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

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

1. The Wellcome Trust 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. Correspondence to: nw17@sanger.ac.uk; lars.barquist@uni-wuerzburg.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 integrate a method for scoring the functional impact of mutations with a random forest classifier, and apply this to the classification of *Salmonella enterica* 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. By training our random forest classifier to discriminate gastrointestinal and invasive serovars of *Salmonella*, using a small and well-characterised training dataset, we are able to additionally discriminate recently emerged *Salmonella* Enteritidis and Typhimurium lineages associated with invasive disease in immunocompromised populations in sub-Saharan Africa. Importantly, our classifier produces interpretable lists of gene variants associated with extraintestinal disease. This approach accurately identifies patterns of gene degradation 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 (Frank and Schmid-Hempel 2008; Fauci and Morens 2012). 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 (Pallen and Wren 2007; Loman and Pallen 2015). 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 (McNally et al. 2016; The et al. 2016; Merhej et al. 2013; Reuter et al. 2014). These new niches can be the result of virulence factor acquisition providing access to a previously inaccessible niche in a so-called foothold moment (Reuter et al. 2014), or the emergence of new host niches driven by chronic disease (Marvig et al. 2015; Klemm et al. 2016; Feasey et al. 2012). 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 (Rabsch et al. 2002; Feasey et al. 2012). Host-restricted, extraintestinal variants of *Salmonella enterica* have evolved independently multiple times from gastrointestinal ancestors (Bäumler and Fang 2013), and show a greater degree of gene degradation compared to their generalist relatives (Parkhill et al. 2001; McClelland et al. 2004; Thomson et al. 2008). 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 gut (Nuccio and Bäumler 2014; Langridge et al. 2015), a pattern with parallels in other host-adapting enteropathogens (McNally et al. 2016).

Identifying these signals of parallel evolution has been challenging, relying mainly on manual annotation and comparison of pseudogenes (Nuccio and Bäumler 2014; Langridge et al. 2015). Detection of pseudogenes in particular relies on ad-hoc criteria to identify large truncations, deletions, or frameshifts (Lerat and Ochman 2005; Kuo and Ochman 2010). 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 (Nuccio and Bäumler 2014). 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 (Wheeler et al. 2016), 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 (Wheeler et al. 2016). 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 (Kingsley et al. 2013), and that a subset of genes in host-adapted strains appear to accumulate large DeltaBS values (Wheeler et al. 2016). Additionally, others have observed similar changes in DeltaBS distributions during adaptation of *Salmonella* to a single immunocompromised host (Klemm et al. 2016). We generally assume that a large DeltaBS value is indicative of a decay in protein function. We cannot rule out that a large DeltaBS may rather indicate a change in protein function, though we expect this to be relatively rare.

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 (Wheeler et al. 2016) 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 (Dutilh et al. 2013; Pappu and Pardalos 2014), and have the potential to detect functional relationships (i.e. epistasis) between genes with a decision tree structure (Touw et al. 2013; Wei et al. 2014). They have been applied successfully in the past to predict microbial phenotype using gene presence/absence data (Bayjanov et al. 2012), and SNPs already known to be associated with phenotype (Laabei et al. 2014; Alam et al. 2014). 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 (Supplemental Table S1), drawing on the extensive curation of orthology relationships performed by Nuccio and Bäumler (2014). 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 (Nuccio and Bäumler 2014), 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 (Huerta-Cepas et al. 2016), designed to capture patterns of sequence variation typically seen in the protein coding genes of Gammaproteobacteria (see Methods).

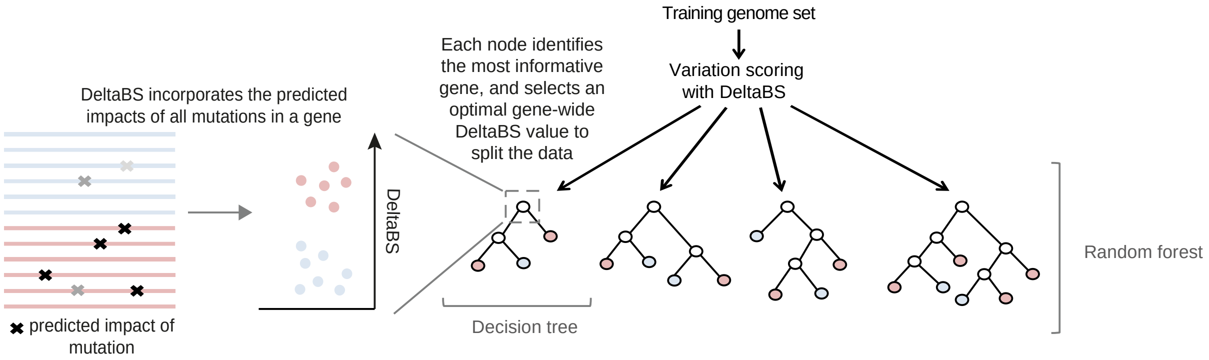
Training genome set
Variation scoring with DeltaBS
Each node identifies the most informative gene, and selects an optimal gene-wide DeltaBS value to split the data
DeltaBS incorporates the predicted impacts of all mutations in a gene
DeltaBS
✖ predicted impact of mutation
Decision tree
Random forest

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 (Breiman 2001), 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). The final model used 196 of the original 6,438 genes for prediction (Supplemental Table S2). This model additionally achieved perfect classification accuracy on an independent set of genomes of the same serovars as our training data (Supplemental Fig S1).

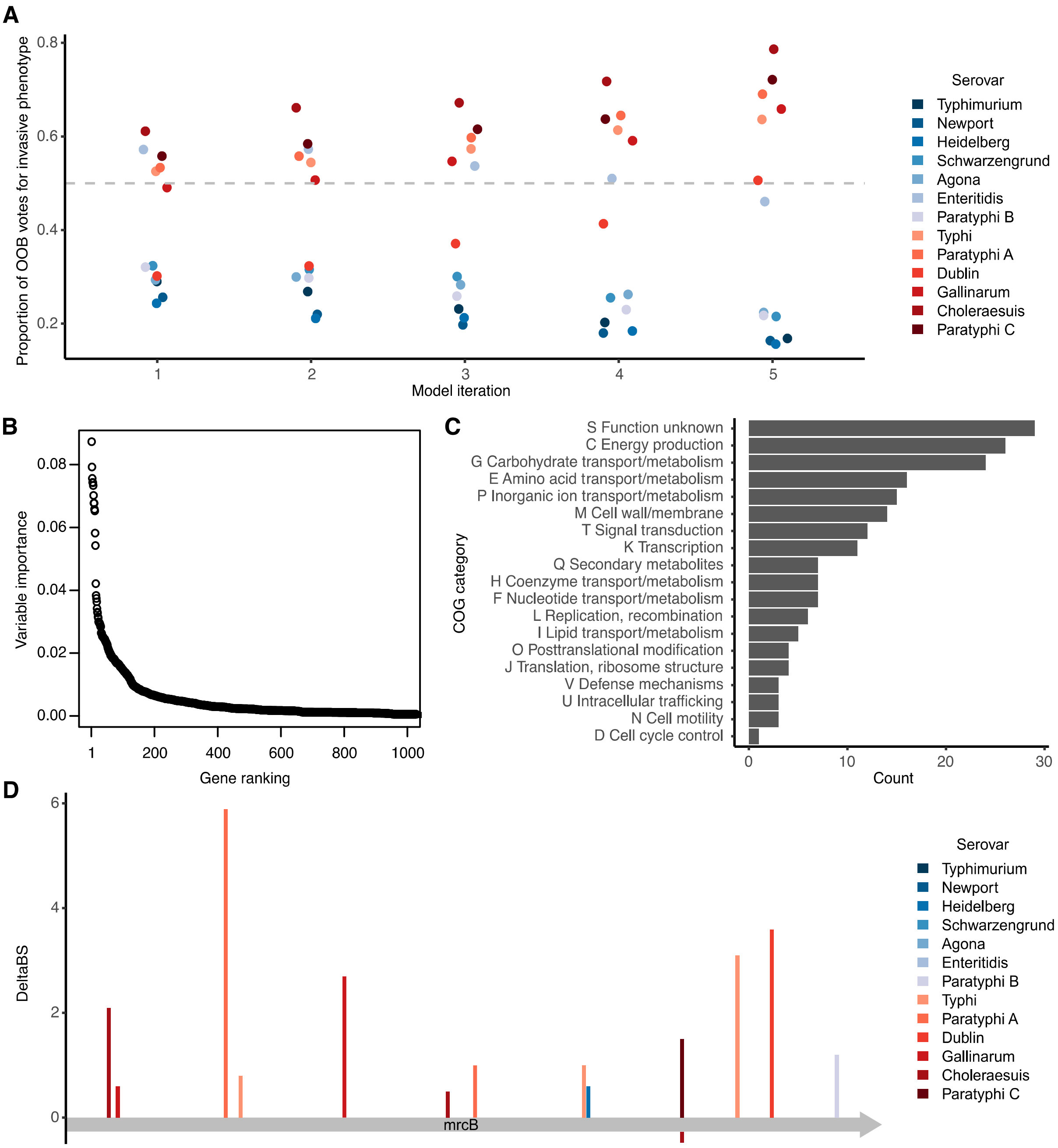

Fig 2 | A subset of *Salmonella* genes are strongly indicative of invasive potential

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 being deleterious to 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 (Kthiri et al. 2010; Le et al. 2012).

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 (Roth et al. 1996). These operons have previously been identified as key degraded pathways in invasive isolates (Thomson et al. 2008; Nuccio and Bäumler 2014; Langridge et al. 2015), 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 (Nuccio and Bäumler 2014; Langridge et al. 2015). 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.

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. 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 (Typas et al. 2011). PBP1b in particular extends and crosslinks peptidoglycan chains during cell division. While PBP1b is not essential, it has been shown to be synthetically lethal with PBP1a and is important for competitive survival of extended stationary phase, osmotic stress (Pepper et al. 2006), and — in *Salmonella* Typhi — growth in the presence of bile (Langridge et al. 2009). Bile is an important environmental challenge for *Salmonella*, particularly for extraintestinal serovars which colonize the gall bladder (Crawford et al. 2010). 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 (Langridge et al. 2009).

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 (Breiman 2001). 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. *S.* Gallinarum was classified much more readily than *S*. Entereitidis 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 (Langridge et al. 2015; Blondel et al. 2009). These loci have been implicated in host-adaptation and survival during extraintestinal infection (Blondel et al. 2013; Mulder et al. 2012), 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 (Langridge et al. 2015). 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 (Nuccio and Bäumler 2014; Langridge et al. 2015), suggesting a lower degree of host adaptation than other invasive isolates. Indeed, *S*. Dublin is more promiscuous in its host range, primarily infecting cattle (Kingsley and Bäumler 2000) while still causing sporadic human disease (Harvey et al. 2017). 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 (Kingsley et al. 2009; Okoro et al. 2012; Feasey et al. 2016) in populations with a high prevalence of immunosuppressive illness such as HIV, malaria, and malnutrition (Uche et al. 2017). 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 (Ao et al. 2015). 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 (Kingsley et al. 2009; Feasey et al. 2012; Okoro et al. 2012, 2015; Feasey et al. 2016).

Two iNTS-associated lineages have recently been described within serovar Enteritidis (Feasey et al. 2016), 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 (Feasey et al. 2016), suggesting adaptation to a new host niche. Similarly, two iNTS disease associated lineages have been described in serovar Typhimurium (Okoro et al. 2012), 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 (Okoro et al. 2012). Laboratory characterization of Lineage II strains has shown that they are not host-restricted (Parsons et al. 2013; Ramachandran et al. 2017), but do appear to possess characteristics suggestive of adaptation to an invasive lifestyle (Ramachandran et al. 2015; Carden et al. 2015; Singletary et al. 2016; Carden et al. 2017).

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) (Feasey et al. 2016).

Our model gave iNTS-associated *S*. Enteritidis strains a higher invasiveness index than the globally distributed isolates (Fig 3A,B, Supplemental Table S3), 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. (2016) 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

(2014) 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 (Feasey et al. 2016).

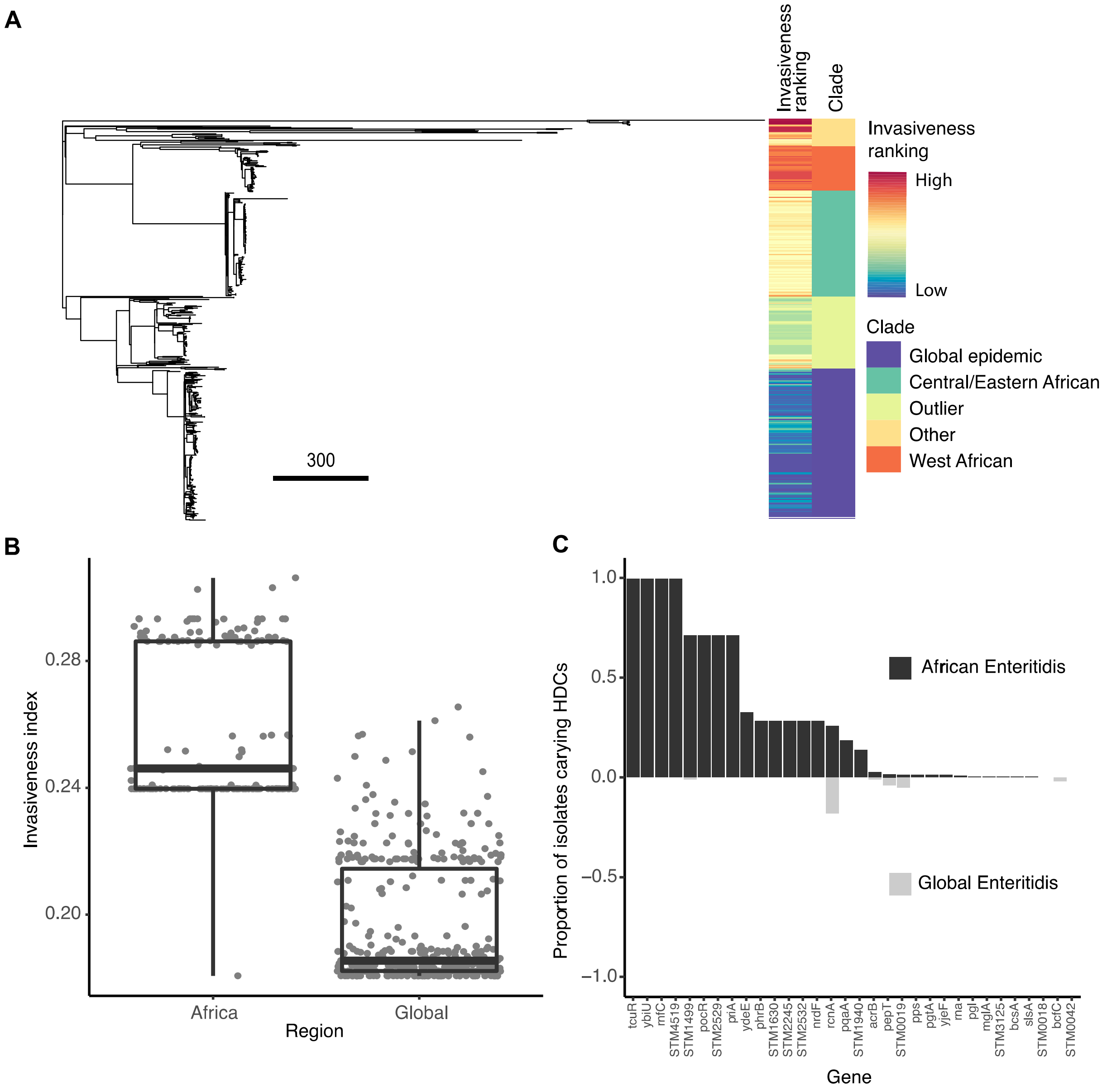

Fig 3 | Voting of the model on African iNTS and global gastrointestinal isolates

A: Maximum likelihood phylogeny of all *S*. Enteritidis isolates included in the study, annotated with invasiveness ranking and clade.

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 disrupted coding sequence (HDC, as 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.

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) (Okoro et al. 2015; Ashton et al. 2017). Similarly to iNTS associated *S*. Enteritidis isolates, *S*. Typhimurium ST313 isolates has a higher invasiveness index than isolates from other STs (Supplemental Fig S2, Supplemental Table S4). 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 (Supplemental Fig S2 & S3). Interestingly, ST313 has recently been shown not to be entirely restricted to Africa, with isolation reported in Brazil (Almeida et al. 2017) and the UK (Ashton et al. 2017). We included a collection of UK ST313 strains (Ashton et al. 2017) 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 some of the changes we are detecting are ancestral to ST313 as a whole (Supplemental Fig S3).

To test whether we could detect a recent case of accelerated adaptation over the course of a single infection, 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 (Klemm et al. 2016). We found a significant positive correlation between invasiveness index and duration of carriage (r=0.96, n=6, *P*=0.002, Supplemental Fig S4).

Discussion

Parallel evolution appears to be common in niche adaptation, which allows us to identify genes that are important for survival in different environments. 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 lead to independent losses of the *ttr*, *cbi* and *pdu* genes, important for anaerobic metabolism during intestinal infection (McNally et al. 2016). Within genuses, parallelism has been observed when distinct lineages acquire similar virulence factors leading to similar phenotypes, as with *Yersinia pseudotuberculosis* and *enterocolitica* (Reuter et al. 2014), or the repeated emergence of the *Shigella* phenotype within the *Escherichia* (The et al. 2016). 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 (Marvig et al. 2015), or a hypermutator strain of *Salmonella* adapting to an immunocompromised host (Klemm et al. 2016). With pathogen sequencing for disease surveillance becoming increasingly routine (Quick et al. 2016; Aanensen et al. 2016; Schürch and Schaik 2017), 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 (Wheeler et al. 2016) 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 (Lippert et al. 2011) or words (Lees et al. 2016; Earle et al. 2016) 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 (Chen and Shapiro 2015). 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 (Chen and Shapiro 2015), particularly in phenotypes that are highly polygenic and are not under strong positive selection (Power et al. 2017). This can be corrected by increasing the sample size of the study, but increasing sample size can make measurement of complex phenotypes infeasible (Dutilh et al. 2013).

DeltaBS 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 (Lippert et al. 2011), or used in a hybrid LMM-random forest based approach (Stephan et al. 2015) to preserve the ability of the metric to detect epistasis between genes (Wei et al. 2014).

Methods

Genome data and identification of orthologs

Genomes for 13 *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 (2014). Ortholog calls were also taken from the Supplementary Material of Nuccio and Bäumler (2014).

Measuring the divergence of genes from predicted sequence constraints

Profile hidden Markov models (HMMs) for Gammaproteobacterial proteins were retrieved from the eggNOG database (Huerta-Cepas et al. 2016). We chose this source of HMMs because it is publically 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 (Wheeler et al. 2016). 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” (Liaw and Wiener 2002) was used to build random forest classifiers using a variety of parameters to assess which were best for accuracy. Prediction accuracy, as measured by out-of-bag (OOB) error rate, stabilised at 1000 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, (Breiman et al. 1984)) stabilised at ~92%.

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 Table S1). Orthologs to the top genes identified by our model were identified using phmmer from the HMMER3.0 package (http://hmmer.org).

Invasive non-typhoidal Salmonella analysis

Read data from Feasey et al. (2016) and Klemm et al. (2016) was mapped to the reference genome *S*. Enteritidis P125109. Reads from Okoro et al. (2015) and Ashton et al. (2017) 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 (Li and Durbin 2009) and regions near indels were realigned using GATK (McKenna et al. 2010). 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 (Li 2011), 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 (Croucher et al. 2015) to exclude recombination events, and RAxML v8.2.8 (Stamatakis 2014) to build maximum likelihood trees using a GTR + Gamma model.

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 (Liaw and Wiener 2002). 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 (Hadfield et al. 2017).

Data access

All genome sequence data are publically 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 publically available at github.com/UCanCompBio/invasive_salmonella.

Funding information

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”.

Acknowledgements

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 Nick Feasey, Nick Thomson and John Crump for their helpful feedback.

References

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

Alam MT, Petit RA 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, Satola SW, Read TD. 2014. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome Biol Evol* **6**: 1174–1185.

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

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

Ashton PM, Owen SV, Kaindama L, Rowe WPM, Lane C, Larkin L, Nair S, Jenkins C, de Pinna E, Feasey N, et al. 2017. Salmonella enterica Serovar Typhimurium ST313 Responsible For Gastroenteritis In The UK Are Genetically Distinct From Isolates Causing Bloodstream Infections In Africa. *bioRxiv* 139576. doi: 10.1101/139576.

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

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

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

Blondel CJ, Jiménez JC, Leiva LE, Alvarez SA, Pinto BI, Contreras F, Pezoa D, Santiviago CA, Contreras I. 2013. The type VI secretion system encoded in Salmonella pathogenicity island 19 is required for Salmonella enterica serotype Gallinarum survival within infected macrophages. *Infect Immun* **81**: 1207–1220.

Breiman L. 2001. Random Forests. *Mach Learn* **45**: 5–32.

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

Carden SE, Walker GT, Honeycutt J, Lugo K, Pham T, Jacobson A, Bouley D, Idoyaga J, Tsolis RM, Monack D. 2017. Pseudogenization of the Secreted Effector Gene sseI Confers Rapid Systemic Dissemination of S. Typhimurium ST313 within Migratory Dendritic Cells. *Cell Host Microbe* **21**:

182–194.

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

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

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

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**: e15.

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

Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, et al. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol* **1**: 16041.

Fauci AS, Morens DM. 2012. The perpetual challenge of infectious diseases. *N Engl J Med* **366**: 454–461.

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

Feasey NA, Hadfield J, Keddy KH, Dallman TJ, Jacobs J, Deng X, Wigley P, Barquist Barquist L, Langridge GC, Feltwell T, et al. 2016. Distinct Salmonella Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. *Nat Genet* **48**: 1211–1217.

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

Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. 2017. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*. doi:

10.1093/bioinformatics/btx610.

Harvey RR, Friedman CR, Crim SM, Judd M, Barrett KA, Tolar B, Folster JP, Griffin PM, Brown AC. 2017. Epidemiology of Salmonella enterica Serotype Dublin Infections among Humans, United States, 1968–2013. *Emerging Infectious Disease journal* **23**: 1493.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**: D286–93.

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

Kingsley RA, Kay S, Connor T, Barquist L, Sait L, Holt KE, Sivaraman K, Wileman T, Goulding D, Clare S, et al. 2013. Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted Salmonella enterica serovar Typhimurium pathovar. *MBio* **4**: e00565–13.

Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, Gordon MA, Harris D, Clarke L, Whitehead S, Sangal V, et al. 2009. Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res* **19**: 2279–2287.

Klemm EJ, Gkrania-Klotsas E, Hadfield J, Forbester JL, Harris SR, Hale C, Heath JN, Wileman T, Clare S, Kane L, et al. 2016. Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. *Nat Microbiol* **1**: 15023.

Kthiri F, Gautier V, Le H-T, Prère M-F, Fayet O, Malki A, Landoulsi A, Richarme G. 2010. Translational defects in a mutant deficient in YajL, the bacterial homolog of the parkinsonism-associated protein DJ-1. *J Bacteriol* **192**: 6302–6306.

Kuo C-H, Ochman H. 2010. The extinction dynamics of bacterial pseudogenes. *PLoS Genet* **6**. http://dx.doi.org/10.1371/journal.pgen.1001050.

Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, et al. 2014. Predicting the virulence of MRSA from its genome sequence. *Genome Res* **24**: 839–849.

Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HMB, Barquist

L, Stedman A, Humphrey T, et al. 2015. Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proc Natl Acad Sci U S A* **112**: 863–868.

Langridge GC, Phan M-D, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, et al. 2009. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* **19**: 2308–2316.

Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, Marttinen P, Davies MR, Steer AC, Tong SYC, et al. 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun* **7**: 12797.

Le H-T, Gautier V, Kthiri F, Malki A, Messaoudi N, Mihoub M, Landoulsi A, An YJ, Cha S-S, Richarme G. 2012. YajL, prokaryotic homolog of parkinsonism-associated protein DJ-1, functions as a covalent chaperone for thiol proteome. *J Biol Chem* **287**: 5861–5870.

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

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

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

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

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

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

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

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

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

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

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

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

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

Okoro CK, Barquist L, Connor TR, Harris SR, Clare S, Stevens MP, Arends MJ, Hale C, Kane L, Pickard DJ, et al. 2015. Signatures of Adaptation in Human Invasive Salmonella Typhimurium ST313 Populations from Sub-Saharan Africa. *PLoS Negl Trop Dis* **9**: e0003611.

Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E, et al. 2012. Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nat Genet* **44**: 1215–1221.

Pallen MJ, Wren BW. 2007. Bacterial pathogenomics. *Nature* **449**: 835–842.

Pappu V, Pardalos PM. 2014. High-Dimensional Data Classification. In *Clusters, Orders, and Trees: Methods and Applications* (eds. F. Aleskerov, B. Goldengorin, and P.M. Pardalos), *Springer Optimization and Its Applications*, pp. 119–150, Springer New York.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MT, et al. 2001. Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature* **413**: 848–852.

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

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

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

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**: 228–232.

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

Ramachandran G, Panda A, Higginson EE, Ateh E, Lipsky MM, Sen S, Matson CA, Permala-Booth J, DeTolla LJ, Tennant SM. 2017. Virulence of invasive Salmonella Typhimurium ST313 in animal models of infection. *PLoS Negl Trop Dis* **11**: e0005697.

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

Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, Fookes M, Hall ME, Petty NK, Fuchs TM, et al. 2014. Parallel independent evolution of pathogenicity within the genus Yersinia. *Proc Natl Acad Sci U S A* **111**: 6768–6773.

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

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

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

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

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

The HC, Thanh DP, Holt KE, Thomson NR, Baker S. 2016. The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nat Rev Microbiol*. http://dx.doi.org/10.1038/nrmicro.2016.10.

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

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SAFT. 2013. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform* **14**: 315–326.

Typas A, Banzhaf M, Gross CA, Vollmer W. 2011. From the regulation of peptidoglycan synthesis to bacterial growth and morphology. *Nat Rev Microbiol* **10**: 123–136.

Uche IV, MacLennan CA, Saul A. 2017. 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* **11**: e0005118.

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

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