divergence using tBLASTn as indicated.
623 For the 30-valent M protein vaccine, the 180bp hypervariable 5' sequence was extracted from
624 publicly available databases and compared against the N-terminal sequence of SDSE *emm*
625 types represented in the 501 genomes from this report at 70% nucleotide identity and 70%
626 coverage. The representative/type SDSE *emm* sequence (e.g., *stG*62647.0, *stG*840.0) was used
627 for the comparison. For the T antigens, nucleotide sequences of the individual subdomains
628 from different T alleles included in the fusion multivalent vaccine formulations were extracted

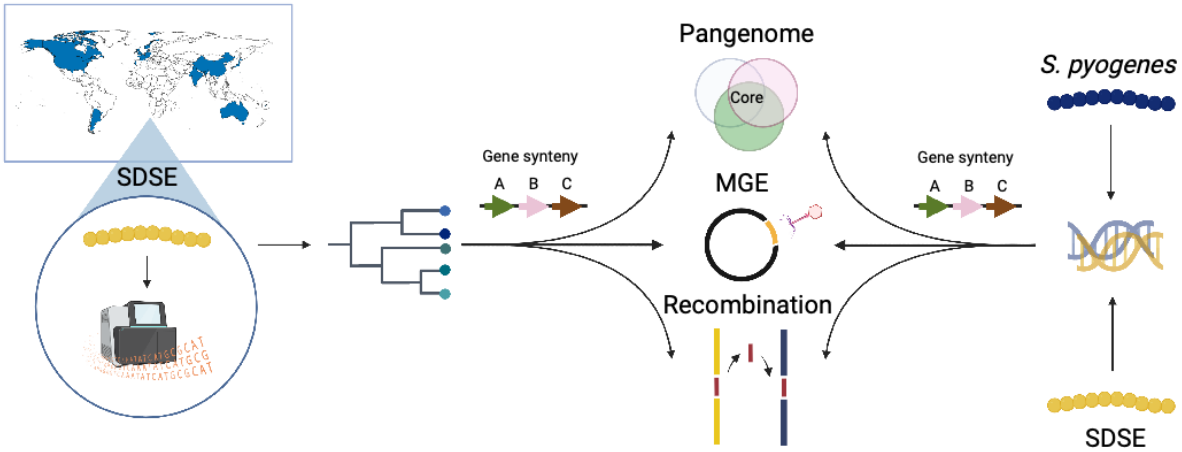
629 and searched at a threshold of 70% nucleotide identity and 70% coverage. Results were
630 presented as presence/absence as M and T antigens are hypervariable.
631 The small peptide antigens J8.0, StrepInCor 'common' overlapping B and T cell epitope, P*17
632 and S2 were screened using a six-frame translation of the target genome and search at 100%
633 identity and coverage. As P*17 had two amino acid substitutions at positions 13 and 17
634 introduced which are not naturally occurring, amino acids at these positions were replaced with
635 a wildcard for the search.
636
637 *Data availability*
638 Accessions for raw sequencing data are available in Supplementary Table 1a. The complete
639 genome sequence of *S. dysgalactiae* subsp. *equisimilis* NS3396 was deposited to GenBank
640 (accession CP128987).
641
642 *Code availability*
643 Supplementary code used to extract and classify accession segments and MGEs is available at
644 https://github.com/OuliXie/Global_SDSE.

645 **Table 1.** Fully assembled mobile genetic elements shared at conserved cross-species insertion regions in the global *Streptococcus dysgalactiae*
646 subsp. *equisimilis* (SDSE) and *S. pyogenes* databases. SDSE insertion sites are mapped to reference genome GGS_124 (NC_012891.1) and *S.*
647 *pyogenes* insertion sites are mapped to reference genome SF370 (NC_002737.2). SDSE insertion site number refers to naming in Figure 3a.

SDSE insertion	S. pyogenes insertion	MGE	SDSE isolates (countries)	*S. pyogenes* isolates (countries)	Max length and identity	Virulence and AMR cargo genes
SDEG_RS03340 – SDEG_RS03345 (insertion region 17)	SPY_RS02725 – SPY_RS02975	Prophage	1 genome USA	2 genomes France	42kb 100% identity, 99% coverage	Exotoxin *speC* Streptodornase *spd1*
SDEG_RS03450 – SDEG_RS03465 (insertion region 19)	SPY_RS03010 – SPY_RS03025	Prophage	1 genome USA	2 genomes UK	42kb 89% identity, 98% coverage	Streptodornase *sda1* allele with 95% nucleotide identity to MGAS5005 M1T1 *sda1* allele.
SDEG_RS08710 – SDEG_RS08720 (insertion region 45)	SPY_RS07175 – SPY_RS07185	Prophage	12 genomes USA, Australia, Denmark, Norway, Canada, and Portugal	104 genomes India, New Zealand, Brazil, Australia, China, Hong Kong, Taiwan, UK, Japan, Trinidad, Canada, Lebanon, USA	41kb 99% identity, 99% coverage	Streptodornase *sda1* on the *S. pyogenes* M1T1 prophage, ϕ5005.3.

SDEG_RS09545 – SDEG_RS09550 (insertion region 49)	SPY_RS01170 – SPY_RS01175	Nested ICE-like and IS/transposon	1 genome India	1 genome Canada	53kb 90% identity, 98% coverage	Macrolide AMR *ermB* Aminoglycoside AMR *ant(6)-Ia* and *aph(3')-IIIa* Additional IS/transposon in SDSE genome with chloramphenicol AMR *catA*.

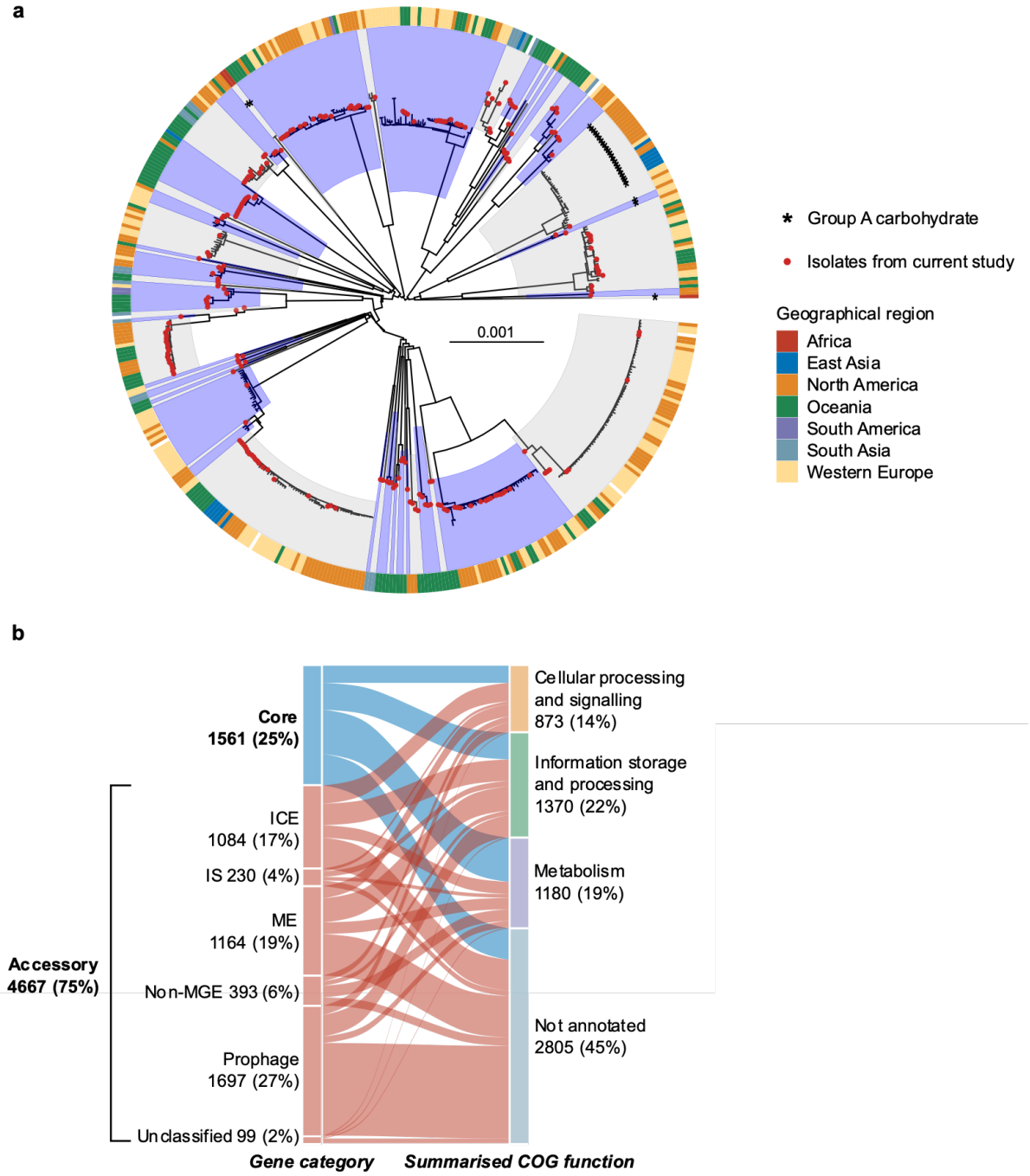
648



649

650 **Figure 1.** Workflow to characterise the global population structure of *Streptococcus*
651 *dysgalactiae* subsp. *equisimilis* (SDSE) and its overlap with *S. pyogenes*. A globally diverse
652 collection of 228 SDSE isolates were whole genome sequenced using Illumina short read
653 sequencing and collated with publicly available genomes to form a database of 501 global
654 sequences. An analysis of the SDSE population structure was undertaken followed by a
655 systematic analysis of the pangenome, evidence of homologous recombination, mobile genetic
656 elements (MGEs) and MGE insertion sites using pangenome gene synteny contextual
657 information. This framework was then compared with a previously published global database
658 of 2,083 *S. pyogenes* genomes by merging pangenomes accounting for shared gene synteny to
659 reveal extensive overlap at the level of shared genes, homologous recombination and MGEs[22].
660 Figure created with BioRender.com.

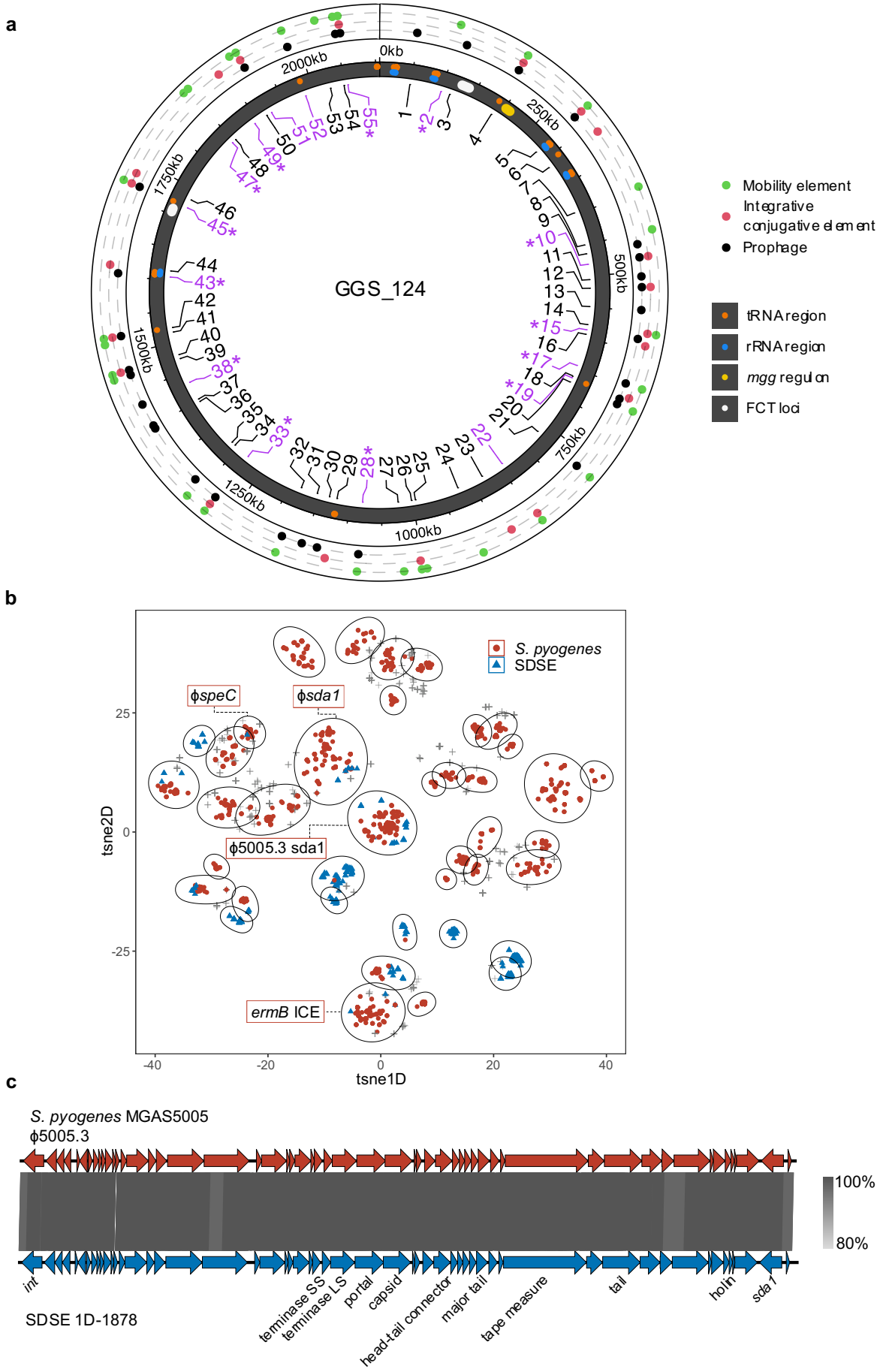
661



662
663

664 **Figure 2.** Global population structure and pangenome composition of *Streptococcus*
665 *dysgalactiae* subsp. *equisimilis* (SDSE). **a)** Minimum evolution phylogenetic tree of 501 SDSE
666 isolates using recombination masked genomic distances. Fifty-nine distinct whole genome
667 clusters are highlighted by alternating blue and grey shades from internal nodes. Newly
668 published sequences from this study are highlighted by red points at the tips of the tree. Isolates

669 carrying the group A carbohydrate are marked by asterisks. The outer ring around the tree
670 denotes geographical region of isolation. **b)** Alluvial plot of the SDSE pangenome with
671 categorisation of core (present in ≥99% of isolates) and accessory genes into summarised COG
672 functional groups. Accessory genes are further characterised based on the type of element they
673 are most frequently found. Genes associated with prophage and prophage-like elements were
674 grouped together. Genes associated with mobile genetic elements (MGE) that could not be
675 classified further were assigned to mobility elements (ME). A small number of unclassified
676 accessory genes were present only adjacent to contig breaks without an integrase and could not
677 confidently be classified into a MGE or non-MGE category. ICE, integrative conjugative
678 element; IS, insertion sequence; non-MGE, non-mobile genetic element.


a
GGS_124
0kb
250kb
500kb
750kb
1000kb
1250kb
1500kb
1750kb
2000kb
Mobility element
Integrative conjugative element
Prophage
tRNA region
rRNA region
mgg regulon
FCT loci
b
S. pyogenes
SDSE
ϕspeC
ϕsda1
ϕ5005.3 sda1
ermB ICE
tsne1D
tsne2D
c
S. pyogenes MGAS5005
ϕ5005.3
100%
80%
int
terminase SS
terminase LS
portal
capsid
head-tail connector
major tail
tape measure
tail
holin
sda1
SDSE 1D-1878

680 **Figure 3.** Clustering and genome localisation of *Streptococcus dysgalactiae* subp. *equisimilis*
681 (SDSE) and *S. pyogenes* mobile genetic elements (MGEs). **a)** Location of SDSE MGE insertion
682 regions relative to the GGS_124 reference genome (NC_012891.1). Insertion regions are
683 labelled from 1 to 55 of which 16 (29%) were shared with *S. pyogenes* (highlighted in purple).
684 Shared insertion regions at which MGE clusters were present and shared across species are
685 highlighted with asterisks. The outer ring indicates the type of element detected at each
686 insertion region in SDSE. tRNA and rRNA regions, the *mgg* regulon containing the *emm* gene,
687 and the two FCT loci in GGS_124 are marked on the genome representation. **b)** Clustering of
688 3,630 MGEs at 16 shared insertion regions across SDSE and *S. pyogenes*. Each cluster is
689 outlined by an ellipsoid. Of the 40 MGE clusters, 10 (25%) were shared across species. Clusters
690 which contained notable examples of near-identical (>80% nucleotide identity and coverage)
691 cross-species MGEs are labelled: the global M1T1 ϕ5005.3 prophage carrying the Sda1
692 streptodornase (104 *S. pyogenes*, 12 SDSE), ϕ*sda1* refers to a prophage carrying a *sda1* allele
693 95% similar to ϕ5005.3 (2 *S. pyogenes*, 1 SDSE), ϕ*speC* refers to a previously described
694 prophage carrying *speC* and *spd1*[18] (2 *S. pyogenes*, 1 SDSE), *ermB* ICE refers to a complex
695 nested ICE and IS/transposon element carrying multiple AMR genes including *ermB* (1 *S.*
696 *pyogenes*, 1 SDSE). These MGEs represent a subset of elements within each respective cluster.
697 **c)** Genome architecture and comparison of the *S. pyogenes* M1T1 prophage ϕ5005.3, with a
698 near identical prophage found in group G SDSE isolate 1D-1878 at insertion site 45. 1D-1878
699 was isolated from a case of invasive disease in Denmark in 2018[53]. Regions of genomic
700 similarity were inferred using BLAST and plotted using Easyfig v2.2.3[76]. The grey gradient
701 indicates the percent identity in the legend. The same prophage was detected in an additional
702 19 SDSE isolates spanning 10 distinct genome clusters and 7 countries indicating significant
703 dispersion of the prophage in the SDSE population.

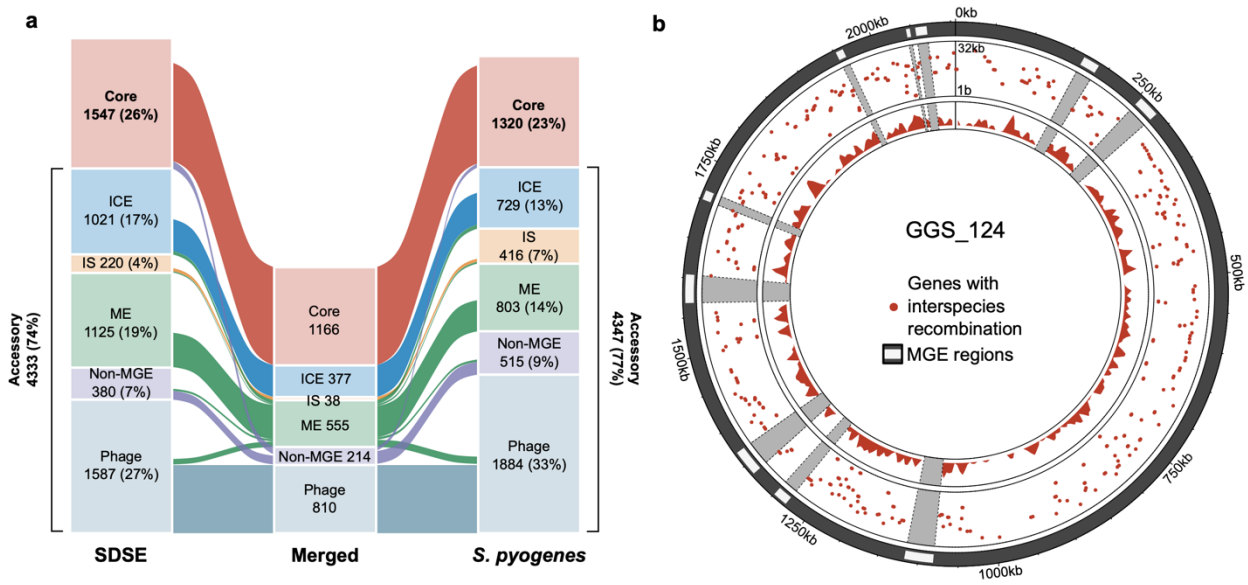
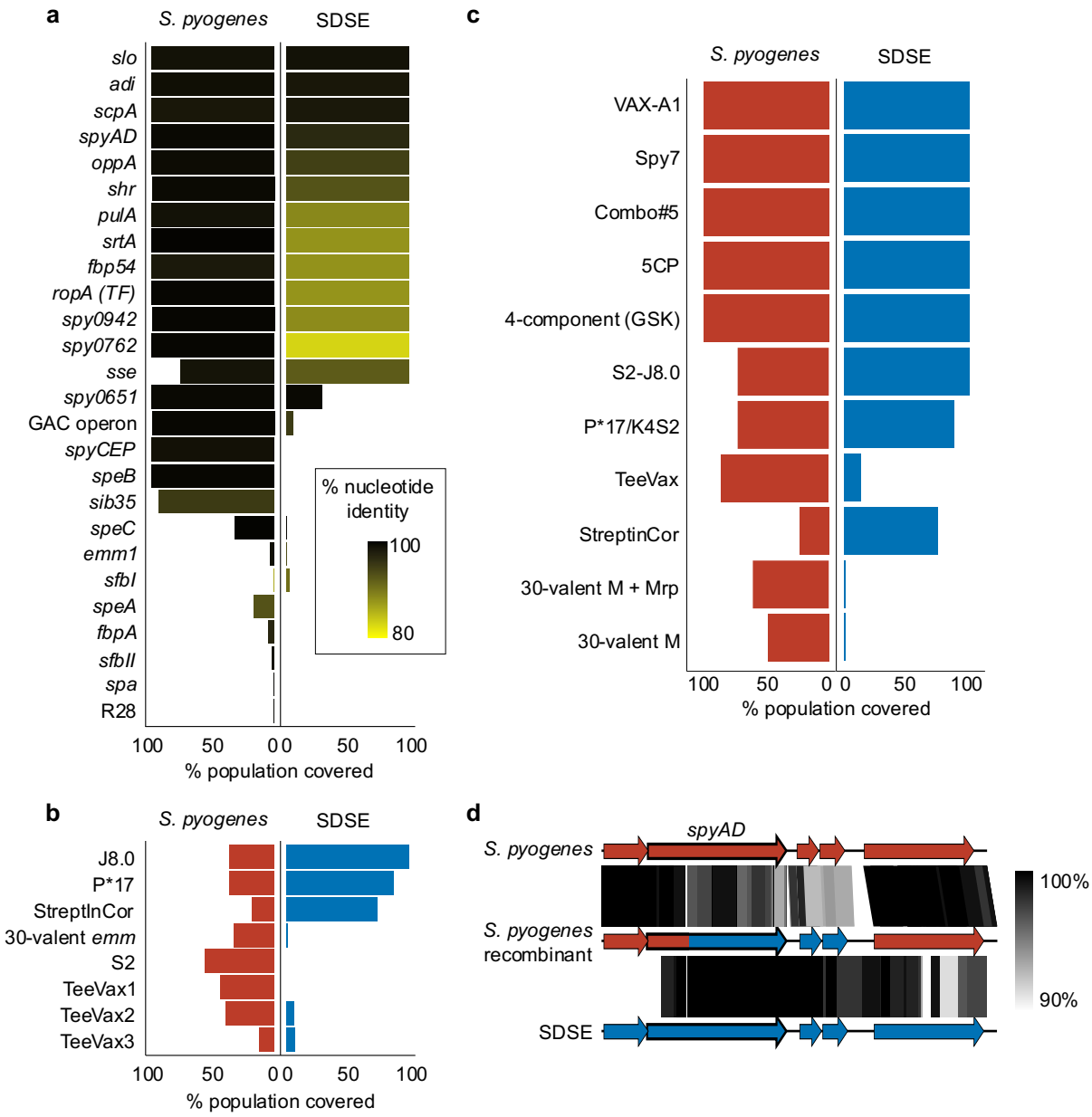


Figure 4. Comparison and recombination signatures between the *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) and *S. pyogenes* pangenome. **a)** Alluvial plot of the shared SDSE (n=501) and *S. pyogenes* (n=2,083) pangenome. Categorisation of core genes (present in ≥99% of isolates in both species) and accessory genes by association with mobile genetic element (MGE). ICE, integrative conjugative element; IS, insertion sequence; ME, mobility element. Genes classified as belonging to different types of MGEs or an MGE/non-MGE combination across the species were binned as ME in the merged pangenome. Unclassified genes were excluded and a small number of genes were merged when overlaying pangenomes resulting in a slightly smaller merged pangenome than the pangenomes of individual species. **b)** Circular rainfall plot of core genes with signatures of recombination flagged by fastGEAR[35] plotted relative to the GGS_124 reference genome (NC_012891.1). Genes flagged with putative interspecies recombination events are highlighted by red points. The distance between each gene with its neighbour within the same category is plotted on a $\log_{10}$ scale between 1 bp to 32 kbp. The innermost track plots the density of genes with evidence of interspecies recombination using a window size of 10,000 bp. MGE regions are masked in grey. The rainfall plot demonstrates that genes flagged as affected by inter-species recombinants are dispersed across the SDSE genome.



722

723 **Figure 5.** Theoretical coverage of preclinical vaccine candidates and multi-component
724 formulations in global *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE, n = 501) and *S.*
725 *pyogenes* (n = 2,083) populations. **a)** Whole gene candidates were screened in both species at
726 70% nucleotide length and identity. Theoretical coverage of the population (% presence) is
727 expressed by length of bar and conservation relative to the *S. pyogenes* query sequence is
728 reflected by gradient fill from 80% similar (yellow) to 100% (black). **b)** Peptides and gene sub-
729 domain candidates were screened without calculation of sequence diversity. Peptide fragments

730 were screened using a 100% match approach and a six-frame translation of the sequence (J8.0,
731 S2, P*17, StreptInCor 'common epitope') and gene sub-domains (30-valent *emm*, T antigen
732 fragments in TeeVax1, TeeVax2, TeeVax3) were screened at 70% nucleotide identity and
733 coverage. **c)** Theoretical coverage of SDSE and *S. pyogenes* isolates in this study by multivalent
734 vaccine candidates. **d)** Exemplar of putative recombination between *S. pyogenes* and SDSE
735 vaccine candidate *spyAD*. *S. pyogenes* isolate NS1140 demonstrated evidence of cross-species
736 recombination around the vaccine candidate *spyAD* gene. Sequence similarity between *S.*
737 *pyogenes* reference genome SF370 NC_002737.2 (top), recombinant *S. pyogenes* isolate
738 NS1140 (middle), and SDSE reference 3836-05 (bottom) demonstrated greater similarity
739 between NS1140 and 3836-05 at position 419 of *spyAD* to 2 ORFs downstream. Regions of
740 genomic similarity were inferred using BLAST and plotted using Easyfig v2.2.3[76].

741 **Author Contributions**

742 OX, DJM and MRD planned the study. RJT, LS, KSS, TH, PG, ACS, MRB, BWB, MDP, MR,
743 DEB, BJC and DJM provided samples and metadata. OX, JMM, AJH, MGJ, JA Lees, NLBZ,
744 OB, SLB, GPC, GTH, LM, JA Lacey, TBJ, SAB, GD, SDB, MJW, SYCT, DJM and MRD
745 designed experimental procedures and generated data. OX, JMM, AJH, MGJ, JA Lees, NLBZ,
746 SLB, GPC, SYCT, DJM and MRD analysed data. OX, JMM, AJH, SLB, GPC, DJM and MRD
747 wrote the manuscript. All authors revised and approved the manuscript.

748

749 **Acknowledgements**

750 The work was supported by the National Health and Medical Research Council of Australia
751 (NHMRC) and The Wellcome Trust, UK. MRD was supported by a NHMRC postdoctoral
752 training fellowship (635250) and a University of Melbourne CR Roper Fellowship. OX was
753 supported by the NHMRC postgraduate scholarship (GNT2013831) and Avant Foundation
754 Doctors in Training Research Scholarship (2021/000017). We acknowledge the assistance of
755 the sequencing and pathogen informatics core teams at the Wellcome Sanger Institute, UK
756 where this work was supported by the Wellcome Trust core grants 206194 and 108413/A/15/D.

757 **References**

758 1. Wright CM, Moorin R, Pearson G et al. Invasive Infections Caused by Lancefield
759 Groups C/G and A *Streptococcus*, Western Australia, Australia, 2000-2018. *Emerg Infect Dis*
760 2022; **28**: 2190-7.

761 2. Couture-Cossette A, Carignan A, Mercier A et al. Secular trends in incidence of
762 invasive beta-hemolytic streptococci and efficacy of adjunctive therapy in Quebec, Canada,
763 1996-2016. *PLoS One* 2018; **13**: e0206289.

764 3. Sylvetsky N, Raveh D, Schlesinger Y et al. Bacteremia due to beta-hemolytic
765 *Streptococcus* group G: increasing incidence and clinical characteristics of patients. *Am J Med*
766 2002; **112**: 622-6.

767 4. Oppegaard O, Mylvaganam H, Kittang BR. Beta-haemolytic group A, C and G
768 streptococcal infections in Western Norway: a 15-year retrospective survey. *Clin Microbiol*
769 *Infect* 2015; **21**: 171-8.

770 5. Lambertsen LM, Ingels H, Schønheyder HC et al. Nationwide laboratory-based
771 surveillance of invasive beta-haemolytic streptococci in Denmark from 2005 to 2011. *Clin*
772 *Microbiol Infect* 2014; **20**: O216-23.

773 6. Harris P, Siew DA, Proud M et al. Bacteraemia caused by beta-haemolytic streptococci
774 in North Queensland: changing trends over a 14-year period. *Clin Microbiol Infect* 2011; **17**:
775 1216-22.

776 7. Shinohara K, Murase K, Tsuchido Y et al. Clonal Expansion of Multidrug-Resistant
777 *Streptococcus dysgalactiae* subspecies *equisimilis* Causing Bacteremia, Japan, 2005-2021.
778 *Emerg Infect Dis* 2023; **29**: 528-39.

779 8. Broyles LN, Van Beneden C, Beall B et al. Population-based study of invasive disease
780 due to beta-hemolytic streptococci of groups other than A and B. *Clin Infect Dis* 2009; **48**: 706-
781 12.

782 9. Brandt CM, Spellerberg B. Human infections due to *Streptococcus dysgalactiae*
783 subspecies *equisimilis*. *Clin Infect Dis* 2009; **49**: 766-72.
784 10. Takahashi T, Sunaoshi K, Sunakawa K et al. Clinical aspects of invasive infections
785 with *Streptococcus dysgalactiae* ssp. *equisimilis* in Japan: differences with respect to
786 *Streptococcus pyogenes* and *Streptococcus agalactiae* infections. *Clin Microbiol Infect* 2010;
787 **16**: 1097-103.
788 11. McMillan DJ, Bessen DE, Pinho M et al. Population genetics of *Streptococcus*
789 *dysgalactiae* subspecies *equisimilis* reveals widely dispersed clones and extensive
790 recombination. *PLoS One* 2010; **5**: e11741.
791 12. McNeilly CL, McMillan DJ. Horizontal gene transfer and recombination in
792 *Streptococcus dysgalactiae* subsp. *equisimilis*. *Front Microbiol* 2014; **5**: 676.
793 13. Vähäkuopus S, Vuento R, Siljander T et al. Distribution of *emm* types in invasive and
794 non-invasive group A and G streptococci. *Eur J Clin Microbiol Infect Dis* 2012; **31**: 1251-6.
795 14. Pinho MD, Melo-Cristino J, Ramirez M. Fluoroquinolone resistance in *Streptococcus*
796 *dysgalactiae* subsp. *equisimilis* and evidence for a shared global gene pool with *Streptococcus*
797 *pyogenes*. *Antimicrob Agents Chemother* 2010; **54**: 1769-77.
798 15. van Sorge NM, Cole JN, Kuipers K et al. The classical lancefield antigen of group A
799 *Streptococcus* is a virulence determinant with implications for vaccine design. *Cell Host*
800 *Microbe* 2014; **15**: 729-40.
801 16. Beres SB, Zhu L, Pruitt L et al. Integrative Reverse Genetic Analysis Identifies
802 Polymorphisms Contributing to Decreased Antimicrobial Agent Susceptibility in
803 *Streptococcus pyogenes*. *mBio* 2022; **13**: e0361821.
804 17. Ahmad Y, Gertz RE, Jr., Li Z et al. Genetic relationships deduced from emm and
805 multilocus sequence typing of invasive *Streptococcus dysgalactiae* subsp. *equisimilis* and *S.*

806 *canis* recovered from isolates collected in the United States. *J Clin Microbiol* 2009; **47**: 2046-
807 54.
808 18. Chochua S, Rivers J, Mathis S et al. Emergent Invasive Group A *Streptococcus*
809 *dysgalactiae* subsp. *equisimilis*, United States, 2015-2018. *Emerg Infect Dis* 2019; **25**: 1543-7.
810 19. Chochua S, Metcalf BJ, Li Z et al. Population and Whole Genome Sequence Based
811 Characterization of Invasive Group A Streptococci Recovered in the United States during 2015.
812 *mBio* 2017; **8**: e01422-17.
813 20. Davies MR, Shera J, Van Domselaar GH et al. A novel integrative conjugative element
814 mediates genetic transfer from group G *Streptococcus* to other beta-hemolytic Streptococci. *J*
815 *Bacteriol* 2009; **191**: 2257-65.
816 21. Palmieri C, Magi G, Creti R et al. Interspecies mobilization of an *ermT*-carrying
817 plasmid of *Streptococcus dysgalactiae* subsp. *equisimilis* by a coresident ICE of the ICESa2603
818 family. *J Antimicrob Chemother* 2013; **68**: 23-6.
819 22. Davies MR, McIntyre L, Mutreja A et al. Atlas of group A streptococcal vaccine
820 candidates compiled using large-scale comparative genomics. *Nature Genetics* 2019; **51**: 1035-
821 43.
822 23. Lother SA, Demczuk W, Martin I et al. Clonal Clusters and Virulence Factors of Group
823 C and G *Streptococcus* Causing Severe Infections, Manitoba, Canada, 2012-2014. *Emerg*
824 *Infect Dis* 2017; **23**: 1079-88.
825 24. Oppegaard O, Mylvaganam H, Skrede S et al. Emergence of a *Streptococcus*
826 *dysgalactiae* subspecies *equisimilis* stG62647-lineage associated with severe clinical
827 manifestations. *Sci Rep* 2017; **7**: 7589.
828 25. Kaci A, Jonassen CM, Skrede S et al. Genomic epidemiology of *Streptococcus*
829 *dysgalactiae* subsp. *equisimilis* strains causing invasive disease in Norway during 2018.
830 *Frontiers in Microbiology* 2023; **14**.

831 26. Dale JB, Walker MJ. Update on group A streptococcal vaccine development. *Curr Opin*
832 *Infect Dis* 2020; **33**: 244-50.
833 27. Vekemans J, Gouvea-Reis F, Kim JH et al. The Path to Group A *Streptococcus*
834 Vaccines: World Health Organization Research and Development Technology Roadmap and
835 Preferred Product Characteristics. *Clin Infect Dis* 2019; **69**: 877-83.
836 28. Davies MR, McMillan DJ, Van Domselaar GH et al. Phage 3396 from a *Streptococcus*
837 *dysgalactiae* subsp. *equisimilis* pathovar may have its origins in *Streptococcus pyogenes*. *J*
838 *Bacteriol* 2007; **189**: 2646-52.
839 29. Lees JA, Harris SR, Tonkin-Hill G et al. Fast and flexible bacterial genomic
840 epidemiology with PopPUNK. *Genome Res* 2019; **29**: 304-16.
841 30. Didelot X, Croucher NJ, Bentley SD et al. Bayesian inference of ancestral dates on
842 bacterial phylogenetic trees. *Nucleic Acids Res* 2018; **46**: e134.
843 31. Croucher NJ, Page AJ, Connor TR et al. Rapid phylogenetic analysis of large samples
844 of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015; **43**:
845 e15.
846 32. Ikebe T, Okuno R, Sasaki M et al. Molecular characterization and antibiotic resistance
847 of *Streptococcus dysgalactiae* subspecies *equisimilis* isolated from patients with streptococcal
848 toxic shock syndrome. *J Infect Chemother* 2018; **24**: 117-22.
849 33. Beres SB, Olsen RJ, Long SW et al. Analysis of the Genomics and Mouse Virulence
850 of an Emergent Clone of *Streptococcus dysgalactiae* subspecies *equisimilis*. *Microbiol Spectr*
851 2023: e0455022.
852 34. Oppegaard O, Mylvaganam H, Skrede S et al. Exploring the arthritogenicity of
853 *Streptococcus dysgalactiae* subspecies *equisimilis*. *BMC Microbiology* 2018; **18**: 17.
854 35. Mostowy R, Croucher NJ, Andam CP et al. Efficient Inference of Recent and Ancestral
855 Recombination within Bacterial Populations. *Mol Biol Evol* 2017; **34**: 1167-82.

856 36. Khedkar S, Smyshlyaev G, Letunic I et al. Landscape of mobile genetic elements and
857 their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Research* 2022; **50**:
858 3155-68.
859 37. McShan WM, McCullor KA, Nguyen SV. The Bacteriophages of *Streptococcus*
860 *pyogenes*. *Microbiol Spectr* 2019; **7**.
861 38. Berbel D, Càmara J, González-Díaz A et al. Deciphering mobile genetic elements
862 disseminating macrolide resistance in *Streptococcus pyogenes* over a 21 year period in
863 Barcelona, Spain. *J Antimicrob Chemother* 2021; **76**: 1991-2003.
864 39. Arredondo-Alonso S, Gladstone RA, Pöntinen AK et al. Mge-cluster: a reference-free
865 approach for typing bacterial plasmids. *NAR Genom Bioinform* 2023; **5**: lqad066.
866 40. Sumby P, Barbian KD, Gardner DJ et al. Extracellular deoxyribonuclease made by
867 group A *Streptococcus* assists pathogenesis by enhancing evasion of the innate immune
868 response. *Proc Natl Acad Sci U S A* 2005; **102**: 1679-84.
869 41. Sumby P, Porcella SF, Madrigal AG et al. Evolutionary origin and emergence of a
870 highly successful clone of serotype M1 group A *Streptococcus* involved multiple horizontal
871 gene transfer events. *J Infect Dis* 2005; **192**: 771-82.
872 42. Harbison-Price N, Rivera-Hernandez T, Osowicki J et al. Current Approaches to
873 Vaccine Development of *Streptococcus pyogenes*. In: Ferretti JJ, Stevens DL, Fischetti VA,
874 eds. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations*. Oklahoma City (OK):
875 University of Oklahoma Health Sciences Center
876 © The University of Oklahoma Health Sciences Center., 2022.
877 43. Porcellato D, Smistad M, Skeie SB et al. Whole genome sequencing reveals possible
878 host species adaptation of *Streptococcus dysgalactiae*. *Sci Rep* 2021; **11**: 17350.
879 44. Diop A, Torrance EL, Stott CM et al. Gene flow and introgression are pervasive forces
880 shaping the evolution of bacterial species. *Genome Biol* 2022; **23**: 239.

881 45. Davies MR, Keller N, Brouwer S et al. Detection of *Streptococcus pyogenes* M1(UK)
882 in Australia and characterization of the mutation driving enhanced expression of superantigen
883 SpeA. *Nat Commun* 2023; **14**: 1051.
884 46. Nordström T, Malcolm J, Magor G et al. In vivo efficacy of a chimeric peptide derived
885 from the conserved region of the M protein against group C and G streptococci. *Clin Vaccine*
886 *Immunol* 2012; **19**: 1984-7.
887 47. Ishihara H, Ogura K, Miyoshi-Akiyama T et al. Prevalence and genomic
888 characterization of Group A *Streptococcus dysgalactiae* subsp. *equisimilis* isolated from
889 patients with invasive infections in Toyama prefecture, Japan. *Microbiol Immunol* 2020; **64**:
890 113-22.
891 48. Shimomura Y, Okumura K, Murayama SY et al. Complete genome sequencing and
892 analysis of a Lancefield group G *Streptococcus dysgalactiae* subsp. *equisimilis* strain causing
893 streptococcal toxic shock syndrome (STSS). *BMC Genomics* 2011; **12**: 17.
894 49. Okumura K, Shimomura Y, Murayama SY et al. Evolutionary paths of streptococcal
895 and staphylococcal superantigens. *BMC Genomics* 2012; **13**: 404.
896 50. Brandt CM, Haase G, Schnitzler N et al. Characterization of blood culture isolates of
897 *Streptococcus dysgalactiae* subsp. *equisimilis* possessing Lancefield's group A antigen. *J Clin*
898 *Microbiol* 1999; **37**: 4194-7.
899 51. Babbar A, Nitsche-Schmitz DP, Pieper DH et al. Draft Genome Sequence of
900 *Streptococcus dysgalactiae* subsp. *equisimilis* Strain C161L1 Isolated in Vellore, India.
901 *Genome Announc* 2017; **5**.
902 52. Wang X, Zhang X, Zong Z. Genome sequence and virulence factors of a group G
903 *Streptococcus dysgalactiae* subsp. *equisimilis* strain with a new element carrying *erm(B)*. *Sci*
904 *Rep* 2016; **6**: 20389.

905 53. Rebelo AR, Bortolaia V, Leekitcharoenphon P et al. One Day in Denmark: Comparison
906 of Phenotypic and Genotypic Antimicrobial Susceptibility Testing in Bacterial Isolates From
907 Clinical Settings. *Front Microbiol* 2022; **13**: 804627.
908 54. Cuénod A, Foucault F, Pflüger V et al. Factors Associated With MALDI-TOF Mass
909 Spectral Quality of Species Identification in Clinical Routine Diagnostics. *Front Cell Infect*
910 *Microbiol* 2021; **11**: 646648.
911 55. Jagne I, Keeley AJ, Bojang A et al. Impact of intra-partum azithromycin on carriage of
912 group A *Streptococcus* in the Gambia: a posthoc analysis of a double-blind randomized
913 placebo-controlled trial. *BMC Infect Dis* 2022; **22**: 103.
914 56. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome*
915 *Biol* 2019; **20**: 257.
916 57. Prjibelski A, Antipov D, Meleshko D et al. Using SPAdes De Novo Assembler. *Curr*
917 *Protoc Bioinformatics* 2020; **70**: e102.
918 58. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; **30**:
919 2068-9.
920 59. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb
921 software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018; **3**: 124.
922 60. Baines SL, Howden BP, Heffernan H et al. Rapid Emergence and Evolution of
923 Staphylococcus aureus Clones Harboring fusC-Containing Staphylococcal Cassette
924 Chromosome Elements. *Antimicrob Agents Chemother* 2016; **60**: 2359-65.
925 61. Page AJ, Ainsworth EV, Langridge GC. socru: typing of genome-level order and
926 orientation around ribosomal operons in bacteria. *Microb Genom* 2020; **6**.
927 62. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and Fast
928 Distance-Based Phylogeny Inference Program. *Mol Biol Evol* 2015; **32**: 2798-800.

929 63. Minh BQ, Schmidt HA, Chernomor O et al. IQ-TREE 2: New Models and Efficient
930 Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530-4.
931 64. Smith MR. Information theoretic generalized Robinson-Foulds metrics for comparing
932 phylogenetic trees. *Bioinformatics* 2020; **36**: 5007-13.
933 65. Nascimento M, Sousa A, Ramirez M et al. PHYLOViZ 2.0: providing scalable data
934 integration and visualization for multiple phylogenetic inference methods. *Bioinformatics*
935 2017; **33**: 128-9.
936 66. Liu B, Zheng D, Zhou S et al. VFDB 2022: a general classification scheme for bacterial
937 virulence factors. *Nucleic Acids Res* 2022; **50**: D912-d7.
938 67. Jespersen MG, Hayes A, Davies MR. Magphi: Sequence extraction tool from FASTA
939 and GFF3 files using seed pairs. *Journal of Open Source Software* 2022; **7**: 4369.
940 68. Feldgarden M, Brover V, Gonzalez-Escalona N et al. AMRFinderPlus and the
941 Reference Gene Catalog facilitate examination of the genomic links among antimicrobial
942 resistance, stress response, and virulence. *Sci Rep* 2021; **11**: 12728.
943 69. Tonkin-Hill G, MacAlasdair N, Ruis C et al. Producing polished prokaryotic
944 pangenomes with the Panaroo pipeline. *Genome Biol* 2020; **21**: 180.
945 70. Cantalapiedra CP, Hernández-Plaza A, Letunic I et al. eggNOG-mapper v2: Functional
946 Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol*
947 *Biol Evol* 2021; **38**: 5825-9.
948 71. Jespersen MG, Hayes A, Davies MR. Corekaburra: pan-genome post-processing using
949 core gene synteny. *Journal of Open Source Software* 2022; **7**: 4910.
950 72. Mistry J, Finn RD, Eddy SR et al. Challenges in homology search: HMMER3 and
951 convergent evolution of coiled-coil regions. *Nucleic Acids Res* 2013; **41**: e121.
952 73. Abby SS, Cury J, Guglielmini J et al. Identification of protein secretion systems in
953 bacterial genomes. *Sci Rep* 2016; **6**: 23080.

954 74. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
955 improvements in performance and usability. *Mol Biol Evol* 2013; **30**: 772-80.
956 75. Fu L, Niu B, Zhu Z et al. CD-HIT: accelerated for clustering the next-generation
957 sequencing data. *Bioinformatics* 2012; **28**: 3150-2.
958 76. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer.
959 *Bioinformatics* 2011; **27**: 1009-10.
960 77. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden
961 features in biological data. *Protein Sci* 2022; **31**: 47-53.
962