Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus *Escherichia* 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]AP-HP, Laboratoire de Génétique Moléculaire, Hôpital 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 genus *Escherichia* is composed of several species and cryptic clades, including *E. coli*, which behave as a vertebrate gut commensal, but also as an opportunistic pathogen involved in both diarrheic and extra-intestinal diseases. To characterize the genetic determinants of extra-intestinal virulence within the genus, we carried out an unbiased genome-wide association study (GWAS) on 370 commensal, pathogenic and environmental strains representative of the *Escherichia* genus phylogenetic diversity and including *E. albertii* (n=7), *E. fergusonii* (n=5), *Escherichia* clades (n=32) and *E. coli* (n=326), 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 other associated genetic factors also related to iron uptake, such as the aerobactin and the *sitABCD* operons. We validated the association *in vivo* by deleting key components of the HPI in *E. coli* strains in two phylogenetic backgrounds, and found that virulence is correlated in *E. coli* with growth in the presence of various stressors including several antimicrobials, which hints at collateral sensitivities associated with intrinsic virulence. This study points to the major role of iron capture systems in the extra-intestinal virulence of the genus *Escherichia* and the collateral effects on cell growth of such systems.

Introduction

Members of the *Escherichia* genus are both commensals of vertebrates[1] and opportunistic pathogens[2] involved in a wide range of intestinal and extra-intestinal infections. Apart from the *E. coli* species, it is composed of the cryptic *Escherichia* clades, and the *E. fergusonii* and *E. albertii* species. The latter taxons are rarely isolated in humans but are more frequently found in the environment and avian species where they can cause intestinal infections[3–5]. In humans, extra-intestinal infections represent a considerable burden[6], with bloodstream infections (bacteraemia) being the most severe with a high attributable mortality of between 10-30%[7–10]. The regular increase over the last 20 years of *E. coli* bloodstream incidence[11] and antibiotic resistance[12] 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 outweigh those directly attributable to the bacterial agent[7–9,13].

Nevertheless, the use of animal models has shown a great variability in the intrinsic extra-intestinal virulence potential of natural *Escherichia* 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[14–18]. 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[19], which has also been shown to reduce the efficacy of innate immune cells to cause oxidative stress[20]. The deletion of the HPI results in a decrease in the intrinsic virulence in the mouse model in a strain-dependent manner[16,17,21], indicating complex interactions between the genetic background of each strain 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)[22–25] allows searching in an unbiased manner for genotypes associated with specific phenotypes such as drug resistance or virulence. In this context, we conducted a GWAS in 370 commensal and extra-intestinal pathogenic strains of *E. coli*, and related *Escherichia* clades, as well as *E. fergusonii* and *E. albertii*, representing the genus phylogenetic diversity, to search for traits associated with virulence in the mouse model of sepsis[26]. Most of the strains were isolated from a human host and are divided between commensals and extra-intestinal pathogenic. Most importantly, many (N=186) of these strains have been recently phenotyped across hundreds of growth conditions, including antibiotics and other chemical and physical stressors[27]. 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 extra-intestinal virulence 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 a 326 strain collection representative of the *E. coli* phylogenetic diversity, with strains belonging to phylogroups A (N=72), B1 (N=41), B2 (N=111), C (N=36), D (N=20), E (N=19), F (N=12) and G (N=15). To have a broader phylogenetic representation, we also included strains from *Escherichia* clades I to V (N=32) and the species *E. albertii* (N=7) and *E. fergusonii* (N=5)[28]. These strains encompass 170 commensal strains and 187 strains isolated in various extra-intestinal infections, mainly urinary tract infections and septicaemia[7,14,29–35]. 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[14,26], expressed as the number of killed mice over the 10 inoculated per strain. In accordance with previous data[14,26,36–38], phylogroup B2 is the most associated with the virulence phenotype ($2E^{-9}$ Wald test p-value, Figure 1A, Supplementary Table 1).

We used a bacterial GWAS method to associate unitigs to the virulence phenotype, allowing us to simultaneously test the contribution of core and accessory genome variation to pathogenicity[24]. 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 causal variants[23,39]. 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 5,214 unitigs and the virulence phenotype, which were mapped back to 81 genes across the strains' pangenome (Figure 1B, Supplementary Table 2, Methods). To understand whether the presence of these genes is directly associated with virulence or if its due to genetic variants such as SNPs we performed a separate association using genes' presence absence patterns, showing that most genes have an odds ratio that far exceeds the required threshold we estimated from simulations, as well as low association p-value (Figure 1C). Furthermore, 48 out of 81 genes with at least an associated unitig mapped to them have a frequency across strains that is highly correlated with that of the

107 associated unitigs (Figure 1D), indicating that it's the presence of those genes to be
108 associated with virulence. Genes belonging to the HPI had the lowest association
109 p-value by far ($<1E^{-28}$); the presence of two additional operons encoding for bacterial
110 siderophores (aerobactin[40] and *sitABCD*[41]) was also found to be associated with
111 virulence (Figure 1E). We found that the HPI structure was highly conserved across the
112 genomes that encode it (Supplementary Figure 2). We also observed that the
113 distribution of known virulence factors didn't match the virulence phenotype as closely
114 as the HPI or the aerobactin and *sitABCD* operons, or had unitigs passing the
115 association threshold (Supplementary Figure 3). The remaining 33 genes have a high
116 frequency in the pangenome, indicating the presence of genetic variants such as SNPs
117 present in core genes; the *mtfA* gene, which is involved in the regulation of
118 carbohydrate metabolism[42], had the lowest association p-value among those genes
119 ($1E^{-14}$).

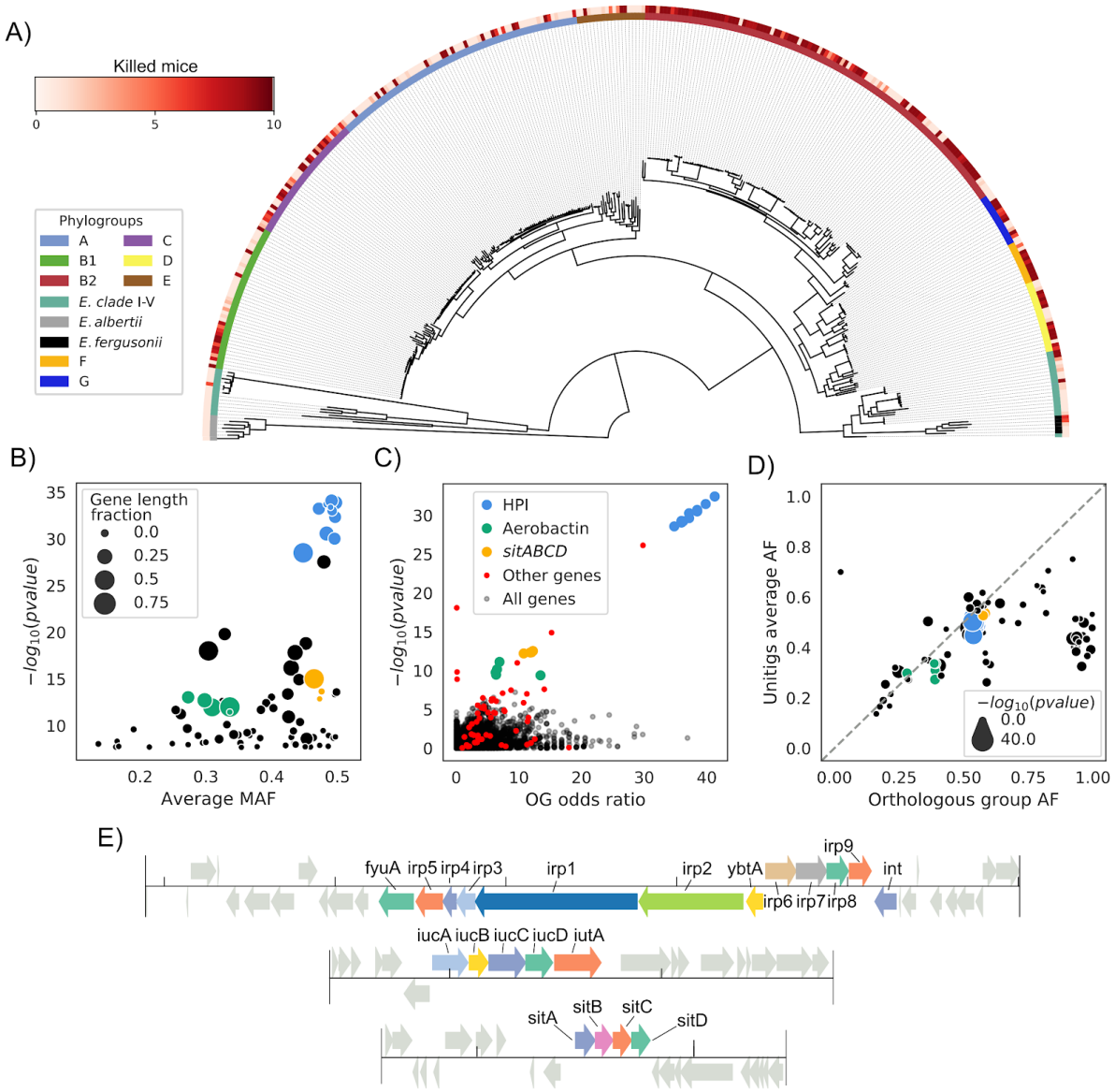


Figure 1. The HPI is strongly associated with the extra-intestinal virulence phenotype assessed in the mouse sepsis assay. A) Core genome phylogenetic tree of the *Escherichia* strains used in this study rooted on *E. albertii* strains. Outer ring reports virulence as the number of killed mice over the 10 inoculated per strain, inner ring the phylogroup or species each strain belongs to. B) Results of the unitigs association analysis: for each gene the minimum association p-value and average minimum allele frequency (MAF) across all mapped unitigs is reported. The gene length fraction is computed by dividing the total length of mapped unitigs by the length of the gene. C) Results of the gene presence/absence association analysis; each gene tested is represented. Genes from the unitigs association analysis are highlighted. OG: orthologous groups. D) Scatterplot of gene frequency versus frequency of associated unitigs; points on the diagonal indicate hits where the association is most likely due to a gene's presence rather than a SNP. E) The structure of the HPI and of the aerobactin and *sitABCD* operons 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[17]. Among B2 phylogroup strains, HPI deletion in the 536 (ST127) strain did not have any effect in the mouse model of sepsis[43] whereas this deletion in the NU14 (ST95) strain dramatically attenuated virulence[17]. Two strains from this study belonging to B2 phylogroup/ST141 (IAI51 and IAI52) deleted in *irp1* have attenuated virulence in the same model[21]. 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 (Sequence Types) frequently involved in human bacteraemia (ST69 and ST10, respectively)[44]. 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 3) 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.

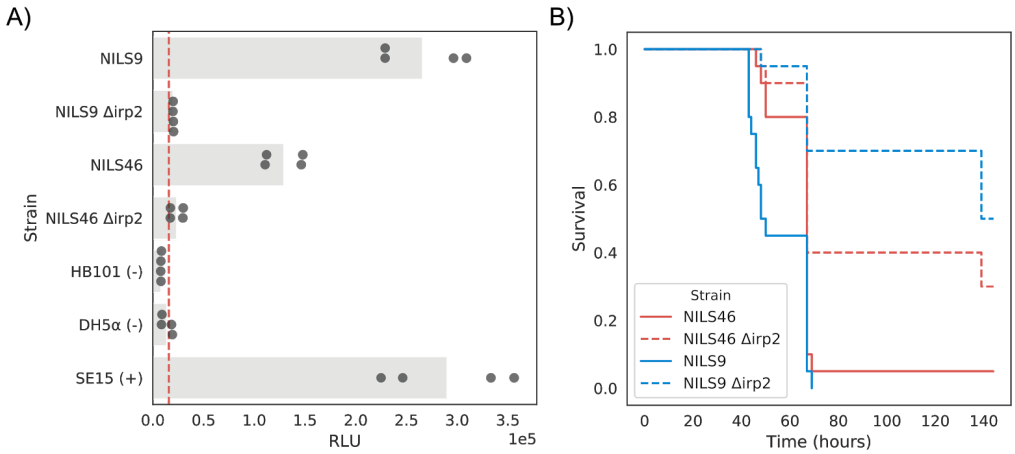


Figure 2. Phenotypic consequences of HPI deletion. A) Deletion of HPI leads to a decrease in production of yersiniabactin. Production of yersiniabactin is measured using a luciferase-based reporter (Methods). Strains marked with a "-" and "+" sign indicate a negative and positive control, respectively. 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 and other iron capture systems function

The main function encoded by the HPI cassette is iron scavenging through the expression of the siderophore yersiniabactin[21], which has been previously validated in *E. coli* through knockout experiments[17]. The aerobactin operon also encodes an iron chelator[40], while the *sitABCD* operon encodes a Mn^{2+}/Fe^{2+} ion transporter[41]. In order to investigate other putative functions of these operons and their relationship with virulence, 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 370)[27]. We observed a relatively high correlation between growth profiles in certain conditions and both virulence and presence of the HPI, aerobactin and *sitABCD* operons (Figure 3A-D, Supplementary Table 4).

As expected, we found a positive correlation between growth on the iron-sequestering agent pentetic acid[45] and both virulence and HPI/aerobactin/*sitABCD* presence (Pearson's r: 0.36, 0.48, 0.23 and 0.39, respectively). We also found that growth in the presence of bipyridyl, an iron chelator, was positively correlated with the presence of aerobactin (bipyridyl + tobramycin, Pearson's r: 0.30). We similarly observed a positive correlation between growth with pyocyanin, a redox-active phenazine compound being able to reduce Fe^{3+} to Fe^{2+}[46], and both HPI/aerobactin/*sitABCD* presence (Pearson's r: 0.35, 0.28, 0.26 and 0.27 respectively). All growth conditions agree with the iron scavenging function of the three gene clusters and their importance for virulence.

Interestingly, we also found similarly strong positive correlations with growth on sub-inhibitory concentrations of several antibiotics, such as rifampicin, ciprofloxacin, tetracycline and ß-lactams such as amoxicillin, oxacillin and meropenem, as well as other antimicrobial agents such as cerulenin and colicin. On the other hand we found that growth in presence of indole at 2 mM either in association with sub-inhibitory concentrations of cefsulodin and tobramycin, or alone at 40°C was negatively correlated with both virulence and HPI/aerobactin/*sitABCD* presence. Similar negative correlation was observed with aerobactin presence and the MAC13243 compound that increases outer membrane permeability[47]. These correlations might be due to the presence/absence of acquired 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[48]. To investigate these two hypotheses, we focused on tetracycline resistance, a common occurrence in the genus[33,37,49], and for which resistance genes can be easily found through sequence homology (Methods). We measured the correlation between the presence of tetracycline resistant genes, found in 26.8% of the strains, and virulence (Pearson's r: 0.16), as well with the presence of

either of the three iron capture systems (Pearson's r: 0.21, 0.33 and 0.24 for HPI, aerobactin and *sitABCD*, respectively), which we found to be all comparable with the direct correlation between growth on sub-inhibitory concentration of tetracycline and the presence of resistance genes (Pearson's r: 0.4). As such, both scenarios seem plausible to explain the relationship between growth in presence of tetracycline and both virulence and presence of HPI/aerobactin/*sitABCD*.

Given the relatively large number of conditions correlated with both virulence and presence of iron uptake systems, we tested whether these features could be predicted from growth profiles. We used the commonly-used random forests machine learning algorithm with appropriate partitioning of input data to tune hyperparameters and reduce overfitting, leading to four classifiers for virulence and presence of the HPI, aerobactin and *sitABCD* operons with high predictive power, with the exception of aerobactin (Figure 3E and 3F and Methods). We noted that prediction of the gene clusters 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 each feature have a higher weight across classifiers (Figure 3G, Supplementary Table 5), 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 for the function of virulence factors.

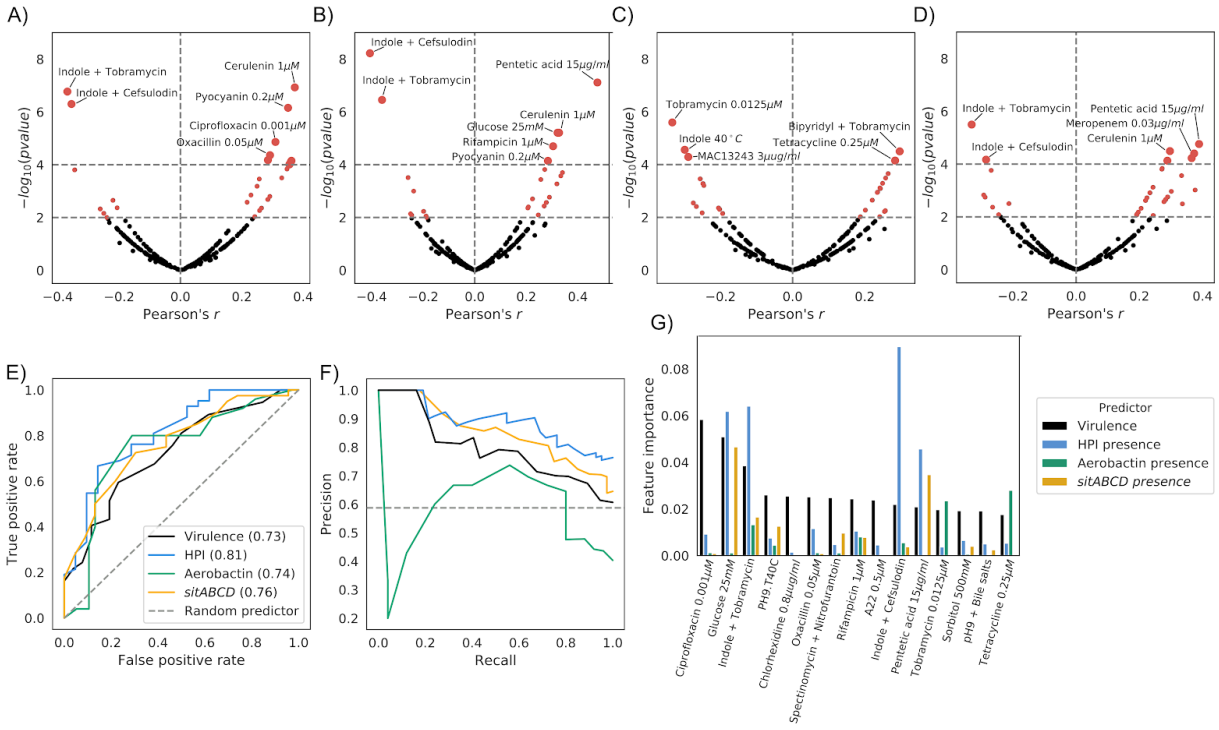


Figure 3. Growth profiles can predict virulence and presence of virulence factors. A-D) Volcano plots for the correlation between the strains' growth profiles and: A) virulence levels, B) presence of the HPI, C) presence of aerobactin, and D) presence of *sitABCD*. E-F) Use of the

strains' growth profiles to build a predictor of virulence levels and presence of the three iron uptake systems. E) Receiver operating characteristic (ROC) curve and F) Precision-Recall curve for the four tested predictors. G) 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 hypotheses 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[14,26] 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[17–19,21,50,51], which showed the strongest signal in both the unitigs and accessory genome association analysis, and whose importance was validated *in vitro* and in an *in vivo*. The distribution of the HPI within the species resulting from multiple horizontal gene transfers via homologous recombination[52] has probably facilitated its identification using GWAS. We associated additional genetic factors to intrinsic virulence, such as the aerobactin and *sitABCD* operons, both related to iron scavenging together with the HPI. We also found mutations in the *mtfA* core gene to be associated with virulence, although its role in virulence remains elusive. Additional 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[53–55], 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 virulence factors - such as the HPI - and previously collected growth data allowed us to generate testable hypotheses on mechanism of pathogenesis and putative additional functions of these factors. In particular we observed a relatively strong correlation between growth on various antimicrobial agents and both virulence and HPI/aerobactin/*sitABCD* presence, which seems to confirm the pressure to acquire resistance for these isolates, as exemplified for tetracycline resistance genes, but also

258 raises the question about the potential role of iron homeostasis on antimicrobial
259 resistance[48]. As we observed several classes of antibiotics with different molecular
260 targets, it is possible that this is due to a pleiotropic effect. *E. coli* Fur mutants, a
261 transcriptional regulator that represses iron uptake systems, which accumulate high
262 level of intracellular iron, have been shown to increase resistance to quinolones,
263 aminoglycoside, tetracycline, rifampicin and amoxicillin[56]. Cell envelope permeability
264 can also be modified in response to the presence of iron via two-component systems,
265 rendering the cell more resistant[48]. Sub-lethal doses of antibiotics have also been
266 shown to behave as a stressor for the cells that engage a general core hormetic stress
267 response consisting of increased energy production and translation, allowing growth
268 when nutrients are available despite the constant presence of the stressor[57]. When the
269 resources are instead limited, cell growth is stopped. It can be hypothesised that iron is
270 part of these limiting nutrients and that iron capture systems allow coping with the
271 stressors, i.e. the antibiotics.

272 The negative correlation between virulence and iron capture systems and growth
273 profiles in the presence of indole associated with stress conditions such as sub-lethal
274 doses of antibiotics or high temperature but not indole alone, points however to the
275 possible deleterious role of iron in such conditions. In lysogeny broth, sub-lethal
276 concentrations of antibiotics increase the endogenous production of indole[57] and, at
277 very high concentration (*e.g.* 5 mM), indole induces the production of reactive oxygen
278 species and is toxic for the cells[58]. A vicious circle is rapidly established as antibiotics
279 increase the production of indole[57], which in turn destabilises the membrane[58], further
280 increasing the penetration of the antibiotics. This toxicity has been shown to be partly
281 iron mediated due to the Fenton reaction, the deletion of TonB, an iron transporter,
282 increasing resistance to the antibiotic[59]. Interestingly, tobramycin is the antibiotic
283 involved in this negative correlation (Fig. 3A to D). Sub-lethal doses of this
284 aminoglycoside lead to an increase of reactive oxygen species in the bacterial cell in
285 relation to intra-cellular iron and Fenton reaction[60]. Thus, cells with increased import of
286 extracellular iron might be more sensitive to sub-lethal doses of antibiotics, suggesting
287 the potential for collateral sensitivities related to both intrinsic virulence and the
288 presence of the iron uptake systems. Altogether, these data bring new light on the
289 "liaisons dangereuses" between iron and antibiotics that could potentially be targeted[48].
290 More generally, they show that the presence of iron capturing systems can be either
291 advantageous or disadvantageous, depending on the growth conditions.

292 In conclusion, we showed the power of bacterial GWAS to identify major virulence
293 determinants in bacteria. Within the *Escherichia* genus, iron capture systems are the
294 main predictors of the intrinsic extra-intestinal virulence. Furthermore, this analysis
295 demonstrates how a data-centric approach can increase our knowledge of complex

296 bacterial phenotypes and guide further empirical work on gene function and its
297 relationship to intrinsic virulence.

298 Materials and methods

299 Strains used

300 The full list of the 370 strains used in the association analysis, together with their main
301 characteristics is reported in Supplementary Table 1. These strains belong to various
302 published collections: ECOR (N=71)[29], IAI (n=81)[14], NILS (N=82)[31], Septicoli (N=39)[10],
303 ROAR (N=30)[32], Guyana (N=12)[30], Coliville (N=8)[33], FN (N=6)[34], COLIRED (N=3)[35],
304 COLIBAFI (N=2)[7], correspond to archetypal strains (N=7) or are miscellaneous strains
305 from our personal collections (N=29). The isolation host is predominantly humans
306 (N=291), followed by animals (N=72) and isolated from the environment (N=6). One
307 hundred and seventy strains were commensal whereas five and 187 were responsible
308 of intestinal and extra-intestinal infections, respectively. The genomes of 295 strains
309 were previously available, while the remaining 75 were sequenced as part of this work
310 by Illumina technology as described previously[35]. The genome sequences of all strains
311 are available through Figshare[61].
312 The construction of the *irp2* deletion mutants of the NILS9 and NILS46 strains was
313 achieved following a strategy adapted from Datsenko and Wanner[62]. Primers used in
314 the study are listed in Supplementary Table 6. In brief, primers used for gene disruption
315 included 44-46 nucleotide homology extensions to the 5'- and 3' regions of the target
316 gene, respectively, and additional 20 nucleotides of priming sequence for amplification
317 of the resistance cassette on the template plasmids pKD4. The PCR product was then
318 transformed into strains carrying the helper plasmid pKOBEG expressing the lambda
319 red recombinase under control of an arabinose-inducible promoter[63]. Kanamycin
320 resistant transformants were selected and further screened for correct integration of the
321 resistance marker by PCR. For elimination of the antibiotic resistance gene, helper
322 plasmid pCP20 was used according to the published protocol. PCR followed by Sanger
323 sequencing of the mutants were performed to verify the deletion and the presence of
324 the expected scar.

325 Yersiniabactin detection assay

326 Production of the siderophore yersiniabactin was detected and quantified using a
327 luciferase reporter assay as described elsewhere[17,64]. Briefly, bacterial strains were
328 cultivated in NBD medium for 24 hours at 37°C. Next, bacteria were pelleted by
329 centrifugation and the supernatant was added to the indicator strain WR 1542
330 harbouring plasmid pACYC5.3L. All the genes necessary for yersiniabactin uptake are
331 located on the plasmid pACYC5.3L, i.e. *irp6, irp7, irp8, fyuA, ybtA*. Furthermore, this
332 plasmid is equipped with a fusion of the *fyuA* promoter region with the luciferase

333 reporter gene. The amount of yersiniabactin can be quantified semi-quantitatively, as
334 yersiniabactin-dependant upregulation of *fyuA* expression is determined by luciferase
335 activity of the *fyuA-luc* reporter fusion.

336 **Mouse virulence assay**

337 Ten female mice OF1 of 14-16 g (4 week-old) from Charles River® (L'Arbresle, France)
338 received a subcutaneous injection of 0.2 ml of bacterial suspension in the neck ($2 \cdot 10^8$
339 colony forming unit). Time to death was recorded during the following 7 days. Mice
340 surviving more than 7 days were considered cured and sacrificed[14]. In each experiment,
341 the *E. coli* CFT073 strain was used as a positive control killing all the inoculated mice
342 whereas the *E. coli* K-12 MG1655 strain was used as a negative control for which all the
343 inoculated mice survive[26]. For the mutant assays, 20 mice per strain were used to
344 obtain statistical relevant data. The data was analysed using the lifeline package
345 v0.21.0[65].

346 **Association analysis**

347 All genome-wide association analysis were carried out using pyseer, version v1.3.4[24].
348 All input genomes were re-annotated using prokka, version v1.14.5[66], to ensure uniform
349 gene calls and excluding contigs whose size was above 200 base pairs. The core
350 genome phylogenetic tree was generated using ParSNP[67] to generate the core genome
351 alignment and gubbins v2.3.5[68] to generate the phylogenetic tree. The strain's
352 pangenome was estimated using roary v3.13.0[69]. Unitigs distributions from the input
353 genome assemblies were computed using unitig-counter v1.0.5. The association
354 between both unitigs and pangenome and phenotype (expressed as number of mice
355 killed post-infection) was carried out using the FastLMM[70] linear mixed-model
356 implemented in pyseer, using a kinship matrix derived from the phylogenetic tree as
357 population structure. For both association analysis we used the number of unique
358 presence/absence patterns to derive an appropriate p-value threshold for the
359 association likelihood ratio test ($2.16E^{-08}$ and $5.45E^{-06}$ for the unitigs and pangenome
360 analysis, respectively). Unitigs significantly associated with the phenotype were mapped
361 back to each input genome using bwa mem v0.7.17-r1188[71] and betools v.2.29.2[72],
362 using the pangenome analysis to collapse gene hits to individual groups of orthologs. A
363 sample protein sequence for each groups of orthologs where at least one unitig with
364 size 20 or higher was mapped was extracted giving priority to strain IAI39 when
365 available, given it was the only strain with a complete genome available[73]; those sample
366 sequences where used to search for homologs in the uniref50 database from uniprot[74]
367 using blast v2.9.0[75]. Each group of orthologs was then given a gene name using both
368 available literature information and the results of the homology search.

369 **Power simulations**

370 Statistical power was estimated using an unrelated set of 548 complete *E. coli* genomes
371 downloaded from NCBI RefSeq using ncbi-genome-download v0.2.9 on May 24th 2018.
372 Each genome was subject to the same processing as the actual ones used in the real
373 analysis (re-annotation, phylogenetic tree construction, pangenome estimation). The
374 gene presence/absence patterns were used to run the simulations, in a similar way as
375 described in the original SEER implementation[23]. Briefly, for each sample size, a
376 random subset of strains was selected, and the likelihood ratio test p-value threshold
377 was estimated by counting the number of unique gene presence/absence patterns in
378 the reduced roary matrix. For each odds ratio tested, a binary case-control phenotype
379 vector was constructed for the strains subset using the following formulae:

380 $$P_{case|variant} = \frac{D_e}{MAF}$$

381 $$P_{case|novariant} = \frac{\frac{S_r}{S_r+1} - D_e}{1 - MAF}$$

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

390 **Correlations with growth profiles**

391 The previously generated phenotypic data[27] for 186 over 370 strains were used to
392 compute correlations with both the number of mice killed after infection and
393 presence/absence of the associated virulence factors. The data was downloaded from
394 the ecoref website (https://evocellnet.github.io/ecoref/download/) and the pearson
395 correlation with the s-scores was computed together with the correlation p-value.
396 Prediction of tetracycline resistance was carried out using staramr v0.7.1 with the
397 ResFinder database[76]. Four predictors, one for virulence (number of killed mice
398 post-infection) and one for presence of the HPI, aerobactin and the *sitABCD* operon
399 were built using the random forest classifier algorithm implemented in scikit-learn
400 v.022.0[77], using the s-scores as predictors. The input was column imputed, and 33% of
401 the observations were kept as a test dataset, using a "stratified shuffle split" strategy.
402 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 v1.17.3[78], scipy v1.4.0[79], biopython v1.75[80,81], pandas v0.25.3[82], pybedtools v0.8.0[83], dendropy 4.4.0[84], ete3 v3.1.1[85], statsmodels v0.10.2[86], matplotlib v3.1.2[87], seaborn v0.9.0[88], jupyterlab v1.2.4[89] and snakemake v5.8.2[90].

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* vol. 8 26–38 (2010).
3. Oaks, J. L. *et al.* Escherichia albertii in wild and domestic birds. *Emerg. Infect. Dis.* **16**, 638 (2010).
4. Clermont, O., Gordon, D. M., Brisse, S., Walk, S. T. & Denamur, E. Characterization of the cryptic Escherichia lineages: rapid identification and prevalence. *Environ. Microbiol.* **13**, 2468–2477 (2011).
5. Blyton, M. D. J. *et al.* Genetic Structure and Antimicrobial Resistance of Escherichia coli and Cryptic Clades in Birds with Diverse Human Associations. *Appl. Environ. Microbiol.* **81**, 5123–5133 (2015).
6. 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).
7. 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* vol. 49 777–783 (2011).
8. Burdet, C. *et al.* Escherichia coli bacteremia in children: age and portal of entry are the main

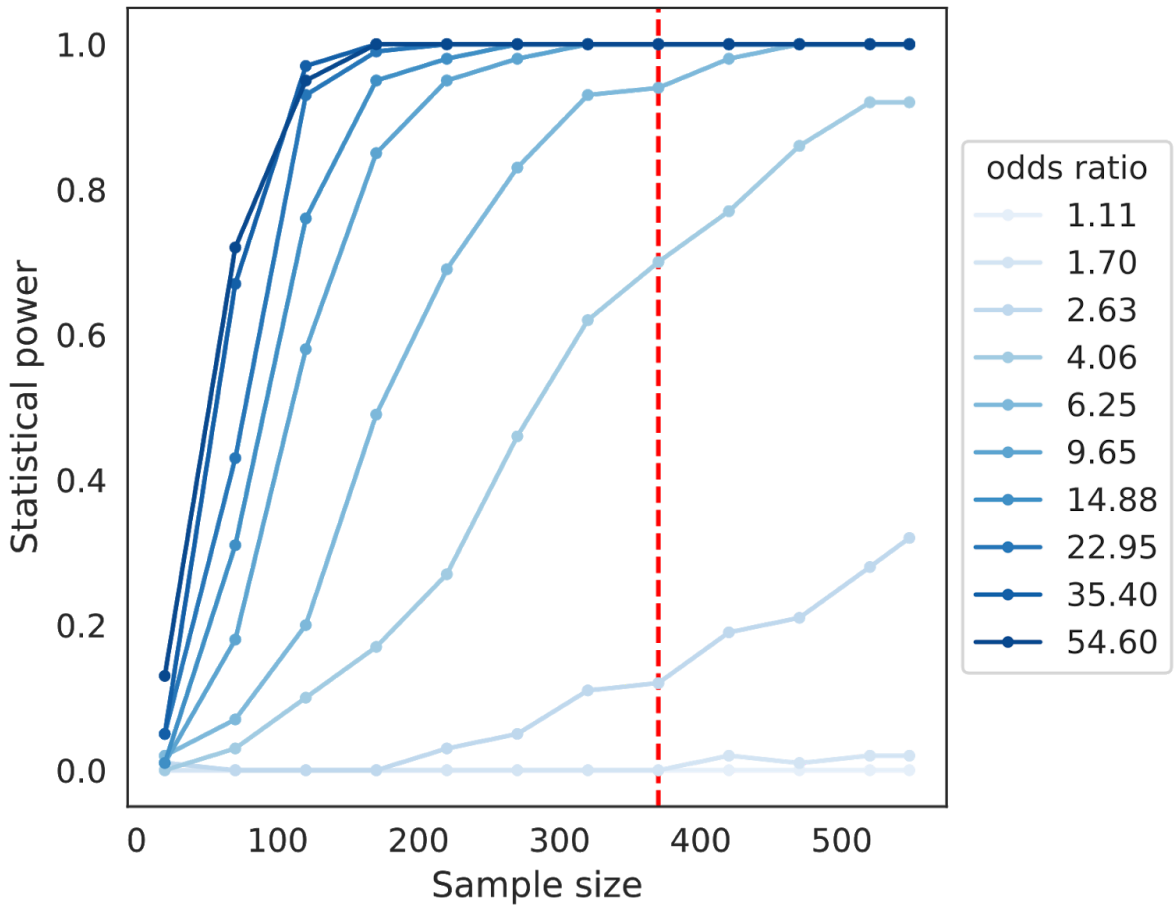
predictors of severity. *Pediatr. Infect. Dis. J.* **33**, 872–879 (2014).
9. Abernethy, J. K. *et al.* Thirty day all-cause mortality in patients with Escherichia coli bacteraemia in England. *Clin. Microbiol. Infect.* **21**, 251.e1–8 (2015).
10. V. de Lastours, C. Laouénan, G. Royer, E. Carbonnelle, R. Lepeule, M. Esposito-Farèse, O. Clermont, X. Duval, B. Fantin, F. Mentré, JW Decousser, E. Denamur, A. Lefort, on behalf of the SEPTICOLI group. Mortality in Escherichia coli bloodstream infections: antibiotic resistance still does not make it. *J. Antimicrob. Chemother.*
11. 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* vol. 18 1138–1149 (2018).
12. 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).
13. 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).
14. Picard, B. *et al.* The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect. Immun.* **67**, 546–553 (1999).
15. Johnson, J. R. & Kuskowski, M. Clonal origin, virulence factors, and virulence. *Infection and immunity* vol. 68 424–425 (2000).
16. 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).
17. 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).
18. Johnson, J. R. & Russo, T. A. Molecular Epidemiology of Extraintestinal Pathogenic Escherichia coli. *EcoSal Plus* **8**, (2018).
19. 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).
20. Paauw, A., Leverstein-van Hall, M. A., van Kessel, K. P. M., Verhoef, J. & Fluit, A. C. Yersiniabactin reduces the respiratory oxidative stress response of innate immune cells. *PLoS One* **4**, e8240 (2009).
21. 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).
22. 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).
23. Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).
24. Lees, J., Galardini, M., Bentley, S. D. & Weiser, J. N. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *bioRxiv* (2018).
25. 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).
26. 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).
27. Galardini, M. *et al.* Phenotype inference in an Escherichia coli strain panel. *Elife* **6**, 1–19 (2017).
28. 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).
29. Ochman, H. & Selander, R. K. Standard reference strains of Escherichia coli from natural populations. *J. Bacteriol.* **157**, 690–693 (1984).
30. Lescat, M. *et al.* Commensal Escherichia coli strains in Guiana reveal a high genetic diversity with host-dependant population structure. *Environ. Microbiol. Rep.* **5**, 49–57 (2013).
31. 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).
32. Skurnik, D. *et al.* Emergence of Antimicrobial-Resistant Escherichia coli of Animal Origin Spreading in Humans. *Mol. Biol. Evol.* **33**, 898–914 (2016).
33. Massot, M. *et al.* Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of Escherichia coli from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology* **162**, 642–650 (2016).
34. Nowrouzian, F. L. *et al.* Escherichia coli B2 Phylogenetic Subgroups in the Infant Gut Microbiota: Predominance of Uropathogenic Lineages in Swedish Infants and Enteropathogenic Lineages in Pakistani Infants. *Appl. Environ. Microbiol.* **85**, (2019).
35. Bourrel, A. S. *et al.* Colistin resistance in Parisian inpatient faecal Escherichia coli as the result of two distinct evolutionary pathways. *J. Antimicrob. Chemother.* **74**, 1521–1530 (2019).
36. Moissenet, D. *et al.* Meningitis caused by Escherichia coli producing TEM-52 extended-spectrum beta-lactamase within an extensive outbreak in a neonatal ward: epidemiological investigation and characterization of the strain. *J. Clin. Microbiol.* **48**, 2459–2463 (2010).
37. Ingle, D. J. *et al.* Biofilm formation by and thermal niche and virulence characteristics of Escherichia spp. *Appl. Environ. Microbiol.* **77**, 2695–2700 (2011).
38. Clermont, O. *et al.* Characterization and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* **21**, 3107–3117 (2019).
39. Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* **18**, 41–50 (2016).
40. Warner, P. J., Williams, P. H., Bindereif, A. & Neilands, J. B. ColV plasmid-specific aerobactin synthesis by invasive strains of Escherichia coli. *Infect. Immun.* **33**, 540–545 (1981).
41. Bearden, S. W., Staggs, T. M. & Perry, R. D. An ABC transporter system of Yersinia pestis allows utilization of chelated iron by Escherichia coli SAB11. *J. Bacteriol.* **180**, 1135–1147 (1998).
42. Becker, A.-K. *et al.* YeeI, a novel protein involved in modulation of the activity of the glucose-phosphotransferase system in Escherichia coli K-12. *J. Bacteriol.* **188**, 5439–5449 (2006).
43. 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).
44. 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.
45. 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).
46. 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).
47. Muheim, C. *et al.* Increasing the permeability of Escherichia coli using MAC13243. *Scientific Reports* vol. 7 (2017).
48. Ezraty, B. & Barras, F. The 'liaisons dangereuses' between iron and antibiotics. *FEMS Microbiol. Rev.* **40**, 418–435 (2016).
49. Mazel, D., Dychinco, B., Webb, V. A. & Davies, J. Antibiotic resistance in the ECOR collection: integrons and identification of a novel aad gene. *Antimicrob. Agents Chemother.* **44**, 1568–1574 (2000).
50. 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).
51. Johnson, J. R. *et al.* Accessory Traits and Phylogenetic Background Predict Escherichia coli Extraintestinal Virulence Better Than Does Ecological Source. *J. Infect. Dis.* **219**, 121–132 (2019).
52. Schubert, S. *et al.* Role of Intraspecies Recombination in the Spread of Pathogenicity Islands within the Escherichia coli Species. *PLoS Pathog.* **5**, e1000257 (2009).
53. Fricke, W. F. & Rasko, D. A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat. Rev. Genet.* **15**, 49–55 (2014).
54. Quainoo, S. *et al.* Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial

Outbreak Analysis. *Clin. Microbiol. Rev.* **30**, 1015–1063 (2017).
55. Tagini, F. & Greub, G. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 2007–2020 (2017).
56. Nichols, R. J. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
57. 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).
58. Garbe, T. R., Kobayashi, M. & Yukawa, H. Indole-inducible proteins in bacteria suggest membrane and oxidant toxicity. *Arch. Microbiol.* **173**, 78–82 (2000).
59. 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).
60. Baharoglu, Z., Krin, E. & Mazel, D. RpoS plays a central role in the SOS induction by sub-lethal aminoglycoside concentrations in Vibrio cholerae. *PLoS Genet.* **9**, e1003421 (2013).
61. Galardini, M. Escherichia coli pathogenicity GWAS: input genome sequences (updated). (2020) doi:10.6084/m9.figshare.11879340.v1.
62. 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).
63. 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).
64. Martin, P. *et al.* Interplay between Siderophores and Colibactin Genotoxin Biosynthetic Pathways in Escherichia coli. *PLoS Pathogens* vol. 9 e1003437 (2013).
65. Davidson-Pilon, C. *et al. CamDavidsonPilon/lifelines: v0.21.0*. (2019). doi:10.5281/zenodo.2638135.
66. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
67. 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).
68. 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).
69. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
70. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* vol. 8 833–835 (2011).
71. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Touchon, M. *et al.* Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).
74. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–12 (2015).
75. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* vol. 215 403–410 (1990).
76. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
77. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
78. 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).
79. Jones, E., Oliphant, T. & Peterson, P. SciPy: Open source scientific tools for Python. *http://www.scipy. org/* (2001).
80. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
81. 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* vol. 13 209 (2012).
82. McKinney, W. & Others. Data structures for statistical computing in Python. in *Proceedings of the 9th*

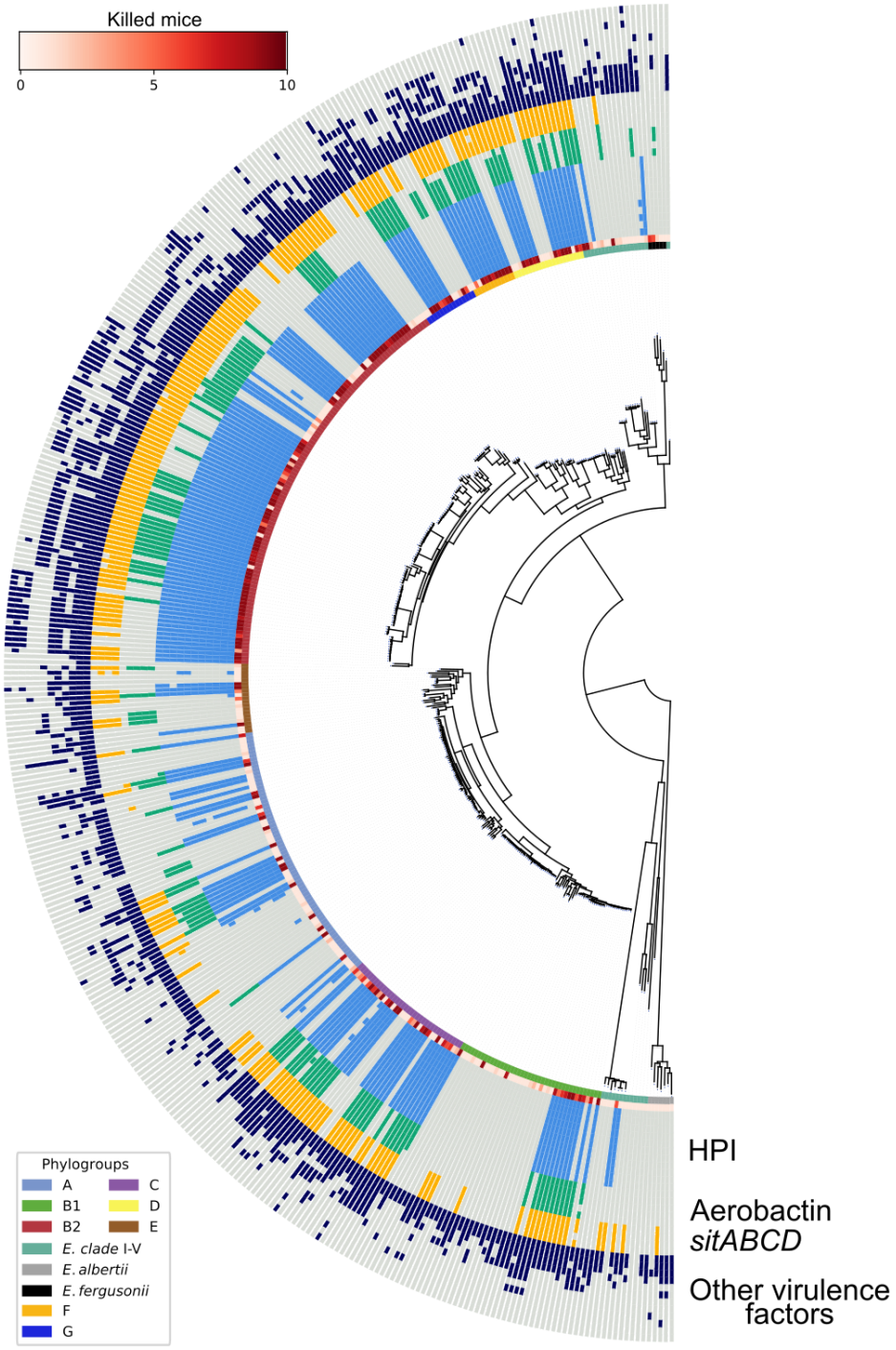
Python in Science Conference vol. 445 51–56 (2010).
83. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
84. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
85. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
86. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* vol. 57 61 (SciPy society Austin, 2010).
87. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**, 90–95 (2007).
88. Waskom, M. *et al. mwaskom/seaborn: v0.9.0 (July 2018)*. (2018). doi:10.5281/zenodo.1313201.
89. Kluyver, T. *et al.* Jupyter Notebooks-a publishing format for reproducible computational workflows. in *ELPUB* 87–90 (2016).
90. 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 indicates the sample size used in the actual analysis.

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



Supplementary Figure 3. Presence/absence patterns of known virulence factors. Solid color indicates presence, light grey indicates absence. Phenotypes (number of killed mice) and phylogroup or species of each strain are reported as in Figure 1A. "Other virulence factors" are (from inside the ring towards the outside): *sfaD*, *sfaE*, *ompT*, *traT*, *hra2*, *papC*, *iha*, *ireA*, *neuC*, *hlyC*, *clbQ* and *cnf1*.

631 **Supplementary Information**

632 **Supplementary Table 1:** Strains' information, including virulence phenotype

633 **Supplementary Table 2:** Summary of the 81 genes with at least one mapped unitig

634 **Supplementary Table 3:** Survival analysis for NILS9 and NILS46 wild-type and HPI
635 mutants

636 **Supplementary Table 4:** Correlation between growth on stress conditions (s-scores)
637 and both virulence and presence of the HPI

638 **Supplementary Table 5:** Feature importance for each growth condition in the random
639 forests predictor for virulence and HPI presence

640 **Supplementary Table 6:** List of PCR primers used in this study