Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*

Short title: Machine learning identifies signatures of bacterial host adaptation

Nicole E. Wheeler[1,2,*], Paul P. Gardner[2,3], Lars Barquist[4,5,*]

1. Wellcome Sanger Institute, Hinxton, United Kingdom.
2. Biomolecular Interaction Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
3. Department of Biochemistry, University of Otago, Dunedin, New Zealand.
4. Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany.
5. Helmholtz Institute for RNA-based Infection Research, Wuerzburg, Germany

* Correspondence to: nw17@sanger.ac.uk; lars.barquist@helmholtz-hiri.de

Keywords: hidden Markov model, random forest, loss-of-function, niche adaptation.

Abstract

Emerging pathogens are a major threat to public health, however understanding how pathogens adapt to new niches remains a challenge. New methods are urgently required to provide functional insights into pathogens from the massive genomic data sets now being generated from routine pathogen surveillance for epidemiological purposes. Here, we measure the burden of atypical mutations in protein coding genes across independently evolved *Salmonella enterica* lineages, and use these as input to train a random forest classifier to identify strains associated with extraintestinal disease. Members of the species fall along a continuum, from pathovars which cause gastrointestinal infection and low mortality, associated with a broad host-range, to those that cause invasive infection and high mortality, associated with a narrowed host range. Our random forest classifier learned to perfectly discriminate long-established gastrointestinal and invasive serovars of *Salmonella*. Additionally, it was able to discriminate recently emerged *Salmonella* Enteritidis and Typhimurium lineages associated with invasive disease in immunocompromised populations in sub-Saharan Africa, and within-host adaptation to invasive infection. We dissect the architecture of the model to identify the genes that were most informative of phenotype, revealing a common theme of degradation of metabolic pathways in extraintestinal lineages. This approach accurately identifies patterns of gene degradation and diversifying selection specific to invasive serovars that have been captured by more labour-intensive investigations, but can be readily scaled to larger analyses.

Introduction

Understanding how bacteria adapt to new niches and hosts and thus emerge or re-emerge as a cause of infectious disease in human and animals is of critical importance to anticipating and preventing epidemic disease [1,2]. With the decreasing cost of genome sequencing, comparative genomics has become a rich source of insight into the origins and

43 movement of bacteria in new pathogenic niches. However, translating whole genome
44 sequence databases into mechanistic and functional insights remains a challenge.
45
46 Early expectations were that pathogen evolution would be driven primarily by the acquisition
47 of virulence factors. However, as whole-genome sequencing has become increasingly
48 routine, a decidedly more complex picture has emerged [3,4]. A pattern of bacterial entrance
49 to a new niche followed by adaptation through the loss of antivirulence loci and reduced
50 metabolic flexibility is now recognised as a paradigm of the emergence of important human
51 pathogens from non-pathogenic bacterial species [5–8]. These new niches can be the result
52 of virulence factor acquisition providing access to a previously inaccessible niche in a so-
53 called foothold moment [8], or the emergence of new host niches driven by chronic disease
54 [9–11]. While pathogen and host requirements for infection vary, there is increasing
55 evidence of parallel evolution in bacteria adapting to the same or similar host niche. This is
56 perhaps nowhere more evident than in the species *Salmonella enterica*.
57
58 *Salmonella enterica* strains that cause disease in warm-blooded mammals lie on a spectrum
59 from those that have a broad host range and cause self-limiting gastrointestinal infection, to
60 those that are more restricted in host range, but cause systemic disease and are typically
61 associated with higher mortality [11,12]. Host-restricted, extraintestinal variants of
62 *Salmonella enterica* have evolved independently multiple times from gastrointestinal
63 ancestors [13], and show a greater degree of gene degradation compared to their generalist
64 relatives [14–16]. There are common patterns in the genes that undergo pseudogenization in
65 invasive *Salmonella*, most obviously an extensive network of genes required for anaerobic
66 metabolism in the inflamed host [17,18], a pattern with parallels in other host-adapting
67 enteropathogens [5].
68
69 Identifying these signals of parallel evolution has been challenging, relying mainly on manual
70 annotation and comparison of pseudogenes [17,18]. Detection of pseudogenes in particular

relies on ad-hoc criteria to identify large truncations, deletions, or frameshifts [19,20]. It is rare that the same genes or complete pathways are pseudogenized in host-adapted species; rather interpretation has relied on identifying overrepresentation of independent pseudogenization events clustered in certain pathways [17]. If pseudogenization leads to pathway attenuation or inactivation, it seems likely that reduced selective pressure will lead to a higher incidence of detrimental mutation fixation in other genes in these pathways. Indeed, we have previously shown that functional variant calling, based on sequence deviation from patterns of conservation observed in deep sequence alignments, shows a similar functional signal in host-restricted *Salmonella enterica* serovar Gallinarum to pseudogene analysis [21], identifying a larger cohort of genes where constraints on drift appear to have been lifted during host-adaptation.

In previous work we developed DeltaBS, a profile hidden Markov model (HMM) based approach to functional variant calling [21]. The basic assumption of this approach is that variation in conserved positions of a protein sequence is more likely to affect protein function than variation in less conserved regions. This approach can integrate information about nonsynonymous mutations, indels, and truncations. We have previously shown that DeltaBS can successfully identify functional changes in genes that would be missed by standard pseudogene analysis [22], and that a subset of genes in host-adapted strains appear to accumulate large DeltaBS values [21]. Additionally, others have observed similar changes in DeltaBS distributions during adaptation of *Salmonella* to a single immunocompromised host [10]. We generally assume that a large DeltaBS value is indicative of a decay in protein function, however a modest increase in DeltaBS associated with a phenotype may instead be indicative of diversifying selection.

Here, we have leveraged these previous observations to identify signatures of mutational burden consistent with adaptation to an invasive lifestyle. We have developed a random forest classifier using delta bitscore (DeltaBS) functional variant calling [21] that can perfectly

separate intestinal *Salmonella* serovars from host-adapted, extraintestinal serovars. We use random forest models because they perform well on datasets with few informative variables [23,24], and the decision tree structure they employ has the potential to detect functional relationships (i.e. epistasis) between genes [25,26]. They have been applied successfully in the past to predict microbial phenotype using gene presence/absence data [27], and SNPs already known to be associated with phenotype [28,29]. We show that these models produce interpretable signatures of host-adaptation, and furthermore that these signatures can be detected in strains of *Salmonella* associated with invasive disease in immunocompromised populations in sub-Saharan Africa.

Results

Constructing a random forest classifier for extraintestinal Salmonellae

The approach taken in this investigation is summarised in Fig 1, and described below. We built our model using a collection of genomes from well-characterised reference strains of gastrointestinal and extraintestinal *Salmonella* serovars (S1 Table), drawing on the extensive curation of orthology relationships performed by Nuccio and Bäumler [17]. These strains were originally characterised as “gastrointestinal” or “extraintestinal” based on common patterns of gene degradation, host restriction and clinical characteristics observed among the extraintestinal strains [17], and we have employed this same categorisation our analysis. We scored the functional importance of sequence variation by comparing the protein coding genes of each serovar to profile HMMs from the eggNOG database [30], designed to capture patterns of sequence variation typically seen in the protein coding genes of Gammaproteobacteria (see Methods).

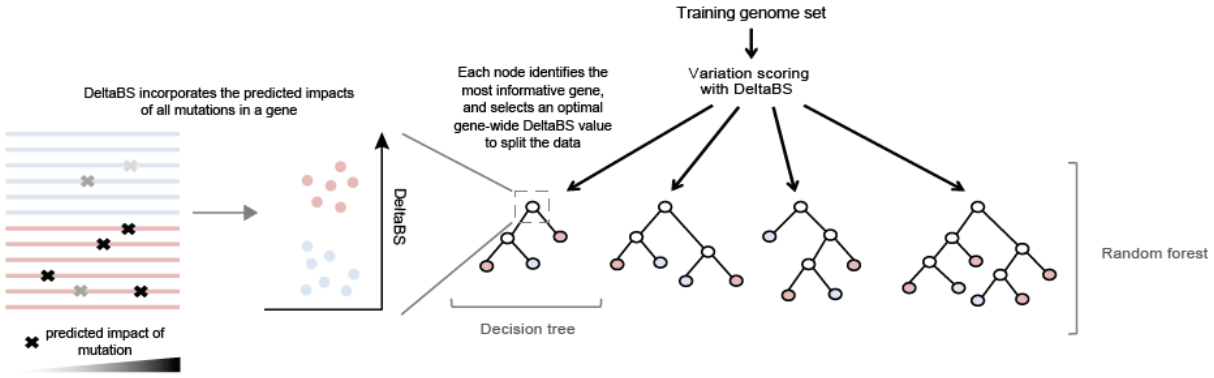


Fig 1 | Overview of the approach employed in this study

For each genome, the functional significance of sequence variation within protein coding genes is quantified using the DeltaBS metric. Following scoring, a bootstrap sampling of genomes are used to train each decision tree. For each node in the tree, a random subset of genes are sampled, and the most informative gene from this set is chosen to split the data. For each node in the tree, the predictive utility of the selected gene (variable importance) is tested by calculating how well the gene separates the samples according to phenotype.

We then employed random forests to identify the genes which were most informative of phenotype when viewed collectively. Random forests work by building an ensemble of decision trees designed to predict a characteristic of the samples [31], in this case adaptation to an extraintestinal, or invasive, niche. For each node in the decision tree, the best gene of a random sampling from the training gene set is selected according to its ability to separate a randomly selected subset of samples by phenotype based on DeltaBS values. The process of building a random forest produces measures of variable importance that can be used to assess the relative utility of different genes in classification of *Salmonella* strains based on lifestyle.

A small subset of genes are strongly predictive of invasiveness in Salmonella

To obtain an indication of the proportion of the genome that shows patterns of unusual sequence variation associated with an invasive phenotype, we trained a random forest model on a set of 6,438 orthologous genes. Accuracy of the model was assessed using out-

of-bag accuracy. This out-of-bag (OOB) measure of accuracy gives us an indication of how well each decision tree in the forest performs at predicting phenotype in a serovar it has never encountered before, using information on DeltaBS differences collected from other serovars. Next, we performed iterative feature selection to improve the performance of the model. This process involved repeated rounds of selecting the top 50% of predictors and re-training the model, until the model achieved perfect OOB predictive performance on the training dataset (Fig 2A). When the full set of filtered orthologous genes was used to build a model, a subset of genes ranked much higher than the others in variable importance (VI) (Fig 2B). We then saw a tailing off of VI, resulting in 4,721 orthologous groups either not being used in the model, or not improving classification accuracy (as indicated by VI = 0). This set of genes was discarded in the the first round of feature selection, and a subsequent 1,521 genes were discarded in the subsequent three rounds. The final model used 196 of the original 6,438 genes for prediction (S2 Table). This model additionally achieved perfect classification accuracy on an independent set of genomes of the same serovars as our training data (Fig S1). We tested for overfitting using permutation tests, and for correlation bias [32] using a variety of alternative model building strategies, and found no evidence for either phenomenon in our model (File S1).

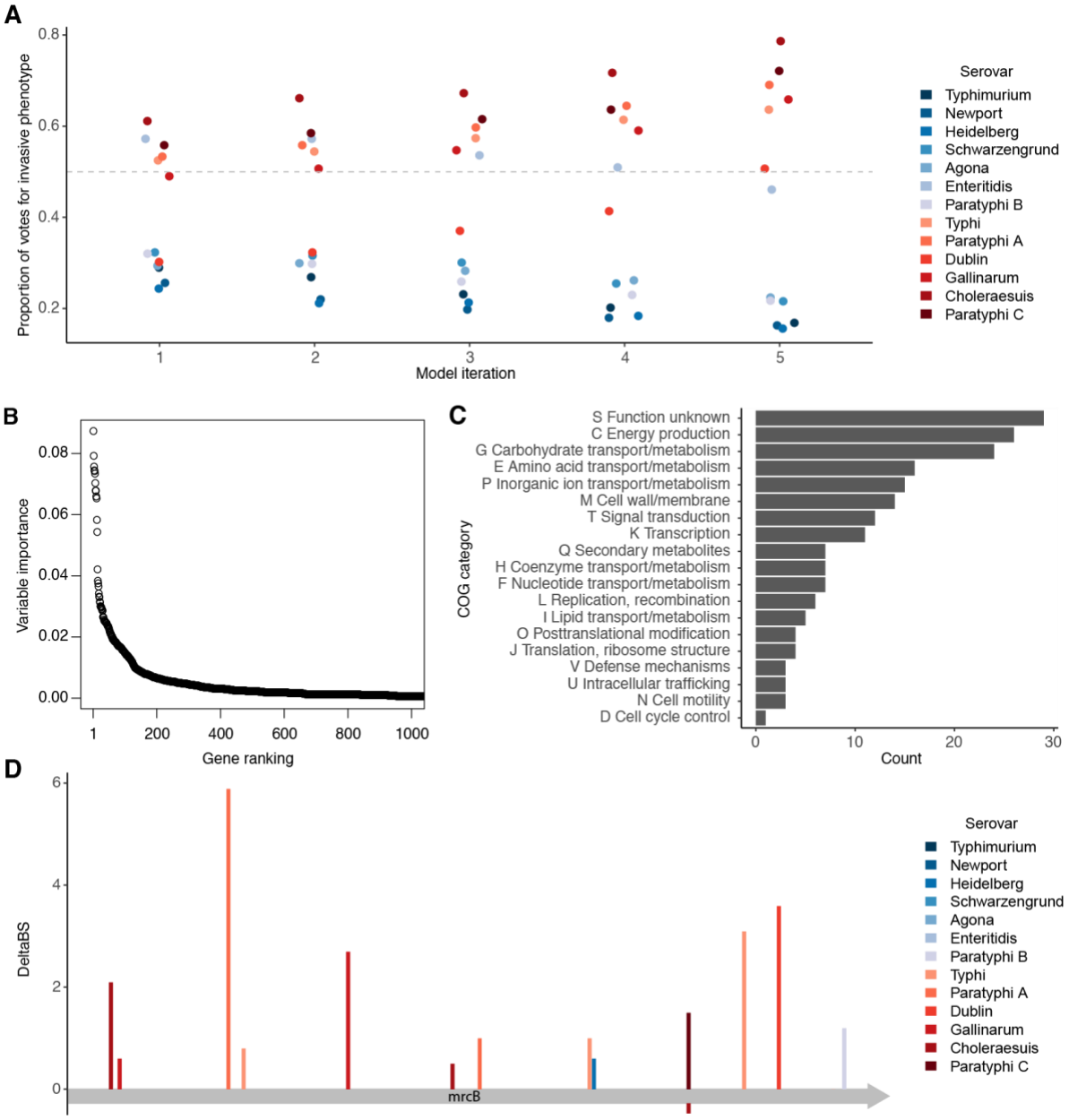


Fig 2 | A subset of *Salmonella* genes are strongly indicative of invasive potential

A: Out-of-bag votes for phenotype of each serovar cast by each model. Model 1 is the model built using all predictor variables, then each successive model was built using sparsity pruning from the previous model's predictor variables. Model 5 is the final model with 100% accuracy. Out-of-bag votes include only those votes cast by trees that were not trained on a given sample. The dashed grey line indicates the voting threshold to classify an isolate as invasive. Invasive serovars are coloured in red and gastrointestinal serovars are coloured in blue.

B: Of all genes used in the original training dataset, a small minority are given high importance in identifying invasive strains. Variable importance is shown for the top 1000 genes used in the original training set. Variable importance was measured as average decrease in Gini index in a random forest model trained on all orthologous groups that met the inclusion criteria (N = 6,438).

C: Functional categories associated with the top predictive genes.

D: Mutations in *mrcB* (penicillin-binding protein 1b), one of the top three predictors. Mutations in different strains are colour-coded, with bars in red indicating a mutation in an extraintestinal strain and bars in blue indicating a mutation in a gastrointestinal strain. An estimate of the effect of the mutation on protein function (DeltaBS) is shown on the y-axis, with positive values indicating higher chance of a mutation impacting protein function. The x-axis represents the length of the protein.

Predictive genes are typically degraded or absent in invasive isolates

We anticipated that the majority of informative genes identified in our study would be genes that showed functional degradation in invasive isolates but not in gastrointestinal isolates. Of the top predictors in our study (N = 196), 154 showed significantly greater mutational burden in extraintestinal strains compared to gastrointestinal strains (Mann-Whitney U test, adjusted *P*-value < 0.05), compared to 9 genes that showed significantly greater mutational burden in gastrointestinal strains. Of the genes that were more conserved in invasive isolates, one was the aldo-keto reductase *yakC*, which was deleted or truncated in all but one gastrointestinal strain and intact in all invasive strains. Another was the chaperone protein *yajL*, which appears to be important for oxidative stress tolerance [33,34].

Among the top predictors were several sets of genes belonging to the same operon (S2 Table). Examples included the *ttr*, *cbi* and *pdu* operons, which are all required for the anaerobic metabolism of 1,2-propanediol [35]. These operons have previously been identified as key degraded pathways in invasive isolates [16–18], and indicate the

agreement of this method with other studies linking loss of gene function to host niche. Overall, a large proportion of the identified genes were involved in metabolism (Fig 2C), consistent with the findings of similar studies [17,18]. Of the 167 central metabolism genes identified by Nuccio and Bäumler (2014) as truncated or deleted in at least one extraintestinal serovar, only one of these was previously reported to be truncated in > 4 serovars. In contrast, we found that 20 of the 167 central metabolism genes were identified by our model as informative of phenotype, indicating that including signal from more subtle forms of loss of function improves our ability to detect parallelism across lineages of invasive *Salmonella*. Of the 13 genes reported to be frequently disrupted by Nuccio and Bäumler, our approach identified 9. The other 4 were either not a match to profile HMMs in our database, or the truncation did not fall within the span of the model. Other major categories affected include proteins involved in cell wall and membrane function, perhaps suggesting changes affecting recognition by the host immune system, and signal transduction, suggesting some degree of consistent regulatory rewiring during adaptation to an extraintestinal niche.

Information provided by multiple genes was often more informative of phenotype than a single gene individually, as was the case for *fimD* and *fimH* (Fig S2). FimD and FimH constitute central components of type 1 pili, and both are required for expression of normal fimbriae [36]. This demonstrates that our approach is capable of identifying epistatic relationships between genes, where a modification in function of one gene masks the functional status of the other.

Sequence changes in key indicator genes involve independent mutations in each serovar, contributing to similar functional outcomes

When examining individual genes that showed differences in mutational burden between invasive and gastrointestinal isolates, we found that most of these mutations had occurred independently, and had occurred at different sites in the protein. Using a permissive threshold (DBS>3), or a conservative threshold (DBS>5), there were close to twice as many

deleterious, independent mutations in the genes of the invasive serovars than those of the gastrointestinal (476:910; 537:991, respectively, see Methods). This phenomenon was even more pronounced when only mutations with DBS over the upper quartile were counted (249:612, Table S3). While the majority of genes identified appeared to be cases of gene degradation in invasive lineages, some genes showed more subtle signs of mutational burden, restricted to nonsynonymous changes of modest predicted functional impact.

An example of this, Fig 2D, illustrates mutation accumulation in one of the top candidate genes, *mrcB*, encoding penicillin-binding protein 1b (PBP1b). Not only does *mrcB* carry more mutations in invasive serovars compared to gastrointestinal serovars, the mutations have occurred independently in different positions within the protein. Penicillin-binding proteins are the major target of β-lactam antibiotics and are important for synthesis and maturation of peptidoglycan [37]. PBP1b in particular extends and crosslinks peptidoglycan chains during cell division. While PBP1b is not essential, it has been shown to be synthetically lethal when the partially redundant *mrcA*/PBP1a is deleted, and is important in *E. coli* for competitive survival of extended stationary phase, osmotic stress [38], and — in *Salmonella* Typhi — growth in the presence of bile [39]. Bile is an important environmental challenge for *Salmonella*, particularly for extraintestinal serovars which colonize the gall bladder [40]. While there are more mutations in invasive than in gastrointestinal serovars, the mutations that occur in this protein are all amino acid substitutions of modest predicted impact. This suggests that sequence changes could result in a modification of protein function, rather than a loss, consistent with the importance of PBP1b for the survival of *S*. Typhi during a typical infection cycle [39].

S. Dublin and S. Enteritidis serovars are more difficult to classify than others

To anticipate the performance of our random forest model on new data we computed out-of-bag (OOB) error. Because random forests train each decision tree on a random subset of the training data, OOB error can be computed by testing the performance of these trees on

data they have not been trained on, providing inbuilt cross-validation [31]. In our case, perfect OOB classifications were only achieved by the fifth iteration of the model. The need for iterative improvement of the model came from difficulty in correctly classifying the reference strains for serovars Enteritidis and Dublin. This is reflective of their relatively recent divergence and niche adaptation compared to other serovars in the study (Fig S3, Langridge et al. 2015). *S.* Gallinarum was classified much more readily than *S*. Enteritidis and *S.* Dublin, despite being closely related to both serovars, perhaps due to its host restriction.

S. Enteritidis was initially mis-classified as invasive, indicating that it shares genomic trends with invasive lineages. Genomic analyses have indicated that the ancestor of *S*. Enteritidis previously possessed intact pathogenicity islands (SPI-6 and SPI-19), each encoding a type six secretion system [18,41]. These loci have been implicated in host-adaptation and survival during extraintestinal infection [42,43], and it has been speculated based on their loss and other evidence that classical *S*. Enteritidis has been adapting towards greater host generalism with respect to its ancestral state [18]. This could explain the greater number of disrupted and deleted genes relative to other gastrointestinal serovars used in this study, and the difficulty in classifying it correctly. Conversely, *S*. Dublin was initially mis-classified as gastrointestinal. In previous studies *S.* Dublin has been shown to possess fewer pseudogenes than related invasive isolates [17,18], suggesting a lower degree of host adaptation than other invasive isolates. Indeed, *S*. Dublin is more promiscuous in its host range, primarily infecting cattle [44] while still causing sporadic human disease [45]. It seems likely that a subset of informative genes identified in early iterations of the model may have been indicators of host restriction or generalism rather than broad extraintestinal adaptation.

275 ***Patterns of gene degradation identified in established invasive lineages are present in***
276 ***novel lineages of S. Typhimurium and S. Enteritidis associated with systemic***
277 ***infection***

278 In recent years there have been reports of novel *S.* Typhimurium and *S.* Enteritidis lineages
279 associated with invasive disease in sub-Saharan Africa [46–48] in populations with a high
280 prevalence of immunosuppressive illness such as HIV, malaria, and malnutrition [49]. These
281 lineages contribute to a staggering burden of invasive non-typhoidal salmonella (iNTS)
282 disease, which is responsible for an estimated 3.4 million cases and circa 680,000 deaths
283 annually [50]. Based on epidemiological analysis, high-throughput metabolic screening of
284 selected strains, and analysis of pseudogenes it has been suggested that these lineages
285 may be rapidly adapting to cause invasive disease in the human niche created by
286 widespread immunosuppressive illness [11,46–48,51].
287
288 Two iNTS-associated lineages have recently been described within serovar Enteritidis [48],
289 geographically restricted to West Africa and Central/East Africa, respectively. Initial
290 observations have demonstrated that a representative isolate of the Central/East African
291 clade has a reduced capacity to respire in the presence of metabolites requiring cobalamin
292 for their metabolism and has lost the ability to colonize a chick infection model [48],
293 suggesting adaptation to a new host niche. Similarly, two iNTS disease associated lineages
294 have been described in serovar Typhimurium [47], both members of sequence type 313
295 (ST313), generally referred to as Lineage I and II in the literature. Lineage II appears to have
296 largely replaced Lineage I since 2004, and it has been suggested this is due to Lineage II
297 possessing a gene encoding chloramphenicol resistance [47]. Laboratory characterization of
298 Lineage II strains has shown that they are not host-restricted [52,53], but do appear to
299 possess characteristics suggestive of adaptation to an invasive lifestyle [54–57], though it is
300 important to note that this is a complex trait and not easily quantified.
301

Given the evidence of adaptation to an invasive niche in these lineages, we asked if genomics signatures of extraintestinal adaptation we had detected previously could be detected in iNTS disease associated lineages. To this end, we applied our predictive model trained on well-characterized extraintestinal strains to calculate an invasiveness index, the fraction of decision trees in the random forest voting for an invasive phenotype. First, we compared isolates from African iNTS-associated clades of *S*. Enteritidis (N=233) to a global collection of isolates generally associated with intestinal infection (N=100) [48].

Our model gave iNTS-associated *S*. Enteritidis strains a higher invasiveness index than the globally distributed isolates (Fig 3A,B, Table S4), indicating the presence of genetic changes paralleling those that have occurred in extraintestinal serovars of *Salmonella*. Similar gene signatures were only rarely observed in the global epidemic clade (Fig 3C). These findings are consistent with the metabolic changes observed by Feasey et al. [48] in the Central/Eastern African clade compared to the global epidemic clade. In particular we found signs of gene sequence variation uncharacteristic of gastrointestinal *Salmonella* across a number of key genomic indicators, including *tcuR, ttrA, pocR, pduW, eutH,* SEN2509 (a putative anaerobic dimethylsulfoxide reductase) and SEN3188 (a putative tartrate dehydratase subunit), all in pathways previously identified by Nuccio and Bäumler [17] as being involved in the utilization of host-derived nutrients in the inflamed gut environment. This indicates that our model is able to identify early signatures of adaptation, even in these recently emerged strains that still retain some capacity to cause enterocolitis [48].

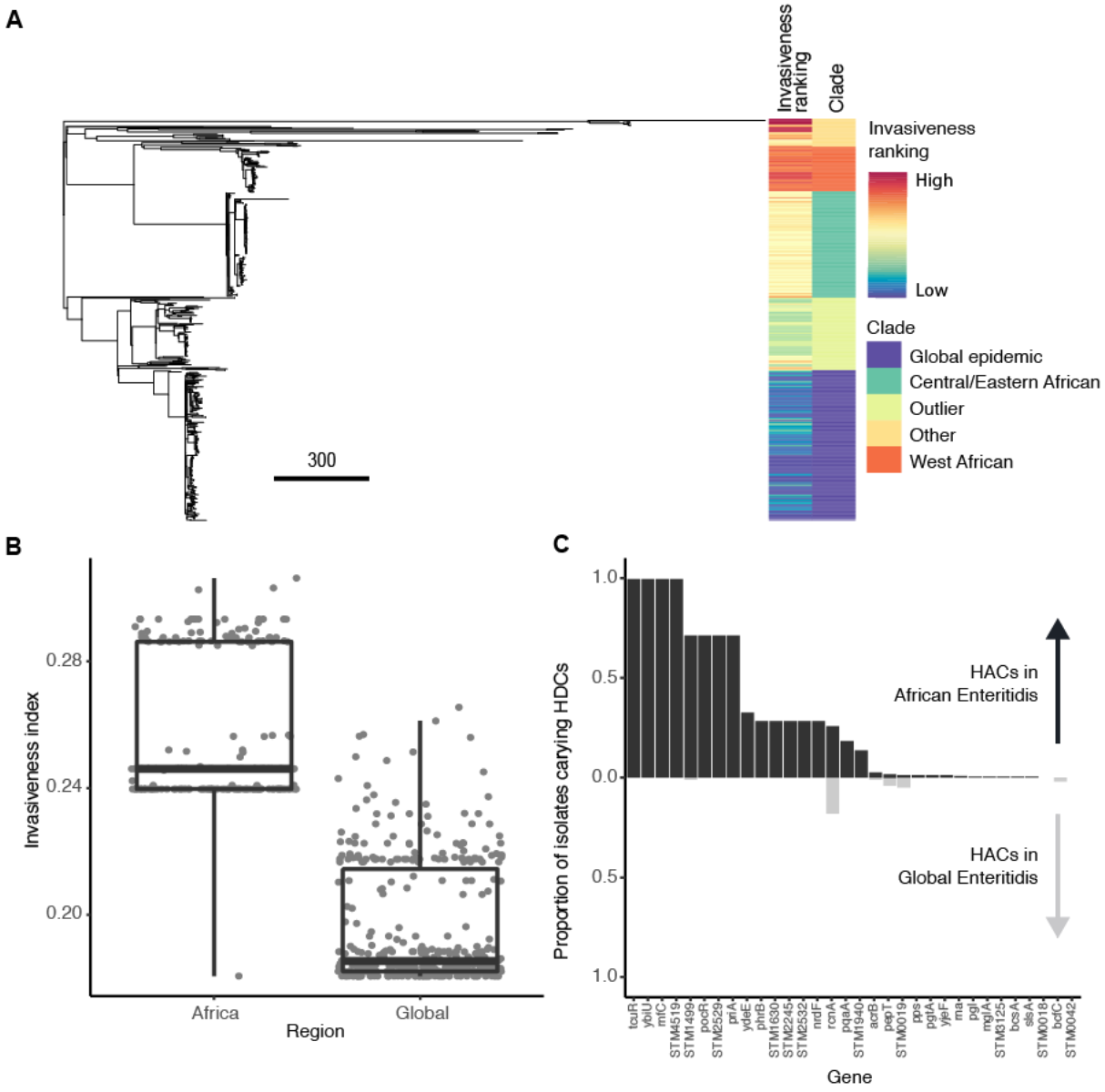


Fig 3 | Voting of the model on African iNTS and global gastrointestinal isolates

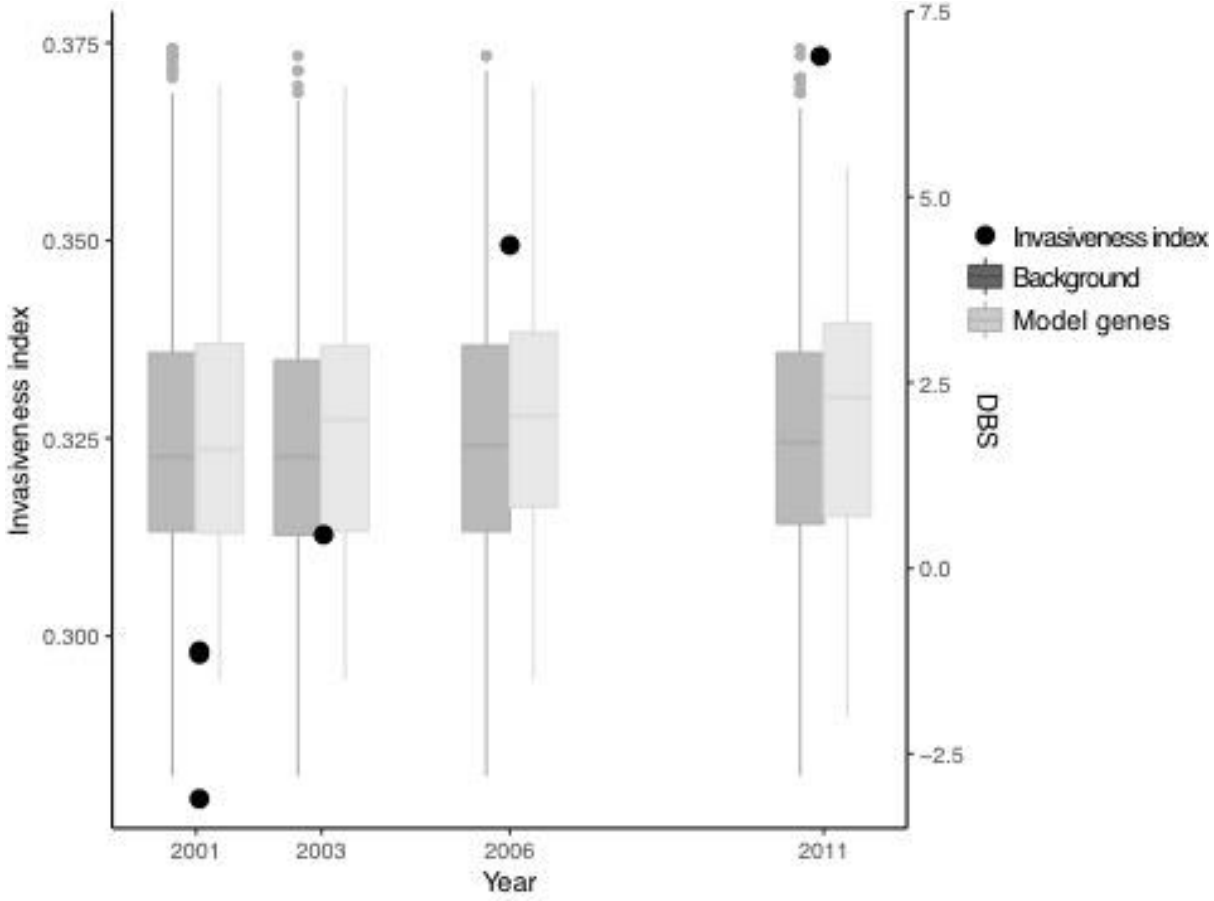
A: Maximum likelihood phylogeny of all *S*. Enteritidis isolates included in the study, annotated with invasiveness ranking and clade (note: Outlier refers to the distinct sister clade of the global epidemic strains identified by [48], while Other refers to strains that don't belong to a named clade).

B: Invasiveness indices for African and non-African clades of *Salmonella*. Lower and upper boundaries of the boxplots correspond to the 25th and 75th quantiles.

C: The proportion of isolates from each tested dataset carrying a hypothetically attenuated coding sequence (HAC, defined by a DeltaBS>3 relative to the reference serovar). Genes

334 are ordered by the amount of degradation observed in African clades. African strains are
335 shown in the positive y-axis in darker grey, global strains are shown in the negative y-axis in
336 lighter grey.
337
338 To confirm this, we performed an additional comparison of *S*. Typhimurium ST313 isolates
339 (N=208), to global isolates from other STs, predominantly ST19, associated with
340 gastroenteritis (N=51) [51,58]. Similarly to iNTS associated *S*. Enteritidis isolates, *S*.
341 Typhimurium ST313 isolates has a higher invasiveness index than isolates from other STs
342 (Fig S4, Table S5). Within ST313, Lineage II scored higher than Lineage I, possibly
343 suggesting differential adaptation to the extraintestinal niche. We found that there were in
344 fact more degraded genes unique to Lineage I than Lineage II, but that these genes were
345 assigned less weight in the model, so did not impact score as strongly (Figure S2 & S3).
346 Interestingly, ST313 has recently been shown not to be entirely restricted to Africa, with
347 isolation reported in Brazil [59] and the UK [58], associated primarily with gastrointestinal
348 disease. We included a collection of UK ST313 strains [58] in our analysis, and found that
349 their invasiveness index tended to be elevated compared to non-ST313 salmonellae, and
350 intermediate between Lineage I and II, suggesting that this adaptation is not restricted to
351 circulating African strains, as it can be seen in strains collected from other countries as well
352 (Fig S5). This observation is consistent with the work of Ashton et al., who noted shared
353 pseudogenes and phenotypic traits in UK and African ST313 isolates. This suggests our
354 model is capturing features here associated with the ability to colonize an extraintestinal
355 niche, rather than enter it in healthy individuals.
356
357 In addition to the iNTS lineages we investigated, some other strains had unusually high
358 invasiveness indices. Among the top scoring isolates outside of the African *S*. Enteritidis
359 lineages are Ratin strains, a rodenticidal lineage used as commercial rat poison before the
360 1960s [60]. In *S*. Typhimurium, a clade containing strains DT99, DT56 and U313 also scored

361 highly. These strains appear to be adapted to birds, and DT99 and DT56 have been
362 reported to be highly virulent in pigeons [12,61–63].
363
364 While the above data suggests that our model is detecting genetic changes associated with
365 extraintestinal survival, it is difficult to infer directionality from large isolate collections. We
366 have addressed this using a unique case of accelerated adaptation over the course of a
367 single infection (Fig 4). We scored the invasiveness index of a collection of hypermutator *S*.
368 Enteritidis isolates collected over a ten year period that were adapting to chronic systemic
369 infection of an immunocompromised patient [10]. We found a significant positive correlation
370 between invasiveness index and duration of carriage (r=0.96, n=6, *P*=0.002). Additionally,
371 there was a significant shift over time in the DeltaBS distribution for the genes in our model
372 as compared to the rest of the genome (*P*=7.576e-05, Mann Whitney U test). This suggests
373 a specific change in selective pressure on genes inferred to be important for extraintestinal
374 survival from established invasive serovars, and provides evidence for parallel adaptation.



375

Figure 4 | Invasiveness indices and DeltaBS (DBS) values for isolates collected during long term invasive infection of an immunocompromised patient (Klemm et al. 2016). Black points show the increase in the invasiveness index over time. Boxplots show a significant shift in DBS distribution over the duration of carriage for genes selected by our model built from well-characterised invasive serovars as compared to the rest of the proteome. Isolates from [10]**.** DBS distributions for 2001 have been pooled, but are representative for all three isolates individually. The y-axis for DBS values has been truncated for better visualisation.

Discussion

Parallel evolution appears to be common in niche adaptation, which allows us to identify genes that are important for survival in different environments [64]. Parallelism has been observed across vastly different time scales in adapting pathogens. Parallel evolution in the distantly related genuses *Salmonella* and *Yersinia* during adaptation to invasive infection of the human host has led to independent losses of the *ttr*, *cbi* and *pdu* genes, important for anaerobic metabolism during intestinal infection [5]. Within genuses, parallelism has been observed when distinct lineages acquire similar virulence factors leading to similar phenotypes, as with *Yersinia pseudotuberculosis* and *enterocolitica* [8], or the repeated emergence of the *Shigella* phenotype within the *Escherichia* [6]. Even on the scale of a single human lifetime, parallel adaptation has been observed in *Pseudomonas aeruginosa* lineages adapting to infection of the lungs of children with cystic fibrosis [9], or a hypermutator strain of *Salmonella* adapting to an immunocompromised host [10]. With pathogen sequencing for disease surveillance becoming increasingly routine [65–67], we have the opportunity to search for signals of parallel evolution as new pathogens emerge, or old pathogens expand into new niches.

Here, we have developed an approach for automatically learning which genes contribute to this parallel adaptation. Leveraging the DeltaBS functional variant scoring approach we developed previously [21] allowed us to construct scores which integrate independent

403 mutations and indels that impact gene function. Using these scores, we were able to
404 construct a classifier model which is able to separate *Salmonella* serovars adapted to an
405 extraintestinal niche from gastrointestinal strains. Importantly, the random forest classifier
406 that we used produces interpretable lists of genes involved in this adaptation, which agree
407 with results in the literature attained through manual curation of pseudogenes. Additionally,
408 we have shown that this classifier is able to identify nascent signatures of adaptation in
409 strains of *Salmonella* which have been evolving in response to large populations of
410 immunocompromised patients in resource-poor nations.
411
412 Other automated approaches to detecting adaptation have been developed which search for
413 SNPs [68] or words [69,70] associated with phenotype. These approaches, termed microbial
414 genome-wide association studies (GWASs), have used techniques adapted from human
415 GWASs, but better cater to methodological issues that arise due to the differences between
416 human and bacterial inheritance patterns. Major differences impacting analyses are stronger
417 linkage disequilibrium (LD) between genetic variants in bacterial genomes, greater
418 population stratification, and often stronger selection for traits [71]. Greater LD and
419 population stratification often result in traits being linked closely with particular lineages, and
420 a large number of variants unique to a lineage being spuriously associated with phenotype.
421 Correction for population stratification allows greater discrimination of true and false positive
422 associations, but results in a substantial loss of power to detect true positives [71],
423 particularly in phenotypes that are highly polygenic and are not under strong positive
424 selection [72]. This can be corrected by increasing the sample size of the study, but
425 increasing sample size can make measurement of complex phenotypes infeasible [23].
426
427 A number of machine learning approaches to predicting phenotype from genotypic
428 information have also been recently developed. A notable example is a Support Vector
429 Machine (SVM) based approach to predicting host range in *Salmonella enterica* and
430 *Escherichia coli* [73], as it has a similar aim of predicting strains with a higher probability of

causing severe disease. We have taken a markedly different approach to other machine learning based studies, primarily in our use of few, distantly related training examples, rather than densely sampled strains across a narrower phylogenetic distance. This is because we wanted to prevent over-fitting of the model through the inclusion of predictors that were informative of phylogeny rather than phenotype, and we wanted an accurate estimation of predictive error, which cannot be achieved using traditional cross-validation when there is a strong correlation structure in a dataset [74]. We have also taken additional steps to examine the genes and criteria used by the model to make predictions, and have presented these in Supplemental Table S2, in order to aid the reader's understanding of how the model makes predictions, and what this teaches us about the biology of this phenotype.

The use of DeltaBS output as training variables differs from current approaches by allowing the estimation of the combined effects of variants, both common and rare, on gene function. The weighting scheme can also combine data on gene presence/absence, indels and SNPs into a single metric. It significantly reduces the number of association tests that need to be performed to comprehensively capture much of the genetic diversity in a species, increasing power to detect associations, and reducing the requirement for such large sample sizes. The approach also aids in identifying genetic variants that are most likely to have a phenotypic effect within LD blocks. The DeltaBS variant scoring approach can be readily applied to large datasets, and could be employed in a linear mixed model (LMM) based association testing framework [68], or used in a hybrid LMM-random forest based approach [75] to preserve the ability of the metric to detect epistasis between genes [26].

Conclusions

In this study, we have demonstrated the insight to be gained by the layering of machine learning approaches to better understand niche adaptation in a bacterial pathogen. Firstly, profile hidden Markov models allow us to capture information on common patterns of

sequence variation in protein families in order to understand the functional significance of specific mutations. Using data on the accumulation of functionally impactful mutations across the proteome as input, random forests then allow us to identify genes that display a difference in selective pressures between lineages with different phenotypes. Not only has this approach proved effective at identifying biological mechanisms behind bacterial niche adaptation, it has also allowed us to detect the emergence of new extraintestinal lineages by searching for these recurrent patterns of mutation accumulation in a way that allows the recognition of novel mutations as cases of the same underlying shift away from the sequence constraints a gene is usually subjected to. We believe this general approach will be broadly applicable to any pathogen where multiple lineages are adapting to the same niche, and will be able to detect signatures of adaptation that are missed by other methods.

Methods

Genome data and identification of orthologs

High quality genomes for 13 well-characterised *Salmonella enterica* serovars were retrieved from the NCBI database (accessions and serovar information can be found in S1 Table). The serovars were divided into gastrointestinal and extraintestinal serovars according to the classifications made by Nuccio and Bäumler [17]. Ortholog calls were also taken from the Supplementary Material of Nuccio and Bäumler [17]. A core gene phylogeny for the strains used to build the model was produced using RAxML [76], based on a core gene alignment created in Roary [77].

Measuring the divergence of genes from predicted sequence constraints

Profile hidden Markov models (HMMs) for Gammaproteobacterial proteins were retrieved from the eggNOG database [30]. We chose this source of HMMs because it is publicly available, allowing for better reproduction of analyses, and we feel it provides a good balance between collecting enough sequence diversity to capture typical patterns of

sequence variation in a protein, without sacrificing sensitivity in the detection of deleterious mutations, as we have observed with Pfam HMMs [21]. Each protein sequence was searched against the HMM database using hmmsearch from the HMMER3.0 package (http://hmmer.org). The top scoring model corresponding to each protein was used for analysis (N = 8,060 groups). Orthologous groups (OGs) with no corresponding eggNOG HMM, or more than one top model hit were excluded from further analysis (N = 1,524). If most genes in an OG had a significant hit (E-value<0.0001) to the same eggNOG model, any genes within this OG that did not were assigned a score of zero, reflecting a loss of the function of that protein. These cases typically reflected a truncation that had occurred early in the protein sequence. Additionally, genes with no variation in bitscore for the match between protein sequences and their respective eggNOG HMM across isolates were excluded (N = 188). After this filtering process, 6,439 orthologous groups remained for analysis. Residue-specific DeltaBS (as in Fig 2D) was calculated by aligning orthologous sequences, choosing a reference sequence (from *S*. Typhimurium), and substituting each variant match state and any accompanying insertions into the reference sequence and calculating the difference in bitscore caused by the substitution.

Training a random forest classifier

The R package “randomForest” [78] was used to build random forest classifiers using a variety of parameters to assess which were best for accuracy. We used out-of-bag (OOB) error rate to measure the performance of the model [31]. Out-of-bag error is calculated automatically by the randomForest R package as the model is built. Briefly, calculations are performed as follows: as each decision tree is trained using a bootstrap sampling of the training genomes, a small number of samples are left aside to test the predictive accuracy of each decision tree on previously unseen samples. For each serovar, votes are collated and accuracy is calculated from only those decision trees that did not include the serovar in their training set. In this application, this step tests whether the genomic signatures of invasiveness captured by the decision trees based on some serovars are present in other

509 serovars, and thus whether the model can detect adaptation to an invasive lifestyle in
510 previously unseen lineages. OOB error rate, stabilised at 10,000 trees, so we chose this as a
511 parameter for optimising the number of genes sampled per node (mtry). mtry values of 1,
512 *p*/10, *p*/5, *p*/3, *p*/2 and *p* (where *p* = the number of predictors) were tested, and we found that
513 at mtry=*p*/10, the number of genes that were either not incorporated into trees, or did not
514 improve the homogeneity of daughter nodes when they were incorporated into trees (as
515 measured by mean decrease in Gini index, [79]) stabilised at ~92%. Training the random
516 forest classifier over five iterations took 55 seconds on a laptop computer. In order to assess
517 how well this method would scale, we trained another model on a larger dataset of *S*.
518 Enteritidis strains (N=677) using the same workflow and site of isolation as a proxy for
519 phenotype, which took 28 minutes.
520
521 To improve the performance of the model, we performed five model building and sparsity
522 pruning cycles. For the first cycle, we built a random forest model using all genes that met
523 the inclusion criteria, and performed sparsity pruning by eliminating all variables that had a
524 mean Gini index (variable importance) of zero or lower (meaning the gene was either not
525 included in the model or did not improve model accuracy when it was). Four successive
526 rounds of model building and sparsity pruning involved building a new model with the pruned
527 dataset, then pruning the genes with the lowest 50% of variable importances. The resulting
528 model had 100% out-of-bag classification accuracy. We also tested the accuracy of the full
529 model on a collection of alternative strains related to the training dataset (see Table S1).
530 Orthologs to the top genes identified by our model were identified using phmmer from the
531 HMMER3.0 package (http://hmmer.org). Additional notes on model building and testing are
532 provided in File S1.
533
534 We tested the top 196 genes for the presence of independent mutations in each serovar by
535 aligning each sequence to the profile HMM representing that protein family. Variation in each
536 sequence with respect to a designated reference sequence from the set (as selected by

Nuccio and Bäumler, 2014) at each site in the HMM was identified and classified as either a mutation unique to a single serovar, or one shared among mutliple serovars. Consecutive deletions or insertions with respect to the HMM consensus sequence were collapsed into single mutational events.

Invasive non-typhoidal Salmonella analysis

Read data from Feasey et al. [48] and Klemm et al [10] was mapped to the reference genome *S*. Enteritidis P125109. Reads from Okoro et al. [51] and Ashton et al. [58] were mapped to the reference genome *S*. Typhimurium LT2. For samples in the Okoro study, if an isolate was sequenced using multiple runs, the most recent run was chosen for analysis. All reads were mapped using BWA mem [80] and regions near indels were realigned using GATK [81]. Picard (http://broadinstitute.github.io/picard) was used to identify and flag optical duplicates generated during library preparation. SNPs and indels were called using samtools v1.2 mpileup [82], and were filtered to exclude those variants with coverage <10 or quality <30. For tree building, a pseudogenome was constructed by substituting high confidence (coverage >4, quality >50) variant sites in the reference genome, and masking any sites with low confidence with an "N". Insertions relative to the reference genome were ignored, and deletions were filled with an "N". Pseudogenome alignments were then used as input to produce trees using Gubbins [83] to exclude recombination events, and RAxML v8.2.8 [76] to build maximum likelihood trees using a GTR + Gamma model. Samples with >10% missing base calls were excluded from the analysis.

Sequences for the 196 genes of interest used in the random forest model were retrieved for each isolate and translated. These were then scored using their respective profile HMMs. Score data was collated, and any missing values were marked as 'NA' and imputed using the na.roughfix function from the randomForest R package [78]. This is a different approach used to that of the training dataset, due to the potentially lower quality of the sequenced genomes leading to gene absence due to low coverage rather than true deletion or severe

564 truncation. The relationship between invasiveness ranking and phylogeny were visualised
565 using Phandango [84].

566 ***Data availability***

567 All genome sequence data are publicly available, and accessions are provided in the
568 appropriate Supplemental Tables. Code and data for reproducing this analysis, performing
569 an equivalent analysis using new data, and assessing the invasiveness index of other
570 *Salmonella* strains is publicly available at
571 http://www.github.com/UCanCompBio/invasive_salmonella.

572 **Funding information**

573 NEW was supported by a PhD scholarship from the University of Canterbury, a Biomolecular
574 Interaction Centre Postdoctoral Fellowship, and the Wellcome Trust grant 206194. LB was
575 supported in part by a Research Fellowship from the Alexander von Humboldt
576 Stiftung/Foundation. NEW and PPG are supported by a Rutherford Discovery Fellowship
577 administered by the Royal Society of New Zealand, the Bioprotection Research Centre and
578 the National Science Challenge “NZ’s Biological Heritage”.

579 **Acknowledgements**

580 We are grateful to Sean Eddy for useful discussions and providing fast, accurate and free
581 software, and to Simon Harris for developing the pipeline used for mapping reads and calling
582 SNPs for the iNTS portion of our analysis. We also thank Julian Parkhill, Nick Feasey, Nick
583 Thomson, Alexander Westermann, Stan Gorski, and John Crump for their helpful feedback.

584 **References**

585 1. Frank SA, Schmid-Hempel P. Mechanisms of pathogenesis and the evolution of
586 parasite virulence. J Evol Biol. 2008;21: 396–404.

587 2. Fauci AS, Morens DM. The perpetual challenge of infectious diseases. N Engl J Med.

2012;366: 454–461.

3. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature. nature.com; 2007;449: 835–842.

4. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015;13: 787–794.

5. McNally A, Thomson NR, Reuter S, Wren BW. “Add, stir and reduce”: Yersinia spp. as model bacteria for pathogen evolution. Nat Rev Microbiol. 2016;14: 177–190.

6. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of Shigella evolution, adaptation and geographical spread. Nat Rev Microbiol. nature.com; 2016; doi:10.1038/nrmicro.2016.10

7. Merhej V, Georgiades K, Raoult D. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. Brief Funct Genomics. 2013;12: 291–304.

8. Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, et al. Parallel independent evolution of pathogenicity within the genus Yersinia. Proc Natl Acad Sci U S A. 2014;111: 6768–6773.

9. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet. 2015;47: 57–64.

10. Klemm EJ, Gkrania-Klotsas E, Hadfield J, Forbester JL, Harris SR, Hale C, et al. Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. Nat Microbiol. 2016;1: 15023.

11. Feasey NA, Dougan G, Kingsley RA, Heyderman RS, Gordon MA. Invasive non-typhoidal salmonella disease: an emerging and neglected tropical disease in Africa. Lancet. 2012;379: 2489–2499.

12. Rabsch W, Andrews HL, Kingsley RA, Prager R, Tschäpe H, Adams LG, et al. Salmonella enterica serotype Typhimurium and its host-adapted variants. Infect Immun. 2002;70: 2249–2255.

13. Bäumler A, Fang FC. Host specificity of bacterial pathogens. Cold Spring Harb Perspect Med. 2013;3: a010041.

14. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, et al. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature. 2001;413: 848–852.

15. McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, et al. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. Nat Genet. 2004;36: 1268–1274.

16. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, et al. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. 2008;18: 1624–1637.

628 17. Nuccio S-P, Bäumler AJ. Comparative Analysis of Salmonella Genomes Identifies a
629 Metabolic Network for Escalating Growth in the Inflamed Gut. MBio. 2014;5: e00929–
630 14–e00929–14.

631 18. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, et al.
632 Patterns of genome evolution that have accompanied host adaptation in Salmonella.
633 Proc Natl Acad Sci U S A. 2015;112: 863–868.

634 19. Lerat E, Ochman H. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids
635 Res. 2005;33: 3125–3132.

636 20. Kuo C-H, Ochman H. The extinction dynamics of bacterial pseudogenes. PLoS Genet.
637 2010;6. doi:10.1371/journal.pgen.1001050

638 21. Wheeler NE, Barquist L, Kingsley RA, Gardner PP. A profile-based method for
639 identifying functional divergence of orthologous genes in bacterial genomes.
640 Bioinformatics. 2016;32: 3566–3574.

641 22. Kingsley RA, Kay S, Connor T, Barquist L, Sait L, Holt KE, et al. Genome and
642 transcriptome adaptation accompanying emergence of the definitive type 2 host-
643 restricted Salmonella enterica serovar Typhimurium pathovar. MBio. 2013;4: e00565–
644 13.

645 23. Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SAFT. Explaining
646 microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics.
647 2013;12: 366–380.

648 24. Pappu V, Pardalos PM. High-Dimensional Data Classification. In: Aleskerov F,
649 Goldengorin B, Pardalos PM, editors. Clusters, Orders, and Trees: Methods and
650 Applications. Springer New York; 2014. pp. 119–150.

651 25. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data
652 mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?
653 Brief Bioinform. 2013;14: 315–326.

654 26. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev
655 Genet. 2014;15: 722–733.

656 27. Bayjanov JR, Molenaar D, Tzeneva V, Siezen RJ, van Hijum SAFT. PhenoLink--a web-
657 tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching
658 for Lactobacillus plantarum strains. BMC Genomics. 2012;13: 170.

659 28. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, et al. Predicting the
660 virulence of MRSA from its genome sequence. Genome Res. 2014;24: 839–849.

661 29. Alam MT, Petit RA 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al.
662 Dissecting vancomycin-intermediate resistance in staphylococcus aureus using
663 genome-wide association. Genome Biol Evol. 2014;6: 1174–1185.

664 30. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al.
665 eggNOG 4.5: a hierarchical orthology framework with improved functional annotations
666 for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44: D286–93.

667 31. Breiman L. Random Forests. Mach Learn. Kluwer Academic Publishers; 2001;45: 5–32.

668 32. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature
669 ranking and solutions. Bioinformatics. 2011;27: 1986–1994.

670 33. Kthiri F, Gautier V, Le H-T, Prère M-F, Fayet O, Malki A, et al. Translational defects in a
671 mutant deficient in YajL, the bacterial homolog of the parkinsonism-associated protein
672 DJ-1. J Bacteriol. 2010;192: 6302–6306.

673 34. Le H-T, Gautier V, Kthiri F, Malki A, Messaoudi N, Mihoub M, et al. YajL, prokaryotic
674 homolog of parkinsonism-associated protein DJ-1, functions as a covalent chaperone
675 for thiol proteome. J Biol Chem. 2012;287: 5861–5870.

676 35. Roth JR, Lawrence JG, Bobik TA. Cobalamin (coenzyme B12): synthesis and biological
677 significance. Annu Rev Microbiol. 1996;50: 137–181.

678 36. Phan G, Remaut H, Wang T, Allen WJ, Pirker KF, Lebedev A, et al. Crystal structure of
679 the FimD usher bound to its cognate FimC-FimH substrate. Nature. 2011;474: 49–53.

680 37. Typas A, Banzhaf M, Gross CA, Vollmer W. From the regulation of peptidoglycan
681 synthesis to bacterial growth and morphology. Nat Rev Microbiol. ncbi.nlm.nih.gov;
682 2011;10: 123–136.

683 38. Pepper ED, Farrell MJ, Finkel SE. Role of penicillin-binding protein 1b in competitive
684 stationary-phase survival of Escherichia coli. FEMS Microbiol Lett. 2006;263: 61–67.

685 39. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous
686 assay of every Salmonella Typhi gene using one million transposon mutants. Genome
687 Res. 2009;19: 2308–2316.

688 40. Crawford RW, Rosales-Reyes R, Ramírez-Aguilar M de la L, Chapa-Azuela O,
689 Alpuche-Aranda C, Gunn JS. Gallstones play a significant role in Salmonella spp.
690 gallbladder colonization and carriage. Proc Natl Acad Sci U S A. 2010;107: 4353–4358.

691 41. Blondel CJ, Jiménez JC, Contreras I, Santiviago CA. Comparative genomic analysis
692 uncovers 3 novel loci encoding type six secretion systems differentially distributed in
693 Salmonella serotypes. BMC Genomics. 2009;10: 354.

694 42. Blondel CJ, Jiménez JC, Leiva LE, Alvarez SA, Pinto BI, Contreras F, et al. The type VI
695 secretion system encoded in Salmonella pathogenicity island 19 is required for
696 Salmonella enterica serotype Gallinarum survival within infected macrophages. Infect
697 Immun. 2013;81: 1207–1220.

698 43. Mulder DT, Cooper CA, Coombes BK. Type VI secretion system-associated gene
699 clusters contribute to pathogenesis of Salmonella enterica serovar Typhimurium. Infect
700 Immun. Am Soc Microbiol; 2012;80: 1996–2007.

701 44. Kingsley RA, Bäumler AJ. Host adaptation and the emergence of infectious disease:
702 the Salmonella paradigm. Mol Microbiol. 2000;36: 1006–1014.

703 45. Harvey RR, Friedman CR, Crim SM, Judd M, Barrett KA, Tolar B, et al. Epidemiology of
704 Salmonella enterica Serotype Dublin Infections among Humans, United States, 1968–
705 2013. Emerging Infectious Disease journal. 2017;23: 1493.

706 46. Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, Gordon MA, et al. Epidemic
707 multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-
708 Saharan Africa have a distinct genotype. Genome Res. 2009;19: 2279–2287.

47. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, et al. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. Nat Genet. 2012;44: 1215–1221.

48. Feasey NA, Hadfield J, Keddy KH, Dallman TJ, Jacobs J, Deng X, et al. Distinct Salmonella Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. Nat Genet. 2016;48: 1211–1217.

49. Uche IV, MacLennan CA, Saul A. A Systematic Review of the Incidence, Risk Factors and Case Fatality Rates of Invasive Nontyphoidal Salmonella (iNTS) Disease in Africa (1966 to 2014). PLoS Negl Trop Dis. 2017;11: e0005118.

50. Ao TT, Feasey NA, Gordon MA, Heddy KH, Angulo FJ, Crump JA. Global Burden of Invasive Nontyphoidal Salmonella Disease, 2010[1]. Emerging Infectious Disease journal. 2015;21: 941.

51. Okoro CK, Barquist L, Connor TR, Harris SR, Clare S, Stevens MP, et al. Signatures of Adaptation in Human Invasive Salmonella Typhimurium ST313 Populations from Sub-Saharan Africa. PLoS Negl Trop Dis. 2015;9: e0003611.

52. Parsons BN, Humphrey S, Salisbury AM, Mikoleit J, Hinton JCD, Gordon MA, et al. Invasive non-typhoidal Salmonella typhimurium ST313 are not host-restricted and have an invasive phenotype in experimentally infected chickens. PLoS Negl Trop Dis. journals.plos.org; 2013;7: e2487.

53. Ramachandran G, Panda A, Higginson EE, Ateh E, Lipsky MM, Sen S, et al. Virulence of invasive Salmonella Typhimurium ST313 in animal models of infection. PLoS Negl Trop Dis. 2017;11: e0005697.

54. Ramachandran G, Perkins DJ, Schmidlein PJ, Tulapurkar ME, Tennant SM. Invasive Salmonella Typhimurium ST313 with naturally attenuated flagellin elicits reduced inflammation and replicates within macrophages. PLoS Negl Trop Dis. 2015;9: e3394.

55. Carden S, Okoro C, Dougan G, Monack D. Non-typhoidal Salmonella Typhimurium ST313 isolates that cause bacteremia in humans stimulate less inflammasome activation than ST19 isolates associated with gastroenteritis. Pathog Dis. 2015;73. doi:10.1093/femspd/ftu023

56. Singletary LA, Karlinsey JE, Libby SJ, Mooney JP, Lokken KL, Tsolis RM, et al. Loss of Multicellular Behavior in Epidemic African Nontyphoidal Salmonella enterica Serovar Typhimurium ST313 Strain D23580. MBio. 2016;7. doi:10.1128/mBio.02265-15

57. Carden SE, Walker GT, Honeycutt J, Lugo K, Pham T, Jacobson A, et al. Pseudogenization of the Secreted Effector Gene sseI Confers Rapid Systemic Dissemination of S. Typhimurium ST313 within Migratory Dendritic Cells. Cell Host Microbe. 2017;21: 182–194.

58. Ashton PM, Owen SV, Kaindama L, Rowe WPM, Lane C, Larkin L, et al. Salmonella enterica Serovar Typhimurium ST313 Responsible For Gastroenteritis In The UK Are Genetically Distinct From Isolates Causing Bloodstream Infections In Africa [Internet]. bioRxiv. 2017. p. 139576. doi:10.1101/139576

59. Almeida F, Seribelli AA, da Silva P, Medeiros MIC, Dos Prazeres Rodrigues D, Moreira CG, et al. Multilocus sequence typing of Salmonella Typhimurium reveals the presence of the highly invasive ST313 in Brazil. Infect Genet Evol. 2017;51: 41–44.

60. Painter JA, Mølbak K, Sonne-Hansen J, Barrett T, Wells JG, Tauxe RV. Salmonella-based rodenticides and public health. Emerg Infect Dis. 2004;10: 985–987.

61. Pasmans F, Van Immerseel F, Hermans K, Heyndrickx M, Collard J-M, Ducatelle R, et al. Assessment of virulence of pigeon isolates of Salmonella enterica subsp. enterica serovar typhimurium variant copenhagen for humans. J Clin Microbiol. 2004;42: 2000–2002.

62. Lawson B, Hughes LA, Peters T, de Pinna E, John SK, Macgregor SK, et al. Pulsed-field gel electrophoresis supports the presence of host-adapted Salmonella enterica subsp. enterica serovar Typhimurium strains in the British garden bird population. Appl Environ Microbiol. 2011;77: 8139–8144.

63. Mather AE, Lawson B, de Pinna E, Wigley P, Parkhill J, Thomson NR, et al. Genomic Analysis of Salmonella enterica Serovar Typhimurium from Wild Passerines in England and Wales. Appl Environ Microbiol. 2016;82: 6728–6735.

64. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013;14: 827–839.

65. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530: 228–232.

66. Aanensen DM, Feil EJ, Holden MTG, Dordel J, Yeats CA, Fedosejev A, et al. Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive Staphylococcus aureus in Europe. MBio. 2016;7. doi:10.1128/mBio.00444-16

67. Schürch AC, Schaik W. Challenges and opportunities for whole-genome sequencing--based surveillance of antibiotic resistance. Ann N Y Acad Sci. Wiley Online Library; 2017;1388: 108–120.

68. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8: 833–835.

69. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016;7: 12797.

70. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1: 16041.

71. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25: 17–24.

72. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017;18: 41–50.

73. Lupolova N, Dallman TJ, Holden NJ, Gally DL. Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. Microbial Genomics. Microbiology Society; 2017; doi:10.1099/mgen.0.000135

74. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic

793 structure. Ecography . Blackwell Publishing Ltd; 2017;40: 913–929.

794 75. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in
795 the presence of population structure. Nat Commun. 2015;6: 7432.

796 76. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
797 large phylogenies. Bioinformatics. 2014;30: 1312–1313.

798 77. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid
799 large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31: 3691–3693.

800 78. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2: 18–
801 22.

802 79. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees.
803 Chapman and Hall/CRC; 1984.

804 80. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
805 transform. Bioinformatics. 2009;25: 1754–1760.

806 81. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
807 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
808 sequencing data. Genome Res. 2010;20: 1297–1303.

809 82. Li H. A statistical framework for SNP calling, mutation discovery, association mapping
810 and population genetical parameter estimation from sequencing data. Bioinformatics.
811 2011;27: 2987–2993.

812 83. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid
813 phylogenetic analysis of large samples of recombinant bacterial whole genome
814 sequences using Gubbins. Nucleic Acids Res. 2015;43: e15.

815 84. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR.
816 Phandango: an interactive viewer for bacterial population genomics. Bioinformatics.
817 2017; doi:10.1093/bioinformatics/btx610

818