RESEARCH ARTICLE

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

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

* nw17@sanger.ac.uk (NEW); lars.barquist@helmholtz-hiri.de (LB)

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.

Author summary

Researchers are now collecting a wealth of genomic data from bacterial pathogens, and this will continue to grow with the introduction of routine sequencing for disease surveillance. However, our ability to use this data to predict how changes in genome sequence lead to differences in disease is limited. Here, we have used machine learning to detect an

OPEN ACCESS

Citation: Wheeler NE, Gardner PP, Barquist L (2018) Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. PLoS Genet 14(5): e1007333. https://doi.org/10.1371/journal.pgen.1007333

Editor: Xavier Didelot, Imperial College London, UNITED KINGDOM

Received: December 1, 2017

Accepted: March 24, 2018

Published: May 8, 2018

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

Funding: NEW was supported by a PhD scholarship from the University of Canterbury, a Biomolecular Interaction Centre Postdoctoral

Fellowship, and the Wellcome Trust grant 206194. LB was supported in part by a Research Fellowship from the Alexander von Humboldt Stiftung/Foundation. NEW and PPG are supported by a Rutherford Discovery Fellowship administered by the Royal Society of New Zealand, the Bioprotection Research Centre and the National Science Challenge "NZ's Biological Heritage". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

enrichment in functionally significant mutations in genes associated with a shift in pathogenic niche. This approach captures convergence in functional outcomes that does not necessarily result in a convergence in sequence, facilitating the inclusion of rare variants of large effect in an analysis, and allowing for complex interactions between genes. We apply this approach to Salmonella, showing that we can detect changes associated with disease phenotype in emerging lineages associated with the HIV epidemic. This approach should be applicable to other bacterial species with lineages independently adapting to similar niches. We provide open-source implementations of both the predictive model, and the workflow used to build it.

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 movement of bacteria in new pathogenic niches. However, translating whole genome sequence databases into mechanistic and functional insights remains a challenge.

Early expectations were that pathogen evolution would be driven primarily by the acquisition of virulence factors. However, as whole-genome sequencing has become increasingly routine, a decidedly more complex picture has emerged [3,4]. A pattern of bacterial entrance to a new niche followed by adaptation through the loss of antivirulence loci and reduced metabolic flexibility is now recognised as a paradigm of the emergence of important human pathogens from non-pathogenic bacterial species [5–8]. These new niches can be the result of virulence factor acquisition providing access to a previously inaccessible niche in a so-called foothold moment [8], or the emergence of new host niches driven by chronic disease [9–11]. While pathogen and host requirements for infection vary, there is increasing evidence of parallel evolution in bacteria adapting to the same or similar host niche. This is perhaps nowhere more evident than in the species *Salmonella enterica*.

Salmonella enterica strains that cause disease in warm-blooded mammals lie on a spectrum from those that have a broad host range and cause self-limiting gastrointestinal infection, to those that are more restricted in host range, but cause systemic disease and are typically associated with higher mortality [11,12]. Host-restricted, extraintestinal variants of *Salmonella enterica* have evolved independently multiple times from gastrointestinal ancestors [13], and show a greater degree of gene degradation compared to their generalist relatives [14–16]. There are common patterns in the genes that undergo pseudogenization in invasive *Salmonella*, most obviously an extensive network of genes required for anaerobic metabolism in the inflamed host [17,18], a pattern with parallels in other host-adapting enteropathogens [5].

Identifying these signals of parallel evolution has been challenging, relying mainly on manual 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 delta bitscore (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 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).

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.

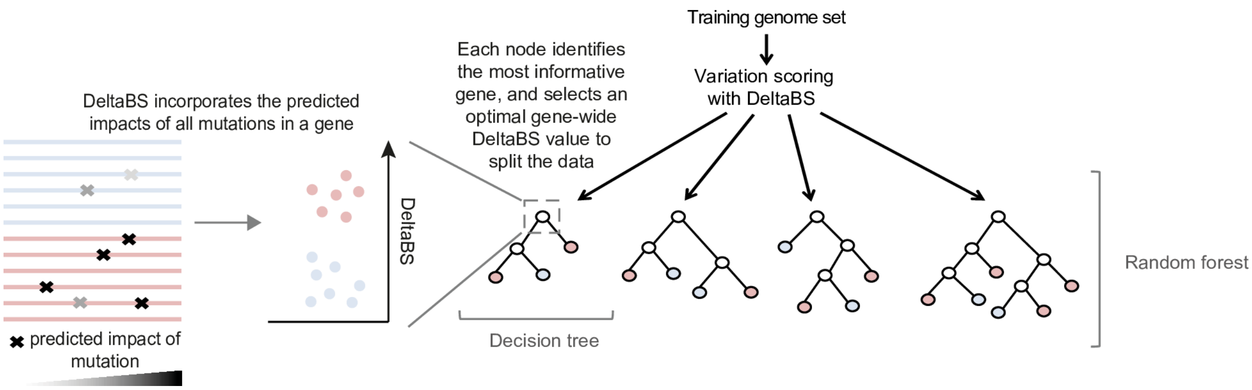
DeltaBS incorporates the predicted impacts of all mutations in a gene
✖ predicted impact of mutation
DeltaBS
Each node identifies the most informative gene, and selects an optimal gene-wide DeltaBS value to split the data
Decision tree
Training genome set
Variation scoring with DeltaBS
Random forest

Fig 1. Overview of the approach employed in this study.

https://doi.org/10.1371/journal.pgen.1007333.g001

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 first round of feature selection, and 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 (S1 Fig). 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 (S1 File).

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

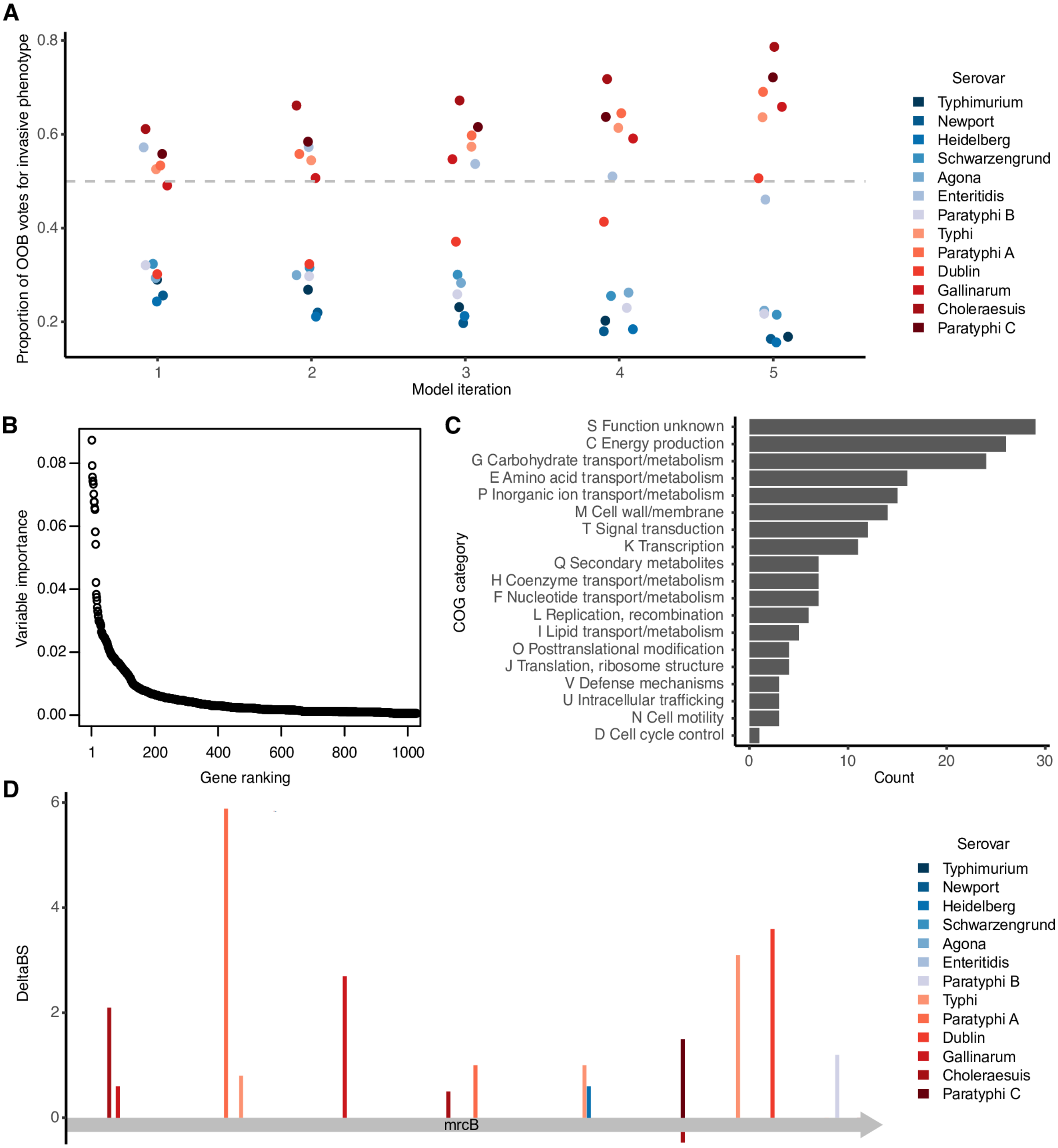

Fig 2. A subset of *Salmonella* genes are strongly indicative of invasive potential. A: Out-of-bag votes for phenotype of each serovar cast by each model. Model 1 is the model built using all predictor variables, then each successive model was built using sparsity pruning from the previous model's predictor 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.

https://doi.org/10.1371/journal.pgen.1007333.g002

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 [17] 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* (S2 Fig). 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, S3 Table). 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 (S3 Fig, [18]). *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.

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

In recent years there have been reports of novel *S*. Typhimurium and *S*. Enteritidis lineages associated with invasive disease in sub-Saharan Africa [46–48] in populations with a high prevalence of immunosuppressive illness such as HIV, malaria, and malnutrition [49]. These lineages contribute to a staggering burden of invasive non-typhoidal salmonella (iNTS) disease, which is responsible for an estimated 3.4 million cases and circa 680,000 deaths annually [50].

Based on epidemiological analysis, high-throughput metabolic screening of selected strains, and analysis of pseudogenes it has been suggested that these lineages may be rapidly adapting to cause invasive disease in the human niche created by widespread immunosuppressive illness [11,46–48,51].

Two iNTS-associated lineages have recently been described within serovar Enteritidis [48], geographically restricted to West Africa and Central/East Africa, respectively. Initial observations have demonstrated that a representative isolate of the Central/East African clade has a reduced capacity to respire in the presence of metabolites requiring cobalamin for their metabolism and has lost the ability to colonize a chick infection model [48], suggesting adaptation to a new host niche. Similarly, two iNTS disease associated lineages have been described in serovar Typhimurium [47], both members of sequence type 313 (ST313), generally referred to as Lineage I and II in the literature. Lineage II appears to have largely replaced Lineage I since 2004, and it has been suggested this is due to Lineage II possessing a gene encoding chloramphenicol resistance [47]. Laboratory characterization of Lineage II strains has shown that they are not host-restricted [52,53], but do appear to possess characteristics suggestive of adaptation to an invasive lifestyle [54–57], though it is important to note that this is a complex trait and not easily quantified.

Given the evidence of adaptation to an invasive niche in these lineages, we asked if genomic 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 and 3B, S4 Table), 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].

To confirm this, we performed an additional comparison of *S*. Typhimurium ST313 isolates (N = 208), to global isolates from other STs, predominantly ST19, associated with gastroenteritis (N = 51) [51,58]. Similarly to iNTS associated *S*. Enteritidis isolates, *S*. Typhimurium ST313 isolates has a higher invasiveness index than isolates from other STs (S4 Fig, S5 Table). Within ST313, Lineage II scored higher than Lineage I, possibly suggesting differential adaptation to the extraintestinal niche. We found that there were in fact more degraded genes unique to Lineage I than Lineage II, but that these genes were assigned less weight in the model, so did not impact score as strongly (S2 Fig & S3 Fig). Interestingly, ST313 has recently been shown not to be entirely restricted to Africa, with isolation reported in Brazil [59] and the UK [58], associated primarily with gastrointestinal disease. We included a collection of UK ST313 strains [58] in our analysis, and found that their invasiveness index tended to be elevated compared to non-ST313 salmonellae, and intermediate between Lineage I and II, suggesting that

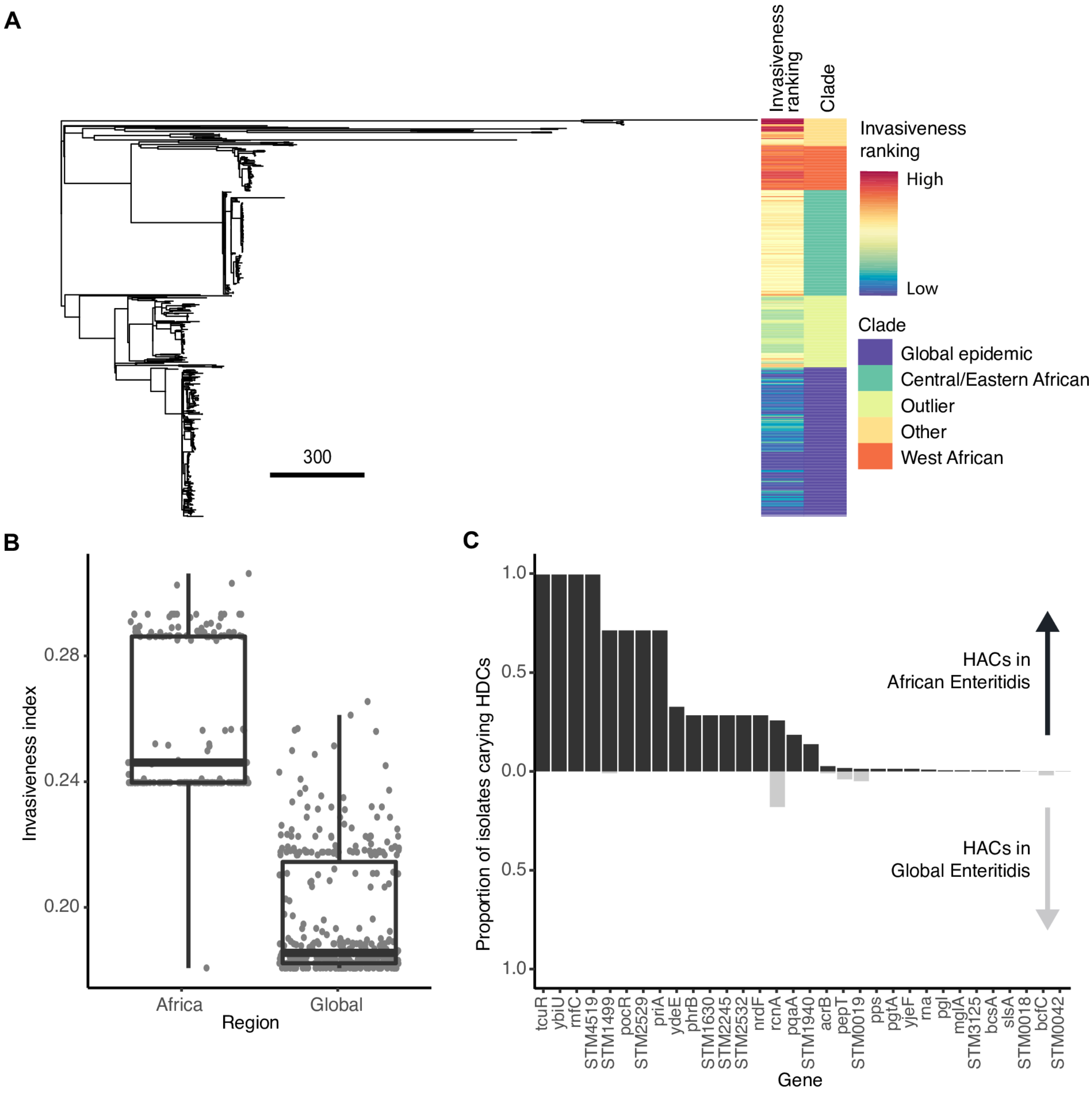

Fig 3. Voting of the model on African iNTS and global gastrointestinal isolates. A: Maximum likelihood phylogeny of all *S*. Enteritidis isolates included in the study, annotated with invasiveness ranking and clade (note: Outlier refers to the distinct sister clade of the global epidemic strains identified 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 are ordered by the amount of degradation observed in African clades. African strains are shown in the positive y-axis in darker grey, global strains are shown in the negative y-axis in lighter grey.

https://doi.org/10.1371/journal.pgen.1007333.g003

this adaptation is not restricted to circulating African strains, as it can be seen in strains collected from other countries as well (S5 Fig). This observation is consistent with the work of Ashton et al. [58], who noted shared pseudogenes and phenotypic traits in UK and African ST313 isolates. This suggests our model is capturing features here associated with the ability to colonize an extraintestinal niche, rather than enter it in healthy individuals.

In addition to the iNTS lineages we investigated, some other strains had unusually high invasiveness indices. Among the top scoring isolates outside of the African *S*. Enteritidis lineages are Ratin strains, a rodenticidal lineage used as commercial rat poison before the 1960s [60]. In *S*. Typhimurium, a clade containing strains DT99, DT56 and U313 also scored highly. These strains appear to be adapted to birds, and DT99 and DT56 have been reported to be highly virulent in pigeons [12,61–63].

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

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

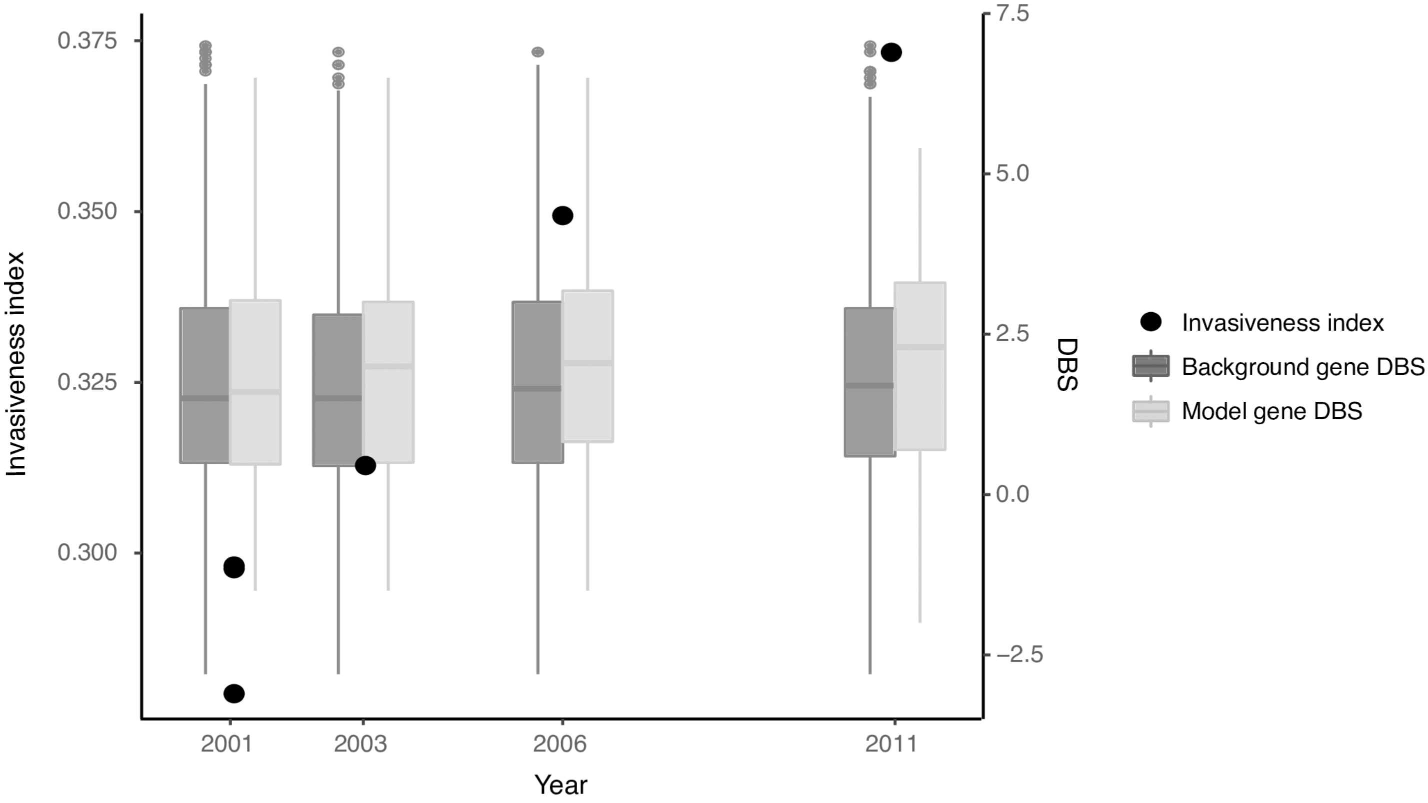

Fig 4. Invasiveness indices and DeltaBS (DBS) values for isolates collected during long term invasive infection of an immunocompromised patient provide evidence for parallel adaptation. Black points show the increase in the invasiveness index over time. Boxplots show a significant shift in DBS distribution 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.

https://doi.org/10.1371/journal.pgen.1007333.g004

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 mutations and indels that impact gene function. Using these scores, we were able to construct a classifier model which is able to separate *Salmonella* serovars adapted to an extraintestinal niche from gastrointestinal strains. Importantly, the random forest classifier that we used produces interpretable lists of genes involved in this adaptation, which agree with results in the literature attained through manual curation of pseudogenes. Additionally, we have shown that this classifier is able to identify nascent signatures of adaptation in strains of *Salmonella* which have been evolving in response to large populations of immunocompromised patients in resource-poor nations.

Other automated approaches to detecting adaptation have been developed which search for SNPs [68] or words [69,70] associated with phenotype. These approaches, termed microbial genome-wide association studies (GWASs), have used techniques adapted from human GWASs, but better cater to methodological issues that arise due to the differences between human and bacterial inheritance patterns. Major differences impacting analyses are stronger linkage disequilibrium (LD) between genetic variants in bacterial genomes, greater population stratification, and often stronger selection for traits [71]. Greater LD and population stratification often result in traits being linked closely with particular lineages, and a large number of variants unique to a lineage being spuriously associated with phenotype. Correction for population stratification allows greater discrimination of true and false positive associations, but results in a substantial loss of power to detect true positives [71], particularly in phenotypes that are highly polygenic and are not under strong positive selection [72]. This can be corrected by increasing the sample size of the study, but increasing sample size can make measurement of complex phenotypes infeasible [23].

A number of machine learning approaches to predicting phenotype from genotypic information have also been recently developed. A notable example is a Support Vector Machine (SVM) based approach to predicting host range in *Salmonella enterica* and *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 S2 Table, 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 serovars, and thus whether the model can detect adaptation to an invasive lifestyle in previously unseen lineages. OOB error rate, stabilised at 10,000 trees, so we chose this as a parameter for optimising the number of genes sampled per node (mtry). mtry values of 1, $p/10$, $p/5$, $p/3$, $p/2$ and p (where p = the number of predictors) were tested, and we found that at mtry = $p/10$, the number of genes that were either not incorporated into trees, or did not improve the homogeneity of daughter nodes when they were incorporated into trees (as measured by mean decrease in Gini index, [79]) stabilised at ~92%. Training the random forest classifier over five iterations took 55 seconds on a laptop computer. In order to assess how well this method would scale, we trained another model on a larger dataset of *S*. Enteritidis strains (N = 677) using the same workflow and site of isolation as a proxy for phenotype, which took 28 minutes.

To improve the performance of the model, we performed five model building and sparsity pruning cycles. For the first cycle, we built a random forest model using all genes that met the inclusion criteria, and performed sparsity pruning by eliminating all variables that had a mean Gini index (variable importance) of zero or lower (meaning the gene was either not included in the model or did not improve model accuracy when it was). Four successive rounds of model building and sparsity pruning involved building a new model with the pruned dataset, then pruning the genes with the lowest 50% of variable importances. The resulting model had 100% out-of-bag classification accuracy. We also tested the accuracy of the full model on a collection of alternative strains related to the training dataset (see S1 Table). Orthologs to the top genes identified by our model were identified using phmmer from the HMMER3.0 package (http://hmmer.org). Additional notes on model building and testing are provided in S1 File.

We tested the top 196 genes for the presence of independent mutations in each serovar by aligning each sequence to the profile HMM representing that protein family. Variation in each 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 multiple 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 truncation. The relationship between invasiveness ranking and phylogeny were visualised using Phandango [84].

Supporting information

S1 Fig. Invasiveness index assigned to validation strains of *Salmonella*.
(TIF)

S2 Fig. Bitscore values from the genes from the genes fimD and fimH combined are better predictors of phenotype than either gene individually.
(TIF)

S3 Fig. Phylogeny of the Salmonella serovars used in this study. The tree was constructed in RAxML using a core gene alignment produced by Roary. Invasive serovars are highlighted in red.
(TIF)

S4 Fig. Invasiveness scores for *S*. Typhimurium ST313 isolates. A: RAxML tree of all ST313 isolates included in the study, annotated with invasiveness ranking and lineage. B: Invasiveness index for all ST313 isolates. C: Proportion of isolates carrying HACs in ST313 compared to other sequence types.
(TIF)

S5 Fig. Genes in S. Typhimurium ST313 above (intact, purple) or below (attenuated, green) bitscore threshold defined by random forest model for detecting gene degradation associated with invasive isolates. Genes for which homology to the reference sequence was not detected (usually due to extreme truncation) are marked in orange.
(TIF)

S1 Table. Accession numbers of *Salmonella* enterica serovars used in this study.
(XLSX)

S2 Table. Top predictor genes.
(XLSX)

S3 Table. Counts of high-impact (DBS in the top quartile) independent mutations in each top predictor gene for each strain.
(XLSX)

S4 Table. Metadata and invasiveness indices for S. Enteritidis iNTS strains.
(XLSX)

S5 Table. Metadata and invasiveness indices for S. Typhimurium iNTS strains.
(XLSX)

S1 File. Additional details on training and testing of the random forest model.
(PDF)

Acknowledgments

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

Author Contributions

Conceptualization: Nicole E. Wheeler, Paul P. Gardner, Lars Barquist.

Data curation: Nicole E. Wheeler.

Formal analysis: Nicole E. Wheeler.

Funding acquisition: Paul P. Gardner.

Investigation: Nicole E. Wheeler.

Methodology: Nicole E. Wheeler, Paul P. Gardner, Lars Barquist.

Software: Nicole E. Wheeler.

Supervision: Paul P. Gardner, Lars Barquist.

Visualization: Nicole E. Wheeler.

Writing – original draft: Nicole E. Wheeler.

Writing – review & editing: Nicole E. Wheeler, Paul P. Gardner, Lars Barquist.

References

1. Frank SA, Schmid-Hempel P. Mechanisms of pathogenesis and the evolution of parasite virulence. J Evol Biol. 2008; 21: 396–404. https://doi.org/10.1111/j.1420-9101.2007.01480.x PMID: 18179516

2. Fauci AS, Morens DM. The perpetual challenge of infectious diseases. N Engl J Med. 2012; 366: 454–461. https://doi.org/10.1056/NEJMra1108296 PMID: 22296079
3. Pallen MJ, Wren BW. Bacterial pathogenomics. Nature. nature.com; 2007; 449: 835–842. https://doi.org/10.1038/nature06248 PMID: 17943120
4. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015; 13: 787–794. https://doi.org/10.1038/nrmicro3565 PMID: 26548914
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. https://doi.org/10.1038/nrmicro.2015.29 PMID: 26876035
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; https://doi.org/10.1038/nrmicro.2016.10 PMID: 26923111
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. https://doi.org/10.1093/bfgp/elt015 PMID: 23814139
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. https://doi.org/10.1073/pnas.1317161111 PMID: 24753568
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. https://doi.org/10.1038/ng.3148 PMID: 25401299
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. https://doi.org/10.1016/S0140-6736(11)61752-2 PMID: 22587967
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. https://doi.org/10.1128/IAI.70.5.2249-2255.2002 PMID: 11953356
13. Bäumler A, Fang FC. Host specificity of bacterial pathogens. Cold Spring Harb Perspect Med. 2013; 3: a010041. https://doi.org/10.1101/cshperspect.a010041 PMID: 24296346
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. https://doi.org/10.1038/35101607 PMID: 11677608
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. https://doi.org/10.1038/ng1470 PMID: 15531882
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. https://doi.org/10.1101/gr.077404.108 PMID: 18583645
17. Nuccio S-P, Bäumler AJ. Comparative Analysis of Salmonella Genomes Identifies a Metabolic Network for Escalating Growth in the Inflamed Gut. MBio. 2014; 5: e00929–14–e00929–14. https://doi.org/10.1128/mBio.00929-14 PMID: 24643865
18. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, et al. Patterns of genome evolution that have accompanied host adaptation in Salmonella. Proc Natl Acad Sci U S A. 2015; 112: 863–868. https://doi.org/10.1073/pnas.1416707112 PMID: 25535353
19. Lerat E, Ochman H. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. 2005; 33: 3125–3132. https://doi.org/10.1093/nar/gki631 PMID: 15933207
20. Kuo C-H, Ochman H. The extinction dynamics of bacterial pseudogenes. PLoS Genet. 2010; 6. https://doi.org/10.1371/journal.pgen.1001050 PMID: 20700439
21. Wheeler NE, Barquist L, Kingsley RA, Gardner PP. A profile-based method for identifying functional divergence of orthologous genes in bacterial genomes. Bioinformatics. 2016; 32: 3566–3574. https://doi.org/10.1093/bioinformatics/btw518 PMID: 27503221
22. Kingsley RA, Kay S, Connor T, Barquist L, Sait L, Holt KE, et al. Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted Salmonella enterica serovar Typhimurium pathovar. MBio. 2013; 4: e00565–13. https://doi.org/10.1128/mBio.00565-13 PMID: 23982073

23. Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SAFT. Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics. 2013; 12: 366–380. https://doi.org/10.1093/bfgp/elt008 PMID: 23625995

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

25. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013; 14: 315–326. https://doi.org/10.1093/bib/bbs034 PMID: 22786785

26. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014; 15: 722–733. https://doi.org/10.1038/nrg3747 PMID: 25200660

27. Bayjanov JR, Molenaar D, Tzeneva V, Siezen RJ, van Hijum SAFT. PhenoLink—a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. BMC Genomics. 2012; 13: 170. https://doi.org/10.1186/1471-2164-13-170 PMID: 22559291

28. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, et al. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014; 24: 839–849. https://doi.org/10.1101/gr.165415.113 PMID: 24717264

29. Alam MT, Petit RA 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. Genome Biol Evol. 2014; 6: 1174–1185. https://doi.org/10.1093/gbe/evu092 PMID: 24787619

30. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016; 44: D286–93. https://doi.org/10.1093/nar/gkv1248 PMID: 26582926

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

32. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 2011; 27: 1986–1994. https://doi.org/10.1093/bioinformatics/btr300 PMID: 21576180

33. Kthiri F, Gautier V, Le H-T, Prère M-F, Fayet O, Malki A, et al. Translational defects in a mutant deficient in YajL, the bacterial homolog of the parkinsonism-associated protein DJ-1. J Bacteriol. 2010; 192: 6302–6306. https://doi.org/10.1128/JB.01077-10 PMID: 20889753

34. Le H-T, Gautier V, Kthiri F, Malki A, Messaoudi N, Mihoub M, et al. YajL, prokaryotic homolog of parkinsonism-associated protein DJ-1, functions as a covalent chaperone for thiol proteome. J Biol Chem. 2012; 287: 5861–5870. https://doi.org/10.1074/jbc.M111.299198 PMID: 22157000

35. Roth JR, Lawrence JG, Bobik TA. Cobalamin (coenzyme B12): synthesis and biological significance. Annu Rev Microbiol. 1996; 50: 137–181. https://doi.org/10.1146/annurev.micro.50.1.137 PMID: 8905078

36. Phan G, Remaut H, Wang T, Allen WJ, Pirker KF, Lebedev A, et al. Crystal structure of the FimD usher bound to its cognate FimC-FimH substrate. Nature. 2011; 474: 49–53. https://doi.org/10.1038/nature10109 PMID: 21637253

37. Typas A, Banzhaf M, Gross CA, Vollmer W. From the regulation of peptidoglycan synthesis to bacterial growth and morphology. Nat Rev Microbiol. ncbi.nlm.nih.gov; 2011; 10: 123–136. https://doi.org/10.1038/nrmicro2677 PMID: 22203377

38. Pepper ED, Farrell MJ, Finkel SE. Role of penicillin-binding protein 1b in competitive stationary-phase survival of Escherichia coli. FEMS Microbiol Lett. 2006; 263: 61–67. https://doi.org/10.1111/j.1574-6968.2006.00418.x PMID: 16958852

39. Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res. 2009; 19: 2308–2316. https://doi.org/10.1101/gr.097097.109 PMID: 19826075

40. Crawford RW, Rosales-Reyes R, Ramírez-Aguilar M de la L, Chapa-Azuela O, Alpuche-Aranda C, Gunn JS. Gallstones play a significant role in Salmonella spp. gallbladder colonization and carriage. Proc Natl Acad Sci U S A. 2010; 107: 4353–4358. https://doi.org/10.1073/pnas.1000862107 PMID: 20176950

41. Blondel CJ, Jiménez JC, Contreras I, Santiviago CA. Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems differentially distributed in Salmonella serotypes. BMC Genomics. 2009; 10: 354. https://doi.org/10.1186/1471-2164-10-354 PMID: 19653904

42. Blondel CJ, Jiménez JC, Leiva LE, Alvarez SA, Pinto BI, Contreras F, et al. The type VI secretion system encoded in Salmonella pathogenicity island 19 is required for Salmonella enterica serotype

Gallinarum survival within infected macrophages. Infect Immun. 2013; 81: 1207–1220. https://doi.org/10.1128/IAI.01165-12 PMID: 23357385

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

44. Kingsley RA, Bäumler AJ. Host adaptation and the emergence of infectious disease: the Salmonella paradigm. Mol Microbiol. 2000; 36: 1006–1014. PMID: 10844686

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

46. Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, Gordon MA, et al. Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. Genome Res. 2009; 19: 2279–2287. https://doi.org/10.1101/gr.091017.109 PMID: 19901036

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. https://doi.org/10.1038/ng.2423 PMID: 23023330

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. https://doi.org/10.1038/ng.3644 PMID: 27548315

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. https://doi.org/10.1371/journal.pntd.0005118 PMID: 28056035

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. https://doi.org/10.1371/journal.pntd.0003611 PMID: 25803844

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. https://doi.org/10.1371/journal.pntd.0002487 PMID: 24130915

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. https://doi.org/10.1371/journal.pntd.0005697 PMID: 28783750

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. https://doi.org/10.1371/journal.pntd.0003394 PMID: 25569606

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. https://doi.org/10.1093/femspd/ftu023 PMID: 25808600

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. https://doi.org/10.1128/mBio.02265-15 PMID: 26933058

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. https://doi.org/10.1016/j.chom.2017.01.009 PMID: 28182950

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. https://doi.org/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. https://doi.org/10.1016/j.meegid.2017.03.009 PMID: 28288927

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. https://doi.org/10.3201/eid1006.030790 PMID: 15207046

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. https://doi.org/10.1128/JCM.42.5.2000-2002.2004 PMID: 15131161

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. https://doi.org/10.1128/AEM.00131-11 PMID: 21948838

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. https://doi.org/10.1128/AEM.01660-16 PMID: 27613688

64. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013; 14: 827–839. https://doi.org/10.1038/nrg3564 PMID: 24166031

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. https://doi.org/10.1038/nature16996 PMID: 26840485

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. https://doi.org/10.1128/mBio.00444-16 PMID: 27150362

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. https://doi.org/10.1111/nyas.13310 PMID: 28134443

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. https://doi.org/10.1038/nmeth.1681 PMID: 21892150

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. https://doi.org/10.1038/ncomms12797 PMID: 27633831

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. https://doi.org/10.1038/nmicrobiol.2016.41 PMID: 27572646

71. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015; 25: 17–24. https://doi.org/10.1016/j.mib.2015.03.002 PMID: 25835153

72. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017; 18: 41–50. https://doi.org/10.1038/nrg.2016.132 PMID: 27840430

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; https://doi.org/10.1099/mgen.0.000135 PMID: 29177093

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

75. Stephan J, Stegle O, Beyer A. A random forest approach to capture genetic effects in the presence of population structure. Nat Commun. 2015; 6: 7432. https://doi.org/10.1038/ncomms8432 PMID: 26109276

76. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

77. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015; 31: 3691–3693. https://doi.org/10.1093/bioinformatics/btv421 PMID: 26198102

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

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

80. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

81. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20: 1297–1303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199

82. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27: 2987–2993. https://doi.org/10.1093/bioinformatics/btr509 PMID: 21903627

83. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015; 43: e15. https://doi.org/10.1093/nar/gku1196 PMID: 25414349

84. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics. 2017; https://doi.org/10.1093/bioinformatics/btx610 PMID: 29028899