

## ARTICLE


<https://doi.org/10.1038/s41467-021-20988-w>

OPEN

# Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*

Leonardos Mageiros<sup>1</sup>, Guillaume Méric<sup>1</sup>, Sion C. Bayliss<sup>1,2</sup>, Johan Pensar<sup>3,4</sup>, Ben Pascoe<sup>1,3</sup>, Evangelos Mourkas<sup>1</sup>, Jessica K. Calland<sup>1</sup>, Koji Yahara<sup>5</sup>, Susan Murray<sup>6</sup>, Thomas S. Wilkinson<sup>7</sup>, Lisa K. Williams<sup>7</sup>, Matthew D. Hitchings<sup>7</sup>, Jonathan Porter<sup>8</sup>, Kirsty Kemmett<sup>9</sup>, Edward J. Feil<sup>1</sup>, Keith A. Jolley<sup>10</sup>, Nicola J. Williams<sup>9</sup>, Jukka Corander<sup>3,4,11</sup> & Samuel K. Sheppard<sup>1,2,10</sup>✉

Chickens are the most common birds on Earth and colibacillosis is among the most common diseases affecting them. This major threat to animal welfare and safe sustainable food production is difficult to combat because the etiological agent, avian pathogenic *Escherichia coli* (APEC), emerges from ubiquitous commensal gut bacteria, with no single virulence gene present in all disease-causing isolates. Here, we address the underlying evolutionary mechanisms of extraintestinal spread and systemic infection in poultry. Combining population scale comparative genomics and pangenome-wide association studies, we compare *E. coli* from commensal carriage and systemic infections. We identify phylogroup-specific and species-wide genetic elements that are enriched in APEC, including pathogenicity-associated variation in 143 genes that have diverse functions, including genes involved in metabolism, lipopolysaccharide synthesis, heat shock response, antimicrobial resistance and toxicity. We find that horizontal gene transfer spreads pathogenicity elements, allowing divergent clones to cause infection. Finally, a Random Forest model prediction of disease status (carriage vs. disease) identifies pathogenic strains in the emergent ST-117 poultry-associated lineage with 73% accuracy, demonstrating the potential for early identification of emergent APEC in healthy flocks.

<sup>1</sup>The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK. <sup>2</sup>MRC Cloud Infrastructure for Microbial Bioinformatics (CLIMB) Consortium, London, UK. <sup>3</sup>Department of Biostatistics, University of Oslo, Oslo, Norway. <sup>4</sup>Department of Mathematics and Statistics, Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland. <sup>5</sup>Antimicrobial Resistance Research Centre, National Institute of Infectious Diseases, Tokyo, Japan. <sup>6</sup>Uppsala University, Department for medical biochemistry and microbiology, Uppsala University, Uppsala, Sweden. <sup>7</sup>Swansea University Medical School, Institute of Life Science, Swansea SA2 8PP, UK. <sup>8</sup>National Laboratory Service, Environment Agency, Starcross, UK. <sup>9</sup>Department of Epidemiology and Population Health, Institute of Infection & Global Health, University of Liverpool, Leahurst Campus, Wirral, UK. <sup>10</sup>Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. <sup>11</sup>Parasites and Microbes, Wellcome Sanger Institute, Cambridge, UK. ✉email: [s.k.sheppard@bath.ac.uk](mailto:s.k.sheppard@bath.ac.uk)

A seemingly insatiable human appetite for poultry meat and eggs has resulted in modern livestock farming on a colossal scale. Today there are over 26 billion chickens worldwide, with poultry constituting around 70% of all bird biomass on earth<sup>1</sup>. Advances in selective breeding and husbandry have greatly increased productivity in the last 50 years but this level of agricultural intensification brings significant challenges for animal health, welfare and safe sustainable food production. Of particular concern are the opportunities created for the spread of livestock diseases and the emergence of zoonotic pathogens<sup>2,3</sup>.

Among the most common bacterial diseases of chickens reared for egg and meat production is colibacillosis<sup>4</sup> caused by avian pathogenic *Escherichia coli* (APEC). Like other forms of extra-intestinal pathogenic *E. coli* (ExPEC)<sup>5</sup>, APEC exists as a commensal component of the avian gut microbiota but emerges to cause a variety of systemic infections. Diseases of chickens and other birds range from epidermal, yolk sac and common respiratory tract (aerosacculitis) infections, to severe pericarditis, perihepatitis, omphalitis and septicemia<sup>6,7</sup>. In some cases mortality can reach 20%, condemning whole flocks, leading to suffering for millions of farmed birds and multimillion pound losses to the worldwide poultry industry<sup>4,6</sup>. The problem is exacerbated by the rise of antimicrobial resistance occurring across global transmission networks<sup>8–12</sup>, and the recognition that APEC may cause human infections<sup>13–16</sup> highlights the need to control this bacterium for both animal and human health.

Risk factors for colibacillosis have been identified, and include chicken immunological immaturity and stress<sup>17</sup>, but the disease has proved difficult to control, not least because no single gene, plasmid, phage or pathogenicity island has been exclusively associated with the emergence of virulent APEC from a background of harmless gut-dwelling *E. coli*<sup>18,19</sup>. Technical advances in high-throughput whole-genome sequencing offer opportunities to investigate the population genomics of pathogen evolution<sup>20</sup> but understanding the spread of APEC remains challenging for two reasons. First, there is uncertainty about the extent to which disease results from the transmission of a few globally distributed epidemic clones<sup>4,6,7</sup> or a diverse assemblage of disease-causing lineages<sup>21,22</sup>. Second, while a considerable body of knowledge has been gathered<sup>10,23–26</sup>, the genes contributing to APEC virulence are less well described than in human ExPEC pathotypes<sup>27,28</sup>. Pathogenicity is often linked to the presence of plasmids that confer a range of virulence-associated traits<sup>10,23–26</sup>, such as aerobactin production, complement resistance and iron acquisition<sup>7,10,22</sup>. However, no one gene is known to be essential for the development of extraintestinal infection in birds<sup>10,29–31</sup> and pathogenicity appears to be linked to a heterogeneous mix of plasmid and chromosomal genes involved in bacterial adhesion, invasion, toxicity, antibiotic resistance, survival and metabolism under stress<sup>22,31–33</sup>.

With colibacillosis set to increase in line with expanding poultry production there is a pressing need to monitor the emergence of APEC within genetically diverse commensal populations and identify strains that are predisposed to pathogenicity because of the genetic elements harboured in their genome. Here we take a large-scale comparative genomics approach to investigate the genetic basis of APEC pathogenicity that is agnostic to pre-existing assumptions about putative virulence determinants. Using a genome-wide association study (GWAS) approach<sup>34,35</sup>, we analyse 568 *E. coli* genomes from commercial poultry farms, including isolates from healthy chickens and those from various systemic infection body sites, and identify genes and genetic elements associated with avian pathogenicity (Fig. 1). Finally, having described an evolutionary context for understanding pathogen emergence, we use a machine learning approach to identify risk genotypes, that with further validation,

could form a basis of diagnostics and interventions to improve animal health.

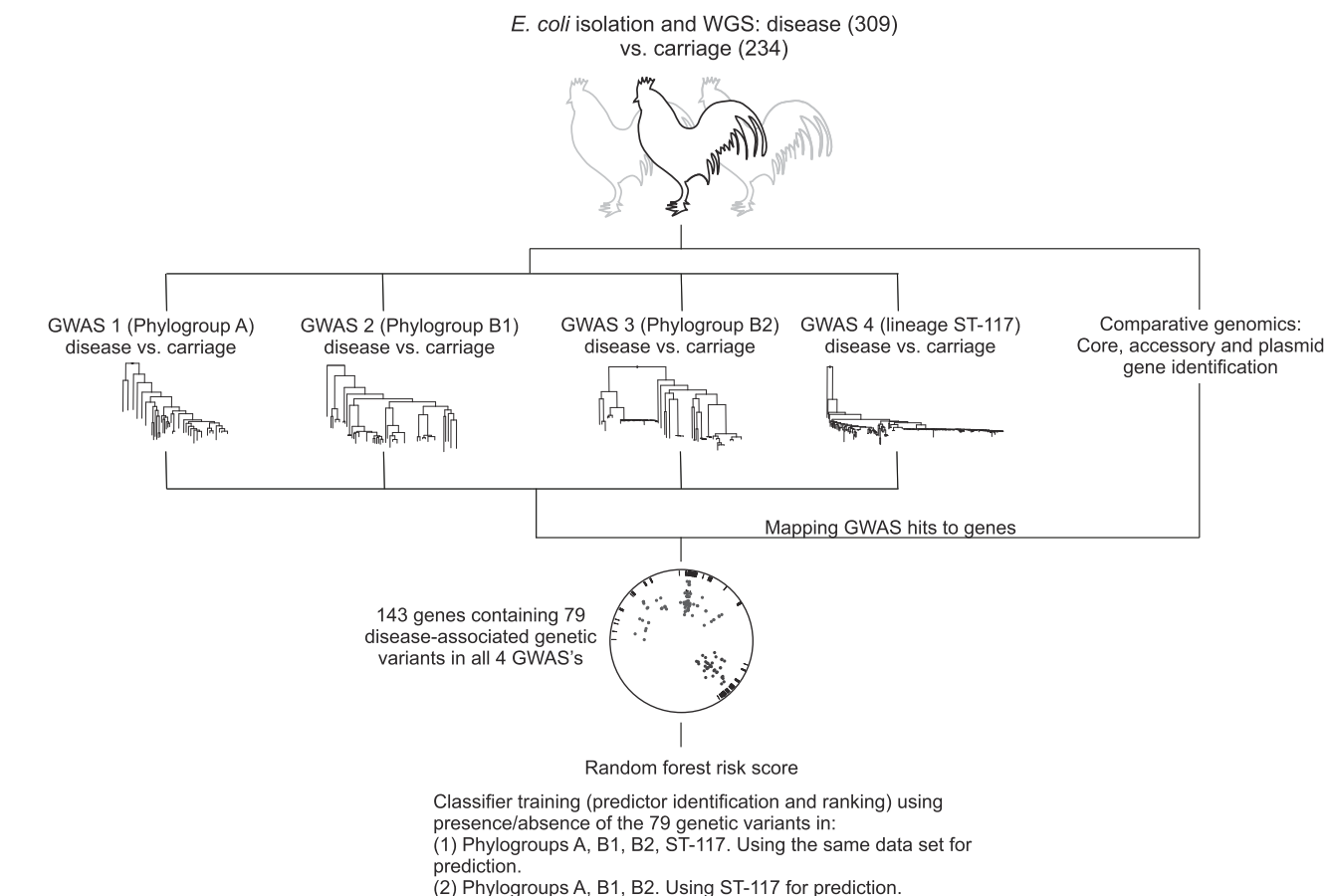
## Results

**Core and accessory genome variation in avian *E. coli*.** The pangenome of the 568 avian *E. coli* isolate dataset (309 disease-associated and 234 asymptomatic carriage strains) comprised 15,281 unique genes, with an average of 4115 genes per isolate. These included 3094 genes present in at least 95% of the dataset, which corresponded to 75% of the average genome size, consistent with previous *E. coli* core genome estimates<sup>28,36</sup>. The rate of accessory gene discovery did not plateau as the sampling increased (Supplementary Fig. 1), consistent with widespread acquisition of genes through horizontal gene transfer (HGT). While only 15.5% of all annotated genes from the reference avian *E. coli* strain APEC\_O1 were of unknown function, this number increased to 65.8% for the whole pangenome. All the assembled genomes analysed in this study are available via Figshare (<https://doi.org/10.6084/m9.figshare.12011811>) and raw sequence data has been deposited in the sequence read archive (SRA) associated with BioProject PRJNA592536.

**Quantitative analysis of APEC plasmid genes does not fully explain pathogenicity.** The emergence of APEC has been widely linked to the acquisition of plasmids containing virulence genes<sup>10,24–26,32</sup>. Therefore, we first quantified the presence of putative plasmid genes in isolates associated with disease and asymptomatic carriage using a gene-by-gene approach<sup>37,38</sup>. Putative plasmid genes were widely distributed among APEC and commensal *E. coli* strains in chickens and there was evidence that the average number of plasmid genes per isolate was greater among commensal strains (Supplementary Fig. 2). These results provide initial evidence that emergence of APEC virulence is not entirely dependent on the presence of specific defined plasmids, as in some *E. coli* pathotypes<sup>39</sup>. In fact, rather than a pattern of complete plasmid (as detailed in the reference sequences) presence/absence, there was evidence for mosaicism of plasmid genes found together in different combinations. Furthermore, these analyses do not discriminate the context of putative plasmid-associated genes that may be plasmid-borne or integrated in the chromosome. One explanation for the high numbers of putative plasmid genes harboured among commensal isolates is that some contribute to avian adaptation rather than *sensu stricto* virulence. For example, plasmid-borne antimicrobial resistance genes<sup>40</sup> may promote persistence in intensive livestock systems, where antimicrobials may have been used for prophylaxis or treatment, but are not directly associated with invasive disease. However, while some putative plasmid genes were more common in APEC compared to other *E. coli*, there was little evidence of complete segregation that would be indicative of direct causation (Fig. 2 and Supplementary Fig. 2). These findings suggest that a full understanding of the genetic determinants of APEC emergence requires consideration of homologous sequence variation rather than simple plasmid gene presence/absence analysis.

## Avian pathogenic strains emerge from multiple *E. coli* lineages.

A maximum-likelihood phylogeny constructed from a concatenated gene-by-gene core genome alignment (3,094 genes) revealed a highly structured population (Fig. 3a). Inclusion of isolates from the well-described ECOR collection allowed contextualisation of avian isolates among known *E. coli* phylogroups and multi-locus sequence types (STs)<sup>41,42</sup>. Poultry isolates from our dataset (Supplementary Data 1) clustered within six out of the eight known *E. coli* phylogroups (A, B1, B2, D, E, F)<sup>43</sup>, and were absent in phylogroups C and G<sup>44,45</sup>. There was evidence for



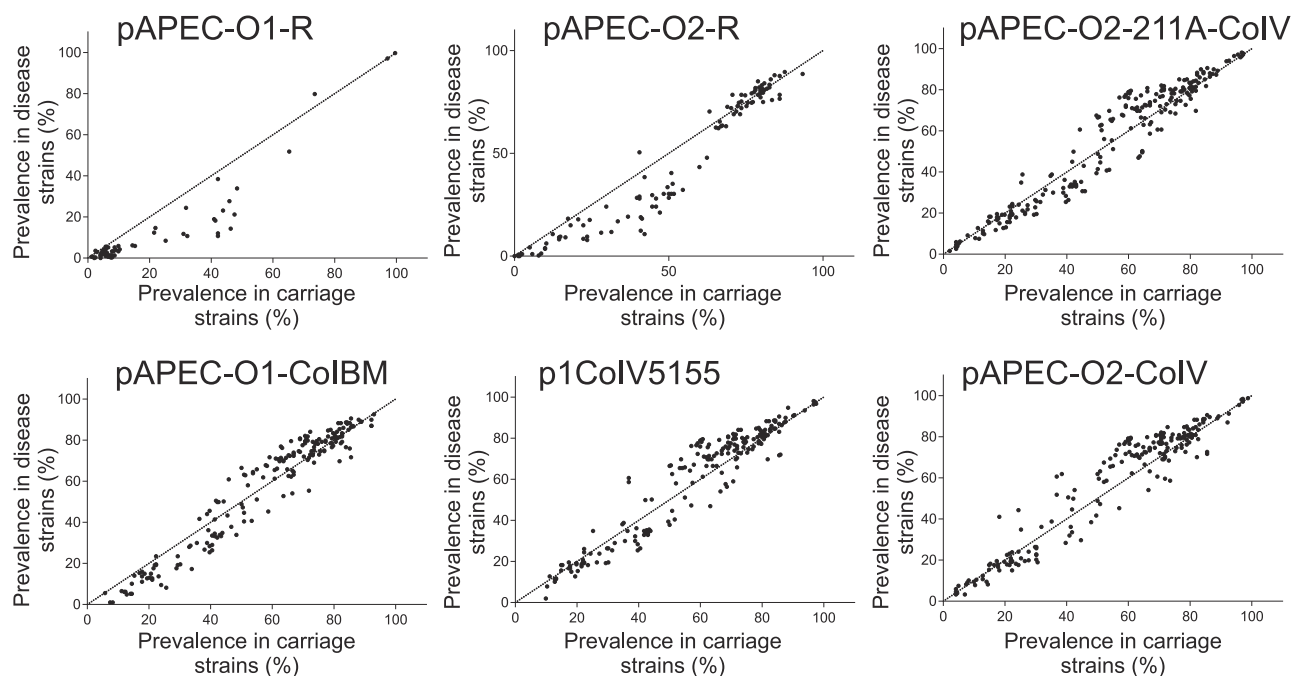
**Fig. 1 Avian Pathogenic *E. coli* (APEC) GWAS and risk prediction.** Genome-wide association studies (GWAS) can identify multiple genetic variants associated with complex traits but these can be difficult to interpret. For example, pathogenicity is a multifactorial phenotype, potentially involving genes that affect phenotypes like toxicity, antimicrobial resistance, immune evasion etc. Furthermore, the role of certain genes may be poorly defined, especially in bacteria with large accessory genomes. We developed a method in which 4 GWAS experiments (carriage vs. disease isolates) were conducted and the disease-associated genetic variants (core genome SNPs, accessory genes, fission/fusion, duplications and accessory gene alleles) were mapped to genes within the pan genome. Disease-associated elements identified in all four lineage-specific GWAS (phylogroups A, B1, B2 and ST-117) included 143 genes, containing 79 species-wide genetic variants. Patterns of presence and absence of these variants were used as classifiers in two different random forest models to identify the best predictors of APEC disease.

variation in the distribution of poultry strains among lineages. For example, 211 isolates (39%) belonged to a single sequence type (ST-117) which, together with isolates in the B2, B1 and A phylogroups, constituted 93% of the poultry isolates in our dataset. Of the remaining isolates, the most common STs were: ST-1618, ST-95, ST-919 and ST-429 (phylogroup B2); ST-101, ST-155 and ST-469 (phylogroup B1); ST-38 and ST-69 (phylogroup D); and ST-10 (phylogroup A). Phylogroups E and F were exclusively populated with ST-350 and ST-648 strains respectively. While variation in the frequency of poultry isolates in different lineages may reflect natural abundance within host populations, this does not necessarily indicate pathogenicity. To assess this, the ratio of invasive to commensal strains in the common lineages was determined for each common phylogroup and ranged from 61% (B2) to 31% (E). While it remains possible that lineages with enhanced pathogenicity may exist or emerge in the future, among known *E. coli* diversity, isolates from all major phylogenetic groups were represented in both the asymptomatic and the disease isolate collections. This reflects the emergence of pathogenic clones from multiple genetic backgrounds.

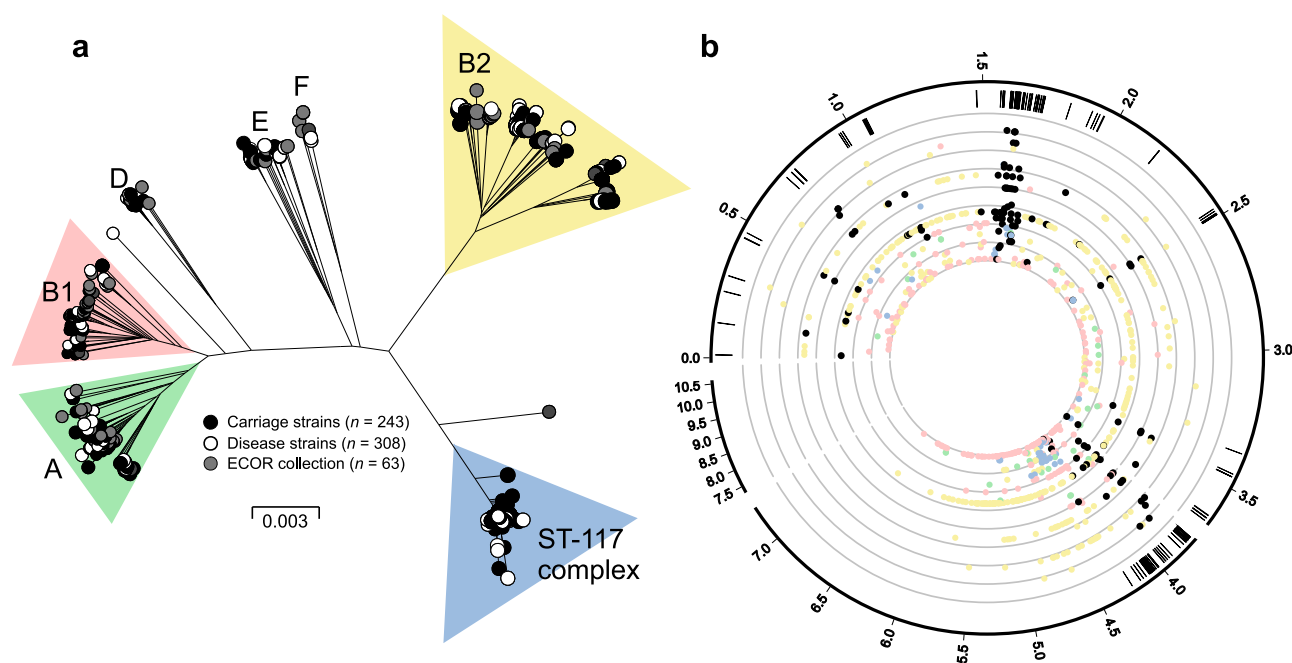
### Pangenome-wide association study reveals pathogenicity-associated genes.

GWAS was performed for each of the four

most common lineages in the dataset, namely on phylogroups A, B1, B2 and ST-117. In each case, disease isolates were compared to those from asymptomatic carriage within the same phylogroup. The GWAS approach incorporated a ClonalFrameML phylogeny that accounts for the impact of recombination, thereby reducing the effect of population structure and maximising the chance of identifying elements associated with a switch from commensal to pathogenic lifestyle. These independent GWAS analyses identified 11,947, 15,670, 43,980 and 7110 associated genetic elements with a  $p$ -value  $< 0.05$ , that mapped to 1925, 2099, 3946 and 554 infection-associated genes in phylogroup A, B1, B2 and the ST-117 lineage respectively. Nonetheless, only 896 genes contained associated genetic elements with  $p$ -value  $< 0.01$  (dots on Fig. 3b) and only these were considered for further downstream analysis. Out of the 896 pathogenicity-associated genes, 753 were phylogroup- or lineage-specific, suggesting multiple independent pathways to pathogenicity (Supplementary Fig. 3), with some variation in prevalence based upon extra-intestinal isolation source (Supplementary Fig. 5a/Supplementary Data 7). However, 143 genes were flagged as pathogenicity-associated in all four GWAS analyses (Supplementary Data 2 and Supplementary Fig. 3). Of these, 65 were core genes (45.5%) and 78 accessory genes (54.5%), and had diverse predicted functions, including genes involved in metabolism, lipopolysaccharide

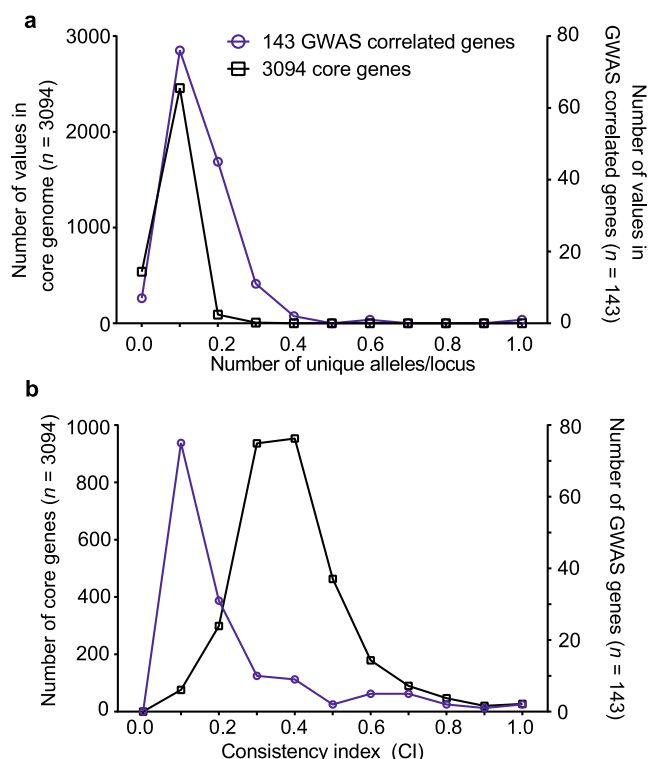


**Fig. 2 Segregation of genes in six known APEC plasmids among carriage and disease isolates.** For every gene (represented with a dot) within six known APEC plasmids, we calculated the prevalence (%) among carriage ( $n = 234$ ) and disease-associated ( $n = 309$ ) *E. coli* genomes. The dashed line represents equal prevalence in each population. Source data are provided as a Source Data file.



**Fig. 3 Population structure and genome-wide association study of avian *E. coli*.** **a** Phylogenetic tree of 568 avian *E. coli* strains, reconstructed using a maximum-likelihood algorithm (IQ-TREE) from a core genome alignment ( $n = 3094$  genes shared at least by 95% of all the isolates). Isolates are labelled according to source: disease isolates (white); carriage isolates (black); ECOL collection (grey). The letters designate the different *E. coli* phylogroups. **b** Pangenomic position of GWAS results. The outer ring represents the pangenomic position of genes in the *E. coli* reference strain APEC\_O1, the rest of the pangenome inferred in this study and a group of low frequency accessory genes that were excluded from the GWAS. Black ticks in the second ring show the position of genes containing disease-associated genetic variants in all four distinct GWAS. Coloured circles are shown for the most statistically associated (lowest  $p$ -value) elements in a given gene for GWAS in phylogroups A (green), B1 (pink), B2 (yellow), ST-117 complex (blue) and species-wide disease-associated elements (black). The threshold for significance was  $p$ -value = 0.01 (inner circle). Statistical significance was determined by the treeWAS algorithm. Concentric rings emanating from this threshold correspond to incremental reductions to a  $p$ -value of 0.000001 (outer ring). The numbers in the outer ring denote the length of the pangenome in Mbp. Source data are provided as a Source Data file.





**Fig. 4 Comparison of allelic variation and consistency index for core genes and genes containing disease-associated elements.** **a** The average number of alleles per locus and **b** consistency indices to a core phylogeny, were calculated for each gene alignment for core genes and 143 genes containing pathogenicity-associated elements using R and the phangorn package. The left y-axis indicates the number of core genes (black line), the right y-axis indicates the number of genes containing pathogenicity-associated elements (blue line). For the consistency index, the two distributions were significantly different (two-tailed Mann-Whitney test;  $p$ -value = 0.0001, Mann-Whitney  $U = 11,366$ ). Source data are provided as a Source Data file.

synthesis, heat shock response, antimicrobial resistance and toxicity. A total of 58 (74.4%) of the accessory genes were putatively of plasmid or phage origin (Supplementary Data 2). Finally, within the 143 species-wide pathogenicity-associated genes, we identified 79 genetic elements that segregated by disease/carriage ( $p$ -value < 0.05) in all the GWAS, including 66 core genome SNPs, 3 accessory genes, 1 fission/fusion, 4 duplications and 5 accessory gene alleles (Supplementary Data 3). These stringently defined elements constitute robust candidates for disease risk prediction.

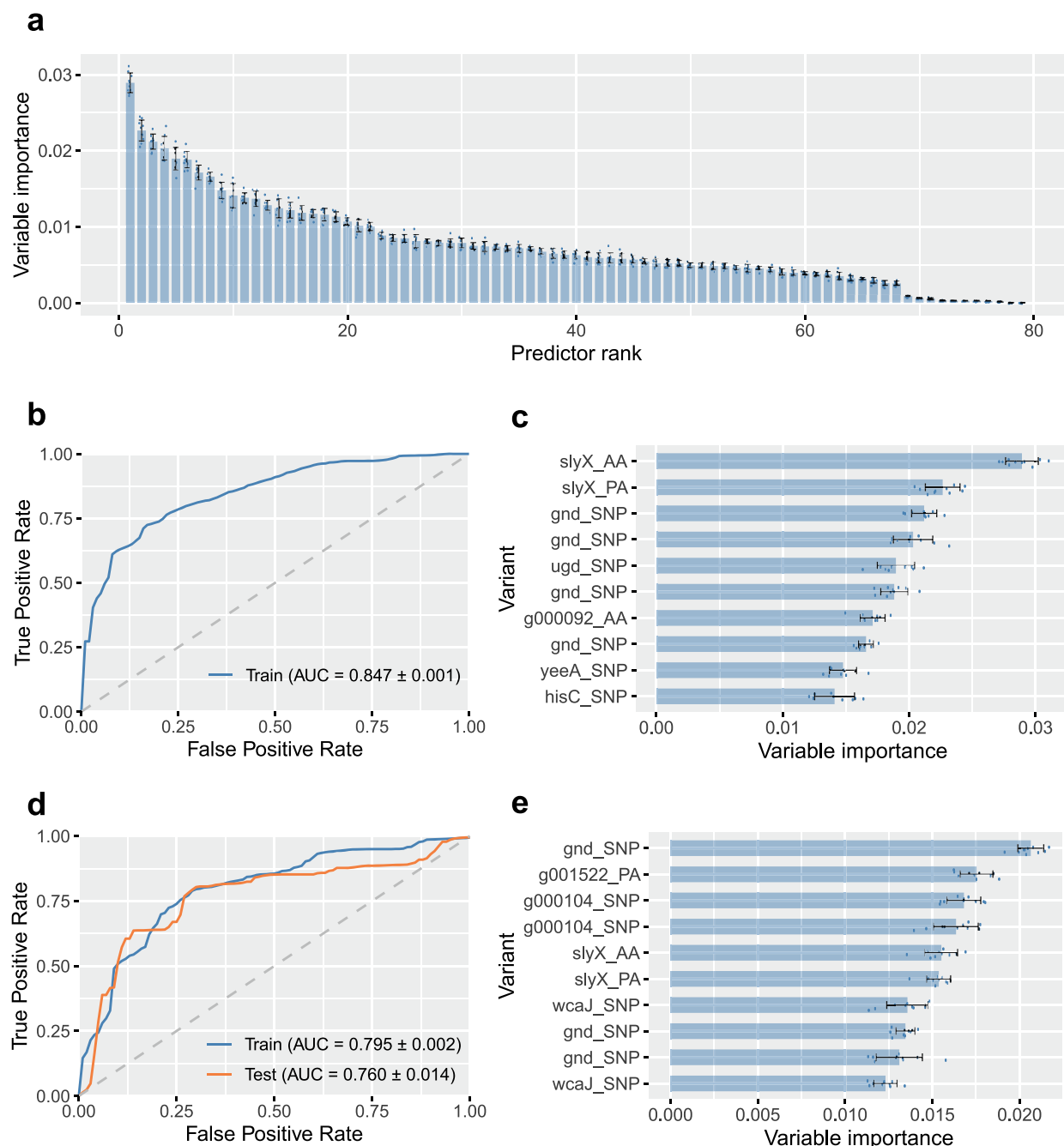
**Pathogenicity-associated genes recombine among *E. coli* lineages.** The acquisition of pathogenicity-associated elements among divergent lineages and the importance of potentially mobile elements (plasmid and phage genes) suggest a role for HGT in the emergence of avian pathogenicity. There was a significant increase (Mann-Whitney test;  $U = 80516$ ,  $p$ -value < 0.0001) in allelic variation among genes associated with pathogenicity (Fig. 4a), with an average of  $0.156 (\pm 0.106)$  unique alleles per locus for the 143 pathogenicity-associated genes, and  $0.08 (\pm 0.035)$  for 3094 core genes. This could be due to the accumulation of deleterious mutations resulting from *E. coli* range expansion into the pathogenic niche<sup>46</sup> but an equally likely explanation is that these substitutions result from elevated recombination among pathogenicity-associated genes. Evidence for this comes from calculation of the mean consistency index (CI) that was significantly lower

(Mann-Whitney test;  $U = 11,366$ ,  $p$ -value < 0.0001) among pathogenicity-associated genes ( $0.2226 \pm 0.2037$ ) compared with other core genes ( $0.3895 \pm 0.01186$ ; Fig. 4b). This suggests that the clonal mode of descent is disrupted in pathogenicity-associated genes consistent with elevated HGT.

**Machine learning identifies disease risk genotypes.** Quantitative determination of species-wide pathogenicity risk markers was carried out using a Random Forest (RF) classifier approach based on the presence/absence of the 79 genomic variants, that were associated with disease isolates in all four lineage-specific GWAS analyses (Supplementary Data 3). Using disease-associated elements found in all major phylogroups maximised the likelihood of capturing generalised predictors of APEC pathogenesis and limited the possible lineage specific effects. The estimated risk score was defined as the probability of an isolate coming from disease given a certain profile of the 79 genetic elements. The relative predictive power of each of these elements was estimated by ranking them according to their estimated importance as classifiers in the model (Fig. 5a). Next, the diagnostic ability of the classifier system at varying discrimination thresholds (receiver operating characteristic - ROC curves) and the importance of the 10 highest ranked predictors were investigated in two analyses (Fig. 5b–e). In the first analysis, in which the training data contained all isolates from the four phylogroups, the RF model reached an out-of-sample classification accuracy of 76.9% for predicting infection status of *E. coli* strains (healthy carriage vs. disease). SNPs within the *gnd* gene, involved in inter-strain transfer and recombination<sup>47</sup>, accounted for 4 of the 10 most important predictors. These 10 predictors achieved a classification accuracy of 73.5% on their own, potentially offering simple targets for investigation of *E. coli* pathogenicity risk.

In a second analysis, we tested the ability of the model trained on data from phylogroup A, B1 and B2 isolates, to predict infection status (healthy carriage vs. disease) in the emergent ST-117 lineage that is thought to be virulent in birds and hold zoonotic potential<sup>48–50</sup>. Replicate analyses with ST-117 isolates included (‘Train’) and excluded (‘Test’) from the training data returned seven of the same top ten ranked predictors as in the first analysis (Fig. 5c, e) and gave average RF out-of-sample accuracy of 75% and 73% respectively. The slight reduction in the accuracy when moving outside the training domain of the RF model (area under the ROC 0.79 to 0.76, Fig. 5d), indicated that the model may generalise to other *E. coli* data using existing predictors. Achieving this level of prediction accuracy has clear potential for the development of pathogenicity biomarkers in a farm setting, particularly as the power of the model is limited by the input data. Specifically, as samples from asymptomatic chicken may include strains that have the potential to cause future infection (as well as those that do not), the commensal strain training dataset likely includes some pathogenicity elements. While it is difficult to predict which strains these are, especially as host factors may influence infection, broader sampling may increase the numbers of representative commensal strains that do not have the potential to cause disease, and elevate prediction accuracy beyond 75–77%.

**Prevalence of APEC-associated genetic variants in *E. coli* isolates from different infection sites and other host sources.** The prevalence of APEC-associated variants, used in the RF model, was investigated in isolates sampled from other host niches. Specifically, we analysed the *E. coli* reference collection (ECOR)<sup>42</sup>, 175 human ExPEC strains<sup>28</sup>, 14 disease-associated strains from dogs<sup>51</sup> and 521 strains from healthy cattle<sup>52</sup>. The APEC-associated genetic variants also occurred in humans and other

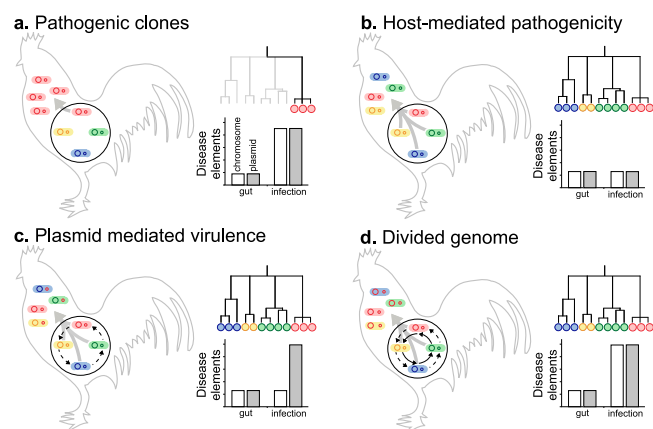


**Fig. 5 Identification of predictive genotypes for pathogenicity in avian *E. coli* using random forest (RF) models.** **a** The importance of the predictors derived from the four GWAS using the primary classifier model (trained using data from the four lineages A, B1, B2, ST-117); **b** Receiver operating characteristic (ROC) curve showing the overall performance of the primary classifier model; **c** Importance of the top 10 predictors in the primary classifier model; **d** ROC curve showing the overall performance of the follow-up classifier model (trained using data from the four phylogroups A, B1, B2 and predicting in ST-117); **e** The importance of the top 10 predictors in the follow-up classifier model. Data are presented as mean values  $\pm$  SD from  $n = 10$  repeated analyses. Source data are provided as a Source Data file.

animals implying that host species does not constitute a complete gene-pool barrier between the niches. A Kruskal–Wallis test identified significant differences in the prevalence of genetic markers in human ExPEC compared to the ECOL collection ( $p$ -value  $< 0.0001$ ) and healthy bovine isolates ( $p$ -value  $< 0.0001$ ), indicating that APEC-associated genetic variants can be found in other animals but overall they are significantly less abundant in human ExPEC strains (Supplementary Fig. 5b). However, analysing the individual prevalence of specific APEC-associated genetic variants in each of these additional *E. coli* groups

(Supplementary Data 6) revealed that many of the APEC-associated elements were also common among human ExPEC (Supplementary Fig. 5c/Supplementary Data 6). This may suggest shared adaptations to establishing extraintestinal infection in both avian and human hosts<sup>13</sup>.

Additionally, we investigated the prevalence of the same genetic variants in isolates from different infection sites within our dataset (Supplementary Fig. 5a/Supplementary Data 7). A Kruskal–Wallis test revealed significant differences between asymptomatic carriage and bone marrow samples



**Fig. 6 Evolutionary scenarios for APEC infection and predicted variation in isolate phylogenies and disease-associated elements.** Panels summarise models for the spread of *E. coli* from the primary commensal gut niche (black circle) to extraintestinal tissue, and the effect on the population of *E. coli* clones (blue, red, green and yellow circles) and their genomes (internal circles) which may be enriched for putative pathogenicity-associated genes (red). Conceptual genealogies are given for isolates sampled from extraintestinal disease sites and the estimated prevalence of chromosomal (white) and plasmid (grey) disease determinants in the genome of isolates from the gut and systemic infection are shown. Evolutionary scenarios include: **a** extraintestinal spread of pathogenic clones with genomes enriched for disease elements, seen as one (or few) lineages on the tree; **b** host-mediated pathogenicity in which multiple diverse clones spread systemically as a result of host factors, irrespective of disease elements (no significant difference); **c** plasmids transfer between lineages (dashed black arrow) and clones harbouring plasmids spread extraintestinally; **d** Horizontal gene transfer reassorts plasmid and chromosomal disease elements into multiple genomic backgrounds leading to the emergence of multiple APEC clones.

( $p$ -value < 0.0001), between asymptomatic carriage and liver isolates ( $p$ -value = 0.0003) and between bone marrow and heart isolates ( $p$ -value = 0.009). While prevalence is distinct from GWAS-association, this suggests differences between infection types and provides evidence that the elements that underly pathogenicity may vary with different pathologies.

## Discussion

*E. coli* are simultaneously ubiquitous in healthy animal guts and a major cause of diverse intestinal and extraintestinal infections. Clearly, interpreting these contrasting lifestyles requires an understanding of the factors that promote pathogenicity in this, typically commensal, bacterium. By definition, extraintestinal disease requires migration from the gut and proliferation within the pathogenic niche. While host factors and dysbiosis may be important for this<sup>27</sup>, disease also depends on the ability of the invasive strains to colonise a new habitat where the conditions are different.

Sampling *E. coli* from both healthy chicken guts and infected systemic tissues (APEC) allowed characterisation of the genotypes and gene pools associated with different sites, and health states. It is therefore possible to consider different evolutionary scenarios for the migration of pathogenic strains from the commensal gut niche and proliferation in extraintestinal tissues (Fig. 6). First, the emergence of dominant pathogenic APEC clones from the background gut population. Second, host-mediated pathogenicity, where all poultry-associated *E. coli* are able to cause extraintestinal infection. Third, plasmid-mediated virulence in which multiple lineages proliferate extraintestinally as a result of the acquisition of specific plasmid-borne virulence genes. Finally, a

divided genome scenario<sup>53</sup>, where HGT introduces disease-associated chromosomal and plasmid genes into multiple genetic backgrounds allowing colonisation of the extraintestinal niche.

It is known that certain *E. coli* lineages are predisposed to extraintestinal pathogenicity, as evidenced by the global spread of pandemic ExPEC clones<sup>5</sup>. In a simple infection model, where only dominant clones can cause disease, all isolates recovered from infected tissues will belong to discrete clusters of genetically related pathogenicity-associated strains (Fig. 6a). This was not observed among APEC isolates. In fact, disease isolates were distributed across the phylogeny within all six previously described phylogroups and the ST-117 complex lineage (Fig. 3a). One interpretation of this genetic structuring is that given particular host factors, such as gut perturbation or dysbiosis, all *E. coli* lineages are equally able to cause disease by mass action rather than specific pathogenicity<sup>5</sup> (Fig. 6a). If this were the case, then genome analyses would not identify enrichment of pathogenicity-associated elements within the genome of disease strains. However, the GWAS identified numerous pathogenicity-associated elements which mapped to genes known to be associated with pathogenicity. This is consistent with enrichment for sequence that encodes traits associated with pathogenicity.

Plasmid carriage is an important factor in the emergence of *E. coli* pathotypes<sup>39</sup>. The mobility of these elements makes them ideal candidates for spreading pathogenicity genes among APEC, potentially conferring multiple virulence phenotypes in a single evolutionary step. This could explain the emergence of pathogenic strains from divergent genetic backgrounds (Fig. 6c). However, our population-scale analysis revealed that known APEC plasmid genes were no more abundant in disease compared to commensal isolates from poultry (Supplementary Fig. 2). Furthermore, rather than containing a discrete compendium of genes associated with a given plasmid, genes were present at varying frequencies suggesting a mosaic of putative plasmid elements found in different combinations in the genome (Fig. 2 and Supplementary Fig. 2). While these findings are inconsistent with a simple model of plasmid-mediated virulence (Fig. 6c), the ubiquity of plasmid genes implies a role in poultry adaptation or the emergence of APEC as described in various studies<sup>32,40,54</sup>.

Investigating the genetic basis of pathogenicity beyond the role of plasmids requires consideration of the putative function of all core and accessory genes. Pathogenicity is multifactorial and there is evidence that different traits can contribute in different *E. coli* lineages (Supplementary Fig. 3). Population-wide genomic screening and analysis approaches, such as GWAS and machine learning, deliver a deluge of potentially useful information describing the genetic basis of complex traits. If correctly integrated with laboratory microbiology, this can underpin rigorous confirmatory tests of gene function that satisfy Molecular Koch's postulates<sup>55</sup>. While functional genomic genotype-phenotype maps are required for a full understanding of pathogenicity, evidence for common APEC disease determinants comes from genes that are enriched in APEC strains from all *E. coli* phylogroups – consistent with convergent evolution of pathogenicity traits. Species-wide pathogenicity-associated genes included those associated with generic Gram-negative virulence factors, such as O-antigen chain length regulation (*wzzB*<sup>56</sup>) and a host killing toxins (*hokA* and *hokC*<sup>57</sup>), as well as known avian virulence factors including outer membrane proteins (*ompT*<sup>58,59</sup>), pilus chaperones (*papD* and *fimD*<sup>60–62</sup>), cell envelope integrity (*wcaJ*), and general secretory pathways (*gspO*) responsible for transport of toxins from the periplasm to the extracellular medium<sup>63</sup>. Antimicrobial resistance (*YeeO*<sup>64</sup> and *evgS*<sup>65</sup>) and central metabolism (*gnd*<sup>47</sup>, *glts*<sup>66</sup> and *hisBCDGH*<sup>67</sup>) genes also contained pathogenicity-associated elements, potentially conferring a survival advantage in the stress conditions of the infection niche<sup>68</sup>.

Bacterial GWAS relies on whole-genome sequencing rather than standardised genotyping arrays and therefore typically has lower sample size than human GWAS (often >100,000 samples<sup>69</sup>). This can impact on the statistical ability to detect rare associated variants, but also on the ability to generalise results outside of the sampled dataset. Because of the lack of standardisation of bacterial GWAS results<sup>70</sup>, and the scarcity of other comparable bacterial GWAS results, replication meta-analyses can be extremely challenging. Nevertheless, as for human GWAS, robustly associated variants in one study can always potentially highlight important functions and mechanisms associated with a trait<sup>71,72</sup>, here pathogenicity in APEC. Therefore, any comparison with previously described virulence determinants can provide useful confirmatory evidence of a potential role for APEC associated genetic variation.

Cross-referencing with a recent authoritative review<sup>73</sup>, identified GWAS associations in genes that have been linked to pathogenicity in other studies. For example, *eae* - an attaching and effacing gene that encodes intimin, and *ompT* - that encodes a protease able to cleave colicin. However, there was little overlap with some other studies<sup>30</sup>. It is inevitable that the findings of our population-wide approach will not exactly match data from existing microbiology studies for three important reasons. First, pathogenicity-associated elements are ranked based upon a significance score. With experimental design targeting ubiquitous APEC genomic signatures, this will inevitably flag pathogenicity elements found in multiple lineages and pathologies. Therefore, variation associated with a specific infection type may have lower significance. For example, while the *hokA* and *hokC* genes encoding components of toxin-antitoxin systems are not APEC virulence determinants in the strict sense, they may indicate the importance of wide-spread mobile genetic elements that are linked to virulence genes that vary by infection type. Second, many of the most significant hits were SNPs in core genes that may be linked to other genes (Supplementary Data 3). For example, while the chaperone-encoding genes *papD* and *fimD* are among the pathogenicity-associated genes, other P and type 1 fimbrial determinants are not, despite their likely role in virulence<sup>74</sup>. The reason for this is that the variation within the genes encoding fimbrial determinants does not segregate as strongly as the chaperone genes based on the binary pathogenicity phenotype in this study. This suggests that multiple homologous sequence variations within these genes can underlie pathogenicity when particular SNPs are present in the chaperone genes. Finally, factors promoting host colonisation, such as adhesion, can benefit commensal strains as well as representing a step towards bacterial pathogenicity in ExPEC<sup>75-77</sup>. In this case, while there may be a specific role in extraintestinal spread, the underlying genomic signature may not achieve statistical significance. For these reasons, the population-wide GWAS approach can be considered a platform on which to develop further functional genomic characterisation rather than a definitive road-map to pathogenicity, highlighting segregating variation in genes, such as *fimD*, that are linked to operons or pathways that are known to relate to *E. coli* virulence<sup>61,73,74</sup>.

The enrichment of putative virulence determinants among disease isolates suggests that pathogenic strains are a subset of the commensal gut population that possess genetic elements that may promote migration and colonisation of extraintestinal sites (Fig. 6d). This presents something of a theoretical conundrum. Specifically, within the competitive milieu of the gut microbiota (primary niche), genes that are only beneficial in the secondary (extraintestinal) niche will impose a fitness cost on the bacterium. Therefore, strains with many pathogenicity adaptations will be less competitive than those with few. A clue to explaining this comes from the observation that 74% of the

pathogenicity-associated accessory genes were putatively of plasmid or phage origin and are therefore mobile among lineages. HGT is known to be a major force in *E. coli* evolution<sup>78,79</sup>, with the acquisition of genes through recombination potentially conferring adaptations associated with pathogenicity<sup>80</sup>. Therefore, virulence genes that are maintained at low frequencies in different lineages in the primary niche can be reassorted and come together in a common genetic background, potentially allowing successful extraintestinal colonisation.

Evidence of the importance of HGT in APEC comes from the divergent position of strains across the population phylogeny (Fig. 3a), as well as the lower mean consistency index of individual pathogenicity-associated gene trees compared with core genes. This suggests homoplasmy and the horizontal spread of adaptive genes through the population, consistent with a gene-specific selective sweep, or divided genome, model of bacterial evolution<sup>53,81,82</sup>. In this scenario, as migration from the gut occurs, HGT will increase the rate at which positively selected genes sweep through the invasive population. The speed and efficiency of adaptation are therefore increased by recombination combining multiple selected plasmid and chromosomal genes into a common genomic background (Fig. 6d). For diverse commensal *E. coli* populations this can potentially promote the emergence of pathogens at the boundary between commensal and extraintestinal niches and rapid adaptation to life in the extraintestinal environment. Furthermore, studies of in vivo bacterial populations have demonstrated recombination among strains within the gut of humans and chickens<sup>83,84</sup>. Knowing that *E. coli* can be found at high concentrations in chickens (mean log<sub>10</sub>*E. coli* of 4.15 colony forming units per ml of faeces<sup>85</sup>), with an estimated doubling time of around 3 to 15 h in natural populations<sup>86,87</sup>, it is also possible that in chronic APEC infections, lasting more than 2 days from experimental infection to the death of the bird<sup>88</sup>, virulence determinants could accumulate and reassort among strains.

In an animal health setting, early identification of pathogens has great potential to improve livestock welfare and reduce economic losses resulting from disease. It is evident that targeting individual clones based upon traditional molecular typing methods<sup>89</sup> has limitations because the putative virulence determinants are mobile between lineages. It follows, therefore, that quantifying pathogenicity-associated genes may allow the identification of carriage strains that pose a disease risk. However, this is complicated by two factors. First, there is evidence of lineage-specific and species-wide pathogenicity-associated genes so simple diagnosis may be challenging. Second, extraintestinal spread and systemic infection involves numerous colonisation and virulence factors that may be associated with progression of different types of infection, such as common respiratory tract infections and septicaemia. Statistically significant GWAS correlation with different infection types would require larger numbers of samples from each extraintestinal source but varying prevalence of pathogenicity-associated elements among isolation sources was consistent with multiple pathways to infection.

To achieve more accurate risk prediction of this complex disease, we developed a Random Forest machine learning approach to quantify the power of different combinations of pathogenicity-associated elements (classifiers) to predict the source of isolates (carriage/disease) from their genome. A simple analysis, in which all isolates were used to train the model, achieved a classification accuracy of 76.9%. Among the classifiers that provided the most accurate prediction were elements in a gene associated with polymyxin resistance (*ugd*<sup>90</sup>) and a gene participating in the oxidative pentose phosphate metabolic pathway (*gnd*<sup>47,91</sup>) located in the highly recombinant *rfb* region associated with the avoidance of host defence systems<sup>92</sup>. This may be explained by the use



of polymyxin in the treatment of colibacillosis in poultry production<sup>93</sup> and HGT among strains, and provides clues about the functional basis of the pathogenicity-associated elements. However, as APEC emerges in multiple *E. coli* lineages, generalising the risk prediction method required that the training and test datasets were phylogenetically distinct. Focussing on the ST-117 lineage, that is thought to be an emergent bird pathogen<sup>49</sup> achieved a risk prediction accuracy of 72.7% when the model was trained on phylogroups A, B1 and B2. This suggests that relatively few 'global' pathogenicity markers may provide a basis for risk prediction and that model training on ever larger reference genome datasets may have potential for early identification emergent APEC in healthy flocks to inform targeted interventions.

Pathogens remain a major threat to sustainable livestock production. Control of highly pathogenic avian influenza is typically achieved through early diagnosis, flock isolation and bird culling. However, these measures are not applicable for some common diseases, such as colibacillosis, because *E. coli* are found in all chicken guts and a full understanding of the factors responsible for the emergence of APEC has remained elusive. Population-scale comparative genomic analyses take us a step closer to characterising the differences between commensal gut *E. coli* and APEC that have acquired genetic elements that promote extraintestinal infection. Untangling the web of interacting genes that underly pathogenicity is extremely difficult without full functional genomic characterisation and it is inevitable that some lineage- or pathology-specific genetic variation is missed when targeting species-wide markers of infection. However, there is clear utility for the development of APEC molecular diagnostics and targeted antibiotic therapy - where pathogens have a different resistance profile<sup>94,95</sup>. Among the most promising applications are those involving the development of guided antimicrobials that can selectively target a gene, cellular process, or strain of choice. Where the pathogen sequence is known, nucleic acid-based antibacterials, peptides, bacteriophage therapies, bacteriocins, and anti-virulence compounds may be able to exclusively target disease-causing bacteria<sup>94,96</sup>. For example, CRISPR-Cas technology can be used for sequence-specific bacterial killing through guided nucleases that recognise known DNA sequences<sup>97,98</sup>. Despite the complexity of pathogenicity phenotypes, and the mobility of the underlying genes, these techniques show considerable potential. With further functional genomic validation, a firm understanding of the genetics of pathogenicity could pave the way to early diagnosis of risk, more effective control and improved animal welfare.

## Methods

**Bacterial sampling.** The isolate collection comprised 568 avian *E. coli* isolates sampled from a variety of sources (Supplementary Data 1). These included 152 previously published isolates<sup>50,58,99</sup>, 414 isolates sequenced as part of this study and the reference strains APEC01<sup>100</sup> and APEC078<sup>101</sup>. In total, 482 isolates were sampled at slaughter from broiler chicken (*Gallus gallus domesticus*), 44 isolates were from commercial turkey (*Meleagris sp.*), 12 isolates were from avian wildfowl and 5 isolates were from gulls. A total of 234 faecal isolates were from asymptomatic carriage, sampled post mortem from the gut when no symptoms could be observed in the bird, and 309 isolates were considered disease-associated APEC based on their extraintestinal site of isolation, and/or symptoms in the bird at autopsy. Specifically, out of the disease-associated isolates 35 were isolated from the bone marrow, 70 from the liver (perihepatitis), 23 from the heart (pericarditis), 40 from the peritoneum (peritonitis), 26 from blood (septicaemia), 19 from yolk sac infections and 96 from various other infection sites. Finally, 25 isolates were not phenotypically characterised as they were isolated from the poultry farm environment. Isolates from the poultry farm environment were only used for the phylogenetic and pangenome analysis of this study.

**Genomic DNA extraction, sequencing and archiving.** DNA was extracted using the QIAamp DNA Mini Kit (QIAGEN; cat. number: 51306), using

manufacturer's instructions. DNA was quantified using a Nanodrop spectrophotometer, as well as the Quant-iT DNA Assay Kit (Life Technologies, Paisley, UK). High-throughput genome sequencing was performed on a MiSeq (Illumina, San Diego, CA, USA), using the Nextera XT Library Preparation Kit with standard protocols. Libraries were sequenced using 2 × 250 bp paired end v3 reagent kit (Illumina), following manufacturer's protocols. Short read paired-end data was assembled using the de novo assembly algorithm, SPAdes (version 3.10.0)<sup>102</sup>. The average number of contigs was 336 (range: 11–1373) for an average total assembled sequence size of 5.16 Mbp (range: 4.42–5.79). All 414 genomes sequenced in this study have been deposited in GenBank, associated with BioProject PRJNA592536. Accession numbers for all genomes, including those previously sequenced can be found in Supplementary Data 1. Genome assemblies for the entire collection 568 can be downloaded from figshare: <https://doi.org/10.6084/m9.figshare.12011811>. An overview of the assembly information is provided on Supplementary Data 4.

**Core and accessory genome characterisation.** All unique genes present in at least one isolate (the pangenome) were identified by automated annotation using PROKKA (version 1.13)<sup>103</sup> and PIRATE<sup>104</sup> - a pangenomics tool which allows for orthologous gene clustering in divergent bacterial species. We defined genes in PIRATE using a wide range of amino acid percentage sequence identity thresholds for Markov Cluster algorithm (MCL) clustering (45, 50, 60, 70, 80, 90, 95, 98). Additional APEC reference genomes and APEC associated plasmids were included in the pangenome to maintain locus nomenclature and identify plasmid carriage, including APEC\_O1 (accession: GCA\_000014845.1)[6], APEC\_O78 (accession: GCF\_000332755.1)[26], APEC\_IMT5155 (accession: GCF\_000813165.1)[32] and *E. coli* 789 (accession: GCF\_000819645.1)[25]. Genes in the pangenome were ordered initially using the APEC\_O1 reference followed by the order defined in PIRATE based on gene synteny and frequency. To perform core and accessory pangenome variation analyses a matrix was produced summarising the presence/absence and allelic diversity of every gene in the pangenome list<sup>105,106</sup>. Core genes were defined as present in 95% of the genomes and accessory genes were present in at least one isolate. The number of genes detected in each strain was calculated by PIRATE and can be found in Supplementary Data 4. Using this approach, the amount of coding sequences detected per strain were not significantly affected by the quality of the assembled genomes (Supplementary Fig. 4).

**Phylogenetic analysis.** Phylogenies were constructed by mapping pseudoreads simulated from assembled genomes to the *E. coli* O157 RefSeq reference genome (accession: NZ\_CP015831.1; 5,831,209 bp)<sup>107</sup> using snippy<sup>108</sup>.

This well-characterised closed reference genome, from Phylogroup E, was absent from the collection under investigation. This minimised and standardised bias caused by reference strain selection<sup>109</sup>. As all isolates were mapped to a single reference, this also allowed for a tree to be constructed from all isolates and comparison of genomic/alignment regions used for tree building during the recombination analysis. Other references, including APEC\_O1, were included in the pan genome analysis in order to provide points of reference for comparison of genes to well-characterised reference genomes. Pseudo-reads were created as a part of the SNIPPY pipeline used for variant calling. Contigs passed to SNIPPY were split into 250 bp single-end read pairs at a simulated ~20x coverage of the reference genome. These pseudo-reads were mapped against the reference genome O157 in the same manner as trimmed reads to retain order and ensure all data were comparable. Maximum-likelihood phylogenies were constructed separately for phylogroups A, B1, B2 and ST117, and the complete collection using a GTR + I + G substitution model and ultra-fast bootstrapping (1000 bootstraps) implemented in IQtree (version 1.6.8)<sup>110</sup> and visualised on Microreact<sup>111</sup>. Putative recombination sites were inferred using ClonalFrameML<sup>112</sup> and masked using cfml-maskrc (<https://github.com/kwong/cfml-maskrc>). Recombination-masked alignments were used to build midpoint rooted trees, which were used in treeWAS to weight associations, accounting for lineage effects<sup>113</sup> (Supplementary Fig. 3).

**Pangenome-wide association study of infection-associated genes.** Four GWAS analyses were performed to identify pathogenicity-associated core and accessory genome variation in the most common *E. coli* lineages, specifically phylogroups A ( $n = 71$ ), B1 ( $n = 85$ ) and B2 ( $n = 152$ ), and the ST-117 lineage ( $n = 220$ ). The remaining phylogroups contained too few isolates for reliable GWAS analysis. GWAS were performed using treeWAS<sup>35</sup> incorporating core and accessory genome variation identified by PIRATE<sup>104</sup>. We used the core and accessory genes to investigate associations representing segregation (disease vs. carriage isolates) of: (i) core SNPs; (ii) accessory gene presence/absence; (iii) gene fissions/fusions; (iv) gene duplications; (v) accessory gene alleles. Fission/fusion genes are identified by PIRATE<sup>104</sup> as genes which, due to nonsense mutation or frameshifts, comprise a single ORF in at least one isolate but two or more distinct ORFs in other isolates within the collection. Variants with an allele frequency <0.01 were excluded from the GWAS. The treeWAS algorithm performs three statistical association tests to calculate terminal, simultaneous and subsequent association scores. Infection-associated variants ( $p$ -value < 0.05) in any of these three tests were further

investigated<sup>114</sup>, and elements that were significantly associated in all four GWAS experiments ( $p$ -value < 0.01) were identified<sup>35</sup>.

**Quantitative analysis of putative plasmid genes.** Plasmid-associated genes are known to be important in APEC virulence<sup>31–33</sup>. We therefore identified putative plasmid replicon sequences in all the isolates using PlasmidFinder 2.1<sup>115</sup> (Supplementary Data 5). However, these genes can be located on a plasmid or the bacterial chromosome, for example in genomic islands. This means that analyses that focus on entire plasmids may miss the mosaic of possible plasmid and chromosomal genes that are known to recombine extensively in avian *E. coli*<sup>116</sup>. Confirmation of the plasmidic context of specific genes may require that analyses are performed separately on miniprep plasmid extractions or the use of long read sequencing<sup>117</sup>. However, the prevalence of putative APEC plasmid genes among isolates from carriage and invasive disease can be determined using the gene-by-gene approach<sup>37,38,118</sup> employed here. Additionally, annotated gene lists were obtained for the following plasmids: pAPEC-O1-R<sup>119</sup>; pAPEC-O2-R<sup>120</sup>; pAPEC-O2-211A-ColV<sup>121</sup>; pAPEC-O1-ColBM<sup>54</sup>; p1ColV5155<sup>122</sup>; pAPEC-O2-ColV<sup>10</sup>. To detect additional genes of plasmid origin among the infection-associated genes we used the PlasmidSeeker database (latest update on 12/07/2017)<sup>123</sup>. Specifically, all plasmid genes were annotated for putative function using the PlasmidSeeker database and PROKKA 1.13<sup>103</sup> to create a database compatible with Abriicate 0.8.13<sup>124</sup>, which was interrogated in relation to the list of the GWAS infection-associated genes. Putative plasmid genes were identified in all isolate genomes as a BLAST match of >70% over >50% of the gene length. For every plasmid, the number of genes present in each isolate was quantified and the distribution among asymptomatic and disease isolates was compared using an unpaired  $t$ -test.

**Horizontal gene transfer among infection-associated genes.** Population genetic analyses was undertaken to compare molecular variation among the 143 genes that contained infection-associated elements and the core genome of the dataset in this study. The number of alleles per locus was determined using a whole-genome MLST approach<sup>37</sup>, and the consistency of the phylogenetic trees to patterns of variation in sequence alignments for each gene was calculated<sup>125,126</sup> to infer the minimum amount of homoplasy in infection-associated and core genome genes. We defined the number of alleles per locus as the number of unique alleles per gene divided by the total number of isolates. Consistency indices (CI) for each single-gene alignment of 143 infection-associated genes to a phylogeny constructed using an alignment of 3094 core genes shared by 570 isolates, were calculated using the CI function of the R *Phangorn* package<sup>127</sup>. The average CI of these shared genes was compared with the CI of the genes containing pathogenicity-associated elements.

**Risk calculation.** Invasive infections caused by APEC are associated with multiple factors and large numbers of disease-associated elements within the bacterial genome. Therefore, while GWAS improves understanding of genome evolution and pathogen emergence, translating these findings for practical risk prediction models can be challenging. To achieve this, we trained a Random Forest (RF) classifier<sup>128</sup> using the GWAS output. This allowed us to capture the potentially complex, non-linear relationship between presence/absence patterns of disease-associated elements and phenotype, and rank the features according to their power to predict the isolate source (invasive disease vs. carriage). Analyses were conducted in R<sup>129</sup> using RandomForest<sup>130</sup>, ROCR<sup>131</sup> and ggplot2<sup>132</sup> software using the 510 isolates (291 invasive disease, 219 carriage) used in the GWAS with 79 binary presence/absence species-wide predictors used to train the RF model (Supplementary Data 3). In separate RF analyses, the classifiers were trained with (i) all 510 isolates and (ii) 294 isolates from phylogroups A ( $n = 70$ ), B1 ( $n = 80$ ) and B2 ( $n = 144$ ) with ST-117 ( $n = 216$ ) isolates as a test set. Based on the training data, a RF model with 1000 trees estimated the importance of the predictors with model criteria for feature selection. To estimate the out-of-sample accuracy of the model within its training domain (as specified by the phylogroups), the out-of-bag (OOB) predictions were used. In addition, in the second analysis the model was evaluated on the test set, which contained isolates from outside the training domain of the RF model. The predictive performance of the models was evaluated by predictive accuracy and area under the receiver operating characteristic (ROC) curve. Each analysis was repeated ten times and the reported results are the average over the ten independent runs. To investigate the prevalence of the APEC-associated genetic variants used in the RF model, in isolates sampled from other host niches we used four additional *E. coli* collections. Specifically, we analysed the *E. coli* reference collection (ECOR)<sup>42</sup>, 175 human ExPEC strains<sup>28</sup>, 14 disease-associated strains from dogs<sup>51</sup> and 521 strains from healthy cattle<sup>52</sup>. Published genomes were uploaded to BIGSdb<sup>133</sup> and presence-absence matrices were created for the APEC-associated SNPs and accessory genes (excluding fission/fusion and accessory gene alleles). The number and prevalence (%) of APEC-associated variants were compared for the four additional groups (Supplementary Fig. 5).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Short-read sequence data for all isolates sequenced in this study are deposited in the sequence read archive (SRA) and can be found associated with BioProject #PRJNA592536. Assembled genomes are also available on Figshare (<https://doi.org/10.6084/m9.figshare.12011811>). NCBI genome accession numbers for isolates in the validation dataset are included in Supplementary Data 1. Source data are provided for this paper.

Received: 20 April 2020; Accepted: 4 January 2021;

Published online: 03 February 2021

## References

- Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl Acad. Sci. USA* **115**, 6506–6511 (2018).
- Robbins, J. A., von Keyserlingk, M. A., Fraser, D. & Weary, D. M. INVITED REVIEW: Farm size and animal welfare. *J. Anim. Sci.* **94**, 5439–5455 (2016).
- Oakley, B. B. et al. The poultry-associated microbiome: network analysis and farm-to-fork characterizations. *PLoS ONE* **8**, e57190 (2013).
- Lutful Kabir, S. M. Avian colibacillosis and salmonellosis: a closer look at epidemiology, pathogenesis, diagnosis, control and public health concerns. *Int J. Environ. Res. Public Health* **7**, 89–114 (2010).
- Manges, A. R. et al. Global extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. *Clin. Microbiol. Rev.* <https://doi.org/10.1128/CMR.00135-18> (2019).
- Dho-Moulin, M. & Fairbrother, J. M. Avian pathogenic *Escherichia coli* (APEC). *Vet. Res.* **30**, 299–316 (1999).
- Huja, S. et al. Genomic avenue to avian colisepticemia. *mBio* <https://doi.org/10.1128/mBio.01681-14> (2015).
- Subedi, M. et al. Antibiotic resistance pattern and virulence genes content in avian pathogenic *Escherichia coli* (APEC) from broiler chickens in Chitwan, Nepal. *BMC Vet. Res.* **14**, 113 (2018).
- Xu, X., Sun, Q. & Zhao, L. Virulence factors and antibiotic resistance of avian pathogenic *Escherichia coli* in Eastern China. *J. Vet. Res.* **63**, 317–320 (2019).
- Johnson, T. J., Siek, K. E., Johnson, S. J. & Nolan, L. K. DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains. *J. Bacteriol.* **188**, 745–758 (2006).
- da Silva, G. J. & Mendonça, N. Association between antimicrobial resistance and virulence in *Escherichia coli*. *Virulence* **3**, 18–28 (2012).
- Lima-Filho, J. V. et al. Zoonotic potential of multidrug-resistant extraintestinal pathogenic *Escherichia coli* obtained from healthy poultry carcasses in Salvador. *Braz. J. Infect. Dis.* **17**, 54–61 (2013).
- Johnson, T. J. et al. Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens. *Appl. Environ. Microbiol.* **74**, 7043–7050 (2008).
- Tivendale, K. A. et al. Avian-pathogenic *Escherichia coli* strains are similar to neonatal meningitis *E. coli* strains and are able to cause meningitis in the rat model of human disease. *Infect. Immun.* **78**, 3412–3419 (2010).
- Mora, A. et al. Poultry as reservoir for extraintestinal pathogenic *Escherichia coli* O45:K1:H7-B2-ST95 in humans. *Vet. Microbiol.* **167**, 506–512 (2013).
- Vincent, C. et al. Food reservoir for *Escherichia coli* causing urinary tract infections. *Emerg. Infect. Dis.* **16**, 88–95 (2010).
- Collingwood, C., Kemmett, K., Williams, N. & Wigley, P. Is the concept of avian pathogenic *Escherichia coli* as a single pathotype fundamentally flawed? *Front. Vet. Sci.* **1**, 5 (2014).
- Hiki, M. et al. Phylogenetic grouping, epidemiological typing, analysis of virulence genes, and antimicrobial susceptibility of *Escherichia coli* isolated from healthy broilers in Japan. *Ir. Vet. J.* **67**, 14 (2014).
- Johnson, T. J. et al. Plasmid replicon typing of commensal and pathogenic *Escherichia coli* isolates. *Appl. Environ. Microbiol.* **73**, 1976–1983 (2007).
- Sheppard, S. K., Guttman, D. S. & Fitzgerald, J. R. Population genomics of bacterial host adaptation. *Nat. Rev. Genet.* **19**, 549–565 (2018).
- Cordoni, G. et al. Comparative genomics of European avian pathogenic *E. coli* (APEC). *BMC Genomics* **17**, 960 (2016).
- Cummins, M. L. et al. Whole genome sequence analysis of Australian avian pathogenic *Escherichia coli* that carry the class 1 integrase gene. *Microb. Genom.* <https://doi.org/10.1099/mgen.0.000250> (2019).
- Johnson, T. J. et al. Sequence analysis and characterization of a transferable hybrid plasmid encoding multidrug resistance and enabling zoonotic potential for extraintestinal *Escherichia coli*. *Infect. Immun.* **78**, 1931–1942 (2010).
- Binns, M. M., Mayden, J. & Levine, R. P. Further characterization of complement resistance conferred on *Escherichia coli* by the plasmid genes traT of R100 and iss of ColV,I-K94. *Infect. Immun.* **35**, 654–659 (1982).

25. Nolan, L. K. et al. Resistance to serum complement, iss, and virulence of avian *Escherichia coli*. *Vet. Res. Commun.* **27**, 101–110 (2003).
26. Johnson, T. J., Wannemuehler, Y. M. & Nolan, L. K. Evolution of the iss gene in *Escherichia coli*. *Appl. Environ. Microbiol.* **74**, 2360–2369 (2008).
27. Kohler, C. D. & Dobrindt, U. What defines extraintestinal pathogenic *Escherichia coli*? *Int. J. Med. Microbiol.* **301**, 642–647 (2011).
28. McNally, A. et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* **12**, e1006280 (2016).
29. Tivendale, K. A., Allen, J. L., Ginns, C. A., Crabb, B. S. & Browning, G. F. Association of iss and iucA, but not tsh, with plasmid-mediated virulence of avian pathogenic *Escherichia coli*. *Infect. Immun.* **72**, 6554–6560 (2004).
30. Delicato, E. R., de Brito, B. G., Gaziri, L. C. & Vidotto, M. C. Virulence-associated genes in *Escherichia coli* isolates from poultry with colibacillosis. *Vet. Microbiol.* **94**, 97–103 (2003).
31. Rodriguez-Siek, K. E., Giddings, C. W., Doetkott, C., Johnson, T. J. & Nolan, L. K. Characterizing the APEC pathotype. *Vet. Res.* **36**, 241–256 (2005).
32. Olsen, R. H., Christensen, H. & Bisgaard, M. Comparative genomics of multiple plasmids from APEC associated with clonal outbreaks demonstrates major similarities and identifies several potential vaccine-targets. *Vet. Microbiol.* **158**, 384–393 (2012).
33. Barbieri, N. L. et al. FNR regulates the expression of important virulence factors contributing to the pathogenicity of avian pathogenic *Escherichia coli*. *Front. Cell Infect. Microbiol.* **7**, 265 (2017).
34. Sheppard, S. K. et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl Acad. Sci. USA* **110**, 11923–11927 (2013).
35. Collins, C. & Didelot, X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.* **14**, e1005958 (2018).
36. Rasko, D. A. et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–6893 (2008).
37. Sheppard, S. K., Jolley, K. A. & Maiden, M. C. A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* **3**, 261–277 (2012).
38. Meric, G. et al. Lineage-specific plasmid acquisition and the evolution of specialized pathogens in *Bacillus thuringiensis* and the *Bacillus cereus* group. *Mol. Ecol.* **27**, 1524–1540 (2018).
39. Johnson, T. J. & Nolan, L. K. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* **73**, 750–774 (2009).
40. Fricke, W. F. et al. Antimicrobial resistance-conferring plasmids with similarity to virulence plasmids from avian pathogenic *Escherichia coli* strains in *Salmonella enterica* serovar Kentucky isolates from poultry. *Appl. Environ. Microbiol.* **75**, 5963–5971 (2009).
41. Chaudhuri, R. R. & Henderson, I. R. The evolution of the *Escherichia coli* phylogeny. *Infect. Genet. Evol.* **12**, 214–226 (2012).
42. Patel, I. R. et al. Draft genome sequences of the *Escherichia coli* reference (ECOR) collection. *Microbiol. Resour. Annu.* <https://doi.org/10.1128/MRA.01133-18> (2018).
43. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **5**, 58–65 (2013).
44. Clermont, O. et al. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infect. Genet. Evol.* **11**, 654–662 (2011).
45. Clermont, O. et al. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* **21**, 3107–3117 (2019).
46. Bosshard, L. et al. Accumulation of deleterious mutations during bacterial range expansions. *Genetics* **207**, 669–684 (2017).
47. Nelson, K. & Selander, R. K. Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (gnd) in enteric bacteria. *Proc. Natl Acad. Sci. USA* **91**, 10227–10231 (1994).
48. Dissanayake, D. R., Octavia, S. & Lan, R. Population structure and virulence content of avian pathogenic *Escherichia coli* isolated from outbreaks in Sri Lanka. *Vet. Microbiol.* **168**, 403–412 (2014).
49. Mora, A. et al. Emerging avian pathogenic *Escherichia coli* strains belonging to clonal groups O111:H4-D-ST2085 and O111:H4-D-ST117 with high virulence-gene content and zoonotic potential. *Vet. Microbiol.* **156**, 347–352 (2012).
50. Ronco, T. et al. Spread of avian pathogenic *Escherichia coli* ST117 O78:H4 in Nordic broiler production. *BMC Genomics* **18**, 13 (2017).
51. Valat, C. et al. Pathogenic *Escherichia coli* in dogs reveals the predominance of ST372 and the human-associated ST73 extra-intestinal lineages. *Front. Microbiol.* **11**, 580 (2020).
52. Arimizu, Y. et al. Large-scale genome analysis of bovine commensal *Escherichia coli* reveals that bovine-adapted *E. coli* lineages are serving as evolutionary sources of the emergence of human intestinal pathogenic strains. *Genome Res.* **29**, 1495–1505 (2019).
53. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
54. Johnson, T. J., Johnson, S. J. & Nolan, L. K. Complete DNA sequence of a ColBM plasmid from avian pathogenic *Escherichia coli* suggests that it evolved from closely related ColV virulence plasmids. *J. Bacteriol.* **188**, 5975–5983 (2006).
55. Falkow, S. Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* **10**, S274–S276 (1988).
56. Murray, G. L., Attridge, S. R. & Morona, R. Regulation of *Salmonella typhimurium* lipopolysaccharide O antigen chain length is required for virulence; identification of FepE as a second Wzz. *Mol. Microbiol.* **47**, 1395–1406 (2003).
57. Poulsen, L. K., Larsen, N. W., Molin, S. & Andersson, P. A family of genes encoding a cell-killing function may be conserved in all gram-negative bacteria. *Mol. Microbiol.* **3**, 1463–1472 (1989).
58. Jorgensen, S. L. et al. Diversity and population overlap between avian and human *Escherichia coli* belonging to sequence type 95. *mSphere* <https://doi.org/10.1128/mSphere.00333-18> (2019).
59. Najafi, S., Rahimi, M. & Nikousefat, Z. Extra-intestinal pathogenic *Escherichia coli* from human and avian origin: detection of the most common virulence-encoding genes. *Vet. Res. Forum* **10**, 43–49 (2019).
60. Kariyawasam, S., Johnson, T. J. & Nolan, L. K. The pap operon of avian pathogenic *Escherichia coli* strain O1:K1 is located on a novel pathogenicity island. *Infect. Immun.* **74**, 744–749 (2006).
61. Janben, T. et al. Virulence-associated genes in avian pathogenic *Escherichia coli* (APEC) isolated from internal organs of poultry having died from colibacillosis. *Int. J. Med. Microbiol.* **291**, 371–378 (2001).
62. Klemm, P. & Hedegaard, L. Fimbriae of *Escherichia coli* as carriers of heterologous antigenic sequences. *Res. Microbiol.* **141**, 1013–1017 (1990).
63. Tuntufye, H. N., Lebeer, S., Gwakisa, P. S. & Goddeeris, B. M. Identification of Avian pathogenic *Escherichia coli* genes that are induced in vivo during infection in chickens. *Appl. Environ. Microbiol.* **78**, 3343–3351 (2012).
64. Hayashi, M., Tabata, K., Yagasaki, M. & Yonetani, Y. Effect of multidrug-efflux transporter genes on dipeptide resistance and overproduction in *Escherichia coli*. *FEMS Microbiol. Lett.* **304**, 12–19 (2010).
65. Hirakawa, H., Nishino, K., Hirata, T. & Yamaguchi, A. Comprehensive studies of drug resistance mediated by overexpression of response regulators of two-component signal transduction systems in *Escherichia coli*. *J. Bacteriol.* **185**, 1851–1856 (2003).
66. Szvetnik, A., Gal, J. & Kalman, M. Membrane topology of the GltS Na<sup>+</sup>/glutamate permease of *Escherichia coli*. *FEMS Microbiol. Lett.* **275**, 71–79 (2007).
67. Grisolia, V., Carlomagno, M. S., Nappo, A. G. & Bruni, C. B. Cloning, structure, and expression of the *Escherichia coli* K-12 hisC gene. *J. Bacteriol.* **164**, 1317–1323 (1985).
68. Peleg, A. et al. Identification of an *Escherichia coli* operon required for formation of the O-antigen capsule. *J. Bacteriol.* **187**, 5259–5266 (2005).
69. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).
70. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50 (2017).
71. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
72. Cano-Gamez, E. & Trynka, G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.* **11**, 424 (2020).
73. Nolan, L. K., Vaillancourt, J.-P., Barbieri, N. L. & Logue, C. M. Colibacillosis. In *Diseases of Poultry* 14th edn (eds Swayne, D. E. et al.) Ch. 18 (John Wiley & Sons, Ltd, 2020).
74. Klemm, P. & Christiansen, G. The fimD gene required for cell surface localization of *Escherichia coli* type 1 fimbriae. *Mol. Gen. Genet.* **220**, 334–338 (1990).
75. Law, D. Adhesion and its role in the virulence of enteropathogenic *Escherichia coli*. *Clin. Microbiol. Rev.* **7**, 152–173 (1994).
76. Hagberg, L. et al. Adhesion, hemagglutination, and virulence of *Escherichia coli* causing urinary tract infections. *Infect. Immun.* **31**, 564–570 (1981).
77. Mainil, J. *Escherichia coli* virulence factors. *Vet. Immunol. Immunopathol.* **152**, 2–12 (2013).
78. Schubert, S. et al. Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. *PLoS Pathog.* **5**, e1000257 (2009).
79. Rodriguez-Beltran, J. et al. High recombinant frequency in extraintestinal pathogenic *Escherichia coli* strains. *Mol. Biol. Evol.* **32**, 1708–1716 (2015).
80. Didelot, X. & Maiden, M. C. Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315–322 (2010).



81. Takeuchi, N., Cordero, O. X., Koonin, E. V. & Kaneko, K. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.* **13**, 20 (2015).
82. Shapiro, B. J. et al. Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48–51 (2012).
83. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
84. Card, R. M. et al. An In vitro chicken gut model demonstrates transfer of a multidrug resistance plasmid from *Salmonella* to commensal *Escherichia coli*. *mBio* <https://doi.org/10.1128/mBio.00777-17> (2017).
85. Ervin, J. S. et al. Characterization of fecal concentrations in human and other animal sources by physical, culture-based, and quantitative real-time PCR methods. *Water Res.* **47**, 6873–6882 (2013).
86. Gibson, B., Wilson, D. J., Feil, E. & Eyre-Walker, A. The distribution of bacterial doubling times in the wild. *Proc. Biol. Sci.* <https://doi.org/10.1098/rspb.2018.0789> (2018).
87. Wiser, M. J., Ribbeck, N. & Lenski, R. E. Long-term dynamics of adaptation in asexual populations. *Science* **342**, 1364–1367 (2013).
88. Antao, E. M. et al. The chicken as a natural model for extraintestinal infections caused by avian pathogenic *Escherichia coli* (APEC). *Microb. Pathog.* **45**, 361–369 (2008).
89. La Ragione, R. M. & Woodward, M. J. Virulence factors of *Escherichia coli* serotypes associated with avian colisepticaemia. *Res. Vet. Sci.* **73**, 27–35 (2002).
90. Lacour, S., Bechet, E., Cozzzone, A. J., Mijakovic, I. & Grangeasse, C. Tyrosine phosphorylation of the UDP-glucose dehydrogenase of *Escherichia coli* is at the crossroads of colanic acid synthesis and polymyxin resistance. *PLoS ONE* **3**, e3053 (2008).
91. Zhao, J., Baba, T., Mori, H. & Shimizu, K. Global metabolic response of *Escherichia coli* to *gnd* or *zwf* gene-knockout, based on <sup>13</sup>C-labeling experiments and the measurement of enzyme activities. *Appl. Microbiol. Biotechnol.* **64**, 91–98 (2004).
92. Tarr, P. I. et al. Acquisition of the *rfb-gnd* cluster in evolution of *Escherichia coli* O55 and O157. *J. Bacteriol.* **182**, 6183–6191 (2000).
93. Roth, N. et al. The application of antibiotics in broiler production and the resulting antibiotic resistance in *Escherichia coli*: a global overview. *Poult. Sci.* **98**, 1791–1804 (2019).
94. Marquardt, R. R. & Li, S. Antimicrobial resistance in livestock: advances and alternatives to antibiotics. *Anim. Front.* **8**, 30–37 (2018).
95. Paharik, A. E., Schreiber, H. L. T., Spaulding, C. N., Dodson, K. W. & Hultgren, S. J. Narrowing the spectrum: the new frontier of precision antimicrobials. *Genome Med.* **9**, 110 (2017).
96. de la Fuente-Nunez, C., Torres, M. D., Mojica, F. J. & Lu, T. K. Next-generation precision antimicrobials: towards personalized treatment of infectious diseases. *Curr. Opin. Microbiol.* **37**, 95–102 (2017).
97. Bikard, D. & Barrangou, R. Using CRISPR-Cas systems as antimicrobials. *Curr. Opin. Microbiol.* **37**, 155–160 (2017).
98. Kiga, K. et al. Development of CRISPR-Cas13a-based antimicrobials capable of sequence-specific killing of target bacteria. *Nat. Commun.* **11**, 2934 (2020).
99. Kemmett, K. et al. A longitudinal study simultaneously exploring the carriage of APEC virulence associated genes and the molecular epidemiology of faecal and systemic *E. coli* in commercial broiler chickens. *PLoS ONE* **8**, e67749 (2013).
100. Johnson, T. J. et al. The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J. Bacteriol.* **189**, 3228–3236 (2007).
101. Mangiamale, P. et al. Complete genome sequence of the avian pathogenic *Escherichia coli* strain APEC O78. *Genome Announc.* **1**, e0002613 (2013).
102. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
103. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
104. Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K. & Feil, E. J. PIRATE: a fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* <https://doi.org/10.1093/gigascience/giz119> (2019).
105. Pascoe, B. et al. Enhanced biofilm formation and multi-host transmission evolve from divergent genetic backgrounds in *Campylobacter jejuni*. *Environ. Microbiol.* **17**, 4779–4789 (2015).
106. Meric, G. et al. Ecological overlap and horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Genome Biol. Evol.* **7**, 1313–1328 (2015).
107. Ronner, A. B. & Cliver, D. O. Isolation and characterization of a coliphage specific for *Escherichia coli* O157:H7. *J. Food Prot.* **53**, 944–947 (1990).
108. Seemann, T. *Snippy: fast bacterial variant calling from NGS reads* (2015).
109. Bush, S. J. et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *GigaScience* <https://doi.org/10.1093/gigascience/giaa007> (2020).
110. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
111. Argimon, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genom.* **2**, e000093 (2016).
112. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
113. Earle, S. G. et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.* **1**, 16041 (2016).
114. Read, T. D. & Massey, R. C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med.* **6**, 109 (2014).
115. Carattoli, A. & Hasman, H. PlasmidFinder and In Silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol. Biol.* **2075**, 285–294 (2020).
116. Boyd, E. F. & Hartl, D. L. Recent horizontal transmission of plasmids between natural populations of *Escherichia coli* and *Salmonella enterica*. *J. Bacteriol.* **179**, 1622–1627 (1997).
117. Gonzalez-Escalona, N., Allard, M. A., Brown, E. W., Sharma, S. & Hoffmann, M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS ONE* **14**, e0220494 (2019).
118. Brehony, C. et al. An MLST approach to support tracking of plasmids carrying OXA-48-like carbapenemase. *J. Antimicrob. Chemother.* **74**, 1856–1862 (2019).
119. Hall, R. M. Antibiotic resistance gene cluster of pAPEC-O1-R. *Antimicrob. Agents Chemother.* **51**, 3461–3462 (2007).
120. Johnson, T. J., Siek, K. E., Johnson, S. J. & Nolan, L. K. DNA sequence and comparative genomics of pAPEC-O2-R, an avian pathogenic *Escherichia coli* transmissible R plasmid. *Antimicrob. Agents Chemother.* **49**, 4681–4688 (2005).
121. Nielsen, D. W. et al. Complete genome sequence of avian pathogenic *Escherichia coli* strain APEC O2-211. *Microbiol. Resour. Announc.* <https://doi.org/10.1128/MRA.01046-18> (2018).
122. Zhu, Ge, X. et al. Comparative genomic analysis shows that avian pathogenic *Escherichia coli* isolate IMT5155 (O2:K1:H5; ST complex 95, ST140) shares close relationship with ST95 APEC O1:K1 and human ExPEC O18:K1 strains. *PLoS ONE* **9**, e112048 (2014).
123. Roosaare, M., Puustusmaa, M., Mols, M., Vaher, M. & Remm, M. PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. *PeerJ* **6**, e4588 (2018).
124. Lee, R. S. et al. The changing landscape of vancomycin-resistant *Enterococcus faecium* in Australia: a population-level genomic study. *J. Antimicrob. Chemother.* **73**, 3268–3278 (2018).
125. Meric, G. et al. Disease-associated genotypes of the commensal skin bacterium *Staphylococcus epidermidis*. *Nat. Commun.* **9**, 5034 (2018).
126. Mourkas, E. et al. Gene pool transmission of multidrug resistance among *Campylobacter* from livestock, sewage and human disease. *Environ. Microbiol.* **21**, 4597–4613 (2019).
127. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
128. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
129. R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2010).
130. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
131. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCr: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
132. Wickham, H. *ggplot2: elegant graphics for data analysis* (Springer New York, 2009).
133. Jolley, K. A. & Maiden, M. C. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform.* **11**, 595 (2010).

## Acknowledgements

This work was supported by Wellcome Trust grants 088786/C/09/Z and Medical Research Council (MRC) grants MR/M501608/1 and MR/L015080/1 awarded to S.K.S.

## Author contributions

S.K.S., L.M. and G.M. conceived and designed the study. L.M., G.M., B.P., S.K.S., S.M., J.C., T.S.W., and L.K.W. carried out Laboratory work. B.P., K.A.J., and S.K.S. supported data archiving. L.M., G.M., S.B., K.Y., E.M., J.P. and J.C. analysed the data. E.J.F., S.C.B., M.D.H., J.P., K.K., N.J.W. and J.C. contributed to data interpretation. S.K.S. L.M. and G.M. wrote the paper.

## Competing interests

The authors declare no competing interests.



### Additional information

The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-20988-w>.

**Correspondence** and requests for materials should be addressed to S.K.S.

**Peer review information** *Nature Communications* thanks Ola Brynildsrud and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021