Major role of the high-pathogenicity island (HPI) in the intrinsic extra-intestinal virulence of *Escherichia coli* revealed by a genome-wide association study

Marco Galardini[1,6,*], Olivier Clermont[2], Alexandra Baron[2], Bede Busby[3], Sara Dion[2], Sören Schubert[4], Pedro Beltrao[1], Erick Denamur[2,5,*]

[1]EMBL-EBI, Wellcome Genome Campus, Cambridge, United Kingdom
[2] Université de Paris, IAME, UMR1137, INSERM, Paris, France
[3]Genome Biology Unit, EMBL, Heidelberg, Germany
[4]Max von Pettenkofer Institute of Hygiene and Medical Microbiology, Faculty of Medicine, LMU Munich, Germany
[5]APHP, Hôpitaux Universitaires Paris Nord Val-de-Seine, Site Bichat, Paris, France
[6]Current address: Biological Design Center, Boston University, Boston, MA 02215, USA

[*]**Corresponding authors:** mgala@bu.edu and erick.denamur@inserm.fr

Abstract:

The bacterium *Escherichia coli* is not only an important gut commensal, but also a common pathogen involved in both diarrheic and extra-intestinal diseases. To characterize the genetic determinants of extra-intestinal virulence we carried out an unbiased genome-wide association study (GWAS) on 234 commensal and extra-intestinal pathogenic strains representative of the species phylogenetic diversity, tested in a mouse model of sepsis. We found that the high-pathogenicity island (HPI), a ~35 kbp gene island encoding the yersiniabactin siderophore, is highly associated with death in mice, surpassing all other genetic factors by far. We validated the association *in vivo* by deleting key components of the HPI in strains in two phylogenetic backgrounds, and found that virulence is correlated with growth in the presence of various compounds including several antimicrobials, which hints at collateral sensitivities associated with intrinsic pathogenicity. This study points at the power of unbiased genetic approaches to uncover virulence determinants and the use of phenotypic data to generate new hypothesis on pathogenicity and phenotypic characteristics associated with it.

Introduction

Escherichia coli is both a commensal of vertebrates[1] and an opportunistic pathogen[2] involved in a wide range of intestinal and extra-intestinal infections. Extra-intestinal infections in humans represent a considerable burden[3], bloodstream infections (bacteraemia) being the most severe with a high attributable mortality of between 10-30%[4–6]. The regular increase over the last 20 years of *E. coli* bloodstream incidence[7] and antibiotic resistance[8] is particularly worrisome. The factors associated with high mortality are mainly linked to host conditions such as age, the presence of underlying diseases and to the portal of entry, with the urinary origin being more protective. These factors outweighing those directly attributable to the bacterial agent[4–6,9].

Nevertheless, the use of animal models has shown a great variability in the intrinsic extra-intestinal virulence potential of natural *E. coli* isolates. In a mouse model of sepsis where bacteria are inoculated subcutaneously, it has been clearly shown that the intrinsic virulence quantified by the number of animal deaths over the number of inoculated animals for a given strain is dependant on the number of virulence factors such as adhesins, toxins, protectins and iron capture systems[10–13]. One of the most relevant virulence factors is the so-called high-pathogenicity island (HPI), a 36 to 43 kb region encoding the siderophore yersiniabactin, a major bacterial iron uptake system[14]. The deletion of the HPI results in a decrease in the intrinsic virulence in the mouse model but in a strain-dependent manner[13–16], indicating complex interactions between the genetic background of the strains and the HPI.

The limitation of these gene KO studies is that they target specific candidate genes. Recently, the development of new approaches in bacterial genome-wide association studies (GWAS)[17–20] allows searching in an unbiased manner for genotypes associated to specific phenotypes such as drug resistance or virulence. In this context, we conducted a GWAS in 234 commensal and extra-intestinal pathogenic strains of *E. coli*, representing the species phylogenetic diversity, to search for traits associated to virulence in the mouse model of sepsis[21]. The strains belong to three large strain collections that span the species' major phylogroup diversity; the ECOR[22], IAI[10] and NILS[23] collections. All three collections contain commensal as well as extra-intestinal pathogenic *E. coli* (ExPEC), being defined as strains that possessed currently recognized extra-intestinal virulence factors and/or demonstrated enhanced virulence in an appropriate animal model of extra-intestinal infection[24]. Most importantly, strains from these collections have been recently sequenced and phenotyped across hundreds of growth conditions, including antibiotics and other chemical and physical stressors[25]. This data could then be used to find phenotype associations with virulence and to generate hypotheses on the function of genetic variants associated with the ExPEC phenotype and potential collateral sensitivities associated with them.

Results

GWAS identifies the high-pathogenicity island as the strongest driver of the extra-intestinal virulence phenotype

We studied three strain collections representative of the *E. coli sensu lato* phylogenetic diversity, i.e., *Escherichia* clade I in addition to phylogroup A, B1, C, D, E and F strains[26]. These strains encompass 90 commensal strains and 144 strains isolated in various extra-intestinal infections, mainly urinary tract infections and septicaemia[10,22,23]. To avoid any bias linked to host conditions, we assessed the strain virulence as its intrinsic extra-intestinal pathogenic potential using a well-calibrated mouse model of sepsis[10,21], expressed as the number of killed mice over the 10 inoculated per strain. In accordance with previous data, phylogroups B2, D and F have a higher proportion of virulent strains, as compared to phylogroups A and B1 (Figure 1A, Supplementary Table 1).

We used a bacterial GWAS method to associate *k*-mers to the virulence phenotype, allowing us to simultaneously test the contribution of core and accessory genome variation to pathogenicity[19]. It is generally understood that such methods require large sample sizes to have sufficient power, partly due to the need to break the long clonal frames typical of bacterial genomes; the appropriate sample size is also a function of the penetrance of the associated variants[18,27]. We ran simulations with an unrelated set of complete *E. coli* genomes and verified that our sample size was appropriate for variants with high penetrance (i.e. odds ratio above 5, Supplementary Figure 1, Methods). We reasoned that the genetic determinants of virulence are likely to have a relatively high penetrance, and that the strains used were genetically diverse, enough to break the clonal frame.

We uncovered a statistically significant association between 47,598 *k*-mers and the virulence phenotype, which were mapped back to 86 genes across the strains' pangenome (Figure 1B, Methods). A separate association using genes' presence absence patterns showed that the genes to which the associated *k*-mers mapped to have an odds ratio that far exceeds the required threshold we estimated from simulations (Figure 1C). Since the average minimum allele frequency (MAF) of associated k-mers is consistently around 36% (Figure 1B) and the distance between the genes they map to across all strains is around 1 kbp (Figure 1D), we concluded that the virulence phenotype is associated to the presence of a gene island. In fact, all genes belonging to the HPI had the vast majority of associated *k*-mers mapped to them (normalized hits >= 0.1, Figure 1E). Moreover, we found that the HPI structure was highly conserved across the 151 genomes that encode it (Supplementary Figure 2). We also observed that the distribution of known virulence factors doesn't match the virulence phenotype as closely as the HPI or has *k*-mers passing the association

threshold, further reinforcing the association results that the HPI is one of the main genetic factors behind virulence across phylogroups (Supplementary Figure 3).

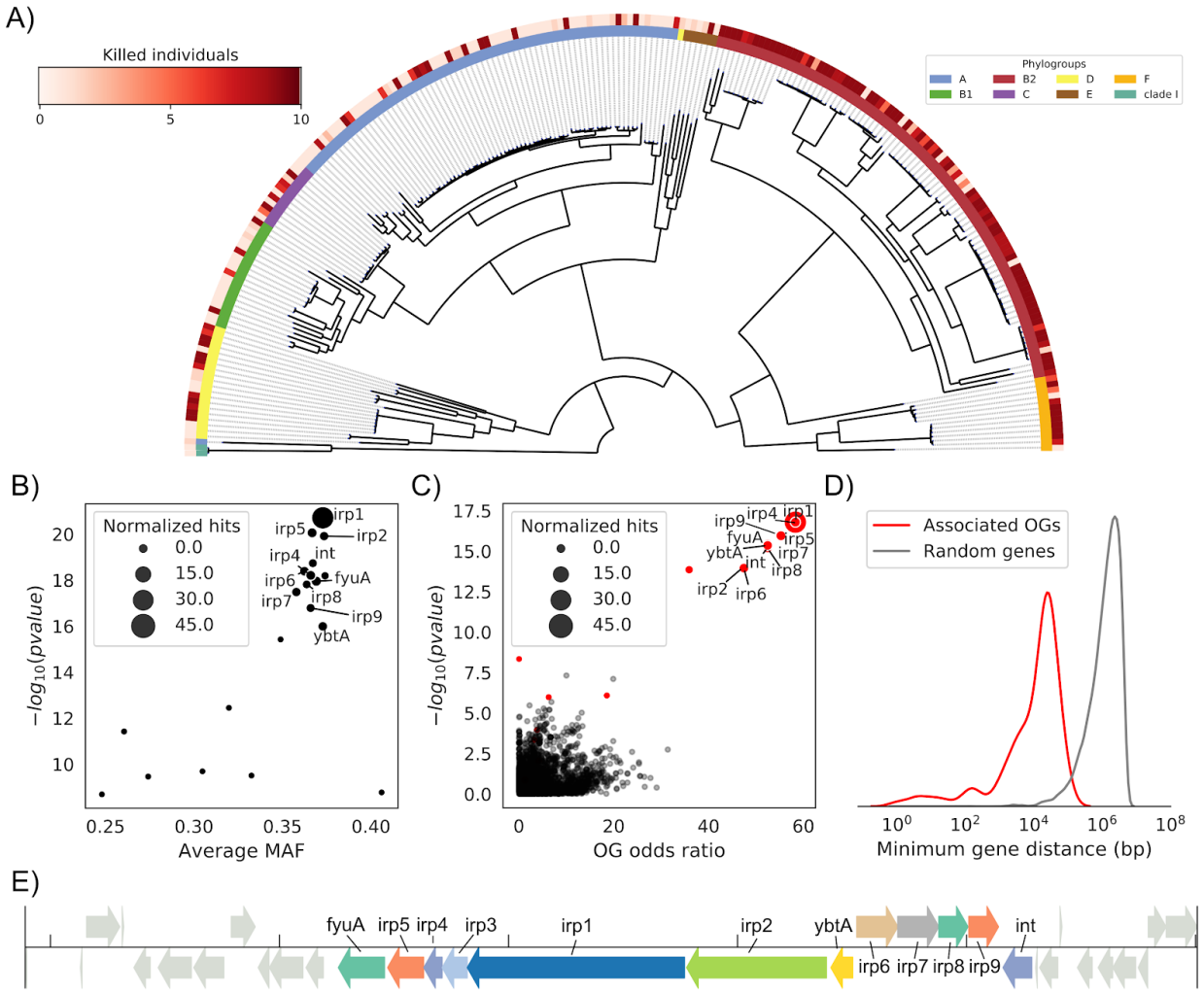


Figure 1. The HPI is strongly associated with the extra-intestinal virulence phenotype assessed in the mouse sepsis assay. A) Core genome phylogenetic tree of the *E. coli* strains used in this study rooted on *Escherichia* clade I strains. Outer ring reports virulence as the number of killed mice over the 10 inoculated per strain, inner ring the phylogroup each strain belongs to. B) Results of the k-mer association analysis: for each gene the minimum association p-value and average minimum allele frequency (MAF) across all mapped k-mers are reported. The normalized hits are computed by dividing the number of mapped kmers by the length of the gene. C) Results of the gene association analysis; each gene tested is represented. Genes from the k-mer association analysis are highlighted in red. D) The associated genes (normalized hits >= 0.1) belong to a gene cassette. OGs: orthologous groups. E) The HPI gene cassette structure in strain IAI39; all associated genes are highlighted.

KO gene experiments validate the role of the HPI in the extra-intestinal phenotype

The studies on the role of the HPI in experimental virulence gave contrasting results according to the strains' genetic background[13]. Among B2 phylogroup strains, HPI deletion in the 536 (ST127) strain did not have any effect in the mouse model of sepsis[28] whereas this deletion in the NU14 (ST95) strain dramatically attenuated virulence[13]. Two strains from this study belonging to B2 phylogroup/ST141 (IAI51 and IAI52) deleted in *irp1* have attenuated virulence in the same model[15]. To have a broader view of the role of the HPI in various genetic backgrounds, we constructed *irp2* deletion gene mutants in two strains of phylogroup D (NILS46) and A (NILS9) belonging to STs frequently involved in human bacteraemia (ST69 and ST10, respectively)[29]. We first verified that the wild-type strains strongly produced yersiniabactin, whereas the *irp2* mutants did not (Figure 2A). We then tested them in the mouse sepsis model and saw an increase in survival for both mutated strains (log-rank test p-value < 0.0001 and 0.0217 for strain NILS9 and NILS46, respectively, Figure 2B, Supplementary Table 2) with no significant difference between the survival profiles for the two mutants (log-rank test p-value > 0.1). We have therefore validated *in vivo* the causal link between the HPI and the virulence phenotype detected by the means of an unbiased association approach, which demonstrates the power and accuracy of bacterial GWAS.

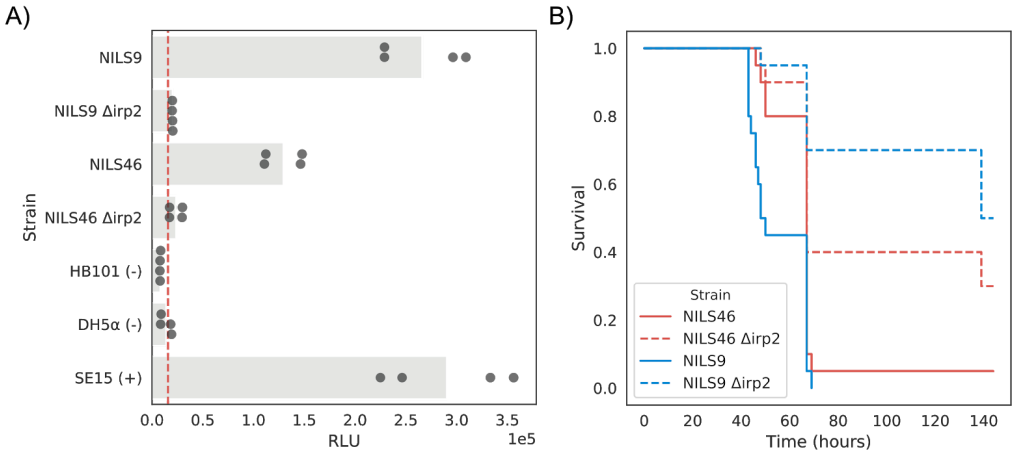


Figure 2. Phenotypic consequences of HPI's deletion. A) Deletion of HPI leads to a decrease in production of yersiniabactin. Production of yersiniabactin is measured using a luciferase-based reporter (Methods). Strains marked with a "-" and "+" sign indicates a negative and positive control, respectively. The red dashed line indicates an arbitrary threshold for yersiniabactin production, derived from the average signal recorded from the negative controls plus two standard deviations. RLU, relative light units. B) Deletion of HPI leads to an increase in survival after infection. Survival curves for wild-type strains and the corresponding *irp2* deletion mutant, built after infection of 20 mice for each strain.

High-throughput phenotypic data sheds light on HPI's function

The main function encoded by the HPI cassette is iron scavenging through the expression of the siderophore yersiniabactin[15], which has been previously validated in *E. coli* through knockout experiments[13]. In order to investigate other putative functions, we leveraged a previously-generated high-throughput phenotypic screening in an *E. coli* strain panel that largely overlaps with the strains used here (186 over 234)[25]. We observed a relatively high correlation between growth profiles in certain conditions and both virulence and presence of the HPI cassette (Figure 3A and 3B, Supplementary Table 3); given the strong association between the two, we observed similar conditions being correlated.

As expected, we found a correlation between growth on the iron-sequestering agent pentetic acid[30] and both HPI presence and virulence (Pearson's r: 0.47 and 0.36, respectively). We similarly, observed a correlation between pyocyanin, a redox-active phenazine compound being able to reduce Fe^{3+} to Fe^{2+} [31], and both HPI presence and virulence (Pearson's r: 0.36 and 0.29, respectively)

Interestingly, we found similarly strong correlations with growth on sub-inhibitory concentrations of several antibiotics, such as rifampicin, ciprofloxacin, amoxicillin and oxacillin, as well as other antimicrobial agents such as cerulenin and colicin. This might be due to the presence of resistance alleles and/or genes that are strongly associated with pathogenic strains, or might point to the role of iron homeostasis in intrinsic resistance to antibiotics[32]. As an example, quinolones bind Fe^{3+} on its pyridione ring, which is also involved in the interaction with its target, DNA gyrase[33]. Cell envelope permeability can also be modified in response to the presence of iron via two-component systems, rendering the cell more resistant[32]. On the other hand we found that growth in presence of indole at 2 mM in association with sub-inhibitory concentrations of cefsulodin and tobramycin antibiotics, but not with each of these compounds alone, was negatively correlated with both HPI presence and virulence. This might indicate a synergy between antibiotics and indole. In lysogeny broth, sub lethal concentrations of antibiotics increased the endogenous production of indole by the cells[34] and, at very high concentration (5 mM), indole induces the production of reactive oxygen species and is toxic for the cells[35]. This toxicity has been shown to be partly iron mediated due to the Fenton reaction[36], explaining that cells with increased import of extracellular iron due to the HPI might be more sensitive to these compounds. These associations suggest the potential for collateral sensitivities related to both intrinsic pathogenicity and the presence of the HPI.

Given the relatively large number of conditions correlated with both pathogenicity and HPI presence, we tested whether both features could be predicted from growth data. We used the commonly-used random forests machine learning algorithm with appropriate partitioning of input data to tune hyperparameters and reduce overfitting,

leading to two classifiers for virulence and presence of the HPI cassette with high predictive power (Figure 3C and 3D and Methods). We noted that prediction of HPI presence performs slightly better than virulence, possibly reflecting the complex nature of the latter phenotype. As expected, we found that conditions with relatively high correlation with both features have a higher weight in both classifiers (Figure 3E, Supplementary Table 4), which suggests that a subset of phenotypic tests might be sufficient to classify pathogenic strains. These results show how phenotypic data can be used to generate hypotheses over gene function and pathogenesis.

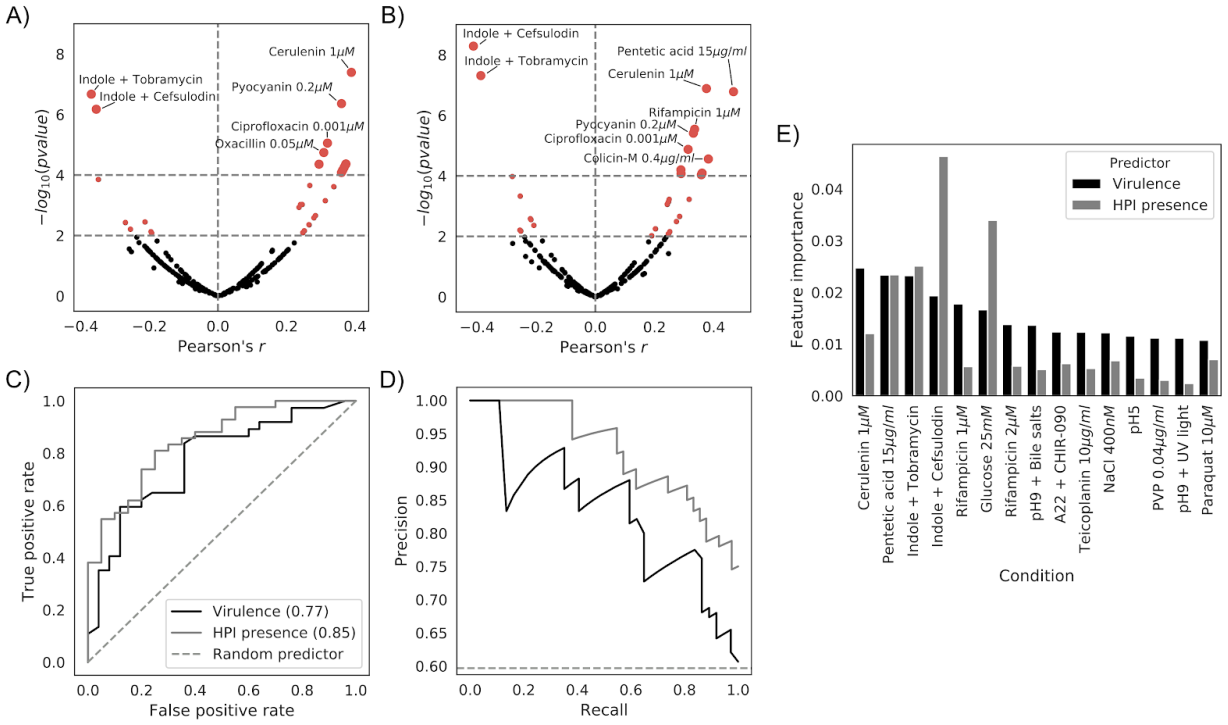


Figure 3. Growth profiles can predict virulence and HPI presence. A-B) Volcano plots for the correlation between the strains' growth profiles and: A) virulence levels and B) presence of the HPI. C-D) Use of the strains' growth profiles to build a predictor of virulence levels and presence of the HPI. C) Receiver operating characteristic (ROC) curve and D) Precision-Recall curve for the two tested predictors. E) Feature importance for the predictors, showing the top 15 conditions contributing to the virulence levels predictor.

Discussion

With the steady decline in the price of genomic sequencing and the increasing availability of molecular and phenotypic data for bacterial isolates, it has finally become possible to use statistical genomics approaches such as GWAS to uncover the genetic determinants of relevant phenotypes. Such approaches have the advantage of being unbiased, and can then be used to confirm previous targeted findings and potentially

uncover new factors, given sufficient statistical power. The accumulation of other molecular and phenotypic data can on the other hand uncover variables correlated with phenotype, which can be used to generate testable hypothesis on the function of genomic hits and potential collateral sensitivities associated with them. Given the rise of both *E. coli* extra-intestinal infections and antimicrobial resistance, we reasoned that the intrinsic virulence assessed in a calibrated mouse model of sepsis[10,21] is a phenotype worth exploring with such an unbiased approach.

We were able to confirm earlier reports on the importance of the HPI in extra-intestinal virulence[13–16,37], which showed the strongest signal in both the *k*-mer and accessory genome association analysis, and whose importance was validated *in vitro* and in an *in vivo* model. The distribution of the HPI within the species resulting from multiple horizontal gene transfers via homologous recombination[38] has probably facilitated its identification using GWAS. Additional genetic factors might have been overlooked by this analysis, due to the relatively small sample size; we however estimate that those putative additional factors might have a relatively low penetrance, based on our simulations in an independent dataset. As sequencing of bacterial isolates is becoming more common in clinical settings[39–41], we expect to be able to uncover these additional genetic factors in future studies.

The association between both the intrinsic virulence phenotype and the presence of the HPI and previously collected growth data allowed us to generate testable hypotheses on mechanism of pathogenesis and putative additional functions of the HPI. In particular we observed a relatively strong correlation between growth on various antimicrobial agents and both pathogenicity and HPI presence, which confirms the pressure to acquire resistance for these isolates, but also on the potential role of HPI and iron homeostasis on antimicrobial resistance[32]. *E. coli* mutants of *fur*, a transcriptional regulator that represses iron uptake systems, which accumulate high level of intracellular iron, have been shown to increase resistance to quinolones, aminoglycoside, tetracycline, rifampicin and amoxicillin[42]. The negative correlation with growth profiles in the presence of the indole associated to antibiotics points to the possible deleterious role of iron in the effect of sublethal doses of antibiotics. A vicious circle is rapidly established as antibiotics increase the production of indole[34], which in turn destabilise the membrane[35], further increasing the penetration of the antibiotics. The deletion of TonB, an iron transporter, increase resistance to the antibiotic, showing the role of reactive oxygen species generated by the Fenton reaction in the presence of iron[36]. Altogether, these data bring new light on the “liaisons dangereuses” between iron and antibiotics that could potentially be targeted[32].

We also demonstrate how growth data across several conditions can accurately distinguish pathogenic from non-pathogenic isolates, which could lead to the development of growth-based tests, which could complement and validate existing

diagnostic tools based on molecular and phenotypic data[43–45]. Taken together this analysis demonstrates how a data-centric approach can increase our knowledge of complex bacterial phenotypes and guide further empirical work on gene function and its relationship to intrinsic pathogenicity.

Materials and methods

Strains used

The full list of the 234 strains used in the association analysis, together with their main characteristics is reported in Supplementary Table 1. The genome sequences of all 234 strains is available through Figshare[46].

The construction of the *irp2* deletion mutants of the NILS9 and NILS46 strains was achieved following a strategy adapted from Datsenko and Wanner[47]. Primers used in the study are listed in Supplementary Table 5. In brief, primers used for gene disruption included 44-46 nucleotide homology extensions to the 5'- and 3' regions of the target gene, respectively, and additional 20 nucleotides of priming sequence for amplification of the resistance cassette on the template plasmids pKD4. The PCR product was then transformed into strains carrying the helper plasmid pKOBEG expressing the lambda red recombinase under control of an arabinose-inducible promoter[48]. Kanamycin resistant transformants were selected and further screened for correct integration of the resistance marker by PCR. For elimination of the antibiotic resistance gene, helper plasmid pCP20 was used according to the published protocol. PCR followed by Sanger sequencing of the mutants were performed to verify the deletion and the presence of the expected scar.

Yersiniabactin detection assay

Production of the siderophore yersiniabactin was detected and quantified using a luciferase reporter assay as described elsewhere[13,49]. Briefly, bacterial strains were cultivated in NBD medium for 24 hours at 37°C. Next, bacteria were pelleted by centrifugation and the supernatant was added to the indicator strain WR 1542 harbouring plasmid pACYC5.3L. All the genes necessary for yersiniabactin uptake are located on the plasmid pACYC5.3L, i.e. *irp6, irp7, irp8, fyuA, ybtA*. Furthermore, this plasmid is equipped with a fusion of the *fyuA* promoter region with the luciferase reporter gene. The amount of yersiniabactin can be quantified semi-quantitatively, as yersiniabactin-dependant upregulation of *fyuA* expression is determined by luciferase activity of the *fyuA-luc* reporter fusion.

Mouse virulence assay

Ten female mice OF1 of 14-16 g (4 week-old) from Charles River® (L'Arbresle, France) received a subcutaneous injection of 0.2 ml of bacterial suspension in the neck ($2{\cdot}10^8$

colony forming unit). Time to death was recorded during the following 7 days. Mice surviving more than 7 days were considered cured and sacrificed[10]. In each experiment, the *E. coli* CFT073 strain was used as a positive control killing all the inoculated mice whereas the *E. coli* K-12 MG1655 strain was used as a negative control for which all the inoculated mice survive[21]. For the mutant assays, 20 mice per strain were used to obtain statistical relevant data. The data was analysed using the lifeline package v0.21.0[50].

Association analysis

All genome-wide association analysis were carried out using pyseer, version v1.2.0[19]. All input genomes were re-annotated using prokka, version v1.13.3[51], to ensure uniform gene calls and excluding contigs whose size was above 200 base pairs. The core genome phylogenetic tree was generated using ParSNP[52] to generate the core genome alignment and gubbins v2.3.1[53] to generate the phylogenetic tree. The strain's pangenome was estimated using roary v3.12.0[54]. K-mers distributions from the input genome assemblies were computed using fsm-lite[18], with a minimum and maximum *k* value of 9 and 100, respectively. The association between both k-mers and pangenome and phenotype (expressed as number of mice killed post-infection) was carried out using the FastLMM[55] linear mixed-model implemented in pyseer, using a kinship matrix derived from the phylogenetic tree as population structure. For both association analysis we used the number of unique presence/absence patterns to derive an appropriate p-value threshold for the association likelihood ratio test ($2.90E^{-09}$ and $7.03E^{-06}$ for the *k*-mers and pangenome analysis, respectively). *K*-mers significantly associated with the phenotype were mapped back to each input genome using bwa mem v0.7.17-r1188[56] and betools v.2.27.1[57], using the pangenome analysis to collapse gene hits to individual groups of orthologs. A sample protein sequence for each groups of orthologs where at least on *k*-mer with size 20 or higher was mapped was extracted giving priority to strain IAI39 when available, given it was the only strain with a complete genome available; those sample sequences where used to search for homologs in the uniref50 database from uniprot[58] using blast v2.7.1+[59]. Each group of orthologs was then given a gene name using both available literature information and the results of the homology search. Distances between each pair of associated groups of orthologs was computed using the annotation files, using an equal number of random pairs as background.

Power simulations

Statistical power was estimated using an unrelated set of 548 complete *E. coli* genomes downloaded from NCBI RefSeq using ncbi-genome-download v0.2.9 on May 24th 2018. Each genome was subject to the same processing as the actual ones used in the real analysis (re-annotation, phylogenetic tree construction, pangenome estimation). The

gene presence/absence patterns were used to run the simulations, in a similar way as described in the original SEER implementation[18]. Briefly, for each sample size, a random subset of strains was selected, and the likelihood ratio test p-value threshold was estimated by counting the number of unique gene presence/absence patterns in the reduced roary matrix. For each odds ratio tested, a binary case-control phenotype vector was constructed for the strains subset using the following formulae:

$$P_{case|variant} = \frac{D_e}{MAF}$$

$$P_{case|novariant} = \frac{\frac{S_r}{S_r+1} - D_e}{1 - MAF}$$

Were S_r is the ratio of case/controls (set at 1 in these simulations), MAF as the minimum allele frequency of the target gene in the strains subset, and D_e the number of cases. pyseer's LMM model was then applied to the presence/absence vector of the target gene and the likelihood ratio test p-value was compared with the empirical threshold. The randomization was repeated 100 times and power was defined as the proportion of randomizations for each sample size and odds ratio whose p-value was below the threshold. The *pks2* and *fabG* genes were used as gene targets in the simulations, and both gave very similar results.

Correlations with growth profiles

The previously generated phenotypic data[25] for 186 over 234 strains were used to compute correlations with both the number of mice killed after infection and presence/absence of the HPI. The data was downloaded from the ecoref website (https://evocellnet.github.io/ecoref/download/) and the pearson correlation with the s-scores was computed together with the correlation p-value. Two predictors, one for virulence (number of killed mice post-infection) and one for presence of the HPI were built using the random forest classifier algorithm implemented in scikit-learn v.020.2[60], using the s-scores as predictors. The input was column imputed, and 33% of the observations were kept as a test dataset, using a "stratified shuffle split" strategy. The remainder was used to train the classifier, using a grid search to select the number of trees and the maximum number of features used, through 10 rounds of stratified shuffle split with validation set size of 33% the training set and using the F1 measure as score. The performance of the classifiers on the test set were assessed by computing the area under the receiver operating characteristic curve (ROC-curve).

Code and data availability

All input data and code used to run the analysis and generate the plots is available online at https://github.com/mgalardini/2018_ecoli_pathogenicity. Code is mostly based on the Python programming language and the following libraries: numpy v1.16.1[61], scipy v1.2.1[62], biopython v1.71[63,64], pandas v0.24.1[65], pybedtools v0.8.0[66], dendropy 4.4.0[67], ete3 v3.1.1[68], statsmodels v0.9.0[69], matplotlib v3.0.2[70], seaborn v0.9.0[71], jupyterlab v0.34.11[72] and snakemake v4.5.0[73].

Ethics statement

All animal experimentations were conducted following European (Directive 2010/63/EU on the protection of animals used for scientific purposes) and national recommendations (French Ministry of Agriculture and French Veterinary Services, accreditation A 75-18-05). The protocol was approved by the Animal Welfare Committee of the Veterinary Faculty in Lugo, University of Santiago de Compostela (AE-LU-002/12/INV MED.02/OUTROS 04).

Acknowledgements

We are grateful to Ivan Matic for discussion on the effect of indole. This work was partially supported by the “Fondation pour la Recherche Médicale” (Equipe FRM 2016, grant number DEQ20161136698).

References

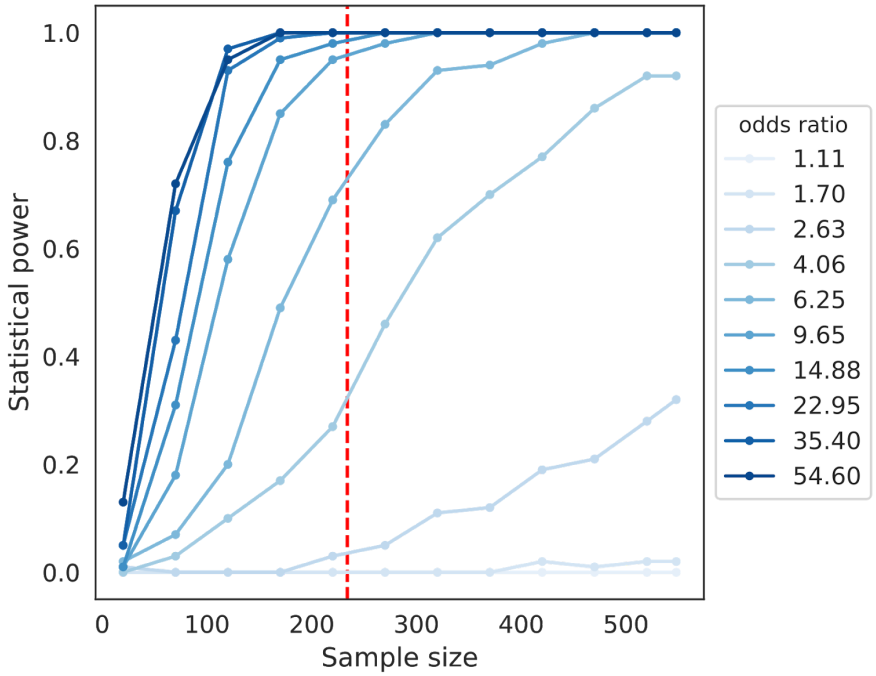
1. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal Escherichia coli. *Nat. Rev. Microbiol.* **8**, 207–217 (2010).
2. Croxen, M. A. & Brett Finlay, B. Molecular mechanisms of Escherichia coli pathogenicity. *Nature Reviews Microbiology* **8**, 26–38 (2010).
3. Russo, T. A. & Johnson, J. R. Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. *Microbes Infect.* **5**, 449–456 (2003).
4. Lefort, A. *et al.* Host Factors and Portal of Entry Outweigh Bacterial Determinants To Predict the Severity of Escherichia coli Bacteremia. *Journal of Clinical Microbiology* **49**, 777–783 (2011).
5. Burdet, C. *et al.* Escherichia coli bacteremia in children: age and portal of entry are the main predictors of severity. *Pediatr. Infect. Dis. J.* **33**, 872–879 (2014).
6. Abernethy, J. K. *et al.* Thirty day all-cause mortality in patients with Escherichia coli bacteraemia in England. *Clin. Microbiol. Infect.* **21**, 251.e1–8 (2015).
7. Vihta, K.-D. *et al.* Trends over time in Escherichia coli bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998–2016: a study of electronic health records. *The Lancet Infectious Diseases* **18**, 1138–1149 (2018).
8. Cassini, A. *et al.* Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **19**, 56–66 (2019).
9. Baudron, C. R. *et al.* Escherichia coli bacteraemia in adults: age-related differences in clinical and bacteriological characteristics, and outcome. *Epidemiology & Infection* **142**, 2672–2683 (2014).
10. Picard, B. *et al.* The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect. Immun.* **67**, 546–553 (1999).
11. Johnson, J. R. & Kuskowski, M. Clonal origin, virulence factors, and virulence. *Infect. Immun.* **68**,

424–425 (2000).
12. Tourret, J., Diard, M., Garry, L., Matic, I. & Denamur, E. Effects of single and multiple pathogenicity island deletions on uropathogenic Escherichia coli strain 536 intrinsic extra-intestinal virulence. *Int. J. Med. Microbiol.* **300**, 435–439 (2010).
13. Smati, M. *et al.* Strain-specific impact of the high-pathogenicity island on virulence in extra-intestinal pathogenic Escherichia coli. *Int. J. Med. Microbiol.* **307**, 44–56 (2017).
14. Schubert, S., Cuenca, S., Fischer, D. & Heesemann, J. High-pathogenicity island of Yersinia pestis in enterobacteriaceae isolated from blood cultures and urine samples: prevalence and functional expression. *J. Infect. Dis.* **182**, 1268–1271 (2000).
15. Schubert, S., Picard, B., Gouriou, S., Heesemann, J. & Denamur, E. Yersinia high-pathogenicity island contributes to virulence in Escherichia coli causing extraintestinal infections. *Infect. Immun.* **70**, 5335–5337 (2002).
16. Johnson, J. R. & Russo, T. A. Molecular Epidemiology of Extraintestinal Pathogenic Escherichia coli. *EcoSal Plus* **8**, (2018).
17. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* **1**, 1–8 (2016).
18. Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).
19. Lees, J., Galardini, M., Bentley, S. D. & Weiser, J. N. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *bioRxiv* (2018).
20. Jaillard, M. *et al.* A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* **14**, e1007758 (2018).
21. Johnson, J. R. *et al.* Experimental mouse lethality of Escherichia coli isolates, in relation to accessory traits, phylogenetic group, and ecological source. *J. Infect. Dis.* **194**, 1141–1150 (2006).
22. Ochman, H. & Selander, R. K. Standard reference strains of Escherichia coli from natural populations. *J. Bacteriol.* **157**, 690–693 (1984).
23. Bleibtreu, A. *et al.* The rpoS gene is predominantly inactivated during laboratory storage and undergoes source-sink evolution in Escherichia coli species. *J. Bacteriol.* **196**, 4276–4284 (2014).
24. Russo, T. A. & Johnson, J. R. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of Escherichia coli: ExPEC. *J. Infect. Dis.* **181**, 1753–1754 (2000).
25. Galardini, M. *et al.* Phenotype inference in an Escherichia coli strain panel. *Elife* **6**, 1–19 (2017).
26. 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).
27. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50 (2016).
28. Diard, M. *et al.* Pathogenicity-associated islands in extraintestinal pathogenic Escherichia coli are fitness elements involved in intestinal colonization. *J. Bacteriol.* **192**, 4885–4893 (2010).
29. Kallonen, T. *et al.* Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res.* (2017). doi:10.1101/gr.216606.116
30. Pippard, M. J., Jackson, M. J., Hoffman, K., Petrou, M. & Modell, C. B. Iron chelation using subcutaneous infusions of diethylene triamine penta-acetic acid (DTPA). *Scand. J. Haematol.* **36**, 466–472 (1986).
31. Cornelis, P. & Dingemans, J. Pseudomonas aeruginosa adapts its iron uptake strategies in function of the type of infections. *Front. Cell. Infect. Microbiol.* **3**, 75 (2013).
32. Ezraty, B. & Barras, F. The 'liaisons dangereuses' between iron and antibiotics. *FEMS Microbiol. Rev.* **40**, 418–435 (2016).
33. Uivarosi, V. Metal complexes of quinolone antibiotics and their applications: an update. *Molecules* **18**, 11153–11197 (2013).
34. Mathieu, A. *et al.* Discovery and Function of a General Core Hormetic Stress Response in E. coli Induced by Sublethal Concentrations of Antibiotics. *Cell Rep.* **17**, 46–57 (2016).
35. Garbe, T. R., Kobayashi, M. & Yukawa, H. Indole-inducible proteins in bacteria suggest membrane and oxidant toxicity. *Arch. Microbiol.* **173**, 78–82 (2000).

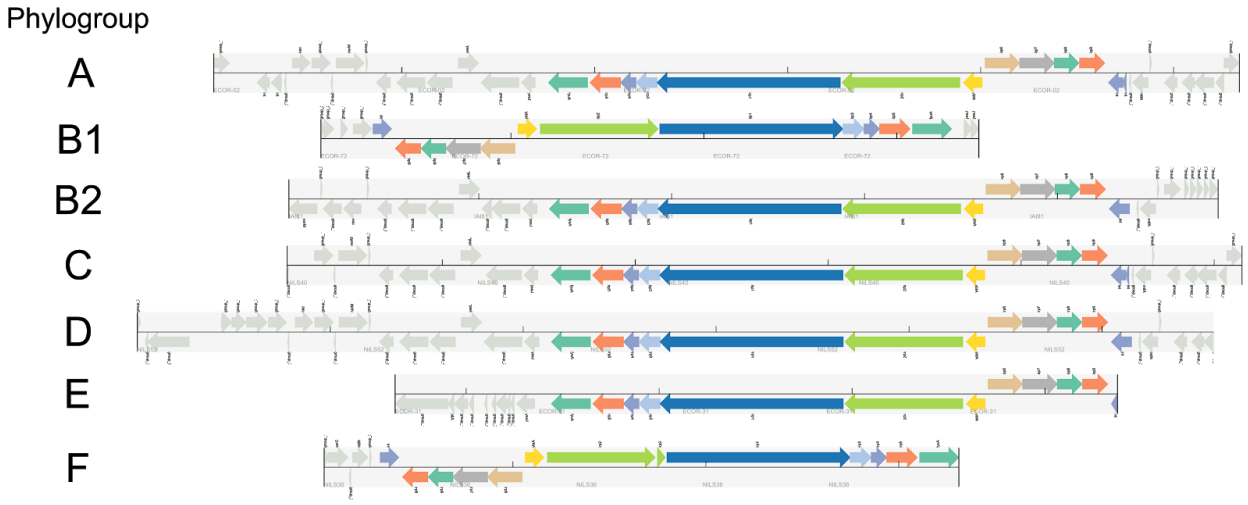
36. Giroux, X., Su, W.-L., Bredeche, M.-F. & Matic, I. Maladaptive DNA repair is the ultimate contributor to the death of trimethoprim-treated cells under aerobic and anaerobic conditions. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11512–11517 (2017).
37. Johnson, J. R. *et al.* Contribution of yersiniabactin to the virulence of an Escherichia coli sequence type 69 ('clonal group A') cystitis isolate in murine models of urinary tract infection and sepsis. *Microb. Pathog.* **120**, 128–131 (2018).
38. Schubert, S. *et al.* Role of Intraspecies Recombination in the Spread of Pathogenicity Islands within the Escherichia coli Species. *PLoS Pathog.* **5**, e1000257 (2009).
39. Fricke, W. F. & Rasko, D. A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat. Rev. Genet.* **15**, 49–55 (2014).
40. Quainoo, S. *et al.* Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin. Microbiol. Rev.* **30**, 1015–1063 (2017).
41. Tagini, F. & Greub, G. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 2007–2020 (2017).
42. Nichols, R. J. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
43. Tsalik, E. L., Bonomo, R. A. & Fowler, V. G., Jr. New Molecular Diagnostic Approaches to Bacterial Infections and Antibacterial Resistance. *Annu. Rev. Med.* **69**, 379–394 (2018).
44. Břinda, K., Callendrello, A., Cowley, L. & Charalampous, T. Lineage calling can identify antibiotic resistant clones within minutes. *bioRxiv* (2018).
45. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6**, 10063 (2015).
46. Galardini, M. Escherichia coli pathogenicity GWAS: input genome sequences. (2019). doi:10.6084/m9.figshare.8866259.v1
47. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6640–6645 (2000).
48. Chaveroche, M. K., Ghigo, J. M. & d'Enfert, C. A rapid method for efficient gene replacement in the filamentous fungus Aspergillus nidulans. *Nucleic Acids Res.* **28**, E97 (2000).
49. Martin, P. *et al.* Interplay between Siderophores and Colibactin Genotoxin Biosynthetic Pathways in Escherichia coli. *PLoS Pathogens* **9**, e1003437 (2013).
50. Davidson-Pilon, C. *et al. CamDavidsonPilon/lifelines: v0.21.0*. (2019). doi:10.5281/zenodo.2638135
51. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
52. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 524 (2014).
53. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
54. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
55. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
58. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–12 (2015).
59. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).
60. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
61. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
62. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open source scientific tools for Python. *http://www.scipy. org/* (2001).
63. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and

bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
64. Talevich, E., Invergo, B. M., Cock, P. J. & Chapman, B. a. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* **13**, 209 (2012).
65. McKinney, W. & Others. Data structures for statistical computing in Python. in *Proceedings of the 9th Python in Science Conference* **445**, 51–56 (2010).
66. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
67. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
68. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
69. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* **57**, 61 (SciPy society Austin, 2010).
70. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**, 90–95 (2007).
71. Waskom, M. *et al. mwaskom/seaborn: v0.9.0 (July 2018)*. (2018). doi:10.5281/zenodo.1313201
72. Kluyver, T. *et al.* Jupyter Notebooks-a publishing format for reproducible computational workflows. in *ELPUB* 87–90 (2016).
73. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600 (2018).

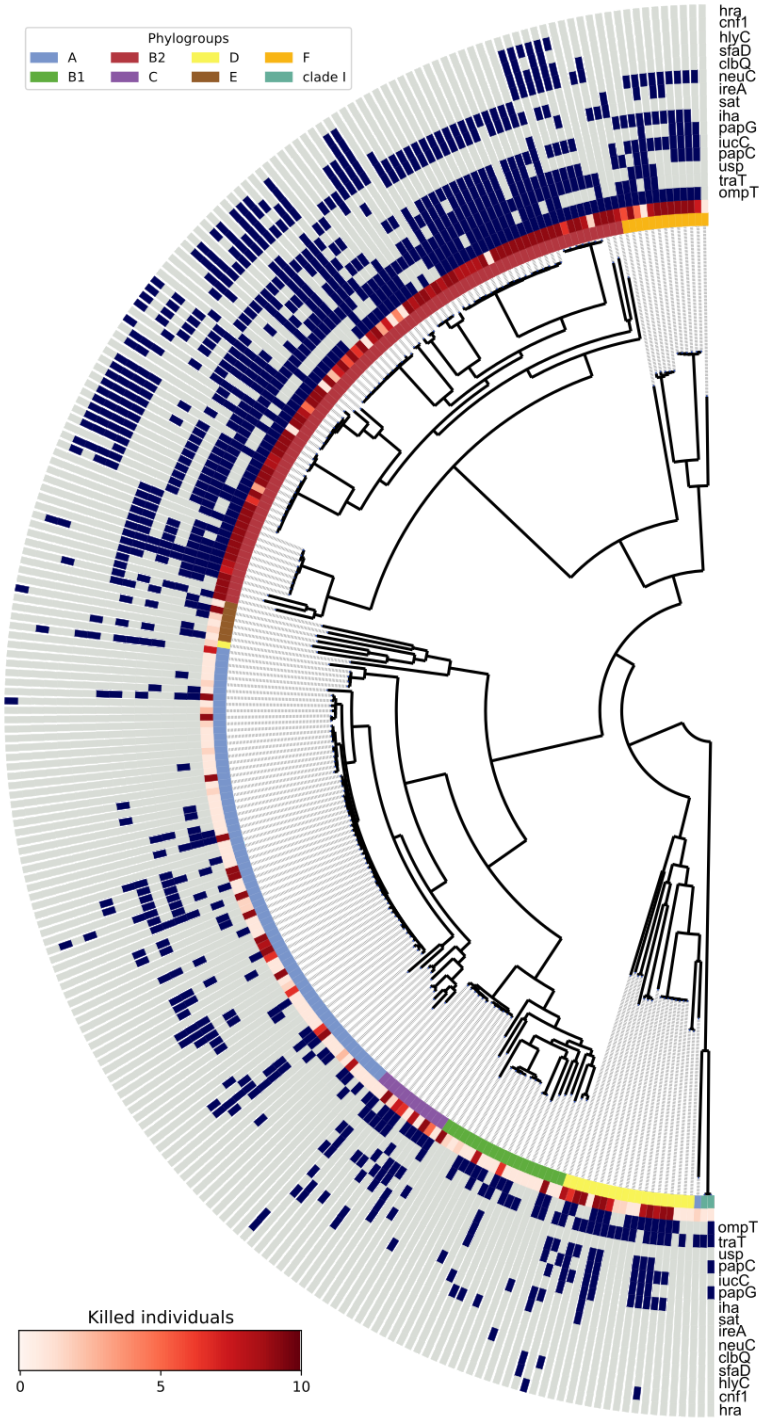
Supplementary Figures



Supplementary Figure 1. Simulations of statistical power on an unrelated set of complete *E. coli* genomes, using the *pks2* gene as target. The dotted red line indicate the sample size used in the actual analysis.



Supplementary Figure 2. HPI structure conservation across strains. One strain per phylogroup is shown, using the same color scheme as Figure 1E for each gene.



Supplementary Figure 3. Presence/absence patterns of known virulence factors other than genes belonging to the HPI. Blue indicates presence, light grey indicates absence. Phenotypes (number of killed mice) and phylogroup of each strain are reported as in Figure 1A.

Supplementary Information

Supplementary Table 1: Strains' information, including virulence phenotype

Supplementary Table 2: Survival analysis for NILS9 and NILS46 wild-type and HPI mutants

Supplementary Table 3: Correlation between growth on stress conditions (s-scores) and both virulence and presence of the HPI

Supplementary Table 4: Feature importance for each growth condition in the random forests predictor for virulence and HPI presence

Supplementary Table 5: List of PCR primers used in this study