Description and validation of a new set of PCR markers predictive of avian pathogenic *Escherichia coli* virulence

Camille Lucas[a], Sabine Delannoy[b], Catherine Schouler[c], Rozenn Souillard[a], Laetitia Le Devendec[a], Pierrick Lucas[a], Alassane Keita[a], Patrick Fach[b], Julie Puterflam[d]*, Stéphanie Bougeard[a]* and Isabelle Kempf[a]*&

*contributed equally to this work

[a]ANSES (The French Agency for Food, Environmental and Occupational Health & Safety), Ploufragan-Plouzané-Niort Laboratory, 22440 Ploufragan, France

[b]ANSES (The French Agency for Food, Environmental and Occupational Health & Safety), Laboratory for Food Safety, Unit of 'Pathogenic E. coli (COLiPATH) & Genomics platform 'IdentyPath' (IDPA), 94701 Maisons-Alfort, France

[c]INRAE, Université de Tours, ISP, 37380 Nouzilly, France

[d]ITAVI, 22440 Ploufragan, France

[&]Corresponding Author: Isabelle Kempf, French Agency for Food, Environmental and Occupational Health & Safety (ANSES), Ploufragan-Plouzané-Niort Laboratory, Zoopole les croix, 22440 Ploufragan, France

Phone: 33 296 01 62 81

Fax: 33 2 96 01 62 73

Email: Isabelle.kempf@anses.fr

Abstract

Avian colibacillosis is the main bacterial infectious disease in poultry and is caused by avian pathogenic *Escherichia coli* (APEC). However, *E. coli* strains are very diverse, and not all are pathogenic for poultry. A straightforward scheme for identifying APEC is crucial to better control avian colibacillosis. In this study, we combined high-throughput PCR and a machine learning procedure to identify relevant genetic markers associated with APEC. Markers related to phylogroup, serotype and 66 virulence factors were tested on a large number of *E. coli* strains isolated from environmental, faecal or colibacillosis lesion samples in 80 broiler flocks. Nine classification methods and a machine learning procedure were used to differentiate 170 strains presumed non-virulent (obtained from farm environments) from 203 strains presumed virulent (obtained from colibacillosis cases on chicken farms) and to develop a prediction model to evaluate the pathogenicity of isolates. The model was then validated on 14 isolates using a chick embryo lethality assay. The selected and validated model based on the bootstrap aggregating tree method relied on a scheme of 13 positive or negative markers associated with phylogroups (*arpA*), H4 antigen and virulence markers (*aec4*, ETT2.2, $frz_{orf4,}$ *fyuA, iha, ireA, iroN, iutA1, papA, tsh,* and *vat*). It had a specificity of 84% and a sensitivity of 85%, and was implemented as an online tool. Our scheme offers an easy evaluation of the virulence of avian *E. coli* isolates on the basis of the presence/absence of these 13 genetic markers, allowing for better control of avian colibacillosis**.**

Keywords: Avian pathogenic *Escherichia coli*; virulence; machine learning, high-throughput PCR

Introduction

Colibacillosis is the main bacterial infectious disease of poultry worldwide (Nolan et al., 2020). *Escherichia coli* can infect birds of various species or production types (broilers, breeders, layers, turkeys, etc.) at different ages (Nolan et al., 2020; Souillard et al., 2019). The disease leads to high economic losses and animal welfare issues. Methods to control infections include biosecurity measures, vaccination and medical treatment, including antimicrobials. Antimicrobials in poultry are most often administered by the oral route via drinking water and lead to the selection of antimicrobial-resistant pathogenic and commensal *E. coli* strains, not to mention the other poultry intestinal species, some of which are zoonotic, e.g. *Campylobacter* and *Salmonella*. Non-pathogenic antimicrobial-resistant *E. coli* of poultry origin may then transfer their resistance genes to human *E. coli*, because a subset of avian *E. coli* is pathogenic for humans (Nolan et al., 2020). Better management and mitigation of avian colibacillosis is thus essential for economic, animal welfare and public health reasons.

Like other authors (Nolan et al., 2020), we recently reported very high genetic diversity in *E. coli* strains isolated from diseased and healthy chickens and their environment (Delannoy et al., 2020). In many colibacillosis cases, several strains are isolated from internal organs, but it is highly likely that not all of them are actually pathogenic. For example, using the one-day-old chick lethality test, Schouler et al. (2012)demonstrated that some strains isolated from colibacillosis lesions of diseased birds are non-virulent. In the field, veterinarians need to know the pathogenic potential of isolates to identify avian pathogenic *E. coli* (APEC) strains to control colibacillosis, either by use of conventional vaccines, endogenous vaccines or antimicrobials. Several authors have therefore proposed diagnostic strategies to predict the pathogenicity of isolates (Table 1). In 2003, Skyberg et al. proposed a multiplex amplification protocol targeting four genes (*iss*, increased serum survival gene; *tsh*, temperature-sensitive haemagglutinin gene; *cvi*, ColV immunity gene; and *iucC*, a gene of the aerobactin operon) and

described a relationship between the number of genes present in the 20 tested strains and their virulence for chicken embryos. Another typing scheme is based on eight virulence-associated factors, including – in addition to *iss, tsh, cva/cvi* and *iucD – pap*C (P-fimbriae gene), *irp2* (iron-acquisition system yersiniabactin gene), *ast*A (enteroaggregative heat-stable toxin gene) and *vat* (vacuolating autotransporter toxin gene) (Ewers et al., 2005). Strains isolated from septicaemic poultry harboured four to eight of these genes, contrary to strains from the faeces of healthy chickens. Subsequently, a set of five plasmid-linked virulence genes: *iss*, *iutA* (aerobactin siderophore receptor gene), *hlyF* (putative avian haemolysin gene), *iroN* (salmochelin siderophore receptor gene) and *ompT* (episomal outer membrane protease gene) was developed and validated for use in a multiplex PCR to differentiate APEC, which possess on average 4.0 of these virulence-associated genes (VAG) compared with 1.3 VAG for faecal isolates (Johnson et al., 2008). Moreover, these markers correspond to the mortality and lesions observed with 124 isolates that were experimentally inoculated in chickens. The above-cited Schouler et al. (Schouler et al., 2012) study included the characterization of 1491 isolates and a scheme including *iutA*, P(F11) (fimbriae-encoding gene), frz_{orf4} (metabolic operon gene), *sitA* (iron and manganese transport gene) and *aec26* (gene coding for a putative membrane protein component of a type VI secretion system) and the serogroup O78, yielding a diagnostic strategy based on the definition of four genetic patterns for pathogenic strains: A [$iutA^+$, $P(F11)^+$], B [$iutA^+$, $P(F11)^-$, $frz_{orf4}{}^+$], C [$iutA^+$, $P(F11)^-$, $frz_{orf4}{}^-$, $O78^+$], and D [$iutA^-$, $sitA^+$, $aec26^+$] (N.B. P(F11) means positive for *felA*, *papC* and a variant of *papG*). However, these different APEC characterization tools were developed using a relatively small number of isolates, or could identify only 70.2% (Schouler et al., 2012) to 85.4% (Johnson et al., 2008) of pathogenic isolates. Thus, we took advantage of recent tools, such as high-throughput PCR and machine-learning procedures (Hastie et al., 2009; Mitchell, 1997), to study a large number of *E. coli*

isolates obtained from poultry or their environment (Delannoy et al., 2020) with the aim of developing an improved scheme for differentiating APEC from non-APEC strains.

Materials and methods

Strains and high-throughput PCR analysis

The bacterial isolates were previously described in (Delannoy et al., 2020). Briefly, they were obtained as part of an epidemiological study of colibacillosis in 80 broiler flocks in western France. The strains were isolated from farm environments before chick arrival, day-old chicks, chick transport boxes and from internal organs of birds exhibiting typical colibacillosis symptoms and lesions. Early colibacillosis was defined as a flock of up to 10 days of age with a daily mortality rate higher than 0.3% and suspect clinical signs and lesions, whereas late colibacillosis was defined as birds older than 10 days with a daily mortality rate higher than 0.1% on two consecutive days and suspect clinical signs and lesions. Isolates from early colibacillosis were isolated from joints, the liver, the pericardium, the spleen, air sacs and the vitellus and those from late colibacillosis from the same organs except the vitellus. Colibacillosis episodes were recorded in 31 flocks (Delannoy et al., 2020).

A high-throughput microfluidic real-time PCR (qPCR) system, the BioMark™ real-time PCR system (Fluidigm, San Francisco, USA), was used to screen for genetic markers related to 23 antigens (O1 (2 variants), O2 (3 variants), O6, O8, O11, O18, O23, O25, O35, O45(S88), O78, O88, O153, H4, H7, H8, H21, H25, K1, and K5), five phylogroups (Clermont et al., 2019) and 66 virulence markers (Delannoy et al., 2020).

Statistical analysis

The original data involved 1,050 *E. coli* strains on which 68 genetic variables were measured (i.e., phylogroup, serotype and 66 virulence markers). Our aim was to predict the virulence status (i.e., virulent/non-virulent) of the strains; therefore, statistical analyses were applied to a

subset of 373 *E. coli* strains (i.e., 170 presumed non-virulent strains from no-colibacillosis farm environments (controls), and 203 presumed virulent strains obtained from colibacillosis lesions from actual colibacillosis cases on chicken farms). The 203 presumed virulent strains had been isolated during early colibacillosis cases (138 isolates) or late colibacillosis cases (65 isolates). The prediction (response) variable is Boolean, designed to predict the virulence status (i.e., virulent/non-virulent). To give the same weight to all genetic variables in the analysis, phylogroup markers and serotypes – both of which have numerous levels – were recoded as multiple Boolean variables (i.e., 0/1). Model selection was carried out in four steps. First, weakly informative variables (i.e., less than 1% of variability) were discarded. Second, nine classification models were applied with a machine-learning procedure, namely the bootstrap aggregating tree, random forest, elastic net logistic regression, regularized logistic regression, boosting logistic regression, PLS discriminant analysis, K-nearest neighbours, support vector machine and neural network models. The tuning parameters of each model were optimized with a 10-fold cross-validation procedure repeated 10 times (Stone, 1974). Thus, the prediction ability of each model was evaluated based on four criteria (i.e., sensibility, specificity, false positive and false negative rates) within a 2-fold cross-validation procedure (i.e., training and test data) repeated 30 times. Third, the best predictive model was selected (i.e., highest sensibility and specificity with a threshold set to 87.5%, lowest false positive and negative rates with a threshold set to 10%). Indices of the informativeness of each variable were used to select the most relevant variables − among all phylogroup markers, serotypes and virulence markers − for prediction. To help with interpretation and to determine the way each marker influences the prediction, univariate logistic regressions were applied for each marker of interest. Finally, the best predictive model with the selected relevant variables was used to predict the status (i.e., virulent/non-virulent) of new strains. This procedure is illustrated in Figure 1. The overall procedure was coded in R, based on the ‘caret’ package (Kuhn, 2008).

Chick embryo pathogenicity of selected isolates

To validate the sets of genetic markers, the virulence of 14 isolates was determined in a chicken embryo lethality assay (CELA). The isolates were randomly chosen, based on the four combinations of origin (farm environment or colibacillosis lesions from cases) and predictions of pathogenicity or non-pathogenicity, according to the best predictive model with the selected relevant variables (i.e., the bootstrap aggregating tree method with the set of the 13 selected markers). Insofar as possible, isolates were chosen from different flocks.

For inoculation, we used a protocol adapted from Trotereau and Schouler (2019). Briefly, for each strain, one colony was suspended in 5 mL of Mueller Hinton (MH) broth and the culture was incubated at 37°C for 5 hours. Then a 0.5 MacFarland suspension was prepared in MH broth and diluted 1:100 in endotoxin-free Dulbecco's phosphate-buffered saline (PBS). The titres of the 1:100 diluted suspensions were determined by plating decimal dilutions onto MH agar plates and the suspensions were stored at 5°C for 24 hours until the day of inoculation. On the day of inoculation, the colonies were counted, the titres were calculated and the 1:100 suspensions were diluted in PBS in order to obtain suspensions containing approximately 10^3 colony-forming units (CFU)/mL for inoculating eggs. Then, 100 µL of these suspensions were inoculated into the allantoic cavity of 10 eleven-day-old specific-pathogen-free chicken embryos (poultry experimental unit at the ANSES Ploufragan Laboratory). Moreover, 10 eggs were non-inoculated and 10 eggs were inoculated with 100 µL of sterile PBS. The eggs were then candled daily to monitor mortality up to the fourth day after inoculation, as described in (Wooley et al., 2000). Each day, for each strain, tissues from one of the dead embryos were placed on MacConkey agar plates, and plates were incubated at 37°C to check for the presence of *E. coli* colonies. Because our inoculation conditions were not strictly those of Wooley et al. (2000), we set the following criteria: strains resulting in mortality for up to 3 embryos were

considered non-virulent, those with mortality for 7 or more strains were considered virulent, and the others were classified as of intermediate virulence.

Results

Identification of relevant genetic markers associated with APEC

The PCR results obtained on the subset of 373 *E. coli* strains (i.e., 170 (presumed) non-virulent strains from control-farm environment, and 203 (presumed) virulent strains obtained from colibacillosis lesions on case-farm chickens) are given in the Table S1. The nine phylogroups and the 34 different serotypes were recoded into Boolean variables. The serotype variables O25B and O153 were removed because they were similar to two other explanatory variables (O25 and O153alter). The variables with zero or very low variability (i.e., leading to unstable models) were not included in the final statistical analysis. Thus, 26 variables were removed, and the final number of variables in the statistical study was 16 serotypes, 5 phylogroups and 48 virulence markers, thus 69 variables in total.

Several markers (O1a, O18, O23, O153alter, *fli*CH21, *fli*CH25, K5, *clb*N and *cld*B) were detected in less than 5% of control and case strains. On the contrary, the markers *ecp*D1, *fep*A3, *iss1*, *iss2*, *iss3*, *iss5*, *nir*C and *pab*B were present in more than 85% of control and of cases strains. For 15 markers (*fli*CH4, *arpA*, TspE4.C2, *aec4*, ETT2.2, frz_{orf4}, *fyuA*, *iha*, *ireA*, *iroN*, *iutA1*, *tkt1*, *tsh*, *vat* and *yqi*C), the percentages of positive strains differed by more than 30% between control strains and case strains, with higher percentages for case strains except for *arp*A, Tspe4.c2 and ETT2.2.

The performances obtained from the 30 simulations for the nine methods showed that, with the full set of 69 markers, four methods met the criterion thresholds (sensitivity and specificity greater than 87.5% and false positive and false negative rates less than 10%): K-nearest neighbours, neural networks, random forest, and bootstrap aggregating tree (Table 2).

Comparing the variables for each method, thirteen markers stood out as significant for all four methods (Table 3); 5, 10 and 13 markers were relevant for respectively three, two and one of the four methods. Twenty-eight markers were not relevant for any method.

The K-nearest neighbours method seemed unstable when there was a reduction in the number of markers used. Indeed, the sensitivity for the prediction with 18 markers was rather low (i.e., 54%), although it was compensated by a high specificity in prediction (i.e., 98%) (results not shown). Nevertheless, the other three methods seemed robust (even with 13 markers) and allowed good predictions of cases and controls. The bootstrap aggregating tree method applied with the 13 selected markers yielded prediction scores of 85% for cases, 84% for controls, with 17% of false negatives and 13% of false positives. The scheme is now available as a free application (https://sbougeard.shinyapps.io/applishinypcr/) and is presented in Figure S1and supplementary material "DataPCR.xlsx").

The 13 selected markers influence colibacillosis in different ways. Univariate logistic regressions were applied to determine how each marker influences virulence (Table 4). Results showed that 11 markers (*fli*CH4, *aec4*, frz_{orf4}, *fyuA*, *iha*, *ireA*, *iroN*, *iutA1*, *papA*, *tsh* and *vat*) were significantly associated with the colibacillosis strains and two markers (*arpA* and ETT2.2) were significantly associated with a non-virulent status.

Validation of the model

To validate our prediction model based on the bootstrap aggregating tree method and the use of 13 genetic markers, the virulence of the 373 strains was calculated. The virulence of 14 isolates was assessed using CELA. The selection of these 14 test isolates was based on their origin and prediction probabilities that were respectively higher than 0.95 or 0 for virulence or non-virulence. Thus, we included randomly selected isolates obtained from colibacillosis cases and predicted to be virulent (CV group: 5 isolates) or non-virulent (CN group: 3 isolates). Inversely, isolates obtained from farm environments and predicted non-virulent (EN group: 4

222 isolates) or **v**irulent (EV group: 2 isolates) were included. The randomly selected isolates had
223 probabilities of virulence of 0.96 for EV strain #5 and 1 for the four other CV and the two EV
224 strains, and of 0 for the three CN and four EN strains. Only two EV isolates were available and
225 both were included. Two isolates from flock L1 were included, one CV strain of phylogroup
226 D/E and one EN strain of phylogroup B1.
227 The titres of the inocula from 14 tested strains and the observed dead embryos are given in
228 Table 5. Titres ranged from 28 to 136 CFU/egg. No mortality was detected in non-inoculated
229 eggs or in eggs inoculated with sterile PBS. CV strains yielded mortalities from 6 to 9 of the 10
230 inoculated eggs (Table 5). The two EV strains killed 7 or 8 out of 10 embryos. The three CN
231 strains killed 1 to 6 of the 10 inoculated embryos. Finally, the mortalities recorded for the four
232 EN strains ranged from 0/10 to 3/10. All cultures from dead embryos yielded abundant pure *E.*
233 *coli* cultures.
234
235 **Discussion**
236 Initially, 95 genetic markers were investigated, including 23 associated with serotypes, 5 with
237 phylogroups and 66 related to virulence. Ultimately, we developed a predictive model using the
238 bootstrap aggregating tree method and based on a scheme of 13 positive or negative genetic
239 markers, including one associated with an antigen (*fli*CH4), one associated with a phylogroup
240 (*arpA*) and 11 associated with virulence (*tsh, iutA, iroN, papA, fyuA, aec4,* frz_{ofr4}*, iha, ireA,*
241 *cma, vat* and ETT2.2).
242 Our 13-marker scheme led to the prediction of 84% of cases and 84% of controls, with 17% of
243 false negatives and 13% of false positives. Some of the 13 colibacillosis markers had already
244 been included in previously published sets of predictors (Table 1) and are well known to be
245 involved in colibacillosis pathogenesis.

The role of *tsh, iroN* and the aerobactin cluster (*iucABCDiutA*) in *in vivo* persistence and development of lesions in respiratory and deeper tissues in inoculated chickens has been demonstrated using selective capture of transcribed sequences and *in vivo* virulence studies in chickens inoculated with mutant strains (Dozois et al., 2003). Our scheme included the *tsh* gene, also included in two other sets (Ewers et al., 2005; Skyberg et al., 2003), and the *iroN* gene, one of the predictors in the Johnson et al. (2008) scheme; *tsh* and *iroN* are borne by the pColV plasmid. The *tsh* gene encodes a temperature-sensitive haemagglutinin and *iroN* a siderophore receptor. The *iutA* gene is borne on plasmids or sometimes on the chromosome (Schouler et al., 2012). It was present in our and in two previous schemes (Johnson et al., 2008; Schouler et al., 2012), and is part of the five genes of the aerobactin operon. Gao et al. (2015) showed that aerobactin-defective mutants of an O2 *E. coli* have significantly decreased pathogenicity in challenged chickens. Another study detected the *iutA* and the *iroN* genes in respectively 91.9% and 79.9% of the strains isolated from colibacillosis lesions from 60 commercial broiler farms in Korea (Kim et al., 2020). In a collection of isolates obtained in the USA in 2018 from poultry diagnosed with colibacillosis, *iroN* was identified in 93.4% of isolates (Newman et al., 2021).

The frz_{orf4} marker (cf. Schouler et al., 2012), is a chromosomal gene associated with sugar metabolism and fitness under stress conditions: it is significantly associated with virulence for one-day-old chicks (Rouquet et al., 2009; Schouler et al., 2012).

The *vat* gene can induce intracellular vacuoles and *vat* mutants exhibit no or reduced virulence in infection models of disease in broiler (Parreira and Gyles, 2003). This gene is also included in the Ewers et al. scheme (2005).

Adhesins including fimbriae are important for colonization of respiratory tissues (Guabiraba and Schouler, 2015), and the *papC* gene is included in the Ewers et al. scheme (2005). For example, one study demonstrated greatly attenuated *in vivo* virulence of a Pap mutant in an APEC O1:K1:H7 strain (Kariyawasam and Nolan, 2009). Several studies have described a

higher prevalence of *papA* or *papC* genes in pathogenic strains or in isolates obtained from colibacillosis compared with non-pathogenic, fluff or faecal isolates (Wang et al., 2015; Zhao et al., 2019) in line with the inclusion of *papA* in our model.

The *irp2* (iron repressible gene associated with yersiniabactin synthesis) was not included in our VAG detection array, but knowing that the *irp2* and the *fyuA* (yersiniabactin receptor) iron-acquisition genes are closely associated with the high pathogenicity island (Wang et al., 2015), we tested our isolates for the presence of the *fyuA* gene, which was then included in our model. The inclusion of *fyuA* in our model confirms that both *irp2* and *fyuA* genes are significantly more frequently detected in highly pathogenic strains than in low and non-pathogenic ones, as suggested previously (Wang et al., 2015).

Four novel markers, not present in the above-mentioned schemes were included *iha, ireA, aec4* and *fli*CH4. Regarding the bifunctional enterobactin receptor adhesin protein *iha* gene, a previous study did not find a significant relationship between the APEC pathotype (high, intermediate, or low pathogenicity) and *iha* gene prevalence; a significant association was detected between the presence of *iha* and APEC phylogroups, but not with those of avian faecal *E. coli* (Johnson et al., 2008). The lethality score for one-day-old chicks of isolates from broiler chickens with colisepticaemia in Brazil tended to be associated with the presence of the *iha* gene ($p = 0.054$) (Barbieri et al., 2015).

The gene *ireA* (iron regulated outer membrane protein gene) is involved in iron acquisition, because it encodes a TonB-dependent receptor (Zhang et al., 2020). According to Johnson et al. (2008), there are significant associations between the prevalence of the chromosomal *ireA* gene and APEC pathotypes (high, intermediate, or low pathogenicity), APEC strains or avian *E. coli* strains of faecal origin, and APEC phylogroups or avian faecal *E. coli*. Conversely, Barbieri et al. (2015) demonstrated a non-significant relationship between the chick lethality

score and *ireA*. *ireA* was detected in 73% of 15 *E. coli* strains involved in vertebral osteomyelitis and arthritis in broilers in Brazil (Braga et al., 2016).

The *aec4* gene is chromosomally located between genes *xseE* and *yfgK* on a genomic island that is homologous to the island CS54, which is involved in intestinal colonization and persistence of *Salmonella* serotype Typhimurium strain ATCC 14028 (Kingsley et al., 2003). The *aec4* gene is significantly associated with serogroup O1 and O2 compared with serogroup O78, and present in some human ExPEC strains (Schouler et al., 2009). Furthermore, another study detected the *aec4* gene in 46.9% of 352 pathogenic strains, but in only 13.9% of 108 non-pathogenic ones (Schouler et al., 2012). *aec4* was initially included in the Schouler et al. (2012) scheme, but the addition of *aec4* in the D pattern in their final scheme did not improve results . The serotype-related gene *fli*CH4 was also identified to be a marker of colibacillosis strains. Little has been published about the importance of H4 in APEC. A recent study identified H4 in a high proportion (36%) of genomes of 125 strains isolated from chickens and ducks with obvious colibacillosis clinical symptoms in China (Chen et al., 2021). The emergence of two APEC strains belonging to clonal groups O111:H4-D-ST2085 and O111:H4-D-ST117 with high virulence-gene content and zoonotic potential has been reported in Spain (Mora et al., 2012). Moreover the human pandemic clone O25:H4-B2-ST131 has also been reported in diseased broilers by several authors in different countries (Ahmed et al., 2013; Barbieri et al., 2015).Two of the markers included in our scheme were associated with a colibacillosis protective effect. This was the case for the *arpA* gene, which is present in all *E. coli* except those belonging to phylogroups B2 and F (phylogroups including ExPEC strains), and common to all non-virulent *E. coli* (Clermont et al., 2019). A similar protective effect was found for the ATPase gene *eivC* that is part of the ETT2 (type III secretion system 2) of *E. coli*. According to Wang et al. (2016), ETT2 is found in the majority of *E. coli* strains, but multiple genetic mutations and deletions resulting in various isoforms are detected in APEC: all of the ETT2 loci in serotype O78 isolates were degenerate, whereas an intact ETT2 locus was present primarily in O1 and O2 serotype strains. Even though some studies indicated that *eiv*C is linked

to pathogenesis of APEC strains, the functional effects of ETT2 remain unclear (Fox et al., 2020).Several markers, such as *iss, hlyF, ompT, cva/cvi, sitA* or O78 serotype selected in other above-cited schemes, were not selected in our machine learning approach, because they were not found to be informative on the basis of the PCR results obtained with our set of isolates (for more information on these markers, see Sup file 1).

Overall, our set of 13 markers is consistent with the Stromberg et al. (2017) study which inoculated suspensions of chicken faecal isolates via the air sacs of adult chickens. At 2 days post-challenge, internal organs were cultured. Bacterial counts for three out of nine tested isolates exceeded those for the negative control strain MG1655. These three isolates possessed 3, 4 or 5 genes out of the 7 (*iha, papA, fyuA, ireA, iroN, iutA* and *tsh*) in common with our scheme. The six isolates for which bacterial counts did not significantly differ from the control strain had a mean of 1.8 of these genes.

The validation of our prediction model was based on the evaluation of the CELA of a limited number of strains. According to Gibbs et al. (2001), the CELA assay can discriminate between virulent and non-virulent avian *E. coli* isolates and is simpler to implement than other *in vivo* challenge models. Strains were chosen among the four categories based on their origin (colibacillosis or not) and predicted virulence/non-virulence status. Results showed that the seven predicted virulent strains, which were also predicted virulent according to Johnson et al. marker scheme, killing 6 to 9 embryos, meaning that the virulence predictions were correct. Six of the seven predicted non-virulent strains killed no more than three embryos, in line with our non-virulence predictions, but one killed six embryos, meaning that our prediction for this latter isolate was wrong. It should be noted that the three tested CN strains, of which two proved non-pathogenic for embryos, possessed 4 or 5 of the Johnson et al. markers. Thus, overall, according to the embryo lethality test, our virulence predictions were accurate for 13 out of 14 (93%) isolates.

Noteworthily, all of our isolates were obtained from the same geographic region (i.e., the western part of France) and during the same two-year period. APEC strains from diverse geographic regions in Europe may present notable different genetic characteristics (Cordoni et al., 2016). Moreover, there may be various groups of APECs, because APEC strains that cause diverse forms of colibacillosis, such as swollen head syndrome, coliform cellulitis, salpingitis or respiratory diseases, may not share the same virulence patterns, according to the pathogenesis of infection (Maturana et al., 2011). The combination of virulence genes reflects the adaptation to a specific niche, with particular adhesion, metabolism, transmission, survival, stress resistance properties. It is thus important to stress the fact that our model was based on a set of strains isolated from broilers and may not be appropriate for colibacillosis in poultry from other types production sectors (breeders, layers, turkeys, ducks, etc.). For ethical reasons, we could not validate the model with a large number of strains tested on a larger number of embryos. However, three strains (two CV (#1 and #2) and one CN (#9)) were used in a subcutaneous chick inoculation model and the results (clinical signs, lesions and mortality) confirmed that the two CV strains were pathogenic for chicks contrary to the CN strain (manuscript submitted).

Conclusion

Based on the high-throughput PCR characterization of a large number of isolates obtained in 80 broiler flocks (environment, faecal isolates or colibacillosis lesions), we used a machine learning procedure to develop a marker-based prediction scheme to evaluate the pathogenicity of isolates. The model, based on the bootstrap aggregating tree method and the presence/absence of 13 genetic markers associated with phylogroups, flagellar antigen and 11 virulence markers, has a specificity of 84% and a sensitivity of 85%. This online tool ("Prediction of Avian Pathogenic Escherichia coli based on 13 genetic/PCR markers", https://sbougeard.shinyapps.io/applishinypcr/) can help veterinarians, laboratory staff and

researchers to easily evaluate the pathogenicity of avian *E. coli* isolates on the basis of the presence/absence of these 13 virulence markers, which can be screened for using conventional PCR or high-throughput PCR, or can be deduced from whole-genome sequencing analysis. A tool allowing the identification of APEC strains is a prerequisite for implementing a better control of avian colibacillosis, with a judicious and rational use of antimicrobials in poultry**.**

Declarations of interest: none

Funding

This work was supported by the French Ministry of Agriculture (General Education and Research Department, grant ITAVI-DGER 16/110), and the French broiler industry association (*Comité interprofessionnel du poulet de chair*).

Acknowledgements

The authors are grateful to the veterinarians and farmers who participated in the study and the Finalab, Labocea, and Resalab laboratories for isolating the strains. They also thank Isabelle Pierre and Audrey Schmitz (ANSES, Ploufragan-Plouzané-Niort Laboratory) for help with the chicken embryo lethality test.

References

Ahmed, A.M., Shimamoto, T., Shimamoto, T., 2013. Molecular characterization of multidrug-resistant avian pathogenic *Escherichia coli* isolated from septicemic broilers. Internat. J. Med. Microbiol. 303, 475-483.

Barbieri, N.L., de Oliveira, A.L., Tejkowski, T.M., Pavanelo, D.B., Matter, L.B., Pinheiro, S.R., Vaz, T.M., Nolan, L.K., Logue, C.M., de Brito, B.G., Horn, F., 2015. Molecular characterization and clonal relationships among *Escherichia coli* strains isolated from broiler chickens with colisepticemia. Foodborne Pathog. Dis. 12, 74-83.

Braga, J.F.V., Chanteloup, N.K., Trotereau, A., Baucheron, S., Guabiraba, R., Ecco, R., Schouler, C., 2016. Diversity of *Escherichia coli* strains involved in vertebral osteomyelitis and arthritis in broilers in Brazil. BMC Vet. Res. 12, 140.

Chen, X., Liu, W., Li, H., Yan, S., Jiang, F., Cai, W., Li, G., 2021. Whole genome sequencing analysis of avian pathogenic *Escherichia coli* from China. Vet. Microbiol. 259.

Clermont, O., Dixit, O.V.A., Vangchhia, B., Condamine, B., Dion, S., Bridier-Nahmias, A., Denamur, E., Gordon, D., 2019. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. Environ. Microbiol. 21, 3107-3117.

Cordoni, G., Woodward, M.J., Wu, H., Alanazi, M., Wallis, T., La Ragione, R.M., 2016. Comparative genomics of European avian pathogenic *E. coli* (APEC). BMC Genomics 17.

Delannoy, S., Schouler, C., Souillard, R., Yousfi, L., Le Devendec, L., Lucas, C., Bougeard, S., Keita, A., Fach, P., Galliot, P., Balaine, L., Puterflam, J., Kempf, I., 2020. Diversity of *Escherichia coli* strains isolated from day-old broiler chicks, their environment and colibacillosis lesions in 80 flocks in France. Vet. Microbiol. 252,108923. https://doi.org/10.1016/j.vetmic.2020.108923

413 Dozois, C.M., Daigle, F., Curtiss, R., 3rd, 2003. Identification of pathogen-specific and conserved
414 genes expressed in vivo by an avian pathogenic *Escherichia coli* strain. Proc. Natl. Acad. Sci.
415 USA 100, 247-252.
416 Ewers, C., Janßen, T., Kießling, S., Philipp, H.C., Wieler, L.H., 2005. Rapid detection of virulence-
417 associated genes in avian pathogenic *Escherichia coli* by multiplex polymerase chain reaction.
418 Avian Dis. 49, 269-273.
419 Fox, S., Goswami, C., Holden, M., Connolly, J.P.R., Mordue, J., O'Boyle, N., Roe, A., Connor, M.,
420 Leanord, A., Evans, T.J., 2020. A highly conserved complete accessory *Escherichia coli* type III
421 secretion system 2 is widespread in bloodstream isolates of the ST69 lineage. Sci. Rep. 10,
422 4135.
423 Gao, Q., Jia, X., Wang, X., Xiong, L., Gao, S., Liu, X., 2015. The avian pathogenic *Escherichia coli* O2
424 strain E058 carrying the defined aerobactin-defective *iucD* or *iucDiutA* mutation is less
425 virulent in the chicken. Infect. Genet. Evol. 30, 267-277.
426 Gibbs, P.S., Petermann, S.R., Wooley, R.E., 2004. Comparison of several challenge models for studies
427 in avian colibacillosis. Avian Dis. 48, 751-758.
428 Guabiraba, R., Schouler, C., 2015. Avian colibacillosis: still many black holes. FEMS Microbiol.
429 Lett. 362, fnv118. https://doi.org/10.1093/femsle/fnv118
430 Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining,
431 Inference, and Prediction. https://web.stanford.edu/~hastie/ElemStatLearn/, 763 pages p.
432 Johnson, T.J., Wannemuehler, Y., Doetkott, C., Johnson, S.J., Rosenberger, S.C., Nolan, L.K., 2008.
433 Identification of minimal predictors of avian pathogenic *Escherichia coli* virulence for use as a
434 rapid diagnostic tool. J. Clin. Microbiol. 46, 3987-3996.
435 Kariyawasam, S., Nolan, L.K., 2009. Pap mutant of avian pathogenic *Escherichia coli* O1, an O1:K1:H7
436 strain, is attenuated in vivo. Avian Dis. 53, 255-260.

437 Kim, Y.B., Yoon, M.Y., Ha, J.S., Seo, K.W., Noh, E.B., Son, S.H., Lee, Y.J., 2020. Molecular
438 characterization of avian pathogenic *Escherichia coli* from broiler chickens with colibacillosis.
439 Poultry Sci. 99, 1088-1095.
440 Kingsley, R.A., Humphries, A.D., Weening, E.H., De Zoete, M.R., Winter, S., Papaconstantinopoulou,
441 A., Dougan, G., Baumler, A.J., 2003. Molecular and phenotypic analysis of the CS54 island of
442 *Salmonella enterica* serotype typhimurium: identification of intestinal colonization and
443 persistence determinants. Infec.Immun. 71, 629-640.
444 Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Software, 28, 1-26.
445 Maturana, V.G., de Pace, F., Carlos, C., Mistretta Pires, M., Amabile de Campos, T., Nakazato, G.,
446 Guedes Stheling, E., Logue, C.M., Nolan, L.K., Dias da Silveira, W., 2011. Subpathotypes of
447 Avian Pathogenic *Escherichia coli* (APEC) exist as defined by their syndromes and virulence
448 traits. Open Microbiol. J. 5, 55-64.
449 Mitchell, T., 1997. Machine Learning, McGraw Hill Edition, 414 p.
450 Mora, A., López, C., Herrera, A., Viso, S., Mamani, R., Dhabi, G., Alonso, M.P., Blanco, M., Blanco, J.E.,
451 Blanco, J., 2012. Emerging avian pathogenic *Escherichia coli* strains belonging to clonal
452 groups O111: H4-D-ST2085 and O111: H4-D-ST117 with high virulence-gene content and
453 zoonotic potential. Vet. Microbiol. 156, 347-352.
454 Newman, D.M., Barbieri, N.L., de Oliveira, A.L., Willis, D., Nolan, L.K., Logue, C.M., 2021.
455 Characterizing avian pathogenic *Escherichia coli* (APEC) from colibacillosis cases, 2018. Peer J
456 9. https://doi.org/10.7717/peerj.11025
457 Nolan, L.K., Vaillancourt, J.P., Barbieri, N.L., Logue, C.M. 2020. Colibacillosis, In: Swayne, D.E.,
458 Boulianne, M., Logue, C.M., McDougald, L.R., Nair, V., Suarez, D.L. (Eds.) Diseases of Poultry,
459 14th edn. Wiley-Blackwell, Hoboken, NJ, 770-830.
460 Parreira, V.R., Gyles, C.L., 2003. A novel pathogenicity island integrated adjacent to the thrW tRNA
461 gene of avian pathogenic *Escherichia coli* encodes a vacuolating autotransporter toxin. Infect.
462 Immun. 71, 5087-5096.

463 Rouquet, G., Porcheron, G., Barra, C., Reperant, M., Chanteloup, N.K., Schouler, C., Gilot, P., 2009. A
464 metabolic operon in extraintestinal pathogenic *Escherichia coli* promotes fitness under
465 stressful conditions and invasion of eukaryotic cells. J. Bacteriol. 191, 4427-4440.
466 Schouler, C., Schaeffer, B., Bree, A., Mora, A., Dahbi, G., Biet, F., Oswald, E., Mainil, J., Blanco, J.,
467 Moulin-Schouleur, M., 2012. Diagnostic strategy for identifying avian pathogenic *Escherichia*
468 *coli* based on four patterns of virulence genes. J. Clin. Microbiol. 50, 1673-1678.
469 Schouler, C., Taki, A., Chouikha, I., Moulin-Schouleur, M., Gilot, P., 2009. A genomic island of an
470 extraintestinal pathogenic *Escherichia coli* strain enables the metabolism of
471 fructooligosaccharides, which improves intestinal colonization. J. Bacteriol. 91, 388-393.
472 Skyberg, J.A., Horne, S.M., Giddings, C.W., Wooley, R.E., Gibbs, P.S., Nolan, L.K., 2003. Characterizing
473 Avian *Escherichia coli* Isolates with multiplex Polymerase Chain Reaction. Avian Dis. 47, 1441-
474 1447.
475 Souillard, R., Allain, V., Toux, J.Y., Lecaer, V., Lahmar, A., Tatone, F., Amenna-Bernard, A., Le Bouquin,
476 S., 2019. Synthèse des pathologies aviaires observées en 2018 par le Réseau National
477 d'Observations Épidémiologiques en Aviculture (RNOEA). Bulletin épidémiologique, santé
478 animale et alimentation 88, 1-5.
479 Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Stat. Soc 36.

480 Stromberg, Z.R., Johnson, J.R., Fairbrother, J.M., Kilbourne, J., Van Goor, A., Curtiss, R., 3rd, Mellata,
481 M., 2017. Evaluation of *Escherichia coli* isolates from healthy chickens to determine their potential
482 risk to poultry and human health. PLoS ONE 12. https://DOI: 10.1371/journal.pone.0180599

483 Trotereau, A., Schouler, C. 2019. Use of a Chicken Embryo Lethality Assay to Assess the Efficacy of
484 Phage Therapy. In Methods in Molecular Biology (Humana Press Inc.), 199-205.
485 Wang, J., Tang, P., Tan, D., Wang, L., Zhang, S., Qiu, Y., Dong, R., Liu, W., Huang, J., Chen, T., Ren, J., Li,
486 C., Liu, H.J., 2015. The pathogenicity of chicken pathogenic *Escherichia coli* is associated with

the numbers and combination patterns of virulence-associated genes. Open J. Vet. Med. 5, 243-254.

Wang, S., Liu, X., Xu, X., Zhao, Y., Yang, D., Han, X., Tian, M., Ding, C., Peng, D., Yu, S., 2016. *Escherichia coli* type III secretion system 2 (ETT2) is widely distributed in avian pathogenic *Escherichia coli* isolates from Eastern China. Epidemiol. Infect. 144, 2824-2830.

Wooley, R.E., Gibbs, P.S., Brown, T.P., Maurer, J.J., 2000. Chicken embryo lethality assay for determining the virulence of avian *Escherichia coli* isolates. Avian Dis. 44, 318-324.

Zhang, Z., Jiang, S., Liu, Y., Sun, Y., Yu, P., Gong, Q., Zeng, H., Li, Y., Xue, F., Zhuge, X., Ren, J., Dai, J., Tang, F., 2020. Identification of *ireA*, 0007, 0008, and 2235 as *TonB*-dependent receptors in the avian pathogenic *Escherichia coli* strain DE205B. Vet. Res. 51. https://doi.org/10.1186/s13567-020-0734-z

Zhao, S., Wang, C.L., Chang, S.K., Tsai, Y.L., Chou, C.H., 2019. Characterization of *Escherichia coli* isolated from day-old chicken fluff in Taiwanese hatcheries. Avian Dis. 63, 9-16.

Table 1: Four pathogenicity schemes for avian pathogenic *E. coli* (APEC)

Scheme (reference)	*iss*	*tsh*	*cvi* or *cva/cvi*	*iucC* or *iucD*	*iutA*	*papC*	*irp2*	*Vat*	*astA*	*hlyF*	*iroN*	*ompT*	P(F11)	frz_{orf4}	*sitA*	*aec26*	Serogroup O78
Skyberg *et al.*, 2003	+	+	+	+													
Ewers et al, 2005	+	+	+	+		+	+	+	+								
Johnson et al, 2008	+				+					+	+	+					
Schouler et al, 2012					+								+	+	+	+	+

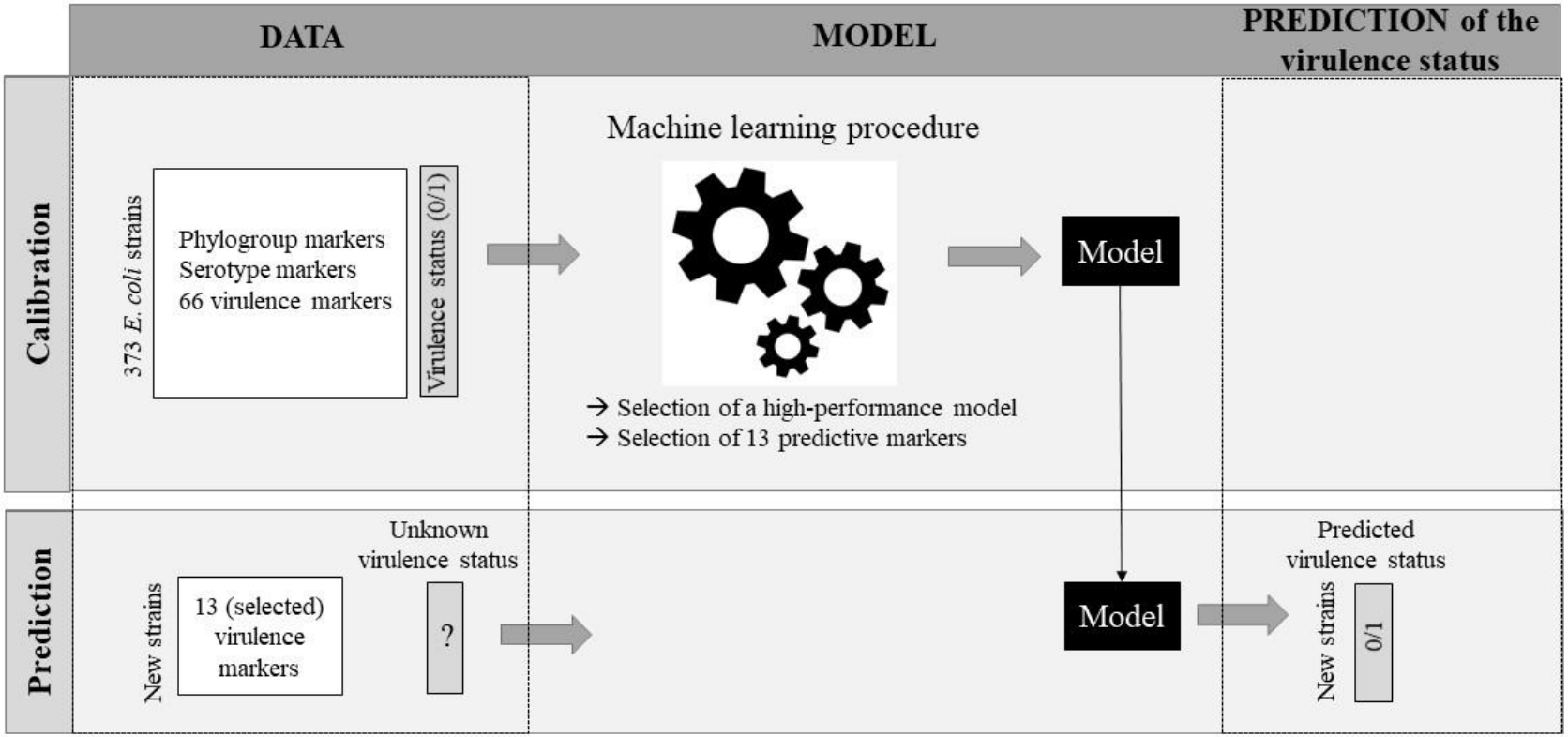


Figure 1. Diagram of the overall process of prediction of *E. coli* strain virulence status with respect to colibacillosis.

Table S1: PCR results obtained for the 373 strains –including 170 (presumed) non-virulent strains from control-farm environment, and 203 (presumed) virulent strains obtained from colibacillosis lesions from case-farm chickens

	Control (203)		Cases (170)		% positive for control strains	%positive for cases strains
	negative	positive	negative	positive		
O1a	169	1	195	8	0.6	3.9
O2	161	9	173	30	5.3	14.8
O8	152	18	199	4	10.6	2.0
O18	162	8	197	6	4.7	3.0
O23	166	4	203	0	2.4	0.0
O25	170	0	187	16	0.0	7.9
O78	170	0	178	25	0.0	12.3
O88	170	0	172	31	0.0	15.3
O153alter	167	3	202	1	1.8	0.5
*fli*CH4	163	7	115	88	4.1	43.3
*fli*CH7	146	24	189	14	14.1	6.9
*fli*CH8	165	5	181	22	2.9	10.8
*fli*CH21	164	6	195	8	3.5	3.9
*fli*CH25	166	4	203	0	2.4	0.0
K1	165	5	155	48	2.9	23.6
K5	168	2	198	5	1.2	2.5
arpA	26	144	117	86	84.7	42.4
*chu*AB2	150	20	143	60	11.8	29.6
*chu*AD	145	25	135	68	14.7	33.5
yjaA	128	42	127	76	24.7	37.4
TspE4.C2	79	91	157	46	53.5	22.7
astA	145	25	165	38	14.7	18.7
aec35	164	6	184	19	3.5	9.4
aec4	160	10	111	92	5.9	45.3
cba	156	14	170	33	8.2	16.3
clbN	170	0	197	6	0.0	3.0
cldB	170	0	197	6	0.0	3.0
clpV non sakai	120	50	175	28	29.4	13.8
cma	130	40	123	80	23.5	39.4
*csg*A1	34	136	70	133	80.0	65.5
*csg*A2	142	28	140	63	16.5	31.0
csgA3	157	13	196	7	7.6	3.4
*ecp*D1	16	154	0	203	90.6	100.0
ETT2.2	42	128	137	66	75.3	32.5
*fep*C	52	118	107	96	69.4	47.3
*fim*A1	93	77	91	112	45.3	55.2
*fim*A2	45	125	45	158	73.5	77.8

*fep*A3	15	155	30	173	91.2	85.2
frz$_{orf4}$	142	28	79	124	16.5	61.1
fyuA	119	51	68	135	30.0	66.5
hcp	98	72	147	56	42.4	27.6
hlyF	43	127	0	203	74.7	100.0
hra	151	19	154	49	11.2	24.1
ibeA	156	14	151	52	8.2	25.6
iha	147	23	94	109	13.5	53.7
ireA	150	20	107	96	11.8	47.3
iroN	54	116	2	201	68.2	99.0
iss1	20	150	0	203	88.2	100.0
iss2	19	151	0	203	88.8	100.0
iss3	20	150	0	203	88.2	100.0
iss4	49	121	3	200	71.2	98.5
iss5	14	156	2	201	91.8	99.0
iutA1	84	86	38	165	50.6	81.3
*nir*C	14	156	0	203	91.8	100.0
ompT1	55	115	20	183	67.6	90.1
ompT2	46	124	0	203	72.9	100.0
ompT3	43	127	0	203	74.7	100.0
pabB	14	156	21	182	91.8	89.7
papA	168	2	154	49	1.2	24.1
papG-allele II	166	4	159	44	2.4	21.7
pic	166	4	146	57	2.4	28.1
sat2	52	118	92	111	69.4	54.7
sitA	31	139	0	203	81.8	100.0
tia	165	5	170	33	2.9	16.3
tkt1	152	18	85	118	10.6	58.1
traT	28	142	14	189	83.5	93.1
tsh	113	57	26	177	33.5	87.2
vat	112	58	26	177	34.1	87.2
*yqi*C	124	46	74	129	27.1	63.5

Table 2: Prediction performance criteria according to the number of markers used for the four best performing methods to explain the presence of colibacillosis (standard deviation).

	Method	Sensitivity	Specificity	False negative	False positive
69 markers	kknn	0.93 (0.03)	0.88 (0.04)	0.09 (0.03)	0.09 (0.03)
	nnet	0.92 (0.03)	0.88 (0.06)	0.09 (0.03)	0.10 (0.04)
	rf	0.92 (0.03)	0.91 (0.05)	0.09 (0.03)	0.08 (0.03)
	treebag	0.92 (0.03)	0.91 (0.05)	0.08 (0.03)	0.07 (0.03)
13 markers	kknn	0.87 (0.04)	0.70 (0.09)	0.18 (0.04)	0.22 (0.05)
	nnet	0.85 (0.05)	0.82 (0.06)	0.18 (0.04)	0.15 (0.04)
	rf	0.84 (0.05)	0.85 (0.05)	0.18 (0.04)	0.13 (0.04)
	treebag	0.85 (0.04)	0.84 (0.06)	0.17 (0.04)	0.13 (0.04)

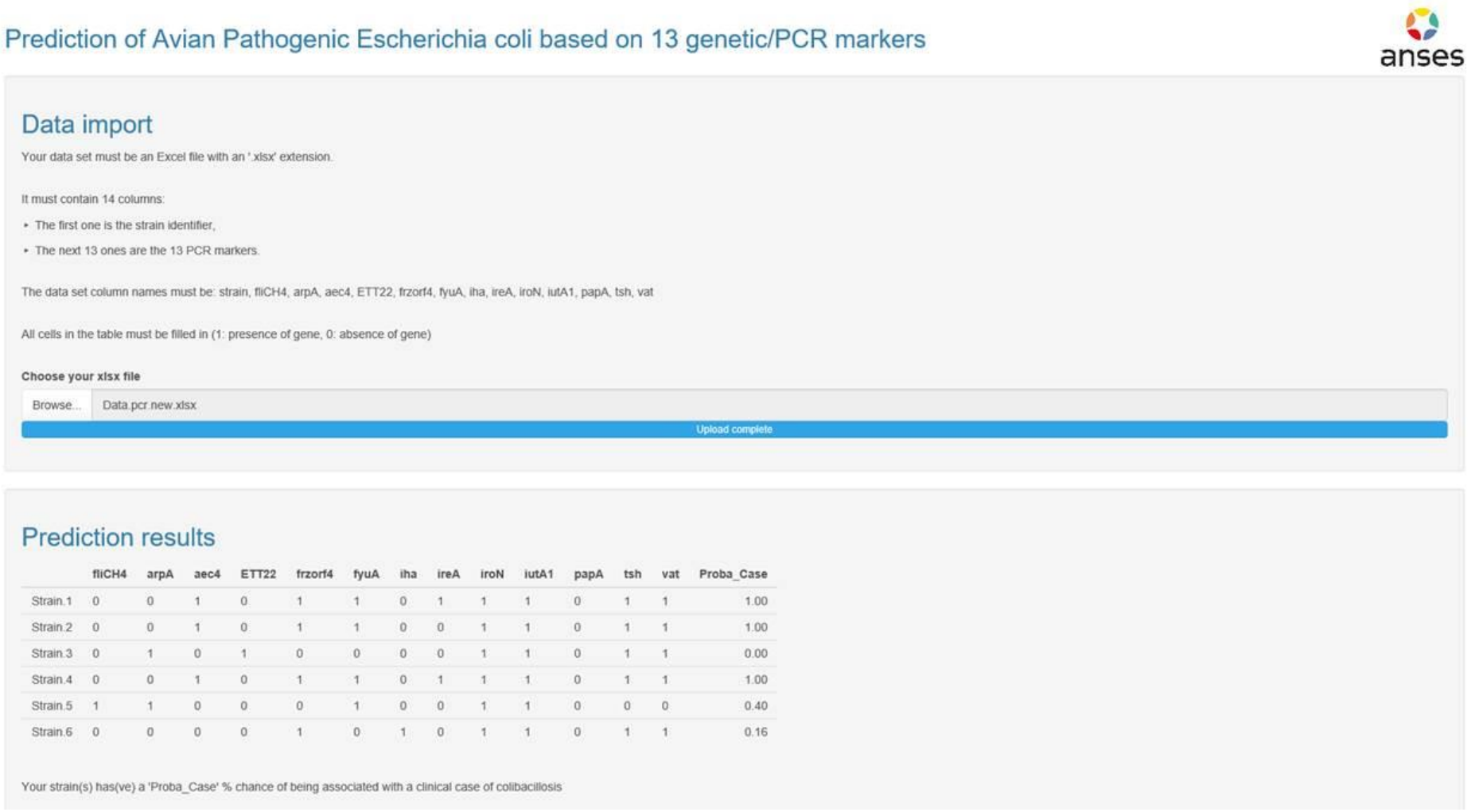
kknn: K-nearest neighbours; nnet: neural networks; rf: random forest; treebag: bootstrap aggregating tree

Table 3: Relevance of genetic variables for the four best performing methods

		Methods				
Variable type	Marker	kknn	nnet	rf	treebag	Total
Serotype	fliCH4	1*	1	1	1	4
	O88	0	1	1	1	3
Phylogroup	*arpA*	1	1	1	1	4
Virulence marker	*aec4*	1	1	1	1	4
	ETT2.2	1	1	1	1	4
	frz$_{orf4}$	1	1	1	1	4
	fyuA	1	1	1	1	4
	iha	1	1	1	1	4
	ireA	1	1	1	1	4
	iroN	1	1	1	1	4
	iutA1	1	1	1	1	4
	papA	1	1	1	1	4
	tsh	1	1	1	1	4
	vat	1	1	1	1	4
	cma	0	1	1	1	3
	hra	0	1	1	1	3
	sat2	0	1	1	1	3
	tkt1	1	0	1	1	3

*1: marker relevant for the method, 0, marker not informative for the method; kknn: K-nearest neighbours; nnet: neural networks; rf: random forest; treebag: bootstrap aggregating tree

Figure S1 : Presentation of the online tool for prediction of APEC based on 13 genetic / PCR markers



Prediction of Avian Pathogenic Escherichia coli based on 13 genetic/PCR markers

anses

Data import

Your data set must be an Excel file with an '.xlsx' extension.

It must contain 14 columns:

- The first one is the strain identifier,
- The next 13 ones are the 13 PCR markers.

The data set column names must be: strain, fliCH4, arpA, aec4, ETT22, frzorf4, fyuA, iha, ireA, iroN, iutA1, papA, tsh, vat

All cells in the table must be filled in (1: presence of gene, 0: absence of gene)

Choose your xlsx file

Browse... Data.pcr.new.xlsx

Upload complete

Prediction results

	fliCH4	arpA	aec4	ETT22	frzorf4	fyuA	iha	ireA	iroN	iutA1	papA	tsh	vat	Proba_Case
Strain.1	0	0	1	0	1	1	0	1	1	1	0	1	1	1.00
Strain.2	0	0	1	0	1	1	0	0	1	1	0	1	1	1.00
Strain.3	0	1	0	1	0	0	0	0	1	1	0	1	1	0.00
Strain.4	0	0	1	0	1	1	0	1	1	1	0	1	1	1.00
Strain.5	1	1	0	0	0	1	0	0	1	1	0	0	0	0.40
Strain.6	0	0	0	0	1	0	1	0	1	1	0	1	1	0.16

Your strain(s) has(ve) a 'Proba_Case' % chance of being associated with a clinical case of colibacillosis

Model performance

- Se (sensitivity) = 85% ; ability to predict strains associated with clinical cases of colibacillosis (= predicts clinical case for a case isolate)
- Sp (specificity) = 84% ; ability to predict strains associated with non-cases of colibacillosis (= predicts non-case for non-case isolate
- FN (false negative) = 17% ; prediction error rate of strains associated with non-cases of colibacillosis (= predicts non-case for case isolate)
- FP (false positive) = 13% ; prediction error rate of strains associated with clinical cases of colibacillosis (= predicts clinical case for non-case isolate)

Reference article

Lucas, C., Delannoy, S., Schouler, C., Souillard, R., Le Devendec, L., Lucas, P., Keita, A., Fach, P., Puterflam, J., Bougeard, S. & Kempf (Submitted). Description and validation of a new set of virulence PCR markers for Avian Patho

Author(s): Bougeard, S. & Lucas, C. (stephanie.bougeard@anses.fr)

Date: January 2022

Table 4: Results from univariate logistic regressions for each of the 13 selected markers

	Strains from the farm environment (control)		Strains from colibacillosis lesions (cases)		Coefficient	*p*-value
Marker	+	-	+	-		
*fli*CH4	7	163	88	115	2.88	0
arpA	144	26	86	117	-2.02	0
aec4	10	160	92	111	2.585	0
ETT2.2	128	42	66	137	-1.845	0
frz_{ofr4}	28	142	124	79	2.074	≈ 0
fyuA	51	119	135	68	1.533	≈ 0
Iha	23	147	109	94	2.003	≈ 0
ireA	20	150	96	107	1.906	≈ 0
iroN	116	54	201	2	3.846	≈ 0
iutA1	86	84	165	38	1.445	≈ 0
papA	2	168	49	154	3.286	≈ 0
tsh	57	113	177	26	2.602	≈ 0
vat	58	112	177	26	2.576	≈ 0

Table 5: Result of the chicken embryo lethality assay (CELA): characteristics of strains, titres of inocula and embryo mortality

Group	Code (flock)	Serotype	Phylogenetic group[a]	Detected markers	Prediction Johnson[b]	Titre[c]	Mortality[d] Day 0- Day 4
CV	1 (L1)	O2:K1	B2	*aec4-celB2-csgA2-ecpD1-ecpA1-ecpA2-fimA2-fepA3-frzorf4-fyuA-hlyF-ibeA-ireA-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT1-ompT2-ompT3-pabB-phoB-rstA-sat2-sitA-tkt1-traT-tsh-vat-yjjQ-YqiC*	5	28	8
	2 (L32)	O?:H4	F	*aec4-celB2-cma-csgA1-ecpD1-ecpA1-ecpA2-fimA1-fimA2-fepA3-frzorf4-fyuA-hlyF-iha-ireA-iroN-iss1-iss2-iss3-iss4-iss5-nirC-ompT1-ompT2-ompT3-pabB-phoB-pic-rstA-sat2-sitA-tkt1-traT-tsh-vat-yjjQ-yqiC*	4	54	8
	3 (L11)	Undetermined	D/E	*aec35-celB2-clpVnonsakai-cma-csgA3-ecpD1-ecpA1-ecpA2-fepC-fimA2-fepA3-*	5	136	9

				frzorf4-hcp-hlyF-ireA-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT2-ompT3-pabB-phoB-rstA-sitA-traT-tsh-vat-yjjQ-yqiC			
	4 (P8)	O1:K1:H7	B2	*astA-aec4-celB2-csgA2-ecpD1-ecpA1-ecpA2-fepC-fimA2-fepA3-frzorf4-fyuA-hlyF-hra-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT1-ompT2-ompT3-pabB-phoB-rstA-sat2-sitA-tia-tkt1-traT-tsh-vat-yjjQ-yqiC*	5	80	8
	5 (L24)	Undetermined	F	*aec4-celB2-csgA1-ecpD1-ecpA1-ecpA2-fimA1-fimA2-fepA3-frzorf4-hlyF-iha-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT1-ompT2-ompT3-pabB-phoB-pic-rstA-sat2-sitA-tkt1-traT-tsh-vat-yjjQ-yqiC*	5	58	6
EV	7 (L10)	Undetermined	A/C/Cl1	*celB2-clpVnonsakai-csgA1-csgA3-ecpD1-ecpA1-ecpA2*-ETT2.2-*fepC-fepA3-fyuA-hcp-hlyF-iha-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-*	5	81	8

				nirC-ompT1-ompT2-ompT3-pabB-phoB-rstA-sitA-traT-tsh-vat-yjjQ-			
	8 (P11)	O?:H4	F	*aec4-celB2-cma-csgA1-ecpD1-ecpA1-ecpA2-fimA1-fimA2-fepA3-frzorf4-fyuA-hlyF-ireA-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT1-ompT2-ompT3-pabB-phoB-papGalleleII-rstA-sat2-sitA-tia-tkt1-traT-tsh-vat-yjjQ-yqiC*	5	67	7
CN	9 (L5)	Undetermined	B1	*celB2-clpVnonsakai-csgA1-ecpD1-ecpA1-ecpA2*-ETT2.2-*fepC-fimA1-fimA2-fepA3-hcp-hlyF-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT1-ompT2-ompT3-pabB-phoB-rstA-sat2-sitA-traT-tsh-vat-yjjQ-*	4	113	1
	10 (P14)	Undetermined	B1	*celB2-cma-csgA1-ecpD1-ecpA1-ecpA2-*ETT2.2-*fepC-fimA1-fimA2-fepA3-hlyF-iroN-iss1-iss2-iss3-iss4-iss5-nirC-ompT1-*	4	90	2

				ompT2-ompT3-pabB-phoB-rstA-sat2-sitA-traT-yjjQ			
	11 (L11)	O?:H21	B1	*cba-celB2-cma-csgA1-ecpD1-ecpA1-ecpA2-*ETT2.2*-fepC-fepA3-hlyF-iha-iroN-iss1-iss2-iss3-iss4-iss5-iutA1-nirC-ompT1-ompT2-ompT3-pabB-phoB-rstA-sitA-traT-tsh-vat-yjjQ-*	5	60	6
EN	12 (P7)	Undetermined	A/C/Cl1	*celB2-csgA1-ecpD1-ecpA1-ecpA2*-ETT2.2-*fepC-fimA2-fepA3--iss5-nirC-pabB-phoB-rstA-traT-yjjQ-*	1	106	3
	13 (L20)	Undetermined	D/E	*celB2-csgA3-ecpD1-ecpA1-ecpA2*-ETT2.2-*fepC-fimA2-fepA3--iss5-nirC-ompT1-ompT3-pabB-phoB-rstA-sat2-traT-yjjQ-yqiC*	2	87	2
	14 (P30)	O?:H21	B1	*celB2-clpVnonsakai-csgA1-ecpD1-ecpA1-ecpA2*-ETT2.2-*fepC-fepA3-hcp--nirC-pabB-phoB-rstA-sat2-traT-yjjQ-*	0	74	0

15 (P34)	Undetermined	B1	*celB2-clpVnonsakai-csgA1-ecpD1-ecpA1-ecpA2*-ETT2.2-*fepC-fimA2-fepA3-hcp-iss1-iss2-iss3-iss5-ompT1-pabB-phoB-rstA-sat2-yjjQ-*	2	87	1

CV: isolated from colibacillosis, predicted virulent; EV: isolated from environment, predicted virulent; CN: isolated from colibacillosis, predicted non-virulent; EN: isolated from environment, predicted non-virulent

[a]Phylogroups, based on presence or absence of *arpA, chuA,* TspE4 and *yjaA* according to Clermont et al., 2019)

[b]Prediction Johnson: number of predictor markers of pathogenicity (among *hlyF, iroN, iss, iutA* and *ompT*) detected by PCR (Johnson et al., 2008)

[c]Titre of the inoculum (CFU/egg)

[d]Mortality: number of dead embryos out of 10 inoculated embryos, from the day of inoculation to the 4th day after inoculation