

1 **Genome-wide association reveals host-specific genomic traits in *Escherichia coli***
2

3 Sumeet K. Tiwari^{1*}, Boas C.L. van der Putten^{2,3*}, Thilo M. Fuchs⁴, Trung N. Vinh⁵, Martin
4 Bootsma⁶, Rik Oldenkamp², Roberto La Ragione⁷, Sébastien Matamoros³, Ngo T. Hoa^{5,8,9},
5 Christian Berens⁴, Joy Leng⁷, Julio Álvarez^{10,11}, Marta Ferrandis-Vila⁴, Jenny M. Ritchie¹²,
6 Angelika Fruth¹³, Stefan Schwarz¹⁴, Lucas Domínguez^{10,11}, María Ugarte-Ruiz¹⁰, Astrid
7 Bethe¹⁴, Charlotte Huber¹⁵, Vanessa Johanns¹⁵, Ivonne Stamm¹⁶, Lothar H. Wieler¹⁷, Christa
8 Ewers¹⁸, Amanda Fivian-Hughes¹², Herbert Schmidt¹⁹, Christian Menge⁴, Torsten Semmler^{1#},
9 Constance Schultsz^{2,3#}

10 ¹Robert Koch Institute, Genome sequencing and genomic epidemiology, Berlin, Germany

11 ²Amsterdam UMC, Department of Global Health, Amsterdam, Netherlands

12 ³Amsterdam UMC, Department of Medical Microbiology, Amsterdam, Netherlands

13 ⁴Friedrich-Loeffler-Institut, Institute of Molecular Pathogenesis, Jena, Germany

14 ⁵Oxford University Clinical Research Unit, Vietnam

15 ⁶UMC Utrecht, Utrecht, Netherlands

16 ⁷Department of Pathology and Infectious Diseases, School of Veterinary Medicine,
17 University of Surrey, Guildford, UK

18 ⁸Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK

19 ⁹Microbiology Department and the Micro-Parasitology Unit of the Center for Bio-Medical
20 Research, Pham Ngoc Thach University of Medicine, Ho Chi Minh City, Vietnam

21 ¹⁰VISAVET Health Surveillance Centre, Complutense University of Madrid, Madrid, Spain

22 ¹¹Department of Animal Health, Faculty of Veterinary Medicine, Complutense University of
23 Madrid, Madrid, Spain

24 ¹²School of Biosciences and Medicine, University of Surrey, Guildford, UK

25 ¹³Robert Koch Institute, Enteropathogenic Bacteria and Legionella, Wernigerode, Germany

26 ¹⁴Institute of Microbiology and Epizootics, Freie Universität Berlin, Berlin, Institute of
27 Microbiology and Epizootics, Berlin, Germany

28 ¹⁵Robert Koch Institute, Advanced Light and Electron Microscopy, Berlin, Germany

29 ¹⁶Vet Med Labor GmbH, Division of IDEXX Laboratories, Ludwigsburg, Germany

30 ¹⁷Robert Koch Institute, Berlin, Germany

31 ¹⁸Institute of Hygiene and Infectious Diseases of Animals, Giessen, Germany

32 ¹⁹Institute of Food Science and Biotechnology, Department of Food Microbiology and
33 Hygiene, University of Hohenheim, Stuttgart, Germany

34 *These authors contributed equally to this work

35 #Corresponding authors

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52 **Abstract:**

53

54 *Escherichia coli* is an opportunistic pathogen that can colonize or infect various host species.

55 There is a significant gap in our understanding to what extent genetic lineages of *E. coli* are

56 adapted or restricted to specific hosts. In addition, genomic determinants underlying such host

57 specificity are unknown. By analyzing a randomly sampled collection of 1198 whole-genome

58 sequenced *E. coli* isolates from four countries (Germany, UK, Spain, and Vietnam), obtained

59 from five host species (human, pig, cattle, chicken, and wild boar) over 16 years, from both

60 healthy and diseased hosts, we demonstrate that certain lineages of *E. coli* are frequently

61 detected in specific hosts. We report a novel *nan* gene cluster, designated *nan-9*, putatively

62 encoding acetylesterases and determinants of uptake and metabolism of sialic acid, to be

63 associated with the human host as identified through genome wide association studies. *In silico*

64 characterization predicts *nan-9* to be involved in sialic acid (Sia) metabolism. *In vitro* growth

65 experiments with a representative Δ *nan* *E. coli* mutant strain, using sialic acids 5-*N*-acetyl

66 neuraminic acid (Neu5Ac) and *N*-glycolyl neuraminic acid (Neu5Gc) as the sole carbon source,

67 indicate an impaired growth behaviour compared to the wild-type. In addition, we identified

68 several additional *E. coli* genes that are potentially associated with adaptation to human, cattle

69 and chicken hosts, but not for the pig host. Collectively, this study provides an extensive

70 overview of genetic determinants which may mediate host specificity in *E. coli*. Our findings

71 should inform risk analysis and epidemiological monitoring of (antimicrobial resistant) *E. coli*.

72

73

74

75

76

77 **Introduction:**

78 *Escherichia coli* is a Gram-negative bacterium which has been isolated from various host
79 species, including humans, cattle, chickens and pigs(1). Because *E. coli* can colonize or infect
80 multiple host species, this bacterium can act as a reservoir for genes encoding antimicrobial
81 resistance (AMR)(2) that can be transmitted between different host species. The likelihood that
82 *E. coli* and its AMR encoding genes persist in a new host after transmission depends on
83 multiple factors(3,4). For example, small changes in metabolic pathways may enable *E. coli* to
84 colonize or infect a host more efficiently(1). Several studies have suggested that highly
85 successful *E. coli* clones, such as the sequence type 131 (ST131) clone(5,6) or clonal complex
86 87 (ST58 and ST155) *E. coli* facilitate the spread of AMR *E. coli* in the human population(7)
87 whilst other studies have shown that different lineages of AMR *E. coli* vary in their ability to
88 spread(8). These findings both indicate that AMR genes, at least to some extent, hitchhike on
89 bacterial strains that are specifically equipped to colonize a given host. Beyond classical
90 virulence or adhesion factors, genetic and functional traits defining different degrees of host
91 adaptation(3,9) and thereby indirectly impacting on the spread of AMR between host species,
92 have not been identified thus far.

93

94 Comparative genomic analysis of bacterial populations from multiple hosts has revealed
95 signatures of host-adaptation in bacterial genomes(10). The emergence of large-scale bacterial
96 genome-wide association studies (GWAS) allowed for the detection of genes or genomic
97 variants that are associated with resistance, pathogenicity, and host adaptive traits(11–13).
98 Here, we have applied population-based bacterial GWAS to identify host-associated genomic
99 determinants in a diverse panel of 1,198 *E. coli* isolates, irrespective of their AMR pattern.
100 Isolates were recovered from five different host species, including healthy and diseased
101 individuals from four different countries in two continents over 16 years. The *pan*-genome was

102 analyzed for specific host association followed by a *k-mer* based bacterial GWAS approach to
103 identify host-specific genomic determinants and their potential role in host-adaptation.

104

105 **Material and Methods**

106 **a) Sampling strategy**

107 A panel of 1213 *E. coli* isolates from four countries (Germany, UK, Spain, and Vietnam),
108 obtained from five host species (human, pig, cattle, chicken, and wild boar) during three time
109 periods (2003-2007, 2008-2012 and 2013-2018) from both healthy and diseased hosts were
110 selected randomly from existing strain collections and newly collected isolates. Out of 120
111 possible strata (defined as a unique combination of country, host, time-period, and host health
112 status), 42 strata contained isolates. We included all isolates available per stratum if there were
113 less than 30 isolates and performed a random selection of up to a maximum of 30 isolates if
114 more were available. Potentially duplicate isolates that were part of an outbreak, isolated at a
115 single location within a short timeframe, or from a single farm or a single individual were
116 excluded. Only one isolate per individual was included in the analyses. Isolates included per
117 stratum are shown in Table S1.

118 **b) DNA extraction and sequencing**

119 The DNA of the *E. coli* isolates from Germany was extracted using the QIAamp DNA Mini
120 Kit (Qiagen) following the manufacturer's instructions. The DNA concentration was evaluated
121 fluorometrically by using Qubit™ 2.0 fluorometer (Invitrogen, USA) and the associated
122 Qubit™ dsDNA HS Assay Kit (0.2-100ng) and Qubit™ BR Assay Kit (2-1000ng),
123 respectively. The libraries were generated using Nextera DNA library preparation (Illumina,
124 <https://www.illumina.com>). The sequencing was performed using the Illumina MiSeq and
125 HiSeq systems, generating 2 × 250 bp and 2 × 150 bp reads, respectively.

126 The DNA of the *E. coli* isolates from the UK was purified using a Promega DNA Wizard®
127 genomic purification kit and quantified using Nanodrop. Libraries were generated using
128 Nextera XT technology (Illumina), and DNA sequencing of isolates was performed at the
129 Animal and Plant Health Agency (APHA, Surrey, UK, <https://www.gov.uk/government/organisations/animal-and-plant-healthagency>) using an Illumina MiSeq system generating 2 ×
130 150 bp reads.

132 For *E. coli* isolates from Spain, DNA was extracted using the DNA blood and tissue Qiagen
133 kit according to the manufacturer's instruction. The total amount of DNA was quantified using
134 a Qubit fluorometer and frozen at -20°C until further analysis. Libraries were prepared using
135 Nextera XT DNA Library preparation (Illumina), and DNA samples were sequenced using a
136 MiSeq platform (2 × 300 cycle V3 Kit).

137 The DNA of the *E. coli* isolates from Vietnam was extracted using the Wizard Genomic DNA
138 purification kit (Promega, Madison, WI, USA) following the manufacturer's instructions. The
139 concentration of the DNA was measured fluorometrically by using picogreen (Invitrogen). The
140 sequencing was performed using an Illumina HiSeq 4000 system, which generates 2 × 150 bp
141 reads.

142 **c) Quality control**

143 Adapter sequences were removed from raw reads using flexbar v3.0.3(14,15) with trimming
144 mode (-ae) ANY. Low-quality bases within raw reads (Phred score value <20) were trimmed
145 using a sliding window approach (-q WIN). FastQC v0.11.7(16) and MultiQC v1.6(17) were
146 used for quality control before and after processing steps.

147 **d) Genome assembly and annotation**

148 Adapter-trimmed reads were assembled using SPAdes v3.13.1(18) using read correction.
149 Scaffolds smaller than 500bp were discarded. QUAST v5.0.0(19) was used to assess assembly

150 quality using default parameters. Draft assemblies were excluded if the N50 was below an
151 arbitrary value of 30 kbp or consisted of more than 900 contigs. Draft genomes were annotated
152 using prokka v1.13(20) with a genus-specific blast for *Escherichia*. Phylogroups were
153 predicted using ClermonTyper v1.4.1(21), and sequence types (STs) of the isolates were
154 identified *in silico* using the Achtman seven gene MLST scheme using mlst
155 (<https://github.com/tseemann/mlst>).

156 **e) Pan-genome and phylogenetic analysis**

157 Roary v3.12.0(22) was used to define the *pan*-genome of the population, using paralog
158 splitting. The core genes were aligned using prank(23) on default parameters. The core gene
159 alignment was used to construct the phylogenetic tree using RaxML 8.2.4(24) with 100
160 bootstraps under a General Time Reversible (GTR) substitution model with the Gamma model
161 of rate heterogeneity and Lewis ascertainment bias correction(25). The core gene phylogeny
162 was corrected for recombination using ClonalFrameML(26) using default parameters.
163 Phylogenetic Clusters (or BAPS clusters) within the dataset were defined using
164 hierBAPS(27,28) based on the core gene alignment. The accessory gene clustering was
165 performed using package Rtsne v0.15(29,30) with 5000 iterations and perplexity 15 in R
166 v3.6.1. iTOL(31) and Microreact(32) were used to visualize the population structure in the
167 context of available metadata. The function chisq.test from the MASS library(33) (v7.3-51.1)
168 was used in R(34) (v3.5.2) to perform X²-tests of independence between phylogenetic clusters
169 and host species. Tests were carried out on the full dataset (14 phylogenetic clusters vs. five
170 hosts and nine phylogroups vs. five host species).

171 **f) Genome-wide association study (GWAS)**

172 We excluded the wild boar *E. coli* isolates from the GWAS analysis, because of their low
173 number (n=29). GWAS was performed to screen *k-mers* for associations with their host (pig,

174 human, chicken, and cattle). Assemblies were shredded into *k-mers* of 9-100 bases using FSM-
175 lite (<https://github.com/nvalimak/fsm-lite>). The association between *k-mers* and host
176 phenotype was carried out using Fast-LMM linear mixed model implemented in pyseer(35)
177 using a pairwise similarity matrix derived from the phylogenetic tree as population correction.
178 A GWAS analysis was carried out for each host (pig, human, chicken, and cattle). To reduce
179 false-positive associations, isolates from the host of interest were compared with an equal
180 number of isolates from each of the other hosts, designated control isolates. This analysis was
181 repeated 100 times per host of interest by selecting the control strains from other hosts per
182 iteration(36). The selection of control isolates was random and with replacement except for
183 stratification by phylogenetic clusters to minimize phylogenetic bias. The statistical
184 significance threshold was estimated based on the number of unique *k-mers* patterns for each
185 run(35). *K-mers*, which were significantly associated with 90% of the runs per host, were
186 retained and mapped to reference genomes (Table S2) using a fastmap algorithm in bwa(35,37).
187 An arbitrary cut-off of a minimum of 10 *k-mers* mapped per gene was chosen for further
188 analysis to reduce false-positives. *In silico* characterization and gene ontology (GO)
189 assignment was performed using Blast2GO(38), and Clusters of Orthologous Groups (COGs)
190 were assigned using CD-search(39,40).

191 **g) Prevalence of a human-associated *nan* gene cluster**

192 All available *E. coli* genome assemblies in NCBI RefSeq were downloaded on Nov 29th, 2019,
193 using NCBI-genome-download (<https://github.com/kblin/ncbi-genome-download>). Using a
194 custom ABRicate (<https://github.com/tseemann/abricate>) database, consisting of the nine genes
195 of the novel human-associated *nan* gene cluster, all downloaded genomes (n=17994) were
196 scanned. STs for all the genomes were assigned as described above.

197 **h) Construction of mutants and phenotypic experiments**

198 Mutants $\Delta nan-9$ (Amp^R) and $\Delta nanRATEK$ of extra-intestinal pathogenic *E. coli* (ExPEC) strain
199 IMT12185 (ST131; RKI 20-00501; Amp^R) were constructed using the Datsenko-Wanner
200 method(41). The genomic DNA of the wild-type and the mutant strains was isolated using a
201 QIAamp DNA Mini Kit (QIAGEN). Libraries were prepared using the Nextera XT DNA
202 Library preparation kit (Illumina), and MinION one-dimensional (1D) libraries were
203 constructed using the SQK-RBK004 kit (Nanopore technologies, Oxford, UK) and loaded
204 according to the manufacturer's instructions onto an R9.4 flow cell. MinIon sequencing data
205 were collected for 48 h and the paired-end Illumina sequencing was performed using MiSeq.
206 Hybrid assembly using Illumina and MinION reads was performed using unicycler v0.4.8(42)
207 with default parameters to complete both strains' genomes. The absence of the desired genes
208 was confirmed based on the assembly followed by annotation using prokka v1.13(20).
209 Carbon utilization and chemical sensitivity of the deletion mutants and their parental strain
210 were tested using a Biolog Phenotypic Array system, using the PM1 MicroPlate and the Gen
211 III MicroPlate according to the manufacturer's instructions.

212 **i) Growth curve analysis**

213 *E. coli* strains were grown at 37°C aerobically in lysogeny broth (LB) (10 g/l tryptone, 5 g/l
214 yeast extract, 5 g/l NaCl, pH 7.5) or in minimal medium (MM). MM is M9 mineral medium
215 (33.7 mM Na₂HPO₄, 22.0 mM KH₂PO₄, 8.55 mM NaCl, 9.35 mM NH₄Cl) supplemented with
216 2 mM MgSO₄ and 0.1 mM CaCl₂. As carbon and energy source, either 27.8 mM [0.5% w/v]
217 glucose, 6.47 mM [0.2% w/v] 5-N-acetyl neuraminic acid (Neu5Ac), or 6.15 mM [0.1% w/v]
218 N-glycolylneuraminic acid (Neu5Gc) (all purchased from Sigma-Aldrich, Taufkirchen,
219 Germany) was added. If appropriate, the following antibiotics were used: ampicillin sodium
220 salt (150 µg/ml) or kanamycin (50 µg/ml). For solid media, 1.5% agar (w/v) was added. For
221 all growth experiments, bacterial strains were grown in LB medium overnight at 37°C, washed

222 twice in PBS and then adjusted to an optical density at 600 nm (OD_{600}) of 0.005 in the desired
223 liquid growth medium, or streaked on agar plates. Growth curves were obtained from bacterial
224 cultures incubated at 37°C with gentle agitation in 96-well microtitre plates containing 200 μ l
225 medium. The OD_{600} was measured by an automatic reader (Epoch2T; BioTek, Bad
226 Friedrichshall, Germany) at appropriate time intervals as indicated.

227

228 **Results**

229 **Data collection**

230 After WGS quality control, 14 isolates were excluded because of poor quality sequences. One
231 additional isolate was excluded since this isolate was identified as *Escherichia marmotae*
232 (formerly cryptic clade V)(43,44), a species commonly mistaken for *E. coli*. Our final
233 collection comprised 1198 *E. coli* whole-genome sequences with metadata (Table S1), which
234 also contained 8 cryptic clade I isolates, which were included as *E. coli* based on the
235 recommended species cut-off of 95-96% average nucleotide identity(43). Our collection
236 consisted of 22.1% (n=265) cattle, 28.1% (n=337) chicken, 27.3% (n=327) human, 20.3%
237 (n=240) pigs and 2.4% (n=29) wild boar isolates (Fig. S1A). Fifty-one percent (n=612), 19.4%
238 (n=233), 14.5% (n=174) and 14.9% (n=179) of these isolates were from Germany, Spain, the
239 UK, and Vietnam, respectively (Fig. S1A). Chicken isolates were from all four countries,
240 human isolates from Germany, the UK and Vietnam, pig isolates from Germany, Spain and
241 Vietnam, cattle isolates from Germany and Spain and only Spain provided wild boar isolates.
242 In total, 35.5% (n=426) of the isolates were from hosts with reported disease, whereas 62.0%
243 (n=743) were from hosts without reported disease, while host health status was unknown for
244 the wild boar isolates (2.4%, n=29). Of the 1198 isolates analyzed, 1140 were grouped into 358
245 different STs, and 58 could not be assigned to any known ST. The population structure of the
246 collection closely resembles that of the ECOR collection(45), indicating that it represents most
247 of the known diversity of *E. coli* *sensu stricto* (Fig. S2).

248 **Pan-genome analysis**

249 The *pan*-genome of the 1198 *E. coli* isolates consisted of 77130 genes, of which 1956 genes
250 belonged to the core genome (i.e., present in at least 99% of the isolates). The population
251 structure of the collection based on core genome single-nucleotide polymorphisms (SNPs) was

252 defined using Bayesian analysis of population structure (BAPS), which assigns isolates to
253 discrete clusters. Most of the isolates were assigned to phylogroups B1 (n=366, 30.55%), A
254 (n=313, 26.12%) and B2 (n=213, 17.77%). The remaining isolates were distributed among
255 phylogroups D (n=97, 8.09%), E (n=55, 4.59%), G (n=49, 4.09%), F (n=35, 2.92%), C (n=60,
256 5.0%), and clade I (n=8, 0.6%). A comparison of phylogenetic clusters, phylogroups, country,
257 host, and a maximum likelihood (ML) tree based on 110920 core-genome SNPs is shown in
258 Fig 1. The χ^2 -tests for independence revealed a positive correlation between host status and
259 phylogenetic clusters (at $p < 2.26e^{-16}$, df=52) and between phylogroups and hosts ($p < 2.2e^{-16}$,
260 df=32). This indicates that specific phylogenetic clusters (Fig. S1 B&C) and phylogroups, such
261 as B1 (cattle), A (pig), B2 (human and chicken), and G (chicken) were enriched within different
262 hosts in our collection (Fig. S1D).

263 Clustering of isolates based both on core gene alignment and on accessory gene profile
264 appeared to be correlated with phylogroups. The interactive visualization of data is also
265 available on Microreact (<https://microreact.org/project/ouDOdcFxc>). A minimum spanning
266 tree was built on the allelic profiles of 358 (n=1,140 isolates) known STs and 58 isolates
267 belonging to unknown STs using GrapeTree(46) along with the host distribution (Fig. S3).
268 Several sequence types, of which at least ten isolates were available, appeared to be linked with
269 certain host species. ST33 (n= 10/10, 10 human isolates out of all 10 isolates), ST73 (n=11/17),
270 ST131 (n=37/42) and ST1193 (n=12/12) were associated with a human host. ST131 was also
271 found in chickens (n=4/42) and pigs (n=1/42) in this collection. ST23 (n=18/22), ST95
272 (n=25/31), ST115 (n=11/11), ST117 (n=30/33), ST140 (n=19/20) and ST752 (n=29/30) were
273 associated with the chicken host.

274

275

276

277 **GWAS**

278 The genome-wide association analysis was performed on 1169 *E. coli* isolates from cattle,
279 chickens, humans, and pigs. The 29 wild boar isolates were excluded because of their small
280 group size. Genome-wide association analysis revealed the positive association ($\beta > 0$) of
281 27,854, 16,164, and 69,307 *k-mers* with *E. coli* isolates from humans, cattle, and chickens at a
282 likelihood ratio test *p-value* less than 1.87×10^{-9} , 2.16×10^{-9} , and 1.9×10^{-9} respectively (reported
283 as “lrt-pvalue”). There were no *k-mers* significantly associated with the pig host. The
284 significant *k-mers* accounted for 426, 179, and 915 bacterial genes associated with isolation
285 from human, cattle, and chicken hosts, respectively (Fig 2 and Table S3). An arbitrary cut-off
286 of at least 10 *k-mers* mapped per gene was chosen to select genes for *in silico* functional
287 characterization as well as COG assignment using Blast2GO(38) (Table S4) and CD-
288 search(39,40) (Fig. S4).

289 **Association of novel *nan* genes with human host**

290 GWAS revealed a strong association of nine contiguous genes, assigned to the group of *nan*
291 genes with the human host (Fig 2b). Seven of these genes were annotated *in silico* as *nan* genes
292 (Fig 3a) and the remaining two genes were annotated as being similar to *axeA1* of *Prevotella*
293 *ruminicola* ATCC 19189 (Uniprot accession D5EV35). However, the amino acid sequences of
294 the products of these *axeA1*-like genes only shared 19-20% similarity with AxeA1. Further
295 investigation with EggNOG and CD search revealed an acetylesterase/lipase-encoding region
296 (COG0657) in both genes and confirmed *nan* gene annotations. Previous evidence and the
297 genomic location (i.e., between the *nan* genes; Fig 3a) suggest that these genes encode potential
298 acetylesterases and may be analogous to sialyl esterases (NanS)(47). Hence, these nine novel
299 *nan* genes are collectively termed “human-associated *nan* gene cluster (*nan-9*)” (Fig 3a).

300

301 Distinct *nan* genes are present in *E. coli* and are also known as the sialoregulon (*nanRATEK*-
302 *yhcH*, *nanXY* [*yjhBC*], and *nanCMS*; Fig 3a)(48). The sialoregulon is known to be involved in
303 metabolism of sialic acids(49–51), a diverse group of nine-carbon sugars, abundant in the
304 glycocalyx of many animal tissues(52,53). Sialic acids present on mucin proteins in the human
305 gut are an essential energy source for many intestinal bacteria(54). The proteins encoded by
306 the seven genes of *nan-9* (i.e. *nanAKTCMRS*) share 45-64% similarity with the corresponding
307 *nan* genes of the sialoregulon in *E. coli* or the recently described phage-encoded *nanS-p* genes
308 of enterohemorrhagic *E. coli*(55). Both the human-associated *nan* gene cluster and the
309 sialoregulon are located on the bacterial chromosome. The human-associated *nan* gene cluster
310 was found in 7% of our isolate collection, whereas the genes comprising the sialoregulon were
311 more common. In our collection, *nanXY* was identified in ~15% of isolates, *nanCMS* in ~93%
312 of isolates, whilst *nanRATEKyhcH* was found in almost all (>99%) isolates.

313

314 The *nan-9* cluster was detected in 86 isolates, mainly from phylogroups B2 and D (Fig 3b) and
315 predominantly in isolates belonging to ST131, ST73, and ST69, both in our collection as well
316 as across 17,994 RefSeq *E. coli* genomes (Fig 3c). The order and orientation of genes in the
317 human-associated *nan* gene cluster were found to be identical in 82 out of 86 isolates (Fig. S5).
318 In 63 isolates, insertion sequence (IS) 682 was found upstream, and in 23 isolates, IS2 was
319 found downstream of this novel gene cluster (Fig. S5).

320

321 To further explore the function of the human-associated *nan-9* gene cluster, the entire cluster
322 was knocked-out from strain IMT12185 (ST131), yielding strain IMT12185 Δ *nan-9*. For
323 comparison, an additional mutant, which lacked the *nanRATEK* locus from the sialoregulon
324 (IMT12185 Δ *nanRATEK*) was constructed from wild-type IMT12185. Correct gene deletion in
325 both mutants was confirmed through WGS. No significant differences in carbon utilization and

326 chemical sensitivity were observed between wild-type strain IMT12185 and its mutant
327 IMT12185 Δ *nan-9* in Biolog phenotyping array experiments (PM1 and Gen III MicroPlates).

328

329 Deletion mutant IMT12185 Δ *nan-9* was grown in MM with 0.2% 5-N-acetylneuraminic acid
330 (Neu5Ac) or with 0.1% N-glycolylneuraminic acid (Neu5Gc) as sole carbon and energy
331 source. Neu5Ac is the most common sialic acid of the glycocalyx of both humans and other
332 mammals, whereas Neu5Gc is absent in humans. In the presence of Neu5Ac, mutant
333 IMT12185 Δ *nan-9* grew to a maximal OD₆₀₀ of 1.34 comparable to that of parental strain
334 IMT12185 (OD₆₀₀ = 1.37). However, the mutant exhibited a delayed growth start of
335 approximately three hours (Fig 4a). When Neu5Gc was offered as substrate, the mutant not
336 only showed a similar growth start retardation, but also a slower growth rate and a lower
337 maximal OD₆₀₀ (1.31) in comparison with strain IMT12185 (OD₆₀₀ = 1.43) (Fig 4b). Both
338 Neu5Ac and Neu5Gc are degraded by the enzymatic activities of the enzymes NanRATEK, of
339 which four, namely NanRATK, are encoded by redundant genes located on the determinants
340 *nanRATEK* and *nan-9*. Deletion mutant IMT12185 Δ *nanRATEK* was unable to grow with
341 Neu5Ac (Fig 4c), demonstrating that *nan-9* alone is not sufficient for sialic acid degradation,
342 probably due to a lack of *nanE* in the *nan-9* gene cluster. To exclude a pleiotropic effect of the
343 *nan-9* deletion, parental strain IMT12185 and its mutant IMT12185 Δ *nan-9* were grown in LB
344 medium. No significant difference was observed between the two growth curves (Fig 4d).
345 These data demonstrate that the *nan-9* determinant of strain IMT12185 is biologically
346 functional and contributes to the degradation of the sialic acids Neu5Ac and Neu5Gc.

347

348

349 **Other genes associated with the human host**

350 Several other genes associated with the human host were identified in the GWAS analysis,
351 such as the *sat* gene encoding a serine protease autotransporter vacuolating toxin (Fig 2b)(56).
352 This gene was detected in 22.9% (n=75/327) of the human isolates in our collection and in only
353 0.59% (n=5/891) of the strains isolated from other hosts (Table S5). This gene was mainly
354 detected in isolates belonging to specific lineages such as ST131, ST1193, and ST73 (Table
355 S5). In addition, we found an association with two distinct homologs of the *macB* gene that
356 encodes an ABC transporter(57) and is involved in many diverse processes, such as resistance
357 to macrolides(58), lipoprotein trafficking(59), and cell division(60).

358 **Association of distinct Omptins with the cattle and chicken hosts**

359 We detected homologs of the *ompT* (encoding outer-membrane protease VII) gene, a member
360 of the omptin family of proteases, in our dataset (Fig 2a & Fig 2c). Two homologs, *ompP*
361 (UniProt accession P34210, sharing 70% amino acid identity with OmpT) and *arlC* (also
362 referred to as *ompTp*, UniProt accession Q3L7I1, sharing 74% amino acid identity with
363 OmpT), were found to be associated with the cattle and chicken hosts, respectively (Fig 5). In
364 our collection, *ompP* was predominant in phylogroup B1 (n=68), whereas *arlC* was found in
365 distinct phylogroups (such as B2, B1 and G) (Fig 5) and in isolates belonging to ST95 and
366 ST117 (Table S6). A similar association was observed in 17,994 public *E. coli* genomes from
367 RefSeq (Table S6). Previous studies have reported an increased prevalence of *arlC*
368 (erroneously reported there as *ompT*) in a cluster of uropathogenic *E. coli* (UPEC) and avian
369 pathogenic *E. coli* (APEC) classified as ST95(61). Notably, *arlC* is associated with increased
370 degradation of antimicrobial peptides (AMPs) in UPEC isolates(62). OmpP is also able to
371 degrade AMPs and displays a AMP cleavage specificity different from that of OmpT(63).

372

373 **Association of genes involved in metal acquisition with the chicken host**

374 GWAS analysis revealed an association of the *iroBCDEN* gene cluster (C) with the chicken
375 host, but not with other host species included in this study. The prevalence of the *iro* gene
376 cluster was 24.3% (n=291/1198) in our collection, of which 61.5% (n=179/291) were from the
377 chicken host. The gene cluster was found in different STs and with higher prevalence in STs
378 such as ST117, ST95, ST23, and ST140 (Table S7). The chromosomal *iroBCDEN* gene cluster
379 was first described in *Salmonella enterica* and is involved in uptake of catecholate-type
380 siderophores, high-affinity iron-chelating molecules contributing to bacterial survival during
381 infection by sequestering iron(64). In *E. coli*, this gene cluster has mainly been described in
382 uropathogenic (UPEC) and avian pathogenic *E. coli* (APEC) and is regarded as a virulence
383 factor(65). The cluster has been reported on a chromosomal pathogenicity island, although in
384 ExPEC, the cluster can also be located on ColV or ColBM virulence plasmids(66,67). In
385 addition, homologs of genes involved in zinc catabolism (*znuB*) and iron metabolism (*fes*) were
386 found to be associated with the chicken host (Fig 2c).

387 **Discussion**

388 *Escherichia coli* can colonize many different ecological niches in a diverse range of host
389 species, ranging from a commensal lifestyle to intra- or extra-intestinal infections. Presence of
390 certain adhesin and other virulence-associated genes is well known to correlate with the relative
391 ability of *E. coli* strains to colonize the intestinal tract of certain hosts (e.g., *ecp* for humans(68),
392 F9 fimbriae and H7 flagellae for cattle(69,70) or Stg fimbriae for chickens(71)). Variations in
393 host adaptation levels and their molecular basis in *E. coli* strains presumptively realizing a
394 commensal-like lifestyle in the reservoir host are rarely described and poorly understood as of
395 yet(72). Commensal *E. coli* strains may be carriers of AMR and a source of mobile genetic
396 elements conferring AMR to other bacteria including pathogenic strains in a shared
397 microbiome, e.g. in the intestinal tract of animals including humans. We therefore collated an
398 extensive and diverse dataset to identify genetic determinants of *E. coli* host adaptation. We
399 observed significant enrichment of specific hosts within some phylogroups and STs in our
400 collection. Furthermore, we unveiled correlations between the likelihood of genetically related
401 isolates having been isolated from a certain host with the possession of distinctive genetic traits.
402 Some of these traits, e.g. the *iroBCDEN* gene cluster, have been linked to *E. coli* and
403 *Salmonella* virulence before, while others, in particular the human-associated *nan* gene cluster,
404 are novel traits and have not been implicated in the infection and colonization process of *E. coli*.
405 Of note, the latter gene cluster encodes for metabolic properties which have received little
406 attention in bacterial infectious disease research. Specific metabolic properties have been
407 linked to the relative ability of Shiga toxin-encoding *E. coli* (STEC) to asymptotically
408 colonize cattle, their reservoir host(73). Unraveling the nutrient and energy flows in the
409 complex interplay of intestinal bacteria, the surrounding microbiome and the host may open
410 novel avenues to control the persistence and transmission of pathogenic and/or antimicrobial
411 resistant bacteria(74).

412

413 We employed a *k-mer* based bacterial GWAS, applied in previous studies to associate multiple
414 types of genetic variation with phenotypes(75,76). In our study, we were able to associate a
415 phenotype (i.e., isolates obtained from a certain host species) with the presence of specific
416 genes, but not with sequence variation at the level of single nucleotide polymorphisms between
417 genes. This lack of associations found at the SNP level could possibly be explained by the fact
418 that through our filtering approach to prevent false positive hits, we might have excluded *k-*
419 *mers* that captured host-associated SNP variation. Secondly, it might be possible that since
420 *E. coli* is genetically diverse, host-associated SNP variation is challenging to capture between
421 unrelated strains. Finally, the absence of host-associated SNPs might be a biological
422 observation, indicating that colonization of particular hosts is determined by gene presence or
423 absence rather than minimal genetic variation within genetic elements. However, we were able
424 to confirm previously published host associations, indicating the validity of our approach. For
425 example, carriage of the salmochelin operon encoded by *iroBCDEN* and involved in iron
426 metabolism was previously identified as associated with increased ability of *E. coli* strains to
427 colonize chickens(65,77).

428 In addition to *iroBCDEN*, we found an association of omptin proteins (OmpP and ArlC) with
429 chickens and cattle as hosts, respectively. Earlier studies using UPEC strains had demonstrated
430 that these proteins are associated with cleavage and inactivation of cationic antimicrobial
431 peptides (AMPs)(62). Because AMPs are secreted as part of the host's innate immune
432 response(78–80), these proteins may play a vital role in colonization. AMPs are also
433 increasingly used as alternatives to antimicrobial agents in animal farming(81–83), further
434 investigation into the contribution of these Omp variants to host colonization as well as to
435 resistance to exogenous AMPs is warranted.

436

437 We did not identify any significant associations of *k-mers* with the pig host. Bacterial
438 colonization of the porcine intestine by edema-disease *E. coli* (EDEC) is mediated by the ability
439 of these bacteria to adhere to villous epithelial cells via their cytoadhesive F18 fimbriae(84).
440 The expression of receptors for these fimbriae on the apical enterocyte surface is inherited as a
441 dominant trait among pigs and determines susceptibility to diseases caused by F18-fimbriated
442 pathogenic *E. coli*(85). Enterotoxigenic *E. coli* (ETEC) express F4 or F5 fimbriae with similar
443 consequences(86). However, we found only three, four and six isolates harbouring genes for
444 F4, F5 and F18 fimbriae, respectively. Thus, we might not have had all *E. coli* pathovars
445 associated with pig host sufficiently present in our collection, although we did observe an
446 association between phylogroup A and pig colonization. An alternative reason might be that
447 the association between phylogroup A and pig colonization complicated the identification of
448 statistically significant *k-mers*. GWAS corrects for population structure, which means that if
449 there is a strong association between lineage and phenotype, the genes harbored by that lineage
450 will not be reported as having a strong association with the phenotype under study(87).

451
452 We identified a novel human host-associated *nan* gene cluster, distinct from the previously
453 reported sialic acid (Sia) metabolic operon (*nanRATEK-yhcH*, *nanXY*, and *nanCMS*)(48). This
454 novel cluster is conserved and abundant in ExPEC lineages, such as ST131, ST73, and ST69.
455 The gene cluster is flanked by insertion sequences which might play a role in the horizontal
456 exchange between different *E. coli* lineages. Knock-out *in vitro* studies indicated that this novel
457 *nan-9* gene cluster contributes to catabolism of the sialic acids Neu5Ac and Neu5Gc, although
458 it cannot replace the function of the *nanRATEK* locus which is abundant in *E. coli*. Hence, we
459 hypothesize that *E. coli* harboring the *nan-9* gene cluster have an evolutionary advantage
460 through either more efficient access to sialic acids or through access to more diverse sialic
461 acids. The genes annotated as acetylxyran esterases are expected to represent novel sialyl

462 esterases, as known sialyl esterases (*nanS* variants) have previously been mistaken for
463 acetylxyran esterases(47). Additional sialyl esterases – possibly with alternative deacetylation
464 specificity – might provide a more efficient catabolism of acetylated sialic acids. Future studies
465 should investigate the role of the human-associated *nan-9* gene cluster in the catabolism of
466 differentially acetylated sialic acids and their relevance for the human host.

467

468 Approximately one-third of the isolates in our dataset were obtained from diseased hosts, while
469 the remaining isolates were from healthy hosts. Many of the isolates in our dataset that originate
470 from healthy hosts belong to ExPEC lineages which are typically considered to be pathogenic.
471 In fact, the locus most strongly associated with the human host, the *nan-9* gene cluster, is
472 abundant in ExPEC lineages. This does not necessarily mean that the *nan-9* gene cluster is
473 associated with pathogenicity. In fact, this observation primarily supports the notion that these
474 pathogenic *E. coli* are highly efficient colonizers of the human intestine(72). Based on our
475 results, we hypothesize that the human-associated *nan-9* gene cluster is one of the factors
476 driving the adaptation of ExPEC to the human intestine.

477

478 Finally, we observed an association between the *sat* gene and human host colonization. *Sat*
479 contributes to the pathogenicity of *E. coli* in the urinary tract(56). The high prevalence of *sat*
480 in previously studied *E. coli* isolates from the feces of healthy individuals suggests it may not
481 act as a virulence factor in the human gut(88). However, in our isolate collection, the *sat* gene
482 was found in *E. coli* strains belonging to phylogroups A, B2, D, and F, which had been isolated
483 from both healthy and diseased hosts (Table S5). Understanding the role of *Sat* in the
484 colonization and adaptation of *E. coli* in healthy humans warrants further investigation.

485

486

487 **Conclusion**

488 Our study identified several distinct genetic determinants that may influence *E. coli* adaptation
489 to different host species and provide an adaptive advantage. These findings are important as
490 they aid the better understanding of the potential outcome of transmission events of *E. coli*
491 between host species. This is particularly relevant for the control of the spread of antimicrobial
492 resistant commensal and zoonotic *E. coli* strains within and across human and animal
493 populations. The data generated here can also be used in risk analysis and for diagnostic and
494 monitoring purposes. More importantly, our study identified biological processes, including
495 sialic acid catabolism, that should be investigated in more detail to better understand *E. coli*
496 host adaptation.

497

498 **Data availability:**

499 The raw-reads of the 1090 *E. coli* isolates sequenced in this study were submitted to NCBI
500 SRA with the Bioproject accession number PRJNA739205 and the SRA accession of 108
501 isolates, that were taken from other studies, were provided in supplement table S1.

502 **Acknowledgements:**

503 The HECTOR research project was supported under the framework of the JPIAMR - Joint
504 Programming Initiative on Antimicrobial Resistance – through the 3rd joint call, thanks to the generous
505 funding by the Netherlands Organisation for Health Research and Development (ZonMw, grant number
506 547001012), the Federal Ministry of Education and Research (BMBF/DLR grant numbers 01KI1703A,
507 01KI1703C and 01KI1703B), the State Research Agency (AEI) of the Ministry of Science, Innovation
508 and Universities (MINECO, grant number PCIN-2016-096), and the Medical Research Council (MRC,
509 grant number MR/R002762/1).

510 **Competing Interests:**

511 The authors declare no competing interests.

512 **References:**

- 513 1. Alteri CJ, Mobley HLT. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host
514 microenvironments. Vol. 15, *Current Opinion in Microbiology*. 2012.
- 515 2. Ewers C, Bethe A, Semmler T, Guenther S, Wieler LH. Extended-spectrum β -lactamase-producing and
516 AmpC-producing *Escherichia coli* from livestock and companion animals, and their putative impact on
517 public health: A global perspective. *Clinical Microbiology and Infection*. 2012.
- 518 3. Bonnet R, Beyrouty R, Haenni M, Nicolas-Chanoine M-H, Dalmasso G, Madec J-Y. Host
519 Colonization as a Major Evolutionary Force Favoring the Diversity and the Emergence of the
520 Worldwide Multidrug-Resistant *Escherichia coli* ST131. *MBio*. 2021;12(4).
- 521 4. Lopatkin AJ, Meredith HR, Srimani JK, Pfeiffer C, Durrett R, You L. Persistence and reversal of
522 plasmid-mediated antibiotic resistance. *Nat Commun*. 2017;8(1).
- 523 5. Pitout JDD, DeVinney R. *Escherichia coli* ST131: A multidrug-resistant clone primed for global
524 domination. *F1000Research*. 2017.
- 525 6. Nicolas-Chanoine MH, Bertrand X, Madec JY. *Escherichia coli* st131, an intriguing clonal group. *Clin
526 Microbiol Rev*. 2014;
- 527 7. Skurnik D, Clermont O, Guillard T, Launay A, Danilchanka O, Pons S, et al. Emergence of
528 antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Mol Biol Evol*. 2016;
- 529 8. Riley LW. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. Vol. 20, *Clinical
530 Microbiology and Infection*. 2014.
- 531 9. Cohen E, Azriel S, Austeri O, Gal A, Zitronblat C, Mikhlin S, et al. Pathoadaptation of the passerine-
532 associated *Salmonella enterica* serovar *Typhimurium* lineage to the avian host. *PLoS Pathog*.
533 2021;17(3).
- 534 10. Toft C, Andersson SGE. Evolutionary microbial genomics: Insights into bacterial host adaptation.
535 *Nature Reviews Genetics*. 2010.
- 536 11. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association
537 study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad
538 Sci U S A*. 2013;
- 539 12. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, et al. Genome evolution and the
540 emergence of pathogenicity in avian *Escherichia coli*. *Nat Commun*. 2021;12(1).
- 541 13. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, et al. Large-scale
542 genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res*. 2015;
- 543 14. Dodd M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR-flexible barcode and adapter processing for next-
544 generation sequencing platforms. *Biology (Basel)*. 2012;
- 545 15. Roehr JT, Dieterich C, Reinert K. Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics*.
546 2017;
- 547 16. Andrews S, Krueger F, Seconds-Pichon A, Biggins F, Wingett S. FastQC. A quality control tool for
548 high throughput sequence data. Babraham Bioinformatics. Babraham Institute. 2015.
- 549 17. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools
550 and samples in a single report. *Bioinformatics*. 2016;
- 551 18. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new
552 genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;
- 553 19. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies.
554 *Bioinformatics*. 2013;
- 555 20. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. 2014;
- 556 21. Beghain J, Bridier-Nahmias A, Nagard H Le, Denamur E, Clermont O. ClermonTyping: An easy-to-use
557 and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb Genomics*. 2018;
- 558 22. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale
559 prokaryote pan genome analysis. *Bioinformatics*. 2015;
- 560 23. Löytynoja A. Phylogeny-aware alignment with PRANK. *Methods Mol Biol*. 2014;
- 561 24. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
562 phylogenies. *Bioinformatics*. 2014;
- 563 25. Lewis PO. A likelihood approach to estimating phylogeny from discrete morphological character data.
564 *Syst Biol*. 2001;
- 565 26. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial
566 Genomes. *PLoS Comput Biol*. 2015;
- 567 27. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering
568 of DNA sequences with BAPS software. *Mol Biol Evol*. 2013;
- 569 28. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. RhierBAPs: An R implementation of the
570 population clustering algorithm hierbaps [version 1; referees: 2 approved]. *Wellcome Open Res*. 2018;

571 29. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. *J Mach Learn Res.* 2008;

572 30. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res.* 2015;

573 31. Letunic I, Bork P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and 574 annotation. *Bioinformatics.* 2007;

575 32. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and 576 sharing data for genomic epidemiology and phylogeography. *Microb genomics.* 2016;

577 33. Venables WN, Ripley BD. Modern Applied Statistics with S Fourth edition by. Vol. 53, World. 2002.

578 34. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for 579 Statistical Computing, Vienna, Austria. 2020.

580 35. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: A comprehensive tool for microbial 581 pangenome-wide association studies. *Bioinformatics.* 2018;

582 36. Epping L, Walther B, Piro RM, Knüver MT, Huber C, Thürmer A, et al. Genome-wide insights into 583 population structure and host specificity of *Campylobacter jejuni*. *Sci Rep.* 2021;11(1).

584 37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. 585 *Bioinformatics.* 2009;

586 38. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for 587 annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 2005;

588 39. Marchler-Bauer A, Bryant SH. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* 589 2004;

590 40. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales R, et al. CDD / SPARCLE : the conserved 591 domain database in 2020. *2020;48(November 2019):265–8.*

592 41. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using 593 PCR products. *Proc Natl Acad Sci U S A.* 2000;

594 42. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short 595 and long sequencing reads. *PLoS Comput Biol.* 2017;

596 43. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, et al. Proposed minimal standards 597 for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol.* 2018;

598 44. Liu S, Jin D, Lan R, Wang Y, Meng Q, Dai H, et al. *Escherichia marmotae* sp. nov., isolated from faeces 599 of *Marmota himalayana*. *Int J Syst Evol Microbiol.* 2015;

600 45. Ochman H, Selander RK. Standard reference strains of *Escherichia coli* from natural populations. *J 601 Bacteriol.* 1984;

602 46. Zhou Z, Alikhan N, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree : visualization of 603 core genomic relationships among 100 , 000 bacterial pathogens. 2018;1395–404.

604 47. Steenbergen SM, Jirik JL, Vimr ER. YjhS (NanS) is required for *Escherichia coli* to grow on 9-O- 605 acetylated N-acetylneurameric acid. *J Bacteriol.* 2009;

606 48. Kalivoda KA, Steenbergen SM, Vimr ER. Control of the *Escherichia coli* sialoregulon by transcriptional 607 repressor NanR. *J Bacteriol.* 2013;

608 49. Vimr ER, Kalivoda KA, Deszo EL, Steenbergen SM. Diversity of Microbial Sialic Acid Metabolism. 609 *Microbiol Mol Biol Rev.* 2004;

610 50. Vimr ER, Troy FA. Identification of an inducible catabolic system for sialic acids (nan) in *Escherichia* 611 *coli*. *J Bacteriol.* 1985;

612 51. Bell A, Severi E, Lee M, Monaco S, Latousakis D, Angulo J, et al. Uncovering a novel molecular 613 mechanism for scavenging sialic acids in bacteria. *J Biol Chem.* 2020;295(40).

614 52. Vimr ER. Unified Theory of Bacterial Sialometabolism: How and Why Bacteria Metabolize Host Sialic 615 Acids. *ISRN Microbiol.* 2013;

616 53. Severi E, Hood DW, Thomas GH. Sialic acid utilization by bacterial pathogens. *Microbiology.* 2007.

617 54. Haines-menges BL, Whitaker WB, Lubin JB, Boyd EF. Host Sialic Acids: A Delicacy for the Pathogen 618 with Discerning Taste. In: *Metabolism and Bacterial Pathogenesis.* 2015.

619 55. Saile N, Voigt A, Kessler S, Stressler T, Klumpp J, Fischer L, et al. *Escherichia coli* O157:H7 strain 620 EDL933 harbors multiple functional prophage-associated genes necessary for the utilization of 5-N- 621 acetyl-9-O-acetyl neurameric acid as a growth substrate. *Appl Environ Microbiol.* 2016;82(19).

622 56. Guyer DM, Henderson IR, Nataro JP, Mobley HLT. Identification of Sat, an autotransporter toxin 623 produced by uropathogenic *Escherichia coli*. *Mol Microbiol.* 2000;

624 57. Kobayashi N, Nishino K, Yamaguchi A. Novel macrolide-specific ABC-type efflux transporter in 625 *Escherichia coli*. *J Bacteriol.* 2001;

626 58. Tikhonova EB, Devroy VK, Lau SY, Zgurskaya HI. Reconstitution of the *Escherichia coli* macrolide 627 transporter: The periplasmic membrane fusion protein MacA stimulates the ATPase activity of MacB. 628 *Mol Microbiol.* 2007;

629 59. Khwaja M, Ma Q, Saier MH. Topological analysis of integral membrane constituents of prokaryotic 630

631 60. ABC efflux systems. *Res Microbiol.* 2005;

632 60. Yakushi T, Masuda K, Narita SI, Matsuyama SI, Tokuda H. A new ABC transporter mediating the
633 detachment of lipid-modified proteins from membranes. *Nat Cell Biol.* 2000;

634 61. Johnson TJ, Wannemuehler Y, Johnson SJ, Stell AL, Doekott C, Johnson JR, et al. Comparison of
635 extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset
636 representing potential zoonotic pathogens. *Appl Environ Microbiol.* 2008;

637 62. Desloges I, Taylor JA, Leclerc JM, Brannon JR, Portt A, Spencer JD, et al. Identification and
638 characterization of OmpT-like proteases in uropathogenic *Escherichia coli* clinical isolates.
639 *Microbiologyopen.* 2019;

640 63. Hwang BY, Varadarajan N, Li H, Rodriguez S, Iverson BL, Georgiou G. Substrate specificity of the
641 *Escherichia coli* outer membrane protease OmpP. *J Bacteriol.* 2007;

642 64. Ratledge C, Dover LG. Iron Metabolism in Pathogenic Bacteria. *Annu Rev Microbiol.* 2000;

643 65. Caza M, Lépine F, Milot S, Dozois CM. Specific roles of the iroBCDEN genes in virulence of an avian
644 pathogenic *Escherichia coli* O78 strain and in production of salmochelins. *Infect Immun.* 2008;

645 66. Sorsa LJ, Dufke S, Heesemann J, Schubert S. Characterization of an iroBCDEN gene cluster on a
646 transmissible plasmid of uropathogenic *Escherichia coli*: Evidence for horizontal transfer of a
647 chromosomal virulence factor. *Infect Immun.* 2003;

648 67. Dobrindt U, Blum-Oehler G, Nagy G, Schneider G, Johann A, Gottschalk G, et al. Genetic structure and
649 distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic *Escherichia coli*
650 strain 536. *Infect Immun.* 2002;

651 68. Rendón MA, Saldaña Z, Erdem AL, Monteiro-Neto V, Vázquez A, Kaper JB, et al. Commensal and
652 pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization. *Proc
653 Natl Acad Sci U S A.* 2007;104(25).

654 69. Low AS, Dziva F, Torres AG, Martinez JL, Rosser T, Naylor S, et al. Cloning, expression, and
655 characterization of fimbrial operon F9 from enterohemorrhagic *Escherichia coli* O157:H7. *Infect
656 Immun.* 2006;74(4).

657 70. Mahajan A, Currie CG, Mackie S, Tree J, Mcateer S, Mckendrick I, et al. An investigation of the
658 expression and adhesin function of H7 flagella in the interaction of *Escherichia coli* O157: H7 with
659 bovine intestinal epithelium. *Cell Microbiol.* 2009;11(1).

660 71. Lymberopoulos MH, Houle S, Daigle F, Léveillé S, Brée A, Moulin-Schouleur M, et al.
661 Characterization of Stg fimbriae from an avian pathogenic *Escherichia coli* O78:K80 Strain and
662 assessment of their contribution to colonization of the chicken respiratory tract. *J Bacteriol.*
663 2006;188(18).

664 72. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*.
665 *Nature Reviews Microbiology.* 2010.

666 73. Barth SA, Weber M, Schaufler K, Berens C, Geue L, Menge C. Metabolic traits of bovine shiga toxin-
667 producing *escherichia coli* (STEC) strains with different colonization properties. *Toxins (Basel).*
668 2020;12(6).

669 74. Stecher B. Establishing causality in *Salmonella*-microbiota-host interaction: The use of gnotobiotic
670 mouse models and synthetic microbial communities. *Int J Med Microbiol.* 2021;311(3).

671 75. Ma KC, Mortimer TD, Hicks AL, Wheeler NE, Sánchez-Busó L, Golparian D, et al. Adaptation to the
672 cervical environment is associated with increased antibiotic susceptibility in *Neisseria gonorrhoeae*. *Nat
673 Commun.* 2020;

674 76. Gröschel MI, Meehan CJ, Barilar I, Diricks M, Gonzaga A, Steglich M, et al. The phylogenetic
675 landscape and nosocomial spread of the multidrug-resistant opportunist *Stenotrophomonas maltophilia*.
676 *Nat Commun.* 2020;

677 77. Gao Q, Wang X, Xu H, Xu Y, Ling J, Zhang D, et al. Roles of iron acquisition systems in virulence of
678 extraintestinal pathogenic *Escherichia coli*: Salmochelin and aerobactin contribute more to virulence
679 than heme in a chicken infection model. *BMC Microbiol.* 2012;

680 78. McPhee JB, Small CL, Reid-Yu SA, Brannon JR, Moual H Le, Coombes BK. Host defense peptide
681 resistance contributes to colonization and maximal intestinal pathology by Crohn's disease-associated
682 adherent-invasive *Escherichia coli*. *Infect Immun.* 2014;

683 79. Fjell CD, Jenssen H, Fries P, Aich P, Griebel P, Hilpert K, et al. Identification of novel host defense
684 peptides and the absence of α -defensins in the bovine genome. *Proteins Struct Funct Genet.* 2008;

685 80. Lynn DJ, Higgs R, Gaines S, Tierney J, James T, Lloyd AT, et al. Bioinformatic discovery and initial
686 characterisation of nine novel antimicrobial peptide genes in the chicken. *Immunogenetics.* 2004;

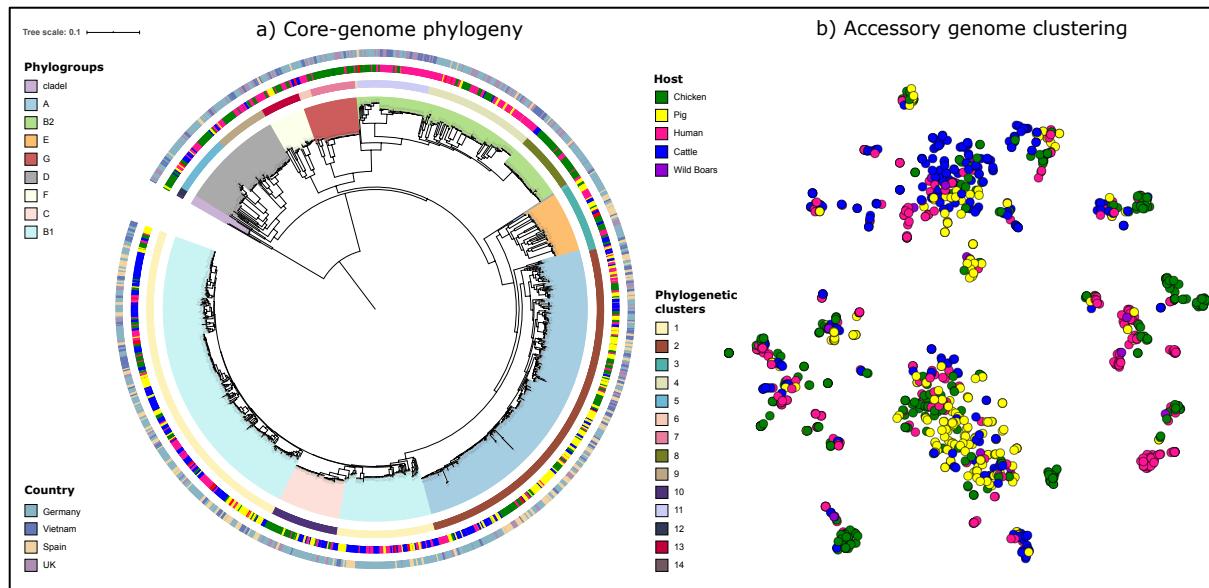
687 81. Li Z, Hu Y, Yang Y, Lu Z, Wang Y. Antimicrobial resistance in livestock: Antimicrobial peptides
688 provide a new solution for a growing challenge. *Anim Front.* 2018;

689 82. Liu Q, Yao S, Chen Y, Gao S, Yang Y, Deng J, et al. Use of antimicrobial peptides as a feed additive
690 for juvenile goats. *Sci Rep.* 2017;

691 83. Xiao H, Shao F, Wu M, Ren W, Xiong X, Tan B, et al. The application of antimicrobial peptides as
692 growth and health promoters for swine. *Journal of Animal Science and Biotechnology*. 2015.
693 84. Barth S, Schwanitz A, Bauerfeind R. Polymerase chain reaction-based method for the typing of f18
694 fimbriae and distribution of f18 fimbrial subtypes among porcine shiga toxin-encoding escherichia coli
695 in Germany. *J Vet Diagnostic Investig*. 2011;23(3).
696 85. Frydendahl K, Jensen TK, Andersen JS, Fredholm M, Evans G. Association between the porcine
697 Escherichia coli F18 receptor genotype and phenotype and susceptibility to colonisation and
698 postweaning diarrhoea caused by *E. coli* O138:F18. *Vet Microbiol*. 2003;93(1).
699 86. Barth S, Tscholshiew A, Menge C, Weiß R, Baljer G, Bauerfeind R. Virulence and fitness gene patterns
700 of Shiga toxin-encoding Escherichia coli isolated from pigs with edema disease or diarrhea in Germany.
701 *Berl Munch Tierarztl Wochenschr*. 2007;120(7–8).
702 87. Power RA, Parkhill J, De Oliveira T. Microbial genome-wide association studies: lessons from human
703 GWAS. *Nat Rev Genet*. 2016;18(1):41–50.
704 88. Toloza L, Giménez R, Fábrega MJ, Alvarez CS, Aguilera L, Cañas MA, et al. The secreted
705 autotransporter toxin (Sat) does not act as a virulence factor in the probiotic *Escherichia coli* strain
706 Nissle 1917. *BMC Microbiol*. 2015;
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737

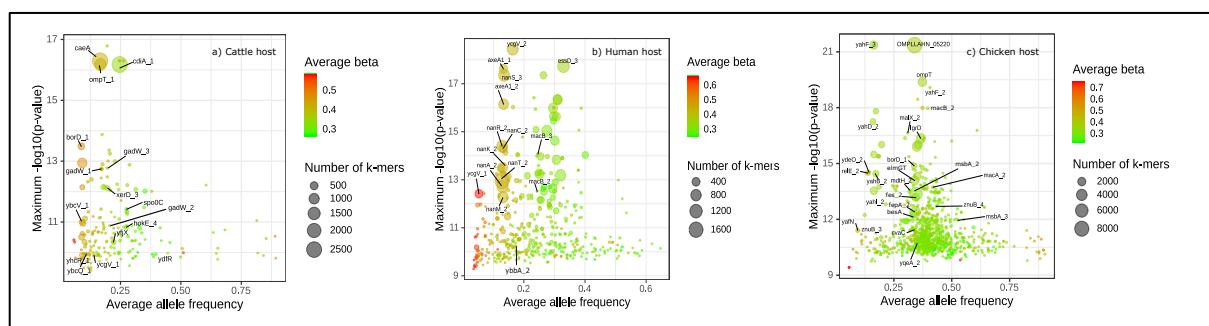
738 Figure Legends:

739 Fig 1: Distribution of 1198 isolates with host species by a) core-genome phylogeny and b) clustering based on
740 accessory gene content (right). Clades on the phylogeny represent phylogroups, inner-ring represents
741 phylogenetic clusters, middle-ring represents host-species, and outer ring indicates the geographical
742 region.



743

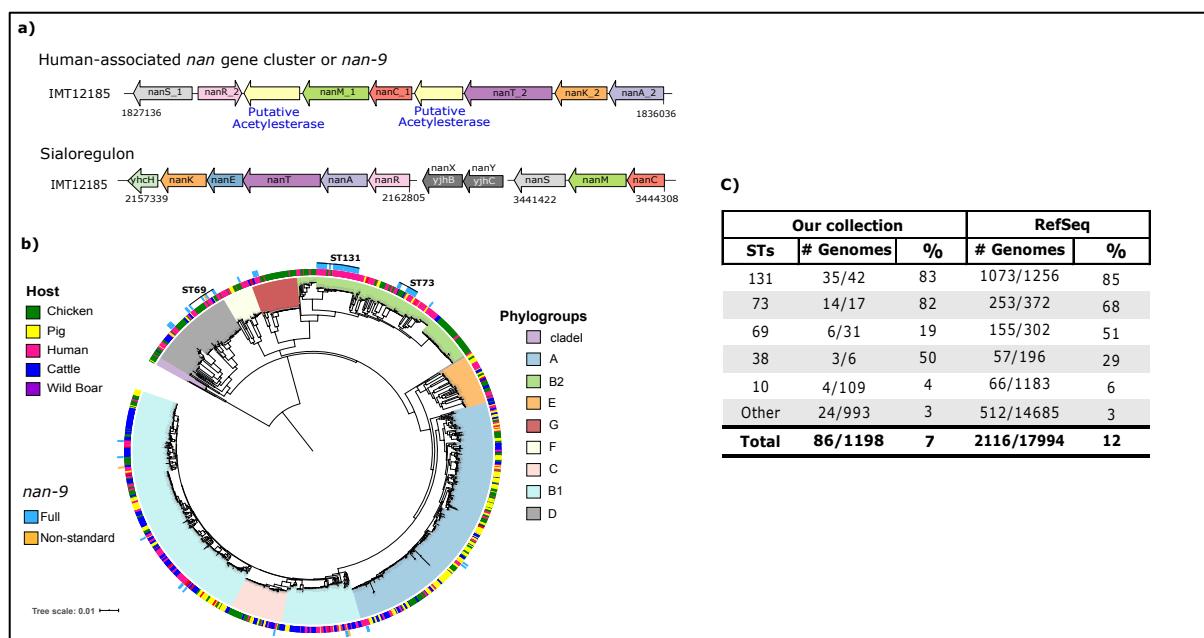
744 Fig 2: Plots representing the *E. coli* genes or gene variants associated with the a) Cattle host, b) Human host, and
745 c) Chicken host. The bubble size represents the number of k-mers mapped to a specific gene, and the color
746 gradient represents the effect size (β).



747

748

749 Fig 3: a) Genetic architecture of the human-associated *nan* gene cluster (*nan-9*) and the sialoregulon on the
750 complete genome of the strain IMT12185. The strain lacks the *nanXY* genes of the sialoregulon. b)
751 Distribution of the *nan-9* cluster on core-genome phylogeny marked with STs with higher prevalence. c)
752 The table indicates the prevalence of the *nan-9* gene cluster in different STs in our collection and in the
753 RefSeq *E. coli* genomes.

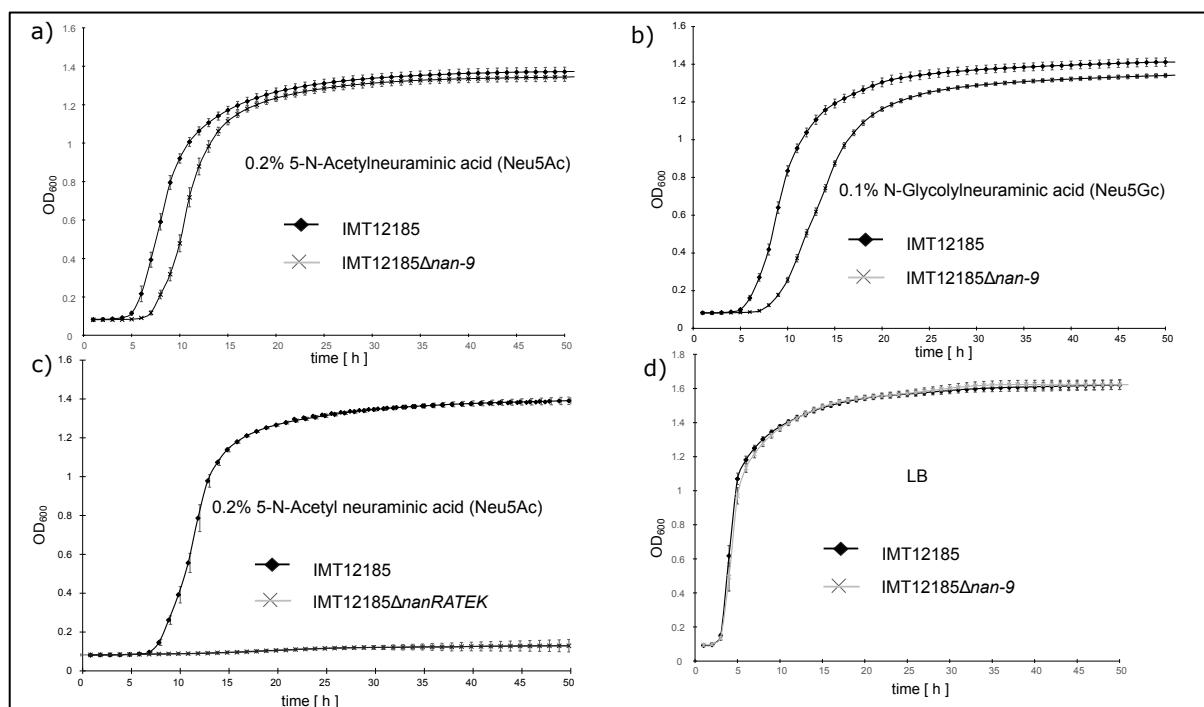


754

755

756

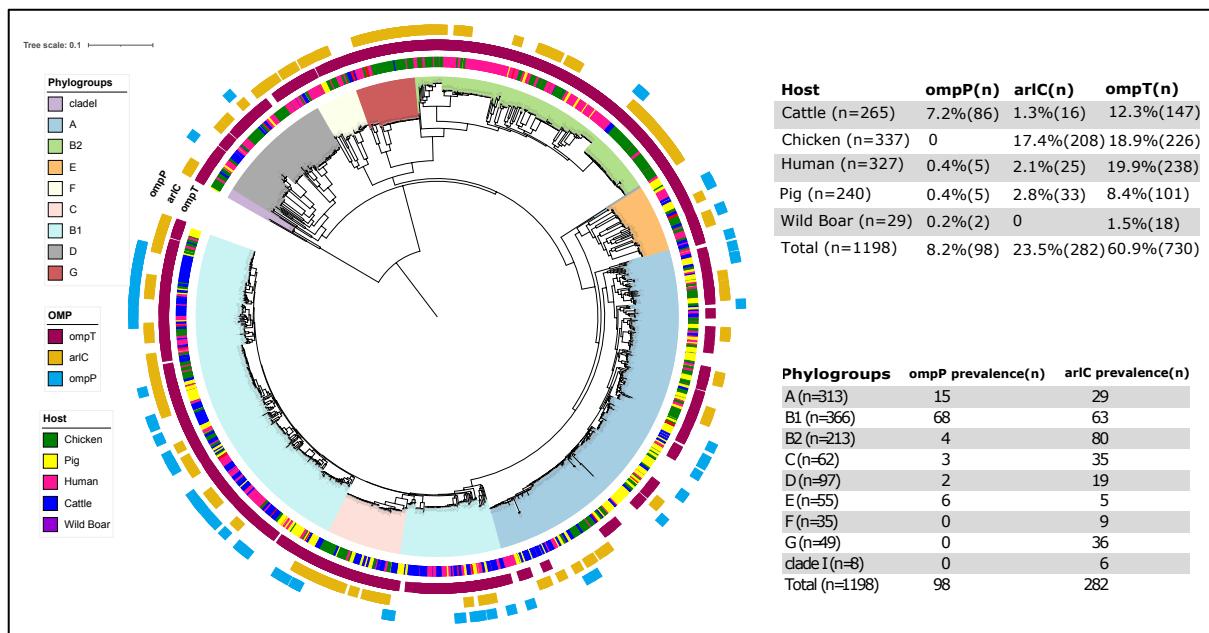
757 Fig 4: Growth curves of *E. coli* IMT12185 and its mutant derivatives in various media. a) Growth of IMT12185
 758 and IMT12185 Δ nan-9 in M9 minimal medium with 0.2% 5-N-Acetylneuraminic acid (Neu5Ac) b)
 759 Growth of IMT12185 and IMT12185 Δ nan-9 in M9 minimal medium with 0.1% 5-N-Glycolylneuraminic acid
 760 (Neu5Gc) c) Growth of IMT12185 and IMT12185 Δ nanRATEK in M9 minimal medium with 0.2% 5-N-
 761 Acetylneuraminic acid (Neu5Ac) d) Growth of IMT12185 and IMT12185 Δ nan-9 in lysogeny broth (LB).



762

763

764 Fig 5: Distribution of *ompP*, *arlC*, and *ompT* genes in phylogroups and host across the phylogeny and their
765 estimated prevalence.



766

767

768